Sufficient Statistics

A sample is a sequence $X_1, \dots, X_n$ of iid random variables.

Convention: We use capital letters to denote the sample as random variables and lower case letters for the observations. Thus the sample $X_1, \dots, X_n$ consists of random variables while $x_1, \dots, x_n$ denotes the observed (measured) values of the samples.

Assume the common distribution of $X_1, X_2, \dots, X_n$ depends on a parameter θ.

Here θ can be one parameter or multiple parameters.

Def A statistics (or estimator of θ) is a function of the sample space:

$$T = u(X_1, \dots, X_n)$$

Def We call T a sufficient statistics if the conditional distribution of the sample given T does not depend on the parameter θ.

In other words, if $f(x;\theta)$ is the distribution of X (either pmf or pdf), then

$$\frac{f(x_1;\theta) \cdots f(x_n;\theta)}{f_T(u(x_1, \dots, x_n);\theta)} = \underbrace{H(x_1, \dots, x_n)}_{\text{independent of } \theta}$$

Alternatively we can formulate this as

$(X_1, \dots, X_n) \mid T = t$ does not depend on θ.

This, from the definition of the conditional expectation is

$$\frac{f_{X_1}(x_1;\theta) \cdots f_{X_n}(x_n;\theta)}{f_T(t;\theta)}$$ is independent of θ if $T(x_1, \dots, x_n) = t$.

This is a canonical way of defining a sufficient statistics.
An alternative way of discribing this is the following theorem

Th (Neyman) $T = u(X_1, \dots, X_n)$ is a sufficient statistics if and only if

$$f(x_1;\theta) \cdots f(x_n;\theta) = k_1(u(x_1, \dots, x_n);\theta)\, k_2(x_1, \dots, x_n)$$

where k_2 is a function which depends only on $x_1, \dots, x_n$ and not on θ.

Pf We will do a proof in the case of discrete r.v.
The direct implication is straightforward because

$$f(x_1;\theta) \cdots f(x_n;\theta) = f_T(u(x_1, \dots, x_n);\theta)\, H(x_1, \dots, x_n).$$

For the reverse if we have

$$f(x_1;\theta) \cdots f(x_n;\theta) = k_1(u(x_1, \dots, x_n);\theta)\, k_2(x_1, \dots, x_n)$$

with $u(x_1, \dots, x_n) = t$, we have that

$$f_{X \mid T}(x_1, \dots, x_n) = \frac{f(x_1;\theta) \cdots f_{X_n}(x_n;\theta)}{f_T(t;\theta)} = \frac{k_1(t;\theta)}{f_T(t;\theta)} \cdot k_2(x_1, \dots, x_n).$$

This means that for fixed $x_1, \dots, x_n$ such that $\frac{x_1 + \dots + x_n}{n} = t$

$\frac{k_1(t;\theta)}{f_T(t;\theta)}$ does not depend on θ. Thus

$$\frac{k_1(t;\theta)}{f_T(t;\theta)} = g(t) \text{ and thus}$$

$$f(x_1;\theta)\cdots f(x_n;\theta) = f_T(t;\theta)\, k_3(x_1,\dots x_n)$$

Def We say that an estimator T is unbiased if $E[T] = \theta$.

The estimator is called MVUE if for any other unbiased estimator T' we have

$$Var(T) \le Var(T')$$

Given a sufficient statistics and an unbiased estimator T, then we can find another one with a smaller variance.

Th (Rao-Blackwell) If T_1 is a sufficient statistics and T_2 is an unbiased est, then $T_3 = E[T_2 / T_1]$ is actually also unbiased and has smaller variance than T_1.

Pf We know that $E[T_3] = E[T_2] = \theta$. Thus T_3 is unbiased. On the other hand, (see the Appendix here)

$$Var(T_2) = Var(T_3) + E[(T_2 - T_3)^2] \ge$$
$$\ge Var(T_3).$$

Equality is attained iff $T_3 = T_2$ or in other words if T_2 is a function of T_1. If we prevent this, then T_3 has a stricly smaller variance.

Completeness and Uniqueness

Def A variable Z from $(f(x;\theta))_\theta$ is called complete if any function u such that $E_\theta[u(Z)]=0$ $\forall\theta$ implies $u\equiv 0$ a.s..

Ex Take $Z\sim \text{Poisson}(\theta)$. Then $f(x;\theta)=e^{-\theta}\dfrac{\theta^x}{x!}$ is a complete family

Sol For a function u, we have

$$E[u(Z)]=\sum_{x=0}^{\infty} u(x)P_\theta(Z=x)$$
$$=\sum_{x\geq 0} u(x)\frac{\theta^x}{x!}e^{-\theta}=0 \quad \forall\theta$$

Therefore $\sum_{x\geq 0} u(x)\dfrac{\theta^x}{x!}=0$ $\forall\theta>0$, so $u\equiv 0$

Theorem (Lehmann-Scheffé)
If $X_1,..,X_n$ is a sample and a sufficient statistics $Y_1=u(X_1,..,X_n)$ has a family $(f_{Y_1}(y;\theta))$ which is complete, and there is a function $Y_2=\varphi(Y_1)$ which is an unbiased estimator of θ, then Y_2 is the unique MVUE.

Pf From Rao-Blackwell we know that we can restrict our search for $Y_3=\psi(Y_1)$.
If Y_3 is also unbiased, then $E[\psi(Y_1)]=\theta=$
$=E[\varphi(Y_1)]$ and thus $E[\varphi(Y_1)-\psi(Y_1)]=0$. From completeness we get that $\varphi=\psi$ a.s., thus $Y_2=Y_3$

Slang We say that Y_1 is a complete sufficient statistic.

Appendix Conditioning

Typically if we have X and Y, two random variables then the distribution of X given Y is defined via the conditional pmf or pdf.

$$p_{X|Y}(x|y) = \frac{P(X=x, Y=y)}{P(Y=y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

Inspired by this, if (X,Y) have a joint density, then

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

This is good if we do have a joint density.

Ex What would by the conditional dist of X given X?

$$p_{X|X}(x|y) = \frac{\mathbb{P}(X=x, X=y)}{\mathbb{P}(X=y)} = \begin{cases} 0 & \text{if } x \neq y \\ 1 & \text{if } x = y \end{cases}$$

What this means is that the distribution of X given $X=y$ is actually concentrated at y.

This is the case in the discrete situation, though in the continuous case is even more problematic because $f_{X,X}(x,y)$ does not really make sense.

However the conditional expectation of any two random variables actually makes sense.

More precisely if we have X and Y, we can try to define $\mathbb{E}[X|Y]$ as the best approximation of X in terms of Y. In other words we want

$\mathbb{E}[X|Y] = \varphi(Y)$ such that

$\mathbb{E}[(X-\varphi(Y))^2]$ has the smallest possible value among all possible functions φ.

This leads to some properties of $\varphi(Y)$.

<u>Properties</u> ① $\mathbb{E}[(X-\varphi(Y))\psi(Y)]=0 \ \forall \psi$

In particular $\mathbb{E}[X]=\mathbb{E}[\varphi(Y)]$

② $Var(\varphi(Y)) \leq Var(X)$

<u>Pf</u> ① If we assume a φ which minimizes $\mathbb{E}[(X-\varphi(Y))^2]$ over all choices of φ then by replacing φ by $\varphi+\varepsilon\psi$ and taking derivatives w.r.t ε at $\varepsilon=0$ we get

$$\mathbb{E}[(X-\varphi(Y))\psi(Y)]=0$$

Alternatively we can consider the Hilbert space of all functions $\varphi(Y)$ such that $\mathbb{E}[\varphi^2(Y)]<\infty$

Then one instance of $\varphi(Y)$ is obtained by projecting X from $L^2(\Omega,\mathcal{F},P)$ onto this subspace.

② $Var(X) = \mathbb{E}[(X-\mathbb{E}[X])^2] =$

$$= \mathbb{E}[(X-\varphi(Y)+\varphi(Y)-\mathbb{E}[X])^2]$$

$$= \mathbb{E}[(X-\varphi(Y))^2] + 2\underbrace{\mathbb{E}[(X-\varphi(Y))(\varphi(Y)-\mathbb{E}[X])]}_{\substack{=0 \\ \text{from the first part}}}$$

$$+ \mathbb{E}[(\varphi(Y)-\mathbb{E}[X])^2]$$

Because $\mathbb{E}[X] = \mathbb{E}[\varphi(Y)]$ we actually get

$$\mathrm{Var}(X) = \mathbb{E}[(X - \varphi(Y))^2] + \mathrm{Var}(\varphi(Y)).$$

In particular we actually get something better, namely $\mathrm{Var}(X) \geq \mathrm{Var}(\varphi(Y))$ with equality iff $X = \varphi(Y)$ which means that X is a function of Y.

<u>Def</u> The $\varphi(Y)$ defined above is denoted

$$\mathbb{E}[X|Y].$$

The function φ itself is denoted as

$$\mathbb{E}[X|Y=y] = \varphi(y)$$

<u>Ex</u> Compute $\mathbb{E}[X|X^2]$ if $X \sim N(0,1)$.

<u>Sol</u> $\mathbb{E}[X/X^2] = \varphi(X^2)$ and to find this φ we can use the first property above, namely

$$\mathbb{E}[X\psi(X^2)] = \mathbb{E}[\varphi(X^2)\psi(X^2)] \text{ for any } \psi.$$

In particular

$$\underbrace{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x\,\psi(x^2)\,e^{-x^2/2}\,dx}_{\substack{\shortparallel \\ 0}} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\varphi(x^2)\psi(x^2)e^{-x^2/2}\,dx$$

thus $\mathbb{E}[X|X^2] = 0$.

Interestingly, $\mathbb{E}[X^2|X] = X^2$ because X^2 is already a function of X.

Using this interpretation of the conditional expectation we actually have also that a formal definition can be put forward in the form:

Def $E[X|Y) = Z$ such that

$$E[Z\,\psi(Y)] = E[X\,\psi(Y)]$$

for any ψ continuous and bounded.

Th If X is integrable, then $E[X|Y]$ always exists.

In the case X, Y have a joint pdf then

$Z = \varphi(Y)$ with

$$\varphi(y) = \int x \frac{f_{X,Y}(x,y)}{f_Y(y)}\,dx$$

because in this case

$$E[\varphi(Y)\psi(Y)] = \iint \frac{x f_{X,Y}(x,y)}{\cancel{f_Y(y)}}\,dx\,\psi(y)\cancel{f_Y(y)}\,dy$$

$$= \iint x\,\psi(y) f_{X,Y}(x,y)\,dx\,dy$$

$$= E[X\psi(Y)].$$

thus we fall back onto the well known case.

Exercise A rigourous way of formulating that $(X_1,\dots,X_n)\,|\,u(X_1,\dots,X_n)$ does not depend on θ is to state that

$$\mathbb{E}\left[\xi(X_1,\dots,X_n)\,|\,u(X_1,\dots,X_n)\right] = \varphi_{\xi}(u(X_1,\dots,X_n))$$ does not on θ for any choice of $X_1,\dots,X_n$.

By definition, this is the same as saying that

$$\mathbb{E}\left[\xi(X_1,\dots,X_n)\,\psi(u(X_1,\dots,X_n))\right] = \mathbb{E}\left[\varphi_{\xi}(u(X_1,\dots,X_n))\,\psi(u(X_1,\dots,X_n))\right]$$

for any choice of ψ.

The right hand side is the same as

$$\int \xi(x_1,\dots,x_n)\,\psi(u(x_1,\dots,x_n);\theta)\,f(x_1;\theta)\cdots f(x_n;\theta)\,dx_1\dots dx_n =$$

$$= \int \varphi_{\xi}(t)\,\psi(t;\theta)\,f_T(t;\theta)\,dt$$

Taking $\psi(t;\theta) = \dfrac{\psi(t)}{f_T(t;\theta)}$ we get that

$$\int \xi(x_1,\dots,x_n)\,\psi(u(x_1,\dots,x_n))\,\frac{f(x_1;\theta)\cdots f(x_n;\theta)}{f_T(u(x_1,\dots,x_n);\theta)}\,dx_1\dots dx_n$$

$$= \int \varphi_{\xi}(t)\,\psi(t)\,dt$$

Since ξ is an arbitrary function of $x_1,\dots,x_n$ and φ_{ξ}, ψ do not depend on θ, it must be that

$$\frac{f(x_1;\theta)\cdots f(x_n;\theta)}{f_T(u(x_1,\dots,x_n);\theta)}$$ does not depend on θ.