Sufficient Statistic

- Intuitively, a sufficient statistic for a parameter is a statistic that captures all the information about a given parameter contained in the sample.
- Sufficiency Principle: If $T(\boldsymbol{X})$ is a sufficient statistic for θ, then any inference about θ should depend on the sample $\boldsymbol{X}$ only through the value of $T(\boldsymbol{X})$.
- That is, if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $T(\boldsymbol{x}) = T(\boldsymbol{y})$, then the inference about θ should be the same whether $\boldsymbol{X} = \boldsymbol{x}$ or $\boldsymbol{X} = \boldsymbol{y}$.
- Definition: A statistic $T(\boldsymbol{x})$ is a sufficient statistic for θ if the conditional distribution of the sample $\boldsymbol{X}$ given $T(\boldsymbol{x})$ does not depend on θ.

- Definition: Let X_1, X_2, … X_n denote a random sample of size n from a distribution that has a pdf $f(x,\theta)$, $\theta \varepsilon \Omega$.

  Let $Y_1=u_1(X_1, X_2, \ldots X_n)$ be a statistic whose pdf or pmf is $f_{Y1}(y_1,\theta)$. Then Y_1 is a sufficient statistic for θ if and only if

$$\frac{f\left(x_1;\theta\right) f\left(x_2;\theta\right)\cdots f\left(x_n;\theta\right)}{f_{Y_1}\left[u_1\left(x_1, x_2, \ldots x_n\right);\theta\right]} = H\left(x_1, x_2, \ldots x_n\right)$$

- Example: Normal sufficient statistic:

  Let X_1, X_2, … X_n be independently and identically distributed $N(\mu,\sigma^2)$ where the variance is known. The sample mean

$$T(\underline{X}) = \bar{X} = {}^{1}\!/_{n} \sum_{i=1}^{n} X_i$$

is the sufficient statistic for μ.

- Starting with the joint distribution function

$$f\left(\underline{x}\middle|\mu\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right]$$

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

- Next, we add and subtract the sample average yielding

$$f\left(\underline{x}\middle|\mu\right)=\frac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\sum_{i=1}^{n}\frac{\left(x_i-\bar{x}+\bar{x}-\mu\right)^2}{2\sigma^2}\right]$$

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

- Where the last equality derives from

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(\bar{x}-\mu\right)=\left(\bar{x}-\mu\right)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)=0$$

- Given that the distribution of the sample mean is

$$q\left(T\left(\underline{X}\right)\middle|\theta\right)=\frac{1}{\left(2\pi\,\sigma^2/n\right)^{1/2}}\exp\left[-\frac{n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]$$

- The ratio of the information in the sample to the information in the statistic becomes

$$\frac{f\left(\underline{x}\middle|\theta\right)}{q\left(T\left(\underline{x}\right)\middle|\theta\right)}=\frac{\dfrac{1}{\left(2\pi\sigma^2\right)^{n/2}}\exp\left[-\dfrac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]}{\dfrac{1}{\left(2\pi\,\sigma^2/n\right)^{1/2}}\exp\left[-\dfrac{n\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right]}$$

$$\frac{f\left(\underline{x}\middle|\theta\right)}{q\left(T\left(\underline{x}\right)\middle|\theta\right)} = \frac{1}{n^{1/2}\left(2\pi\sigma^2\right)^{n-1/2}} \exp\left[-\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{2\sigma^2}\right]$$

which is a function of the data X_1, X_2, … X_n only, and does not depend on μ. Thus we have shown that the sample mean is a sufficient statistic for μ.

- Theorem (**Factorization Theorem**) Let $f(\boldsymbol{x}|\theta)$ denote the joint pdf or pmf of a sample $\boldsymbol{X}$. A statistic $T(\boldsymbol{X})$ is a sufficient statistic for θ if and only if there exists functions $g(t|\theta)$ and $h(\boldsymbol{x})$ such that, for all sample points *x* and all parameter points θ

$$f\left(\underset{\sim}{x}\middle|\theta\right)=g\left(T\left(\underset{\sim}{x}\right)\middle|\theta\right)h\left(\underset{\sim}{x}\right)$$

Posterior Distribution Through Sufficient Statistics

Theorem: The **posterior distribution** depends only on **sufficient statistics.**

Proof: let T(**X**) be a sufficient statistic for θ, then

$$f(\mathbf{x}\,|\,\theta) = f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})$$

$$f(\theta\,|\,\mathbf{x}) = \frac{f(\theta)f(\mathbf{x}\,|\,\theta)}{\int f(\theta)f(\mathbf{x}\,|\,\theta)d\theta} = \frac{f(\theta)f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})}{\int f(\theta)f(T(\mathbf{x})\,|\,\theta)H(\mathbf{x})d\theta}$$

$$= \frac{f(\theta)f(T(\mathbf{x})\,|\,\theta)}{\int f(\theta)f(T(\mathbf{x})\,|\,\theta)d\theta} = f(\theta\,|\,T(\mathbf{x}))$$

Posterior Distribution Through Sufficient Statistics

Example: Posterior for Normal distribution mean (with known variance)

Now, instead of using the entire sample, we can derive the posterior distribution using the sufficient statistic

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

Exercise: Please derive the posterior distribution using this approach.