Lecture 4
Sufficient Statistics. Introduction to Estimation

1 Sufficient statistics

Let $f(x|\theta)$ with $\theta \in \Theta$ be some parametric family. Let $X = (X_1, ..., X_n)$ be a random sample from distribution $f(x|\theta)$. Suppose we would like to learn parameter value θ from our sample. The concept of sufficient statistic allows us to separate information contained in X into two parts. One part contains all the valuable information as long as we are concerned with parameter θ, while the other part contains pure noise in the sense that this part has no valuable information. Thus, we can ignore the latter part.

Definition 1. Statistic $T(X)$ is sufficient for θ if the conditional distribution of X given $T(X)$ does not depend on θ.

Let $T(X)$ be a sufficient statistic. Consider the pair $(X, T(X))$. Obviously, $(X, T(X))$ contains the same information about θ as X alone, since $T(X)$ is a function of X. But if we know $T(X)$, then X itself has no value for us since its conditional distribution given $T(X)$ is independent of θ. Thus, by observing X (in addition to $T(X)$), we cannot say whether one particular value of parameter θ is more likely than another. Therefore, once we know $T(X)$, we can discard X completely.

Example Let $X = (X_1, ..., X_n)$ be a random sample from $N(\mu, \sigma^2)$. Suppose that σ^2 is known. Thus, the only parameter is μ ($\theta = \mu$). We have already seen that $T(X) = \overline{X}_n \sim N(\mu, \sigma^2/n)$. Let us calculate the conditional distribution of X given $T(X) = t$. First, note that

$$\begin{aligned}
\sum_{i=1}^n (x_i - \mu)^2 - n(\overline{x}_n - \mu)^2 &= \sum_{i=1}^n (x_i - \overline{x}_n + \overline{x}_n - \mu)^2 - n(\overline{x}_n - \mu)^2 \\
&= \sum_{i=1}^n (x_i - \overline{x}_n)^2 + 2\sum_{i=1}^n (x_i - \overline{x}_n)(\overline{x}_n - \mu) \\
&= \sum_{i=1}^n (x_i - \overline{x}_n)^2.
\end{aligned}$$

Therefore

$$\begin{aligned} f_{X|T(X)}(x|T(X)=T(x)) &= \frac{f_X(x)}{f_T(T(x))} \\ &= \frac{\exp\{-\sum_{i=1}^n (x_i-\mu)^2/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n)}{\exp\{-n(\overline{x}_n-\mu)^2/(2\sigma^2)\}/((2\pi)^{1/2}\sigma/n^{1/2})} \\ &= \exp\{-\sum_{i=1}^n (x_i-\overline{x}_n)^2/(2\sigma^2)\}/((2\pi)^{(n-1)/2}\sigma^{n-1}/n^{1/2}), \end{aligned}$$

which is independent of μ. We conclude that $T(X)=\overline{X}_n$ is a sufficient statistic for our parametric family. Note, however, that $\overline{X}_n$ is not sufficient if σ^2 is not known.

2 Factorization Theorem

The Factorization Theorem gives a general approach for how to find a sufficient statistic:

Theorem 2 (Factorization Theorem)**.** *Let $f(x|\theta)$ be the pdf of X. Then $T(X)$ is a sufficient statistic if and only if there exist functions $g(t|\theta)$ and $h(x)$ such that $f(x|\theta)=g(T(x)|\theta)h(x)$.*

Proof. Let $l(t|\theta)$ be the pdf of $T(X)$.

Suppose $T(X)$ is a sufficient statistic. Then $f_{X|T(X)}(x|T(X)=T(x)) = f_X(x|\theta)/l(T(x)|\theta)$ does not depend on θ. Denote it by $h(x)$. Then $f(x|\theta)=l(T(x)|\theta)h(x)$. Denoting l by g yields the result in one direction.

In the other direction we will give a "sloppy" proof. Denote $A(x)=\{y: T(y)=T(x)\}$. Then

$$l(T(x)|\theta)=\int_{A(x)} f(y|\theta)dy=\int_{A(x)} g(T(y)|\theta)h(y)dy = g(T(x)|\theta)\int_{A(x)} h(y)dy.$$

So

$$\begin{aligned} f_{X|T(X)}(x|T(X)=T(x)) &= \frac{f(x|\theta)}{l(T(x)|\theta)} \\ &= \frac{g(T(x)|\theta)h(x)}{g(T(x)|\theta)\int_{A(x)} h(y)dy} \\ &= \frac{h(x)}{\int_{A(x)} h(y)dy}, \end{aligned}$$

which is independent of θ. We conclude that $T(X)$ is a sufficient statistic. □

Example Let us show how to use the factorization theorem in practice. Let $X_1, ..., X_n$ be a random sample from $N(\mu, \sigma^2)$ where both μ and σ^2 are unknown, i.e. $\theta = (\mu, \sigma^2)$. Then

$$\begin{aligned} f(x|\theta) &= \exp\{-\sum_{i=1}^{n}(x_i - \mu)^2/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n) \\ &= \exp\{-[\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n). \end{aligned}$$

Thus, $T(X) = (\sum_{i=1}^{n} X_i^2, \sum_{i=1}^{n} X_i)$ is a sufficient statistic (here $h(x) = 1$ and g is the whole thing). Note that in this example we actually have a pair of sufficient statistics. In addition, as we have seen before,

$$\begin{aligned} f(x|\theta) &= \exp\{-[\sum_{i=1}^{n}(x_i - \overline{x}_n)^2 + n(\overline{x}_n - \mu)^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n) \\ &= \exp\{-[(n-1)s_n^2 + n(\overline{x}_n - \mu)^2]/(2\sigma^2)\}/((2\pi)^{n/2}\sigma^n). \end{aligned}$$

Thus, $T(X) = (\overline{X}_n, s_n^2)$ is another sufficient statistic. Yet another sufficient statistic is $T(X) = (X_1, ..., X_n)$. Note that $\overline{X}_n$ is not sufficient in this example.

Example A less trivial example: let $X_1, ..., X_n$ be a random sample from $U[\theta, 1+\theta]$. Then $f(x|\theta) = 1$ if $\theta \le \min_i X_i \le \max_i X_i \le 1+\theta$ and 0 otherwise. In other words, $f(x|\theta) = I\{\theta \le X_{(1)}\}I\{1+\theta \ge X_{(n)}\}$. So $T(X) = (X_{(1)}, X_{(n)})$ is sufficient.

3 Minimal Sufficient Statistics

Could we reduce sufficient statistic $T(X)$ in the previous example even more? Suppose we have two statistics, say, $T(X)$ and $T^\star(X)$. We say that $T^\star$ is not bigger than T if there exists some function r such that $T^\star(X) = r(T(X))$. In other words, we can calculate $T^\star(X)$ whenever we know $T(X)$. In this case when T^* changes its value, statistic T must change its value as well. In this sense T^* does not give less of an information reduction than T.

Definition 3. A sufficient statistic $T^\star(X)$ is called *minimal* if for any sufficient statistic $T(X)$ there exists some function r such that $T^\star(X) = r(T(X))$.

Thus, in some sense, the minimal sufficient statistic gives us the greatest data reduction without a loss of information about parameters. The following theorem gives a characterization of minimal sufficient statistics:

Theorem 4. *Let $f(x|\theta)$ be the pdf of X and $T(X)$ be such that, for any x, y, statement $\{f(x|\theta)/f(y|\theta)$ does not depend on $\theta\}$ is equivalent to statement $\{T(x) = T(y)\}$. Then $T(X)$ is minimal sufficient.*

We will leave this statement unproven here.

Example Let us now go back to the example with $X_1, ..., X_n \sim U[\theta, 1+\theta]$. Ratio $f(x|\theta)/f(y|\theta)$ is independent of θ if and only if $x_{(1)} = y_{(1)}$ and $x_{(n)} = y_{(n)}$ which is the case if and only if $T(x) = T(y)$. Therefore $T(X) = (X_{(1)}, X_{(n)})$ is minimal sufficient.

Example Let $X_1, ..., X_n$ be a random sample from the Cauchy distribution with parameter θ, i.e. the distribution with the pdf $f(x|\theta) = 1/(\pi(x-\theta)^2)$. Then $f(x_1, ..., x_n|\theta) = 1/(\pi^n \prod_{i=1}^n (x_i - \theta)^2)$. By the theorem above, $T(X) = (X_{(1)}, ..., X_{(n)})$ is minimal sufficient.

4 Estimators. Properties of estimators.

An estimator is a function of the data (statistic). If we have a parametric family with parameter θ, then an estimator of θ is usually denoted by $\hat{\theta}$.

Example For example, if $X_1, ..., X_n$ is a random sample from some distribution with mean μ and variance σ^2, then sample average $\hat{\mu} = \overline{X}_n$ is an estimator of the population mean, and sample variance $\hat{\sigma}^2 = s^2 = \sum_{i=1}^n (X_i - \overline{X}_n)^2/(n-1)$ is an estimator of the population variance.

4.1 Unbiasness

Let X be our data. Let $\hat{\theta} = T(X)$ be an estimator where T is some function.

We say that $\hat{\theta}$ is *unbiased* for θ if $E_\theta[T(X)] = \theta$ for all possible values of θ where E_θ denotes the expectation when θ is the true parameter value. The *bias* of $\hat{\theta}$ is defined by $\text{Bias}(\hat{\theta}) = E_\theta[\hat{\theta}] - \theta$.

Thus, the concept of unbiasness means that we are on average correct. For example, if X is a random sample $X_1, ..., X_n$ from some distribution with mean μ and variance σ^2, then, as we have already seen, $E[\hat{\mu}] = \mu$ and $E[s^2] = \sigma^2$. Thus, sample average and sample variance are unbiased estimators of population mean and population variance correspondingly.

4.2 Efficiency: MSE

Another of the concepts that evaluates performance of estimators is the MSE (Mean Squared Error). By definition, $\text{MSE}(\hat{\theta}) = E_\theta[(\hat{\theta} - \theta)^2]$. The theorem below gives a useful decomposition for MSE:

Theorem 5. *$MSE(\hat{\theta}) = Bias^2(\hat{\theta}) + V(\hat{\theta})$.*

Proof.

$$\begin{aligned} E[(\hat{\theta} - \theta)^2] &= E[(\hat{\theta} - E[\hat{\theta}] + E[\hat{\theta}] - \theta)^2] \\ &= E[(\hat{\theta} - E[\hat{\theta}])^2 + (E[\hat{\theta}] - \theta)^2 + 2(\hat{\theta} - E[\hat{\theta}])(E[\hat{\theta}] - \theta)] \\ &= V(\hat{\theta}) + \text{Bias}^2(\hat{\theta}) + 2E[\hat{\theta} - E[\hat{\theta}]](E[\hat{\theta}] - \theta) \\ &= V(\hat{\theta}) + \text{Bias}^2(\hat{\theta}). \end{aligned}$$

□

Estimators with smaller MSE are considered to be better, or more *efficient*.

4.3 Connection between efficiency and sufficient statistics

Let $X = (X_1, ..., X_n)$ be a random sample from distribution f_θ. Let $\hat{\theta} = \delta(X)$ be an estimator of θ. Let $T(X)$ be a sufficient statistic for θ. As we have seen already, an MSE provides one way to compare the quality of different estimators. In particular, estimators with smaller MSE are said to be more efficient. On the other hand, once we know $T(X)$, we can discard X. How do these concepts relate to each other? The theorem below shows that for any estimator $\hat{\theta} = \delta(X)$, there is another estimator which depends on data X only through $T(X)$ and is at least as efficient as $\hat{\theta}$:

Theorem 6 (Rao-Blackwell)**.** *In the setting above, define $\phi(T) = E[\delta(X)|T]$. Then $\hat{\theta}_2 = \phi(T(X))$ is an estimator for θ and $MSE(\hat{\theta}_2) \leq MSE(\hat{\theta})$. In addition, if $\hat{\theta}$ is unbiased, then $\hat{\theta}_2$ is unbiased as well.*

Proof. To show that $\hat{\theta}_2$ is an estimator, we have to check that it does not depend on θ. Indeed, since T is sufficient for θ, the conditional distribution of X given T is independent of θ. So the conditional distribution of $\delta(X)$ given T is independent of θ as well. In particular, the conditional expectation $E[\delta(X)|T]$ does not depend on θ. Thus, $\phi(T(X))$ depends only on the data X and $\hat{\theta}_2$ is an estimator.

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \hat{\theta}_2 + \hat{\theta}_2 - \theta)^2] \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + 2E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] + E[(\hat{\theta}_2 - \theta)^2] \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + 2E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] + \text{MSE}(\hat{\theta}_2) \\
&= E[(\hat{\theta} - \hat{\theta}_2)^2] + \text{MSE}(\hat{\theta}_2),
\end{aligned}$$

where in the last line we used

$$\begin{aligned}
E[(\hat{\theta} - \hat{\theta}_2)(\hat{\theta}_2 - \theta)] &= E[(\delta(X) - \phi(T(X)))(\phi(T(X)) - \theta)] \\
&= E[E[(\delta(X) - \phi(T(X)))(\phi(T(X)) - \theta)|T]] \\
&= E[(\phi(T(X)) - \theta)E[(\delta(X) - \phi(T(X)))|T]] \\
&= E[(\phi(T(X)) - \theta) \cdot (E[\delta(X)|T] - \phi(T(X)))] \\
&= 0,
\end{aligned}$$

since $E[\delta(X)|T] = \phi(T(X))$.

To show the last result, we have

$$E[\phi(T(X))] = E[E[\delta(X)|T]] = E[\delta(X)] = \theta$$

by the law of iterated expectations. □

Example Let $X_1, ..., X_n$ be a random sample from Binomial(p, k), i.e. $P\{X_j = m\} = (k!/(m!(k-m)!))p^m(1-p)^{k-m}$ for any integer $m \geq 0$. Suppose our parameter of interest is the probability of one success, i.e. $\theta = P\{X_j = 1\} = kp(1-p)^{k-1}$. One possible estimator is $\hat{\theta} = \sum_{i=1}^n I(X_i = 1)/n$. This estimator is unbiased, i.e. $E[\hat{\theta}] = \theta$. Let us find a sufficient statistic. The joint density of the data is

$$\begin{aligned} f(x_1, ..., x_n) &= \prod_{i=1}^{n} (k!/(x_i!(k-x_i)!))p^{x_i}(1-p)^{k-x_i} \\ &= \text{function}(x_1, ..., x_n)p^{\sum x_i}(1-p)^{nk-\sum x_i}. \end{aligned}$$

Thus, $T = \sum_{i=1}^{n} X_i$ is sufficient. In fact, it is minimal sufficient.

Using the Rao-Blackwell theorem, we can improve $\hat{\theta}$ by considering its conditional expectation given T. Let $\phi = E[\hat{\theta}|T]$ denote this estimator. Then, for any nonnegative integer t,

$$\begin{aligned} \phi(t) &= E[\sum_{i=1}^{n} I(X_i = 1)/n| \sum_{i=1}^{n} X_i = t] \\ &= \sum_{i=1}^{n} P\{X_i = 1| \sum_{j=1}^{n} X_j = t\}/n \\ &= P\{X_1 = 1| \sum_{j=1}^{n} X_j = t\} \\ &= \frac{P\{X_1 = 1, \sum_{j=1}^{n} X_j = t\}}{P\{\sum_{j=1}^{n} X_j = t\}} \\ &= \frac{P\{X_1 = 1, \sum_{j=2}^{n} X_j = t-1\}}{P\{\sum_{j=1}^{n} X_j = t\}} \\ &= \frac{P\{X_1 = 1\}P\{\sum_{j=2}^{n} X_j = t-1\}}{P\{\sum_{j=1}^{n} X_j = t\}} \\ &= \frac{kp(1-p)^{k-1} \cdot (k(n-1))!/((t-1)!(k(n-1)-(t-1))!)p^{t-1}(1-p)^{k(n-1)-(t-1)}}{(kn)!/(t!(kn-t)!)p^t(1-p)^{kn-t}} \\ &= \frac{k(k(n-1))!/((t-1)!(k(n-1)-(t-1))!)}{(kn)!/(t!(kn-t)!)} \\ &= \frac{k(k(n-1))!(kn-t)!t}{(kn)!(kn-k+1-t)!} \end{aligned}$$

where we used the fact that X_1 is independent of $(X_2, ..., X_n)$, $\sum_{i=1}^{n} X_i \sim \text{Binomial}(kn, p)$, and $\sum_{i=2}^{n} X_i \sim \text{Binomial}(k(n-1), p)$. So our new estimator is

$$\hat{\theta}_2 = \phi(X_1, ..., X_n) = \frac{k(k(n-1))!(kn - \sum_{i=1}^{n} X_i)!(\sum_{i=1}^{n} X_i)}{(kn)!(kn-k+1-\sum_{i=1}^{n} X_i)!}.$$

By the theorem above, it is unbiased and at least as efficient as $\hat{\theta}$. The procedure we just applied is sometimes informally referred to as Rao-Blackwellization.

Note on implementation. One may say “this is too complicated”. We have derived $\phi(t) = E(\hat{\theta}| \sum_{i=1}^{n} X_i = t)$ analytically in order to calculate a new estimate $\hat{\theta}_2 = \phi(T) = \phi(\sum_{i=1}^{n} X_i)$, but in real life you may just do this with Monte-Carlo simulations. Note, that we do not need to calculate the whole $\phi(t)$ function we need only $\phi(T)$, that is evaluated in one point (the realized value of T). Note also, that the result does not depend on p, so we are free to choose any p (choosing p close to $T/(kn)$ will give faster calculations).

Choose some $p \in (0,1)$. For $b = 1, ..., B$ repeat the following:

- Draw $X_{1b}^*, ..., X_{nb}^*$ as independent variables from Binomial (p, k), if $\sum_{i=1}^n X_{ib}^* \neq T$, discard this sample. Repeat drawing samples until you get $\sum_{i=1}^n X_{ib}^* = T$.
- Calculate $Y_b = \frac{1}{n} \sum_{i=1}^n I(X_{ib}^* = 1)$.

The new estimator is $\hat{\theta}_2 \approx \frac{1}{B} \sum_{b=1}^B Y_b$. The accuracy is better for larger number of simulation B.