Ancillary Statistics

Suppose $X \sim P_\theta$, $\theta \in \Theta$.

A statistic is <u>ancillary</u> if its distn. does not depend on θ.

More precisely,
A statistic $S(X)$ is ancillary for Θ if its distn. is the same for all $\theta \in \Theta$.
That is, $P_\theta(S(X) \in A)$ is constant for $\theta \in \Theta$ for any set A.

<u>Example:</u> $X = (X_1, \ldots, X_n)$ iid $N(\mu, \sigma^2)$.

Let $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$.

We know $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

or equivalently $s^2 \sim \frac{\sigma^2}{n-1} \chi^2_{n-1}$

so that the distn. of s^2 depends upon σ^2 but <u>not</u> on μ.

Thus s^2 is ancillary for

$$\Theta_1 = \{(\mu, \sigma^2) : \sigma^2 = \sigma_0^2\},$$

but is <u>not</u> ancillary for

$$\Theta_2 = \{(\mu, \sigma^2) : \sigma^2 > 0\}.$$

Let $\psi(x)$ be a fixed density.

Location family (LF) of densities

$$f(x|\theta) = \psi(x-\theta) \quad (-\infty < \theta < \infty)$$

Scale Family (SF) of densities

$$f(x|\theta) = \frac{1}{\theta}\psi\left(\frac{x}{\theta}\right) \quad (\theta > 0)$$

Location-Scale Family (LSF) of densities

$$f(x|\underbrace{\mu,\sigma}_{\theta}) = \frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right) \quad \left(\begin{array}{c}\sigma>0,\\ -\infty<\mu<\infty\end{array}\right)$$

If $X \sim f(\cdot|\theta)$ and $Z \sim \psi(\cdot)$, then

(LF) $X \stackrel{d}{=} Z + \theta \quad (X - \theta \stackrel{d}{=} Z)$

(SF) $X \stackrel{d}{=} \theta Z \quad (X/\theta \stackrel{d}{=} Z)$

(LSF) $X \stackrel{d}{=} \sigma Z + \mu \quad ((X-\mu)/\sigma \stackrel{d}{=} Z)$

If $\underset{\sim}{X} = (X_1, \ldots, X_n)$ is iid $\sim f(\cdot|\theta)$ and $\underset{\sim}{Z} = (Z_1, \ldots, Z_n)$ is iid $\sim \psi(\cdot)$, then

(LF) $\underset{\sim}{X} \stackrel{d}{=} \underset{\sim}{Z} + \theta \underset{\sim}{1}$	$\underset{\sim}{X} - \theta \stackrel{d}{=} \underset{\sim}{Z}$
(SF) $\underset{\sim}{X} \stackrel{d}{=} \theta \underset{\sim}{Z}$	$\underset{\sim}{X}/\theta \stackrel{d}{=} \underset{\sim}{Z}$
(LSF) $\underset{\sim}{X} \stackrel{d}{=} \sigma \underset{\sim}{Z} + \mu \underset{\sim}{1}$	$(\underset{\sim}{X} - \mu)/\sigma \stackrel{d}{=} \underset{\sim}{Z}$

Examples of Location families:

- Uniform$(\theta, \theta + 1)$ distributions ($\theta \in \Theta = \mathbb{R}$) with pdf $f(x \mid \theta) = I(\theta \leq x \leq \theta + 1)$.
- Cauchy location family with pdf
$$f(x \mid \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}.$$
- $N(\mu, \sigma_0^2)$ distributions ($\mu \in \mathbb{R}$ unknown, $\sigma^2 = \sigma_0^2$ known)

Examples of Scale families:

- Uniform$(0, \theta)$ distributions ($\theta > 0$ unknown) with pdf $f(x \mid \theta) = \theta^{-1} I(0 \leq x \leq \theta)$
- Cauchy scale family with pdf
$$f(x \mid \theta) = \frac{1}{\theta\pi\,[1 + (x/\theta)^2]}.$$
- $N(0, \sigma^2)$ distributions with $\sigma^2 > 0$ unknown.
- Exponential(β) distributions ($\beta > 0$ unknown) with pdf $f(x \mid \beta) = \beta^{-1} e^{-x/\beta} I(x \geq 0)$.

Examples of Location-Scale families:

- Uniform(α, β), $-\infty < \alpha < \beta < \infty$ (all uniform distns).
- $N(\mu, \sigma^2)$, $\mu \in \mathbb{R}$, $\sigma^2 > 0$ (all normal distns).

Facts :

① If $\underset{\sim}{X} = (X_1, \dots, X_n)$ is iid from a LF and $S(\underset{\sim}{x})$ is a location invariant function, [$S(\underset{\sim}{x} + c\underset{\sim}{1}) = S(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathbb{R}^n$ and $c \in \mathbb{R}$] then $S(\underset{\sim}{X})$ is ancillary.

② If $\underset{\sim}{X} = (X_1, \dots, X_n)$ is iid from a SF and $S(\underset{\sim}{x})$ is a scale invariant function, [$S(c\underset{\sim}{x}) = S(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathbb{R}^n$ and $c > 0$] then $S(\underset{\sim}{X})$ is ancillary.

③ If $\underset{\sim}{X} = (X_1, \dots, X_n)$ is iid from a LSF and $S(\underset{\sim}{x})$ is location-scale invariant, [$S(a\underset{\sim}{x} + b\underset{\sim}{1}) = S(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathbb{R}^n$, $a > 0$, $b \in \mathbb{R}$.] then $S(\underset{\sim}{X})$ is ancillary.

Proofs:

Let $\underset{\sim}{X} = (X_1, \ldots, X_n)$ be iid $\sim f(\cdot|\theta)$

and $\underset{\sim}{Z} = (Z_1, \ldots, Z_n)$ be iid $\sim \psi(\cdot)$.

① Since $\underset{\sim}{X} \overset{d}{=} \underset{\sim}{Z} + \theta \underset{\sim}{1}$ we have

$$P(S(\underset{\sim}{X}) \in A) = P(S(\underset{\sim}{Z} + \theta \underset{\sim}{1}) \in A)$$

(does not involve θ) $$= P(S(\underset{\sim}{Z}) \in A)$$

by location invariance of S.

② Since $\underset{\sim}{X} \overset{d}{=} \theta \underset{\sim}{Z}$ we have

$$P(S(\underset{\sim}{X}) \in A) = P(S(\theta \underset{\sim}{Z}) \in A)$$

$$= P(S(\underset{\sim}{Z}) \in A)$$

by scale invariance of S.

③ Since $\underset{\sim}{X} \overset{d}{=} \sigma \underset{\sim}{Z} + \mu \underset{\sim}{1}$ we have

$$P(S(\underset{\sim}{X}) \in A) = P(S(\sigma \underset{\sim}{Z} + \mu \underset{\sim}{1}) \in A)$$

$$= P(S(\underset{\sim}{Z}) \in A)$$

by location-scale invariance of S.

Alternate Proof for 3:

$$P(S(\underset{\sim}{X}) \in A) = P\left(S\left(\frac{\underset{\sim}{X} - \mu}{\sigma}\right) \in A\right)$$

since $S(\underset{\sim}{X}) = S\left(\frac{\underset{\sim}{X} - \mu}{\sigma}\right)$
by location-scale invariance of S

$$= P(S(\underset{\sim}{Z}) \in A)$$

since $\frac{\underset{\sim}{X} - \mu}{\sigma} \overset{d}{=} \underset{\sim}{Z}$.

But this does not depend on $\theta = (\mu, \sigma)$. Since A is arbitrary, this shows that $\mathcal{L}(S(\underset{\sim}{X}))$ does not depend on θ.

Location Invariant Statistics

$\rightarrow$ $S(X) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ is location invariant:

$$S(X+c) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i + c - (\bar{X}+c))^2 = S(X)$$

since the c's cancel.

Here we are using the fact that

$$\bar{X}(X+c) = \frac{1}{n} \sum_{i=1}^{n} (X_i + c)$$

$$= \left(\frac{1}{n} \sum_{i=1}^{n} X_i\right) + c = \bar{X}(X) + c$$

$\longrightarrow$ $S(X) = \sum_{i=1}^{n} | X_i - \text{Median}(X) |$ is location invariant:

$$S(X+c) = \sum_{i=1}^{n} |X_i + c - \underbrace{\text{Median}(X+c)}_{= \text{Median}(X)+c}|$$

$= S(X)$ since the c's cancel.

$\longrightarrow$ $S(X) = \max_i X_i - \min_i X_i = X_{(n)} - X_{(1)}$ is location invariant:

$$S(X+c) = \max(X_i + c) - \min(X_i + c)$$
$$= (\max X_i) + c - ((\min X_i) + c)$$

$= S(X)$ since the c's cancel

→ The vector $S(X) =$

$$(X_2 - X_1, X_3 - X_1, \ldots, X_n - X_1)$$

is location invariant by a similar argument.

Scale Invariant Statistics

→ $t = \dfrac{\bar{X} - 0}{S/\sqrt{n}}$ is scale invariant:

$$t(cx) = \frac{c\bar{x}}{cs/\sqrt{n}} = t(x)$$ since the c's cancel.

Here we have used:

$$\bar{X}(cx) = \frac{1}{n}\sum_{i=1}^{n} cx_i = c\left(\frac{1}{n}\sum x_i\right)$$
$$= c\bar{X}(x),$$

$$S(cx) = \sqrt{\frac{1}{n-1}\sum (cx_i - c\bar{x})^2}$$
$$= c\sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$
$$= c\, S(x).$$

$\rightarrow$ $S(X) = \dfrac{\bar{X}}{X_{(n)}}$ is scale invariant:

$$S(cX) = \frac{c\bar{X}}{cX_{(n)}} = S(X) \text{ for } c>0$$

since the c's cancel.

<u>Note</u>: $S(cX) \neq S(X)$ for $c \leq 0$.

$\rightarrow$ $S(X) = \left(\dfrac{X_1}{\Sigma X_i}, \dfrac{X_2}{\Sigma X_i}, \ldots, \dfrac{X_n}{\Sigma X_i}\right)$

is scale invariant.

$\rightarrow$ $S(X) = \left(\dfrac{X_1}{S}, \dfrac{X_2}{S}, \ldots, \dfrac{X_n}{S}\right)$

is scale invariant

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

<u>Location-Scale Invariant Statistics</u>

$\rightarrow$ sample skewness is proportional to

$$S(X) = \frac{\Sigma(X_i-\bar{X})^3}{\left[\Sigma(X_i-\bar{X})^2\right]^{3/2}}.$$

→ sample kurtosis is proportional to

$$K(X) = \frac{\sum(x_i-\bar{x})^4}{\left[\sum(x_i-\bar{x})^2\right]^2}.$$

They are both location-scale invariant.

It suffices to show:

(1) $S(aX) = S(X)$ for $a>0$, and

(2) $S(X+b) = S(X)$ for all b.

Part (2) follows from

$$(x_i+b)-(\bar{x}+b) = x_i - \bar{x}.$$

Part (1) follows from

$$\sum(cx_i - c\bar{x})^m = c^m \sum(x_i-\bar{x})^m.$$

→ The standardized residuals

$$Z = (z_1, z_2, \ldots, z_n) \text{ with } z_i = \frac{x_i - \bar{x}}{s}$$

are location-scale invariant.