Regular models. Either (i) all P_θ are continuous with density $f(.|\theta)$, or (ii) all P_θ are discrete with mass function $p(x|\theta)$ and there exists a countable set $\mathcal{S} = \{x_1, x_2, \ldots\}$ independent of θ such that $\sum_{i=1}^{\infty} p(x_i|\theta) = 1$.

For example, the following models are not regular under our definition.

Example. Let X be the weight of a randomly chosen product from a population, and assume $X \sim N(\theta, 10^2)$. However the weighing device cannot show weights above a certain level c, so it fixes (censors) such weights at c. Then the observed weight, Y has the distribution, $Y = \begin{cases} X & \text{if } X < c; \\ c & \text{if } X \geq c. \end{cases}$. Then Y has a continuous part as well as a point mass (at c).

Example. Let X be discrete with pmf $p(x|\theta) = \begin{cases} 1/2 & \text{if } x = \theta; \\ 1/2 & \text{if } x = \theta + 1, \end{cases}$ where $\theta \in \mathcal{R}$. Here $\mathcal{S}$ is not countable.

Sufficient Statistics

Statistical inference is our objective, i.e., making inferences about unknown parameters in the model. The first step in this direction is data reduction or compression – condensing all the useful information and removing all the irrelevant pieces. This will allow modeling only the informative parts of the data. For example, suppose the mean yield of fruit in a farm is of interest, and a random sample of trees is investigated for this purpose. Note that the sample data may look different depending on the order in which one records the yields from the trees in the sample, but this order of observations is not relevant for inferential purposes.

Let $\mathbf{X}$ denote sample data or the list of observations. Then any real or vector-valued function $T(\mathbf{X})$ is called a statistic. Examples are:
$T(X_1, \ldots, X_n) = \bar{X}$,
$T(X_1, \ldots, X_n) = \sum_{i=1}^{n} (X_i - \bar{X})^2$,
$T(X_1, \ldots, X_n) = (\bar{X}, \sum_{i=1}^{n} (X_i - \bar{X})^2)$.
Intuitively, a statistic $T(\mathbf{X})$ is sufficient if it contains all the useful information about the quantities of interest.

Definition. A statistic $T(\mathbf{X})$ is called sufficient for a parameter θ, or sufficient for a family of distributions P_θ indexed by θ if the conditional distribution of $\mathbf{X}$ given $T(\mathbf{X}) = t$ does not involve θ. i.e.,
$P_\theta(a < X \leq b | T(X) = t)$ is independent of θ for all a, b if X is a random variable.

Note. When $T(\mathbf{X})$ is sufficient, if you know its value, you don't care what

$\mathbf{X}$ is anymore.

Example. Suppose $X_1, \ldots, X_n$ are i.i.d. Poisson(λ).
Claim: $S = \sum_{i=1}^n X_i$ is sufficient for λ.
Note that

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n|\lambda) = \prod_{i=1}^n \exp(-\lambda)\frac{\lambda^{x_i}}{x_i!} = \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}, \text{ and}$$

$$f_S(s|\lambda) = \exp(-n\lambda)\frac{(n\lambda)^s}{s!}.$$

Therefore,

$$\begin{aligned}
f_{(X_1,\ldots,X_n)|S=s}(x_1, \ldots, x_n|\lambda) &= \frac{P_\lambda(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = x_n, S = s)}{P_\lambda(S = s)} \\
&= \frac{P_\lambda(X_1 = x_1, \ldots, X_{n-1} = x_{n-1}, X_n = s - \sum_{i=1}^{n-1} x_i)}{P_\lambda(S = s)} \quad \text{if } \textstyle\sum_{i=1}^n x_i = s \\
&= \frac{\exp(-n\lambda) \ \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}}{\exp(-n\lambda) \ \frac{(n\lambda)^s}{s!}} \quad \text{if } \textstyle\sum_{i=1}^n x_i = s \\
&= \frac{s!}{\prod_{i=1}^n x_i!}\left(\frac{1}{n}\right)^s \quad \text{if } \textstyle\sum_{i=1}^n x_i = s \ ,
\end{aligned}$$

which is free of λ. In fact, the conditional distribution above is Multinomial($s; \frac{1}{n}, \ldots, \frac{1}{n}$).

One needs to guess T for using the above definition. Instead, there is the following useful and important result.

Factorization Theorem (Neyman-Fisher). Let $f(\mathbf{x}|\theta)$ be the density of $\mathbf{X}$ under the probability model $P_\theta, \theta \in \Theta$. Then, if the model is regular, a statistic $T(\mathbf{X})$ is sufficient for θ iff there exists a function $g(t, \theta)$ and a function $h(\mathbf{x})$ such that

$$f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x}).$$

i.e., one is able to factor f into two parts, one involving θ and data through T, and the other involving data only. If $\mathbf{x} \in \mathcal{R}^n$, then we have
$T : \mathcal{R}^n \to I \subseteq \mathcal{R}^k, k \leq n$,
$g : I \times \Theta \to \mathcal{R}^+$,
$h : \mathcal{R}^n \to \mathcal{R}^+$. g and h are not unique.

Proof. For the discrete case only. The continuous case is similar, but a rigorous proof requires measure theoretical arguments.

Let $\mathcal{S} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots\}$ be the (countable) sample space, the set of all possible values of $\mathbf{X}$. Let $t = T(\mathbf{x})$. Then $T = T(\mathbf{X})$ is discrete and $\sum_{i=1}^{\infty} f_T(t_i|\theta) = P_\theta(T = t_i, 1 \le i < \infty) = 1$ for all θ.
if part: There exist g and h such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. Need to show that T is sufficient. i.e., show that $P_\theta[\mathbf{X} = \mathbf{x}_j|T(\mathbf{X}) = t_i]$ does not involve θ for each i and j. Since $P_\theta[\mathbf{X} = \mathbf{x}_j|T(\mathbf{X}) = t_i]$ is defined when $P_\theta[T = t_i] > 0$ only, it is enough to show that $P_\theta[\mathbf{X} = \mathbf{x}_j|T(\mathbf{X}) = t_i]$ is independent of θ when $\theta \in \Omega_i = \{\theta : P_\theta[T = t_i] > 0\}$, $i = 1, 2, \ldots$ Now, note

$$P_\theta[T = t_i] = \sum_{\{\mathbf{x}:T(\mathbf{x})=t_i\}} f(\mathbf{x}|\theta) = g(t_i, \theta) \sum_{\{\mathbf{x}:T(\mathbf{x})=t_i\}} h(\mathbf{x}).$$

If $\theta \in \Omega_i$, then

$$\begin{aligned} P_\theta[\mathbf{X} = \mathbf{x}_j|T(\mathbf{X}) = t_i] &= \frac{P_\theta[\mathbf{X} = \mathbf{x}_j, T(\mathbf{X}) = t_i]}{P_\theta[T = t_i]} \\ &= \begin{cases} \frac{f(\mathbf{x}_j|\theta)}{P_\theta[T=t_i]} & \text{if } T(\mathbf{x}_j) = t_i; \\ 0 & \text{if } T(\mathbf{x}_j) \neq t_i. \end{cases} \end{aligned}$$

Note that $P_\theta[\mathbf{X} = \mathbf{x}_j, T(\mathbf{X}) = t_i] = 0$ if $T(\mathbf{x}_j) \neq t_i$. Using the fact that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$, whenever $T(\mathbf{x}_j) = t_i$, we get

$$\begin{aligned} \frac{f(\mathbf{x}_j|\theta)}{P_\theta[T = t_i]} &= \frac{g(t_i, \theta)h(\mathbf{x}_j)}{g(t_i, \theta)\sum_{\{\mathbf{x}:T(\mathbf{x})=t_i\}} h(\mathbf{x})} \\ &= \frac{h(\mathbf{x}_j)}{\sum_{\{\mathbf{x}:T(\mathbf{x})=t_i\}} h(\mathbf{x})}, \end{aligned}$$

which is independent of θ. Thus $T = T(\mathbf{X})$ is sufficient for θ.

only if: Now we have that $T = T(\mathbf{X})$ is sufficient for θ. We want to find functions, g and h such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x}), \theta)h(\mathbf{x})$. Define $g(t_i, \theta) = P_\theta[T = t_i]$ and $h(\mathbf{x}) = P_\theta[\mathbf{X} = \mathbf{x}|T = T(\mathbf{x})]$. Then $h(\mathbf{x})$ is independent of θ by definition (of sufficiency). Also,

$$\begin{aligned} f(\mathbf{x}|\theta) &= P_\theta[\mathbf{X} = \mathbf{x}] = P_\theta[\mathbf{X} = \mathbf{x}, T = T(\mathbf{x})] \\ &= P_\theta[\mathbf{X} = \mathbf{x}|T = T(\mathbf{x})]P_\theta[T = T(\mathbf{x})] \\ &= h(\mathbf{x})g(T(\mathbf{x}), \theta), \end{aligned}$$

noting that $P_\theta[T = T(\mathbf{x})]$ is a function of $T(\mathbf{x})$ and θ only.