Principles of Data Reduction

Statistical Theory

Victor Panaretos
Ecole Polytechnique Fédérale de Lausanne

Statistical Models and The Problem of Inference

Recall our setup:

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

The Problem of Point Estimation

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

Statistics

Definition (Statistic)

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

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

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

Example

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

Example

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

Example

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

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

Definition (Ancillary Statistic)

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

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

Example

Suppose that $X_1, ..., X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where μ unknown but σ^2 known). Let $T(X_1, ..., X_n) = X_1 - X_2$; then T has a Normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for the unknown parameter μ. If both μ and σ^2 were unknown, T would not be ancillary for $\theta = (\mu, \sigma^2)$.

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

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

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

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

- $X = (X_1, \ldots, X_n) \overset{iid}{\sim} F_\theta$ and $T(X)$ a statistic.
- The *fibres* or *level sets* or *contours* of T are the sets

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

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

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

- Any realization of 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(X) on an fibre A_t: $f_{X|T=t}(x)$

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

$\Longrightarrow$ Then X contains no information about θ on the set A_t

$\Longrightarrow$ In other words, X is ancillary for θ on A_t

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

Definition (Sufficient Statistic)

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

Example (Bernoulli Trials)

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

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

- 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

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

Theorem (Fisher-Neyman Factorization Theorem)

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

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

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,

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

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

Proof of Neyman-Fisher Theorem - Discrete Case.

Suppose first that T is sufficient. Then

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

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

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

which does not depend on θ. □

- Saw that sufficient statistic keeps what is important and leaves out irrelevant information.
- How much info can we through away? Is there a “necessary” statistic?

Definition (Minimally Sufficient Statistic)

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

$$T(X) = g(S(X)).$$

Lemma

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

Theorem

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

Proof.

Assume for simplicity that $f(x;\theta) > 0$ for all $x \in \mathbb{R}^n$ and $\theta \in \Theta$. [sufficiency part] Let $\mathcal{T} = \{T(y) : y \in \mathbb{R}^n\}$ 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 $y_t \in A_t$. Notice that for any x, $y_{T(x)}$ is in the same level set as x, so that

$$f(x;\theta)/f(y_{T(x)};\theta)$$

does not depend on θ by assumption. Let $g(t,\theta) := f(y_t;\theta)$ and notice

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

and the claim follows from the factorization theorem.

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

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

Since ratio does not depend on θ, we have by assumption $T(x) = T(y)$. Hence T is a function of T'; so is minimal by arbitrary choice of T'. $\square$

Example (Bernoulli Trials)

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

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

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

Complete Statistics

- Ancillary Statistic $\to$ Contains no info on θ
- Minimally Sufficient Statistic $\to$ 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 $T(X)$. 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 $X_1, \dots, X_n \overset{iid}{\sim} Bern(\theta)$, $\theta \in (0,1)$, 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 θ 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 $g(t) = 0$, $t = 1, \dots, n$. Hence $\mathbb{P}[h(T) = 0] = 1$ for all $\theta \in (0,\infty)$.

Complete Statistics

↪ Why is completeness relevant to data reduction?

Lemma

If T is complete, then $h(T)$ is ancillary for θ 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 θ. Hence $\mathbb{E}[h(T)] = c$, for some constant c, regardless of θ. Equivalently, $\mathbb{E}[h(T) - c] = 0$ for all θ. By completeness of T, $\mathbb{P}[h(T) = c] = 1$. □

- (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,

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

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

and observe that:

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

and so h does not depend on θ.

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

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

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

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 X have density $f(x;\theta)$. If $T(X)$ is sufficient and complete for θ then T is minimally sufficient.

Proof.

First of all we show that a minimally sufficient statistic exists. Define an equivalence relation as $x \equiv x'$ if and only if $f(x;\theta)/f(x';\theta)$ is independent of θ. 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 θ since S sufficient). Consider:

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

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

(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. $\square$

One can also prove:

Theorem

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