Statistics 612: Regular Parametric Models and Likelihood Based Inference

Moulinath Banerjee

December 6, 2006

A parametric model is a family of probability distributions $\mathcal{P}$, such that there exists some (open) subset of a finite dimensional Euclidean space, say Θ, such that $\mathcal{P}$ can be written as $\{P_\theta : \theta \in \Theta \subset \mathbb{R}^k\}$. In other words, we can associate each distribution in $\mathcal{P}$ with a $\theta \in \Theta$. When this tagging/correspondence is one-one, we say that the parameter is identifiable; in other words, the parameter uniquely specifies the distribution. For meaningful statistical inference, this is usually a requirement.

In what follows, identifiability will be implicitly assumed. Note that we are really interested in the class of probability distributions (this may be our postulated model for observed data) and not the parameter space itself. So what does a parametrization buy us? For meaningful inference, the parameter describes an integral feature of the probability distribution it is associated with, so that knowledge about the parameter translates easily to knowledge about the features of the distribution. Hence, to obtain meaningful results, one requires adequate *regularity conditions* that govern the behavior of the distribution functions or density functions in terms of θ, in a mathematically tractable manner. We will usually write parametric models as $\{p(x, \theta) : \theta \in \Theta\}$ where $p(x, \theta)$ is the density of P_θ with respect to some dominating measure μ, and x assumes values in the range space of the random variable/vector. The log-density $\log p(x, \theta)$ is denoted by $l(x, \theta)$.

Estimation procedures for θ can be many and varied. A ubiquitous method is maximum likelihood, which, as you know has many desirable properties. Under appropriate regularity conditions on the parametric model, it is consistent for θ, and asymptotically normal, with an asymptotic variance that is the best possible among the class of so-called "regular" estimators. We will talk about "regularity" in some detail later. Roughly, it requires fairly nasty scenarios to render maximum likelihood impotent. Furthermore, the likelihood ratio statistic for testing $\theta = \theta_0$ is asymptotically χ^2, so that confidence sets for θ may be obtained by inversion. Likelihood ratio based confidence sets in many cases have better finite sample properties than their Wald type counterparts based on the asymptotic distribution of the MLE's, since they are more data-driven and adapt nicely to the skewness in the underlying distribution.

The regularity conditions under which maximum likelihood works well can be found in any standard text (see, for example, Chapter 7 of Lehmann (Elements of Large Sample Theory) or Chapter 11 of Keener's notes or Chapter 4 of Wellner's notes). The smoothness of the log-density in θ is a key-requirement. Before, we proceed further, a brief discussion on the term "maximum likelihood estimator" is warranted. Recall that the MLE $\hat{\theta}_n$ is defined as:

$$\hat{\theta}_n = \text{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} l(X_i, \theta) \equiv \text{argmax}_{\theta in \Theta}\, l_n(\theta; \underline{X} = (X_1, X_2, \ldots, X_n))$$

Things are ideal when this exists, is unique and is an interior point of the parameter space, in which case it also satisfies the "score equation"

$$\frac{1}{n} \sum_{i=1}^{n} \dot{l}(X_i, \theta) = 0\,.$$

Here, $\dot{l}(x, \theta)$, the score function is simply $(\partial / \partial\, \theta)\, l(x, \theta)$, written as a column vector. However, it is possible that the MLE may not be unique; furthermore there may be local solutions to the score equation that may not globally maximize the log-likelihood function $l_n(\theta, \underline{X})$. However, under appropriate regularity conditions (see Page 5 of Chapter 4 of Wellner's notes), it can be shown that with probability converging to 1, there exist solutions $\tilde{\theta}_n$ of the score-equations, such that $\tilde{\theta}_n$ converges in probability to θ_0, when θ_0 is true. Henceforth, when we refer to the MLE $\hat{\theta}_n$, we will mean such a consistent sequence of solutions of the score-equation.

0.1 Likelihood ratio, score and Wald statistics

The main result on asymptotic normality of $\hat{\theta}_n$, the MLE of θ can be stated as:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1})\,,$$

where $I(\theta_0)$ is the Fisher Information matrix for θ at the point θ_0 and $p(x, \theta_0)$ is the density from which the i.i.d. observations $X_1, X_2, \ldots, X_n$ are generated. Recall that

$$I(\theta) = E\left[\dot{l}(X, \theta)\, \dot{l}(X, \theta)^T\right] = -E_\theta\left(\ddot{l}(X, \theta)\right),$$

where $\ddot{l}(X, \theta) = (\partial^2 / \partial\, \theta\ \partial\, \theta^T)\, l(X, \theta)$. Here X generically denotes a random variable that follows density $p(x, \theta)$.

Furthermore, we can consider different tests of hypothesis for the parameter θ. Thus, we seek to test $H_0 : \theta = \theta_0$, based on three different statistics. These are:

(a) **Likelihood ratio statistic:** This is given by:

$$2 \log\, \lambda_n = 2\left(\sum_{i=1}^{n} l(X_i, \hat{\theta}_n) - \sum_{i=1}^{n} l(X_i, \theta_0)\right).$$

(b) **Score statistic:** Set $Z_n = n^{-1/2} \sum_{i=1}^n \dot{l}(X_i, \theta_0)$. Define the score statistic as, $R_n = Z_n^T \, I^{-1}(\theta_0) \, Z_n$. Another version of the score statistic is obtained by replacing $I(\theta_0)$ by $I(\hat{\theta}_n)$ above, or even, $\hat{I}_n(\hat{\theta}_n)$, where $\hat{I}_n(\theta) = -n^{-1} \sum_{i=1}^n \ddot{l}(X_i, \theta)$.

(c) **Wald statistic:** The Wald statistic measures the (square of the) distance between the estimate $\hat{\theta}_n$ and the hypothesized parameter θ_0, with respect to an appropriate metric. We define it as: $W_n = n(\hat{\theta}_n - \theta_0)^T \, I(\theta_0) \, (\hat{\theta}_n - \theta_0)$. As before $I(\theta_0)$ can be replaced by $\hat{I}_n(\hat{\theta}_n)$.

Proposition: When the null hypothesis $H_0 : \theta = \theta_0$ holds true, all the three displayed statistics above are asymptotically distributed as a χ^2_k random variable, where k is the number of dimensions of θ.

The proofs of (b) and (c) are left as homework. The proof of (a) will be sketched below.

We first indicate how the asymptotic normality of $\hat{\theta}_n$ comes about. This is a proof-sketch but contains the essentials of the argument. For technical rigour, see any standard text (Chapter 4 of Wellner's notes, for example). Also, the proof below is for one–dimensional θ; the extension to multidimensional θ is routine, albeit at the cost of more complicated notation.

Observe that $\sum_{i=1}^n \dot{l}(X_i, \hat{\theta}_n) = 0$. A standard Taylor series expansion yields that

$$\sum_{i=1}^n \dot{l}(X_i, \theta_0) + (\hat{\theta}_n - \theta_0) \sum_{i=1}^n \ddot{l}(X_i, \theta_0) + \frac{(\hat{\theta}_n - \theta_0)^2}{2} \sum_{i=1}^n l'''(X_i, \theta_n^\star) = 0 \,,$$

for a random point $\theta_n^\star$ that lies between $\hat{\theta}_n$ and θ_0. Some rearrangement yields

$$-\sqrt{n}(\hat{\theta}_n - \theta_0) \, \frac{1}{n} \sum_{i=1}^n \ddot{l}(X_i, \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}(X_i, \theta_0) + \frac{\sqrt{n} \, (\hat{\theta}_n - \theta_0)^2}{2 \, n} \sum_{i=1}^n l'''(X_i, \theta_n^\star) \tag{0.1}$$

whence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \left[-\frac{1}{n} \sum_{i=1}^n \ddot{l}(X_i, \theta_0) - (\hat{\theta}_n - \theta_0) \, \frac{1}{2n} \sum_{i=1}^n l'''(X_i, \theta_n^\star) \right] = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}(X_i, \theta_0) \,. \tag{0.2}$$

A typical assumption is a boundedness condition on the third derivative of $l(x, \theta)$ with respect to θ. More formally, one assumes that $\mid l'''(x, \theta) \mid \leq M(x)$ for all θ in some neighborhood of θ_0, for almost every x, with $E_{\theta_0}(M(X)) < \infty$. This, in conjunction with (0.2) and the consistency of $\hat{\theta}_n$ for θ_0 implies that $\sqrt{n} \, (\hat{\theta}_n - \theta_0)$ is $O_p(1)$ (WHY?). By the strong law of large numbers $-n^{-1} \sum_{i=1}^n \ddot{l}(X_i, \theta_0)$ converges almost surely to $I(\theta_0)$. It follows now from (0.1) that:

$$I(\theta_0) \, \sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{l}(X_i, \theta_0) + o_p(1) \,.$$

Another typical assumption is that $E_\theta \, \dot{l}(X, \theta) = 0$ for all θ (this happens in models where one can differentiate the density under the integral sign, which is a key assumption for the Cramer-Rao inequality). The CLT now implies that:

$$I(\theta_0) \sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0))$$

whence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1}) \, .$$

Furthermore, $\hat{\theta}_n$ has an asymptotically linear representation as:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I(\theta_0)^{-1} \, \dot{l}(X_i, \theta_0) + o_p(1) \, .$$

We now establish the limit distribution of the likelihood ratio statistic. This argument also relies on Taylor expansion. The likelihood ratio statistic, recall, is given by:

$$2 \, \log \, \lambda_n = 2 \left[\sum_{i=1}^{n} l(X_i, \hat{\theta}_n) - \sum_{i=1}^{n} l(X_i, \theta_0) \right] .$$

Expanding the second term within square brackets around $\hat{\theta}_n$, and using the fact that $\sum_{i=1}^{n} \dot{l}(X_i, \hat{\theta}_n) = 0$, we get:

$$2 \, \log \, \lambda_n = -(\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^{n} \ddot{l}(X_i, \hat{\theta}_n) + \frac{(\hat{\theta}_n - \theta_0)^3}{3} \sum_{i=1}^{n} l'''(X_i, \theta_n^\star)$$

for some intermediate point between $\hat{\theta}_n$ and θ_0. Once again, the term involving the third derivative is $o_p(1)$ (by the boundedness condition on the third derivative in terms of the function M; show this rigorously), whence

$$2 \, \log \, \lambda_n = -(\hat{\theta}_n - \theta_0)^2 \sum_{i=1}^{n} \ddot{l}(X_i, \theta_0) + o_p(1) \, .$$

I want you to justify this last step too (Hint: a Taylor expansion will do it). Conclude that the random variable on the left side of the above display converges to χ_1^2.

Food for thought: The above derivations required a Taylor expansion up to the third order; we needed to do this in order to exploit the boundedness assumption on the third derivative. However, it is possible to get away with an expansion up to the second order, without any assumptions on the third derivative of the log-likelihood, provided we make some assumptions on the second derivative. Suppose that for some neighborhood Θ_0 of θ_0 we have:

$$\sup_{\theta \in \Theta_0} \, | \, n^{-1} \sum_{i=1}^{n} \ddot{l}(X_i, \theta) - E_{\theta_0} \, \ddot{l}(X_1, \theta) \, | \rightarrow_{P_{\theta_0} \, a.s.} 0 \, .$$

This is a version of the strong law of large numbers holding uniformly over a class of functions $\{\ddot{l}(X,\theta) : \theta \in \Theta_0\}$. If this happens, the class of functions is referred to as a *Glivenko-Cantelli* class. Glivenko-Cantelli functions play an important role in the modern theory of empirical processes, and consequently, in modern statistical theory.

Can you reconstruct the proofs of the asymptotic normality of $\hat{\theta}$ and the limit distribution of the likelihood ratio statistic using the Glivenko-Cantelli phenomenon above? (Actually, the almost sure convergence in the last display can be weakened to convergence in probability) This is an exercise. Remember, we have forsaken all assumptions about the third derivative now, so you'll need play tricks on a Taylor series expansion up to the quadratic term.

0.2 Confidence sets for θ_0

We indicate how large sample level $1-\alpha$ confidence sets for θ_0 can be constructed. The **likelihood ratio based confidence set** is given by

$$\{\theta : 2 \log \lambda_n(\theta) \leq q_{\chi^2_k,\alpha}\},$$

where $\lambda_n(\theta)$ is the likelihood ratio for the data computed under the null hypothesis $H_{0,\theta}$ (that stipulates that the true data–generating parameter] is θ) and $q_{\chi^2_k,\alpha}$ is the upper α'th quantile of the χ^2_k distribution. If $p(x,\theta)$ is log-concave in θ for almost every x (as happens with exponential family models), then this is guaranteed to be an interval. (WHY?) On the other hand, the confidence set based on the Wald statistic is given by:

$$\{\theta : n\,(\hat{\theta}_n - \theta)^T\, I_n(\hat{\theta}_n)\,(\hat{\theta}_n - \theta) \leq q_{\chi^2_k,\alpha}\}.$$

This is an ellipsoid centered at the MLE $\hat{\theta}_n$ and is necessarily convex. The $I_n(\hat{\theta}_n)$ can be replaced by $I(\hat{\theta}_n)$ or $I(\theta)$ even; asymptotically these are all level $1-\alpha$ confidence sets, though finite sample properties will differ. When $k=1$ this reduces to an interval centered around $\hat{\theta}_n$ and is given by

$$[\hat{\theta}_n - n^{-1/2}\, z_{\alpha/2}\, \sqrt{I(\hat{\theta}_n)}, \hat{\theta}_n + n^{-1/2}\, z_{\alpha/2}\, \sqrt{I(\hat{\theta}_n)}],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$'th quantile of the standard normal distribution (VERIFY). This is identical to the confidence interval that one derives using the asymptotic normality of $\hat{\theta}_n$: $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, I(\theta_0)^{-1})$ when θ_0 is the data–generating parameter. Finally, letting $Z_n(\theta) \equiv n^{-1/2} \sum_{i=1}^n \dot{l}(X_i,\theta)$, the confidence set based on the score–statistic is given by:

$$\{\theta : Z_n(\hat{\theta}_n)^T\, I^{-1}(\hat{\theta}_n)\, Z_n(\hat{\theta}_n) \leq q_{\chi^2_k,\alpha}\}.$$

Of course $I^{-1}(\hat{\theta}_n)$ can be replaced by $\hat{I}_n^{-1}(\hat{\theta}_n)$.

Variance stabilizing transformations: Apart from the likelihood ratio based confidence set, the construction of the others involves estimation of the information matrix (or its inverse),

which typically introduces more variability into the confidence set. The problem gets trickier in semiparametric models, where one has to contend with an infinite dimensional nuisance parameter, in the presence of which inference on θ has to be made, and what needs to be estimated is an *efficient information matrix* that involves projection operators onto infinite dimensional spaces, thereby rendering the whole procedure computationally quite complex. This is why I (and many others) like the likelihood ratio based method of inference, because it turns out to have parameter–free limit distributions in many different statistical settings. This allows construction of confidence sets without having to estimate asymptotic variances of estimators, which in my book is an unmixed blessing. A significant body of my own research has centered around the use of likelihood ratios in shape–restricted inference, and at least, in the context of estimating monotone functions, we now know that the likelihood ratio statistic for testing pointwise hypotheses about the monotone function does exhibit the same pleasing behavior, as in the parametric models of our current discussion. The limit distribution is no longer a χ^2_1 owing to certain basic differences in the asymptotics involved, but nonetheless, nuisance parameters – those satans of inference–are effectively exorcised. See for example, Banerjee and Wellner (2001) (Annals of Statistics), and Banerjee (2004) (Tech. Report, 414, University of Michigan, Department of Statistics). But enough digression, and enough of beating one's own drum.

It turns out that for one dimensional θ there is often a somewhat different way of exorcising nuisance parameters in the limit distribution, thereby facilitating the construction of confidence sets. This is the method of variance stabilizing transformations. By the Delta Method, we know that for any continuously differentiable function g from $\mathbb{R}$ to $\mathbb{R}$, $\sqrt{n}\,(g(\hat{\theta}_n) - g(\theta_0)) \rightarrow_d N(0, g'(\theta_0)^2\, I(\theta_0)^{-1})$. Now, the functional form for $I(\theta)$ is known in many parametric models. Our goal is to choose $g(\theta)$ in such a way that the limiting variance $g'(\theta)^2\, I(\theta)^{-1}$ is a constant – say, 1, without loss of generality. This means that $g'(\theta)$ can be chosen either to be $\sqrt{I(\theta)}$ or $-\sqrt{I(\theta)}$; to keep things simple, choose the first, whence $g(\theta)$ is simply the primitive of $\sqrt{I(\theta)}$. Of course, this whole business really works in practice if the primitive has an analytic closed form expression. We now have, $\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \rightarrow_d N(0,1)$, whence a level $1-\alpha$ large sample confidence set for $g(\theta)$ is given by: $[g(\hat{\theta}_n) - n^{-1/2}\, z_{\alpha/2}, g(\hat{\theta}_n + n^{-1/2}\, z_{\alpha/2}]$. Since g is strictly increasing, transforming the above C.I. by g^{-1} yields a level $1-\alpha$ large sample C.I. for θ.

An illustration: We illustrate the above concepts for the Poisson distribution, along with some numerical simulations. Consider i.i.d. data $X_1, X_2, \ldots, X_n$ from the Poisson distribution with parameter θ, where θ is also the mean of the distribution. Thus:

$$p(x, \theta) = \frac{e^{-\theta}\, \theta^x}{x!}\,.$$

Here $x = 0, 1, 2, \ldots$ takes values in the space of non–negative integers. For this model, $l(x, \theta) = -\theta + x \log \theta + \log x!$ whence $\dot{l}(x, \theta) = -1 + x/\theta$, and the score equation $\sum_{i=1}^n \dot{l}(X_i, \theta) = 0$ has the unique solution $\hat{\theta}_n = \overline{X}$ (check!). This is indeed the MLE (the log–concavity of the density implies uniqueness) and is strongly consistent for the parameter value θ by the strong law of large numbers. For this model $\ddot{l}(x, \theta) = -x/\theta^2$. Check that for this model $I_n(\hat{\theta}_n) = I(\hat{\theta}_n) = 1/\hat{\theta}_n$.

It is easy to check that the Glivenko-Cantelli property is satisfied for the class of functions $\{\ddot{l}(x,\theta) : \theta \in (\theta_0 - \epsilon, \theta_0 + \epsilon)\}$; in other words,

$$\sup_{\theta \in [\theta_0 - \epsilon, \theta_0 + \epsilon]} \left| \frac{1}{n} \sum_{i=1}^{n} \left(-\frac{X_i}{\theta^2} \right) + \frac{E_{\theta_0}(X)}{\theta^2} \right| \rightarrow_{P_{\theta_0}\ a.s.}\ 0\,.$$

The left side of the above display is dominated by $(\theta_0 - \epsilon)^{-2} \mid n^{-1} \sum_{i=1}^{n} X_i - \theta_0 \mid$ which converges to 0 almost surely by the usual strong law of large numbers. Hence, for this model, one could Taylor–expand only up to the quadratic, in order to derive the asymptotics of the MLE and the likelihood ratio statistic.

The information in this model is $I(\theta) = \theta^{-1}$, and hence $\sqrt{n}\,(\hat{\theta}_n - \theta) \rightarrow_d N(0, \theta)$, which also follows as a direct consequence of the CLT. The likelihood ratio statistic for testing $\theta = \theta_0$ is given by:

$$2 \log \lambda_n \equiv 2 \left[\sum_{i=1}^{n} (X_i \log \hat{\theta}_n - \hat{\theta}_n) - \sum_{i=1}^{n} (X_i \log \theta_0 - \theta_0) \right].$$

Based on this, the large sample level $1 - \alpha$ likelihood ratio based C.I. for θ is given by:

$$\left\{ \theta :\ n\,\theta - \log \theta \sum_{i=1}^{n} X_i \leq \frac{q_{\chi^2_1\,\alpha}}{2} + n\,\hat{\theta}_n - \log \hat{\theta}_n \sum_{i=1}^{n} X_i \right\}.$$

This set can be obtained numerically either by bisection methods, or grid search.

For this model, $Z_n(\theta) = n^{-1/2} \sum_{i=1}^{n} (X_i/\theta - 1)$. The score statistic leads to two different kinds of confidence sets. One is based on the asymptotic pivot $Z_n(\theta)^T\, I^{-1}(\hat{\theta}_n)\, Z_n(\theta)$, the other on $Z_n(\theta)^T\, I^{-1}(\theta)\, Z_n(\theta)$. Noting that $I^{-1}(\theta) = \theta$, the two different kinds of confidence sets are respectively obtained as:

$$\left\{ \theta : \frac{n\,\hat{\theta}_n\,(\hat{\theta}_n - \theta)^2}{\theta^2} \leq q_{\chi^2_1,\alpha} \right\}$$

and

$$\left\{ \theta : \frac{n\,(\hat{\theta}_n - \theta)^2}{\theta} \leq q_{\chi^2_1,\alpha} \right\}.$$

The Wald statistic also leads to two different confidence sets for θ. Since, under parameter value θ_0, both $n\,(\hat{\theta}_n - \theta_0)^2\, I(\theta_0)$ and $n\,(\hat{\theta}_n - \theta_0)^2\, I_n(\hat{\theta}_n)$ converge to a χ^2_1 distribution, we get two different confidence sets:

$$\{\theta : n\,(\hat{\theta}_n - \theta)^2/\hat{\theta}_n \leq q_{\chi^2_1,\alpha}\}$$

and

$$\{\theta : n\,(\hat{\theta}_n - \theta)^2/\theta \leq q_{\chi^2_1,\alpha}\}\,.$$

The second Wald–type confidence set and second score type confidence set coincide for this model. There is nothing systemtatic here; it just works out that way in this example.

Table 1: Coverage and average confidence interval length, Poisson data

	LRT		S1		S2		W		VS	
n	C	L	C	L	C	L	C	L	C	L
20	.953	.872	.943	1.087	.941	.889	.929	.873	.935	.869
30	.959	.711	.956	.817	.949	.720	.910	.710	.942	.710
50	.949	.553	.956	.598	.942	.557	.926	.554	.934	.553
100	.954	.390	.951	.406	.951 c	.392	.948	.391	.954	.391
150	.950	.319	.951	.327	.948	.320	.942	.320	.950	.320
200	.942	.276	.937	.282	.935	.277	.943	.278	.938	.277

Last, but not the least, we consider the confidence set based on variance stablization in this problem. Here $I(\theta) = 1/\theta$, so $g'(\theta) = \sqrt{I(\theta)} = 1/\sqrt{\theta}$, hence $g(\theta) = 2\sqrt{\theta}$. This gives $[2\sqrt{\hat{\theta}_n} - n^{-1/2} z_{\alpha/2}, 2\sqrt{\hat{\theta}_n} + n^{-1/2} z_{\alpha/2}] \equiv [a_n, b_n]$ as an asymptotic level $1-\alpha$ confidence set for $g(\theta_0)$, whence $[a_n^2/4, b_n^2/4]$ (using the inverse transformation of g) gives the corresponding C.I. for θ_0.

Numerical simulations comparing the different methods: For each value of n displayed in Table 1, 1000 replicates from the distribution of $(X_1, X_2, \ldots, X_n)$ were generated, where the X_i's are i.i.d. Poisson(1). For each replicate, a 95% confidence interval using 5 different methods – likelihood ratio, score statistic based (1'st kind), score statistic based (2'nd kind), Wald statistic based (1'st kind) and variance stabilization based – were generated. The average lengths of the different C.I's (over the 1000 replicates) and the empirical coverage are displayed. Note the improvement effected by variance stabilization (over the Wald type interval) especially for small samples, and the overall pleasing behavior of the likelihood ratio based C.I's.

0.3 Problems

(1) Repeat the above simulation exercise but with $X_1, X_2, \ldots, X_n$ i.i.d. Bernoulli(θ) for some θ between 0 and 1. I advocate trying this for θ in the vicinity of 0.5 and also in the vicinity of the boundaries, and reporting your conclusions for various sample sizes. You will need to find an appropriate variance stablizing transformation for the Bernoulli variance in this problem.

(2) **One parameter full–rank exponential families:** Consider a one parameter full rank exponential family model nautrally parametrized. Denote the natural parameter by η. The density can be written as:

$$q(x, \eta) = \exp[\eta\, T(x) - B(\eta)]\, h(x)\,,$$

with respect to an appropriate dominating measure, for a strictly convex function B that is infinitely differentiable. The sufficient statistic is $T(x)$, with $E_\eta\,(T(X)) = B'(\eta)$ and

$\mathrm{Var}_\eta(T(X)) = B''(\eta)$. Verify this using the fact that if $W(X)$ has finite expectation, then $\psi(\eta) = E_\eta W(X)$ can be differentiated infinitely many times under the integral sign. Check that the information in this model for η, say $I(\eta) = B''(\eta)$.

Given a sample $X_1, X_2, \ldots, X_n$ from such a density, show that the MLE $\hat{\theta}_n$ is the unique solution to the score equation and is given by $\hat{\eta}_n = H(\overline{T(X)})$, where H is the inverse function of B' and $\overline{T(X)}$ is $n^{-1}(T(X_1) + \ldots + T(X_n))$. Use this to deduce consistency of $\hat{\eta}$ from first principles and establish that $\sqrt{n}(\hat{\eta} - \eta) \to_d N(0, 1/B''(\eta))$. Also, from first principles deduce that the likelihood ratio statistic for testing $\eta = \eta_0$ converges to the χ^2_1 distribution when the null hypothesis is indeed satisfied.

The above formulation includes normal (with known fixed variance), binomial and poisson as special cases. Consider for example $X_1, X_2, \ldots, X_n$ i.i.d. Bernoulli(θ), where $0 < \theta < 1$ and θ is the probability of success in a single trial. Note that $\theta = E_\theta(X_1)$. However, this is not the natural parametrization for this family. Identify the nautral parameter and B for this family. Do the same for the Poisson, where $X_1, X_2, \ldots, X_n$ are i.i.d. Poisson(θ) random variables, with θ denoting the expectation of the Poisson. The natural parameter η, in either case, will be a strictly monotone function of the *usual* parameter θ (with the function depending on the model).

(3) **Unbiased estimators may not always exist** (i) Show that there does not exist any unbiased estimator of the odds ratio $\theta/(1-\theta)$ based on i.i.d. data $X_1, X_2, \ldots, X_n$ following Bernoulli(θ).

(ii) Let $\mathcal{F}$ denote a class of densities $\{f(x, \theta) : \theta > 0\}$ with mean θ^{-1} and variance θ^{-2}, that satisifies the conditions of the information inequality. For what family is the Fisher information $I(\theta)$ minimized over $\mathcal{F}$?

(4) (i) Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. vectors with values in $\mathbb{R}^k$, with $E\, X_1 = \mu$ and $E(X_1^T X_1) < \infty$, so that $\Sigma = E\,(X_1 - \mu)\,(X_1 - \mu)^T$ is well-defined. Let g be a real–valued function on $\mathbb{R}^k$ and suppose that g is continuously differentiable in a neighborhood of μ; thus $\bigtriangledown g$ exists and is continuous in a neighborhood of μ. Show then that $g(\overline{X}_n)$ is *asymptotically linear* at μ, i.e.

$$\sqrt{n}\,(g(\overline{X}_n) - g(\mu)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i) + o_p(1)\,,$$

for some function $\psi(x)$ that you need to identify.

(ii) Show that in a regular parametric model, with θ_0 denoting the true underlying parameter and $\hat{\theta}_n$ denoting the MLE and q being a continuously differentiable map, $q(\hat{\theta}_n)$ is

asymptotically linear, i.e.

$$\sqrt{n}(q(\hat{\theta}_n) - q(\theta_0)) = n^{-1/2} \sum_{i=1}^{n} \psi_{\theta_0}(X_i) + o_p(1) \ ,$$

for some function $\psi_{\theta_0}(\cdot)$ that you need to identify.

4 **Nonregular parametric models:** A classic example of a nonregular parametric model is the set of uniform distributions on $(0, \theta)$ where $\theta > 0$. Consider i.i.d. observations $X_1, X_2, \ldots, X_n$ from $U(0, \theta)$. It is well known (derive for yourself if not comfortable with this fact) that the MLE of θ is $X_{(n)}$. However, unlike regular parametric models, $X_{(n)}$ is not asymptotically normal. The following exercises reveal certain facts about $X_{(n)}$.

(a) Show that there exist sequences of constants $\{a_n\}$ and $\{b_n\}$, possibly depending upon θ such that $(X_{(n)} - a_n)/b_n$ converges to a limiting distribution. Identify the limit. Is $X_{(n)}$ consistent for θ?

(b) Find the UMVUE of θ. Call this T_n. Compute the mean squared errors of T_n, $X_{(n)}$ and $R_n \equiv 2\,\overline{X}_n$ as estimates of θ and comment on their relative behavior for fixed and increasing n.

(c) Use both the exact and the limit distributions of $X_{(n)}$ to construct level $1 - \alpha$ confidence sets (exact and asymptotic respectively) for θ.

5 **Exchangeability and conditional independence for Bernoulli random variables.** The history of this problem dates back to the 1700's and originates according to Stigler in the work of Bayes (see Chapter 3 on Inverse Probability by Stigler – The History of Statistics, pages 122 – 131 for a detailed discussion).

Let $X_1, X_2, \ldots,$ be (a possibly infinite sequence of) random variables defined on a common probability space. Call this sequence exchangeable if for all n and for all permutations Π of $\{1, 2, \ldots, n\}$, the joint distribution of $(X_1, X_2, \ldots, X_n)$ is the same as the joint distribution of $(X_{\Pi(1)}, X_{\Pi(2)} \ldots, X_{\Pi(n)})$. Thus exchangeability amounts to the invariance of the joint distribution under permutations of the data.

In addition to the above, suppose that each X_i is a Bernoulli random variable with parameter p_i.

(a) Show that the p_i's are all equal; in other words the random variables are identically distributed.

(b) Let Θ be a random variable on $(0, 1)$ with distribution denoted by G. Conditional on Θ, generate $X_1, X_2, \ldots,$ as i.i.d. Bernoulli(θ). Show that the joint distribution of the X_i's is exchangeable. What is the common p_i for all these X_i's?

(c) Here is a much stronger fact (the converse of (b)). Suppose we have a sequence of exchangeable Bernoulli random variables $X_1, X_2, \ldots,$. Then there exists a random variable Θ assuming values between 0 and 1, such that given Θ, $X_1, X_2, \ldots$ are all i.i.d. Bernoulli(θ). Thus, infinite exchangeability amounts to conditional independence. This is essentially a version of the *De Finetti representation theorem*. This part is not for credit, but if you've already taken the 620's you should give it a shot.

(d) Consider an infinite sequence of exchangeable Bernoulli random variables, $X_1, X_2, \ldots,$. By (c), without loss of generality you can assume that these are conditionally independent given some Θ, assuming values in (0,1). Suppose that you know that the chance that each of the first n trials ends in a success is $1/n + 1$ for any n, i.e.

$$P(X_1 = 1, X_2 = 1, \ldots, X_n = 1) = \frac{1}{n+1}\ , \quad n = 1, 2, 3, \ldots\ .$$

Show that this implies that for each $1 \leq n < \infty$,

$$P(k\ out\ of\ the\ first\ n\ trials\ are\ successes) \equiv P(S_n = k) = \frac{1}{n+1}\ , \quad k = 0, 1, \ldots, n\,.$$

In the above display, $S_n = X_1 + X_2 + \ldots + X_n$.

Hint: Can you say anything about the distribution of Θ in this problem?

(e) Now consider a finite sequence of exchangeable Bernoulli random variables $X_1, X_2, \ldots, X_N$ and assume that assume that,

$$P(X_1 = 1, X_2 = 1, \ldots, X_n = 1) = \frac{1}{n+1}\ , \quad n = 1, 2, \ldots, N\,.$$

Then $S_n \equiv X_1 + X_2 + \ldots + X_n$, the number of successes in the first n trials has a discrete uniform distribution for each $n \leq N$; i.e.

$$P(k\ out\ of\ the\ first\ n\ trials\ are\ successes) \equiv P(S_n = k) = \frac{1}{n+1}\ , \quad k = 0, 1, \ldots, n\,.$$

The proof I have uses induction, but other proofs would be welcome. There is a subtle difference between scenarios (d) and (e). In (e) we only have finitely many random variables and therefore we cannot conclude that they are conditionally i.i.d. given some fixed Θ. The De Finetti representation can only be invoked for an infinite sequence. The assertion in (d) can be established using the distribution of Θ; in (e) the exchangeability hypothesis needs to be used more fundamentally. Indeed, as (e) shows, the fact that the distribution of the total number of successes S_n has a discrete uniform distribution is fundamentally a consequence of exchangeability, rather than the De Finetti representation. While the result in (d) seems to be well-known, the result in (e) does not seem to be. Thomas Richardson and myself stumbled on this fact a few years ago. It does not seem substantial enough for a paper, but certainly makes a very good homework problem.