9

Limits of Experiments

A sequence of experiments is defined to converge to a limit experiment if the sequence of likelihood ratio processes converges marginally in distribution to the likelihood ratio process of the limit experiment. A limit experiment serves as an approximation for the converging sequence of experiments. This generalizes the convergence of locally asymptotically normal sequences of experiments considered in Chapter 7. Several examples of nonnormal limit experiments are discussed.

9.1 Introduction

This chapter introduces a notion of convergence of statistical models or "experiments" to a limit experiment. In this notion a sequence of models, rather than just a sequence of estimators or tests, converges to a limit. The limit experiment serves two purposes. First, it provides an absolute standard for what can be achieved asymptotically by a sequence of tests or estimators, in the form of a "lower bound": No sequence of statistical procedures can be asymptotically better than the "best" procedure in the limit experiment. For instance, the best limiting power function is the best power function in the limit experiment; a best sequence of estimators converges to a best estimator in the limit experiment. Statements of this type are true irrespective of the precise meaning of "best." A second purpose of a limit experiment is to explain the asymptotic behaviour of sequences of statistical procedures. For instance, the asymptotic normality or (in)efficiency of maximum likelihood estimators.

Many sequences of experiments converge to normal limit experiments. In particular, the local experiments in a given locally asymptotically normal sequence of experiments, as considered in Chapter 7, converge to a normal location experiment. The asymptotic representation theorem given in the present chapter is therefore a generalization of Theorem 7.10 (for the LAN case) to the general situation. The importance of the general concept is illustrated by several examples of non-Gaussian limit experiments.

In the present context it is customary to speak of "experiment" rather than model, although these terms are interchangeable. Formally an *experiment* is a measurable space $(\mathcal{X}, \mathcal{A})$, the *sample space*, equipped with a collection of probability measures $(P_h : h \in H)$. The set of probability measures serves as a statistical model for the observation, written as X. In this chapter the parameter is denoted by h (and not θ), because the results are typically applied to "local" parameters (such as $h = \sqrt{n}(\theta - \theta_0)$). The experiment is denoted

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

by $(\mathcal{X}, \mathcal{A}, P_h : h \in H)$ and, if there can be no misunderstanding about the sample space, also by $(P_h : h \in H)$.

Given a fixed parameter $h_0 \in H$, the *likelihood ratio process* with base h_0 is formed as

$$\left(\frac{dP_h}{dP_{h_0}}(X)\right)_{h\in H} \equiv \left(\frac{p_h}{p_{h_0}}(X)\right)_{h\in H}.$$

Each likelihood ratio process is a (typically infinite-dimensional) vector of random variables $dP_h/dP_{h_0}(X)$. According to the results of section 6.1, the right side of the display is P_{h_0}-almost surely the same for any given densities p_h and p_{h_0} with respect to any measure μ. Because we are interested only in the laws under P_{h_0} of finite subvectors of the likelihood processes, the nonuniqueness is best left unresolved.

9.1 Definition. A sequence $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ of experiments *converges to a limit experiment* $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$ if, for every finite subset $I \subset H$ and every $h_0 \in H$,

$$\left(\frac{dP_{n,h}}{dP_{n,h_0}}(X_n)\right)_{h\in I} \overset{h_0}{\rightsquigarrow} \left(\frac{dP_h}{dP_{h_0}}(X)\right)_{h\in I}.$$

The objects in this display are random vectors of length $|I|$. The requirement is that each of these vectors converges in law, under the assumption that h_0 is the true parameter, in the ordinary sense of convergence in distribution in $\mathbb{R}^I$. This type of convergence is sometimes called *marginal weak convergence*: The finite-dimensional marginal distributions of the likelihood processes converge in distribution to the corresponding marginals in the limit experiment.

Because a weak limit of a sequence of random vectors is unique, the marginal distributions of the likelihood ratio process of a limit experiment are unique. The limit experiment itself is not unique; even its sample space is not uniquely determined. This causes no problems. Two experiments of which the likelihood ratio processes are equal in marginal distributions are called *equivalent* or of the same type. Many examples of equivalent experiments arise through sufficiency.

9.2 ***Example (Equivalence by sufficiency).*** Let $S : \mathcal{X} \mapsto \mathcal{Y}$ be a statistic in the statistical experiment $(\mathcal{X}, \mathcal{A}, P_h : h \in H)$ with values in the measurable space $(\mathcal{Y}, \mathcal{B})$. The experiment of image laws $(\mathcal{Y}, \mathcal{B}, P_h \circ S^{-1} : h \in H)$ corresponds to observing S. If S is a sufficient statistic, then this experiment is equivalent to the original experiment $(\mathcal{X}, \mathcal{A}, P_h : h \in H)$. This may be proved using the Neyman factorization criterion of sufficiency. This shows that there exist measurable functions g_h and f such that $p_h(x) = g_h\big(S(x)\big)f(x)$, so that the likelihood ratio $p_h/p_{h_0}(X)$ is the function $g_h/g_{h_0}(S)$ of S. The likelihood ratios of the measures $P_h \circ S^{-1}$ take the same form.

Consequently, if $(P_h : h \in H)$ is a limit experiment, then so is $(P_h \circ S^{-1} : h \in H)$. A very simple example that we encounter frequently is as follows: For a given invertible matrix J the experiments $\big(N(Jh, J) : h \in \mathbb{R}^d\big)$ and $\big(N(h, J^{-1}) : h \in \mathbb{R}^d\big)$ are equivalent. □

9.2 Asymptotic Representation Theorem

In this section it is shown that a limit experiment is always statistically easier than a given sequence. Suppose that a sequence of statistical problems involves experiments

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

$\mathcal{E}_n = (P_{n,h} : h \in H)$ and statistics T_n. For instance, the statistics are test statistics for testing certain hypotheses concerning the parameter h, or estimators of some function of h. Most of the quality measures of the procedures based on the statistics T_n can be expressed in their laws under the different parameters. For simplicity we assume that the sequence of statistics T_n converges under a given parameter h in distribution to a limit L_h, for every parameter h. Then the asymptotic quality of the sequence T_n may be judged from the set of limit laws $\{L_h : h \in H\}$. According to the following theorem the only possible sets of limit laws are the laws of randomized statistics in the limit experiment: Every weakly converging sequence of statistics converges to a statistic in the limit experiment. One consequence is that asymptotically no sequence of statistical procedures can be better than the best procedure in the limit experiment. This is true for every meaning of "good" that is expressible in terms of laws. In this way the limit experiment obtains the character of an asymptotic lower bound.

We assume that the limit experiment $\mathcal{E} = (P_h : h \in H)$ is *dominated*: This requires the existence of a σ-finite measure μ such that $P_h \ll \mu$ for every h. Recall that a *randomized statistic* T in the experiment $(\mathcal{X}, \mathcal{A}, P_h : h \in H)$ with values in $\mathbb{R}^k$ is a measurable map $T : \mathcal{X} \times [0, 1] \mapsto \mathbb{R}^k$ for the product σ-field $\mathcal{A} \times$ Borel sets on the space $\mathcal{X} \times [0, 1]$. Its law under h is to be computed under the product measure $P_h \times \text{uniform}[0, 1]$.

9.3 *Theorem.* *Let $\mathcal{E}_n = (\mathcal{X}_n, \mathcal{A}_n, P_{n,h} : h \in H)$ be a sequence of experiments that converges to a dominated experiment $\mathcal{E} = (\mathcal{X}, \mathcal{A}, P_h : h \in H)$. Let T_n be a sequence of statistics in $\mathcal{E}_n$ that converges in distribution for every h. Then there exists a randomized statistic T in $\mathcal{E}$ such that $T_n \overset{h}{\rightsquigarrow} T$ for every h.*

Proof. The proof of the theorem starting from the definition of convergence of experiments is long and can best be broken up into parts of independent interest. This goes beyond the scope of this book.

The proof for the case of local asymptotic normal sequences of experiments is given in Chapter 7. (It is shown in Theorem 9.4 that such a sequence of experiments converges to a Gaussian location experiment.) Many other examples can be treated by the same method of proof.† ■

9.3 Asymptotic Normality

As in much of statistics, normal limits are of prime importance. In Chapter 7 a sequence of statistical models $(P_{n,\theta} : \theta \in \Theta)$ indexed by an open subset $\Theta \subset \mathbb{R}^d$ is defined to be locally asymptotically normal at θ if the log likelihood ratios $\log dP_{n,\theta+r_n^{-1}h_n}/dP_{n,\theta}$ allow a certain quadratic expansion. This is shown to be valid in the case that $P_{n,\theta}$ is the distribution of a sample of size n from a smooth parametric model. Such experiments converge to simple normal limit experiments if they are reparametrized in terms of the "local parameter" h. This follows from the following theorem.

9.4 *Theorem.* *Let $\mathcal{E}_n = (P_{n,h} : h \in H)$ be a sequence of experiments indexed by a subset H of $\mathbb{R}^d$ (with $0 \in H$) such that*

$$\log \frac{dP_{n,h}}{dP_{n,0}} = h^T \Delta_n - \frac{1}{2} h^T J h + o_{P_{n,0}}(1),$$

† For a proof of the general theorem see, for instance, [141].

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

for a sequence of statistics Δ_n that converges weakly under $h = 0$ to a $N(0, J)$-distribution. Then the sequence $\mathcal{E}_n$ converges to the experiment $\big(N(Jh, J) : h \in H\big)$.

Proof. The log likelihood ratio process with base h_0 for the normal experiment has coordinates

$$\log \frac{dN(Jh, J)}{dN(Jh_0, J)}(X) = (h - h_0)^T X - \frac{1}{2}h^T Jh + \frac{1}{2}h_0^T Jh_0.$$

If J is nonsingular, then this follows by simple algebra, because the left side is the quotient of two normal densities. The case that J is singular perhaps requires some thought.

By the assumption combined with Slutsky's lemma, the sequence $\log p_{n,h}/p_{n,0}$ is under $h = 0$ asymptotically normal with mean $-\frac{1}{2}h^T Jh$ and variance $h^T Jh$). This implies contiguity of the sequences of measures $P_{n,h}$ and $P_{n,0}$ for every h, by Example 6.5. Therefore, the probability of the set on which one of $p_{n,0}$, $p_{n,h}$, or p_{n,h_0} is zero converges to zero. Outside this set we can write

$$\log \frac{p_{n,h}}{p_{n,h_0}} = \log \frac{p_{n,h}}{p_{n,0}} - \log \frac{p_{n,h_0}}{p_{n,0}}.$$

Because this is true with probability tending to 1, the difference between the left and the right sides converges to zero in probability. Apply the (local) asymptotic normality assumption twice to obtain that

$$\log \frac{p_{n,h}}{p_{n,h_0}} = (h - h_0)^T \Delta_n - \frac{1}{2}h^T Jh + \frac{1}{2}h_0^T Jh_0 + o_{P_{n,h_0}}(1).$$

On comparing this to the expression for the normal likelihood ratio process, we see that it suffices to show that the sequence Δ_n converges under h_0 in law to X: In that case the vector $(p_{n,h}/p_{n,h_0})_{h\in I}$ converges in distribution to $\big(dN(Jh, J)/dN(0, J)(X)\big)_{h\in I}$, by Slutsky's lemma and the continuous-mapping theorem.

By assumption, the sequence $(\Delta_n, h_0^T \Delta_n)$ converges in distribution under $h = 0$ to a vector $(\Delta, h_0^T \Delta)$, where Δ is $N(0, J)$-distributed. By local asymptotic normality and Slutsky's lemma, the sequence of vectors $(\Delta_n, \log p_{n,h_0}/p_{n,0})$ converges to the vector $(\Delta, h_0^T \Delta - \frac{1}{2}h_0^T Jh_0)$. In other words

$$\left(\Delta_n, \log \frac{p_{n,h_0}}{p_{n,0}}\right) \overset{0}{\rightsquigarrow} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}h_0^T Jh_0 \end{pmatrix}, \begin{pmatrix} J & Jh_0 \\ h_0^T J & h_0^T Jh_0 \end{pmatrix}\right).$$

By the Gaussian form of Le Cam's third lemma, Example 6.5, the sequence Δ_n converges in distribution under h_0 to a $N(Jh_0, J)$-distribution. This is equal to the distribution of X under h_0. ■

9.5 ***Corollary.*** *Let Θ be an open subset of $\mathbb{R}^d$, and let the sequence of statistical models $(P_{n,\theta} : \theta \in \Theta)$ be locally asymptotically normal at θ with norming matrices r_n and a nonsingular matrix I_θ. Then the sequence of experiments $(P_{n,\theta+r_n^{-1}h} : h \in \mathbb{R}^d)$ converges to the experiment $\big(N(h, I_\theta^{-1}) : h \in \mathbb{R}^d\big)$.*

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

9.4 Uniform Distribution

The model consisting of the uniform distributions on $[0, \theta]$ is not differentiable in quadratic mean (see Example 7.9.) In this case an asymptotically normal approximation is impossible. Instead, we have convergence to an exponential experiment.

9.6 *Theorem.* *Let P_θ^n be the distribution of a random sample of size n from a uniform distribution on $[0, \theta]$. Then the sequence of experiments $(P_{\theta - h/n}^n : h \in \mathbb{R})$ converges for each fixed $\theta > 0$ to the experiment consisting of observing one observation from the shifted exponential density $z \mapsto e^{-(z-h)/\theta} 1\{z > h\}/\theta$.*[†]

Proof. If Z is distributed according to the given exponential density, then

$$\frac{dP_h^Z}{dP_{h_0}^Z}(Z) = \frac{e^{-(Z-h)/\theta} 1\{Z > h\}/\theta}{e^{-(Z-h_0)/\theta} 1\{Z > h_0\}/\theta} = e^{(h-h_0)/\theta} 1\{Z > h\},$$

almost surely under h_0, because the indicator $1\{z > h_0\}$ in the denominator equals 1 almost surely if h_0 is the true parameter.

The joint density of a random sample $X_1, \ldots, X_n$ from the uniform $[0, \theta]$ distribution can be written in the form $(1/\theta)^n 1\{X_{(n)} \le \theta\}$. The likelihood ratios take the form

$$\frac{dP_{\theta - h/n}^n}{dP_{\theta - h_0/n}^n}(X_1, \ldots, X_n) = \frac{(\theta - h/n)^{-n} 1\{X_{(n)} \le \theta - h/n\}}{(\theta - h_0/n)^{-n} 1\{X_{(n)} \le \theta - h_0/n\}}.$$

Under the parameter $\theta - h_0/n$, the maximum of the observations is certainly bounded above by $\theta - h_0/n$ and the indicator in the denominator equals 1. Thus, with probability 1 under $\theta - h_0/n$, the likelihood ratio in the preceding display can be written

$$\left(e^{(h-h_0)/\theta} + o(1)\right) 1\left\{-n(X_{(n)} - \theta) \ge h\right\}.$$

By direct calculation, $-n(X_{(n)} - \theta) \overset{h_0}{\rightsquigarrow} Z$. By the continuous-mapping theorem and Slutsky's lemma, the sequence of likelihood processes converges under $\theta - h_0/n$ marginally in distribution to the likelihood process of the exponential experiment. ■

Along the same lines it may be proved that in the case of uniform distributions with both endpoints unknown a limit experiment based on observation of two independent exponential variables pertains. These types of experiments are completely determined by the discontinuities of the underlying densities at their left and right endpoints. It can be shown more generally that exponential limit experiments are obtained for any densities that have jumps at one or both of their endpoints and are smooth in between. For densities with discontinuities in the middle, or weaker singularities, other limit experiments pertain.

The convergence to a limit experiment combined with the asymptotic representation theorem, Theorem 9.3, allows one to obtain asymptotic lower bounds for sequences of estimators, much as in the locally asymptotically normal case in Chapter 8. We give only one concrete statement.

[†] Define P_θ arbitrarily for $\theta < 0$.

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

9.7 ***Corollary.*** *Let T_n be estimators based on a sample $X_1, \ldots, X_n$ from the uniform distribution on $[0, \theta]$ such that the sequence $n(T_n - \theta)$ converges under θ in distribution to a limit L_θ, for every θ. Then for Lebesgue almost-every θ we have $\int |x|\, dL_\theta(x) \geq \mathrm{E}|Z - \operatorname{med} Z|$ and $\int x^2\, dL_\theta(x) \geq \mathrm{E}(Z - \mathrm{E}Z)^2$ for the random variable Z exponentially distributed with mean θ.*

Proof (Sketch). By Lemma 8.10, the estimator sequence T_n is automatically almost regular in the sense that $n(T_n - \theta + h/n)$ converges under $\theta - h/n$ in distribution to L_θ for Lebesgue almost every θ and h, at least along a subsequence. Thus, it is matched in the limit experiment by an equivariant-in-law estimator for almost every θ. More precisely, for almost every θ there exists a randomized statistic T_θ such that the law of $T_\theta(Z + h, U) - h$ does not depend on h (if Z is exponentially distributed with mean θ). By classical statistical decision theory the given lower bounds are the (constant) risks of the best equivariant-in-law estimators in the exponential limit experiment in terms of absolute error and mean-square error loss functions, respectively. ■

In view of this lemma, the maximum likelihood estimator $X_{(n)}$ is asymptotically inefficient. This is not surprising given its bias downwards, but it is encouraging for the present approach that the small bias, which is of the order $1/n$, is visible in the "first-order" asymptotics. The bias can be corrected by a multiplicative factor, which, unfortunately, must depend on the loss function. The sequences of estimators

$$\frac{n + \log 2}{n} X_{(n)} \quad \text{and} \quad \frac{n+1}{n} X_{(n)}$$

are asymptotically efficient in terms of absolute value and quadratic loss, respectively.

9.5 Pareto Distribution

The Pareto distributions are a two-parameter family of distributions on the real line with parameters $\alpha > 0$ and $\mu > 0$ and density

$$x \mapsto \frac{\alpha \mu^\alpha}{x^{\alpha+1}} 1\{x > \mu\}.$$

This density is smooth in α, but it resembles a uniform distribution as discussed in the preceding section in its dependence on μ. The limit experiment consists of a combination of a normal experiment and an exponential experiment.

The likelihood ratios for a sample of size n from the Pareto distributions with parameters $(\alpha + g/\sqrt{n}, \mu + h/n)$ and $(\alpha + g_0/\sqrt{n}, \mu + h_0/n)$, respectively, is equal to

$$\left(\frac{\alpha + g/\sqrt{n}}{\alpha + g_0/\sqrt{n}}\right)^n \frac{(\mu + h/n)^{n\alpha + \sqrt{n}g}}{(\mu + h_0/n)^{n\alpha + \sqrt{n}g_0}} \left(\prod_{i=1}^n X_i\right)^{(g_0 - g)/\sqrt{n}} 1\left\{X_{(1)} > \mu + \frac{h}{n}\right\}$$

$$= \exp\left((g - g_0)\Delta_n - \frac{1}{2}\frac{g^2 + g_0^2}{\alpha^2} + o(1)\right)\left(e^{(h - h_0)\alpha/\mu} + o(1)\right) 1\{Z_n > h\}.$$

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

Here, under the parameters $(\alpha + g_0/\sqrt{n}, \mu + h_0/n)$, the sequence

$$\Delta_n = -\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\log \frac{X_i}{\mu} - \frac{1}{\alpha} \right)$$

converges weakly to a normal distribution with mean g_0/α^2 and variance $1/\alpha^2$; and the sequence $Z_n = n(X_{(1)} - \mu)$ converges in distribution to the (shifted) exponential distribution with mean $\mu/\alpha + h_0$ and variance $(\mu/\alpha)^2$. The two sequences are asymptotically independent. Thus the likelihood is a product of a locally asymptotically normal and a "locally asymptotically exponential" factor. The local limit experiment consists of observing a pair (Δ, Z) of independent variables Δ and Z with a $N(g, \alpha^2)$-distribution and an $\exp(\alpha/\mu) + h$-distribution, respectively.

The maximum likelihood estimators for the parameters α and μ are given by

$$\hat{\alpha}_n = \frac{n}{\sum_{i=1}^{n} \log(X_i/X_{(1)})}, \quad \text{and} \quad \hat{\mu}_n = X_{(1)}.$$

The sequence $\sqrt{n}(\hat{\alpha}_n - \alpha)$ converges in distribution under the parameters $(\alpha + g/\sqrt{n}, \mu + h/n)$ to the variable $\Delta - g$. Because the distribution of Z does not depend on g, and Δ follows a normal location model, the variable Δ can be considered an optimal estimator for g based on the observation (Δ, Z). This optimality is carried over into the asymptotic optimality of the maximum likelihood estimator $\hat{\alpha}_n$. A precise formulation could be given in terms of a convolution or a minimax theorem.

On the other hand, the maximum likelihood estimator for μ is asymptotically inefficient. Because the sequence $n(\hat{\mu}_n - \mu - h/n)$ converges in distribution to $Z - h$, the estimators $\hat{\mu}_n$ are asymptotically biased upwards.

9.6 Asymptotic Mixed Normality

The likelihood ratios of some models allow an approximation by a two-term Taylor expansion without the linear term being asymptotically normal and the quadratic term being deterministic. Then a generalization of local asymptotic normality is possible. In the most important example of this situation, the linear term is asymptotically distributed as a mixture of normal distributions.

A sequence of experiments $(P_{n,\theta} : \theta \in \Theta)$ indexed by an open subset Θ of $\mathbb{R}^d$ is called *locally asymptotically mixed normal* at θ if there exist matrices $\gamma_{n,\theta} \to 0$ such that

$$\log \frac{dP_{n,\theta+\gamma_{n,\theta}h_n}}{dP_{n,\theta}} = h^T \Delta_{n,\theta} - \frac{1}{2} h^T J_{n,\theta} h + o_{P_{n,\theta}}(1),$$

for every converging sequence $h_n \to h$, and random vectors $\Delta_{n,\theta}$ and random matrices $J_{n,\theta}$ such that $(\Delta_{n,\theta}, J_{n,\theta}) \overset{\theta}{\rightsquigarrow} (\Delta_\theta, J_\theta)$ for a random vector such that the conditional distribution of Δ_θ given that $J_\theta = J$ is normal $N(0, J)$.

Locally asymptotically mixed normal is often abbreviated to LAMN. Locally asymptotically normal, or LAN, is the special case in which the matrix J_θ is deterministic. Sequences of experiments whose likelihood ratios allow a quadratic approximation as in the preceding display (but without the specific limit distribution of $(\Delta_{n,\theta}, J_{n,\theta})$) and that are

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

such that $P_{n,\theta+\gamma_{n,\theta}h} \lhd\rhd P_{n,\theta}$ are called *locally asymptotically quadratic*, or LAQ. We note that LAQ or LAMN requires much more than the mere existence of two derivatives of the likelihood: There is no reason why, in general, the remainder would be negligible.

9.8 ***Theorem.*** *Assume that the sequence of experiments $(P_{n,\theta}: \theta \in \Theta)$ is locally asymptotically mixed normal at θ. Then the sequence of experiments $(P_{n,\theta+\gamma_{n,\theta}h}: h \in \mathbb{R}^d)$ converges to the experiment consisting of observing a pair (Δ, J) such that J is marginally distributed as J_θ for every h and the conditional distribution of Δ given J is normal $N(Jh, J)$.*

Proof. Write $P_{\theta,h}$ for the distribution of (Δ, J) under h. Because the marginal distribution of J does not depend on h and the conditional distribution of Δ given J is Gaussian

$$\frac{dP_{\theta,h}}{dP_{\theta,h_0}}(\Delta, J) = \frac{dN(Jh, J)}{dN(Jh_0, J)}(\Delta) = e^{(h-h_0)^T\Delta - \frac{1}{2}h^TJh + \frac{1}{2}h_0^TJh_0}.$$

By Slutsky's lemma and the assumptions, the sequence $dP_{n,\theta+\gamma_{n,\theta}h}/dP_{n,\theta}$ converges under θ in distribution to $\exp(h^T\Delta_\theta - \frac{1}{2}h^TJ_\theta h)$. Because the latter variable has mean one, it follows that the sequences of distributions $P_{n,\theta+\gamma_{n,\theta}h}$ and $P_{n,\theta}$ are mutually contiguous. In particular, the probability under θ that $dP_{n,\theta+\gamma_{n,\theta}h}$ is zero converges to zero for every h, so that

$$\begin{aligned}\log\frac{dP_{n,\theta+\gamma_{n,\theta}h}}{dP_{n,\theta+\gamma_{n,\theta}h_0}} &= \log\frac{dP_{n,\theta+\gamma_{n,\theta}h}}{dP_{n,\theta}} - \log\frac{dP_{n,\theta+\gamma_{n,\theta}h_0}}{dP_{n,\theta}} + o_{P_{n,\theta}}(1)\\ &= (h-h_0)^T\Delta_{n,\theta} - \frac{1}{2}h^TJ_{n,\theta}h + \frac{1}{2}h_0^TJ_{n,\theta}h_0 + o_{P_{n,\theta}}(1).\end{aligned}$$

Conclude that it suffices to show that the sequence $(\Delta_{n,\theta}, J_{n,\theta})$ converges under $\theta + \gamma_{n,\theta}h_0$ to the distribution of (Δ, J) under h_0.

Using the general form of Le Cam's third lemma we obtain that the limit distribution of the sequence $(\Delta_{n,\theta}, J_{n,\theta})$ under $\theta + \gamma_{n,\theta}h$ takes the form

$$L_h(B) = \mathrm{E}1_B(\Delta_\theta, J_\theta)e^{h^T\Delta_\theta - \frac{1}{2}h^TJ_\theta h}.$$

On noting that the distribution of (Δ, J) under $h = 0$ is the same as the distribution of $(\Delta_\theta, J_\theta)$, we see that this is equal to $\mathrm{E}_0 1_B(\Delta, J)\, dP_{\theta,h}/dP_{\theta,0}(\Delta, J) = \mathrm{P}_h\big((\Delta, J) \in B\big)$. ■

It is possible to develop a theory of asymptotic "lower bounds" for LAMN models, much as is done for LAN models in Chapter 8. Because conditionally on the ancillary statistic J, the limit experiment is a Gaussian shift experiment, the lower bounds take the form of mixtures of the lower bounds for the LAN case. We give only one example, leaving the details to the reader.

9.9 ***Corollary.*** *Let T_n be an estimator sequence in a LAMN sequence of experiments $(P_{n,\theta}: \theta \in \Theta)$ such that $\gamma_{n,\theta}^{-1}\big(T_n - \psi(\theta + \gamma_{n,\theta}h)\big)$ converges weakly under every $\theta + \gamma_{n,\theta}h$ to a limit distribution L_θ, for every h. Then there exist probability distributions M_j (or rather a Markov kernel) such that $L_\theta = \mathrm{E}N(0, \dot\psi_\theta J_\theta^{-1}\dot\psi_\theta^T) * M_{J_\theta}$. In particular,* $\mathrm{cov}_\theta\, L_\theta \geq \mathrm{E}\dot\psi_\theta J_\theta^{-1}\dot\psi_\theta^T$.

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

We include two examples to give some idea of the application of local asymptotic mixed normality. In both examples the sequence of models is LAMN rather than LAN due to an explosive growth of information, occurring at certain supercritical parameter values. The second derivative of the log likelihood, the information, remains random. In both examples there is also (approximate) Gaussianity present in every single observation. This appears to be typical, unlike the situation with LAN, in which the normality results from sums over (approximately) independent observations. In explosive models of this type the likelihood is dominated by a few observations, and normality cannot be brought in through (martingale) central limit theorems.

9.10 ***Example (Branching processes).*** In a Galton-Watson branching process the "nth generation" is formed by replacing each element of the $(n-1)$-th generation by a random number of elements, independently from the rest of the population and from the preceding generations. This random number is distributed according to a fixed distribution, called the *offspring distribution*. Thus, conditionally on the size X_{n-1} of the $(n-1)$th generation the size X_n of the nth generation is distributed as the sum of X_{n-1} i.i.d. copies of an offspring variable Z. Suppose that $X_0 = 1$, that we observe $(X_1, \ldots, X_n)$, and that the offspring distribution is known to belong to an exponential family of the form

$$\mathrm{P}_\theta(Z = z) = a_z\, \theta^z c(\theta), \quad z = 0, 1, 2, \ldots,$$

for given numbers $a_0, a_1, \ldots$. The natural parameter space is the set of all θ such that $c(\theta)^{-1} = \sum_z a_z \theta^z$ is finite (an interval). We shall concentrate on parameters in the interior of the natural parameter space such that $\mu(\theta) := \mathrm{E}_\theta Z > 1$. Set $\sigma^2(\theta) = \mathrm{var}_\theta\, Z$.

The sequence $X_1, X_2, \ldots$ is a Markov chain with transition density

$$p_\theta(y \mid x) = \mathrm{P}_\theta(X_n = y \mid X_{n-1} = x) = \overbrace{a * \cdots * a}^{x \text{ times}}\, \theta^y c(\theta)^x.$$

To obtain a two-term Taylor expansion of the log likelihood ratios, let $\ell_\theta(y \mid x)$ be the log transition density, and calculate that

$$\dot\ell_\theta(y \mid x) = \frac{y - x\mu(\theta)}{\theta}, \qquad \ddot\ell_\theta(y \mid x) = -\frac{y - x\mu(\theta)}{\theta^2} - \frac{x\dot\mu(\theta)}{\theta}.$$

(The fact that the score function of the model $\theta \mapsto \mathrm{P}_\theta(Z = z)$ has derivative zero yields the identity $\mu(\theta) = -\theta(c/\dot c)(\theta)$, as is usual for exponential families.) Thus, the Fisher information in the observation $(X_1, \ldots, X_n)$ equals (note that $\mathrm{E}_\theta(X_j \mid X_{j-1}) = X_{j-1}\mu(\theta)$)

$$\begin{aligned} -\mathrm{E}_\theta \sum_{j=1}^n \ddot\ell_\theta(X_j \mid X_{j-1}) &= \mathrm{E}_\theta \sum_{j=1}^n X_{j-1} \frac{\dot\mu(\theta)}{\theta} \\ &= \frac{\dot\mu(\theta)}{\theta} \sum_{j=1}^n \mu(\theta)^{j-1} = \frac{\dot\mu(\theta)}{\theta} \frac{\mu(\theta)^n - 1}{\mu(\theta) - 1}. \end{aligned}$$

For $\mu(\theta) > 1$, this converges to infinity at a much faster rate than "usually." Because the total information in $(X_1, \ldots, X_n)$ is of the same order as the information in the last observation X_n, the model is "explosive" in terms of growth of information. The calculation suggests the rescaling rate $\gamma_{n,\theta} = \mu(\theta)^{-n/2}$, which is roughly the inverse root of the information.

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

A Taylor expansion of the log likelihood ratio yields the existence of a point θ_n between θ and $\theta + \gamma_{n,\theta} h$ such that

$$\log \prod_{j=1}^{n} \frac{p_{\theta+\gamma_{n,\theta}h}}{p_\theta}(X_j \mid X_{j-1})$$
$$= \frac{h}{\mu(\theta)^{n/2}} \sum_{j=1}^{n} \dot{\ell}_\theta(X_j \mid X_{j-1}) + \frac{1}{2} \frac{h^2}{\mu(\theta)^n} \sum_{j=1}^{n} \ddot{\ell}_{\theta_n}(X_j \mid X_{j-1}).$$

This motivates the definitions

$$\Delta_{n,\theta} = \frac{1}{\mu(\theta)^{n/2}} \sum_{j=1}^{n} \frac{X_j - \mu(\theta) X_{j-1}}{\theta}$$
$$J_{n,\theta} = \frac{1}{\mu(\theta)^n} \sum_{j=1}^{n} \left[\frac{X_j - \mu(\theta) X_{j-1}}{\theta^2} + \frac{X_{j-1} \dot{\mu}(\theta)}{\theta} \right].$$

Because $\mathrm{E}_\theta(X_n \mid X_{n-1}, \ldots, X_1) = X_{n-1}\mu(\theta)$, the sequence of random variables $\mu(\theta)^{-n} X_n$ is a martingale under θ. Some algebra shows that its second moments are bounded as $n \to \infty$. Thus, by a martingale convergence theorem (e.g., Theorem 10.5.4 of [42]), there exists a random variable V such that $\mu(\theta)^{-n} X_n \to V$ almost surely. By the Toeplitz lemma (Problem 9.6) and again some algebra, we obtain that, almost surely under θ,

$$\frac{1}{\mu(\theta)^n} \sum_{j=1}^{n} X_j \to \frac{\mu(\theta)}{\mu(\theta) - 1} V, \qquad \frac{1}{\mu(\theta)^n} \sum_{j=1}^{n} X_{j-1} \to \frac{1}{\mu(\theta) - 1} V.$$

It follows that the point θ_n in the expansion of the log likelihood can be replaced by θ at the cost of adding a term that converges to zero in probability under θ. Furthermore,

$$J_{n,\theta} \mapsto \frac{\dot{\mu}(\theta)}{\theta\big(\mu(\theta) - 1\big)} V, \qquad P_\theta\text{-almost surely.}$$

It remains to derive the limit distribution of the sequence $\Delta_{n,\theta}$. If we write $X_j = \sum_{i=1}^{X_{j-1}} Z_{j,i}$ for independent copies $Z_{j,i}$ of the offspring variable Z, then

$$\Delta_{n,\theta} = \frac{1}{\theta \mu(\theta)^{n/2}} \sum_{j=1}^{n} \sum_{i=1}^{X_{j-1}} \big(Z_{j,i} - \mu(\theta)\big) = \frac{1}{\theta \mu(\theta)^{n/2}} \sum_{i=1}^{\nu_n} \big(Z_i - \mu(\theta)\big),$$

for independent copies Z_i of Z and $\nu_n = \sum_{i=1}^{n} X_{j-1}$. Even though $Z_1, Z_2, \ldots$ and the total number ν_n of variables in the sum are dependent, a central limit theorem applies to the right side: conditionally on the event $\{V > 0\}$ (on which $\nu_n \to \infty$), the sequence $\nu_n^{-1/2} \sum_{i=1}^{\nu_n} \big(Z_i - \mu(\theta)\big)$ converges in distribution to $\sigma(\theta)$ times a standard normal variable G. Furthermore, if we define G independent of V, conditionally on $\{V > 0\}$,[†]

$$(\Delta_{n,\theta}, J_{n,\theta}) \rightsquigarrow \left(\frac{\sigma(\theta)}{\theta} \sqrt{\frac{V}{\mu(\theta) - 1}} G, \frac{\dot{\mu}(\theta)}{\theta\big(\mu(\theta) - 1\big)} V \right). \tag{9.11}$$

[†] See the appendix of [81] or, e.g., Theorem 3.5.1 and its proof in [146].

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

It is well known that the event $\{V = 0\}$ coincides with the event $\{\lim X_n = 0\}$ of extinction of the population. (This occurs with positive probability if and only if $a_0 > 0$.) Thus, on the set $\{V = 0\}$ the series $\sum_{j=1}^{\infty} X_j$ converges almost surely, whence $\Delta_{n,\theta} \to 0$. Interpreting zero as the product of a standard normal variable and zero, we see that again (9.11) is valid. Thus the sequence $(\Delta_{n,\theta}, J_{n,\theta})$ converges also unconditionally to this limit. Finally, note that $\sigma^2(\theta)/\theta = \dot{\mu}(\theta)$, so that the limit distribution has the right form.

The maximum likelihood estimator for $\mu(\theta)$ can be shown to be asymptotically efficient, (see, e.g., [29] or [81]). □

9.12 ***Example (Gaussian AR).*** The canonical example of an LAMN sequence of experiments is obtained from an explosive autoregressive process of order one with Gaussian innovations. (The Gaussianity is essential.) Let $|\theta| > 1$ and $\varepsilon_1, \varepsilon_2, \ldots$ be an i.i.d. sequence of standard normal variables independent of a fixed variable X_0. We observe the vector $(X_0, X_1, \ldots, X_n)$ generated by the recursive formula $X_t = \theta X_{t-1} + \varepsilon_t$.

The observations form a Markov chain with transition density $p(\cdot \mid x_{t-1})$ equal to the $N(\theta x_{t-1}, 1)$-density. Therefore, the log likelihood ratio process takes the form

$$\log \frac{p_{n,\theta+\gamma_{n,\theta}h}}{p_{n,\theta}}(X_0, \ldots, X_n) = h\,\gamma_{n,\theta} \sum_{t=1}^{n} (X_t - \theta X_{t-1}) X_{t-1} - \frac{1}{2} h^2 \gamma_{n,\theta}^2 \sum_{t=1}^{n} X_{t-1}^2.$$

This has already the appropriate quadratic structure. To establish LAMN, it suffices to find the right rescaling rate and to establish the joint convergence of the linear and the quadratic term. The rescaling rate may be chosen proportional to the Fisher information and is taken $\gamma_{n,\theta} = \theta^{-n}$.

By repeated application of the defining autoregressive relationship, we see that

$$\theta^{-t} X_t = X_0 + \sum_{j=1}^{t} \theta^{-j} \varepsilon_j \to V := X_0 + \sum_{j=1}^{\infty} \theta^{-j} \varepsilon_j,$$

almost surely as well as in second mean. Given the variable X_0, the limit is normally distributed with mean X_0 and variance $(\theta^2 - 1)^{-1}$. An application of the Toeplitz lemma (Problem 9.6) yields

$$\frac{1}{\theta^{2n}} \sum_{t=1}^{n} X_{t-1}^2 \to \frac{V^2}{\theta^2 - 1}.$$

The linear term in the quadratic representation of the log likelihood can (under θ) be rewritten as $\theta^{-n} \sum_{t=1}^{n} \varepsilon_t X_{t-1}$, and satisfies, by the Cauchy-Schwarz inequality and the Toeplitz lemma,

$$\mathrm{E}\left| \frac{1}{\theta^n} \sum_{t=1}^{n} \varepsilon_t X_{t-1} - \frac{1}{\theta^n} \sum_{t=1}^{n} \varepsilon_t \theta^{t-1} V \right| \le \frac{1}{|\theta|^n} \sum_{t=1}^{n} |\theta|^{t-1} \left(\mathrm{E}(\theta^{-t+1} X_{t-1} - V)^2\right)^{1/2} \to 0.$$

It follows that the sequence of vectors $(\Delta_{n,\theta}, J_{n,\theta})$ has the same limit distribution as the sequence of vectors $\left(\theta^{-n} \sum_{t=1}^{n} \varepsilon_t \theta^{t-1} V,\ V^2/(\theta^2 - 1)\right)$. For every n the vector $\left(\theta^{-n} \sum_{t=1}^{n} \varepsilon_t\right.$

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

$\theta^{t-1}, V)$ possesses, conditionally on X_0, a bivariate-normal distribution. As $n \to \infty$ these distributions converge to a bivariate-normal distribution with mean $(0, X_0)$ and covariance matrix $I/(\theta^2 - 1)$. Conclude that the sequence $(\Delta_{n,\theta}, J_{n,\theta})$ converges in distribution as required by the LAMN criterion. □

9.7 Heuristics

The asymptotic representation theorem, Theorem 9.3, shows that every sequence of statistics in a converging sequence of experiments is matched by a statistic in the limit experiment. It is remarkable that this is true under the present definition of convergence of experiments, which involves only marginal convergence and is very weak.

Under appropriate stronger forms of convergence more can be said about the nature of the matching procedure in the limit experiment. For instance, a sequence of maximum likelihood estimators converges to the maximum likelihood estimator in the limit experiment, or a sequence of likelihood ratio statistics converges to the likelihood ratio statistic in the limit experiment. We do not introduce such stronger convergence concepts in this section but only note the potential of this argument as a heuristic principle. See section 5.9 for rigorous results.

For the maximum likelihood estimator the heuristic argument takes the following form. If $\hat{h}_n$ maximizes the likelihood $h \mapsto dP_{n,h}$, then it also maximizes the likelihood ratio process $h \mapsto dP_{n,h}/dP_{n,h_0}$. The latter sequence of processes converges (marginally) in distribution to the likelihood ratio process $h \mapsto dP_h/dP_{h_0}$ of the limit experiment. It is reasonable to expect that the maximizer $\hat{h}_n$ converges in distribution to the maximizer of the process $h \mapsto dP_h/dP_{h_0}$, which is the maximum likelihood estimator for h in the limit experiment. (Assume that this exists and is unique.) If the converging experiments are the local experiments corresponding to a given sequence of experiments with a parameter θ, then the argument suggests that the sequence of local maximum likelihood estimators $\hat{h}_n = r_n(\hat{\theta}_n - \theta)$ converges, under θ, in distribution to the maximum likelihood estimator in the local limit experiment, under $h = 0$.

Besides yielding the limit distribution of the maximum likelihood estimator, the argument also shows to what extent the estimator is asymptotically efficient. It is efficient, or inefficient, in the same sense as the maximum likelihood estimator is efficient or inefficient in the limit experiment. That maximum likelihood estimators are often asymptotically efficient is a consequence of the fact that often the limit experiment is Gaussian and the maximum likelihood estimator of a Gaussian location parameter is optimal in a certain sense. If the limit experiment is not Gaussian, there is no a priori reason to expect that the maximum likelihood estimators are asymptotically efficient.

A variety of examples shows that the conclusions of the preceding heuristic arguments are often but not universally valid. The reason for failures is that the convergence of experiments is not well suited to allow claims about maximum likelihood estimators. Such claims require stronger forms of convergence than marginal convergence only.

For the case of experiments consisting of a random sample from a smooth parametric model, the argument is made precise in section 7.4. Next to the convergence of experiments, it is required only that the maximum likelihood estimator is consistent and that the log density is locally Lipschitz in the parameter. The preceding heuristic argument also extends to the other examples of convergence to limit experiments considered in this chapter. For instance, the maximum likelihood estimator based on a sample from the uniform distribution on $[0, \theta]$

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010

is asymptotically inefficient, because it corresponds to the estimator Z for h (the maximum likelihood estimator) in the exponential limit experiment. The latter is biased upwards and inefficient for every of the usual loss functions.

Notes

This chapter presents a few examples from a large body of theory. The notion of a limit experiment was introduced by Le Cam in [95]. He defined convergence of experiments through convergence of all finite subexperiments relative to his *deficiency distance*, rather than through convergence of the likelihood ratio processes. This deficiency distance introduces a "strong topology" next to the "weak topology" corresponding to convergence of experiments. For experiments with a finite parameter set, the two topologies coincide. There are many general results that can help to prove the convergence of experiments and to find the limits (also in the examples discussed in this chapter). See [82], [89], [96], [97], [115], [138], [142] and [144] for more information and more examples. For nonlocal approximations in the strong topology see, for example, [96] or [110].

PROBLEMS

1. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the normal $N(h/\sqrt{n}, 1)$ distribution, in which $h \in \mathbb{R}$. The corresponding sequence of experiments converges to a normal experiment by the general results. Can you see this directly?
2. If the nth experiment corresponds to the observation of a sample of size n from the uniform $[0, 1-h/n]$, then the limit experiment corresponds to observation of a shifted exponential variable Z. The sequences $-n(X_{(n)} - 1)$ and $\sqrt{n}(2\overline{X}_n - 1)$ both converge in distribution under every h. According to the representation theorem their sets of limit distributions are the distributions of randomized statistics based on Z. Find these randomized statistics explicitly. Any implications regarding the quality of $X_{(n)}$ and $\overline{X}_n$ as estimators?
3. Let the nth experiment consist of one observation from the binomial distribution with parameters n and success probability h/n with $0 < h < 1$ unknown. Show that this sequence of experiments converges to the experiment consisting of observing a Poisson variable with mean h.
4. Let the nth experiment consists of observing an i.i.d. sample of size n from the uniform $[-1 - h/n, 1 + h/n]$ distribution. Find the limit experiment.
5. Prove the asymptotic representation theorem for the case in which the nth experiment corresponds to an i.i.d. sample from the uniform $[0, \theta - h/n]$ distribution with $h > 0$ by mimicking the proof of this theorem for the locally asymptotically normal case.
6. **(Toeplitz lemma.)** If a_n is a sequence of nonnegative numbers with $\sum a_n = \infty$ and $x_n \to x$ an arbitrary converging sequence of numbers, then the sequence $\sum_{j=1}^n a_j x_j / \sum_{j=1}^n a_j$ converges to x as well. Show this.
7. Derive a limit experiment in the case of Galton-Watson branching with $\mu(\theta) < 1$.
8. Derive a limit experiment in the case of a Gaussian AR(1) process with $\theta = 1$.
9. Derive a limit experiment for sampling from a $U[\sigma, \tau]$ distribution with both endpoints unknown.
10. In the case of sampling from the $U[0, \theta]$ distribution show that the maximum likelihood estimator for θ converges to the maximum likelihood estimator in the limit experiment. Why is the latter not a good estimator?
11. Formulate and prove a local asymptotic minimax theorem for estimating θ from a sample from a $U[0, \theta]$ distribution, using $\ell(x) = x^2$ as loss function.

Downloaded from https://www.cambridge.org/core. Columbia University - Law Library, on 13 Feb 2019 at 00:35:31, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/CBO9780511802256.010