Asymptotic Equivalence of Statistical Experiments

Michael Nussbaum *

Cornell University

Abstract

The idea of approximating a sequence of statistical experiments by a gaussian family goes back to Wald (1943), but has been fully developed by Lucien Le Cam, who introduced the term "local asymptotic normality". This theoretical framework has become a standard tool for proving efficiency of tests and estimators (in particular of the maximum likelihood estimator) in an asymptotic sense. It suffices to note that the initial model is approximately normal and thus inherits, in an asymptotic sense, the simple structure of normal models. The passage to the limit in the sense of the whole model is stronger and richer in consequences than the results on limit laws of various individual functions of the sample which are consequences of the central limit theorem.

1 Basic definitions and first examples

In this article we try to give an elementary introduction to asymptotic theory of statistical experiments, a theory closely associated with the name of Lucien Le Cam (see [15], [16]). Furthermore, we discuss related developments in nonparametric statistics that have recently expanded the scope of applications of this theory.

The basic object in statistics is a family of laws $(P_\vartheta,\ \vartheta \in \Theta)$ on a measurable space $(\Omega, \mathcal{A})$. This structure could be called a statistical model; but we will use the term statistical experiment. An experiment $\mathcal{P}$ is defined as a collection

$$\mathcal{P} = (\Omega, \mathcal{A}, (P_\vartheta, \vartheta \in \Theta)).$$

We observe a random variable X defined on Ω distributed according to P_ϑ; and the value of ϑ is unknown. Because the measurable space is already subsumed in the definition of the law P_ϑ, we can write $\mathcal{P} = (P_\vartheta,\ \vartheta \in \Theta)$.

Here is a direct quotation by L. Le Cam that nicely describes the starting point (see [13]): *En général, la famille $\mathcal{P}$ est compliquée. On voudrait alors l'approcher par une famille plus simple.*

The method used to choose a "simpler" family is provided by the sufficient statistics. These allow one to reduce the data dimensionality in a general sense by a specified transformation

*Research supported by the National Science Foundation under grant DMS0306497

without loss of information. Let us begin with one of the most elementary examples in which we can find a sufficient statistic. In the following, a n-sample is a vector of i.i.d. observations.

Example 1 (Normal n-sample location family) *Let X_i, $i = 1, \dots, n$, be independent $N(\vartheta, 1)$. The sample mean $\bar{X}_n$ is a sufficient statistic. The law of $\bar{X}_n$ is $\mathcal{L}(\bar{X}_n) = N(\vartheta, n^{-1})$.*

In this case, the initial experiment is given by the family of joint distribution of the n-sample, ϑ being unknown ($\vartheta \in \Theta$ say, with $\Theta \subset \mathbb{R}$). If one observes only the value of the sufficient statistic $\bar{X}_n$, one is dealing with its family of distributions $N(\vartheta, n^{-1})$ which depends on the unknown parameter ϑ, as does the family of distributions of the entire sample. Thus we obtain two experiments indexed by the same ϑ. If we believe that a sufficient statistic contains all the information available about ϑ, we are led to the intuitive notion of equivalent experiment.

Proposition 1 *Suppose that $\Theta \subset \mathbb{R}$. Then the experiments given by the observations*

$$X_i = \vartheta + \xi_i,\ i = 1, \dots, n,\ \xi_i \sim N(0,1),\ \textit{independent},\ \vartheta \in \Theta$$
$$Y = \vartheta + n^{-1/2}\xi,\ \ \xi \sim N(0,1),\ \vartheta \in \Theta$$

are equivalent.

The exact definition of equivalence will be given later. For now, we will settle with the intuitive notion that the two experiments contain the same information about ϑ.

Example 2 (Normal n-sample scale family)
Let X_i, $i = 1, \dots, n$, be independent $N(0, \sigma^2)$. The sample variance $S_n^2 = n^{-1}\Sigma_{i=1}^n X_i^2$ is a sufficient statistic. S_n^2 is distributed as $n^{-1}\sigma^2\chi_n^2$, where χ_n^2 is a chi-square random variable with n degrees of freedom.

Suppose that the unknown parameter σ^2 satisfies that $\sigma^2 \in \Theta$, where $\Theta \in (0, \infty)$. Here the reduction in the dimension of the model took place, but we could try a further simplication, wondering if the family of laws $\mathcal{L}(n^{-1}\sigma^2\chi_n^2)$, $\sigma^2 \in \Theta$ could admit an extra simplication through the central limit theorem. Indeed, it can

$$\sqrt{n}(S_n^2 - \sigma^2) \overset{\mathcal{L}}{\Longrightarrow} N(0, 2\sigma^4). \tag{1}$$

If the law of a sufficient statistic has a normal limit law, as is the case here, one is tempted to conclude that also the family law (initial experiment) converges to a Gaussian experiment. We would rewrite (1) as

$$S_n^2 \approx N\left(\sigma^2, 2\sigma^4\right) \tag{2}$$

without specifying the relation $\approx$, and the experiment would limit the family

$$\left(N\left(\sigma^2, 2\sigma^4\right), \sigma^2 \in \Theta\right). \tag{3}$$

However, since this convergence in law is a weak convergence, we need a stronger version of (1). Recall that the total variation distance $\|\cdot\|_{TV}$ of the laws P, Q is defined as

$$\|P - Q\|_{TV} = 2 \sup_{A \text{ mesurable}} |P(A) - Q(A)| = \int |p - q|\, d\mu$$

where p, q are the densities with respect to μ of P and Q, respectively. It is well-known, under certain regularity conditions, that the central limit theorem applies to the stronger sense of total variation (cf. van der Vaart [22], 2.31). Let us explain why. In our case the law χ^2_n is continuous and very regular, so the distribution of $\sqrt{n}(S_n^2 - \sigma^2)$ admits a density function. Since this law converges itself to the law of $N(0, 2\sigma^4)$, it is natural that the density also converges. For densities, the central limit theorem is known as the local limit theorem. If this theorem is applied point by point (pointwise?) for a sequence of densities, it follows by Scheffé's lemma that for the two densities ($p_{\sigma,n}$ and q_σ, say), we have

$$\int |p_{\sigma,n} - q_\sigma| \to 0, \text{ as } n \to \infty,$$

and therefore, for the total variation distance,

$$\left\| \mathcal{L}(\sqrt{n}(S_n^2 - \sigma^2)) - N(0, 2\sigma^4) \right\|_{TV} \to 0, \text{ as } n \to \infty.$$

The application (mapping?) $x \mapsto n^{-1/2}x + \sigma^2$ is measurable and bijective. As a consequence,

$$\left\| \mathcal{L}(S_n^2) - N(\sigma^2, 2n^{-1}\sigma^4) \right\|_{TV} \to 0, \text{ as } n \to \infty.$$

For an additional argument, we deduce that this convergence is uniform on any parameter set of the form $\sigma^2 \in \Theta \subset (a, b)$, $a > 0$.

Thus, the law of the sufficient statistic converges in total variation, uniformly. Since the normal approximation can be made for all the events, uniformly on the unknown parameter, it is legitimate to describe the initial experiment with a normal law (asymptotically).

Hence, we can formulate the concept of *asymptotic equivalence* as follows: it is a comparison of experiments intended to contain asymptotically, as $n \to \infty$, the same amount of information about the unknown parameter σ^2. Clearly, if we want to clarify the precise meaning of the foregoing, it will be closely related with a concept to define asymptotic sufficiency.

Proposition 2 *Suppose that $\sigma^2 \in \Theta \subset (a, b)$, $a > 0$. Then the experiments given by the observations*

$$X_i = \sigma\xi_i, \ i = 1, \ldots, n, \ \xi_i \sim N(0, 1), \textit{independents}, \ \sigma^2 \in \Theta$$
$$Y = \sigma^2 + n^{-1/2}\sqrt{2}\sigma^2\xi, \quad \xi \sim N(0, 1), \ \sigma^2 \in \Theta$$

are asymptotically equivalent.

Since we want the simplest approximation of the initial experiment, the second model above is not entirely satisfactory. Although normal expectation of σ^2, the second experiment is heteroskedastic, that is to say that the variance will also depend on σ^2.

For all the problems of statistical inference, a homoscedatic Gaussian model would be preferable. To obtain homoscedasticity, we used a procedure to stabilize the variance. Remember the principle of this idea well known in statistics. Returning to equation 1 and note that for a regular function g (twice differentiable, say), it leads

$$\sqrt{n}(g(S_n^2) - g(\sigma^2)) \overset{\mathcal{L}}{\Longrightarrow} N(0, 2\sigma^4(g'(\sigma^2))^2). \tag{4}$$

The function $g(x) = \log x$ has derivative $g'(x) = x^{-1}$; we obtained

$$\sqrt{n}(\log S_n^2 - \log \sigma^2)) \overset{\mathcal{L}}{\Longrightarrow} N(0, 2n^{-1}),$$

and we could rewrite (2) as

$$\log S_n^2 \approx N\left(\log \sigma^2, 2\right). \tag{5}$$

In this case, $S_n^2 \mapsto \log(S_n^2)$ is the transformation for stabilizing variance. This simple reasoning is good for the convergence in law, to justify at the level of experiences, we must use convergence in total variation.

Proposition 3 *Suppose that $\sigma^2 \in \Theta \subset (a, b)$, $a > 0$. Then, the experiments given by the observations*

$$X_i = \sigma\xi_i,\ i = 1, \ldots, n,\ \xi_i \sim N(0, 1),\ \textit{independent},\ \sigma^2 \in \Theta$$
$$Y = \log \sigma^2 + n^{-1/2}\sqrt{2}\xi, \quad \xi \sim N(0, 1),\ \sigma^2 \in \Theta$$

are asymptotically equivalents.

With the second experiment above was pushed further simplification, by obtaining a simple Gaussian model of translation. But the disadvantage now is that the unknown parameter (the average) is not the original σ^2, but the transformed $\log \sigma^2$, which makes it more complicated procedures such as statistical estimation of σ^2. However, this reduced model is not without interest, and is the main subject of this presentation.

Example 3 (Poisson n-sample) *Let $X_i,\ i = 1, \ldots, n$ be Poisson $\mathrm{Po}(\vartheta)$. Again, the sample mean $\bar{X}_n$ is a sufficient statistique.*

By the central limit theorem, we have

$$\sqrt{n}(\bar{X}_n - \vartheta) \overset{\mathcal{L}}{\Longrightarrow} N(0, \vartheta). \tag{6}$$

Here, however, convergence in total variation does not hold because the law of $\mathcal{L}(\bar{X}_n)$ is discrete (we have $\mathcal{L}(n\bar{X}_n) = \mathrm{Po}(n\vartheta)$). But from what we know about the limit theorems, convergence in total variation could take place after appropriate smoothing. This idea leads us reasonable to reconsider the concept of statistical equivalence.

Recall the equivalence of Example 1 (n-sample Gaussian by sufficient statistic): set $P_{n,\vartheta} = \mathcal{L}(X_1, \ldots, X_n)$ and $Q_{n,\vartheta} = \mathcal{L}(\bar{X}_n) = N(0, n^{-1}\vartheta)$. Moreover, $P_{n,\vartheta}(\cdot|\bar{X}_n = x)$ is the conditional distribution of $(X_1, \ldots, X_n)$ given $\bar{X}_n$. According to the definition of a sufficient statistic, $P_{n,\vartheta}(\cdot|\bar{X}_n = x)$ does not depend on the parameter ϑ so to speak.

$$P_{n,\vartheta}(A|\bar{X}_n = x) = P_{n,\cdot}(A|\bar{X}_n = x) \tag{7}$$

for any Borel set A of $\mathbb{R}^n$. But regardless of completeness, in this very regular conditional law may be chosen as the Markov kernel (transition function).

$$K_\vartheta(A, x) = P_{n,\vartheta}(A|\bar{X}_n = x)$$

to restore the law $P_{n,\vartheta}$ from the law $Q_{n,\vartheta}$; if the mapping $K_\vartheta : Q_{n,\vartheta} \mapsto K_\vartheta Q_{n,\vartheta}$ is defined as

$$K_\vartheta Q_{n,\vartheta}(A) = \int K_\vartheta(A,x) Q_{n,\vartheta}(dx) = \int P_{n,\vartheta}(A|\bar{X}_n = x) Q_{n,\vartheta}(dx), \quad A \in \mathcal{A},$$

so we have $K_\vartheta Q_{n,\vartheta} = P_{n,\vartheta}$. The sufficiency (7) implies that the kernel K_ϑ can be chosen independently of ϑ; so there exists a Markovian kernel K such that

$$KQ_{n,\vartheta} = P_{n,\vartheta} \text{ for all } \vartheta \in \Theta. \tag{8}$$

Regarding the converse relationship, it is clear that there is a another Markovian kernel K' which yields $Q_{n,\vartheta}$ from $P_{n,\vartheta}$. Let $X = (X_1, \ldots, X_n)$ and consider the non-random application $t(X) = \bar{X}_n$; the trivially defined Markovian kernel

$$K'(B,x) = \mathbf{1}_B(t(x))$$

for all Borelian B of $\mathbb{R}$ is one the satisfies that

$$K'P_{n,\vartheta}(B) = \int \mathbf{1}_B(t(x)) P_{n,\vartheta}(dx) = \int_B Q_{n,\vartheta}(dx) = Q_{n,\vartheta}(B).$$

We therefore have

$$K'P_{n,\vartheta} = Q_{n,\vartheta} \text{ for all } \vartheta \in \Theta. \tag{9}$$

The two relations (8) and (9) which are satisfied within the framework of a sufficient statistic inspire the following definition

Definition 1 (Δ-distance of Le Cam) *Let $\mathcal{P} = (P_\vartheta,\ \vartheta \in \Theta)$ and $\mathcal{Q} = (Q_\vartheta,\ \vartheta \in \Theta)$ be two experiments indexed by the same parameter Θ, but possibly with different sample spaces. The deficiency of $\mathcal{P}$ with respect to $\mathcal{Q}$ is*

$$\delta(\mathcal{P}, \mathcal{Q}) = \inf_K \sup_{\vartheta \in \Theta} \|Q_\vartheta - KP_\vartheta\|_{TV}$$

*(*inf *is taken over all the Markovian kernels) and the Δ-distance is*

$$\Delta(\mathcal{P}, \mathcal{Q}) = \max(\delta(\mathcal{P}, \mathcal{Q}), \delta(\mathcal{Q}, \mathcal{P})).$$

This is a simplified definition, which is valid under the following conditions of regularity. For both experiments, $\mathcal{P} = (\Omega_\mathcal{P}, \mathcal{A}_\mathcal{P}, (P_\vartheta, \vartheta \in \Theta))$ (resp., $\mathcal{Q} = (\Omega_\mathcal{Q}, \mathcal{A}_\mathcal{Q},\ (Q_\vartheta, \vartheta \in \Theta))$), space observations $\Omega_\mathcal{P}$ (resp., $\Omega_\mathcal{Q}$) is a Polish space (separable metric space and complete) and $\mathcal{A}_\mathcal{P}$ (resp., $\mathcal{A}_\mathcal{Q}$) is the corresponding Borelian. In addition, the family $(P_\vartheta, \vartheta \in \Theta)$ (resp., $(Q_\vartheta, \vartheta \in \Theta)$) is dominated by a σ-finite measure. If these conditions are not satisfied, the definition involves more abstract objects as generalized Markov kernels (cf. Le Cam [15], Chap. 2.3 or van der Vaart [23], Chap. 8).

In examples 1-3, if $\mathcal{P} = (P_\vartheta, \vartheta \in \Theta)$ is the family of origin, T is a sufficient statistic and $\mathcal{Q} = (Q_\vartheta\ ,\ \vartheta \in \Theta)$ is the family $Q_\vartheta = \mathcal{L}(T|P_\vartheta)$, then we have $\Delta(\mathcal{P}, \mathcal{Q}) = 0$. The latter relationship is interpreted as the equivalence of experiments (in strict sense).

Definition 2 *Supppose that $\mathcal{P}_n = (P_{n,\vartheta}\ ,\ \vartheta \in \Theta)$ and $\mathcal{Q}_n = (Q_{n,\vartheta},\ \vartheta \in \Theta)$ are two sequence of experiments, indexed by n, and such that the space of observations can also depend on n. The sequences $\mathcal{P}_n, \mathcal{Q}_n$ are asymptotically equivalent if $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$.*

To clarify the statistical significance of the deficiency, recall the classical setting of decision theory. For a parameter space Θ and a measurable space of decisions $(E, \mathcal{E})$, $W : E \times \Theta \mapsto [0, \infty)$ is a loss function such that for all $\vartheta \in \Theta$, $W(\vartheta, \cdot)$ is measurable with respect to E. A decision rule in the randomized experiment $\mathcal{P} = (\Omega, \mathcal{A}, (P_\vartheta, \vartheta \in \Theta))$ is a Markovian kernel $t(\cdot, \omega)$ which associates to each $\omega \in \Omega$ a probability measure on $(E, \mathcal{E})$. The risk of t at ϑ is defined as

$$r_t(\vartheta) = \int W(e, \vartheta) t(de, \omega) P_\vartheta(d\omega).$$

Proposition 4 (Characterization of deficiency) *Two experiments $\mathcal{P}, \mathcal{Q}$ satisfy the relation $\delta(\mathcal{P}, \mathcal{Q}) \leq \varepsilon$ if and only if for all $\varepsilon > 0$, for all decision problem with loss function W such that $0 \leq W \leq 1$, for all decision function t avaiable in $\mathcal{Q}$, there exists a decision function t^* available in $\mathcal{P}$ such that*

$$r_{t^*}(\vartheta) \leq r_t(\vartheta) + \varepsilon, \quad \vartheta \in \Theta$$

(the decision rule t^* is as good as t, ε near).

Note that this characterization concerns the deficiency $\delta(\mathcal{P}, \mathcal{Q})$ of $\mathcal{P}$ with respect to $\mathcal{Q}$ and that δ is not symmetric. Therefore, if the symmetrized expression (the *Delta*-distance) is such that $\delta(\mathcal{P}, \mathcal{Q}) \leq \varepsilon$, then the risk found in $\mathcal{Q}$ are also available in $\mathcal{P}$, ε close, and vice versa.

Let's go back to the Example 3, where the law of the sufficient statistic $\tilde{X}_n = n\bar{X}_n = \sum_{i=1}^n X_i$ is a Poisson distribution $\mathrm{Po}(n\vartheta)$. First, in (6), we would deduce an approximation by Gaussian experiment

$$\bar{X}_n \approx N\left(\vartheta, n^{-1}\vartheta\right). \tag{10}$$

This experiment, however, is heteroscedastic. Here, the function $g(x) = x^{1/2}$ is a variance stabilizing transform since, in analogy with (4), we obtain $g'(x) = (4x)^{-1/2}$ and as a consequence

$$\sqrt{n}(g(\bar{X}_n) - g(\vartheta)) \stackrel{\mathcal{L}}{\Longrightarrow} N(0, \vartheta(g'(\vartheta))^2) = N(0, 1/4).$$

Thus, we could rewrite (10) in the form

$$\bar{X}_n^{1/2} \approx N\left(\vartheta^{1/2}, (4n)^{-1}\right), \tag{11}$$

and thus obtain a simple Gaussian approximation, it remains to establish the same result in the strong sense of Markovian kernels. The most elegant method fot this has recently been obtained in [2], Theorem 4. Let U be a uniform random variable on $[-1/2, 1/2)$, independent of $\tilde{X}_n$. We set

$$Z_n = \mathrm{sgn}\left(\tilde{X}_n + U\right)\sqrt{\left|\tilde{X}_n + U\right|} \tag{12}$$

and we show that

$$\left\|\mathcal{L}(Z_n) - N((n\vartheta)^{1/2}, 1/4)\right\|_{TV} \leq C(n\vartheta)^{-1/2}$$

where C is independent on ϑ and n. Here, the Markovian kernel is given by the operation "smoothing" $\tilde{X}_n \mapsto \tilde{X}_n + U$ which is also invertible: the value of $\tilde{X}_n + U$ is identified with that of $\tilde{X}_n$ which takes integers values. Therefore, we obtain two kernels K, K' which are converging to 0 in both deficiencies $\delta(\mathcal{P}, \mathcal{Q})$ and $\delta(\mathcal{Q}, \mathcal{P})$.

Proposition 5 *Suppose that $\vartheta \in \Theta \subset (a,b)$, $a > 0$. Then the experiments given by the observations*

$$X_i,\ i = 1, \ldots, n,\ X_i \sim \mathrm{Po}(\vartheta), \textit{independent},\ \vartheta \in \Theta$$
$$Y = \vartheta^{1/2} + \frac{1}{2} n^{-1/2} \xi, \quad \xi \sim N(0,1), \vartheta \in \Theta$$

are asymptotically equivalent.

Similarly in Proposition 3, the approximation is obtained by a Gaussian translation experiment, the parameter here, however, is $\vartheta^{1/2}$.

2 Parametric model: Local Asymptotic Normality

Let $\mathcal{P}_n = (P_{n,\vartheta},\ \vartheta \in \Theta)$ be a sequence of regular parametric models, $\Theta \subset \mathbb{R}^k$, generated by n equally distributed independent variables, in which $P_{n,\vartheta}$ is a product law P_ϑ^n. Suppose that the the maximum likelihood estimate $\hat{\vartheta}_n$ satisfies that

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta) \stackrel{\mathcal{L}}{\Longrightarrow} N_k(0, J_\vartheta^{-1}),$$

where J_ϑ is the Fisher's information matrix at the point ϑ. Often the MLE is a sufficient statistic, or at least sufficient in an assymptotic sense. From what we saw (cf. (2), (3), (10)), one is tempted to seek an approximation of the experiment $\mathcal{P}_n$ by the family

$$\left(N_k(\vartheta, n^{-1} J_\vartheta^{-1}), \vartheta \in \Theta\right) \tag{13}$$

which is a Gaussian heteroskedastic experiment, generalizaing that of Proposition 2 (for Y). This idea was developed by Le Cam, who also reported that the approximation (13) is not attractive from the standpoint of decision theory. Indeed, the presence of the parameter ϑ in the covariance matrix and the structure of J_ϑ^{-1}, $\vartheta \in \Theta$, essentially arbitrary, do not allow us to consider (13) as a simplification.

2.1 The local method

A more promising approximation is provided by the local method. We consider a limited series of experiments where the parameter ϑ varies only in a neighborhood $\Theta_n(\vartheta_0)$ of a known value ϑ_0, and where the diameter of the neighborhood is on the order of $n^{-1/2}$. A such restriction may be justified by two arguments: Firstly, for each decision rule, the risk on the restricted set of parameters $\Theta_n(\vartheta_0)$ provides lower bounds for the not restricted experiments. Moreover, these lower bounds are often reasonable because ϑ_0, which is supposde to be known, may be replaced by a prior estimator that would identify the true ϑ with an accuray of about $n^{-1/2}$. Therefore, choose $\vartheta_0 \in \Theta$ (localization center) and let $\vartheta = \vartheta_0 + n^{-1/2} h$, where h is a local parameter. We have then for all compact $K \subset \mathbb{R}^k$

$$\sqrt{n}(\hat{\vartheta}_n - \vartheta_0) \stackrel{\mathcal{L}}{\Longrightarrow} N(h, J_{\vartheta_0}^{-1}),\ h \in K\ . \tag{14}$$

Here, if ϑ_0 is assumed known and $\hat{\vartheta}_n$ is complete, was again a result of complete statistics which converges in distribution to the family

$$\left(N(h, J_{\vartheta_0}^{-1}),\ h \in \mathbb{R}^k\right). \tag{15}$$

Such a family of Gaussian translation offers all the benefits of simplicity. In order to obtain a lower bound of the risk in the estimation of ϑ, we make a change of variable accordingly: For the risk and the quadratic estimator $\tilde{\vartheta}_n$ of any ϑ we have that

$$\begin{aligned} n\ E_\vartheta\left(\tilde{\vartheta}_n - \vartheta\right)^2 &= n\ E_\vartheta\left(\tilde{\vartheta}_n - \vartheta_0 + n^{-1/2}h\right)^2 = E_\vartheta\left(n^{1/2}(\tilde{\vartheta}_n - \vartheta_0) - h\right)^2 \\ &= E_\vartheta\left(\tilde{h}_n - h\right)^2 \end{aligned} \tag{16}$$

where we set $\tilde{h}_n = n^{1/2}(\tilde{\vartheta}_n - \vartheta_0)$; $\tilde{h}_n$ is interpreted as an estimator of h.

It remains to prove rigorously the convergence of the localized and restricted experiment to the Gaussian limit(15). For this, Le Cam has developed a direct method from the weak convergence (14), by Markov kernels which is similar to smoothing (12) (see [15], section 11.8, also M(u)ller [17] for a comparison).

2.2 The Likelihood Processes

Another method, much more efficient, is based on the deep connection between the equivalence of experiments and the likelihood processes. The latter is defined as follows. Let $\mathcal{P} = (P_\vartheta, \vartheta \in \Theta)$ be a family defined on $(\Omega, \mathcal{A})$, dominated by one of its elements P_{ϑ_0}, where $\vartheta_0 \in \Theta$. The density $\Lambda(\vartheta)(\omega) = dP_\vartheta/dP_{\vartheta_0}(\omega)$ generates a random variable $\Lambda(\vartheta)$, if the argument ω follows the law $\mathcal{L}(\omega) = P_{\vartheta_0}$. The set of random variables

$$\Lambda_\mathcal{P} = (\Lambda(\vartheta),\ \vartheta \in \Theta)$$

all of them defined on the probability space $(\Omega, \mathcal{A}, P_{\vartheta_0})$ form a stochastic process indexed by ϑ, which is called the likelihood process of the experiment $\mathcal{P}$. The law $\mathcal{L}(\Lambda_\mathcal{P})$ of this process is the set of finite marginal laws. Another key result of Le Cam is: if $\Lambda_\mathcal{P}$ and $\Lambda_\mathcal{Q}$ are the likelihood processes associated with the experiments $\mathcal{P}$, $\mathcal{Q}$, then we get

$$\Delta(\mathcal{P}, \mathcal{Q}) = 0 \text{ if and only if } \mathcal{L}(\Lambda_\mathcal{P}) = \mathcal{L}(\Lambda_\mathcal{Q}). \tag{17}$$

To explain this result in heuristic, note first that the process $\Lambda_\mathcal{P}$ is a sufficient statistic. To be more precise, if ω is the data, we define a statistic $T(\omega)$ with values in a high dimensional space as whole

$$T(\omega) = (\Lambda(\vartheta)(\omega), \vartheta \in \Theta)$$

(this takes values in the space $\mathbb{R}^\Theta$). According to the factorization criterion of Neyman, if g_ϑ is the projection $\mathbb{R}^\Theta \mapsto \mathbb{R}$ whose value is the corresponding coordinated of ϑ, we have

$$\Lambda(\vartheta)(\omega) = dP_\vartheta/dP_{\vartheta_0}(\omega) = g_\vartheta\left(T(\omega)\right),$$

and as a consequence T is sufficient. The familly of laws

$$(\mathcal{L}(T|P_\vartheta), \vartheta \in \Theta)$$

is an equivalent experiment to $\mathcal{P}$. It suffices to remark that this family is already determined by one of its elements, i.e. $\mathcal{L}(T|P_{\vartheta_0}) = \mathcal{L}(\Lambda_{\mathcal{P}})$. Now we have for each bounded measurable function h which only depends on a finite number of coordinates h

$$E_\vartheta h(T) = E_{\vartheta_0} h(T) \frac{dP_\vartheta}{dP_{\vartheta_0}} = E_{\vartheta_0} h(T) g_\vartheta(T),$$

which is a functional of the law $\mathcal{L}(T|P_{\vartheta_0})$. We have thus proved (17) loosely, assuming that the family $\mathcal{P}$ is dominated by one of its elements. The criterion (17) suggests a similar approximating result

$$\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0 \text{ if and only if } \mathcal{L}(\Lambda_{\mathcal{P}_n}) \overset{\mathcal{L}}{\approx} \mathcal{L}(\Lambda_{\mathcal{Q}_n}) \text{ for } n \to 0 \tag{18}$$

where the relation $\overset{\mathcal{L}}{\approx}$ means approximation in law, we will precise this concept. If a limit experiment is designated as in (15), following $\mathcal{Q}_n$ is constant and the approximation is reduced to convergence in law.

2.3 Local asymptotic normality (LAN) in the case of an n-sample

In the case of independent variables, the convergence in law of the likelihood process can be checked in the following way. First, the translation Gaussian model (15), takes the form

$$\Lambda_{\mathcal{Q}}(h) = \exp\left(h^\top J_{\vartheta_0}^{1/2} \xi - \frac{1}{2} h^\top J_{\vartheta_0} h\right)$$

where ξ is a standard Gaussian normal in $\mathbb{R}^k$. In the family $\mathcal{P}_n = (P_{n,\vartheta},\ \vartheta \in \Theta)$, $\Theta \subset \mathbb{R}^k$, where $P_{n,\vartheta}$ is a product measure P_ϑ^n, we performed the localization $\vartheta = \vartheta_0 + n^{-1/2}h$ introducing the new local parameter h. Let f_h be the density of $P_{\vartheta_0 + n^{-1/2}h}$ with respect to the Lebesgue measure. In the regular cases where f_h is differentiable at h, we obtain $f_h/f_0 - 1 \approx n^{-1/2}$ and the log-likelihood can be written as

$$\begin{aligned} \log \Lambda_{\mathcal{P}_n}(h) &= \sum_{i=1}^n \log \frac{f_h}{f_0}(X_i) \approx \sum_{i=1}^n \left(\frac{f_h}{f_0}(X_i) - 1\right) - \frac{1}{2} \sum_{i=1}^n \left(\frac{f_h}{f_0}(X_i) - 1\right)^2 \\ &\overset{\mathcal{L}}{\Longrightarrow} h^\top J_{\vartheta_0}^{1/2} \xi - \frac{1}{2} h^\top J_{\vartheta_0} h \end{aligned} \tag{19}$$

by a sort of central limit theorem and a law of large numbers. The convergence in law from $\Lambda_{\mathcal{P}_n}$ to $\Lambda_{\mathcal{Q}}$ is the verified through the logarithm.

Recall that the Gaussian limit experiment (15) has been established in a local setting, around ϑ_0, assuming implicitly that ϑ_0 can be estimated with an accuracy of order $n^{-1/2}$. The theory developed around the idea of Le Cam may be called the paradigm LAN (local asymptotic normality). The scope of applications has been much more extensive than that of the independent variables, see Strasser [21] Genon-Catalot and Picard [7] van der Vaart [22], Shiryaev and Spokoiny [20].

The usefulness of this approach is limited in the non parametric case. Suppose that all P_ϑ are laws on [0, 1], that the family $\mathcal{P}_n$ of product laws $P_{n,\vartheta} = P_\vartheta^n$ is parameterized by the

Lebesgue density $\vartheta = f$ of P_ϑ and that Θ is identified with a set of densities Σ of infinite dimension. Again, the location around a central density f_0 by $f - f_0 \approx n^{-1/2}$ and the LAN property is often possible. But typically the center f_0 can not be estimated with precision $n^{-1/2}$, because the problem is ill posed (in the analytical sense). Indeed, for the empirical distribution function $\hat{F}_n$ we have

$$\hat{F}_n(t) - \int_0^t f(t)dt = O_p(n^{-1/2})$$

but the mapping $f \mapsto \int_0^t f(t)dt$ has no continuous inverse, so the speed of estimating $n^{-1/2}$ is not possible to f. Since the boundaries of risk was obtained through the LAN method based on renormalization by $n^{1/2}$ (cf. (16)), it follows that the LAN paradigm is not adequate to estimate the overall density.

3 Non-parametric asymptotic equivalence

In the case of n independent variables equidistributed on $\mathbb{R}$, the empirical distribution function $\hat{F}_n$ is always a sufficient statistic. The starting point may be the central limit theorem for this statistic:

$$\sqrt{n}(\hat{F}_n(t) - F(t)) \stackrel{\mathcal{L}}{\Longrightarrow} B \circ F(t), \text{ for a Brownian bridge } B.$$

First, we would derive an approximating Gaussian experiment similar to (10) and (13), which now could take the form

$$dy(t) = f(t)dt + n^{-1/2} f^{1/2}(t)dW(t), \ t \in [0,1]$$

("signal f observed in white noise"). But this heteroskedastic model is not valid in a statistical sense (i.e. it is trivial) since the laws of the process $y(t)$, $t \in [0,1]$ are orthogonal if f are different. This effect is caused by the presence of the factor $f^{1/2}(t)$ in $dW(t)$, that is to say, in the diffusion coefficient. To solve this problem of heteroskedasticity, a stabilization of the variance, similar to (11), would be desirable. In the case of (11) we have used the square root transformation and now we could take into account the special role of the root density $f^{1/2}$ related to the Hellinger distance. This suggests as a valid Gaussian approximation of the laws of family $\mathcal{P}_n$ a signal model with homoscedastic white noise where the signal is $f^{1/2}$.

3.1 Approximation by a signal model with white noise

Taking the corresponding statement of the theorem (see [18]). Consider for $\alpha \in (0,1)$, $M > 0$ a class of Hölder functions

$$\mathcal{H}^\alpha(M) = \{f : \ |f(x) - f(y)| \leq M \ |x - y|^\alpha\}. \tag{20}$$

For α, M, and given $\varepsilon > 0$, define the parameter set

$$\Sigma_d(\alpha, M, \varepsilon) = \mathcal{H}^\alpha(M) \cap \{\text{densities on } [0,1], \text{ bounded from below by } \varepsilon\}.$$

Theorem 1 *Let $\Sigma = \Sigma_d(\alpha, M, \varepsilon)$ for $\varepsilon > 0$, $M > 0$ and $\alpha > 1/2$ fixed. Then, the experiments given by the observations*

$$\begin{aligned} &X_i,\ i = 1, \ldots, n \text{ independent, with density } f \\ &dy(t) = f^{1/2}(t)dt + \frac{1}{2}n^{-1/2}dW(t),\ t \in [0,1], \end{aligned} \tag{21}$$

where $f \in \Sigma$, are asymptotically equivalent.

The proof of this result depends on the relation (18) on the likelihood process, where it must be clarified the concept of approximation in law $\overset{\mathcal{L}}{\approx}$ for two sequences. For this we use the coupling, that is to say, the construction of processes on the same probability space that are close to each other in a metric sense. Markov kernels carrying the asymptotic equivalence are not given explicitly, the method is very indirect. More recent work (Carter [4], Brown, Carter, Low and Zhang [2]) have managed to redo the proof by exhibiting Markov kernels that could carry out a smoothing of the empirical process as in (12).

It should be noted that a precursor of Theorem 1 for the parametric models has already been shown by Le Cam in [14]. This is related to the densities of sets Σ which are finite dimensional in the Hellinger metric. This case essentially boils down to that of a parametric family of densities f_ϑ, $\vartheta \in \Theta \subset \mathbb{R}^k$, and the model of white Gaussian noise (21) with signal $f_\vartheta^{1/2}$ could be understood as a result of stabilization of the variance in (13). There are other variants of the overall Gaussian approximation, see Pfanzagl [19].

A more immediate precursor of Theorem 1 was the result of Brown and Low [1] on the relation between a template signal with white noise and its discretized version, that is to say with the Gaussian nonparametric regression. Let f be a function on $[0,1]$ belongs to a class Hölder $\mathcal{H}^\alpha(M)$ (cf. (20)) where $\alpha > 1/2$. Then the asymptotic equivalence occurs between the two models

$$\begin{aligned} &Y_i,\ i = 1, \ldots, n \text{ independent, with law } N(f(i/n), 1) \\ &dy(t) = f(t)dt + n^{-1/2}dW(t),\ t \in [0,1]. \end{aligned}$$

The elegant proof of this result is based on the sufficiency of Gaussian models and it is constructive in the sense of Markov kernels. Enfin, Brown, Zhang [3] have shown that for Hölder classes, the bound $\alpha > 1/2$ is true in Theorem 1 in the nonparametric Gaussian regression above, giving counterexamples in the case $\alpha = 1/2$. The case of non-Gaussian regression has been discussed in [10], [11]. Moreover, it has been shown that the result of Theorem 1 is reproduced in the case where the model of independent equidistributed variables is replaced by a diffusion process (see [8]).

Let $y(1), \ldots, y(n)$ be n observations of a stationary Gaussian process such that $Ey(1) = 0$, with autocovariance function

$$\gamma(h) = Ey(t)y(t+h) = \int_{-\pi}^{\pi} \exp(\mathrm{i}\, h\omega)\, f(\omega)\, d\omega$$

where f is the spectral density defined on $[-\pi, \pi]$. The function f is nonnegative, symmetric ($f(\omega) = f(-\omega)$) and we suppose in addition that $f \in L_2[-\pi, \pi]$. Again, we consider the

nonparametric case, that is to say, we assume that $f \in \Sigma$, a class of regular functions of infinite dimension. The traditional topics such as estimation of f, the speed of convergence optimality etc. have been studied in detail. In addition, the property of local asymptotic normality (LAN) was established in the parametric framework, see Davies [5] and Dzhaparidze [6] where the Fisher information (the term J_ϑ in (19)) is determined as follows. Consider a family of regular spectral densities: f_ϑ, $\vartheta \in \Theta \subset \mathbb{R}$: then

$$J_\vartheta = \frac{1}{4\pi} \int_{-\pi}^{\pi} \left(\frac{\partial}{\partial \vartheta} \log f_\vartheta(\omega) \right)^2 d\omega.$$

This suggests a model of signal with white noise

$$dZ_\omega = \log f_\vartheta(\omega) d\omega + 2\pi^{1/2} n^{-1/2} dW_\omega,\ \omega \in [-\pi, \pi]$$

where $\vartheta \in \Theta \subset \mathbb{R}$, for all the regular parametric families. In fact, the latter model has the same asymptotic Fisher information.

Taking the statement of the corresponding theorem (see [9]). Consider, for $\alpha \in (0,1)$ and $M > 0$, a class of Hölder functions (20). For α, M and $\varepsilon \in (0,1)$ given, define a parameter space

$$\Sigma_s(\alpha, M, \varepsilon) = \mathcal{H}^\alpha(M) \cap \left\{ \text{functions } f \text{ on } [0,1] \text{ with values in } (\varepsilon, \varepsilon^{-1}) \right\}.$$

Theorem 2 *Let $\Sigma = \Sigma_s(\alpha, M, \varepsilon)$ for $\varepsilon \in (0,1)$, $M > 0$ and $\alpha > 1/2$ given. Let ω_j, $j = 1, \ldots, n$ be a grid of points equally spaced in $[-\pi, \pi]$. Then the three experiments given by the observations*

$$\begin{aligned}
&Y_i,\ i = 1, \ldots, n,\ \textit{stationary, centered, Gaussian with spectral density } f \\
&Z_i,\ i = 1, \ldots, n\ \textit{independents, with the law } N(0, f(\omega_i)) \\
&dZ_\omega = \log f(\omega) d\omega + 2\pi^{1/2} n^{-1/2} dW_\omega, \omega \in [-\pi, \pi],
\end{aligned}$$

where $f \in \Sigma$, are asymptotically equivalent.

This result has two components. The first one makes the reduction of the stationary sequence to a model of independent Gaussian variables. This is a non-parametric Gaussian scale model of Example 2. In this context, the second part of the theorem is similar to Proposition 3. The known proofs so far are indirect in the sense of the existence of Markov kernels (see [9]).

References

[1] Brown, L. D. and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384-2398

[2] Brown, L. D., Carter, A. V., Low, M. G. and Zhang, C.-H. (2004). Equivalence theory for density estimation, Poisson processes and Gaussian white noise with drift. To appear, *Ann. Statist.* **32** (5)

[3] Brown, L. D. and Zhang, C.-H. (1998). Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist.* **26**, 279-287.

[4] Carter, A. (2002). Deficiency distance between multinomial and multivariate normal experiments. *Ann. Statist.* **30** 708-730

[5] Davies, R.B. (1973). Asymptotic inference in stationary Gaussian time-series, *Adv. Appl. Probab.* **5**, 469–497.

[6] Dzhaparidze K. (1986). *Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series*. Springer-Verlag, New York Inc

[7] Genon-Catalot, V. et Picard, D. (1993). *Eléments de Statistique Asymptotique*. Mathématiques et Applications 11, Springer Verlag, Paris

[8] Genon-Catalot, V., Larédo, C., Nussbaum, M. (2002). Asymptotic equivalence of estimating a Poisson intensity and a positive diffusion drift. *Ann. Statist.* **30** 731-753

[9] Golubev, G., Nussbaum, M. and Zhou. H. (2004). Asymptotic equivalence of spectral density estimation and Gaussian white noise. En préparation.

[10] Grama, I. and Nussbaum, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Prob. Theor. Rel. Fields*, **111**, 167-214

[11] Grama, I and Nussbaum, M., (2002). Asymptotic equivalence for nonparametric regression. *Math. Meth. Statist.* **11** (1) 1-36

[12] Brown, L. D., and Low, M. (1996). Asymptotic equivalence of nonparametric regression and white noise, *Ann. Statist.* **24** 2384-2398 (1996)

[13] Le Cam , L. (1969). *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal.

[14] Le Cam, L. (1985). Sur l'approximation de familles de mesures par des familles gaussiennes. *Ann. Inst. Henri Poincaré* **21** (3) 225-287

[15] Le Cam , L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

[16] Le Cam, L. and Yang, G. (2000). *Asymptotics in Statistics, 2nd ed.*. Springer-Verlag, New-York.

[17] Müller, D. W. (1981). The increase of risk due to inaccurate models. *Symposia Mathematica*. Instituto Nazionale di Alta Mathematica, Vol. 25.

[18] Nussbaum, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24**, 2399–2430.

[19] Pfanzagl, J. (1995). On local and global asymptotic normality. *Math. Meth. Statist.* **4** 115-136

[20] Shiryaev, A. N. and Spokoiny, V. (2000). *Statistical Experiments and Decisions: Asymptotic Theory*. World Scientific, Singapore.

[21] Strasser, H. (1985). *Mathematical Theory of Statistics*. de Gruyter, Berlin.

[22] van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

[23] van der Vaart, A. W. (2002). The statistical work of Lucien Le Cam. *Ann. Statist.* **30** 631-682

[24] Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Trans. Amer. Math. Soc.* 54 426-482

Department of Mathematics
Malott Hall
Cornell University
Ithaca, NY 14853-4201
USA
e-mail nussbaum@math.cornell.edu