

1.6 Inference

Inference

The two main tools in statistics to make conclusions are

- **confidence intervals:** the width of a confidence interval provides a measure of the precision of the point estimates
- **hypothesis testing:** it compares how well two related models fit the data. The logic can be summarised as follows:
 - specify a model M_0 corresponding to H_0 and a more general model M_1 corresponding to H_1
 - fit model M_0 and compute measure of goodness of fit, G_0 ; repeat for M_1 to obtain G_1
 - calculate the improvement in fit
 - test the null hypothesis $G_0 = G_1$
 - if $G_0 = G_1$ is not rejected, then H_0 is not rejected and M_0 is the preferred model

For confidence intervals and hypothesis testing, sampling distributions are required.

- for **normally distributed r.v.**, the sampling distribution can be determined exactly
- for **other distributions** we need to rely on large-sample asymptotic results based on the CLT

The basic idea is that, under appropriate conditions (i.i.d and S being a sum), the statistic of interest S is

$$\frac{S - \mathbb{E}(S)}{\sqrt{\text{Var}(S)}} \sim \mathcal{N}(0, 1) \quad (1.6.1)$$

or equivalently

$$\frac{[S - \mathbb{E}(S)]^2}{\text{Var}(S)} \sim \chi_1^2 \quad (1.6.2)$$

and, in case, of p -multivariate statistics

$$[\mathbf{S} - \mathbb{E}(\mathbf{S})]^\top \mathbf{V}^{-1} [\mathbf{S} - \mathbb{E}(\mathbf{S})] \sim \chi_p^2 \quad (1.6.3)$$

Sampling distribution for score statistics

Suppose Y_1, \dots, Y_N are independent random variables from a distribution which belongs to the exponential family, with parameter $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)$.

The **score statistics** are such that

$$E[U_j] = 0, \quad \text{for all } j = 1, \dots, p. \quad (1.6.4)$$

The **variance-covariance matrix of the score statistics** is the information matrix \mathcal{I} , with elements

$$\mathcal{I}_{jk} = \mathbb{E}[U_j U_k] \quad (1.6.5)$$

If $p = 1$, the score statistic has the asymptotic sampling distribution

$$\frac{U}{\sqrt{\mathcal{I}}} \dot{\sim} \mathcal{N}(0, 1),$$

where the $\dot{\sim}$ symbol means "approximately distributed as". Equivalently we can write

$$\frac{U^2}{\mathcal{I}} \dot{\sim} \chi_1^2$$

If $p > 1$,

$$\mathbf{U} \dot{\sim} MVN(\mathbf{0}, \mathcal{I})$$

or, equivalently,

$$\mathbf{u}^\top \mathcal{I}^{-1} \mathbf{u} \dot{\sim} \chi_p^2$$

Example: Binomial distribution

If $Y \sim \text{Bin}(n, p)$, the log-likelihood function is

$$\ell(p; y) = y \log p + (n - y) \log(1 - p) + \log \binom{n}{y}$$

and the score statistic is

$$U = \frac{d\ell}{dp} = \frac{Y}{p} - \frac{n - Y}{1 - p} = \frac{Y - np}{p(1 - p)}$$

Since $\mathbb{E}(Y) = np$, $\mathbb{E}(U) = 0$ as expected.

Since $\mathbb{V}\text{ar}(Y) = np(1 - p)$, the information is

$$\mathcal{I} = \mathbb{V}\text{ar}(U) = \frac{1}{p^2(1 - p)^2} \mathbb{V}\text{ar}(Y) = \frac{n}{p(1 - p)}$$

and, hence,

$$\frac{U}{\sqrt{\mathcal{I}}} = \frac{Y - np}{\sqrt{np(1 - p)}} \stackrel{\sim}{\sim} \mathcal{N}(0, 1)$$

This is known as the *normal approximation to the binomial distribution*.

Taylor approximation

To obtain the asymptotic sampling distributions for various statistics, it is useful to use **Taylor approximations** for generic functions f in a neighbourhood of t

$$f(x) = f(t) + (x - t) \left[\frac{df}{dx} \right]_{x=t} + \frac{1}{2}(x - t)^2 \left[\frac{d^2 f}{dx^2} \right]_{x=t} + \dots$$

For a **log-likelihood**, the first *three* terms are

$$\ell(\theta) = \ell(\hat{\theta}) + (\theta - \hat{\theta})U(\hat{\theta}) - \frac{1}{2}(\theta - \hat{\theta})^2 \mathcal{I}(\hat{\theta}) \quad (1.6.6)$$

where $\hat{\theta}$ is the MLE of θ .

For a p -dimensional vector $\boldsymbol{\theta}$

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathbf{u}(\hat{\boldsymbol{\theta}}) - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \quad (1.6.7)$$

Similarly, for the **score function** of a one-dimensional parameter θ , the first *two* terms of the Taylor approximation are

$$U(\theta) \approx U(\hat{\theta}) + (\theta - \hat{\theta})U'(\hat{\theta}) = U(\hat{\theta}) - (\theta - \hat{\theta})\mathcal{I}(\hat{\theta}) \quad (1.6.8)$$

and the score function of a p -dimensional parameter $\boldsymbol{\theta}$ becomes

$$\mathbf{u}(\boldsymbol{\theta}) \approx \mathbf{u}(\hat{\boldsymbol{\theta}}) - \mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \quad (1.6.9)$$

Sampling distribution of the MLE

Let's define the MLE as $\hat{\boldsymbol{\theta}}$.

By definition, the MLE is the estimator which maximises $\ell(\hat{\boldsymbol{\theta}})$, i.e. $\mathbf{u}(\hat{\boldsymbol{\theta}}) = \mathbf{0}$

$$\begin{aligned}\mathbf{u}(\boldsymbol{\theta}) &\approx -\mathcal{I}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ -\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{u}(\boldsymbol{\theta}) &\approx (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})\end{aligned}$$

Properties:

- **consistency:** since $\mathbb{E}(\mathbf{u}) = \mathbf{0}$

$$\begin{aligned}\mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) &= \mathbf{0} \longrightarrow \mathbb{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} \\ \mathbb{E}(\mathbf{u}) &= \mathbb{E}(\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})) \\ \mathbf{0} &= \mathcal{I}(\hat{\boldsymbol{\theta}})\mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\ \mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}\mathbf{0} &= \mathbb{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\end{aligned}$$

- **variance-covariance matrix**

$$\begin{aligned}\mathbb{E}\left[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top\right] &= \mathbb{E}[\mathcal{I}^{-1}\mathbf{u}\mathbf{u}^\top\mathcal{I}^{-1}] \\ &= \mathcal{I}^{-1}\mathbb{E}[\mathbf{u}\mathbf{u}^\top]\mathcal{I}^{-1} = \mathcal{I}^{-1}\end{aligned}$$

- **asymptotic sampling distribution**

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \sim \chi^2(p)$$

Remarks

- $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^\top \mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is also known as **Wald statistics**
- for one-dimensional parameter, you can write $\hat{\theta} \sim \mathcal{N}(\theta, \mathcal{I}^{-1})$
- if the response variable is normally distributed, the results are exact; for other GLM, the results are asymptotic