STAT 713 sp 2023 Lec 01 slides

Data reduction part 1: Sufficiency

Karl B. Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.

Setup and notation

- Let $\mathbf{X} = (X_1, \ldots, X_n)$ represent the set of rvs from an experiment.
- Use $\mathbf{x} = (x_1, \ldots, x_n)$ to represent a specific set of values for the rvs in $\mathbf{X}$.
- Let $f(\mathbf{x}; \theta)$ or $p(\mathbf{x}; \theta)$ denote the joint pdf or pmf of the rvs in $\mathbf{X}$, resp.
- The distribution of $\mathbf{X}$ depends on a parameter (or some parameters) $\theta \in \Theta$.
- A function $T(\mathbf{X})$ of the rvs in $\mathbf{X}$ is called a *statistic*.

Goal: Learn about θ from a realization of $\mathbf{X}$ via the value of a statistic $T(\mathbf{X})$.

Key concepts of data reduction in the rough: $\underset{\sim}{X} = (\underline{X_1, \ldots, X_n}) \to T(\underset{\sim}{X})$

1. A *sufficient statistic* carries all the information about θ from $\mathbf{X}$.
2. A *minimal sufficient statistic* carries the above and no more than this.
3. An *ancillary statistic* carries no information about θ.
4. A *complete statistic* cannot be used to construct an unbiased estimator of 0.

We think of computing a statistic $T(\mathbf{X})$ as "reducing" or summarizing the data.

Example: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Exponential}(\lambda)$, $\lambda > 0$, and consider the statistics

$$T_1(\mathbf{X}) = X_{(1)}, \qquad T_2(\mathbf{X}) = \bar{X}_n, \qquad T_3(\mathbf{X}) = S_n^2,$$

$$T_4(\mathbf{X}) = (\textstyle\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2), \quad T_5(\mathbf{X}) = X_{(1)}/X_{(n)}, \quad T_6(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)}).$$

Sufficiency, *minimality*, *ancillarity*, and *completeness* address (resp) the questions

1. Which reductions of the data do not discard any information about λ?
2. Which ones keep all relevant information about λ, but discard all else?
3. Which ones discard all information about λ?
4. Which ones cannot be used to construct an unbiased estimator of 0?

Sufficient statistic

A statistic $T(\mathbf{X})$ is *sufficient for* θ if the joint pdf/pmf of $\mathbf{X}$, conditional on the value of $T(\mathbf{X})$, does not depend on θ.

Affect of θ on the distribution of $\mathbf{X}$ is expressed fully in the value $T(\mathbf{X})$.

Two samples $\mathbf{X}_1, \mathbf{X}_2$ carry same info about θ if $T(\mathbf{X}_1) = T(\mathbf{X}_2)$.

Example: For the Bernoulli(p) distribution, it seems like the random samples

$$\mathbf{X}_1 = (0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1)$$
$$\mathbf{X}_2 = (1, 0, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0)$$

$T(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$

should lead to the same inferences about $p \in (0, 1)$. Why?

all possible data sets

More notation

rv with support $\mathcal{T}$

- Denote by $\mathcal{X}$ the support of $\mathbf{X}$.
- Let $\mathcal{T} = \{t : T(\mathbf{x}) = t \text{ for some } \mathbf{x} \in \mathcal{X}\}$ be the support of $T(\mathbf{X})$.
- Use subscript θ in $\mathbb{E}_\theta$ and P_θ to indicate dependence on θ.

For the discrete case, if $T(\mathbf{X})$ is a sufficient statistic, then

$$P(\mathbf{X} = \mathbf{x} | T(\mathbf{X}) = t)$$

is free of θ for every $t \in \mathcal{T}$ and $\mathbf{x} \in \mathcal{X}$.

Definition of sufficiency for discrete $\underset{\sim}{X}$.

Theorem (Checking for sufficiency, cf. Thm 6.2.2 in CB)

For $\mathbf{X}$ *with joint pdf/pmf* $f(\mathbf{x};\theta)$, $T(\mathbf{X})$ *is a sufficient statistic for* θ *if*

$$\frac{f(\mathbf{x};\theta)}{f_T(T(\mathbf{x});\theta)} \left(= \frac{P_\theta(\underset{\sim}{X} = x)}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))} \text{ for discrete} \right)$$

is free of θ *for all* $\mathbf{x} \in \mathcal{X}$, *where* f_T *is the pdf/pmf of* $T = T(\mathbf{X})$.

Exercise: Prove for the discrete case.

Sufficiency of $T(\underset{\sim}{X})$ for θ means

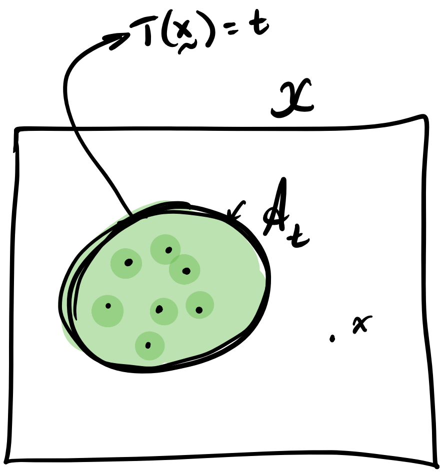
$$P(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) \quad \text{is free of } \theta.$$

Write

$$P(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) = \frac{P(\underset{\sim}{X} = \underset{\sim}{x} \wedge T(\underset{\sim}{X}) = t)}{P(T(\underset{\sim}{X}) = t)}$$

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

$$= P(T(\underset{\sim}{X}) = t \mid \underset{\sim}{X} = \underset{\sim}{x}) \, \frac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = t)}$$



free of θ

$$= \begin{cases} 0, & \underset{\sim}{x} \notin A_t \\ \boxed{\dfrac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))}} & \underset{\sim}{x} \in A_t \end{cases}$$

By assumption is free of θ.

$$A_t = \{ \underset{\sim}{x} \in \mathcal{X} : T(\underset{\sim}{x}) = t \}$$

Note: If $\underset{\sim}{x} \in A_t$, then $T(\underset{\sim}{x}) = t$.

Check if $\dfrac{p(\underset{\sim}{x};\lambda)}{p_T(T(\underset{\sim}{x});\lambda)}$ is free of λ, where p_T is the pmf of $T(\underset{\sim}{X})$.

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \quad x = 0,1,2,\ldots$$

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Poisson}(\lambda)$, where $\lambda > 0$.

1. Check whether $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for λ...
2. Interpret.

$$p(\underset{\sim}{x};\lambda) = \prod_{i=1}^n \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} = \underline{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\prod_{i=1}^n \frac{1}{x_i!}}$$

Find pmf of $T(\underset{\sim}{X})$.

$$M_{T(\underset{\sim}{X})}(t) = M_{\Sigma X_i}(t) = \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n e^{\lambda(e^t-1)} = e^{n\lambda(e^t-1)},$$

so $T(\underset{\sim}{X}) \sim \text{Poisson}(n\lambda)$.

So $\quad p_T(t;\lambda) = \dfrac{e^{-n\lambda}(n\lambda)^t}{t!}$

Now write

$$\frac{p(\underset{\sim}{x};\lambda)}{p_T(T(\underset{\sim}{x});\lambda)} = \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}\prod_{i=1}^n \frac{1}{x_i!}}{e^{-n\lambda}(n\lambda)^{T(\underset{\sim}{x})} / T(\underset{\sim}{x})!}$$

$$= \frac{\cancel{e^{-n\lambda}}\,\cancel{\lambda^{\sum x_i}}\prod_{i=1}^n \frac{1}{x_i!}}{\cancel{e^{-n\lambda}}(n\cancel{\lambda})^{\sum x_i} / (\sum x_i)!}$$

$$= \frac{\prod_{i=1}^n \frac{1}{x_i!}}{n^{\sum x_i} / (\sum x_i)!}.$$

Since λ canceled out, $\quad T(\underset{\sim}{x}) = \sum_{i=1}^n x_i$ is a suff. stat. for λ.

Remember: Check if $\dfrac{f(\underset{\sim}{x};\theta)}{f_T(T(\underset{\sim}{x});\theta)}$ is free of θ, where $f_T(t;\theta)$ is pdf of $T(\underset{\sim}{X}) = X_{(n)}$.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Uniform}(0, \theta)$, where $\theta > 0$.

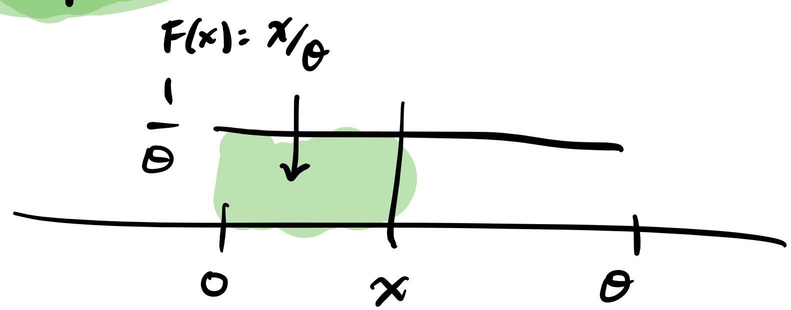
Check whether $T(\mathbf{X}) = X_{(n)}$ is a sufficient statistic for θ.

$$f(x;\theta) = \frac{1}{\theta} \mathbb{1}(0 < x < \theta)$$

$$f(\underset{\sim}{x};\theta) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}(0 < x_i < \theta) = \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} \mathbb{1}(0 < x_i < \theta)$$

$$f_T(t;\theta) = n\left[F(t;\theta)\right]^{n-1} f(t;\theta) = n\left[\frac{t}{\theta}\right]^{n-1} \frac{1}{\theta} \mathbb{1}(0 < t < \theta)$$

$$= \frac{n}{\theta^n} t^{n-1} \mathbb{1}(0 < t < \theta)$$



Now, with

$$\frac{f(\underset{\sim}{x};\theta)}{f_T(T(\underset{\sim}{x});\theta)} = \frac{\left(\frac{1}{\theta}\right)^n \prod_{i=1}^n \mathbb{1}(0 < x_i < \theta)}{\frac{n}{\theta^n}\left[T(\underset{\sim}{x})\right]^{n-1} \mathbb{1}(0 < T(\underset{\sim}{x}) < \theta)}$$

$x_{(1)} > 0$ and $x_{(n)} < \theta$

$$= \frac{\cancel{\left(\frac{1}{\theta}\right)^n} \boxed{\prod_{i=1}^n \mathbb{1}(0 < x_i < \theta)}}{\frac{n}{\cancel{\theta^n}} x_{(n)}^{n-1} \mathbb{1}(0 < x_{(n)} < \theta)} \qquad \boxed{\prod_{i=1}^n \mathbb{1}(0 < x_i < \theta)} = \begin{cases} 1 & \text{if } 0 < x_{(1)},\ x_{(n)} < \theta \\ 0 & \text{o.w.} \end{cases}$$

$$= \frac{\mathbb{1}(x_{(1)} > 0)\ \mathbb{1}(x_{(n)} < \theta)}{n\, x_{(n)}^{n-1}\ \mathbb{1}(0 < x_{(n)} < \theta)}$$

$$= \frac{\mathbb{1}(x_{(1)} > 0)\ \cancel{\mathbb{1}(x_{(n)} < \theta)}}{n\, x_{(n)}^{n-1}\ \mathbb{1}(x_{(n)} > 0)\ \cancel{\mathbb{1}(x_{(n)} < \theta)}}$$

$$= \frac{\mathbb{1}(x_{(1)} > 0)}{n\, x_{(n)}^{n-1}\ \mathbb{1}(x_{(n)} > 0)}.$$

Since this is free of θ, $T(\underset{\sim}{X}) = X_{(n)}$ is a suff. stat for θ.

Theorem (Easier check for suff. by factorization, cf. Thm 6.2.6 in CB)

For $\mathbf{X}$ *with joint pdf/pmf* $f(\mathbf{x};\theta)$,

1. $T = T(\mathbf{X})$ *is a suff. stat. for* θ *iff there exist non-neg fns* $g(t,\theta)$ *and* $h(\mathbf{x})$ *st*

   $\Leftrightarrow$
   $$f(\mathbf{x};\theta) = g(T(\mathbf{x});\theta)\cdot h(\mathbf{x})$$

   for all $\theta \in \Theta$ *and all* $\mathbf{x} \in \mathcal{X}$.

2. *As a consequence, the pdf/pmf* $f_T(t;\theta)$ *of* T *is given by*

   $$f_T(t;\theta) = g(t;\theta)\tilde{h}(t)$$

   for some function $\tilde{h}(t) \geq 0$ *not depending on* θ.

Can be used to *find* a sufficient statistic.

Exercise: Prove the result for the discrete case.

$T(\underset{\sim}{X})$ is suff for θ $\iff$ $f(\underset{\sim}{x};\theta) = g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x})$

Stick to discrete case, $f(\underset{\sim}{x};\theta) = P_\theta(\underset{\sim}{X} = \underset{\sim}{x})$.

"$\Rightarrow$"

Suppose $T(\underset{\sim}{X})$ is suff for θ.

Then $\dfrac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))}$ does not depend on θ.

So we can write

$$P_\theta(\underset{\sim}{X} = \underset{\sim}{x}) = \underbrace{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))}_{g(T(\underset{\sim}{x});\theta)} \underbrace{\boxed{\frac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))}}}_{= h(\underset{\sim}{x}),\ \text{does not depend on } \theta}$$

"$\Leftarrow$"

Suppose $\boxed{P_\theta(\underset{\sim}{X} = \underset{\sim}{x}) = g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x})}$.

Then ... I want to show $P(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t)$ is free of θ.

We have

$$P(\underset{\sim}{X} = \underset{\sim}{x} \mid T(\underset{\sim}{X}) = t) = \frac{P(T(\underset{\sim}{X}) = t \mid \underset{\sim}{X} = \underset{\sim}{x})\, P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = t)}$$

$P(A|B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(B|A)P(A)}{P(B)}$

$$A_t = \{ \underset{\sim}{x} \in \mathcal{X} : T(\underset{\sim}{x}) = t \}$$

$$= \begin{cases} 0 & \underset{\sim}{x} \notin A_t \\ \boxed{\dfrac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))}} & \underset{\sim}{x} \in A_t \end{cases}$$

Still don't know if this depends on θ.

$\underset{\sim}{x} \in A_t \iff T(\underset{\sim}{x}) = t$

Write

$$\frac{P_\theta(\underset{\sim}{X} = \underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X}) = T(\underset{\sim}{x}))} = \frac{g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x})}{\displaystyle\sum_{\underset{\sim}{y} \in \mathcal{X} : T(\underset{\sim}{y}) = T(\underset{\sim}{x})} P\left(T(\underset{\sim}{X}) = T(\underset{\sim}{x}) \cap \underset{\sim}{X} = \underset{\sim}{y}\right)}$$

Could be multiple data sets giving same value of stat.

$$= \frac{g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x})}{\displaystyle\sum_{\underset{\sim}{y} \in \mathcal{X} : T(\underset{\sim}{y}) = T(\underset{\sim}{x})} \underbrace{P\left(T(\underset{\sim}{X}) = T(\underset{\sim}{x}) \mid \underset{\sim}{X} = \underset{\sim}{y}\right)}_{=1} \underbrace{P_\theta(\underset{\sim}{X} = \underset{\sim}{y})}_{g(T(\underset{\sim}{y});\theta)\, h(\underset{\sim}{y})}}$$

$$= \frac{g(\cancel{T(\underset{\sim}{x});\theta})\, h(\underset{\sim}{x})}{\displaystyle\sum_{\underset{\sim}{y} \in \mathcal{X} : T(\underset{\sim}{y}) = T(\underset{\sim}{x})} g(\cancel{T(\underset{\sim}{y});\theta})\, h(\underset{\sim}{y})}$$

$$= \frac{h(\underset{\sim}{x})}{\displaystyle\sum_{\underset{\sim}{y} \in \mathcal{X} : T(\underset{\sim}{y}) = T(\underset{\sim}{x})} h(\underset{\sim}{y})}$$

does not depend on θ.

So $T(\underset{\sim}{X})$ is suff for θ.

$$f(x;\theta) = \frac{\Gamma(1+\theta)}{\Gamma(1)\Gamma(\theta)} x^{1-1}(1-x)^{\theta-1}\mathbb{1}(0<x<1) = \frac{\overset{\theta\Gamma(\theta)}{\Gamma(\theta+1)}}{\Gamma(\theta)}(1-x)^{\theta-1}\mathbb{1}(0<x<1)$$

$$= \theta(1-x)^{\theta-1}\mathbb{1}(0<x<1)$$

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Beta}(\overset{\alpha}{1}, \overset{\beta}{\theta})$, where $\theta > 0$.

Check whether $T(\mathbf{X}) = \prod_{i=1}^n (1 - X_i)$ is a sufficient statistic for θ.

Joint pdf of $X_1, \dots, X_n$ is

$$(1-x_1)^{\theta-1}\cdot(1-x_2)^{\theta-1}\cdot\ldots\cdot(1-x_n)^{\theta-1} = [(1-x_1)(1-x_2)\cdot\ldots\cdot(1-x_n)]^{\theta-1}$$

$$f(\underset{\sim}{x};\theta) = \prod_{i=1}^n \theta(1-x_i)^{\theta-1}\mathbb{1}(0<x_i<1)$$

$$= \underbrace{\theta^n\left[\prod_{i=1}^n(1-x_i)\right]^{\theta-1}}\ \underbrace{\prod_{i=1}^n \mathbb{1}(0<x_i<1)}$$

$$\overset{?}{=} g(T(\underset{\sim}{x});\theta)\cdot h(\underset{\sim}{x})$$

yes :) .

$$f(x;\mu) = \frac{1}{\pi} \frac{1}{1+(x-\mu)^2}, \quad x \in \mathbb{R}$$

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} f_X(x;\mu) = \pi^{-1}[1+(x-\mu)^2]^{-1}$, where $\mu \in \mathbb{R}$.

Check whether $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for μ.

$$f(\underset{\sim}{x};\mu) = \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1+(x_i-\mu)^2} = \left(\frac{1}{\pi}\right)^n \prod_{i=1}^n \frac{1}{1+(x_i-\mu)^2},$$

Cannot find factorization so $T(\underset{\sim}{X}) = \sum X_i$ is NOT a suff stat.

But $T(\underset{\sim}{X}) = (X_{(1)}, \dots, X_{(n)})$. Then

$$f(\underset{\sim}{x};\mu) = \left(\frac{1}{\pi}\right)^n \prod_{i=1}^n \frac{1}{1+(x_{(i)}-\mu)^2} = g((x_{(1)},\dots,x_{(n)});\mu).$$

$h(\underset{\sim}{x}) = 1$

Let $X_1, \dots, X_n \sim f_X(x;\alpha) = \dfrac{\alpha}{x^{\alpha+1}} \mathbb{1}(x>1), \quad \alpha>0.$

Find a suff stat. for α.

Write

$$f(\underset{\sim}{x};\alpha) = \prod_{i=1}^{n} \frac{\alpha}{x_i^{\alpha+1}} \mathbb{1}(x_i>1)$$

$$= \underbrace{\alpha^n \Big(\underbrace{\prod_{i=1}^{n} x_i}_{T(\underset{\sim}{x})}\Big)^{-(\alpha+1)}}_{g(T(\underset{\sim}{x});\alpha)} \underbrace{\prod_{i=1}^{n} \mathbb{1}(x_i>1)}_{h(\underset{\sim}{x})}$$

So $T(\underset{\sim}{X}) = \prod_{i=1}^{n} X_i$ is a suff. stat. for α.

Corollary (1:1 function of a sufficient statistic)

If $T(\mathbf{X})$ is a sufficient statistic for θ and $a(t)$ is a 1:1 function not depending on θ, then $a(T(\mathbf{X}))$ is a sufficient statistic for θ.

Exercise: Prove result.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$, where $p \in (0,1)$. Check sufficiency of

1. $T(\mathbf{X}) = \sum_{i=1}^n X_i$
2. $T'(\mathbf{X}) = \bar{X}_n$

Result: $T(\underset{\sim}{X})$ is suff for θ, a is 1:1.

$\Rightarrow$ $a(T(\underset{\sim}{X}))$ is suff for θ.

Proof of result:

$T(\underset{\sim}{X})$ is suff stat for θ means

$$f(\underset{\sim}{x};\theta) = g(T(\underset{\sim}{x});\theta)\, h(\underset{\sim}{x})$$

$$= g(a^{-1}(a(T(x)));\theta)\quad h(\underset{\sim}{x})$$

$$= \tilde{g}(a(T(x));\theta)\, h(\underset{\sim}{x}),$$

when $\tilde{g}(\cdot;\theta) = g(a^{-1}(\cdot);\theta)$.

So by factorization, $a(T(\underset{\sim}{X}))$ is suff stat. for θ.

$X_1,\dots,X_n \overset{ind}{\sim}$ Bernoulli(p). Consider $T_1(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$

$T_2(\underset{\sim}{X}) = \bar{X}_n$.

$$f(\underset{\sim}{x};p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}\, \mathbb{1}(x_i \in \{0,1\})$$

$$= \underbrace{p^{\overset{n\bar{x}_n}{\sum_{i=1}^{n} x_i}}(1-p)^{n-\overset{n\bar{x}_n}{\sum_{i=1}^{n} x_i}}}_{g(T_1(\underset{\sim}{X});p)}\ \underbrace{\prod_{i=1}^{n}\mathbb{1}(x_i \in \{0,1\})}_{h(\underset{\sim}{x})}$$

$T_1(\underset{\sim}{X}) = \sum_{i=1}^{n} X_i$.

$T_2(\underset{\sim}{X}) = \bar{X}_n = \frac{1}{n}\sum X_i = \frac{1}{n}T_1(\underset{\sim}{X})$, which 1:1. So $\bar{X}_n$ is also suff for p.

A statistic may consist of multiple functions of the data:

$$T(\mathbf{X}) = (T_1(\mathbf{X}), \dots, T_k(\mathbf{X})), \quad k \geq 1.$$

A parameter θ may consist of several values $\theta = (\theta_1, \dots, \theta_d)$, $d \geq 1$.

We can still establish sufficiency using the factorization theorem.

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Normal}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma > 0$. Check sufficiency of

1. $T(\mathbf{X}) = (\bar{X}_n, S_n^2)$ 1:1 $\quad f(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
2. $T'(\mathbf{X}) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$

$$f(\underset{\sim}{x}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[- \frac{\sum_{i=1}^{n} (x_i-\mu)^2}{2\sigma^2} \right]$$

$$= \underbrace{(2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[- \frac{(n-1)s_n^2 + n(\bar{x}_n-\mu)^2}{2\sigma^2} \right]}_{g(\bar{x}_n, s_n^2; \mu, \sigma^2)} \cdot \underbrace{1}_{h(\underset{\sim}{x})}$$

$$\sum_{i=1}^{n} (x_i-\mu)^2 = \sum_{i=1}^{n} \left((x_i-\bar{x}_n) + (\bar{x}_n-\mu) \right)^2$$

$$= \sum_{i=1}^{n} (x_i-\bar{x}_n)^2 - 2 \underbrace{\sum_{i=1}^{n} (x_i-\bar{x}_n)(\bar{x}_n-\mu)}_{=0} + n(\bar{x}_n-\mu)^2$$

$$= (n-1)s_n^2 + n(\bar{x}_n-\mu)^2$$

Theorem (Cannot make a suff. statistic from a non-sufficient statistic)

Let $T(\mathbf{X})$ and $U(\mathbf{X})$ be statistics such that $T(\mathbf{X}) = a(U(\mathbf{X}))$ for a function a not depending on θ. If $T(\mathbf{X})$ is sufficient for θ, $U(\mathbf{X})$ is also sufficient for θ.

If you start with a statistic that is not sufficient, no function of it will be sufficient.

If you reduce the data too much, you cannot recover any info you have discarded.

Note that the entire sample $T(\mathbf{X}) = \mathbf{X}$ is always sufficient!

Exercise: Prove the result.

If $T(\underset{\sim}{x}) = a(U(\underset{\sim}{x}))$ and $T(\underset{\sim}{x})$ is suff. then $U(\underset{\sim}{x})$ is suff.

Prm:
$$f(\underset{\sim}{x};\theta) = g(T(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x})$$
$$= g(a(U(\underset{\sim}{x}));\theta)\,h(\underset{\sim}{x})$$
$$= \tilde{g}(U(\underset{\sim}{x});\theta)\,h(\underset{\sim}{x}), \quad \tilde{g}(\cdot;\theta) = g(a(\cdot);\theta)$$

$A \Rightarrow B$ [Crossed bridge $\Rightarrow$ Crossed River
$A^c \Leftarrow B^c$ didn't cross bridge $\Leftarrow$ didn't cross River]

Exercise: Let $Y_1, \ldots, Y_n$ be ind. rvs and $\boxed{x_1, \ldots, x_n \text{ real numbers}}$ such that

$$Y_i \sim \text{Poisson}(\lambda_i), \quad \text{where} \quad \log(\lambda_i) = \beta_0 + \beta_1 x_i, \quad \text{for } i = 1, \ldots, n.$$

Show that we lose no information about $(\beta_0, \beta_1) \in \mathbb{R}^2$ by reducing $Y_1, \ldots, Y_n$ to

$$\lambda_i = e^{\beta_0 + \beta_1 x_i}$$

$$T(\mathbf{Y}) = \left(\sum_{i=1}^n Y_i, \sum_{i=1}^n x_i Y_i \right).$$

$$f(y_i; \lambda_i) = \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!} = \frac{e^{-e^{\beta_0 + \beta_1 x_i}} \left(e^{\beta_0 + \beta_1 x_i} \right)^{y_i}}{y_i!}$$

$$f(\underset{\sim}{y}; \overbrace{\beta_0, \beta_1}^{\theta}) = \prod_{i=1}^{n} \frac{e^{-e^{\beta_0+\beta_1 x_i}} \left(e^{\beta_0+\beta_1 x_i}\right)^{y_i}}{y_i!}$$

$$= \underbrace{e^{-\sum_{i=1}^{n} e^{\beta_0+\beta_1 x_i}} \, \underbrace{e^{\sum_{i=1}^{n} y_i(\beta_0+\beta_1 x_i)}}_{e^{\beta_0 \sum_{i=1}^{n} y_i + \beta_1 \sum_{i=1}^{n} x_i y_i}}}_{g\left(\sum y_i, \sum x_i y_i; \beta_0, \beta_1\right)} \; \underbrace{\frac{1}{\prod_{i=1}^{n} y_i!}}_{h(\underset{\sim}{y})}$$

$$\Rightarrow \quad T(\underset{\sim}{y}) = \left(\sum_{i=1}^{n} y_i, \; \sum_{i=1}^{n} x_i y_i\right)$$

Exercise: Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} f_X(x;\theta)$, $\theta \in \Theta$.

Show that $T(\mathbf{X}) = (X_{(1)}, \dots, X_{(n)})$ is sufficient for θ.

$$f(\underset{\sim}{x};\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \underbrace{\prod_{i=1}^{n} f(x_{(i)};\theta)}_{g(x_{(1)},\dots,x_{(n)};\theta)} \cdot \underbrace{1}_{h(\underset{\sim}{x})}$$

$$f(x;\theta) = h(x)\,c(\theta)\, e^{\omega_1(\theta)t_1(x) + \dots + \omega_k(\theta)t_k(x)}$$

Theorem (Find suff. stat. in exp. family, cf. Thm 6.2.10 in CB)

If $X_1, \dots, X_n \overset{\text{ind}}{\sim} f(x;\theta)$, *where* $f(x;\theta)$ *belongs to an exponential family such that*

$$f(x;\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^k w_j(\theta)t_j(x)\right), \quad x \in \mathbb{R}, \quad \theta \in \Theta,$$

then a sufficient statistic for θ *is given by*

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

Exercise: Prove the result.

$$f(\underset{\sim}{x};\theta) = \prod_{i=1}^n h(x_i)\,c(\theta)\exp\left[\omega_1(\theta)t_1(x_i) + \dots + \omega_k(\theta)t_k(x_i)\right]$$

$$= \underbrace{\left(\prod_{i=1}^{n} h(x_i)\right)}_{\tilde{h}(\underline{x})} \underbrace{\left(c(\theta)\right)^n \exp\left[\omega_1(\theta)\sum_{i=1}^{n} t_1(x_i) + \dots + \omega_k(\theta)\sum_{i=1}^{n} t_k(x_i)\right]}_{g\left(\sum_{i=1}^{n} t_1(x_i), \dots, \sum_{i=1}^{n} t_k(x_i);\ \theta\right)}$$

$$\underline{}$$

Suff by factorization

Exercise: Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(\alpha, \beta)$, where $\alpha > 0$, $\beta > 0$.

Find a sufficient statistic for (α, β) using the exponential family result.

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \mathbb{1}(0 < x < 1)$$

$$= \mathbb{1}(0 < x < 1) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \exp\left(\log\left(x^{\alpha-1}(1-x)^{\beta-1}\right)\right)$$

$$= \mathbb{1}(0 < x < 1) \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \exp\left((\alpha-1)\log x + (\beta-1)\log(1-x) \right)$$

$w_1(\alpha,\beta) = \alpha - 1$ $\qquad$ $w_2(\alpha,\beta) = \beta - 1$

$t_1(x) = \log x$ $\qquad$ $t_2(x) = \log(1-x)$

$$T(\underset{\sim}{X}) = \left(\sum_{i=1}^{n} \log X_i , \sum_{i=1}^{n} \log(1-X_i) \right)$$

$$e^{\sum \log X_i} = \prod e^{\log X_i} = \prod X_i$$

$$\tilde{T}(\underset{\sim}{X}) = \left(\prod_{i=1}^{n} X_i , \prod_{i=1}^{n} (1-X_i) \right).$$