MA 3140: Statistical Inference

Dr. Sameen Naqvi
Department of Mathematics, IIT Hyderabad
Email id: sameen@math.iith.ac.in

Example 2: Location Family[1] Ancillary Statistic

$Z_1, \ldots, Z_n$: iid observations from $F(x)$.

Define $X_i = Z_i + \theta$. Then, $X_1, \ldots, X_n$ are iid observations from $F(x - \theta)$, $-\infty < \theta < \infty$.

Claim: $R = X_{(n)} - X_{(1)}$ is an ancillary statistic where, $X_{(1)}, \ldots, X_{(n)}$ are order statistics from the sample.

---

[1] Let U be a random variable with a fixed distribution F. If a constant a is added to U, the resulting variable $X = U + a$ has distribution $P(X \leq x) = F(x - a)$.
The totality of such distributions, for fixed F and as a varies from $-\infty$ to ∞, is said to constitute a location family.
Examples: Normal Distribution with unknown μ and $\sigma = 1$, Cauchy Distribution.

Example 2 cont'd

The cdf of the range statistic, R is

$$\begin{aligned} F_R(r) &= P(R \leq r) \\ &= P(X_{(n)} - X_{(1)} \leq r) \\ &= P((Z_{(n)} + \theta) - (Z_{(1)} + \theta) \leq r) \\ &= P(Z_{(n)} - Z_{(1)} + \theta - \theta \leq r) \\ &= P(Z_{(n)} - Z_{(1)} \leq r). \end{aligned}$$

The last probability does not depend on θ.

Thus, the cdf of R does not depend on θ, and hence, R is an ancillary statistic.

Example 3: Scale Family[2] Ancillary Statistic

$Z_1, \ldots, Z_n$: iid observations from $F(x)$.

Define $X_i = \sigma Z_i$. Then, $X_1, \ldots, X_n$ are iid observations from $F(x/\sigma)$, $\sigma > 0$.

Claim:

$$\frac{X_1 + \ldots + X_n}{X_n} = \frac{X_1}{X_n} + \ldots + \frac{X_{n-1}}{X_n} + 1$$

is an ancillary statistic.

[2] A scale family is generated by the transformations $X = bU$, $b > 0$, and has the form $P(X \leq x) = F(x/b)$.
Examples: Exponential Distribution, Gamma Distribution with α fixed and β unknown, Normal Distribution with $\mu = 0$ and σ unknown.

Example 3 cont'd

The joint cdf of $\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}$ is

$$
\begin{aligned}
F(y_1, \ldots, y_{n-1}) &= P\Big(\frac{X_1}{X_n} \le y_1, \ldots, \frac{X_{n-1}}{X_n} \le y_{n-1}\Big) \\
&= P\Big(\frac{\sigma Z_1}{\sigma Z_n} \le y_1, \ldots, \frac{\sigma Z_{n-1}}{\sigma Z_n} \le y_{n-1}\Big) \\
&= P\Big(\frac{Z_1}{Z_n} \le y_1, \ldots, \frac{Z_{n-1}}{Z_n} \le y_{n-1}\Big)
\end{aligned}
$$

The last probability does not depend on σ.

Thus, the distribution of $\frac{X_1}{X_n}, \ldots, \frac{X_{n-1}}{X_n}$ does not depend on σ, and so is the distribution of any function of these.

Example 4: Normal Distribution

Let $X_1, \ldots, X_n \sim N(\mu, 1)$

Claim: $V = (X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ is an ancillary statistic.

Here, V does not depend on μ, and hence, is an ancillary statistic.

Question: Check whether $W = (X_1 - \overline{X}, X_2 - \overline{X}, \ldots, X_n - \overline{X})$ is an ancillary statistic or not.

MSS vs. Ancillary Statistic

- **MSS:** achieves maximum amount of data reduction while still retaining all the information about the parameter θ.
  - It eliminates all the extraneous information in the sample, retaining only the piece with info about θ.
- **Ancillary statistic:** the distribution does not depend on θ.
- One might suspect that MSS is unrelated to an ancillary statistic. However, this is not necessarily true (see the next example).

Example: Revisiting Uniform Distribution

Let $X_1, \ldots, X_n \sim U(\theta, \theta + 1)$, $-\infty < \theta < \infty$.

We showed that $X_{(n)} - X_{(1)}$ is an ancillary statistic.

Also recall that $(X_{(1)}, X_{(n)})$ is MSS, and so is $(X_{(n)} - X_{(1)}, (X_{(1)} + X_{(n)})/2)$.

Thus, in this case, the ancillary statistic is an important component of the MSS, and hence, both are not independent.

Example 2: Ancillary statistic reveals information about θ.

Let X_1 and X_2 be iid observations from the discrete distribution such that

$$P_\theta(X = \theta) = P_\theta(X = \theta + 1) = P_\theta(X = \theta + 2) = \frac{1}{3},$$

where θ, the unknown parameter, is an integer.

Let $X_{(1)}$ and $X_{(2)}$ be the O.S. for the sample.

- It can be easily shown that (R, M) is MSS, where $R = X_{(2)} - X_{(1)}$ and $M = (X_{(1)} + X_{(2)})/2$.
- Further, since this is a location parameter family, R is an ancillary statistic.

Example 2 cont'd

See how R might give information about θ:

Consider a sample point (r, m), where m is an integer.

First consider only m. For m to have positive probability, θ must be one of the 3 values, i.e., $\theta = m$ or $\theta = m - 1$ or $\theta = m - 2$.

With only the info that $M = m$, all 3 values of θ are possible.

Example 2 cont'd

Now, suppose you get additional information that $R = 2$.

Then, in this case, $X_{(1)} = m - 1$ and $X_{(2)} = m + 1$.

With this info, the only possible value for θ is $m - 1$.

Thus, the knowledge of the value of the ancillary statistic R has increased our knowledge about θ.

Of course, the knowledge of R alone would not give us any information about θ.

Completeness

Complete Statistic

- X: a random variable with probability distribution $P_\theta, \theta \in \Theta$.
- $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$: family of probability distributions
- $\mathcal{P}$ is said to be complete if

$$
\begin{aligned}
& E_\theta[g(X)] = 0, \quad \forall\ \theta \in \Theta, \text{ and a function } g \\
\Longrightarrow\ & P_\theta[g(X) = 0] = 1, \quad \forall\ \theta \in \Theta.
\end{aligned}
$$

- A statistic $T(X)$ is said to be complete if the family of distributions of T is complete.

Example 1: Binomial Family

Let $X \sim Bin(n, p)$, n is known, $0 < p < 1$. Check whether the Binomial family is complete.

Solution:

$$
\begin{aligned}
E_p[g(X)] &= 0, \quad \forall\ p \in (0,1) \\
\implies \sum_{x=0}^{n} \underbrace{g(x)\binom{n}{x}}_{h(x)} p^x (1-p)^{n-x} &= 0, \quad \forall\ p \in (0,1) \\
\implies \quad \sum_{x=0}^{n} h(x) s^x &= 0, \quad \forall\ s > 0, \ \left(s = \frac{p}{1-p} > 0\right) \\
\implies \quad h(x) &= 0, \quad \forall\ x = 0, 1, 2, \ldots, n \\
\implies \quad g(x) &= 0, \quad \forall\ x = 0, 1, 2, \ldots, n \\
\implies \quad P_p[g(X) = 0] &= 1, \quad \forall\ p \in (0,1)
\end{aligned}
$$

Thus, the family $\{Bin(n, p) : 0 < p < 1\}$ is complete.

Example 1': Binomial Family

Let $X_1, X_2, \ldots, X_n \sim Bin(1, p)$, n is known, $0 < p < 1$. Then, we know that $T = \sum X_i \sim Bin(n, p)$ is a sufficient statistic. Check whether T is also complete.

Solution:

$$\begin{aligned}
& E_p[g(T)] = 0, \quad \forall\, p \in (0,1) \\
\Longrightarrow \sum_{t=0}^{n} \underbrace{g(t)\binom{n}{t}}_{h(t)} p^t (1-p)^{n-t} &= 0, \quad \forall\, p \in (0,1) \\
\Longrightarrow \quad \sum_{t=0}^{n} h(t) s^t &= 0, \quad \forall\, s > 0, \ \left(s = \frac{p}{1-p} > 0\right) \\
\Longrightarrow \quad h(t) &= 0, \quad \forall\, t = 0, 1, 2, \ldots, n \\
\Longrightarrow \quad g(t) &= 0, \quad \forall\, t = 0, 1, 2, \ldots, n \\
\Longrightarrow \quad P_p[g(T) = 0] &= 1, \quad \forall\, p \in (0,1)
\end{aligned}$$

Thus, $T = \sum X_i$ is a complete statistic.

Example 2: Poisson Family

Let $X \sim P(\lambda)$, $\lambda > 0$. Check whether the Poisson family is complete.

Solution:

$$
\begin{aligned}
& E_\lambda[g(X)] = 0, \quad \forall\ \lambda > 0 \\
\Longrightarrow\ & \sum_{x=0}^{\infty} g(x) \frac{e^{-\lambda}\lambda^x}{x!} = 0, \quad \forall\ \lambda > 0 \\
\Longrightarrow\ & \sum_{x=0}^{\infty} g^*(x)\lambda^x = 0, \quad \forall\ \lambda > 0, \quad \left(g^*(x) = \frac{g(x)}{x!}\right) \\
\Longrightarrow\ & g^*(x) = 0, \quad \forall\ x = 0, 1, 2, \ldots \\
\Longrightarrow\ & g(x) = 0, \quad \forall\ x = 0, 1, 2, \ldots \\
\Longrightarrow\ & P_\lambda[g(X) = 0] = 1, \quad \forall\ \lambda > 0
\end{aligned}
$$

Thus, the family $\{P(\lambda) : \lambda > 0\}$ is complete.

Example 2': Poisson Family

Let $X \sim P(\lambda)$, $\lambda > 0$. Then, we know that $T = \sum X_i \sim P(n\lambda)$ is a sufficient statistic. Check whether T is also complete.

Solution:

$$
\begin{aligned}
& E_\lambda[g(T)] = 0, \quad \forall\ \lambda > 0 \\
\Longrightarrow & \sum_{t=0}^{\infty} g(t) \frac{e^{-n\lambda}(n\lambda)^t}{t!} = 0, \quad \forall\ \lambda > 0 \\
\Longrightarrow & \sum_{t=0}^{\infty} \frac{g(t)}{t!}(n\lambda)^t = 0, \quad \forall\ \lambda > 0 \\
\Longrightarrow & \frac{g(t)}{t!} = 0, \quad \forall\ t = 0, 1, 2, \ldots \\
\Longrightarrow & g(t) = 0, \quad \forall\ t = 0, 1, 2, \ldots \\
\Longrightarrow & P_\lambda[g(T) = 0] = 1, \quad \forall\ \lambda > 0
\end{aligned}
$$

Thus, $T = \sum X_i$ is a complete statistic.

Some Integral Transforms

(i) A unilateral Laplace Transform of $f(x)$ is
$\phi_U(t) = \int\limits_0^\infty e^{-tx} f(x)dx$.

(ii) A bilateral Laplace Transform of $f(x)$ is
$\phi_B(t) = \int\limits_{-\infty}^\infty e^{-tx} f(x)dx$.

(iii) Melin's Transform of $f(x)$ is $\phi_M(t) = \int\limits_0^\infty x^{t-1} f(x)dx$.

- If $f(x) \equiv 0$ then $\phi_U, \phi_B, \phi_M = 0$.
- If $\phi_U, \phi_B, \phi_M = 0$ then $f(x) = 0$.

Example 3: Normal Family

Let $X \sim N(\mu, 1)$, $\mu \in \mathbb{R}$. Check whether the Normal family is complete.

Solution:

$$
\begin{aligned}
& E_\mu[g(X)] = 0, \ \forall \ \mu \in \mathbb{R} \\
\implies \ & \int_{-\infty}^{\infty} g(x) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2} dx = 0, \ \forall \ \mu \in \mathbb{R} \\
\implies \ & \int_{-\infty}^{\infty} g(x) e^{-\frac{x^2}{2}} e^{\mu x} dx = 0, \ \forall \ \mu \in \mathbb{R} \\
\implies \ & g(x) e^{-\frac{x^2}{2}} = 0, \ \forall \ x \in \mathbb{R} \ \text{(using Bilateral LT)} \\
\implies \ & g(x) = 0, \ \forall \ x \in \mathbb{R} \\
\implies \ & P_\mu[g(X) = 0] = 1, \ \forall \ \mu \in \mathbb{R}
\end{aligned}
$$

Thus, the family $\{N(\mu, 1) : \mu \in \mathbb{R}\}$ is complete.

Example 3': Normal Family

Let $X \sim N(\mu, 1)$, $\mu \in \mathbb{R}$. Then, we know that $T = \overline{X} \sim N(\mu, \frac{1}{n})$. Check whether T is complete.

Solution:

$$
\begin{aligned}
& & E_\mu[g(T)] &= 0, \ \forall\ \mu \in \mathbb{R} \\
&\Longrightarrow & \frac{\sqrt{n}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(t) e^{-\frac{\sqrt{n}}{2}(t-\mu)^2} dt &= 0, \ \forall\ \mu \in \mathbb{R} \\
&\Longrightarrow & \int_{-\infty}^{\infty} g(t) e^{-\frac{\sqrt{n}}{2}t^2 + \sqrt{n}\mu t} dt &= 0, \ \forall\ \mu \in \mathbb{R} \\
&\Longrightarrow & g(t) e^{-\frac{\sqrt{n}}{2}t^2} &= 0, \ \forall\ t \in \mathbb{R} \quad \text{(using Bilateral L} \\
&\Longrightarrow & g(t) &= 0, \ \forall\ t \in \mathbb{R} \\
&\Longrightarrow & P_\mu[g(T) = 0] &= 1, \ \forall\ \mu \in \mathbb{R}
\end{aligned}
$$

Thus, $T = \overline{X}$ is a complete statistic.

Example 4

Let $X_1, \ldots, X_n \sim U(0, \theta)$, $\theta > 0$. Check whether $T = X_{(n)}$ is complete or not, where the pdf of $X_{(n)}$ is

$$f_T(t) = \begin{cases} \frac{nt^{n-1}}{\theta^n}, & 0 < t < \theta \\ 0, & \text{o/w} \end{cases}$$

For completeness, consider

$$\begin{aligned} & E_\theta[g(T)] = 0, \ \forall \ \theta > 0 \\ \implies & \frac{n}{\theta^n} \int_0^\theta g(t) t^{n-1} dt = 0, \ \forall \ \theta > 0 \\ \implies & \int_0^\theta g(t) t^{n-1} dt = 0, \ \forall \ \theta > 0 \\ \implies & g(\theta)\theta^{n-1} = 0, \ \forall \ \theta \quad \text{(differentiating both sides)} \\ \implies & g(\theta) = 0, \ \forall \ \theta \end{aligned}$$

Thus, $T = X_{(n)}$ is a complete statistic.

Incomplete Family

Example: Let $X \sim N(0, \theta)$, $\theta > 0$. Check whether the Normal family is complete.

$$
\begin{aligned}
E_\theta[g(X)] &= 0, \ \forall \ \theta > 0 \\
\Longrightarrow \ \frac{1}{\sqrt{\theta}\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-\frac{x^2}{\theta}} dx &= 0, \ \forall \ \theta > 0 \\
\Longrightarrow \qquad \int_{-\infty}^{\infty} g(x) e^{-\frac{x^2}{\theta}} dx &= 0, \ \forall \ \theta > 0
\end{aligned}
$$

which will hold true for any odd function $g(x) = x, x^3, \ldots$.

Thus, the family $\{N(0, \theta) : \theta > 0\}$ is not complete.

Sufficient statistic which is not complete

Example: Let $X_1, \ldots, X_m \sim N(\mu, \sigma_1^2)$ and $Y_1, \ldots, Y_n \sim N(\mu, \sigma_2^2)$ be two independent samples such that $\sigma_1^2 \neq \sigma_2^2$.

The joint pdf of $(X_1, \ldots, X_m, Y_1, \ldots, Y_n)$ is

$$
\begin{aligned}
f(\boldsymbol{x}, \boldsymbol{y}, \mu, \sigma_1^2, \sigma_2^2) &= \frac{1}{(\sqrt{2\pi})^{m+n}\sigma_1^m \sigma_2^n} e^{-\frac{1}{2\sigma_1^2}\sum(x_i-\mu)^2 - \frac{1}{2\sigma_2^2}\sum(y_j-\mu)^2} \\
&= \frac{1}{(\sqrt{2\pi})^{m+n}\sigma_1^m \sigma_2^n} e^{-\frac{\sum x_i^2}{2\sigma_1^2} + \frac{m\mu\bar{x}}{\sigma_1^2} - \frac{m\mu^2}{2\sigma_1^2} - \frac{\sum y_j^2}{2\sigma_2^2} + \frac{n\mu\bar{y}}{\sigma_2^2} - \frac{n\mu^2}{2\sigma_2^2}}
\end{aligned}
$$

Example cont'd

$$\frac{f(\boldsymbol{x}, \boldsymbol{y}, \mu, \sigma_1^2, \sigma_2^2)}{f(\boldsymbol{x}', \boldsymbol{y}', \mu, \sigma_1^2, \sigma_2^2)} = e^{\frac{1}{2\sigma_1^2}(\sum {x_i'}^2 - \sum x_i^2) + \frac{1}{2\sigma_2^2}(\sum {y_j'}^2 - \sum y_j^2)}$$

$$.e^{\frac{\mu}{\sigma_1^2}(\sum x_i - \sum x_i') + \frac{\mu}{\sigma_2^2}(\sum y_j - \sum y_i')}$$

This is independent of $(\mu, \sigma_1^2, \sigma_2^2)$ iff $(\sum x_i, \sum x_i^2, \sum y_j, \sum y_j^2) = (\sum x_i', \sum {x_i'}^2, \sum y_j', \sum {y_j'}^2)$.

Thus $T = (\sum X_i, \sum X_i^2, \sum Y_j, \sum Y_j^2)$ is MSS.

Example cont'd

However, T is not complete.

Let $g(T) = \frac{\sum X_i}{m} - \frac{\sum Y_j}{n}$.

Then, $E[g(T)] = 0, \quad \forall\ (\mu, \sigma_1^2, \sigma_2^2)$.

But, $P[g(T) \neq 0] = 1$.

Thus, MSS is not complete.

Basu' Theorem

Let $T(\mathbf{X})$ be a complete and sufficient statistic and $V(\mathbf{X})$ be ancillary for θ.

Then, $T(\mathbf{X})$ and $V(\mathbf{X})$ are independently distributed.

It allows us to conclude that the two statistics are independent without ever finding the joint distribution of the two statistics.

Proof

Since $V(\mathbf{X})$ is an ancillary statistic, $P(V(\mathbf{X}) = v)$ does not depend on θ.

Also, the conditional probability,

$$P(V(\mathbf{X}) = v | T(\mathbf{X}) = t) = P(\mathbf{X} \in \{\mathbf{x} : V(\mathbf{x}) = v\} | T(\mathbf{X}) = t),$$

does not depend on θ because $T(\mathbf{X})$ is a sufficient statistic.

Thus, to show that $V(\mathbf{X})$ and $T(\mathbf{X})$ are independent, it suffices to show that

$$P(V(\mathbf{X}) = v | T(\mathbf{X}) = t) = P(V(\mathbf{X}) = v) \qquad (1)$$

for all possible values of $t \in \mathcal{T}$.

Proof cont'd

Now,

$$P(V(\mathbf{X}) = v) = \sum_{t \in \mathfrak{T}} P(V(\mathbf{X}) = v | T(\mathbf{X}) = t) P_\theta(T(\mathbf{X}) = t). \tag{2}$$

Furthermore, since $\sum_{t \in \mathfrak{T}} P_\theta(T(\mathbf{X}) = t) = 1$, we can write

$$P(V(\mathbf{X}) = v) = \sum_{t \in \mathfrak{T}} P(V(\mathbf{X}) = v) P_\theta(T(\mathbf{X}) = t). \tag{3}$$

Therefore, if we define the statistic

$$g(t) = P(V(\mathbf{X}) = v | T(\mathbf{X}) = t) - P(V(\mathbf{X}) = v),$$

the above two equations show that

$$E_\theta g(T) = \sum_{t \in \mathfrak{T}} g(t) P_\theta(T(\mathbf{X}) = t) = 0, \quad \forall\ \theta. \tag{4}$$

Proof cont'd

Explanation of (4):

$$
\begin{aligned}
& E_\theta g(T) \\
=& \sum_{t \in \mathcal{T}} [P(V(\mathbf{X}) = v | T(\mathbf{X}) = t) - P(V(\mathbf{X}) = v)] P_\theta(T(\mathbf{X}) = t) \\
=& \sum_{t \in \mathcal{T}} [P(V(\mathbf{X}) = v | T(\mathbf{X}) = t)\ P_\theta(T(\mathbf{X}) = t) - P(V(\mathbf{X}) = v) P_\theta(T(\mathbf{X}) = \\
=& P(V(\mathbf{X}) = v) - P(V(\mathbf{X}) = v) \\
=& 0.
\end{aligned}
$$

Since $T(\mathbf{X})$ is a complete statistic, this implies that $g(t) = 0$ for all possible values $t \in \mathcal{T}$. Hence, (1) is verified.

Example 1

Let $X_1, \ldots, X_n \sim N(\mu, 1)$.

Define $V = (X_1 - X_1, X_3 - X_1, \ldots, X_n - X_1)$ and $T = \sum X_i$.

Here, V is an ancillary statistic and T is a complete and sufficient statistic.

Thus, V and T are independently distributed.

Thanks for your patience!