1 Parametric Models.

To begin with on a space X we have a family P_θ of probability distributions. In practice X will be either a countable set of points $\{x\}$ and P_θ specified by the individual probabilities $p(\theta, x)$ for $x \in X$. They will of course satisfy

$$\sum_x p(\theta, x) \equiv 1$$

Or one may have the situation where X is some finite dimensional space R^d (or perhaps a suset of one). P_θ in such situations will be specified by probability densities $f(\theta, x)$ with respect to the Lebesgue measure on R^d. In such a case

$$\int_{R^d} f(\theta, x)dx \equiv 1$$

The set Θ of possible values of the parameters is also usually a subset of some R^k. The integer k represents the number of parameters in the problem.

Identifiability. One assumes that if $\theta_1 \neq \theta_2$ are two distinct values of the parameter from Θ then the two probability measures P_{θ_1} and P_{θ_2} are different.

In statistical terminology an observation is a point $x \in X$ that has been observed. The true value of the parameter is unknown, although it is known that it comes from the set Θ.

The Model. The basic philosophy is that P_θ descibes the probability distribution of the result x of an experiment. The real underlying physical situation that generated the result is identified by the parameter θ. The model is specified by decribing how the probabilities of various events concerning the result of the experiment are determined by the parameter $\theta \in \Theta$, i.e. by P_θ.

Inference. Inference or Statistical Inference is a statement regarding the true value of the parameter based on the observation or 'evidence' x.

Inference in mathematics is based on logic, and presumably infallible atleast when correctly applied. Statistical inference on the other hand is based on probabilities. One is rarely certain about the inference. But one has a certain level of 'confidence' that can be expressed quantitaively. As more evidence is gathered the level confidence increases and approaches certainty only asymptotically.

Random Sample and Sample Size. If we have just a single observation, since most of the time any result or observation is compatible with any value of the parameter, i.e. for the observed x, $p(\theta, x) > 0$ or $f(\theta, x) > 0$ for all θ, one can never be sure of what θ is, based on a single observation x. However if we have repeated observations, we can gather more evidence and will be able to make more confident inference. If we have n independent observations under the same model, i.e. if we have a random smaple of size n, X gets replaced by its n-fold product $X^{(n)}$ and

$$p(\theta, x_1, \ldots, x_n) = \Pi_{i=1}^n p(\theta, x_i)$$

$$f(\theta, x_1, \ldots, x_n) = \Pi_{i=1}^n f(\theta, x_i)$$

Examples.

1. The Normal family $\{f(\mu, ; x)\}$ with $\mu \in R$ is given by the densities

$$f(\mu, ; x) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{(x-\mu)^2}{2}] \tag{1.1}$$

   is a one parameter family of probability densities on R.

2. The Normal family $\{f(\mu, \theta\,; x)\}$ with $\mu \in R$ and $\theta > 0$ is given by the densities

$$f(\mu, \theta\,; x) = \frac{1}{\sqrt{2\pi\theta}} \exp[-\frac{(x-\mu)^2}{2\theta}] \tag{1.2}$$

   is two parameter family of probability densities on R.

3. The family of Gamma distributions

$$f(\alpha, p\,; x) = \frac{\alpha^P}{\Gamma(p)} \exp[-\alpha x] x^{p-1} \tag{1.3}$$

   for $\alpha, p > 0$ is a two parameter family of densities on $[0, \infty)$

4. The family of Beta distributions

$$f(p, q\,; x) = \frac{1}{\beta(p,q)} x^{p-1}(1-x)^{q-1} \tag{1.4}$$

   for $p, q > 0$ is a two parameter family of densities on $[0, 1]$

5. The family of Cauchy distributions

$$f(\mu\,; x) = \frac{1}{2\pi} \frac{1}{1 + (x-\mu)^2} \tag{1.5}$$

   is a one parameter family densities on R.

6. For any density f on R, the family

$$f(\mu\,; x) = f(x - \mu) \tag{1.6}$$

   is a one parameter family of densities on R and $\mu \in R$ is called the location parameter.

7. For $\theta \in R$ the two sided exponential family is defined by

$$f(\theta,\, x) = \frac{1}{2} \exp[-|x - \theta|] \tag{1.7}$$

8. For any density f on R, the family

$$f(\mu, \sigma\,; x) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma}) \tag{1.8}$$

is a two parameter family of densities on R. $\mu \in R$ is called the location parameter and $\sigma > 0$ is called the scale parameter.

We now provide some examples of discrete distributions.

9. For any positive integer k, the Binomial family of probabilities on $X = \{0, 1, 2, \ldots, k\}$ are given by

$$p(\theta\,; x) = \binom{k}{x} \theta^x (1-\theta)^{k-x} \tag{1.9}$$

for $0 \leq \theta \leq 1$

10. The Poisson family on $X = \{0, 1, 2, \ldots\ldots\}$ has probabilities

$$p(\lambda\,; x) = \frac{e^{-\lambda} \lambda^x}{x!} \tag{1.10}$$

11. The uniform distributions

$$f(a, b\,; x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \tag{1.11}$$

is a two parameter family where the set of possible values of x depends on the parameters $-\infty < a < b < \infty$.

12. The uniform distributions

$$f(\theta\,; x) = \begin{cases} \frac{1}{\theta} & \text{for } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{1.12}$$

is a one parameter family where the set of possible values of x depends on the parameter $0 < \theta < \infty$. This is a subfamily of the previous example.

13. Finally, the family of multivariate normal distributions on R^d is parametrized by their mean $\mu \in R^d$ and covariance A, a symmetric positive definite $d \times d$ matrix. The densities are given by

$$f(\mu, A\,; x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{|A|^{\frac{1}{2}}} \exp[-\frac{1}{2} < (x-\mu), A^{-1}(x-\mu) >] \tag{1.13}$$

2 Decision Theory.

We have a family of models indexed by a parameter θ from the set Θ of all possible values of the parameter. The model has generated an observation and the model in particular specfies the probability distribution on the set Ω of observations ω, that depends on the parameter θ. There is a set D of possible decisions or actions d. How good or bad a decision d is, will depend on the true value of the parameter, which we do not know. The only information available, upon which we can base our decision, is the observation ω. A decision rule is then a function $d(\omega)$ that determines the decision as function of the observation. We have a loss function $L(\theta,\, d)$ that meausres how bad the decision d is, if the value of the parameter is θ. The expected loss, called Risk, is the function

$$R(d, \theta) = E_\theta[L(\theta,\, \omega)] \tag{2.1}$$

The object of the game is to look for $d \in D$ that minimizes $R(d, \theta)$. This is hard to do. Since for each $d \in D$, $R(d, \theta)$ is a function on Θ, to compare d_1 and d_2 we must compare $R(d_1, \theta)$ and $R(d_2, \theta)$ for all values of θ. A decision rule d is said to be *inadmissible* if there is a d' such that $R(d',\, \theta) \leq R(d,\, \theta)$ for all θ with strict inequality holding for atleast one θ. Anything that is not inadmissible is admissible. We could try to guard against the worst situation by trying

$$\inf_{d \in D} \sup_{\theta \in \Theta} R(d,\, \theta) \tag{2.2}$$

Or, if we believe in some prior distribution μ of possible values of the true parameter, we can average over the values of θ and try

$$\inf_{d \in D} \int_\Theta R(d,\, \theta) d\mu(\theta) \tag{2.3}$$

These are called respectively, the minimax and Bayes solutions.

3 Estimation.

A particular decision problem is to arrive at the true value of the parameter. In this case $D = \Theta$ and the decision rule $d(\omega)$ is really a map from $\Omega \to \Theta$. Such a map $\hat{\theta} = U(\omega)$ is called an estimator of θ. L is usually a measure of how far apart θ and $\hat{\theta}$ are. For example if $\Theta = R$ then $L(\theta,\, \hat{\theta})$ could be $|\theta - \hat{\theta}|^2$

We will look at the situation where we have n independent observations from a probability distribution P_θ specified either by the probabilities $p(\theta,\, x)$ in the discrete case, or by a density function $f(\theta,\, x)$ in the continuious case. The set X of possible values of the observation is either the integers or some similar finite or countable set (when the probabilities are specified as functions of θ), or the real line or d-dimensional Euclidean space (when the density is specified as a function of θ).

An estimator based on n observations is a function $U(x_1, \ldots, x_n)$ that is a fuction of $(x_1, \ldots, x_n)$ only. That is to say it can not depend on θ. Since

we will always assume that the observations are independent, the probability distribution for the n observations is the product probabilities or the product density as the case may be.

An estimator of $U(x_1, \dots, x_n)$ is called an *unbised estimator* of θ if

$$E_\theta[U(x_1, \dots, x_n)] \equiv \theta \tag{3.1}$$

Here we will use the notation E_θ to denote either

$$E_\theta[U(x_1, \dots, x_n)] = \sum_{x_1, \dots, x_n} U(x_1, \dots, x_n) p(\theta, x_1) \dots p(\theta, x_n)$$

or

$$E_\theta[U(x_1, \dots, x_n)] = \int_{R^n} U(x_1, \dots, x_n) f(\theta, x_1) \dots f(\theta, x_n) dx_1 \cdots dx_n$$

depending on the circumstance. Different unbiased estimators can be compared with the help of their variances, which is the risk corresponding to the loss function $|\theta - \hat{\theta}|^2$

It is useful to introduce the name of likelihood function. It is either $p(\theta, x)$ or $f(\theta, x)$ and is denoted by $L(\theta, x)$. For n obsrvations it is the product function. But the likelihood function will more often than not be viewed as a function of θ rather than a function of the observations.

The following result gives a lower bound on the variance of any unbiased estimator. Let $I(\theta)$ be defined by

$$I(\theta) = E_\theta\left[\left[\frac{\partial \log L(\theta, x)}{\partial \theta}\right]^2\right] \tag{3.2}$$

From the relation $\sum_x p(\theta, x) \equiv 1$ or $\int f(\theta, x) dx \equiv 1$, one can conclude easily that

$$E_\theta\left[\left[\frac{\partial \log L(\theta, x)}{\partial \theta}\right]\right] \equiv 0 \tag{3.3}$$

From independence, for the likelihood function $L(\theta, x_1 \dots, x_n)$ based on n observations, we can conclude from equations (3.2) and (3.3) that

$$I_n(\theta) = E_\theta\left[\left[\frac{\partial \log L(\theta, x_1, \dots, x_n)}{\partial \theta}\right]^2\right] = nI(\theta) \tag{3.4}$$

If U is unbiased, by differentiating equation (3), we get

$$E_\theta[U(x_1, \dots, x_n) \frac{\partial \log L(\theta, x_1 \dots, x_n)}{\partial \theta}] \equiv 1$$

which, because of equation (3.3) can be rewritten as

$$E_\theta\big[[U(x_1, \dots, x_n) - \theta] \frac{\partial \log L(\theta, x_1 \dots, x_n)}{\partial \theta}\big] \equiv 1 \tag{3.5}$$

We use Schwartz's inequality to derive from equations (3.4) and (3.5), the *Cramér-Rao Lower Bound*

$$E_\theta\left[\left[U(x_1,\ldots,x_n)\right]^2\right] \geq \frac{1}{nI(\theta)} \tag{3.6}$$

In particular if we find an unbiased estimator U that matches the Cramér-Rao Lower Bound, it is hard to beat and it is called a MVUB estimator (for minimum variance unbiased estimator)

Examples.

1. The family is the Normal family given by 1.1. The sample mean $U(x_1,\ldots,x_n) = \frac{x_1+\ldots+x_n}{n}$ is clearly an unbised estimator of μ. Its variance is $\frac{1}{n}$. The information function $I(\mu)$ is calculated easily. $\frac{\partial \log L}{\partial \mu} = (x-\mu)$ and $I(\mu) \equiv 1$ and the Cramér-Rao Lower Bound is attained.

2. Let us look at the Binomial example (1.9). $E_\theta[\frac{x}{k}] = \theta$ and the variance of $\frac{x}{k}$ is equal to $\frac{\theta(1-\theta)}{k}$. The likelihood function is given by

$$L(\theta) = \log\binom{k}{x} + x\log\theta + (k-x)\log(1-\theta)$$

$$\frac{\partial \log L}{\partial \theta} = \frac{x}{\theta} - \frac{(k-x)}{1-\theta} = \frac{x-k\theta}{\theta(1-\theta)}$$

   and the information function is calculated to be $\frac{k}{\theta(1-\theta)}$.

If we examine a little closely when equality can hold in Schwartz's inequality, it is clear that for the Cramér-Rao lower bound to be attained, we must have

$$\frac{\partial \log L}{\partial \theta} = k(\theta)(U(x_1,\ldots,x_n) - \theta)$$

which leads to the form

$$\log L = a(\theta)U(x_1,\ldots,x_n) + b(\theta) + W(x_1,\ldots,x_n)$$

If the the likelihood function of a single observation looks like

$$\log L = a(\theta)V(x) + b(\theta) + W(x)$$

then

$$U(x_1,\ldots,x_n) = \sum_{i=1}^{n} V(x_i)$$

is a MVUB estimator of its expectation. We can reparametrize and put the likelihood in a more canonical form

$$L = b(\theta)W(x)\exp[\theta V(x)] \tag{3.7}$$

It is a simple calculation that

$$E_\theta[V(x)] = \frac{b'(\theta)}{b(\theta)}$$

We now turn our attention to the example given by (1.12). For any given x, the likelihood function has a discontinuity at $\theta = x$. It is zero if $\theta < x$ and equals $\frac{1}{\theta}$ for $\theta > x$. Now all bets are off. One can see that with $U(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n)$

$$P_\theta[U \leq x] = \frac{x^n}{\theta^n}$$

One can verify that $V = \frac{n+1}{n}U$ is unbiased and has a variance that behaves like $\frac{C}{n^2}$, much smaller than any Cramér-Rao lower bound.

4 Consistency.

What distinguishes a good estimator from a bad one? Ideally, as we obtain more observations, we have more information and our estiamtor should become more accurate. This is not a statement about a single estiamator, but one about a sequence $U_n(x_1, \dots, x_n)$ of estimators.

The sequence $U_n(x_1, \dots, x_n)$ is said to be a *consistent* estiamtor of θ if, for every $\delta > 0$,

$$\lim_{n\to\infty} P_\theta\big[(x_1, \dots, x_n) : |U_n(x_1, \dots, x_n) - \theta| \geq \delta\big] = 0 \tag{4.1}$$

Let us denote by $m_n(\theta)$ and $\sigma_n^2(\theta)$ the mean and variance

$$m_n(\theta) = E_\theta\Big[U_n(x_1, \dots, x_n)\Big]$$
$$\sigma_n^2(\theta) = E_\theta\Big[\big[U_n(x_1, \dots, x_n) - m_n(\theta)\big]^2\Big]$$

We can be sure of consistency provided $\lim_{n\to\infty} m_n(\theta) = \theta$ and $\lim_{n\to\infty} \sigma_n^2(\theta) = 0$. If for some k we have an unbiased estimator $V_k(x_1, \dots, x_k)$ we can construct a consistent sequence of estimators by taking

$$U_n(x_1, \cdots, x_n) = \frac{1}{\ell}\sum_{j=1}^{\ell} V_k(x_{(j-1)k+1}, \dots, x_{jk})$$

for $\ell k \leq n < (\ell+1)k$. The law of large numbers will guarantee the validity of (4.1) even if the variance does not exist.

One way to generate consistent estimators is to find a function $U(x)$ of a single observation such that $E_\theta[U(x)] = f(\theta)$ which is a one-to-one function of θ with a continuous inverse $g(\theta) = f^{-1}(\theta)$. Then it is not hard to see that $g(\frac{1}{n}\sum_{i=1}^n U(x_i))$ is consistent for θ.

Let us apply this to the two parameter normal family. $E_{\mu,\theta}[x] = \mu$ and $E_{\mu,\theta}[x^2] = \mu^2 + \theta$. Clearly $\bar{x}_n = \frac{x_1+\dots+x_n}{n}$ and $t_n = \frac{x_1^2+\dots+x_n^2}{n}$ are consistent for μ and $\mu^2 + \theta$. Then $\bar{x}_n$ and $U_n = t_n - \bar{x}_n^2$ are consistent for μ and θ. Such an approach of using x^k successively with $k = 1, 2, \dots$ is called the method of moments. We need as many moments as there are parameters in order to successfully invert.

5 Sufficiency.

If we have data or observations $(x_1, \dots, x_n)$ generated from a distribution P_θ it is conceivable that we might throw away some information with out any real loss, especially if the information can be recreated. Any information that can be recreated must have been worthless to begin with! If the probability distribution of an observation is known exactly, with out any unknown parameters, then

such an observation can be created at any time by simulation and is irrelevent. Although it seems that we are only stating the obvious, there is more here than meets the eye.

Let us look at two scenarios. There are n identical coins all having the same probability p of coming up head. In the first scenario they are tossed one after another and the sequence of outcomes is recorded as word of length n each letter being H or T. In the second scenario they are tossed at the same time and only the total number of heads is noted. It is clear that the first scenario offers more information. We can always count the number of heads, but cannot tell the sequence from the total number. But if the total number of heads is k, there are $\binom{n}{k}$ sequences with the same number k of heads and they all have the same probability $p^k(1-p)^{n-k}$. Conditionally, given k, they all have the probability $\frac{1}{\binom{n}{k}}$. Therefore if we know k, the order can be randomly chosen, without knowing p, and no one can say if the sequence so generated is different, statistically speaking, from the one generated by tossing the coins one after another. Since the order can be recreated it could not have been relevent. What ever any one can do from the full sequence we can accomplish it with just knowing the total number of heads. So even if some one presented us with the full sequence we must just count the total number of heads and forget the rest. This is the philosophy behind sufficiency.

We have n independent observations $(x_1, x_2, \dots, x_n)$ from a population P_θ with an unknown parameter $\theta \in \Theta$. A collection $U_1, \dots, U_k$ of k functions of $(x_1, x_2, \dots, x_n)$, are said to be jointly sufficient for θ if the conditional probabilities

$$P_\theta\Big[(x_1, x_2, \dots, x_n) \in A | U_1, \dots, U_k\Big] = p(U_1, \dots, U_k\,; A) \tag{5.1}$$

are independent of $\theta \in \Theta$. Of course θ can be a set of m parameters and k can be any number, usually m, sometimes more than m, but rarely less than m. It is clear that what is relevent is the σ-field generated by $U_1, \dots, U_k$ which should be thought of as the information contaned in $U_1, \dots, U_k$. We will look at some examples. First we make a general observation.

If the likelihood function $L(\theta,\, x_1, \dots, x_n)$ factors into

$$L(\theta,\, x_1, \dots, x_n) = k(\theta) G(\theta,\, U_1, \dots, U_k) H(x_1, \dots, x_n)$$

then in the discrete case

$$\begin{aligned} &P_\theta\Big[(x_1, x_2, \dots, x_n) \in A | U_1 = u_1, \dots, U_k = u_k\Big] \\ &= \frac{\sum_{A \cap \{x : U_1 = u_1, \dots, U_k = u_k\}} L(\theta,\, x_1, \dots, x_n)}{\sum_{\{x : U_1 = u_1, \dots, U_k = u_k\}} L(\theta,\, x_1, \dots, x_n)} \\ &= \frac{\sum_{A \cap \{x : U_1 = u_1, \dots, U_k = u_k\}} H(x_1, \dots, x_n)}{\sum_{\{x : U_1 = u_1, \dots, U_k = u_k\}} H(x_1, \dots, x_n)} \end{aligned}$$

which is independent of θ proving sufficiency. In the continuous case the summation has to be replaced by integration. There are some technical details, but the end result is the same.

Examples.

1. If we have n observations from the binomial population (1.9) with parameter p, the likelihood function is given by

$$L(p,\, x_1, \dots, x_n) = \Pi_{I=1}^{n} \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}$$

   The function $U(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ is seen to be sufficient.

2. If we have n observations from the Poisson population (1.10) with parameter λ, the likelihood function is given by

$$L(\lambda,\, x_1, \dots, x_n) = \Pi_{i=1}^{n} e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$

   The function $U(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ is again seen to be sufficient.

3. If we have n observations from the Normal population (1.1) with mean μ and variance 1, the likelihood function is given by

$$\begin{aligned} L(\mu,\, x_1, \dots, x_n) &= \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp[-\frac{(x_i-\mu)^2}{2}] \\ &= [\frac{1}{\sqrt{2\pi}}]^n \exp[-\frac{\sum_i x_i^2}{2}] \exp[-\frac{n\mu^2}{2}] \exp[\mu \sum_i x_i] \end{aligned}$$

   The function $U(x_1, \dots, x_n) = \sum_{i=1}^{n} x_i$ is again seen to be sufficient.

4. A similar calculation with the two parameter Normal family (1.2) with mean μ and variance θ shows the sufficiency of $(\sum_i x_i\ , \sum_i x_i^2)$.

5. A calculation with the two parameter Gamma family (1.3) with parametrs α and variance p shows the sufficiency of $(\sum_i x_i\ , \Pi_i x_i)$.

6. Finally, for the example (1.12) of uniform distribution on $[0, \theta]$ the likelihood function takes the form

$$L(\theta,\, x_1, \dots, x_n) = \Pi_{i=1}^{n} \mathbf{1}_{[\theta \geq x_i]} \frac{1}{\theta} = \frac{1}{\theta^n} \mathbf{1}_{[\theta \geq \max(x_1, \dots, x_n)]}$$

   proving the sufficiency of $U(x_1, \dots, x_n) = \max(x_1, \dots, x_n)$.

If we are given an unbiased estimator $W(x_1, \dots, x_n)$ and there is a sufficient statistic $U(x_1, \dots, x_n)$ around we can replace W by a new estimator

$$\widehat{W}(U) = E_\theta\big[W(x_1, \dots, x_n)|U\big] \tag{5.2}$$

The properties of conditional expectation assures us that

$$E_\theta\big[\widehat{W}(U)\big] \equiv E_\theta\big[W(x_1,\dots,x_n)\big] \equiv \theta$$

and

$$Var_\theta\big[\widehat{W}(U)\big] \le Var_\theta\big[W(x_1,\dots,x_n)\big]$$

We do better by replacing W with $\widehat{W}$. Sufficiency of the estimator is crucial and ensures that $\widehat{W}(U)$ depends only on the observations and is independent of the unknown parameter θ. A set $(U_1, U_2, \dots, U_k)$ that is sufficient for a family P_θ is said to be complete if whenever

$$E_\theta\big[F(U_1, U_2, \dots, U_k)\big] \equiv 0$$

for all $\theta \in \Theta$, it follows that $F \equiv 0$. It is a simple argument to show that if $(U_1, U_2, \dots, U_k)$ is sufficient and complete for any family, then any function $F(U_1, U_2, \dots, U_k)$ is a MVUB estimator of its expectation

$$E_\theta\big[F(U_1, U_2, \dots, U_k)\big] = f(\theta)$$

If W is any unbiased estimator of $f(\theta)$ by sufficiency it can be replaced by $\widehat{W}$ which is better no matter what θ is. On the other hand $\widehat{W}$ and F are two unbiased estimators of the same $f(\theta)$ and the diffrence $\widehat{W} - F$ has mean 0 for all θ. Since it is a function of $U_1, U_2, \dots, U_k$ that are complete and sufficient $\widehat{W} - F$ must be 0. Therefore F which is the same as $\widehat{W}$ must be better than W for all θ. Since W was arbitrary F is a MVUB estimator.

A minimum variance unbiased estimator, if it exists is unique. Suppose U_1 and U_2 are both MVUB estimators with variance $D(\theta)$ for the same parameter, then $U = \frac{U_1+U_2}{2}$ is an unbiased estimator as well. If we denote by $W = \frac{U_1-U_2}{2}$ the identity

$$\mathrm{Var}(U) + \mathrm{Var}(W) = \frac{\mathrm{Var}(U_1) + \mathrm{Var}(U_2)}{2} = D(\theta) \ge \mathrm{Var}(U)$$

forces $\mathrm{Var}(W)$ to be 0.

Examples.

1. In the case of the normal family (1.1) with mean μ and variance 1, $\bar{x}$ is sufficient. Its distribution is given by the density

$$f_n(\mu, \bar{x}) = \frac{\sqrt{n}}{\sqrt{2\pi}} \exp[-\frac{n(\bar{x}-\mu)^2}{2}]$$

   If $F(\bar{x})$ has mean 0 for all then

$$\int_{-\infty}^{\infty} F(y) \exp[-\frac{ny^2}{2}] \exp[n\mu y] dy \equiv 0 \tag{5.3}$$

   From the uniqueness theorem of Laplace trnasforms it follows that $F \equiv 0$.

2. For the family of uniform distribution (1.12) on $[0, \theta]$ we saw that the sufficient statistic $U = \max(x_1, \ldots, x_n)$ has the distribution

$$P_\theta\big[U \leq x\big] = \frac{x^n}{\theta^n}$$

Completeness is just a consequence of the fact, that if

$$\frac{n}{\theta^n}\int_0^\theta F(U)U^{n-1}dU \equiv 0$$

then $F(U) \equiv 0$. We conclude that $\frac{(n+1)U}{n}$ is the MVUB estiamtor of θ. Since $E_\theta[x_1] = \frac{\theta}{2}$, $2x_1$ is an unbiased estimator. Do you see why

$$E_\theta\big[2x_1|U\big] = \frac{(n+1)U}{n} \ ?$$

6 Maximum Likelihood Estimators.

Perhaps the most important aspect of parametric estimation is the method of maximum likelihood. It is a method that has three attractive qualities. It is simple to state, if not always to carry out. It is universal. It is asymptotically optimal. Given the likelihood function $L(\theta,\, x_1, \ldots, x_n)$ the maximum likelihood estimator $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$ is the value of θ that maximizes the likelihood function $L(\theta,\, x_1, \ldots, x_n)$ as a function of θ for the given set $(x_1, \ldots, x_n)$ of observations. In other words

$$\hat{\theta}(x_1, \ldots, x_n) = \text{argmax}_\theta L(\theta,\, x_1, \ldots, x_n) \tag{6.1}$$

As we can see this is a simple and universal recipe. Often, in practice, the MLE is obtained by solving, in the one parameter case the likelihood equation

$$\frac{\partial \log L}{\partial \theta} = \sum_i \frac{\partial \log L(\theta,\, x_i)}{\partial \theta} = 0 \tag{6.2}$$

or more genrally, in the multi parameter case, by solving the set of k simultaneous equations

$$\frac{\partial \log L}{\partial \theta_j} = \sum_i \frac{\partial \log L(\theta,\, x_i)}{\partial \theta_j} = 0 \tag{6.3}$$

for $j = 1, \ldots, k$ to obtain $\{\hat{\theta}_j(x_1, \ldots, x_n)\}$. Let us look at a few examples.

Examples.

1. As usual we start with the Normal family (1.2).

$$\log L = -\frac{n}{2}[\log 2\pi] - \frac{n}{2}\log\theta - \frac{1}{2\theta}\sum_i (x_i - \mu)^2$$

Solving the likelihood equations yield $\hat{\mu} = \bar{x} = \frac{\sum_i x_i}{n}$ and $\hat{\theta} = \frac{\sum_i (x_i - \bar{x})^2}{n}$, i.e. the sample mean and the sample variance as the two estimators for μ and θ.

2. For the family of uniform distributions on $[0, \theta]$ (1.12) the likelihood function is maximized by the smallest admissible value of θ namely $\max(x_1, \ldots, x_n)$ which is then the MLE.

3. For the family of Cauchy distributions (1.5) the likelihood equation is

$$\sum_i \frac{x_i - \theta}{1 + (x_i - \theta)^2} = 0$$

which is a mess and cannot be solved explcitly. The method still works, but the argmax can only be numerically computed. We will later see some ways of handling this.

7 Consistency of the MLE.

Under fairly general conditions one can prove that the MLE provides a consistent estimate as $n \to \infty$. A fairly simple and elegant proof can be given under somewhat stronger assumptions than what is necessary. We will do that now. If we have two probability distributions $p_1(x)$ and $p_2(x)$ on a countable set X of points x, the Kullback-Leibler information number is defined by

$$\begin{aligned} H(p_2(\cdot)|p_1(\cdot)) &= \sum_x p_2(x) \log \frac{p_2(x)}{p_1(x)} \\ &= \sum_x \big[\frac{p_2(x)}{p_1(x)} \log \frac{p_2(x)}{p_1(x)}\big] p_1(x) \\ &= \sum_x \big[\frac{p_2(x)}{p_1(x)} \log \frac{p_2(x)}{p_1(x)} - \frac{p_2(x)}{p_1(x)} + 1\big] p_1(x) \\ &= \sum_x \Phi\big(\frac{p_2(x)}{p_1(x)}\big) p_1(x) \end{aligned} \tag{7.1}$$

where $\Phi(u) = u \log u - u + 1$ is nonnegative and strictly positive if $u \neq 1$. We have assumed here that $p_2(x)$ is not positive when $p_1(x) = 0$ and used the fact that $\sum_x \frac{p_2(x)}{p_1(x)} p_1(x) = \sum_x p_2(x) = 1 = \sum_x p_1(x)$. A similar formula holds when

the distributions are given by densities

$$\begin{aligned} H(f_2(\cdot)|f_1(\cdot)) &= \int f_2(x)\log\frac{f_2(x)}{f_1(x)}dx \\ &= \int \big[\frac{f_2(x)}{f_1(x)}\log\frac{f_2(x)}{f_1(x)}\big]f_1(x)dx \\ &= \int \big[\frac{f_2(x)}{f_1(x)}\log\frac{f_2(x)}{f_1(x)} - \frac{f_2(x)}{f_1(x)} + 1\big]f_1(x)dx \\ &= \int \Phi\big(\frac{f_2(x)}{f_1(x)}\big)f_1(x)dx \end{aligned} \tag{7.2}$$

If $f_2(x)$ or $p_2(x)$ is positive when $f_1(x)$ or $p_1(x)$ is 0, the natural definition for $H(\cdot\,|\,\cdot)$ is $+\infty$. For a family of densities or probabilities depending on a parameter θ let us define

$$H(\theta_1|\theta_2) = \begin{cases} H(f(\theta_1,\,\cdot)|f(\theta_2,\,\cdot)) & \text{continuous case} \\ H(p(\theta_1,\,\cdot)|p(\theta_2,\,\cdot)) & \text{discrete case} \end{cases} \tag{7.3}$$

Let us suppose that the Kullback-Leibler function is well defined for all $\theta_1, \theta_2 \in \Theta$. Then $H \geq 0$ and let us suppose that it is 0 only on the diagonal $\theta_1 = \theta_2$. This is the assumption of identifiability of the parameter. We will now make the additional assumption that Θ is a closed bounded subset of R^k for some k and that the log-likelihood function $\log L$ is a continuous function of $\theta \in \Theta$. If we denote by θ_0 the true unknown value of the parameter, the expectation

$$F_0(\theta) = E_{\theta_0}[\log L(\theta,\, x_i)]$$

has the property that $F_0(\theta_0) \geq F_0(\theta)$ for all $\theta \in \Theta$ with equality hollding only when $\theta = \theta_0$. Note that the difference $F_0(\theta_0) - F_0(\theta)$ is equal to $H(\theta_0|\theta_1)$. The function $\frac{1}{n}\log L(\theta,\, x_1, \ldots, x_n)$ is the average of n random functions $\log L(\theta,\, x_i)$ and should converge by the strong law of large numbers to its expectation $F_0(\theta)$. Since $F_0(\theta)$ has a unique maximum at $\theta = \theta_0$ the approximating functions will have all their global maxima close to θ_0 with probability close to 1. Therefore the argmax of L is close to θ_0 with probability nearly one. This proves the consistency of the MLE. The technical assumption that will make this proof work is that $\log L(\theta,\, x_i)$ is continuous in θ for every x, and for each $\theta_0 \in \Theta$,

$$E_{\theta_0}\big[\sup_{\theta\in\Theta}|\log L(\theta,\, x_i)|\big] < \infty \tag{7.4}$$

In practice the parameter space is all of R^k or if it is a subset, most often it is not closed. If there exists an increasing family Θ_λ of closed bounded subsets of Θ, that exhausts Θ as $\lambda \to \infty$ satisfying

$$\lim_{\lambda\to\infty} E_{\theta_0}\big[\sup_{\theta\in\Theta_\lambda^c}\log L(\theta,\, x)\big] = -\infty \tag{7.5}$$

equation (7.4) can be replaced by the weaker condition,

$$E_{\theta_0}\big[\sup_{\theta\in\Theta_\lambda} |\log L(\theta, x_i)|\big] < \infty \tag{7.6}$$

for every $\theta_0 \in \Theta$ and $\lambda < \infty$.

The proof proceeds roughly along the following lines. We have the random function

$$X_n(\theta, x_1, \dots, x_n) = \frac{1}{n}\sum_i \log L(\theta, x_i)$$

defined on Θ. It converges uniformly on every Θ_λ to $F_0(\theta)$ which has a unique maximum at θ_0. If we can rule out getting stuck with a maximum in Θ_k^c for some (large) k, then the MLE has to be close to θ_0. This means showing with P_{θ_0} probability tending to 1,

$$\sup_{\theta\in\Theta_k^c} X_n(\theta, x_1 \dots, x_n) < F_0(\theta_0) - 1$$

for some large but fixed k. Since

$$\sup_{\theta\in\Theta_k^c} X_n(\theta, x_1 \dots, x_n) \le \frac{1}{n}\sum_i \sup_{\theta\in\Theta_k^c} L(\theta, x_i)$$

and the law of large numbers assures the convergence of the right hand side to $E_{\theta_0}\big[\sup_{\theta\in\Theta_k^c} L(\theta, x_i)\big]$, that is close to $-\infty$ far large k, we are done.

Let us apply this to the example of the Gamma family (1.3) restricted to $\alpha = 1$.

$$\log L(p, x) = -\log\Gamma(p) - x + (p-1)\log x$$

It is clear that (7.6) is valid. $E_{p_O}[x]$ is a finite constant. $\Gamma(p) \simeq p^{-1}$ for small p and $\log\Gamma(p) \simeq c - (p-1) + (p-\frac{1}{2})\log(p-1)$ for large p, by Stirling's formula.

The calculation of

$$f(a, x) = \sup_{p\ge a}[p(1+\log x) - p\log p]$$

yields

$$f(a,x) = \begin{cases} a & \text{if } x \ge a \\ a(1+\log x) - a\log a & \text{if } x \le a \end{cases}$$

It is easily seen that $E_{p_0}[f(x,a)] \to -\infty$ as $a \to \infty$.

8 Asymptotic Normality of MLE.

We saw in the last section that under some regularity assumption on the model, any MLE is consistent. In principle MLE need not be unique. For instance for

the two sided exponential family, (1.7) the MLE is not unique if the number n of observations is even. They are all consistent in that if for some $(x_1, \ldots, x_n)$ there are more than 1 MLE, they are very close to each other.

The next step is to investigate what the asymptotic distribution of the diffrence $\hat{\theta}_n - \theta_0$ is. Under more regularity assumptions we will show that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normally distributed with a limiting normal distribution with mean 0 and variance $\frac{1}{I(\theta_0)}$. In some sense asymptotically, the Cramér-Rao lower bound is achieved.

The proof is surprisingly simple. We need to assume that the maximum of the likelihood function is attained at an interior point so that the estimator $\hat{\theta}_n$ is a solution of the likelihood equation (6.2).

$$\sum_i \frac{\partial \log L}{\partial \theta}(\hat{\theta}_n,\, x_i) = 0$$

If denote by

$$Y_n(\theta,\, x_1, \ldots, x_n) = \sum_i \frac{\partial \log L}{\partial \theta}(\theta,\, x_i)$$

then by the mean value theorem

$$Y_n(\hat{\theta}_n,\, x_1, \ldots, x_n) - Y_n(\theta_0,\, x_1, \ldots, x_n) = (\hat{\theta}_n - \theta_0) \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i) \quad (8.1)$$

where $\tilde{\theta}_n$ is some value of θ near $\hat{\theta}_n$ and θ_0. Under P_{θ_0}, by consistency, both $\hat{\theta}_0$ and $\tilde{\theta}_n$ are close to the true value θ_0. We rewrite equation (8.1) as

$$-\frac{1}{\sqrt{n}} Y_n(\theta_0,\, x_1, \ldots, x_n) = [\sqrt{n}(\hat{\theta}_n - \theta_0)][\frac{1}{n} \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i)]$$

By the central limit theorem, he distribution of left hand side converges to a Normal distribution with mean zero and variance $I(\theta_0)$. By the law of large numbers $[\frac{1}{n} \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i)]$ converges to the constant $-I(\theta_0)$. Therefore $[\sqrt{n}(\hat{\theta}_n - \theta_0)]$ has a limiting distribution which is normal with mean zero and variance $\frac{1}{I(\theta_0)}$. We have assumed here that $\frac{\partial^2 \log L}{\partial \theta^2}$ exists and is continuous. We have also made the calculation

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{1}{L} \frac{\partial^2 L}{\partial \theta^2} - \frac{1}{L^2} [\frac{\partial L}{\partial \theta}]^2$$

and taking expectaions

$$E_{\theta_0}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = E_{\theta_0}\left[\frac{1}{L} \frac{\partial^2 L}{\partial \theta^2}\right] - E_{\theta_0}\left[\frac{1}{L^2} [\frac{\partial L}{\partial \theta}]^2\right] = -I(\theta_0)$$

9 Asymptotic Normality of MLE.

We saw in the last section that under some regularity assumption on the model, any MLE is consistent. In principle MLE need not be unique. For instance for the two sided exponential family, (1.7) the MLE is not unique if the number n of observations is even. They are all consistent in that if for some $(x_1, \dots, x_n)$ there is more than one MLE, they are very close to each other.

The next step is to investigate what the asymptotic distribution of the difference $\hat{\theta}_n - \theta_0$ is. Under more regularity assumptions we will show that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normally distributed with a limiting normal distribution with mean 0 and variance $\frac{1}{I(\theta_0)}$. In some sense, asymptotically, the Cramér-Rao lower bound is achieved.

The proof is surprisingly simple. We need to assume that the maximum of the likelihood function is attained at an interior point so that the estimator $\hat{\theta}_n$ is a solution of the likelihood equation (6.2).

$$\sum_i \frac{\partial \log L}{\partial \theta}(\hat{\theta}_n,\, x_i) = 0$$

If we denote by

$$Y_n(\theta,\, x_1, \dots, x_n) = \sum_i \frac{\partial \log L}{\partial \theta}(\theta,\, x_i)$$

then by the mean value theorem

$$Y_n(\hat{\theta}_n,\, x_1, \dots, x_n) - Y_n(\theta_0,\, x_1, \dots, x_n) = (\hat{\theta}_n - \theta_0) \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i) \quad (9.1)$$

where $\tilde{\theta}_n$ is some value of θ near $\hat{\theta}_n$ and θ_0. Under P_{θ_0}, by consistency, both $\hat{\theta}_0$ and $\tilde{\theta}_n$ are close to the true value θ_0. We rewrite equation (9.1) as

$$-\frac{1}{\sqrt{n}} Y_n(\theta_0,\, x_1, \dots, x_n) = [\sqrt{n}(\hat{\theta}_n - \theta_0)][\frac{1}{n} \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i)]$$

By the central limit theorem, the distribution of the left hand side converges to a Normal distribution with mean zero and variance $I(\theta_0)$. By the law of large numbers $[\frac{1}{n} \sum_i \frac{\partial^2 \log L}{\partial \theta^2}(\tilde{\theta}_n,\, x_i)]$ converges to the constant $-I(\theta_0)$. Therefore $[\sqrt{n}(\hat{\theta}_n - \theta_0)]$ has a limiting distribution which is normal with mean zero and variance $\frac{1}{I(\theta_0)}$. We have assumed here that $\frac{\partial^2 \log L}{\partial \theta^2}$ exists and is continuous. We have also made the calculation

$$\frac{\partial^2 \log L}{\partial \theta^2} = \frac{1}{L} \frac{\partial^2 L}{\partial \theta^2} - \frac{1}{L^2} [\frac{\partial L}{\partial \theta}]^2$$

and taking expectaions

$$E_{\theta_0}\left[\frac{\partial^2 \log L}{\partial \theta^2}\right] = E_{\theta_0}\left[\frac{1}{L} \frac{\partial^2 L}{\partial \theta^2}\right] - E_{\theta_0}\left[\frac{1}{L^2} [\frac{\partial L}{\partial \theta}]^2\right] = -I(\theta_0)$$

10 Efficiency.

It is possible to construct an estimate that does better than the Cramér-Rao lower bound at one point. Of course it cannot be unbiased. For instance, if we are estimating any parameter θ the estimator $U(x_1, \ldots, x_n) \equiv c$ is obviously excellent if the true value of θ is c, in which case the variance is 0. One could make the case that this estimate is not only biased, but hopelessly so, and is in fact not even consistent. We would like to have a theorem of the form if U_n is an estimator that is consistent for all values of θ and if $\sqrt{n}(U_n - \theta)$ has an asymptotic distribution F, then the variance

$$\sigma^2(F) = \int x^2 dF - (\int x dF)^2 \geq \frac{1}{I(\theta)} \tag{10.1}$$

It is easy to demonstrate that this is false. Look at the estimation of the unknown mean θ of a normal distribution with variance 1. The standard estimate is $\bar{x}_n$, the sample mean. Let us try to improve it at one point, say $\theta = 0$ with out making it worse elsewhere. We define $U_n = \bar{x}_n$ if $|\bar{x}_n| \geq n^{-\frac{1}{4}}$ and $U_n = 0$ if $|\bar{x}_n| < n^{-\frac{1}{4}}$. Clearly if $\theta = 0$,

$$E[U_n] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} y e^{-\frac{ny^2}{2}} dy = 0$$
$$Var[U_n] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} y^2 e^{-\frac{ny^2}{2}} dy = o(\frac{1}{n})$$

If $\theta \neq 0$,

$$E[U_n] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} y e^{-\frac{n(y-\theta)^2}{2}} dy = \theta + o(\frac{1}{n})$$
$$E[(U_n - \theta)^2] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} (y-\theta)^2 e^{-\frac{n(y-\theta)^2}{2}} dy = \frac{1}{n} + o(\frac{1}{n})$$

We seem to have gained something at 0 without losing anything anywhere else. But there is no free lunch. Suppose, $\theta = \frac{a}{n}$, which is a possibility that is real, because as n is large but finite, θ can be nonzero but small. In such a case

$$E[U_n] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} y e^{-\frac{n(y-\frac{a}{\sqrt{n}})^2}{2}} dy = o(\frac{1}{n})$$
$$E[(U_n)^2] = \frac{1}{\sqrt{2\pi n}} \int_{|y| \geq n^{-\frac{1}{4}}} y^2 e^{-\frac{n(y-\frac{a}{\sqrt{n}})^2}{2}} dy = o(\frac{1}{n})$$
$$E[(U_n - \theta)^2] = \frac{a^2}{n} + o(\frac{1}{n})$$

For $a >> 1$ there is a considerable loss. Let us now formulate a correct theorem.

Theorem 10.1. *Let U_n be an estimator such that for some θ_0 and for every sequence $\theta_n \to \theta_0$,*

$$\lim_{n \to \infty} P_{\theta_n}[\sqrt{n}(U_n - \theta_n) \leq x] = F(x)$$

exists. Then

$$\int x^2 dF(x) \geq \sigma^2(F) \geq \frac{1}{I(\theta_0)}$$

Proof. We will give a proof under the additional assumption that F has a nice density f although it is not needed.

$$\lim_{n\to\infty} P_{\theta_0+\frac{a}{\sqrt{n}}}[\sqrt{n}(U_n - \theta_n - \frac{a}{\sqrt{n}}) \leq x] = F(x)$$

or

$$\lim_{n\to\infty} P_{\theta_0+\frac{a}{\sqrt{n}}}[\sqrt{n}(U_n - \theta_n - \frac{a}{\sqrt{n}}) \leq x + a] = F(x)$$

and therefore

$$\lim_{n\to\infty} P_{\theta_0+\frac{a}{\sqrt{n}}}[\sqrt{n}(U_n - \theta_n - \frac{a}{\sqrt{n}}) \leq x] = F(x - a)$$

It follows from Jensen's inequality that

$$\int \log \frac{f_n(u)}{f_n(u-a)} f_n(u)du \leq n \int \log \frac{L_n(x)}{L_n(\theta_0 + \frac{a}{\sqrt{n}}, x)} L_n(x)dx$$

Where f_n is the density of $\sqrt{n}(U_n - \theta_n)$ under P_{θ_n}. From the lower semi continuity of the Kullback-Lebler information number and the smoothness of the log-likelihood function we conclude that

$$\int \log \frac{f(u)}{f(u-a)} f(u)du \leq \frac{1}{2}a^2 I(\theta_0)$$

Dividing by a^2 and passing to the limit yields

$$I = \int \frac{(f'(u))^2}{f(u)} du \leq I(\theta_0)$$

For the estimation of the location parameter a in the family $\{f(x-a)\}$, the information function is the constant I, and in this context $x - \int uf(u)du$ is an unbiased estimator based on a sample of size 1. Its variance is $\sigma^2(F)$ and by the Cramér- Rao lower bound

$$\sigma^2(F) \geq \frac{1}{I} \geq \frac{1}{I(\theta_0)}$$

□

11 Order Statistics.

If we have n observations $x_1, \ldots, x_n$ and arrange them in increasing order $x_{(1)} \leq \ldots \leq x_{(n)}$, then $\{x_{(i)} : 1 \leq i \leq n\}$ are called the order statistics. They are

clearly sufficient set of statistics. The reduction of data is not much. The only information that is lost are the original labels. A function of the original observations is a function of the order statistics if and only if it is a symmetric function of the n observations. $x_{(k)}$ is called the k-th order statistic. The median is defined as $x_{(\frac{n+1}{2})}$ if n is odd. If n is even, it is a bit ambiguous, any value between $x_{(\frac{n}{2})}$ and $x_{(\frac{n+2}{2})}$ being reasonable. In practice one often takes it as $\frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n+2}{2})})$.

Let us suppose that our observations are drawn from the uniform distribution in the interval $[0, 1]$. The distribution function of the k-th order statistic from a sample of size n is not hard to evaluate. $x_{(k)} \leq x$ if and only if there are atleast k observations that are less than or equal to x. Its distribution function is therefore given by

$$F_{k,n}(x) = \sum_{j \geq k} \binom{n}{j} x^j (1-x)^{n-j}$$

and its density by

$$\begin{aligned} f_{k,n}(x) = \frac{dF_{k,n}(x)}{dx} &= \sum_{j \geq k} \binom{n}{j} [j x^{j-1}(1-x)^{n-j} - (n-j)(1-x)^{n-j-1}] \\ &= n \binom{n-1}{k-1} x^{k-1}(1-x)^{n-k} \end{aligned}$$

Let us suppose that $k = k_n$ is such that $\frac{k_n}{n} \to p$ where $0 < p < 1$. We are interested in the asymptotic behavior of the density $\frac{1}{\sqrt{n}} f_{k_n,n}(p + \frac{y}{\sqrt{n}})$ of $y = \sqrt{n}(x_{(k_n)} - p)$. Using Stirling's approximation it is not hard to see that

$$\lim_{\substack{n \to \infty \\ \frac{k_n}{n} \to p}} \sqrt{n} \binom{n-1}{k-1} (p + \frac{y}{n})^{k_n - 1} (1 - p - \frac{y}{n})^{n - k_n} = \frac{1}{\sqrt{2\pi p(1-p)}} e^{-\frac{y^2}{2p(1-p)}}$$

proving that the normalized order statistic $y = \sqrt{n}(x_{(k_n)} - p)$ has asymptotically a normal distribution with mean 0 and variance $p(1-p)$ provided $\frac{k_n}{n} \to p$. In particular for the median $p = \frac{1}{2}$ and $p(1-p) = \frac{1}{4}$.

If $F(x)$ is a distribution function the solution $x = x_p$ of $F(x) = p$ is called the p-th quantile. If F is continuous and strictly increasing it exists and is unique. If F has density $f(x)$ that is positive, then all the quantiles are well defined.

Another useful fact in dealing with order statistics is that if F is any continuous distribution, then the distibution $y = F(x)$ where x is distributed according to $F(\cdot)$, is uniform on $[0, 1]$. Let us suppose that F, in addition, is strictly increasing. If it continuous, but has flat regions we can approximate it by continuous strictly increasing F's. Let $0 < t < 1$. Then

$$P[F(x) \leq t] = P[x \leq F^{-1}(t)] = F(F^{-1}(t)) = t$$

From our earlier discussion, and the fact that the nondecreasing function F preserves order, we conclude that if $\frac{k_n}{n} \to p$, then the distribution of $\sqrt{n}(F(x_{(k_n)}) - p)$ is asymptotically normal with mean 0 and variance $p(1-p)$. Since $x_{k_n} = F^{-1}(F(x_{(k_n)}))$ the asymptotic distribution of $\sqrt{n}(x_{(k_n)} - x_p)$ is agian normal, with mean 0 and variance $p(1-p)[\frac{dF^{-1}(p)}{dp}]^2$. If we assume that F has a density that is continuous and positive at $x = x_p$, then $\frac{dF^{-1}(p)}{dp} = \frac{1}{f(x_p)}$. So the asymptotic variance of the normalized sample p-th quantile is $\frac{p(1-p)}{(f(x_p))^2}$. For the sample median it is $\frac{1}{4(f(x_p))^2}$. For the normal diistribution with variance σ^2, $f(x_{\frac{1}{2}}) = \frac{1}{\sqrt{2\pi}\sigma}$ and the asymptotic variance of the normalized median is $\frac{\pi\sigma^2}{2}$. It is of course larger than the variance of the sample mean, which is optimal, and the ratio of the two variances is $\frac{2}{\pi}$. This is called the efiiciency of the median.

Let us consider the Cauchy family $\frac{1}{\pi}\frac{1}{1+(x-\theta)^2}$. The asymptotic variance of the normalized median is $\frac{\pi^2}{4}$. That of the normalized MLE is $\frac{1}{I(\theta)}$, where

$$I(\theta) = \frac{1}{\pi}\int \frac{1}{1+(x-\theta)^2}\frac{4(x-\theta)^2}{(1+(x-\theta)^2)^2}dx = \frac{4}{\pi}\int \frac{x^2}{(1+x^2)^3}dx = c$$

that is independent of θ and can be evaluated.

The MLE is the solution $\hat{\theta}_n$ of

$$\psi(\theta) = \frac{1}{n}\sum_i \frac{x_i - \theta}{1+(x_i-\theta)^2} = 0$$

If we denote by U_n the median,

$$0 = \psi(\hat{\theta}_n) = \psi(U_n) + (U_n - \hat{\theta}_n)\psi'(U_n) \simeq \psi(U_n) - c(U_n - \hat{\theta}_n)$$

Therefore,

$$U_n' = U_n - \frac{\psi(U_n)}{c}$$

is a better approximation to the MLE than the median and because of the rapidity of the convergence of Newton's method, after one iteration U_n' is as good as the MLE.

In other situations where $I(\theta)$ is not a constant, but depends on θ, one can estimate $I(\theta)$ by $\frac{1}{n}\sum_{i=1}^n [\frac{\partial \log L}{\partial \theta}(U_n, x_i)]^2$, where U_n is a preliminary consistent estimate. In general, the second approximation is given by

$$U_n' = U_n - \frac{\sum_{i=1}^n \frac{\partial \log L}{\partial \theta}(U_n, x_i)}{\sum_{i=1}^n [\frac{\partial \log L}{\partial \theta}(U_n, x_i)]^2}$$