Sufficient statistics and factorization theorem

- Suppose that $X = (X_1, \ldots, X_n)$ is a sample and the distribution of $(X_1, \ldots, X_n)$ is determined by a parameter vector θ, where θ is in some space Θ. For a statistic $T(X)$, if the conditional distribution of X given $T(X) = t$ does not depend on θ for all t (in the range of $T(X)$), then $T(X)$ is called a sufficient statistic for θ. We can think that the data $X_1, \ldots, X_n$ are generated in two steps:

  - Step 1. Generate $T(X)$ according the distribution of $T(X)$.
  - Step 2. Suppose that we obtain $T(X) = t$ from Step 1. Generate $X = (X_1, \ldots, X_n)$ according to the conditional distribution of X given $T(X) = t$.

  $T(X)$ is sufficient for θ means that in Step 2, the way X is generated does not depend on θ as long as $T(X) = t$ is given. Thus we can estimate θ based on $T(X)$ only (instead of based on X) without lossing information.

- Factorization Theorem. Suppose that $X = (X_1, \ldots, X_n)$ is a sample and X has PDF (or PMF) f_θ, where $\theta \in \Theta$. For a statistic $T(X)$, $T(X)$ is a sufficient statistic for θ if and only if there exist functions g and h such that

$$f_\theta(x) = g(T(x), \theta)h(x) \text{ for all } x \tag{1}$$

  for all $\theta \in \Theta$. Note that h does not depend on θ.

- Proof of the factorization theorem under some conditions. Here we assume that X is discrete and the set of possible values of X and the set of possible values of $T(X)$ do not depend on θ. Suppose that t is a possible value of $T(X)$ ($P(T(X) = t) > 0$ for all θ). Let p_t be the conditional PMF of X given $T(X) = t$ and let

$$S_t = \{x : T(x) = t\},$$

  then

$$\begin{aligned} p_t(x) &= P(X = x | T(X) = t) \\ &= \frac{P(X = x \text{ and } T(X) = t)}{P(T(X) = t)} \\ &= \frac{P(X = x) I_{S_t}(x)}{P(T(X) = t)} \end{aligned} \tag{2}$$

  for all x.

  - Proof of the "only if" direction. Suppose that $T(X)$ is sufficient for θ, then the PMF p_t does not depend on θ. For x that is a possible value of X, take $t = T(x)$ in (2) and we have

$$P(X = x) = p_{T(x)}(x) P(T(X) = T(x)),$$

    so (1) holds with $h(x) = p_{T(x)}(x)$ and $g(T(x), \theta) = P(T(X) = T(x))$.

– Proof of the "if" direction. Suppose that (1) holds. Then the conditional PMF of X given $T(X) = t$ at x is

$$\begin{aligned} p_t(x) &= \frac{P(X = x)I_{S_t}(x)}{P(T(X) = t)} \\ &= \frac{f_\theta(x)I_{S_t}(x)}{\sum_{x':x'\in S_t} f_\theta(x')} \\ &= \frac{g(T(x),\theta)h(x)I_{S_t}(x)}{\sum_{x':T(x')=t} g(T(x'),\theta)h(x')} \\ &= \frac{g(t,\theta)h(x)I_{\{x:T(x)=t\}}(x)}{\sum_{x':T(x')=t} g(t,\theta)h(x')} \\ &= \frac{h(x)I_{S_t}(x)}{\sum_{x':x'\in S_t} h(x')}, \end{aligned}$$

which does not depend on θ. Thus $T(X)$ is sufficient for θ.

- Note. Suppose that $(X_1, \ldots, X_n)$ is a random sample and the distribution of X_1 is $\mathcal{D}$, then we say that $(X_1, \ldots, X_n)$ is a random sample from $\mathcal{D}$.

- Example 1. Suppose that $(X_1, \ldots, X_n)$ is a random sample from $N(\mu, 1)$, where $\mu \in (-\infty, \infty)$. Let $\bar{X} = \sum_{i=1}^n X_i/n$. Show that $\bar{X}$ is a sufficient statistic of μ.

  Sol. For $\mu \in (-\infty, \infty)$, define the function f_μ on R^n as follows:

$$f_\mu(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2}$$

  for $(x_1, \ldots, x_n) \in R^n$. Then, f_μ is a PDF of $(X_1, \ldots, X_n)$. Since

$$\begin{aligned} f_\mu(x_1, \ldots, x_n) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-(x_i-\mu)^2/2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-(\sum_{i=1}^n (x_i-\bar{x})^2 + n(\bar{x}-\mu)^2)/2}, \end{aligned}$$

  where $\bar{x} = \sum_{i=1}^n x_i/n$. Take $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i/n$,

$$g(t, \mu) = e^{-n(t-\mu)^2/2}$$

  and

$$h(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\sum_{i=1}^n (x_i-\bar{x})^2/2},$$

  then

$$\begin{aligned} f_\mu(x_1, \ldots, x_n) &= g(\bar{x}, \mu)h(x_1, \ldots, x_n) \\ &= g(T(x_1, \ldots, x_n), \mu)h(x_1, \ldots, x_n) \end{aligned}$$

for all $(x_1, \ldots, x_n) \in R^n$ for all $\mu \in (-\infty, \infty)$. By the factorization theorem, $\sum_{i=1}^n X_i/n$ is a sufficient statistic of μ.

- Example 2. Suppose that $(X_1, \ldots, X_n)$ is a sample from $N(\mu, \sigma^2)$, where $\mu \in (-\infty, \infty)$ and $\sigma > 0$. Let $\bar{X} = \sum_{i=1}^n X_i/n$ and $S = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)}$. Show that $(\bar{X}, S)$ is a sufficient statistic of (μ, σ).

  The proof is left as an exercise.

- The data $(X_1, \ldots, X_n)$ in Example 1 can be generated in two steps:
  - Step 1. Generate T from $N(\mu, 1/n)$.
  - Step 2. Suppose that we obtain $T = t$ from Step 1. Generate $(X_1, \ldots, X_{n-1})$ from $N(\boldsymbol{\mu}_t, \Sigma)$, where

$$\boldsymbol{\mu}_t = (t, \ldots, t)^T$$

    is a $(n-1) \times 1$ column vector and Σ is an $(n-1) \times (n-1)$ matrix whose (i, j)-th element is

$$\Sigma_{i,j} = \begin{cases} 1 - 1/n & \text{if } i = j; \\ -1/n & \text{if } i \neq j. \end{cases} \tag{3}$$

    Take $X_n = nt - (X_1 + \cdots + X_{n-1})$ and we have $(X_1, \ldots, X_n)$.

  Remarks.

  - In Step 2, the conditional distribution of $(X_1, \ldots, X_{n-1})$ given $T = t$ is $N(\boldsymbol{\mu}_t, \Sigma)$ for all $t \in (-\infty, \infty)$.
  - Suppose that $Y_1, \ldots, Y_n$ are IID $N(\mu, 1)$ and $\bar{Y} = \sum_{i=1}^n Y_i/n$. Then $N(\boldsymbol{\mu}_t, \Sigma)$ is also the conditional distribution of $(Y_1, \ldots, Y_{n-1})$ given $\bar{Y} = t$, which can be found by applying Fact 5 in the handout "Multivariate normal distributions" given last semester, which is stated below.

    Fact. *Suppose that $\boldsymbol{X} = (X_1, \ldots, X_m)^T$, $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, and the distribution of $(\boldsymbol{X}^T, \boldsymbol{Y}^T)$ is a multivariate normal distribution. Let Y_i^* be the best linear predictor of Y_i based on $\boldsymbol{X}$ for $i = 1, \ldots, n$, and let $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)^T$, then (i) and (ii) hold.*

    (i) $\boldsymbol{Y} - \boldsymbol{Y}^*$ and $\boldsymbol{X}$ are independent.

    (ii) Let $B\boldsymbol{X} + \boldsymbol{a} = \boldsymbol{Y}^*$. If the covariance matrix of $(\boldsymbol{X}^T, \boldsymbol{Y}^T)$ is invertible, then a conditional PDF of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ is the continuous PDF of $N(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = B\boldsymbol{x} + \boldsymbol{a}$ and $\Sigma = E(\boldsymbol{Y} - \boldsymbol{Y}^*)(\boldsymbol{Y} - \boldsymbol{Y}^*)^T$.

    Here the best linear predictor of Y_i based on $\bar{Y} = \sum_{i=1}^n Y_i/n$ is $\bar{Y}$ for each i and $Cov(Y_i - \bar{Y}, Y_j - \bar{Y})$ is the $\Sigma_{i,j}$ in (3) for each (i, j).

    The handout "Multivariate normal distributions" is at

```
https://stat.walkup.tw/teaching/math_stat_under/handouts/C03_5_mnormal.pdf
```

  - The joint distribution of $(X_1, \ldots, X_{n-1}, T)$ can be determined by the conditional distribution of of $(X_1, \ldots, X_{n-1})$ given $T = t$ for all t and the marginal distribution of T. See the handout "Finding a joint PDF using conditional and marginal PDFs" given last semester for more details. The handout is at

```
https://stat.walkup.tw/teaching/math_stat_under/handouts/condi_extra.pdf
```

- Generating a random vector X with distribution $N(\mu, \Sigma)$.
  - To generate a random vector X with distribution $N(\mu, \Sigma)$, we can first compute the spectral decomposition of Σ to obtain $\Sigma = PDP^T$, where P is a matrix of eigen vectors of Σ such that $PP^T = I$ and D is a diagonal matrix whose diagonal elements are eigen values of Σ. Then, generate a random vector U from $N(0, D)$ and take $X = \mu + PU$, then $X \sim N(\mu, \Sigma)$.
  - The following R function `rmnorm` returns a random vector `X` generated from N(`mu`, `Sig`) with input `mu` and `Sig`. The spectral decomposition of `Sig` is computed using the R command `eigen(Sig)`. The `P` and `diag.D` computed in the function are P and the vector of diagonal elements of D respectively so that `Sig`= PDP^T and $PP^T = I$.

```
rmnorm <- function(mu, Sig){
  Sig.eigen <- eigen(Sig)
  P <- Sig.eigen$vectors      #P: matrix of eigen vectors of Sig
  D.diag <- Sig.eigen$values  #Sig.eigen$values: vector of eigen values of Sig
                              #D.diag: a vector of diagonal elements of D
  k <- length(mu)
  U <- rnorm(k, mean=rep(0, k), sd=sqrt(D.diag))   #U~N(0,D)
  X <- mu + P%*%U
  return(X)
}
```

- MLE's can be computed based on sufficient statistics.

  Fact 1. *Suppose that $X = (X_1, \ldots, X_n)$ is a sample and X has PDF (or PMF) f_θ, where $\theta \in \Theta$. Suppose that $T(X)$ is a sufficient statistic for θ. Then the MLE of θ can be computed based on $T(X)$.*

  The proof of Fact 1 is based on the "only if" part of the factorization theorem.