STAT-510 **2017 Fall**

Lecture 3: Suffciency and Completeness

Instructor: Xiaohui Chen Scribe: Yutong Dai Last Modified: 2018-03-13

Motivation: We want to summarize the information contained in data by determining a few key features of the sample values.

Notation: Denote $X = (X_1, ..., X_n)$, where $X_i \overset{iid}{\sim} f(x;\theta)$ and x denotes the realization of random samples. x also represents the variable of a function $f(.)$ and we will use X and x interchangeably.

Data reduction in terms of a particular statistic can be thought of a partition of the sample space $\mathcal{X}$. Let $\mathcal{T} = \{t; t \in T(\mathcal{X})\}$. Then T induces a partition of $\mathcal{X}$, A_t, which is defined as $A_T = \{x \in \mathcal{X}; T(x) = t, t \in \mathcal{T}\}$. The statistic $T(x)$ summarizes the data in that, rather than reporting the entire sample x, it reports only that $T(x) = t$ or, equivalently, $x \in A_T$. For example, if $T(x) = x_1 + ... + x_n$, then $T(x)$ does not report the actual sample values but only the sum. There may be many different sample points that have the same sum.

We are interested in methods of data reduction that do not discard important information about the unknown parameter θ and methods that successfully discard information that is irrelevant as far as gaining knowledge about θ is concerned.

3.1 The Suffciency Principle

A sufficient statistic for a parameter θ is a statistic that, in a certain sense, captures all the information about θ contained in the sample. Any additional information in the sample, besides the value of the sufficient statistic, does not contain any more information about θ.

Definition 3.1 (SUFFICIENCY PRINCIPLE) If $T(X)$ is a sufficient statistic(SS) for θ, then any inference about θ should depend on the sample X only through the value $T(X)$. That is, if x and y are two sample points such that $T(x) = T(y)$, then the inference about θ should be the same whether $X = x$ or $Y = y$ is observed.

In this subsection we investigate some aspects of sufficient statistics and the Sufficiency Principle.

3.1.1 Sufficient Statistics

Definition 3.2 (SUFFICIENT STATISTICS) A statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.

To understand definition 3.2, let t be a possible value of $T(X)$. Of course we are interested in the conditional probability $P_\theta(X = x|T(X) = t)$. If the sample x is a point such that $t \neq T(x)$, $P_\theta(X = x|T(X) = t) = 0$. Thus, we restrict $t = T(x)$, for some $x \in \mathcal{X}$. Then, by definition

$$P_\theta(X = x|T(X) = T(x))$$

does not depend on θ, hence $P_\theta(X = x|T(X) = T(x)) \stackrel{\Delta}{=} P(X = x|T(X) = T(x))$.

Remark

1. When $T(X)$ has a continuous distribution, then $P_\theta(T(X) = t) = 0$ for all values of t. A more sophisticated notion of conditional probability is needed. To simplify, we only focus on discrete case and point out analogous results that are true in the continuous case.
2. A satistic $T(X)$ is not sufficient if $P_\theta(X = x|T(X) = T(x))$ depends on θ, for some x.

Theorem 3.3 Suppose $p(x;\theta)$ is the joint pmf of the sample X and $q(t;\theta)$ is the pmf of $T(X)$. Then $T(X)$ is a sufficient statistic for $\theta \iff \frac{p(x;\theta)}{q(T(x);\theta)}$ is constant as a function of θ.

Proof: "⇒":

$$P_\theta(X = x|T(X) = T(x)) = \frac{P_\theta(X = x, T(X) = T(x))}{P_\theta(T(X) = T(x))} \tag{3.1}$$

$$= \frac{P_\theta(X = x)}{P_\theta(T(X) = T(x))} \tag{3.2}$$

$$= \frac{p(x;\theta)}{q(T(x);\theta)} \tag{3.3}$$

where the second equality due to the fact that $\{X; X = x\} \subset \{X; T(X) = T(x)\}$. Since $T(X)$ is a sufficient statistic for θ, $\frac{p(x;\theta)}{q(T(x);\theta)}$ is constant as a function of θ.

"⇐" Just inverse the proof of the sufficiency. ■

Remark

1. It is still appropriate to use the above criterion to determine if $T(X)$ is a sufficient statistic for θ when X and $T(X)$ have continuous distributions.
2. Sufficient statistics are not unique. Any one-to-one mappings of a sufficient statistic is also a sufficient statistic.

Example 1(Discrete case): Let $X_i \stackrel{iid}{\sim}$ Bernoulli(θ). Then the $T(X) = \sum_{i=1}^n X_i$ is a sufficient statistic for θ. Note that $T(X)$ follows Binomial(n, θ). Then by Theorem 3.3, we know that

$$\frac{p(x;\theta)}{q(T(x);\theta)} = \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{(1-x_i)}}{C_n^t \theta^t (1-\theta)^{(n-t)}} \qquad (t = \sum_i x_i) \tag{3.4}$$

$$= \frac{1}{C_t^n} \tag{3.5}$$

Example 2(Continuous case): Let $X_i \stackrel{iid}{\sim} N(\mu, \sigma^2)$ and σ^2 is known. Then the $T(X) = \bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic for μ. Since $\bar{X}_n$ follows $N(\mu, \sigma^2/n)$, we have

$$\frac{p(x;\mu)}{q(T(x);\mu)} = \frac{1}{\sqrt{n}}(2\pi\sigma^2)^{-(n-1)/2} \exp((n\bar{x}_n^2 - \sum_{i=1}^n x_i^2)/(2\sigma^2))$$

Then we show there exists a example that substantial data reduction is impossible.

Example 3(No reduction): Let $X_i \overset{iid}{\sim} f(x)$, where we can not specify more information about f. Then the ss can be $T(X) = (X_{\pi(1)}, X_{\pi(2)}, ..., X_{\pi(n)})$, where $\pi : \{1, 2, ..., n\}| \to \{1, 2, ..., n|\}$ is a permutation.

Theorem 3.4 (Factorization Theorem) Let $f(x;\theta)$ denote the joint pdf or pmf of a sample X. A statistic $T(X)$ is a sufficient for $\theta \iff$ there exists functions $g(t;\theta)$ and $h(x)$ such that for all sample points x and all parameter points θ,

$$f(x;\theta) = h(x)g(T(x);\theta).$$

Theorem 3.5 (SS for Exponential family) Let $X_1, ..., X_n \overset{iid}{\sim} f(x;\boldsymbol{\theta})$ that belongs to an exponential family given by

$$f(x;\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})\exp(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)),$$

where $\boldsymbol{\theta} = (\theta_1, ..., \theta_d), d \leq k$. Then

$$T(X) = (\sum_{i=1}^{n} t_1(X_i), ..., \sum_{i=1}^{n} t_k(X_i))$$

is a sufficient statistic for $\boldsymbol{\theta}$.

3.1.2 Minimal Sufficient Statistics

We have seen that SS are not unique, we introduce minimal sufficient statistics(MSS) to answer the questions that whether one SS is better than the other.

Definition 3.6 A sufficient statistics $M(X)$ is called a minimal sufficient statistic if for any other sufficient statistic $T(X)$ there exists a functional $g(.)$ that $M(X) = g(T(X))$.

Remark We interpret the definition of MSS through the partitions on the sample space $\mathcal{X}$. Both $T(X)$ and $M(X)$ can induce partitions on $\mathcal{X}$, and we denote them as $\{B_t = \{x \in \mathcal{X}; T(x) = t\}, t \in T(\mathcal{X})\}$ and $\{A_m = \{x \in \mathcal{X}; M(x) = m\}, m \in M(\mathcal{X})\}$ receptively. $\forall t \in T(\mathcal{X}), x \in B_t, \exists m \in M(\mathcal{X})$ such that $g(T(x)) = m$, hence $B_t \subset A_m$. That is any set in the partition induced by T is the subset of some set in the partition induced by M. Intuitively, M corresponds to the coarsest possible partition for $\mathcal{X}$, hence achieving the greatest possible data reduction for a SS.

Theorem 3.7 Let $f(x;\theta)$ be the pmf or the pdf of a sample X. Suppose there exists a function $T(x)$ such that, for every two sample points x and y, the ratio $\frac{f(x;\theta)}{f(y;\theta)}$ 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.

Proof: Without the loss of generality, we assume that $f(x;\theta) > 0, \quad \forall x \in \mathcal{X}, \theta \in \Theta$.

Actually, T defines a partition on set $\mathcal{X}$. Let $\mathcal{T} = \{t; t \in T(\mathcal{X})\}$. Any fixed $t \in \mathcal{T}$ corresponds to a set $A_t = \{x \in \mathcal{X}; T(x) = t\}$. Define $g(t;\theta) = \sum_{x \in A_t} P_\theta(X = x) = \sum_{x \in A_t} f(x;\theta)$. Now, we use factorization theorem to show that $T(X)$ is a SS.

Note that

$$f(x;\theta) = g(T(x);\theta)\frac{f(x;\theta)}{g(T(x);\theta)} \tag{3.6}$$

$$= g(t;\theta)\frac{f(x;\theta)}{\sum_{y\in A_{T(x)}} f(y;\theta)} \tag{3.7}$$

$$= g(T(x);\theta)h(x) \tag{3.8}$$

Since $h(x)$ is a constant in θ, $T(X)$ is a SS. Suppose there exists another SS, $T'(X)$. By factorization theorem, we can find $h'(X)$, $g'(t;\theta)$ such that $f(x;\theta) = g'(T'(x);\theta)h'(x)$. Let x and y be any two sample points with $T'(x) = T'(y)$, then

$$\frac{f(x;\theta)}{f(y;\theta)} = \frac{g'(T'(x);\theta)h'(x)}{g'(T'(y);\theta)h'(y)} = \frac{h'(x)}{h'(y)}.$$

Since T' is a SS, so h' does not depend on θ, hence the ratio is a constant in θ. By assumption, it implies that $T(x) = T(y)$. So $T(X)$ is the MSS.

■

Example 1: Let $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ and both σ^2, μ are unknown. Let x and y denote two sample points and let $(\bar{x}_n, S_x^2)$ and $(\bar{y}_n, S_y^2)$ be the sample mean and variance respectively. Since

$$f(x;\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}[n(\bar{x}_n-\mu)^2 + \sum_i (x_i-\bar{x}_n)^2]).$$

Set $g(\boldsymbol{t};\mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}exp(-\frac{1}{2\sigma^2}[n(t_1-\mu)^2 + (n-1)t_2])$ and $h(x) = 1$. We have $f(x;\mu,\sigma^2) = g(T_1(x), T_2(x);\mu,\sigma^2)h(x)$, where $T_1(x) = \bar{x}_n$ and $T_2(x) = S_x^2 = \frac{1}{n-1}\sum_i (x_i-\bar{x}_n)^2]$. Therefore,

$$\frac{f(x;\mu,\sigma^2)}{f(y;\mu,\sigma^2)} = \exp(\frac{-n(\bar{x}_n-\bar{y}_n) + 2n\mu(\bar{x}-\bar{y}) - (n-1)(S_x^2-S_y^2)}{2\sigma^2}).$$

This ratio will become constant in μ and $\sigma^2 \iff \bar{x} = \bar{y}$ and $S_x^2 = S_y^2$.

Example 2: Suppose $X_i \overset{iid}{\sim} U[\theta, \theta+1], \theta \in R$.Then the joint distribution of a sample point x is

$$f(x;\theta) = \begin{cases} 1, & \theta \le x_i \le \theta+1, \forall i \\ 0, & otherwise \end{cases}$$

$\frac{f(x;\theta)}{f(y;\theta)} = \frac{\prod_{i=1}^n I(x_i \le \theta+1)I(x_i \ge \theta)}{\prod_{i=1}^n I(y_i \le \theta+1)I(y_i \ge \theta)}$ is constant in θ , $\forall x, y \iff X_{(1)} = Y_{(1)}$ and $X_{(n)} = Y_{(n)}$.

And thus $T(X) = (X_{(1)}, X_{(n)})$.

3.2 Ancillary Statistics

In the preceding sections, we consider sufficient statistics, which contain all the information about θ that is available. Now, we consider another statistic - ancillary statistic.

Definition 3.8 A statistic $T(x)$ whose distribution dosen't depend on θ is called an ancillary statistic.

From the Definition 3.8, we know that an ancillary statistic contains no information about θ.

Example 1: Suppose $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$ and (μ, σ^2) are both unknown. The statistic $T(X) = \frac{X_1 - X_n}{\sqrt{S^2}}$ is an ancillary statistic, where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. This is because

$$T(X) = \frac{Z_1 - Z_2}{\sqrt{\sum_{i=1}^{n}(Z_i - \bar{Z}_n)^2}}$$

doesn't depend on θ, where $Z_i = \frac{X_i - \mu}{\sigma}$. The location information is cancelled through taking difference and the scale information is eliminated by taking ratio.

Example 2: (Location-Scale Family) Suppose $X_i \overset{iid}{\sim} \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$ and (μ, σ^2) are both unknown. The statistic $T(X) = (\frac{X_1 - X_n}{\sqrt{S^2}}, ..., \frac{X_{n-1} - X_n}{\sqrt{S^2}})$ is an ancillary statistic.

However, following two examples will show that, even an ancillary statistic contains no information about the θ, once we combine it with another statistic, which has some information about θ, then we can have all information abouth θ. Most suprisingly, we can use two ancillary statistic to derive a sufficient statistic.

Example 3: (Uniform ancillary) $X_i \overset{iid}{\sim} U[\theta, \theta + 1], \theta \in R$. Let $X_{(1)}, ..., X_{(n)}$ be the ordered statistic. Then $R_n = X_{(n)} - X_{(1)}$ is the ancillary statistic.

Example 4: $X_1, X_2 \overset{iid}{\sim} N(\theta, 1), \theta \in R$. $T_1 = X_1 - X_2$ is ancillary statistic. and $T_2 = X_2$ isn't a sufficient statistic. But (T_1, T_2) is a sufficient statistic. Consider a one-to-one function $f : (X_1 - X_2, X_2) \quad | \to (X_1, X_2)$. Since $X_1 - X_2$ is a sufficient statistic, (T_1, T_2) is also a sufficient statistic.

Example 5: Consider

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}).$$

Them $T_1 = T_1(X, Y) = X, T_2 = T_2(X, Y) = Y$ are two ancillary statistic since their distributions don't depend on ρ. But (X, Y) is the a sufficient statistic for ρ.

3.3 Complete Statistic

On the one hand, a minimal sufficient statistic contains all information about the parameter θ, hencing achieving the maximum data reduction. On the other hand, an ancillary statistci contains no information about the parameter θ. So, intuitively, any minimal sufficient statistic should be functionally independent of any ancillary staistic. However, this claim doesn't necessary hold without any other contraint.

For example, consider $X_i \overset{iid}{\sim} U[\theta, \theta + 1], \theta \in R$.Then a minimal sufficient statistic is $(X_{(n)}, X_{(1)})$. Clearly, the $(X_{(1)} - X_{(n)}, \frac{X_{(1)} + X_{(n)}}{2})$ is another minimal sufficient statistic. However, in this case $X_{(n)} - X_{(1)}$ is ancillary and is an part of a minimal sufficieny statistic. Certainly, the anciallry and the minimal sufficient statistic are not functionally independent.

This example motivates us that we need put addiitonal requirements to ensure the independence.

Definition 3.9 (Induced family of distributions) Suppose $X \sim f(x; \theta), \theta \in \Theta$ and $T(X) \sim g(t; \theta)$. The collection of pmf/pdf denoted by $\{g(t; \theta); \theta \in \Theta\}$ is called the family of distribution induced by the statistic $T(X)$.

Example: $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$,$\sigma^2$ is given and $T(X) = \bar{X}_n = \frac{1}{n}\sum_i X_i \sim N(\mu, \sigma^2/n)$. Then the induced family of distribution induced by $T(X)$ is

$$\{g(t;\mu) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp(\frac{-n(x-\mu)^2}{\sigma^2}); \mu \in R\}.$$

Definition 3.10 (Completeness) The family of distribution $\{g(t;\theta); \theta \in \Theta\}$ induced by the statistic $T(X)$ is complete if the following condition holds:

1. $\forall$ real-valued function $h(.)$, we have

$$E_\theta(|h(T)|) = \int |h(T(x))| f_X(x;\theta)dx < \infty \qquad \forall\theta \in \Theta.$$

2. If $\forall\theta \in \Theta, E_\theta(h(T)) = 0$ implies $\forall\theta \in \Theta, P_\theta[h(T) = 0] = 1$.

Moreover, a statistic $T(X)$ is called complete iff the induced family of distribution induced by $T(X)$ is complete.

Example 1: (Uniform complete) Suppose $X_i \overset{iid}{\sim} U[0,\theta], \theta > 0$. $T(X) = \max_i X_i$. Note that

$$F_T(t) = P(T(X) \le t) = \begin{cases} 0 & ,t < 0 \\ (t/\theta)^n & ,0 \le t \le \theta \\ 1 & ,t > \theta, \end{cases}$$

we have

$$f_T(t) = \begin{cases} 0 & ,t < 0 \\ n\theta^{-n}t^{n-1} & ,0 \le t \le \theta \\ 0 & ,t > \theta. \end{cases}$$

Suppose $\forall\theta \in \Theta$,

$$0 = E_\theta[h(T)] = \int_0^\theta h(t) n\theta^{-n} t^{n-1} dt,$$

which implies that $\int_0^\theta h(t)t^{n-1}dt = 0, \forall\theta \in \Theta$. Set derivative with respect to θ on both side, we have $h(\theta)\theta^{n-1} = 0, \forall\theta \in \Theta$. This means that $h(\theta) = 0, \forall\theta \in \Theta$, and further more we have $h(t) = 0, t \in (0,\theta], \forall\theta \in \Theta$ and $\forall\theta \in \Theta, P_\theta[h(T) = 0] = 1$.

Example 2: (Possion complete) Let $X_1, ..., X_n$ denote a random sample from a Poisson distribution with parameter $\lambda > 0$. Then $T(X) = \sum X_i \sim p(n\lambda)$. For any function $h(.)$ and any λ,

$$0 = E_\lambda(h(T(X)) = \sum_{t=0}^{+\infty} e^{-n\lambda} \frac{(n\lambda)^t}{t!} h(t).$$

This implies $h(t) = 0,\ t = 0, 1, 2, ...$ and therefore $P_\lambda(h(T) = 0) = 1$.

- In general a complete statistic is not necessary a sufficient statistic.

Example 3: (continued with **Example 2**) The joint distribution of $(X_1, ..., X_n)$ is

$$f(x_1, x_2, ..., x_n; \lambda) = \left(e^{-n\lambda}\lambda^{\Sigma x_i}\right) \times \left(\frac{1}{x_1!x_2!...x_n!}\right) = g(T(x); \theta)h(x_1, x_2, ..., x_n).$$

By the factorization theorem we know $T(X)$ is a sufficient statistic. Now, consider $T'(X) = X_1 + X_2$. By the analysis in **Example 2**, we know that $T'(X)$ is a complete statistic but not a sufficient statistic.

- However, a sufficient statistic is also not necessary a complete statistic.

Example 4: Let X_1, X_2 denote a random sample from a Poisson distribution with parameter $\lambda > 0$. Then $T(X) = (X_1, X_2)$ is sufficient but not complete. Consider a function $h(x_1, x_2) = x_1 - x_2$, then

$$E_\lambda(T(X)) = \sum_{x_1=0}^{+\infty}\sum_{x_2=0}^{+\infty}(x_1 - x_2)e^{-2\lambda}\frac{\lambda^{(x_1+x_2)}}{(x_1 + x_2)!} = 0,$$

which implies $P_\lambda(h(T) = 0) < 1$. (Not that rigorous.)

Next, we introduce a theorem on the complete statistic for exponential family.

Theorem 3.11 (exponential family) Let $X_1, ..., X_n \overset{iid}{\sim} f(x; \theta) = h(x)c(\theta)\exp(\sum_{j=1}^k w_j(\theta)t_j(x))$, where $\theta = (\theta_1, ..., \theta_k)^T$. Then the statistic $T(X) = (\sum_{i=1}^n t_1(X_i), ..., \sum_{i=1}^n t_k(X_i))$ is a complete sufficient statistic, if the parameter space $\{w_1(\theta), ...w_k(\theta); \theta \in \Theta\}$ contains a open subset.

Remark:

1. From the Theorem 3.11, we know that for curved exponential family, the sufficient statistic is typically not complete.
2. Since the transformation $T(.)$ is usually continues, (not that rigorous) the condition that the parameter space $\{w_1(\theta), ...w_k(\theta); \theta \in \Theta\}$ contains a open subset is equivalent to the one that Θ contains a open subset.

Example : (Simple Linear Regression) Consider the following linear regression setting:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, σ^2 is known and $\theta = (\beta_0, \beta_1)$. The joint pdf of $(y_1, ..., y_n)$ is

$$f(y_1, ..., y_n; \theta) = (2\pi\sigma^2)^{-n/2}\exp(\frac{-\sum_{i=1}^n(\beta_0 + \beta_1 x_i)^2}{2\sigma^2})\exp(\frac{-\sum_{i=1}^n y_i^2}{2\sigma^2})\exp(\frac{2\sum_{i=1}^n(\beta_0 + \beta_1 x_i)y_i}{2\sigma^2}) \quad (3.9)$$

$$= (2\pi\sigma^2)^{-n/2}\exp(\frac{-\sum_{i=1}^n(\beta_0 + \beta_1 x_i)^2}{2\sigma^2})\exp(\frac{-\sum_{i=1}^n y_i^2}{2\sigma^2})\exp(\frac{\beta_0}{\sigma^2}\sum_{i=1}^n y_i)\exp(\frac{\beta_1}{\sigma^2}\sum_{i=1}^n x_i y_i) \quad (3.10)$$

$$= h(x)c(\theta)\exp(w_1(\theta)t_1(y))\exp(w_2(\theta)t_2(y)). \quad (3.11)$$

Since $\{(\frac{\beta_0}{\sigma^2}, \frac{\beta_1}{\sigma^2}); \beta_0, \beta_1 \in R\}$ contains a open subset, we know $T(Y) = (\sum_{i=1} Y_i, \sum_{i=1} x_i Y_i)$ is a complete sufficient statistic for (β_0, β_1).

Remark: Recall that the LS estimator for (β_0, β_1) is

$$\begin{pmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{pmatrix} = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2\end{pmatrix}^{-1}\begin{pmatrix}\sum_i Y_i \\ \sum_i x_i Y_i\end{pmatrix} = AT(Y).$$

Since A is inevitable, whchich means A is a one-to-one transformation, we conclude that $(\hat{\beta}_0, \hat{\beta}_1)$ is a complete sufficient statistic estimator. This implies that LS estimator is really good in the sense of completeness and sufficiency.

Now, with the help of completeness, Basu gives the following theorem to show that when the sufficient statistic is independent of anciallary statistic.

Theorem 3.12 (Basu's Theorem) If $T(X)$ is a complete and sufficient statistic, then $T(X)$ si independent of every anciallary statistic.

Proof: We only give the proof of discrete case.

Let $S(X)$ be any anciallary statistic. We want to prove that $\forall s, t$,

$$P_\theta(S(X) = s|T(X) = t) = P_\theta(S(X) = s) = P(S(X) = s),$$

where the last equality holds as $S(X)$ is ancillary statistic. Since the $T(X)$ is the sufficient, we know that

$$P_\theta(S(X) = s|T(X) = t) = P(S(X) = s|T(X) = t).$$

So the only thing left is to prove that $\forall s, t$

$$P(S(X) = s|T(X) = t) - P(S(X) = s) = 0.$$

Define a funtion

$$g_s(t) = P(S(X) = s|T(X) = t) - P(S(X) = s),$$

which doesn't depend on θ. Take expectation with resepct to $T(X)$, we have

$$\begin{aligned} E_\theta[g_s(T)] &= \sum_t g_s(t)P_\theta(T(X) = t) && (3.12)\\ &= \sum_t P(S(X) = s|T(X) = t)P_\theta(T(X) = t) - \sum_t P(S(X) = s)P_\theta(T(X) = t) && (3.13)\\ &= P(S(X) = s) - P(S(X) = s) && (3.14)\\ &= 0 && (3.15)\end{aligned}$$

By the completeness of $T(X)$, we know that $P_\theta(g_s(T) = 0) = 1, \forall\theta$, which completes the proof. ■

Exmaple : $X_i \overset{iid}{\sim} N(0, \sigma^2), \sigma^2 > 0$ but is unknown. Let $S^2 = \frac{(\sum_{i=1}^n x_i)^2}{\sum_{i=1}^n x_i^2}$. What is the value of $E(S^2)$?

Consider $T(X) = \sum_{i=1}^n x_i^2$, then it's easy to see that $T(X)$ is complete and suffcient. Since S^2 belongs to the scale family, hence it is ancillary. By Theorem 3.12, we know that $T(X) \perp S(X)$. Then $E(T(X))E(S^2) = E[(\sum_{i=1}^n x_i)^2]$, which implies that $E(S^2) = 1$.

3.4 Appendix

3.4.1 pdfs of order statistic

Suppose the cdf of X is $F_X(x)$ and $X_{(1)}, ..., X_{(n)}$ are the ordered statistic, then

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}[F_X(x)]^{k-1}[1-F_X(x)]^{n-k}f_X(x) \tag{3.16}$$

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}\times$$
$$[F_X(x)]^{j-1}[F_X(y)-F_X(x)]^{k-1-j}[1-F_X(y)]^{n-k}f_X(x)f_X(y) \text{ where } x \le y \tag{3.17}$$

$$f_{X_{(1)},...,X_{(n)}}(x_1,\ldots,x_n) = n!f_X(x_1)\cdots f_X(x_n) \text{ where } x_1 \le x_2 \le \cdots \le x_n. \tag{3.18}$$