Why Does the Factorization Theorem Work?

For reference, here are the definition of sufficient statistic and the statement of the Factorization Theorem:

<u>Definition.</u> A statistic *T* is sufficient if the conditional distribution of the data, given *T*, does not depend upon the parameters: that is, $f(x_1,...,x_n \mid T)$ does not depend upon the parameters.

The Factorization Theorem: If *T* is a statistic and θ is a parameter, then *T* is sufficient for θ if the joint density $f(x_1,...,x_n;\theta)$ of the data can be factored into a function $g(T,\theta)$ of *T* and θ alone and a function $h(x_1,...,x_n)$ of the data but not of θ.

Remember the mathematical intuition behind the definition of sufficiency: By itself alone, *T* provides a sufficient basis for an inference about θ if the complete dataset $x_1,...,x_n$ no longer depends on θ after you know *T*. That says there is no information about θ left in the data after *T* is known. The purpose of the Factorization Theorem is to provide an easy way to find a statistic that is sufficient. It may not be immediately apparent why the Factorization Theorem works. In this note, I provide the justification.

The short version of why the Factorization Theorem works is that the conditional density function $f(x_1,...,x_n \mid T;\theta)$ is a ratio $\dfrac{f(x_1,...,x_n,T;\theta)}{f(T;\theta)}$ that appears to depend upon θ; but the Factorization Theorem permits factoring the numerator of the ratio as

$$f(x_1,...,x_n \mid T;\theta) = \frac{f(x_1,...,x_n,T;\theta)}{f(T;\theta)} = \frac{g(T,\theta)h(x_1,...,x_n)}{f(T;\theta)}$$

And it turns out that the numerator function $g(T,\theta)$ and the denominator function $f(T;\theta)$ are proportional to each other, so they cancel out of the ratio, leaving only parts that do not depend upon θ.

But the details require some work. More than some. They are a small part of the argument, but they are very technical in nature and require disproportionate space. In order to see why the Factorization Theorem is true, suppose that $X_1, X_2,..., X_n$ are random variables with joint density function $f(x_1, x_2,...,x_n;\theta)$ that depends upon the parameter θ (which could be a vector) and that $T = T(x_1, x_2,...,x_n)$ is a statistic (which could also possibly be a vector). Suppose that the joint density of the data factors in the manner hypothesized by the Factorization Theorem:

$$f(x_1,...,x_n;\theta) = g(T(x_1,x_2,...,x_n),\theta)\cdot h(x_1,x_2,...,x_n) \quad \text{[Eqn 1]}$$

in which $g(T(x_1,x_2,...,x_n),\theta)$ is a function of *T* and θ alone [i.e., *g* depends upon the data $(x_1,...,x_n)$ only through *T*], and $h(x_1,...,x_n)$ is a function of the data but not of θ. Using the factorization, I will demonstrate that *T* is sufficient for θ. To do that, I must show that the conditional density of the data given *T* does not depend upon θ. I do that now.

By definition, the conditional density of the data given *T* is:

$$f(x_1,...,x_n \mid T(x_1,...,x_n);\theta) = \frac{f(x_1,...,x_n,T(x_1,...,x_n);\theta)}{m(T(x_1,...,x_n);\theta)} \quad \text{[Eqn 2]}$$

where $f(x_1,...,x_n,T(x_1,...,x_n);\theta)$ is the joint density of the data and *T*, and $m(T(x_1,...,x_n);\theta)$ is the (marginal) density of *T*. The key step will be to rewrite Eqn 2 so that $g(t,\theta)$ appears in both numerator and denominator of the conditional density and therefore cancels out, leaving only pieces that do not depend upon θ.

Note that the joint density of the data and *T*, namely $f(x_1,...,x_n,T(x_1,...,x_n);\theta)$, is *formally* a function of *n*+1 random variables, whereas the joint density of the data, namely $f(x_1,...,x_n;\theta)$, is a function of only *n* random variables; so these two densities are not immediately even comparable. However, the $n+1^{st}$ argument of the numerator, namely $T(x_1,...,x_n)$, is a function of the other *n* arguments $x_1,...,x_n$; so in reality both densities are functions of $x_1,...,x_n$. Part of the argument will be to show that, given $x_1,...,x_n$, the two densities are equal: $f(x_1,...,x_n,T(x_1,...,x_n);\theta) = f(x_1,...,x_n;\theta)$, so that the factorization may be applied to the numerator. It is easier to show this for discrete than for continuous distributions.

Let me consider two cases: First, the data distribution is discrete; second, the data distribution is continuous.

Case 1 (discrete data). In the discrete case, the densities are probabilities. That makes the details easier mathematically. Let me first find the marginal density of *T*, which is the denominator of Eqn 2: For any given *t* (possibly a vector), we have $m(t;\theta) = P(T=t;\theta) =$

(*) $\displaystyle\sum\sum\cdots\sum_{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}} P(X_1=z_1,...,X_n=z_n;\theta) = \sum\sum\cdots\sum_{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}} f(z_1,...,z_n;\theta) =$

(**) $\displaystyle\sum\sum\cdots\sum_{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}} g(T(z_1,z_2,...,z_n),\theta)\cdot h(z_1,z_2,...,z_n) =$

$$\sum\sum\cdots\sum_{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}} g(t,\theta)\cdot h(z_1,z_2,...,z_n) =$$

(***) $\displaystyle g(t,\theta)\cdot \sum\sum\cdots\sum_{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}} h(z_1,z_2,...,z_n)$. Here is the explanation for the starred steps:

(*) The probability that *T* = *t* is the sum of the probabilities of all data $(z_1,z_2,...,z_n)$ for which $T(z_1,z_2,...,z_n)=t$.

(**) This step applies the factorization.

(***) $g(t,\theta)$ may be pulled outside the summation because it has a constant value for all $(z_1,z_2,...,z_n)$ involved in the summation.

Now for the numerator of Eqn 2: Given data values $(x_1, x_2,..., x_n)$ and a t such that $T(x_1, x_2,..., x_n) = t$, the numerator of Eqn 2 is $f(x_1,..., x_n, T(x_1,..., x_n);\theta) =$ $P(X_1 = x_1,..., X_n = x_n, T(x_1,..., x_n) = t;\theta) =$ $P(X_1 = x_1,..., X_n = x_n;\theta) =$ $f(x_1,..., x_n;\theta)$. This is true because the condition $T(x_1, x_2,..., x_n) = t$ is redundant in the joint probability of $(x_1, x_2,..., x_n)$ and $T = t$ since we are considering only the specific set of data values $(x_1, x_2,..., x_n)$ that satisfy $T(x_1, x_2,..., x_n) = t$. Indeed, given $x_1,..., x_n$ such that $T(x_1, x_2,..., x_n) = t$, the event $\{X_1 = x_1,..., X_n = x_n, T(x_1,..., x_n) = t\}$ is the same as the event $\{X_1 = x_1,..., X_n = x_n\}$ and so these two events have the same probability. Furthermore, using the factorization, I then have $f(x_1,..., x_n;\theta) =$ $g(t,\theta)\cdot h(x_1, x_2,..., x_n)$.

I have now found expressions for the numerator and denominator of Eqn 2 in which $g(t,\theta)$ is a common factor. Substitute them into Eqn 2 for given data values $(x_1, x_2,..., x_n)$ and a t such that $T(x_1, x_2,..., x_n) = t$, we have:

$$f(x_1,..., x_n \mid T(x_1,..., x_n);\theta) = \frac{g(t,\theta)\cdot h(x_1, x_2,..., x_n)}{g(t,\theta)\cdot \underset{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}}{\sum\sum\cdots\sum} h(z_1, z_2,..., z_n)} =$$

$$\frac{h(x_1, x_2,..., x_n)}{\underset{\{(z_1,...,z_n);T(z_1,...,z_n)=t\}}{\sum\sum\cdots\sum} h(z_1, z_2,..., z_n)}$$

. By the factorization hypothesis, h does not depend on θ. Thus, neither does the preceding ratio. Since this is true for any such data values, it is true in general. This completes the argument for the discrete case.

Case 2 (continuous data). In the continuous case, the densities are not probabilities. This complicates the details. The argument requires a transformation of variables in multivariable calculus. First, let me again find the marginal density of *T*, which is the denominator of Eqn 2: The marginal density of T may be found by integrating out a set of helper functions $s_2,..., s_n$ in a transformation from the joint space of $(x_1, x_2,..., x_n)$ to a joint space $(t, s_2,..., s_n)$. Suppose I have defined a one-to-one differentiable transformation $\boldsymbol{t = t(x_1, x_2,..., x_n)}$, $s_2 = s_2(x_1, x_2,..., x_n)$, $s_3 = s_3(x_1, x_2,..., x_n)$, … , $s_n = s_n(x_1, x_2,..., x_n)$, not involving θ, with differentiable inverse transformation $x_1 = x_1(t, s_2,..., s_n)$, $x_2 = x_2(t, s_2,..., s_n)$, $x_3 = x_3(t, s_2,..., s_n)$, … , $x_n = x_n(t, s_2,..., s_n)$,which has Jacobian $|J|$. The Jacobian of the transformation is the absolute value of the determinant of the matrix of all partial derivatives

$$\begin{vmatrix} \partial x_1/\partial t & \partial x_1/\partial s_2 & \cdots & \partial x_1/\partial s_n \\ \partial x_2/\partial t & \partial x_2/\partial s_2 & \cdots & \partial x_2/\partial s_n \\ \vdots & \vdots & \ddots & \vdots \\ \partial x_n/\partial t & \partial x_n/\partial s_2 & \cdots & \partial x_n/\partial s_n \end{vmatrix}$$

, which accounts for the stretching of space done by the transformation and may be a function of $(t, s_2,..., s_n)$, rather than just be a constant.

Note that the Jacobian also does not depend on θ since the transformations do not, and therefore their derivatives do not. In setting up the helper transformations, you should choose $t = t(x_1, x_2,..., x_n)$ to be the putatively sufficient statistic (e.g., $t = x_1 + x_2 + ... + x_n$) – unless the sufficient statistic is a vector – and you should choose the helper functions to be as easy as possible, consistent with being one-to-one, invertible and differentiable (e.g., $s_2 = x_2, s_3 = x_3,..., s_n = x_n$ are easy, one-to-one, invertible and differentiable). Remember that I am trying to find the marginal density of *T* by integrating out the helper variables from the joint density of $(t, s_2,..., s_n)$. So I want this transformation to be as easy as possible.

Now, the joint density of $(t, s_2,..., s_n)$ is found by substituting the inverse transformations into *f* and multiplying by |*J*|, and integrating on $(s_2,..., s_n)$ for each given value of *T* = *t* – except that I will first factor *f* by the Factorization Theorem and then proceed. So let $T(x_1, x_2,..., x_n) = t$ be given. Then the joint density of $(x_1, x_2,..., x_n)$ is

$$f(x_1,..., x_n;\theta) = g(T(x_1, x_2,..., x_n),\theta) \cdot h(x_1, x_2,..., x_n) = g(t,\theta) \cdot h(x_1, x_2,..., x_n)$$

So the joint density of $(t, s_2,..., s_n)$ is, upon substitution,

$$j(t, s_2,..., s_n;\theta) = g(t,\theta) \cdot h(x_1(t, s_2,..., s_n), x_2(t, s_2,..., s_n),..., x_n(t, s_2,..., s_n)) \,|\, J \,|$$

Therefore, the marginal density of *T* is $m(t;\theta) = \int\int\cdots\int j(t, s_2,..., s_n;\theta)\, ds_2 ds_3 \cdots ds_n =$

$$\int\int\cdots\int g(t,\theta) \cdot h(x_1(t, s_2,..., s_n), x_2(t, s_2,..., s_n),..., x_n(t, s_2,..., s_n)) \,|\, J \,|\, ds_2 ds_3 \cdots ds_n$$

It is logically important to note that $T(x_1, x_2,..., x_n) = t$ is one given value; so although the $(x_1, x_2,..., x_n)$'s in $g(T(x_1, x_2,..., x_n),\theta)$ are functions of $(s_2,..., s_n)$ as well as of *t*, when substituted into $T(x_1, x_2,..., x_n)$, they result in one value $T(x_1, x_2,..., x_n) = t$. Thus, in the (*n*-1)-fold integral immediately above, $g(t,\theta)$ is a constant over the range of integration on $(s_2,..., s_n)$ and so may be brought outside the integrals:

$$m(t;\theta) = g(t,\theta)\int\int\cdots\int h(x_1(t, s_2,..., s_n), x_2(t, s_2,..., s_n),..., x_n(t, s_2,..., s_n)) \,|\, J \,|\, ds_2 ds_3 \cdots ds_n$$

Now for the numerator of Eqn 2: The numerator of Eqn 2 is the joint density function of the *n*+1 random variables $X_1, X_2, \cdots, X_n, T(X_1, X_2, \cdots, X_n)$. Recall that the joint density factors into the product of the conditional density times the marginal density – in general, $g(u, v) = g(u)g(v \mid u)$ where *u* and/or *v* may be vectors – so that the numerator of Eqn 2 factors into $f(x_1,..., x_n, t;\theta) =$ $f(x_1,..., x_n;\theta) f(t \mid x_1,..., x_n;\theta)$. Now consider the conditional density $f(t \mid x_1,..., x_n;\theta)$. Given $x_1,..., x_n$, the value of the random variable *T* must be $T(x_1,..., x_n)$ - i.e., a constant. That is, the conditional distribution of *T* given $x_1,..., x_n$ puts all of its probability on the constant $t = T(x_1,..., x_n)$. Thus, the conditional density $f(t \mid x_1,..., x_n;\theta)$ can be nonzero for only one value, namely when $t = T(x_1,..., x_n)$.[1] Thus, $f(t \mid x_1,..., x_n; \theta) = 1$ if $t = T(x_1,..., x_n)$ and $f(t \mid x_1,..., x_n; \theta) = 0$ otherwise.[2] Therefore, $f(x_1,..., x_n, t; \theta) = f(x_1,..., x_n; \theta) f(t \mid x_1,..., x_n; \theta) = f(x_1,..., x_n; \theta) \cdot 1 = f(x_1,..., x_n; \theta)$ if $t = T(x_1,..., x_n)$; otherwise $f(x_1,..., x_n, t; \theta) = 0$ if $t \neq T(x_1,..., x_n)$. What this means is that for all $x_1,..., x_n$ for which $T(x_1,..., x_n) = t$, the numerator of Eqn 2 = $f(x_1,..., x_n; \theta)$, which is the joint density of the data. The premise of the Factorization Theorem is that the joint density of the data factors into $f(x_1,...,x_n; \theta) = g(t, \theta) \cdot h(x_1, x_2,..., x_n)$.

Putting it all together, using the numerator and denominator, for given $(x_1, x_2,..., x_n)$ such that $T(x_1, x_2,..., x_n) = t$:

$$f(x_1,...,x_n \mid T = t; \theta) = \frac{f(x_1,...,x_n, t; \theta)}{m(t;\theta)} =$$
$$\frac{g(t,\theta) \cdot h(x_1, x_2,..., x_n)}{g(t,\theta) \int\int \cdots \int h(x_1(t, s_2,..., s_n), x_2(t, s_2,..., s_n),..., x_n(t, s_2,..., s_n)) \mid J \mid ds_2 ds_3 \cdots ds_n} =$$
$$\frac{h(x_1, x_2,..., x_n)}{\int\int \cdots \int h(x_1(t, s_2,..., s_n), x_2(t, s_2,..., s_n),..., x_n(t, s_2,..., s_n)) \mid J \mid ds_2 ds_3 \cdots ds_n} .$$

Neither numerator nor denominator in the expression immediately above depends upon θ . Thus the conditional density $f(x_1,...,x_n \mid T = t; \theta) = f(x_1,...,x_n \mid T = t)$ also does not depend upon θ . This completes the argument for the continuous case. Thus *T* is sufficient by the definition.

The converse. The Factorization Theorem also provides a characterization of sufficiency, for its converse is also true. That is,

Suppose that *T* is a statistic and θ is a parameter. *T* is sufficient for θ **if and only if** the joint density $f(x_1,...,x_n; \theta)$ of the data can be factored into a function $g(T, \theta)$ of *T* and θ alone and a function $h(x_1,...,x_n)$ of the data but not of θ .

To see why the converse is true, suppose that *T* is sufficient for θ . Let $(x_1,...,x_n)$ and *t* be given such that $T(x_1,...,x_n) = t$. By definition of sufficiency, the conditional density $f(x_1,...,x_n \mid T = t; \theta) = f(x_1,...,x_n \mid T = t)$ does not depend upon θ . But $f(x_1,...,x_n \mid T = t)$ =

---

[1] To be sure, for a different specific $x_1,..., x_n$, *T* could have a different value (but still only one value for that specific $x_1,..., x_n$). But conditioned on $x_1,..., x_n$, *T* can have only one value. Once $x_1,..., x_n$ are given, *T* is fixed.

[2] It may seem odd that you can get a discrete conditional density $f(t \mid x_1,..., x_n; \theta)$ when the data distribution is continuous, but it is true! Here is a simple example that may make this more plausible: Suppose *X* is standard normal. Define *Y* = 2*X*. Given *X* = 3, the value of *Y* must be 6. So the conditional density of *Y* given *X* = 3 equals 1 if *Y* = 6 and is zero otherwise.

$\dfrac{f(x_1,\ldots,x_n,T(x_1,\ldots,x_n);\theta)}{m(t;\theta)} = \dfrac{f(x_1,\ldots,x_n;\theta)}{m(t;\theta)}$ since $T(x_1,\ldots,x_n)$ is superfluous in the joint density as long as $T(x_1,\ldots,x_n) = t$.[3] Multiplying both sides by $m(t;\theta)$ yields $f(x_1,\ldots,x_n;\theta) =$ $m(t;\theta)\cdot f(x_1,\ldots,x_n \mid T = t)$. Now set $g(t;\theta) = m(t;\theta)$ and $h(x_1,\ldots,x_n) = f(x_1,\ldots,x_n \mid T = t)$ to get the factorization.

[3] See the argument of the Factorization Theorem for rigorous justification of this statement in the discrete and continuous cases.