Chapter 6
Principle of Data Deduction

Sufficiency

Outline

Introduction

Even a fairly simple experiment can have an enormous number of outcomes. For example, flip a coin 333 times. Then the number of outcomes is more than a google (10^{100}) - a number at least 100 quintillion times the number of elementary particles in the known universe. Our parameter estimate or hypothesis may be sufficiently addressed by an analysis that does not considers separately every possible outcome but rather some simpler concept like the number of heads or the longest run of tails.

Definition. A statistic $T(\mathbf{X})$ is a function of the sample $\mathbf{X} = (X_1 \ldots, X_n)$. The probability distribution of a $T(\mathbf{X})$ is called the sampling distribution.

A statistic can not be a function of the unknown parameter.

Thus, if $X_1 \ldots, X_n \sim N(\mu, 1)$ are independent, then $\bar{X} + 1$ *is* a statistic and $\bar{X} + \mu$ *is not* a statistic

Statistics

Exercise. For $\mathbf{X} = (X_1, \ldots, X_n)$, name some statistics.

$$\sum_{i=1}^{n} X_i,\ \bar{X},\ X_{(k)},\ S^2,\ X_1X_2 + 3X_3,\ \prod_{i=1}^{n} \sin(X_i),\ \frac{1}{n}\sum_{i=1}^{n}((X_i - \bar{X})/S)^3,$$

Denote

- $\mathcal{X}$ - the state space for $\mathbf{X}$.
- $\mathcal{T}$ - the image of $\mathcal{X}$ under T, i.e.,

$$\mathcal{T} = \{t; t = T(\mathbf{x}) \text{ for some } \mathbf{x} \in \mathcal{X}\}.$$

- $A_t = \{\mathbf{x}; T(\mathbf{x}) = t\}$.

Statistics

Any function T induces an equivalence relation, $\sim$, on its domain $\mathcal{X}$, namely,

$$\mathbf{x}_1 \sim \mathbf{x}_2 \text{ if and only if } T(\mathbf{x}_1) = T(\mathbf{x}_2).$$

Exercise. Check that $\sim$ is an equivalence relation.
Show reflexivity, symmetry, transitivity

The equivalence relation induces a partition of $\mathcal{X}$,

$$\mathcal{A} = \{A_t; t \in \mathcal{T}\}$$

.

Exercise. Let $\mathbf{X} = (X_1, X_2, X_3)$ be Bernoulli trials and $T(\mathbf{x}) = x_1 + x_2 + x_3$. Define $\mathcal{X}$, $\mathcal{T}$, and $\mathcal{A}$.

$$\mathcal{X} = \{0,1\}^3, \quad \mathcal{T} = \{0,1,2,3\}$$

.

Statistics

Exercise. Let S be a one-to-one mapping from $\mathcal{T}$. What is the resulting induced partition of $S \circ T$? $\mathcal{A}$ again.

Desired properties for $T(\mathbf{X})$:

- Contains all the information necessary for parameter estimation from $\mathbf{X} = (X_1 \ldots, X_n)$. (related to sufficiency)
- Summarizes information in the most parsimoniously way possible (related to minimal sufficiency)
- The dimension of the image of $\dim \operatorname{image}(T(\mathbf{X})) < n$, the sample size. (guaranteed for exponential families)

Conditional Probabilities

For $\mathbf{X}$, a discrete random vector,

$$\begin{aligned} f_{\mathbf{X}|T(\mathbf{X})}(x|t) &= \frac{P\{\mathbf{X}=\mathbf{x}, T(\mathbf{X})=t\}}{P\{T(\mathbf{X})=t\}} I_{\{t\}}(T(\mathbf{x})) \\ &= \frac{P\{\mathbf{X}=\mathbf{x}, T(\mathbf{X})=T(\mathbf{x})\}}{P\{T(\mathbf{X})=T(\mathbf{x})\}} I_{\{t\}}(T(\mathbf{x})) \\ &= \frac{P\{\mathbf{X}=\mathbf{x}\}}{P\{T(\mathbf{X})=T(\mathbf{x})\}} I_{\{t\}}(T(\mathbf{x})) \\ &= \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{T(\mathbf{X})}(T(\mathbf{x}))} I_{\{t\}}(T(\mathbf{x})) \end{aligned}$$

Note that $\{\mathbf{X}=\mathbf{x}\} \subset \{T(\mathbf{X})=T(\mathbf{x})\}$.

The corresponding formula holds for $\mathbf{X}$, a continuous random vector.

Definition of Sufficiency

Now consider a family of distributions indexed by a parameter $\theta \in \Theta$ and rewrite the formula on the previous slide for the parameter set.

$$f_{\mathbf{X}|T(\mathbf{x})}(x|T(\mathbf{x}),\theta) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{f_{T(\mathbf{x})}(T(\mathbf{x})|\theta)} = h_\theta(\mathbf{x})$$

Definition. T is sufficient if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X})$ does not depend on θ.

In other words, the function $h_\theta(\mathbf{x})$ does not depend on θ and given $T(\mathbf{X}) = T(\mathbf{x})$, one can generate the conditional distribution of $\mathbf{X}$ without any knowledge of the parameter θ. Thus, if T is sufficient, then

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = f_{T(\mathbf{x})}(T(\mathbf{x})|\theta)h(\mathbf{x}).$$

So all the information about θ is contained in the distribution of $T(\mathbf{x})$.

Example

Example. Let $\mathbf{X}$ be n independent $Ber(\theta)$ random variables and set $T(\mathbf{x}) = \sum_{i=1}^n x_i$. Then,

$$f_{\mathbf{X}|T(\mathbf{x})}(x|T(\mathbf{x}),\theta) = \frac{P_\theta\{\mathbf{X} = \mathbf{x}\}}{P_\theta\{T(\mathbf{X}) = T(\mathbf{x})\}} = \frac{\theta^t(1-\theta)^{n-T(\mathbf{x})}}{\binom{n}{T(\mathbf{x})}\theta^t(1-\theta)^{n-T(\mathbf{x})}} = \binom{n}{T(\mathbf{x})}^{-1} = h(\mathbf{x}),$$

a distribution that is uniformly distributed on the $\binom{n}{T(\mathbf{x})}$ sequences for the given value of $T(\mathbf{x})$ and, in particular, does not depend on θ. Also,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= f_{T(\mathbf{x})}(T(\mathbf{x})|\theta)h(\mathbf{x}) = \binom{n}{T(\mathbf{x})}\theta^{T(\mathbf{x})}(1-\theta)^{n-T(\mathbf{x})} \cdot \binom{n}{T(\mathbf{x})}^{-1} \\ &= \theta^{T(\mathbf{x})}(1-\theta)^{n-T(\mathbf{x})}. \end{aligned}$$

Example

For a simulation,

```
> n<-25;t<-rbinom(1,n,0.6)
> sample(c(rep(0,n-t),rep(1,t)))
 [1] 0 0 1 1 0 0 0 0 1 0 0 1 0 1 1 0 1 1 1 0 1 1 1 1 1
> sample(c(rep(0,n-t),rep(1,t)))
 [1] 0 0 1 1 0 1 1 0 1 1 1 1 0 0 0 1 1 0 1 1 0 0 0 1 1
```

The order statistics are always sufficient. In fact, the empirical cumulative distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i).$$

is sufficient and is equivalent to knowing the order statistic.

Sufficiency

Notice that if we can choose $T(\mathbf{x})$ so that the ratio

$$f_{\mathbf{X}}(\mathbf{x}|\theta) \Big/ f_{T(\mathbf{x})}(T(\mathbf{x})|\theta)$$

does not depend on θ, then $T(\mathbf{x})$ is a sufficient statistic.

Example. For $\mathbf{X} = (X_1, \ldots, X_n)$ independent $\Gamma(\alpha_0, \beta)$, α_0 known, $T(\mathbf{X}) = \sum_{i=1}^n X_i \sim \Gamma(n\alpha_0, \beta)$. Then, the joint density

$$f_{\mathbf{X}}(\mathbf{x}|\beta) = \frac{\beta^{\alpha_0}}{\Gamma(\alpha_0)} x_1^{\alpha_0 - 1} e^{-\beta x_1} \cdots \frac{\beta^{\alpha_0}}{\Gamma(\alpha_0)} x_n^{\alpha_0 - 1} e^{-\beta x_n} = \frac{\beta^{n\alpha_0}}{\Gamma(\alpha_0)^n} \left(\prod_{i=1}^n x_i \right)^{\alpha_0 - 1} e^{-\beta T(\mathbf{x})}$$

$$f_{T(\mathbf{x})}(T(\mathbf{x})|\beta) = \frac{\beta^{n\alpha_0}}{\Gamma(n\alpha_0)} T(\mathbf{x})^{n\alpha_0 - 1} e^{-\beta T(\mathbf{x})}$$

The ratio does not depend on β.

$$\frac{\Gamma(n\alpha_0)}{\Gamma(\alpha_0)^n} \frac{\left(\prod_{i=1}^n x_i\right)^{\alpha_0 - 1}}{T(\mathbf{x})^{n\alpha_0 - 1}}.$$

Neyman-Fisher Factorization Theorem

Theorem. Neyman-Fisher Factorization Theorem. The statistic T is sufficient for the parameter θ if and only if functions g and $\mathbf{h}$ can be found such that

$$\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))$$

The central idea in proving this theorem can be found in the case of discrete random variables.

Proof. Because T is a function of $\mathbf{x}$,

$$\mathbf{f}_X(\mathbf{x}|\theta) = f_{X,T(X)}(\mathbf{x}, T(\mathbf{x})|\theta) = \mathbf{f}_{X|T(X)}(\mathbf{x}|T(\mathbf{x}), \theta)f_{T(X)}(T(\mathbf{x})|\theta).$$

If we assume that T is sufficient, then $\mathbf{f}_{X|T(X)}(\mathbf{x}|T(\mathbf{x}), \theta)$ is not a function of θ and we can set it to be $\mathbf{h}(\mathbf{x})$. The second term is a function of $T(\mathbf{x})$ and θ. We will write it $g(\theta, T(\mathbf{x}))$.

Neyman-Fisher Factorization Theorem

If we assume the factorization

$$\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x}) g(\theta, T(\mathbf{x}))$$

then, by the definition of conditional expectation,

$$P_\theta\{X = \mathbf{x} | T(X) = t\} = \frac{P_\theta\{X = \mathbf{x}, T(X) = t\}}{P_\theta\{T(X) = t\}}.$$

$$\mathbf{f}_{X|T(X)}(\mathbf{x}|t, \theta) = \frac{\mathbf{f}_{X,T(X)}(\mathbf{x}, t|\theta)}{f_{T(X)}(t|\theta)}.$$

The numerator is 0 if $T(\mathbf{x}) \neq t$ and is

$$\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x}) g(\theta, t)$$

otherwise.

Neyman-Fisher Factorization Theorem

The denominator

$$f_{T(X)}(t|\theta) = \sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{f}_X(\tilde{\mathbf{x}}|\theta) = \sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t} \mathbf{h}(\tilde{\mathbf{x}})g(\theta,t).$$

The ratio

$$\begin{aligned} \mathbf{f}_{X|T(X)}(\mathbf{x}|t,\theta) &= \frac{\mathbf{f}_{X,T(X)}(\mathbf{x},t|\theta)}{f_{T(X)}(t|\theta)} = \frac{\mathbf{h}(\mathbf{x})g(\theta,t)}{\sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t}\mathbf{h}(\tilde{\mathbf{x}})g(\theta,t)} \\ &= \frac{\mathbf{h}(\mathbf{x})}{\sum_{\tilde{\mathbf{x}}:T(\tilde{\mathbf{x}})=t}\mathbf{h}(\tilde{\mathbf{x}})}, \end{aligned}$$

which is independent of θ and, therefore, T is sufficient.

Neyman-Fisher Factorization Theorem

Example. Let $\mathbf{X} = (X_1, \dots, X_n)$ be independent $Pois(\lambda)$ random variables. Then,

$$f_{\mathbf{X}}(\mathbf{x}|\lambda) = \frac{\lambda^{x_1}}{x_1!}e^{-\lambda} \cdots \frac{\lambda^{x_n}}{x_n!}e^{-\lambda} = \frac{e^{-n\lambda}}{x_1! \cdots x_n!}\lambda^{T(\mathbf{x})}$$

and $T(\mathbf{x}) = \sum_{i=1}^n x_i$ is sufficient.

Example. Let $\mathbf{X} = (X_1, \dots, X_n)$ be independent $Beta(\alpha, \beta)$ random variables. Then,

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\alpha, \beta) &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x_1^{\alpha-1}(1-x_1)^{\beta-1} \cdots \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x_n^{\alpha-1}(1-x_n)^{\beta-1} \\ &= \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right)^n T_1(\mathbf{x})^{\alpha-1} T_2(\mathbf{x})^{\beta-1} \end{aligned}$$

and $T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = (\prod_{i=1}^n x_i, \prod_{i=1}^n (1 - x_i))$ is sufficient.

Neyman-Fisher Factorization Theorem

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent $Unif(0, \theta)$ random variables. Then,

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\lambda) &= \frac{1}{\theta} I_{[0,\theta]}(x_1) \cdots \frac{1}{\theta} I_{[0,\theta]}(x_n) \\
&= \begin{cases} \frac{1}{\theta^n} & \text{if all } x_i \leq \theta \\ 0 & \text{if some } x_i > \theta \end{cases} \\
&= \frac{1}{\theta^n} I_{[0,\theta]}(T(\mathbf{x}))
\end{aligned}
$$

and $T(\mathbf{x}) = \max_i x_i$ is sufficient.

Neyman-Fisher Factorization Theorem

Example. Let $\mathbf{X} = (X_1, \dots, X_n)$ be independent $N(\mu, \sigma^2)$ random variables.

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\lambda) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_1-\mu)^2}{2\sigma^2}\right) \cdots \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_n-\mu)^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k-\mu)^2\right) \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{k=1}^{n}(x_k^2 - 2\mu x_k + \mu^2)\right) \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{n}{2\sigma^2}(\overline{x^2} - 2\bar{x} + \mu^2)\right) \\
&= \frac{1}{(\sigma\sqrt{2\pi})^n} \exp\left(-\frac{n}{2\sigma^2} T_2(\mathbf{x}) - 2T_1(\mathbf{x}) + \mu^2)\right)
\end{aligned}
$$

and $T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = (\bar{x}, \overline{x^2})$ is sufficient.

Neyman-Fisher Factorization Theorem

Example. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be independent random variables from an exponential family, the probability density functions can be expressed in the form

$$\mathbf{f}_X(\mathbf{x}|\eta) = \prod_{j=1}^{n} h(x_i) \cdot \exp\left(\sum_{j=1}^{n} \langle \eta, \mathbf{t}(x_j) \rangle\right) e^{-nA(\eta)}, \quad x \in S.$$

Then, take

$$\mathbf{h}(\mathbf{x}) = \prod_{j=1}^{n} h(x_i)$$

and $T(\mathbf{x}) = \sum_{j=1}^{n} \mathbf{t}(x_j)$ is sufficient.

Transformations of Sufficient Statistics

If T is sufficient for θ, and $U = u(T)$ for u one-to-one, then U is also sufficient.

Examples.

- If $T(\mathbf{x}) = \sum_{i=1}^{n} x_i$ is sufficient then so is $U(\mathbf{x}) = \bar{x}$.
- If $T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = (\prod_{i=1}^{n} x_i, \prod_{i=1}^{n}(1 - x_i))$ is sufficient then so is $U(\mathbf{x}) = (U_1(\mathbf{x}), U_2(\mathbf{x})) = (\overline{\ln x}, \overline{\ln(1-x)})$.
- If $T(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x})) = (\bar{x}, \overline{x^2})$ is sufficient then so is $U(\mathbf{x}) = (U_1(\mathbf{x}), U_2(\mathbf{x})) = (\bar{x}, s^2)$.

If T is sufficient for θ, and $T = c(U)$ a function of some other statistic U, then U, is also sufficient.

- If $T(\mathbf{x}) = U_1(\mathbf{x}) + U_2(\mathbf{x})$ is sufficient then so is $U(\mathbf{x}) = (U_1(\mathbf{x}), U_2(\mathbf{x}))$.
- If $T(\mathbf{x})$ is sufficient then so is $U(\mathbf{x}) = (T(\mathbf{x}), U_2(\mathbf{x}))$ for any statistic U_2.

Bayesian Sufficiency

Definition. T is Bayesian sufficient for every prior density π, there exist posterior densities $f_{\Psi|\mathbf{X}}$ and $f_{\Psi|T(\mathbf{X})}$ so that

$$f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) = f_{\Psi|T(\mathbf{X})}(\theta|T(\mathbf{x})).$$

The posterior density is a function of the sufficient statistic.

Theorem. If T is sufficient in the classical sense, then T is sufficient in the Bayesian sense.

Proof. Bayes formula states the posterior density

$$f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) = \frac{f_{\mathbf{X}}(\mathbf{x}|\theta)\pi(\theta)}{\int_\Theta f_{\mathbf{X}}(\mathbf{x}|\psi)\pi(\psi)\nu(d\psi)}$$

By the Neyman-Fisher factorization theorem, $\mathbf{f}_X(\mathbf{x}|\theta) = \mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))$

$$\begin{aligned} f_{\Psi|\mathbf{X}}(\theta|\mathbf{x}) &= \frac{\mathbf{h}(\mathbf{x})g(\theta, T(\mathbf{x}))\pi(\theta)}{\int_\Theta \mathbf{h}(\mathbf{x})g(\psi, T(\mathbf{x}))\pi(\psi)\nu(d\psi)} \\ &= \frac{g(\theta, T(\mathbf{x}))\pi(\theta)}{\int_\Theta g(\theta, T(\mathbf{x}))\pi(\psi)\nu(d\psi)} = f_{\Psi|T(\mathbf{X})}(\theta|T\mathbf{x})) = f_{\Psi|T(\mathbf{X})}(\theta|T(\mathbf{x})) \end{aligned}$$