TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
Asymptotic Distribution of MLE
Information Inequality

LV Nr. 105.692
Summer Semester 2021

- The method of maximum likelihood estimation is remarkable in that we can determine the asymptotic distribution of estimators that are defined only implicitly
  - the maximizer of the log likelihood frequently can only be calculated by computer optimization.
- To derive the asymptotic distribution of the MLE we need to first derive *asymptotics for log likelihood derivatives*

Log Likelihood Derivatives

- Let $f_\theta(x)$ denote the pdf (pmf) of X. Then,

$$\int f_\theta(x)dx = 1$$

or the analogous identity with summation replacing integration for the discrete case.

- We assume we can differentiate with respect to θ under the integral sign, which gives

$$\frac{d}{d\theta}\int f_\theta(x)dx = \int \frac{d}{d\theta} f_\theta(x)dx = 0$$

- Since

$$\ell'(\theta) = \frac{d}{d\theta}\log f_\theta(x) = \frac{1}{f_\theta}\frac{d}{d\theta}f_\theta(x)$$

we have

$$\frac{d}{d\theta}f_\theta(x) = \ell'(\theta)f_\theta(x)$$

so that

$$0 = \int \ell'(\theta)f_\theta(x)dx = \mathbb{E}_\theta(\ell'(\theta))$$

Log Likelihood Derivatives

Next, we consider the second derivative:

$$\int \frac{d^2}{d\theta^2} f_\theta(x)dx = 0$$

$$\ell''(\theta) = \frac{d}{d\theta}\frac{1}{f_\theta(x)}\frac{d}{d\theta}f_\theta(x)$$

$$= \frac{1}{f_\theta(x)}\frac{d^2}{d\theta^2}f_\theta(x) - \frac{1}{f_\theta^2(x)}\left(\frac{d}{d\theta}f_\theta(x)\right)^2$$

$$\frac{d^2}{d\theta^2}f_\theta(x) = \ell''(\theta)f_\theta(x) + \ell'(\theta)^2 f_\theta(x) \quad \text{so that}$$

$$\mathbb{E}_\theta(\ell''(\theta)) + \mathbb{E}_\theta(\ell'(\theta)^2) = 0$$

$$\mathbb{V}\text{ar}_\theta(\ell'(\theta)) = -\mathbb{E}_\theta(\ell''(\theta))$$

Log Likelihood Derivatives

If differentiation under the integral sign is valid,

- First log-likelihood derivative identity:

$$\mathbb{E}_\theta(\ell_n'(\theta)) = 0$$

  The identity holds when the θ in $\ell_n'(\theta)$ and in $\mathbb{E}_\theta$ are the same.
- Second log likelihood derivative identity:

$$\mathbb{V}\mathrm{ar}_\theta(\ell_n'(\theta)) = -\mathbb{E}_\theta(\ell_n''(\theta))$$

  The identity holds when the θ in $\ell_n'(\theta)$ and in $\mathbb{V}\mathrm{ar}_\theta$ are the same.

Fisher Information

- Either side of the second log likelihood derivative identity is called Fisher information

$$I_n(\theta) = \mathbb{V}\mathrm{ar}_\theta(\ell_n'(\theta)) = -\mathbb{E}_\theta(\ell_n''(\theta))$$

- When the data are iid, then the log likelihood and its derivatives are the sum of iid terms

$$\ell_n(\theta) = \sum_{i=1}^n \log f_\theta(x_i)$$

$$\ell_n'(\theta) = \sum_{i=1}^n \frac{d}{d\theta} \log f_\theta(x_i)$$

$$\ell_n''(\theta) = \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f_\theta(x_i)$$

- That is,

$$I_n(\theta) = nI_1(\theta)$$

Asymptotics for Log Likelihood Derivatives

- Also,

$$n^{-1}\ell_n'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\theta} \log f_\theta(x_i)$$

$$n^{-1}\ell_n''(\theta) = \frac{1}{n}\sum_{i=1}^{n} \frac{d^2}{d\theta^2} \log f_\theta(x_i)$$

  and the LLN and CLT apply to them.

Asymptotics for Log Likelihood Derivatives

LLN:

$$\mathbb{E}_\theta\left(\frac{d}{d\theta}\log f_\theta(x_i)\right) = \mathbb{E}_\theta(\ell_1'(\theta)) = 0$$

$$\mathbb{E}_\theta\left(\frac{d^2}{d\theta^2}\log f_\theta(x_i)\right) = \mathbb{E}_\theta(\ell_1''(\theta)) = -I_1(\theta)$$

- Hence the LLN applied to log likelihood derivatives yields

$$n^{-1}\ell_n'(\theta) \xrightarrow{P} 0$$

$$n^{-1}\ell_n''(\theta) \xrightarrow{P} -I_1(\theta)$$

when $X_1, X_2, \ldots$ are iid with pdf (pmf) f_θ (i.e. θ is the true parameter value)

Asymptotics for Log Likelihood Derivatives

CLT:

$$\mathbb{E}_\theta\left(\frac{d}{d\theta}\log f_\theta(x_i)\right) = \mathbb{E}_\theta(\ell_1'(\theta)) = 0$$
$$\mathbb{V}\mathrm{ar}_\theta\left(\frac{d}{d\theta}\log f_\theta(x_i)\right) = \mathbb{V}\mathrm{ar}_\theta(\ell_1'(\theta)) = I_1(\theta)$$

- Hence the CLT is

$$\sqrt{n}\left(n^{-1}\ell_n'(\theta) - 0\right) \xrightarrow{d} \mathcal{N}(0, I_1(\theta))$$

or,

CLT for MLE

$$n^{-1/2}\ell_n'(\theta) \xrightarrow{d} \mathcal{N}(0, I_1(\theta))$$

when $X_1, X_2, \ldots$ are iid with pdf (pmf) f_θ (i.e. θ is the true parameter value)

MLE Asymptotics

- The MLE $\hat{\theta}_n$ is a local maximizer (at least) of the log likelihood, hence it satisfies

$$\ell_n'(\hat{\theta}_n) = 0$$

- Taylor series expansion of the log likelihood about the true known parameter value, say θ_0, is

$$\ell_n'(\theta) = \ell_n'(\theta_0) + \ell_n''(\theta_0)(\theta - \theta_0) + \text{higher order terms}$$

or,

$$n^{-1/2}\ell_n'(\theta) = n^{-1/2}\ell_n'(\theta_0) + n^{-1}\ell_n''(\theta_0)n^{1/2}(\theta - \theta_0) + \text{higher order terms}$$

MLE Asymptotics

- Assuming the higher order terms are negligible when $\hat{\theta}_n$ is plugged in for θ,

$$0 = n^{-1/2}\ell_n'(\theta_0) + n^{-1}\ell_n''(\theta_0)n^{1/2}(\hat{\theta}_n - \theta_0) + o_p(1)$$

- This implies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{n^{-1/2}\ell_n'(\theta_0)}{n^{-1}\ell_n''(\theta_0)} + o_p(1)$$

- By Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -\frac{Y}{I_1(\theta_0)}$$

where

$$Y \sim \mathcal{N}(0, I_1(\theta_0))$$

and thus

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, I_1^{-1}(\theta_0))$$

Theorem

Let $\{f(x|\theta) : \theta \in \Theta\}$ be a parametric model, where $\theta \in \mathbb{R}$ is a single parameter. Let $X_1, \ldots, X_n$ be a random sample with pdf $f(x|\theta_0)$, for $\theta_0 \in \Theta$, and let $\hat{\theta}_n$ be the MLE based on the sample. Suppose certain regularity conditions hold, including

1. all pdfs/pmfs $f(x|\theta)$ in the model have the same support
2. θ_0 is an interior point of Θ
3. the log likelihood $l(\theta) = l(\theta|x)$ is differentiable in θ
4. $\hat{\theta}$ is the unique value of $\theta \in \Theta$ that solves the equation $\frac{d}{d\theta} l(\theta|x) = 0$.

Then $\hat{\theta}$ is consistent and asymptotically normal with

$$\sqrt{n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{1}{I(\theta_0)}\right)$$

as $n \to \infty$, where $I(\theta)$ is the Fisher information defined by the two equivalent expressions

$$\begin{aligned} I(\theta) &= \mathbb{V}\mathrm{ar}\left(z(X, \theta)\right) \\ &= -\mathbb{E}\left(z'(X, \theta)\right), \end{aligned} \tag{1}$$

where the variance and expectation are taken with respect to $X \sim f(x|\theta)$ and

$$z(x, \theta) = \frac{\partial}{\partial \theta} \log f(x|\theta) \qquad \text{and} \qquad z'(x, \theta) = \frac{\partial^2}{\partial \theta^2} \log f(x|\theta).$$

The function $z(x, \theta)$ is called the *score function*.

Asymptotics for MLE

Without knowing anything about the functional form of the MLE, we have derived its asymptotic distribution!
That is,

$$\hat{\theta}_n \approx \mathcal{N}\left(\theta_0, \frac{I_1^{-1}(\theta_0)}{n}\right)$$

or,

$$\hat{\theta}_n \approx \mathcal{N}(\theta_0, I_n^{-1}(\theta_0))$$

The regularity conditions required in Theorem 1 are very technical. They relate to differentiability of the density and the ability to interchange differentiation and integration.

Example: Poisson

Suppose $X_1, \ldots, X_n$ are iid from $\mathcal{P}oi(\lambda)$ with unknown parameter $\lambda > 0$. Thus, each X_i has pmf

$$f(x|\lambda) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda}, & x = 0, 1, 2, \ldots \\ 0, & \text{otherwise} \end{cases} = \frac{\lambda^x}{x!} e^{-\lambda} \cdot \mathbb{1}_{(x \in \mathbb{N}_0)}.$$

The likelihood given x is

$$L(\lambda|x) = \prod_{i=1}^{n} f(x_i|p) = \prod_{i=1}^{n} \left(\frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \right) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \cdot e^{-n\lambda}.$$

The log likelihood given x is

$$\ell(\lambda|x) = \log L(\lambda|x) = \left(\sum_{i=1}^{n} x_i \right) \cdot \log \lambda - \sum_{i=1}^{n} \log(x_i!) - n\lambda, \qquad \lambda > 0.$$

Example: Poisson

The MLE satisfies the first-order condition for an interior maximum $\frac{d}{d\lambda}\ell(\lambda|x)\Big|_{\lambda=\hat{\lambda}} = 0$. Thus, by solving

$$\frac{d}{dp}\ell(p|x)\Big|_{p=\hat{p}} = \Big(\sum_{i=1}^{n} x_i\Big)\cdot\frac{1}{\hat{\lambda}} - n = 0$$

we obtain

$$\hat{\lambda} = \frac{1}{n}\cdot\sum_{i=1}^{n} x_i = \bar{x}.$$

This unique solution to the first-order condition is a maximizer because

$$\frac{d^2}{d\lambda^2}\ell(\lambda|x)\Big|_{\lambda=\hat{\lambda}} = -\Big(\sum_{i=1}^{n} x_i\Big)\cdot\frac{1}{\hat{p}^2} < 0.$$

Example: Poisson

To see how accurate the MLE is, we recall

$$\mathbb{E}(\hat{\lambda}) = \mathbb{E}(\bar{X}) = \lambda \qquad \text{and} \qquad \mathbb{V}\text{ar}(\hat{\lambda}) = \mathbb{V}\text{ar}(\bar{X}) = \frac{\lambda}{n}.$$

Thus, $\hat{\lambda}$ is unbiased with variance $\frac{\lambda}{n}$. Additionally, for large n, by the LLN, the sample mean $\bar{X} \xrightarrow{p} \lambda$ as $n \to \infty$, and by the CLT

$$\frac{\bar{X} - \lambda}{\frac{\lambda}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

as $n \to \infty$. Therefore, for large n, we expect $\hat{\lambda}$ to be close to λ and the sampling distribution of $\hat{\lambda}$ is approximately $\mathcal{N}(\lambda, \frac{\lambda}{n})$.

Example: Poisson

Let's apply the theorem:

$$\log f(x|\lambda) = \log\left(\frac{\lambda^x}{x!}\, e^{-\lambda}\right) = x\log\lambda - \lambda - \log(x!).$$

Then the score function and its derivative are given by

$$z(x,\lambda) = \frac{\partial}{\partial\lambda}\, \log f(x|\lambda) = \frac{x}{\lambda} - 1, \qquad z'(x,\lambda) = \frac{\partial^2}{\partial\lambda^2}\, \log f(x|\lambda) = -\frac{x}{\lambda^2}.$$

The Fisher information is

$$I(\lambda) = -\mathbb{E}(z'(X,\lambda)) = -\frac{1}{\lambda^2}\,\mathbb{E}(X) = \frac{1}{\lambda}.$$

Thus,

$$\sqrt{n}\,(\hat{\lambda} - \lambda) = \sqrt{n}\,(\bar{X} - \lambda) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \frac{1}{I(\lambda)}\right) = \mathcal{N}(0,\lambda).$$

This is the same result as what we obtained by applying the CLT directly. This normal approximation will be used later in the course, for example to obtain a confidence interval for $\hat{\lambda}$.

Example: Normal

X_i iid normal with known mean μ and unknown variance $\nu = \sigma^2$

$$\begin{aligned} I_n(\nu) &= -\mathbb{E}(\ell_n''(\nu)) \\ &= -\mathbb{E}\left(\frac{n}{2\nu^2} - \frac{1}{\nu^3}\sum_{i=1}^{n}(X_i - \mu)^2\right) \\ &= -\frac{n}{2\nu^2} + \frac{1}{\nu^3}n\nu = \frac{n}{2\nu^2} \end{aligned}$$

Thus,

$$\hat{\nu}_n \approx \mathcal{N}(\nu, I_n^{-1}(\nu)) = \mathcal{N}(\nu, \frac{2\nu^2}{n})$$

or,

$$\widetilde{S}_n^2 \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^2}{n}\right)$$

Example: One-parameter Gamma

X_i iid $Gamma(\alpha, \lambda)$ with known rate λ and unknown shape α

$$
\begin{aligned}
L_n(\alpha) &= \prod_{i=1}^{n} \frac{\lambda^\alpha}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\lambda x_i} \\
&= \left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \right)^n \prod_{i=1}^{n} x_i^{\alpha-1} e^{-\lambda x_i} \\
&= \left(\frac{\lambda^\alpha}{\Gamma(\alpha)} \right)^n \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \exp\left(-\lambda \sum_{i=1}^{n} x_i \right)
\end{aligned}
$$

Example: One-parameter Gamma

The log-likelihood is

$$\ell_n(\alpha) = n\alpha - n\log(\Gamma(\alpha)) + (\alpha - 1)\log\left(\prod_{i=1}^{n} x_i\right)$$

Every term except $n\log(\Gamma(\alpha))$ is linear in α and hence has second derivative with respect to α equal to zero.
Therefore,

$$\ell_n''(\alpha) = -n\frac{d^2}{d\alpha^2}\log\Gamma(\alpha)$$

and

$$I_n(\alpha) = n\frac{d^2}{d\alpha^2}\log\Gamma(\alpha)$$

Example: One-parameter Gamma

The second derivative of the logarithm of the gamma function is called the trigamma function, which can be calculated in R.

It satisfies

$$\frac{d^2}{d\alpha^2}\,\ell(\hat{\alpha}) = -\frac{d^2}{d\alpha^2}\log\Gamma(\hat{\alpha}) < 0$$

In sum, even though there is no closed form expression for the MLE, its asymptotic distribution is

$$\hat{\alpha}_n \approx \mathcal{N}\left(\alpha, \frac{1}{n\ trigamma(\alpha)}\right)$$

Asymptotic variance of the MLE

- Since θ_0 is unknown, the Fisher information $I_1(\theta_0)$ is also unknown, that is the asymptotic variance of the MLE is unknown.
- We estimate it using the *plug-in* method:
  - If $I_1(\theta)$ is a continuous function, then

$$I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0)$$

    by the continuous mapping theorem.
  - By Slutsky's theorem,

$$\sqrt{n}\frac{\hat{\theta}_n - \theta_0}{I_1(\hat{\theta}_n)^{-1/2}} = (\hat{\theta}_n - \theta_0)I_1(\hat{\theta}_n)^{1/2} \xrightarrow{d} \mathcal{N}(0,1)$$

Fisher Information

$I(\theta_0)$ quantifies the *amount of information that each observation* X_i *contains about the unknown parameter*.

- If $I(\theta_0)$ is small, then a small change in θ does not affect the log likelihood $l(\theta)$ much, and the data do not provide much information that the true value of θ is close to θ_0.
- If $I(\theta_0)$ is large, then a small change in θ can lead to a large decrease in $l(\theta)$ an therefore the data provide more information about the true value of θ.

Observed Fisher Information

- If the expectation involved in calculating Fisher information is too hard, we use the observed Fisher information

$$J_n(\theta) = -\ell_n''(\theta)$$

  - The LLN says

$$n^{-1}\ell_n''(\theta_0) \xrightarrow{P} -I_1(\theta_0)$$

  - Then, $J_n(\theta) \approx I_n(\theta)$ so that

$$J_n(\hat{\theta}_n) \approx I_n(\hat{\theta}_n) \approx I_n(\theta_0)$$

    should also hold (we do not prove this).

The Information Inequality

Theorem

Consider a parametric model $\{f(x|\theta) : \theta \in \Theta\}$ *satisfying certain mild regularity assumptions, where* $\theta \in \mathbb{R}$ *is a single parameter. Let* $\hat{\theta}_n = T(X_1, \ldots, X_n)$ *be any unbiased estimator of* θ *based on a random sample* $X_1, \ldots, X_n$ *from a population with pdf (pmf)* $f(x|\theta)$*. Then*

$$\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}_n) \geqslant \frac{1}{nI(\theta)} = I_n(\theta)^{-1} \tag{2}$$

which is called the information inequality or the Cramér-Rao lower bound.

Proof

$$\begin{aligned}\mathbb{C}\mathrm{ov}_\theta(\hat{\theta}, \ell'(\theta)) &= \mathbb{E}_\theta(\hat{\theta}, \ell'(\theta)) - \mathbb{E}_\theta(\hat{\theta})\mathbb{E}_\theta(\ell'(\theta)) \\ &= \mathbb{E}_\theta(\hat{\theta}, \ell'(\theta))\end{aligned}$$

since $\mathbb{E}_\theta(\ell'(\theta)) = 0$. Also,

$$\begin{aligned}\mathbb{E}_\theta(\hat{\theta}, \ell'(\theta)) &= \int \hat{\theta} \left(\frac{1}{f_\theta(x)} \frac{d}{d\theta} f_\theta(x) \right) f_\theta(x) dx \\ &= \int \hat{\theta} \frac{d}{d\theta} f_\theta(x) dx \\ &= \frac{d}{d\theta} \int \hat{\theta} f_\theta(x) dx \\ &= \frac{d}{d\theta} \mathbb{E}_\theta(\hat{\theta})\end{aligned}$$

assuming differentiation under the integral sign is valid.

Proof (ctd.)

Since $\hat{\theta}$ is unbiased,

$$\begin{aligned} \mathbb{E}_\theta(\hat{\theta}) &= \theta \\ \frac{d}{d\theta}\mathbb{E}_\theta(\hat{\theta}) &= 1 \quad \text{so that} \\ \mathbb{C}\mathrm{ov}_\theta(\hat{\theta}, \ell'(\theta)) &= 1 \end{aligned}$$

From the correlation inequality,

$$\begin{aligned} 1 &\geqslant \mathrm{cor}^2_\theta(\hat{\theta}, \ell'(\theta)) \\ &= \frac{\mathbb{C}\mathrm{ov}_\theta(\hat{\theta}, \ell'(\theta))}{\mathbb{V}\mathrm{ar}_\theta(\hat{\theta})\mathbb{V}\mathrm{ar}_\theta(\ell'(\theta)} \\ &= \frac{1}{\mathbb{V}\mathrm{ar}_\theta(\hat{\theta})I(\theta)} \end{aligned}$$

from which the inequality follows.

The Information Inequality (ctd.)

- The information inequality says no unbiased estimator can be more efficient than the MLE. But what about biased estimators? They can be more efficient.
- An estimator that is better than the MLE in the ARE sense is called superefficient, and such estimators do exist.
- The Hájek convolution theorem says no estimator that is asymptotically unbiased in a certain sense can be superefficient.
- The Le Cam convolution theorem says no estimator can be superefficient except at a set of true unknown parameter points of measure zero.

The Information Inequality (ctd.)

- In summary, the MLE is as about as efficient as an estimator can be.
- Exact theory: no unbiased estimator can be superefficient.
- Asymptotic theory: no estimator can be superefficient except at a negligible set of true unknown parameter values.

Multiparameter MLE

Basic ideas are the same:

- conditions for local and global maxima
- log likelihood derivative identities
- Fisher information
- asymptotics and plug-in

Local Maxima

Suppose W is an open region of $\mathbb{R}^p$ and $f : W \to \mathbb{R}$ is a twice differentiable function. A necessary condition for a point $\mathbf{x} \in W$ to be a local maximum of f is

$$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right)^T = 0$$

and a sufficient condition for $\mathbf{x}$ to be a local maximum is

$$\nabla^2 f(\mathbf{x}) = \begin{pmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_p} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_p \partial x_1} & \frac{\partial^2 f(\mathbf{x})}{\partial x_p \partial x_2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_p^2} \end{pmatrix} < 0$$

Local Maxima

- For the first-order condition for a local maximum, set all first partial derivatives to zero and solve for the variables.
- The second-order condition is hard when done by hand: we have to verify that

$$\sum_{i=1}^{p} \sum_{j=1}^{p} w_i w_j \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} < 0$$

  for all real numbers $w_1, \ldots, w_p$.
- The computer check that all eigenvalues are negative is easy.

Example: two-parameter normal

- The log likelihood for the two-parameter normal model is

$$\ell_n(\mu, \nu) = -\frac{n}{2}\log(\nu) - \frac{n\tilde{s}_n^2}{2\nu} - \frac{n(\bar{x}_n - \mu)^2}{2\nu}$$

- The first partial derivatives are

$$\frac{\partial \ell_n(\mu, \nu)}{\partial \mu} = \frac{n(\bar{x}_n - \mu)}{\nu}$$
$$\frac{\partial \ell_n(\mu, \nu)}{\partial \mu} = -\frac{n}{2\nu} - \frac{n\tilde{s}_n^2}{2\nu^2} - \frac{n(\bar{x}_n - \mu)^2}{2\nu^2}$$

Example: two-parameter normal

- The second partial derivatives a

$$\frac{\partial^2 \ell_n(\mu, \nu)}{\partial \mu^2} = -\frac{n}{\nu}$$
$$\frac{\partial^2 \ell_n(\mu, \nu)}{\partial \mu \partial \nu} = -\frac{n(\bar{x}_n - \mu)}{\nu^2}$$
$$\frac{\partial^2 \ell_n(\mu, \nu)}{\partial \nu^2} = \frac{n}{2\nu^2} - \frac{n\tilde{s}_n^2}{\nu^3} - \frac{n(\bar{x}_n - \mu)^2}{\nu^3}$$

Example: two-parameter normal

- Setting the first partial derivative with respect to μ equal to zero and solving for μ gives
$$\hat{\mu} = \bar{x}_n$$
- Plugging that into the first partial derivative with respect to ν and set equal to zero gives
$$-\frac{n}{2\nu} + \frac{n\tilde{s}_n^2}{2\nu^2} = 0$$
and solving for ν gives
$$\hat{\nu} = \hat{\sigma}^2 = \tilde{s}_n^2$$
- Plugging the MLE into the second partial derivatives gives
$$\frac{\partial^2 \ell_n(\hat{\mu}, \hat{\nu})}{\partial \mu^2} = -\frac{n}{\hat{\nu}}$$
$$\frac{\partial^2 \ell_n(\hat{\mu}, \hat{\nu})}{\partial \mu \partial \nu} = 0$$
$$\frac{\partial^2 \ell_n(\hat{\mu}, \hat{\nu})}{\partial \nu^2} = \frac{n}{2\hat{\nu}^2} - \frac{n\tilde{s}_n^2}{\hat{\nu}^3} = -\frac{n}{2\hat{\nu}^2}$$
- Hence the Hessian matrix is diagonal, and is negative definite since each of the diagonal terms is negative: the MLE is a local maximizer of the log-likelihood.

Global Maxima

A region W of $\mathbb{R}^p$ is *convex* if

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in W$$

for $\mathbf{x}, \mathbf{y} \in W$ and $0 < \lambda < 1$.
Suppose W is an open convex region of $\mathbb{R}^p$ and $f : W \to \mathbb{R}$ is a twice-differentiable function. If

$$\nabla^2 f(\mathbf{y}) < 0 \quad \text{for all } \mathbf{y} \in W$$

then f is called strictly concave. In this case

$$\nabla f(\mathbf{x}) = 0$$

is a sufficient condition for $\mathbf{x}$ to be the unique global maximum.

Log Likelihood Derivative Identities

The same differentiation under the integral sign argument applied to partial derivatives results in

$$\mathbb{E}_\theta\left(\frac{\partial \ell_n(\theta)}{\partial \theta_i}\right) = 0$$

$$\mathbb{E}_\theta\left(\frac{\partial \ell_n(\theta)}{\partial \theta_i}\frac{\partial \ell_n(\theta)}{\partial \theta_j}\right) = -\mathbb{E}_\theta\left(\frac{\partial^2 \ell_n(\theta)}{\partial \theta_i \partial \theta_j}\right)$$

or,

$$\mathbb{E}_\theta\left(\nabla \ell_n(\theta)\right) = 0$$

$$\mathbb{V}\mathrm{ar}_\theta\left(\nabla \ell_n(\theta)\right) = -\mathbb{E}_\theta\left(\nabla^2 \ell_n(\theta)\right)$$

Fisher Information

- As in the uniparameter case, either side of the second log likelihood derivative identity is called Fisher information

$$\mathbf{I}_n(\boldsymbol{\theta}) = \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}\left(\nabla \ell_n(\boldsymbol{\theta})\right) = -\mathbb{E}_{\boldsymbol{\theta}}\left(\nabla^2 \ell_n(\boldsymbol{\theta})\right)$$

- Being a variance matrix, the Fisher information matrix is symmetric and positive semi-definite.
- It is usually positive definite, and we will assume this in the course.

Example (cont.)

1. Back to the two-parameter normal model

$$\mathbb{E}_{\mu,\nu}\left(\frac{\nabla^2 \ell_n(\mu,\nu)}{\partial \mu^2}\right) = -\frac{n}{\nu}$$

$$\mathbb{E}_{\mu,\nu}\left(\frac{\nabla^2 \ell_n(\mu,\nu)}{\partial \mu \partial \nu}\right) = -\frac{n\mathbb{E}_{\mu,\nu}(\bar{X}_n - \mu)}{\nu^2} = 0$$

$$\begin{aligned}\mathbb{E}_{\mu,\nu}\left(\frac{\nabla^2 \ell_n(\mu,\nu)}{\partial \nu^2}\right) &= \frac{n}{2\nu^2} - \frac{n\mathbb{E}_{\mu,\nu}(\widetilde{S}_n^2)}{\nu^3} - \frac{n\mathbb{E}_{\mu,\nu}(\bar{X}_n - \mu)^2}{\nu^3} \\ &= \frac{n}{2\nu^2} - \frac{(n-1)\mathbb{E}_{\mu,\nu}(S_n^2)}{\nu^3} - \frac{n\mathbb{V}\mathrm{ar}_{\mu,\nu}(\bar{X}_n)}{\nu^3} \\ &= -\frac{n}{2\nu^2}\end{aligned}$$

2. Thus,

$$\mathbf{I}_n(\theta) = \begin{pmatrix} n/\nu & 0 \\ 0 & n/2\nu^2 \end{pmatrix}$$

Multivariate MLE CLT

1. The same CLT argument applied to the gradient vector gives

$$n^{-1/2}\nabla\ell_n(\theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_1(\theta_0))$$

and the same LLN argument applied to the Hessian matrix gives

$$-n^{-1}\nabla^2\ell_n(\theta_0) \xrightarrow{P} \mathbf{I}_1(\theta_0)$$

2. The same argument, i.e., expand the gradient of the log likelihood in a Taylor series, assume terms after the first two are negligible, and apply Slutsky, used in the univariate case gives

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathbf{I}_1(\theta_0)^{-1})$$

or,

$$\widehat{\theta}_n \approx \mathcal{N}\left(\theta_0, \mathbf{I}_1(\theta_0)^{-1}\right)$$

Example (ctd.)

1. Returning to the two-parameter normal model, inverse Fisher information is

$$\mathbf{I}_n(\theta)^{-1} = \begin{pmatrix} \nu/n & 0 \\ 0 & 2\nu^2/n \end{pmatrix} = \begin{pmatrix} \sigma^2/n & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}$$

2. Since the asymptotic covariance is zero, the two components of the MLE are asymptotically independent (we know they are for any n) and their asymptotic distributions are

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\widetilde{S}_n^2 \approx \mathcal{N}\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

Example: Gamma(α, λ)

Let $X_i \sim$ i.i.d. $Gamma(\alpha, \lambda)$). The log-likelihood is

$$\begin{aligned}
\ell_n(\alpha, \lambda) &= n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \log \prod_{i=1}^{n} x_i - \lambda \sum_{i=1}^{n} x_i \\
&= n\alpha \log \lambda - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log x_i - \lambda \sum_{i=1}^{n} x_i \\
&= n\alpha \log \lambda - n \log \Gamma(\alpha) + n(\alpha - 1)\bar{y}_n - n\lambda \bar{x}_n
\end{aligned}$$

where $\bar{y}_n = \sum_i \log x_i / n$.

Example: Gamma(α, λ)

$$\frac{\partial \ell_n(\alpha, \lambda)}{\partial \alpha} = n \log \lambda - n\text{digamma}(\alpha) + n\bar{y}_n$$
$$\frac{\partial^2 \ell_n(\alpha, \lambda)}{\partial \lambda} = \frac{n\alpha}{\lambda} - -n\lambda\bar{x}_n$$
$$\frac{\partial \ell_n(\alpha, \lambda)}{\partial \alpha^2} = n \text{ trigamma}(\alpha)$$
$$\frac{\partial^2 \ell_n(\alpha, \lambda)}{\partial \lambda \partial \lambda} = \frac{n}{\lambda}$$
$$\frac{\partial \ell_n(\alpha, \lambda)}{\partial \lambda^2} = -\frac{n\alpha}{\lambda^2}$$

There is no closed-form solution for the first order equations!
The MLE can only be found numerically, maximizing the log likelihood for particular data.

But we know the asymptotic distribution of the MLE

$$\begin{pmatrix} \hat{\alpha}_n \\ \hat{\lambda}_n \end{pmatrix} \approx \mathcal{N}\left(\begin{pmatrix} \alpha_n \\ \lambda_n \end{pmatrix}, \mathbf{I}_n(\theta)^{-1} \right)$$

where

$$\mathbf{I}_n(\theta) = \begin{pmatrix} n \text{ trigamma}(\alpha) & -n/\lambda \\ -n/\lambda & n\alpha/\lambda^2 \end{pmatrix}$$

Since we don't know θ, we must use a plug-in estimate for asymptotic variance. As in the uniparameter case, we can use either expected Fisher information,

$$\widehat{\theta}_n \approx \mathcal{N}(\theta, \mathbf{I}_n(\widehat{\theta})^{-1})$$

or the observed Fisher information,

$$\widehat{\theta}_n \approx \mathcal{N}(\theta, \mathbf{J}_n(\widehat{\theta})^{-1})$$

where $\mathbf{J}_n(\widehat{\theta}) = -\nabla^2 \ell_n(\theta)$.

Starting Points for Optimization

1. When a maximum likelihood problem is not concave, there can be more than one local maximum.
2. Math Stats: one of the local maxima is the efficient estimator which has inverse Fisher information for its asymptotic variance. The rest of the local maxima are no good.
3. How to find the right one? We need a starting point for optimization that is a " a root n consistent" estimator, that is,

$$\widetilde{\theta}_n = \theta_0 + O_p(n^{-1/2})$$

4. Any CAN estimator: for example, method of moments estimators and sample quantiles. such that

Invariance of Maximum Likelihood

1. If $\psi = g(\theta)$ is an invertible change of parameter and $\widehat{\theta}_n$ is the MLE for θ, then $\widehat{\psi}_n = g(\widehat{\theta}_n)$ is the MLE for ψ.
2. Let's see this for the univariate case:

$$\begin{aligned}
\ell_n(\theta) &= \tilde{\ell}_n(g(\theta)) \\
\ell_n'(\theta) &= \tilde{\ell}_n'(g(\theta))g'(\theta) \\
\ell_n''(\theta) &= \tilde{\ell}_n''(g(\theta))(g'(\theta))^2 + \tilde{\ell}_n'(g(\theta))g''(\theta)
\end{aligned}$$

Taking expectations, the second term in the second derivative is zero by the first log likelihood derivative identity, which obtains the one-parameter case of what was to be proved.

Invariance of Maximum Likelihood

1. This invariance does not extend to derivatives of the log likelihood.
2. If $\mathbf{I}(\theta)$ is the Fisher information matrix for θ and $\widetilde{\mathbf{I}}(\psi)$ is the Fisher information matrix for ψ, then the chain rule and log likelihood derivative identities give

$$\mathbf{I}_n(\theta) = (\nabla g(\theta))^T \widetilde{\mathbf{I}}_n(\psi) (\nabla g(\theta))$$

Example

- Using the invariant property, we obtain that the MLE of μ^2, the square of a normal mean, given a random sample $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$, is

$$\hat{\mu}^2 = \bar{X}^2.$$

- Also, the MLE of $\sqrt{p(1-p)}$, where p is a Binomial probability, is given by

$$\sqrt{\hat{p}(1-\hat{p})}.$$

Example: Pareto distribution

The Pareto (x_0, θ) distribution for $x_0 > 0$ and $\theta > 1$ is a continuous distribution over the interval $[x_0, +\infty)$, with pdf

$$f(x|x_0, \theta) = \begin{cases} \theta\, x_0^{\theta}\, x^{-\theta-1}, & x \geqslant x_0 \\ 0, & x < x_0. \end{cases}$$

The Pareto distribution is commonly used in economics as a model for the distribution of income.

The value x_0 represents the minimum possible income. Let us assume here that x_0 is known and equal to 1. We then have a one parameter model (the shape parameter θ).

Example: Pareto distribution

- The expectation and variance of the Pareto distribution are

$$\mathbb{E}(X) = \int_1^{+\infty} \frac{\theta}{x^\theta}\, dx = \frac{\theta}{\theta - 1} \quad \text{and} \quad \mathbb{V}\mathrm{ar}(X) = \frac{\theta}{(\theta-2)(\theta-1)^2}, \text{ for } \theta > 2.$$

  For $\theta \leqslant 2$ the variance of X is infinite (HW).

- We can estimate the mean income by $\frac{\hat{\theta}}{\hat{\theta}-1}$, where $\hat{\theta}$ is the MLE.

- To compute $\hat{\theta}$ from the sample $X_1, \ldots, X_n$, the log likelihood function given $x = (x_1, \ldots, x_n)$ is

$$\begin{aligned}\ell(\theta|x) &= \log \prod_{i=1}^n f(x_i|1,\theta) = \sum_{i=1}^n \log(\theta\, x_i^{-\theta-1}) \\ &= \sum_{i=1}^n (\log\theta - (\theta+1)\log x_i) = n\log\theta - (\theta+1)\sum_{i=1}^n \log x_i.\end{aligned}$$

Example: Pareto distribution

Then, from $\frac{d}{d\theta}\ell(\theta|x)\Big|_{\theta=\hat{\theta}} = 0$ and $\frac{d^2}{d\theta^2}\ell(\theta|x) = -\frac{n}{\theta^2} < 0$ we obtain the MLE

$$\hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log X_i}.$$

We next compute the Fisher information. For $x \geqslant 1$ we have

$$\log f(x|1,\theta) = \log(\theta\, x^{-\theta-1}) = \log\theta - (\theta+1)\log x$$
$$z(x,\theta) = \frac{\partial}{\partial\theta}\log f(x|1,\theta) = \frac{1}{\theta} - \log x$$
$$z'(x,\theta) = \frac{\partial^2}{\partial\theta^2}\log f(x|1,\theta) = -\frac{1}{\theta^2}$$

and the Fisher information is given by

$$I(\theta) = -\mathbb{E}\left(z'(x,\theta)\right) = \frac{1}{\theta^2}.$$

Example: Pareto distribution

We know,

$$\frac{\hat{\theta} - \theta}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}\Big(0, \frac{1}{I(\theta)}\Big) = \mathcal{N}(0, \theta^2).$$

For $g(\theta) = \frac{\theta}{\theta - 1}$ we have $g'(\theta) = -\frac{1}{(\theta-1)^2} \neq 0$. The Delta method then implies

$$\frac{g(\hat{\theta}) - g(\theta)}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}\Big(0, \frac{g'(\theta)^2}{I(\theta)}\Big) = \mathcal{N}\Big(0, \frac{\theta^2}{(\theta-1)^4}\Big).$$

If for example we take a data set with $n = 10000$ income values with MLE value $\hat{\theta} = 3.5$. Then we can estimate the mean income as $\frac{\hat{\theta}}{\hat{\theta}-1} = 1.4$, and estimate the standard error by $\sqrt{\frac{\hat{\theta}^2}{n(\hat{\theta}-1)^4}} \approx 0.0875$.

Example: Pareto distribution

Another statistic that could be used to estimate the mean income can be the sample mean $\bar{X}$. Then, for $\theta > 2$, by the CLT

$$\frac{\bar{X} - \frac{\theta}{\theta-1}}{\frac{1}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0, \mathbb{V}\mathrm{ar}(X_1)) = \mathcal{N}\Big(0, \frac{\theta}{(\theta-1)^2(\theta-2)}\Big).$$

Since $\frac{\theta}{\theta-2} = 1 + \frac{2}{\theta-2} > 1$,

$$\mathbb{V}\mathrm{ar}(\bar{X}) > \mathbb{V}\mathrm{ar}(g(\hat{\theta})) = \frac{\theta^2}{n(\theta-1)^4}.$$

That is, if the Pareto model for income is correct, then the estimate $g(\hat{\theta}) = \frac{\hat{\theta}}{\hat{\theta}-1}$ is more accurate for the mean income than the sample mean $\bar{X}$. The reason for this is because the Pareto distribution is heavy-tailed, and the sample mean $\bar{X}$ is heavily influenced by rare but extremely large data values. On the other hand, the MLE $\hat{\theta}$ is computed through the values $\log X_i$ and is not as heavily influenced by extremely large data values as $\bar{X}$ is.

General Information inequality

Theorem

(**Cramér-Rao Inequality**) For a parametric model $\{f(x|\theta) : \theta \in \Theta\}$ satisfying certain regularity assumptions, where θ is a single parameter, let g be any function differentiable on all of Θ, and let T be any unbiased estimator of $g(\theta)$ based on a random sample $X_1, \ldots, X_n$ from a population with pdf $f(x|\theta)$. Then,

$$\mathbb{V}\mathrm{ar}_\theta(T) \geqslant \frac{g'(\theta)^2}{nI(\theta)}. \tag{3}$$

Proof.

The proof is the same as the proof of the simple Cramér-Rao bound. The equation $\theta = \mathbb{E}_\theta(T)$ should be replaced by $g(\theta) = \mathbb{E}_\theta(T)$. After differentiation with respect to θ, we obtain $g'(\theta) = \mathbb{E}_\theta(TZ) = \mathbb{C}\mathrm{ov}_\theta(T, Z)$. The rest remains the same. □

HW

1. Let $X_1, \ldots, X_n$ be i.i.d. from uniform $(0, \theta)$ distribution, where $\theta > 0$ is unknown parameter. Consider

$$\hat{\theta} = \frac{n+1}{n} \max_{1 \leqslant i \leqslant n} X_i.$$

   Check if $\hat{\theta}$ is unbiased for θ. Is it consistent? Show that the Rao-Cramér inequality does not work in this example, i.e. show that $\mathbb{V}\mathrm{ar}_\theta(\hat{\theta}) \leqslant \frac{1}{nI(\theta)}$. The reason for this is that the support of the uniform distribution depends on θ, i.e. the assumption of common support is violated.

2. Let $X_1, \ldots, X_n$ be a random sample from a population with Rayleigh distribution

$$f(x|\theta) = \begin{cases} \frac{x}{\theta^2} e^{-\frac{x^2}{2\theta^2}}, & x \geqslant 0 \\ 0, & x < 0 \end{cases}$$

   with unknown parameter $\theta > 0$. Find the method of moments estimator of θ.
   Find the MLE of θ and its asymptotic variance.

HW

1. Let $X_1, \ldots, X_n$ be a random sample from a population with pdf

$$f(x|\theta) = \begin{cases} (\theta+1)\, x^{\theta} & 0 \leqslant x \leqslant 1 \\ 0, & \text{otherwise} \end{cases}.$$

with unknown parameter $\theta > 0$. Find the method of moments estimate of θ.
Find the MLE of θ and the asymptotic variance of the MLE.

2. Let $X_1, \ldots, X_n$ be a random sample from a population with an exponential distribution $\exp(\frac{1}{\tau})$, i.e. with pdf

$$f(x|\theta) = \begin{cases} \frac{1}{\tau} e^{-\frac{x}{\tau}} & x \geqslant 0 \\ 0, & \text{otherwise} \end{cases}.$$

and unknown scale parameter $\tau > 0$.

   1. Find the MLE of τ. What is the exact sampling distribution of the MLE? Use the central limit theorem to find a normal approximation to the sampling distribution.
   2. Show that the MLE is unbiased and find its exact variance. Is there any other unbiased estimate with smaller variance? *Hint:* The sum of the X_i follows a gamma distribution.

Exponential Family

A family of pdfs or pmfs is called an *exponential family of distributions* if it can be represented in the form

$$f(x|\theta) = h(x)\, c(\theta)\; e^{\sum_{i=1}^{k} w_i(\theta)\, t_i(x)}. \tag{4}$$

where

- $h(x) \geqslant 0$ and $t_1(x), \ldots, t_k(x)$ are real-valued functions of the observation x (they cannot depend on θ)
- $c(\theta) \geqslant 0$ and $w_1(\theta), \ldots, w_k(\theta)$ are real-valued functions of the possibly vector-valued parameter θ (they cannot depend on x).

Exponential Family

Alternatively, a statistical model is called an exponential family if the log likelihood has the form

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{k} w_i(\boldsymbol{\theta})\, t_i(\mathbf{x}) - a(\boldsymbol{\theta}) + u(\mathbf{x})$$

and the last term can be dropped.

Natural Statistic and Natural Parameters

Let

$$y_i = t_i(\mathbf{x})$$

and

$$\psi_i = w_i(\theta)$$

The log-likelihood in terms of natural parameters and statistics has the simple form

$$\ell(\boldsymbol{\psi}) = \mathbf{y}^T\boldsymbol{\psi} - b(\boldsymbol{\psi})$$

with

$$\nabla\ell(\boldsymbol{\psi}) = \mathbf{y} - \nabla b(\boldsymbol{\psi})$$
$$\nabla^2\ell(\boldsymbol{\psi}) = -\nabla^2 b(\boldsymbol{\psi})$$

MLE

The log likelihood derivative identities give

$$\mathbb{E}_{\boldsymbol{\psi}}(\mathbf{Y}) = \nabla b(\boldsymbol{\psi})$$
$$\mathbb{V}\mathrm{ar}_{\boldsymbol{\psi}}(\mathbf{Y}) = -\nabla^2 b(\boldsymbol{\psi})$$

- The MLE is a method of moments estimator that sets the observed value of the natural statistic vector equal to its expected value.
- The second derivative of the log likelihood is always nonrandom, so observed and expected Fisher information for the natural parameter vector are the same, and
- the log likelihood for the natural parameter is always concave and strictly concave unless the distribution of the natural statistic is degenerate.
- Hence any local maximizer of the log likelihood is the unique global maximizer.

Exponential Family Distributions

- By invariance of maximum likelihood, the property that any local maximizer is the unique global maximizer holds for any parametrization.

- Distributions that are exponential families:
  1. Bernoulli,
  2. binomial,
  3. Poisson,
  4. geometric,
  5. negative binomial (p unknown, r known),
  6. normal,
  7. exponential,
  8. gamma,
  9. beta,
  10. multinomial,
  11. multivariate normal.

Example: Binomial

Let n be a positive integer and consider $X \sim bin(n,p)$ with $0 < p < 1$.

$$\begin{aligned} P(X=x) &= \binom{n}{x} e^{\log\left(p^x(1-p)^{n-x}\right)} = \binom{n}{x} e^{x(\log p - \log(1-p)) + n\log(1-p)} \\ &= \binom{k}{x} \cdot (1-p)^n e^{x \log \frac{p}{1-p}} \\ &= h(x) \cdot c(p)\, e^{w_1(p)\, t_1(x)}, \end{aligned}$$

where we set

$$\begin{aligned} h(x) &= \begin{cases} \binom{n}{x}, & x = 0,1,\ldots,n \\ 0, & \text{otherwise} \end{cases} \\ c(p) &= (1-p)^n, \quad 0 < p < 1, \\ w_1(p) &= \log \frac{p}{1-p} \\ t_1(x) &= x \end{aligned}$$

If $n = 1$, $bin(1,p) = \text{Bernoulli}(p)$, for $0 < p < 1$, and thus we conclude that Bernoulli distribution belongs to the exponential family.

Example: Binomial natural form

The log likelihood is

$$\ell_n(p) = x \log(p) + (n - x) \log(1 - p) = x \left(\log(p) - \log(1 - p) \right) + n \log(1 - p)$$

hence it is an exponential family with natural statistic $y = x$ and natural parameter

$$\psi = \operatorname{logit}(p) = \log(p) - \log(1 - p) = \log \frac{p}{1 - p}$$

Suppose $X_1, \ldots, X_n$ are iid from an exponential family distribution with log likelihood for sample size one

$$\ell_1(\theta) = \sum_{i=1}^{k} w_i(\theta)\, t_i(\mathbf{x}) - a(\theta)$$

Then the log likelihood for sample size n is

$$\ell_n(\theta) = \sum_{i=1}^{k} \left(\sum_{j=1}^{n} t_i(\mathbf{x}_j) \right) w_i(\theta) \; - n \, a(\theta)$$

hence the sample size n distribution is also an exponential family with the same natural parameter vector as for sample size one and natural statistic vector y with components

$$y_i = \sum_{j=1}^{n} t_i(\mathbf{x}_j)$$

Sum of Bernoulli random variables

Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli p distribution, $0 < p < 1$.

We showed that a Bernoulli (p) distribution belongs to the exponential family (4) with $k = 1$, $c(p) = 1 - p$, $w_1(p) = \log \frac{p}{1-p}$ and $t_1(x) = x$.

Then,

$$t_1(X_1, \ldots, X_n) = X_1 + \cdots + X_n.$$

We know that t_1 has a binomial distribution $t_1 \sim bin(n, p)$, which also belongs to the exponential family with the same w_1 and $c(p) = (1 - p)^n$.

Two-parameter normal

For the two-parameter normal, the log likelihood for sample size one is

$$\begin{aligned}
\ell_1(\mu, \sigma^2) &= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x - \mu)^2 \\
&= -\frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2) \\
&= -\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2}\log(\sigma^2) - \frac{\mu^2}{2\sigma^2}
\end{aligned}$$

Two-parameter normal

Since this is a two-parameter family, the natural parameter and statistic must also be two dimensional.
Let

$$\mathbf{y} = \begin{pmatrix} x \\ x^2 \end{pmatrix}$$

$$\boldsymbol{\psi} = \begin{pmatrix} \mu/\sigma^2 \\ -1/(2\sigma^2) \end{pmatrix} = \begin{pmatrix} \psi_1 \\ \psi_2 \end{pmatrix}$$

$$b(\boldsymbol{\psi}) = \frac{1}{2}\log(\sigma^2) + \frac{\mu^2}{2\sigma^2}$$

For sample size n, the natural statistics are

$$y_1 = \sum_{i=1}^{n} x_i$$

$$y_2 = \sum_{i=1}^{n} x_i^2$$

Two-parameter normal

$$\mu = -\frac{\psi_1}{2\psi_2}$$
$$\sigma^2 = -\frac{1}{2\psi_2}$$
$$b(\boldsymbol{\psi}) = \frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}$$
$$= \frac{1}{2}\log(-\frac{1}{2\psi 2}) - \frac{\psi_1^2}{4\psi_2}$$

$$\frac{\partial b(\boldsymbol{\psi})}{\partial \psi_1} = -\frac{\psi_1}{2\psi_2} = \mu$$
$$\frac{\partial b(\boldsymbol{\psi})}{\partial \psi_2} = -\frac{1}{2\psi_2} + \frac{\psi_1^2}{4\psi_2^2} = \sigma^2 + \mu^2$$

Two-parameter normal

Since for a random sample of size n, $\mathbb{E}_{\boldsymbol{\psi}}(\mathbf{Y}) = n\nabla b(\boldsymbol{\psi})$,

$$\mathbb{E}(Y_1) = \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = n\frac{\partial b(\boldsymbol{\psi})}{\partial \psi_1} = n\mu$$

$$\mathbb{E}(Y_2) = \mathbb{E}\left(\sum_{i=1}^{n} x_i^2\right) = \frac{\partial b(\boldsymbol{\psi})}{\partial \psi_2} = n(\sigma^2 + \mu^2)$$

as expected since X_i are iid $\mathcal{N}(\mu, \sigma^2)$.