1 Ancillary statistics

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

Definition 1. *A statistics is ancillary if its distribution does not depend on θ. More precisely, a statistic $S(X)$ is ancillary for Θ it its distribution 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.*

Example: $X = (X_1, \ldots, X_n)$ iid $\mathrm{N}(\mu, \sigma^2)$. Let

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

We know

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \Leftrightarrow S^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$$

so that the distribution of S^2 depends upon σ^2 but not on μ. Thus S^2 is ancillary for

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

but is not ancillary for

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

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

1. Location Family (LF) of densities: $f(x \mid \theta) = \psi(x - \theta), -\infty < \theta < \infty$.
2. Scale Family (SF) of densities: $f(x \mid \theta) = \frac{1}{\theta}\psi(\frac{x}{\theta}), \theta > 0$.
3. Location-Scale Family (LSF) of densities: $f(x \mid \mu, \sigma) = \frac{1}{\sigma}\psi(\frac{x-\mu}{\sigma}), (\sigma > 0, -\infty < \mu < \infty)$.

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

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

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

3. (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 $f(\cdot \mid \theta)$ and $\underset{\sim}{Z} = (Z_1, \ldots, Z_n)$ iid $\psi(\cdot)$, then

1. (LF) $\underset{\sim}{X} \stackrel{d}{=} \underset{\sim}{Z} + \theta \underset{\sim}{1} \quad (\underset{\sim}{X} - \theta \underset{\sim}{1} \stackrel{d}{=} \underset{\sim}{Z})$

2. (SF) $\underset{\sim}{X} \stackrel{d}{=} \theta \underset{\sim}{Z} \quad (\underset{\sim}{X}/\theta \stackrel{d}{=} \underset{\sim}{Z})$

3. (LSF) $\underset{\sim}{X} \stackrel{d}{=} \sigma \underset{\sim}{Z} + \mu \underset{\sim}{1} \quad ((\underset{\sim}{X} - \mu \underset{\sim}{1})/\sigma \stackrel{d}{=} \underset{\sim}{Z})$

1. Examples of Location families:
   (a) Unif$(\theta, \theta+1)$ distributions ($\theta \in \Theta = \mathbb{R}$) with pdf $f(x \mid \theta) = I(\theta \le x \le \theta + 1)$
   (b) Cauchy location family with pdf
   $$f(x \mid \theta) = \frac{1}{\pi\{1 + (x-\theta)^2\}}.$$
   (c) N(μ, σ_0^2) distributions with $\mu \in \mathbb{R}$ unknown, σ_0^2 known.
2. Examples of Scale families:
   (a) Unif$(0, \theta)$ distributions ($\theta > 0$ unknown) with pdf $f(x \mid \theta) = \theta^{-1} I(0 \le x \le \theta)$
   (b) Cauchy scale family with pdf
   $$f(x \mid \theta) = \frac{1}{\theta\pi\{1 + (x/\theta)^2\}}.$$
   (c) N$(0, \sigma^2)$ distributions with $\sigma^2 > 0$ unknown.
   (d) Exp(β) distributions ($\beta > 0$ unknown) with pdf $f(x \mid \beta) = \beta^{-1} e^{-x/\beta} I(x \ge 0)$.
3. Examples of Location-Scale families:
   (a) Unif$(\alpha, \beta), -\infty < \alpha < \beta < \infty$(all uniform distributions)
   (b) N$(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0$ (all normal distributions).

1.1 Facts

1. If $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ 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 $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ 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 $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ is iid from a LSF and $S(\underset{\sim}{x})$ is a location-scale invariant function, ($S(a\underset{\sim}{x} + b\underset{\sim}{1}) = S(\underset{\sim}{x})$ for all $\underset{\sim}{x} \in \mathbb{R}^n$ and $a > 0, b \in \mathbb{R}$), then $S(\underset{\sim}{X})$ is ancillary.

Proof. Let $\underset{\sim}{X} = (X_1, X_2, \ldots, X_n)$ be iid $f(\cdot \mid \theta)$ and $\underset{\sim}{Z} = (Z_1, \ldots, Z_n)$ be iid $\psi(\cdot \mid \theta)$.

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

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

   which does not involve θ by the location invariance of S.

2. Since $\underset{\sim}{X} \stackrel{d}{=} \theta\underset{\sim}{Z}$, we have

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

   which does not involve θ by the scale invariance of S.

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

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

   which does not involve μ, σ by the location-scale invariance of S.

□

1.2 Location Invariant Statistics

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

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

Here $\overline{X+c} = (1/n)\sum_{i=1}^{n}(X_i + c) = \bar{X} + c$.

2. $S(X) = \sum_{i=1}|X_i - \text{median}(X)|$ is location invariant:

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

3. $S(X) = \max X_i - \min X_i = X_{(n)} - X_{(1)}$ is location invariant:

$$S(X+c) = \max(X_i + c) - \min(X_i + c) = \max(X_i) + c - \min(X_i) - c = X_{(n)} - X_{(1)}.$$

4. The vector $S(X) = (X_2 - X_1, X_3 - X_2, \ldots, X_n - X_1)$ is location invariant by a similar argument.

1.3 Scale Invariant Statistics

1. $t = \frac{\bar{x}-0}{S/\sqrt{n}}$ is scale invariant as:

$$t(cx) = \frac{c\bar{x}}{cs/\sqrt{n}} = t(x)$$

since the c's cancel. Here we have used

$$\begin{aligned} \overline{cx} &= \frac{1}{n}\sum_{i=1}^{n} cx_i = c\frac{1}{n}\sum_{i=1}^{n} x_i = c\bar{x}, \\ S(cx) &= \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(cx_i - c\bar{x})^2} = c\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} = cS(x). \end{aligned}$$

2. $S(X) = \frac{\bar{X}}{X_{(n)}}$ is scale invariant:

$$S(cX) = \frac{c\bar{X}}{cX_{(n)}} = S(X)$$

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

3.
$$S(X) = \left(\frac{X_1}{\sum X_i}, \frac{X_2}{\sum X_i}, \ldots, \frac{X_n}{\sum X_i} \right)$$
is scale invariant.

1.4 Scale Invariant Statistics

1. Sample skewness is proportional to
$$S(X) = \frac{\sum (X_i - \bar{X})^3}{[\sum (X_i - \bar{X})^2]^{3/2}}.$$

2. Sample kurtosis is proportional to
$$S(X) = \frac{\sum (X_i - \bar{X})^4}{[\sum (X_i - \bar{X})^2]^2}.$$
They are location-scale invariant.

Proof. It suffices to show:

(a) $S(aX) = S(X)$ for $a > 0$, and
(b) $S(X + b) = S(X)$ for all b.

Part (b) follows from
$$(X_i + b) - (\bar{X} + b) = X_i - \bar{X}$$
Part (a) follows from
$$\sum (cx_i - c\bar{x})^m = c^m \sum (x_i - \bar{x})^m$$
□

3. The standardized residuals
$$z = (z_1, z_2, \ldots, z_n), z_i = \frac{x_i - \bar{x}}{S}$$
are location-scale invariant.

General comment: An ancillary statistic by itself can tell us nothing about θ, but when combined with other statistics, it may give information about θ.
Example: $X = (X_1, X_2, \ldots, X_n)$ iid Unif$(\theta, \theta + 1)$. We know $(X_{(1)}, X_{(n)})$ is MSS. Any 1-1 function of a MSS is also MSS. Therefore $(X_{(1)}, X_{(n)} - X_{(1)})$ is MSS. We cannot drop $X_{(n)} - X_{(1)}$ without losing information about θ. But $X_{(n)} - X_{(1)}$ is ancillary ! It is ancillary because Unif$(\theta, \theta + 1)$ is a location family, and $X_{(n)} - X_{(1)}$ is a location invariant statistic.