check the binomial case (Ex6.27, LNp.55)

$(X_1, \cdots, X_r)$

- **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.

$\boxed{4/1}$

❖ **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. under same statistical modeling
- Among these estimators, how to choose a better ones? $\boxed{\text{criteria}=?}$
- What properties, that we have defined and discussed for estimators, can be used to evaluate estimators?

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

$E_\theta(\hat\theta - \theta) = 0 \Leftrightarrow E_\theta(\hat\theta) = \theta$ $\qquad \hat\theta_n \xrightarrow[n\to\infty]{P} \theta$

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)

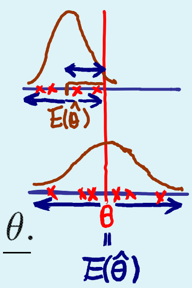
The **mean square error** of an estimator $\hat\theta$ at θ is defined as

$E(Y - g(X))^2$, LN, CH1~6, P.53 $\xleftrightarrow{\text{cf.}}$ $MSE_\theta(\hat\theta) = E_\theta[(\overbrace{\hat\theta - \theta}^{-E(\hat\theta)+E(\hat\theta)})^2]$. ← a function of θ

error

Note that $\boxed{\text{r.v.}}$ LN, CHI~6, p.42, item5

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, $MSE_\theta(\hat\theta) = Var_\theta(\hat\theta)$

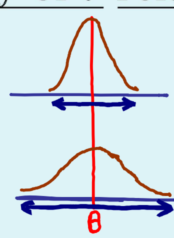


Definition 6.19 (relative efficiency, TBp.298)

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

a function of θ → $\text{eff}_\theta(\hat\theta, \tilde\theta) = \text{Var}_\theta(\tilde\theta)/\text{Var}_\theta(\hat\theta)$, $\left\{ \begin{array}{l} >1 \\ =1 \\ <1 \end{array} \right.$

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



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 θ s.t. $\text{eff}_\theta(\hat\theta,\tilde\theta) > 1$, $\hat\theta$ has smaller variance than $\tilde\theta$ on the θ
   dispersion

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

   ☆ $\boxed{\text{For good estimators, usually: ① } \text{Var} \sim O(n^{-1}) \text{ ② } (\text{bias})^2 \sim o(n^{-1}) \text{ (e.g., LNp.40)}}$

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

   ☆ $\boxed{c_1, c_2\text{: functions of } \theta.}$

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

   i.e. the sample size m s.t. $\text{Var}(\hat\theta_n) \approx \text{Var}(\tilde\theta_m)$ $\boxed{\text{a function of } n}$

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

   depends on θ

   $1 \approx \text{Var}(\tilde\theta_m)/\text{Var}(\hat\theta_n) \approx \dfrac{c_2 m^{-1}}{c_1 n^{-1}} \Rightarrow \dfrac{c_2}{c_1} \approx \dfrac{m}{n}$

   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) ← (cont. Ex6.16, LNp.27)

- $\tilde\alpha = 3\overline{X}$: method of moments estimator
- MLE $\hat\alpha$ solves $\sum_{i=1}^n X_i/(1+\hat\alpha X_i) = 0$ ← no close form
- $\text{Var}(\tilde\alpha) = 9\text{Var}(\overline{X}) = (3-\alpha^2)/n$ ← $\boxed{n^{-1}}$
- $\text{Var}(\hat\alpha) \approx [nI(\alpha)]^{-1}$ ← check LNp.41, Note 1

$\boxed{\text{When sample size is large}}$

$\boxed{\text{Does not mean } \hat\alpha \text{ is always closer to } \alpha \text{ than } \tilde\alpha \text{ for any observed data.}}$

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

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

$\boxed{\text{LNp.60, } \underline{2.}\ \alpha=\beta=-1,\ c_2/c_1 \approx 1/2;\ \underline{3.}\ m/n \approx 1/2}$

- The asymptotic relative efficiency is thus

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

- The following table gives this efficiency for various values of α

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

$\boxed{\leq 1}$

- The MLE is not much better than method of moments estimator for α close to 0, but does increasingly better as α 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. on MSE

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 θ. Then, under smooth assumptions on $f(x|\theta)$,

$$\text{MSE}_\theta(T) = \text{Var}_\theta(T) \geq \boxed{\frac{1}{n\, I_{X_1}(\theta)}}$$

lower bound of MSE for unbiased estimators

called C.-R. bound (C.f., asymptotic variance of MLE)

Proof : Let

score function, It's a r.v., but not a statistic

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

$\ell'(\theta)$

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

Check the proof of Thm 6.6 (LNp.40)

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

$$0 \leq \rho_{T,Z}^2 = \frac{\text{cov}^2(T,Z)}{\text{Var}(T)\text{Var}(Z)} \leq 1$$

But

$$\text{Cov}_\theta(T, Z) = \text{E}_\theta(TZ) - \underbrace{\text{E}_\theta(T)\text{E}_\theta(Z)}_{=0}$$

$$= \int \cdots \int T(x_1, x_2, \ldots, x_n) \underbrace{\left[\sum_{i=1}^n \frac{\frac{\partial}{\partial\theta} f(x_i|\theta)}{f(x_i|\theta)}\right]}_{Z} \left[\prod_{j=1}^n f(x_j|\theta)\right] dx_1 dx_2 \cdots dx_n$$

$$= \int \cdots \int T(x_1, x_2, \ldots, x_n) \left[\frac{\partial}{\partial\theta} \prod_{i=1}^n f(x_i|\theta)\right] dx_1 dx_2 \cdots dx_n$$

$$= \left[\tfrac{\partial}{\partial\theta} f(x_1|\theta)\right] f(x_2|\theta) \times \cdots \times f(x_n|\theta) + f(x_1|\theta)\left[\tfrac{\partial}{\partial\theta} f(x_2|\theta)\right] \times \cdots \times f(x_n|\theta) + \cdots$$

$$= \frac{\partial}{\partial\theta} \int \cdots \int T(x_1, x_2, \ldots, x_n) \left[\prod_{i=1}^n f(x_i|\theta)\right] dx_1 dx_2 \cdots dx_n$$

$\because T$ is unbiased

$$= \frac{\partial}{\partial\theta} \text{E}_\theta(T) = \frac{\partial}{\partial\theta}\theta = 1 \Rightarrow 1 \leq \text{Var}_\theta(T)\text{Var}_\theta(Z) \Rightarrow \text{Var}_\theta(T) \geq \frac{1}{\text{Var}_\theta(Z)} = \frac{1}{nI_{X_1}(\theta)}$$

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*.

   It's called UMVUE

   $MSE = 1/nI_{X_1}$

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

   Note $MSE = \text{Var}(\hat{\theta}) + \text{bias}^2$