Information in Data

Sufficiency, Ancillarity, Minimality, and Completeness

Important properties of statistics that determine the usefulness of those statistics in statistical inference.

These general properties often can be used as guides in seeking optimal statistical procedures.
Let $X$ be a sample from a population $P \in \mathcal{P}$. A statistic $T(X)$ is **sufficient** for $P \in \mathcal{P}$ if and only if the conditional distribution of $X$ given $T$ does not depend on $P$.

In general terms, this involves the conditional independence from the parameter of the distribution of any other function of the random variable, given the sufficient statistic.

Sufficiency depends on the family of distributions, $\mathcal{P}$, w.r.t. which $E$ is defined. If a statistic is sufficient for $\mathcal{P}$, it may not be sufficient for a larger family, $\mathcal{P}_1$, where $\mathcal{P} \subset \mathcal{P}_1$.

Sufficiency may allow reduction of data without sacrifice of information.
Sufficiency

We can establish sufficiency by the factorization criterion.

A necessary and sufficient condition for a statistic $T$ to be sufficient for a family $\mathcal{P}$ of distributions of a sample $X$ dominated by a $\sigma$-finite measure $\nu$ is that there exist nonnegative Borel functions $g_P$ and $h$, where $h$ does not depend on $P$, such that

$$
\frac{dP}{d\nu}(x) = g_P(T(x))h(x) \quad \text{a.e. } \nu.
$$

An important consequence of sufficiency in an estimation problem with convex loss is the Rao-Blackwell theorem.

When the density can be written in the separable form $c(\theta)f(x)$, unless $c(\theta)$ is a constant, the support must be a function of $\theta$, and a sufficient statistic must be an extreme order statistic. When the support depends on the parameter, the extreme order statistic(s) at the boundary of the support determined by the parameter carry the full information about the parameter.
Ancillarity

Ancillarity is, in a way, the opposite of sufficiency: A statistic $U(X)$ is called ancillary for $P$ (or $\theta$) if the distribution of $U(X)$ does not depend on $P$ (or $\theta$).

If $\mathbb{E}(U(X))$ does not depend on $P$ (or $\theta$), then $U(X)$ is said to be first-order ancillary for $P$ (or $\theta$).

Often a probability model contains a parameter of no interest for inference. Such a parameter is called a nuisance parameter. A statistic to be used for inferences about the parameters of interest should be ancillary for a nuisance parameter.
Minimal Sufficiency

Let $T$ be a given sufficient statistic for $P \in \mathcal{P}$. The statistic $T$ is *minimal sufficient* if for any sufficient statistic for $P \in \mathcal{P}$, $S$, there is a measurable function $h$ such that $T = h(S)$ a.s. $\mathcal{P}$.

Minimal sufficiency has a heuristic appeal: it relates to the greatest amount of data reduction.

An easy way of establishing minimality when the range does not depend on the parameter is by use of the following facts:

Let $\mathcal{P}$ be a family with densities $p_0, p_1, \ldots, p_k$, all with the same support. The statistic

$$T(X) = \left( \frac{p_1(X)}{p_0(X)}, \ldots, \frac{p_k(X)}{p_0(X)} \right)$$

is minimal sufficient.
This follows from the following corollary of the factorization theorem:

A necessary and sufficient condition for a statistic $T$ to be sufficient for a family $\mathcal{P}$ of distributions of a sample $X$ dominated by a $\sigma$-finite measure $\nu$ is that for any two densities $p_1$ and $p_2$ in $\mathcal{P}$, the ratio $p_1(x)/p_2(x)$ is a function only of $T(x)$.

Then, the other important fact is

Let $\mathcal{P}$ be a family of distributions with the common support, and let $\mathcal{P}_0 \subset \mathcal{P}$. If $T$ is minimal sufficient for $\mathcal{P}_0$ and is sufficient for $\mathcal{P}$, then it is minimal sufficient for $\mathcal{P}$.

We see this by considering any statistic $U$ that is sufficient for $\mathcal{P}$. It must also be sufficient for $\mathcal{P}_0$, and since $T$ is minimal sufficient for $\mathcal{P}_0$, $T$ is a function of $U$. 
Completeness

A sufficient statistic $T$ is particularly useful in a complete family or a boundedly complete family of distributions.

In this case, for every Borel (bounded) function $h$ that does not involve $P$,

$$\mathbb{E}_P(h(T)) = 0 \forall P \in \mathcal{P} \Rightarrow h(t) = 0 \text{ a.e. } \mathcal{P}.$$  

In a complete family, we often refer to the completeness of a statistic, rather than the completeness of the family.

We often call a sufficient statistic in a complete family, a “complete sufficient” statistic.
Complete Sufficiency

Complete sufficiency depends on $\mathcal{P}$, the family of distributions w.r.t. which $E$ is defined. If a statistic is complete and sufficient with respect to $\mathcal{P}$, and if it is sufficient for $\mathcal{P}_1$, where $\mathcal{P} \subset \mathcal{P}_1$ and all distributions in $\mathcal{P}_1$ have common support, then it is complete and sufficient for $\mathcal{P}_1$, because in this case, the condition a.s. $\mathcal{P}$ implies the condition a.s. $\mathcal{P}_1$. 
Complete Sufficiency

Complete sufficiency is useful for establishing independence using Basu’s theorem, and in estimation problems in which we seek an unbiased estimator that has minimum variance uniformly (UMVUE).

It is important to remember that completeness and sufficiency are different properties; you can have either one without the other.

There is always a sufficient statistic: the sample itself.

There may not be a complete statistic within a given family.
Complete Sufficiency

If there is a complete statistic and it is sufficient, then it is minimal sufficient.

Complete sufficiency implies minimal sufficiency, but minimal sufficiency does not imply completeness.
Basu’s Theorem

Complete sufficiency, ancillarity, and independence are related.

Basu’s theorem (Theorem 2.4 in Shao) states that if $T$ is a boundedly complete sufficient statistic for $\theta$, and if $U$ is ancillary for $\theta$, then $T$ and $U$ are independent.
Consider \( U(\theta - 1/2, \theta + 1/2) \)

Let \( X_1, \ldots, X_n \), with \( n \geq 2 \), be i.i.d. as \( U(\theta - 1/2, \theta + 1/2) \). It is clear that \( T = \{X_{(1)}, X_{(n)}\} \) is sufficient; in fact, it is minimal sufficient. Now consider \( U = X_{(n)} - X_{(1)} \), which we easily see is ancillary. It is clear that \( T \) and \( U \) are not independent (\( U \) is a function of \( T \)).

This shows the importance of completeness in Basu’s theorem and also shows that minimality does not imply completeness.

Writing \( U = h(T) \), where \( h \) is a measurable function, we can see the \( T \) is not complete (although it is minimal.)

If \( T \) were complete, then Basu’s theorem would say that \( T \) and \( U \) are independent.
Sufficiency, Minimality, and Completeness in Various Families

We can use general properties of specific families of distributions to establish properties of statistics quickly and easily.
Approaches to Statistical Inference

- use of the empirical cumulative distribution function (ECDF) for example, method of moments

- use of a likelihood function for example, maximum likelihood

- fitting expected values for example, least squares

- fitting a probability distribution for example, maximum entropy

- definition and use of a loss function for example, uniform minimum variance unbiased estimation.
Approaches to Statistical Inference

These approaches are associated with various philosophical/scientific principles, sometimes explicitly stated and sometimes not.

Some of these principles, such as the likelihood principle, inform a major class of statistical methods, while other principles, such as the bootstrap principle, are more local in application.
Decision Theoretic Approach

Comprehensive.

Bayesian methods

UMVUE
Approaches Based on the Likelihood
The Likelihood Function

Given a sample $x_1, \ldots, x_n$ from distributions with probability densities $p_i(x)$ with respect to a common $\sigma$-finite measure, the likelihood function is defined as

$$ L_n(p_i ; x) = c \prod_{i=1}^{n} p_i(x_i), $$

where $c$ is any positive constant; that is, the likelihood function is any member of an equivalence class $\{cL : c \in \mathbb{R}_+\}$.

The likelihood principle (below) justifies this view of the likelihood function.

It is common to speak of $L_n^1(p_i ; x) = \prod_{i=1}^{n} p_i(x_i)$ as “the” likelihood function.

Methods based on the likelihood function are often chosen because of their asymptotic properties, and so it is common to use the $n$ subscript.
The Likelihood Function

The domain of the likelihood function is some class of distributions specified by their probability densities, \( \mathcal{P} = \{ p_i(x) \} \), where all PDFs are with respect to a common \( \sigma \)-finite measure.

In applications, often the PDFs are of a common parametric form, so equivalently, we can think of the domain of the likelihood function as being a parameter space, say \( \Theta \), so the family of densities can be written as \( \mathcal{P} = \{ p_\theta(x) \} \) where \( \theta \in \Theta \), the known parameter space.

We often write the likelihood function as

\[
L(\theta ; x) = \prod_{i=1}^{n} p(x_i ; \theta).
\]

Although we have written \( L(\theta ; x) \), the expression \( L(p_\theta ; x) \) may be more appropriate because it reminds us of an essential ingredient in the likelihood, namely a PDF.
The Likelihood Function

Example 1 likelihood in an exponential family
Consider the exponential family of distributions with parameter \( \lambda \). The PDF is

\[
p_X(x; \lambda) = \lambda e^{-\lambda x} I_{\mathbb{R}^+}(x),
\]

for \( \lambda \in \mathbb{R}_+ \).

Given a single observation \( x \), the likelihood is

\[
L(\lambda; x) = \lambda e^{-\lambda x} I_{\mathbb{R}^+}(\lambda).
\]

See the figure.
\[ \text{Likelihood } L(\theta; x) \]

PDF \[ p_x(x; \theta) \]
What Likelihood Is Not

A likelihood is neither a probability nor a probability density.

Notice, for example, that the definite integral over $\mathbb{R}_+$ of the likelihood in Example 1 is $x^{-2}$; that is, the likelihood does not integrate to 1.

It is not appropriate to refer to the “likelihood of an observation”. We use the term “likelihood” in the sense of the likelihood of a model or the likelihood of a distribution *given observations*. 
The Log-Likelihood Function

The *log-likelihood function*,

\[ l_L(\theta ; x) = \log L(\theta ; x), \]

is a sum rather than a product.

We often denote the log-likelihood without the “\(L\)” subscript.

The notation for the likelihood and the log-likelihood varies with authors.

We will often work with either the likelihood or the log-likelihood as if there is only one observation.
Likelihood Principle

The *likelihood principle* in statistical inference asserts that all of the information which the data provide concerning the relative merits of two hypotheses (two possible distributions that give rise to the data) is contained in the likelihood ratio of those hypotheses and the data.

An alternative statement of the likelihood principle is that if for \( x \) and \( y \),

\[
\frac{L(\theta ; x)}{L(\theta ; y)} = c(x, y) \quad \forall \theta,
\]

where \( c(x, y) \) is constant for given \( x \) and \( y \), then any inference about \( \theta \) based on \( x \) should be in agreement with any inference about \( \theta \) based on \( y \).
Likelihood Principle

Although at first glance, we may think that the likelihood principle is so obviously the right way to make decisions, an example may cause us to think more critically about this principle.

The likelihood principle asserts that for making inferences about a probability distribution, the overall data-generating process need not be considered; only the observed data are relevant.
Example of an Implication of the Likelihood Principle

Example 2 The likelihood principle in sampling in a Bernoulli distribution
Consider (again!) the problem of making inferences on the parameter $\pi$ in a family of Bernoulli distributions.

One approach is to take a random sample of size $n$, $X_1, \ldots, X_n$ from the Bernoulli($\pi$), and then use $T = \sum X_i$, which has a binomial distribution with parameters $n$ and $\pi$.

Another approach is to take a sequential sample, $X_1, X_2, \ldots$, until a fixed number $t$ of 1’s have occurred. The size of the sample $N$ is random and the random variable $N$ has a negative binomial distribution with parameters $t$ and $\pi$. 
Now, suppose we take the first approach with $n = 12$ and we observe $t = 3$; and then we take the second approach with $t = 2$ and we observe $n = 12$. Using the PDFs of the binomial and the negative binomial, we get the likelihoods

$$L_B(\pi) = \binom{12}{3} \pi^3 (1 - \pi)^9$$

and

$$L_{NB}(\pi) = \binom{11}{2} \pi^3 (1 - \pi)^9.$$ 

Because $L_B(\pi)/L_{NB}(\pi)$ does not involve $\pi$, in accordance with the likelihood principle, any decision about $\pi$ based on a binomial observation of 5 out of 10 should be the same as any decision about $\pi$ based on a negative binomial observation of 10 for 5 1’s. This seems reasonable.

**However:**
variance estimates and significance tests will be different!
Maximum Likelihood Estimation

Let us assume a parametric model; that is, a family of densities \( \mathcal{P} = \{p(x ; \theta)\} \) where \( \theta \in \Theta \), a known parameter space.

For a sample \( X_1, \ldots, X_n \) from a distribution with probability density \( p(x ; \theta) \), we write the likelihood function as a function of a variable in place of the parameter:

\[
L(t ; x) = \prod_{i=1}^{n} p(x_i ; t).
\]

Note the reversal in roles of variables and parameters.

I like to write the likelihood as function of a variable of something other than the parameter, which I think of as fixed, but I usually write it like everyone else; that is, \( L(\theta ; x) = \prod_{i=1}^{n} p(x_i ; \theta) \).
The data, that is, the realizations of the variables in the density function, are considered as fixed and the parameters are considered as variables of the optimization problem,

$$\max_{\theta} L(\theta ; x).$$

It is important to specify the domain of the likelihood function.
If \( \Theta \) is the domain of \( L \), we want to maximize \( L \) for \( t \in \Theta \).

There may be difficulties with this maximization problem, however, because of open sets.

The first kind of problem is because the parameter space may be open.

The second kind of open set may be the region over which the likelihood function is positive.
Definition and Examples

Let $L(\theta; x)$ be the likelihood of $\theta \in \Theta$ for the observations $x$ from a distribution with PDF with respect to a $\sigma$-finite measure $\nu$.

A maximum likelihood estimate, or MLE, of $\theta$, written $\hat{\theta}$, is defined as

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta; x),$$

if it exists.

There may be more than solution; any one is an MLE.
The estimate $\hat{\theta}$ is a function of the observations, $x$, and if $\hat{\theta}(x)$ is a Borel function a.e. w.r.t. $\nu$, then $\hat{\theta}$ is called a maximum likelihood estimator of $\theta$.

While I like to use the “hat” notation to mean an MLE, I also sometimes use it to mean any estimate or estimator.

We use “MLE” to denote maximum likelihood estimate or estimator, or the method of maximum likelihood estimation.
Notice that finding an MLE means to solve a constrained optimization problem.

In simple cases, the constraints may not be active.

In even simpler cases, the likelihood is differentiable, and the MLE occurs at a stationary point in the interior of the constraint space.

In these happy cases, the MLE can be identified by differentiation.
Example 3 MLE in an exponential family (continuation of Example 1)

In the exponential family of Example 1, with a sample $x_1, \ldots, x_n$, the likelihood is

$$L(\theta; x) = \theta^{-n} e^{-\sum_{i=1}^{n} x_i/\theta} I_{\mathbb{R}^+}(\theta),$$

whose derivative w.r.t $\theta$ is

$$\left(-n\theta^{-n-1} e^{-\sum_{i=1}^{n} x_i/\theta} + \theta^{-n-2} \sum_{i=1}^{n} x_i e^{-\sum_{i=1}^{n} x_i/\theta}\right) I_{\mathbb{R}^+}(\theta).$$

Equating this to zero, we obtain

$$\tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i / n.$$
Checking the second derivative, we find it is $-n$ at $\tilde{\theta}$, and so we conclude that $\tilde{\theta}$ is the MLE of $\theta$, and it is the only maximizer.

In many cases, of course, we cannot find an MLE by just differentiating the likelihood.
The Likelihood and the Log-Likelihood Functions

If \( \hat{\theta} \) exists, we also have

\[
\hat{\theta} = \arg \max_{\theta \in \Theta} l_L(\theta ; x),
\]

that is, the MLE can be identified either from the likelihood function or from the log-likelihood.
Example 4 MLE of Bernoulli parameter

Consider the Bernoulli family of distributions with parameter $\pi$.

In the usual definition of this family, $\pi \in \Pi = ]0, 1[.$

Suppose we take a random sample $X_1, \ldots, X_n$.

The log-likelihood is

$$l_L(\pi ; x) = \sum_{i=1}^{n} x_i \log(\pi) + \left( n - \sum_{i=1}^{n} x_i \right) \log(1 - \pi).$$

This is a concave differentiable function, so we can get the maximum by differentiating and setting the result to zero. We obtain

$$\hat{\pi} = \frac{\sum_{i=1}^{n} x_i}{n}.$$ 

If $\sum_{i=1}^{n} x_i = 0$ or if $\sum_{i=1}^{n} x_i = n$, $\hat{\pi} \notin \Pi$, but $\hat{\pi} \in \overline{\Pi}$ so $\hat{\pi}$ is the MLE of $\pi$. 
Note that in this case, the MLE corresponds to a Bayes estimator and to the UMVUE.
Allowing an MLE to be in $\overline{\Theta} - \Theta$ is preferable to saying that an MLE does not exist. It does, however, ignore the question of continuity of $L(\theta; x)$ over $\overline{\Theta}$, and it allows an estimated PDF that is degenerate.

We have encountered this situation before in the case of UMVUEs.

Another type of problem may arise that the definition does not take care of.

This is the case in which the maximum does not exist because the likelihood is unbounded from above. In this case the argmax does not exist, and the maximum likelihood estimate does not exist.
Example 5 nonexistence of MLE
Consider the normal family of distributions with parameters $\mu$ and $\sigma^2$. Suppose we have one observation $x$. The log-likelihood is

$$l_L(\mu, \sigma^2 ; x) = -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(x - \mu)^2}{2\sigma^2},$$

which is unbounded when $\mu = x$ and $\sigma^2$ approaches zero; hence, it is clear that no MLE exists.
While the open parameter space in Example 4 would lead to a problem with existence of the MLE unless we changed the definition of MLE, an open support can likewise lead to a problem.

Consider a distribution with Lebesgue PDF

$$p_X(x) = h(x, \theta)I_{S(\theta)}(x)$$

where $S(\theta)$ is open. In this case, the likelihood has the form

$$L(\theta; x) = h(x, \theta)I_{R(x)}(\theta),$$

where $R(x)$ is open.

It is quite possible that $\sup L(\theta; x)$ will occur on $\overline{R(x)} - R(x)$. 
Example 6 MLE when support is open
Consider $X_1, \ldots, X_n$ i.i.d. from $U(0, \theta)$ that is, with PDF

$$p_X(x) = \frac{1}{\theta} I_{[0, \theta]}(x),$$

where $\Theta = \mathbb{R}_+$. 

The likelihood is

$$L(\theta ; x) = \frac{1}{\theta} I_{[x_{(n)}, \infty]}(\theta).$$

This is discontinuous and it does not have a maximum, as we see in the figure.
In this case the maximum of the likelihood does not exist, but the supremum of the likelihood occurs at $x(n)$ and it is finite. We would like to call $x(n)$ the MLE of $\theta$.

We can reasonably do this by modifying the definition of the family of distributions by adding a zero-probability set to the support.

We redefine the family to have the Lebesgue PDF

$$p_X(x) = \frac{1}{\theta} I_{[0, \theta]}(x),$$

Now, the open interval $]x(n), \infty[$ where the likelihood was positive before becomes a half-closed interval $[x(n), \infty[$, and the maximum of the likelihood occurs at $x(n)$.

The UMVUE of $\theta$ is $(1 + 1/n)x(n)$. 
The approach in Example 6 is cleaner than solving the logical problem by defining the MLE in terms of the sup rather than the max.

A definition in terms of the sup may not address problems that could arise due to various types of discontinuity of $L(\theta; x)$ at the boundary of $S(\theta)$.

In the case of underlying normal probability distribution, estimation of the mean based on least squares is the same as MLE.
Example 7 MLE in a linear model

Let

\[ Y = X\beta + E, \]

where \( Y \) and \( E \) are \( n \)-vectors with \( \mathbb{E}(E) = 0 \) and \( \mathbb{V}(E) = \sigma^2 I_n \), \( X \) is an \( n \times p \) matrix whose rows are the \( x_i^\top \), and \( \beta \) is the \( p \)-vector parameter. A least squares estimator of \( \beta \) is

\[
\begin{align*}
    b^* &= \arg\min_{b \in B} \| Y - Xb \|^2 \\
    &= (X^\top X)^{-1} X^\top Y.
\end{align*}
\]

Even if \( X \) is not of full rank, in which case the least squares estimator is not unique, we found that the least squares estimator has certain optimal properties for estimable functions.

Of course at this point, we could not use MLE — we do not have a distribution.
Then we considered the additional assumption in the model that

\[ E \sim \mathcal{N}_n(0, \sigma^2 I_n), \]

or

\[ Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n). \]

In that case, we found that the least squares estimator yielded the unique UMVUE for any estimable function of \( \beta \) and for \( \sigma^2 \).

We could define a least squares estimator without an assumption on the distribution of \( Y \) or \( E \), but for an MLE we need an assumption on the distribution. Again, let us assume

\[ Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n), \]

yielding, for an observed \( y \), the log-likelihood

\[ l_L(\beta, \sigma^2, y, X) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta). \]

So an MLE of \( \beta \) is the same as a least squares estimator of \( \beta \).
Estimation of $\sigma^2$, however, is different.

In the case of least squares estimation, with no specific assumptions about the distribution of $E$ in the model, we have no basis for forming an objective function of squares to minimize.

Even with an assumption of normality, however, instead of explicitly forming a least squares problem for estimating $\sigma^2$, we used the least squares estimator of $\beta$, which is complete and sufficient, and we used the distribution of $(y - Xb^*)^T(y - Xb^*)$ to form a UMVUE of $\sigma^2$,

$$s^2 = (Y - Xb^*)^T(Y - Xb^*)/(n - r),$$

where $r = \text{rank}(X)$. 
In the case of maximum likelihood, we directly determine the value of $\sigma^2$ that maximizes the expression.

This is an easy optimization problem.

The solution is

$$\hat{\sigma}^2 = (y - X\hat{\beta})^T(y - X\hat{\beta})/n$$

where $\hat{\beta} = b^*$ is an MLE of $\beta$.

Compare the MLE of $\sigma^2$ with the least squares estimator, and note that the MLE is biased. Recall that we have encountered these two estimators in the simpler cases.