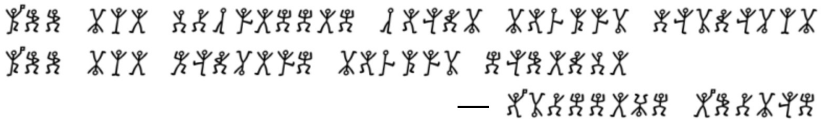
Lecture 3. Admissibility, Unbiasedness & UMVUE

[1] *School of Data Science, The Chinese University of Hong Kong, Shenzhen (CUHK-Shenzhen)*

1. Admissibility



Definition 1.1 (♣**Admissible**). An estimator δ is said to be inadmissible if there exists another estimator δ' which dominates δ (that is, such that $R(\theta, \delta') \leq R(\theta, \delta)$ for all $\theta \in \Theta$, with strict inequality for some $\theta \in \Theta$.) and admissible if no such estimator δ' exists.

We are longing for finding an admissible estimator, and the following information inequality can sometimes provides help.

1.1. Convex Loss Function

Theorem 1.2 (♣ **Rao-Blackwell Theorem**). *Let X be a radom variable with distribution $f_\theta \in \mathcal{F} = \{f_{\theta'}, \theta' \in \Theta\}$, and let T be a sufficient statistic for $\mathcal{F}$. Let δ be an estimator of an estimand $g(\theta)$, and let the loss function $L(g(\theta), \delta(x))$ be a strictly convex function of $\delta(x)$. Then, if δ has finite expectation and risk,*

$$R(g(\theta), \delta) = \mathbb{E}\Big[L\Big(g(\theta), \delta(X)\Big)\Big] < \infty,$$

and if for arbitrary t, we define

$$\eta(t) = \mathbb{E}\Big[\delta(X) | T = t\Big],$$

the risk of the estimator $\eta(T)$ statisfies

$$R\big(g(\theta), \eta\big) < R\big(g(\theta), \delta\big) \tag{1.1}$$

unless $\delta(X) = \eta(T)$ with probability 1.

Proof of Theorem.1.2. Since L is strictly convex in the second coordinate, so by Jensen's inequality, we have

$$L\left\{g(\theta),\eta(t)\right\} = L\Big\{g(\theta),\mathbb{E}\Big[\delta(X)|T=t\Big]\Big\} < \mathbb{E}\Big\{L\Big[g(\theta),\delta(X)\Big]\Big|T=t\Big\},$$

where the expectation is taking with repect to the condition distribution of $X|T=t$, and according to the Jensen's inequality, we know the equality sign holds only when $\eta(t)=\delta(x)$. Since t is arbitrary, so taking the expectation on both sides of this inequality yields

$$R\big(g(\theta),\eta\big) = \mathbb{E}\left\{L\big[g(\theta),\eta\big]\right\} < \mathbb{E}\Big\{\mathbb{E}\Big\{L\Big[g(\theta),\delta(X)\Big]\Big|T\Big\}\Big\} = R\big(g(\theta),\delta\big),$$

and the equality sign holds unless $\delta(X)=\eta(T)$ with probability 1. □

Some points concerning this result are worth noting.

1. Sufficiency of T is used in the proof only to ensure that $\eta(T)$ does not depend on θ and hence is an estimator.
2. If the loss function is convex but not strictly convex, the theorem remains true provided the inequality sign in (1.1) is replaced by "$\leq$". Even in that case, the theorem still provides information because it shows that the particular eatimator $\eta(T)$ is at least as good as $\delta(X)$.
3. The theorem is not true if the convexity assumption is dropped.

Theorem 1.3 (Uniqueness of Admissible Estimator Under Strictly Convex Loss Function). *If L is strictly convex and δ is an admissible estimator of $g(\theta)$, and if δ' is another estimator with the same risk function, that is satisfying $R(g(\theta),\delta)=R(g(\theta),\delta')$ for all θ, then $\delta'=\delta$ with probability 1.*

Proof of Theorem.1.3. Define $\delta^*=\frac{1}{2}(\delta+\delta')$, then by the strictly convex property of the loss function, we have

$$R(\theta,\delta*) < \frac{1}{2}[R(g(\theta),\delta)+R(g(\theta),\delta')] = R(g(\theta),\delta) \tag{1.2}$$

unless $\delta'=\delta$ with probability 1, and (1.2) contradicts the admissibility of δ. □

• *Example* 1.4. Consider two unbiased estimation δ,δ' of $g(\theta)$, if δ is admissible and $\mathrm{Var}(\delta)=\mathrm{Var}(\delta')$ for $\forall\theta\in\Theta$, then $\delta=\delta'$ with probability 1.

Answer. Apparently, $\mathrm{MSE}(\delta)=\mathbb{E}(\delta-g(\theta))^2=\mathrm{Var}(\delta)+bias(\delta)^2=\mathrm{Var}(\delta)$, and $L(g(\theta),\delta)=(\delta-g(\theta))^2$, i.e., the quadratic loss, is a strictly convex function, so by Theorem.1.3 we have $\mathbb{P}(\delta=\delta')=1$. □

This example tells that, when we use MSE as our risk function, i.e., using quadratic loss function, and when we restrict our attention to unbiased estimator, all we have to evaluate is the variance of the estimator in order to find the minimal risk estimator (if it do exist).

1.2. Shannon Entropy, Relative Entropy and Kullback-Leibler Divergence

Here, for simplicity of our argument, we assume the pdf. or pmf. appeared in this subsection, i.e., $p(x|\theta)$, $q(x|\theta)$, etc., have support $\mathbb{R}$, i.e., $p(x|\theta) > 0$ and $q(x|\theta) > 0$ for all $x \in \mathbb{R}$.

Definition 1.5 (♣ **Shannon Entropy**). The Shannon entropy of a random variable X with pdf. or pmf. $p(x|\theta)$ is defined as $H(X) = -\mathbb{E}\big[\log p(X|\theta)\big]$.

Similar to Shannon entropy, we may define an entropy between two distribution function $p(x|\theta)$ and $q(x|\theta)$, which is commonly called the relative entropy, or Kullback-Leibler divergence (KL-divergence),

Definition 1.6 (♣ **KL-Divergence**). Assume p and q are two distribution functions, the KL-divergence of q from p, or the relative entropy of q with respect to p, is defined as

$$DL(p\|q) = -\int \left(\log \frac{q(x|\theta)}{p(x|\theta)}\right) \cdot p(x|\theta)dx = -\mathbb{E}_p\left[\log\left(\frac{q(X|\theta)}{p(X|\theta)}\right)\right],$$

where the subscript p in $\mathbb{E}_p$ indicates that the random variable X appeared in the form follows distribution $p(x|\theta)$.

Since logarithm function $\log(\cdot)$ is concave, and $-\log(\cdot)$ is convex, so by Jensen's inequality, we conclude that

$$\begin{aligned} DL(p\|q) &= \mathbb{E}_p\left[-\log\left(\frac{q(X|\theta)}{p(X|\theta)}\right)\right] \geq -\log\left(\mathbb{E}_p\left[\frac{q(X|\theta)}{p(X|\theta)}\right]\right) \\ &= -\log\left(\int \frac{q(x|\theta)}{p(x|\theta)} \cdot p(x|\theta)dx\right) = -\log\left(\int q(x|\theta)dx\right) = 0. \end{aligned} \tag{1.3}$$

The equality sign holds iff $p \equiv q$ a.s.. The inequality (1.3) or its variational from (1.4) in the following

$$H(X) = -\mathbb{E}_p\big[\log p(X|\theta)\big] \leq -\mathbb{E}_p\big[\log q(X|\theta)\big] \quad \text{when } X \sim p(x|\theta). \tag{1.4}$$

is called the information inequality. The KL divergence measures the difference between the two distributions by calculating the expected logarithmic difference between the probabilities assigned by p and q to each outcome x. More importantly, under the Fisher Information regularity condition (recall notes 1) of the family $\mathcal{F} = \{p(x|\theta), \theta \in \Theta\}$ and assume that

$$\frac{\partial^2}{\partial\theta^2}\int p(x|\theta)dx = \int \frac{\partial^2}{\partial\theta^2} p(x|\theta)dx \quad \text{holds for } \forall\theta \in \Theta.$$

We then have the second Bartlett's Identities holds for the distribution function of X, i.e., $p(x|\theta)$. If we further assume that

$$\frac{\partial^2}{\partial\theta^2}\int \log p(x|\theta)dx = \int \frac{\partial^2}{\partial\theta^2} \log p(x|\theta)dx \quad \text{holds for } \forall\theta \in \Theta.$$

Thus for arbitrary $\theta_0 \in \Theta$, we have

$$I(\theta_0) = -\mathbb{E}_p \left[\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} \right] \bigg|_{\theta=\theta_0} = \frac{\partial^2}{\partial \theta^2} DL(p_\theta \| p_{\theta_0}) \Big|_{\theta=\theta_0}.$$

1.3. Fisher Information and Cramér-Rao Lower Bound

One important application of the Fisher information is its appearence in the Cramér-Rao lower bound, which is a useful benchmark to measure the performance of a given unbiased estimator δ. If the variance of δ is close to the Cramér-Rao lower bound for all θ, not much further improvement is possible. In general, the Cramér-Rao lower bound is not sharp (see Example.3.9), but it's relatively simpler to calculate compare to other lower bounds (such as the local minimum variance unbiased estimator).

The idea of Cramér-Rao lower bound start from a simple covariance property. For any unbiased estimator δ of $g(\theta)$ which we want to evaluate, and any function $\psi(x, \theta)$ with a finite second moment, the Cauchy inequality states that

$$\mathrm{Var}(\delta) \geq \frac{[\mathrm{Cov}(\delta, \psi)]^2}{\mathrm{Var}(\psi)} \tag{1.5}$$

In general, this inequality is not helpful since the right side also involves δ. However, when $\mathrm{cov}(\delta, \psi)$ depends on δ only through $E_\theta(\delta) = g(\theta)$, (1.5) does provide a lower bound for the variance of all unbiased estimators of $g(\theta)$. As a matter of fact, when we assume

$$\frac{\partial}{\partial \theta} \int \delta(x) \cdot p(x|\theta) dx = \int \frac{\partial}{\partial \theta} \delta(x) \cdot p(x|\theta) dx \quad \text{holds for } \forall \theta \in \Theta. \tag{1.6}$$

Then under the Fisher Information regularity condition and by putting $\psi = s(\theta|x) = \partial \log p(x|\theta)/\partial\theta$, we have $\mathrm{Var}(\psi) = I(\theta)$ and

$$\begin{aligned}\mathrm{Cov}(\delta, \psi) &= \mathbb{E}\big(\delta \cdot s(\theta|X)\big) - \mathbb{E}\big(\delta\big) \cdot \mathbb{E}\big(s(\theta|X)\big) = \mathbb{E}\big(\delta \cdot s(\theta|x)\big) \\ &= \int \delta(x) \cdot \frac{1}{p(x|\theta)} \cdot \frac{\partial p(x|\theta)}{\partial \theta} \cdot p(x|\theta) dx = \frac{\partial}{\partial \theta} \mathbb{E}\big(\delta(X)\big) = g'(\theta).\end{aligned}$$

Hence the following theorem,

Theorem 1.7 (♣ **Cramér-Rao Lower Bound**)**.** *Assume the pdf. or pmf. of X satisfy the Fisher Information regularity condition and an unbiased estimator $\delta(X)$ of $g(\theta)$ satisfy (1.6), then*

$$\mathrm{Var}(\delta) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

1.3.1. Multi-Dimensional Cramér-Rao Lower Bound

Notice that, the above argument may easily extend to multi-dimensional case. We first give the definition of two matrix $A \geq B$. Suppose $A = (a_{ij})_{k\times k}$, $B = (b_{ij})_{k\times k}$ are two $k \times k$ matrix, then we say $A \geq B$ if $A - B$ is an non-negative definite matrix, i.e.,

$$a^T(A-B)a \geq 0 \quad \text{for arbitrary } \ a^T = (a_1, \cdots, a_k) \in \mathcal{M}_{k\times 1}.$$

Now, for a $k-$dimensional parametric distribution family $\mathcal{F} = \{f_\theta, \theta \in \Theta\}$, we have $\theta = (\theta_1, \cdots, \theta_k)$, and a sample $X = (X_1, \cdots, X_n)$ from $f(x, \theta)$. Suppose $\delta = \delta(X) = (\delta_1(X), \cdots, \delta_s(X))$ is an unbiased estimator of $g(\theta) = (g_1(\theta), \cdots, g_s(\theta))$, and we denote the Jacobian matrix of $\partial g/\partial\theta$ as J, i.e.,

$$J = \begin{bmatrix} \frac{\partial g}{\partial\theta_1} & \cdots & \frac{\partial g}{\partial\theta_k} \end{bmatrix} = \begin{bmatrix} \nabla^T g_1 \\ \vdots \\ \nabla^T g_s \end{bmatrix} = \begin{bmatrix} \frac{\partial g_1}{\partial\theta_1} & \cdots & \frac{\partial g_1}{\partial\theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial g_s}{\partial\theta_1} & \cdots & \frac{\partial g_1}{\partial\theta_s} \end{bmatrix} \in \mathcal{M}_{s\times k}$$

then we have the multi-dimensional CR Inequality given as

$$\mathrm{Cov}_\theta(\delta) \geq J \cdot (I(\theta))^{-1} \cdot J^T$$

Where $I(\theta) \in \mathcal{M}_{k\times k}$ is a positive definite Fisher information matrix, with

$$I_{ij}(\theta) = E_\theta\left[\left(\frac{\partial \log f(X,\theta)}{\partial\theta_i}\right)\left(\frac{\partial \log f(X,\theta)}{\partial\theta_j}\right)\right]$$

for $i, j = 1, \cdots, k$.

The behavior of the multi-dimensional CR inequality is a little bit tricky compare to the one-dimensional case, even for special distribution like exponential family (e.g., Normal distribution, see Example.3.9) may not reach it's CR lower bound.

1.3.2. Reparameterization and Additivity

It is important to realize that $I(\theta)$ depends on the particular parametrization chosen. In fact, if $\xi = h(\theta)$ and h^{-1} is differentiable, the information that X contains about ξ is given by

$$I(\xi) = \mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\xi}\right]^2 = \mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta} \cdot \frac{\partial\theta}{\partial\xi}\right]^2 = \left(\frac{\partial h^{-1}}{\partial\xi}\right) \cdot I(\theta) \cdot \left(\frac{\partial h^{-1}}{\partial\xi}\right)^T.$$

But unlike the Fisher information who change along with the reparameterization, the Cramér-Rao lower bound does not.

Meanwhile, if X (with distribution function p) is independent of Y (with distribution function q), then the information of θ contained in $Z = (X, Y)$ is

$$I_Z(\theta) = \mathbb{E}\left[\frac{\partial \log\big(p(x|\theta)\cdot q(y|\theta)\big)}{\partial\theta}\right]^2$$

$$
\begin{aligned}
=&\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right]^2+\mathbb{E}\left[\frac{\partial \log q(y|\theta)}{\partial\theta}\right]^2+\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right]\mathbb{E}\left[\frac{\partial \log q(y|\theta)}{\partial\theta}\right]\\
=&\mathbb{E}\left[\frac{\partial \log p(x|\theta)}{\partial\theta}\right]^2+\mathbb{E}\left[\frac{\partial \log q(y|\theta)}{\partial\theta}\right]^2=I_X(\theta)+I_Y(\theta).
\end{aligned}
$$

Thus, for an i.i.d sample $X=\{X_1,\cdots,X_n\}$, we have $I_X(\theta)=n\cdot I_{X_1}(\theta)$.

- *Example* 1.8. Suppose $X_1,\cdots,X_n \overset{i.i.d}{\sim} Poisson(\lambda)$, please calculate the information $X=\{X_1,\cdots,X_n\}$ contains about λ, $g_1(\lambda)=\log\lambda$, and $g_2(\lambda)=\sqrt{\lambda}$.

Answer. Let's denote $f(x|\lambda)$ the likelihood of X_1, so

$$
s(\lambda|x)=\frac{\partial\log f(x|\lambda)}{\partial\lambda}=\frac{\partial\log\left[\lambda^x e^{-\lambda}\cdot\mathbb{1}(x\in\mathbb{N})/x!\right]}{\partial\lambda}=\frac{x}{\lambda}-1
$$

therefore,

$$
I_X(\lambda)=n\cdot\mathbb{E}\Big[s(\lambda|x)\Big]^2=n\cdot\mathbb{E}\Big[\frac{x}{\lambda}-1\Big]^2=\frac{n}{\lambda}.
$$

and accordingly,

$$
\begin{aligned}
I_X(g_1(\lambda))=I_X(\log\lambda)=I_X(\lambda)/(g_1'(\lambda))^2=\lambda n,\\
I_X(g_2(\lambda))=I_X(\sqrt{\lambda})=I_X(\lambda)/(g_2'(\lambda))^2=4n.
\end{aligned}
$$

□

Interestingly, we see that $I(\lambda)$ is a decreasing function of λ, $I(\log\lambda)$ is a increasing function of λ, while $I(\sqrt{\lambda})$ whose behavior is intermediate between that of $I(\lambda)$ and $I(\log\lambda)$, is independent of λ.. Thus, the information in X about λ is inversely proportional to that about $\log\lambda$. In particular, for large values of λ, it seems that the parameter $\log\lambda$ can be estimated quite accurately, although the converse is true for λ. This conclusion is correct and is explained by the fact that $\log\lambda$ changes very slowly when λ is large. Hence, for large λ, even a large error in the estimate of λ will lead to only a small error in $\log\lambda$, whereas the situation is reversed for λ near zero where $\log\lambda$ changes very rapidly.

- *Example* 1.9. Consider a location family $\mathcal{F}=\{F(x-\theta):\theta\in\Theta\}$, with F being a known distribution such that $F''(x)=f'(x)$ exists and has support $\mathbb{R}$. Please calculate the Fisher Information $I(\theta)$.

Answer. Notice that

$$
I(\theta)=\mathbb{E}\Big[\frac{\partial\log f(x-\theta)}{\partial\theta}\Big]^2=\int\left(\frac{f'(x-\theta)}{f(x-\theta)}\right)^2 f(x-\theta)dx=\int\frac{\big[f'(x)\big]^2}{f(x)}dx.
$$

which is invariant over the value of θ. □

1.4. Attainment of Cramér-Rao Lower Bound

We now look at the attainment of Cramér-Rao lower bound in the case of one-parameter distribution family.

Theorem 1.10 (♣ **Attainment of Cramér-Rao Lower Bound**)**.** *Under the Fisher Information regularity condition, and assume that δ is a statistic with finite variance and $\mathbb{E}_\theta \delta = g(\theta)$. Then δ attains the lower bound*

$$\operatorname{Var}(\delta) = \big(g'(\theta)\big)^2 \cdot I(\theta)^{-1}$$

for all $\theta \in \Theta$ iff there exists a continuously differentiable function $\phi(\theta)$ such that the joint distribution of the sample is given by

$$f(x|\theta) = h(x)C(\theta)\exp\Big(\phi(\theta)\cdot\delta(x)\Big)$$

for suitably chosen $C(\theta)$ and $h(x)$, i.e., $f(x|\theta)$ constitutes an exponential family.

A proof of Theorem.1.10 is postponed to Supplement.

*1.5. *Randomized estimator*

Definition 1.11 (♣**Randomized Estimator**). For a sample $X = \{X_1, \cdots, X_n\}$, and a sequence of unobservable $i.i.d$ uniform$(0,1)$ random variables $U = \{U_i\}_{i\geq 1}$ which is independent of the sample X. A randomized estimator is defined as a function of X and U, i.e., $\delta^*(X,U)$, who does not depend on unknown parameters. Apparently, δ^* is a statistic with extra randomness brought by U.

Theorem 1.12 (**Minimal Risk Estimator in Strictly Convex Loss Function**)**.** *Given any randomized estimator of $g(\theta)$, there exists a nonrandomized estimator which is uniformly better if the loss function is strictly convex and at least as good when it is convex.*

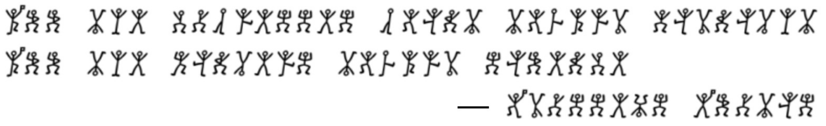
Proof of Theorem.1.12. Define $\delta(X) = \mathbb{E}\big[\delta^*(X,U)|X\big]$, since X, the whole sample, is a sufficient statistic of $g(\theta)$ and the loss function is strictly convex. Therefore, according to Rao-Blackwell Theorem, we have

$$R(g(\theta),\delta(X)) < R(g(\theta),\delta^*(X,U))$$

unless $\delta(X) = \delta^*(X,U)$ with probability one. The inequality sign changes to "$\leq$" when the loss function is convex instead of strictly convex. Thus finishes the proof. □

2. Unbiasedness

Even though bias is not a loss function, it is some property we commonly desire in estimation. With law of large numbers, we know when used repeatedly, an unbiased estimator in the long run will estimate the right value "on the average".



Definition 2.1 (♣**U-estimable**). Generally, for a parameter $g(\theta)$, we call $g(\theta)$ U-estimable if there exists some unbiased estimator $\delta(X)$ of $g(\theta)$.

We first give an obvious characterization of the totality of unbiased estimator.

Lemma 2.2. *If δ_0 is any unbiased estimator of $g(\theta)$, the totality of unbiased estimators is given by $\{\delta : \delta = \delta_0 - U, \text{ where } E_\theta(U) = 0 \text{ for all } \theta \in \Theta\}$, i.e., U is any unbiased estimator of zero.*

- *Example* 2.3 (♣**U-estimable in Geometric Distribution**). Let X take on the values $\{-1\} \cup \mathbb{N}$ with probabilities

$$P(X = -1) = p, \quad \text{and} \quad P(X = k) = q^2 p^k, \quad k = 0, 1, \cdots,$$

where $0 < p < 1$ and $q = 1 - p$. In order for U to be an unbiased estimator of zero, please show that $U(k) = -kU(-1)$ for $k = 0, 1, \cdots$, or equivalently,

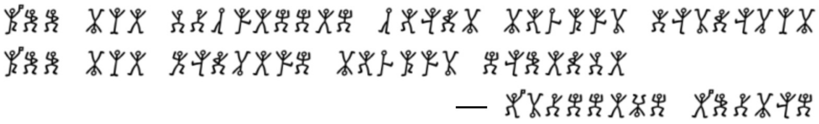
$$U(k) = ak, \quad \text{for some } a \in \mathbb{R} \text{ and } k = -1, 0, 1, \cdots. \tag{2.1}$$

- *Example* 2.4 (♣**Existence of UMVUE**). Let us continue with Example.2.3 and consider of the problems of estimating p and q^2. Please find an unbiased estimator δ_{p_0} for p (and another one $\delta_{q_0^2}$ for q^2) such that this unbiased estimator has minimal variance among all unbiased estimators when $p = p_0$ (i.e., $q = 1 - p_0 = q_0$).

Notably, δ_{p_0} depends on the value p_0 while $\delta_{q_0^2}$ does not depend on the value $q_0 = 1 - p_0$. Generally, we characterizing this feature more rigorously using the following definition.

3. UMVUE

3.1. Basic Definition



Definition 3.1 (♣**UMVUE & LMVUE**). An unbiased estimator $\delta(x)$ of $g(\theta)$ is called the uniform minimum variance unbiased estimator (UMVUE) of $g(\theta)$ if $\operatorname{Var}_\theta \delta(x) \leq \operatorname{var}_\theta \delta'(x)$ for all $\theta \in \Theta$ and for any other unbiased estimator $\delta'(x)$ of $g(\theta)$. Besides, $\delta(x)$ is locally minimum variance unbiased estimator (LMVUE) at $\theta = \theta_0$ if $\operatorname{var}_{\theta_0} \delta(x) \leq \operatorname{var}_{\theta_0} \delta'(x)$ for any other unbiased estimator $\delta'(x)$.

For e.g., according to definition.3.1, we clearly have δ_{p_0} is a LMVUE of p at $p = p_0$ while $\delta_{q_0^2}$ is a UMVUE.

Besides, when using MSE as our risk function, we have UMVUE is admissible and therefore by Theorem.1.3, we know the UMVUE is unique if it exists. For the existence, uniqueness, and characterization of LMVUE, we refer to Barankin (1950) and Stein (1950) for interested readers.

3.2. Methods of Finding UMVUE

3.2.1. Methods of Solving for Unbiased Estimators of 0

Theorem 3.2 (♣**Necessary and Sufficient Condition of Being an UMVUE**). *If δ is an unbiased estimator of $g(\theta)$ with $\operatorname{Var}_\theta(\delta) < \infty$ for $\forall\ \theta \in \Theta$, then δ is the UMVUE of $g(\theta)$ iff δ is uncorrelated with all unbiased estimators of 0, i.e., δ is the UMVUE of $g(\theta)$ iff : for arbitrary statistic U satisfy $\mathbb{E}_\theta(U) = 0$ for $\forall\ \theta \in \Theta$, we have $\operatorname{Cov}_\theta(\delta, U) = 0$ for $\forall \theta \in \Theta$.*

Remark 3.3.

1. The basic intuition of this theorem is that if an estimator could be improved by adding random noise to it, the estimator probably is defective.
2. This theorem is sometime useful in determining that an estimator is NOT UMVUE.
3. This theorem indicates a way to improve upon a given unbiased estimator.

Proof of Theorem.3.2.

(i) Necessity. Since δ is unbiased of $g(\theta)$, so for arbitrary statistic U being unbiased to 0, we have $\delta + t \cdot U$ is an unbiased estimator of $g(\theta)$ for arbitrary $t \in \mathbb{R}$. And since δ is the UMVUE of $g(\theta)$, so

$$\mathrm{Var}_\theta(\delta) \leq \mathrm{Var}_\theta(\delta + t \cdot U) = \mathrm{Var}_\theta(\delta) + 2t \cdot \mathrm{Cov}_\theta(\delta, U) + t^2 \cdot \mathrm{Var}_\theta(U), \quad \text{for } \forall t \in \mathbb{R}.$$

or equivalently,

$$f(t) = \mathrm{Var}_\theta(U) \cdot t^2 + 2\,\mathrm{Cov}_\theta(\delta, U) \cdot t \geq 0, \quad \text{for } \forall t \in \mathbb{R}$$

Since the $f(t)$ above is a quadratic function of t which is positive for the whole real line, hence

$$\Delta = 4\left[\mathrm{Cov}_\theta(\delta, U)\right]^2 \leq 0 \quad \Rightarrow \quad \mathrm{Cov}_\theta(\delta, U) = 0.$$

As U is arbitrary, so δ is uncorrelated with all unbiased estimators of 0.

(ii) Sufficiency. We now know δ is uncorrelated with all unbiased estimators of 0 and δ is an unbiased estimator of $g(\theta)$. For any other unbiased estimator δ' of $g(\theta)$, we define $U = \delta' - \delta$ who is an unbiased estimator of 0, so

$$\mathrm{Var}_\theta(\delta') = \mathrm{Var}_\theta(\delta + U) = \mathrm{Var}_\theta(\delta) + 2 \cdot \mathrm{Cov}_\theta(\delta, U) + \mathrm{Var}_\theta(U) \geq \mathrm{Var}_\theta(\delta)$$

And since δ' is arbitrary, we proved that δ is the UMVUE of $g(\theta)$.

□

3.2.2. Methods of Conditioning

Theorem 3.4 (♣Lehmann-Scheffé Theorem). *Let T be a sufficient complete statistic of θ, then for any statistic $\phi(T)$, it is the UMVUE of $g(\theta) = \mathbb{E}_\theta(\phi(T))$.*

Remark 3.5. Lehmann-Scheffé Theorem provides the most useful way to find UMVUE, general procedure will be:

1. Find a complete sufficient statistic T of θ.
2. Find *any* unbiased estimator δ for $g(\theta) = \mathbb{E}(\delta)$ of our interests.
3. $\phi(T) = E[\delta|T]$ is the UMVUE of $g(\theta)$.

Proof of Theorem.3.4.

(i) For arbitrary U-estimable $g(\theta)$, there exist one and only one function of T, i.e., $\phi(T)$, who is unbiased to $g(\theta)$. Since $g(\theta)$ is U-estimable, so there exist an unbiased estimator δ of $g(\theta)$, and hence $\phi(T) = \mathbb{E}\big[\delta|T\big]$ is a function of T and is unbiased to $g(\theta)$. Now if there exist another $\phi'(T)$ being unbiased to $g(\theta)$. Define $\psi(T) = \phi'(T) - \phi(T)$, we have $\mathbb{E}\psi(T) = 0$ for $\forall \theta \in \Theta$ and since T is complete, so

$$\mathbb{P}(\psi(T) = 0) = 1 \quad \Rightarrow \quad \phi(T) = \phi'(T) \text{ with probability 1.}$$

Hence $\phi(T)$ exists and is unique.

(ii) UMVUE is the function of T. Now, for arbitrary unbiased estimator δ' of $g(\theta)$, since $\mathbb{E}[\delta'|T]$ is unbiased to $g(\theta)$, so $\mathbb{E}[\delta'|T] = \phi(T)$. Meanwhile, according to Rao-Blackwell Theorem, we have

$$\mathrm{Var}_\theta(\delta') \geq \mathrm{Var}_\theta\{\mathbb{E}[\delta'|T]\} = \mathrm{Var}_\theta(\phi(T)), \quad \text{for } \forall\theta \in \Theta.$$

Hence $\phi(T)$ is the UMVUE of $g(\theta)$.

□

3.2.3. Methods of Cramér-Rao Lower Bound

Apparently, if we have some unbiased estimator δ of $g(\theta) = \mathbb{E}\delta$ attaining the Cramér-Rao Lower Bound, which is an lower bound for all unbiased estimator, then it must be an UMVUE.

3.2.4. Some Examples of UMVUE

• *Example* 3.6 (♣**UMVUE in Bernoulli Distribution**). Let $X_1, \cdots, X_n$ be a random sample from Bernoulli(p). Please find the UMVUE for $g(p) = p(1-p)$.

• *Example* 3.7 (♣**UMVUE in Exponential Distribution**). Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} \mathrm{Exp}(\lambda)$, with p.d.f given as

$$f(x) = \lambda e^{-\lambda x} \cdot \mathbb{1}(x \geq 0)$$

Please find the UMVUE of $\mathbb{P}(X_1 \leq x)$.

- *Example* 3.8 (♣**UMVUE in Normal Distribution**). $X_1, \cdots, X_n \overset{i.i.d.}{\sim} \mathrm{N}(\mu, \sigma^2)$. Please find the UMVUE for μ and σ^2.

- *Example* 3.9 (♣**Gap between UMVUE and the Cramér Rao Lower Bound**). Let's continue with Example.3.8, please varify whether the UMVUE $(\bar{X}, S^2)$ attained the Cramér-Rao lower bound.

This is a typical example showing how loose the Cramér-Rao lower bound is, especially for multi-dimensional case. It can not even be reached for well behaved distribution like normal distribution.

- *Example* 3.10 (♣**UMVUE in Normal Distribution**). Let's continue with Example.3.8, please find the UMVUE of σ^r for $n > -r + 1$.

- *Example* 3.11 (♣**UMVUE in Uniform Distribution**). $\{X_i\}_{i=1}^n \overset{i.i.d.}{\sim} \mathrm{U}(0, \theta)$. Please find the UMVUE of $g(\theta)$, where g is a known function with g' exists, and g and g' are well defined on $\mathbb{R}$.

• *Example* 3.12 (♣**UMVUE in Poisson Distribution**). Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} Poisson(\lambda)$.

1. Please find the UMVUE for λ.
2. Please find the UMVUE for λ^r, where $r \in \mathbb{N}^+$.
3. Please find the UMVUE for $\mathbb{P}(X_1 = x)$, where $x \in \mathbb{N}$.
4. Please find the UMVUE for $e^{-\lambda}$.
5. Please calculate $E[S^2 \mid \bar{X}]$ and try to show that $\text{Var}(S^2) > \text{Var}(\bar{X})$.

♣ ! Similar questions could be asked and answered for $X_1, \dots, X_n \overset{i.i.d.}{\sim} \text{Binomial}(k, p)$ with k being known.

Answer. We derive the sufficient complete statistic first. The joint p.m.f is

$$\begin{aligned}
\mathbb{P}(X = x|\lambda) &= \prod_{i=1}^{n} \mathbb{P}(X_i = x_i|\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \cdot \mathbb{1}(x_i \in \mathbb{N}) \\
&= \exp\left\{ \log\lambda \cdot \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!) \right\} \cdot \left[\prod_{i=1}^{n} \mathbb{1}(x_i \in \mathbb{N}) \right] \\
&= \exp\left\{ \log\lambda \cdot \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \log(x_i!) \right\} \cdot \mathbb{1}(x_i \in \mathbb{N}, i = 1, \cdots, n)
\end{aligned}$$

If we set

$$C(\lambda) = e^{-n\lambda}, \quad T_1(x) = \sum_{i=1}^{n} x_i, \quad \omega_1(\lambda) = \log\lambda\ , \quad h(x) = \prod_{i=1}^{n} \frac{\mathbb{1}(x_i \in \mathbb{N})}{x_i!}.$$

Since the natural parameter space $\Theta = \{\log(\lambda) : \lambda > 0\}$ has an inner point in $\mathbb{R}$, hence $T = \sum_{i=1}^{n} X_i$ is sufficient and complete statistic for λ. Now, we try to apply Lehmann-scheffé Theorem for each question.

1. For λ, notice that we simply have

$$\mathbb{E}\left(\frac{T}{n}\right) = \mathbb{E}(X_1) = \sum_{x=0}^{+\infty} x \cdot \frac{\lambda^x}{x!} e^{-\lambda} = \lambda \sum_{x=1}^{+\infty} \frac{\lambda^{x-1}}{(x-1)!} e^{-\lambda} = \lambda \sum_{y=0}^{+\infty} \frac{\lambda^y}{y!} e^{-\lambda} = \lambda$$

   Hence $\phi(T) = T/n$ is an unbiased estimator while it's an function of the sufficient complete statistic T, thus, by Lehmann-scheffé Theorem, we have $\phi(T) = T/n$ is the UMVUE of λ.

2. Notice for λ^r, $r \in \mathbb{N}^+$, guessing an unbiased estimator is not that easy.
   - **Method 1.** But if you are really familiar with the calculation of the moment of Poisson distribution, we can still figure out the possible form, which leads to our first method. Based on the additivity of the Poisson distribution, we have

$$T = \sum_{i=1}^{n} X_i \sim \text{Poisson}(n\lambda)$$

And since we have the expectation of T is given as $n\lambda$ and what we are looking for is λ^r, so the first instinct would be we should see how T^r performed. And based on the calculation above, we know it's easier to calculate the expectation of $g(T) = T(T-1)\cdots(T-r+1)$ rather than calculate that of T^r directly. So we have

$$\begin{aligned}
\mathbb{E}\Big(g(T)\Big) &= \sum_{t=0}^{+\infty} \Big[t(t-1)\cdots(t-r+1)\Big] \cdot \frac{(n\lambda)^t}{t!} e^{-n\lambda} \\
&= \sum_{t=r}^{+\infty} \Big[t(t-1)\cdots(t-r+1)\Big] \cdot \frac{(n\lambda)^t}{t!} e^{-n\lambda} \\
&= \sum_{t=r}^{+\infty} \frac{(n\lambda)^t}{(t-r)!} e^{-n\lambda} = (n\lambda)^r \cdot \sum_{t=r}^{+\infty} \frac{(n\lambda)^{t-r}}{(t-r)!} e^{-n\lambda} \\
&= (n\lambda)^r \cdot \sum_{y=0}^{+\infty} \frac{(n\lambda)^y}{y!} e^{-n\lambda} = (n\lambda)^r.
\end{aligned}$$

Hence $g(T)/n^r$ is a unbiased estimator and it is a function of the sufficient complete statistic T, so it is the UMVUE for λ^r.

- **Method 2.** Still, notice that $T = \sum_{i=1}^n X_i \sim$Poisson$(n\lambda)$, and we just assume $\phi(T)$ is the unbiased estimator of λ^r since the form isn't easy to guess. Now, by unbiasedness,

$$\mathbb{E}(\phi(T)) = \sum_{t=0}^{+\infty} \phi(t) \frac{(n\lambda)^t}{t!} e^{-n\lambda} = \lambda^r, \quad \text{holds for } \forall\lambda > 0.$$

which is

$$\sum_{t=0}^{+\infty} \phi(t) \frac{(n\lambda)^t}{t!} = \lambda^r e^{n\lambda}, \quad \text{holds for } \forall\lambda > 0.$$

Now by take Taylor's expansion for the right hand side term, we get

$$\sum_{t=0}^{+\infty} \frac{\phi(t)n^t}{t!} \cdot \lambda^t = \lambda^r e^{n\lambda} = \sum_{l=0}^{+\infty} \frac{n^l \lambda^{r+l}}{l!} = \sum_{t=r}^{+\infty} \frac{n^{t-r}}{(t-r)!} \cdot \lambda^t, \quad \text{holds for } \forall\lambda > 0.$$

Since both sides are power series of λ, and the equation holds for $\lambda \in (0, +\infty)$, so we may simply ask (♣ We are trying to construct the form of this $\phi(T)$, so we just simply admitting the form which fit our requirements) the coefficients corresponding to the same order of λ be the same, i.e.,

$$\begin{aligned}
&\frac{\phi(t)n^t}{t!} = 0\ , \quad \text{for } t = 0, 1, \cdots, r-1 \\
\text{and} \quad &\frac{\phi(t)n^t}{t!} = \frac{n^{t-r}}{(t-r)!}\ , \quad \text{for } t = r, r+1 \cdots
\end{aligned}$$

Which is

$$\phi(t) = 0 \ , \quad \text{for } t = 0, 1, \cdots, r-1$$
$$\text{and} \quad \phi(t) = \frac{t(t-1)\cdots(t-r+1)}{n^r} \ , \quad \text{for } t = r, r+1\cdots$$

Combine the two parts above, we have

$$\phi(T) = \frac{T(T-1)\cdots(T-r+1)}{n^r}$$

is an unbiased estimator of λ^r and is a function of the sufficient complete statistic T, therefore, $\phi(T) = T(T-1)\cdots(T-r+1)/n^r$ is the UMVUE for λ^r.

3. We first follow the last question's idea to show the UMVUE.
   - **Method 1.** Notice that $T = \sum_{i=1}^n X_i \sim$Poisson$(n\lambda)$, and we assume $\phi(T)$ is the unbiased estimator of $\mathbb{P}(X_1 = x)$, $x \in \mathbb{N}$, since the form is not clear. Now, by unbiasedness,

     $$\mathbb{E}(\phi(T)) = \sum_{t=0}^{+\infty} \phi(t) \frac{(n\lambda)^t}{t!} e^{-n\lambda} = \mathbb{P}(X_1 = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

     Which holds for all $\lambda > 0$. Therefore,

     $$\sum_{t=0}^{+\infty} \phi(t) \frac{(n\lambda)^t}{t!} = \frac{\lambda^x}{x!} e^{(n-1)\lambda}.$$

     Now by take Taylor's expansion for the right hand side term,

     $$\frac{\lambda^x}{x!} e^{(n-1)\lambda} = \sum_{l=0}^{+\infty} \frac{(n-1)^l \lambda^l}{l!} \cdot \frac{\lambda^x}{x!} = \sum_{l=0}^{+\infty} \frac{(n-1)^l \lambda^{x+l}}{l! \cdot x!} = \sum_{t=x}^{+\infty} \frac{(n-1)^{t-x} \lambda^t}{x! \cdot (t-x)!},$$

     substitute the form into the equation $\mathbb{E}(\phi(T)) = \mathbb{P}(X_1 = x)$ hence

     $$\sum_{t=0}^{+\infty} \frac{\phi(t) n^t}{t!} \cdot \lambda^t = \sum_{t=x}^{+\infty} \frac{(n-1)^{t-x}}{x! \cdot (t-x)!} \cdot \lambda^t$$

     Since both sides are power series of λ, and the equation holds for $\lambda \in (0, +\infty)$, so we may simply ask the coefficients corresponding to the same order of λ be the same, i.e.,

     $$\frac{\phi(t) n^t}{t!} = 0 \ , \quad \text{for } t = 0, 1, \cdots, x-1,$$
     $$\text{and} \quad \frac{\phi(t) n^t}{t!} = \frac{(n-1)^{t-x}}{x! \cdot (t-x)!} \ , \quad \text{for } t = x, x+1\cdots$$

Which is

$$\phi(t) = 0 \ , \quad \text{for } t = 0, 1, \cdots, x-1,$$
$$\text{and} \quad \phi(t) = \binom{t}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{t-x} \quad , \quad \text{for } t = x, x+1\cdots$$

Combine the two parts above, we have

$$\begin{aligned}\phi(T) &= \binom{T}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{T-x} \cdot \mathbb{1}(T \geq x, x \in \mathbb{N})\\ &= \binom{\sum_{i=1}^n X_i}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{\sum_{i=1}^n X_i - x} \cdot \mathbb{1}(\sum_{i=1}^n X_i \geq x, x \in \mathbb{N})\end{aligned}$$

is an unbiased estimator of $\mathbb{P}(X_1 = x)$ and is a function of the sufficient complete statistic T, therefore, $\phi(T)$ given above is the UMVUE for $\mathbb{P}(X_1 = x)$.

- **Method 2.** For another approach, we have already showed that T is a sufficient and complete statistic for λ, so in order to apply the Lehmann-Scheffé Theorem, what we need is to find an unbiased estimator for $\mathbb{P}(X_1 = x)$, and notice that by definition

$$\mathbb{P}(X_1 = x) = \mathbb{E}(\mathbb{1}(X_1 = x))$$

i.e., $\delta = \mathbb{1}(X_1 = x)$ is an unbiased estimator of $\mathbb{P}(X_1 = x)$, hence

$$\phi(T) = \mathbb{E}(\delta|T) = \mathbb{E}(\mathbb{1}(X_1 = x)|T)$$

is our unique UMVUE by Lehmann-Scheffé Theorem. Before we begin to calculate $\phi(T)$, one thing to notice is that by the additive of Poisson random variable, we have $T = \sum_{i=1}^n X_i$ ~Poisson($n\lambda$) and $\sum_{i=2}^n X_i$ ~Poisson($(n-1)\lambda$). Now

$$\begin{aligned}&\mathbb{E}\Big[\mathbb{1}(X_1 = x)|T = t\Big] = \mathbb{P}\Big(X_1 = x\Big|\sum_{i=1}^n X_i = t\Big)\\ =&\mathbb{P}\Big(X_1 = x, \sum_{i=1}^n X_i = t\Big)\Big/\mathbb{P}\Big(\sum_{i=1}^n X_i = t\Big)\\ =&\mathbb{P}\Big(X_1 = x, \sum_{i=2}^n X_i = t-x\Big)\Big/\mathbb{P}\Big(\sum_{i=1}^n X_i = t\Big)\\ =&\mathbb{P}(X_1 = x)\cdot\mathbb{P}\Big(\sum_{i=2}^n X_i = t-x\Big)\Big/\mathbb{P}\Big(\sum_{i=1}^n X_i = t\Big)\\ =&\left[\frac{\lambda^x e^{-\lambda}}{x!}\cdot\frac{((n-1)\lambda)^{t-x}e^{-(n-1)\lambda}}{(t-x)!}\cdot\mathbb{1}\big(t-x \geq 0, x, t \in \mathbb{N}\big)\right]\end{aligned}$$

$$
\begin{aligned}
&\cdot \left[\frac{(n\lambda)^t e^{-n\lambda}}{t!} \cdot \mathbb{1}\big(t \in \mathbb{N}\big)\right]^{-1} \\
=&\binom{t}{x} \cdot \left(\frac{1}{n}\right)^x \cdot \left(1-\frac{1}{n}\right)^{t-x} \cdot \mathbb{1}\big(t \geq x, x, t \in \mathbb{N}\big)
\end{aligned}
$$

So overall, we have

$$
\begin{aligned}
\phi(T) &= \binom{T}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{T-x} \cdot \mathbb{1}(T \geq x, x \in \mathbb{N}) \\
&= \binom{\sum_{i=1}^n X_i}{x}\left(\frac{1}{n}\right)^x\left(1-\frac{1}{n}\right)^{\sum_{i=1}^n X_i - x} \cdot \mathbb{1}(\sum_{i=1}^n X_i \geq x, x \in \mathbb{N})
\end{aligned}
$$

is an unbiased estimator of $\mathbb{P}(X_1 = x)$ and is a function of the sufficient complete statistic T, therefore, $\phi(T)$ given above is the UMVUE for $\mathbb{P}(X_1 = x)$.

4. Notice that $e^{-\lambda} = \mathbb{P}(X_1 = 0)$ is just a special case of the last question, so we just simply have

$$
\begin{aligned}
\phi(T) =&\binom{\sum_{i=1}^n X_i}{0}\left(\frac{1}{n}\right)^0\left(1-\frac{1}{n}\right)^{\sum_{i=1}^n X_i - 0} \cdot \mathbb{1}(\sum_{i=1}^n X_i \geq 0) \\
=&\left(1-\frac{1}{n}\right)^{\sum_{i=1}^n X_i}
\end{aligned}
$$

is the unique UMVUE of $e^{-\lambda}$.

5. One trivial way to solve this problem is just calculate directly

   - **Method 1.** Recall our notation here, $\bar{X} = \frac{T}{n}$, $T = \sum_{i=1}^n X_i$, and $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, so

$$
\begin{aligned}
\mathbb{E}[S^2|\bar{X}] &= \frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^n (X_i - \bar{X})^2 |\bar{X}\Big] \\
&= \frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^n X_i^2 - 2\bar{X}\sum_{i=1}^n X_i + n\bar{X}^2|\bar{X}\Big] \\
&= \frac{1}{n-1}\mathbb{E}\Big[\sum_{i=1}^n X_i^2 - n\bar{X}^2|\bar{X}\Big] \\
&= \frac{n}{n-1}\left(\mathbb{E}[X_1^2|\bar{X}] - \bar{X}^2\right)
\end{aligned}
$$

   So we calculate $\mathbb{E}[X_1^2|\bar{X} = t]$ below.

$$
\mathbb{E}[X_1^2|\bar{X} = t] = \sum_{x=0}^{+\infty} x^2 \cdot \mathbb{P}(X_1 = x|\bar{X} = t)
$$

$$
\begin{aligned}
=&\sum_{x=0}^{+\infty} \frac{x^2 \cdot \mathbb{P}(X_1 = x, \sum_{i=2}^n X_i = nt - x)}{\mathbb{P}(\sum_{i=1}^n X_i = nt)} \\
=&\sum_{x=0}^{+\infty} \frac{x^2 \cdot \mathbb{P}(X_1 = x) \cdot \mathbb{P}(\sum_{i=2}^n X_i = nt - x)}{\mathbb{P}(\sum_{i=1}^n X_i = nt)} \\
=&\sum_{x=0}^{nt} x^2 \cdot \frac{\lambda^x e^{-\lambda}}{x!} \cdot \frac{((n-1)\lambda)^{nt-x}\, e^{-(n-1)\lambda}}{(nt-x)!} \cdot \left(\frac{(n\lambda)^{nt}\, e^{-n\lambda}}{(nt)!}\right)^{-1} \\
=&\sum_{x=0}^{nt} x^2 \cdot \binom{nt}{x} \cdot \left(\frac{1}{n}\right)^x \cdot \left(1 - \frac{1}{n}\right)^{nt-x} \\
=&\sum_{x=0}^{nt} x(x-1) \cdot \binom{nt}{x} \cdot \left(\frac{1}{n}\right)^x \cdot \left(1 - \frac{1}{n}\right)^{nt-x} \\
&+\sum_{x=0}^{nt} x \cdot \binom{nt}{x} \cdot \left(\frac{1}{n}\right)^x \cdot \left(1 - \frac{1}{n}\right)^{nt-x} \\
=&\frac{nt \cdot (nt-1)}{n^2} \cdot \sum_{x-2=0}^{nt-2} \binom{nt-2}{x-2} \cdot \left(\frac{1}{n}\right)^{x-2} \cdot \left(1 - \frac{1}{n}\right)^{nt-x} \\
&+\frac{nt}{n} \sum_{x-1=0}^{nt-1} \binom{nt-1}{x-1} \cdot \left(\frac{1}{n}\right)^{x-1} \cdot \left(1 - \frac{1}{n}\right)^{nt-x} \\
=&\frac{t \cdot (nt-1)}{n} + \frac{nt}{n} = t^2 + \frac{n-1}{n} t
\end{aligned}
$$

Hence $\mathbb{E}[X_1^2|\bar{X}] = \bar{X}^2 + (n-1)\bar{X}/n$, and we get

$$\mathbb{E}[S^2|\bar{X}] = \frac{n}{n-1}\left(\mathbb{E}[X_1^2|\bar{X}] - \bar{X}^2\right) = \frac{n}{n-1}\left(\bar{X}^2 + \frac{n-1}{n}\bar{X} - \bar{X}^2\right) = \bar{X}.$$

Since S^2 doesn't equal to $\bar{X}$ almost surely, so $\mathbb{E}[\mathrm{Var}(S^2|\bar{X})] > 0$, hence

$$\mathrm{Var}(S^2) = \mathrm{Var}(\mathbb{E}[S^2|\bar{X}]) + \mathbb{E}[\mathrm{Var}(S^2|\bar{X})] = \mathrm{Var}(\bar{X}) + \mathbb{E}[\mathrm{Var}(S^2|\bar{X})] > \mathrm{Var}(\bar{X}).$$

- **Method 2.** We have already proved that $\bar{X} = T/n$ is a sufficient complete statistic of θ. Notice that $\mathbb{E}(\bar{X}) = \frac{1}{n}\sum_{i=1}^n EX_i = \lambda$ and

  $$\mathbb{E}(\mathbb{E}[S^2|\bar{X}]) = \mathbb{E}(S^2) = \mathrm{Var}(X_1) = \lambda$$

  Hence both $\bar{X}$ and $\mathbb{E}[S^2|\bar{X}]$ are the unbiased estimator of λ and are the functions of the sufficient complete statistic, so by Lehmann-Scheffé Theorem, we have both $\bar{X}$ and $\mathbb{E}[S^2|\bar{X}]$ are the unique UMVUE of λ, in other words, we must have

  $$\mathbb{E}[S^2|\bar{X}] = \bar{X}$$

And since S^2 is not a function of $\bar{X}$ almost surely, and the fact that S^2 is an unbiased estimator of λ and $\bar{X}$ is the UMVUE of λ, thus, by definition, we get $\text{Var}(S^2) > \text{Var}(\bar{X})$.

□

- *Example* 3.13 (♣**UMVUE in Regression**). Now, Let's consider an simple linear model

$$Y_{n\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1}$$

We assume we have $\epsilon \sim$N$(0, \sigma^2 I_n)$, i.e., $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T$, and $\epsilon_1, \cdots, \epsilon_n \sim_{i.i.d} N(0, \sigma^2)$. $X_{n\times p}$ is our known design matrix. $\beta_{p\times 1} = (\beta_1, \cdots, \beta_p)^T$ and σ^2 are the unknown parameter, while $Y_{n\times 1} = (Y_1, \cdots, Y_n)^T$ is the sample. Please give the UMVUE of $c^T\beta$ for some known c.

Answer. Based on model, we simply have

$$Y_{n\times 1}|\beta_{p\times 1} = X_{n\times p}\beta_{p\times 1} + \epsilon_{n\times 1} \sim N(X\beta, \sigma^2 I_n)$$

So we can derive the likelihood function

$$\begin{aligned}
f(y|\beta) &= (\sqrt{2\pi\sigma^2})^{-n} \exp\left\{-\frac{1}{2\sigma^2}\|y - X\beta\|_2^2\right\} \\
&= (\sqrt{2\pi\sigma^2})^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (y_i - \sum_{j=1}^p x_{ij}\beta_j)^2\right]\right\} \\
&= (\sqrt{2\pi\sigma^2})^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (y_i^2 - 2\sum_{j=1}^p x_{ij}y_i \cdot \beta_j + (\sum_{j=1}^p x_{ij}\beta_j)^2)\right]\right\} \\
&= (\sqrt{2\pi\sigma^2})^{-n} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n y_i^2 - 2\sum_{i=1}^n\sum_{j=1}^p x_{ij}y_i \cdot \beta_j + \sum_{i=1}^n(\sum_{j=1}^p x_{ij}\beta_j)^2\right]\right\} \\
&= (\sqrt{2\pi\sigma^2})^{-n} \exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2} + \sum_{j=1}^p\left[\frac{\beta_j}{\sigma^2}\sum_{i=1}^n x_{ij}y_i\right]\right\} \cdot \exp\left\{\sum_{i=1}^n(\sum_{j=1}^p x_{ij}\beta_j)^2\right\}
\end{aligned}$$

If for $j = 1, 2, \cdots, p$, we set

$$\omega_1(\theta) = -\frac{1}{2\sigma^2}, \quad \omega_{j+1}(\theta) = \frac{\beta_j}{\sigma^2}, \quad T_1(y) = \sum_{i=1}^n y_i^2, \quad T_{j+1}(y) = \sum_{i=1}^n x_{ij}y_i,$$

$$\text{and} \quad C(\theta) = (\sqrt{2\pi\sigma^2})^{-n} \cdot \exp\left\{\sum_{i=1}^n(\sum_{j=1}^p x_{ij}\beta_j)^2\right\}, \quad h(y) = 1.$$

Since this likelihood is coming from an exponential family and the natural parameter space

$$\Theta = \left\{\left(-\frac{1}{2\sigma^2}, \frac{\beta_1}{\sigma^2}, \cdots, \frac{\beta_p}{\sigma^2}\right) : \sigma^2 > 0, \beta_i \in \mathbb{R}, i = 1, \cdots, p\right\}$$

has an inner point, thus, we know that $T = (T_1, \cdots, T_{p+1})$ is a sufficient and complete statistic for β. Meanwhile, notice

$$\max_{\beta} L(\beta; Y) = \max_{\beta} f(Y|\beta) = \min_{\beta} \|Y - X\beta\|_2^2$$

So we have the MLE of this linear regression (the case where error are white noise with normal distribution) is also the least square estimator, i.e.,

$$\hat{\beta} = (X^T X)^{-1} X^T Y = (X^T X)^{-1} \cdot (T_2(Y), \cdots, T_{p+1}(Y))^T$$

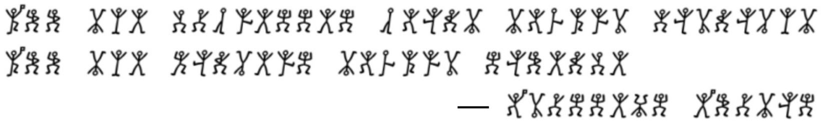
is a function of the sufficient complete statistic, and for arbitrary matrix $c_{p\times k}$, we have

$$\begin{aligned}\mathbb{E}(c^T\hat{\beta}) &= c^T\mathbb{E}((X^T X)^{-1} X^T Y) = c^T (X^T X)^{-1} X^T \mathbb{E}(Y) \\ &= c^T (X^T X)^{-1} X^T X\beta = c^T\beta\end{aligned}$$

Hence $c^T\hat{\beta}$ is an unbiased estimator for $c^T\beta$ and is also a function of the sufficient complete statistic, so by Lehmann-Scheffé Theorem, we have $c^T\hat{\beta}$ is the unique UMVUE (it's the Maximum Likelihood Estimator and also the Least Square Estimator). □

4. Drawback of Unbiasedness

Requiring unbiasedness can sometimes be cruel.



• *Example* 4.1 (♣**Variance Bias Trade-off**). Consider $X_1, \cdots, X_n \sim_{i.i.d} N(\mu, \sigma^2)$ with unknown mean and variance. We already know that S^2 is the UMVUE of σ^2 and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 = \frac{n-1}{n} S^2$ is the MLE of σ^2. Using the quadratic loss as our loss function, we conclude that

$$\mathrm{MSE}(S^2) = \mathrm{Var}(S^2) = \frac{2\sigma^4}{n-1}, \quad \mathrm{MSE}(\hat{\sigma}^2) = \mathrm{Var}\,\hat{\sigma}^2 + \big(\mathrm{Bias}(\hat{\sigma}^2)\big)^2 = \frac{(2n-1)\sigma^4}{n^2}.$$

And it's clear that $\mathrm{MSE}(\hat{\sigma}^2) < \mathrm{MSE}(S^2)$, hence the MLE is preferable compare to the UMVUE even in a simple estimation question with a well behaved distribution family like normal distribution.

• *Example* 4.2 (Misbehaved UMVU estimator). Let X have the Poisson distribution and let $g(\theta) = e^{-a\theta}$, where a is a known constant. Please give the UMVUE of $g(\theta)$.

Answer. Assume the UMVUE is given by $\delta(X)$, then the unbiasedness leads to

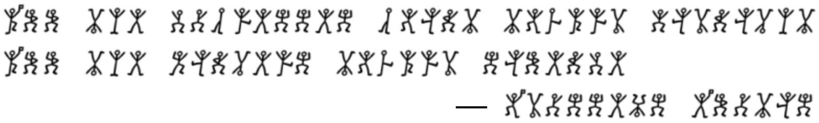
$$\sum_{x=0}^{\infty} \delta(x) \frac{\theta^x}{x!} = e^{(1-a)\theta} = \sum_{x=0}^{\infty} \frac{(1-a)^x \theta^x}{x!},$$

and hence both sides are power series of θ, so we can simply require the coefficients of each order of θ to be the same, i.e.,

$$\delta(X) = (1-a)^X. \tag{4.1}$$

Suppose $a = 7$. Then, $g(\theta) = e^{-7\theta}$, and one would expect an estimator which decreases from 1 to 0 as X goes from 0 to infinity. The ML estimator e^{-7X} meets this expectation. However, the unique unbiased estimator $\delta = (-2)^X$ oscillates wildly between positive and negative values and appears to bear no relation to the problem at hand. As a matter of fact, the issue cames from our restriction of δ as we ask it to be unbiased. □

5. *Lehmann-Scheffé Theorem for General Convex Loss Function



Even though our attention has been restricted to squared error loss, the preceding argument establishes the following result since Rao-Blackwell theorem applies to any convex loss function.

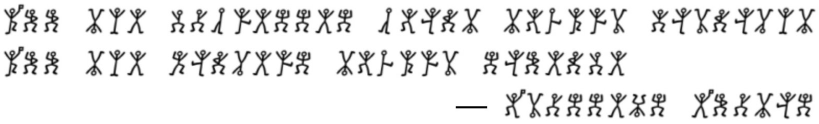
Theorem 5.1 (♣ **Lehmann-Scheffé Theorem for General Convex Loss Function**)**.** *Let X be distributed according to a distribution in $\mathcal{F} = \{P_\theta, \theta \in \Theta\}$, and suppose that T is a complete sufficient statistic for $\mathcal{F}$.*

- *(a) For every U-estimable function $g(\theta)$, there exists an unbiased estimator that uniformly minimizes the risk for any loss function $L(g(\theta), \delta)$ which is convex in its second argument; therefore, this estimator in particular is UMRUE (uniformly minimum risk unbiased estimator).*
- *(b) The UMRUE of $g(\theta)$ is the unique unbiased estimator which is a function of T; it is the unique unbiased estimator with minimum risk, provided its risk is finite and L is strictly convex in δ.*

It is interesting to note that under mild conditions, the existence of a complete sufficient statistic is not only sufficient but also necessary for every U-estimatiable function to have a UMVUE (according to Bahadur (1957)).

6. *Supplement

Here we list some of the proofs of theorems listed before.



Proof of Theorem.1.10. The estimator δ attains the lower bound for $\forall\theta \in \Theta$ is equivalent to the equality sign of Cauchy inequality

$$\mathrm{Var}(\delta) \cdot \mathbb{E}\left[s(\theta|x)\right]^2 = \mathrm{Cov}(\delta, s(\theta|x)) \quad \text{holds for } \forall\theta \in \Theta,$$

which happens only when $\delta(x)/s(\theta|x)$ is invariant of x over the support of $s(\theta|x)$, i.e., there exists some $a(\theta)$ and $b(\theta)$, s.t.,

$$s(\theta|x) = \frac{\partial \log f(x|\theta)}{\partial\theta} = a(\theta)\cdot\delta(x) + b(\theta). \tag{6.1}$$

Solving the ordinary differential equality (6.1) leads us to

$$\log f(x|\theta) = \left(\int a(\theta)d\theta\right)\cdot\delta(x) + \left(\int b(\theta)d\theta\right) + h'(x),$$

hence there exist some differentiable function $\phi(\theta)$ and suitable function $C(\theta)$ and $h(x)$, such that

$$f(x|\theta) = h(x)C(\theta)\exp\Big(\phi(\theta)\cdot\delta(x)\Big).$$

□

References

Bahadur, R. R. (1957). On unbiased estimates of uniformly minimum variance. Sankhyā: The Indian Journal of Statistics (1933-1960), 18(3/4), 211-224.

Barankin, E. W. (1950). Extension of a theorem of Blackwell. The Annals of Mathematical Statistics, 21(2), 280-284.

Stein, C. (1950). Unbiased estimates with minimum variance. The Annals of Mathematical Statistics, 406-415.