SUFFICIENT STATISTICS

Suppose the distribution of a random variable $\underline{X}$ is determined by a parameter $\underline{\theta}$:

$$\underline{X} \sim f_{\underline{\theta}}(\underline{x})$$

The functional form of f is known, but $\underline{\theta}$ is unknown.

[Note: We will use f to denote a pdf, p to denote a pmf, and f when it could be either.]

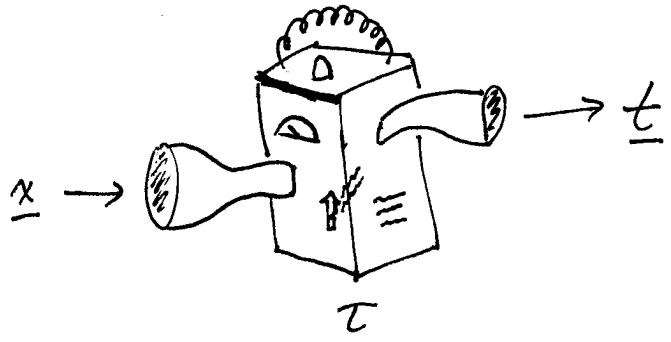
In statistical inference, we observe a realization $\underline{x}$ of $\underline{X}$ and need to answer some question about $\underline{\theta}$, such as

- Is $\underline{\theta} \in \mathcal{H}_1$ or is $\underline{\theta} \in \mathcal{H}_2$? (Detection)
- What is a good guess for $\underline{\theta}$? (Estimation)

If $\underline{x} = [x_1 \cdots x_N]^T$ and $\underline{\theta} = [\theta_1 \cdots \theta_p]^T$ where $p < N$, one might wonder whether it is possible to <u>compress</u> the measurement $\underline{x}$ into a low-dimensional statistic without affecting the quality of the inference about $\underline{\theta}$.

In other words, does there exist $\underline{T} = \tau(\underline{x})$, where the dimension of $\underline{T}$ is $M < N$, such that $\underline{T}$ carries all the useful information about $\underline{\theta}$?

If so, for the purpose of studying $\underline{\theta}$, we could discard the raw measurement $\underline{x}$ and retain only the compressed statistic $\underline{t}$.



Definition Let $\underline{X} \sim f_{\underline{\theta}}(\underline{x})$. The statistic

$\underline{T} = \tau(\underline{X})$ is a <u>sufficient statistic</u> for $\underline{\theta}$ if the conditional distribution of $\underline{X}$ given $\underline{T}$ is independent of $\underline{\theta}$. Equivalently, the functional form of $f(\underline{x}|\underline{t})$ does not involve $\underline{\theta}$.

Interpretations

1. Let $P_{\underline{\theta}}(\underline{x}, \underline{t})$ denote the joint pmf of $(\underline{X}, \underline{T})$. Then

$$P_{\underline{\theta}}(\underline{x}, \underline{t}) = \begin{cases} P_{\underline{\theta}}(\underline{x}) & \text{if } \underline{t} = \tau(\underline{x}) \\ 0 & \text{otherwise} \end{cases}$$

Therefore

$$\begin{aligned} P_{\underline{\theta}}(\underline{x}) &= P_{\underline{\theta}}(\underline{x}, \tau(\underline{x})) \\ &= P_{\underline{\theta}}(\underline{x} \,|\, \tau(\underline{x})) P_{\underline{\theta}}(\tau(\underline{x})) \\ &= p(\underline{x} \,|\, \tau(\underline{x})) P_{\underline{\theta}}(\tau(\underline{x})) \end{aligned}$$

$\Rightarrow$ the dependence of $P_{\underline{\theta}}(\underline{x})$ on θ is manifested entirely in $P_{\underline{\theta}}(\underline{t})$.

[continuous case requires more care, but same conclusion holds – see Scharf]

2. Given $\underline{t} = \tau(\underline{x})$, full knowledge of $\underline{x}$ brings no additional information about $\underline{\theta}$.

3. Any inference strategy based on $f_{\underline{\theta}}(\underline{x})$ may be replaced by a strategy based on $f_{\underline{\theta}}(\underline{t})$.

Example] Bernoulli trials

Suppose we observe $\underline{x} = [x_1, \cdots x_N]^T$ where

$$X_i \overset{iid}{\sim} \text{Bernoulli}(\theta)$$

$\theta \in [0,1]$ is unknown.

Recall

(a) $$p_\theta(x) =$$

Since we can assume $x_i \in \{0,1\}$, we may write

$$p_\theta(x_i) = \theta^{x_i}(1-\theta)^{1-x_i}$$

Therefore

$$p_\theta(\underline{x}) = \prod_{i=1}^{n} p_\theta(x_i)$$

$$= \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^k (1-\theta)^{N-k}$$

where $k = \sum_{i=1}^{n} x_i$.

Claim: K is a sufficient statistic for θ.

We must show $p_\theta(\underline{x} \mid k)$ is independent of θ.

From interpretation #1 we know

$$p_\theta(\underline{x} \mid k) = \frac{P_\theta(\underline{x})}{P_\theta(k)}.$$

Exercise | Complete this argument to establish that K is sufficient for θ.

Solution | K is a Binomial (N, θ) random variable.

Therefore

$$P_\theta(k) = \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

and

$$p(\underline{x} \mid k) = \frac{P_\theta(\underline{x})}{P_\theta(k)}$$

$$= \frac{\theta^k (1-\theta)^{N-k}}{\binom{N}{k} \theta^k (1-\theta)^{N-k}}$$

$$= \frac{1}{\binom{N}{k}}$$

which is independent of θ.

The Fisher-Neyman Factorization Theorem

In the previous example, we had to guess the sufficient statistic and work out the conditional pmf by hand. In general, it is difficult to verify the definition of sufficient statistic directly.

The following theorem allows us to identify and verify sufficient statistics more readily, and can be taken as a working definition of sufficiency.

Theorem Let $f_{\underline{\theta}}(\underline{x})$ be the density or mass function for $\underline{X}$. The statistic $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$ iff there exist functions $g_{\underline{\theta}}(\underline{t})$ and $h(\underline{x})$ such that

$$f_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x})) \cdot h(\underline{x})$$

Note: h is independent of θ.

Example Bernoulli trials revisited

$$P_\theta(\underline{x}) = \prod_{i=1}^{N} \theta^{x_i}(1-\theta)^{1-x_i}$$

$$= \theta^k (1-\theta)^{N-k}$$

$$= g_\theta(k) \cdot h(\underline{x})$$

where

$$g_\theta(k) = \theta^k (1-\theta)^{N-k}$$

$$h(\underline{x}) = 1$$

$\Longrightarrow$ K is sufficient for θ.

Proof of Theorem

We will assume $\underline{X}$ is discrete. The continuous case is slightly more involved - see Scharf or Kay, vol I.

First, assume $\underline{T}$ is sufficient for $\underline{\theta}$. Recall

$$P_{\underline{\theta}}(\underline{x}) = p(\underline{x}/\tau(\underline{x})) \cdot P_{\underline{\theta}}(\tau(\underline{x}))$$

Now take

$$g_{\underline{\theta}}(\underline{t}) = P_{\underline{\theta}}(\underline{t})$$

$$h(\underline{x}) = p(\underline{x}/\tau(\underline{x}))$$ ← Independent of $\underline{\theta}$ by sufficiency

For the other direction, assume $p_{\underline{\theta}}(\underline{x})$ may be written

$$p_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x}))h(\underline{x}).$$

We need to show $\underline{T} = \tau(\underline{X})$ is sufficient for $\underline{\theta}$. That is, we need to show $p_{\underline{\theta}}(\underline{x} \mid \underline{t})$ is independent of $\underline{\theta}$.

Again, we will rely on the identity

$$p_{\underline{\theta}}(\underline{x} \mid \underline{t}) = \frac{p_{\underline{\theta}}(\underline{x})}{p_{\underline{\theta}}(\underline{t})}.$$

Since $\underline{X}$ and $\underline{T}$ are discrete, we have

$$p_{\underline{\theta}}(\underline{t}) = \sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} p_{\underline{\theta}}(\underline{x}').$$

Therefore

$$p_{\underline{\theta}}(\underline{x} \mid \underline{t}) = \frac{g_{\underline{\theta}}(\underline{t}) \cdot h(\underline{x})}{\sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} g_{\underline{\theta}}(\underline{t}) h(\underline{x}')}$$

$$= \frac{h(\underline{x})}{\sum_{\underline{x}' : \tau(\underline{x}') = \underline{t}} h(\underline{x}')} \leftarrow \boxed{\text{Independent of } \underline{\theta}}$$

∎

Note The FNFT gives us a formula for $P(\underline{x}|\underline{t})$, namely

$$P(\underline{x}|\underline{t}) = \frac{h(\underline{x})}{\sum_{\underline{x}': \tau(\underline{x}') = \underline{t}} h(\underline{x}')}$$

Example Bernoulli trials, part III

$h(\underline{x}) = 1$, so

$$p(\underline{x}|k) = \frac{1}{\sum_{\underline{x}': \sum_{i=1}^{N} x_i' = k} 1}$$

$$= \frac{1}{\#\{\underline{x}': \sum_{i=1}^{N} x_i' = k\}}$$

$$= \frac{1}{\binom{N}{k}}$$

Let look at an example of the FNFT for the continuous case:

Example] Gaussian with unknown mean

We are given $\underline{x} = [x_1 \cdots x_N]^T$ where

$$x_i \overset{iid}{\sim} N(\theta, \sigma^2)$$

and σ^2 is known.

$$f_\theta(\underline{x}) = \prod_{i=1}^{N} f_\theta(x_i)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\theta)^2}{2\sigma^2}\right\}$$

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

$$= (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-1}{2\sigma^2}\left[\sum_{i=1}^{N} x_i^2 - 2\theta\sum_{i=1}^{N} x_i + N\theta^2\right]\right\}$$

(b)

where $t =$ $\Rightarrow$ t is sufficient for θ.

Example Gaussian w/ unknown mean and variance.

Now assume

$$x_i \overset{iid}{\sim} N(\theta_1, \theta_2), \quad i = 1, \ldots, N$$

where $\underline{\theta} = [\theta_1 \ \theta_2]^T$ is unknown. Then

$$f_{\underline{\theta}}(\underline{x}) = \underbrace{\left(\frac{1}{2\pi\theta_2}\right)^{\frac{N}{2}} \exp\left\{-\frac{1}{2\theta_2}\left[\sum_{i=1}^{N} x_i^2 - 2\theta_1 \sum_{i=1}^{N} x_i + N\theta_1^2\right]\right\}}_{g_{\underline{\theta}}(\underline{t})} \cdot \underbrace{1}_{h(\underline{x})}$$

where $\underline{t} = \left[\sum_{i=1}^{N} x_i \quad \sum_{i=1}^{N} x_i^2\right]^T$ is sufficient.

If an invertible function is applied to a sufficient statistic, the result is again a sufficient statistic.

For example, in the Gaussian iid model:

- $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is sufficient for θ_1
- $\left[\bar{x} \quad s^2\right]^T$ is sufficient for $[\theta_1, \theta_2]^T$

  ↳ $= \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$

The Rao-Blackwell Theorem

The power and importance of sufficient statistics is reflected in the following famous result:

Theorem Let $\underline{X}$ be a random variable with pdf/pmf $f_{\underline{\theta}}(\underline{x})$ and let $\underline{T} = \tau(\underline{X})$ be a sufficient statistic. Let $\hat{\underline{\theta}}_1(\underline{x})$ be an estimator of $\underline{\theta}$ and define the mean-squared error

$$MSE(\hat{\underline{\theta}}_1) := E\left[\left\|\hat{\underline{\theta}}_1(\underline{X}) - \underline{\theta}\right\|^2\right].$$

Next define

$$\hat{\underline{\theta}}_2(\underline{x}) = E\left[\hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x})\right]$$

Then

$$MSE(\hat{\underline{\theta}}_2) \leq MSE(\hat{\underline{\theta}}_1)$$

with equality iff

$$\hat{\underline{\theta}}_1(\underline{X}) = \hat{\underline{\theta}}_2(\underline{X})$$

with probability one (almost surely).

Observations / Interpretations

1. $\hat{\underline{\theta}}_2$ is a function of the sufficient statistic.

2. Given any estimator $\hat{\underline{\theta}}_1$ that is <u>not</u> a function of a sufficient statistic, there exists a better estimator (with respect to MSE).

3. We may restrict our search for estimators to functions of a sufficient stat.

4. The conditional expectation

$$E\left[\hat{\underline{\theta}}_1(\underline{X}) \mid \underline{T} = \tau(\underline{x})\right]$$

averages out (or removes) non-informative components in $\hat{\underline{\theta}}_1$. We can view the conditional expectation operator as a <u>filter</u> that eliminates unnecessary components of the data.

Proof of Theorem

$$MSE(\underline{\hat{\theta}}_2) = E\left[\|\underline{\hat{\theta}}_2(\underline{X}) - \underline{\theta}\|^2\right]$$

$$= E\left[\|E[\underline{\hat{\theta}}_1(\underline{X}) \mid \underline{T}] - \underline{\theta}\|^2\right]$$

$$= E\left[\|E[\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta} \mid \underline{T}]\|^2\right]$$

$$\leq E\left[E\left[\|\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta}\|^2 \mid \underline{T}\right]\right] \quad \boxed{1}$$

$$= E\left[\|\underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta}\|^2\right] \quad \boxed{2}$$

$$= MSE(\underline{\hat{\theta}}_1)$$

1 Consider the random variable

$$\underline{Y} = \underline{Y}(\underline{t}) = \underline{\hat{\theta}}_1(\underline{X}) - \underline{\theta} \mid T(\underline{X}) = \underline{t}$$

By Jensen's inequality,

$$\varphi(E[\underline{Y}]) \leq E[\varphi(\underline{Y})]$$

where φ is the convex function

$$\varphi(\underline{y}) = \|\underline{y}\|^2 = \underline{y}^T\underline{y}.$$

Moreover, since φ is <u>strictly</u> convex, equality holds iff $\underline{Y}$ is a deterministic function of $\underline{t}$ almost surely.

But then

$$\hat{\underline{\theta}}_1(\underline{X}) = \hat{\underline{\theta}}_2(\underline{X})$$

almost surely.

[2] This follows from the "law of total expectation" which states that

$$E_{\underline{u}}[\psi(\underline{u})] = E_{\underline{v}}[E_{\underline{u}|\underline{v}}[\psi(\underline{u})|\underline{v}]]$$

<u>Remark</u>

The result still holds if we replace $\varphi(\underline{y}) = \underline{y}^T\underline{y}$ with any convex function.

However, the condition on equality may need to be modified if φ is not <u>strictly</u> convex (e.g., $\varphi(\underline{y}) = \sum_{i=1}^{p} |y_i|$)

Minimal Sufficient Statistics

If we observe $\underline{x}$, then $\underline{x}$ itself is a sufficient statistic, albeit not a very interesting one. When is a suff. stat. as compressed as it can possibly be?

Definition A sufficient statistic is *minimal* if it is a function of every other sufficient statistic.

Example

For iid Gaussian observations with unknown mean, the following statistics are sufficient:

- $\underline{x} = [x_1, \cdots, x_n]^T$
- $[x_1 + x_3 + \cdots \;,\; x_2 + x_4 + \cdots]^T$
- $[\bar{x} \;,\; s^2]^T$
- $\bar{x}$

However, the first 3 are not minimal because they are not functions of the 4th.

Since $\bar{x}$ is 1-dimensional, it is minimal.

Remark | When we say "is a function of every other sufficient statistic," we exclude functions that increase dimensionality. Otherwise

$$[\bar{x}, \bar{x}]^T$$

would be sufficient in the previous example.

The dimension of a minimal suff. stat. cannot be less than the dimension of $\underline{\theta}$. If we are lucky the dimensions will be equal. Sometimes, the dim. of a minimal suff. stat. is as large as N. See, for example, the Cauchy distribution.

Proving Minimality

Proposition | $\underline{T} = \tau(\underline{X})$ is a minimal suff. stat. if

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} \text{ is independent of } \underline{\theta} \iff \tau(\underline{x}) = \tau(\underline{y})$$

Proof First let's show $\underline{T}$ is a sufficient stat. under the given assumption.

For each $\underline{t}$ in the range of τ, assign a vector $\underline{y}(\underline{t})$ such that $\tau(\underline{y}(\underline{t})) = \underline{t}$.

Then

$$f_{\underline{\theta}}(\underline{x}) = \frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y}(\tau(\underline{x})))} \cdot f_{\underline{\theta}}(\underline{y}(\tau(\underline{x})))$$

$$= h(\underline{x}) \cdot g_{\theta}(\tau(\underline{x}))$$

$\uparrow$ independent of θ since $\tau(\underline{x}) = \tau(\underline{y}(\tau(\underline{x})))$.

To show $\underline{T}$ is minimal, we must show that for any other suff. stat. $\underline{T}' = \tau'(\underline{X})$, $\underline{T}$ is a function of $\underline{T}'$.

So suppose $\underline{T}'$ takes on the value $\underline{t}'$. We must show that $\underline{T}$ is uniquely determined by $\underline{t}'$.

That is, if $\underline{x}$ and $\underline{y}$ are such that $\tau'(\underline{x}) = \tau'(\underline{y}) = \underline{t}'$, then $\tau(\underline{x}) = \tau(\underline{y})$.

So suppose $\underline{x}$ any $\underline{y}$ are such that $\tau'(\underline{x}) = \tau'(\underline{y}) = \underline{t}'$.

Then

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} = \frac{g'_{\underline{\theta}}(\tau'(\underline{x})) \cdot h'(\underline{x})}{g'_{\underline{\theta}}(\tau'(\underline{y})) \cdot h'(\underline{y})} = \frac{h'(\underline{x})}{h'(\underline{y})}$$

which is independent of $\underline{\theta}$. Therefore $\tau(\underline{x}) = \tau(\underline{y})$ ▨

Exercise | Show that $[\Sigma X_i^2, \Sigma X_i]^T$ is a minimal suff. stat. for $[\mu, \sigma^2]^T$, where $X_i \overset{iid}{\sim} N(\mu, \sigma^2)$.

Solution | $\underline{\theta} = [\mu, \sigma^2]^T$.

$$\frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}}(\underline{y})} = \frac{\exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - 2\mu \sum x_i + n\mu^2\right]\right\}}{\exp\left\{-\frac{1}{2\sigma^2}\left[\sum y_i^2 - 2\mu \sum y_i + n\mu^2\right]\right\}}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left[\sum x_i^2 - \sum y_i^2\right] + \frac{\mu}{\sigma}\left[\sum x_i - \sum y_i\right]\right\}$$

which is independent of $\underline{\theta} \iff \left[\sum x_i^2 \;\; \sum x_i\right]^T = \left[\sum y_i^2 \;\; \sum y_i\right]^T$

$$\iff \tau(\underline{x}) = \tau(\underline{y}).$$

Since $[\bar{X} \;\; S^2]^T$ is a function of $\left[\sum X_i^2 \;\; \sum X_i\right]^T$, it is also minimal. ▨

A second way to show that a suff. stat. is minimal is to show that it is <u>complete</u>.

Complete Sufficient Statistics

Definition A sufficient statistic $T = \tau(\underline{X})$ is **complete** iff for all real-valued functions ϕ

$$\left(E_{\underline{\theta}}[\phi(T)] = 0 \quad \forall \underline{\theta} \right) \Rightarrow \left(P_{\underline{\theta}}[\phi(T) = 0] = 1 \ \forall \underline{\theta} \right)$$

Example Bernoulli trials, part IV

Consider N independent Bernoulli trials

$$X_i \overset{iid}{\sim} \text{Bernoulli}(\theta) \qquad \theta \in [0,1]$$

Recall $K = \sum_{i=1}^{N} X_i$ is sufficient for θ.

Suppose $E_\theta[\phi(K)] = 0 \quad \forall \theta$. But

$$E[\phi(K)] = \sum_{k=0}^{N} \phi(k) \binom{N}{k} \theta^k (1-\theta)^{N-k}$$

$$= \text{poly}(\theta)$$

where $\text{poly}(\theta)$ is an N^{th} degree polynomial.

Then $\text{poly}(\theta) = 0 \quad \forall \theta \in [0,1]$

$\Rightarrow$ $\text{poly}(\theta)$ is the zero polynomial

$\Rightarrow \phi(k) = 0 \quad \forall k$

$\Rightarrow K$ is complete

Note The definition of completeness depends on the parameter space Θ. In the last example we had $\Theta = [0,1]$. What would happen if $\Theta = \{\frac{1}{3}, \frac{2}{3}\}$?

Proposition Under very general conditions, if T is a complete S.S., then T is minimal.

Proof Let $T' = \tau'(X)$ be any S.S.

We need to show T is determined completely by T'.

Define

$$\psi(t') := E_\theta[T \mid T' = t'].$$

We will show in particular that $T = \psi(T')$.

Introduce

$$\rho(t) := E_\theta[\psi(T') \mid T = t]$$

Note that

$$\begin{aligned} E_\theta[T] &= E[E[T \mid T']] \quad \leftarrow \text{(law of total expectation)} \\ &= E[\psi(T')] \\ &= E[E[\psi(T') \mid T]] \\ &= E[\rho(T)] \end{aligned}$$

By completeness, we deduce

$$P_{\underline{\theta}}[\underline{T} = \rho(\underline{T})] = 1 \quad \forall \underline{\theta} \tag{1}$$

This implies

$$\psi(\underline{t}') = E[\underline{T} \mid \underline{T}' = \underline{t}'] = E[\rho(\underline{T}) \mid \underline{T}' = \underline{t}'] \tag{2}$$

Now recall the "law of total variance":

$$\mathrm{Var}[Y] = E[\mathrm{Var}[Y \mid Z]] + \mathrm{Var}[E[Y \mid Z]]$$

for scalar random variables Y and Z. Using the subscript "i" to denote the i^{th} component, we have

$$\mathrm{Var}[\rho_i(\underline{T})] = E[\mathrm{Var}[\rho_i(\underline{T}) \mid \underline{T}']] + \mathrm{Var}[E[\rho_i(\underline{T}) \mid \underline{T}']]$$

$$= \text{"} \quad + \mathrm{Var}[\psi_i(\underline{T}')] \quad \text{(by (2))}$$

$$= \text{"} \quad + E[\mathrm{Var}[\psi_i(\underline{T}') \mid \underline{T}]] + \mathrm{Var}[E[\psi_i(\underline{T}') \mid \underline{T}]]$$

$$= \text{"} \quad + \text{"} \quad + \mathrm{Var}[\rho_i(\underline{T})]$$

Since $\mathrm{Var}[\text{anything}] \geq 0$ we deduce

$$\mathrm{Var}[\rho_i(\underline{T}) \mid \underline{T}'] = 0 \quad \text{with prob. 1}$$

$$\mathrm{Var}[\psi_i(\underline{T}') \mid \underline{T}] = 0 \quad \text{with prob. 1}$$

How does this imply the desired result?

$\rho_i(\underline{I})$ is a deterministic function of $\underline{I}'$

In particular (returning to vector notation)

$$\rho(\underline{I}) = E[\rho(\underline{I}) \mid \underline{I}']$$

$$= \psi(\underline{I}')$$

Since $\underline{I} = \rho(\underline{I})$ with prob. 1, we conclude

$$\underline{I} = \rho(\underline{I}) = \psi(\underline{I}') \qquad \text{w.p. 1}$$

as was to be shown. ▨

The "general conditions" under which the result is valid are that the various means and variances used throughout the proof are well-defined and finite.

The Exponential Family

In general, sufficient statistics, especially ones that are minimal and complete, can be difficult to find (if they even exist).

For a special family of distributions, however, we can immediately identify a complete and minimal suff. stat.

Definition We say the distribution of $\underline{X}$ belongs to the exponential family of distributions if its pdf/pmf can be written

$$f_{\underline{\theta}}(\underline{x}) = a(\underline{\theta})\, b(\underline{x}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}$$

for some a, b, c and τ, where the dimension p of $\underline{\theta}$ is also the dimension of $c(\underline{\theta})$ and $\tau(\underline{x})$.

Note My definition is slightly different from that given in Prof. Hero's notes.

- c instead of $-c$
- defined for vectors, not just scalars

Example | Bernoulli trials, part IV

$$p_\theta(\underline{x}) = \theta^k (1-\theta)^{N-k} \qquad \left[k = \sum_{i=1}^{N} x_i\right]$$

$$= \exp\left\{ \log\left(\theta^k (1-\theta)^{N-k} \right) \right\}$$

$$= \exp\left\{ k \log\theta + (N-k)\log(1-\theta) \right\}$$

$$= \underbrace{\exp\left\{ N \log(1-\theta) \right\}}_{a(\theta)} \cdot \underbrace{\exp\left\{ [\log\theta - \log(1-\theta)]}_{c(\theta)} \cdot \underbrace{k \right\}}_{\tau(x)}$$

$$[b(\underline{x}) = 1]$$

Many common distributions belong to the exponential family, including Gaussian w/ unknown mean and/or variance, Poisson, exponential, gamma, binomial, and multinomial.

Exercise Suppose $\underline{X} = [X_1, \ldots, X_N]^T$ where

$$X_i \overset{iid}{\sim} \text{Gamma}(\alpha, \beta)$$

Recall that if $X \sim \text{Gamma}(\alpha, \beta)$, then

$$f_{\underline{\theta}}(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1} e^{-\beta x}, \quad x \geq 0$$

where $\underline{\theta} = [\alpha \ \beta]^T$. Show that the distribution of $\underline{X}$ belongs to the exponential family.

Solution

$$f_{\underline{\theta}}(\underline{x}) = \prod_{i=1}^{N} f_{\underline{\theta}}(x_i)$$

$$= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)}\right)^{N} \cdot \prod_{i=1}^{N} x_i^{\alpha-1} \exp\left\{-\beta \sum_{i=1}^{N} x_i\right\}$$

$$= \left(\frac{\beta}{\Gamma(\alpha)}\right)^{N} \exp\left\{\log\left[\left(\prod x_i\right)^{\alpha-1}\right] - \beta \sum x_i\right\}$$

$$= \left(\frac{\beta}{\Gamma(\alpha)}\right)^{N} \exp\left\{(\alpha-1) \sum_{i=1}^{N} \log x_i - \beta \sum x_i\right\}$$

$$\Rightarrow = \underbrace{\left(\frac{\beta}{\Gamma(\alpha)}\right)^{N}}_{a(\underline{\theta})} \underbrace{\left(\prod_{i=1}^{N} x_i\right)^{-1}}_{b(\underline{x})} \exp\left\{\underbrace{[\alpha \;\; -\beta]}_{c(\underline{\theta})^T} \underbrace{\begin{bmatrix} \sum \log x_i \\ \sum x_i \end{bmatrix}}_{\tau(\underline{x})}\right\}$$

[Note: representation is not unique.]

Proposition If the distribution of $\underline{X}$ belongs to the exponential family, then $\underline{T} = \tau(\underline{X})$ is a sufficient statistic for $\underline{\theta}$.

Proof | $\underline{T}$ is sufficient for $\underline{\theta}$ by the FNFT:

$$f_{\underline{\theta}}(\underline{x}) = a(\underline{\theta})\, b(\underline{x}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}$$

$$= \underbrace{a(\underline{\theta}) \exp\left\{ c(\underline{\theta})^T \tau(\underline{x}) \right\}}_{g_\theta(\tau(\underline{x}))} \cdot \underbrace{b(\underline{x})}_{h(\underline{x})}$$

Proposition | Under certain "reasonable" conditions, $\underline{T} = \tau(\underline{X})$ is a complete and minimal sufficient statistic for $\underline{\theta}$.

Sketch of proof | Before we argued that the pdf/pmf of $\underline{X}$ depends on $\underline{\theta}$ only through $f_{\underline{\theta}}(\underline{t})$.

Thus

$$f_{\underline{\theta}}(\underline{t}) \propto \exp\left\{ c(\underline{\theta})^T \underline{t} \right\}$$

Suppose ϕ is a real-valued function such that

$$E_{\underline{\theta}}\left\{ \phi(\underline{T}) \right\} = 0 \quad \forall \theta$$

We must show $P_\theta\left\{ \phi(\underline{T}) = 0 \right\} = 1 \quad \forall \theta$.

For each $\underline{\theta}$ we can write

$$0 = E\{\phi(\underline{T})\}$$

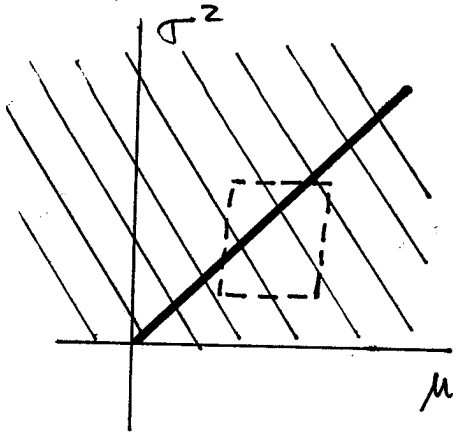
$$= \int \phi(\underline{t}) f_{\underline{\theta}}(\underline{t}) \, d\underline{t}$$

$$\propto \int \phi(\underline{t}) \exp\{c(\underline{\theta})^T \underline{t}\} \, d\underline{t}$$

which is the Laplace transform of ϕ at $c(\underline{\theta})$. Inverting the Laplace transform we find $\phi \equiv 0$. ∎

The "reasonable" conditions under which the above arguments hold are needed to ensure the uniqueness (invertibility) of the Laplace transform:

- The parameter space Θ must contain an open rectangle: think back to Bernoulli example. As another example, in Gaussian case with $\mu = \sigma^2$, $\underline{T}$ is not complete.



- The image of $c(\underline{\theta})$, $\underline{\theta} \in \Theta$, should have full dimensionality.

Note that minimality of $\underline{T} = \tau(\underline{X})$ for the exponential model follows from completeness by a previous result.

How could we show minimality directly? What condition would need to be satisfied?

Summary

- Sufficient statistic $\underline{T}$ for $\underline{\theta}$: contains all the information about $\underline{\theta}$ present in $\underline{X}$.
- Fisher-Neyman Factorization Theorem: $\underline{T}$ is sufficient

$$\iff f_{\underline{\theta}}(\underline{x}) = g_{\underline{\theta}}(\tau(\underline{x})) \cdot h(\underline{x})$$

- Rao-Blackwell Theorem: Can improve any estimator by conditioning on a suff. stat.
- Minimal suff. stat: function of any other suff. stat. Intuitively, maximal compression of information.
- Completeness: technical condition that ensures minimality. Will also be important in our study of minimum variance unbiased estimators.
- Exponential family: Broad class for which a complete and minimal suff. stat. is easily identified.

Key

a. $\begin{cases} \theta & \text{if } x = 1 \\ 1-\theta & \text{if } x = 0 \\ 0 & \text{else} \end{cases}$

b. $\exp\left\{\frac{-1}{2\sigma^2}\left[-2\theta\sum_{i=1}^{N} x_i + N\theta^2\right]\right\}$ $\underbrace{\qquad\qquad}_{g_\theta(t)}$ $(2\pi\sigma^2)^{-\frac{N}{2}} \exp\left\{\frac{-\sum x_i^2}{2\sigma^2}\right\}$ $\underbrace{\qquad\qquad}_{h(\underline{x})}$

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