# Principles of Data Reduction

Statistical Theory

Guillaume Dehaene
Ecole Polytechnique Fédérale de Lausanne

# Statistical Models and The Problem of Inference

Recall our setup:

- Collection of r.v.'s (a random vector) $\mathbf{X} = (X_1, ..., X_n)$
- $\mathbf{X} \sim F_\theta \in \mathcal{F}$
- $\mathcal{F}$ a parametric class with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$

**The Problem of Point Estimation**

1. Assume that $F_\theta$ is known up to the parameter $\theta$ which is unknown
2. Let $(x_1, ..., x_n)$ be a realization of $\mathbf{X} \sim F_\theta$ which is available to us
3. Estimate the value of $\theta$ that generated the sample given $(x_1, ..., x_n)$

The only guide (apart from knowledge of $\mathcal{F}$) at hand is the data:

$\hookrightarrow$ Anything we "do" will be a function of the data $g(x_1, ..., x_n)$

$\hookrightarrow$ Need to study properties of such functions and information loss incurred (any function of $(x_1, .., x_n)$ will carry at most the same information but usually less)

# The data-processing inequality

Key idea: whatever we do with the data, it can't increase our information.

Only new data brings new information.

By transforming the data / projecting it down onto the value of a statistic, at best we preserve the information that is in the data.

# Statistics of the data

# Statistics

## Definition (Statistic)

Let $\mathbf{X}$ be a random sample from $F_\theta$. A *statistic* is a (measurable) function $T$ that maps $\mathbf{X}$ into $\mathbb{R}^d$ and does not depend on $\theta$.

$\hookrightarrow$ Intuitively, any function of the sample alone is a statistic.

$\hookrightarrow$ Any statistic is itself a r.v. with its own distribution.

## Example

$T(\mathbf{X}) = n^{-1}\sum_{i=1}^n X_i$ is a statistic (since $n$, the sample size, is known).

## Example

$T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})$ where $X_{(1)} \leq X_{(2)} \leq \dots X_{(n)}$ are the order statistics of $\mathbf{X}$. Since $T$ depends only on the values of $\mathbf{X}$, $T$ is a statistic.

## Example

Let $T(\mathbf{X}) = c$, where $c$ is a known constant. Then $T$ is a statistic

# Ancillarity

# Statistics and Information About $\theta$

- Evident from previous examples: some statistics are more informative and others are less informative regarding the true value of $\theta$
- Any $T(\mathbf{X})$ that is not "1-1" carries less information about $\theta$ than $\mathbf{X}$
- Which are "good" and which are "bad" statistics?

**Definition (Ancillary Statistic)**

A statistic $T$ is an *ancillary statistic* (for $\theta$) if its distribution does not functionally depend $\theta$

$\hookrightarrow$ So an ancillary statistic has the same distribution $\forall\ \theta \in \Theta$.

# Ancillarity example

### Example

Suppose that $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, 1)$ (only the mean $\mu$ is unknown).

Let $T(X_1, ..., X_n) = X_1 - X_2$.

Then $T$ has a Normal distribution with mean 0 and variance 2. Thus $T$ is ancillary for the unknown parameter $\mu$. If both $\mu$ and $\sigma^2$ were unknown, $T$ would not be ancillary for $\theta = (\mu, \sigma^2)$.

# Statistics and Information about $\theta$

- If $T$ is ancillary for $\theta$ then $T$ contains no information about $\theta$
- In order to contain any useful information about $\theta$, the dist($T$) must depend explicitly on $\theta$.
- Intuitively, the amount of information $T$ gives on $\theta$ increases as the dependence of dist($T$) on $\theta$ increases

## Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$, $S = \min(X_1, \ldots, X_n)$ and $T = \max(X_1, \ldots, X_n)$.

- $f_S(x; \theta) = \frac{n}{\theta}\left(1 - \frac{x}{\theta}\right)^{n-1}, \quad 0 \leq x \leq \theta$
- $f_T(x; \theta) = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1}, \quad 0 \leq x \leq \theta$

$\hookrightarrow$ Neither $S$ nor $T$ are ancillary for $\theta$

$\hookrightarrow$ As $n \uparrow \infty$, $f_S$ becomes concentrated around 0

$\hookrightarrow$ As $n \uparrow \infty$, $f_T$ becomes concentrated around $\theta$ while

$\hookrightarrow$ Indicates that $T$ provides more information about $\theta$ than does $S$.

# Sufficiency

# Statistics and Information about $\theta$

- $\mathbf{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} F_\theta$ and $T(\mathbf{X})$ a statistic.

- The *fibres* or *level sets* or *contours* of $T$ are the sets

$$A_t = \{\mathbf{x} \in \mathbb{R}^n : T(\mathbf{x}) = t\}.$$

(all potential samples that could have given me the value $t$ for $T$)

$\hookrightarrow$ $T$ is constant when restricted to a fibre.

- Any realization of $\mathbf{X}$ that falls in a given fibre is equivalent as far as $T$ is concerned
- Any inference drawn through $T$ will be the same within fibres.
- Look at the dist($\mathbf{X}$) on an fibre $A_t$: $f_{\mathbf{X}|T=t}(\mathbf{x})$

# Statistics and Information about $\theta$

- Suppose $f_{\mathbf{X}|T=t}$ changes depending on $\theta$: we are losing information.
- Suppose $f_{\mathbf{X}|T=t}$ is functionally independent of $\theta$

$\Longrightarrow$ Then $\mathbf{X}$ contains no information about $\theta$ on the set $A_t$
$\Longrightarrow$ In other words, $\mathbf{X}$ is ancillary for $\theta$ on $A_t$

- If this is true for each $t \in \text{Range}(T)$ then $T(\mathbf{X})$ contains the same information about $\theta$ as $\mathbf{X}$ does.
  - $\hookrightarrow$ It does not matter whether we observe $\mathbf{X} = (X_1, ..., X_n)$ or just $T(\mathbf{X})$.
  - $\hookrightarrow$ Knowing the exact value $\mathbf{X}$ in addition to knowing $T(\mathbf{X})$ does not give us any additional information - $\mathbf{X}$ is irrelevant if we already know $T(\mathbf{X})$.

**Definition (Sufficient Statistic)**

A statistic $T = T(\mathbf{X})$ is said to be *sufficient* for the parameter $\theta$ if for all (Borel) sets $B$ the probability $\mathbb{P}[\mathbf{X} \in B | T(\mathbf{X}) = t]$ does not depend on $\theta$.

# Sufficient Statistics

## Example (Bernoulli Trials)

Let $X_1, \dots, X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$ and $T(\mathbf{X}) = \sum_{i=1}^n X_i$. Given $\mathbf{x} \in \{0,1\}^n$,

$$\begin{aligned}
\mathbb{P}[\mathbf{X} = \mathbf{x} | T = t] &= \frac{\mathbb{P}[\mathbf{X} = \mathbf{x}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[\mathbf{X} = \mathbf{x}]}{\mathbb{P}[T = t]} \mathbf{1}\{\Sigma_{i=1}^n x_i = t\} \\
&= \frac{\theta^{\Sigma_{i=1}^n x_i}(1-\theta)^{n - \Sigma_{i=1}^n x_i}}{\binom{n}{t}\theta^t (1-\theta)^{n-t}} \mathbf{1}\{\Sigma_{i=1}^n x_i = t\} \\
&= \frac{\theta^t (1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1}.
\end{aligned}$$

- $T$ is sufficient for $\theta \to$ Given # of tosses that came heads, knowing *which tosses* came heads is irrelevant in deciding if the coin is fair:

0 0 1 1 1 0 1 VS 1 0 0 0 1 1 1 VS 1 0 1 0 1 0 1

## Sufficient Statistics

- Definition hard to verify (especially for continuous variables)
- Definition does not allow easy identification of sufficient statistics

### Theorem (Fisher-Neyman Factorization Theorem)

*Suppose that* $\mathbf{X} = (X_1, \ldots, X_n)$ *has a joint density or frequency function* $f(\mathbf{x}; \theta)$, $\theta \in \Theta$. *A statistic* $T = T(\mathbf{X})$ *is sufficient for* $\theta$ *if and only if*

$$f(\mathbf{x}; \theta) = g(T(\mathbf{x}), \theta) h(\mathbf{x}).$$

### Example

Let $X_1, ..., X_n \overset{iid}{\sim} \mathcal{U}[0, \theta]$ with pdf $f(x; \theta) = \mathbf{1}\{x \in [0, \theta]\}/\theta$. Then,

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\theta^n} \mathbf{1}\{\mathbf{x} \in [0, \theta]^n\} = \frac{\mathbf{1}\{\max[x_1, ..., x_n] \leq \theta\} \mathbf{1}\{\min[x_1, ..., x_n] \geq 0\}}{\theta^n}$$

Therefore $T(\mathbf{X}) = X_{(n)} = \max[X_1, ..., X_n]$ is sufficient for $\theta$.

## Proof of Neyman-Fisher Theorem - Discrete Case.

Suppose first that $T$ is sufficient. Then

$$
\begin{aligned}
f(x;\theta) &= \mathbb{P}[\mathbf{X}=\mathbf{x}] = \sum_t \mathbb{P}[\mathbf{X}=\mathbf{x}, T=t] \\
&= \mathbb{P}[\mathbf{X}=\mathbf{x}, T=T(\mathbf{x})] = \mathbb{P}[T=T(\mathbf{x})]\mathbb{P}[\mathbf{X}=\mathbf{x}|T=T(\mathbf{x})]
\end{aligned}
$$

Since T is sufficient, $\mathbb{P}[\mathbf{X}=\mathbf{x}|T=T(\mathbf{x})]$ is independent of $\theta$ and so $f(x;\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$.
Now suppose that $f(x;\theta) = g(T(\mathbf{x});\theta)h(\mathbf{x})$. Then if $T(\mathbf{x}) = t$,

$$
\begin{aligned}
\mathbb{P}[\mathbf{X}=\mathbf{x}|T=t] &= \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}, T=t]}{\mathbb{P}[T=t]} = \frac{\mathbb{P}[\mathbf{X}=\mathbf{x}]}{\mathbb{P}[T=t]}\mathbf{1}\{T(\mathbf{x})=t\} \\
&= \frac{g(T(\mathbf{x});\theta)h(\mathbf{x})\mathbf{1}\{T(\mathbf{x})=t\}}{\sum_{\mathbf{y}:T(\mathbf{y})=t} g(T(\mathbf{y});\theta)h(\mathbf{y})} = \frac{h(\mathbf{x})\mathbf{1}\{T(\mathbf{x})=t\}}{\sum_{T(\mathbf{y})=t} h(\mathbf{y})}.
\end{aligned}
$$

which does not depend on $\theta$. $\square$

# Minimal Sufficiency

# Minimally Sufficient Statistics

- Saw that sufficient statistic keeps what is important and leaves out irrelevant information.
- How much info can we throw away? Is there a "smallest" sufficient statistic?

**Definition (Minimally Sufficient Statistic)**

A statistic $T = T(\mathbf{X})$ is said to be *minimally sufficient* for the parameter $\theta$ if it is sufficient for $\theta$ and for any other sufficient statistic $S = S(\mathbf{X})$ there exists a function $g(\cdot)$ with

$$T(\mathbf{X}) = g(S(\mathbf{X})).$$

**Lemma**

*If $T$ and $S$ are minimaly sufficient statistics for a parameter $\theta$, then there exists injective functions $g$ and $h$ such that $S = g(T)$ and $T = h(S)$.*

## Theorem

*Let* $\mathbf{X} = (X_1, ..., X_n)$ *have joint density or frequency function* $f(\mathbf{x}; \theta)$ *and* $T = T(\mathbf{X})$ *be a statistic. Suppose that* $f(\mathbf{x}; \theta)/f(\mathbf{y}; \theta)$ *is independent of* $\theta$ *if and only if* $T(\mathbf{x}) = T(\mathbf{y})$*. Then* $T$ *is minimally sufficient for* $\theta$*.*

## Proof.

Assume for simplicity that $f(\mathbf{x}; \theta) > 0$ for all $\mathbf{x} \in \mathbb{R}^n$ and $\theta \in \Theta$.
[sufficiency part] Let $\mathcal{T} = \{T(\mathbf{y}) : y \in \mathbb{R}^n\}$ be the image of $\mathbb{R}^n$ under $T$ and let $A_t$ be the level sets of $T$. For each $t$, choose a representative element $\mathbf{y}_t \in A_t$. Notice that for any $\mathbf{x}$, $\mathbf{y}_{T(\mathbf{x})}$ is in the same level set as $\mathbf{x}$, so that

$$f(\mathbf{x}; \theta)/f(\mathbf{y}_{T(\mathbf{x})}; \theta)$$

does not depend on $\theta$ by assumption. Let $g(t, \theta) := f(\mathbf{y}_t; \theta)$ and notice

$$f(\mathbf{x}; \theta) = \frac{f(\mathbf{y}_{T(\mathbf{x})}; \theta) f(\mathbf{x}; \theta)}{f(\mathbf{y}_{T(\mathbf{x})}; \theta)} = g(T(\mathbf{x}), \theta) h(\mathbf{x})$$

and the claim follows from the factorization theorem.

[minimality part] Suppose that $T'$ is another sufficient statistic. By the factorization thm: $\exists g', h' : \quad f(\mathbf{x};\theta) = g'(T'(\mathbf{x});\theta)h'(\mathbf{x})$.
Let $\mathbf{x}, \mathbf{y}$ be such that $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

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

Since ratio does not depend on $\theta$, we have by assumption $T(\mathbf{x}) = T(\mathbf{y})$. Hence $T$ is a function of $T'$; so is minimal by arbitrary choice of $T'$ because the fibres of T' are subsets of the fibres of T. $\square$

### Example (Bernoulli Trials)

Let $X_1, ..., X_n \overset{iid}{\sim} \text{Bernoulli}(\theta)$. Let $\mathbf{x}, \mathbf{y} \in \{0, 1\}^n$ be two possible outcomes. Then

$$\frac{f(\mathbf{x}; \theta)}{f(\mathbf{y}; \theta)} = \frac{\theta^{\Sigma x_i}(1-\theta)^{n-\Sigma x_i}}{\theta^{\Sigma y_i}(1-\theta)^{n-\Sigma y_i}}$$

which is constant if and only if $T(\mathbf{x}) = \sum x_i = \sum y_i = T(\mathbf{y})$, so that $T$ is minimally sufficient.

### Exercise

Prove that the likelihood $f(\mathbf{X}; \theta)$ (which is a **random function**) is a sufficient statistic.

Let $\theta_0$ be some arbitrary value such that $\forall \mathbf{X} : f(\mathbf{X}; \theta_0) \neq 0$. Prove that the normalized likelihood: $\frac{f(\mathbf{X};\theta)}{f(\mathbf{X};\theta_0)}$ is minimally sufficient.

This exercise shows that a "minimal" statistic can be quite big.

# Completeness

# Complete Statistics

- Ancillary Statistic $\rightarrow$ Contains no info on $\theta$
- Minimally Sufficient Statistic $\rightarrow$ Contains all relevant info and as little irrelevant as possible.
- Should they be mutually independent?

**Definition (Complete Statistic)**

Let $\{g(t;\theta):\theta\in\Theta\}$ be a family of densities (or frequencies) corresponding to a statistic $T(\mathbf{X})$. The statistic $T$ is called *complete* if given any measurable function $h$, the following implication holds

$$\int h(t)g(t;\theta)dt=0 \quad \forall\theta\in\Theta \implies \mathbb{P}[h(T)=0]=1 \quad \forall\theta\in\Theta.$$

Not clear why term "complete" was chosen – one reason might be the resemblance to the notion of *complete system* in a Hilbert space (whose orthogonal complement is the zero space), in reference to $\{g(\cdot;\theta)\}_{\theta\in\Theta}$.

# Complete Statistics

## Example (Bernoulli Trials)

Let $X_1, ..., X_n \stackrel{iid}{\sim} Bern(\theta)$, $\theta \in (0,1)$, and $T = \sum X_i$. Let $h$ be arbitrary.

$$\mathbb{E}[h(T)] = \sum_{t=0}^{n} h(t)\binom{n}{t}\theta^t(1-\theta)^{n-t} = (1-\theta)^n \sum_{t=0}^{n} h(t)\binom{n}{t}\left(\frac{\theta}{1-\theta}\right)^t$$

As $\theta$ ranges in $(0,1)$, the ratio $\theta/(1-\theta)$ ranges in $(0,\infty)$. Thus, assuming $\mathbb{E}[h(T)] = 0$ for all $\theta \in (0,1)$ implies that

$$P(x) = \sum_{t=0}^{n} h(t)\binom{n}{t}x^t = 0 \qquad \forall\, x > 0,$$

i.e. the polynomial $P(x)$ is uniformly zero over the entire positive reals. Hence, its coefficients must be all zero, so $g(t) = 0$, $t = 1, ..., n$. Hence $\mathbb{P}[h(T) = 0] = 1$ for all $\theta \in (0,\infty)$.

# Complete Statistics

$\hookrightarrow$ Why is completeness relevant to data reduction?

**Lemma**

*If $T$ is complete, then $h(T)$ is ancillary for $\theta$ if and only if $h(T) = c$ a.s.*

**Proof.**

One direction is obvious. For the other, let $h(T)$ be ancillary. Then its distribution does not depend on $\theta$. Hence $\mathbb{E}[h(T)] = c$, for some constant $c$, regardless of $\theta$. Equivalently, $\mathbb{E}[h(T) - c] = 0$ for all $\theta$. By completeness of $T$, $\mathbb{P}[h(T) = c] = 1$. $\square$

- (equivalently: only trivial (=constant) functions of $T$ are ancillary)
- In other words, a complete statistic contains no ancillary information
- Contrast to a sufficient statistic:
  - A sufficient statistic keeps all the relevant information
  - A complete statistic throws away all the irrelevant information

# Complete Statistics

**Theorem (Basu's Theorem)**

*A complete sufficient statistic is independent of every ancillary statistic.*

**Proof.**

We consider the discrete case only. It suffices to show that,

$$\mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t] = \mathbb{P}[S(\mathbf{X}) = s]$$

$$\text{Define: } h(t) = \mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t] - \mathbb{P}[S(\mathbf{X}) = s]$$

and observe that:

1. $\mathbb{P}[S(\mathbf{x}) = s]$ does not depend on $\theta$ (ancillarity)
2. $\mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t] = \mathbb{P}[\mathbf{X} \in \{\mathbf{x} : S(\mathbf{x}) = s\} | T = t]$ does not depend on $\theta$ (sufficiency)

and so $h$ does not depend on $\theta$.

Therefore, for any $\theta \in \Theta$,

$$\begin{aligned}
\mathbb{E}h(T) &= \sum_t (\mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t] - \mathbb{P}[S(\mathbf{X}) = s])\mathbb{P}[T(\mathbf{X}) = t] \\
&= \sum_t \mathbb{P}[S(\mathbf{X}) = s | T(\mathbf{X}) = t]\mathbb{P}[T(\mathbf{X}) = t] + \\
&\qquad + \mathbb{P}[S(\mathbf{X}) = s] \sum_t \mathbb{P}[T(\mathbf{X}) = t] \\
&= \mathbb{P}[S(\mathbf{X}) = s] - \mathbb{P}[S(\mathbf{X}) = s] = 0.
\end{aligned}$$

But $T$ is complete so it follows that $h(t) = 0$ for all $t$. QED. $\square$

Basu's Theorem is useful for deducing independence of two statistics:

- No need to determine their joint distribution
- Needs showing completeness (usually hard analytical problem)
- Will see models in which completeness is easy to check

# Completeness and Minimal Sufficiency

## Theorem (Lehmann-Scheffé)

*Let $\mathbf{X}$ have density $f(\mathbf{x};\theta)$. If $T(\mathbf{X})$ is sufficient and complete for $\theta$ then $T$ is minimally sufficient.*

## Proof.

First of all we show that a minimally sufficient statistic exists. Define an equivalence relation as $\mathbf{x} \equiv \mathbf{x}'$ if and only if $f(\mathbf{x};\theta)/f(\mathbf{x}';\theta)$ is independent of $\theta$. If $S$ is any function such that $S = c$ on these equivalent classes, then $S$ is a minimally sufficient, establishing existence (rigorous proof by Lehmann-Scheffé (1950) to assure $S$ measurably constructible). Therefore, it must be the case that $S = g_1(T)$, for some $g_1$. Let $g_2(S) = \mathbb{E}[T|S]$ (does not depend on $\theta$ since $S$ sufficient). Consider:

$$g(T) = T - g_2(S)$$

Write $\mathbb{E}[g(T)] = \mathbb{E}[T] - \mathbb{E}\left\{\mathbb{E}[T|S]\right\} = \mathbb{E}T - \mathbb{E}T = 0$ for all $\theta$.

**(proof cont'd).**

By completeness of $T$, it follows that $g_2(S) = T$ a.s. In fact, $g_2$ has to be injective, or otherwise we would contradict minimal sufficiency of $S$. But then $T$ is 1-1 a function of $S$ and $S$ is a $1-1$ function of $T$. Invoking our previous lemma proves that $T$ is minimally sufficient. $\square$

# Sufficiency and completeness

The log-likelihood is minimally sufficient (if normalized), but not necessarily complete !

## Exercise

Consider the following situation:

- We pick a random number $\mathbb{N} \ni N \sim F_n$
- We gather $N$ IID Gaussian samples $X_1 \dots X_N \sim \mathcal{N}(\mu, 1)$.

1. Write down the normalized log-likelihood function $\mu \to LL(\mu) - LL(0)$ as a function of $N, \mathbf{X}$. This is a **function valued random variable**.
2. Prove that it is minimally sufficient.
   (Note that the log-likelihood $\mu \to LL(\mu)$ is only sufficient, not minimally sufficient)
3. Prove that it is not complete.

# Summary

We looked at how to "summarize" the data by computing the value of a statistic $S(\mathbf{X})$:

- Ancillary: $S$ carries no information.
- Sufficient: $S$ doesn't lose information.
- Minimally sufficient: $S$ doesn't lose information and carries as little ancillary information as possible.
- Complete: $S$ carries no ancillary information.

Most of the time, a minimally sufficient statistic exists: the normalized log-likelihood.
A complete sufficient statistic might not exist.