Data Reduction*

Ricardo Dahis

PUC-Rio, Department of Economics

Summer 2023

*Last updated: January 30, 2023, 08:45h

Contents

Contents

how to summarize the information in the sample?

- if the sample size n is large, then the sample $x_1, \dots, x_n$ is a long list of numbers that may be hard to interpret
- solution: compute a few statistics, e.g., sample mean, variance and quantiles, to determine the key features of the sample values
- any statistic $T(X)$ defines a form of data reduction
- like partitioning the sample space into sets $A_t = \{x : T(x) = t\}$, and so we should be very careful in defining these partitions
- general principle: contrive data reduction methods that **do not discard** important information about the unknown parameter vector θ as well as that **do discard** irrelevant information

example

- experimenter A and B know that some data X has been generated as a normal random variable with mean μ and variance σ^2.
- only experimenter A has access to the data, but tells B the sample average $\bar{X}_n$ and sample variance S_N^2. I.e., tells $T(x)$.
- does experimenter B need any additional information to fully characterize the distribution?
- experimenter B could, for example, generate another stretch of data y such that $T(x) = T(y)$
- we shall see that $\bar{X}_N$ and S_N^2 as sufficient statistics for the normal distribution

Contents

- **definition**:
  (i) $T(X)$ is sufficient for θ if any inference about θ depends on the sample X only through $T(X)$
  (ii) more formally, a statistic $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample X given the value of $T(X)$ does not depend on θ.
- the information of θ in the observation of X is concentrated in that of T. Usually, T is of lower dimension than X. Hence, the observation of T is less costly, though it includes the same amount of information on θ. Usefulness of a sufficient statistics lies in such data reduction.
- equivalently, if x and y are two samples such that $T(x) = T(y)$, then the inference about θ should be the same regardless of whether we observe $X = x$ or $X = y$

characterization via pdf/pmf

- discrete random variables: we can always write

$$\mathbb{P}_\theta(X = x, T(X) = T(x)) \quad = \quad \mathbb{P}_\theta\big(X = x \mid T(X) = T(x)\big) \cdot \mathbb{P}_\theta\big(T(X) = T(x)\big)$$

- if $T(X)$ is a sufficient statistic, then

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

characterization via pdf/pmf

- so to verify whether $T(X)$ is a sufficient statistic, we must check if

$$\mathbb{P}_\theta\big(X = x \,|\, T(X) = T(x)\big) \quad = \quad \frac{\mathbb{P}_\theta(X = x, T(X) = T(x))}{\mathbb{P}_\theta\big(T(X) = T(x)\big)}$$

  does not depend on θ, for all fixed x and t. Finally, we use the fact that $\{X = x\}$ is a subset of $\{T(X) = T(x)\}$

$$\begin{aligned} \mathbb{P}_\theta\big(X = x \,|\, T(X) = T(x)\big) \quad &= \quad \frac{\mathbb{P}_\theta\big(X = x\big)}{\mathbb{P}_\theta\big(T(X) = T(x)\big)} \\ &= \quad \frac{p(x|\theta)}{q\big(T(x)|\theta\big)} \end{aligned}$$

  where p is the pdf of X and q is the pdf of $T(x)$ given θ.

- continuous random variables: analogous with $p(x|\theta)$ and $q(t|\theta)$ denoting the pdfs of X and of the statistic $T(X)$, respectively

binomial sufficient statistic

- **theorem**: if $X_1, \ldots, X_n$ are iid Bernoulli random variables with parameter θ, then $T(X) = X_1 + \ldots + X_n$ is a sufficient statistic for θ
- proof: given that $T(X) \sim \text{Bin}(n, \theta)$, the ratio of pdfs is, defining $t = \sum_{i=1}^n x_i$

$$
\begin{aligned}
\frac{p(x|\theta)}{q(T(x)|\theta)} &= \frac{\prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{\sum_{i=1}^n (1-x_i)}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \\
&= \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}}
\end{aligned}
$$

which does not depend on θ ■

normal sufficient statistic

- theorem: if $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ random variables, then $T(X) = \bar{X}_n$ is a sufficient statistic for μ

- proof: given that $T(X) \sim N(\mu, 1/n)$, the ratio of pdfs is

$$
\begin{aligned}
\frac{f(x|\mu)}{q(\bar{x}_n|\mu)} &= \frac{\prod_{i=1}^n (2\pi)^{-1/2} \exp\left(-(x_i-\mu)^2/2\right)}{(2\pi/n)^{-1/2} \exp\left(-\frac{n}{2}(\bar{x}_n-\mu)^2\right)} \\
&\vdots \\
&= \frac{(2\pi)^{-n/2} \exp\left(-\frac{1}{2}\left[\sum_{i=1}^n (x_i-\bar{x}_n)^2 + n(\bar{x}_n-\mu)^2\right]\right)}{(2\pi/n)^{-1/2} \exp\left(-\frac{n}{2}(\bar{x}_n-\mu)^2\right)} \\
&= n^{-1/2}(2\pi)^{-(n-1)/2} \exp\left(-\frac{1}{2}\sum_{i=1}^n (x_i-\bar{x}_n)^2\right),
\end{aligned}
$$

which does not depend on μ ■

how to find a sufficient statistic

- it may not be the best approach to use the definition directly, since one has to guess the sufficient statistic, find the pmfs, and calculate the ratio.

- fortunately, there is the following theorem

- **factorization theorem** (CB 6.2.6): let $f_X(x|\theta)$ denote the joint pmf/pdf of a sample X, then $T(X)$ is a sufficient statistic for θ if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that, for all sample points x and all parameter points θ,

$$f_X(x|\theta) \quad = \quad g\big(T(x)|\theta\big)\, h(x)$$

- proof ($\Rightarrow$, discrete case): note that, because $T(X)$ is a sufficient statistic, $h(X) = \mathbb{P}\big(X = x \,|\, T(X) = T(x)\big)$ does not depend on θ. Letting then $g\big(t|\theta\big) = \mathbb{P}_\theta\big(T(X) = t\big)$ yields

$$\begin{aligned} f(x|\theta) &= \mathbb{P}_\theta(X = x) \\ &= \mathbb{P}_\theta\big(X = x, T(X) = T(x)\big) \\ &= \mathbb{P}_\theta\big(T(X) = T(x)\big)\mathbb{P}\big(X = x \,|\, T(X) = T(x)\big) \\ &= g\big(T(x)|\theta\big)\, h(x) \end{aligned}$$

■

how to find a sufficient statistic

- proof ($\Leftarrow$, discrete case): assume that factorization

$$f_X(x|\theta) \;=\; g\big(T(x)|\theta\big)\,h(x)$$

exists and examine the ratio $\frac{f_X(x|\theta)}{q(T(x)|\theta)}$, where $q(t|\theta)$ is the pmf of $T(X)$.

$$\frac{f_X(x|\theta)}{q(T(x)|\theta)} \;=\; \frac{g(T(x)|\theta)h(x)}{q(T(x)|\theta)} \;=\; \frac{g(T(x)|\theta)h(x)}{\sum_{\mathcal{A}_{T(x)}} g(T(y)|\theta)h(y)}$$

where $\mathcal{A}_{T(x)} = \{y : T(y) = T(x)\}$. Then

$$\frac{f_X(x|\theta)}{q(T(x)|\theta)} \;=\; \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\sum_{\mathcal{A}_{T(x)}} h(y)} \;=\; \frac{h(x)}{\sum_{\mathcal{A}_{T(x)}} h(y)}$$

The ratio does not depend on θ, so $T(X)$ is a sufficient statistic. ■

how to find a sufficient statistic

- to use the factorization theorem, we factor the joint pdf of the sample into two parts:
  - $h(x)$: one does not depend on θ
  - $g(T(x)|\theta)$: depends on the sample x only through $T(x)$
- let's see some examples...

discrete uniform sufficient statistic

- theorem: if $X_1, \ldots, X_n$ are iid uniform on $1, \ldots, \theta$ then $T(X) = X_{(n)}$ is a sufficient statistic for θ
- proof: the joint pmf of X is

$$f(x|\theta) = \begin{cases} \frac{1}{\theta^n} & \text{if } x_i = 1, \ldots, \theta \text{ for } i = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

write the restriction

$$\{x_i = 1, \ldots, \theta \text{ for } i = 1, \ldots, n\} = \mathcal{I}(x_i \in \mathbb{N}) \cdot \mathcal{I}(x_{(n)} \leq \theta)$$

so

$$f(x|\theta) = \underbrace{\frac{1}{\theta^n} \mathcal{I}(x_{(n)} \leq \theta)}_{g(x_{(n)}|\theta)} \cdot \underbrace{\prod_{i=1}^{n} \mathcal{I}(x_i \in \mathbb{N})}_{h(x)}$$

normal distribution, both parameters unknown

- theorem: let $X_1, \ldots, X_n$ be iid $N(\mu, \sigma^2)$. Then

$$\begin{aligned} T_1(x) &= \bar{x} \\ T_2(x) &= s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \end{aligned}$$

are sufficient statistics.

- proof: the joint pdf of the sample X is

$$\begin{aligned} f(x|\mu,\sigma^2) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left\{-(x_i-\mu)^2/(2\sigma^2)\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\left(\sum_{i=1}^{n}(x_i-\bar{x})^2 + n(\bar{x}-\mu)^2\right)/(2\sigma^2)\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\left((n-1)t_2 + n(t_1-\mu)^2\right)/(2\sigma^2)\right\} \end{aligned}$$

then select $g(T(x)|\theta)$ as this expression and $h(x) = 1$. ■

exponential family

- theorem: if $X_1, \ldots, X_n$ are iid observations from an exponential family then

$$T(X) = \left(\sum_{j=1}^{n} s_1(X_j), \ldots, \sum_{j=1}^{n} s_d(X_j) \right)$$

  is sufficient for θ

- proof: the joint pdf is

$$\begin{aligned} \prod_{i=1}^{n} \left\{ h(x_i) c(\theta) \exp\left(\sum_{j=1}^{k} w_j(\theta) t_j(x_i) \right) \right\} &= c(\theta)^n \exp\left(\sum_{i=1}^{n} \sum_{j=1}^{k} w_j(\theta) t_j(x_i) \right) \cdot \prod_{i=1}^{n} h(x_i) \\ &= \underbrace{c(\theta)^n \exp\left(\sum_{j=1}^{k} w_j(\theta) \sum_{i=1}^{n} t_j(x_i) \right)}_{\equiv g(T(x)|\theta)} \cdot \underbrace{\prod_{i=1}^{n} h(x_i)}_{\equiv h(x)} \end{aligned}$$

sufficient statistics

- not all sufficient statistic achieve a substantial data reduction
- example: if $X_1, \dots, X_n$ are i.i.d. from a pdf f_X, the order statistic

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

  is a sufficient statistic for f_X

- example: the complete sample is a sufficient statistic, since

$$f_X(x|\theta) = f(T(x)|\theta)h(x)$$

  with $h(x) = 1$

- sometimes it is impossible to achieve a substantial data reduction: nonparametric statistics
- it turns out that having a data reduction is a particular property of only a few distributions: outside the exponential family, it is rare to have a sufficient statistic smaller than the sample size

Contents

minimal sufficient statistics

- as we have seen, there are often numerous sufficient statistics. which one is better?
- a criterion: achieving more data reduction is better
- a minimial sufficient statistic is a statistic that has achieved the maximal amount of data reduction while still retaining all the information about the parameter θ

- **theorem**: any bijective function of a sufficient statistic is also a sufficient statistic.
- proof: to see this, let $T(X)$ be a sufficient statistic and $T^*(X) \equiv r(T(X))$. Then

$$f_X(X|\theta) \;=\; g(T(X)|\theta)h(x) \;=\; g(r^{-1}(T^*(X))|\theta)h(x)$$

and defining $g^*(t|\theta) = g(r^{-1}(t)|\theta)$,

$$f(x|\theta) \;=\; g^*(T^*(x)|\theta)h(x)$$

so $T^*(x)$ is a sufficient statistic ∎

minimal sufficient statistics

- definition: a sufficient statistic $T(X)$ is called a **minimal** sufficient statistic if, for any other sufficient statistic $T'(X)$, $T(X)$ is a function of $T'(X)$.
- we defined a sufficient statistic as a partition of the sample space $\mathcal{X}$:
  - let $\mathcal{T} = \{t : t = T(x), x \in \mathcal{X}\}$, the image of $\mathcal{X}$ under $T(x)$
  - $T(x)$ induces a partition of $\mathcal{X}$, $\{\mathcal{A}_t : t \in \mathcal{T}\}$, $\mathcal{A}_t = \{x : T(x) = t\}$
- ... and back to the minimal sufficient statistics:
  - if $T(x)$ is a function of $T'(x)$, then $T'(x) = T'(y) \implies T(x) = T(y)$
  - let $\mathcal{B}_{t'} = \{x | T'(x) = t'\}$. So $\mathcal{B}_{t'} \subseteq \mathcal{A}_t$ for some t
  - this must be true for any sufficient statistic $T'(x)$
  - in other words, the minimal sufficient statistic induces the coarsest partition $\{\mathcal{A}_t | t \in \mathcal{T}\}$ of $\mathcal{X}$ among all sufficient statistics

minimal sufficient statistics

- again, applying this definition may not be too practical. Fortunately, we have the following theorem

- **theorem** (CB 6.2.13): let $f(x|\theta)$ be the pmf or pdf of a sample X. Suppose that there exists a function $T(x)$ such that, for every two sample points x and y, the ratio $\frac{f(x|\theta)}{f(y|\theta)}$ is a constant function of θ if and only if $T(x) = T(y)$. Then $T(X)$ is a minimal sufficient statistic for θ.

- before proving the theorem, let's see an example

minimal sufficient statistics

- example: let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$, both μ and σ^2 unknown. The ratio of densities are

$$\begin{aligned} \frac{f(x|\mu,\sigma^2)}{f(y|\mu,\sigma^2)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left\{-[n(\bar{x}-\mu)^2+(n-1)s_x^2]/(2\sigma^2)\right\}}{(2\pi\sigma^2)^{-n/2}\exp\left\{-[n(\bar{y}-\mu)^2+(n-1)s_y^2]/(2\sigma^2)\right\}} \\ &= \exp\left\{\frac{-n(\bar{x}^2-\bar{y}^2)+2n\mu(\bar{x}-\bar{y})-(n-1)(s_x^2-s_y^2)}{2\sigma^2}\right\} \end{aligned}$$

  which will be a constant function of μ and σ^2 if, and only if, $\bar{x}=\bar{y}$ and $s_x^2 = s_y^2$. Thus, $(\bar{X}, S^2)$ is a minimal sufficient statistic for (μ, σ^2).

- proof ($T(X)$ is a sufficient statistic): let $\mathcal{T} = \{t : t = T(x) \text{ for some } x \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(X)$, along with its partitions $\mathcal{A}_t = \{x : T(x) = t\}$. For each $\mathcal{A}_t$, fix one element $x_t \in \mathcal{A}_t$. So for any $x \in \mathcal{X}$, there is a $x_{T(x)} \in \mathcal{A}_t$. That is, both x and $x_{T(x)} \in \mathcal{A}_t$. Therefore $T(x) = T(x_{T(x)})$ and $\frac{f(x|\theta)}{f\left(x_{T(x)}|\theta\right)}$ is constant in θ.

  Now take $g(t|\theta) = f(x_t|\theta)$ defined on $\mathcal{T}$, and $h(x) = \frac{f(x|\theta)}{f\left(x_{T(x)}|\theta\right)}$ defined on $\mathcal{X}$. And write

$$f(x|\theta) \quad = \quad g(T(x)|\theta)h(x)$$

  By the factorization theorem, $T(X)$ is a sufficient statistic. ■

minimal sufficient statistics

- proof ($T(X)$ is a minimal sufficient statistic): let $T'(X)$ be any other sufficient statistic. By the factorization theorem, there exists functions g' and h' such that $f(x|\theta) = g'(T'(x)|\theta)h'(x)$. Let x and y be any two sample points with $T'(x) = T'(y)$. 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 the ratio does not depend on θ, by assumption $T(x) = T(y)$. Thus, $T(x)$ is a function of $T'(x)$ and $T(x)$ is minimal. ∎

minimal sufficient statistics

- example: let $\{X_1, \ldots, X_n\}$ be a random sample. Find the minimal sufficient statistics for θ with distribution.

$$f(x|\theta) \quad = \quad e^{-(x-\theta)}, \;\; \theta < x < \infty, \;\; -\infty < \theta < \infty$$

- answer:

$$\begin{aligned}
\frac{f(x|\theta)}{f(y|\theta)} &= \frac{\prod_{i=1}^{n}\left(e^{-(x_i-\theta)} \cdot \mathcal{I}(\theta < x_i < \infty)\right)}{\prod_{i=1}^{n}\left(e^{-(y_i-\theta)} \cdot \mathcal{I}(\theta < y_i < \infty)\right)} \\
&= \frac{e^{n\theta} e^{-\sum_i x_i} \prod_{i=1}^{n} \mathcal{I}(\theta < x_i < \infty)}{e^{n\theta} e^{-\sum_i y_i} \prod_{i=1}^{n} \mathcal{I}(\theta < y_i < \infty)} \\
&= \frac{e^{-\sum_i x_i} \mathcal{I}(\theta < \min x_i < \infty)}{e^{-\sum_i y_i} \mathcal{I}(\theta < \min y_i < \infty)}
\end{aligned}$$

which is independent of θ if, and only if, $T(X) = \min\{X_1, \ldots, X_n\}$.

Contents

likelihood function

- **definition**: let $f_X(x|\theta)$ denote the joint pdf/pmf of X, then the likelihood function of θ given $X = x$ is

$$\ell(\theta|x) \quad = \quad f_X(x|\theta)$$

- discrete case: if we compare the likelihood functions at θ_1 and θ_2 and find that

$$P_{\theta_1}(X = x) \quad = \quad \ell(\theta_1|x) \quad > \quad \ell(\theta_2|x) \quad = \quad P_{\theta_2}(X = x),$$

  then the sample we observe is more likely to stem from $\theta = \theta_1$ than from $\theta = \theta_2$.

- in other words, θ_1 is more plausible that θ_2 given $X = x$

- continuous case: remains a basis for comparison

$$\frac{P_{\theta_1}(|X - x| < \epsilon)}{P_{\theta_2}(|X - x| < \epsilon)} \quad \cong \quad \frac{\ell(\theta_1|x)}{\ell(\theta_2|x)} \qquad \text{for small } \epsilon > 0$$

likelihood function

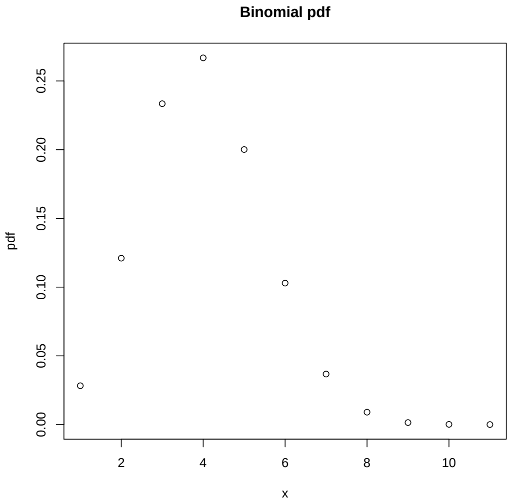
- example: let X have a binomial distribution. The p.d.f. is a function of x, given p,

$$f_X(x|p = 0.3) \quad = \quad \binom{10}{x}(0.3)^x(0.7)^{10-x}$$

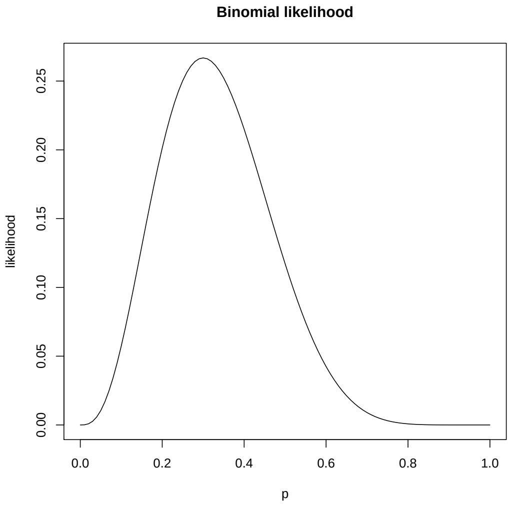
  and the likelihood is a function of p given x

$$\ell(p|x = 3) \quad = \quad \binom{10}{3}p^3(1-p)^{10-3}$$

likelihood function



likelihood function



principle

- **likelihood principle**: if x and y are two sample points such that $\ell(\theta|x)$ is proportional to $\ell(\theta|y)$, that is to say, there exists a constant $c(x, y)$ such that $\ell(\theta|x) = c(x, y)\,\ell(\theta|y)$ for all θ, then they entail the same information about θ
- that is, even if two sample points x and y have only proportional likelihoods, then they contain equivalent information about θ (this is true as long as $c(x, y)$ does not depend on θ)
- we are careful enough to say that θ_1 is more plausible that θ_2 rather than more probable, not only because $\ell(\theta|x)$ is typically not a pdf, but also because θ is usually fixed

Contents

Reference:

- Casella and Berger, Ch. 6

Exercises:

- 6.1–6.6, 6.8, 6.9, 6.16, 6.17.