Summary (formulation of information and data reduction problem, TBp. 305)

- Let $X_1, X_2, \ldots, X_n$ be a sample with joint pdf/pmf $f(\mathbf{x}|\Theta)$, where Θ is unknown parameter. ← statistical modeling ⇒ introduce Θ (unknown systematic pattern)
  - $X_1, X_2, \ldots, X_n$ contains two types of information:
    - ∗ information related to Θ ← "important" (useful) information
    - ∗ information irrelevant to Θ ← useless information.

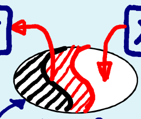
    Fisher information / cf.
  - For example, toss a coin n times, i.e., $X_1, X_2, \ldots, X_n$ are i.i.d. from Bernoulli $B(\theta)$, ← What is important information?
    - ∗ $\overline{X}_n$ or $T = \sum_{i=1}^n X_i$ contains information about θ 3/31

      estimator of θ / non-invertible
    - ∗ When T is known, say $T = t$, the information that at which trials the t head's occur is irrelevant to θ ← place heads in t out of n positions

      what information lost? / useless information
    - ∗ n=5, consider the following possible results:
      - ▷ $(0, 1, 1, 1, 1)$, $T = 4$; $(1, 0, 1, 1, 1)$, $T = 4$; $(1, 1, 0, 1, 1)$, $T = 4$; $(1, 1, 1, 0, 1)$, $T = 4$; $(1, 1, 1, 1, 0)$, $T = 4$

        $P(X_1, \cdots, X_5 | T=4) = 1/5$ ← irrelevant to θ ← true for any t
      - ▷ $(1, 0, 0, 0, 0)$, $T = 1$; $(0, 1, 0, 0, 0)$, $T = 1$; $(0, 0, 1, 0, 0)$, $T = 1$; $(0, 0, 0, 1, 0)$, $T = 1$; $(0, 0, 0, 0, 1)$, $T = 1$

        $P(X_1, \cdots, X_5 | T=1) = 1/5$ ← irrelevant to θ

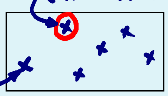


$T \sim$ Binomial$(5, \theta)$ — distribution of T — different θ ⇒ different distributions

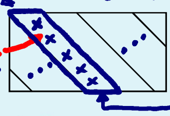
distribution of $X_1, \cdots, X_5 | T$ — different θ ⇒ the same distribution — cf.

- Information about θ is revealed by the different values of T, i.e., larger T, larger θ, and vice versa.

data space — $(X_1, \ldots, X_n)$ — (1,0,1,1,1), (0,1,0,0,0)



$T = \sum_{i=1}^n X_i$ — It's enough to keep t — $T=4$ — data with same value of T ⇒ data reduction — observations that carry same information about θ



- Question. Is there a statistic $T(X_1, X_2, \ldots, X_n)$ which contains all the information in the sample about θ? If so, a reduction of the original data to this statistic without loss of "information" is possible. useful

  充分

Definition 6.13 (sufficient, TBp. 305)

A statistic $T(X_1, X_2, \ldots, X_n)$ is said to be **sufficient** for θ if the conditional distribution of $X_1, X_2, \ldots, X_n$ given $T = t$ does not depend on θ for any value of t. → When T is known (given), the rest (probabilistic) information in $X_1, \cdots, X_n$ is irrelevant to θ. (check graph in LNp.44)

can be a vector

Caution: → Note. It is possible that the joint distribution we assigned to the data is not suitable. ← statistical modeling

1. If T is a sufficient statistic, formally, we can keep only T and throw away all X_i's. Realistically, the X_i's are used to check whether the model did not fit, or that something was fishy about the data. ← exam statistical modeling
2. The definition of "all (important) information" depends on the statistical modeling, i.e., the joint distribution assumption.

Example 6.20 (sufficient statistics of i.i.d. Bernoulli distribution, TBp. 306)

Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli random variables with $P(X_i = 1) = \theta$. Let $T = \sum_{i=1}^{n} X_i$ then $T \sim \text{Binomial}(n, \theta)$

unknown 規律

$$P(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{P(X_1 = x_1, \ldots, X_n = x_n, T = t)}{P(T = t)}$$

$\sum_{i=1}^{n} x_i = t$

$$= \frac{\theta^{\overbrace{t}^{\Sigma x_i}}(1-\theta)^{\overbrace{n-t}^{\Sigma(1-x_i)}}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \frac{1}{\binom{n}{t}},$$

This carries information about where the 1's locate, which is irrelevant to θ

If $\sum_{i=1}^{n} x_i \neq t$, the probability is zero

if $x_1 + \cdots + x_n = t$ and x_i are nonnegative integers, and 0 otherwise. The conditional distribution is independent of θ. Hence T is sufficient for θ.

Theorem 6.10 (factorization theorem, TBp. 306) can be a vector, i.e., $T(\underset{\sim}{X}) = (T_1(\underset{\sim}{X}), \cdots, T_k(\underset{\sim}{X}))$

分解定理

A necessary and sufficient condition for $T(X_1, \ldots, X_n)$ to be sufficient for a parameter θ is that the joint pdf or pmf of $X_1, \ldots, X_n$ factors in the form

$$f(x_1, x_2, \ldots, x_n | \theta) = g\big(T(x_1, x_2, \ldots, x_n), \theta\big)\, h(x_1, x_2, \ldots, x_n)$$

free of θ

$P(X_1 = x_1, \cdots, X_n = x_n, T(X_1, \cdots, X_n) = t)$

multiplication law

intuition: $P(X_1 = x_1, \ldots, X_n = x_n) = P(T = t)P(X_1 = x_1, \ldots, X_n = x_n | T = t)$

Proof: only for discrete case (continuous case requires some regularity conditions, but the basic idea are the same.): $(\Leftarrow)$ Suppose

$$f(x_1, x_2, \ldots, x_n | \theta) = g\big(\underbrace{T(x_1, x_2, \ldots, x_n)}_{t}, \theta\big)\, h(\underbrace{x_1, x_2, \ldots, x_n}_{\mathbf{x}}).$$

Then

$$P(T = t) = \sum_{T(\mathbf{x})=t} P(\mathbf{X} = \mathbf{x}) = g(t, \theta) \sum_{T(\mathbf{x})=t} h(\mathbf{x}),$$

the sum is irrelevant to θ

$T(\mathbf{x}) = t$

$$P(\mathbf{X} = \mathbf{x} | T = t) = \frac{P(\mathbf{X} = \mathbf{x}, T = t)}{P(T = t)} \overset{= P(\mathbf{X}=\mathbf{x}) =}{} \frac{\cancel{g(t,\theta)} \cdot h(\mathbf{x})}{\cancel{g(t,\theta)} \cdot \sum_{T(\mathbf{x})=t} h(\mathbf{x})},$$

for $\mathbf{x}$ s.t. $T(\mathbf{x}) = t$

which does not depend on θ. Hence T is sufficient for θ. $(\Rightarrow)$ Conversely, suppose that the conditional distribution of $\mathbf{X}$ given T is independent of θ. Let

$$g(t, \theta) = P(T = t | \theta), \quad h(\mathbf{x}) = P(\mathbf{X} = \mathbf{x} | T = t).$$

Then $P(\mathbf{X} = \mathbf{x} | \theta) = P(T = t | \theta) P(\mathbf{X} = \mathbf{x} | T = t) = g(t, \theta) h(\mathbf{x})$ as required.

Theorem 6.11 (MLE and sufficient statistics, TBp.309) → check ★ in LNp.46

If T is sufficient for θ, then the maximum likelihood estimate for θ, if unique, is a function of T. Note. for any sufficient statistics

Note. It does not mean that MLE is always sufficient.

Proof. From factorization theorem, the likelihood is $g(t, \theta)h(\mathbf{x})$. To maximize this quantity we only need to maximize $g(t, \theta)$.

Example 6.21 (cont. Ex. 6.20, sufficient statistic of i.i.d. Bernoulli distribution, TBp.309)

Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables

$$P(X_i = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0 \text{ or } 1.$$

Then

$$f(x_1,\dots,x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}$$

$$= \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^n x_i}(1-\theta)^n = g(t,\theta)h(x_1,\dots,x_n)$$

dimension reduction $n \to 1$

where $t = \sum_{i=1}^n x_i$, $g(t,\theta) = \left(\frac{\theta}{1-\theta}\right)^t (1-\theta)^n$, $h(x)=1$

By factorization Thm.

Hence $T = \sum_{i=1}^n X_i$ is sufficient for θ. $\xleftrightarrow{\text{cf.}}$ Ex6.18 (LNp.36-38) $\to$ Note. T and $(X_1,\cdots,X_n)$ yield the same amount of Fisher information

Example 6.22 (sufficient statistics of i.i.d. Normal distribution, TBp.308)

If $X_i \sim N(\mu,\sigma^2)$, $i=1,2,\dots,n$, are i.i.d, where μ,σ are unknown. Then

$\hookrightarrow$ parameters

$$f(x_1,\dots,x_n|\mu,\sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}$$

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

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

dimension reduction $n \to 2$

$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$

$\hat{\mu} = \bar{X}_n$

sufficient $\to (\hat{\mu}, \hat{\sigma}^2)$

$\updownarrow$ one-to-one transformation

and $\left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is a 2-dimensional sufficient statistic for (μ,σ).