6 Principles of Data Reduction

6.1 Statistical Inference & Data Reduction

Suppose data $X = (X_1, \ldots, X_n)$ are from a probability distribution $P$, which is either completely or partially unknown, e.g. Poisson distribution. In most cases, we assume that the distribution is controlled by either one or multiple unknown parameters $\theta$, the object of interest.

Data is used to make inferences on model parameters, or the form of the model itself.

Three main inference problems:

- Point estimation of $\theta$;
• Hypothesis testing (relative to $\theta$ or the form of the model);
• Interval estimation of $\theta$.

Inference is typically based on statistics. A statistic, $T(\mathbf{X})$, is any function of the observable random variables in a sample.

• e.g. sample mean $\bar{X}$, sample variance $S^2$ are statistics.
• $\bar{X} - \theta$ is not a statistic because it involves the unknown parameter $\theta$.
• A statistic could be multi-dimensional, e.g. $T(\mathbf{X}) = (\bar{X}, S^2)$

First step: **data reduction**

• Any statistic $T(\mathbf{X})$ is a data summary, and provides a form of data reduction.
• In many situations, much of the information given by the sample \{\(X_1, \ldots, X_n\)} is unnecessary for inference concerning $\theta$, and some
summary (data reduction) of the sample captures all information in the data about $\theta$.

- In such situations, storing only relevant statistics is more practical.

Example:

- Suppose we observe a random sample of Bernoulli random variables $X_1, \ldots, X_n$, where the success rate $\theta$ is unknown.
- Do we need know the complete sample to make inference about $\theta$?
- Knowing the total number of successes $X_1 + \cdots + X_n$ is enough!

### 6.2 Sufficient Statistics

**Intuitive definition:**

A statistic $T(X)$ is said to be **sufficient** for a parameter $\theta$ if it contains all the information about $\theta$ that is available in the sample.
X = (X_1, X_2, \ldots, X_n). In other words, any inference about \( \theta \) should depend on X only through \( T(X) \).

**Definition.** (Formal definition) Let \( X = (X_1, \ldots, X_n) \) be a random sample from a distribution with parameter \( \theta \). A statistic \( T(X) \) is said to be a **sufficient statistic for** \( \theta \) if the conditional distribution of \( X_1, \ldots, X_n \) given \( T(X) \) does not depend on \( \theta \). I.e.

\[
T(X) \text{ is sufficient for } \theta \iff P_\theta\{X = x | T(X) = T(x)\} \text{ does not depend on } \theta.
\]

**Remarks.** A sufficient statistic for a parameter \( \theta \) has the property that it carries all the information in the sample pertaining to \( \theta \). Once \( T(X) \) is known, no other functions of \( X \) will give additional information about \( \theta \). This means that you can summarize the information in the sample with this statistic without losing any information about \( \theta \).
Theorem 1. (Th6.2.2) If $f(x|\theta)$ is the pdf or pmf of $X$ and $g(t|\theta)$ is the pdf or pmf of $T(X)$. Then $T(X)$ is sufficient for $\theta$ if and only if, for every $x$ in the sample space, the ratio
\[
\frac{f(x|\theta)}{g\{T(x)|\theta\}} \text{ is constant (does not depend on } \theta)\text{.}
\]

Proof. (discrete case) \qed
Example 6.2.1. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli($\theta$). Show that $T(X) = X_1 + \cdots + X_n$ is sufficient for $\theta$.

Proof.
Example 6.2.2. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known. Show that $T(X) = \bar{X} = n^{-1}(X_1 + \cdots + X_n)$ is sufficient for $\mu$.

Proof.
How do we find sufficient estimators? The definition (or Theorem 6.2.2) is not very workable.

- Does not tell which statistic is likely to be sufficient. We need first guess a statistic $T(X)$ that is sufficient.
- Requires us to derive the conditional distribution or the distribution of $T(X)$, which may not be easy.

The next **factorization theorem** is useful because it allows us to determine a sufficient statistic by writing down the joint pdf or pmf and factoring it.
Factorization Theorem

**Theorem 2.** (Neyman-Fisher Factorization Theorem) A statistic $T(X)$ is sufficient statistic for $\theta$ if, and only if there exist functions $g(t|\theta)$ and $h(x)$ such that for all sample points $x$ and parameter values $\theta$, the joint pdf/pmf (likelihood function) can be factored as:

$$f(x|\theta) = g\{T(x)|\theta\} \times h(x).$$

Here

- $g(t|\theta)$ is a function of $\theta$ and it depends on $x$ only through $t$;
- the function $h(x)$ does not depend on $\theta$ but may depend in any way on $x$.

Proof. (discrete case)
Example 6.2.3. Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli($\theta$). Use the Factorization Theorem to show that $\sum_{i=1}^{n} X_i$ is sufficient for $\theta$.

Proof. \qed
Example 6.2.4. Let $X_1, \ldots, X_n$ be i.i.d. Exponential($\theta$). Use the Factorization Theorem to show that $\bar{X}$ and $\sum_{i=1}^{n} X_i$ are both sufficient for $\theta$.

Proof.
Remarks.

• There are many possible sufficient statistics for any parameter.

• Any statistic that is a one-to-one function of a sufficient statistic is itself a sufficient statistic for $\theta$. For example, in Example 6.2.4, $\bar{X}^2$ is also sufficient for $\theta$.

Example 6.2.5. $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known. Find a sufficient statistic for $\mu$. 
Example 6.2.6. $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown. Find a sufficient statistic for $(\mu, \sigma^2)$. 
Be careful when the support of the distribution depends on $\theta$.

**Support**: the set of feasible values for which the pdf/pmf is nonzero.

**Example 6.2.7.** $X_1, \ldots, X_n$ i.i.d. $U(0, \theta)$. Find a sufficient statistic for $\theta$. 
**Note:** Usually, sufficient statistic has the same dimension as the parameter vector but not always!

**Example 6.2.8.** \(X_1, \ldots, X_n\) i.i.d. \(U(\theta, \theta + 1)\). Find a sufficient statistic for \(\theta\).
All of the above examples other than the uniform distribution ones are special cases of an exponential family.

**Theorem 3.** Let $X_1, \ldots, X_n$ be i.i.d. observations from a pdf/pmf $f(x|\theta)$ that belongs to an exponential family given by

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

where $\theta = (\theta_1, \ldots, \theta_d), d \leq k$. Then

$$T(X) = \left(\sum_{j=1}^{n} t_1(X_j), \ldots, \sum_{j=1}^{n} t_k(X_j)\right)$$

is sufficient for $\theta$.

**Proof.** The result follows by applying the Factorization theorem.
Example 6.2.9. \( X_1, \ldots, X_n \) are i.i.d. \( N(\theta, a\theta) \), where \( a > 0 \) is known, i.e. the variance is proportional to the mean. Find a sufficient statistic for \( \theta \).
Remarks.

- The random sample \((X_1, \ldots, X_n)\) itself is sufficient for \(\theta\).
- The set of order statistics \((X_{(1)}, \ldots, X_{(n)})\) is sufficient for \(\theta\).
- But the above two provide no data reduction.
- With many choices of sufficient statistics for a parameter, which one is better? e.g. \(N(\mu, \sigma^2)\) with \(\sigma^2\) known, \(X, \bar{X}, (\bar{X}, S^2), \sum_i X_i\) are all sufficient for \(\mu\), which is better?
- “Best” sufficient statistic: the one that achieves the most data reduction while retaining all the information about \(\theta\).
- The factorization theorem provides the “best” sufficient statistic in the sense that it is a function of every other sufficient statistics. Such a sufficient statistic is the called the **minimal sufficient** statistic (next topic).
6.3 Minimal Sufficient Statistics

**Definition.** A sufficient statistic $T(X)$ is called a **minimal sufficient statistic** if for any other sufficient statistic $T^*(X)$, $T(X)$ is a function of $T^*(X)$.

An equivalent condition: $T$ is minimal sufficient if, given any other sufficient statistic $T^*$ and for any two sample points $x$ and $y$, we have

$$T^*(x) = T^*(y) \Rightarrow T(x) = T(y).$$

**Remarks.**

- The equivalent condition $T^*(x) = T^*(y) \Rightarrow T(x) = T(y)$ implies that $B_{t^*} = \{x : T^*(x) = t^*\}$ is a subset of $A_t = \{x : T(x) = t\}$ for some $t$, i.e., possibly more values of $x$ are mapped to $t$ by $T$ than to $t^*$ by $T^*$.

- That is, the minimal sufficient statistic $T(X)$ has the coarsest partition and thus gives the greatest data reduction.
Example 6.3.1. $X_1$ and $X_2$ are i.i.d. Bernoulli$(\theta)$.

The sufficient statistic $T^*(X) = (X_1, X_2)$ induces the partitions

$$B_{0,0} = \{(0, 0)\}, B_{0,1} = \{(0, 1)\}, B_{1,0} = \{(1, 0)\}, B_{1,1} = \{(1, 1)\}.$$ 

The minimal sufficient statistic $T(X) = X_1 + X_2$ induces the partitions

$$A_0 = \{(0, 0)\}, A_1 = \{(0, 1), (1, 0)\}, A_2 = \{(1, 1)\}.$$

- $T^*(x) = T^*(y) \Rightarrow x = y \Rightarrow T(x) = T(y)$ so $T$ is some function of $T^*$. However, $T(x) = T(y)$ does not imply $T^*(x) = T^*(y)$.

- For $t = 1$, $B_{0,1}$ and $B_{1,0}$ are subsets of $A_1$. So the minimal sufficient statistic $T(X) = X_1 + X_2$ has the coarsest partition and gives more data reduction.
Note: the partition given by another minimal sufficient statistic is also the **minimal sufficient partition** but just with different labels for the set.

In Example 6.3.1, the partition induced by $T(X) = \bar{X}$ is

$$A_0 = \{(0, 0)\}, A_{1/2} = \{(0, 1), (1, 0)\}, A_1 = \{(1, 1)\}.$$
**Example 6.3.2.** $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$ with $\sigma^2$ known. The joint pdf is

$$f(x|\mu) = (2\pi \sigma^2)^{-n/2} \exp \left( -\frac{\sum_i x_i^2}{2\sigma^2} \right) \exp \left( \frac{\mu}{\sigma^2} \sum_i x_i - \frac{n\mu^2}{2\sigma^2} \right).$$

Consider $T(X) = \sum_{i=1}^n X_i$ and $T^*(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$.

- Are both statistics sufficient?
- Which one achieves greater data reduction?
Finding minimal sufficient statistics

**Theorem 4.** (Lehmann-Scheffe) $T(X)$ is a minimal sufficient statistic for $\theta$ if for any two sample points $x$ and $y$,

$$f(x|\theta)/f(y|\theta) \text{ is constant } \iff T(x) = T(y).$$

*Proof.* For simplicity, assume that $f(x|\theta) > 0$ for all $x \in \mathcal{X}$ and $\theta$.

(i) First show that $f(x|\theta)/f(y|\theta) \text{ is constant } \iff T(x) = T(y)$ implies that $T(X)$ is sufficient.

- We introduce some notations. Let $A_t = \{x : T(x) = t\}$. Each set $A_t$ may contain multiple $x$’s. For each set $A_t$, choose and fix one element $x_t \in A_t$. Now choose any $x$ so this $x \in A_{T(x)}$. Then $x_{T(x)}$ is the fixed element that is in the same set $A_t$ as $x$, where $t = T(x)$.

For example, $X_1, X_2, X_3$ are i.i.d. $\text{Bernoulli}(\theta)$. The sufficient
statistic $T(\mathbf{X}) = \sum_{i=1}^{3} X_i$ induces partitions

$A_0 = \{(0, 0, 0)\}, A_1 = \{(1, 0, 0), (0, 1, 0), (0, 0, 1)\}

A_2 = \{(1, 1, 0), (1, 0, 1), (0, 1, 1)\}, A_3 = \{(1, 1, 1)\}.

For each $A_t$, we choose and fix any one element and call it $\mathbf{x}_t$, e.g.

$\mathbf{x}_0 = (0, 0, 0), \mathbf{x}_1 = (1, 0, 0), \mathbf{x}_2 = (1, 1, 0), \mathbf{x}_3 = (1, 1, 1)$.

For say $\mathbf{x} = (0, 1, 0) \in A_1$, $\mathbf{x}_{T(\mathbf{x})} = \mathbf{x}_1 = (1, 0, 0)$.

- With the above notations, for any $\mathbf{x} \in \mathcal{X}$,

  $T(\mathbf{x}) = T(\mathbf{x}_{T(\mathbf{x})}) \Leftrightarrow f(\mathbf{x}|\theta)/f(\mathbf{x}_{T(\mathbf{x})}|\theta)$ is constant.

- Then

  $f(\mathbf{x}|\theta) = \frac{f(\mathbf{x}|\theta)}{f(\mathbf{x}_{T(\mathbf{x})}|\theta)} f(\mathbf{x}_{T(\mathbf{x})}|\theta) = h(\mathbf{x}) g\{T(\mathbf{x})|\theta\}$,
where \( h(x) = f(x|\theta)/f(x_{T(x)}|\theta) \) (not dependent on \( \theta \)) and \( g(t) = f(x_t|\theta) \).

- Thus \( T(X) \) is sufficient by the Factorization Theorem.

(iii) Next show that \( T(X) \) is minimal.

- Let \( T^*(X) \) be any other sufficient statistic.

- By the Factorization Theorem, there exists functions \( g^* \) and \( h^* \) such that

\[
    f(x|\theta) = g^*\{T^*(x)|\theta\}h^*(x).
\]

- Let \( x \) and \( y \) be two sample points such that \( T^*(x) = T^*(y) \).

Then

\[
    \frac{f(x|\theta)}{f(y|\theta)} = \frac{g^*\{T^*(x)|\theta\}h^*(x)}{g^*\{T^*(y)|\theta\}h^*(y)} = \frac{h^*(x)}{h^*(y)} \text{ is constant}
\]

\[\Leftrightarrow \quad T(x) = T(y).\]

- Thus \( T \) is a function of \( T^* \) and \( T \) is minimal sufficient.
Remarks. Minimal sufficient statistic is not unique. Any one-to-one transformation of a minimal sufficient statistic is still minimal sufficient.
Example 6.3.3. \( X_1, \ldots, X_n \) i.i.d. \( \sim N(\mu, \sigma^2) \) with \( \sigma^2 \) known.
Example 6.3.4. $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ unknown.
Example 6.3.5. $X_1, \ldots, X_n$ i.i.d. $\sim U(\theta, \theta + 1)$. 
Example 6.3.6. $X_1, \ldots, X_n$ i.i.d. Cauchy($\theta$) with density

$$f(x|\theta) = 1/\{\pi(x - \theta)^2\}.$$
Minimal Sufficiency in Exponential Family

Theorem 5. If $X_1, \ldots, X_n$ are i.i.d. from an exponential family

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

so that no affine (linear plus constant) relationship exists between $w_1(\theta), \ldots, w_k(\theta)$, then statistic

$$T(X) = \left(\sum_{i=1}^{n} t_1(x_i), \ldots, \sum_{i=1}^{n} t_k(x_i)\right)$$

is minimal sufficient for $\theta = (\theta_1, \ldots, \theta_d)$. 
Example 6.3.7. $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ unknown.
Example 6.3.8. $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, \theta^2)$. 
Example 6.3.9. $X_1, \ldots, X_n$ i.i.d. $\sim \text{Beta}(\alpha, \beta)$. 
6.4 Ancillary Statistics

**Definition.** A statistic $S(X)$ is called ancillary if its distribution does not depend on $\theta$.

- The induced family is completely known.
- An ancillary statistic by itself contains no information about $\theta$.
- This is the opposite of sufficient statistic which contains all the information about $\theta$.
- Functions of ancillary is ancillary.

**Examples**

- $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, 1)$, $T = X_1 - \bar{X}$ is ancillary.
- $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$, $T = (X_1 - \bar{X})/S$ is ancillary.
General Location Family

- Suppose $X_1, \ldots, X_n$ i.i.d. $\sim f(x - \theta)$, a location family.

- If $T$ is a location-invariant statistic, i.e.
  
  \[ T(x_1 + b, \ldots, x_n + b) = T(x_1, \ldots, x_n), \]

  then $T$ is ancillary. e.g. 
  
  \[ (X_1 - \bar{X}, \ldots, X_n - \bar{X}); \] sample sd $S$ (and other estimates of scale, e.g. range).

- Let $Z_1, \ldots, Z_n$ i.i.d $\sim F(x)$. That is, $X_i = Z_i + \theta$. e.g. 
  
  \[ X_i \sim N(\mu, 1), Z_i \sim N(0, 1). \]

- Then the cdf of $R = X_{(n)} - X_{(1)}$ is 

  \[
  F_R(r) = P(R \leq r) = P(\max X_i - \min X_i \leq r) \\
  = P(\max Z_i - \min Z_i \leq r),
  \]

  which does not depend on $\theta$. That is, the range $R = X_{(n)} - X_{(1)}$ is ancillary for the location parameter in any location family.
Scale Family

Let $X_1, \ldots, X_n$ be i.i.d. from a scale family with $f(x/\sigma)$, $\sigma > 0$. Then any statistic that depends on the ratios is ancillary for the scale parameter $\sigma$. For instance,

- $T(X) = X_1/X_n$
- $T(X) = \frac{X_1 + \cdots + X_n}{X_n} = \frac{X_1}{X_n} + \cdots + \frac{X_{n-1}}{X_n} + 1$.

Justification:

- Let $Z_1, \ldots, Z_n$ i.i.d. $\sim F(x)$, i.e. $X_i = \sigma Z_i$.
- Then $T(X) = X_1/X_n = Z_1/Z_n$, whose distribution does not depend on $\sigma$. 
Location-scale Family

Let $X_1, \ldots, X_n$ i.i.d. $\sim \sigma^{-1} f \{(x - \mu)/\sigma\}$, a location-scale family.

- If $T$ is location-scale invariant statistic, i.e.

$$T(ax_1 + b, \ldots, ax_n + b) = T(x_1, \ldots, x_n),$$

then $T$ is ancillary. E.g. $(X_1 - \bar{X}/s, \ldots, X_n - \bar{X}/s)$ is ancillary.

- If $T_1$ and $T_2$ are such that

$$T_1(ax_1 + b, \ldots, ax_n + b) = aT_1(x_1, \ldots, x_n),$$

$$T_2(ax_1 + b, \ldots, ax_n + b) = aT_2(x_1, \ldots, x_n),$$

then $T_1/T2$ is ancillary.
Remarks. An ancillary statistic by itself contains no information about \( \theta \). However, sometimes, an ancillary statistic combined with other statistics may contain information for inference about \( \theta \).

**Example 6.4.1.** \( X_1, \ldots, X_n \) i.i.d. \( U(\theta, \theta + 1) \). Let \( R = X_{(n)} - X_{(1)} \) and \( S = X_{(1)} \).

1. Show that the joint pdf of \((R, S)\) is \( f(r, s | \theta) = n(n - 1)r^{n-2} \), \( \theta < s < r + s < \theta \), i.e. \( 0 < r < 1, \theta < s < \theta + 1 - r \).

2. Verify that \( R \sim \text{Beta}(n - 1, 2) \) and thus is ancillary to \( \theta \).
Remarks. In the previous example,

- since \((X_{(1)}, X_{(n)})\) is minimal sufficient, so is the one-to-one transformation \((R, X_{(1)})\);

- though \(R\) itself gives no information about \(\theta\), combined with \(X_{(1)}\) it is minimal sufficient.
6.5 Completeness

- A minimal sufficient statistic achieves the maximum amount of data reduction while retaining all the information the sample has concerning $\theta$.

- On the other hand, the distribution of an ancillary statistic does not depend on $\theta$ at all.

- It may seem that minimal sufficient and ancillary statistics must be functionally independent, but this isn’t necessarily the case. For example, the $U(\theta, \theta + 1)$ example, $R = X_{(n)} - X_{(1)}$ is ancillary, while $(R, X_{(1)})$ is minimal sufficient, yet the two are dependent.

- For many important situations, however, our intuition is correct. **Completeness** is the needed condition for the minimal sufficient and ancillary statistics to be independent.
• The property of completeness will also be used to construct “good” estimators.

**Definition.** A statistic \( T(X) \) is called **complete** (for the family \( \{ f(x|\theta), \theta \in \Theta \} \)), or equivalently the induced family \( f_T(t|\theta) \) is called **complete** if

\[
E_{\theta}[g(T)] = 0 \text{ for all } \theta
\]

implies that

\[
P_{\theta}\{g(T) = 0\} = 1 \text{ for all } \theta.
\]
Remarks.

- In other words, the completeness of $T$ means, no non-constant function of $T$ can have constant expectation (in $\theta$).
- Completeness depends not only on the statistic, but also on the entire family of distributions indexed by $\theta$, but not a particular distribution with a specific $\theta$.
- In fact, no nontrivial statistic is complete if the family has a specific $\theta$, i.e. is completely known.
- For example, $X \sim N(0, 1)$, consider the identity function $g(x) = x$, $E(X) = 0$ but $P(X = 0) = 0 \neq 1$.
- We shall see that if $X \sim N(\theta, 1)$, no functions of $X$ will have mean zero for all $\theta$ unless the function is zero with probability 1, i.e. $N(\theta, 1)$ is complete.
Example 6.5.1. (Binomial). Suppose $T \sim \text{Binomial}(n, p)$. Suppose $g(T)$ satisfies $E[g(T)] = 0$ for all $p \in (0, 1)$. Show that $g(T) = 0$ with probability 1 for all $p$. 
Example 6.5.2. $X_1, \ldots, X_n$ i.i.d. $\sim$ Poisson($\theta$), $\theta > 0$.

1. Is this model complete? No!

2. Now let $T = X_1 + \cdots + X_n$, a sufficient statistic. Is the model for $T$ complete? Yes.
Example 6.5.3. (Uniform). $X_1, \ldots, X_n$ i.i.d. $\sim U(0, \theta)$, $n > 1$. This model is not complete (no iid model is complete if $n > 1$). Consider $T = X_{(n)}$. Show that the model for $T$ is complete.
**Theorem 6.** Let $X_1, \ldots, X_n$ be i.i.d. from an exponential family with given by

$$f(x|\theta) = h(x)c(\theta) \exp \left[ \sum_{j=1}^{k} w_j(\theta)t_j(x) \right],$$

where $\theta = (\theta_1, \ldots, \theta_k)$. Then

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

is complete if $\{w_1(\theta), \ldots, w_k(\theta) : \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$ (i.e. $d = k$).
Example 6.5.4. $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$ with $\mu$ and $\sigma^2$ unknown. Show that $T(X) = (\sum X_i, \sum X_i^2)$ is complete.
Example 6.5.5. $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, \theta^2)$. Show that $T(X) = (\sum X_i, \sum X_i^2)$ is not complete.
In Example 6.5.5, since

\[ E(\bar{X}^2) = \frac{(n + 1)}{n\theta^2}, \quad E(S^2) = \theta^2, \]

we have \( E(\bar{X}^2 - \frac{(n + 1)}{nS^2}) = 0 \) for all \( \theta \) but \( \bar{X}^2 - \frac{(n + 1)}{nS^2} \) isn’t identically zero.
Example 6.5.6. $X_1, \ldots, X_n$ i.i.d. $\sim N(\theta, \theta)$, $\theta > 0$. Show that $\sum X_i^2$ is a complete sufficient statistic.
Some facts

1. If $T$ is complete, then $S = \Psi(T)$ is also complete, where $\Psi(\cdot)$ is some transformation function.
   - Otherwise there exists a function $g(\cdot) \neq 0$ but $E[g(S)] = E[g\{\Psi(T)\}] = 0$ for all $\theta$. So define $g^*(t) = g\{\Psi(t)\}$, a contradiction.

2. Any ancillary statistic cannot be complete.
   - Otherwise, define $g(T) = T - C$, where $C = E(T)$ does not depend on $\theta$. Then $E[g(T)] = 0$ for all $\theta$ though $T \neq C$ wp1.

3. If a function of $T$ is ancillary, then $T$ cannot be complete.
   - If $T$ is complete, by Fact 1, a function of $T$ is also complete, contradiction with Fact 2.
   - This implies that no complete sufficient statistic exists for the $U(\theta, \theta + 1)$ family. Why?
– $T(X) = (X_{(1)}, X_{(n)} - X_{(1)})$ is minimal sufficient, thus a function of every sufficient statistic
– However, a function of $T$, $X_{(n)} - X_{(1)}$ is ancillary.
– So $T$ cannot be complete ⇒ no sufficient statistic is complete (otherwise $T$ will be complete).

4. If $T$ is complete, then only one unbiased estimator of $\Psi(\theta)$ based on $T$ is possible.

Note: an estimator $g(T)$ for a parameter $\Psi(\theta)$ is unbiased if $E[g(T)] = \Psi(\theta)$. In Chapter 7, we will show that this estimator based on a complete sufficient statistic is the best possible unbiased estimator in a certain sense.
6.6 Basu’s Theorem

- $T$ complete sufficient: carries all relevant information about $\theta$.
- $S$ ancillary: carries no information about $\theta$.
- The following result show that they are statistically independent.

**Theorem 7.** If $T(X)$ is a complete sufficient statistic, then $T(X)$ is independent of every ancillary statistic.

Note: The book says that $T(X)$ is minimal sufficient, but the minimal part is not needed in the proof.

**Proof.** (discrete case)
Remarks. Basu’s Theorem depends heavily on completeness. Completeness condition cannot be dropped, even if $T$ is minimal sufficient. For example, for the $U(\theta, \theta + 1)$ distribution, minimal sufficient statistic $T = (X_{(1)}, X_{(n)})$ is not independent of the ancillary statistic $X_{(n)} - X_{(1)}$. 
Example 6.6.1. Let $X_1, \ldots, X_n$ i.i.d. $\sim \text{Exponential}(\theta)$ with $\theta > 0$. Define $g(X) = X_{(1)}/\sum X_i$, the proportion of the total that is accounted for by the smallest observation.

1. Use Basu’s theorem to show that $g(X)$ is independent of $T(X) = \sum X_i$.

2. Calculate $E[g(X)]$. 
Example 6.6.2. Let $X_1, \ldots, X_n$ i.i.d. $\sim N(\mu, \sigma^2)$ with both $\mu$ and $\sigma^2$ unknown. Use Basu’s theorem to show that $\bar{X}$ is independent of $S^2$. 
6.7 The Likelihood Principle

**Definition.** Let \( f(x|\theta) \) denote the joint pdf/pmf of a sample \( X = (X_1, \ldots, X_n) \). Then given that \( X = x \) is observed, the **likelihood function** is defined as \( L(\theta|x) = f(x|\theta) \).

- We typically view the joint pdf/pmf for a sample \( f(x|\theta) \) as the “probability for observing \( x \)” for a given \( \theta \).
- In contrast, we view the likelihood function as a function of \( \theta \) given the observed sample \( x \). But note we are not saying that \( \theta \) is random, but that different values of \( \theta \) would give different probabilities/likelihood for observing \( X = x \).
- Likelihood can be viewed as the degree of plausibility.
- Suppose \( L(\theta_1|x) > L(\theta_2|x) \) for two possible values of \( \theta \), then the sample \( x \) we observed is more likely to occur if the true parameter is \( \theta_1 \) than \( \theta_2 \). So we may say \( \theta_1 \) is more plausible than \( \theta_2 \).
**Likelihood Principle:** Let \( x \) and \( y \) be two sample points such that 
\[ L(\theta|x) \propto L(\theta|y), \]
that is, there exists a function \( C(x,y) \) not depending on \( \theta \) such that
\[ L(\theta|x) = C(x,y)L(\theta|y). \]

Then the conclusions drawn from \( x \) and \( y \) are identical.

**Key idea:** the likelihood ratios are more important quantities.

Suppose \( L(\theta|x) \propto L(\theta|y) \). Then
\[ \frac{L(\theta_1|x)}{L(\theta_2|x)} = 2 \iff \frac{L(\theta_1|y)}{L(\theta_2|y)} = 2. \]

So observing \( x \) or \( y \) gives the same conclusion: \( \theta_1 \) is twice as plausible as \( \theta_2 \) for observing the sample.
6.8 The Equivariance Principle

**Informal definition:** if a statistic $T(\cdot)$ is **equivariant** and if $T(x) = T(y)$, then the inference made if $x$ is observed should have a certain relationship to the inference made if $y$ is observed. However, the two inferences may not be the same.

**Location Family.** Suppose $X_1, \ldots, X_n$ i.i.d. $\sim f(x - \theta)$. Suppose we change $\theta$ to $\theta + b$, then $(X_1, \ldots, X_n) \rightarrow (X_1 + b, \ldots, X_n + b)$.

- **Location-invariant** statistic:
  
  \[ T(X_1 + b, \ldots, X_n + b) = T(X_1, \ldots, X_n), \]
  
  i.e. the statistic and thus inference about $\theta$ does not change. E.g.
  \[ T(X) = \sum_i (X_i - \bar{X})^2. \]

- **Location-equivariant** statistic:
  \[ T(X_1 + b, \ldots, X_n + b) = T(X_1, \ldots, X_n) + b. \]
  Idea: if $\theta$ changes by $b$, then a good estimator of $\theta$ should do the same. E.g.
\[ T(X) = \bar{X} \text{ or } T(X) = \max\{X_i\}. \]

**Scale Family.** Suppose we change \( \theta \) to \( b\theta \), then
\[(X_1, \ldots, X_n) \rightarrow (bX_1, \ldots, bX_n).\]

- **Scale-invariant** statistic: \( T(bX_1, \ldots, bX_n) = T(X_1, \ldots, X_n) \), e.g. anything based on ratios such as \( T(X) = \frac{X_1}{X_n} \).

- **Scale-equivariant** statistic:
  \[ T(bX_1, \ldots, bX_n) = bT(X_1, \ldots, X_n), \text{ e.g. } \bar{X}, S, \text{ median, } X_{(n)}. \]