Remarks on Big Data and Sufficient Statistics

The way to deal with Big Data is to turn it into Small Data.

There are many ways to do this, and a few ways to do it well.

The spreadsheet provides a conceptual model for most types of data.[1] In a spreadsheet, the data are arranged in columns (variables, fields) and rows (observations, records).

A dataset can be Big because it has many columns, many rows, or both.

A basic way to make a Big dataset Small is to reduce the number of columns, reduce the number of rows, or both. A simple way to do this is to delete columns and/or rows. Arbitrary deletion, however, is probably not a good idea because deletion of data also deletes information. You should delete only unimportant columns and/or rows – those that contain the least information. A more sophisticated way is to combine existing columns or rows to make fewer columns or rows in such a manner as to preserve as much of the information in the original data as possible.

Data vs Information. In this discussion, it is useful to distinguish *data* from *information*. Data are the tangible raw results of observation or experimentation, often in the form of numbers or tallies. Data have not yet been analyzed with the intent of learning what they imply. Information can be thought of as a distilled essence of data, boiled down into a form that human beings can readily understand. Information consists of analysis and interpretation of data. A spreadsheet of sales figures is data. The statement, "There is a 90% chance that sales next year will be between $19,000,000 and $21,000,000", based on the spreadsheet, is information. That information is not explicit in the data. It is implicit and hidden. The data need to be analyzed and interpreted to make the information explicit.

The purpose of turning Big Data into Small Data is to start boiling down the data in order to concentrate the implicit information hidden in it. This prepares the data to become information in the hands of a skilled Data Analyst. You can think of data as encrypted information. The job of the Data Analyst is to decrypt the information in the data.

Reducing the column count. Principal Components Analysis reduces the number of columns in a spreadsheet by combining them. So does Factor Analysis. Multiple regression reduces the dimensionality of the predictor variables (columns) to one by combining the predictors into a single function (a single column).

Reducing the row count. Summary statistics can reduce the number of rows and/or columns, but typically reduce the number of rows. Example: Calculating the mean and variance reduces the number of rows by combining them into one number. Example: Calculating the correlation coefficient reduces the number of columns and rows. Cluster Analysis effectively reduces the number of rows by grouping them.

The real problem with Big Data is that its information is diffuse – spread over many variables and/or observations. To make the information in Big Data useful, it must be concentrated so that we can more easily understand it.

[1] In principle, even a network can be displayed as spreadsheet: The rows and columns can both denote nodes, with cell contents as 0-1 indicators of the existence of row node to column node linkage, or as numbers that measure the strengths of those linkages.

Sufficient Statistics: Generalities

The theory of information sufficiency within statistics assesses how well summary statistics capture the information in data. Example: If the data are randomly sampled from a normal distribution, then it is true that the sample mean and sample variance capture *all* of the information in the data. The consequence is that for making inferences about a normal universe, you can throw away the original data after you have computed the sample mean and sample variance. Potentially millions of records can be reduced to two numbers. The sample mean and sample variance are said to be *sufficient* statistics for a normal universe. As the same suggests, a sufficient statistic is "sufficient" – all you need to know – for making any inference with the data. You do not need to know anything else about the data. Therefore, a sufficient statistic captures all of the information that the data have. It is amazing that sufficiency has a precise mathematical formulation. It is not just a qualitative label based upon opinion or experience.

The notion of sufficiency is one of the most elegant aspects of classical statistics. All maximum likelihood estimators are functions of sufficient statistics. All Bayes estimators are also functions of sufficient statistics. It is hard to think of sensible statistical procedures that cannot begin by reducing the original data to sufficient statistics. Partial exceptions to this general rule are provided by nonparametric and semiparametric data problems, which are increasingly common.

Parametric data problems. In a parametric problem, the data distribution has a given mathematical form that is known except for the values of a few parameters. An example is any problem in which the data distribution is normal: the probability density function (p.d.f.) is $f(x;\mu,\sigma^2)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-\mu)^2}{2\sigma^2}}$, which is known except for the values of the population mean μ and population variance σ^2, which are the *parameters* of this distribution.

Nonparametric data problems. In a nonparametric problem, the density function is not parametric. The distribution cannot be known completely by specifying any finite set of parameters. An example is any problem in which the data distribution has a discrete density function on the nonnegative integers, and the functional form of the density is not given. All that is known about the p.d.f. is that it is an infinite set of nonnegative numbers $f(0), f(1), f(2), f(3),...$ that add up to 1, where $f(0)$ is the unknown probability of getting a 0, $f(1)$ is the unknown probability of getting a 1, etc. The set of values $\{f(i); i=0,1,2,...\}$ could be any numbers between 0 and 1 and summing to 1. In some sense, the parameters for the latter problem are the countably infinite set $\{f(i); i=0,1,2,...\}$ of probabilities themselves. The distribution *f* is not completely determined by any finite subset of $\{f(i); i=0,1,2,...\}$. All must be given.

For a nonparametric problem, not much data reduction is possible if one wants to preserve all of the original information. In the above nonparametric problem $\{f(i); i=0,1,2,...\}$, it turns out that ordering the original data $x_1,...,x_n$ into the set of order statistics $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ is the maximal information-preserving reduction that can be done. In general, the maximal information-preserving reduction is called a *minimal sufficient statistic*. Any further reduction will lose at least some information.

An example of a parametric problem on the nonnegative integers is the Poisson distribution. Its p.d.f. is $f(x;\lambda)=\frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0,1,2,\ldots$, a discrete distribution on the nonnegative integers. This is a parametric distribution because its values are all determined by specifying the value of a single parameter λ, from

which all of the probabilities $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$, $x = 0,1,2,\ldots$, can be calculated once λ is determined. It turns out that the sample mean is a minimal sufficient statistic for the Poisson distribution.

Sufficient Statistics: Technicalities

After the preceding general remarks, I would like to discuss the technical foundation for the notion of sufficiency. First, recall the definition of conditional distribution: $f(y \mid x) = \frac{f(x,y)}{f(x)}$ is the conditional density function of *Y* given *x*. Here, $f(x,y)$ is the joint p.d.f of random variables *X* and *Y*, and $f(x)$ is the marginal p.d.f. of *X*. [2] An important application of conditional distribution is the regression equation, which estimates the mean of the conditional distribution of *Y* as a function of *x* (i.e., $E(Y \mid x)$.

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

Sufficiency is most useful for parametric problems.

A few sentences about the intuition that justifies using the above as the definition of sufficiency: Ordinarily, we look at data to learn things about their distribution. This works because the data should look like their distribution – the data should be plentiful where there is high probability and sparse where there is low probability. As a consequence, the statistics of the data should be reasonably close to the corresponding parameters of the distribution. The sample mean should be close to the population mean; the sample variance should be close to the population variance. If the population mean is 6, then the sample mean should be reasonably close to 6. If the population mean is 12, then the sample mean should be reasonably close to 12. That is, the data values depend upon their distribution. If the parameters change, then the data change. If the population mean changes from 6 to 12, then the sample mean should change from about 6 to about 12. We do not know what the population mean is, but we can look to the sample data to inform us about the change in the population mean.

Now, if *T* is sufficient, then the definition of sufficiency tells us that once *T* is known, the data no longer depend upon their parameters – so the data have nothing further to tell us about their parameters, hence nothing further to help us distinguish one possible form of the distribution from another. There is nothing we can learn about the distribution that is not already captured by *T*. So we may as well throw away the data and just use *T*. So if the population mean changes from 6 to 12, and if *T* is sufficient, then the data have nothing further to teach us about that change once we know the change in *T*.

Comment. If *T* is sufficient, then it is intuitive that any one-to-one function of *T* is also sufficient. For example, $g(T) = 2T$. The intuitive reason is that *T* and $g(T)$ contain the same information, because you can always transform one into the other – for example, by dividing the example $g(T)$ by 2. This intuition is correct and can also be demonstrated mathematically. On the other hand, if T is sufficient, then

[2] In the case that X and Y are discrete, the definition of the conditional p.d.f. is just the definition of conditional probability: $f(y \mid x) = P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}$.

$g(T) = T^2$ is not necessarily sufficient because this $g(T)$ is not one-to-one. If you know $T^2 = 9$, then you do not know whether $T = +3$ or -3.

Here is a more specific example, but without mathematical proof: Suppose you have data $x_1,...,x_n$ that are a Random Sample from the $N(\mu,\sigma^2)$ distribution. Then it turns out to be true that $f(x_1,...,x_n \mid \bar{X}, S^2)$ does not depend upon μ,σ^2. This says that once you know the sample mean and variance, the data no longer have any way to tell you more about μ,σ^2 because the data are no longer sensitive to changes in μ,σ^2. Moreover, it is true that the sample mean and variance are *minimally sufficient*. This means there is no simpler function, no further reduction of the data that is sufficient. So only functions of the sample mean and sample variance need to be considered when you estimate anything about the universe.

Now I will show how to demonstrate sufficiency from the definition for a discrete example and for a continuous example.

The discrete example: Suppose that X_1 and X_2 are independent Poisson variables. Then their joint density function is $f(x_1,x_2;\lambda) = \frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}$ for $x_1, x_2 = 0, 1, 2, 3, \ldots$ I will show that $T = x_1 + x_2$ is sufficient. This will also show that the sample mean $\bar{x} = \frac{x_1 + x_2}{2}$ is sufficient, for $\bar{x}$ and T are one-to-one functions of each other and therefore contain exactly the same information. How to show sufficiency? I will show that $f(x_1,x_2 \mid T)$ does not depend upon λ. This will satisfy the definition of sufficiency.

By definition of conditional probability, $f(x_1,x_2 \mid T) = \frac{P(X_1 = x_1, X_2 = x_2, T = x_1 + x_2)}{P(T = x_1 + x_2)}$. But for any given x_1 and x_2, the event $(X_1 = x_1, X_2 = x_2, T = x_1 + x_2)$ is the same as the event $(X_1 = x_1, X_2 = x_2)$ [for if $X_1 = x_1$ and $X_2 = x_2$, then necessarily $T = x_1 + x_2$] so these two events have the same probability. So $f(x_1,x_2 \mid T) = \frac{P(X_1 = x_1, X_2 = x_2, T = x_1 + x_2)}{P(T = x_1 + x_2)} = \frac{P(X_1 = x_1, X_2 = x_2)}{P(T = x_1 + x_2)} =$

$\frac{f(x_1;\lambda)f(x_2;\lambda)}{P(T = x_1 + x_2)} = \frac{\frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}}{P(T = x_1 + x_2)} = \frac{\frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}}{P(T = x_1 + x_2)}$. The quick way to conclude now is to draw upon the statistical fact that the sum of two independent Poisson random variables is also Poisson, with parameter equal to the sum of the two Poisson parameters. That is, if $U_1 \sim Po(\lambda_1)$ and $U_2 \sim Po(\lambda_2)$ are independent, then $U_1 + U_2 \sim Po(\lambda_1 + \lambda_2)$. Thus $T = X_1 + X_2 \sim Po(2\lambda)$. Hence, $P(T = x_1 + x_2) =$

$\frac{(2\lambda)^{x_1+x_2}e^{-2\lambda}}{(x_1+x_2)!}$. Thus, $f(x_1,x_2\,|\,T)=\frac{\frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}}{P(T=x_1+x_2)}=\frac{\frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}}{\frac{(2\lambda)^{x_1+x_2}e^{-2\lambda}}{(x_1+x_2)!}}=\frac{(x_1+x_2)!}{2^{x_1+x_2}x_1!x_2!}$. This does not depend upon λ. Once *T* is known, the data lose their sensitivity to changes in λ, hence lose their ability to tell us anything further about the value of λ. Therefore *T* is sufficient for λ.

Technical note: In this argument, if you do not know that the sum of two independent Poisson random variables is also Poisson, you can derive the marginal density of *T* from the joint density $f(x_1,x_2;\lambda)$ by a change of variable: Let $t=x_1+x_2$, $s=x_1$. This is a one-to-one invertible mapping from $\{(x_1,x_2);x_1=0,1,2,...,x_2=0,1,2,...\}$ to $\{(s,t);t=0,1,2,...,s=0,1,2,...,t\}$. Thus, the joint density function of (*S*,*T*) is obtained by substituting $x_1=s$, $x_2=t-s$ into $f(x_1,x_2;\lambda)=\frac{\lambda^{x_1+x_2}e^{-2\lambda}}{x_1!x_2!}$ to yield $f(s,t;\lambda)=\frac{\lambda^{s+(t-s)}e^{-2\lambda}}{s!(t-s)!}=\frac{\lambda^t e^{-2\lambda}}{s!(t-s)!}$ on $\{(s,t);t=0,1,2,...,s=0,1,2,...,t\}$. To obtain the marginal density of *T*, pick a generic *t*, and sum the joint density on all *s*: $f_T(t)=\sum_{s=0}^{t}f(s,t;\lambda)=\sum_{s=0}^{t}\frac{\lambda^t e^{-2\lambda}}{s!(t-s)!}=$ $\frac{\lambda^t e^{-2\lambda}}{t!}\sum_{s=0}^{t}\frac{t!}{s!(t-s)!}=\frac{\lambda^t e^{-2\lambda}}{t!}2^t$ by Newton's binomial formula $\left(\sum_{s=0}^{t}\frac{t!}{s!(t-s)!}=2^t\right)$. This is the same form for the p.d.f. that I stated earlier for the sum of two independent Poissons. So the argument can be completed as before.

The continuous example: Suppose that X_1 and X_2 are independent $N(\mu,\sigma^2)$ variables, where μ is unknown but σ^2 is known (to make the example simple). Then the joint density function of X_1 and X_2 is $f(x_1,x_2;\mu)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}=\frac{1}{\sigma^2 2\pi}e^{\frac{-1}{2\sigma^2}\left\{x_1^2+x_2^2-2(x_1+x_2)\mu+2\mu^2\right\}}$ for $-\infty<x_1<\infty,-\infty<x_2<\infty$. I will show that T = x_1+x_2 is sufficient. Once again, this will also show that the sample mean $\bar{x}=\frac{x_1+x_2}{2}$ is sufficient, since $\bar{x}$ and *T* are one-to-one functions of each other and therefore contain exactly the same information. How to do this? I will show that $f(x_1,x_2\,|\,T)$ does not depend upon μ. This will satisfy the definition of sufficiency. A new wrinkle arises in the continuous case that does not arise in the discrete Poisson example. In the discrete example, the densities are legitimate probabilities. Not so in the continuous case: The densities are "likelihoods". Nevertheless, much the same argument goes through: By definition of conditional densities, $f(x_1,x_2\,|\,T)=\frac{f(x_1,x_2,t)}{f_T(t)}$. Since $t=x_1+x_2$, then the value of the joint density $f(x_1,x_2,t)$ must be the same as the value of the joint density $f(x_1,x_2)$ except possibly on a set of (x_1,x_2) that has probability zero. More rigorously, this

is so because $f(x_1,x_2,t) = f(x_1,x_2)f(t \mid x_1,x_2) = f(x_1,x_2)\cdot 1$ since the conditional distribution of *t* given x_1, x_2 puts all of its probability on the single value $t = x_1 + x_2$. Thus, $f(x_1,x_2 \mid T) = \frac{f(x_1,x_2,t)}{f_T(t)} = \frac{f(x_1,x_2)}{f_T(t)} = \frac{\frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{x_1^2+x_2^2-2(x_1+x_2)\mu+2\mu^2\}}}{f_T(t)}$. The quick way to conclude now is to draw upon the statistical fact that the sum of two independent Normal random variables is also Normal, with mean and variance equal to the sum of the two Normal parameters. That is, if $U_1 \sim N(\mu_1,{\sigma_1}^2)$ and $U_2 \sim N(\mu_2,{\sigma_2}^2)$ are independent, then $U_1 + U_2 \sim N(\mu_1+\mu_2, {\sigma_1}^2+{\sigma_2}^2)$. Thus $T = X_1 + X_2 \sim N(2\mu, 2\sigma^2)$. Hence, $f_T(t) = \frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-(t-2\mu)^2}{4\sigma^2}} = \frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-1}{4\sigma^2}\{t^2-4t\mu+4\mu^2\}}$. So $f_T(t = x_1+x_2) = \frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-1}{4\sigma^2}\{(x_1+x_2)^2-4(x_1+x_2)\mu+4\mu^2\}}$ is the marginal density function of *T*. Therefore,

$$f(x_1,x_2 \mid T = x_1+x_2) = \frac{\frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{x_1^2+x_2^2-2(x_1+x_2)\mu+2\mu^2\}}}{f_T(t = x_1+x_2)} = \frac{\frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{x_1^2+x_2^2-2(x_1+x_2)\mu+2\mu^2\}}}{\frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-1}{4\sigma^2}\{(x_1+x_2)^2-4(x_1+x_2)\mu+4\mu^2\}}} =$$

$\frac{1}{\sigma\sqrt{\pi}} e^{\frac{-1}{2\sigma^2}\{x_1^2+x_2^2\}+\frac{1}{4\sigma^2}(x_1+x_2)^2}$, which does not depend upon μ. Once *T* is known, the data lose their sensitivity to changes in μ, hence lose their ability to tell us anything further about the value of μ. Therefore *T* is sufficient for μ.

Very technical note: In the above argument, if you do not know that the sum of two independent Normal random variables is also Normal, you can derive the marginal density of *T* from the joint density $f(x_1,x_2;\mu)$ by a change of variable: Let $t = x_1 + x_2$, $s = x_1$. This is a one-to-one invertible mapping from $\{(x_1,x_2); -\infty < x_1 < \infty, -\infty < x_2 < \infty\}$ to $\{(s,t); -\infty < s < \infty, -\infty < t < \infty\}$. Unlike a change of variables in discrete densities, a one-to-one change of variables in continuous densities may stretch or compress the space on which they are defined. The densities must stretch or compress correspondingly in order to keep the total probability at 1.[3] This is accomplished in multivariable calculus by inserting an adjustment factor called the Jacobian of the transformation $|J|$ into the transformed density. Here, in this example, $|J| = \begin{vmatrix} \partial x_1/\partial s & \partial x_1/\partial t \\ \partial x_2/\partial s & \partial x_2/\partial t \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} = 1$. The Jacobian is the absolute value of the determinant of the matrix of partial derivatives. In this case, the transformation neither stretches nor

[3] For example, if there were a transformation $u = 2v$, then the transformation would stretch the space of *v* two-fold, and the unit interval [0,1] in *v* would become [0,2] in *u*. So the density over [0,1] in *v* would need to be reduced by half over [0,2] in *u* in order to keep the probabilities of the intervals the same.

compresses, since $|J| = 1$. Thus, the joint density function of (*S*,*T*) is obtained by substituting $x_1 = s$, $x_2 = t - s$ and $|J|$ into $f(x_1, x_2; \mu) = \frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{x_1^2 + x_2^2 - 2(x_1+x_2)\mu + 2\mu^2\}}$ to yield $f(s,t;\mu) = \frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{s^2 + (t-s)^2 - 2t\mu + 2\mu^2\}} |J|$ on $\{(s,t); -\infty < s < \infty, -\infty < t < \infty\}$. To obtain the marginal density of *T*, pick a generic *t*, and integrate the joint density on *s*: $f_T(t) = \int_{-\infty}^{\infty} \frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{s^2 + (t-s)^2 - 2t\mu + 2\mu^2\}} |J| ds =$ $\int_{-\infty}^{\infty} \frac{1}{\sigma^2 2\pi} e^{\frac{-1}{2\sigma^2}\{2s^2 - 2ts + t^2 - 2t\mu + 2\mu^2\}} 1 ds$. The key to further progress is a useful integration hack: Reconstruct the integrand to contain a normal density in *s*, which will integrate to the value 1 on account of being a p.d.f, and pull the remaining parts outside the integral:

$$f_T(t) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-2}{2\sigma^2}\left\{s^2 - ts + \frac{t^2}{4}\right\}} e^{\frac{-1}{2\sigma^2}\left\{\frac{t^2}{2} - 2t\mu + 2\mu^2\right\}} ds =$$

$$\frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-1}{2\sigma^2}\left\{\frac{t^2}{2} - 2t\mu + 2\mu^2\right\}} \int_{-\infty}^{\infty} \frac{1}{\frac{\sigma}{\sqrt{2}}\sqrt{2\pi}} e^{\frac{-1}{2\frac{\sigma^2}{2}}\left\{s - \frac{t}{2}\right\}^2} ds$$

. The integrand now has the form of a $N(\frac{t}{2}, \frac{\sigma^2}{2})$ density in the variable *s*, which is integrated over its range. Therefore, the integral = 1. Thus, $f_T(t) = \frac{1}{\sqrt{2}\sigma\sqrt{2\pi}} e^{\frac{-1}{4\sigma^2}\{t^2 - 4t\mu + 4\mu^2\}}$. This is the same form for the p.d.f. that I stated earlier to be the density of the sum of two independent normals. So the argument can be completed as before.

The Case of *n* Random Variables

The Poisson and normal examples discussed above are for the case of two observations, (x_1, x_2). What about the general case of *n* observations, $(x_1, ..., x_n)$? For both the Poisson and normal cases, it remains true that the sum $x_1 + \cdots + x_n$ and (equivalently) the sample mean $\bar{x}$ are sufficient statistics for λ and μ, respectively. Essentially the same arguments may be used to show that $f(x_1, ..., x_n | T)$ does not depend upon the parameters, provided due consideration is given to the transformations. I.e., there must be *n*-1 "helper" transformations. It is convenient to take $T = x_1 + \cdots + x_n$, and $s_1 = x_1, s_2 = x_2, ..., s_{n-1} = x_{n-1}$. In the continuous case, the Jacobian of the transformation is the determinant of the *n* x *n* matrix of all partial derivatives. The helper variables must all be integrated out of the transformed joint density.

I think you will agree from these examples that the brute force demonstration of sufficiency can be technical and hard. Given the complexity of some of the mathematics of sufficiency, sufficiency would be a difficult idea to operationalize without easy ways to identify sufficient statistics. But there is help on the way! Here is the most important helpful result. The Factorization Theorem enables the identification of sufficient statistics with very little calculation – virtually by inspection:

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 [4] and a function $h(x_1,...,x_n)$ of the data but not of θ. [$h(x_1,...,x_n)$ could be a trivial function of the data, like a constant.]

The Factorization is very useful because you usually know what the joint density is. So you write that down, then look at it, and figure out how to split it into the two parts *g* and *h*. Here are two examples of the use of the Factorization Theorem to find sufficient statistics:

Discrete example. Suppose that $X_1, X_2,..., X_n$ are independent Poisson variables. Then their joint density is $f(x_1, x_2,..., x_n;\lambda) = \frac{\lambda^{x_1}e^{-\lambda}}{x_1!}\frac{\lambda^{x_2}e^{-\lambda}}{x_2!}\cdots\frac{\lambda^{x_n}e^{-\lambda}}{x_n!} = \frac{\lambda^{x_1+x_2+\cdots+x_n}e^{-n\lambda}}{x_1!x_2!\cdots x_n!}$. To apply the Factorization Theorem, we can choose $T = x_1 + \cdots + x_n$, $g(T,\lambda) = \lambda^{x_1+x_2+\cdots+x_n}e^{-n\lambda} = \lambda^T e^{-n\lambda}$, and $h(x_1,...,x_n) = \frac{1}{x_1!x_2!\cdots x_n!}$. Then $f(x_1, x_2,..., x_n;\lambda) = \frac{\lambda^{x_1+x_2+\cdots+x_n}e^{-n\lambda}}{x_1!x_2!\cdots x_n!} = g(T,\lambda)h(x_1,...,x_n)$. By the Factorization Theorem, $T = x_1 + \cdots + x_n$ is sufficient for λ.

Continuous example. Suppose that $X_1, X_2,..., X_n$ are independent $N(\mu,\sigma^2)$ variables, where μ is unknown but σ^2 is known (for simplicity). Then their joint density is

$$f(x_1, x_2,..., x_n;\mu) = \frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}}\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_2-\mu)^2}{2\sigma^2}}\cdots\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x_n-\mu)^2}{2\sigma^2}} =$$

$$\frac{1}{(\sigma\sqrt{2\pi})^n}e^{\frac{-(x_1-\mu)^2}{2\sigma^2}\frac{-(x_2-\mu)^2}{2\sigma^2}\cdots\frac{-(x_n-\mu)^2}{2\sigma^2}} = \frac{1}{(\sigma\sqrt{2\pi})^n}e^{\frac{-1}{2\sigma^2}\left\{x_1^2+x_2^2+\cdots+x_n^2-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}} =$$

$$e^{\frac{-1}{2\sigma^2}\left\{-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}}\frac{1}{(\sigma\sqrt{2\pi})^n}e^{\frac{-1}{2\sigma^2}\left\{x_1^2+x_2^2+\cdots+x_n^2\right\}}.$$ To apply the Factorization Theorem, take

$T = x_1 + \cdots + x_n$, $g(T,\mu) = e^{\frac{-1}{2\sigma^2}\left\{-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}} = e^{\frac{-1}{2\sigma^2}\left\{-2\mu T+n\mu^2\right\}}$, and $h(x_1,...,x_n) =$

$\frac{1}{(\sigma\sqrt{2\pi})^n}e^{\frac{-1}{2\sigma^2}\left\{x_1^2+x_2^2+\cdots+x_n^2\right\}}$. Then $f(x_1, x_2,..., x_n;\mu) = \frac{1}{(\sigma\sqrt{2\pi})^n}e^{\frac{-1}{2\sigma^2}\left\{x_1^2+x_2^2+\cdots+x_n^2-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}} =$

$g(T,\mu)h(x_1,..., x_n)$. By the Factorization Theorem, $T = x_1 + \cdots + x_n$ is sufficient for μ.

[4] $g(T,\theta)$ could also be a function of the sample size *n*, which is not considered a parameter in this context.

Continuation. Now suppose in the preceding example that σ^2 is unknown. What then? The Factorization Theorem can still be used to show that the sample mean and sample variance are together sufficient for μ and σ^2. That is, the sufficient statistic is two-dimensional. To see this, take $\mathbf{T} = (t_1, t_2) = (x_1 + \cdots + x_n, x_1^{\,2} + \cdots + x_n^{\,2})$, and $g(\mathbf{T}, \mu, \sigma^2) = g((t_1, t_2), \mu, \sigma^2)$ = $\frac{1}{(\sigma\sqrt{2\pi})^n} e^{\frac{-1}{2\sigma^2}\left\{x_1^{\,2}+x_2^{\,2}+\cdots+x_n^{\,2}-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}} = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{\frac{-1}{2\sigma^2}\left\{t_2-2\mu t_1+n\mu^2\right\}}$, and $h(x_1,...,x_n)$ = 1. Then $f(x_1, x_2, ..., x_n; \mu) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{\frac{-1}{2\sigma^2}\left\{x_1^{\,2}+x_2^{\,2}+\cdots+x_n^{\,2}-2\mu(x_1+x_2+\cdots+x_n)+n\mu^2\right\}} = g(\mathbf{T}, \mu)h(x_1, ..., x_n)$. By the Factorization Theorem, $\mathbf{T} = (t_1, t_2) = (x_1 + \cdots + x_n, x_1^{\,2} + \cdots + x_n^{\,2})$ is sufficient for (μ, σ^2). But any one-to-one function of **T** is also sufficient. In that regard, consider $(\bar{x}, S^2) = \left(\frac{x_1 + \cdots + x_n}{n}, \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \right) =$

$$(\bar{x}, S^2) = \left(\frac{t_1}{n}, \frac{\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n}{n-1} \right) = \left(\frac{t_1}{n}, \frac{t_2 - (t_1)^2 / n}{n-1} \right),$$

which is a one-to-one function of **T**.

Therefore, the sample mean and sample variance are jointly sufficient for μ and σ^2.

These examples show that sufficient statistics often can be obtained easily by inspection of the joint density.

A Caveat

One must be careful about applying the Factorization Theorem. For example, suppose that $X_1, X_2, ..., X_n$ are independent random variables with common density function $f(x;\theta) = e^{-(x-\theta)}$ if $x > \theta$ and is 0 elsewhere. [You can verify that $\int_{-\infty}^{+\infty} f(x;\theta)\, dx = \int_{\theta}^{+\infty} e^{-(x-\theta)} dx = 1$, so that $f(x;\theta)$ is a legitimate density.] The joint density is $f(x_1,...,x_n;\theta) = e^{-(x_1-\theta)} e^{-(x_2-\theta)} \cdots e^{-(x_n-\theta)} = e^{-(x_1+\cdots+x_n)+n\theta}$, which can be "factored" in many different ways. For example, simply take $T = x_1 + \cdots + x_n$, and $g(T, \theta)$ = $e^{-(x_1+\cdots+x_n)+n\theta}$, and $h(x_1,...,x_n)$ = 1. Then $f(x_1,...,x_n;\theta) = e^{-(x_1+\cdots+x_n)+n\theta} = g(T, \theta)\ h(x_1,...,x_n)$ suggests that the sample mean is sufficient for θ by the Factorization Theorem.

Not so! The problem here is that the density is *not* $f(x;\theta) = e^{-(x-\theta)}$. This is the value of the density only if $x > \theta$. The Factorization Theorem must hold for *all x*. In order to bring all *x* into the p.d.f., rewrite the

density as $f(x;\theta) = e^{-(x-\theta)}I(x > \theta)$, where the indicator $I()$ = 1 if the condition inside the parentheses is true and = 0 otherwise. Thus the *true* joint density is
$e^{-(x_1-\theta)}I(x_1 > \theta)e^{-(x_2-\theta)}I(x_2 > \theta)\cdots e^{-(x_n-\theta)}I(x_n > \theta)$ =
$e^{-(x_1+x_2+\cdots+x_n-n\theta)}I(x_1 > \theta)I(x_2 > \theta)\cdots I(x_n > \theta)$ = $e^{-(x_1+x_2+\cdots+x_n-n\theta)}I(\min\{x_1, x_2, \cdots, x_n\} > \theta)$ because the individual indicators are all equal to 1 if and only if all of the *x*'s exceed θ. Now choose $T = \min\{x_1, x_2, \cdots, x_n\}$, $g(T,\theta) = e^{n\theta}I(T > \theta)$, and $h(x_1, x_2, \cdots, x_n) = e^{-(x_1+x_2+\cdots+x_n)}$ to see that the requirements of the Factorization Theorem are met for *all* *x*'s. Thus, $T = \min\{x_1, x_2, \cdots, x_n\}$ is sufficient for θ. Intuitively, until you have seen the smallest *x*, you do not know the upper limit to the possible values of θ. θ cannot exceed any of the *x*'s.

Densities sometimes have an implicit indicator function that restricts the range on which the density is positive. That is the role of $I(x > \theta)$ in the preceding example. The restriction must also be taken into account in the Factorization Theorem. That can be done by making the indicator function an explicit part of the density.

Another example. Suppose that $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$ are *n* independent coordinates drawn at random from the interior of a circle of radius θ and centered at (0,0). Then the density appears to be $f(x, y;\theta) = \frac{1}{\pi\theta^2}$. So the joint density appears to be

$f((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n);\theta) = \left(\frac{1}{\pi\theta^2}\right)^n$, which can be "factored" in many different ways. E.g.,

$\left(\frac{1}{\pi\theta^2}\right)^n = \left(\frac{|x_1|}{\pi\theta^2}\right)^n \frac{1}{|x_1|^n}$, which suggests that $|x_1|$ is sufficient for θ by the Factorization Theorem.

The problem here is that the density is $f(x, y;\theta) = \frac{1}{\pi\theta^2}$ only on the interior of the circle. The density can be rewritten to apply to all coordinates as $f(x, y;\theta) = \frac{1}{\pi\theta^2}I(x^2 + y^2 < \theta^2)$, where $I()$ = 1 if the condition inside the parentheses is true and = 0 otherwise. Thus, the joint density is

$f((x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n);\theta) = \left(\frac{1}{\pi\theta^2}\right)^n I(x_1^2 + y_1^2 < \theta^2)I(x_2^2 + y_2^2 < \theta^2)\cdots I(x_n^2 + y_n^2 < \theta^2)$

$= \left(\frac{1}{\pi\theta^2}\right)^n I(\max\{x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_n^2 + y_n^2\} < \theta^2)$. [All of the squared sample radii are less than θ^2 if and only if the maximum squared sample radius is less than θ^2.] By the Factorization Theorem, $\max\{x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_n^2 + y_n^2\}$ is a sufficient statistic for θ. This makes intuitive sense because the radius of the circle cannot be closer to the origin than the sample point that is farthest from the origin. Until you know that farthest point, your information is incomplete.