Statistics 3858 : Cramer-Rao and Sufficient Statistics

1 Preliminary Comments

One of the methods to compare estimators is based of Mean Square Errors (MSE)

$$MSE(\hat{\theta}) = \mathrm{E}_\theta((\hat{\theta} - \theta)^2) = \mathrm{Var}(\hat{\theta}) + Bias(\hat{\theta})^2$$

An estimator $\hat{\theta}$ is unbiased if and only if for all θ

$$\mathrm{E}_\theta(\hat{\theta}) = \theta \ .$$

Thus for an unbiased estimator $\hat{\theta}$

$$MSE(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) \ .$$

We say an estimator is better (more efficient) than another estimator if and only it has smaller MSE. In the case of comparing unbiased estimators, the better one has smaller variance. Notice these properties are for all θ.

2 Cramér Rao Lower Bound

The Cramér-Rao lower bound (named after two famous statisticians early in the history of statistics Harold Cramér and C. R. Rao) gives a lower bound for the variance of unbiased estimators in the case of iid sampling. Thus *if an estimator* is unbiased and has variance equal to this lower bound it is then impossible to find an unbiased estimator that has smaller variance. In this sense it is the best possible.

We only state the Cramér-Rao lower bound for real parameters, but there is a version for vector valued parameters. One just needs to interpret the inequality in a valid way for variance matrices, in terms of the difference of matrices being positive definite.

Theorem 1 *Suppose X_i , $i = 1, \ldots, n$ are iid with pdf (or pmf) $f(\cdot;\theta)$ and that $\Theta \subset R$. Suppose that $T = T(X_1, \ldots, X_n)$ is an unbiased estimator of θ. Suppose the statistical model satisfies Assumptions I and II, the regularity conditions for MLE (see earlier notes). Then*

$$Var(T) \geq \frac{1}{nI(\theta)}$$

where $I(\theta)$ is Fisher's information.

Proof : (done for pdf case)

Let

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

Then Z is a random variable with mean 0 and variance $nI(\theta)$.

First notice that

$$\begin{aligned}
\frac{\partial \prod_{j=1}^{n} f(x_j;\theta)}{\partial \theta} &= \sum_{i=1}^{n} \frac{\partial \left(f(x_i;\theta)\right)}{\partial \theta} \prod_{j=1;j\neq i}^{n} f(x_j;\theta) \\
&= \sum_{i=1}^{n} \frac{\partial \left(f(x_i;\theta)\right)}{\partial \theta} \frac{1}{f(x_i;\theta)} \prod_{j=1}^{n} f(x_j;\theta) \\
&= \sum_{i=1}^{n} \frac{\partial \log\left(f(x_i;\theta)\right)}{\partial \theta} \prod_{j=1}^{n} f(x_j;\theta)
\end{aligned}$$

Therfore

$$\begin{aligned}
\mathrm{Cov}(Z,T) &= \int_R \cdots \int_R T(x_1,\ldots,x_n) \sum_{i=1}^{n} \frac{\partial \log\left(f(x_i;\theta)\right)}{\partial \theta} \prod_{j=1}^{n} f(x_j;\theta) dx_n \ldots, x_1 \\
&= \int_R \cdots \int_R T(x_1,\ldots,x_n) \frac{\partial \prod_{j=1}^{n} f(x_j;\theta)}{\partial \theta} dx_n \ldots, x_1 \\
&= \frac{\partial}{\partial \theta} \left\{ \int_R \cdots \int_R T(x_1,\ldots,x_n) \prod_{j=1}^{n} f(x_j;\theta) dx_n \ldots, x_1 \right\} \\
&= \frac{\partial \theta}{\partial \theta} \\
&= 1
\end{aligned}$$

Therefore

$$1 = \left(\mathrm{Cov}(Z,T)\right)^2 \leq \mathrm{Var}(T)\mathrm{Var}(Z) = \mathrm{Var}(T) nI(\theta) \ .$$

End of Proof

Recall our earlier discussion of comparing estimators. In that we used properties of the sampling distribution to decide which estimator is better, and in particular if estimators had a normal or approximately normal distribution, and if they were unbiased then the one with smaller variance is more efficient.

Examples : see text section 8.7

For a nontrivial example consider X_i iid Poisson λ. Two possible unbiased estimators are

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

The student should find a calculation or carry out an appropriate calculation to verify that the sample variance in this case is an unbiased estimator; this was actually done earlier in the course.

Which is better? For this in principle we need to find $\mathrm{Var}(S^2)$. The student should consider how to calculate this.

The student should verify that $\mathrm{Var}(\bar{X})$ equals the expression for the Cramér-Rao lower bound; do this at home. Thus we know without any calculations that

$$\mathrm{Var}(S^2) \geq \frac{1}{nI(\lambda)} = \mathrm{Var}(\bar{X}) \ .$$

Thus $\bar{X}$ is more efficient (or at least as efficient) as S^2.

Notice that the Cramér-Rao Lower Bound Theorem allows us to make this conclusion without the need to calculate the variance of the second estimator. Of course lower bound would not be helpful if the estimators were biased.

Aside In the proof of the Cramér-Rao lower bound we used an identity based on correlation ρ.

$$1 \geq \rho^2 = \frac{\mathrm{Cov}^2(X,Y)}{\mathrm{Var}(X)\mathrm{Var}(Y)}$$

Therefore

$$\mathrm{Cov}^2(X,Y) \leq \mathrm{Var}(X)\mathrm{Var}(Y) \ .$$

This is also called the Cauchy-Schwarz inequality, but it is actually a special case of the Cauchy-Schwarz inequality in real

3 Various Properties of Estimators : MLE Is Asymptotically Efficient and Other Properties of Estimators

Suppose that the statistical model satisfies the two regularity Assumptions I and II; review these. Recall that in this case

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N\left(0, \frac{1}{I(\theta)}\right)$$

as $n \to \infty$. However from the Cramér-Rao Lower Bound Theorem we also have for an unbiased estimator

$$\mathrm{Var}\sqrt{n}(\hat{\theta}_n - \theta) = \frac{1}{I(\theta)}$$

Thus the limiting normal variance is the smallest possible.

Another term is also used and we just give its definition here but do not use it much in this course.

An estimator $\hat{\theta}_n$ is said to be asymptotically unbiased if $\mathrm{E}(\hat{\theta}_n) \to \theta$ as $n \to \infty$. (This is actually a property of the sequence of estimators). While this may be a nice property this by itself is not always useful.

Example : X_i iid, say normal. Let

$$\hat{\mu}_{1,n} = X_1 \text{ and } \hat{\mu}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n \ .$$

Verify that both of these estimators are unbiased, and hence asymptotically unbiased. The first is not consistent, but the second is consistent.

Consider the sample variance and the MLE for σ^2 in this example. One is unbiased and one is biased. However both are asymptotically unbiased and both are consistent.

This is also an appropriate place to discuss some further properties for normal approximations. Specifically this is useful for method of moments estimators and for functions of estimators. For the later if we estimate the Poisson parameter λ, then how can we estimate $P(X = 0) = f(0;\lambda)$, where $P(X = k) = f(k;\lambda)$. A natural estimator of $P(X = 0;\lambda)$ is

$$\hat{P}(0) = f(0;\hat{\lambda}) = e^{-\hat{\lambda}} \equiv g(\hat{\lambda}) \ .$$

Here we take $g(x) = e^{-x}$.

Recall

$$\sqrt{n}(\hat{\lambda}_n - \lambda) \Rightarrow N(0, I^{-1}(\lambda)) \equiv N(0,\tau^2)$$

which is one of the conditions to apply the delta method. Here we use the notation $\Rightarrow$ to denote convergence in distribution. If the function g is also differentiable at λ, with non-zero derivative at λ, then the second condition for the delta method is satisfied. Therefore by the delta method

$$\sqrt{n}\left(g(\hat{\lambda}_n) - g(\lambda)\right) = g'(\lambda)\sqrt{n}(\hat{\lambda}_n - \lambda) \Rightarrow N(0, [g'(\lambda)]^2\tau^2) \ .$$

Now we apply the delta method to our estimator $\hat{P}(0)$ above. $g'(x) = -e^{-x}$. Therefore since

$$\sqrt{n}(\hat{\lambda} - \lambda) \Rightarrow N(0, I^{-1}(\lambda)) = N(0,\lambda)$$

then

$$\sqrt{n}(\hat{P}(0) - P(0)) \Rightarrow N(0, (g'(\lambda))^2 I^{-1}(\lambda)) = N(0, \lambda e^{-2\lambda})$$

as $n \to \infty$.

4 Sufficient Statistics

This section introduces a special tool that is very useful for calculation and the simplification of working with maximum likelihood estimators, namely the notion of sufficient statistics. The setting is for parametric models and we continue the same notation as before.

Definition 1 *A statistic $T(X_1,\ldots,X_n)$ is said to be a sufficient statistic if and only if the conditional distribution of $X_1,\ldots,X_n$ given $T = t$ (for any possible t) does not depend on θ.*

This definition means that the conditional distribution is algebraically independent of θ, or equivalently θ does not appear in the expression for this distribution. There are only a few cases where we can

verify that T is a sufficient statistic by directly verifying that it satisfies the definition. Shortly we will see there is an equivalent but easier to use property, called the Factorization Theorem.

Example : Poisson

$X_i, i = 1, \ldots, n$ are iid Poisson. Let $T = X_1 + X_2 + \ldots + X_n$. The student should show that $T \sim \text{Poisson, n}\lambda$.

Do this by using the convolution formula and mathematical induction or using a moment generating function method and an appropriate property of MGFs.

Thus

$$\begin{aligned}
&P_\lambda(X_1 = x_1, \ldots, X_n = x_n | T = t) \\
&= \frac{P_\lambda\left(X_1 = x_1, \ldots, X_n = x_n, T = t\right)}{P_\lambda(T = t)} \\
&= \frac{1}{P_\lambda(T = t)} P_\lambda\left(X_1 = x_1, \ldots, X_n = x_n\right) \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right) \\
&= \frac{t!}{(n\lambda)^t e^{-n\lambda}} \frac{\lambda^{x_1 + x_2 + \ldots + x_n} e^{-n\lambda}}{\prod_{i=1}^n x_i!} \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right) \\
&= \frac{t!}{(n\lambda)^t} \frac{\lambda^t}{\prod_{i=1}^n x_i!} \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right) \\
&= \frac{t!}{n^t \prod_{i=1}^n x_i!} \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right)
\end{aligned}$$

Thus the conditional pmf does not depend on λ (is algebraically independent of λ) therefore by definition T is a sufficient statistic for λ.

Another sufficient statistic is the n dimensional vector $T = (X_1, \ldots, X_n)$. However it is not typically interesting as it does not do any *data reduction*. Another not so interesting sufficient statistic in this problem is

$$T = \left(\sum_{i=1:i \text{ odd}}^{n} X_i \ , \ \sum_{i=1:i \text{ even}}^{n} X_i \right) \ .$$

End of Example

Aside : The student should review the Binomial formula. Do we know that the conditional pmf in the example above is indeed a pmf? It is non negative but does it sum to 1, summing over all possible n tuples $x_1, \ldots, x_n$? There is a multinomial formula that extends the Binomial formula, and the student should look this up.

Notice that if T is a sufficient statistic, then $T_1 = g(T)$ for a 1 to 1 (and hence invertible) function g is also a sufficient statistic. Thus in the Poisson example $\bar{X}$ is also a sufficient statistic. All sufficient statistics are 1 to 1 mappings of another equivalent sufficient statistic. Due to this property we often use terms such as *the sufficient* statistic even though there are many. For the Poisson all *minimal* sufficient statistics are in a 1 to 1 correspondence with $T = \sum_{i=1}^n X_i$. In particular $\bar{X}$, $\sqrt{\bar{X}}$ and $\exp \bar{X}$ are each *the* sufficient statistic. $1/\bar{X}$ is not a sufficient statistic, since for any finite n, $P(\bar{X} = 0) > 0$ and so $1/\bar{X}$ is not even a finite valued random variable.

A minimal sufficient statistic is one that is of smallest possible dimension. In practice this is what we always will look for and try to obtain. In practice the minimal sufficient statistic will of the same dimension as θ.

Why is the notion of sufficient statistic useful? It is most useful if we can make T of as small of a dimension as possible, that is a minimal sufficient statistic.

Consider the likelihood function $L(\theta)$. It is, for $t = T(x_1, \ldots, x_n)$,

$$\begin{aligned} L(\theta) &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \\ &= P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t) P_\theta(T = t) \\ &= h(x_1, x_2, \ldots, x_n) P_\theta(T = t) \end{aligned}$$

where

$$h(x_1, x_2, \ldots, x_n) = P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t)$$

which is algebraically independent of θ. Also h is positive and hence

$$\text{argmax}_{\theta \in \Theta} \text{L}(\theta) = \text{argmax}_{\theta \in \Theta} \text{P}_\theta(\text{T} = \text{t}) \ .$$

This means that for the purpose of calculating, algebraically or numerically, the MLE we may work with $P_\theta(T = t)$ or something proportional to $P_\theta(T = t)$, or equivalently the log of $P_\theta(T = t)$. This makes the algebra and the coding of algorithms much easier as the dimension of t does not change as n changes.

Finally from this we also see that the MLE $\hat{\theta}$ is a function of T, and this may then simplify obtaining the distribution of the MLE or an approximation to the distribution of the MLE.

Finding and verifying sufficient statistics by direct application of the definition is not very easy. Thus it is helpful if we can have an easy way to obtain sufficient statistics. The following theorem is generally quite easy to use in applications.

Theorem 2 (Factorization Theorem) *A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be a sufficient statistic is that the pdf (or pmf) of $X_1, \ldots, X_n$ is of the form*

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

for appropriate functions h and g.

Remark : Appropriate here means functions with domains that are relevant so that the notation makes sense in terms of the number of arguments and the sets to which these arguments belong. The function g has arguments (t, θ), so that in the Theorem the argument t is replaced by a function of the x_i, in particular $T(x_1, x_2, \ldots, x_n)$. The Theorem is called the Factorization Theorem since the pdf or pmf has to be able to factor into the product of two functions of the required types.

We only prove this Theorem in the case of discrete r.v.s, that is for pmf's. The continuous and more general cases require a deeper understanding of the role of derivatives, a topic called Radon-Nikodym derivatives in measure theory, and a topic which is needed for a complete description of conditional expectations. Also notice Theorem 2 is a necessary and sufficient condition statement, so that a proof must be done in both directions.

Proof of Theorem 2

1. Suppose that $T(X_1, \ldots, X_n)$ is a sufficient statistic. Then by definition

$$P_\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t) = h(x_1, x_2, \ldots, x_n)$$

for some function h, and in particular the RHS does not involve θ. Notice that in principle there should be $n+1$ arguments on the RHS, but t is determined by the values of $x_1, \ldots, x_n$.

Therefore the marginal pmf of $X_1, \ldots, X_n$ is given by the function

$$f(x_1, x_2, \ldots, x_n; \theta) = h(x_1, x_2, \ldots, x_n) P(T = t, \theta) \mathrm{I}\left(\mathrm{T}(\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n}) = \mathrm{t}\right)$$

and we take

$$g(t, \theta) = P_T(t; \theta) = P(T(X_1, \ldots, X_n) = t; \theta)$$

is the marginal pmf of T. For simplicity of notation we leave out the indicator function.

2. Suppose the pmf of $(X_1, \ldots, X_n)$ is of the form

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

We need to verify that $T(X_1, \ldots, X_n)$ is a sufficient statistic.

$$\begin{aligned} P(T = t; \theta) &= \sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} f(x_1, x_2, \ldots, x_n; \theta) \\ &= g(t, \theta) \sum_{x_1, \ldots, x_n : T(x_1, \ldots, x_n) = t} h(x_1, x_2, \ldots, x_n) \end{aligned}$$

Therefore

$$\begin{aligned} & P_\theta(X_1 = x_1, \ldots, X_n = x_n | T = t) \\ &= \frac{1}{g(t, \theta) \left\{ \sum_{y_1, \ldots, y_n : T(y_1, \ldots, y_n) = t} h(y_1, y_2, \ldots, y_n) \right\}} g(t; \theta) h(x_1, x_2, \ldots, x_n) \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right) \\ &= \frac{1}{\left\{ \sum_{y_1, \ldots, y_n : T(y_1, \ldots, y_n) = t} h(y_1, y_2, \ldots, y_n) \right\}} h(x_1, x_2, \ldots, x_n) \mathrm{I}\left(\mathrm{x}_1 + \mathrm{x}_2 + \ldots + \mathrm{x}_\mathrm{n} = \mathrm{t}\right) \end{aligned}$$

Thus the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ is algebraically independent of θ.

End of Proof

In general for a parametric model of with d dimensional parameter the sufficient statistic is of dimension d.

Example $X_i \sim \mathrm{Unif}(0, \theta)$ where $\theta > 0$. Then

$$f(x_1, \ldots, x_n; \theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathrm{I}(0 \leq \mathrm{x_i} \leq \theta)$$

Here we take $h = 1$ and

$$g(t, \theta) = \frac{1}{\theta^n} \mathrm{I}(0 \leq \mathrm{t} \leq \theta) \ .$$

Thus if we define $T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\} = x_{(n)}$ then

$$f(x_1, \ldots, x_n; \theta) = g(x_{(n)}, \theta) h(x_1, \ldots, x_n) \ .$$

Thus $X_{(n)}$ is the (minimal) sufficient statistic. *End of Example*

The student should find the minimal sufficient statistic for the iid $N(\mu, \sigma^2)$ sampling model and show that it is equivalent (1 to 1 correspondence) to

$$T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2 \right) \ .$$

In particular the student should show that an equivalent form for the sufficient statistic in this problem is

$$T' = \left(\bar{X}, S^2 \right)$$

the sample mean and sample variance.

Example

Gamma(α, λ)

$$\begin{aligned} L(\theta) &\equiv L(\alpha, \lambda) \\ &= \prod_{i=1}^{n} f(x_i; \alpha, \lambda) \\ &= \prod_{i=1}^{n} \frac{\lambda^\alpha x_i^{\alpha-1}}{\Gamma(\alpha)} e^{-\lambda x_i} \\ &= \frac{\lambda^{n\alpha}}{\Gamma(\alpha)^n} \left\{ \prod_{i=1}^{n} x_i \right\}^{\alpha-1} e^{-\lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

If we set $h(x_1, \ldots, x_n) = \mathrm{I}(\mathrm{x}_1 > 0, \ldots, \mathrm{x}_\mathrm{n} > 0)$ and

$$g(t_1, t_2, \theta) = \frac{\lambda^{n\alpha}}{\Gamma(\alpha)^n} \{t_1\}^{\alpha-1} e^{-\lambda t_2}$$

then we see that $L(\theta) = h(x_1, \ldots, x_n) g(T_1, T_2, \theta)$ when we set

$$T_1 = \prod_{i=1}^{n} X_i \ , \ T_2 = \sum_{i=1}^{n} X_i \ .$$

This means that we can find the MLE by maximizing $g(t_1, t_2, \alpha, \theta)$ with respect to $theta = (\alpha, \lambda)$ and substituting the observed values for t_1, t_2 or correspondingly maximizing

$$\log\left(g(t_1, t_2, \alpha, \lambda)\right) \ .$$

If we choose to work with $\log(g)$ then it is more convenient to use an equivalent sufficient statistic, that is

$$(T_1', T_2') = \left(\sum_{i=1}^{n} \log(X_i), \sum_{i=1}^{n} X_i \right) \ .$$

End of Example

Remark : In coding this MLE example above we only need to work with a function $L(\theta)$ and treat t_1, t_2 as two given numbers (the role of a *parameter* for this maximization problem).

Example Consider the exponential parameter λ model. The student should verify that $T = \sum_{i=1}^{n} X_i$ is the sufficient statistic. Thus when we are maximizing the log likelihood we are working with the function

$$L(\lambda) = n \log(\lambda) - \lambda t$$

where $t = \sum_{i=1}^{n} x_i$ or even more simply

$$\frac{1}{n} L(\lambda) = \log(\lambda) - \lambda \frac{t}{n} \ .$$

One can maximize this analytically and plot the function and find various nice properties of this function.

End of Example

5 Rao-Blackwell Theorem

This theorem is very interesting in that it tells us that any estimator that is not a function of a sufficient statistic can be replaced by another that is a function of a sufficient and that this new estimator has smaller mean square error. Therefore when we construct an estimator it is always best to use one that is a function of the minimal sufficient statistic, such as the maximum likelihood estimator. In practice again we always look for an estimator that is a function of the minimal sufficient statistic.

Theorem 3 (Rao-Blackwell Theorem) *Let $\hat{\theta}$ be an estimator of θ and suppose $E(\hat{\theta}^2) < \infty$. Suppose also that T is a sufficient statistic for θ. Let $\tilde{\theta} = E(\hat{\theta}|T)$ be the conditional expectation of $\hat{\theta}$ given T. Then*

$$E\left[\left(\tilde{\theta} - \theta\right)^2\right] \leq E\left[\left(\hat{\theta} - \theta\right)^2\right] \ .$$

The inequality is strict unless $\hat{\theta} = \tilde{\theta}$.

Remarks : $\hat{\theta} = \tilde{\theta}$ if and only if $\hat{\theta} = \mathrm{E}(\hat{\theta}|T)$, which happens when $\hat{\theta}$ is a function of T. In this case conditioning does not change anything. Otherwise there is a strict improvement. In particular if one conditions on a minimal sufficient statistic and $\hat{\theta}$ is not a function of this minimal sufficient statistic then the new estimator has smaller mean square error. Therefore it is always best to use an estimator that is a function of the minimal sufficient statistic. The MLE satisfies this.

The above remark makes the Rao-Blackwell Theorem one of great theoretical importance. However in practice it is very difficult to calculate the conditional expectation $\tilde{\theta}$, so in general we cannot make use of this when the estimator is not a function of the minimal sufficient statistic. Thus in general we try to find such estimators if possible.

End of Remarks

Proof of Rao-Blackwell Theorem

$$\mathrm{E}(\tilde{\theta}) = \mathrm{E}\left(\mathrm{E}(\hat{\theta}|T)\right) = \mathrm{E}(\hat{\theta}) \ .$$

Also

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(E(\hat{\theta}|T)) + \mathrm{E}\left(\mathrm{Var}(\hat{\theta}|T)\right) \ .$$

Therefore

$$\mathrm{Var}(\hat{\theta}) > \mathrm{Var}(\tilde{\theta})$$

unless $\mathrm{E}\left(\mathrm{Var}(\hat{\theta}|T)\right) = 0$ in which case

$$\mathrm{Var}(\hat{\theta}) = \mathrm{Var}(\tilde{\theta}) \ .$$

Since $\mathrm{Var}(\hat{\theta}|T) \geq 0$ with probability 1, then $\mathrm{E}\left(\mathrm{Var}(\hat{\theta}|T)\right) = 0$ if and only if $\mathrm{Var}(\hat{\theta}|T) = 0$ with probability 1, which happens if and only if $\hat{\theta}$ is constant with respect to T, that is it is a function of T.

If $\mathrm{Var}(\hat{\theta}|T) > 0$ with positive probability then

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{\theta} - \theta\right)^2\right] &= \mathrm{E}\left[\left(\tilde{\theta} - \mathrm{E}(\tilde{\theta}) + \mathrm{E}(\tilde{\theta}) - \theta\right)^2\right] \\
&= \mathrm{E}\left[\left(\tilde{\theta} - \mathrm{E}(\tilde{\theta})\right)^2\right] + \left[\mathrm{E}(\tilde{\theta}) - \theta\right]^2 \\
&= \mathrm{Var}(\tilde{\theta}) + \left[\mathrm{E}(\hat{\theta}) - \theta\right]^2 \\
&< \mathrm{Var}(\hat{\theta}) + \left[\mathrm{E}(\hat{\theta}) - \theta\right]^2 \\
&= \mathrm{E}\left[\left(\hat{\theta} - \theta\right)^2\right] \ .
\end{aligned}$$

This last line is obtained by a calculation similar to the reverse of the first few lines.

This completes the proof.

End of Proof

The student should return to find the form of the method of moments estimator for the Gamma family. Notice these estimators are functions of the first to sample moments, and therefore they are not functions of the minimal sufficient statistic. Thus the method of moments estimator can be improved. On the other hand the MLE is a function of the minimal sufficient statistic, and cannot be improved. Unfortunately the Rao-Blackwell Theorem does not say how these two estimators compare, it only says if they can be improved.