# 点估计

## Point Estimation



- What is point estimation?

**Example 6.1** (current across muscle cell membrane, TBp. 257-258)

- Bevan, Kullberg, and Rice (1979) studied random fluctuations of current across a muscle cell membrane. The cell membrane contained a large number of channels, which opened and closed at random and were assumed to operate independently. The net current resulted from ions flowing through open channels.
- They obtained 49,152 observations of the net current, $x_1, \ldots, x_{49152}$.
- The net current was the sum of a large number of roughly independent small currents.
- It seems appropriate to model the net current data, $X_1, \ldots, X_{49152}$ as i.i.d. $N(\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ represent the mean and variance of net current. Note that the values of $\mu$ and $\sigma^2$ are unknown.
- Question: how to use the observed data, $x_1, \ldots, x_{49152}$ to gain knowledge about the values of $\mu$ and $\sigma^2$?





**Example 6.2** (emission of alpha particles, TBp. 255-256)

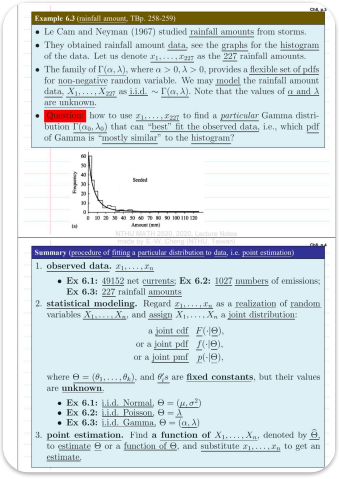
- Berkson (1966) conducted an experiment about emission of alpha particles from radioactive sources. The number of emissions per unit of time is not constant but fluctuates in a random fashion.
- The experimenter recorded 10,220 times between successive emissions. The numbers of emissions, $x_i$, $i = 1, \ldots, 1027$, observed in 1207 time intervals, each of length 10 sec, are summarized in the following table:

| $x_i \in \{0, 1, 2\}$ | $x_i = 3$ | $x_i = 4$ | $\cdots$ |
|---|---|---|---|
| 18 | 28 | 56 | $\cdots$ |

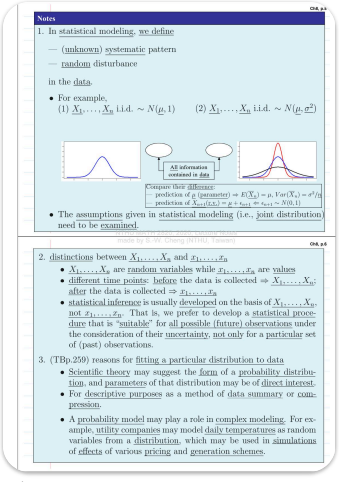
e.g., in 28 of the 1207 intervals, there were 3 counts, etc.

- Assume (1) the underlying rate of emission is constant over the period of observation (2) the particles come from a very large number of independent sources
- It seems appropriate to model the numbers of emissions $X_1, \ldots, X_{1027}$ as i.i.d. $P(\lambda)$, where $\lambda$ represents the underlying rate of emission. Note that the value of $\lambda$ is unknown.
- Question: how to use the observed data, $x_1, \ldots, x_{1207}$, to gain knowledge about the value of $\lambda$?

- $x_1$是随机变量$X_1$所吐出来的值,想知道joint distribution,不过现在已经知道iid条件,可以通过计算marginal distribution找joint distribution
- 先通过已经显性的数据找背后的随机变量,之后统合随机变量去找joint distribution
- distribution都是确定的,以正态分布为例子$\mu$和$\sigma^2$也是固定的,也就是说随机的数据中携带着确定的数据
- 如何利用一堆数据,做transformation,来找到关于参数的信息
- poisson的原因是,因为理论上是很多粒子决定是不是自己会放射,也就是说自己是一个bernoulli,众多伯努利构成二次,正好有n极大,p很小.poisson正好是只有就是对n->正无穷,p趋近0,np->lambda的情况(发生or不发生两种)
- $\lambda$就是单位时间内发生次数
- 考虑的问题就是通过什么transformation能够

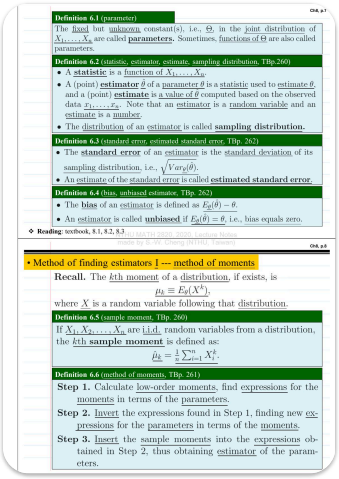



- 直方图->数据变为正无穷,interval分的越来越细,会变为pmf和pdf
- 直方图带着什么信息?
- 三种统计模型有什么不同,前两个是conceptual,最后一个empirical
- 已知是正态分布,已知应该是poisson分布(根据专业知识);不知道雨量应该怎么建模,不知道数据,也不知道怎么建模,要去找到最fit的model
- 数据背后都隐藏了一个随机变量(吐出这些数值),相当于恢复数据的随机性
- 要找的是一个joint distribution
- 这些distribution是一个distribution族,因为参数组有很多可能得数值,所以$\Theta$属于一个参数空间
- 做各种transformation找到各种各样的信息,做一个估计值$\hat{\Theta}$,(其是一个随机变量,因为从理论上这是一堆随机变量X1到Xn的函数,显然也是随机变量)得到参数的估计值




- 统计都是先建模,

- 好分析师"万变不离其宗",拨开随机找系统
- 把规律部分和随机部分分开
- 第二种model的规律性更多,因为误差部分吸带着sigma的信息
- 有很多假设的部分(在建模过程中),要一直检查
- 不是对一组数据,而是对于一系列的随机变量
- 估计:为数据找pdf,pmf
- 理解为去确定自己在哪个平行世界



**Definition 6.1** (parameter)
The fixed but unknown constant(s), i.e., $\Theta$, in the joint distribution of $X_1, \ldots, X_n$ are called **parameters**. Sometimes, functions of $\Theta$ are also called parameters.

**Definition 6.2** (statistic, estimator, estimate, sampling distribution, TBp.260)
- A **statistic** is a function of $X_1, \ldots, X_n$.
- A (point) **estimator** $\hat{\theta}$ of a parameter $\theta$ is a statistic used to estimate $\theta$, and a (point) **estimate** is a value of $\hat{\theta}$ computed based on the observed data $x_1, \ldots, x_n$. Note that an estimator is a random variable and an estimate is a number.
- The distribution of an estimator is called **sampling distribution.**

**Definition 6.3** (standard error, estimated standard error, TBp. 262)
- The **standard error** of an estimator is the standard deviation of its sampling distribution, i.e., $\sqrt{Var_\theta(\hat{\theta})}$.
- An estimate of the standard error is called **estimated standard error.**

**Definition 6.4** (bias, unbiased estimator, TBp. 262)
- The **bias** of an estimator is defined as $E_\theta(\hat{\theta}) - \theta$.
- An estimator is called **unbiased** if $E_\theta(\hat{\theta}) = \theta$, i.e., bias equals zero.

❖ **Reading:** textbook, 8.1, 8.2, 8.3

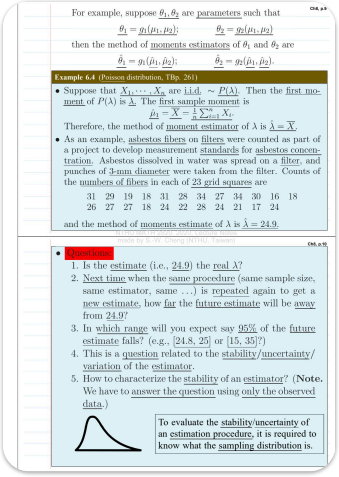
• Method of finding estimators I --- method of moments

**Recall.** The kth moment of a distribution, if exists, is
$$\mu_k \equiv E_\theta(X^k),$$
where $X$ is a random variable following that distribution.

**Definition 6.5** (sample moment, TBp. 260)
If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables from a distribution, the **kth sample moment** is defined as:
$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

**Definition 6.6** (method of moments, TBp. 261)
**Step 1.** Calculate low-order moments, find expressions for the moments in terms of the parameters.
**Step 2.** Invert the expressions found in Step 1, finding new expressions for the parameters in terms of the moments.
**Step 3.** Insert the sample moments into the expressions obtained in Step 2, thus obtaining estimator of the parameters.

- estimator就是一个统计量,是随机变量.是X1到Xn的一个transformation
- estimate是一个数,根据transform出来的estimator:$\hat{\theta}$可以带入已有数据计算
- sampling distribution depends on $\theta$
- standard error实际上是没有随机性的
- estimator相关的var和期望其实都是parameter的函数,但本质上没有随机性
- $\hat{\mu_k}$是一个estimator来预测$\mu_k$,如果无偏则前者期望等于后者,前者是随机变量,后者是数值



For example, suppose $\theta_1, \theta_2$ are parameters such that
$$\theta_1 = g_1(\mu_1, \mu_2); \qquad \theta_2 = g_2(\mu_1, \mu_2)$$
then the method of moments estimators of $\theta_1$ and $\theta_2$ are
$$\hat{\theta}_1 = g_1(\hat{\mu}_1, \hat{\mu}_2); \qquad \hat{\theta}_2 = g_2(\hat{\mu}_1, \hat{\mu}_2).$$

**Example 6.4** (Poisson distribution, TBp. 261)
- Suppose that $X_1, \cdots, X_n$ are i.i.d. $\sim P(\lambda)$. Then the first moment of $P(\lambda)$ is $\lambda$. The first sample moment is
$$\hat{\mu}_1 = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$
Therefore, the method of moment estimator of $\lambda$ is $\hat{\lambda} = \overline{X}$.
- As an example, asbestos fibers on filters were counted as part of a project to develop measurement standards for asbestos concentration. Asbestos dissolved in water was spread on a filter, and punches of 3-mm diameter were taken from the filter. Counts of the numbers of fibers in each of 23 grid squares are

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 31 | 29 | 19 | 18 | 31 | 28 | 34 | 27 | 34 | 30 | 16 | 18 |
| 26 | 27 | 27 | 18 | 24 | 22 | 28 | 24 | 21 | 17 | 24 | |

and the method of moments estimate of $\lambda$ is $\hat{\lambda} = 24.9$.

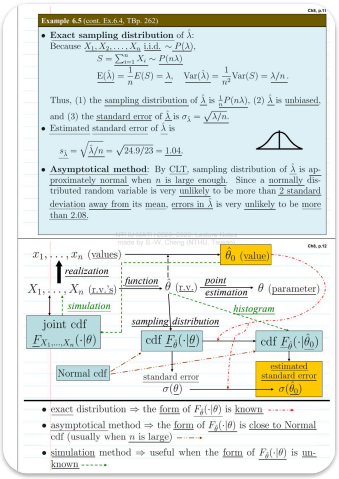
• Questions:
1. Is the estimate (i.e., 24.9) the real $\lambda$?
2. Next time when the same procedure (same sample size, same estimator, same ...) is repeated again to get a new estimate, how far the future estimate will be away from 24.9?
3. In which range will you expect say 95% of the future estimate falls? (e.g., [24.8, 25] or [15, 35]?)
4. This is a question related to the stability/uncertainty/variation of the estimator.
5. How to characterize the stability of an estimator? (**Note.** We have to answer the question using only the observed data.)



To evaluate the stability/uncertainty of an estimation procedure, it is required to know what the sampling distribution is.

- mu_k可以写成两个参数的形式,那么反函数就可以哦那个mu_k来反过来算参数

- poisson:时间(一维)划分小段看事情发生了几次;面积(二维)划分看每一块掉进来几根



**Example 6.6** (Normal distribution, TBp. 263)

- The first and second moments for $N(\mu, \sigma^2)$ are

$$\begin{cases} \mu_1 &= E(X) = \mu \\ \mu_2 &= E(X^2) = \mu^2 + \sigma^2 \end{cases}$$

$$\Rightarrow \begin{cases} \mu &= \mu_1 \\ \sigma^2 &= \mu_2 - \mu_1^2 \end{cases}$$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim N(\mu, \sigma^2)$, then the method of moment estimators of $\mu$ and $\sigma^2$ are

$$\begin{cases} \hat{\mu} &= \overline{X} \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \end{cases}$$

- Sampling distribution of $\overline{X}$ is Normal$(\mu, \frac{\sigma^2}{n})$ and sampling distribution of $\hat{\sigma}^2$ is $\frac{\sigma^2}{n}\chi^2_{n-1}$. Furthermore, $\overline{X}$ and $\hat{\sigma}^2$ are independent.





**Example 6.7** (Gamma distribution, TBp. 263-264)

- The first two moments of the $\Gamma(\alpha, \lambda)$ are

$$\begin{cases} \mu_1 &= \alpha/\lambda \\ \mu_2 &= \alpha(\alpha+1)/\lambda^2 \end{cases} \Rightarrow \begin{cases} \lambda &= \mu_1/(\mu_2 - \mu_1^2) \\ \alpha &= \lambda\mu_1 = \mu_1^2/(\mu_2 - \mu_1^2) \end{cases}$$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim \Gamma(\alpha, \lambda)$, then the method of moment estimators of $\lambda$ and $\alpha$ are

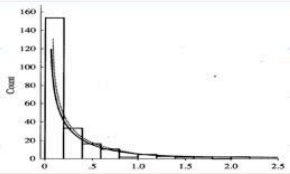
$$\hat{\lambda} = \overline{X}/\hat{\sigma}^2, \qquad \hat{\alpha} = \overline{X}^2/\hat{\sigma}^2$$

where $\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2$.

- As a concrete example, let us consider the fit of the rainfall amounts during 227 storms in Illinois from 1960 to 1964 (the data, listed in Problem 42, Textbook p.414). For the data, $\overline{X} = .224, \hat{\sigma}^2 = .1338$, therefore $\hat{\alpha} = .375$ and $\hat{\lambda} = 1.674$.
- What are the sampling distributions of $\hat{\alpha}$ and $\hat{\lambda}$?
  - Exact distributions are complicated.
  - Asymptotic method: Use central limit theorem and other asymptotic theorems to obtain the limiting distributions.
  - simulation method: bootstrap.

**Definition 6.7** (parametric bootstrap, TBp. 264-265)

Generate many, many samples of size $n$ from a distribution. (True values of parameters in the distribution are unknown, so use their estimates). From each of the samples, compute the estimates of parameters. The histogram of these estimates should give a good idea of the sampling distribution of estimator.

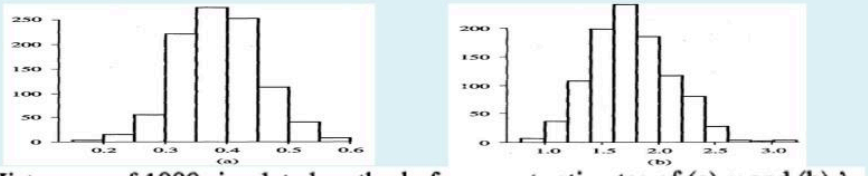


**Example 6.8** (cont. Ex.6.7, bootstrap, TBp. 265-266)

Generate 1000 samples of size 227 from $\Gamma(\hat{\alpha}, \hat{\lambda})$, where $\hat{\alpha}, \hat{\lambda}$ are set to be .375 and 1.674, respectively. From each of 1000 samples, compute the estimates of $\alpha$ and $\lambda$ using the estimators $\hat{\lambda} = \overline{X}/\hat{\sigma}^2$, $\hat{\alpha} = \overline{X}^2/\hat{\sigma}^2$. Denote the 1000 estimates by $\alpha_i^*, \lambda_i^*$, $i = 1, \ldots, 1000$. Histograms of them indicate the variability that is inherent in estimating the parameters from a sample of this size.







Histogram of 1000 simulated method of moment estimates of (a) $\alpha$ and (b) $\lambda$.

We can find from the 1000 $\alpha_i^*$'s and $\lambda_i^*$'s that:

- The histograms looks like normal.
- The histogram suggests that if $\alpha = 0.375$, then it is not very unusual that $\hat{\alpha}$ is in error by 0.1 or more.
- The histograms are centered at 0.375 and 1.674, the (regarded-as-true) parameter values used in the simulation.
- Estimated standard error of $\hat{\alpha}$ is

$$s_{\hat{\alpha}} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}(\alpha_i^* - \overline{\alpha})^2} = 0.06,$$

where $\overline{\alpha}$ is the mean of the 1000 values. Also, in the same way, $s_{\hat{\lambda}} = 0.34$.

❖ **Reading**: textbook, 8.4

## • method of finding estimators II --- Maximum Likelihood Estimator (MLE)

**Questions:**

- Toss a coin 10 times. Let $\theta$ be the probability of getting a head. Suppose that we know
$$\theta \in \{0.1, 0.5, 0.9\}.$$
- When we get 7 heads out of the 10 tosses, which $\theta$ is more plausible to generate the output?
- **Hint.**
$$P(7 \text{ heads}|\theta = 0.1) \approx 0.000,$$
$$P(7 \text{ heads}|\theta = 0.5) \approx 0.117,$$
$$P(7 \text{ heads}|\theta = 0.9) \approx 0.057.$$

**Definition 6.8** (likelihood, log likelihood, TBp. 267, 268)

Suppose random variables $X_1, \ldots, X_n$ have a joint pdf or pmf
$$f(x_1, \ldots, x_n|\Theta).$$
Given the observed values $X_1 = x_1^*, \ldots, X_n = x_n^*$, the **likelihood function** of $\Theta$ is defined as
$$\mathcal{L}(\Theta) = f(x_1^*, x_2^*, \ldots, x_n^*|\Theta),$$
which is a function of $\Theta$. The **log likelihood function** is defined as $\log \mathcal{L}(\Theta)$.





**Notes.**

1. We consider likelihood function as a function of $\theta$ while joint pdf/pmf as a function of $x_i$'s.
2. For discrete case, likelihood function gives the probability of observing the data as a function of $\theta$.

**Definition 6.9** (maximum likelihood estimator, TBp. 267)

The **maximum likelihood estimator (MLE)** of $\theta$ is the value of $\theta$ that maximizes the likelihood.

**Interpretation.** MLE makes the observed data "most probable" or "most likely," i.e., MLE gives the most "plausible" model given the observed data.

**Note.**

1. For i.i.d. case, the likelihood function and the log likelihood function are, respectively,
$$\mathcal{L}(\theta) = \prod_{i=1}^{n} f(x_i^*|\theta), \quad \text{and} \quad l(\theta) = \sum_{i=1}^{n} \log f(x_i^*|\theta).$$
2. Maximizing the likelihood function, $\mathcal{L}(\theta)$, is equivalent to maximizing its natural logarithm, $l(\theta)$, since the logarithm is a monotonic function.

**Theorem 6.1** (invariance property of MLE)

If $\hat{\theta}$ is the MLE of $\theta$, then for any function of $\theta$, denoted by $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

**Proof.** MLE of $\tau(\theta)$ is a solution of the maximization problem

$$\max_{\tau(\theta)} l(\theta).$$

Since $\hat{\theta}$ is the MLE of $\theta$, the maximum is attained when $\theta = \hat{\theta}$, which implies the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

**Example 6.10** (i.i.d Poisson distribution, TBp. 268)

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. $P(\lambda)$. The log likelihood is

$$l(\lambda) = \sum_{i=1}^{n} \log \frac{e^{-\lambda}\lambda^{X_i}}{X_i!} = -n\lambda + \log\lambda \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \log X_i!.$$

Setting $l'(\lambda) = 0$ gives $\frac{1}{\lambda}\sum_{i=1}^{n} X_i - n = 0.$

The MLE is then

$$\hat{\lambda} = \overline{X}.$$





Check that this is a maximum:

$$l''(\lambda) = -\frac{n\overline{X}}{\lambda^2} < 0 \quad \Rightarrow \quad l(\lambda) \text{ is concave.}$$

- **Example for Thm 6.1, LNp.19** $\Rightarrow$ the MLE of $\frac{1}{\lambda}$ is $\frac{1}{\overline{X}}$.

**Example 6.11** (i.i.d normal distribution, TBp. 269)

Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ random variables. The joint density is

$$f(x_1, x_2, \ldots, x_n | \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2\right].$$

The log likelihood is

$$l(\mu, \sigma) = \sum_{i=1}^{n} \left[-\log\sigma - \frac{1}{2}\log(2\pi) - \frac{1}{2}\left(\frac{X_i - \mu}{\sigma}\right)^2\right].$$

Setting

$$\begin{cases} 0 = \frac{\partial l}{\partial \mu} = \sigma^{-2}\sum_{i=1}^{n}(X_i - \mu) \\ 0 = \frac{\partial l}{\partial \sigma} = -n\sigma^{-1} + \sigma^{-3}\sum_{i=1}^{n}(X_i - \mu)^2 \end{cases}$$

The MLE is then

$$\begin{cases} \hat{\mu} &= \overline{X} \\ \hat{\sigma} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2} \end{cases}$$

which is the same as the method of moments estimators.

Check maximum $\Rightarrow$

$$\begin{pmatrix} \frac{\partial^2 l}{\partial \mu^2} & \frac{\partial^2 l}{\partial \sigma \partial \mu} \\ \frac{\partial^2 l}{\partial \mu \partial \sigma} & \frac{\partial^2 l}{\partial \sigma^2} \end{pmatrix} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu) \\ \frac{2}{\sigma^3}\sum_{i=1}^{n}(X_i - \mu) & \frac{3}{\sigma^4}\sum_{i=1}^{n}(X_i - \mu)^2 - \frac{n}{\sigma^2} \end{pmatrix}$$

which is negative definite when $\mu = \hat{\mu}$ and $\sigma = \hat{\sigma}$ and $l \to 0$ as $(\mu, \sigma)$ tends to boundary.

- **Example for Thm 6.1, LNp.19,**
  - MLE of $\mu^2$, the square of a normal mean, is $\overline{X}^2$
  - MLE of $\sigma^2$, the variance, is $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$





**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

$$\begin{aligned} l(\mu) &= -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{n}\left(X_i - \mu\right)^2 \\ &= -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 - \tfrac{n}{2}(\overline{X} - \mu)^2 . \end{aligned}$$

Hence the MLE of $\mu$ is

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

**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 $\theta$ is

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

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

**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 $p_1, p_2, \ldots, p_m$ ($p_i \geq 0$ 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

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

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

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

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

Introduce a Lagrange multiplier $\lambda$, and maximize

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

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

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





Since

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

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

**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 $\theta^2$.

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

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 $n = X_1 + X_2 + X_3$.

$$(X_1, X_2, X_3) \sim \text{multinomial}(n, p_1, p_2, p_3)$$

$$(1-\theta)^2, 2\theta(1-\theta), \theta^2$$

Log likelihood of $\theta$ is

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

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

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

yields the MLE of $\theta$

$$\hat{\theta} = \frac{2X_3 + X_2}{2X_1 + 2X_2 + 2X_3} = \frac{2X_3 + X_2}{2n}.$$

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





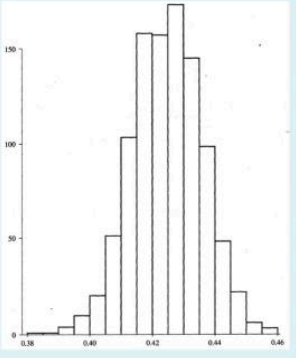
**Chinese population data of Hong Kong in 1937**: ($M$, $N$ are erythrocyte antigens)

| Blood Type | $M$ | $MN$ | $N$ | Total |
|---|---|---|---|---|
| Frequency | 342 | 500 | 187 | 1029 |
| (by Ex.6.14) | 0.333 | 0.486 | 0.182 | |

1. MLE is $\hat{\theta} = 0.4247 \Rightarrow (\hat{p}_1, \hat{p}_2, \hat{p}_3) = (0.331, 0.489, 0.180)$.
2. Approximate the sampling distribution of $\hat{\theta}$ by bootstrap:
   - 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 $\hat{\theta}$: $s_{\hat{\theta}} = 0.011$.



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

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

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

- The mean of $X$ is

$$\mu = \alpha/3 \Rightarrow \alpha = 3\mu.$$

- The moments estimator of $\alpha$ based on a sample $X_1, \cdots, X_n$ is $\hat{\alpha} = 3\overline{X}$.
- The log likelihood of $\alpha$ is

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

Setting the derivative equal to zero, the MLE of $\alpha$ satisfies the *nonlinear equation*

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

- The MLE of $\alpha$ 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

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

- 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

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

- 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.

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

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

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

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

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

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

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

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

**3.** From the histograms of the simulated MLEs:

- The histograms look like normal.

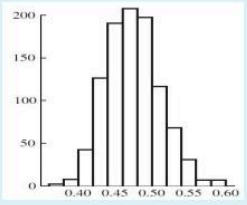


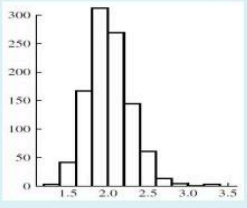


- The histograms are centered at $\hat{\alpha} = 0.471$ and $\hat{\lambda} = 1.97$.
- Estimated standard error of the MLE's are

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

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




**Summary** (advantages of MLE)

1. easy to interpret
2. widely applicable
3. the range of the MLE coincides with the range of the parameter
4. invariance under reparameterizations
5. nice theoretical properties

❖ **Reading**: textbook, 8.5, 8.5.1

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

**Recall:** 1. Law of Large Number
2. Central Limit Theorem
→ for sum/average

**Definition 6.10** (consistent, TBp. 266)

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

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

➢ method of moment estimator

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

The weak law of large numbers implies that

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

If the function relating $\mu_k$ and $\theta_j$ are continuous,
method of moments estimators are consistent.





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

- Recall: In LNp.12, $\sigma_{\hat{\theta}}(\hat{\theta}) \xrightarrow{\text{estimate}} \sigma_{\hat{\theta}}(\theta)$
- Consider the standard error of the form: $\sigma_{\hat{\theta}}(\theta) = \frac{1}{\sqrt{n}}\sigma^*(\theta)$
  - Ex. 6.5 (LNp.11): $\sigma_{\hat{\lambda}}(\lambda) = \frac{1}{\sqrt{n}}\sqrt{\lambda}$
  - Ex. 6.6 (LNp.13): $\sigma_{\hat{\mu}}(\mu, \sigma) = \frac{1}{\sqrt{n}}\sigma$, and $\sigma_{\hat{\sigma}^2}(\mu, \sigma) \approx \frac{1}{\sqrt{n}}\sqrt{2\sigma^2}$
- Let $\theta_0 =$ true parameter, and $\sigma_{\hat{\theta}} \equiv \sigma_{\hat{\theta}}(\theta_0) = \frac{1}{\sqrt{n}}\sigma^*(\theta_0)$.
- Estimate $\sigma_{\hat{\theta}}$ by $s_{\hat{\theta}} \equiv \frac{1}{\sqrt{n}}\sigma^*(\hat{\theta})$.
- If (1) $\sigma^*(\theta)$ is continuous in $\theta$, and (2) $\hat{\theta}$ is consistent $\left(\hat{\theta} \xrightarrow{P} \theta_0\right)$, then

$$\sigma^*(\hat{\theta}) \xrightarrow{P} \sigma^*(\theta_0), \text{ i.e.,}$$
$$\text{as } n \to \infty, \ \frac{s_{\hat{\theta}}}{\sigma_{\hat{\theta}}} \to 1 \text{ in probability} \quad \left(\Rightarrow s_{\hat{\theta}} \xrightarrow{p} \sigma_{\hat{\theta}} \to 0\right)$$

➢ 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
$$l(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta).$$
- The MLE maximizes $l(\theta)$ and is usually obtained by solving the **score equation**
$$0 = \frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta).$$
- $\frac{\partial}{\partial\theta} l(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta)$ is called **score function**.

**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

$$I_{X_1,\ldots,X_n}(\theta) = \mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_1, \cdots, X_n|\theta) \right]^2,$$

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

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





$$I_{X_1,\ldots,X_n}(\theta) = \mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_1, \cdots, X_n|\theta) \right]^2$$

**Theorem 6.4** (TBp. 276)

Let $X_1, \ldots, X_n$ be an i.i.d. sample of size $n$ from a pdf/pmf $f(x|\theta)$.

$$I_{X_1,\cdots,X_n}(\theta) = \mathrm{E}_\theta \left( \frac{\partial}{\partial\theta} \log \left[ \prod_{i=1}^{n} f(X_i|\theta) \right] \right)^2 = \mathrm{E}_\theta \left[ \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f(X_i|\theta) \right]^2$$

$$= \sum_{i=1}^{n} \mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_i|\theta) \right]^2 + 2 \sum_{i<j} \mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_i|\theta) \right] \mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_j|\theta) \right]$$

$$= n\mathrm{E}_\theta \left[ \frac{\partial}{\partial\theta} \log f(X_1|\theta) \right]^2 \equiv n \cdot I_{X_1}(\theta)$$

- $I_{X_1}(\theta)$ is the Fisher information contained in a sample of size one.
- $nI_{X_1}(\theta)$: interpreted as the information of $\theta$ contained in a sample of size $n$ from $f(\cdot|\theta)$.
- The Fisher informations of independent samples are additive.

**Theorem 6.4** (TBp. 276)

Under appropriate smoothness conditions on $f$,

$$I_{X_1}(\theta) \equiv \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 = Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right]$$

**Proof** (for pdf case): Since $\int f(x|\theta)\,dx = 1$ for all $\theta$,

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx = \int \frac{\partial}{\partial\theta} f(x|\theta)\,dx = \int \left[\frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)}\right] f(x|\theta)\,dx \\
&= \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx = E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]. \qquad \cdots(\Delta) \\
&\Rightarrow\ Var_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right] \\
&= \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 - \left\{E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]\right\}^2 \\
&= \mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2.
\end{aligned}$$

$$\begin{aligned}
0 &= \frac{\partial^2}{\partial\theta^2}\int f(x|\theta)\,dx = \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right] f(x|\theta)\,dx \\
&= \int \left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] f(x|\theta)\,dx + \int \left[\frac{\partial}{\partial\theta}\log f(x|\theta)\right]^2 f(x|\theta)\,dx \\
&= E_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right] + E_\theta\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2.
\end{aligned}$$

(need smoothness of $f$ for interchanging integration and differentiation.)





**Note** (TBp. 278, 279).

For i.i.d. case,

$$\begin{aligned}
E_\theta[l'(\theta)^2] &= I_{X_1,\ldots,X_n}(\theta) = n\cdot I_{X_1}(\theta) = -n\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1|\theta)\right] \\
&= -\sum_{i=1}^n \mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta)\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\sum_{i=1}^n \log f(X_i|\theta)\right] = -\mathrm{E}_\theta\left[l''(\theta)\right]
\end{aligned}$$

- **interpretation**: when $|E_\theta[l''(\theta)]|$ is large at $\theta = \theta_0$, $l(\theta)$ is, on average, changing rapidly in a vicnity of $\theta_0$

**Example 6.18** (Fisher information of i.i.d. Bernoulli $B(\theta)$)

Let $X_1, \ldots, X_n$ be i.i.d. from Bernoulli distribution $B(\theta)$ (i.e., the pmf of $X_i$ is, $\theta^x(1-\theta)^{1-x}$, for $x \in \{0,1\}$),

then $E(X_i) = \theta$ and $Var(X_i) = \theta(1-\theta)$.

- For a single observatoin $X_i$, the first and second deratives of its log likelihood are:

$$\begin{aligned}
\log f(x|\theta) &= x\log\theta + (1-x)\log(1-\theta), \\
\partial/\partial\theta \log f(x|\theta) &= x/\theta - (1-x)/(1-\theta) = (x-\theta)/(\theta(1-\theta)), \\
\partial^2/\partial\theta^2 \log f(x|\theta) &= -x/\theta^2 - (1-x)/(1-\theta)^2.
\end{aligned}$$

- The Fisher information of a single observation, say $X_1$, is

$$\begin{aligned} I_{X_1}(\theta) &= E_\theta\left[\frac{X_1-\theta}{\theta(1-\theta)}\right]^2 = \frac{E_\theta[(X_1-\theta)^2]}{\theta^2(1-\theta)^2} \\ &= \frac{Var_\theta(X_1)}{\theta^2(1-\theta)^2} = \frac{\theta(1-\theta)}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}. \\ I_{X_1}(\theta) &= -E_\theta\left[-\frac{X_1}{\theta^2} - \frac{1-X_1}{(1-\theta)^2}\right] \\ &= \frac{\theta}{\theta^2} + \frac{1-\theta}{(1-\theta)^2} = \frac{1}{\theta} + \frac{1}{1-\theta} = \frac{1}{\theta(1-\theta)}. \end{aligned}$$

- The Fisher information of observations $X_1, \ldots, X_n$ is

$$I_{X_1,\cdots,X_n}(\theta) = nI_{X_1}(\theta) = \frac{n}{\theta(1-\theta)}.$$

  Notice that $I_{X_1,\cdots,X_n}(\theta)$

  - increases when $n$ increases,
  - increases when $\theta \downarrow 0$ or $\theta \uparrow 1$,
  - reaches a minimum $4n$ at $\theta = 0.5$.





- Consider a single observation $Y \sim$ Binomial$(n, \theta)$. The pmf of $Y$ is

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, \quad \text{for } y \in \{0, 1, \ldots, n\}.$$

  - The second derative of log likelihood is

$$\partial^2 \log f(y|\theta)/\partial^2\theta = -y/\theta^2 - (n-y)/(1-\theta)^2.$$

  - The Fisher information of $Y$, is

$$I_Y(\theta) = -E_\theta\left[-\frac{Y}{\theta^2} - \frac{n-Y}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta(1-\theta)}.$$

  - Note that $I_Y(\theta)$ is the same as $I_{X_1,\cdots,X_n}(\theta)$.

**Theorem 6.5** (consistency of MLE, TBp. 275)

Under appropriate smoothness conditions of $f$, the MLE from an i.i.d. sample is consistent.

**Proof (sketch, for pdf case):** Denote the true value of $\theta$ by $\theta_0$. The MLE maximizes $\frac{l(\theta)}{n} = \frac{1}{n}\sum_{i=1}^n \log f(X_i|\theta)$.

What is the difference between $\int[\log f(x|\theta)]f(x|\theta_0)dx$ and $\int[\log f(x|\theta)]f(x|\theta)dx$ ?

The weak law of large numbers implies

$$\forall\, \theta, \quad \frac{l(\theta)}{n} \xrightarrow{\mathcal{P}} \mathrm{E}_{\theta_0}[\log f(X|\theta)] = \int [\log f(x|\theta)]\, f(x|\theta_0)\, dx \quad \text{as } n \to \infty.$$

For large $n$, the $\theta$ value that maximizes $l(\theta)$ should be close to the $\theta$ value that maximizes $\mathrm{E}_{\theta_0}[\log f(X|\theta)]$. To maximize $\mathrm{E}_{\theta_0}[\log f(X|\theta)]$, consider

$$\frac{\partial}{\partial\theta}\mathrm{E}_{\theta_0}[\log f(X|\theta)] = \frac{\partial}{\partial\theta}\int [\log f(x|\theta)] f(x|\theta_0)\,dx = \int \frac{\frac{\partial}{\partial\theta}f(x|\theta)}{f(x|\theta)} f(x|\theta_0)\,dx\,.$$

If $\theta = \theta_0$ then

$$\frac{\partial}{\partial\theta}\mathrm{E}_{\theta_0}[\log f(X|\theta)]\Big|_{\theta=\theta_0} = \int \frac{\partial}{\partial\theta} f(x|\theta)\,dx\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\int f(x|\theta)\,dx\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}1 = 0.$$

Thus $\theta_0$ is a stationary point and hopefully a maximizer.

**Theorem 6.6** (asymptotic normality of MLE for one-dimensional parameter, TBp. 277)

Under some regularity conditions on $f$, the probability distribution of

$$\sqrt{nI_{X_1}(\theta_0)}(\hat\theta_{\mathrm{MLE}} - \theta_0)$$

tends to a standard Normal distribution as $n$ tends to infinity, where $\hat\theta_{\mathrm{MLE}}$ is MLE and $\theta_0$ is the true value of $\theta$.

**Proof (sketch):** Denote $\hat\theta_{\mathrm{MLE}}$ by $\hat\theta$. By Taylor expansion,

$$0 = l'(\hat\theta) \approx l'(\theta_0) + (\hat\theta - \theta_0)l''(\theta_0)$$

$$\sqrt{n}(\hat\theta - \theta_0) \approx \frac{-l'(\theta_0)/\sqrt{n}}{l''(\theta_0)/n}$$





Consider the numerator,

$$\mathrm{E}_{\theta_0}\left[l'(\theta_0)\right] = \sum_{i=1}^{n}\mathrm{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right] = 0\,,$$

$$\mathrm{Var}_{\theta_0}\left[l'(\theta_0)\right] = \sum_{i=1}^{n}\mathrm{E}_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right]^2 = nI_{X_1}(\theta_0)\,.$$

And, since $l'(\theta_0) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)$, Central Limit Theorem implies that $l'(\theta_0)/\sqrt{nI_{X_1}(\theta_0)}$ converges in distribution to a standard Normal random variable. For the denominator,

$$\frac{l''(\theta_0)}{n} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0) \xrightarrow{P} \mathrm{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_i|\theta_0)\right] = -I_{X_1}(\theta_0)\,.$$

Hence

$$\sqrt{nI_{X_1}(\theta_0)}(\hat\theta - \theta_0) \xrightarrow{\mathcal{D}} N(0,1).$$

Also,

$$\mathrm{E}_{\theta_0}\left[\sqrt{n}(\hat\theta - \theta_0)\right] \approx 0, \qquad \mathrm{Var}_{\theta_0}(\hat\theta - \theta_0) \approx \frac{1}{nI_{X_1}(\theta_0)}.$$

Thus, MLE ($\hat\theta$) is asymptotically unbiased and its asymptotic variance is

$$\left[nI_{X_1}(\theta_0)\right]^{-1} = \left[I_{X_1,\ldots,X_n}(\theta_0)\right]^{-1} = \left[-\mathrm{E}_{\theta_0}(l''(\theta_0))\right]^{-1}.$$

**Notes** (TBp. 277)

1. Theorem 6.6 (LNp.39) says the large sample distribution of an MLE is approximately normal with
   - mean $\theta_0$ and ($\Rightarrow$ **asymptotically unbiased**)
   - variance $1/(nI_{X_1}(\theta_0))$. ($\Rightarrow$ $1/(nI_{X_1}(\theta_0))$ referred to as **asymptotic variance**)
2. Regularity conditions for Theorem 6.6 (LNp.39):
   (a) True value $\theta_0$ is an interior point of the set of all parameter values (e.g., the theorem would not be expected to apply in Ex 6.16, LNp.27, if $\alpha_0 = 1$.)
   (b) Support of $f(x|\theta)$, i.e., the set of $x$'s for which $f(x|\theta) > 0$, does not depend on $\theta$ (e.g., the theorem would not be expected to apply to estimating $\theta$ from a sample that are uniformly distributed on the interval $[0, \theta]$ (Ex 6.13, LNp.22).)

**Theorem 6.7** (Fisher information under reparameterization)

Under the reparameterization $\tau(\theta)$, the (Fisher) information of $\tau(\theta)$ is

$$I_{X_1}(\tau(\theta)) \equiv \mathrm{E}\left[\frac{\partial}{\partial\tau(\theta)}\log f(X_1|\theta)\right]^2 = \frac{I_{X_1}(\theta)}{\tau'(\theta)^2}.$$





**Proof:**

$$\begin{aligned} I_{X_1}(\tau(\theta)) &= E\left[\frac{\partial}{\partial\tau(\theta)}\log f(X_1|\theta)\right]^2 = E\left[\frac{\partial\theta}{\partial\tau(\theta)}\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 \\ &= \left(\frac{1}{\tau'(\theta)}\right)^2 \cdot E\left[\frac{\partial}{\partial\theta}\log f(X_1|\theta)\right]^2 = \frac{I_{X_1}(\theta)}{\tau'(\theta)^2} \end{aligned}$$

**Theorem 6.8** (asymptotic normality of MLE under reparameterization)

Under the reparameterization $\tau(\theta)$, the MLE of $\tau(\theta)$, $\tau(\hat{\theta})$, is asymptotically Normal with mean $\tau(\theta)$ and variance

$$\frac{1}{nI_{X_1}(\tau(\theta))} = \frac{1}{n}\cdot\frac{\tau'(\theta)^2}{I_{X_1}(\theta)}.$$

**Example 6.19** (information and asymptotic distribution of MLE for Poisson mean, TBp.282)

Let $X_1, \ldots, X_n$ be i.i.d $\sim P(\lambda)$. The MLE of $\lambda$ is $\hat{\lambda} = \overline{X}$. The information of Poisson distribution is:

$$\begin{aligned} I_{X_1}(\lambda) &= \mathrm{E}\left[\frac{\partial}{\partial\lambda}\log f(X|\lambda)\right]^2 = \mathrm{E}\left[\frac{\partial}{\partial\lambda}\log\left(\frac{\lambda^X e^{-\lambda}}{X!}\right)\right]^2 \\ &= \mathrm{E}\left[\frac{\partial}{\partial\lambda}(X\log\lambda - \lambda - \log X!)\right]^2 = \mathrm{E}\left(\frac{X}{\lambda} - 1\right)^2 = \frac{1}{\lambda}. \end{aligned}$$

Or,

$$I_{X_1}(\lambda) = -\mathrm{E}\left[\frac{\partial^2}{\partial\lambda^2}\log f(X|\lambda)\right] = -\mathrm{E}\left[\frac{-X}{\lambda^2}\right] = \frac{1}{\lambda}.$$

Hence, by theorem 6.6, the asymptotic distribution of $\overline{X}$ is Normal with mean $\lambda$ and variance $\lambda/n$.

Exercise: Suppose that we are interested in the parameter $\tau = 1/\lambda$,

- what is the (Fisher) information of $\tau$? (Ans: $\tau^{-3}$)
- what is the MLE of $\tau$? (Ans: $\hat{\tau}_{\mathrm{MLE}} = 1/\overline{X}$)
- what is the asymptotic distribution of the MLE?
- what is its asymptotic variance?

**Theorem 6.9** (multidimensional parameters, information and asymptotic normality of MLE, TBp.279)

Suppose $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)'$, a $k$-dimensional vector. Then, the asymptotic joint distribution of the MLE $\hat{\Theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_k)'$ is multivariate normal with mean vector $\Theta_0$ and covariance matrix $\frac{1}{n}I^{-1}(\Theta_0)$ where $I(\Theta)$ is the $k \times k$ matrix with $ij$-th component

$$\mathrm{E}_\Theta\left[\frac{\partial}{\partial\theta_i}\log f(X_1|\Theta)\,\frac{\partial}{\partial\theta_j}\log f(X_1|\Theta)\right] = -\mathrm{E}_\Theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log f(X_1|\Theta)\right].$$

Here, $I(\Theta)$ is called the **(Fisher) information matrix** for $\Theta$.

❖ **Reading**: textbook, 8.5.2; **Further reading**: Hogg et al., 6.1, 6.2

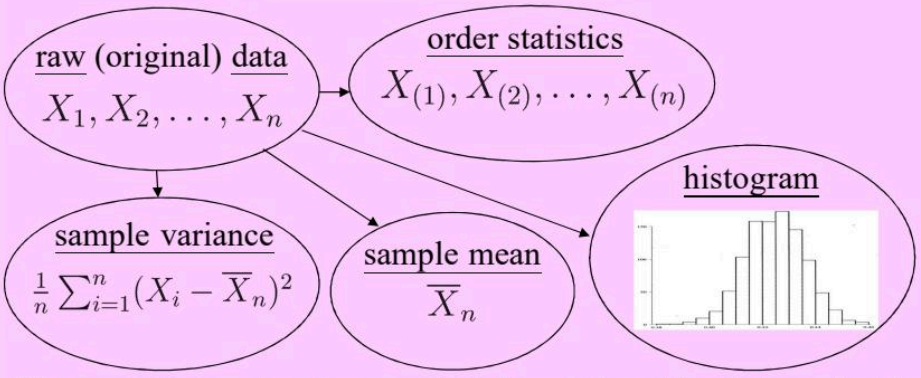




• Data reduction --- the concepts of sufficiency, minimal sufficiency, and completeness

**Question 6.1** (information and data reduction)

(numerical or graphical) transformations of data appear all the time in statistics for offering a summary of information contained in data. For example,




To present or extract concrete information, the data reduction through useful transformations is required. However, non-invertible transformations can cause the loss of information. (Example?) The lost information can be important or worthless to the objective of studying the data.

Q: How to examine whether the important information lost in transformation? Furthermore, what is *important information*?

**Summary** (formulation of information and data reduction problem, TBp. 305)

- Let $X_1, X_2, \ldots, X_n$ be a sample with joint pdf/pmf $f(\mathbf{x}|\Theta)$, where $\Theta$ is unknown parameter.
  - $X_1, X_2, \ldots, X_n$ contains two types of information:
    - information related to $\Theta$
    - information irrelevant to $\Theta$
  - For example, toss a coin $n$ times, i.e., $X_1, X_2, \ldots, X_n$ are i.i.d. from Bernoulli $B(\theta)$,
    - $\overline{X}_n$ or $T = \sum_{i=1}^n X_i$ contains information about $\theta$
    - When $T$ is known, say $T = t$, the information that at which trials the $t$ head's occur is irrelevant to $\theta$
    - $n$=5, consider the following possible results:
      - ▷ (0, 1, 1, 1, 1), $T = 4$; (1, 0, 1, 1, 1), $T = 4$; (1, 1, 0, 1, 1), $T = 4$; (1, 1, 1, 0, 1), $T = 4$; (1, 1, 1, 1, 0), $T = 4$
      - ▷ (1, 0, 0, 0, 0), $T = 1$; (0, 1, 0, 0, 0), $T = 1$; (0, 0, 1, 0, 0), $T = 1$; (0, 0, 0, 1, 0), $T = 1$; (0, 0, 0, 0, 1), $T = 1$





- Information about $\theta$ is revealed by the different values of $T$, i.e., larger $T$, larger $\theta$, and vice versa.

$(X_1, \ldots, X_n)$  | $T = \sum_{i=1}^n X_i$ 

- Question. Is there a statistic $T(X_1, X_2, \ldots, X_n)$ which contains all the information in the sample about $\theta$? If so, a reduction of the original data to this statistic without loss of information is possible.

**Definition 6.13** (sufficient, TBp. 305)

A statistic $T(X_1, X_2, \ldots, X_n)$ is said to be **sufficient** for $\theta$ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ does not depend on $\theta$ for any value of $t$.

**Caution**:

1. If $T$ is a sufficient statistic, formally, we can keep only $T$ and throw away all $X_i$'s. Realistically, the $X_i$'s are used to check whether the model did not fit, or that something was fishy about the data.
2. The definition of "all (important) information" depends on the statistical modeling, i.e., the joint distribution assumption.

**Example 6.20** (sufficient statistics of i.i.d. Bernoulli distribution, TBp. 306)

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with $\mathrm{P}(X_i = 1) = \theta$. Let $T = \sum_{i=1}^n X_i$ then

$$\begin{aligned} \mathrm{P}(X_1 = x_1, \ldots, X_n = x_n | T = t) &= \frac{\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n, T = t)}{\mathrm{P}(T = t)} \\ &= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t} \theta^t (1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}, \end{aligned}$$

if $x_1 + \cdots + x_n = t$ and $x_i$ are nonnegative integers, and 0 otherwise. The conditional distribution is independent of $\theta$. Hence $T$ is sufficient for $\theta$.

**Theorem 6.10** (factorization theorem, TBp. 306)

A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter $\theta$ is that the joint pdf or pmf of $X_1, \ldots, X_n$ factors in the form

$$f(x_1, x_2, \ldots, x_n | \theta) = g\big(T(x_1, x_2, \ldots, x_n), \theta\big)\, h(x_1, x_2, \ldots, x_n)$$

**intuition**: $P(X_1 = x_1, \ldots, X_n = x_n) = P(T = t) P(X_1 = x_1, \ldots, X_n = x_n | T = t)$

**Proof:** only for discrete case (continuous case requires some regularity conditions, but the basic idea are the same.): ($\Leftarrow$) Suppose

$$f(x_1, x_2, \ldots, x_n | \theta) = g\big(T(x_1, x_2, \ldots, x_n), \theta\big)\, h(x_1, x_2, \ldots, x_n).$$





Then

$$\mathrm{P}(T = t) = \sum_{T(\mathbf{x}) = t} \mathrm{P}(\mathbf{X} = \mathbf{x}) = g(t, \theta) \sum_{T(\mathbf{x}) = t} h(\mathbf{x}),$$

$$\mathrm{P}(\mathbf{X} = \mathbf{x} | T = t) = \frac{\mathrm{P}(\mathbf{X} = \mathbf{x}, T = t)}{\mathrm{P}(T = t)} = \frac{g(t, \theta) \cdot h(\mathbf{x})}{g(t, \theta) \cdot \sum_{T(\mathbf{x}) = t} h(\mathbf{x})},$$

which does not depend on $\theta$. Hence $T$ is sufficient for $\theta$. ($\Rightarrow$) Conversely, suppose that the conditional distribution of $\mathbf{X}$ given $T$ is independent of $\theta$. Let

$$g(t, \theta) = \mathrm{P}(T = t | \theta), \quad h(\mathbf{x}) = \mathrm{P}(\mathbf{X} = \mathbf{x} | T = t).$$

Then $\mathrm{P}(\mathbf{X} = \mathbf{x} | \theta) = \mathrm{P}(T = t | \theta) \mathrm{P}(\mathbf{X} = \mathbf{x} | T = t) = g(t, \theta) h(\mathbf{x})$ as required.

**Theorem 6.11** (MLE and sufficient statistics, TBp.309)

If $T$ is sufficient for $\theta$, then the maximum likelihood estimate for $\theta$, if unique, is a function of $T$.

**Proof.** From factorization theorem, the likelihood is $g(t, \theta) h(\mathbf{x})$.
To maximize this quantity we only need to maximize $g(t, \theta)$

**Example 6.21** (cont. Ex. 6.20, sufficient statistic of i.i.d. Bernoulli distribution, TBp.309)

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables

$$\mathrm{P}(X_i = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0 \text{ or } 1.$$

Then

$$f(x_1,\dots,x_n|\theta) = \prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$
$$= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}(1-\theta)^n = g(t,\theta)h(x_1,\dots,x_n)$$

where $t = \sum_{i=1}^{n} x_i$, $g(t,\theta) = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n$, $h(x) = 1$

Hence $T = \sum_{i=1}^n X_i$ is sufficient for $\theta$.

**Example 6.22** (sufficient statistics of i.i.d. Normal distribution, TBp.308)

If $X_i \sim N(\mu,\sigma^2)$, $i = 1,2,\dots,n$, are i.i.d, where $\mu,\sigma$ are unknown. Then

$$f(x_1,\dots,x_n|\mu,\sigma) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$
$$= \frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$
$$= \frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}x_i^2 - 2\mu\sum_{i=1}^{n}x_i + n\mu^2\right)\right\}$$

and $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a 2-dimensional sufficient statistic for $(\mu,\sigma)$.



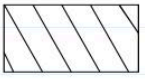
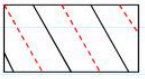


**Question 6.2**

There exist many sufficient statistics. A trivial example is the raw data. However, a large collection of numbers is not as *meaningful* as a few good summary statistics. How can we know whether the data has been reduced *as much as possible* and still keep relelvant information?

**Definition 6.14** (minimal sufficient statistic)

A sufficient statistic $S$ for $\theta$ is called **minimal** if $S$ is a function of $T$ for any sufficient statistic $T$ of $\theta$.

$X_1,\dots,X_n$  $T = g(X_1,\dots,X_n)$  $S = h(T)$ 

**Example 6.23** (minimal sufficient statistics for i.i.d. Uniform distribution $U(\theta, \theta+1)$)

Let $X_1,\dots,X_n$ be i.i.d. from uniform distribution $U(\theta,\theta+1)$. Then, the joint pdf is

$$f(\mathbf{x}|\theta) = \prod_{i=1}^{n} I_{(\theta,\theta+1)}(x_i) = I_{(x_{(n)}-1,x_{(1)})}(\theta),$$

where $I_{(a,b)}(u) = 1$ if $a \le u \le b$ and 0 otherwise. Therefore, $T = (X_{(1)}, X_{(n)})$ is sufficient for $\theta$ by factorization theorem.

Note that

$$x_{(1)} = \sup\{\theta : f(\mathbf{x}|\theta) > 0\}, \quad \text{and}$$
$$x_{(n)} = 1 + \inf\{\theta : f(\mathbf{x}|\theta) > 0\}.$$

For a sufficient statistics $T$, by factorization theorem, $f(\mathbf{x}|\theta) = g(t,\theta)h(\mathbf{x})$. Therefore, for $\mathbf{x}$ such that $h(\mathbf{x}) > 0$,

$$x_{(1)} = \sup\{\theta : g(t,\theta) > 0\} \text{ and } x_{(n)} = 1 + \inf\{\theta : g(t,\theta) > 0\}.$$

We conclude that $(X_{(1)}, X_{(n)})$ is minimal sufficient.

**Example 6.24** (cont. Ex. 6.23, ancillary statistics)

- In Example 6.23, the pdf of $R = X_{(n)} - X_{(1)}$ is
$$n(n-1)r^{n-2}(1-r), \qquad (\Rightarrow \text{Beta}(n-1, 2))$$
for $0 \le r \le 1$, which is **irrelevant** to $\theta$.
  - For any $\theta$, the appearance of the values of $R$ follows the same probabilistic pattern.
  (cf. $T = \sum_{i=1}^n X_i$, where $X_1, \ldots, X_n$ are i.i.d. from Bernoulli($\theta$))
  - The observed values of $R$ do not carry information about $\theta$
  - That is, there exists transformation of the minimum sufficient statistics $X_{(1)}$ and $X_{(n)}$ that contains no information about $\theta$.





- Statistics like $R$ are called **ancillary statistics**, which have distributions free of the parameters and seemingly contain *no* information about the parameters.
- other example of ancillary statistics?

| | |
|---|---|
| $X_1, \cdots, X_n$, i.i.d. $N(\theta, 1)$, | $X_1, X_2$, i.i.d. $Gamma(1, \theta)$, |
| $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$ is ancillary | $Z = \frac{X_1}{X_1+X_2}$ is ancillary |

**Question 6.3**

Note that minimal sufficient statistics may still contain ancillary information. What other property can guarantee sufficient statistics containing no ancillary information?

**Definition 6.15** (completeness, TBp.310)

Let $f(s|\theta)$, $\theta \in \Omega$, be a family of pdfs or pmfs for a statistic $S = S(X_1, \ldots, X_n)$. The family of probability distributions is called **complete** if $\mathrm{E}_\theta[u(S)] = 0$ (or $c$: a constant) for all $\theta \in \Omega$, where $u$ is a function of $S$, implies $u(S) = 0$ (or $c$) with probability 1 for all $\theta \in \Omega$. Equivalently, $S$ is called a **complete statistics**.

$(X_1, \ldots, X_n)$ $S(X_1, \ldots, X_n)$ $u_1(S)$: non-constant function $u_2(S)$: constant function

- $u(S) = c$, $c$ : a constant, is a trivial ancillary statistic and
$$E_\theta[u(S) - c] = 0, \quad \text{for any } \theta.$$
- $S$ is complete $\Leftrightarrow$ $E_\theta[u(S)]$ is a constant for all $\theta$ implies that the trnasformation $u$ is a constant transformation.
- $S$ is complete $\Leftrightarrow$ any transformations of $S$ (except the constant functions) contains *some* information about $\theta$.
- In Example 6.24, $E_\theta(R) = \frac{n-1}{n+1}$. That is,
$$X_{(n)} - X_{(1)} - \frac{n-1}{n+1} = R - E_\theta(R)$$
has mean zero for all $\theta$. $\Rightarrow$ there is a *nonzero* function of $X_{(1)}$ and $X_{(n)}$ whose expectation is zero for all $\theta$.

**Example 6.25** (sufficient and complete statistics of i.i.d. Uniform distribution $U(0,\theta)$)

Let $X_1, \ldots, X_n$ be i.i.d. from Uniform distribution $U(0,\theta)$, $\theta > 0$.

- By factorization theorem, $X_{(n)}$, the largest order statistics, is sufficient.
- The pdf of $X_{(n)}$ is $\frac{nx^{n-1}}{\theta^n} I_{(0,\theta)}(x)$.

  Let $u$ be a function such that $E[u(X_{(n)})] = 0$ for all $\theta$. Then

  which implies $\int_0^\theta u(x)x^{n-1}dx = 0, \quad \text{for all } \theta > 0,$

  $$u(x)x^{n-1} = 0, \quad \text{a.s. for } x \in (0,\infty) \quad \Longrightarrow X_{(n)} \text{ is complete.}$$





**Example 6.26** (sufficient and complete statistic of i.i.d. Poisson distribution)

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from Poisson distribution $P(\lambda)$. Then

$$f(x_1, \ldots, x_n; \lambda) = (e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i})/\prod_{i=1}^n x_i!\,.$$

So $S = \sum_{i=1}^n X_i$ is sufficient for $\lambda$ and $S \sim P(n\lambda)$. If $u(S)$ is a function of $S$ s.t.,

$$0 = \mathrm{E}[u(S)] = e^{-n\lambda}\sum_{s=0}^\infty \frac{u(s)(n)^s}{s!}\lambda^s, \text{ for all } \lambda,$$

then, all coefficients of $\lambda$ are zero and $u(s) = 0$. Hence $S$ is also complete.

**Theorem 6.12**

- A complete and sufficient statistic is minimal sufficient. However, a minimal sufficient statistic is not necessarily complete (e.g., Ex.6.24 in LNp.53).
- If a non-constant function of a sufficient statistic $S = (S_1, \ldots, S_k)$ is ancillary, then $S$ is not complete.

**Definition 6.16** (one-parameter exponential family of probability distributions, TBp.308)

A family of distributions $\{f(x|\theta) : \theta \in \Omega\}$ is a **one-parameter exponential family** if the pdf or pmf is of the form:

$$f(x|\theta) = \begin{cases} \exp[\,c(\theta)\,T(x) + d(\theta) + S(x)] = e^{c(\theta)T(x)}e^{d(\theta)}e^{S(x)}, & x \in A \\ 0, & x \notin A \end{cases}$$

where the set $A$ does not depend on $\theta$.

**Theorem 6.13** (sufficient and complete statistics, one-parameter exponential family, TBp.309)

Suppose $X_1, X_2, \cdots, X_n$ is an i.i.d. sample from a member of the exponential family, the joint probability function is

$$\begin{aligned} f(x_1, \ldots, x_n|\theta) &= \prod_{i=1}^n \exp\left[c(\theta)T(x_i) + d(\theta) + S(x_i)\right] I_A(x_i) \\ &= \exp\left[c(\theta)\sum_{i=1}^n T(x_i) + nd(\theta) + \sum_{i=1}^n S(x_i)\right] \prod_{i=1}^n I_A(x_i), \end{aligned}$$

where $I_A(x_i) = 1$ if $x_i \in A$ and 0 otherwise. Then, $\sum_{i=1}^n T(X_i)$ is a sufficient and complete statistics for $\theta$.

**Example 6.27** (some one-parameter exponential families, TBp.309)

- The pmf of the Bernoulli distribution $B(\theta)$ is
$$P(X = x) = \theta^x(1-\theta)^{1-x} = \exp\left[x \log\left(\frac{\theta}{1-\theta}\right) + \log(1-\theta)\right], \ x = 0, 1.$$
This is a one-parameter exponential family with $T(x) = x$. For i.i.d. $X_1, \ldots, X_n \sim B(\theta)$, $\sum_{i=1}^n X_i$ is sufficient and complete for $\theta$.
- The pmf of the Binomial distribution $B(m, \theta)$ is: for $x \in \{0, \ldots, m\}$,
$$p(X = x) = \binom{m}{x}\theta^x(1-\theta)^{m-x} = \exp\left[x \log\frac{\theta}{1-\theta} + m\log(1-\theta)\right]\binom{m}{x}.$$
This is a one-parameter exponential family with $T(x) = x$. For i.i.d. $X_1, \ldots, X_n \sim B(m, p)$, $\sum_{i=1}^n X_i$ is sufficient and complete for $\theta$.





- The pmf of the Poisson distribution $P(\lambda)$ is
$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp(x \log\lambda - \lambda - \log x!), \quad x = 0, 1, 2, \ldots$$
This is a one-parameter exponential family with $T(x) = x$. For i.i.d. $X_1, \ldots, X_n \sim P(\lambda)$, $\sum_{i=1}^n X_i$ is sufficient and complete for $\lambda$.

**Definition 6.17** (regular $k$-parameter exponential family, TBp.309)

A family of distributions $\{f(x|\Theta) : \Theta \in \Omega \subset \mathbb{R}^k\}$ is called a **regular $k$-parameter exponential family** if the pdf or pmf is of the form

$$f(x;\Theta) = \begin{cases} \exp\left[\sum_{j=1}^k q_j(\Theta)\, T_j(x)\right] c(\Theta)h(x), & x \in A \\ 0, & \text{otherwise} \end{cases}$$

where $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ is $k$-parameter, and the following conditions hold:

1. $A$ does not depend on $\Theta$, and $\Omega$ contains a nonempty, $k$-dimensional open rectangle.
2. $\{(q_1(\Theta), q_2(\Theta), \ldots, q_k(\Theta)) : \Theta \in \Omega\}$ is non-degenerate and $q_j(\Theta)$'s are non-trivial, functionally independent, continuous function of $\Theta$.
3. **(a)** For continuous case, $T_j(x)$'s are linearly independent, continuous functions of $x$ over $A$; **(b)** For discrete case, $T_j(x)$'s are nontrivial functions of $x$, and none is a linear function of the others.

**Theorem 6.14** (sufficient and complete statistics for regular $k$-parameter exponential family)

Let $X_1, X_2, \cdots, X_n$ be an i.i.d. sample from a regular $k$-parameter exponential family, then

$$S_1 = \sum_{i=1}^{n} T_1(X_i), \quad S_2 = \sum_{i=1}^{n} T_2(X_i), \ldots, S_k = \sum_{i=1}^{n} T_k(X_i)$$

is a minimal set of complete and sufficient statistics for $\theta_1, \theta_2, \ldots, \theta_k$.

**Example 6.28** (some regular $k$-parameter exponential families)

- The pdf of the Normal distribution $N(\mu, \sigma^2)$ is

$$\begin{aligned} f(x|\mu,\sigma) &= 1/(\sqrt{2\pi}\sigma)\exp\left[-(x-\mu)^2/(2\sigma^2)\right] \\ &= \exp\left[\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2 - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)\right] \end{aligned}$$

  This is a regular two-parameter exponential family with $T_1(x) = x$ and $T_2(x) = x^2$. Consequently, for an i.i.d. sample $X_1, \ldots, X_n$ from $N(\mu, \sigma^2)$, $(S_1 = \sum_{i=1}^n X_i, S_2 = \sum_{i=1}^n X_i^2)$ is a minimal set of sufficient and complete statistics for $(\mu, \sigma^2)$. Since the relations

$$\frac{S_1}{n} = \overline{X} \equiv \hat{\mu} \quad \text{and} \quad \frac{S_2 - S_1^2/n}{n-1} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} \equiv \hat{\sigma}^2$$

  define a one-to-one transformation, $(\hat{\mu}, \hat{\sigma}^2)$ are also sufficient and complete for $(\mu, \sigma^2)$.





- Exercise: Show that Multinomial distribution, $Multinomial(n, p_1, \ldots, p_r)$, is a regular $(r-1)$-parameter exponential family and find its sufficient and complete statistics. [Hint: subsitute $X_r$ by $n - X_1 - X_2 - \cdots - X_{r-1}$ and $p_r$ by $1 - p_1 - p_2 - \cdots - p_{r-1}$]

Exercise: Check whether other distributions given in LN, Ch1-6, p.58-85, belong to exponential family.

❖ **Reading**: textbook, 8.8, 8.8.1; **Further reading**: Hogg et al., 7.2, 7.4, 7.5, 7.7, 7.8, 7.9

## • criteria for evaluating estimators

**Question 6.4** (choice among different estimators of the same parameter, TBp.298)

- In most statistical estimation problems, there are a variety of possible parameter estimators.
- Among these estimators, how to choose a better ones?
- What properties, that we have defined and discussed for estimators, can be used to evaluate estimators?

  **1.** unbiased **2.** consistency (large sample criterion)

  Note that the two criteria not directly compare the dispersion of estimator. Any other criteria?

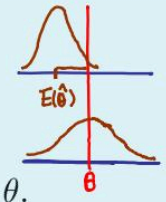
**Question 6.5** (TBp.298)

**criteria related to dispersion of estimator:** conceptually, perfer the estimator whose sampling distribution is *most concentrated* around the true parameter value. How to construct an operational definition, i.e., how to specify a quantitative measure of the dispersion?

**Definition 6.18** (mean square error, TBp.298)

The **mean square error** of an estimator $\hat{\theta}$ at $\theta$ is defined as

$$MSE_\theta(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2].$$



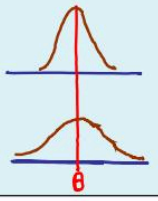

Note that

1. $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta}) + [Bias(\hat{\theta})]^2$, where $Bias(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$.
2. If $\hat{\theta}$ is unbiased, $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta})$

**Definition 6.19** (relative efficiency, TBp.298)

Suppose $\hat{\theta}$ and $\tilde{\theta}$ are two estimators of parameter $\theta$. The **efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$** at $\theta$ is defined as:

$$\text{eff}_\theta(\hat{\theta}, \tilde{\theta}) = \text{Var}_\theta(\tilde{\theta})/\text{Var}_\theta(\hat{\theta}),$$




which is most meaningful when $\hat{\theta}$ and $\tilde{\theta}$ are **both unbiased**.
Note that $\text{eff}(\tilde{\theta}, \hat{\theta}) = 1/\text{eff}(\hat{\theta}, \tilde{\theta})$.





**Notes** (interpretation of relative efficiency, TBp. 298)

1. Finite sample case. $\text{eff}_\theta(\hat{\theta}, \tilde{\theta})$: accuracy of $\hat{\theta}$ relative to accuracy of $\tilde{\theta}$.
   $\Rightarrow$ For $\theta$ s.t. $\text{eff}_\theta(\hat{\theta}, \tilde{\theta}) > 1$, $\hat{\theta}$ has smaller variance than $\tilde{\theta}$ on the $\theta$

2. Large sample case. When $\text{Var}(\hat{\theta}_n) = c_1 n^{-\alpha}(1 + o(1))$ and $\text{Var}(\tilde{\theta}_n) = c_2 n^{-\beta}(1 + o(1))$, where $n$ is the sample size, then

   $$\textbf{asymptotic relative efficiency} \equiv \lim_{n\to\infty} \text{eff}_\theta(\hat{\theta}, \tilde{\theta}) = \begin{cases} c_2/c_1, & \text{if } \alpha = \beta, \\ 0, & \text{if } \alpha < \beta, \\ \infty, & \text{if } \alpha > \beta. \end{cases}$$

3. Relative sample size. When $\text{Var}(\hat{\theta}_n) = c_1 n^{-1}(1 + o(1))$ and $\text{Var}(\tilde{\theta}_m) = c_2 m^{-1}(1 + o(1))$, where $n$ and $m$ are the sample sizes. For $n$ fixed, let $m$ be the smallest sample size such that $\text{Var}(\hat{\theta}_n) \geq \text{Var}(\tilde{\theta}_m)$. Then

   $$\lim_{n\to\infty} \frac{m}{n} \approx \lim_{n\to\infty} \text{eff}_\theta(\hat{\theta}_n, \tilde{\theta}_n) = \frac{c_2}{c_1}.$$

   That is, $\text{eff}_\theta(\hat{\theta}_n, \tilde{\theta}_n)$ is approximately the ratio of sample sizes necessary to obtain the same variance for $\hat{\theta}_n$ and $\tilde{\theta}_m$.

**Example 6.29** (Muon Decay, TBp.299)

- $\tilde{\alpha} = 3\overline{X}$: method of moments estimator
- MLE $\hat{\alpha}$ solves $\sum_{i=1}^{n} X_i/(1+\hat{\alpha}X_i) = 0$
- $\mathrm{Var}(\tilde{\alpha}) = 9\mathrm{Var}(\overline{X}) = (3-\alpha^2)/n$
- $\mathrm{Var}(\hat{\alpha}) \approx [nI(\alpha)]^{-1}$

$$I(\alpha) = \mathrm{E}_\alpha\left[\frac{\partial}{\partial\alpha}\log f(X|\alpha)\right]^2 = \int_{-1}^{1}\frac{x^2}{(1+\alpha x)^2}\left(\frac{1+\alpha x}{2}\right)dx$$

$$= \begin{cases} \frac{1}{2\alpha^3}\left[\log\frac{1+\alpha}{1-\alpha} - 2\alpha\right], & -1<\alpha<1, \alpha\neq 0 \\ \frac{1}{3}, & \alpha = 0 \end{cases}$$

- The asymptotic relative efficiency is thus

$$\lim_{n\to\infty}\mathrm{eff}_\alpha(\tilde{\alpha},\hat{\alpha}) = \lim_{n\to\infty}\frac{\mathrm{Var}(\hat{\alpha})}{\mathrm{Var}(\tilde{\alpha})} = \frac{2\alpha^3}{3-\alpha^2}\left[\log\left(\frac{1+\alpha}{1-\alpha}\right) - 2\alpha\right]^{-1}, \text{ for } \alpha\neq 0.$$

- The following table gives this efficiency for various values of $\alpha$

| $\alpha$ | 0 | 0.1 | 0.2 | 0.3 | 0.4 | $\cdots$ | 0.9 | 0.95 |
|---|---|---|---|---|---|---|---|---|
| $\lim_{n\to\infty}\mathrm{eff}_\alpha(\tilde{\alpha},\hat{\alpha})$ | 1 | 0.997 | 0.989 | 0.975 | 0.953 | $\cdots$ | 0.582 | 0.464 |

- The MLE is not much better than method of moments estimator for $\alpha$ close to 0, but does increasingly better as $\alpha$ tends to 1.





**Question 6.6** (TBp.300)

Is there a lower bound for the $MSE$ of *any* estimator? If such a lower bound exists, what can it help us on the comparison and choice of estimators?

**Ans:** 1. It would function as a benchmark against which estimator could be compared.

2. For estimators achieve the lower bound, they cannot be "improved" upon.

**Theorem 6.15** (Cramer-Rao inequality, TBp.300)

Suppose $X_1,\ldots,X_n$ are i.i.d. with pdf/pmf $f(x|\theta)$, and $T = T(X_1,\ldots,X_n)$ is an unbiased estimator of $\theta$. Then, under smooth assumptions on $f(x|\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{1}{n\, I_{X_1}(\theta)}.$$

**Proof :** Let

$$Z = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(X_i|\theta) = \sum_{i=1}^{n}\frac{\frac{\partial}{\partial\theta}f(X_i|\theta)}{f(X_i|\theta)}.$$

Then $\mathrm{E}_\theta(Z) = 0$, $\mathrm{Var}_\theta(Z) = nI_{X_1}(\theta)$. By the Cauchy-Schwarz inequality,

$$\mathrm{Cov}_\theta^2(T,Z) \leq \mathrm{Var}_\theta(T)\mathrm{Var}_\theta(Z).$$

But

$$\begin{aligned}
\mathrm{Cov}_\theta(T,Z) &= \mathrm{E}_\theta(TZ) - \mathrm{E}_\theta(T)\mathrm{E}_\theta(Z) \\
&= \int\cdots\int T(x_1,x_2,\ldots,x_n)\left[\sum_{i=1}^n \frac{\frac{\partial}{\partial\theta}f(x_i|\theta)}{f(x_i|\theta)}\right]\left[\prod_{j=1}^n f(x_j|\theta)\right] dx_1dx_2\cdots dx_n \\
&= \int\cdots\int T(x_1,x_2,\ldots,x_n)\left[\frac{\partial}{\partial\theta}\prod_{i=1}^n f(x_i|\theta)\right] dx_1dx_2\cdots dx_n \\
&= \frac{\partial}{\partial\theta}\int\cdots\int T(x_1,x_2,\ldots,x_n)\left[\prod_{i=1}^n f(x_i|\theta)\right] dx_1dx_2\cdots dx_n \\
&= \frac{\partial}{\partial\theta}\mathrm{E}_\theta(T) = \frac{\partial}{\partial\theta}\,\theta = 1 \Rightarrow 1 \le \mathrm{Var}_\theta(T)\mathrm{Var}_\theta(Z) \Rightarrow \mathrm{Var}_\theta(T) \ge \frac{1}{\mathrm{Var}_\theta(Z)} = \frac{1}{nI_{X_1}(\theta)}
\end{aligned}$$

and we have completed the proof.

**Notes**

1. An unbiased estimator whose variance achieves this lower bound will have the smallest $MSE$ among *all unbiased estimators.*

2. Theorem 6.15 does not preclude the possibility that there is a biased estimator of $\theta$ that has a smaller $MSE$ than the unbiased estimator that achieve the lower bound.





**Definition 6.20** (efficiency, efficient, asymptotically efficient, TBp.302)

- The **efficiency** of an unbiased estimator $\hat\theta$ of $\theta$ is
$$\mathrm{eff}_\theta(\hat\theta) = \frac{1/nI_{X_1}(\theta)}{\mathrm{Var}_\theta(\hat\theta)} = \frac{1}{nI_{X_1}(\theta)\mathrm{Var}_\theta(\hat\theta)}.$$
- An unbiased estimator whose variance achieves the lower bound is called **efficient**.
- If the asymptotic variance of an estimator equals the lower bound, the estimator is said to be **asymptotically efficient**.

**Notes** (TBp.302)

1. MLE is asymptotically efficient.
2. For a finite sample size, MLE may not be efficient.
3. MLEs are not the only asymptotically efficient estimators.
4. There may exist biased and super efficient estimators. For example, if $X_1,\ldots,X_n$ is an iid sample $\sim$ Normal$(\mu,1)$. Let $\tilde\mu = \overline{X}\, I(|\overline{X}| > n^{-1/4})$. Then
$$\sqrt{n}(\tilde\mu - \mu) \xrightarrow{D} N(0,\sigma^2(\mu))$$
where $\sigma^2(\mu) = 1$ if $\mu \neq 0$ and $\sigma^2(0) = 0$.

**Example 6.30** (Poisson distribution, TBp.302)

For the Poisson distribution, $I_{X_1}(\lambda) = 1/\lambda$. For any unbiased estimator $T$ of $\lambda$, based on a sample of i.i.d. Poisson variables $X_1, X_2, \ldots, X_n$,

$$\mathrm{Var}_\lambda(T) \geq \lambda/n\,.$$

MLE of $\lambda$ is $\overline{X} = S/n$ and

$$\mathrm{Var}_\lambda(\overline{X}) = \lambda/n.$$

Hence $\overline{X}$ attains the Cramer-Rao lower bound. (**Note.** There may exist a biased estimator of $\lambda$ that has a smaller MSE than $\overline{X}$.)

**Theorem 6.16** (generalization of Cramer-Rao inequality)

Suppose $X_1, \ldots, X_n$ are i.i.d. with pdf/pmf $f(x|\theta)$, and $T = T(X_1, \ldots, X_n)$ is an unbiased estimator of $\tau(\theta)$. Then, under smooth assumptions on $f(x|\theta)$,

$$\mathrm{Var}_\theta(T) \geq \frac{\tau'(\theta)^2}{nI_{X_1}(\theta)}\,.$$

**Example 6.31** (generalized negative binomial, TBp.302-305)

- The generalized negative binomial distribution has the pmf:

$$f(x|m,k) = \left(\frac{k}{m+k}\right)^k \frac{\Gamma(k+x)}{x!\Gamma(k)} \left(\frac{m}{m+k}\right)^x, \quad x = 0, 1, \ldots.$$

  Note that its mean is $\mu = m$ and its variance is $\sigma^2 = m + m^2/k$





$$\left.\begin{array}{l} m: \ \#\text{ of black balls} \\ k: \ \#\text{ of white balls} \end{array}\right] \Longrightarrow X = \begin{array}{l}\text{draw with replacement} \\ \#\text{ of black balls that will be drawn before} \\ \text{the } k\text{th white ball is drawn}\end{array}$$

- Generalized negative binomial model arises in several cases.
  1. If $k$ is an integer, then the number of failures up to the $k$th success in a sequence of Bernoulli trails with probability of success $p = k/(m+k)$ follows a generalized negative binomial distribution.
  2. Hierarchical model: If $\Lambda \sim \mathrm{Gamma}(k, k/m)$, and $X|\Lambda = \lambda \sim \mathrm{Poisson}(\lambda)$, then $X \sim$ generalized negative binomial.
  3. Biological model (Solomon, 1983): If $N$, the number of clusters/colonies, follows a Poisson($\lambda$) distribution and $X_i$, $i = 1, \ldots, N$, the numbers of individuals in each colonies, follows a logarithmic series distribution with success probability $p$

$$P(X_i = t) = \frac{-1}{\log(p)} \frac{(1-p)^t}{t}, \ t = 1, 2, \cdots.$$

     Assume that the $X_i$'s are independent. Then the distribution of the total number of individuals is generalized negative binomial with $k = -\lambda/\log(p)$ and $p = m/(m+k)$.
  4. birth, death rate = constant; immigration rate = constant
     Population size $\sim$ generalized negative binomial

- Estimation of $m$ and $k$

  method 1: method of moments estimators

  $$\hat{m} = \overline{X}, \quad \hat{k} = \frac{\overline{X}^2}{\hat{\sigma}^2 - \overline{X}}$$

  method 2:

  $n_0$ = number of zeros out of a sample of size $n$

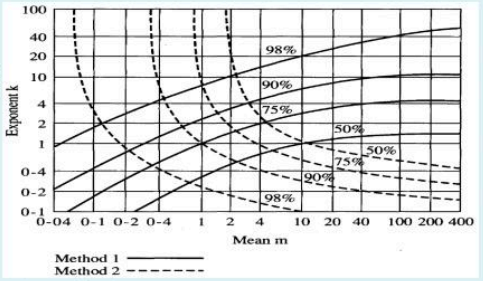
  $p_0 = \left(\frac{k}{m+k}\right)^k$ = probability of zero count

  If $m$ is estimated by $\overline{X}$, then $k$ can be estimated by solving the equation

  $$\frac{n_0}{n} = \left(\frac{\hat{k}}{\hat{k}+\overline{X}}\right)^{\hat{k}} = \left(1+\frac{\overline{X}}{\hat{k}}\right)^{-\hat{k}}.$$

  method 3: MLE is difficult to compute. MLE of $m$ is the sample mean $\overline{X}$, but MLE of $k$ is the solution of a nonlinear equation.

- Asymptotic relative efficiencies of estimators (**Figure 8.11 in textbook**)
  - — Method 2 is quite efficient when $p_0$ close to 1 or 0.
  - — Method 1 becomes more efficient as $k$ increases.








- **Insect counts data, Bliss and Fisher (1953)**
  - 25 leaves from each of 6 apple trees in an sprayed orchard
  - number of adult female red mites on each of the leaves

| number per leaf | observed count | Poisson fit | Negative Binomial fit |
|---|---|---|---|
| 0 | 70 | 47.7 | 69.5 |
| 1 | 38 | 54.6 | 37.6 |
| 2 | 17 | 31.3 | 20.1 |
| 3 | 10 | 12.0 | 10.7 |
| 4 | 9 | 3.4 | 5.7 |
| 5 | 3 | 0.75 | 3.0 |
| 6 | 2 | 0.15 | 1.6 |
| 7 | 1 | 0.03 | 0.85 |
| $8^+$ | 0 | 0.00 | 0.95 |

  - Recursive algorithm for pmf of generalized negative binomial:

    $$f(0|m,k) = \left(1+\frac{m}{k}\right)^{-k}, \quad f(n|m,k) = \frac{k+n-1}{n}\left(\frac{m}{k+m}\right) f(n-1|m,k)$$

  - Poisson fit is not good. Heterogeneous: the rates of insect infection might be different on different trees and at different locations on the same tree. (For the Poisson fit, polling the last 3 cells: $\chi^2 = 82.4$, df=5, extremely small $p$-value; For the Negative Binomial fit, pooling last 2 cells: $\chi^2 = 2.48$, df=5, $p$-value=0.22)

❖ **Reading**: textbook, 8.7; **Further reading**: Hogg et al., 6.2

## • method of finding estimator III --- UMVUE (or MVUE)

**Question 6.7** ("best" unbiased estimator)

Among *all* unbiased estimators, how to find the "best" estimator, i.e., the one that has the smallest $MSE$?

(**Recall.** [1]. $MSE_\theta(\hat{\theta}) = Var_\theta(\hat{\theta})$ for any unbiased estimator $\hat{\theta}$;
[2]. Cramer-Rao bound)

**Definition 6.21** (UMVUE)

An estimator $T^*$ of $\tau(\theta)$ is called a **(uniformly) minimum variance unbiased estimator (UMVUE or MVUE)** of $\tau(\theta)$ if

1. $T^*$ is unbiased for $\tau(\theta)$, and
2. for any unbiased estimator $T$ of $\tau(\theta)$, $\mathrm{Var}_\theta(T^*) \le \mathrm{Var}_\theta(T)$ for all $\theta \in \Omega$.

**Example 6.32** (cont. Ex. 6.30, UMVUE of Poisson mean, TBp.302)

- The Cramer-Rao bound for $\tau(\theta) = \theta$ is $\theta/n$.
- On the other hand, $\mathrm{E}(\overline{X}) = \theta$ and $\mathrm{Var}(\overline{X}) = \theta/n$.
- Hence, $\overline{X}$ is a UMVUE of $\theta$.





**Question 6.8**

A UMVUE may not achieve the Cramer-Rao lower bound. Are there other *systematic procedures* that can be used to derive/identify a UMVUE among a large number of unbiased estimators?

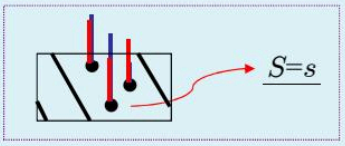
**Theorem 6.17** (Rao-Blackwell theorem, TBp.310)

Suppose $X_1, \ldots, X_n$ have a joint pdf or pmf $f(x_1, \ldots, x_n; \theta)$, and

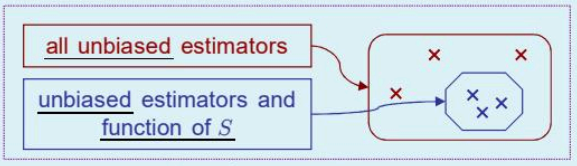
$$S = (S_1, \ldots, S_k)$$

is sufficient for $\theta$. If $T$ is any estimator of $\tau(\theta)$ with $\mathrm{E}(T^2) < \infty$, let

$$T^* = \mathrm{E}_\theta(\, T \mid S \,).$$




Then,




(1). $\mathrm{E}_\theta(T^*) = \mathrm{E}_\theta(T)$ for any $\theta$,

(2). $T^*$ is a function of $S$ alone and does not involve $\theta$, i.e., $T^*$ is a statistics,

(3). $E_\theta(T^* - \tau(\theta))^2 \le E_\theta(T - \tau(\theta))^2$, i.e., $MSE_\theta(T^*) \le MSE_\theta(T)$, for every $\theta$ and the equality is strict unless $T = T^*$.
(Q: Why $T^*$ better?)

**Proof.**

1. $\mathrm{E}(T^*) = \mathrm{E}[\mathrm{E}(T|S)] = \mathrm{E}(T)$
2. $S$ is sufficient for $\theta \Rightarrow$ distribution of $X_1, \ldots, X_n | S$ not involve $\theta$
   $\Rightarrow$ conditional distribution $T|S$ does not involve $\theta$
   $\Rightarrow T^* = \mathrm{E}(T|S)$ is a function of $S$ only
3. $$\begin{aligned} \mathrm{Var}(T) &= \mathrm{Var}[\mathrm{E}(T|S)] + \mathrm{E}[\mathrm{Var}(T|S)] \\ &\geq \mathrm{Var}[\mathrm{E}(T|S)] = \mathrm{Var}(T^*) \end{aligned}$$
   with equality if and only if $\mathrm{E}[\mathrm{Var}(T|S)] = 0$, i.e. $\mathrm{Var}(T|S) = 0$ with probability 1, which is the case only if $T$ is a function of $S$, which would imply $T = T^*$.

**Question 6.9** (uniqueness)

Suppose that two estimates $T_1$ and $T_2$ are unbiased (or have same bias). Which of $E(T_1|S)$ and $E(T_2|S)$, where $S$ is sufficient, has smaller variance? Is it possible that there exists only one *unique* function of $S$ that is unbiased (or has same bias)?

**Theorem 6.18**

If the distribution of $S$ is complete and $\mathrm{E}[u_1(S)] = \tau(\theta)$, $\mathrm{E}[u_2(S)] = \tau(\theta)$ for all $\theta$, then $\mathrm{E}[u_1(S) - u_2(S)] = 0$ for all $\theta$, which implies that $u_1(S) = u_2(S)$ with probability 1 for any $\theta$.





**Theorem 6.19** (Lehmann-Scheffe theorem)

Suppose that $X_1, X_2, \ldots, X_n$ have joint pdf/pmf $f(x_1, x_2, \ldots, x_n; \theta)$ and $S$ is a complete and sufficient statistic for $\theta$. If $T^* = T^*(S)$ is a function of $S$ and unbiased for $\tau(\theta)$. Then $T^*$ is the unique UMVUE of $\tau(\theta)$.

**Proof.** Suppose $T$ is any other statistic with $\mathrm{E}(T) = \tau(\theta)$ for all $\theta$. Then $\mathrm{E}(T|S)$ is also unbiased for $\tau(\theta)$ and a function of $S$. Completeness implies that any statistic that is a function of $S$ and unbiased for $\tau(\theta)$ must be equal to $T^*$ with probability 1. Hence, $T^* = \mathrm{E}(T|S)$ with probability 1 and

$$\mathrm{Var}(T^*) \leq \mathrm{Var}(T) \text{ for all } \theta.$$

Therefore $T^*$ is the unique UMVUE of $\tau(\theta)$.

**Summary** (methods of finding UMVUE)

1. If $\mathrm{E}(T) = \tau(\theta)$ for all $\theta$ and $\mathrm{Var}(T)$ achieves the Cramer-Rao lower bound. Then $T$ is a UMVUE of $\tau(\theta)$. (**Note.** A UMVUE may not attain Cramer-Rao lower bound.)
2. If $S$ is a complete sufficient statistic, and $u(S)$ satisfies $\mathrm{E}[u(S)] = \tau(\theta)$ for all $\theta$. Then $u(S)$ is a UMVUE of $\tau(\theta)$.
3. If $T$ is an unbiased estimator for $\tau(\theta)$ and $S$ is a complete sufficient statistic for $\theta$. Then $\mathrm{E}(T|S)$ is a UMVUE of $\tau(\theta)$.

**Example 6.33** (cont. Ex. 6.32, UMVUE of Poisson mean)

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from Poisson($\lambda$) distribution. Then $S = \sum_{i=1}^n X_i$ is a sufficient and complete statistics for $\lambda$. Because $\overline{X} = S/n$ is a function of $S$ and $E(\overline{X}) = \lambda$, $\overline{X}$ is UMVUE of $\lambda$.

**Example 6.34** (UMVUE of Normal mean and variance)

Let $X_1, \ldots, X_n$ be i.i.d. random variables from Normal distribution $N(\mu, \sigma^2)$, then

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

are sufficient and complete statistics for $(\mu, \sigma^2)$. Clearly, $E(\overline{X}) = \mu$ and $E(S^2) = \sigma^2$ (**Note.** $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1} \Rightarrow E((n-1)S^2/\sigma^2) = n-1$), so $\overline{X}$ and $S^2$ are unbiased estimator of $\mu$ and $\sigma^2$, respectively. Since they depend only on the sufficient and complete statistics, they are UMVUE.

**Example 6.35** (UMVUE may not attain Cramer-Rao lower bound)

Let $X \sim$ Poisson($\lambda$). Let $T(X) = 1$ if $X = 0$ and $T(X) = 0$ otherwise. Then, $T$ is UMVUE of $\tau(\lambda) = e^{-\lambda}$ and $\text{Var}(T) = e^{-\lambda}(1 - e^{-\lambda})$. The Cramer-Rao lower bound for $e^{-\lambda}$ is $e^{-2\lambda}/I(\lambda) = e^{-2\lambda}\lambda$. Hence

$$\text{Var}(T) - e^{-2\lambda}\lambda = e^{-\lambda}(1 - e^{-\lambda} - e^{-\lambda}\lambda) = e^{-\lambda}P(X \geq 2) \geq 0$$

and Cramer-Rao lower bound is not attained.





**Example 6.36** (UMVUE of Poisson zero probability)

Suppose $X_1, \ldots, X_n$ is a sample from a Poisson($\lambda$) distribution. Then $S = \sum_{i=1}^n X_i = n\overline{X}$ is complete and sufficient for $\lambda$. To find a UMVUE of $e^{-\lambda}$, start with $I(X_1 = 0)$, which is an unbiased estimator. Since

$$\begin{aligned} \text{E}[I(X_1 = 0)|S = s] &= 1 \cdot P(X_1 = 0|S = s) = \frac{P(X_1 = 0, \sum_{i=2}^n X_i = s)}{P(\sum_{i=1}^n X_i = s)} \\ &= \frac{e^{-\lambda}(e^{-(n-1)\lambda}[(n-1)\lambda]^s/s!)}{e^{-n\lambda}(n\lambda)^s/s!} = (1 - \frac{1}{n})^s, \end{aligned}$$

the UMVUE of $e^{-\lambda}$ is $(1 - n^{-1})^{n\overline{X}}$.

**Example 6.37** (cont. Ex. 6.25, UMVUE of Uniform upper bound, c.f. Ex. 6.13)

Suppose $X_1, \ldots, X_n$ is an i.i.d. sample from Uniform distribution $U(0, \theta)$. Because $X_{(n)}$ is sufficient and complete, and $E(X_{(n)}) = \frac{n}{n+1}\theta$, $\frac{n+1}{n}X_{(n)}$ is unbiased and is the UMVUE of $\theta$.

**Theorem 6.20**

Suppose that $S_1$ and $S_2$ are sufficient statistics for $\theta$, and $S_2 = g(S_1)$. If $U$ is an unbiased estimator for $\tau(\theta)$, let $V_1 = E(U|S_1)$ and $V_2 = E(U|S_2)$ then

$$\text{Var}(V_2) \leq \text{Var}(V_1).$$

❖ **Reading**: textbook, 8.8.2; **Further reading**: Hogg et al., 7.1, 7.3, 7.6