MATH3911 Summary

Alexander Kwok

August 14, 2014

Lecture 1: The Subject of Statistical Inference

Statistical Models

In general, we can view the statistical model as the triplet ($\mathcal{X}$, $\mathcal{P}$, Θ) where:

- $\mathcal{X}$ is the sample space (i.e. the set of all possible realizations $\mathbf{X} = (X_1, X_2, \cdots, X_n)$)
- $\mathcal{P}$ is a family of model functions $P_\theta(\mathbf{X})$ that depend on the unknown parameter θ
- Θ is the set of possible θ- values, i.e. the parameter space indexing the model

Lecture 2: Sufficient statistic

Definition 1. (sufficient statistic) *$P(X = x|T = t)$ is a function of x and t only. (i.e. is not a function of θ).* The intuition behind the sufficient statistic concept is that it contains all the information necessary for estimating θ.

Definition 2. *Let $X_1, \cdots, X_n$ be iid RVs whose distribution is the pdf f_X, or the pmf p_{x_i}. The likelihood function is the product of the pdfs or pmfs*

$$\mathrm{L}(x_1, \cdots, x_n|\theta) = \begin{cases} \prod_{i=1}^n f_{X_i}(x_i) & \text{if } X_i \text{ is a continuous RV} \\ \prod_{i=1}^n p_{X_i}(x_i) & \text{if } X_i \text{ is a discrete RV} \end{cases}$$

The likelihood function is sometimes viewed as a function of $x_1, \cdots, x_n$ (fixing θ) and sometimes as a function of θ (fixing $x_1, \cdots, x_n$). In the latter case, the likelihood is sometimes denoted $\mathrm{L}(\theta)$.

Definition 3. (sufficiency principle) The sufficiency principle implies that if T is sufficient for θ, then if x and y are such that $T(x) = T(y)$, then inference about θ should be the same whether $X = x$ or $Y = y$ is observed.

Theorem 1 (Neyman Fisher Factorization Criterion) *T is a sufficient statistic for θ if the likelihood factorizes into the following form*

$$L(x_1, \cdots, x_n|\theta) = g(\theta, T(x_1, \cdots, x_n)) \cdot h(x_1, \cdots, x_n)$$

for some functions g,h.

Proof. (For the discrete case)

$$p(x_1, \cdots, x_n|T(x_1, \cdots, x_n)) = \frac{p(x_1, \cdots, x_n, T(x_1, \cdots, x_n))}{p(T(x_1, \cdots, x_n))} = \frac{p(x_1, \cdots, x_n)}{\sum_{y:T(y)=T(x)} p(y_1, \cdots, y_n)} = \frac{h(x_1, \cdots, x_n)}{\sum_{y:T(y)=T(x)} h(y_1, \cdots, y}$$

which is not a function of θ. Conversely, assume that T is a sufficient statistic for θ. Then

$$L(x_1, \cdots, x_n|\theta) = p(x_1, \cdots, x_n|T(x_1, \cdots, x_n), \theta) = h(x_1, \cdots, x_n)g(T(x_1, \cdots, x_n), \theta)$$

Example (*normal population, unknown mean, known variance*) The joint density is

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

Since σ^2 is known, we can let

$$h(x) = (2\pi)^{-\frac{n}{2}}\sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}x_i^2\right)$$

and

$$g(T(X), \mu) = \exp\left(-\frac{n\mu^2}{2\sigma^2} + \frac{\mu}{\sigma^2}\sum_{i=1}^{n}x_i\right)$$

By the factorization theorem this shows that $\sum_{i=1}^{n} X_i$ is a sufficient statistic. It follows that the sample mean $\bar{X}_n$ is also a sufficient statistic.

Example (*Uniform distribution*) Suppose that X_i are uniformly distributed on $[0,\ \theta]$ where θ is unknown. Then the joint density is

$$f(x_1, \cdots, x_n|\theta) = \theta^{-n}1(x_i \leq \theta, i = 1, 2, \cdots, n)$$

Here 1(E) is an indicator function. It is 1 if the event E holds, 0 if it does not. Now $x_i \leq \theta$ for $i = 1, 2, \cdots, n$ if and only if $\max\{x_1, x_2, \cdots, x_n\} \leq \theta$. So we have

$$f(x_1, \cdots, x_n|\theta) = \theta^{-n}1(max\{x_1, x_2, \cdots, x_n\} \leq \theta)$$

By the factorization theorem this shows that

$$T = max\{X_1, X_2, \cdots, X_n\}$$

is a sufficient statistic.

Definition 4. (Minimal sufficient statistic) *A sufficient statistic allows the greatest data reduction without loss of information on θ.* A sufficient statistic is not uniquely defined. From the factorization theorem it is easy to see that (i) the identity function $T(x_1, \cdots, x_n) = (x_1, \cdots, x_n)$ is a sufficient statistic vector and (ii) if T is a sufficient statistic for θ then so any 1-1 function of T. A function that is not 1-1 of a sufficient static may or may not be a sufficient statistic. This leads to the notion of a minimal sufficient statistic.

Example: Since $T = (\sum X_i, \sum X_i^2)$ are jointly sufficient statistics for $\theta = (\mu, \sigma^2)$ for normally distributed data $X_1, \cdots, X_n \sum N(\mu, \sigma^2)$, then so are $(\bar{X}, S^2)$ which are a 1-1 function of $(\sum X_i, \sum X_i^2)$.

Theorem 2. (Simplified version of Theorem by Lehmann- Scheffe) Suppose there exists a function **T(x)** such that, for two sample points **x** and **y**, the ratio $\frac{L(x,\theta)}{L(y,\theta)}$ is constant as a function of θ if and only if **T(x)=T(y)**. Then **T(X)** is a minimal sufficient statistic for θ.

Example: T=$(\sum X_i, \sum X_i^2)$ is a minimal sufficient statistic for the Normal distribution since the likelihood ratio is not a function of θ iff T(x)= T(y).

$$\frac{L(x_1, \cdots, x_n|\theta)}{L(y_1, \cdots, y_n|\theta)} = e^{-\frac{1}{2\sigma^2}\sum(x_i-\mu)^2-(y_i-\mu)^2} = e^{-\frac{1}{2\sigma^2}(\sum x_i^2 - \sum y_i^2) + \frac{\mu}{\sigma^2}(\sum x_i - \sum y_i)}$$

When T(x)=T(y), it is not a function of θ nor zero function. Since $(\bar{X},\ S^2)$ is a function of T, it is minimal sufficient statistic as well.

Definition 5. (One parameter exponential family densities)

$$f(x, \theta) = a(\theta)b(x)\exp(c(\theta)d(x))$$

with $c(\theta)$ strictly monotone and $T = \sum_{i=1}^{n} d(X_i)$ is minimal sufficient.

Lecture 3: Maximum Likelihood Inference

Definition 1 (Likelihood principle) In the inference about θ, after x is observed, all relevant experimental information is contained in the likelihood function for the observed x. In particular,

"Data sets with proportional likelihood functions should lead to identical conclusions"

Definition 2: (Likelihood function) The likelihood of a set of parameter values, θ, given some observed outcomes, x, is equal to the probability of those observed outcomes given those parameter values, i.e.

$$\mathcal{L}(\theta|x) = P(x|\theta)$$

Definition 3: (Maximum Likelihood Estimation) The Maximum Likelihood Estimator (MLE) is defined as

$$\hat{\theta} = \arg[\sup_{\theta\in\Theta} L(x,\theta)]$$

Remark: It is more convenient to maximize not the function $L(x,\theta)$ but its logarithm. The function $\log L(x,\theta)$ is called the Log-likelihood. Since the logarithm is a monotone increasing function, maximization of $L(x,\theta)$ or of $\log L(x,\theta)$ is achieved for the same value of the argument $\hat{\theta}$.

Definition 4: (Normed likelihood) When comparing different models,

$$R(x,\theta) = \frac{L(x,\theta)}{L(x,\hat{\theta})}$$

which has a range in [0, 1]. An even more often used measure is the deviance $D(\theta)$ which is defined as

$$D(\theta) = -\log R(\theta) = -2[\log L(x,\theta) - \log L(x,\hat{\theta})]$$

The deviance is a non- negative number which can be attached to each model indexed by the parameter θ. The larger the deviance, the further the model from the "most likely" model.

Definition 5: (Fisher Information) can be defined as the variance of the score, or as the expected value of the observed information.

- We define $V(\mathbf{X},\theta) = \frac{\partial}{\partial\theta}\log L(\mathbf{X},\theta)$ to be the score function for non-i.i.d. random variable where $L(\mathbf{X},\theta)$ is the joint density. This indicates how sensitively a likelihood function $L(\theta X)$ depends on its parameter θ.
- In the case where $\mathbf{X}= (X_1, X_2, \cdots, X_n)$ and the X_i are i.i.d. with a density $f(x,\theta)$ then $V(\mathbf{X},\theta) = \sum_{i=1}^{n} \frac{\partial/\partial\theta f(X_i,\theta)}{f(X_i,\theta)}$

Properties:

- If $\hat{\theta}$ is the MLE then $V(x,\hat{\theta}) = 0$ holds. (This holds because $\hat{\theta}$ maximises $\log L(\mathbf{X},\theta)$ with respect to θ.)
- $E_\theta(V(\mathbf{X},\theta)) = 0$ holds under suitable regularity conditions.

Definition 6: (Expected Fisher Information) about θ contained in the vector $\mathbf{X}$:
It is denoted by $I_X(\theta)$ and is defined as

$$I_X(\theta) = Var_\theta(V(\mathbf{X},\theta)) = E_\theta\left\{\frac{\partial}{\partial\theta}\ln L(\mathbf{X},\theta)\right\}^2$$

(where we utilized the fact that $E_\theta(V,\mathbf{X},\theta)) = 0$)

<u>Proof:</u>

$$\mathbb{E}(V|\theta) = \int_{-\infty}^{\infty} \frac{\frac{\partial f(x;\theta)}{\partial\theta}}{f(x;\theta)} f(x;\theta)dx = \int_{-\infty}^{\infty} \frac{\partial f(x;\theta)}{\partial\theta}dx$$

If certain differentiability conditions are met, the integral may be rewritten as

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty} f(x;\theta)dx = \frac{\partial}{\partial\theta}1 = 0$$

Properties:

1. Additivity over independent samples: if X and Y are independent random variables whose densities depend on θ then for the information in the vector $\mathbf{Z}$= (X, Y) we have

$$I_{\mathbf{Z}}(\theta) = I_{(X,Y)}(\theta) = I_X(\theta) + I_Y(\theta)$$

In particular, when sampling n times, the information in the sample about the parameter equals n times the information in a single observation about the parameter: If $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ then

$$I_X(\theta) = nI_{X_1}(\theta)$$

2. If $T(X)$ is sufficient for θ then $I_T(\theta) = I_X(\theta)$

3. Under regularity conditions: $I_X(\theta) = -E(\frac{\partial^2}{\partial\theta^2} \ln L(X, \theta))$

4. For any statistics $T(X)$ it holds: $I_T(\theta) \leq I_X(\theta)$ with equality if and only if T is sufficient for θ This property most clearly underlines the importance of sufficiency when we try to perform data reduction without loss of information about the parameter.

Sketch of proofs

1. Starting with $L_{(X,Y)}(x, y; \theta) = L_X(x; \theta)L_Y(y; \theta)$. We take logarithms of both sides first and then calculate partial derivatives with respect to θ of both sides. In the resulting equality, we square both sides and take expected values. This gives us:

$$E_\theta \left[\frac{\partial}{\partial\theta} \log L_{(X,Y)}(X, Y, \theta)\right]^2 = I_X(\theta) + I_Y(\theta) + 2E_\theta[V(X, \theta)V(Y, \theta)]$$

Since X and Y are independent:

$$E_\theta[V(X, \theta)V(Y, \theta)] = E_\theta V(X, \theta)E_\theta V(Y, \theta) = 0$$

2. (for the discrete case) First, let us note that because of the sufficiency,

$$f_T(t, \theta) = \sum_{x:T(x)=t} f_X(x, \theta) = \sum_{x:T(x)=t} g(T(x), \theta)h(x) = g(t, \theta) \sum_{x:T(x)=t} h(x)$$

holds and hence $E_\theta \left[\frac{\partial}{\partial\theta} \log f_T(T; \theta)\right]^2 = E_\theta \left[\frac{\partial}{\partial\theta} \log g_T(T; \theta)\right]^2$ holds. Then

$$I_T(\theta) = E_\theta \left[\frac{\partial}{\partial\theta} \log f_T(T; \theta)\right]^2 = E_\theta \left[\frac{\partial}{\partial\theta} \log g(T; \theta)\right]^2 =$$

$$E_\theta \left[\frac{\partial}{\partial\theta} (\log g(T; \theta) + \log h(X))\right]^2 = E_\theta \left[\frac{\partial}{\partial\theta} \log L(X; \theta)\right]^2 = I_X(\theta)$$

3. If $f(x, \theta)$ denotes the density of a single observation and under suitable differentiability conditions, we can write:

$$\frac{\partial^2}{\partial\theta^2}(\log f(x, \theta)) = \frac{\partial}{\partial\theta}\left(\frac{f'(x, \theta)}{f(x, \theta)}\right) = \frac{\frac{\partial^2}{\partial\theta^2} f(x, \theta)}{f(x, \theta)} - \left[\frac{\frac{\partial}{\partial\theta} f(x, \theta)}{f(x, \theta)}\right]^2$$

Under suitable regularity conditions that allow for exchange of order of integration and differentiation, we have

$$E_\theta \left[\frac{\frac{\partial^2}{\partial\theta^2} f(x, \theta)}{f(x, \theta)}\right] = \int \frac{\frac{\partial^2}{\partial\theta^2} f(x, \theta)}{f(x, \theta)} \cdot f(x, \theta)dx = \frac{\partial^2}{\partial\theta^2} \int f(x, \theta)dx = \frac{\partial^2}{\partial\theta^2} 1 = 0$$

holds.

4. For random variables Z and Y and a function g(z) we can write

$$\begin{aligned} E(g(Z)Y|Z=z) &= g(z)E(Y|Z=z) \\ E(Y) &= E_Z(E(Y|Z=z)) \end{aligned}$$

Since the expected value of the square of any random variable is non0 negative, we know that:

$$0 \le E\left\{\frac{\partial}{\partial\theta}\log L(X,\theta) - \frac{\partial}{\partial\theta}\log f_T(T,\theta)\right\}^2 = I_X(\theta) + I_T(\theta) - 2E\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}f_T(T,\theta)\right]$$

Consider

$$E\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\frac{\partial}{\partial\theta}f_T(T,\theta)\right] = E_T\left[\frac{\partial}{\partial\theta}\log f_T(t,\theta)E\left(\frac{\partial}{\partial\theta}\log L(X,\theta)|T=t\right)\right]$$

Hence $0 \le I_X(\theta) + I_T(\theta) - 2I_T(\theta)$.

$$I_T(\theta) \le I_X(\theta)$$

Lecture 4: Classical Estimation Theory

Definition 1: (Bias) Suppose $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ are i.i.d. from $f(x,\theta)$, $\theta \in \Theta \in \mathcal{R}$ and we use a statistic $T_n(\mathbf{X})$ to estimate θ. If $E_\theta(T_n) = \theta + b_n(\theta)$ then the quantity

$$b_n(\theta) = E_\theta(T_n) - \theta$$

We call the estimator unbiased if $b_n(\theta) \equiv 0$ for $\theta \in \Theta$.

- Interpretation of unbiasedness: when used repeatedly, an unbiased estimator, in the long run, will estimate the true value on average.

Caution: Some families an unbiased estimators may not exist, or even when they exist, may not be very useful. For example, if using a single observation x from geometrix distribution $f(x,\theta) = \theta(1-\theta)^{x-1}$, $x = 1, 2, \cdots$ an unbiased estimator of θ, say, $T(x)$ must satisfy $\sum_{x=1}^{\infty} T(x)\theta(1-\theta)^{x-1} = \theta$ for all $\theta \in [0,1]$. Cancelling θ on both sides, and setting $\tilde{\theta} = 1 - \theta$ we get

$$\sum_{x=1}^{\infty} T(x)\tilde{\theta}^{x-1} = 1, \text{ for all } \tilde{\theta} \in [0,1]$$

Hence, by a polynomial expansion, the only estimator satisfying this requirement would be $T(1) = 1$, $T(x) = 0$ if $x \ge 2$. SUch an estimator is neither very reliable, nor very useful.

Definition 2: (Mean Squared Error)

$$MSE_\theta(T_n) = E_\theta(T_n - \theta)^2$$

Remark: The following property holds:

$$MSE_\theta(T_n) = Var_\theta T_n + (b_n(\theta))^2$$

Indeed,

$$\begin{aligned} MSE_\theta(T_n) &= E_\theta[(T_n - E_\theta T_n + E_\theta T_n - \theta]^2 \\ &= E_\theta(T_n - E_\theta T_n)^2 - 2E_\theta[(T_n - E_\theta T_n)(\theta - E_\theta T_n)] + E_\theta(E_\theta T_n - \theta)^2 \\ &= Var(T_n) + (b_n(\theta))^2 \end{aligned}$$

Keeping the MSE as small as possible is more important than just to ask for unbiasedness.

Definition 3: (Uniformly minimum variance unbiased estimator (UMVUE) An unbiased estimator with the smallest MSE(=Var) for all θ values

Theorem 1: (Cramer- Rao theorem) The bound states that the variance of any unbiased estimator is at least as high as the inverse of the Fisher information. An unbiased estimator which achieves this lower bound is said to be (fully) efficient.

Consider an unbiased estimator $W(\mathbf{X})$ of $\tau(\theta)$, i.e. $E_\theta W(\mathbf{X}) = \tau(\theta)$. Suppose, in addition, that $L(\mathbf{X}, \theta)$ satisfies:

$$\frac{\partial}{\partial\theta} \int \cdots \int h(\mathbf{X}) L(\mathbf{X}, \theta) dX_1 \cdots dX_n = \int \cdots \int h(\mathbf{X}) \frac{\partial}{\partial\theta} L(\mathbf{X}, \theta) dX_1 \cdots dX_n$$

for any function $h(\mathbf{X})$ with $E_\theta |h(\mathbf{X})| < \infty$. Then:

$$Var_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{I_X(\theta)}$$

for all θ holds.

Corollary for i.i.d. case If $\mathbf{X}$=$(X_1, X_2, \cdots, X_n)$ are i.i.d. with $d(x, \theta)$ then L($\mathbf{X}$, θ)= $\prod_{i=1}^n f(X_i, \theta)$; $I_X(\theta) = nI_{X_1}(\theta)$ and the CR Inequality becomes:

$$Var_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{nI_{X_1}(\theta)}$$

Proof:
Cauchy-Schwartz Inequality: If Z and Y are two random variables with finite variances Var(Z) and Var(Y) then

$$[Cov(Z, Y)]^2 = \{E[(Z - E(Z))(Y - E(Y))]\}^2 \leq Var(Z)Var(Y)$$

holds. To prove the Cramer- Rao Theorem, we choose W to be the Z- variable, and the score V to be the Y- variable in the Cauchy- Schwartz Inequality. Since $E_\theta V(\mathbf{X}, \theta) = 0$ holds for the score, we have that

$$[Cov_\theta(WV)]^2 = [E_\theta(WV)]^2 \leq Var_\theta(W)Var_\theta(V)$$

Substituting the definition of the score, we get:

$$Cov_\theta(WV) = E_\theta(WV) = \int \cdots \int W(\mathbf{X}) \frac{\frac{\partial}{\partial\theta} L(\mathbf{X}, \theta)}{L(\mathbf{X}, \theta)} L(\mathbf{X}, \theta) d\mathbf{X}$$

where $d\mathbf{X} = dX_1 dX_2 \cdots dX_n$ is used as shorthand Now if we utilize condition, we can continue to get:

$$Cov_\theta(WV) = \frac{\partial}{\partial\theta} E_\theta W = \frac{\partial}{\partial\theta}\tau(\theta)$$

Then, Inequality implies:

$$Var_\theta(W(\mathbf{X})) \geq \frac{(\frac{\partial}{\partial\theta}\tau(\theta))^2}{I_X(\theta)}$$

Note: The Cramer- Rao(CR) Inequality was stated for continuous random variables. By interchanging differentiation and summation (instead of differentiation and integration) one can formula this for discrete random variables.

Comments on applying the CR Inequality in the search of the UMVUE

1. In case where there exists an unbiased estimator of $\tau(\theta)$ whose variance is equal to the lower bound given by CR Inequality, this will be the UMVUE of $\tau(\theta)$.

2. CR Theorem often happens that it is not satisfied. A typical situation is when the range of the random variables X_i, $i = 1, 2, \cdots, n$ depends on θ, for example, in the case of random sample from uniform $[0, \theta)$ observations. According to the general Leibnitz' rule for differentiation of parameter- dependent integrals:

$$\frac{\partial}{\partial\theta}\int_{a(\theta)}^{b(\theta)} f(x,\theta)dx = f(b(\theta),\theta)\frac{\partial}{\partial\theta}b(\theta) - f(a(\theta),\theta)\frac{\partial}{\partial\theta}a(\theta) + \int_{a(\theta)}^{b(\theta)}\frac{\partial}{\partial\theta}f(x,\theta)dx$$

   holds and we see that, if a and b were genuine functions of θ, on the RHS there would be some additional non- zero terms included.

3. Even in cases where the CR Theorem is applicable, there is no guarantee that the lower bound on the variance is attainable. The bound is attainable if and only if we have equality in the Cauchy- Schwartz Inequality which means that the score $V(\mathbf{X}.\theta)$ must have a representation of the form

$$V(\mathbf{X}.\theta) = k_n(\theta)[W(\mathbf{X}) - \tau(\theta)]$$

   If $V(\mathbf{X}, \theta)$ can not be written in this form then no unbiased estimator of $\tau(\theta)$ would have a variance equal to the one given by the CR bound and in this case the CR Inequality would be of no use when searching for UMVUE.

   Example: (Attainable) Estimating the parameter θ in a Poisson(θ) distribution. In this case the CR bound is attainable and the unbiased estimator that attains the bound is $\hat{\theta} = \bar{X}$.
   Ans: The score $V(\mathbf{X}, \theta) = -n + \frac{n\bar{X}}{\theta} = -\frac{n}{\theta}(\theta - \bar{X})$ where $k_n(\theta) = -\frac{n}{\theta}$, $W(\mathbf{X}) = \theta$ and $\tau(\theta) = \bar{X}$.

   Example: (Not attainable) Estimating the function $\tau(\theta 0 = \exp(-\theta)$ from a sample of Poisson (θ) distribution. In this case no unbiased estimator of $\tau(\theta)$ has variance equal to the bound.

Theorem 2: If under the regularity conditions of CR Theorem there is an estimator of $\tau(\theta)$ which attins the lower bound, it should be the MLE of $\tau(\theta)$.

Theorem 3: (Rao- Blackwell Theorem) Let W be any unbiased estimator of $\tau(\theta)$ and let T be a sufficient statistic for θ. Define $\tau(\hat{T}) = E(W|T)$. Then $E_{\theta\hat{\tau}}(T) = \tau(\theta)$ and $Var_{\theta\hat{\tau}}(T) \leq Var_{\theta}W$ for all $\theta \in \Theta$, i.e. $\hat{\tau}(T)$ is uniformly better than W as an estimator of $\tau(\theta)$.

Theorem 4: (Uniqueness of UMVUE) If an estimator W is UMVUE for $\tau(\theta)$, then W is unique. Moreover, W is UMVUE if and only if W is uncorrelated with all unbiased estimators of zero.

Definition 4: (Complete Statistics) If we have a sufficient statistic that optimally summaries the data, then there shouldn't be an ancillary statistic that is function of that static.

- somewhat stronger and easier to check we can insist that there is no function of T whose expectation does not depend on θ.

Theorem 5: (Completeness of a family of distributions) Let $\tilde{f}(t, \theta)$, $\theta \in \Theta$ be a family of distributions for statistic T($\mathbf{X}$). The family is called complete if $E_{\theta g}(T) = 0$ for all $\theta \in \Theta$ imples $P_\theta(g(T) = 0) = 1$ for all $\theta \in \Theta$. Equivalently, T(X) is called a complete statistic for θ.

- if $h(T)$ is ancillary then it has some expectation and we can form

$$g(T) = h(T) - E(h(T))$$

  which has a zero expectation and thus T is not compete.

Theorem 6: If P is an exponential family of full rank with p.d.f.'s given by

$$f(x,\theta)=a(\theta)b(x)\exp(c(\theta)d(x))$$

then $T(X)$ is complete and sufficient for θ.
Theorem 7: (*Basu's Theorem) In a complete family, evern ancillary statistic is independent of the minimal sufficient statistic.

1. $\bar{x}$ is independent of s^2
2. $U_{(n)}$ is independent of R or $U_{(n)}/U_{(1)}$

Proof: Take T to be a complete sufficient statistic and S be an ancillary statistic. For any event A, the $\mathbb{P}\{S\in A\}$ does not depend on θ because S is ancillary. $\mathbb{P}\{S\in A|T=t\}$ also does not depend on θ because T is sufficient. Take for our g function

$$g(t)=\mathbb{P}\{S\in A\}-\mathbb{P}\{S\in A|T=t\}$$

Tkae the expectation of the function g,

$$\mathbb{E}_\theta g(t)=\mathbb{P}\{S\in A\}-\mathbb{E}_\theta\mathbb{P}\{S\in A|T\}=0$$

for all θ. By the assumption of completeness, this implies that $\mathbb{P}\{g(t)=0\}=1$ and

$$\mathbb{P}\{S\in A\}=\mathbb{P}\{S\in A|T=t\}$$

a.s. T. Thus

$$\mathbb{P}\{S\in A,T\in B\}=\mathbb{P}\{S\in A\}\mathbb{P}\{T\in B\}$$

for any sets A and B and therefore S and T are independent.

Theorem 7: (Theorem of Lehmann- Scheffe) Let T be a complete sufficient statistic for a parameter θ and W be any unbiased estimator of $\tau(\theta)$. Then $\hat{\tau}(T)=E(W|T)$ is unique UMVUE of $\tau(\theta)$.

Lecture 5: Asymptotic properties of estimators

Convergence concepts in asymptotics

An estimator T_n of the parameter θ is said to be

1. consistent (or weakly consistent) if

   $$\lim_{n\to\infty}P_\theta(|T_n-\theta|>\epsilon)=0$$

   for all $\theta\in\Theta$ and for every fixed $\epsilon>0$. We denote this by $T_n\xrightarrow{P}\theta$.
2. strong consistent if $P_\theta(\lim_{n\to\infty}T_n=\theta)=1$ for all $\theta\in\Theta$.
3. mean- square consistent if $MSE_\theta(T_n)\to_{n\to\infty}0$ for all $\theta\in\Theta$.

Strong consistency implies weak consistency

Mean- square consistency and consistency: If the estimator is mean- square consistent then it is consistent.We can use the Chebyshev Inequality. It states that for any random variable X and any $\epsilon>0$ it holds for the k-th moment:

$$P(|X|>\epsilon)\leq\frac{E(|X|)^k}{\epsilon^k}$$

Applying this inequality for X being $T_n - \theta$ and $k = 2$ we get

$$0 \leq P(|T_n - \theta| > \epsilon) \leq \frac{MSE_\theta(T_n)}{\epsilon^2}$$

Therefore, if an estimator T_n is mean- square consistent and the RHS tends to zero, the LHS will also tend to zero thus implying consistency.

Convergence in distribution: Assume that the sequence of random variances $X_1, X_2, \cdots, X_n$ have cumulative distribution functions $F_1, F_2, \cdots, F_n$ respectively. Assume the continuous random variable X has a cdf F and that it holds for each argument $x \in R$ that $\lim_{n\to\infty} F_n(x) = F(x)$. Then we say that the sequence of random variables $\{X_n\}$, $n = 1, 2, \cdots$ converges weakly (or in distribution) to X and denote this fact by $X_n \xrightarrow{d} X$.

Consistency and asymptotic normality of MLE

Let X= $(X_1, X_2, \cdots, X_n)$ be i.i.d. from $f(x, \theta)$, $\theta \in \Theta \in R^1$. Θ - open interval. Assume, following regularity conditions are satisfied:

1. $\frac{\partial f}{\partial \theta}(x, \theta)$, $\frac{\partial^2 f}{\partial \theta^2}(x, \theta)$, $\frac{\partial^3 f}{\partial \theta^3}(x, \theta)$ exist for all x and all $\theta \in \Theta$.
2. $\frac{\partial}{\partial \theta} \int f(x, \theta) dx \int \frac{\partial}{\partial \theta} f(x, \theta)$; $\frac{\partial^2}{\partial \theta^2} \int f(x, \theta) dx \int \frac{\partial^2}{\partial \theta^2} f(x, \theta)$;
3. $0 < I(\theta) = E_\theta \left(\frac{\partial lnf}{\partial \theta}(x, \theta)^2 \right) < \infty$ for all $\theta \in \Theta$
4. $\left| \frac{\partial^3 lnf}{\partial \theta^3}(x, \theta) \right| \leq H(x)$ for all $\theta \in \Theta$ with $E_\theta H(X) = \int H(x) f(x, \theta) dx \leq C$, C not depending on $\theta \in \Theta$.

Let θ_0 be the "true" value of θ. Then the MLE $\hat{\theta}_n$ of θ_0 is strongly consistent and asymptotically normal, i.e.

1. $P_{\theta_0}(X : \hat{\theta}_n \to \theta_0) = 1$
2. $\sqrt{n}(\hat{\theta}_n \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$

Proof:

(a) Notice that

$$\frac{1}{n} \log L(X, \hat{\theta}_n) \geq \frac{1}{n} \log(X, \theta_0)$$

where $L(.,.)$ denotes the joint density of n independent identically distributed (i.i.d.) observations, each with a density $\int(.,.)$.

Notice that by Jensen's Inequality:

$$E_{\theta_0} \left[\log \frac{L(X, \theta)}{L(X, \theta_0)} \right] < \log E_{\theta_0} \left[\frac{L(X, \theta)}{L(X, \theta_0)} \right] = 0$$

which implies

$$E_{\theta_0} \left[\frac{1}{n} \log L(X, \theta) \right] < \log E_{\theta_0} \left[\frac{1}{n} L(X, \theta_0) \right]$$

The Law of Large numbers implies then that for a fixed $\theta \neq \theta_0$ we should have

$$P_{\theta_0} \left\{ \lim_{n\to\infty} \frac{1}{n} \log L(X, \theta) < \lim_{n\to\infty} \frac{1}{n} \log L(X, \theta_0) \right\} = 1$$

We see that we need to have $P_{\theta_0}(\hat{\theta}_n \to \theta_0) = 1$

(b) Since

$$0 = \frac{\partial}{\partial\theta}\log L(X,0)_{|\theta=\hat{\theta}_n} = \sum_{i=1}^{n} \log f(X_i,\theta)_{|\theta=\hat{\theta}_n}$$

holds, after Taylor expansion around θ_0 and recollection of terms we get

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{(-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta} f(x_i,\theta_0))/[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)]}{1+\frac{1}{2}(\hat{\theta}_n-\theta_0)\frac{\frac{1}{n}\sum_{i=1}^n \eta_i H(x_i)}{\frac{1}{n}\sum i=1^n \frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)}}$$

Here η_i are intermediate values: $|\eta_i| < 1$

- the Law of large numbers regarding the term $[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta^2}\log f(x_i,\theta_0)]$
- the central limit theorem regarding the term $-\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial}{\partial\theta} f(x_i,\theta_0)$
- the uniform bound assumption

$$\frac{1}{n}\sum_{i=1}^{n}|\eta_i H(X_i)| \le \frac{1}{n}\sum_{i=1}^{n} H(X_i) \to E_{\theta_0}H(X_i) \le C$$

Comments on asymptotic properties of MLE

- The asymptotic normality of the MLE and the form of the asymptotic variance show that $\lim_{n\to\infty} Var(\hat{\theta}_{mle})\cdot [nI_{X_1}(\theta) = 1]$ for all $\theta \in \Theta$ which means that MLE "asymptotically achieve the CR bound on variance".

Definition 1: (Transformation invariance property of MLE) $h(\hat{\theta})$ is the MLE of $h(\theta)$.

Definition 2: (Delta method) Since the transformation is assumed to be smooth, we can expand $h(\hat{\theta})$ around the true parameter θ_0:

$$h(\hat{\theta}_{mle}) = h(\theta_0) + (\hat{\theta}_{mle}-\theta_0)\frac{\partial h}{\partial\theta}(\theta_0) + \frac{1}{2}(\hat{\theta}_{mle}-\theta_0)^2\frac{\partial^2 h(\theta_0)}{\partial\theta^2} + \cdots$$

We get the convergence in distribution

$$\sqrt{n}(h(\hat{\theta}_{mle}) - h(\theta_0)) \xrightarrow{d} N\left(0, \left[\frac{\partial h}{\partial\theta}(\theta_0)\right]^2 I^{-1}(\theta_0)\right)$$

Roughly, we shall also say that the distribution of $h(\hat{\theta}_{mle})$ can be approximated by

$$h(\hat{\theta}_{mle}) \approx N\left(h(\theta_0), \frac{1}{n}\left[\frac{\partial h}{\partial\theta}(\theta_0)\right]^2 I^{-1}(\theta_0)\right)$$

Definition 3: (Fisher Inofrmation matrix)

$$N\left(h(\vec{\theta_0}), \nabla h(\vec{\theta_0})' I_X(\vec{\theta_0})^{-1}\nabla h(\vec{\theta_0})\right)$$

Exact and asymptotic distributions for the deviance statistic

- Chi square approximation of the deviance. Assume that $\theta = \theta_0$ is the "true" value of the population parameter. Expand the Log- likelihood in Taylor series around $\theta = \hat{\theta}_{mle}$:

$$\log L(X.\theta_0) = \log L(X,\hat{\theta}_{mle}) + (\theta_0 - \hat{\theta}_{mle})\frac{\partial \log L(x,\hat{\theta}_{mle})}{\partial\theta} + \frac{1}{2}(\hat{\theta}_{mle}-\theta_0)^2\frac{\partial^2 \log L(X,\hat{\theta}_{mle})}{\partial\theta^2} + \cdots$$

Because the second summand in the RHS vanishes at $\hat{\theta}_{mle}$, ignoring higher order terms, we get:

$$D(\theta_0) \approx (\hat{\theta}_{mle} - \theta_0)^2 \left\{ -\frac{\partial^2 \log L(X, \hat{\theta}_{mle})}{\partial \theta^2} \right\}$$

Using the results about the normal approximation to the distribution of MLE, we get the deviance has an asymptotic χ^2 distribution with one degree of freedom.

$$(\theta_0 - \hat{\theta}_{mle})' \left\{ -\frac{\partial^2 \log L(X, \hat{\theta}_{mle})}{\partial \theta^2} \right\} (\theta_0 - \hat{\theta}_{mle}) \sim \chi^2_p$$

Lecture 6: Hypothesis testing

Simple hypothesis: $H_0 : \theta = \theta_0$ versus a single alternative $H_1 : \theta = \theta_1$. A test

$$\varphi(x) = P(\text{reject } H_0 | X = x)$$

Based on the observations we calculate $\varphi(x)$ and according to its value, we reject H_0 (if it happens that $\varphi(x) = 1$) or do not reject it (if $\varphi(x) = 0$).

Types of errors:

- To reject H_0 given that H_0 is correct (first type of error/ false alarm)
- To accept H_0 given that H_1 is correct (second type of error)

The corresponding proabilities are denoted as follows:

$$P(\text{reject } H_0 | H_0 \text{ correct}) = \text{level of the test (significance)}$$

$$P(\text{accept } H_0 | H_1 \text{ correct}) = 1 - (\text{power of thet test})$$

Randomized allowing $\varphi(x)$ to take any value in $[0, 1]$

$$\begin{aligned} P(\text{reject } H_0 | H_0 \text{ correct}) &= \int \cdots \int P(\text{reject } H_0 | X = x) L(x, \theta_0) dx \\ &= \int \cdots \int \varphi(x) L(x, \theta_0) dx = E_{\theta_0} \varphi \\ P(\text{accept } H_0 | H_1 \text{ true}) &= \text{similarly to the argument above} = 1 - E_{\theta_1} \varphi \end{aligned}$$

Neyman- Pearson lemma states that when performing a hypothesis test between two point hypothesis $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$, then the likelihood- ratio test which rejects H_0 in favour of H_1 when

$\Lambda(X) = \dfrac{L(\theta_0|x)}{L(\theta_1|x)} \leq C$, when $P(\Lambda(X) \leq X | H_0) = \alpha$ is the most powerful test of size α for a constant C

1. For every $\alpha \in (0, 1)$ there exists a constant C and a test

$$\varphi* = \begin{cases} 1 & \text{if } x \in S = \{x : L(x, \theta_1)/L(x, \theta_0) > C\}, \\ \gamma & \text{if } x \in R = \{x : L(x, \theta_1)/L(x, \theta_0) = C\}, \\ 0 & \text{if } x \in A = \{x : L(x, \theta_1)/L(x, \theta_0) < C\} \end{cases}$$

with $E_{\theta_0} \varphi* = 1 \times P_{\theta_0}(X \in S) + \gamma \times P_{\theta_0}(X \in R) = \cdots = \alpha$. The constant $\gamma \in (0, 1)$ in the definition of the test is equal to $\gamma = \frac{\alpha - P_{\theta_0}(S)}{P_{\theta_0}(R)}$;

2. $\varphi*$ is the best $\alpha-$ test, i. e. $E_{\theta_1} \varphi*$ is maximal among all tests $\varphi \in \phi_\theta = \{\varphi | E_{\theta_0} \varphi \leq \alpha\}$.

3. $\varphi*$ is essentially unique, i.e. all other "best" α- tests in the sense of 2. must coincide with $\varphi*$ on S and A.

Proof

1. Given α, we define C to be the smallest value on the real line for which $P_{\theta_0}\{\frac{L(X,\theta_1)}{L(X,\theta_0)} > C\}$ is still $\leq \alpha$. The constant C is called the upper $\alpha * 100\%$- point of the distribution of $\frac{L(X,\theta_1)}{L(X,\theta_0)}$ when θ_0 is the true parameter.

2. Take any other α- test φ and divide the sampe space $\mathcal{X}$ into $\mathcal{X} = \mathcal{X}^+ \cup \mathcal{X}^- \cup \mathcal{X}^=$ with:
$$\mathcal{X}^+ = \{X : \varphi * (X) - \varphi(X) > 0\}$$
$$\mathcal{X}^- = \{X : \varphi * (X) - \varphi(X) < 0\}$$
$$\mathcal{X}^= = \{X : \varphi * (X) - \varphi(X) = 0\}$$
Analyzing the expression $Z(X) = (\varphi * (X) - \varphi(X))(L(X,\theta_1) - CL(X,\theta_0))$ separately for values of $X \in \mathcal{X}^+$, $X \in \mathcal{X}^-$ and $X \in \mathcal{X}^=$, we see that always $Z(X) \geq 0$ holds.

   But
$$\int_{\mathcal{X}} (\varphi * (X) - \varphi(X))(L(X,\theta_1) - CL(X,\theta_0))dX \geq 0$$
$$E_{\theta_1}\varphi* \geq E_{\theta_1}\varphi$$
Since φ was arbitrarily chosen in the set of α- tests, this implies that $\varphi*$ can not be improved with respect to power, that is, it is the best α- size test.

3. If $\bar{\varphi}$ is another "best" α test (that is $E_{\theta_1}\varphi* = E_{\theta_1}\bar{\varphi}$ holds) then we necessarily need to have $Z(X) \equiv 0$. Since $Z(X)$ is a product of two factors, we either have one of the factos being zero. Hence, always when X is not in R but in S or in A, we must have $\varphi * (X) = \bar{\varphi}(X)$.

Example: Suppose X is a single observation from a normal population with unknown mean μ and known standard deviation $\sigma = \frac{1}{3}$. Then, we can apply the Nehman Pearson Lemma when testing the simple null hypothesis $H_0 : \mu = 3$ against the simple alternative hypothesis $H_1 : \mu = 4$.

Example: Suppose $X_1, X_2, \cdots, X_n$ is a random sample from a normal population with mean μ and variance 16. Find the test with the best critical region, that is, find the most powerful test, with a sample size of $n = 16$ and a significance level $\alpha = 0.05$ to test the simple null hypothesis $H_0 : \mu = 10$ against the simple alternative hypothesis $H_1 : \mu = 15$.

$$\frac{L(10)}{L(15)} = \frac{(32\pi)^{-16/2}\exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i - 10)^2\right]}{(32\pi)^{-16/2}\exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i - 15)^2\right]} \leq k$$

Simplifying we get:

$$\exp\left[-\left(\frac{1}{32}\right)\left(\sum_{i=1}^{16}(x_i - 10)^2 - \sum_{i=1}^{16}(x_i - 15)^2\right)\right] \leq k$$
$$\exp\left[-\left(\frac{1}{32}\right)\left(\sum_{i=1}^{16}x_i^2 - 20\sum_{i=1}^{16}x_i + 1600 - \sum_{i=1}^{16}x_i^2 + 30\sum_{i=1}^{16}x_i - 3600\right)\right] \leq k$$
$$-10\sum_{i=1}^{16}x_i + 2000 \leq 32ln(k)$$
$$\frac{1}{16}\sum_{i=1}^{16}x_i \geq -\frac{1}{160}(32\ln(k) - 2000)$$

That is, the Neyman Pearson Lemma tells us that the rejection region for the most powerful test of testing $H_0 : \mu = 10$ against $H_1 : \mu = 15$, under the normal probability model, is of the form:

$$\bar{x} \geq k^*$$

where k^* is selected so that the size of the critical region is $\alpha = 0.05$. Under the null hypothesis, the sample mean is normally distribution with mean 10 and standard deviation 4/4=1. Therefore the critical value k^* is deemed to be 11.645.

The power of such a test when $\mu = 15$ is:

$$P(\bar{X} > 11.645 \text{ when } \mu = 15) = P\left(Z > \frac{11.645 - 15}{\sqrt{16}/\sqrt{16}}\right) = P(Z > -3.36) = 0.9996$$

Example: Similiar to the example above, find the uniformly most powerful test to test the simple null hypothesis $H_0 : \mu = 10$ against the composite alternative hypothesis $H_1 : \mu > 10$.

$$\frac{L(10)}{L(\mu_\alpha)} = \frac{(32\pi)^{-16/2} \exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i - 10)^2\right]}{(32\pi)^{-16/2} \exp\left[-\frac{1}{32}\sum_{i=1}^{16}(x_i - \mu_\alpha)^2\right]} \leq k$$

Simplying we get:

$$-2(\mu_\alpha - 10)\sum_{i=1}^{16} x_i + 16(\mu_\alpha^2 - 10^2) \leq 32\ln(k)$$

$$\frac{1}{16}\sum_{i=1}^{16} x_i \geq -\frac{1}{16(2(\mu_\alpha - 10))}(32\ln(k) - 16(\mu_\alpha^2 - 10^2))$$

Therefore, the best critical region of size α for testing $H_0 : \mu = 10$ against each simple alternative $H_1 : \mu = \mu_\alpha$ where $\mu_\alpha > 10$, is given by:

$$C = \{(x_1, x_2, \cdots, x_n) : \bar{x} > k^*\}$$

where k^* is selected such that the probability of committing a Type I error is α, that is:

$$\alpha = P(\bar{X} \geq k^*) \text{ when } \mu = 10$$

Uniformly most powerful (UMP) α- size test is a hypothesis test which rejects $H_0 : \theta = \theta_0$ in favour of $H_1 : \theta < \theta_0$. It has the greatest power $1 - \beta$ among all possible tests of a given size α. **Definition 1**: (Monotone likelihood ratio (MLR)) The family $L(x, \theta)$, $\theta \in R$ has a monotone likelihood ratio in the statistic $T(X)$ if for any fixed θ' and θ" such that $\theta' < \theta$", it holds that $\frac{L(x,\theta")}{L(x,\theta')}$ is a non- decreasing function of $T(x) = T(x_1, x_2, \cdots, x_n)$.

Note typical examples are form the one- parameter exponential family: if $f(x, \theta) = a(\theta)b(x) \exp(c(\theta)d(x))$ and $c(\theta)$ is strictly monotone then

$$\frac{L(x,\theta")}{L(x,\theta')} = \frac{a^n(\theta")}{a^n(\theta')} \cdot \exp\left\{ [c(\theta") - c(\theta')] \sum_{i=1}^{n} d(x_i) \right\}$$

and this family has a MLR in $T(X) = \sum_{i=1}^{n} d(X_i)$.

Theorem 1: (Theorem of Blackwell & Girshick) Suppose $X \sim L(x, \theta)$ and the family is with MLR is $T(X)$. Then for testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, the α- test $\varphi*$ with the structure:

$$\varphi^* = \begin{cases} 1 & \text{if } T(x) > k, \\ \gamma & \text{if } T(x) = k, \\ 0 & \text{if } T(x) < k \end{cases}$$

The test rejects $H_0 : \theta \geq \theta_0$ in favour of $H_1 : \theta < \theta_0$ when $T(x) < k(\alpha = P_{\theta_0}(T < k) + \gamma P_{\theta_0}(T = k))$ is the UMP α- test.

Definition 1: A test φ of $H_0 : \theta \in \Theta_0$ $(\Theta_0 \subset \Theta)$ versus $H_1 : \theta \in \Theta \backslash \Theta_0$ is an unbiased size α- test if $E_\theta \varphi \leq \alpha$ all $\theta \in \Theta_0$ and $E_\theta \varphi \geq \alpha$ for all $\theta \in \Theta \backslash \Theta_0$.

The definition of unbiasedness ensures that there exist no alternatives for which acceptance of the hypothesis is more probale than in cases when the null hypothesis is true.

Theorem 2 Suppose $X \sim L(x, \theta)$ with $L(x, \theta) = (a(\theta))^n \prod_{i=1}^{n} b(x_i) \exp[c(\theta) \cdot \sum_{i=1}^{n} d(x_i)]$ and $T(X) = \sum_{i=1}^{n} d(x_i)$. Then, for testing $H_0 : \theta_1 \leq \theta \leq \theta_2$ versus $H_1 : \theta < \theta_1$ or $\theta > \theta_2$. the test $\varphi*$ is:

$$\varphi^*(x) = \begin{cases} 1 & \text{if } T(x) \notin [c_1, c_2], \\ \gamma_i & \text{if } T(x) = c_i, i = 1, 2, \text{ where } c_1, \gamma_1, c_2, \gamma_2 \\ 0 & \text{if } c_1 < T(x) < c_2 \end{cases}$$

the conditions $E_{\theta_1} \varphi* = E_{\theta_2} \varphi* = \alpha$. Moreover, the power function has a minimum somewhere within (θ_1, θ_2) and is monotone outside (θ_1, θ_2).

Example: Assume a "sample of one observation (n=1) from an exponential family with density $f(x, \theta) = \frac{1}{\theta} \exp\left(-\frac{1}{\theta}\right)$, $x > 0$ is available. The parameter $\theta > 0$ is to be tested. One would like to test $H_0 : \theta = 1$ versus $H_1 : \theta \neq 1$.

The UMPU α- test $\varphi*$ has the structure:

$$\varphi^*(x) = \begin{cases} 1 & \text{if } T < C_1 \text{ or } T(x) > C_2, \\ 0 & \text{if } C_1 \leq T(x) \leq C_2 \end{cases}$$

where T(x)=x in this case(one-parameter exponential family, $d(x) = x$, $n = 1$). We only need to find C_1 and C_2 in order to uniquely specify the above test. Note that $E_\theta \varphi * P_\theta(x \notin (C_1, C_2)) = 1 - \exp(-C_1/\theta) + \exp(-C_2/\theta)$ (since the cdf is $F(x, \theta) = 1 - \exp(-x/\theta), x > 0)$. The two conditions on $E_\theta \varphi*$ are:

- $E_\theta \varphi * |_{\theta=1} = \alpha = 1 - \exp(-C_1) + \exp(-C_2)$

-

Theorem 3 Consider the same family like in the previous Theorem 2. Then, for testing $H_0 : \theta = \theta_0$ versus $H_2 : \theta \neq \theta_0$ an UMPU α-test exists with the structure:

$$\varphi^{*}(x) = \begin{cases} 1 & \text{if } T(x) < c_1 \text{ or } T(x) > c_2 \ , \\ \gamma_i & \text{if } T(x) = c_i, i = 1, 2, \text{ The constants } c_i, \gamma_i, \text{ satisfy: Power}(\theta) \\ 0 & \text{if } c_1 < T(x) < c_2 \end{cases}$$

the conditions $E_{\theta_0}\varphi*$ and $\frac{\partial}{\partial\theta}$ Power $(\theta_0) = 0 = \frac{\partial}{\partial\theta} E_\theta \varphi * |_{\theta=\theta_0}$.