Chapter 2: Data Reduction

In this chapter we explore the concept of data reduction and and three principles of data reduction.

1 Introduction

Reading Assignment: Section 6.1.

Assume that a sample $X_1, \dots, X_n$ is available to make inference about a parameter θ. How do we summarize the information contained in this sample so (i) we do not need to deal with the original n-dimensional data, and (ii) we have an effective way to extra information from the original data? This task is called "data reduction" and it becomes essential if n is large (Why?).

Answer: Any statistic $T(X_1, \dots, X_n)$ defines a form of data reduction because for any two sets of observed sample values, $\mathbf{x_n} \equiv (x_1, \dots, x_n)$ and $\mathbf{y_n} \equiv (y_1, \dots, y_n)$ will be considered equal (i.e. carrying equal information) if $T(\mathbf{x_n}) = T(\mathbf{y_n})$.

Note: Try to digest the concept that a function of the data is a summary of the data which leads to data reduction. This concept will be helpful in later sections to understand the concepts of and various forms and properties of sufficiency.

Rule of Thumb for Data Reduction: Retain information that are crucial to make inference on the parameter θ, only discard information that is irrelevant to infer θ.

Example 2.1.1 $T(X_1, \dots, X_n) = \sum_{i=1}^n X_i$ seems to be a reasonable reduction for the population mean μ, and $T(X_1, \dots, X_n) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ seems a good reduction for $\theta = (\mu, \sigma^2)$. (Why?)

However, the above is true for certain but not all models. In general, one needs to be cautious about data reduction as we may have inadvertently discarded useful information. - Note that T can be real or vector valued. ■

In the rest of the chapter we will discuss three principles of data reduction: sufficiency, likelihood, and invariance principle.

2 The Sufficiency Principle

Reading Assignment: Section 6.2 and Section 5.4 (Order Statistics).

We will begin with an example to gain intuition.

Example 2.2.1 Let $X_1, \dots, X_n \sim Bin\ (1, \theta)$. To estimate θ, we only need to know $T = \sum_{i=1}^n X_i$, the total number of successes. (Why?) So T summarizes all the information contained in the sample $X_1, \dots, X_n$ and is called a "sufficient statistics" for θ.

Because $f_\theta(\mathbf{X_n} = \mathbf{x_n} \mid T = t) = \dfrac{1}{\binom{n}{t}}$, does not depend on θ. ■

For big data, data reduction is important and sufficient statistics provide such a vehicle.

2.1 Sufficient Statistics

Definition 2.2.1 Let $\mathbf{X_n} \equiv (X_1, \dots, X_n)$ be a random sample from $\{f(x \mid \theta) : \theta \in \Theta\}$. A statistic $T(\mathbf{X_n})$ is called a sufficient statistic for θ, if the conditional distribution of $\mathbf{X_n} = (X_1, \dots, X_n)$ given $T = t$ does not depend on θ, (i.e. It is the same for all θ). ■

Remark 2.2.1: A sufficient statistic T captures all the information about θ but the distribution of T itself depends on θ (Why?)

Otherwise, T would contain no information about θ.

The concept of sufficiency is attributed to Fisher (1922) but his original definition is slightly different but equivalent to the one in Definition 2.2.1.

Example 2.2.2 Let $X_1, \dots, X_n \sim U(0, \theta)$, then the largest order statistic $X_{(n)}$ is sufficient!

We will prove this later by the Factorization Theorem but try to understand intuitively why this is the true. ■

There are two ways to check sufficiency:
(a) through definition (not always easy - try the $U(0, \theta)$ example above)
(b) apply the Factorization Theorem (Theorem 2.2.1) below.

Let $\mathbf{X_n} \equiv (X_1, \dots, X_n)$ be a random sample from $f(x \mid \theta)$, $\theta \in \Theta$ and $f_n(\mathbf{x_n} \mid \theta)$ denote the joint p.d.f. or p.m.f. of $\mathbf{X_n}$. Let $T(X_1, \dots, X_n)$ be a statistic whose p.d.f. or p.m.f. is denoted by $f_T(t \mid \theta)$.

Lemma 2.2.1 A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if, for every possible values $\mathbf{x_n}$ of $\mathbf{X_n}$ in the sample space, the ratio $\frac{f_n(\mathbf{x_n}|\theta)}{f_T(T(\mathbf{x_n})|\theta)}$ is constant as a function of θ. (i.e. the value of the ratio does not involve θ).

Proof: The proof follows directly from the definition of sufficient statistics but involves careful handling of the event $T(\mathbf{X_n}) = T(\mathbf{x_n})$ because its probability may be zero for a continuous distribution. Details are provided on page 273 of the textbook.

To verify this we need to show that, for any fixed value $\mathbf{x_n}$ and t, the conditional probability $P_\theta(\mathbf{X_n} = \mathbf{x_n} \mid T(\mathbf{X_n}) = t)$ does not depend on θ. Luckily we only need to show this for $t = T(\mathbf{x_n})$ as the probability would be zero otherwise.

For $t = T(\mathbf{x_n})$, observe that the event $\{\mathbf{X_n} = \mathbf{x_n}\}$ is a subset of the event $\{T(\mathbf{X_n}) = T(\mathbf{x_n})\}$, hence

$$\begin{aligned}
P_\theta(\mathbf{X_n} = \mathbf{x_n} \mid T(\mathbf{X_n}) = T(\mathbf{x_n})) &= \frac{P_\theta(\mathbf{X_n} = \mathbf{x_n}, T(\mathbf{X_n}) = T(\mathbf{x_n}))}{P_\theta(T(\mathbf{X_n} = T(\mathbf{x_n}))} \\
&= \frac{P_\theta(\mathbf{X_n} = \mathbf{x_n} \mid \theta)}{P_\theta(T(\mathbf{X_n}) = T(\mathbf{x_n}))} \\
&= \frac{f_n(\mathbf{x_n} \mid \theta)}{f_T(T(\mathbf{x_n}) \mid \theta)},
\end{aligned}$$

Which does not involve θ. Hence T is sufficient. ■

Theorem 2.2.1 (Fisher - Neymann Factorization Theorem):
A statistic $T(X_1, \dots, X_n)$ is sufficient for θ if and only if $f_n(\mathbf{x_n} \mid \theta) = g(T(\mathbf{x_n}) \mid \theta)\, h(\mathbf{x_n})$, where h does not depend on θ and g depends on $\mathbf{x_n}$ only through $T(\mathbf{x_n})$.

- i.e. sufficiency ⇔ the likelihood function $L(\theta) = f_n(\mathbf{x_n} \mid \theta)$ (as a function of θ for a fixed $\mathbf{x_n}$ is proportional to a function (g) that depends on the data ($\mathbf{x_n}$) only through T.

Proof: ⇒) If T is sufficient $f_{\mathbf{x_n}|T}(\mathbf{x_n} \mid t) = h(\mathbf{x_n})$, $f_T(t \mid \theta) = g(T(\mathbf{x_n}), \theta)$.

⇐) If Factorization Theorem holds, we will prove the discrete case, the continuous case is similar but more subtle. Let $A_{T(\mathbf{x_n})} = \{\mathbf{y_n} \in \mathbb{R}^n, T(\mathbf{y_n}) = T(\mathbf{x_n})\}$, then

$$\begin{aligned}
\frac{f_n(\mathbf{X_n} \mid \theta)}{f_T(T(\mathbf{x_n}) \mid \theta)} &= \frac{h(\mathbf{x_n})g(T(\mathbf{x_n}) \mid \theta)}{\sum_{\mathbf{y_n} \in A_{T(\mathbf{x_n})}} h(\mathbf{y_n})g(T(\mathbf{y_n}) \mid \theta)} \\
&= \frac{h(\mathbf{x_n})g(T(\mathbf{x_n}) \mid \theta)}{g(T(\mathbf{x_n}) \mid \theta) \sum_{\mathbf{y_n} \in A_{T(\mathbf{x_n})}} h(\mathbf{y_n}) \mid \theta)} \\
&= \frac{h(\mathbf{x_n})}{\sum_{\mathbf{y_n} \in A_{T(\mathbf{x_n})}} h(\mathbf{y_n})},
\end{aligned}$$

which does not depend on θ. So Lemma 6.2.1 implies that T is sufficient. ■

Example 2.2.1 (continued) $f_n(\mathbf{x_n} \mid \theta) = \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}$.

$\Rightarrow h = 1, g(t \mid \theta) = \theta^t (1-\theta)^{n-t}$; where $t = \sum_i x_i$. ■

Example 2.2.2 (continued)

$$\begin{aligned} y &= \begin{cases} \frac{1}{\theta^n}, & \text{if } 0 \le x_i \le \theta, \forall i \Leftrightarrow 0 \le x_{(1)} \le x_{(n)} \le \theta \\ 0, & \text{otherwise} \end{cases} \\ &= \frac{1}{\theta^n} 1_{[0,\theta]}(x_{(n)}) \overbrace{1_{(0,\infty)}(x_{(1)})}^{h}. \end{aligned}$$

■

Example 2.2.3 $X_1, \ldots, X_n \sim N(\theta, \sigma^2)$, σ^2 is known.

$$\begin{aligned} f_n(\mathbf{x_n} \mid \theta) &= \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_i (x_i-\theta)^2} \\ &= \underbrace{\frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{1}{2\sigma^2}\sum_i x_i^2}}_{h} \underbrace{e^{\frac{2\theta}{2\sigma^2}\sum_i x_i - \frac{n\theta^2}{2\sigma^2}}}_{g(\sum_i x_i \mid \theta) \text{ or } g(\bar{x} \mid \theta)} \end{aligned}$$

$\Rightarrow \sum_i X_i$ and $\bar{X}$ are both sufficient, and $\frac{1}{2}\bar{X}$ is also sufficient. (Why?) ■

Theorem 2.2.2 If T is sufficient, any one-to-one function $\gamma(T) \equiv T^*$ of T is also sufficient.
Proof: Since $g(T(\mathbf{x_n}) \mid \theta) = g(\gamma^{-1}(T^*(\mathbf{x_n})) = g^*(T^*(\mathbf{x_n}))$, where $g^* = g(\gamma^{-1})$. The Factorization Theorem applies to the same h and g^*. Hence T^* is sufficient. ■

Example 2.2.4 Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ is known.
$\Rightarrow T = \sum_i (X_i - \mu)^2$ is sufficient for σ^2. ($u(x) = 1$).
$\Rightarrow$ Both $\frac{1}{n}\sum_i (X_i - \mu)^2$ and $\frac{1}{n-1}\sum_i (X_i - \mu)^2$ are sufficient for σ^2. ■

Example 2.2.5 Let $X_1, \ldots, X_n \sim N(\theta, \theta^2)$.

$$\begin{aligned} f_n(x \mid \theta) &= \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n e^{-\frac{1}{2\theta^2}\sum_i (x_i-\theta)^2} \\ &= \left(\frac{1}{\sqrt{2\pi}\theta}\right)^n e^{-\frac{1}{2\theta^2}[\sum_i x_i^2 - 2\theta \sum_i x_i]} e^{-\frac{n}{2}}. \end{aligned}$$

There does not exist a single sufficient statistic! We need both $\sum_i X_i^2$ and $\sum_i X_i$.

Likewise, if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, there is no single sufficient statistic for $\theta = (\mu, \sigma)$.
We need $\sum_i X_i$ and $\sum_i X_i^2$. ■

This leads to the concept of joint sufficient statistics when more than one sufficient statistic is needed, that is, when T is vector valued.

2.2 Jointly Sufficient Statistics

Definition 2.2.2 Let $X_1, \ldots, X_n \sim f(x \mid \theta)$, θ may be in $\mathbb{R}$ or $\mathbb{R}^p$. Then $T_i(X_1, \ldots, X_n), i = 1, \ldots, k$, are called jointly sufficient for $\theta \Leftrightarrow$ the conditional distribution of $\mathbf{X_n} = (X_1, \ldots, X_n)$ given $T = (T_1, \ldots, T_k)$ does not depend on θ.

The Factorization Theorem still holds and any one-to-one function of $(T_1, \ldots, T_k)$ is sufficient.

Factorization Theorem $T_1, \ldots, T_k$ are jointly sufficiently for θ

$\Leftrightarrow f_n(\mathbf{x_n} \mid \theta) = g(T_1(\mathbf{x_n}), \ldots, T_k(\mathbf{x_n}) \mid \theta)\ h(\mathbf{x_n})$.

Example 2.2.5 (continued) For $N(\theta, \theta^2)$, $\sum_i X_i$ and $\sum_i X_i^2$ are jointly sufficiently for θ by the Facterization Theorem.

Map $\begin{array}{cccccc} (\sum_i X_i, & \sum_i X_i^2) & \rightarrow & (\bar{X}, & \sum_i (X_i - \bar{X})^2) \\ T_1 & T_2 & & T_1' & T_2' = T_2 - \frac{1}{n} T_1^2 \end{array}$

This is a one-to-one function with $g(a, b) = (\frac{1}{n} a, b - \frac{a^2}{n})$.

Proof: If $\frac{1}{n} a_1 = \frac{1}{n} a_2$ and $b_1 - \frac{a_1^2}{n} = b_2 - \frac{a_2^2}{n}$,

$\Rightarrow a_1 = a_2 \Rightarrow b_1 = b_2$, since $a_1 = a_2$.

So, $(\bar{X}, \sum_i (X_i - \bar{X})^2)$ is also jointly sufficient for θ.

Likewise, $(\sum_i X_i, \frac{1}{n-1} \sum_i (X_i - \bar{X})^2)$ or $(\bar{X}, \frac{1}{n} \sum_i (X_i - \bar{X})^2)$ are all jointly sufficient because they are one-to-one functions of jointly sufficient statistics.

The derivation of jointly sufficient statistics for the case of $N(\mu, \sigma^2)$ is similar to the case for $N(\theta, \theta^2)$ and all the above jointly sufficient statistics are jointly sufficient here as well. ■

Many parametric distributions, such as the normal distribution $N(\mu, \sigma^2)$ and binomial distribution $Bin(n, \theta)$, belong to a larger class of distributions termed the "exponential family", which has the benefit that data reduction is readily available. To see this, we first introduce the exponential family of distributions.

Definition 2.2.3 An exponential family of distribution, $\{f(x \mid \theta), \theta \in \Theta\}$, is a class of distributions whose p.d.f. or p.m.f. takes the form:

$$f(x \mid \theta) = h(x)\ c(\theta)\ e^{\sum_{j=1}^{k} w_j(\theta)\ T_j(x)}. \tag{1}$$

■

It now follows from the Factorization Theorem above that $T(\mathbf{x_n}) = (\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_k(X_i))$ is a sufficient statistic for θ. This is a very attractive properties of the exponential family as it facilitates dimension reduction intrinsinctly.

When the data do not come from an exponential family or even a parametric family, data reduction might be challenged. Below we show that the order statistics from a random sample are always sufficient regardless of its form.

Example 2.2.6 Let $X_1, \ldots, X_n \sim f(x \mid \theta)$ and $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ be their order statistics, then $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$ is always sufficient for θ (meaning one can just record data in ascending order without loosing any information).
Intuition: the order of X's does not carry information.
Proof: $f_n(\mathbf{x_n} \mid \theta) = \prod_i f(x_i \mid \theta) = \prod_{i=[j]} f(x_{[j]} \mid \theta) = \prod_i f(x_{(i)} \mid \theta)$.

An Easier Proof: $f_n(\mathbf{x_n} \mid x_{(1)}, \ldots, x_{(n)}) = \frac{1}{n!}$, all $n!$ arrangements are equal likely. ■

Remark: Example 2.2.5 requires no particular form of f. In fact, there may be no finite-dimensional parameter involved at all, so the parameter itself could be the p.d.f. (or p.m.f.) itself. The bottom line is that the order statistics are always sufficient. Sometime, they are the only way to reduce data, especially if the sample does not come from an exponential family. Exercise 6.8. on page 301 of the textbook provides such an example. However, often the order statistics contain more information than is needed to be a sufficient statistic. This leads to the concept of minimal sufficiency.

2.3 Minimal Sufficiency

Consider $N(\mu, \sigma^2)$, we have many sufficient statistics including $(X_{(1)}, \ldots, X_{(n)})$ which however does not summarize the information succinctly as we can summarize it further. The goal is to have the simplest set of sufficient statistic(s) as sufficiency is about data reduction.

Definition 2.2.4 A statistic $T = (T_1, \ldots, T_k)$ is a minimal sufficient statistic of θ iff T is sufficient and it is a function of every other sufficient statistic.

$\Leftrightarrow$ T cannot be reduced further without destroying the property of sufficiency.

$\Leftrightarrow$ Any function $\psi(T)$ of a minimal sufficient statistic T is sufficient $\Leftrightarrow$ ψ is one-to-one.

Remark: Sufficient statistics obtained from the Factorization Theorem in the "most concise" way are usually minimal sufficient.
For instance, $(\sum_i X_i, \sum_i X_i^2)$ is jointly sufficient for a random sample from $N(\mu, \sigma^2)$
but $(X_1 + X_2, X_3 + \cdots + X_n, \sum_i X_i^2)$ is not minimal sufficient.

How do we prove minimal sufficiency? One way is to go through the maximum likelihood or Bayes method (to be elaborated later). Below we introduce a general way to proactively find minimal sufficient statistics.

Theorem 2.2.3 (Lehmann-Scheffé Theorem) Let $f(x \mid \theta)$ be the p.d.f. and p.m.f. of a sample $\mathbf{X_n}$. A statistic $T(\mathbf{X_n})$ is minimal sufficient if it has the following property:

for every two sample points $\mathbf{x_n}$ and $\mathbf{y_n}$, the ratio $\frac{f(\mathbf{x_n})|\theta)}{f(\mathbf{y_n})|\theta)}$ is constant as a function of θ if and only if $T(\mathbf{x_n}) = T(\mathbf{y_n})$.

Proof: The proof is quite involved but the special case when $f(\mathbf{x_n} \mid \theta) > 0$, for all $\mathbf{x_n}$ and θ, is provided in the proof of Theorem 6.2.13 on page 281 of the textbook. ■

Example 2.2.7 Let $\mathbf{X_n}$ be a random sample from a uniform distribution on the interval $[\theta, \theta+1]$, $-\infty < \theta < \infty$. Find a minimal sufficient statistic for θ.

Solution: The joint p.d.f. of $\mathbf{X_n}$ is

$$f(\mathbf{x_n} \mid \theta) = \begin{cases} 1, & \text{if } \theta \le x_i \le \theta + 1, \forall i, \\ 0, & \text{otherwise,} \end{cases}$$

which is equivalent to

$$f(\mathbf{x_n} \mid \theta) = \begin{cases} 1, & \text{if } \max_i (x_i - 1) \le \theta \le \min_i x_i, \\ 0, & \text{otherwise.} \end{cases}$$

Here for any two $\mathbf{x_n}$ and $\mathbf{y_n}$ the ratio $\frac{f(\mathbf{x_n})|\theta)}{f(\mathbf{y_n})|\theta)}$ will not involve θ if and only if $\min_i x_i = \min_i y_i$ and $\max_i x_i = \max_i y_i$. (Why?)

Thus, $X_{(1)}$ and $X_{(n)}$ are minimal (joint) sufficient statistics. ■

Remark: (i) In the above example, the minimal sufficient statistic has dimension two while the parameter is only one dimensional. We have seen a similar example before (Example 2.2.5) so this should not be a surprise but its cause is still worth thinking.

(ii) A minimal sufficient statistic is not unique as any one-to-one function of it carries exactly the same information so is also minimal sufficient.

(iii) In addition to data reduction a sufficient statistic can be used to improve an estimator. We will learn this in Chapter 7.

(iv) A cautionary remark is in order here. Although a sufficient statistic contains all the information in data it does not mean that we should discard the original data once we have extracted the information of the sufficient statistic. One reason is that in order to do statistical inference, or for uncertainty quantification, we may need information beyond sufficient statistics. See Cox (1971) for details. In the next two subsections we provide additional evidence that a statistic may be useful even if it contains no information of the parameter.

2.4 Ancillary Statistics

So far, we have focused on sufficient statistics as they seem to contain all the information in the data. In this and the next subsection we will learn another type of statistics that complements sufficient statistics.

Definition 2.2.5: A statistic $T(\mathbf{X_n})$ whose distribution does not depend on the parameter θ is called an ancillary statistic. ■

There are several definitions for ancillary statistic (very confusing) and the concept in Definition 2.2.5 is attributed to Fisher but he did not provide a former definition. This definition suggests that an ancillary statistic for θ does not carry any useful information about θ. However, surprisingly, an ancillary statistic could contribute valuable information about θ when used in conjunction with other statistics. We will explore this in a homework problem (Exercise 6.12) and further in the next subsection. Below we show two examples of ancillary statistics.

Example 2.2.8 (Location Family) Let $\mathbf{X_n}$ be a random sample from a location family with c.d.f. (cumulative distribution function) $F(x-\theta)$, $-\infty < \theta < \infty$. Intuitively, the range $R \equiv X_{(n)} - X_{(1)}$ contains no information on the location θ and hence is ancillary. To see this, let $Z_1, \ldots, Z_n$ be a random sample from $F(x)$, i.e. $\theta = 0$. Then notice that $Z_i + \theta$, $1 \leq i \leq n$ have the same joint distribution as $X_1, \ldots, X_n$. Clearly, the range of X_i is the same as the range of Z_i, the latter carries no information on θ. Hence, the range R is an ancillary statistic for the location parameter. For another proof, see Example 6.2.18 on page 283 of the textbook. ■

Example 2.2.9 (Scale Family) Let $\mathbf{X_n}$ be a random sample from a location family with c.d.f. (cumulative distribution function) $F(x/\theta), \theta > 0$. It is intuitively clear (and can be shown rigorously in a homework problem) that X_i/X_n, $1 \leq i \leq n-1$ contain no information on θ. Hence any statistic that depend on the sample $\mathbf{X_n}$ only through X_i/X_n, $1 \leq i \leq n-1$ is an ancillary statistic. ■

Homework 2.1 (i) Show that the $n-1$ variables, X_i/X_n, $1 \leq i \leq n-1$, contain no information on θ. Note that you need to show this for the $n-1$ variables jointly.

(ii) Provide two more examples of ancillary statistic in addition to $\bar{X}/X_n$. One of them should involve the sample median.

We close this subsection with an example that serves as a cautionary remark not to dismiss an ancillary statistic automatically.

Example 2.2.9 Let N be a discrete r.v. with $P(N = j) = p_j, j = 1, 2, ...,$ where p_j does not involve θ. An experiment will be performed in two stages, where in the first stage a random N will be selected. In the next stage after having observed that $N = n$, a sequence of independent Bernoulli, $Bin(1, \theta)$, trials will be performed and X denote the number of success in these n trials.

Exercise 6.12 on page 301 of the textbook reveals that a minimal sufficient statistic for this experiment is (X, N), which contains an ancillary statistic (N in this case). This seems paradoxical but the moral is an ancillary statistic alone may not contain any information but when it is combined with another statistic it becomes useful. We will further elaborate on this in the next subsection. ■

2.5 Sufficient, Ancillary, and Complete Statistics

Since a minimal sufficient statistic is supposed to have achieved the maximal amount of data reduction to retain relevant information about θ and the distribution of an ancillary statistics does not involve θ we expect that a minimal sufficient statistic to be independent of any ancillary statistic. However, this is not always the case as we have seen in Example 2.2.9. The question is: When will a minimal sufficient statistic be independent of any ancillary statistic? The answer is provided by D. Basu in his 1955 paper in Sankhya. To present this we first need the following concept of completeness.

Definition 2.2.6 Let $f_T(t \mid \theta), \theta \in \Theta$, be a family of p.d.f.s or p.m.f.s for a random variable or vector T. This family of probability distributions is called complete if $E_\theta\ g(T) = 0$ for all $\theta \in \Theta$ implies that $P_\theta(g(T) = 0) = 1$ for all θ, i.e. g(T) is zero (with probability one) regardless of which θ generates its distribution.
When T is a statistic $T(\mathbf{X_n})$ from a random sample, we say that $T(\mathbf{X_n})$ is a complete statistic if the family if its sampling distribution $f_T(t \mid \theta)$ is complete. ■

Remark: The concept of completeness seems abstract (**Why is it called "complete"?**) and it is a property of a "family" of distribution, not of a particular distribution.

Example 2.2. 10 Let $T \sim Bin(n, \theta)$, with $0 < \theta < 1$. Assume that $E_\theta\ g(T) = 0$, for all $0 < \theta < 1$. It can be shown that $E_\theta\ g(T)$ is a polynomial of degree n in $\gamma \equiv \frac{\theta}{1-\theta}$, where the coefficient of γ^k is

$g(k)\binom{n}{k}$, for $k = 0, 1, \ldots, n$.

In order for this polynomial to be identical to zero all of its coefficients must be zero, which implies that $g(k) = 0$, for all k. Since T takes the values $\{0, 1, \ldots, n\}$ with probability one, this forces $P_\theta(g(T) = 0) = 1$, for all θ. Hence $\{Bin(n, \theta) : 0 < \theta < 1\}$ is a complete family of distributions and T is complete if $T = \sum_{i=1}^n X_i$, where $X_i \sim Bin(1, \theta)$ are i.i.d. ■

Theorem 2.2. 4 (Basu's Theorem) If $T(\mathbf{X_n})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X_n})$ is independent of every ancillary statistic.
Proof: The proof for the discrete case is provided in the textbook (page 287) ■

Remark: Basu's Theorem is very useful but to show completeness of a statistic is not always easy as we have seen in Example 2.2.10. Luckily, it is very easy to show the completeness of an exponential family using the following theorem.

Theorem 2.2.5 Let $X_1, \ldots, X_n$ be a random sample from an exponential family of distribution with order k, whose p.d.f. or p.m.f. takes the form:

$$f(x \mid \theta) = h(x)\; c(\theta)\; e^{\sum_{j=1}^{K} w_j(\theta)\; T_j(x)}. \tag{2}$$

Then the statistic $T(\mathbf{X_n}) = (\sum_{i=1}^n T_1(X_i), \ldots, \sum_{i=1}^n T_K(X_i))$ is complete as long as the parameter space Θ contains an open set in $\mathbb{R}_K$.
Proof: The proof follows from the uniqueness of a Laplace transform and will be omitted. ■

This theorem implies that the family $N(\theta, \theta^2)$ in Example 2. 2.5 is not complete as the parameter space does not contain any two-dimensional open set. (Why?) In fact, its parameter space contains only points on a parabola. Such an exponential family is termed "curved" exponential family as it is an exponential family but its parameter space is a curve.

Example 2.2.11 The exponential distributions with p.d.f. $\frac{1}{\theta}e^{-\frac{x}{\theta}}, x \geq 0$ belong to an exponential family with $T(\mathbf{X_n}) = \sum_{i=1}^n X_i$. Hence, It follows from Theorem 2.2.5 that $T(\mathbf{X_n}) = \sum_{i=1}^n X_i$ is complete and sufficient for θ since $\Theta = (0, \infty)$. Given that the exponential family is also a scale family (Why?), it can be shown easily that $T^*(\mathbf{X_n}) = \frac{\mathbf{X_j}}{\sum_{i=1}^n X_i}$ is an ancillary statistic. Basu's Theorem then implies that $T(\mathbf{X_n})$ and $T^*(\mathbf{X_n})$ are independent.

Since $\theta = E_\theta[X_j] = E_\theta[T(\mathbf{X_n})T^*(\mathbf{X_n})] = E_\theta[T(\mathbf{X_n})]E_\theta[T^*(\mathbf{X_n})] = n\theta E_\theta[T^*(\mathbf{X_n})]$, this leads to the result that $E_\theta T^*(\mathbf{X_n}) = \frac{1}{n}$, which is also a consequence of the memoryless property of an exponential distribution. ■

Example 2.2.11 illustrate an elegant application of Basu's Theorem. A second application of Basu's Theorem is to show the independence of sample mean and sample variance for a normal $N(\mu, \sigma^2)$ family. We have shown this independence in Chapter 1 but Basu's Theorem provides another way to proof it. This alternative way is more intuitive as the sample variance is an ancillary statistic for the mean μ and the sample mean is a complete and minimal sufficient statistic. However, there are details that need to be filled and you can find a complete argument in Example 6.2.27 of the textbook (page 289). Please go through the argument on page 289 as it may appear on an exam. In a nut shell, Basu's Theorem may provide an elegant way to show the independence of two statistics.

A fundamental property of a complete statistic is that it must be minimal as implied by the following theorem.

Theorem 2.2. 6 If a minimal sufficient statistic exists, then any complete statistic is also a minimal sufficient statistic.

Theorem 2.2.6 implies that the assumption of "minimal" in Basu's Theorem is not needed.

Summary: In this section we learn the concept of sufficiency, minimal sufficiency, and completeness. Sufficiency of a statistic can be checked through the Fisher-Neyman Factorization Theorem and minimal sufficiency can be checked through The Lehmann-Scheffè. Completeness is concept for a family of distributions and for an exponential family of order k it hinges on whether its parameter space contains an open set in $\mathbb{R}_K$. Basu's Theorem implies that a complete and sufficient statistic is independent of any ancillary statistic, hence it accomplishes the ultimate goal of data reduction.

3 The Likelihood Function

Reading Assignment: Section 6.3.

The likelihood function is a very useful tool in statistics, which we will learn repeatedly in this course. We have already encountered it before as the joint distribution $f(\mathbf{x_n} \mid \theta)$ of a random sample but that view treats it as a function of the data $\mathbf{x_n}$ for a given parameter θ. In this section we reverse the role of $\mathbf{x_n}$ and θ and regard it as function of θ given the data.

Definition 2.3.1 Let $f(\mathbf{x_n} \mid \theta)$ be the joint p.d.f. of the sample $\mathbf{X_n}$ and that $\mathbf{X_n} = \mathbf{x_n}$ has been observed. The likelihood function, defined as

$$L(\theta \mid \mathbf{x_n}) = f(\mathbf{x_n} \mid \theta) \tag{3}$$

is a function of θ given that $\mathbf{X_n} = \mathbf{x_n}$ is observed. ■

If $\mathbf{X_n}$ is a discrete random vector, then $L(\theta \mid \mathbf{x_n}) = P_\theta(\mathbf{X_n} = \mathbf{x_n})$. So if $P_{\theta_1}(\mathbf{X_n} = \mathbf{x_n}) = L(\theta_1 \mid \mathbf{x_n}) > L(\theta_2 \mid \mathbf{x_n}) = P_{\theta_2}(\mathbf{X_n} = \mathbf{x_n})$, then the sample is more likely to have come from θ_1 than θ_2. This concept can be extended to continuous distribution with a little modification. For a continuous r.v. X with p.d.f. $f(x \mid \theta)$, $P(x - \theta < X < x + \theta)$ is approximately $2\epsilon f(x \mid \theta) = 2\epsilon L(\theta \mid x)$. Hence,

$$\frac{P_{\theta_1}(x - \epsilon < X < x + \epsilon)}{P_{\theta_2}(x - \epsilon < X < x + \epsilon)} \approx \frac{L(\theta_1 \mid x)}{L(\theta_2 \mid x)},$$

and the sample is more likely to have come from θ_1 than θ_2 if $L(\theta_1 \mid x) > L(\theta_2 \mid x)$.

Thus, $L(\theta \mid \mathbf{x_n})$ represent the information provided by the likelihood function. The likelihood principle below specifies how the likelihood function should be used as a data reduction device.

Likelihood Principle: If $\mathbf{x_n}$ and $\mathbf{y_n}$ are two sample points such that $L(\theta \mid \mathbf{x_n})$ is proportional to $L(\theta \mid \mathbf{y_n})$, i.e. there exists a constant (meaning that it does not involve θ) $C(\mathbf{x_n}, \mathbf{y_n})$ such that

$$L(\theta \mid \mathbf{x_n}) = C(\mathbf{x_n}, \mathbf{y_n}) L(\theta \mid \mathbf{y_n}) \text{ for all } \theta, \tag{4}$$

then the conclusions drawn from $\mathbf{x_n}$ and $\mathbf{y_n}$ should be identical. ■

Here we allow the likelihood ratio to be a constant $C(\mathbf{x_n}, \mathbf{y_n})$ (as a function of θ) rather than insisting $C(\mathbf{x_n}, \mathbf{y_n}) = 1$ because the likelihood function assesses the "plausibility" of the various parameter values. (You need to think hard about this. Why do we not insists $C = 1$?)

Here the term "plausibility" is used in lieu of "probability" because θ is considered a fixed (but unknown) constant. There is another school called the "fiducial statistics" that takes a more flexible approach to regard θ as random even though $L(\theta \mid \mathbf{x_n})$ as a function of θ is not a p.d.f. or p.m.f.

Fiducial Inference: One form of inference that is between the frequentist and Bayesian approach, is the fiducial approach for statistical inference. (Fiducial means founded on faith or trust, or taken as standard of reference) The approach, which dates back to Fisher but is not subscribed by most statisticians, made a recent come back among a small but growing number of statisticians.

The approach interprets likelihood as probability for θ, thereby multiplying $L(\theta \mid \mathbf{x_n})$ by $M(\mathbf{x_n}) = (\int L(\theta \mid \mathbf{x_n}) d\theta)^{-1}$ (i.e. treating $L(\theta \mid \mathbf{x_n})$ as a pseudo p.d.f. or p.m.f.) so $M(\mathbf{x_n})L(\theta \mid \mathbf{x_n})$ is interpreted as a p.d.f. for θ (provided $M(\mathbf{x_n})$ is finite). Note that the fiducial approach is different from the Bayesian approach, which also treats θ as a random quantity but assign a prior distribution to it. We will discuss the Bayesian approach for inference in the next chapter.

Example 2.3.1 (Example 6.3.3. on page 291 here)

Summary: In this section we define the likelihood function as a function of θ given the observed value $\mathbf{x_n}$. We then explore the concept of likelihood principle for data reduction and claim that one parameter is more "plausible" then another if the likelihood function evaluated at this parameter value is larger than the one evaluated at the other parameter value. The concept of fiducial statistic is also introduced.

Question: Why is the likelihood principle a data reduction approach? Where is the reduction on data?

4 The Equivariance Principle

Reading Assignment: Section 6.4.

In this section we explore a different concept of data reduction that involves two different types of equivariance concept. The situation typically arises when one is making an inference (estimation, testing etc.) and the principle(s) dictates what must be abide by for related sample points.

Measurement Equivariance: The inference made should not depend on the measurement scale that is used. That is, if two people carry out the experiment using the same inference procedure but different scales, e.g. one uses the metric system and the other uses the imperial system, then they should reach the same conclusion. ■

This principle seems very natural and almost everyone would agree that it is reasonable.

Formal equivariance/invariance: If two inference problems have the same formal structure, then the same inference procedure should be used in both problems. Here the same formal structure refers to both the model (both are of the form $\{f(x \mid \theta) : \theta \in \Theta\}$) and the set of allowable inferences and consequences of wrong inferences (which should be the same too). ■

For example, if the goal is to estimate θ then the set of allowable inference might be Θ and the consequence of wrong inference could be measured as the mean square error. Note that formal invariance is only concern with the mathematical entities involved, not the physical description of the experiment. For instance, Θ may be $\Theta = \{\theta : 0 \leq \theta \leq 4\}$ in both problems but in one problem θ may be the average GPA of UCD Statistics majors but in another problem it may be the average time to complete a task in a factory. If they have the same formal structure in terms of the model and rules as describe above, then the formal invariance principle requires the same inference procedure be employed to both.

Finally, note that this form of equivariance is really an invariance.

Equivariance Principle: If $\mathbf{Y_n} = g(\mathbf{X_n})$ is a change of measurement scale such that the model for $\mathbf{Y_n}$ has the same "formal" structure as the model for $\mathbf{X_n}$, then an inference procedure should be both measurement equivariant and formally equivariant. ■

Several example of the equivariance principle are provided in the textbook. The measurement equivariance principle is intuitively reasonable but requires a choice of the transformations. We will postpone this topic till later. The formal equivariance/invariance principle is more controversial. Nevertheless, the equivariance principle is a data reduction techniques that restrict inference by prescribing rules for related sample points.