2. Point Estimation

Data $x_1, x_2, \ldots, x_n \quad \rightarrow$ inference about parameter θ, assume to be realisations of random variables $X_1, X_2, \ldots, X_n$ from $f(\mathbf{x}, \theta)$

Denote the expectation with respect to $f(\mathbf{x}, \theta)$ by E_θ, and the variance by Var_θ.

Estimate θ by a function $t(x_1, \ldots, x_n)$ of the data (a *point estimate*); $T = t(X_1, \ldots, X_n) = t(\mathbf{X})$ is called an *estimator* (random)

For example, a sufficient statistic is an estimator.

Properties of estimators

T is *unbiased* for θ if $E_\theta(T) = \theta$ for all θ; otherwise T is *biased*.

The *bias* of T is

$$Bias(T) = Bias_\theta(T) = E_\theta(T) - \theta.$$

Example: Sample mean, sample variance.

$X_1, \dots, X_n$ i.i.d. with unknown mean μ; unknown variance σ^2

Estimate μ by

$$T = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then

$$E_{\mu,\sigma^2}(T) = \frac{1}{n} \sum_{i=1}^{n} \mu = \mu$$

so unbiased.

Recall that

$$Var_{\mu,\sigma^2}(T) = Var_{\mu,\sigma^2}(\overline{X}) = E_{\mu,\sigma^2}\{(\overline{X} - \mu)^2)\} = \frac{\sigma^2}{n}.$$

Estimate σ^2 by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Then

$$\begin{aligned}
&E_{\mu,\sigma^2}(S^2) \\
&= \frac{1}{n-1}\sum_{i=1}^{n} E_{\mu,\sigma^2}\{(X_i - \mu + \mu - \overline{X})^2\} \\
&= \frac{1}{n-1}\sum_{i=1}^{n} \left\{E_{\mu,\sigma^2}\{(X_i-\mu)^2\} + 2E_{\mu,\sigma^2}(X_i-\mu)(\mu-\overline{X})\right. \\
&\quad \left. + E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\}\right\} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\sigma^2 - 2\frac{n}{n-1}E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\} + \frac{n}{n-1}E_{\mu,\sigma^2}\{(\overline{X}-\mu)^2\} \\
&= \sigma^2\left(\frac{n}{n-1} - \frac{2}{n-1} + \frac{1}{n-1}\right) = \sigma^2,
\end{aligned}$$

so unbiased. *Note*: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is **not** unbiased.

Another criterion: small *mean square error* (MSE)

$$MSE(T) = MSE_\theta(T) = E_\theta\{(T-\theta)^2\} = Var_\theta(T) + (Bias_\theta(T))^2$$

Note: $MSE(T)$ is a function of θ and in general therefore cannot be zero everywhere.

Example:

$\hat{\sigma}^2$ has smaller MSE than S^2 (see *Casella and Berger, p.304*) but is biased.

If one has two estimators at hand, one being slightly biased but having a smaller MSE than the second one, which is, say, unbiased, then one may well prefer the slightly biased estimator. Exception: If the estimate is to be combined linearly with other estimates from independent data.

The efficiency of an estimator is defined as

$$Efficiency_\theta(T) = \frac{\mathrm{Var}_\theta(T_0)}{\mathrm{Var}_\theta(T)},$$

where T_0 has minimum possible variance.

Cramér-Rao Inequality

Under regularity conditions on $f(\mathbf{x}, \theta)$, it holds that for any unbiased

T,

$$\mathrm{Var}_\theta(T) \geq (I(\theta))^{-1}$$

(*Cramér-Rao Inequality, Cramér-Rao lower bound*) where

$$I(\theta) := I_n(\theta) = E_\theta\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right]$$

is the *expected Fisher information* of the sample.

Calculation:

$$
\begin{aligned}
I_n(\theta) &= E_\theta\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right] \\
&= \int f(\mathbf{x}, \theta)\left[\left(\frac{\partial \log f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] d\mathbf{x} \\
&= \int f(\mathbf{x}, \theta)\left[\frac{1}{f(\mathbf{x}, \theta)}\left(\frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] d\mathbf{x} \\
&= \int \frac{1}{f(\mathbf{x}, \theta)}\left[\left(\frac{\partial f(\mathbf{x}, \theta)}{\partial \theta}\right)^2\right] d\mathbf{x}.
\end{aligned}
$$

Consequences:

For any unbiased estimator T,

$$Efficiency_\theta(T) = \frac{1}{I(\theta)\mathrm{Var}_\theta(T)}.$$

Assume that T is unbiased. T is called *efficient* (or a *minimum variance unbiased estimator*) if it has the minimum possible variance. An unbiased estimator T is efficient if $\mathrm{Var}_\theta(T) = (I(\theta))^{-1}$.

Often: $T = T(X_1, \ldots, X_n)$ efficient at $n \to \infty$: *asymptotically efficient*

Regularity: conditions on the partial derivatives of $f(\mathbf{x}, \theta)$ with respect to θ; domain may not depend on θ; for example $\mathcal{U}[0, \theta]$ violates the regularity conditions

Under more regularity: the first three partial derivatives of $f(\mathbf{x}, \theta)$ with respect to θ are integrable with respect to x; domain may not depend on θ; then

$$I_n(\theta) = E_\theta \left[-\frac{\partial^2 \ell(\theta, \mathbf{X})}{\partial \theta^2} \right]$$

Notation: We shall often omit the subscript in $I_n(\theta)$, when it is clear whether we refer to a sample of size 1, or to a sample of size n.

For a random sample,

$$I_n(\theta) = n I_1(\theta).$$

Example: Normal distribution, known variance

$\mathcal{N}(\mu, \sigma^2)$, where σ^2 known, $\theta = \mu$

$$
\begin{aligned}
\ell(\theta) &= -\frac{n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \\
\frac{\partial \ell}{\partial \theta} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = \frac{n}{\sigma^2}(\overline{x} - \mu)
\end{aligned}
$$

and

$$
\begin{aligned}
I(\theta) &= E_\theta\left[\left(\frac{\partial \ell(\theta, \mathbf{X})}{\partial \theta}\right)^2\right] \\
&= \frac{n^2}{\sigma^4}E_\theta(\overline{X} - \mu)^2 = \frac{n}{\sigma^2}
\end{aligned}
$$

Note $\mathrm{Var}_\theta(\overline{X}) = \frac{\sigma^2}{n}$, so $\overline{X}$ is an efficient estimator for μ.

NB:

$$
\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{n}{\sigma^2}.
$$

In future we shall often omit the subscript θ in the expectation and in the variance.

Maximum Likelihood Estimation

θ may be a vector

A *maximum likelihood estimate*, denoted $\hat{\theta}(\mathbf{x})$, is a value of θ at which the likelihood $L(\theta, \mathbf{x})$ is maximal. The estimator $\hat{\theta}(\mathbf{X})$ is called *MLE* (also, $\hat{\theta}(\mathbf{x})$ is sometimes called mle).

An mle is a parameter value at which the observed sample is most likely.

Often: easier to maximise log likelihood: **if** derivatives exist, then set first (partial) derivative(s) with respect to θ to zero, check that second (partial) derivative(s) with respect to θ less than zero.

An mle is a function of a sufficient statistic:

$$L(\theta, \mathbf{x}) = f(\mathbf{x}, \theta) = g(t(\mathbf{x}), \theta)h(\mathbf{x})$$

by the factorisation theorem, and maximizing in θ depends on $\mathbf{x}$ only through $t(\mathbf{x})$.

An mle is usually efficient as $n \to \infty$.

Invariance property: An mle of a function $\phi(\theta)$ is $\phi(\hat{\theta})$ (Casella + Berger p.294). That is, if we define the likelihood induced by ϕ as

$$L^*(\lambda, x) = \sup_{\theta:\phi(\theta)=\lambda} L(\theta, x),$$

then one can calculate that for $\hat{\lambda} = \phi(\hat{\theta})$,

$$L^*(\hat{\lambda}, x) = L(\hat{\theta}, x).$$

Examples: Uniforms, normal

1. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[0,\theta]$:

$$L(\theta) = \theta^{-n} \mathbf{1}_{[x_{(n)},\infty)}(\theta),$$

where $x_{(n)} = \max_{1\le i\le n} x_i$; so $\hat{\theta} = X_{(n)}$

2. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{U}[\theta-\frac{1}{2}, \theta+\frac{1}{2}]$, then any $\theta \in [x_{(n)}-\frac{1}{2}, x_{(1)}+\frac{1}{2}]$

maximises the likelihood (*Exercise*)

3. $X_1, \ldots, X_n$ i.i.d. $\sim \mathcal{N}(\mu, \sigma^2)$, then (*Exercise*) $\hat{\mu} = \overline{X}, \hat{\sigma}^2 =$

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

so $\hat{\sigma}^2$ is biased, but $Bias(\hat{\sigma}^2) \to 0$ as $n \to \infty$.

Iterative computation of MLEs

Sometimes the likelihood equations are difficult to solve. Suppose $\hat{\theta}^{(1)}$ is an initial approximation for $\hat{\theta}$. Use Taylor:

$$0 = \ell'(\hat{\theta}) \approx \ell'(\hat{\theta}^{(1)}) + (\hat{\theta} - \hat{\theta}^{(1)})\ell''(\hat{\theta}^{(1)})$$

so

$$\hat{\theta} \approx \hat{\theta}^{(1)} - \frac{\ell'(\hat{\theta}^{(1)})}{\ell''(\hat{\theta}^{(1)})}$$

Iterate (*Newton-Raphson method*)

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - (\ell''(\hat{\theta}^{(k)}))^{-1}\ell'(\hat{\theta}^{(k)}), \quad k = 2, 3, \ldots$$

until $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}| < \epsilon$ for some small ϵ

As $E\left\{-\ell''(\hat{\theta}^{(1)})\right\} = I(\hat{\theta}^{(1)})$ we could instead iterate

$$\hat{\theta}^{(k+1)} \;=\; \hat{\theta}^{(k)} + I^{-1}(\hat{\theta}^{(k)})\ell'(\hat{\theta}^{(k)}), \quad k = 2, 3, \ldots$$

until $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}| < \epsilon$ for some small ϵ.

This is *Fisher's modification of the Newton-Raphson method.*

Repeat with different starting values to reduce the risk of finding just a local maximum.

Example: $Binomial(n, \theta)$. Observe x

$$\ell(\theta) = x\ln(\theta) + (n-x)\ln(1-\theta) + \log\binom{n}{x}$$

$$\ell'(\theta) = \frac{x}{\theta} - \frac{n-x}{1-\theta} = \frac{x-n\theta}{\theta(1-\theta)}$$

$$\ell''(\theta) = -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}$$

$$I(\theta) = \frac{n}{\theta(1-\theta)}$$

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

guess $\hat{\theta}^{(0)} = 0.55$

Newton-Raphson:

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

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

Fisher scoring:

$$I^{-1}(\theta)\ell'(\theta) \;=\; \frac{x - n\theta}{n} = \frac{x}{n} - \theta$$

and so

$$\theta + I^{-1}(\theta)\ell'(\theta) \;=\; \frac{x}{n}$$

for all θ.

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

Profile likelihood

Often $\theta = (\psi, \lambda)$ where ψ contains the parameters of interest and λ contains the other unknown parameters: *nuisance parameters*.

Let $\hat{\lambda}_\psi$ be the MLE for λ for a given value of ψ. Then the *profile likelihood* for ψ is

$$L_P(\psi) = L(\psi, \hat{\lambda}_\psi).$$

(in $L(\psi, \lambda)$ replace λ by $\hat{\lambda}_\psi$); the *profile log-likelihood* is $\ell_P(\psi) = \log[L_P(\psi)]$.

For point estimation, maximizing $L_P(\psi)$ with respect to ψ gives the same estimator $\hat{\psi}$ as maximizing $L(\psi, \lambda)$ with respect to both ψ and λ (but possibly different variances)

Example: Normal distribution.

$X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$ with μ and σ^2 unknown. Given μ, $\hat{\sigma}^2_\mu = (1/n) \sum (x_i - \mu)^2$, and given σ^2, $\hat{\mu}_{\sigma^2} = \overline{x}$. Hence the profile likelihood for μ is

$$\begin{aligned} L_P(\mu) &= (2\pi\hat{\sigma}^2_\mu)^{-n/2} \exp\left\{ -\frac{1}{2\hat{\sigma}^2_\mu} \sum_{i=1}^{n} (x_i - \mu)^2 \right\} \\ &= \left[\frac{2\pi e}{n} \sum (x_i - \mu)^2 \right]^{-n/2}, \end{aligned}$$

which gives $\hat{\mu} = \overline{x}$; and the profile likelihood for σ^2 is

$$L_P(\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum (x_i - \overline{x})^2 \right\},$$

gives (Exercise)

$$\hat{\sigma}^2_\mu = ??$$

Method of Moments (M.O.M)

Idea: match population moments to sample moments in order to obtain estimators

Suppose $X_1, \ldots, X_n$ i.i.d. $\sim f(x; \theta_1, \ldots, \theta_p)$ Denote by $\mu_k = E(X^k)$ the k^{th} moment and by $M_k = \frac{1}{n}\sum(X_i)^k$ the k^{th} sample moment. In general, $\mu_k = \mu_k(\theta_1, \ldots, \theta_p)$.

Equate μ_k to M_k for $k = 1, 2, \ldots,$ until there are sufficient equations to solve for $\theta_1, \ldots, \theta_p$ (usually p equations for the p unknowns)

The solutions $\tilde{\theta}_1, \ldots, \tilde{\theta}_p$ are the *method of moments estimators*

Often not as efficient as MLEs, but may be easier to calculate

Could be used as initial estimates in an iterative calculation of MLEs

Example: Normal distribution.

$X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$; μ and σ^2 unknown

$$\mu_1 = \mu;\ M_1 = \overline{X} \text{ so } \tilde{\mu} = \overline{X}$$

and

$$\mu_2 = \sigma^2 + \mu^2;\ M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

so

$$\tilde{\sigma}^2 = M_2 - M_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(not unbiased)

Example: Gamma distribution. $X_1, \dots, X_n$ i.i.d. $\Gamma(\psi, \lambda)$;

$$f(x; \psi, \lambda) = \frac{1}{\Gamma(\psi)} \lambda^{\psi} x^{\psi - 1} e^{-\lambda x} \qquad \text{for } x \geq 0.$$

Then $\mu_1 = EX = \psi/\lambda$ and

$$\mu_2 = EX^2 = \psi/\lambda^2 + (\psi/\lambda)^2$$

Solve

$$M_1 = \psi/\lambda, \qquad M_2 = \psi/\lambda^2 + (\psi/\lambda)^2$$

for ψ and λ; gives

$$\begin{aligned} \tilde{\psi} &= \overline{X}^2 / [n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2], \\ \tilde{\lambda} &= \overline{X} / [n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2]. \end{aligned}$$

Bias and variance approximations: the delta method

Sometimes T is a function of one or more averages whose means and variances can be calculated exactly

$\rightarrow$ simple approximations for mean and variance of T:

Suppose $T = g(S)$ where $ES = \beta$ and $\operatorname{Var} S = V$. Taylor

$$T = g(S) \approx g(\beta) + (S - \beta)g'(\beta).$$

Taking the mean and variance of the r.h.s.:

$$ET \approx g(\beta), \qquad \operatorname{Var} T \approx [g'(\beta)]^2 V.$$

If S is an average so that the central limit theorem (CLT) applies to it, i.e., $S \approx N(\beta, V)$, then

$$T \approx N(g(\beta), [g'(\beta)]^2 V)$$

for large n

if $V = v(\beta)$, then it is possible to choose g so that T has approximately constant variance in θ: solve $[g'(\beta)]^2 v(\beta) = \text{constant}$

Example: Exponential distribution.

$X_1, \ldots, X_n$ i.i.d. $\sim exp(\frac{1}{\mu})$, mean μ. Then $S = \overline{X}$ has mean μ and variance μ^2/n. If $T = \log \overline{X}$ then $g(\mu) = \log(\mu)$, $g'(\mu) = \mu^{-1}$, and so $\operatorname{Var} T \approx n^{-1}$, independent of μ: *variance stabilization*

If the Taylor expansion is carried to the second-derivative term, we obtaion

$$ET \approx g(\beta) + \frac{1}{2} V g''(\beta).$$

In practice we use numerical estimates for β and V if unknown.

When S, β vectors (V a matrix), with T still a scalar:

$\big(g'(\beta)\big)_i = \partial g / \partial \beta_i$ and $g''(\beta)$ matrix of second derivatives,

$$\operatorname{Var} T \approx [g'(\beta)]^T V g'(\beta)$$

and

$$ET \approx g(\beta) + \frac{1}{2}\,\mathrm{trace}[g''(\beta)V].$$

Example: Exponential family models

A one-parameter (i.e., scalar θ) exponential family density has the form

$$f(x;\theta) = \exp\{a(\theta)b(x) + c(\theta) + d(x)\}, \quad x \in A.$$

Examples: binomial, Poisson, normal (known mean, or known variance), gamma (known α, or known λ (including exponential) distributions

Example: Binomial (n, θ)

For $x = 0, 1, ..., n,$

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

Choose

$$
\begin{aligned}
a(\theta) &= \log\left(\frac{\theta}{1-\theta}\right) \\
b(x) &= x \\
c(\theta) &= n\log(1-\theta) \\
d(x) &= \log\left(\binom{n}{x}\right) \\
A &= \{1, \ldots, n\}.
\end{aligned}
$$

Choosing θ and x to make $a(\theta) = \theta$ and $b(x) = x$: *canonical form*

$$f(x;\theta) = \exp\{\theta x + c(\theta) + d(x)\}.$$

For the canonical form

$$EX = \mu(\theta) = -c'(\theta), \qquad \mathrm{Var}\, X = \sigma^2(\theta) = -c''(\theta)$$

Exercise: Prove the mean and variance results by calculating the moment-generating function $E\exp(tX) = \exp\{c(\theta) - c(t+\theta)\}$. Recall that you obtain mean and variance by differentiating the moment-generating function (how exactly?)

Example: Binomial (n, p)

Above we derived the exponential family form with

$$\begin{aligned} a(p) &= \log\left(\frac{p}{1-p}\right) \\ b(x) &= x \\ c(p) &= n\log(1-p) \\ d(x) &= \log\left(\binom{n}{x}\right) \\ A &= \{1,\ldots,n\}. \end{aligned}$$

To write the density in canonical form we put

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

(this transformation is called the *logit* transformation); then

$$p = \frac{e^\theta}{1+e^\theta}$$

and

$$
\begin{aligned}
a(\theta) &= \theta \\
b(x) &= x \\
c(\theta) &= -n\log(1+e^{\theta}) \\
d(x) &= \log\left(\binom{n}{x}\right) \\
A &= \{1,\ldots,n\}
\end{aligned}
$$

gives the canonical form. We calculate the mean

$$
-c'(\theta) = n\frac{e^{\theta}}{1+e^{\theta}} = \mu(\theta) = np
$$

and the variance

$$
\begin{aligned}
-c''(\theta) &= n\left\{\frac{e^{\theta}}{1+e^{\theta}} - \frac{e^{2\theta}}{(1+e^{\theta})^2}\right\} \\
&= \sigma^2(\theta) = np(1-p).
\end{aligned}
$$

Suppose $X_1, \ldots, X_n$ are i.i.d., canonical density. Then

$$\begin{aligned} \ell(\theta) &= \theta \sum x_i + nc(\theta) + \sum d(x_i), \\ \ell'(\theta) &= \sum x_i + nc'(\theta) = n(\overline{x} + c'(\theta)). \end{aligned}$$

Since $\mu(\theta) = -c'(\theta)$,

$$\ell'(\theta) = 0 \iff \overline{x} = \mu(\hat{\theta})$$

and $\ell''(\theta) = nc''(\theta)$, so $I_n(\theta) = E(-\ell''(\theta)) = -nc''(\theta)$. If μ is invertible, then

$$\hat{\theta} = \mu^{-1}(\overline{x}).$$

Example: Binomial(m, p). With $\theta = \log\left(\frac{p}{1-p}\right)$ we have $\mu(\theta) = m\frac{e^\theta}{1+e^\theta}$, and we calculate

$$\mu^{-1}(t) = \log\left(\frac{\frac{t}{m}}{1 - \frac{t}{m}}\right).$$

Note that here $n = 1$, we have a sample, x, of size 1. This gives

$$\hat{\theta} = \log\left(\frac{\frac{x}{m}}{1 - \frac{x}{m}}\right),$$

as expected from the invariance of mle's.

The CLT applies to $\overline{X}$ so, for large n,

$$\overline{X} \approx \mathcal{N}(\mu(\theta), -c''(\theta)/n)$$

With the delta-method, $S \approx \mathcal{N}(a, b)$ implies that

$$g(S) \approx \mathcal{N}\left(g(a), b[g'(a)]^2\right)$$

for continuous g, and small b. For $S = \overline{X}$, with $g(\cdot) = \mu^{-1}(\cdot)$ we have $g'(\cdot) = (\mu'(\mu^{-1}(\cdot))^{-1}$, thus

$$\hat{\theta} \approx \mathcal{N}\big(\theta, I_n^{-1}(\theta)\big)$$

giving the **Asymptotic Normality of the M.L.E.**.

Note: approximate variance equals the Cramér-Rao lower bound: *quite generally the MLE is asymptotically efficient.*

Example: Logistic regression. Linear model for log odds of binary response Y on predictor x; we are interested in

$$P(Y_i = 1|x) = \pi(x|\beta).$$

The outcome for each experiment is in $[0, 1]$; in order to apply some normal regression model we use the logit transform,

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

which is now spread over the whole real line. The ratio $\frac{p}{1-p}$ is also called the *odds*.

(Generalized linear) model

$$logit(\pi(x|\beta)) \ = \ \log\left(\frac{\pi(x|\beta)}{1-\pi(x|\beta)}\right) = x^T\beta.$$

The coefficients β describe how the odds for π change with change in the explanatory variables. The model can now be treated like an ordinary linear regression, X is the design matrix, β is the vector of coefficients. Transforming back,

$$P(Y_i = 1|x) \ = \ \exp(x^T\beta)\Big/\left(1+\exp(x^T\beta)\right).$$

Invariance property $\rightarrow$ MLE of $\pi(x|\beta)$, for any x, is $\pi(x|\hat{\beta})$, where $\hat{\beta}$ is the MLE obtained in the ordinary linear regression from a sample of responses $y_1, \ldots, y_n$ with associated covariate vectors $x_1, \ldots, x_n$.

(i) If β scalar: Calculate that

$$
\begin{aligned}
&\frac{\partial}{\partial \beta}\pi(x_i|\beta) \\
&= \frac{\partial}{\partial \beta}\exp(x_i\beta)\Big/ (1+\exp(x_i\beta)) \\
&= x_i e^{x_i\beta}\left(1+\exp(x_i\beta)\right)^{-1} \\
&\quad - \left(1+\exp(x_i\beta)\right)^{-2} x_i e^{x_i\beta} e^{x_i\beta} \\
&= x_i\pi(x_i|\beta) - x_i(\pi(x_i|\beta))^2 \\
&= x_i\pi(x_i|\beta)(1-\pi(x_i|\beta))
\end{aligned}
$$

and the likelihood is

$$
\begin{aligned}
L(\beta) &= \prod_{i=1}^{n} \pi(x_i|\beta) \\
&= \prod_{i=1}^{n} \exp(x_i\beta)\Big/ (1+\exp(x_i\beta))
\end{aligned}
$$

Hence the log likelihood has derivative

$$\begin{aligned}\ell'(\beta) &= \sum_{i=1}^{n} \frac{1}{\pi(x_i|\beta)} x_i \pi(x_i|\beta)(1-\pi(x_i|\beta)) \\ &= \sum_{i=1}^{n} x_i(1-\pi(x_i|\beta))\end{aligned}$$

so that

$$\ell''(\beta) = -\sum_{i=1}^{n} x_i^2 \pi(x_i|\beta))(1-\pi(x_i|\beta)).$$

Thus $\hat{\beta} \approx \mathcal{N}(\beta, I^{-1}\big(\beta)\big)$ where $I(\beta) = \sum x_i^2 \pi_i(1-\pi_i)$ with $\pi_i =$

$\pi(x_i|\beta)$

The delta method with $g(\beta) = e^{\beta x}/(1+e^{\beta x})$, gives

$$g'(\beta) = x g(\beta)(1-g(\beta))$$

and $\pi = \pi(x|\hat{\beta}) \approx \mathcal{N}\big(\pi, \pi^2(1-\pi)^2 x^2 I^{-1}(\beta)\big)$.

(ii) If β vector: Similarly it is possible to calculate that $\hat{\beta} \approx \mathcal{N}\big(\beta, I^{-1}(\beta)\big)$ where $[I(\beta)]_{kl} = E\left(-\partial^2 \ell / \partial\beta_k \partial\beta_l\right)$

Vector version of the delta method:

$$\pi(x|\hat{\beta}) \approx \mathcal{N}\left(\pi, \pi^2(1-\pi)^2 x^T I^{-1}(\beta) x\right)$$

with $\pi = \pi(x|\beta)$ and $I(\beta) = X^T R X$, where X is the design matrix, and

$$R = Diag\left(\pi_i(1-\pi_i),\ i = 1, \ldots, n\right)$$

where $\pi_i = \pi(x_i|\beta)$. Note that this normal approximation is likely to be poor for π near zero or one.

Excursion: Minimum Variance Unbiased Estimation MVUE.

There is a pretty theory about how to construct minimum variance unbiased estimators based on sufficient statistics. The key underlying result is the *Rao-Blackwell Theorem* (Casella+Berger p.316). We do not have time to go into detail during lectures, but you may like to read up on it.

Excursion: a more general method of moments

Consider statistics of the form $\frac{1}{n}\sum_{i=1}^{n} h(X_i)$

Find the expected value as a function of θ

$$\frac{1}{n}\sum_{i=1}^{n} Eh(X_i) = r(\theta)$$

Solve $r(\theta) = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$ for θ.