Statistic used	**Meaning in plain english**	**Reduction ratio**
$T(\underline{X}) = [X_1, \ldots, X_n]^T$,	entire data sample	$\mathrm{RR} = 1$
$T(\underline{X}) = [X_{(1)}, \ldots, X_{(n)}]^T$,	rank ordered sample	$\mathrm{RR} = 1/2$
$T(\underline{X}) = \overline{X}$,	sample mean	$\mathrm{RR} = 1/(2n)$
$T(\underline{X}) = [\overline{X}, \mathsf{s}^2]^T$,	sample mean and variance	$\mathrm{RR} = 1/n$

A natural question is: what is the maximal reduction ratio one can get away with without loss of information about $\underline{\theta}$? The answer is: the ratio obtained by compression to a quantity called a *minimal sufficient statistic*. But we are getting ahead of ourselves. We first need to define a plain old sufficient statistic.

3.5.2 DEFINITION OF SUFFICIENCY

Here is a warm up before making a precise definition of sufficiency. $T = T(\underline{X})$ is a **sufficient statistic** (SS) for a parameter $\underline{\theta}$ if it captures all the information in the data sample useful for inferring the value of $\underline{\theta}$. To put it another way: once you have computed a sufficient statistic you can store it and throw away the original sample since keeping it around would not add any useful information.

More concretely, let $\underline{X}$ have a cumulative distribution function (CDF) $F_{\underline{X}}(\underline{x}; \underline{\theta})$ depending on $\underline{\theta}$. A statistic $T = T(\underline{X})$ is said to be sufficient for $\underline{\theta}$ if the conditional CDF of $\underline{X}$ given $T = t$ is not a function of $\underline{\theta}$, i.e.,

$$F_{\underline{X}|T}(\underline{x}|T = t, \underline{\theta}) = G(\underline{x}, t), \tag{16}$$

where G is a function that does not depend on $\underline{\theta}$.

Specializing to a discrete valued $\underline{X}$ with probability mass function $p_{\underline{\theta}}(\underline{x}) = P_{\underline{\theta}}(\underline{X} = \underline{x})$, a statistic $T = T(\underline{X})$ is sufficient for $\underline{\theta}$ if

$$P_{\underline{\theta}}(\underline{X} = \underline{x}|T = t) = G(\underline{x}, t). \tag{17}$$

For a continuous r.v. $\underline{X}$ with pdf $f(\underline{x}; \underline{\theta})$, the condition (16) for T to be a sufficient statistic (SS) becomes:

$$f_{\underline{X}|T}(\underline{x}|t; \underline{\theta}) = G(\underline{x}, t). \tag{18}$$

Sometimes the only sufficient statistics are vector statistics, e.g. $T(\underline{X}) = \underline{T}(\underline{X}) = [T_1(\underline{X}), \ldots, T_K(\underline{X})]^T$. In this case we say that the T_k's are *jointly sufficient* for $\underline{\theta}$

The definition (16) is often difficult to use since it involves derivation of the conditional distribution of $\underline{X}$ given T. When the random variable $\underline{X}$ is discrete or continuous a simpler way to verify sufficiency is through the Fisher factorization (FF) property [33]

Fisher factorization (FF): $T = T(\underline{X})$ is a sufficient statistic for $\underline{\theta}$ if the probability density $f_{\underline{X}}(\underline{x}; \underline{\theta})$ of $\underline{X}$ has the representation

$$f_{\underline{X}}(\underline{x}; \underline{\theta}) = g(T, \underline{\theta})\ h(\underline{x}), \tag{19}$$

for some non-negative functions g and h. The FF can be taken as the operational definition of a sufficient statistic T. An important implication of the Fisher Factorization is that when the density function of a sample $\underline{X}$ satisfies (19) then the density $f_T(t;\underline{\theta})$ of the sufficient statistic T is equal to $g(t,\underline{\theta})$ up to a $\underline{\theta}$-independent constant $q(t)$ (see exercises at end of this chapter):

$$f_T(t;\underline{\theta}) = g(t,\underline{\theta})q(t).$$

Examples of sufficient statistics:

Example 1 *Entire sample*

$\underline{X} = [X_1, \ldots, X_n]^T$ is sufficient but not very interesting

Example 2 *Rank ordered sample*

$X_{(1)}, \ldots, X_{(n)}$ is sufficient when X_i's i.i.d.

Proof: Since X_i's are i.i.d., the joint pdf is

$$f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n f_\theta(x_{(i)}).$$

Hence sufficiency of the rank ordered sample $X_{(1)}, \ldots, X_{(n)}$ follows from Fisher factorization. $\diamond$

Example 3 *Binary likelihood ratios*

Let $\underline{\theta}$ take on only two possible values $\underline{\theta}_0$ and $\underline{\theta}_1$. Then, as $f(\underline{x};\underline{\theta})$ can only be $f(\underline{x};\underline{\theta}_0)$ or $f(\underline{x};\underline{\theta}_1)$, we can reindex the pdf as $f(\underline{x};\theta)$ with the scalar parameter $\theta \in \Theta = \{0,1\}$. This gives the binary decision problem: "decide between $\theta = 0$ versus $\theta = 1$." If it exists, i.e. it is finite for all values of $\underline{X}$, the "likelihood ratio" $\Lambda(\underline{X}) = f_1(\underline{X})/f_0(\underline{X})$ is sufficient for θ, where $f_1(\underline{x}) \stackrel{\text{def}}{=} f(\underline{x};1)$ and $f_0(\underline{x}) \stackrel{\text{def}}{=} f(\underline{x};0)$.

Proof: Express $f_\theta(\underline{X})$ as function of θ, f_0, f_1, factor out f_0, identify Λ, and invoke FF

$$\begin{aligned} f_{\underline{\theta}}(\underline{X}) &= \theta f_1(\underline{X}) + (1-\theta) f_0(\underline{X}) \\ &= \left(\underbrace{\theta \Lambda(\underline{X}) + (1-\theta)}_{g(T,\theta)} \right) \underbrace{f_0(\underline{X})}_{h(\underline{X})}. \end{aligned}$$

$\diamond$

Therefore to discriminate between two values $\underline{\theta}_1$ and $\underline{\theta}_2$ of a parameter vector $\underline{\theta}$ we can throw away all data except for the scalar sufficient statistic $T = \Lambda(\underline{X})$

Example 4 *Discrete likelihood ratios*

Let $\Theta = \{\underline{\theta}_1, \ldots, \underline{\theta}_p\}$ and assume that the vector of $p-1$ *likelihood ratios*

$$T(\underline{X}) = \left[\frac{f_{\theta_1}(\underline{X})}{f_{\theta_p}(\underline{X})}, \ldots, \frac{f_{\theta_{p-1}}(\underline{X})}{f_{\theta_p}(\underline{X})}\right] = [\Lambda_1(\underline{X}), \ldots, \Lambda_{p-1}(\underline{X})]$$

is finite for all $\underline{X}$. Then this vector is sufficient for θ. An equivalent way to express this vector is as the sequence $\{\Lambda_\theta(\underline{X})\}_{\theta\in\Theta} = \Lambda_1(\underline{X}), \ldots, \Lambda_{p-1}(\underline{X})$, and this is called the *likelihood trajectory over* θ.

Proof

Define the $p-1$ element selector vector $\underline{u}(\theta) = \underline{e}_k$ when $\theta = \theta_k$, $k = 1, \ldots, p-1$ (recall that $\underline{e}_k = [0, \ldots, 0, 1, 0, \ldots 0]^T$ is the k-th column of the $(p-1) \times (p-1)$ identity matrix). Now for any $\theta \in \Theta$ we can represent the j.p.d.f. as

$$f_\theta(\underline{x}) = \underbrace{\underline{T}^T \underline{u}(\theta)}_{g(\underline{T},\theta)} \underbrace{f_{\theta_p}(\underline{x})}_{h(\underline{x})},$$

which establishes sufficiency by the FF. $\diamond$

Example 5 *Likelihood ratio trajectory*

When Θ is a set of scalar parameters θ the likelihood ratio trajectory over Θ

$$\Lambda(\underline{X}) = \left\{\frac{f_\theta(\underline{X})}{f_{\theta_0}(\underline{X})}\right\}_{\theta\in\Theta}, \tag{20}$$

is sufficient for θ. Here θ_0 is an arbitrary reference point in Θ for which the trajectory is finite for all $\underline{X}$. When θ is not a scalar (20) becomes a likelihood ratio surface, which is also a sufficient statistic.

3.5.3 MINIMAL SUFFICIENCY

What is the maximum possible amount of reduction one can apply to the data sample without losing information concerning how the model depends on $\underline{\theta}$? The answer to this question lies in the notion of a minimal sufficient statistic. Such statistics cannot be reduced any further without loss in information. In other words, any other sufficient statistic can be reduced down to a minimal sufficient statistic without information loss. Since reduction of a statistic is accomplished by applying a functional transformation we have the formal definition.

Definition: T_{min} is a minimal sufficient statistic if it can be obtained from any other sufficient statistic T by applying a functional transformation to T. Equivalently, if T is any sufficient statistic there exists a function q such that $T_{min} = q(T)$.

Note that minimal sufficient statistics are not unique: if $T_{\min}$ is minimal sufficient $h(T_{min})$ is also minimally sufficient for h any invertible function. Minimal sufficient statistics can be found in a variety of ways [26, 3, 22]. One way is to find a *complete sufficient statistic*; under broad conditions this statistic will also be minimal [22]. A sufficient statistic T is complete if

$$E_{\underline{\theta}}[g(T)] = 0, \quad \text{for all } \underline{\theta} \in \Theta$$

implies that the function g is identically zero, i.e., $g(t) = 0$ for all values of t. However, in some cases there are minimal sufficient statistics that are not complete so this is not a failsafe procedure. Another way to find a minimal sufficient statistic is through reduction of the data to the likelihood ratio surface.

As in Example 5, assume that there exists a reference point $\underline{\theta}_o \in \Theta$ such that the following likelihood-ratio function is finite for all $\underline{x} \in \mathcal{X}$ and all $\theta \in \Theta$

$$\Lambda_{\underline{\theta}}(\underline{x}) = \frac{f_{\underline{\theta}}(\underline{x})}{f_{\underline{\theta}_o}(\underline{x})}.$$

For given $\underline{x}$ let $\Lambda(\underline{x})$ denote the set of likelihood ratios (a likelihood ratio trajectory or surface)

$$\Lambda(\underline{x}) = \{\Lambda_{\underline{\theta}}(\underline{x})\}_{\underline{\theta}\in\Theta}.$$

Definition 1 *We say that a (θ-independent) function of $\underline{x}$, denoted $\tau = \tau(\underline{x})$, indexes the likelihood ratios Λ when both*

1. $\Lambda(\underline{x}) = \Lambda(\tau)$, *i.e., Λ only depends on $\underline{x}$ through $\tau = \tau(\underline{x})$.*
2. $\Lambda(\tau) = \Lambda(\tau')$ *implies $\tau = \tau'$, i.e., the mapping $\tau \to \Lambda(\tau)$ is invertible.*

Condition 1 is an equivalent way of stating that $\tau(\underline{X})$ is a sufficient statistic for $\underline{\theta}$.

Theorem:If $\tau = \tau(\underline{x})$ indexes the likelihood ratios $\Lambda(\underline{x})$ then $T_{min} = \tau(\underline{X})$ is minimally sufficient for $\underline{\theta}$.

Proof:

We prove this only for the case that $\underline{X}$ is a continuous r.v. First, condition 1 in Definition 1 implies that $\tau = \tau(\underline{X})$ is a sufficient statistic. To see this use FF and the definition of the likelihood ratios to see that $\Lambda(\underline{x}) = \Lambda(\tau)$ implies: $f_{\underline{\theta}}(\underline{X}) = \Lambda_{\underline{\theta}}(\tau) f_{\underline{\theta}_o}(\underline{X}) = g(\tau; \underline{\theta}) h(\underline{x})$. Second, let T be any sufficient statistic. Then, again by FF, $f_{\underline{\theta}}(\underline{x}) = g(T, \underline{\theta})\, h(\underline{x})$ and thus

$$\Lambda(\tau) = \left\{ \frac{f_{\underline{\theta}}(\underline{X})}{f_{\underline{\theta}_o}(\underline{X})} \right\}_{\underline{\theta}\in\Theta} = \left\{ \frac{g(T, \underline{\theta})}{g(T, \underline{\theta}_o)} \right\}_{\underline{\theta}\in\Theta}.$$

so we conclude that $\Lambda(\tau)$ is a function of T. But by condition 2 in Definition 1 the mapping $\tau \to \Lambda(\tau)$ is invertible and thus τ is itself a function of T. $\diamond$

Another important concept in practical applications is that of finite dimensionality of a sufficient statistic.

Definition: a sufficient statistic $T(\underline{X})$ is said to be **finite dimensional** if its dimension is not a function of the number of data samples n.

Frequently, but not always (see Cauchy example below), minimal sufficient statistics are finite dimensional.

Example 6 *Minimal sufficient statistic for mean of Gaussian density.*

Assume $X \sim \mathcal{N}(\mu, \sigma^2)$ where σ^2 is known. Find a minimal sufficient statistic for $\theta = \mu$ given the iid sample $\underline{X} = [X_1, \ldots, X_n]^T$.

Solution: the j.p.d.f. is

$$
\begin{aligned}
f_\theta(\underline{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)} \\
&= \underbrace{e^{-\frac{n\mu^2}{2\sigma^2}} e^{\mu/\sigma^2 \overbrace{\sum_{i=1}^n x_i}^{T(\underline{x})}}}_{g(\underline{T},\theta)} \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-1/(2\sigma^2)\sum_{i=1}^n x_i^2}}_{h(\underline{x})}
\end{aligned}
$$

Thus by FF

$$
T = \sum_{i=1}^n X_i
$$

is a sufficient statistic for μ. Furthermore, as $q(T) = n^{-1}T$ is a 1-1 function of T

$$
\underline{S} = \overline{X}
$$

is an equivalent sufficient statistic.

Next we show that the sample mean is in fact minimal sufficient by showing that it indexes the likelihood ratio trajectory $\Lambda(\underline{x}) = \{\Lambda_\theta(\underline{x})\}_{\theta\in\Theta}$, with $\theta = \mu$, $\Theta = \mathbb{R}$. Select the reference point $\theta_o = \mu_o = 0$ to obtain:

$$
\Lambda_\mu(\underline{x}) = \frac{f_\mu(\underline{x})}{f_0(\underline{x})} = \exp\left(\mu/\sigma^2 \sum_{i=1}^n x_i - \tfrac{1}{2} n\mu^2/\sigma^2\right).
$$

Identifying $\tau = \sum_{i=1}^n x_i$, condition 1 in Definition 1 is obviously satisfied since $\Lambda_\mu(\underline{x}) = \Lambda_\mu(\sum x_i)$ (we already knew this since we showed that $\sum_{i=1}^n X_i$ was a sufficient statistic). Condition 2 in Definition 1 follows since $\Lambda_\mu(\sum x_i)$ is an invertible function of $\sum x_i$ for any non-zero value of μ (summation limits omitted for clarity). Therefore the sample mean indexes the trajectories, and is minimal sufficient.

Example 7 *Minimal sufficient statistics for mean and variance of Gaussian density.*

Assume $X \sim \mathcal{N}(\mu, \sigma^2)$ where both μ and σ^2 are unknown. Find a minimal sufficient statistic for $\underline{\theta} = [\mu, \sigma^2]^T$ given the iid sample $\underline{X} = [X_1, \ldots, X_n]^T$.

Solution:

$$
\begin{aligned}
f_{\underline{\theta}}(\underline{x}) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right)}
\end{aligned}
$$

$$= \underbrace{\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{n\mu^2}{2\sigma^2}} \; e^{[\mu/\sigma^2,\; -1/(2\sigma^2)]\overbrace{\left[\sum_{i=1}^n x_i,\; \sum_{i=1}^n x_i^2\right]^T}^{\underline{T}(\underline{x})}}}_{g(\underline{T},\underline{\theta})} \underbrace{1}_{h(\underline{x})}$$

Thus

$$\underline{T} = \left[\underbrace{\sum_{i=1}^n X_i}_{T_1},\; \underbrace{\sum_{i=1}^n X_i^2}_{T_2}\right]$$

is a (jointly) sufficient statistic for μ, σ^2. Furthermore, as $q(\underline{T}) = [n^{-1}T_1, (n-1)^{-1}(T_2 - T_1^2)]$ is a 1-1 function of $\underline{T}$ ($\underline{T} = [T_1, T_2]^T$)

$$\underline{S} = \left[\overline{X},\; \mathsf{s}^2\right]$$

is an equivalent sufficient statistic.

Similarly to Example 6, we can show minimal sufficiency of this statistic by showing that it indexes the likelihood ratio surface $\{\Lambda_{\underline{\theta}}(\underline{X})\}_{\underline{\theta}\in\Theta}$, with $\theta = [\mu, \sigma^2]$, $\Theta = \mathbb{R} \times \mathbb{R}^+$. Arbitrarily select the reference point $\underline{\theta}_o = [\mu_o, \sigma_o^2] = [0, 1]$ to obtain:

$$\Lambda_{\underline{\theta}}(\underline{x}) = \frac{f_\theta(\underline{x})}{f_{\theta_o}(\underline{x})} = \left(\frac{\sigma_o}{\sigma}\right)^n e^{-n\mu^2/(2\sigma^2)} \; e^{[\mu/\sigma^2,\; -\delta/2]\left[\sum_{i=1}^n x_i,\; \sum_{i=1}^n x_i^2\right]^T},$$

where $\delta = \frac{\sigma_o^2 - \sigma^2}{\sigma^2\sigma_o^2}$. Identifying $\underline{\tau} = \left[\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2\right]$, again condition 1 in Definition 1 is obviously satisfied. Condition 2 in Definition 1 requires a bit more work. While $\Lambda_{\underline{\theta}}(\underline{\tau})$ is no longer an invertible function of τ for for any *single value* of $\underline{\theta} = [\mu, \sigma^2]$, we can find two values $\underline{\theta} \in \{\underline{\theta}_1, \underline{\theta}_2\}$ in Θ for which the vector function $[\Lambda_{\underline{\theta}_1}(\underline{\tau}), \Lambda_{\underline{\theta}_2}(\underline{\tau})]$ of $\underline{\tau}$ is invertible in $\underline{\tau}$. Since this vector is specified by $\Lambda(\underline{x})$, this will imply that $\underline{\tau}$ indexes the likelihood ratios.

To construct this invertible relation denote by $\underline{\lambda} = [\lambda_1, \lambda_2]^T$ an observed pair of samples $[\Lambda_{\underline{\theta}_1}(\underline{\tau}), \Lambda_{\underline{\theta}_2}(\underline{\tau})]^T$ of the surface $\Lambda(\underline{x})$. Now consider the problem of determining $\underline{\tau}$ from the equation $\underline{\lambda} = [\Lambda_{\underline{\theta}_1}(\underline{\tau}), \Lambda_{\underline{\theta}_2}(\underline{\tau})]^T$. Taking the log of both sides and rearranging some terms, we see that this is equivalent to a 2×2 linear system of equations of the form $\underline{\lambda}' = \mathbf{A}\underline{\tau}$, where $\mathbf{A}$ is a matrix involving $\underline{\theta}_o, \underline{\theta}_1, \underline{\theta}_2$ and $\underline{\lambda}'$ is a linear function of $\ln \underline{\lambda}$. You can verify that with the selection of $\underline{\theta}_o = [0, 1]$, $\underline{\theta}_1 = [1, 1]$, $\underline{\theta}_2 = [0, 1/2]$ we obtain $\delta = 0$ or 1 for $\underline{\theta} = \underline{\theta}_1$ or $\underline{\theta}_2$, respectively, and $\mathbf{A} = \text{diag}(1, -1/2)$, an invertible matrix. We therefore conclude that the vector [sample mean, sample variance] indexes the trajectories, and this vector is therefore minimal sufficient.

Example 8 *Sufficient statistic for the location of a Cauchy distribution*

Assume that $X_i \sim f(x;\theta) = \frac{1}{\pi} \frac{1}{1+(x-\theta)^2}$ and, as usual, $\underline{X} = [X_1, \ldots, X_n]^T$ is an i.i.d. sample. Then

$$f(\underline{x};\theta) = \prod_{i=1}^n \frac{1}{\pi} \frac{1}{1 + (x_i - \theta)^2} = \frac{1}{\pi^n} \frac{1}{\prod_{i=1}^n (1 + (x_i - \theta)^2)}.$$

Here we encounter a difficulty: the denominator is a $2n$-degree polynomial in θ whose roots depend on all cross products $x_{i_1} \ldots x_{i_p}$, $p = 1, 2, \ldots, n$, of $x_i's$. Thus no sufficient statistic exists having dimension less than that (n) of the entire sample. Therefore, the minimal sufficient statistic is the ordered statistic $[X_{(1)}, \ldots, X_{(n)}]^T$ and no finite dimensional sufficient statistic exists.

3.5.4 EXPONENTIAL FAMILY OF DISTRIBUTIONS

Let $\underline{\theta} = [\theta_1, \ldots, \theta_p]^T$ take values in some parameter space Θ. The distribution $f_{\underline{\theta}}$ of a r.v. X is a member of the p-parameter exponential family if for all $\underline{\theta} \in \Theta$

$$f_{\underline{\theta}}(x) = a(\underline{\theta}) b(x) e^{-\underline{c}^T(\underline{\theta})\underline{t}(x)}, \quad -\infty < x < \infty \tag{21}$$

for some scalar functions a, b and some p-element vector functions $\underline{c}, \underline{t}$. Note that for any $f_{\underline{\theta}}$ in the exponential family its support set $\{x : f_{\underline{\theta}}(x) > 0\}$ does not depend on $\underline{\theta}$. Note that, according to our definition, for $f_{\underline{\theta}}$ to be a member of the p-parameter exponential family the dimension of the vectors $\underline{c}(\underline{\theta})$ and $\underline{t}(x)$ must be exactly p. This is to guarantee that the sufficient statistic has the same dimension as the parameter vector $\underline{\theta}$. While our definition is the most standard [23, 26, 3], some other books, e.g., [31], allow the dimension of the sufficient statistic to be different from p. However, by allowing this we lose some important properties of exponential families [3].

The parameterization of an exponential family of distributions is not unique. In other words, the exponential family is invariant to changes in parameterization. For example, let f_θ, $\theta > 0$, be an the exponential family of densities with θ a positive scalar. If one defines $\alpha = 1/\theta$ and $g_\alpha = f_{1/\theta}$ then g_α, $\alpha > 0$, is also in the exponential family, but possibly with a different definition of the functions $a(\cdot), b(\cdot), \underline{c}(\cdot)$ and $\underline{t}(\cdot)$. More generally, if $f_{\underline{\theta}}(\underline{x})$ is a member of the p-dimensional exponential family then transformation of the parameters by any invertible function of $\underline{\theta}$ preserves membership in the exponential family.

There exists a special parameterization of distributions in the exponential family, called the natural parameterization, that has important advantages in terms of ease of estimation of these parameters.

Definition: Let the random variable X have distribution $f_{\underline{\theta}}(x)$ and assume that $f_{\underline{\theta}}$ is in the exponential family, i.e., it can be expressed in the form (21). $f_{\underline{\theta}}$ is said to have a *natural parameterization* if for all $\underline{\theta} \in \Theta$: $E_{\underline{\theta}}[\underline{t}(X)] = \underline{\theta}$.

In particular, as we will see in the next chapter, this means that having a natural parameterization makes the statistic $\underline{T} = \underline{t}(X)$ an unbiased estimator of $\underline{\theta}$.

Examples of distributions in the exponential family include: Gaussian with unknown mean or variance, Poisson with unknown mean, exponential with unknown mean, gamma, Bernoulli with unknown success probability, binomial with unknown success probability, multinomial with unknown cell probabilities.

Distributions which *are not* from the exponential family include: Cauchy with unknown median, uniform with unknown support, Fisher-F with unknown degrees of freedom.

When the statistical model is in the exponential family, sufficient statistics for the model parameters have a particularly simple form:

$$\begin{aligned} f_{\underline{\theta}}(\underline{x}) &= \prod_{i=1}^n a(\underline{\theta}) b(x_i) e^{-\underline{c}^T(\underline{\theta})\underline{t}(x_i)} \\ &= \underbrace{a^n(\underline{\theta})\, e^{-\underline{c}^T(\underline{\theta}) \overbrace{\sum_{i=1}^n \underline{t}(x_i)}^{\underline{T}}}}_{g(\underline{T},\underline{\theta})} \underbrace{\prod_{i=1}^n b(x_i)}_{h(\underline{x})} \end{aligned}$$