Lectures 3–4 – Sufficiency and Minimal Sufficiency

Last Modified: September 3, 2026

Outline:

3.1 Data reduction

Recall the statistical framework from Lecture 1.

- A **statistical model** is a family of distributions

$$\{f_\theta(x) : \theta \in \Theta\}.$$

  The unknown parameter θ indexes the possible distributions. Suppose x lives in $\mathcal{X}$ (e.g. $\mathcal{X} = \mathbb{R}$ for a normal distribution and $\mathcal{X} = \{0, 1\}$ for a Bernoulli). We observe a realization $\underline{x} = (x_1, \dots, x_n) \in \mathcal{X}^n$ of a random sample

$$\underline{X} = (X_1, \dots, X_n), \qquad X_1, \dots, X_n \overset{\text{iid}}{\sim} f_\theta.$$

  By independence, the joint pmf or pdf of the sample is

$$f_{\underline{X},\theta}(\underline{x}) = \prod_{i=1}^{n} f_\theta(x_i).$$

- The goal of **statistical inference** is to learn about the unknown θ from the observed sample. For example, we may construct a point estimate, test a hypothesis, or form a confidence interval. In every case, our conclusion is a function of the data $\underline{X}$.

To make inference more manageable, we first try to discard extraneous information. We replace the full sample by a statistic $T(\underline{X})$ that summarizes the part of the data relevant to θ. This is called **data reduction**.

Example 1 (Bernoulli motivation)**.** Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$, where p is the unknown probability of success. Suppose $n = 3$ and we observed

$$\underline{x} = (1, 1, 0).$$

Intuitively, this is more information than we need: it is enough to know that two of the three observations are successes to learn about p. This information is recorded by

$$T(\underline{x}) = \overline{x} = \frac{1}{3}\sum_{i=1}^{3} x_i = \frac{2}{3}.$$

The samples $(1,1,0)$ and $(0,1,1)$ should lead to the same conclusion about p.

More generally, any statistic T partitions the sample space. For every possible value t of T, define the **fiber**

$$A_t = \{\underline{x} \in \mathcal{X}^n : T(\underline{x}) = t\}.$$

The statistic value t tells us which fiber A_t contains the observed sample, but not which particular point in that fiber was observed.
For $n = 3$ Bernoulli observations and $T(\underline{x}) = \overline{x}$, the partition is shown in Figure 3.1.

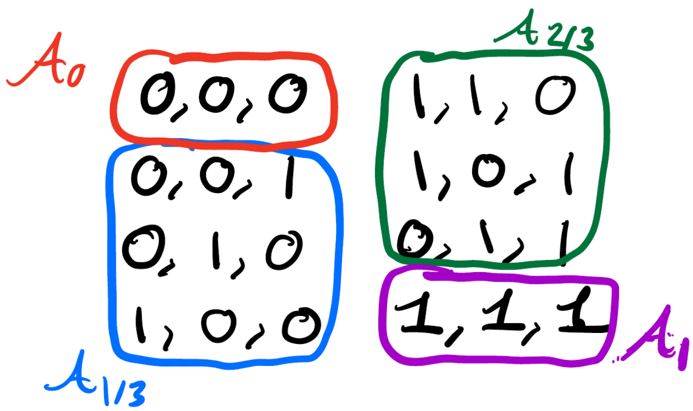
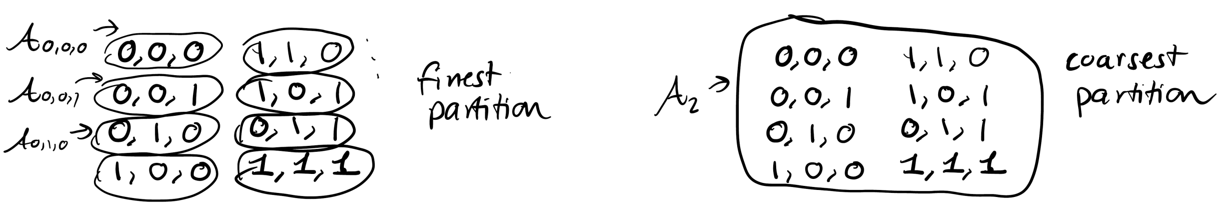


Figure 3.1: The fibers of the sample mean for three Bernoulli observations.

The finer the partition, the less data reduction we perform. For example, $T(\underline{x}) = \underline{x}$ gives singleton fibers and performs no reduction. The coarser the partition, the more reduction we perform, and we risk losing information. A constant statistic such as $T(\underline{x}) = 2$ puts every sample point in one fiber and throws away all the information. Figure 3.2 illustrates these two extremes.



No data reduction All information is lost

Figure 3.2: The finest and coarsest possible partitions of a sample space.

Our goal is to choose a statistic in the middle between these two: it should *reduce* the data *without losing information* about θ.

3.2 Sufficient statistics

The key intuition is that we may discard any part of the data whose distribution is unaffected by θ. A sufficient statistic retains all of the θ-dependence in the sample; once its value is known, the remaining variation in the data carries no further information about θ.

Definition 1 (Sufficient statistic)**.** A statistic $T(\underline{X})$ is **sufficient** for θ if, for every possible value t, the conditional distribution of $\underline{X}$ given $T(\underline{X}) = t$ does not depend on θ.

To understand Definition 1, think of the full sample as being generated in two stages:

1. Draw t from the distribution of $T(\underline{X})$.
2. Given $T(\underline{X}) = t$, draw $\underline{X}$ from its conditional distribution.

If T is sufficient, then θ appears only in the first stage, whereas the second stage carries no extra information about θ.
The two-stage structure is reflected in the pmf of $\underline{X}$. Let us first understand this in the discrete case. Fix a sample point $\underline{x}$ and write $t = T(\underline{x})$. Since

$$\{\underline{X} = \underline{x}\} \subseteq \{T(\underline{X}) = t\},$$

we have

$$\begin{aligned}\mathbb{P}_\theta\big(\underline{X} = \underline{x} \mid T(\underline{X}) = t\big) &= \frac{\mathbb{P}_\theta\big(\underline{X} = \underline{x}, T(\underline{X}) = t\big)}{\mathbb{P}_\theta\big(T(\underline{X}) = t\big)} \\ &= \frac{\mathbb{P}_\theta(\underline{X} = \underline{x})}{\mathbb{P}_\theta\big(T(\underline{X}) = t\big)}.\end{aligned}$$

Equivalently,

$$f_{\underline{X},\theta}(\underline{x}) = \mathbb{P}_\theta(\underline{X} = \underline{x}) = \underbrace{\mathbb{P}_\theta\big(T(\underline{X}) = t\big)}_{\text{Stage 1: sample } T} \underbrace{\mathbb{P}_\theta\big(\underline{X} = \underline{x} \mid T(\underline{X}) = t\big)}_{\text{Stage 2: sample } \underline{X} \mid T}. \tag{3.1}$$

The decomposition (3.1) is true whether or not T is sufficient. If T *is* sufficient, then only the first factor depends on θ.

Example 2 (A sufficient statistic for a Bernoulli sample)**.** Let $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$ and $T(\underline{X}) = \overline{X}$. The distribution of T is

t	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$\mathbb{P}_p(T = t)$	$(1-p)^3$	$3p(1-p)^2$	$3p^2(1-p)$	p^3.

This distribution depends on p, so $T(\underline{X})$ carries information about p. To check sufficiency, we must show that $\underline{X}$ given T does *not* carry additional information about p.
Consider $T(\underline{X}) = 2/3$. Then $\underline{X}$ could equal $(1,1,0)$, $(1,0,1)$, or $(0,1,1)$, and

$$\mathbb{P}_p\big(\underline{X} = (1,1,0) \mid T = 2/3\big) = \frac{p^2(1-p)}{3p^2(1-p)} = \frac{1}{3}.$$

The same calculation holds for the other two sample points, so

$\underline{x}$	$(1,1,0)$	$(1,0,1)$	$(0,1,1)$
$\mathbb{P}_p(\underline{X} = \underline{x} \mid T = 2/3)$	$1/3$	$1/3$	$1/3$.

These probabilities do not depend on p. The same is true on every fiber, so T is sufficient for p.

Example 3 (An insufficient statistic for a Bernoulli sample)**.** Let $X_1, X_2, X_3 \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$ and $T(\underline{X}) = X_1$. The distribution of T is

t	0	1
$\mathbb{P}_p(T = t)$	$1-p$	p.

Thus T carries some information about p, but $\underline{X} \mid T$ does too. For example,

$$\mathbb{P}_p\big(\underline{X} = (0,1,0) \mid T = 0\big) = \mathbb{P}_p(X_2 = 1, X_3 = 0) = p(1-p),$$

which depends on p. Therefore $T(\underline{X}) = X_1$ is not sufficient for p.

3.3 The Neyman–Fisher factorization theorem

Definition 1 tells us how to verify whether a proposed statistic is sufficient, but it does not tell us how to find one. The following theorem gives a constructive criterion: start with the joint pmf or pdf and collect all dependence on θ through a statistic.

Theorem 1 (Neyman–Fisher factorization theorem)**.** A statistic T is sufficient for θ in the statistical model $\{f_\theta(x) : \theta \in \Theta\}$ if and only if there exist nonnegative functions $g(t,\theta)$ and $h(\underline{x})$ such that

$$f_{\underline{X},\theta}(\underline{x}) = g\big(T(\underline{x}),\theta\big)h(\underline{x})$$

for every $\underline{x}$ and θ, where h does not depend on θ.

Proof in the discrete case. First suppose that T is sufficient. For $t = T(\underline{x})$, write

$$\begin{aligned} f_{\underline{X},\theta}(\underline{x}) &= \mathbb{P}_\theta(\underline{X} = \underline{x}) \\ &= \mathbb{P}_\theta\big(T(\underline{X}) = t\big)\mathbb{P}_\theta\big(\underline{X} = \underline{x} \mid T(\underline{X}) = t\big). \end{aligned}$$

Define

$$g(t,\theta) = \mathbb{P}_\theta\big(T(\underline{X}) = t\big)$$

and

$$h(\underline{x}) = \mathbb{P}_\theta\big(\underline{X} = \underline{x} \mid T(\underline{X}) = T(\underline{x})\big).$$

Because T is sufficient, the conditional probability defining h does not depend on θ. Hence the required factorization holds.

Conversely, suppose

$$f_{\underline{X},\theta}(\underline{x}) = g\big(T(\underline{x}),\theta\big)h(\underline{x}).$$

For a value t with positive probability and any $\underline{x} \in A_t$,

$$\begin{aligned} \mathbb{P}_\theta\big(\underline{X} = \underline{x} \mid T(\underline{X}) = t\big) &= \frac{g(t,\theta)h(\underline{x})}{\sum_{\underline{y}\in A_t} g(t,\theta)h(\underline{y})} \\ &= \frac{h(\underline{x})}{\sum_{\underline{y}\in A_t} h(\underline{y})}. \end{aligned}$$

The last expression does not depend on θ, so T is sufficient. □

Example 4 (Bernoulli model)**.** Let $X_1,\dots,X_n \overset{\text{iid}}{\sim} \mathsf{Ber}(p)$. The joint pmf is

$$\begin{aligned} f_{\underline{X},p}(\underline{x}) &= \prod_{i=1}^n \big[p^{x_i}(1-p)^{1-x_i}\mathbb{1}_{\{x_i\in\{0,1\}\}}\big] \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}\mathbb{1}_{\{\underline{x}\in\{0,1\}^n\}}. \end{aligned}$$

Taking

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

we obtain the factorization

$$g(t,p) = p^t(1-p)^{n-t}, \qquad h(\underline{x}) = \mathbb{1}_{\{\underline{x}\in\{0,1\}^n\}}.$$

Therefore T is sufficient for p. We can also verify this directly from Definition 1. Since $T \sim \mathsf{Bin}(n,p)$, for any $\underline{x} \in \{0,1\}^n$ with $\sum_i x_i = t$,

$$\begin{aligned} \mathbb{P}_p\big(\underline{X} = \underline{x} \mid T = t\big) &= \frac{p^t(1-p)^{n-t}}{\binom{n}{t} p^t (1-p)^{n-t}} \\ &= \frac{1}{\binom{n}{t}}, \end{aligned}$$

which does not depend on p.

Example 5 (Normal model)**.** Let

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \sigma^2), \qquad \theta = (\mu, \sigma^2),$$

so both μ and σ^2 are unknown. Then

$$\begin{aligned} f_{\underline{X},\theta}(\underline{x}) &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right) \right\}. \end{aligned}$$

All dependence on the sample enters through

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

Indeed, with $t = (t_1, t_2)$, take

$$g\big(t, (\mu, \sigma^2)\big) = (2\pi\sigma^2)^{-n/2} \exp\left\{ -\frac{t_2 - 2\mu t_1 + n\mu^2}{2\sigma^2} \right\}, \qquad h(\underline{x}) = 1.$$

Theorem 1 shows that T is sufficient for (μ, σ^2).

Example 6 (Uniform model)**.** Let

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta), \qquad \theta > 0.$$

Write $x_{(1)} = \min_i x_i$ and $x_{(n)} = \max_i x_i$. The joint density is

$$\begin{aligned} f_{\underline{X},\theta}(\underline{x}) &= \prod_{i=1}^n \frac{1}{\theta} \mathbb{1}_{\{0 < x_i < \theta\}} \\ &= \theta^{-n} \mathbb{1}_{\{x_{(n)} < \theta\}} \mathbb{1}_{\{x_{(1)} > 0\}}. \end{aligned}$$

Thus, for

$$T(\underline{X}) = X_{(n)},$$

we may take

$$g(t, \theta) = \theta^{-n} \mathbb{1}_{\{t < \theta\}}, \qquad h(\underline{x}) = \mathbb{1}_{\{x_{(1)} > 0\}}.$$

Therefore the sample maximum is sufficient for θ. Once the maximum is known, the locations of the remaining observations below it contain no additional information about θ.

3.4 Sufficiency and further data reduction

The full sample $T^*(\underline{X}) = \underline{X}$ is always sufficient, but it performs no data reduction. More generally, adding information to a sufficient statistic preserves sufficiency.

Theorem 2. Suppose T is sufficient for θ and T^* is another statistic. If there is a function r such that $T = r(T^*)$, then T^* is also sufficient for θ.

Remark 1. Knowing $T^* = t^*$ gives us the value of T, namely $r(t^*)$. On the other hand, knowing $T = t$ does not generally determine T^*: it could be any t^* such that $r(t^*) = t$. Thus T^* contains at least as much information about the sample as T.

Proof. Since T is sufficient, Theorem 1 gives

$$f_{\underline{X},\theta}(\underline{x}) = g\big(T(\underline{x}), \theta\big)h(\underline{x}).$$

Substituting $T = r(T^*)$ gives

$$f_{\underline{X},\theta}(\underline{x}) = g\big(r(T^*(\underline{x})), \theta\big)h(\underline{x}) = g^*\big(T^*(\underline{x}), \theta\big)h(\underline{x}),$$

where $g^*(t^*, \theta) = g(r(t^*), \theta)$. Another application of Theorem 1 shows that T^* is sufficient. □

Example 7 (Several sufficient statistics for the normal model)**.** For the model in Example 5, we have shown that

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

is sufficient. Each of the following more detailed statistics is also sufficient because T can be recovered from it.
First, let

$$T_1(\underline{X}) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n X_i^2\right).$$

Then $r_1(t_1, t_2) = (nt_1, nt_2)$.
Second, let

$$T_2(\underline{X}) = \left(\overline{X}, \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2\right).$$

Using

$$\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \overline{X}^2,$$

we recover T through

$$r_2(t_1, t_2) = \big(nt_1, n(t_2 + t_1^2)\big).$$

For $n \geq 5$, we may add completely extraneous coordinates:

$$T_3(\underline{X}) = \left(\sum_i X_i, \sum_i X_i^2, \sum_i X_i^3, X_5\right), \qquad r_3(t_1, t_2, t_3, t_4) = (t_1, t_2).$$

Finally, the entire sample is sufficient:

$$T_4(\underline{X}) = \underline{X}, \qquad r_4(\underline{x}) = \left(\sum_i x_i, \sum_i x_i^2\right).$$

Example 7 emphasizes that sufficiency alone does not identify the statistic that gives the greatest possible data reduction.

3.5 From sufficiency to minimal sufficiency

Example 7 exposes the limitation of sufficiency by itself: it does not tell us how much to reduce the data. This leaves us with a new question:

Among all sufficient statistics, which one reduces the data as far as possible?

Definition 2 (Minimal sufficient statistic)**.** A statistic T is **minimal sufficient** for θ if

1. T is sufficient for θ, and
2. T is a function of every other sufficient statistic: if T^* is sufficient, then there is a function r such that
$$T = r(T^*).$$

The second condition says that every sufficient statistic contains enough information to recover T, hence T has the *least* amount of information. Equivalently, T has the *coarsest partition into fibers* among all sufficient statistics.
To see why, note that if $T = r(T^*)$, then different values of T^* can be mapped to the same value of T. Distinct fibers of T^* can therefore collapse into a single fiber of T, as illustrated in Figure 3.3.

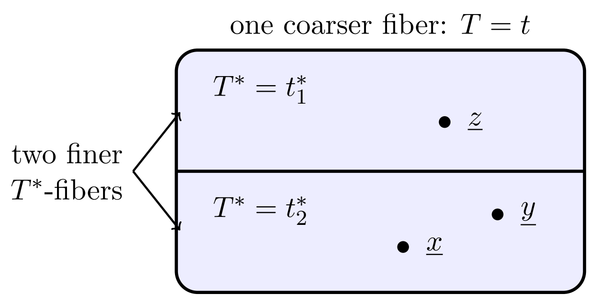


Figure 3.3: If $T = r(T^*)$ and $t = r(t_1^*) = r(t_2^*)$, then the t fiber of T contains both the t_1^* and t_2^* fibers of T^*.

3.6 The likelihood viewpoint

The definition of minimal sufficiency is not easy to check directly. In this section, we give a simpler criterion. To explain the criterion, we need to take a detour through *likelihoods*.

For a fixed value of θ, $f_{\underline{X},\theta}(\underline{x})$ is a function of the possible sample point $\underline{x}$; it describes the sampling distribution of $\underline{X}$. We can read the same expression by fixing $\underline{x}$ and allowing θ to vary.

Definition 3 (Likelihood function)**.** For a fixed sample point $\underline{x}$, the **likelihood function** is the function of θ defined by
$$L_{\underline{x}}(\theta) = f_{\underline{X},\theta}(\underline{x}), \qquad \theta \in \Theta.$$

Thus the same formula has two different readings:

$$\begin{aligned} &\text{sample distribution:} && \underline{x} \longmapsto f_{\underline{X},\theta}(\underline{x}), && \text{with } \theta \text{ fixed;} \\ &\text{likelihood function:} && \theta \longmapsto L_{\underline{x}}(\theta) = f_{\underline{X},\theta}(\underline{x}), && \text{with } \underline{x} \text{ fixed.} \end{aligned}$$

The likelihood is *not* a pmf or pdf for θ. It compares the probability of the observed sample under different possible values of θ.

Example 8 (Likelihood for a Bernoulli sample)**.** Return to the observed sample $\underline{x} = (1, 1, 0)$ in Example 1. As a function of p, its joint pmf is the likelihood

$$L_{(1,1,0)}(p) = \mathbb{P}_p(X_1 = 1, X_2 = 1, X_3 = 0) = p^2(1-p), \qquad 0 \le p \le 1.$$

What sufficiency says about likelihoods

Let S be any sufficient statistic. By Theorem 1,

$$f_{\underline{X},\theta}(\underline{x}) = g\big(S(\underline{x}), \theta\big)h(\underline{x}).$$

Take two samples $\underline{x}, \underline{y}$ in the same fiber, so that $S(\underline{x}) = S(\underline{y}) = s$. Then the g-terms are identical. At any value of θ for which the denominator below is nonzero,

$$\frac{L_{\underline{x}}(\theta)}{L_{\underline{y}}(\theta)} = \frac{f_{\underline{X},\theta}(\underline{x})}{f_{\underline{X},\theta}(\underline{y})} = \frac{h(\underline{x})}{h(\underline{y})}.$$

The right-hand side does not depend on θ. Thus the two likelihood functions have the same shape as θ varies; one is just a vertical rescaling of the other. In other words, *they are proportional to one another as functions of θ.*

Example 9 (Proportional Poisson likelihoods)**.** Suppose $X_1, X_2 \overset{\text{iid}}{\sim} \mathsf{Poisson}(\theta)$, and consider $\underline{x} = (1, 1)$ and $\underline{y} = (2, 0)$. Their likelihoods are

$$L_{\underline{x}}(\theta) = \theta^2 e^{-2\theta}, \qquad L_{\underline{y}}(\theta) = \frac{1}{2}\theta^2 e^{-2\theta}.$$

Thus $L_{\underline{x}}(\theta) = 2L_{\underline{y}}(\theta)$ for every $\theta > 0$. The two likelihoods are proportional. Figure 3.4 shows their common shape and different vertical scales.

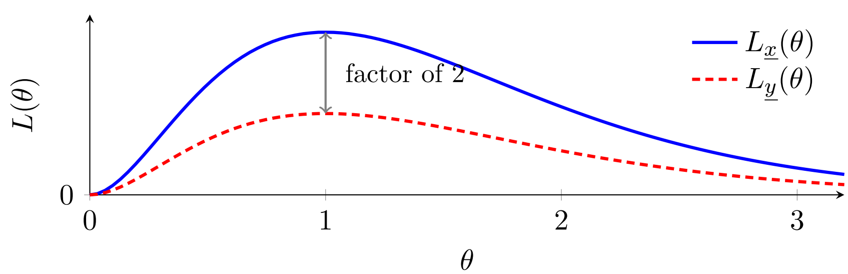


Figure 3.4: Multiplying the dashed likelihood by 2 gives the solid likelihood.

It turns out this is another equivalent characterization of sufficient statistics:

A sufficient statistic is a statistic such that all samples $\underline{x}, \underline{y}$ in the same fiber have likelihoods $L_{\underline{x}}(\theta), L_{\underline{y}}(\theta)$ that are proportional as functions of θ.

Let us define "proportional as functions of θ" a bit more carefully, in a way that avoids issues with dividing by zero.

Definition 4. Two samples $\underline{x}$ and $\underline{y}$ have **proportional likelihoods** if there is a constant $c(\underline{x}, \underline{y}) > 0$, not depending on θ, such that

$$L_{\underline{x}}(\theta) = c(\underline{x}, \underline{y})L_{\underline{y}}(\theta) \qquad \text{for every } \theta \in \Theta.$$

In particular, this means that for all θ's for which $L_{\underline{x}}(\theta) = 0$ we must have $L_{\underline{y}}(\theta) = 0$ and vice versa.Usually we can verify this definition by taking the ratio $L_{\underline{x}}(\theta)/L_{\underline{y}}(\theta)$ and seeing whether θ disappears. Comparing the functions directly is safer when a likelihood can equal zero. This happens in Example 12 below.

3.7 The Lehmann–Scheffé criterion

We now have the bridge from sufficiency to minimal sufficiency. Sample points in the same fiber of a sufficient statistic must have proportional likelihoods. But points in different fibers may also have proportional likelihoods, in which case those fibers can be merged without losing information. In Figure 3.3, for instance, $\underline{z}$ may have a likelihood proportional to those of $\underline{x}$ and $\underline{y}$. Then the two displayed T^*-fibers should be merged.
We keep merging until the resulting fibers have the property that points in distinct fibers do *not* have proportional likelihoods. These are the coarsest possible fibers we can construct, so they are the fibers of a minimal sufficient statistic.

Theorem 3 (Lehmann–Scheffé criterion for minimal sufficiency)**.** Let T be a statistic such that, for every pair of samples $\underline{x}$ and $\underline{y}$,

$$T(\underline{x}) = T(\underline{y}) \quad \text{if and only if} \quad \underline{x} \text{ and } \underline{y} \text{ have proportional likelihoods.}$$

Then T is minimal sufficient for θ.

The two directions of the equivalence do different jobs:

- If $T(\underline{x}) = T(\underline{y})$, then the likelihoods are proportional. This shows that T is *sufficient*.
- If $T(\underline{x}) \neq T(\underline{y})$, then the likelihoods are not proportional. This shows *minimality*: distinct fibers cannot be collapsed any further.

This logical distinction is the heart of the criterion.

3.8 Examples of minimal sufficiency

Example 10 (Poisson model)**.** The concrete pair in Example 9 suggests what matters for a general Poisson sample. Let

$$X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Poisson}(\theta), \qquad \theta > 0.$$

For $\underline{x} \in \{0, 1, 2, \ldots\}^n$,

$$L_{\underline{x}}(\theta) = f_{\underline{X},\theta}(\underline{x}) = \frac{\theta^{\sum_i x_i} e^{-n\theta}}{\prod_i x_i!}.$$

For two sample points $\underline{x}$ and $\underline{y}$,

$$\frac{L_{\underline{x}}(\theta)}{L_{\underline{y}}(\theta)} = \left(\prod_{i=1}^{n} \frac{y_i!}{x_i!} \right) \theta^{\sum_i x_i - \sum_i y_i}.$$

This ratio is independent of θ if and only if

$$\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i.$$

Hence, by Theorem 3, $T(\underline{X}) = \sum_{i=1}^{n} X_i$ is minimal sufficient for θ.

Example 11 (Normal model with both parameters unknown)**.** In Example 5, we proved that

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

is sufficient for (μ, σ^2). We now check that it is minimal sufficient. Let $T_1(\underline{x}) = \sum_{i=1}^n x_i$ and $T_2(\underline{x}) = \sum_{i=1}^n x_i^2$, so that $T(\underline{X}) = (T_1(\underline{X}), T_2(\underline{X}))$. Using the joint density in Example 5,

$$\begin{aligned}\frac{L_{\underline{x}}(\mu,\sigma^2)}{L_{\underline{y}}(\mu,\sigma^2)} &= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_i (x_i-\mu)^2 - \sum_i (y_i-\mu)^2\right]\right\} \\ &= \exp\left\{\frac{\mu}{\sigma^2}(T_1(\underline{x}) - T_1(\underline{y})) - \frac{1}{2\sigma^2}(T_2(\underline{x}) - T_2(\underline{y}))\right\}.\end{aligned}$$

If $T(\underline{x}) = T(\underline{y})$, then the ratio equals 1. Next, suppose $T(\underline{x}) \neq T(\underline{y})$. Considering the ratio as a function of σ^2, we can write it as $\exp(c/\sigma^2)$, where $c = \mu(T_1(\underline{x}) - T_1(\underline{y})) - \frac{1}{2}(T_2(\underline{x}) - T_2(\underline{y}))$. Now just choose any μ for which $c \neq 0$; this is possible since either $T_1(\underline{x}) \neq T_1(\underline{y})$, or $T_2(\underline{x}) \neq T_2(\underline{y})$, or both. But now if $c \neq 0$ then e^{c/σ^2} is not a constant function, and we are done. Therefore

$$\frac{L_{\underline{x}}(\mu,\sigma^2)}{L_{\underline{y}}(\mu,\sigma^2)} \text{ is parameter-free} \quad \iff \quad T(\underline{x}) = T(\underline{y}).$$

By Theorem 3, T is minimal sufficient for (μ, σ^2).

Example 12 (Uniform model: likelihoods can be zero)**.** Example 6 showed that the sample maximum is sufficient when $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathsf{Unif}(0, \theta)$. Write

$$M(\underline{x}) = x_{(n)} = \max_i x_i.$$

For a sample point $\underline{x} \in (0, \infty)^n$, the joint density from Example 6, viewed as a likelihood, is

$$L_{\underline{x}}(\theta) = \theta^{-n} \mathbb{1}_{\{M(\underline{x}) < \theta\}}.$$

Figure 3.5 shows two such likelihoods (i.e. for two samples $\underline{x}$ and $\underline{y}$) when $n = 2$, $M(\underline{x}) = 1$, and $M(\underline{y}) = 2$.

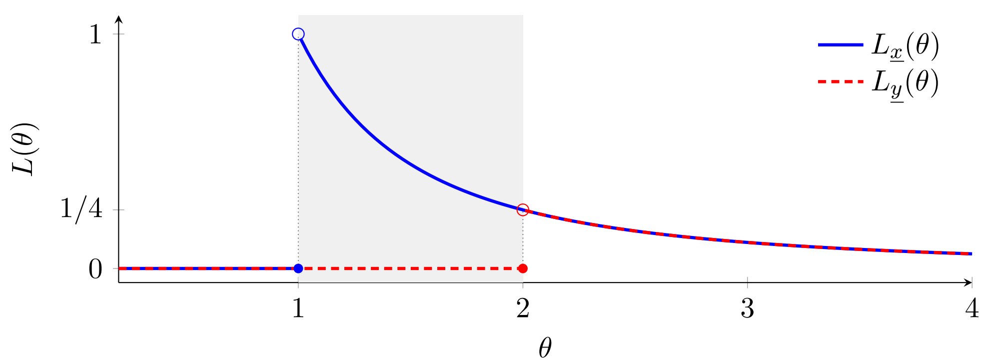


Figure 3.5: Two Uniform likelihoods when $M(\underline{x}) < M(\underline{y})$. Between the two maxima, one likelihood is positive and the other is zero. The open circles show the jumps at the maxima.

If $M(\underline{x}) = M(\underline{y})$, the two likelihood functions are identical. Next, suppose $M(\underline{x}) \neq M(\underline{y})$. Without loss of generality, take $M(\underline{x}) < M(\underline{y})$ and choose θ strictly between the two maxima. Then

$$L_{\underline{x}}(\theta) > 0, \qquad L_{\underline{y}}(\theta) = 0,$$

so the two likelihood functions cannot be proportional. Therefore

$$L_{\underline{x}} \text{ and } L_{\underline{y}} \text{ are proportional} \quad \iff \quad M(\underline{x}) = M(\underline{y}).$$

By Theorem 3, $X_{(n)}$ is minimal sufficient for θ.

This model explains why comparing likelihood functions directly is sometimes safer than always taking their ratio: the denominator may equal zero.

3.9 Uniqueness and the limits of data reduction

Minimal sufficient statistics are not literally unique. For example, knowing T is equivalent to knowing $2T + 7$. The correct statement is that they are unique up to one-to-one transformations.

Theorem 4 (One-to-one transformations)**.** If T is minimal sufficient and r is one-to-one on the range of T, then $T^* = r(T)$ is also minimal sufficient.

Proof. Because r is one-to-one, it has an inverse on its range. Thus

$$T = r^{-1}(T^*),$$

and since T is sufficient, so is T^*. Now we show T^* is minimal. Let S be any sufficient statistic. Since T is minimal sufficient, $T = q(S)$ for some function q. Therefore

$$T^* = r(T) = r\big(q(S)\big),$$

so T^* is a function of every sufficient statistic. Hence T^* is minimal sufficient. □

Theorem 4 means that the important thing is how the statistic groups sample points, not the labels given to those groups. Replacing T by a one-to-one transformation changes the labels but does not change which sample points are grouped together.
Minimal sufficiency also does not guarantee a dramatic reduction in dimension.

Example 13 (Cauchy location model)**.** Let

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} \mathsf{Cauchy}(\theta, 1),$$

with density

$$f_\theta(x) = \frac{1}{\pi\big[1 + (x - \theta)^2\big]}.$$

One can show that the vector of order statistics

$$T(\underline{X}) = \big(X_{(1)}, X_{(2)}, \dots, X_{(n)}\big)$$

is minimal sufficient. These are the n values X_i arranged from smallest to largest. The only information being thrown out is the ordering of the data: if $\underline{y}$ is a permutation of $\underline{x}$, then they both have the same ordering from smallest to largest.
This is an interesting case in which we cannot throw out all but a low-dimensional statistic. The minimal sufficient statistic is n-dimensional.

3.10 Summary

The full progression is

data reduction: replace the sample by a statistic

$\Downarrow$

sufficiency: retain all information about θ

$\Downarrow$

minimal sufficiency: reduce as far as sufficiency allows

$\Downarrow$

Lehmann–Scheffé criterion: merge exactly proportional likelihoods.

For sufficiency calculations, write the joint pmf or pdf and use Theorem 1 to collect all dependence on θ through a statistic T.

For minimal sufficiency calculations, view the joint pmf or pdf as the likelihood $L_{\underline{x}}(\theta)$ and compare the likelihoods of two arbitrary sample points $\underline{x}$ and $\underline{y}$. Determine exactly when they are proportional as functions of θ, then define T so that $T(\underline{x}) = T(\underline{y})$ exactly in that case.

The most important logical distinction in Theorem 3 is that one direction proves that T is sufficient, while the other proves that it is the *coarsest* sufficient statistic.