3. In some senses sufficient statistics contain all the information about θ that is available in the sample, here we consider a different sort of statistic that has a complementary purpose.

   Ancillary Statistics: A statistic $S(\mathbf{x})$ whose distribution does not depend on θ is called an *ancillary statistic*.

   Examples: (1). If $T(\mathbf{X})$ is sufficient, then $P(\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x}))$ is an ancillary statistic.

   (2) $X_1, \cdots, X_n$ i.i.d. from location parameter family with cdf $F(x - \theta)$, $-\infty < \theta < \infty$. Then the range $R = X_{(n)} - X_{(1)}$ is an ancillary statistic for θ.

   Proof: Let $Z_i = X_i - \theta$ for $i = 1, \ldots, n$. Then $P(Z_i \le z) = P(X_i \le \theta + z) = F(\theta + z - \theta) = F(z)$ for all z. Note that the Z_i's are not statistics and cannot be used to construct statistical procedures (since they are unobservable), but their distributions do not depend on θ. Hence,

$$\begin{aligned} P(R \le r) &= P(\max_i X_i - \min_i X_i \le r) \\ &= P(\max_i (Z_i + \theta) - \min_i (Z_i + \theta) \le r) \qquad Z_1, \cdots, Z_n \text{ i.i.d. from } F(x) \\ &= P(\max_i Z_i - \min_i Z_i \le r), \end{aligned}$$

   which does not depend on θ since the distribution of theh Z_i's does not depend on θ.

   (3) $X_1, \cdots, X_n$ i.i.d. from scale family with cdf $F(x/\theta)$. Then $(X_1/X_n, X_2/X_n, \cdots, X_{n-1}/X_n)$ is ancillary for θ, and so is any function of these quantities. In particular, $\frac{X_n}{X_1 + \cdots + X_n}$ is an ancillary statistic.

   Remark: While an ancillary statistic **alone** would give us no information about θ, it can sometimes give important information (with other statistics).

4. A minimal sufficient statistic is a statistic that has achieved the maximal amount of data reduction possible while still retaining all the information about the parameter θ. Intuitively, one may expect that a minimal sufficient statistic eliminates all the extraneous information in the sample, retaining only that piece of information about θ, and thus one may suspect that the minimal sufficient statistic is unrelated to ancillary statistics. However, this is not necessarily true. This leads to the definition of complete statistic.

   Complete Statistics: Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability functions is called *complete* if $E_\theta(g(T)) = 0$ for **all** θ implies that $P_\theta(g(T) = 0) = 1$ for **all** θ. We also say that $T(\mathbf{X})$ is a *complete statistic*.

   Examples: (1) Suppose $T \sim Binomial(n, p)$ with $0 < p < 1$. Show that T is a complete statistic.

   Proof: Suppose g satisfies $E_p(g(T)) = 0$ for all $0 < p < 1$. Then

$$0 = E_p(g(T)) = \sum_{t=0}^{n} g(t) \binom{n}{t} p^t (1-p)^{n-t} = (1-p)^n \sum_{t=0}^{n} g(t) \binom{n}{t} (\frac{p}{1-p})^t.$$

   This holds for all $0 < p < 1$ if and only if $g(t) = 0$ for all $t = 0, 1, \cdots, n$. (why???) Hence T is a complete statistic. □

   (2) We showed that $T(\mathbf{X}) = X_{(n)}$ is a sufficient statistic in the $U(0, \theta)$ family, $\theta > 0$. Is it complete?

   Solution: Note that $X_{(n)}$ has the cdf $F_{X_{(n)}}(u) = (u/\theta)^n$ for $0 \le u \le \theta$, and has pdf

$$f_{X_{(n)}}(u|\theta) = \begin{cases} n\theta^{-n}u^{n-1}, & \text{if } 0 < u < \theta; \\ 0, & \text{otherwise.} \end{cases}$$

   Suppose g satisfies $E_\theta g(X_{(n)}) = 0$ for all $\theta > 0$. Then

$$\int_0^\theta g(u)u^{n-1}du = 0 \quad \text{for all } \theta > 0.$$

   Applying the result of differentiation of an integral yields that $g(\theta)\theta^{n-1} = 0$ almost everywhere for $\theta \ge 0$. Hence $g(u) = 0$ almost everywhere. Therefore, $X_{(n)}$ is complete and sufficient for $\theta \in (0, \infty)$.

   (3) **Example 6 (cont.)** We showed that $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is a minimal sufficient statistic in the $U(\theta, \theta + 1)$ family. Is it complete when $n \ge 2$?

   <u>**Solution:**</u> Note that $U(\theta, \theta+1)$ form a location parameter family. Let $Z_1, \cdots, Z_n$ be i.i.d. $U(0, 1)$, and denote $c_1 = EZ_{(1)} (= \frac{1}{n+1})$ and $c_2 = EZ_{(n)} (= \frac{n}{n+1})$. Then c_1 and c_2 are two constants which only depend on n and do not depend on θ. Thus, $E_\theta(X_{(1)} - c_1) = \theta$ and $E_\theta(X_{(n)} - c_2) = \theta$. This suggests us to consider $g(t_1, t_2) = t_1 - t_2 - c_1 + c_2$. Then

$$E_\theta g(T(\mathbf{X})) = E_\theta(X_{(1)} - X_{(n)} - c_1 + c_2) = E_\theta((\theta + Z_{(1)}) - (\theta + Z_{(n)}) - c_1 + c_2) = 0,$$

but $P_\theta(g(T) = 0) = P(Z_{(1)} - Z_{(n)} = c_1 - c_2) \neq 1$. Thus $T(\mathbf{X}) = (X_{(1)}, X_{(n)})$ is **not complete**, although it is a minimal sufficient statistic. $\square$

What if $n = 1$, i.e., if a random variable $X \sim U(\theta, \theta + 1)$, is X complete?

Solution: No. Consider $g(x) = sin(2\pi x)$, then

$$E_\theta(g(X)) = \int_\theta^{\theta+1} sin(2\pi x)dx = 0.$$

So X itself is not a complete statistic for θ. $\square$

(4) Suppose that X_1, X_2 are independent $N(\theta, 1)$, and consider $T = (X_1, X_2)$, the pair itself. Then T is not a complete statistic (though it is sufficient).

To see this, let $g(t_1, t_2) = t_1 - t_2$. Then $E_\theta(g(T)) = E_\theta(X_1 - X_2) = \theta - \theta = 0$, for all θ, but $g \neq 0$.

Theorem 6.2.25 (complete statistic in the exponential family). Assume $X_1, \cdots, X_n$ i.i.d. from an exponential family with pdf or pmf of the form

$$f_\theta(x) = h(x)c(\theta) \exp\Big(\sum_{j=1}^{k} w_j(\theta)t_j(x)\Big).$$

If $\{w_1(\theta), \cdots, w_k(\theta) : \theta \in \Theta\}$ contains an open set in $\mathcal{R}^k$, then the statistic

$$T(\mathbf{X}) = \Big(\sum_{i=1}^{n} t_1(X_i), \cdots, \sum_{i=1}^{n} t_k(X_i)\Big)$$

is complete.

Remark: Note that this theorem does not apply to the family of $N(\mu, \mu^2)$ in Example 5, as (θ, θ^2) does not include a two-dimensional open set.

5. The following theorem is useful to deduce the independence of two statistics without ever finding the joint distribution of the two statistics.

 Basu's Theorem: If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic $S(\mathbf{X})$.

 See Theorem 6.2.24 on page 287 for the detailed proof.

 Solution: The essential idea: for fixed s, consider the function

$$g(t) = P(S(\mathbf{X}) = s|T(\mathbf{X}) = t) - P(S(\mathbf{X}) = s) \qquad \text{does not depend on } \theta$$

 Then $E_\theta(g(T(\mathbf{X}))) = 0$ for all θ. Since T is complete, this implies that $g(t) = 0$ for all possible values of t. Hence S and T are independent. □

 Remark: the word "minimal" is redundant in the statement, as any "complete sufficient" statistic is also a "minimal sufficient" statistic if the latter exists.

 Example: Let $X_1, \cdots, X_n$ be i.i.d. with exponential distribution with pdf

$$f_\theta(x) = \frac{1}{\theta}\exp(-\frac{x}{\theta})I(x > 0)$$

 for $\theta > 0$. Use Basu's Theorem to show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ and $g(\mathbf{X}) = \frac{X_n}{X_1+\cdots+X_n}$ are independent.

 Solution: First, the exponential distributions form a scale parameter family and thus, by our previous results, $g(\mathbf{X})$ is an ancillary statistic.
 Second, the exponential distributions also form an exponential family with $t(x) = x$ and $w(\theta) = 1/\theta$. Thus, combining the theorem for sufficient statistic in the exponential family with the theorem for complete statistic in the exponential family yields that $T(\mathbf{X})$ is a sufficient and complete statistic. It is also easy to verify that $T(\mathbf{X})$ is minimal.
 Hence, by Basu's Theorem, $T(\mathbf{X}) = \sum_{i=1}^n X_i$ and $g(\mathbf{X}) = \frac{X_n}{X_1+\cdots+X_n}$ are independent.
 Furthermore, we can use this independence to calculate $E_\theta(g(\mathbf{X}))$. To see this,

$$\theta = E_\theta(X_n) = E_\theta(T(\mathbf{X})g(\mathbf{X})) = E_\theta(T(\mathbf{X}))E_\theta(g(\mathbf{X})) = (n\theta)E_\theta(g(\mathbf{X})),$$

 which implies that $E_\theta(g(\mathbf{X})) = n^{-1}$ for any θ. □

In summary:

- A **sufficient** statistic retains at least enough information about θ from the data
- A **complete** statistic retains no irrelevant information about θ (it is possible a complete statistic may retain no information).

In a given statistical problem, the minimal sufficient statistic most likely exists, but the complete sufficient statistic may or may not exist.

The Relationship between minimal sufficient statistics and complete sufficient statistics can be summarized by the following two statements:

- If some sufficient statistic is complete, then so is any minimal sufficient statistic.
- A sufficient statistic which is not minimal cannot be complete (i.e., "A complete statistic is also minimal sufficient.")

Hence, in a given statistical problem, either **"no complete sufficient statistics exist"**, or **"The class of complete sufficient statistics is identical with the class of minimal sufficient statistics"**.

Thus we can always answer the question of whether or not there is a complete sufficient statistic by seeing whether or not a minimal sufficient statistic is complete.

Therefore, in a typical problem, you will be asked first to find a minimal sufficient statistic, and then to verify whether it is complete or not.