Example 6.12 (i.i.d restricted normal distribution)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from $N(\mu, 1)$ with $0 \le \mu < \infty$. The log likelihood is ← statistical modeling

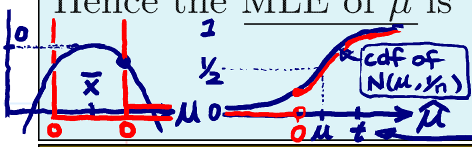
$\boxed{\text{from the } \ell(\mu,\sigma) \text{ in LNp 20}, \sigma = 1}$

$$l(\mu) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(X_i \underbrace{- \mu}_{-\overline{X}+\overline{X}})^2$$

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

$\boxed{\text{sampling distribution=? Note. } \overline{X} \sim N(\mu, 1/n)}$

Hence the MLE of μ is always falls in $(0,\infty)$ (Why?)



$$\hat{\mu} = \begin{cases} \overline{X}, & \text{if } \overline{X} \ge 0 \\ 0, & \text{if } \overline{X} < 0 \end{cases}$$

← $\boxed{\text{cf.}}$ → $\boxed{\text{moment estimator } \hat{\mu} = \overline{X} \text{ (Ec)}}$

reasonable?

$P(\hat{\mu} \le t) = P(\overline{X} \le t)$ if $t > 0$.

$\boxed{3/13}$

Example 6.13 (i.i.d uniform(0, θ) distribution)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $U(0, \theta)$, where $\theta > 0$. Then the likelihood of θ is ← statistical modeling

$X_{(n)}$

0 $\frac{\theta}{2}$ θ

$$\mathcal{L}(\theta) = \begin{cases} \theta^{-n}, & \text{if } 0 \le X_i \le \theta, i = 1, \ldots, n \\ 0, & \text{otherwise} \end{cases}$$

$$= \begin{cases} \theta^{-n}, & \text{if } \theta \ge \max_{1 \le i \le n} X_i = X_{(n)} \ge 0 \\ 0, & \text{otherwise} \end{cases}$$

$\boxed{\text{Q: Are } X_{(n)}, 2\overline{X} \text{ reasonable estimators?}}$

$\boxed{\text{c.f. moment estimator } \hat{\theta} = 2\overline{X} \text{ (Ec)} \Rightarrow \text{unbiased} \Rightarrow \text{but } 2\overline{X} \text{ might be} < X_{(n)}}$

$\boxed{\text{sampling distribution=?}}$

$\boxed{\text{alternative estimator } \frac{n+1}{n}X_{(n)} = (1+\frac{1}{n})X_{(n)}}$

Because $\mathcal{L}(\theta)$ decreases when θ increases, the MLE of θ is $X_{(n)}$.

① $E(X_{(n)}) \underset{\text{(Ec)}}{=} \frac{n}{n+1}\theta$ ⇐ biased ② always underestimate

Example 6.14 (multinomial distribution, TBp. 272)

Suppose $X_1, X_2, \ldots, X_m$ are counts in cells $1, 2, \ldots, m$ and follow a multinomial distribution with total count n and cell probabilities $p_1, p_2, \ldots, p_m$ ($p_i \ge 0$ for $i = 1, 2, \ldots, m$, and $p_1 + p_2 + \cdots + p_m = 1$). The joint pmf of $X_1, X_2, \ldots, X_m$ is

$\boxed{\text{statistical modeling}}$ → $X_i \sim \text{binomial}(n, p_i)$, but $X_1, \ldots, X_m$ not independent

$$f(x_1, x_2, \ldots, x_m | p_1, p_2, \ldots, p_m) = \frac{n!}{\prod_{i=1}^{m} x_i!}\prod_{i=1}^{m} p_i^{x_i},$$

$\boxed{\text{dimension of parameter space} = m-1}$

where $x_1 + x_2 + \cdots + x_m = n$. For n given, the log likelihood is

$$l(p_1, p_2, \ldots, p_m) = \log n! - \sum_{i=1}^{m}\log X_i! + \sum_{i=1}^{m} X_i \log p_i.$$

$\boxed{\text{MLE}}$ ⇒ maximize $l(p_1, p_2, \ldots, p_m)$ subject to $\sum_{i=1}^{m} p_i = 1$.

Introduce a Lagrange multiplier λ, and maximize

$$l(p_1, p_2, \ldots, p_m, \lambda) \equiv \underbrace{\log n! - \sum_{i=1}^{m}\log X_i! + \sum_{i=1}^{m} X_i \log p_i}_{l(p_1,\cdots,p_m)} + \lambda\underbrace{\left(\sum_{i=1}^{m} p_i - 1\right)}.$$

Setting $\frac{\partial l}{\partial p_i} = \frac{X_i}{p_i} + \lambda = 0$, $i = 1, 2, \ldots, m$ gives

$$\hat{p}_j = -\frac{X_j}{\lambda}, \quad j = 1, 2, \ldots, m.$$

$\boxed{\frac{\partial l}{\partial \lambda} = \sum_{i=1}^{m} p_i - 1 = 0}$

Since

$$1 = \sum_{i=1}^{m} \hat{p}_i = \sum_{i=1}^{m} -\frac{X_i}{\lambda}, \quad 1 = -\frac{n}{\lambda}$$

$\boxed{\text{Sampling distribution} = ?}$

we have $\lambda = -n$. Hence, $\hat{p}_i = X_i/n$, $i = 1, 2, \ldots, m$. reasonable?

Example 6.15 (Hardy-Weinberg Equilibrium, TBp. 273)

Hardy-Weinberg law: if gene frequencies are in equilibrium, the genotypes AA, Aa, and aa occur in a population with frequencies $(1-\theta)^2, 2\theta(1-\theta)$, and θ^2. ← $\theta \in (0,1)$ & $(1-\theta)^2 + 2\theta(1-\theta) + \theta^2 = 1$

prob. of getting A
$1\cdot(1-\theta^2) + \frac{1}{2}(2\theta(1-\theta)) + 0\cdot\theta^2 = 1-\theta$
prob. of getting a
$0\cdot(1-\theta^2) + \frac{1}{2}(2\theta(1-\theta)) + 1\cdot\theta^2 = \theta$

		Mother	
		A $[1-\theta]$	a $[\theta]$
Father	A $[1-\theta]$	AA $[(1-\theta)^2]$	$\boxed{Aa\ [\theta(1-\theta)]}$
	a $[\theta]$	$\boxed{Aa\ [\theta(1-\theta)]}$	aa $[\theta^2]$

Question: If we sample n (a fixed number) persons from the population, and let X_1, X_2, and X_3 (random variables) denote the counts in the three cells (AA, Aa, aa), what is a suitable statistical model (i.e., joint distribution) for (X_1, X_2, X_3)? Notice that $n = X_1 + X_2 + X_3$.

large population, without replacement ≈ with replacement

$\boxed{\text{statistical modeling}}$ → $(X_1, X_2, X_3) \sim \text{multinomial}(n, p_1, p_2, p_3)$ $\boxed{P_1+P_2+P_3=1,\ dim=2}$

→ $X_i \sim \text{binomial}(n, p_i)$
but X_1, X_2, X_3 not independent

$(1-\theta)^2, 2\theta(1-\theta), \theta^2$ $\boxed{\text{∵ genetics knowledge}}$ $\boxed{dim=1}$

Log likelihood of θ is

$\boxed{\text{from the log-likelihood in LNp.23}}$

$$\begin{aligned} l(\theta) &= \log n! - \sum_{i=1}^{3} \log X_i! \\ &\quad + X_1 \log (1-\theta)^2 + X_2 \log [2\theta(1-\theta)] + X_3 \log \theta^2 \\ &= \log n! - \sum_{i=1}^{3} \log X_i! \\ &\quad + (2X_1 + X_2) \log (1-\theta) + (2X_3 + X_2) \log \theta + X_2 \log 2. \end{aligned}$$

Setting $\frac{d}{d\theta} l(\theta) = 0$,

$$-\frac{2X_1 + X_2}{1-\theta} + \frac{2X_3 + X_2}{\theta} = 0$$

$\boxed{\text{sampling distribution} = ?}$

yields the MLE of θ reasonable?

$$\hat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3} = \frac{2X_3 + X_2}{2n}. = \hat{P}_3 + \frac{1}{2}\hat{P}_2$$

$\theta^2 + \frac{1}{2}(2\theta(1-\theta)) = \theta$

$\boxed{\text{can be written as a function of } X_1 \text{ and } X_2}$

Question: What is the difference between Ex. 6.14 and Ex. 6.15?

Chinese population data of Hong Kong in 1937: (M, N are erythrocyte antigens)

Assume Hardy-Weinsberg Equilibrium?

statistical modeling = ?

Blood Type	M	MN	N	Total
Frequency	342	500	187	$\boxed{1029}$
(by Ex.6.14)	0.333	0.486	0.182	

$\boxed{\text{cf.}}$ $\boxed{\text{dim}=2}$

$\overset{\shortparallel}{342/1029}$

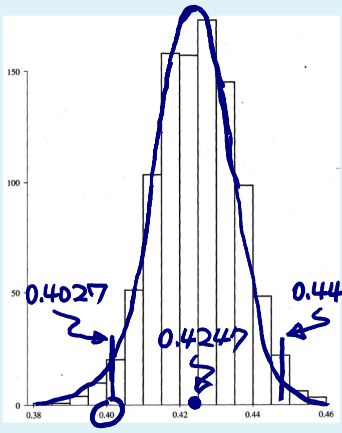
1. MLE is $\hat{\theta} = 0.4247 \Rightarrow (\hat{p}_1, \hat{p}_2, \hat{p}_3) = (0.331, 0.489, 0.180)$. $\boxed{\text{dim}=1}$
2. Approximate the sampling distribution of $\hat{\theta}$ by bootstrap: $\left\{\begin{array}{l}\text{exact method}\\ \text{asymptotic method}\end{array}\right.$

   simulation method $\boxed{\text{cf}}$

   - Generate 1000 random counts from multinomial with $n = 1029$ and cell probabilities 0.331, 0.489, and 0.180.
   - From each of the 1000 experiments, a MLE value $\hat{\theta}^*$ was determined.

   From the histogram of the 1000 estimates (**Figure 8.7 in Textbook**), which should approximate the sampling distribution of $\hat{\theta}$.

   - Looks like Normal.
   - Standard deviation of the 1000 values gives estimated standard error of $\hat{\theta}$: $s_{\hat{\theta}} = 0.011$. $0.4247 \pm 2\times(0.011) = [0.4027, 0.4467]$



Example 6.16 (Muon Decay, TBp. 266 & 271)

- Let Θ be the angle at which electrons are emitted in muon decay.
- Let $X = \cos(\Theta)$. It has a distribution with pdf

$\boxed{\text{data}}$

$$f(x|\alpha) = \frac{1}{2}(1+\alpha x), \quad -1 \le x \le 1, -1 \le \alpha \le 1.$$

$\boxed{a \text{ pdf ?}}$ parameter

- The mean of X is $\mu = \alpha/3 \Rightarrow \alpha = 3\mu$. $\boxed{\text{Ec}}$

  1st moment
- The moments estimator of α based on a sample $X_1, \cdots, X_n$ is $\hat{\alpha} = 3\overline{X}$. $\boxed{\text{i.i.d.} \sim f(x|\alpha)}$ $\boxed{\text{unbiased}}$
- The log likelihood of α is

$$l(\alpha) = \sum_{i=1}^{n} \log(1+\alpha X_i) - n\log 2.$$

Setting the derivative equal to zero, the MLE of α satisfies the *nonlinear equation*

asymptotic method (LNp.39)

$\boxed{\text{sampling distribution} = ?}$

simulation method

$$0 = \frac{d}{d\alpha} l(\alpha) = \sum_{i=1}^{n} \frac{X_i}{1+\alpha X_i}.$$

$\boxed{\text{Dahlquist \& Björck (1974) Chapter 6}}$

- The MLE of α has **no easy close-form solution**.

  $\Rightarrow$ can use an iterative method to numerically solve for MLE.

  $\Rightarrow$ method of moments estimate could be used as a starting value.

Example 6.17 (i.i.d. Gamma distribution, TBp. 270)

- Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $\Gamma(\alpha, \lambda)$. The joint pdf is

statistical modeling

$$f(x_1, x_2, \ldots, x_n | \alpha, \lambda) = \prod_{i=1}^n \frac{1}{\Gamma(\alpha)} \lambda^\alpha x_i^{\alpha-1} e^{-\lambda x_i}$$

- The log likelihood is

$$\begin{aligned} l(\alpha, \lambda) &= \sum_{i=1}^n [\alpha \log \lambda + (\alpha - 1) \log X_i - \lambda X_i - \log \Gamma(\alpha)] \\ &= n\alpha \log \lambda + (\alpha - 1) \sum_{i=1}^n \log X_i - \lambda \sum_{i=1}^n X_i - n \log \Gamma(\alpha). \end{aligned}$$

- Setting

$$\begin{cases} 0 = \frac{\partial l}{\partial \alpha} = n \log \lambda + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \\ 0 = \frac{\partial l}{\partial \lambda} = \frac{n\alpha}{\lambda} - \sum_{i=1}^n X_i \end{cases}$$

- The MLE then satisfies

$(\hat{\alpha}, \hat{\lambda})$

$$\begin{cases} \hat{\lambda} = \hat{\alpha}/\overline{X} \\ n \log \hat{\alpha} - n \log \overline{X} + \sum_{i=1}^n \log X_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0 \end{cases}$$

- 2nd part is a nonlinear equation $\Rightarrow$ **no easy closed-form solution**.
  $\Rightarrow$ can use iterative method to find (approximate) the solution
  $\Rightarrow$ method of moments estimates can be used as initial value.

LNp.14

- **Rainfall amount data** (Ex 6.7-6.8, LNp.14-16):

  1. Take the initial value as the method of moments estimates

  $$\hat{\alpha} = 0.375, \quad \hat{\lambda} = 1.674.$$

  Q: Which estimate is closer to the true values of parameters?

  By an iterative procedure, the MLE's are computed:

  c.f.

  $$\hat{\alpha} = 0.441, \quad \hat{\lambda} = 1.96$$

  asymptotic method (LNp.43)

  $\Rightarrow$ of little practical difference from the moment estimates.

  check histograms in LNp.16

  2. Exact sampling distribution of the MLE is intractable $\Rightarrow$ can use simulation to approximate:

  ∵ no close form

  – Generate many, say 1000, samples of size 227 from Gamma with $\alpha = 0.441$, $\lambda = 1.96$.
  – Form MLE of α, λ for each sample.
  – Construct histogram of the 1000 MLE's.

  3. From the histograms of the simulated MLEs:

  – The histograms look like normal.

$\hat{\alpha}$ might be biased

- The histograms are centered at $\hat{\alpha} = 0.471$ and $\hat{\lambda} = 1.97$.

The histograms in LNp.16 cf.

  - Estimated standard error of the MLE's are

moment estimator $S_{\hat{\alpha}} = 0.06$, $S_{\hat{\lambda}} = 0.34$ (LNp.16) c.f.

$$s_{\hat{\alpha}} = 0.04, \quad s_{\hat{\lambda}} = 0.28.$$

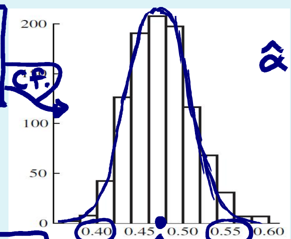
  - Sampling distribution of MLE's are less dispersed than those of the method of moments estimates.

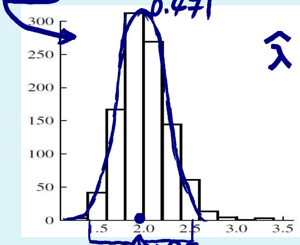
Q: Which estimator will you prefer?

$0.471 \pm 2\times(0.04) = [0.391, 0.551]$

$1.97 \pm 2\times(0.28) = [1.41, 2.53]$

cf.





Summary (advantages of MLE)

1. easy to interpret ← LNp.18, definition
2. widely applicable Why?
3. the range of the MLE coincides with the range of the parameter ← check Ex.6.12, 6.13 (LNp.22)
4. invariance under reparameterizations ← Thm 6.1 (LNp.19)
5. nice theoretical properties ← e.g. asymptotic properties (Thm 6.5 ~ 6.9, LNp 38 ~ 43)

❖ **Reading**: textbook, 8.5, 8.5.1

Large sample (asymptotic) theory for method of moment estimator and MLE

more data, more information, less random

Recall: 1. Law of Large Number
2. Central Limit Theorem → for sum/average ← estimator $\hat{\theta}$

Definition 6.10 (consistent, TBp. 266)

Let $\hat{\theta}_n$ be an estimator of a parameter θ based on a sample of size n. Then $\hat{\theta}_n$ is called **consistent in probability** if $\hat{\theta}_n$ converges in probability to θ as n tends to infinity, i.e. for any $\epsilon > 0$,

$$P\left(|\hat{\theta}_n - \theta| > \epsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty.$$

sample size is large

LLN cf. ↳ for $\bar{X}_n$

$\hat{\theta}_n \xrightarrow{P} \theta$ ⇕ $\hat{\theta}_n \xrightarrow{d} \theta$ ∵ θ is a constant

➤ method of moment estimator

Theorem 6.2 (consistency of method of moment estimator, TBp. 266)

The weak law of large numbers implies that

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \equiv \hat{\mu}_k \longrightarrow \mu_k \text{ in probability as } n \to \infty.$$

FYI, $\dfrac{\sqrt{n}(\hat{\mu}_k - \mu_k)}{\sqrt{\mu_{2k} - \mu_k^2}} \xrightarrow{d} N(0,1)$

the g_i's in LNp.9

Y_i → $E(Y_i) = \mu_k$

If the function relating μ_k and θ_j are continuous, method of moments estimators are consistent.

By Thm 5.1, item 4, LN, Ch1-6, p.89

Theorem 6.3 (justification for estimating standard errors, TBp. 266-267)

- Recall: In LNp.12, $\sigma_{\hat\theta}(\hat\theta) \xrightarrow{\text{estimate}} \sigma_{\hat\theta}(\theta)$ 代入法 standard error (as a function of θ) of sampling distribution of $\hat\theta$: usually $\sigma_{\hat\theta}(\theta) \to 0$ as $n \to \infty$
  estimated standard error
- Consider the standard error of the form: $\sigma_{\hat\theta}(\theta) = \frac{1}{\sqrt{n}}\sigma^*(\theta)$ not a linear function of n, $O(n^{-1/2})$, irrelevant to n
  - Ex. 6.5 (LNp.11): $\sigma_{\hat\lambda}(\lambda) = \frac{1}{\sqrt{n}}\sqrt{\lambda}$ iid ~ $P(\lambda)$
  - Ex. 6.6 (LNp.13): $\sigma_{\hat\mu}(\mu,\sigma) = \frac{1}{\sqrt{n}}\sigma$, and $\sigma_{\hat\sigma^2}(\mu,\sigma) \approx \frac{1}{\sqrt{n}}\sqrt{2}\sigma^2$ iid ~ $N(\mu,\sigma^2)$
- Let θ_0 = true parameter, and $\sigma_{\hat\theta} \equiv \sigma_{\hat\theta}(\theta_0) = \frac{1}{\sqrt{n}}\sigma^*(\theta_0)$. true standard error of sampling distribution; a fixed value
- Estimate $\sigma_{\hat\theta}$ by a r.v. $s_{\hat\theta} \equiv \frac{1}{\sqrt{n}}\sigma^*(\hat\theta)$. $\sigma^*(\cdot)$
- If (1) $\sigma^*(\theta)$ is continuous in θ, and (2) $\hat\theta$ is consistent $\left(\hat\theta \xrightarrow{P} \theta_0\right)$, then e.g., moment estimator, MLE.

  By Thm 5.1, item 4, LN, Ch1-6, p. 89 $\sigma^*(\hat\theta) \xrightarrow{P} \sigma^*(\theta_0)$, i.e.,

  as $n \to \infty$, $\frac{s_{\hat\theta}}{\sigma_{\hat\theta}} \to 1$ in probability $\left(\Rightarrow s_{\hat\theta} \xrightarrow{p} \sigma_{\hat\theta} \to 0\right)$

  Why not say $s_{\hat\theta} \xrightarrow{P} \sigma_{\hat\theta}$? $= \sigma^*(\hat\theta)/\sigma^*(\theta_0)$ same convergence rate when n is large

➢ MLE

Note. the following discussion is mainly for (1) the case of i.i.d. sample, and (2) one-dimensional parameter.

Definition 6.11 (score equation, score function)

- The log likelihood of an i.i.d. sample of size n from a pdf/pmf $f(x|\theta)$ is
  $$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$
  score function of n obs. = sum of score function of ith obs

  (Exercise) Check the previous MLE examples and identify the score equation & score function.
- The MLE maximizes $l(\theta)$ and is usually obtained by solving the **score equation**
  $$0 = \frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta).$$
  sum of scores = 0 at $\hat\theta_{MLE}$ Score of ith observation X_i
- $\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$ is called **score function**. a function of observations X_i's and parameter θ

3/18

Definition 6.12 (Fisher information for one-dimensional parameter, TBp.263)

Let $X_1, \ldots, X_n$ be a sample of size n with a joint pdf/pmf f. Define
$$I_{X_1,\ldots,X_n}(\theta) = E_\theta\left[\frac{\partial}{\partial\theta} \log f(X_1, \cdots, X_n|\theta)\right]^2$$
This is a function of unknown parameter θ; can allow not to be i.i.d.; average (w.r.t. the parameter θ); $\partial/\partial\theta \log f(\underset{\sim}{X}|\theta)$ is a r.v., but not statistic; treated as random variable; weighted average of score²

which is called the **(Fisher) information** of θ contained in $X_1, \ldots, X_n$.

Question: What *information* does $I_{X_1,\cdots,X_n}(\theta)$ offer?