MATHEMATISCHE STATISTIEK

A.W. van der Vaart

VOORWOORD

Dit is een syllabus bij het college Mathematische Statistiek. Het behandelt een aantal onderwerpen uit de theorie van de asymptotische statistiek. Wiskundig gezien bewijzen we limietstellingen voor het geval dat het aantal waarnemingen naar oneindig convergeert. Enerzijds verkrijgen we accurate, maar eenvoudige, benaderingen voor statistische procedures; anderzijds vergelijken we de relatieve efficientie van verschillende procedures. De benaderingen zijn praktisch relevant in die gevallen waar het aantal waarnemingen niet te klein is.

De inhoud van de tentamenstof is ongeveer gelijk aan de inhoud van deze syllabus, maar wordt tijdens het college definitief bekend gemaakt. In het kader van de internationalisering gaan we in het volgende hoofdstuk over op het Engels. Achterin het dictaat is een lijst van vertalingen van technische woorden opgenomen.

Anders dan het college Algemene Statistiek, waarvan we het begrippenkader bekend veronderstellen, is het college Mathematische Statistiek wiskundig exact. Dit maakt op enkele plaatsen het gebruik van begrippen uit de maattheorie noodzakelijk. Deze worden in een appendix kort verklaard (zie ook de syllabus Inleiding Waarschijnlijkheidsrekening), en worden uitvoerig behandeld in het College Maattheorie (vijfde semester).

We gebruiken de volgende notatie, mogelijk in afwijking van andere colleges.

Voor kwantielen van verdelingen nemen we altijd het bovenkwantiel. Bijvoorbeeld ξ_α is gedefineerd als het getal zodanig dat $\mathrm{P}(G \geq \xi_\alpha) = \alpha$ voor een $N(0,1)$-verdeelde grootheid G; evenzo $\chi^2_{k,\alpha}$ en $t_{k,\alpha}$. De standaardafwijking en variantie van een stochastische grootheid worden aangeduid met $\operatorname{sd} X$ en $\operatorname{var} X$.

Soms gebruiken we $\int g(x)\, dP_X(x)$ voor $\mathrm{E}g(X)$. Dus $\int g(x)\, dP_X(x)$ is gelijk aan $\int g(x)\, f(x)\, dx$ als X continu verdeeld is met dichtheid f, en $\sum_x g(x)\mathrm{P}(X = x)$ als X discreet verdeeld is.

Het woord "meetbaar" ("measurable") wordt in de appendix verklaard. Desgewenst kun je er ook gewoon overheen lezen: we komen geen objecten tegen die niet meetbaar zijn. De afkorting "i.i.d." betekent "independent and identically distributed".

Parijs, December 1995,

Amsterdam, Januari 1997, November 1997, December 1998 (herzieningen),

A.W. van der Vaart

CONTENTS

1
Stochastic Convergence

1.1 Basic Theory

A sequence of random vectors X_n is said to *converge in distribution* to a random vector X if

$$\mathrm{P}(X_n \leq x) \to \mathrm{P}(X \leq x),$$

for every x at which the distribution function $x \to \mathrm{P}(X \leq x)$ is continuous. Alternative names are *weak convergence* and *convergence in law*. As the last name suggests, the convergence only depends on the induced laws of the vectors and not on the probability spaces on which they are defined. Weak convergence is denoted by $X_n \rightsquigarrow X$; if X has distribution L, or a distribution with a standard code, such as $N(0,1)$, then also by $X_n \rightsquigarrow L$ or $X_n \rightsquigarrow N(0,1)$. The restriction to "continuity points" of $x \to \mathrm{P}(X \leq x)$ in the definition is a bit odd, but will be seen to be reasonable in examples. A distribution function on $\mathbb{R}$ can have at most countably many jump points, which is a very small subset of $\mathbb{R}$, so that the definition does require convergence at "almost all" points x.

Let $d(x,y)$ be a distance function on $\mathbb{R}^k$ that generates the usual topology. For instance, the Euclidean distance

$$d(x,y) = \|x-y\| = \Big(\sum_{i=1}^{k}(x_i - y_i)^2\Big)^{1/2}.$$

A sequence of random variables X_n is said to *converge in probability* to X if for all $\varepsilon > 0$

$$\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \to 0.$$

This is denoted by $X_n \xrightarrow{\mathrm{P}} X$. In this notation convergence in probability is the same as $d(X_n, X) \xrightarrow{\mathrm{P}} 0$.

As we shall see, convergence in probability is stronger than convergence in distribution. An even stronger mode of convergence is almost-sure convergence. The sequence X_n is said to *converge almost surely* to X if $d(X_n, X) \to 0$ with probability one:

$$\mathrm{P}\big(\lim d(X_n, X) = 0\big) = 1.$$

This is denoted by $X_n \xrightarrow{\mathrm{as}} X$. We shall use this mode of convergence only occasionally.

Note that convergence in probability and convergence almost surely only make sense if each of X_n and X are defined on the same probability space. For convergence in distribution this is not necessary.

The two main limit theorems of probability theory can be expressed in terms of these convergence concepts. We shall use these theorems frequently, but do not prove them in this course.

1.1 Example (Law of large numbers). Let $\bar{Y}_n$ be the average of the first n of a sequence of independent, identically distributed random variables $Y_1, Y_2, \ldots$ whose expectation $\mathrm{E}Y_1$ exists. Then $\bar{Y}_n \xrightarrow{\mathrm{P}} \mathrm{E}Y_1$ by the *weak law of large numbers*. Actually, under the same condition it is also true that $\bar{Y}_n \xrightarrow{\mathrm{as}} \mathrm{E}Y_1$, which is called the *strong law of large numbers*. As the name indicates this is a stronger result, as is proved in Theorem 1.11(i). □

1.2 Example (Central limit theorem). Let $\bar{Y}_n$ be the average of the first n of a sequence of independent, identically distributed random variables $Y_1, Y_2, \ldots$. If $\mathrm{E}Y_1^2 < \infty$, then $\sqrt{n}(\bar{Y}_n - \mathrm{E}Y_1) \rightsquigarrow N(0, \operatorname{var} Y_1)$ by the *central limit theorem*. □

1.3 Example. Suppose that X_n is uniformly distributed on the points $1/n, 2/n, \ldots, n/n$, i.e. $\mathrm{P}(X_n = i/n) = 1/n$ for $i = 1, 2, \ldots, n$. Then $X_n \rightsquigarrow$ uniform$[0, 1]$. This can be proved directly from the definition.

Note that $\mathrm{P}(X_n \in \mathbb{Q}) = 1$ for every n, but $\mathrm{P}(X \in \mathbb{Q}) = 0$ for the uniform limit variable X. Thus in general $X_n \rightsquigarrow X$ does not imply that $\mathrm{P}(X_n \in B) \to \mathrm{P}(X \in B)$ for every set B. □

1.4 Example. Let $Y_1, \ldots, Y_n$ be a random sample from the uniform distribution on $[0, 1]$. Then $X_n = \max(Y_1, \ldots, Y_n)$ satisfies, for $x > 0$ and $n \to \infty$,

$$\mathrm{P}\big(-n(X_n - 1) > x\big) = \Big(1 - \frac{x}{n}\Big)^n \to e^{-x}.$$

This implies that the sequence $-n(X_n - 1)$ converges in distribution to an exponential distribution with parameter 1. □

In a statistical context a different name for convergence in probability is "asymptotically consistent". Given a statistical model indexed by a parameter θ, a sequence of estimators T_n is defined to be *asymptotically consistent* for estimating $g(\theta)$ if $T_n \xrightarrow{P_\theta} g(\theta)$ for every θ in the parameter set. Here the extra θ in $\xrightarrow{P_\theta}$ should be understood as indicating that the probabilities must be calculated assuming that the value θ is the "true" value.

1.5 Example. If the observations $Y_1, \ldots, Y_n$ are a random sample from a distribution with finite mean μ, then the sample mean $\bar{Y}_n$ is an asymptotically consistent estimator for the parameter μ, by the law of large numbers. □

1.6 Example. If $Y_1, \ldots, Y_n$ are a random sample from the uniform distribution on $[0, \theta]$, then both $2\bar{Y}_n$ and the maximum $Y_{n(n)}$ are asymptotically consistent estimators of θ. □

The continuous-mapping theorem asserts that stochastic convergence is retained under application of continuous maps: if the sequence of random vectors X_n converges to X and g is continuous, then $g(X_n)$ converges to $g(X)$. This is true for each of the three modes of stochastic convergence. It is not necessary that g be continuous everywhere, but it should be continuous "at all points where the limit X takes its values".

1.7 Theorem (Continuous mapping). *Let $g: \mathbb{R}^k \to \mathbb{R}^m$ be measurable and continuous at every point of a set C such that $\mathrm{P}(X \in C) = 1$.*
(i) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$;
(ii) If $X_n \xrightarrow{\mathrm{P}} X$, then $g(X_n) \xrightarrow{\mathrm{P}} g(X)$;
(iii) If $X_n \xrightarrow{\mathrm{as}} X$, then $g(X_n) \xrightarrow{\mathrm{as}} g(X)$.

Proof. (i). This is difficult to prove without measure theory, as it is hard to relate the distribution functions of X_n and $g(X_n)$. Therefore, we omit the proof.

(ii). Fix an arbitrary $\varepsilon > 0$. For each $\delta > 0$ let B_δ be the set of x for which there exists y with $d(x, y) < \delta$, but $d\big(g(x), g(y)\big) > \varepsilon$. If $X \notin B_\delta$ and $d\big(g(X_n), g(X)\big) > \varepsilon$, then $d(X_n, X) \geq \delta$. Consequently, if $d\big(g(X_n), g(X)\big) > \varepsilon$, then either $X \in B_\delta$ or $d(X_n, X) \geq \delta$, whence

$$\mathrm{P}\Big(d\big(g(X_n), g(X)\big) > \varepsilon\Big) \leq \mathrm{P}(X \in B_\delta) + \mathrm{P}\big(d(X_n, X) \geq \delta\big).$$

The second term on the right converges to zero as $n \to \infty$ for every fixed $\delta > 0$. Since $B_\delta \cap C$ is decreasing when δ decreases to 0 and $\cap_\delta B_\delta \cap C = \emptyset$ by the continuity of g, the first term converges to zero as $\delta \downarrow 0$, by Lemma 6.1(i).

Assertion (iii) follows immediately from the definitions of almost sure convergence and continuity. ■

Any random vector X is *tight*: for every $\varepsilon > 0$ there exists a constant M such that $\mathrm{P}\big(\|X\| > M\big) < \varepsilon$. This follows since the limit as $M \to \infty$ of these probabilities is zero by Lemma 6.1(i). The minimal value of M will depend both on ε and on X. A set of random vectors $\{X_\alpha : \alpha \in A\}$ is called *uniformly tight* if M can be chosen the same for every X_α: for every $\varepsilon > 0$ there exists a constant M such that

$$\sup_\alpha \mathrm{P}\big(\|X_\alpha\| > M\big) < \varepsilon.$$

Thus, there exists a compact set to which all X_α give probability "almost one". Another name for "uniformly tight" is *bounded in probability*. It is not hard to see, that every weakly converging sequence X_n is uniformly tight. More surprisingly, the converse of this statement is almost true: every uniformly tight sequence contains a weakly converging subsequence. For deterministic sequences, this is (a consequence of) the Heine-Borel theorem, which says that a closed, bounded subset of $\mathbb{R}^k$ is compact. (Furthermore, every sequence contained in a compact set possesses a converging subsequence.) For random vectors the result is known as Prohorov's theorem.

1.8 Theorem (Prohorov's theorem). *Let X_n be random vectors in $\mathbb{R}^k$.*
(i) If $X_n \rightsquigarrow X$ for some X, then $\{X_n : n \in \mathbb{N}\}$ is uniformly tight;
(ii) If X_n is uniformly tight, then there is a subsequence with $X_{n_j} \rightsquigarrow X$ as $j \to \infty$, for some X.

Proof. (i). This is left as an exercise.

(ii). By Helly's lemma (below) there exists a subsequence F_{n_j} of the sequence of cumulative distribution functions $F_n(x) = \mathrm{P}(X_n \le x)$ that converges weakly to a possibly "defective distribution function" F. It suffices to show that F is a proper distribution function: $F(x) \to 0, 1$ if $x_i \to -\infty$ for some i, or $x \to \infty$. By the uniform tightness there exists M such that $F_n(M) > 1 - \varepsilon$ for all n. By making M larger, if necessary, it can be ensured that M is a continuity point of F. Then $F(M) = \lim F_{n_j}(M) \ge 1 - \varepsilon$. Conclude that $F(x) \to 1$ as $x \to \infty$. That the limits at $-\infty$ are zero can be seen in a similar manner. ∎

The crux of the proof of Prohorov's theorem is Helly's lemma. This asserts, that any given sequence of distribution functions contains a subsequence that converges weakly to a possibly "defective distribution function". A *defective distribution function* is a function that has all the properties of a cumulative distribution function with the exception that it has limit < 1 at ∞ and/or > 0 at $-\infty$. Thus, a defective distribution function on $\mathbb{R}$ is a function $F : \mathbb{R} \to [0, 1]$ that is nondecreasing and continuous from the right.

1.9 Lemma (Helly's lemma). *Each given sequence F_n of cumulative distribution functions on $\mathbb{R}^k$ possesses a subsequence F_{n_j} with the property that $F_{n_j}(x) \to F(x)$ at each continuity point x of a possibly defective distribution function F.*

Proof. For simplicity we give the proof for $k = 1$. Let $\mathbb{Q} = \{q_1, q_2, \ldots\}$ be the rational numbers (or another countable dense set), ordered in an arbitrary manner. Since the sequence $F_n(q_1)$ is contained in the interval $[0, 1]$, it has a converging subsequence by the Heine-Borel theorem. Call the indexing subsequence $\{n_j^1\}_{j=1}^{\infty}$ and the limit $G(q_1)$. Next, extract a further subsequence $\{n_j^2\} \subset \{n_j^1\}$ along which $F_n(q_2)$ converges to a limit $G(q_2)$, a further subsequence $\{n_j^3\} \subset \{n_j^2\}$ along which $\cdots$, etc.. The 'tail' of the diagonal sequence $n_j = n_j^j$ belongs to every sequence n_j^i. Hence $F_{n_j}(q_i) \to G(q_i)$ for every $i = 1, 2, \ldots$. Since each F_n is nondecreasing, $G(q) \le G(q')$ if $q \le q'$. Define

$$F(x) = \inf_{q>x} G(q).$$

Then F is nondecreasing. It is also right continuous at every point x, because for every $\varepsilon > 0$ there exists $q > x$ with $G(q) - F(x) < \varepsilon$, which implies $F(y) - F(x) < \varepsilon$ for every $x \le y \le q$.

Continuity of F at x means that, for every $\varepsilon > 0$ there exists $\delta > 0$ such that $\big|F(x) - F(y)\big| < \varepsilon$ for every $|x - y| < \delta$. This implies the existence of $q < x < q'$ such that $G(q') - G(q) < 2\varepsilon$. By monotonicity, we have $G(q) \le F(x) \le G(q')$, and

$$G(q) = \lim F_{n_j}(q) \le \liminf F_{n_j}(x) \le \limsup F_{n_j}(x) \le \lim F_{n_j}(q') = G(q').$$

Conclude that $|\liminf F_{n_j}(x) - F(x)| < 2\varepsilon$ and the same for lim sup. Since this is true for every $\varepsilon > 0$, it follows that $F_{n_j}(x) \to F(x)$ at every continuity point of F. ∎

1.10 Example. A sequence X_n of random variables with $\mathrm{E}|X_n|^r = O(1)$ for some $r > 0$ is uniformly tight. This follows since by Markov's inequality

$$\mathrm{P}\big(|X_n| > M\big) \le \frac{\mathrm{E}|X_n|^r}{M^r}.$$

The right side can be made arbitrarily small uniformly in n, by choosing sufficiently large M.

Since the second moment is the sum of the variance and the square of the mean, an alternative sufficient condition for uniform tightness is: $\mathrm{E}X_n = O(1)$ and $\operatorname{var} X_n = O(1)$. □

Consider some of the relationships between the three modes of convergence. Convergence in distribution is weaker than convergence in probability, which is in turn weaker than almost sure convergence.

1.11 Theorem. *Let X_n, X and Y_n be random vectors. Then*
(i) $X_n \xrightarrow{\text{as}} X$ *implies* $X_n \xrightarrow{\text{P}} X$;
(ii) $X_n \xrightarrow{\text{P}} X$ *implies* $X_n \rightsquigarrow X$;
(iii) $X_n \xrightarrow{\text{P}} c$ *for a constant* c *if and only if* $X_n \rightsquigarrow c$;
(iv) if $X_n \rightsquigarrow X$ *and* $d(X_n, Y_n) \xrightarrow{\text{P}} 0$, *then* $Y_n \rightsquigarrow X$;
(v) if $X_n \rightsquigarrow X$ *and* $Y_n \xrightarrow{\text{P}} c$ *for a constant* c, *then* $(X_n, Y_n) \rightsquigarrow (X, c)$;
(vi) if $X_n \xrightarrow{\text{P}} X$ *and* $Y_n \xrightarrow{\text{P}} Y$, *then* $(X_n, Y_n) \xrightarrow{\text{P}} (X, Y)$.

Proof. (i). For every fixed $\varepsilon > 0$ the sequence of sets

$$A_n = \cup_{m \geq n} \big\{ d(X_m, X) > \varepsilon \big\}$$

is decreasing: $A_1 \supset A_2 \supset \cdots$. If $X_n(\omega) \to X(\omega)$ for some ω, then there exists n such that $d\big(X_m(\omega), X(\omega)\big) \leq \varepsilon$ for $m \geq n$ and hence $\omega \notin A_n$. Therefore $\cap_n A_n$ can contain only ω such that $X_n(\omega)$ does not converge to $X(\omega)$ and hence is a set with probability zero by assumption. Conclude that $\mathrm{P}(A_n) \to 0$ and hence that $\mathrm{P}\big(d(X_n, X) > \varepsilon\big) \leq \mathrm{P}(A_n) \to 0$.

(iv). We give the result for random variables only. The proof for the vector case is similar. For every $\varepsilon > 0$,

$$\begin{aligned} \mathrm{P}(Y_n \leq x) &\leq \mathrm{P}\big(Y_n \leq x, d(X_n, Y_n) \leq \varepsilon\big) + \mathrm{P}\big(d(X_n, Y_n) > \varepsilon\big) \\ &\leq \mathrm{P}(X_n \leq x + \varepsilon) + o(1). \end{aligned}$$

If $x + \varepsilon$ is a continuity point of the distribution function of X, then the right side converges to $\mathrm{P}(X \leq x + \varepsilon)$ and we conclude that $\limsup \mathrm{P}(Y_n \leq x) \leq \mathrm{P}(X \leq x + \varepsilon)$. This is true for all $\varepsilon > 0$ except at most the countably many values such that $x + \varepsilon$ is a jump point of $x \to \mathrm{P}(X \leq x)$. In particular, it is true for a sequence $\varepsilon_m \downarrow 0$ and we conclude that

$$\limsup \mathrm{P}(Y_n \leq x) \leq \lim_{m \to \infty} \mathrm{P}(X \leq x + \varepsilon_m) = \mathrm{P}(X \leq x).$$

This gives one half of the proof. By arguing in an analogous manner, we can prove that $\limsup \mathrm{P}(Y_n > x) \leq \mathrm{P}(X > x - \varepsilon)$ for every x and $\varepsilon > 0$ and hence that $\limsup \mathrm{P}(Y_n > x) \leq \mathrm{P}(X \geq x)$. For x a continuity point of the distribution function of X, the right side is equal to $\mathrm{P}(X > x)$. Taking complements we obtain that $\liminf \mathrm{P}(Y_n \leq x) \geq \mathrm{P}(X \leq x)$.

The two inequalities combined yield that $\mathrm{P}(Y_n \leq x) \to \mathrm{P}(X \leq x)$ for every continuity point of the distribution function of X.

(ii). Since $d(X_n, X) \xrightarrow{\text{P}} 0$, and trivially $X \rightsquigarrow X$, it follows that $X_n \rightsquigarrow X$ by (iv).

(iii). The "only if" part is a special case of (ii). For the converse, we consider only the one-dimensional case and suppose that $X_n \rightsquigarrow c$. The distribution function of the limit variable is 0 to the left of c and 1 to the right of c and everywhere continuous except at the point c. Therefore $\mathrm{P}(X_n \leq x)$ converges to 0 for $x < c$ and converges to 1 for $x > c$. Then $\mathrm{P}\big(d(X_n, c) \geq \varepsilon\big) = \mathrm{P}(X_n \leq c - \varepsilon) + \mathrm{P}(X_n \geq c + \varepsilon) \to 0$ for every $\varepsilon > 0$.

(v). First note that $d\big((X_n, Y_n), (X_n, c)\big) = d(Y_n, c) \xrightarrow{\mathrm{P}} 0$. Thus, according to (iv), it suffices to show that $(X_n, c) \rightsquigarrow (X, c)$. The distribution function $\mathrm{P}(X_n \le x, c \le y)$ is equal to 0 for $y < c$ and equal to $\mathrm{P}(X_n \le x)$ if $y \ge c$; the same is true for X_n replaced by X. Thus, we have that $\mathrm{P}(X_n \le x, c \le y) \to \mathrm{P}(X \le x, c \le y)$ for every (x, y) such that either $y < c$, or both $y \ge c$ and $x \to \mathrm{P}(X_n \le x)$ is continuous at x. This includes the continuity points of $(x, y) \to \mathrm{P}(X \le x, c \le y)$.

(vi). This follows from $d\big((x_1, y_1), (x_2, y_2)\big) \le d(x_1, x_2) + (y_1, y_2)$. ∎

According to the last assertion of the lemma, convergence in probability of a sequence of vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ is equivalent to convergence of every one of the sequences of components $X_{n,i}$ separately. The analogous statement for convergence in distribution is false: convergence in distribution of the sequence X_n is stronger than convergence of every one of the sequences of components $X_{n,i}$. The point is that the distribution of the components $X_{n,i}$ separately does not determine their joint distribution: they might be independent or dependent in many ways. We speak of *joint convergence* in distribution versus *marginal convergence*.

Assertion (v) of the lemma has some useful consequences. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $(X_n, Y_n) \rightsquigarrow (X, c)$. Consequently, by the continuous-mapping theorem, $g(X_n, Y_n) \rightsquigarrow g(X, c)$, for every map g that is continuous at the set $\mathbb{R}^k \times \{c\}$ where the vector (X, c) takes its values. Thus, for every g such that

$$\lim_{x \to x_0, y \to c} g(x, y) = g(x_0, c), \qquad \text{every } x_0.$$

Some particular applications of this principle are known as Slutsky's lemma.

1.12 Lemma (Slutsky). *Let X_n, X and Y_n be random vectors or variables. If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$ for a constant c, then*

(i) $X_n + Y_n \rightsquigarrow X + c$;
(ii) $Y_n X_n \rightsquigarrow cX$;
(iii) $Y_n^{-1} X_n \rightsquigarrow c^{-1} X$ *provided* $c \neq 0$.

In (i) the "constant" c must be a vector of the same dimension as X, and in (ii) c is probably initially understood to be a scalar. However, (ii) is also true if every Y_n and c are matrices (which can be identified with vectors, for instance by aligning rows, to give a meaning to the convergence $Y_n \rightsquigarrow c$), simply because matrix multiplication $(x, y) \to yx$ is a continuous operation. Even (iii) is valid for matrices Y_n and c and vectors X_n provided $c \neq 0$ is understood as c being invertible, because taking an inverse is also continuous.

1.13 Example (t-statistic). Let $Y_1, Y_2, \ldots$ be independent, identically distributed random variables with $\mathrm{E}Y_1 = 0$ and $\mathrm{E}Y_1^2 < \infty$. Then the t-statistic $\sqrt{n}\bar{Y}_n / S_n$, where $S_n^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ is the sample variance, is asymptotically standard normal.

To see this, first note that by two applications of the weak law of large numbers and the continuous-mapping theorem for convergence in probability

$$S_n^2 = \frac{n}{n-1}\Big(\frac{1}{n}\sum_{i=1}^n Y_i^2 - \bar{Y}_n^2\Big) \xrightarrow{\mathrm{P}} 1\big(\mathrm{E}Y_1^2 - (\mathrm{E}Y_1)^2\big) = \operatorname{var} Y_1.$$

Again by the continuous-mapping theorem, S_n converges in probability to $\operatorname{sd} Y_1$. By the central limit theorem $\sqrt{n}\bar{Y}_n$ converges in law to a normal distribution. Finally, Slutsky's lemma gives that the sequence of t-statistics converges in distribution to $N\big(0, \operatorname{var} Y_1\big)/\operatorname{sd} Y_1 = N(0,1)$. □

1.14 Example (Confidence intervals). Let T_n and S_n be sequences of statistical estimators satisfying

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, \sigma^2), \qquad\qquad S_n^2 \xrightarrow{\mathrm{P}} \sigma^2,$$

for certain parameters θ and σ^2 depending on the underlying distribution, for every distribution in the model. Then $\theta = T_n \pm S_n/\sqrt{n}\,\xi_\alpha$ is a confidence interval for θ of asymptotic level $1 - 2\alpha$. More precisely, we have

$$\mathrm{P}\Big(T_n - \frac{S_n}{\sqrt{n}}\xi_\alpha \le \theta \le T_n + \frac{S_n}{\sqrt{n}}\xi_\alpha\Big) \to 1 - 2\alpha.$$

This is a consequence of the fact that the sequence $\sqrt{n}(T_n - \theta)/S_n$ is asymptotically standard normally distributed. □

1.2 Stochastic o and O Symbols

It is convenient to have short expressions for terms that converge in probability to zero or are uniformly tight. The notation $o_P(1)$ ('small "oh-P-one"') is short for a sequence of random vectors that converges to zero in probability. The expression $O_P(1)$ ('big "oh-P-one"') denotes a sequence that is bounded in probability. More generally, for a given sequence of random variables R_n,

$$\begin{aligned} X_n = o_P(R_n) \quad &\text{means} \quad X_n = Y_n R_n \text{ and } Y_n \xrightarrow{\mathrm{P}} 0; \\ X_n = O_P(R_n) \quad &\text{means} \quad X_n = Y_n R_n \text{ and } Y_n = O_P(1). \end{aligned}$$

This expresses that the sequence X_n converges in probability to zero, or is bounded in probability, at 'rate' R_n. For deterministic sequences X_n and R_n, the stochastic oh-symbols reduce to the usual o and O from calculus.

There are many rules of calculus with o and O symbols, which we apply without comment. For instance,

$$\begin{aligned} o_P(1) + o_P(1) &= o_P(1) \\ o_P(1) + O_P(1) &= O_P(1) \\ O_P(1)o_P(1) &= o_P(1) \\ \big(1 + o_P(1)\big)^{-1} &= O_P(1) \\ o_P(R_n) &= R_n o_P(1) \\ O_P(R_n) &= R_n O_P(1). \end{aligned}$$

To see the validity of these 'rules' it suffices to restate them in terms of explicitly named vectors, where each $o_P(1)$ and $O_P(1)$ should be replaced by a different sequence of vectors, that converges to zero or is bounded in probability. In this way the first rule says: if $X_n \xrightarrow{\mathrm{P}} 0$ and $Y_n \xrightarrow{\mathrm{P}} 0$, then $Z_n = X_n + Y_n \xrightarrow{\mathrm{P}} 0$. This is an example of the continuous-mapping theorem. The third rule is short for: if X_n is bounded in probability and $Y_n \xrightarrow{\mathrm{P}} 0$, then $X_n Y_n \xrightarrow{\mathrm{P}} 0$. If X_n would also converge in distribution, then this would be statement (ii) of Slutsky's lemma (with $c = 0$). But by Prohorov's theorem, X_n converges in distribution 'along subsequences' if it is bounded in probability, so that the third rule can still be deduced from Slutsky's lemma by 'arguing along subsequences'.

Note that both rules are in fact implications, and should be read from left to right, even though they are stated with the help of the equality "=" sign. Similarly, while it is true that $o_P(1) + o_P(1) = 2o_P(1)$, writing down this rule does not reflect understanding of the o_P-symbol.

Two more complicated rules are given by the following lemma.

1.15 Lemma. *Let R be a function defined on a neighbourhood of $0 \in \mathbb{R}^k$ such that $R(0) = 0$. Let X_n be a sequence of random vectors that converges in probability to zero.*
(i) if $R(h) = o(\|h\|)$ as $h \to 0$, then $R(X_n) = o_P(\|X_n\|)$;
(ii) if $R(h) = O(\|h\|)$ as $h \to 0$, then $R(X_n) = O_P(\|X_n\|)$.

Proof. Define $g(h)$ as $g(h) = R(h)/\|h\|$ for $h \neq 0$ and $g(0) = 0$. Then $R(X_n) = g(X_n)\|X_n\|$.

(i). Since the function g is continuous at zero by assumption, $g(X_n) \xrightarrow{\mathrm{P}} g(0) = 0$ by the continuous mapping theorem.

(ii). By assumption there exist M and $\delta > 0$ such that $\big|g(h)\big| \leq M$ whenever $\|h\| \leq \delta$. Thus $\mathrm{P}\big(\big|g(X_n)\big| > M\big) \leq \mathrm{P}\big(\|X_n\| > \delta\big) \to 0$, and the sequence $g(X_n)$ is tight. ∎

It should be noted, that the rule expressed by the lemma is not a simple plug-in rule. For instance, it is not true that $R(h) = o(\|h\|)$ implies that $R(X_n) = o_P(\|X_n\|)$ for every sequence of random vectors X_n.

Problems

1. Let $\mathrm{P}(X_n = i/n) = 1/n$ for every $i = 1, 2, \ldots, n$. Show that $X_n \rightsquigarrow X$ for a uniform variable X.

2. Suppose that $\mathrm{P}(X_n \le x) = (nx \vee 0) \wedge 1$.
(i) Draw the graph of $x \to \mathrm{P}(X_n \le x)$. Does $\mathrm{P}(X_n \le x)$ converge for every x?
(ii) Show that X_n converges in distribution.

3. If $\mathrm{P}(X_n = x_n) = 1$ for every n and numbers such that $x_n \to x$, then $X_n \rightsquigarrow X$ for X a random variable such that $\mathrm{P}(X = x) = 1$. Prove this in two ways: (i) by considering distribution functions; (ii) by using Theorem 1.11.

4. If every X_n and X possess a discrete distribution supported on a finite set of integers, show that $X_n \rightsquigarrow X$ if and only if $\mathrm{P}(X_n = x) \to \mathrm{P}(X = x)$ for every x.

5. Show that "finite" in the preceding problem is unnecessary.

6. Let X_n be binomially distributed with parameters n and p_n. If $n \to \infty$ and $np_n \to \lambda > 0$, then $X_n \rightsquigarrow \text{Poisson}(\lambda)$.

7. Find an example of a sequence of random variables such that $X_n \rightsquigarrow 0$, but $\mathrm{E}X_n \to \infty$.

8. In what sense is a χ^2-distribution with n degrees of freedom approximately a normal distribution?

9. Let X_n be the maximum of a random sample $Y_1, \ldots, Y_n$ from the density $2(1-x)$ on $[0,1]$. Find constants a_n and b_n such that $b_n(X_n - a_n)$ converges in distribution to a nondegenerate limit.

10. Let $Y_{n(1)}$ and $Y_{n(n)}$ be the minimum and maximum of a random sample $Y_1, \ldots, Y_n$ from the uniform distribution on $[0,1]$. Show that the sequence $n\big(Y_{n(1)}, 1 - Y_{n(n)}\big)$ converges in distribution to (U, V) for two independent exponential variables.

11. Find an example of sequences $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$, but the joint sequence (X_n, Y_n) does not converge in law.

12. If X_n and Y_n are independent random vectors for every n, then $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ implies that $(X_n, Y_n) \rightsquigarrow (X, Y)$, where X and Y are independent.

13. Suppose that $\mathrm{P}(X_n \le \xi_n) = p$ for every n and $\mathrm{P}(X \le \xi - \varepsilon) < p = \mathrm{P}(X \le \xi) < \mathrm{P}(X \le \xi + \varepsilon)$ for every $\varepsilon > 0$. Show that $\xi_n \to \xi$ if $X_n \rightsquigarrow X$.

14. Suppose that $X_1, \ldots, X_n$ is a random sample from the Poisson distribution with mean θ. Find an asymptotically consistent sequence of estimators for estimating $\mathrm{P}_\theta(X_1 = 0)$.

15. Suppose that $X_1, \ldots, X_n$ are a random sample from the uniform distribution on $[-\theta, \theta]$. Show that the maximum, $X_{n(n)}$, minus the minimum, $-X_{n(1)}$, and $\frac{1}{2}(X_{n(n)} - X_{n(1)})$ are asymptotically consistent estimators of θ.

16. Define a *median* of a distribution function F as a point m such that $F(m) = \frac{1}{2}$. Suppose that F possesses a positive density.
(i) Show that the median of F is unique;
(ii) Show that the sample median, based on a random sample of size n from F, is an asymptotically consistent estimator of the median of F.

17. Suppose that $X_n \rightsquigarrow N(\mu, \sigma^2)$. Find the limit distribution of the sequence $bX_n + a$.

18. Suppose that the sequence of random vectors $X_n = (X_{n,1}, \ldots, X_{n,k})$ converges in distribution to a vector $X = (X_1, \ldots, X_k)$ whose coordinates are independent standard normal variables. Prove that $\|X_n\|^2 \rightsquigarrow \chi^2_k$. (Here $\|\cdot\|$ denotes the Euclidean norm.)

19. If $\mathrm{E}X_n \to \mu$ and $\operatorname{var} X_n \to 0$, then $X_n \xrightarrow{\mathrm{P}} \mu$. Prove this.

20. Suppose that X_n is $N(\mu_n, \sigma_n^2)$-distributed.
(i) Suppose that $\sigma_n = 1$ for every n. Show that the sequence X_n is uniformly tight if and only if $\mu_n = O(1)$;
(ii) Suppose that $\mu_n = 0$ for every n. For which sequences σ_n is the sequence X_n uniformly tight?

21. Prove that for any sequence of random vectors X_n the following statements are equivalent:
(i) the sequence X_n is uniformly tight;
(ii) for every $\varepsilon > 0$ there exist M and N such that $\mathrm{P}\big(\|X_n\| > M\big) < \varepsilon$ for every $n \geq N$;
(iii) $\mathrm{P}\big(\|X_n\| > M_n\big) \to 0$ as $n \to \infty$ for every sequence $M_n \to \infty$.

22. Prove the "simple" part of Prohorov's theorem: if $X_n \rightsquigarrow X$, then the sequence X_n is uniformly tight.

23. If $X_n \rightsquigarrow N(0,1)$ and $Y_n \xrightarrow{\mathrm{P}} \sigma$, then $X_n Y_n \rightsquigarrow N(0, \sigma^2)$. Show this.

24. A random variable X_n is said to possess the t-distribution with n degrees of freedom if it is distributed as $\sqrt{n}Z/(Z_1^2 + \cdots + Z_n^2)^{1/2}$ for independent standard normal variables $Z, Z_1, \ldots, Z_n$. Show that $X_n \rightsquigarrow N(0,1)$ as $n \to \infty$.

25. If $\sqrt{n}(T_n - \theta)$ converges in distribution, then T_n converges in probability to θ. Prove this.

26. Let $X_1, \ldots, X_n$ be i.i.d. with density $f_{\lambda,a}(x) = \lambda e^{-\lambda(x-a)} 1\{x \geq a\}$, where the parameters $\lambda > 0$ and $a \in \mathbb{R}$ are unknown. Calculate the maximum likelihood estimator $(\hat{\lambda}_n, \hat{a}_n)$ of (λ, a) and show that $(\hat{\lambda}_n, \hat{a}_n) \xrightarrow{\mathrm{P}} (\lambda, a)$.

27. Let $X_1, \ldots, X_n$ be a random sample from a distribution with finite second moment and mean μ. Show that the interval $\mu = \bar{X}_n \pm S_n/\sqrt{n}\, t_{n-1,\alpha}$, for S_n^2 the sample variance, is a confidence interval for μ of asymptotic confidence level $1 - 2\alpha$. What can you say in addition if the observations are $N(\mu, \sigma^2)$-distributed?

28. Let X_n be binomially distributed with parameters (n, p). Prove that $p = X_n/n \pm \xi_\alpha/\sqrt{X_n(1 - X_n)/n)}$ is an asymptotic confidence interval of level $1 - 2\alpha$.

29. Let $X_1, \ldots, X_m$ and $Y_1 \ldots, Y_n$ be two independent random samples from distributions with means and variances equal to (μ_1, σ_1^2) and (μ_2, σ_2^2), respectively. Let

$$T_{m,n} = \frac{\bar{X}_m - \bar{Y}_n}{S_{m,n}}, \qquad S_{m,n}^2 = \frac{S_X^2}{m} + \frac{S_Y^2}{n},$$

for S_X^2 and S_Y^2 the sample variances of the two samples. Show that the test that rejects $H_0\colon \mu_1 = \mu_2$ if $|T_{m,n}| > \xi_\alpha$ is of asymptotic level 2α

(i) if $m = n \to \infty$;

(ii) if $m, n \to \infty$.

30. State the rule $O_P(1)o_P(1) = o_P(1)$ in terms of random variables X_n, Y_n and Z_n, and deduce it from Slutsky's lemma by "arguing along subsequences".

31. In what sense is it true that $o_P(1) = O_P(1)$. Is it true that $O_P(1) = o_P(1)$?

32. State the rule $O_P(1) + O_P(1) = O_P(1)$ in terms of random variables and X_n, Y_n and Z_n and prove it. Does it follow from this that $o_P(1) + O_P(1) = O_P(1)$?

33. The rule given by Lemma 1.15 is not a simple plug-in rule. Give an example of a function R with $R(h) = o(\|h\|)$ as $h \to 0$ and a sequence of random variables X_n such that $R(X_n)$ is not equal to $o_P(X_n)$.

34. Find an example of a sequence of random variables such that $X_n \xrightarrow{\mathrm{P}} 0$, but X_n does not converge almost surely.

35. If $\sum_{n=1}^{\infty} \mathrm{P}(|X_n| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then X_n converges almost surely to zero. Prove this.

36. Let X_n be the maximum of a random sample from the uniform distribution on $[0, 1]$. Show that $X_n \xrightarrow{\mathrm{as}} 1$.

2
Multivariate-Normal Distribution

2.1 Covariance Matrices

The covariance of two random variables X and Y is defined as $\operatorname{cov}(X,Y) = \mathrm{E}(X-\mathrm{E}X)(Y-\mathrm{E}Y)$, if these expectations exist. The variance of X is equal to $\operatorname{var} X = \operatorname{cov}(X,X)$. Recall that the expectation is linear: $\mathrm{E}(\alpha X+\beta Y) = \alpha\mathrm{E}+\beta\mathrm{E}Y$; and the covariance is symmetric and bilinear: $\operatorname{cov}(\alpha X+\beta Y, Z) = \alpha\operatorname{cov}(X,Z) + \beta\operatorname{cov}(Y,Z)$.

The *expectation* and *covariance matrix* of a random vector $(X_1,\dots,X_k)$ are the vector and matrix

$$\mathrm{E}X = \begin{pmatrix} \mathrm{E}X_1 \\ \mathrm{E}X_2 \\ \vdots \\ \mathrm{E}X_k \end{pmatrix}, \qquad \operatorname{Cov}(X) = \begin{pmatrix} \operatorname{cov}(X_1,X_1) & \cdots & \operatorname{cov}(X_1,X_k) \\ \operatorname{cov}(X_2,X_1) & \cdots & \operatorname{cov}(X_2,X_k) \\ \vdots & & \vdots \\ \operatorname{cov}(X_k,X_1) & \cdots & \operatorname{cov}(X_k,X_k) \end{pmatrix}.$$

For $k=1$ these are simply the expectation and variance of X_1. The following lemma gives some basic properties.

2.1 Lemma. *For every matrix A, vector b and random vector X:*
(i) $\mathrm{E}(AX+b) = A\mathrm{E}X + b$;
(ii) $\operatorname{Cov}(AX) = A(\operatorname{Cov}X)A^T$;
(iii) $\operatorname{Cov}X$ *is symmetric and nonnegative definite;*
(iv) $\mathrm{P}\big(X \in \mathrm{E}X + \operatorname{range}(\operatorname{Cov}X)\big) = 1$.

Proof. Properties (i) and (ii) follow easily from the linearity and bilinearity of the expectation and covariance of real random variables. For (iii) and (iv), we note that by (ii), $\operatorname{var} a^TX = a^T\operatorname{Cov}(X)a$ for any vector a. A variance

is nonnegative. Furthermore, for any a that is contained in the kernel of $\operatorname{Cov} X$, we have $\operatorname{var} a^T X = 0$, which implies that $a^T(X - \mathrm{E}X) = 0$ with probability one. The exceptional set of probability zero may depend on a, but we still have that $X - \mathrm{E}X \perp a$ for every a in a given countable subset of the kernel of $\operatorname{Cov} X$ with probability one, since a countable union of null sets is a null set. Since we can choose this countable subset dense, it follows that $X - \mathrm{E}X$ is orthogonal to the kernel of $\operatorname{Cov} X$. Since $\operatorname{range}(A^T) = N(A)^\perp$ for any matrix A, this means that X is in the range of $(\operatorname{Cov} X)^T = \operatorname{Cov} X$ with probability one. ∎

2.2 Definition and Basic Properties

For given numbers $\mu \in \mathbb{R}$ and $\sigma > 0$, a random variable is normally $N(\mu, \sigma^2)$-distributed if it has probability density

$$x \to \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2/\sigma^2}.$$

In addition to this we define a random variable X to be $N(\mu, 0)$-distributed if $\mathrm{P}(X = \mu) = 1$. This is the natural extension to the case that $\sigma = 0$, because in every case we now have $\mathrm{E}X = \mu$ and $\operatorname{var} X = \sigma^2$.

We wish to generalize the definition to higher dimensions. Let μ and Σ be an arbitrary vector and a nonnegative, symmetric $(k \times k)$-matrix, respectively. Every such Σ can be written as

$$\Sigma = LL^T,$$

for a $(k \times k)$-matrix L. In fact, there are several possible choices for L. Every choice can be used in the following. One possible choice derives from the transformation to an orthonormal basis of eigenvectors of Σ. Relative to this basis, the linear transformation Σ is given by the diagonal matrix D of the eigenvalues of Σ, and $\Sigma = ODO^T$ for the orthogonal matrix O that represents the change of basis. (Hence $O^T = O^{-1}$.) We could define $L = OD^{1/2}O^T$ for $D^{1/2}$ the diagonal matrix of square roots of eigenvalues of Σ, since

$$OD^{1/2}O^T\, OD^{1/2}O^T = OD^{1/2}D^{1/2}O^T = \Sigma.$$

Thus, this choice of L has the desired property. It is a nonnegative symmetric matrix, just as Σ, and is known as the "positive square root of Σ".

2.2 Definition. A random vector X is said to be *multivariate-normally* distributed with parameters μ and Σ, notation $N_k(\mu, \Sigma)$, if it has the same distribution as the vector $\mu + LZ$, for a matrix L with $\Sigma = LL^T$ and $Z = (Z_1, \ldots, Z_k)^T$ a vector whose coordinates are independent $N(0,1)$-distributed variables.

The notation $N_k(\mu, \Sigma)$ suggests that the distribution of X depends only on μ and Σ. This, indeed, is the case, although this is not clear from its definition as the distribution of $\mu + LZ$, which appears to depend on μ and L. It can be seen from Lemmas 2.3 and 2.4 below, that the distribution of $\mu + LZ$ depends on L only through LL^T as claimed.

The parameters μ and Σ are precisely the mean and covariance matrix of the vector X since, by Lemma 2.1,

$$\mathrm{E}X = \mu + L\mathrm{E}Z = \mu, \qquad \mathrm{Cov}\, X = L\,\mathrm{Cov}\, ZL^T = \Sigma.$$

The multivariate normal distribution with $\mu = 0$ and $\Sigma = I$, the identity matrix, is called *standard normal*. The coordinates of a standard normal vector X are independent $N(0,1)$-distributed variables.

If Σ is singular, then the multivariate-normal distribution $N_k(\mu, \Sigma)$ does not have a density. (This corresponds to the case $\sigma^2 = 0$ in dimension one.) We can see this from Lemma 2.1, which implies that $X - \mathrm{E}X$ takes its values in the range of Σ, a lower dimensional subspace of $\mathbb{R}^k$. It also follows directly from the definition: if Σ is singular, then so is L, and clearly $X - \mu$ takes its values in the range of L. On the other hand if Σ is nonsingular, then the multivariate normal distribution $N_k(\mu, \Sigma)$ has a density.

2.3 Lemma. *A vector X is multivariate-normally distributed for a nonsingular matrix Σ if and only if it has density*

$$x \to \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}.$$

Proof. The density of $Z = (Z_1, \ldots, Z_k)$ is the product of standard normal densities. Thus, for every vector b,

$$\mathrm{P}\big(\mu + LZ \le b\big) = \int_{z:\mu+Lz\le b} \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}\, dz$$

Apply the change of variables $\mu + Lz = x$. The Jacobian $\partial z/\partial x$ of this linear transformation is L^{-1} and has determinant $\det L^{-1} = (\det \Sigma)^{-1/2}$. Furthermore, $\sum z_i^2 = z^T z = (x-\mu)^T\Sigma^{-1}(x-\mu)$. It follows that the integral can be rewritten as

$$\int_{x:x\le b} \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}\, (\det \Sigma)^{-1/2}\, dx.$$

This being true for every b implies the result. ∎

It is frequently useful to "reduce" vectors to dimension 1 by taking linear combinations of its coordinates. By Lemma 6.3 in the appendix the distribution of a vector X is completely determined by the distributions of all linear combinations $a^T X$. For the normal distribution this takes the following particularly attractive form.

2.4 Lemma. *The vector $X = (X_1, \ldots, X_k)$ is $N_k(\mu, \Sigma)$-distributed if and only if $a^T X$ is $N_1(a^T\mu, a^T\Sigma a)$-distributed for every $a \in \mathbb{R}^k$.*

Proof. $\Rightarrow$. The parameters $a^T\mu$ and $a^T\Sigma a$ are correct, because they are the mean and variance of $a^T X$. It suffices to show that $a^T X$ is normally distributed. Since X is distributed as $\mu + LZ$, the variable $a^T X$ is distributed as $a^T\mu + (L^T a)^T Z$. The latter variable is a constant plus a linear combination $b^T Z$ of independent $N(0,1)$-distributed variables (for $b = L^T a$). It is well known, that such a linear combination is normally distributed.

For completeness we give a proof of this fact. Assume without loss of generality that $\|b\| = 1$. There exist vectors $b_2, \ldots, b_k$ such that $\{b, b_2, \ldots, b_k\}$ forms an orthonormal base of $\mathbb{R}^k$. If B is the matrix with first row b and ith row b_i, then $b^T Z$ is the first coordinate of BZ. The distribution function of BZ is

$$\mathrm{P}\big(BZ \le b\big) = \int_{z:Bz\le b} \prod_{i=1}^{k} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}\, dz.$$

Apply the change of variables $Bz = x$. Since B is orthonormal, the Jacobian has determinant 1, and $\sum z_i^2 = \|z\|^2 = \|x\|^2$. Conclude that BZ has the standard normal density, whence $(BZ)_1, \ldots, (BZ)_k$ are i.i.d. $N(0,1)$-variables.

$\Leftarrow$. If $a^T X$ is normally $N_1(a^T\mu, a^T\Sigma a)$-distributed, then by the argument that we just gave, it is distributed as $a^T Y$ for a $N_k(\mu, \Sigma)$-distributed vector Y. If this is true for every a, then X and Y are equal in distribution by Lemma 6.3. Thus X is $N_k(\mu, \Sigma)$-distributed. ∎

2.5 Corollary. *If the vector $X = (X_1, \ldots, X_k)$ is $N_k(\mu, \Sigma)$-distributed and $A\colon \mathbb{R}^k \to \mathbb{R}^m$ is an arbitrary matrix, then AX is $N_m(A\mu, A\Sigma A^T)$-distributed.*

Proof. The parameters $A\mu$ and $A\Sigma A^T$ are correct, because they are the mean and covariance of AX. It suffices to prove the normality. For every vector a, we have $a^T(AX) = (A^T a)^T X$. This variable is one-dimensional normally distributed by the preceding lemma. Thus AX is multivariate-normally distributed by the preceding lemma in the other direction. ∎

The preceding theorem and lemma show that the marginal distributions of a multivariate-normal distribution are normal. The converse is false:

if each of $X_1, \ldots, X_k$ is normally distributed, then the vector $(X_1, \ldots, X_k)$ is not necessarily multivariate-normally distributed. (See the problems for examples.)

We close with a surprising property of multivariate normal vectors. Independent random variables are always uncorrelated, but the converse can fail easily. If the vector $X = (X_1, \ldots, X_k)$ is multivariate-normally distributed, then the converse is true!

2.6 Lemma. *The vector $X = (X_1, \ldots, X_k)$ is multivariate-normally distributed with Σ a diagonal matrix if and only if $X_1, \ldots, X_k$ are independent and marginally normally distributed.*

Proof. A diagonal, symmetric nonnegative definite matrix Σ can be written as $\Sigma = LL^T$ for L the diagonal matrix with the square roots of the diagonal elements of Σ. Then, by definition, if X is $N_k(\mu, \Sigma)$-distributed, it is distributed as $\mu + LZ = (\mu_1 + L_{11}Z_1, \ldots, \mu_k + L_{kk}Z_k)$ for independent standard normal variables $Z_1, \ldots, Z_k$. Hence its coordinates are independent and normally distributed.

Conversely, if $X_1, \ldots, X_k$ are independent and $N(\mu_i, \sigma_i^2)$-distributed, then X is distributed as $(\mu_1 + \sigma_1 Z_1, \ldots, \mu_k + \sigma_k Z_k) = \mu + LZ$, for L the diagonal matrix with diagonal $(\sigma_1, \ldots, \sigma_k)$. Thus X is $N(\mu, LL^T)$-distributed, where LL^T is a diagonal matrix. ∎

2.3 Multivariate Central Limit Theorem

The 'ordinary' central limit theorem asserts that, given a sequence $Y_1, Y_2, \ldots$ of i.i.d. random variables with finite mean μ and finite variance σ^2, the sequence $\sqrt{n}(\bar{Y}_n - \mu)$ converges in distribution to a $N_1(0, \sigma^2)$-distribution. The central limit theorem is true in every dimension.

We define $\bar{Y}_n$ as $1/n$ times the sum of the n vectors $Y_1, \ldots, Y_n$. This is identical to the vector with as coordinates the "averages taken separately over the k coordinates".

2.7 Theorem. *Let $Y_1, Y_2, \ldots$ be i.i.d. random vectors in $\mathbb{R}^k$ with finite mean μ and finite covariance matrix Σ. Then the sequence $\sqrt{n}(\bar{Y}_n - \mu)$ converges in distribution to the $N_k(0, \Sigma)$-distribution.*

Proof. For every a we have that $a^T\sqrt{n}(\bar{Y}_n - \mu) = \sqrt{n}(\overline{a^T Y_n} - a^T\mu)$, where $\overline{a^T Y_n}$ is the average of the variables $a^T Y_1, \ldots, a^T Y_n$. These have mean $a^T\mu$ and variance $a^T\Sigma a$. Thus by the ordinary central limit theorem, we have that the sequence $a^T\sqrt{n}(\bar{Y}_n - \mu)$ converges to the $N(0, a^T\Sigma a)$-distribution, for every a.

Next we note that, for every n,

$$\mathrm{E}\big\|\sqrt{n}(\bar{Y}_n-\mu)\big\|^2=\sum_{i=1}^k\mathrm{E}\big(\sqrt{n}(\bar{Y}_{n,i}-\mu_i)\big)^2=\sum_{i=1}^k n\operatorname{var}\bar{Y}_{n,i}=\sum_{i=1}^k\Sigma_{i,i}<\infty.$$

By Markov's inequality, the sequence $X_n=\sqrt{n}(\bar{Y}_n-\mu)$ is bounded in probability. By Prohorov's theorem, every subsequence n_i has a further subsequence $n_{i(j)}$ along which $X_{n_{i(j)}}\rightsquigarrow X$, as $j\to\infty$, for some X. By the continuous-mapping theorem, $a^TX_{n_{i(j)}}\rightsquigarrow a^TX$ for every a. In view of the preceding paragraph a^TX possesses the $N(0,a^T\Sigma a)$-distribution for every a and hence, by Lemma 2.4, the vector X is $N_k(0,\Sigma)$-distributed.

Thus every subsequence of the sequence $\sqrt{n}(\bar{Y}_n-\mu)$ has a further subsequence that converges in distribution to the $N_k(0,\Sigma)$-distribution. Then the whole sequence converges. ∎

2.4 Quadratic Forms

The *chisquare distribution* with k degrees of freedom is (by definition) the distribution of $\sum_{i=1}^k Z_i^2$ for independent $N(0,1)$-distributed variables $Z_1,\ldots,Z_k$. Note that the sum of squares is the squared norm $\|Z\|^2$ of the standard normal vector $Z=(Z_1,\ldots,Z_k)$. The following lemma gives a characterization of the distribution of the norm of a general zero-mean normal vector.

2.8 Lemma. *If the vector X is $N_k(0,\Sigma)$-distributed, then $\|X\|^2$ is distributed as $\sum_{i=1}^k\lambda_iZ_i^2$ for independent $N(0,1)$-distributed variables $Z_1,\ldots,Z_k$ and $\lambda_1,\ldots,\lambda_k$ the eigenvalues of Σ.*

Proof. There exists an orthogonal matrix O such that $O\Sigma O^T=\operatorname{diag}(\lambda_i)$. By Corollary 2.5, the vector OX is $N_k\big(0,\operatorname{diag}(\lambda_i)\big)$-distributed. This is also the distribution of the vector $(\sqrt{\lambda_1}Z_1,\ldots,\sqrt{\lambda_k}Z_k)$. Consequently, $\|X\|^2=\|OX\|^2$ has the same distribution as $\sum(\sqrt{\lambda_i}Z_i)^2$. ∎

The distribution of a quadratic form of the type $\sum_{i=1}^k\lambda_iZ_i^2$ is complicated in general. However, in the case that every λ_i is either 0 or 1, it reduces to a chisquare distribution. If this is not naturally the case in an application, then a statistic is often transformed to achieve this desirable situation. In the following section we give some examples.

2.5 Chisquare Tests

Suppose that we observe a vector $X_n = (X_{n,1}, \ldots, X_{n,k})$ with the multinomial distribution corresponding to n trials, and with k classes having probabilities $p = (p_1, \ldots, p_k)$. The *Pearson statistic* for testing the null hypothesis $H_0: p = a$ is given by

$$C_n^2 = \sum_{i=1}^{k} \frac{(X_{n,i} - na_i)^2}{na_i}.$$

We shall show that the sequence C_n^2 converges in distribution to a chisquare distribution if the null hypothesis is true. The practical relevance is, that we can use the chisquare table to find critical values for the test. The proof shows why Pearson divided the squares by na_i, and did not propose the simpler statistic $\|X_{n,i} - na_i\|^2$.

2.9 Theorem. *If the vectors X_n are multinomially distributed with parameters n and $a = (a_1, \ldots, a_k) > 0$, then the sequence C_n^2 converges in distribution to the χ^2_{k-1}-distribution.*

Proof. The vector X_n can be thought of as the sum $X_n = \sum_{i=1}^n Y_i$ of n independent multinomial vectors $Y_1, \ldots, Y_n$ with parameters 1 and $a = (a_1, \ldots, a_k)$. We know (or calculate) that

$$\mathrm{E}Y_i = a; \qquad \operatorname{Cov} Y_i = \begin{pmatrix} a_1(1-a_1) & -a_1a_2 & \cdots & -a_1a_k \\ -a_2a_1 & a_2(1-a_2) & \cdots & -a_2a_k \\ \vdots & \vdots & & \vdots \\ -a_ka_1 & -a_ka_2 & \cdots & a_k(1-a_k) \end{pmatrix}.$$

By Theorem 2.7, the sequence $n^{-1/2}(X_n - na)$ converges in distribution to the $N_k(0, \operatorname{Cov} Y_1)$-distribution. If D is the diagonal matrix with diagonal elements $1/\sqrt{a_i}$, then $Dn^{-1/2}(X_n - na) \rightsquigarrow N_k(0, D \operatorname{Cov} Y_1 D^T)$ by the continuous-mapping theorem. With $\sqrt{a}$ the vector with coordinates $\sqrt{a_i}$, we can rewrite this assertion as

$$\Big(\frac{X_{n,1} - na_1}{\sqrt{na_1}}, \ldots, \frac{X_{n,k} - na_k}{\sqrt{na_k}}\Big) \rightsquigarrow N(0, I - \sqrt{a}\sqrt{a}^T).$$

Since $\sum a_i = 1$, the matrix $I - \sqrt{a}\sqrt{a}^T$ has eigenvalue 0, of multiplicity 1 (with eigenspace spanned by $\sqrt{a}$), and eigenvalue 1, of multiplicity $(k-1)$ (with eigenspace equal to the orthocomplement of $\sqrt{a}$). An application of the continuous-mapping theorem and next Lemma 2.8 conclude the proof. ∎

Chisquare tests are used quite often, but usually to test more complicated hypotheses. A case of particular interest is the testing for independence of two categories. Suppose that each element of a population can be

N_{11}	N_{12}	$\cdots$	N_{1r}	$N_{1.}$
N_{21}	N_{22}	$\cdots$	N_{1r}	$N_{2.}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
N_{k1}	N_{k2}	$\cdots$	N_{1r}	$N_{k.}$
$N_{.1}$	$N_{.2}$	$\cdots$	$N_{.r}$	N

Table 2.1. Classification of a population of N elements according to two categories, N_{ij} elements having value i on the first category and value j on the second. The borders give the sums over each row and column, respectively.

classified according to two characteristics, having k and r levels, respectively. The full information concerning the classification can be given by a $(k \times r)$-*table* of the form given in Table 2.1.

Often the full information is not available, but we do know the classification $X_{n,ij}$ for a random sample of size n from the population. The matrix $X_{n,ij}$, which can also be written in the form of a $(k \times r)$-table, is multinomially distributed with parameters n and probabilities $p_{ij} = N_{ij}/N$. The null *hypothesis of independence* asserts that the two categories are independent, i.e. $H_0: p_{ij} = a_i b_j$ for (unknown) probability vectors a_i and b_j. Since the null hypothesis does not specify the values of the probabilities p_{ij}, the Pearson statistic as defined previously cannot be used. A natural modification is to replace the unknown probabilities by the estimates $\hat{a}_i\hat{b}_j$ for $\hat{a}_i = X_{n,i.}/n$ and $\hat{b}_j = X_{n,.j}/n$. These estimates are reasonable if the the null hypothesis is true. We reject the null hypothesis for large values of

$$D_n^2 = \sum_{i=1}^{k}\sum_{j=1}^{r} \frac{(X_{n,ij} - n\hat{a}_i\hat{b}_j)^2}{n\hat{a}_i\hat{b}_j}.$$

This sequence of statistics is still asymptotically chisquare distributed, but we have to "pay" for using the estimated probabilities by a "loss" of degrees of freedom.

2.10 Theorem. *If the vectors X_n are multinomially distributed with parameters n and $p_{ij} = a_i b_j > 0$, then the sequence D_n^2 converges in distribution to the $\chi^2_{(k-1)(r-1)}$-distribution.*

* **Proof.** By the multivariate central limit theorem, the sequence of matrices $n^{-1/2}(X_{n,ij} - na_ib_j)$ converges to a normally distributed matrix (X_{ij}), as before. (Write the matrix as a vector in any way you like to give a meaning to a "multivariate-normal matrix".) This motivates the decomposition

$$X_{n,ij} - n\hat{a}_i\hat{b}_j = X_{n,ij} - na_ib_j - \big(X_{n,i.} - na_i\big)b_j - a_i\big(X_{n,.j} - nb_j\big) - \frac{1}{n}\big(X_{n,i.} - na_i\big)\big(X_{n,.j} - nb_j\big).$$

The last term on the right is of lower order than the other terms, and is asymptotically negligible. By the continuous-mapping theorem and Slutsky's lemma,

$$\frac{X_{n,ij} - n\hat{a}_i\hat{b}_j}{\sqrt{n\hat{a}_i\hat{b}_j}} \rightsquigarrow \frac{X_{ij}}{\sqrt{a_ib_j}} - X_{i\cdot}\sqrt{\frac{b_j}{a_i}} - X_{\cdot j}\sqrt{\frac{a_i}{b_j}}.$$

The convergence is jointly in all coordinates (i, j) of the matrices. By the continuous-mapping theorem the squared norm of the left side converges in distribution to the squared norm of the right side. It suffices to show that the latter has a chisquare distribution with $(k-1)(r-1)$ degrees of freedom.

For simplicity we give the proof for the case that $k = r = 2$. Then the matrix whose coordinates are on the right can be written as the vector of dimension 4

$$\begin{pmatrix} X_{11}\sqrt{\frac{1}{a_1b_1}} \\ X_{12}\sqrt{\frac{1}{a_1b_2}} \\ X_{21}\sqrt{\frac{1}{a_2b_1}} \\ X_{22}\sqrt{\frac{1}{a_2b_2}} \end{pmatrix} - \begin{pmatrix} X_{1\cdot}\sqrt{\frac{b_1}{a_1}} \\ X_{1\cdot}\sqrt{\frac{b_2}{a_1}} \\ X_{2\cdot}\sqrt{\frac{b_1}{a_2}} \\ X_{2\cdot}\sqrt{\frac{b_2}{a_2}} \end{pmatrix} - \begin{pmatrix} X_{\cdot 1}\sqrt{\frac{a_1}{b_1}} \\ X_{\cdot 2}\sqrt{\frac{a_1}{b_2}} \\ X_{\cdot 1}\sqrt{\frac{a_2}{b_1}} \\ X_{\cdot 2}\sqrt{\frac{a_2}{b_2}} \end{pmatrix}.$$

The relations $\sum_{i,j} X_{n,ij} \equiv n$ have their limiting counterpart $\sum_{i,j} X_{ij} \equiv 0$. Thus $X_{2\cdot} = -X_{1\cdot}$ and $X_{\cdot 2} = -X_{\cdot 1}$, and the preceding display can be rewritten in the form

$$\begin{pmatrix} X_{11}\sqrt{\frac{1}{a_1b_1}} \\ X_{12}\sqrt{\frac{1}{a_1b_2}} \\ X_{21}\sqrt{\frac{1}{a_2b_1}} \\ X_{22}\sqrt{\frac{1}{a_2b_2}} \end{pmatrix} - X_{1\cdot}\begin{pmatrix} \sqrt{\frac{b_1}{a_1}} \\ \sqrt{\frac{b_2}{a_1}} \\ -\sqrt{\frac{b_1}{a_2}} \\ -\sqrt{\frac{b_2}{a_2}} \end{pmatrix} - X_{\cdot 1}\begin{pmatrix} \sqrt{\frac{a_1}{b_1}} \\ -\sqrt{\frac{a_1}{b_2}} \\ \sqrt{\frac{a_2}{b_1}} \\ -\sqrt{\frac{a_2}{b_2}} \end{pmatrix}.$$

The first vector we have studied before, and we have seen that it is distributed as $Y = (I - P)Z$ for a standard normal vector Z and P the orthogonal projection on the one-dimensional space spanned by the vector with coordinates $\sqrt{a_ib_j}$. The last two vectors in this display are orthogonal to the vector $\sqrt{a_ib_j}$ and are also orthogonal to each other. Now check that they are the orthogonal projections of the first vector in these directions. (See the notes on projections following the proof.) In other words, the display has the form $Y - P_1Y - P_2Y$ for P_1 and P_2 the orthogonal projections on the one-dimensional spaces spanned by the second and third vector of the display. Thus, the display is distributed as $(I - P_1 - P_2)(I - P)Z$, which is normally distributed with mean zero and covariance matrix $(I-P_1-P_2)(I-P)(I-P_1-P_2)^T = I-P-P_1-P_2$, by the

orthogonality of the three projection subspaces. The latter is a matrix with eigenvalue zero of multiplicity 3 and one of multiplicity 1. By Lemma 2.8 the square norm of the display is chisquare distributed with 1 degree of freedom. ■

In the preceding proof we use the following projection ideas. The orthogonal projection of a vector y on the 1-dimensional subspace spanned by a vector w is λw for the scalar λ determined such that $y - \lambda w$ is orthogonal to w. (Draw a picture.) Simple algebra shows that $\lambda = w^T y / w^T w$. Thus, the projection map $y \to P_w y = (w^T y / w^T w) w$ is linear, and is given by the matrix

$$P_w = \frac{1}{w^T w} w w^T.$$

If the vectors v and w are orthogonal, then the product of their projections vanishes: $P_v P_w = 0$. This follows from the algebraic formula, but is also clear from the geometric picture. Under the same orthogonality condition, the sum $P_v + P_w$ is the orthogonal projection on the 2-dimensional space spanned by v and w. The matrix of an orthogonal projection P is symmetric, and idempotent: $P^2 = P$.

Problems

1. Suppose that the random variables X and Y are uncorrelated and have variance 1. Find the covariance matrix of the vector $(X - Y, X + Y)^T$.

2. If the random vectors X and Y are independent, then $\operatorname{Cov}(X + Y) = \operatorname{Cov} X + \operatorname{Cov} Y$. Prove this.

3. Determine the covariance matrix of a random vector X with the multinomial distribution with parameters n and $p = (p_1, \ldots, p_k)$. Is it singular?

4. The *Laplace transform* of a random vector X is the map $z \to \mathrm{E} \exp z^T X$ (which could take the value ∞). Show that the Laplace transform of a $N_k(\mu, \Sigma)$-distributed vector is given by $z \to \exp(z^T \mu + \frac{1}{2} z^T \Sigma z)$.

5. Give an explicit formula for the multivariate-normal density of a vector (X, Y), whose coordinates have mean zero, variances σ_1^2 and σ_2^2, and correlation ρ.

6. Let X and Y be independent standard normal variables. Show that $X + Y$ and $X - Y$ are independent. Does this surprise you at least a little? If not can you think of two other independent variables whose sum and difference are independent?

7. Suppose that X is symmetrically distributed around 0 and has finite second moment. Show that X and $Y = X^2$ are uncorrelated. Are they every independent?

8. Let $(X_1, \ldots, X_k, Y_1, \ldots, Y_l)$ be multivariate-normally distributed. Show that $(X_1, \ldots, X_k)$ and $(Y_1, \ldots, Y_l)$ are independent if and only if $\operatorname{cov}(X_i, Y_j) = 0$ for every i and j.

9. Let $X_1, \ldots, X_n$ be independent standard normal variables. Show that $\bar{X}_n$ and $(X_1 - \bar{X}_n, \ldots, X_n - \bar{X}_n)$ are independent.

10. Let $X_1, \ldots, X_n$ be independent standard normal variables Show that $\bar{X}_n$ and $n^{-1} \sum |X_i - \bar{X}_n|$ are independent.

11. Suppose that the vector (X, Y) has probability density

$$(x, y) \to \frac{1}{\pi} e^{-\frac{1}{2}(x^2+y^2)} 1\{xy > 0\}.$$

(Note the domain!) Does (X, Y) possess a multivariate-normal distribution? Find the marginal distributions.

12. If the random variables X and Y are (marginally) normally distributed and $\operatorname{cov}(X, Y) = 0$, then (X, Y) is 2-dimensional normally distributed. If you think that this statement is true, prove it. Otherwise, give a counterexample.

13. Let $X_1, \ldots, X_n$ be i.i.d. with finite fourth moment. Find constants a, b and c_n such that the sequence $c_n(\bar{X}_n - a, \overline{X_n^2} - b)$ converges in distribution. Determine the limit law. Here $\bar{X}_n$ and $\overline{X_n^2}$ are the averages of the X_i and the X_i^2, respectively.

14. Let $Z_1, \ldots, Z_n$ be independent standard normal variables. Show that the vector $U = (Z_1, \ldots, Z_n)/N$, where $N^2 = \sum_{i=1}^n Z_i^2$, is uniformly distributed over the unit sphere S^{n-1} in $\mathbb{R}^n$ in the sense that U and OU are identically distributed for every orthogonal transformation O of $\mathbb{R}^n$.

15. For each n let U_n be uniformly distributed over the unit sphere S^{n-1} in $\mathbb{R}^n$. Show that the vectors $\sqrt{n}(U_{n,1}, U_{n,2})$ converge in distribution to a pair of independent standard normal variables as $n \to \infty$. [Use the previous problem.]

16. Suppose that T_n and S_n are sequences of estimators such that

$$\sqrt{n}(T_n - \theta) \rightsquigarrow N_k(0, \Sigma), \qquad S_n \xrightarrow{\mathrm{P}} \Sigma,$$

for a certain vector θ and a nonsingular matrix Σ. Show that S_n is nonsingular with probability tending to one and that $\left\{\theta \colon n(T_n - \theta)^T S_n^{-1}(T_n - \theta) \le \chi^2_{k,\alpha}\right\}$ is a confidence ellipsoid of asymptotic confidence level $1 - \alpha$.

17. (Cochran's theorem.) Let X be $N_k(0, I)$-distributed and let $H_0 \subset H \subset \mathbb{R}^k$ be linear subspaces. Let $H^\perp$ the set of all $x \in \mathbb{R}^k$ such that $\langle x, h\rangle = 0$ for every $h \in H$.

(i) Show that the orthogonal projections $P_0 X$ of X onto H_0 and $(I - P)X$ of X onto $H^\perp$ are independent;

(ii) Find the joint distribution of $\left(\|P_0 X\|^2, \|(I - P)X\|^2\right)$.

18. Suppose that Y is $N_n(\mu, \sigma^2 I)$ distributed, where μ and σ^2 are parameters that are known to belong a linear subspace H and $(0, \infty)$, respectively.

(i) Show that the maximum likelihood estimator of μ is given by the orthogonal projection $P_0 Y$ of Y onto H_0.

(ii) Find the maximum likelihood estimator of σ^2.

19. Suppose that Y is $N_n(\mu, \sigma^2 I)$ distributed, where μ and σ^2 are parameters that are known to belong a linear subspace H and $(0, \infty)$, respectively. We wish to test $H_0: \mu \in H_0$ for a linear subspace $H_0 \subset H$. As a test statistic we use $T = \|P_0 Y - PY\|^2 / \|(I - P)Y\|^2$. Show how an appropriate critical value can be obtained from a table of an F-distribution.

20. Let $Y = (Y_1, \ldots, Y_n)$ be $N_n(\mu, \sigma^2 I)$ distribution for $\mu_i = \alpha + \beta x_i$ for unknown parameters α and β and known constants $x_1, \ldots, x_n$ (the linear regression model with normal measurement errors). Take H the linear space spanned by the vectors $(1, 1, \ldots, 1)$ and $x = (x_1, \ldots, x_n)$ and H_0 the subspace spanned by $(1, 1, \ldots, 1)$. Construct a test for testing $H_0: \beta = 0$, using the preceding problem.

21. Show that the estimators $\hat{a}_i$ and $\hat{b}_j$ in Section 2.5 are the maximum likelihood estimators of a_i and b_j under the model given by the null hypothesis.

22. Suppose that X_n is binomial(n, p)-distributed. To test $H_0: p = a$ for a given value a the test statistic $|X_n - na| / \sqrt{na(1-a)}$ is reasonable.

(i) Find the critical value such that the test is asymptotically of level α;

(ii) Show that this test is equivalent to Pearson's test based on the multinomial vector $(X_n, n - X_n)$.

23. Suppose that X_m and Y_n are independent with the binomial(m, p_1) and binomial(n, p_2) distributions. To test $H_0: p_1 = p_2 = a$ for some fixed a we could use the test statistic

$$C_{m,n}^2 = \frac{|X_m - ma|^2}{ma(1-a)} + \frac{|Y_n - na|^2}{na(1-a)}.$$

(i) Find the limit distribution of $C_{m,n}^2$ as $m, n \to \infty$;

(ii) How would you modify the test if a were unknown? Can you guess the limit distribution of the modification?

3
Delta-Method

3.1 Main Result and Examples

Suppose an estimator T_n for a parameter θ is available, but the quantity of interest is $\phi(\theta)$ for some known function ϕ. A natural estimator is $\phi(T_n)$. How do the asymptotic properties of $\phi(T_n)$ follow from those of T_n?

A first result is an immediate consequence of the continuous-mapping theorem. If the sequence T_n converges in probability to θ and ϕ is continuous at θ, then $\phi(T_n)$ converges in probability to $\phi(\theta)$. Of greater interest is a similar question concerning limit distributions. In particular, if $\sqrt{n}(T_n - \theta)$ converges weakly to a limit distribution, is the same true for $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$? If ϕ is differentiable, then the answer is affirmative. Informally, we have

$$\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \approx \phi'(\theta)\,\sqrt{n}(T_n - \theta),$$

where $\phi'(\theta)$ is the derivative of ϕ at θ. If $\sqrt{n}(T_n - \theta) \rightsquigarrow T$ for some variable T, then we expect that $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'(\theta)\,T$. In particular, if $\sqrt{n}(T_n - \theta)$ is asymptotically normal $N(0, \sigma^2)$, then we expect that $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big)$ is asymptotically normal $N(0, \phi'(\theta)^2\sigma^2)$. This is proved in greater generality in the following theorem.

In the preceding paragraph it was silently understood that T_n is real-valued, but we are more interested in considering statistics $\phi(T_n)$ that are formed out of several more basic statistics. Thus, consider the situation that $T_n = (T_{n,1}, \ldots, T_{n,k})$ is vector-valued, and that $\phi\colon \mathbb{R}^k \to \mathbb{R}^m$ is a given function defined at least on a neighbourhood of θ. Recall that ϕ is *differentiable* at θ if there exists a linear map (matrix) $\phi'_\theta\colon \mathbb{R}^k \to \mathbb{R}^m$ such that

$$\phi(\theta + h) - \phi(\theta) = \phi'_\theta(h) + o(\|h\|), \qquad h \to 0.$$

All the expressions in this equation are vectors of length m, and $\|h\|$ is the Euclidean norm. The linear map $h \to \phi'_\theta(h)$ is sometimes called a "total derivative", as opposed to partial derivatives. A sufficient condition for ϕ to be (totally) differentiable is that all partial derivatives $\partial\phi_j(x)/\partial x_i$ exist for x in a neighbourhood of θ and are continuous at θ. (Just existence of the partial derivatives is not enough.) In any case, the total derivative is found from the partial derivatives. If ϕ is differentiable, then it is partially differentiable, and the derivative map $h \to \phi'_\theta(h)$ is matrix multiplication by the matrix

$$\phi'_\theta = \begin{pmatrix} \frac{\partial\phi_1}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_1}{\partial x_k}(\theta) \\ \vdots & & \vdots \\ \frac{\partial\phi_m}{\partial x_1}(\theta) & \cdots & \frac{\partial\phi_m}{\partial x_k}(\theta) \end{pmatrix}.$$

If the dependence of the derivative ϕ'_θ on θ is continuous, then ϕ is called continuously differentiable.

It is better to think of a derivative as a linear approximation $h \to \phi'_\theta(h)$ to the function $h \to \phi(\theta + h) - \phi(\theta)$, than as a set of partial derivatives. Thus the derivative at a point θ is a linear map. If the range space of ϕ is the real line (so that the derivative is a horizontal vector), then the derivative is also called the *gradient* of the function.

Note. What is usually called the derivative of a function $\phi\colon \mathbb{R} \to \mathbb{R}$, does not completely correspond to the present derivative. The derivative at a point, usually written $\phi'(\theta)$, is written here as ϕ'_θ. While $\phi'(\theta)$ is a number, the second object ϕ'_θ is identified with the map defined by $h \to \phi'_\theta(h) = \phi'(\theta)\, h$. Thus in the present terminology the usual derivative function $\theta \to \phi'(\theta)$ is a map from $\mathbb{R}$ into the set of linear maps from $\mathbb{R} \to \mathbb{R}$, not a map from $\mathbb{R} \to \mathbb{R}$. Graphically the "affine" approximation $h \to \phi(\theta) + \phi'_\theta(h)$ is the tangent to the function ϕ at θ.

Here is the Delta-method in higher dimensions.

3.1 Theorem. *Let $\phi\colon \mathbb{R}^k \to \mathbb{R}^m$ be a measurable map defined on a subset of $\mathbb{R}^k$ and differentiable at θ. Let T_n be random vectors taking their values in the domain of ϕ. If $r_n(T_n - \theta) \rightsquigarrow T$ for numbers $r_n \to \infty$, then $r_n\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \phi'_\theta(T)$. Moreover, the difference between $r_n\big(\phi(T_n) - \phi(\theta)\big)$ and $\phi'_\theta\big(r_n(T_n - \theta)\big)$ converges to zero in probability.*

Proof. Because $r_n \to \infty$, we have by Slutsky's lemma $T_n - \theta = (1/r_n) r_n(T_n - \theta) \rightsquigarrow 0T = 0$ and hence $T_n - \theta$ converges to zero in probability. Define a function g by

$$g(h) = \begin{cases} \frac{\phi(\theta+h) - \phi(\theta) - \phi'_\theta(h)}{\|h\|} & \text{if } h \neq 0, \\ 0 & \text{if } h = 0. \end{cases}$$

Then g is continuous at 0 by the differentiability of ϕ. Therefore, by the continuous mapping theorem, $g(T_n - \theta) \xrightarrow{\mathrm{P}} 0$ and hence, again by Slutsky's

lemma and the continuous mapping theorem $r_n\|T_n-\theta\|g(T_n-\theta) \xrightarrow{\mathrm{P}} \|T\|0 = 0$. Consequently,

$$r_n\big(\phi(T_n) - \phi(\theta) - \phi'_\theta(T_n-\theta)\big) = r_n\|T_n-\theta\|g(T_n-\theta) \xrightarrow{\mathrm{P}} 0.$$

This yields the last statement of the theorem. Since matrix multiplication is continuous, $\phi'_\theta\big(r_n(T_n-\theta)\big) \rightsquigarrow \phi'_\theta(T)$ by the continuous-mapping theorem. Finally, apply Slutsky's lemma to conclude that the sequence $r_n\big(\phi(T_n) - \phi(\theta)\big)$ has the same weak limit. ∎

A common situation is that $\sqrt{n}(T_n-\theta)$ converges to a multivariate normal distribution $N_k(\mu,\Sigma)$. Then the conclusion of the theorem is that the sequence $\sqrt{n}\big(\phi(T_n)-\phi(\theta)\big)$ converges in law to the $N_m\big(\phi'_\theta\mu, \phi'_\theta\Sigma(\phi'_\theta)^T\big)$ distribution.

3.2 Example (Sample variance). The sample variance of n observations $X_1,\ldots,X_n$ is defined as $S^2 = n^{-1}\sum_{i=1}^n (X_i-\bar{X})^2$, and can be written as $\phi(\bar{X},\overline{X^2})$ for the function $\phi(x,y) = y-x^2$. (For simplicity of notation, we divide by n rather than $n-1$.) Suppose that S^2 is based on a sample from a distribution with finite first to fourth moment $\alpha_1,\alpha_2,\alpha_3,\alpha_4$. By the multivariate central limit theorem

$$\sqrt{n}\left(\begin{pmatrix}\bar{X}\\ \overline{X^2}\end{pmatrix} - \begin{pmatrix}\alpha_1\\ \alpha_2\end{pmatrix}\right) \rightsquigarrow N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\alpha_2-\alpha_1^2 & \alpha_3-\alpha_1\alpha_2\\ \alpha_3-\alpha_1\alpha_2 & \alpha_4-\alpha_2^2\end{pmatrix}\right).$$

The map ϕ is differentiable at the point $\theta = (\alpha_1,\alpha_2)^T$, with derivative $\phi'_{(\alpha_1,\alpha_2)} = (-2\alpha_1, 1)$. Thus if the vector $(T_1,T_2)'$ possesses the normal distribution in the last display, then

$$\sqrt{n}\big(\phi(\bar{X},\overline{X^2}) - \phi(\alpha_1,\alpha_2)\big) \rightsquigarrow -2\alpha_1T_1 + T_2.$$

The latter variable is normally distributed with zero mean and a variance that can be expressed in $\alpha_1,\ldots,\alpha_4$. In case $\alpha_1 = 0$, this variance is simply $\alpha_4-\alpha_2^2$. The general case can be reduced to this case, because S^2 does not change if the observations X_i are replaced by the centred variables $Y_i = X_i-\alpha_1$. Write $\mu_k = \mathrm{E}Y_i^k$ for the *central moments* of the X_i. Noting that $S^2 = \phi(\bar{Y},\overline{Y^2})$ and that $\phi(\mu_1,\mu_2) = \mu_2$ is the variance of the original observations, we obtain

$$\sqrt{n}\big(S^2-\mu_2\big) \rightsquigarrow N(0,\mu_4-\mu_2^2).$$

In view of Slutsky's lemma, the same result is valid for the unbiased version $n/(n-1)S^2$ of the sample variance, because $\sqrt{n}\big(n/(n-1)-1\big) \to 0$. □

3.3 Example (Level of the χ^2-test). As an application of the preceding example, consider the χ^2-test for testing variance. Normal theory prescribes

to reject the null hypothesis H_0: $\mu_2 = 1$ for values of nS^2 exceeding the upper α-point $c_{n,\alpha}$ of the χ^2_{n-1}-distribution. If the observations are sampled from a normal distribution, then the test has exactly level α. Is this still approximately the case if the underlying distribution is not normal? Unfortunately, the answer is negative.

For large values of n, this can be seen with the help of the preceding result. The central limit theorem and the preceding example yield the two statements

$$\frac{\chi^2_{n-1} - (n-1)}{\sqrt{2n-2}} \rightsquigarrow N(0,1); \qquad \sqrt{n}\Big(\frac{S^2}{\mu_2} - 1\Big) \rightsquigarrow N(0, \kappa - 1),$$

where $\kappa = \mu_4/\mu_2^2$ is the *kurtosis* of the underlying distribution. The first statement implies that $\big(c_{n,\alpha} - (n-1)\big)/\sqrt{2n-2}\big)$ converges to the upper α-point ξ_α of the standard normal distribution. Thus the level of the χ^2-test satisfies

$$\mathrm{P}_{\mu_2=1}\Big(nS^2 > c_{n,\alpha}\Big) = \mathrm{P}\Big(\sqrt{n}\Big(\frac{S^2}{\mu_2} - 1\Big) > \frac{c_{n,\alpha} - n}{\sqrt{n}}\Big) \to 1 - \Phi\Big(\frac{\xi_\alpha\sqrt{2}}{\sqrt{\kappa - 1}}\Big).$$

The asymptotic level reduces to $1 - \Phi(\xi_\alpha) = \alpha$ if and only if the kurtosis of the underlying distribution is 3. This is the case for normal distributions. On the other hand, heavy-tailed distributions have a much larger kurtosis. If the kurtosis of the underlying distribution is "close to" infinity, then the asymptotic level is close to $1 - \Phi(0) = 1/2$. We conclude that the level of the χ^2-test is nonrobust against departures of normality that affect the value of the kurtosis. At least this is true if the critical values of the test are taken from the chisquare distribution with $(n-1)$ degrees of freedom. If, instead, we would use a normal approximation to the distribution of $\sqrt{n}(S^2/\mu_2 - 1)$ the problem would not arise, provided the asymptotic variance $\kappa - 1$ is estimated accurately. □

Laplace	0.12
$0.95\,N(0,1) + 0.05\,N(0,9)$	0.12

Table 3.1. Level of the test that rejects if nS^2/μ_2 exceeds the 0.95-quantile of the χ^2_{19}-distribution. (Approximations based on simulation of 10000 samples.)

In the preceding example the asymptotic distribution of $\sqrt{n}(S^2 - \sigma^2)$ was obtained by the Delta-method. Actually, it can also and more easily be derived by a direct expansion. Write

$$\sqrt{n}\big(S^2 - \sigma^2\big) = \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 - \sigma^2\Big) - \sqrt{n}(\bar{X} - \mu)^2.$$

The second term converges to zero in probability, while the first term is asymptotically normal by the central limit theorem. The whole expression is asymptotically normal by Slutsky's lemma.

Thus it is not always a good idea to apply general theorems. However, in many examples the Delta-method is a good way to package the mechanics of Taylor expansions in a transparent way.

3.4 Example. Consider the joint limit distribution of the sample variance S^2 and the t-statistic $\bar{X}/S$. Again for the limit distribution it does not make a difference whether we use a factor n or $n-1$ to standardize S^2. For simplicity we use n. Then $(S^2, \bar{X}/S)$ can be written as $\phi(\bar{X}, \overline{X^2})$ for the map $\phi\colon \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$\phi(x,y) = \Big(y - x^2, \frac{x}{(y-x^2)^{1/2}}\Big).$$

The joint limit distribution of $\sqrt{n}(\bar{X} - \alpha_1, \overline{X^2} - \alpha_2)$ is derived in the preceding example. The map ϕ is differentiable at $\theta = (\alpha_1, \alpha_2)$ provided $\sigma^2 = \alpha_2 - \alpha_1^2$ is positive, with derivative

$$\phi'_{(\alpha_1,\alpha_2)} = \begin{pmatrix} -2\alpha_1 & 1 \\ \frac{\alpha_1^2}{(\alpha_2-\alpha_1^2)^{3/2}} + \frac{1}{(\alpha_2-\alpha_1^2)^{1/2}} & \frac{-\alpha_1}{2(\alpha_2-\alpha_1^2)^{3/2}} \end{pmatrix}.$$

It follows that the sequence $\sqrt{n}\big(S^2 - \sigma^2, \bar{X}/S - \alpha_1/\sigma\big)$ is asymptotically bivariate-normally distributed, with zero mean and covariance matrix,

$$\phi'_{(\alpha_1,\alpha_2)} \begin{pmatrix} \alpha_2 - \alpha_1^2 & \alpha_3 - \alpha_1\alpha_2 \\ \alpha_3 - \alpha_1\alpha_2 & \alpha_4 - \alpha_2^2 \end{pmatrix} \big(\phi'_{(\alpha_1,\alpha_2)}\big)^T.$$

It is easy, but uninteresting to compute this explicitly. □

3.5 Example (Skewness). The sample *skewness* of a sample $X_1, \ldots, X_n$ is defined as

$$l_n = \frac{n^{-1}\sum_{i=1}^n (X_i - \bar{X})^3}{\big(n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2\big)^{3/2}}.$$

Not surprisingly it converges in probability to the skewness of the underlying distribution, defined as the quotient $\lambda = \mu_3/\sigma^3$ of the third central moment and the third power of the standard deviation of one observation. The skewness of a symmetric distribution, such as the normal distribution, equals zero and the sample skewness may be used to test this aspect of normality of the underlying distribution. For large samples a critical value may be determined from the normal approximation for the sample skewness.

The sample skewness can be written as $\phi(\bar{X}, \overline{X^2}, \overline{X^3})$ for the function ϕ given by

$$\phi(a,b,c) = \frac{c - 3ab + 2a^3}{(b-a^2)^{3/2}}.$$

The sequence $\sqrt{n}\big(\bar{X} - \alpha_1, \overline{X^2} - \alpha_2, \overline{X^3} - \alpha_3\big)$ is asymptotically mean zero normal by the central limit theorem, provided $\mathrm{E}X_1^6$ is finite. The value $\phi(\alpha_1, \alpha_2, \alpha_3)$ is exactly the population skewness. The function ϕ is differentiable at the point $(\alpha_1, \alpha_2, \alpha_3)$ and application of the Delta-method is straightforward. We can save work by noting that the sample skewness is location and scale invariant. With $Y_i = (X_i - \alpha_1)/\sigma$, the skewness can also be written as $\phi(\bar{Y}, \overline{Y^2}, \overline{Y^3})$. With $\lambda = \mu_3/\sigma^3$ denoting the skewness of the the underlying distribution, the Y's satisfy

$$\sqrt{n}\begin{pmatrix} \bar{Y} \\ \overline{Y^2} - 1 \\ \overline{Y^3} - \lambda \end{pmatrix} \rightsquigarrow N\left(0, \begin{pmatrix} 1 & \lambda & \kappa \\ \lambda & \kappa - 1 & \mu_5/\sigma^5 - \lambda \\ \kappa & \mu_5/\sigma^5 - \lambda & \mu_6/\sigma^6 - \lambda^2 \end{pmatrix}\right).$$

The derivative of ϕ at the point $(0, 1, \lambda)$ equals $(-3, -3\lambda/2, 1)$. Hence if T possesses the normal distribution in the display, then $\sqrt{n}(l_n - \lambda)$ is asymptotically normal distributed with mean zero and variance equal to $\mathrm{var}(-3T_1 - 3\lambda T_2/2 + T_3)$. If the underlying distribution is normal, then $\lambda = \mu_5 = 0$, $\kappa = 3$ and $\mu_6/\sigma^6 = 15$. In that case the sample skewness is asymptotically $N(0, 6)$-distributed.

An approximate level α test for normality based on the sample skewness could read: reject normality if $\sqrt{n}|l_n| > \sqrt{6}\,\xi_{\alpha/2}$. Table 3.2 gives the level of this test for different values of n. □

n	10	20	30	50
level	0.02	0.03	0.03	0.05

Table 3.2. Level of the test that rejects if $\sqrt{n}|l_n|/\sqrt{6}$ exceeds the 0.975-quantile of the normal distribution, in the case that the observations are normally distributed. (Approximations based on simulation of 10000 samples.)

3.2 Variance Stabilizing Transformations

Given a sequence of statistics T_n with $\sqrt{n}(T_n - \theta) \overset{\theta}{\rightsquigarrow} N\big(0, \sigma^2(\theta)\big)$ for a range of values of θ, asymptotic confidence intervals for θ are given by

$$\Big(T_n - \xi_\alpha \frac{\sigma(\theta)}{\sqrt{n}}, T_n + \xi_\alpha \frac{\sigma(\theta)}{\sqrt{n}}\Big).$$

These are asymptotically of level $1 - 2\alpha$ in that the probability that θ is covered by the interval converges to $1 - 2\alpha$ for every θ. Unfortunately, as stated these intervals are useless, because of their dependence on the unknown θ. One solution is to replace the unknown standard deviations $\sigma(\theta)$ by estimators. If the sequence of estimators is chosen consistent, then

the resulting confidence interval will still have asymptotic level $1-2\alpha$. Another approach is to use a variance stabilizing transformation, which will often lead to a better approximation.

The idea is that no problem arises if the asymptotic variances $\sigma^2(\theta)$ are independent of θ. While this fortunate situation is rare, it is often possible to transform the parameter into a different parameter $\eta = \phi(\theta)$, for which this idea can be applied. The natural estimator for η is $\phi(T_n)$. If ϕ is differentiable, then

$$\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \overset{\theta}{\rightsquigarrow} N\big(0, \phi'(\theta)^2\sigma^2(\theta)\big).$$

For ϕ chosen such that $\phi'(\theta)\sigma(\theta) \equiv 1$, the asymptotic variance is constant and finding an asymptotic confidence interval for $\eta = \phi(\theta)$ is easy. The solution

$$\phi(\theta) = \int \frac{1}{\sigma(\theta)}\, d\theta$$

is a *variance stabilizing transformation*. If it is well-defined, then it is automatically monotone, so that a confidence interval for η can be transformed back into a confidence interval for θ.

3.6 Example (Correlation). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a sample from a bivariate normal distribution with correlation coefficient ρ. The *sample correlation coefficient* is defined as

$$r_n = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\big\{\sum_{i=1}^n (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2\big\}^{1/2}}.$$

With the help of the Delta-method, it is possible to derive that $\sqrt{n}(r-\rho)$ is asymptotically zero-mean normal, with variance depending on the (mixed) third and fourth moments of (X, Y). This is true for general underlying distributions, provided the fourth moments exist. Under the normality assumption the asymptotic variance can be expressed in the correlation of X and Y. Tedious algebra gives

$$\sqrt{n}(r_n - \rho) \rightsquigarrow N\big(0, (1-\rho^2)^2\big).$$

It does not work very well to base an asymptotic confidence interval directly on this result. The transformation

$$\phi(\rho) = \int \frac{1}{1-\rho^2}\, d\rho = \tfrac{1}{2} \log \frac{1+\rho}{1-\rho} = \operatorname{arctanh} \rho$$

is variance stabilizing. Thus, the sequence $\sqrt{n}\big(\operatorname{arctanh} r - \operatorname{arctanh} \rho\big)$ converges to a standard normal distribution for every ρ. This leads to the asymptotic confidence interval for the correlation coefficient ρ given by

$$\Big(\tanh\big(\operatorname{arctanh} r - \xi_\alpha/\sqrt{n}\big), \tanh\big(\operatorname{arctanh} r + \xi_\alpha/\sqrt{n}\big)\Big).$$

$n\backslash\rho$	0	0.2	0.4	0.6	0.8
15	0.92	0.92	0.92	0.93	0.92
25	0.93	0.94	0.94	0.94	0.94

Table 3.3. Coverage probability of the asymptotic 95 % confidence interval for the correlation coefficient, for two values of n and and five different values of the true correlation ρ. (Approximations based on simulation of 10000 samples.)

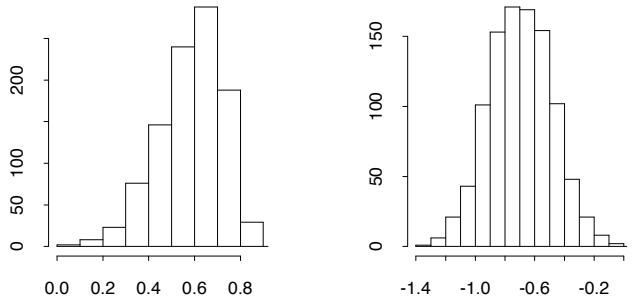

Figure 3.1. Histogram of 1000 sample correlation coefficients, based on 1000 independent samples of the the bivariate normal distribution with correlation 0.6, and histogram of the arctanh of these values.

Table 3.3 gives an indication of the accuracy of this interval. Besides stabilizing the variance the arctanh transformation has the benefit of symmetrizing the distribution of the sample correlation coefficient. □

3.3 Moment Estimators

Let $X_1, \ldots, X_n$ be a sample from a distribution that depends on a parameter θ, ranging over some set Θ. The method of moments proposes to estimate θ by the solution of a system of equations

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = \mathrm{E}_\theta f_j(X), \qquad j = 1, \ldots, k,$$

for given functions $f_1, \ldots, f_k$. Thus the parameter is chosen such that the sample moments (on the left side) match the theoretical moments. If the parameter is k-dimensional one would usually try and match k moments in

this manner. The choice $f_j(x) = x^j$ leads to the method of moments in its simplest form.

Moment estimators are not necessarily the best estimators, but under reasonable conditions they have convergence rate $\sqrt{n}$ and are asymptotically normal. This is a consequence of the Delta-method. Write the given functions in vector notation $f = (f_1, \ldots, f_k)$, and let $e\colon \Theta \to \mathbb{R}^k$ be the vector-valued expectation $e(\theta) = \mathrm{E}_\theta f(X)$. Then the moment estimator $\hat{\theta}_n$ solves the system of equations

$$\bar{f}_n \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i) = e(\theta) \equiv \mathrm{E}_\theta f(X).$$

For existence of the moment estimator, it is necessary that the vector $\bar{f}_n$ be in the range of the function e. If e is one-to-one, then the moment estimator is uniquely determined as $\hat{\theta}_n = e^{-1}(\bar{f}_n)$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \sqrt{n}\big(e^{-1}(\bar{f}_n) - e^{-1}(\mathrm{E}_{\theta_0}\bar{f}_n)\big).$$

If $\bar{f}_n$ is asymptotically normal and e^{-1} is differentiable, then the right side is asymptotically normal by the Delta-method.

The derivative of e^{-1} at $e(\theta_0)$ is the inverse $e'^{-1}_{\theta_0}$ of the derivative e'_{θ_0} of e at θ_0. Since the function e^{-1} is often not explicit, it is convenient to ascertain its differentiability from the differentiability of e. This is possible by the inverse function theorem. According to this theorem a map that is (continuously) differentiable throughout an open set with nonsingular derivatives, is locally one-to-one, is of full rank and has a differentiable inverse. Thus we obtain the following theorem.

3.7 Theorem. *Suppose that $e(\theta) = \mathrm{E}_\theta f(X)$ is one-to-one on an open set $\Theta \subset \mathbb{R}^k$, and continuously differentiable at θ_0 with nonsingular derivative e'_{θ_0}. Moreover, assume that $\mathrm{E}_{\theta_0}\|f(X)\|^2 < \infty$. Then moment estimators $\hat{\theta}_n$ exist with probability tending to one and satisfy*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{\theta_0}{\rightsquigarrow} N\Big(0, e'^{-1}_{\theta_0}\Sigma_{\theta_0}\big(e'^{-1}_{\theta_0}\big)^T\Big),$$

where Σ_{θ_0} is the covariance matrix of the vector $f(X)$ under θ_0.

Proof. By the inverse function theorem, there exists an open neighbourhood V of θ_0 on which $e\colon V \to e(V)$ is a bijection with a differentiable inverse $e^{-1}\colon e(V) \to V$. The range $e(V)$ is an open neighbourhood of $e(\theta_0)$. By the law of large numbers $\bar{f}_n \xrightarrow{\mathrm{P}} e(\theta_0)$, whence $\bar{f}_n$ is contained in $e(V)$ with probability tending to one. As soon as this is the case, the moment estimator $e^{-1}(\bar{f}_n)$ exists.

The central limit theorem guarantees the asymptotic normality of the sequence $\sqrt{n}\big(\bar{f}_n - \mathrm{E}_{\theta_0}\bar{f}_n\big)$. Finally, we use Theorem 3.1 on the display preceding the statement of the theorem. ∎

3.8 Example. Let $X_1, \ldots, X_n$ be a random sample from the Beta-distribution: the common density is equal to (with $\alpha, \beta > 0$),

$$x \to \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} 1_{0<x<1}.$$

The moment estimator for (α, β) is the solution of the system of equations

$$\begin{aligned} \bar{X}_n &= \mathrm{E}_{\alpha,\beta} X_1 = \frac{\alpha}{\alpha+\beta}, \\ \overline{X_n^2} &= \mathrm{E}_{\alpha,\beta} X_1^2 = \frac{(\alpha+1)\alpha}{(\alpha+\beta+1)(\alpha+\beta)}. \end{aligned}$$

The right hand side is a smooth and regular function of (α, β), and the equations can be solved explicitly. Hence the moment estimators exist and are asymptotically normal. □

Problems

1. Let $\hat{\lambda}_n$ be the maximum likelihood likelihood estimator of λ based on a random sample $X_1, \ldots, X_n$ from the exponential distribution with parameter λ.
(i) Find the limit distribution of the sequence $\sqrt{n}(\hat{\lambda}_n - \lambda)$;
(ii) Construct an asymptotic confidence interval based on this.

2. Find the limit distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$ for $\hat{\theta}_n$ the maximum likelihood estimator of $\theta = \mathrm{P}(X_i \leq 10)$ based on a random sample $X_1, \ldots, X_n$ from the $N(\mu, \sigma^2)$-distribution,
(i) when σ^2 is known;
(ii) when both μ and σ^2 are unknown.

3. Find the joint limit distribution of $\big(\sqrt{n}(\bar{X} - \mu), \sqrt{n}(S^2 - \sigma^2)\big)$ if $\bar{X}$ and S^2 are based on a sample of size n from a distribution with finite fourth moment. Under what condition on the underlying distribution are $\sqrt{n}(\bar{X} - \mu)$ and $\sqrt{n}(S^2 - \sigma^2)$ asymptotically independent?

4. The Pareto distribution possesses the density function, with $\alpha, \mu > 0$,

$$p_{\alpha,\mu}(x) = \frac{\alpha \mu^\alpha}{x^{\alpha+1}} 1_{x \geq \mu}.$$

Determine the limit distribution of $\sqrt{n}(\hat{\alpha}_n - \alpha)$ for $\hat{\alpha}_n$ the maximum likelihood estimator of α based on a sample of size n, when
(i) μ is known;
(ii) μ is unknown.

5. In (ii) of the preceding problem find the joint limit distribution of $\sqrt{n}(\hat{\alpha}_n - \alpha)$ and $n(\hat{\mu}_n - \mu)$. [This is hard.]

6. Find the asymptotic distribution of $\sqrt{n}(r_n - \rho)$ if r_n is the correlation coefficient of a sample of n bivariate vectors with finite fourth moments. [This is quite a bit of work. It helps to assume that means and variances equal 0 and 1, respectively.]

7. Investigate the asymptotic robustness of the level of the t-test for testing the mean that rejects $H_0: \mu \leq 0$ if $\sqrt{n}\bar{X}/S$ is larger than the upper α-quantile of the t_{n-1}-distribution.

8. Find the limit distribution of the sample kurtosis $k_n = n^{-1}\sum_{i=1}^n (X_i - \bar{X})^4/S^4$, and design an asymptotic level α test for normality based on k_n. [Warning: at least 500 observations are needed to make the normal approximation work in this case.]

9. Design an asymptotic level α test for normality based on the sample skewness and kurtosis jointly.

10. Let $X_n = (X_{n1}, X_{n2}, X_{n3})$ be multinomially distributed with parameters n and $(p_1, p_2, p_3) \in (0,1)^3$.

(i) Show that the correlation coefficient ρ of $X_1 + X_2$ and $X_2 + X_3$ is given by $\rho = -\sqrt{p_1/(1-p_1)}\sqrt{p_3/(1-p_3)}$;

(ii) Find the limit distribution of $\sqrt{n}(\hat{\rho}_n - \rho)$ for $\hat{\rho}_n$ the maximum likelihood estimator of ρ.

11. Let $X_1, \ldots, X_n$ be a random sample with expectation μ and variance 1. Find constants such that $a_n(\bar{X}_n^2 - b_n)$ converges in distribution if $\mu = 0$ or $\mu \neq 0$.

12. Let $X_1, \ldots, X_n$ be a random sample from the Poisson distribution with mean θ. Find a variance stabilizing transformation for the sample mean and construct a confidence interval for θ based on this.

13. Let $X_1, \ldots, X_n$ be i.i.d. with expectation 1 and finite variance.

(i) Find the limit distribution of $\sqrt{n}(\bar{X}_n^{-1} - 1)$.

(ii) If the random variables are sampled from a density f which is bounded and strictly positive in a neighbourhood of zero, show that $\mathrm{E}|\bar{X}_n^{-1}| = \infty$ for every n. [The density of $\bar{X}_n$ will be bounded away from zero in a neighbourhood of zero for every n.]

14. Let $X_1, \ldots, X_n$ be a sample from the uniform $[-\theta, \theta]$ distribution. Find the moment estimator of θ based on $\overline{X^2}$. Is it asymptotically normal? Can you think of an estimator for θ that converges faster to the parameter?

15. Let $X_1, \ldots, X_n$ be a random sample from the distribution function $x \to p\Phi(x - \mu) + (1-p)\Phi\big((x-\nu)/\sigma\big)$. The parameters $p \in [0,1]$, $\mu, \nu \in \mathbb{R}$ and $\sigma \in (0, \infty)$ are unknown. Construct a moment estimator for (p, μ, ν, σ) and show that it is asymptotically normal.

16. Let $X_1, \ldots, X_n$ be a sample from the 1-dimensional *exponential family* with density

$$p_\theta(x) = c(\theta)\, h(x)\, e^{\theta t(x)}.$$

Here h and t are known functions. The *natural parameter set* is the set

$$\Theta = \Big\{\theta \in \mathbb{R}\colon \int h(x) e^{\theta t(x)}\, dx < \infty\Big\}.$$

(If X is discrete, replace the integral by a sum.)

(i) Show that Θ is an interval;

(ii) Show that the solution $\hat{\theta}_n$ to the likelihood equation is a moment estimator; [Remember that a score function $\dot{\ell}_\theta = \partial/\partial\theta \log p_\theta$ satisfies $\mathrm{E}_\theta \dot{\ell}_\theta(X_i) = 0$.]

(iii) Show that $\partial/\partial\theta \mathrm{E}_\theta t(X) = \mathrm{var}_\theta\, t(X)$ and hence is strictly positive unless $t(X)$ is degenerate;

(iv) Show that $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically normal.

17. Extend the preceding problem to k-dimensional exponential families of the form

$$p_\theta(x) = c(\theta)\, h(x)\, e^{Q(\theta)^T t(x)},$$

where we assume that $\theta \to Q(\theta)$ is one-to-one from $\mathbb{R}^k$ to $\mathbb{R}^k$ and has a differentiable inverse.

18. Work out the details for Example 3.8.

19. Suppose that $\phi\colon \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable at θ with $\phi'(\theta) = 0$ and $\phi''(\theta) \neq 0$. Suppose that $\sqrt{n}(T_n - \theta) \rightsquigarrow N(0, 1)$.

(i) Show that $\sqrt{n}\big(\phi(T_n) - \phi(\theta)\big) \xrightarrow{\mathrm{P}} 0$;

(ii) Show that $n\big(\phi(T_n) - \phi(\theta)\big) \rightsquigarrow \chi^2_1$.

20. Let $X_1, \ldots, X_n$ be i.i.d. with density $f_{\lambda,a}(x) = \lambda e^{-\lambda(x-a)} 1\{x \geq a\}$ where the parameters $\lambda > 0$ and $a \in \mathbb{R}$ are unknown. Calculate the maximum likelihood estimator of $(\hat{\lambda}_n, \hat{a}_n)$ of (λ, a) and derive the asymptotic properties.

4
Z- and M-Estimators

Suppose that we are interested in a parameter (or "functional") θ attached to the distribution of observations $X_1, \ldots, X_n$. A popular method for finding an estimator $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ is to maximize a criterion function of the type

$$\theta \to M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_\theta(X_i). \tag{4.1}$$

Here $m_\theta \colon \mathcal{X} \to \bar{\mathbb{R}}$ are known functions. An estimator *m*aximizing $M_n(\theta)$ over Θ is called an *M-estimator*. In this chapter we investigate the asymptotic behaviour of a sequence of M-estimators.

Often the maximizing value will be sought by setting a derivative (or the set of partial derivatives in the multidimensional case) equal to zero. Therefore, the name M-estimator is also used for estimators satisfying systems of equations of the type

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = 0. \tag{4.2}$$

Here ψ_θ are known vector-valued maps. For instance, if θ is k-dimensional, then ψ_θ would typically have k coordinate functions $\psi_\theta = (\psi_{\theta,1}, \ldots, \psi_{\theta,k})$, and (4.2) is shorthand for the system of equations

$$\sum_{i=1}^{n} \psi_{\theta,j}(X_i) = 0, \qquad j = 1, 2, \ldots, k.$$

Even though in many examples $\psi_{\theta,j}$ is the jth partial derivative of some function m_θ, this is irrelevant for the following. Equations, such as (4.2), defining an estimator are called *estimating equations*, and the estimators they define are *Z-estimators* from zero, although they are often also referred to as M-estimators.

Sometimes the maximum of the criterion function M_n is not taken or the estimating equation does not have an exact solution. Then it is natural to use as estimator a value that almost maximizes the criterion function or is a near zero. This yields an *approximate M-estimator*. Estimators that are sufficiently close to being a point of maximum or a zero often have the same asymptotic behaviour.

An operator notation for taking expectations simplifies the formulas in this chapter. We write P for the marginal law of the observations $X_1, \ldots, X_n$, which we assume to be identically distributed. Furthermore, we denote by $\mathbb{P}_n$ the *empirical distribution*, which is defined as the (random) discrete distribution that puts mass $1/n$ at every of the observations $X_1, \ldots, X_n$. Next we write Pf for the expectation $\mathrm{E}f(X) = \int f\,dP$ and abbreviate the average $n^{-1}\sum_{i=1}^n f(X_i)$ to $\mathbb{P}_n f$. Thus the criterion functions take the forms

$$M_n(\theta) = \mathbb{P}_n m_\theta = \frac{1}{n}\sum_{i=1}^n m_\theta(X_i), \qquad \Psi_n(\theta) = \mathbb{P}_n \psi_\theta = \frac{1}{n}\sum_{i=1}^n \psi_\theta(X_i).$$

4.1 Example (Maximum likelihood estimators). Suppose $X_1, \ldots, X_n$ have a common density p_θ. Then the *maximum likelihood estimator* maximizes the likelihood $\prod_{i=1}^n p_\theta(X_i)$, or equivalently the log likelihood

$$\theta \to \sum_{i=1}^n \log p_\theta(X_i).$$

Thus, a maximum likelihood estimator is an M-estimator as in (4.1) with $m_\theta = \log p_\theta$. If the density is partially differentiable with respect to θ for each fixed x, then the maximum likelihood estimator also solves an equation of type (4.2) with ψ_θ equal to the vector of partial derivatives $\dot\ell_{\theta,j} = \partial/\partial\theta_j \log p_\theta$. The vector-valued function $\dot\ell_\theta$ is known as the *score function* of the model.

The definition (4.1) of an M-estimator may apply in cases where (4.2) does not. For instance, if $X_1, \ldots, X_n$ are i.i.d. according to the uniform distribution on $[0, \theta]$, then it makes sense to maximize the log likelihood

$$\theta \to \sum_{i=1}^n \Big(\log 1_{[0,\theta]}(X_i) - \log\theta\Big).$$

(Define $\log 0 = -\infty$.) However, this function is not smooth in θ and there exists no natural version of (4.2). Thus, the definition as the location of a maximum can be more fundamental than the definition as a zero. □

4.2 Example (Location estimators). Let $X_1, \ldots, X_n$ be a random sample of real-valued observations. Suppose we want to estimate the location of their distribution. "Location" is a vague term; it could be made precise by

defining it as the mean or median, or the centre of symmetry of the distribution if this happens to be symmetric. Two examples of *location estimators* are the sample mean and the sample median. Both are M-estimators, because they solve the equations

$$\sum_{i=1}^{n}(X_i-\theta)=0; \qquad \text{and} \qquad \sum_{i=1}^{n}\operatorname{sign}(X_i-\theta)=0,$$

respectively.† Both estimating equations involve functions of the form $\psi(x-\theta)$ for a function ψ that is monotone and odd around zero. It seems reasonable to study estimators that solve a general equation of the type

$$\sum_{i=1}^{n}\psi(X_i-\theta)=0.$$

We could call an M-estimator defined by this equation a "location" estimator, because it has the desirable property of location equivariance: if the observations X_i are shifted by a fixed amount α, then so is the estimate (in the sense that $\hat\theta+\alpha$ solves $\sum_{i=1}^{n}\psi(X_i+\alpha-\theta)=0$ if $\hat\theta$ solves the original equation).

Popular examples are the *Huber estimators* corresponding to the functions

$$\psi(x)=\begin{cases}-k & \text{if } x\le -k,\\ x & |x|\le k,\\ k & x\ge k.\end{cases}$$

The Huber estimators were motivated by studies in *robust statistics* concerning the influence of extreme data points on the estimate. The exact values of the largest and smallest observations have very little influence on the value of the median, but a proportional influence on the mean. Therefore, the sample mean is considered non-robust against outliers. If the extreme observations are thought to be rather unreliable, it is certainly of advantage to limit their influence on the estimate, although the median may be too successful in this respect. Depending on the value of k, the Huber estimators behave more like the mean (large k) or more like the median (small k), and thus bridge the gap between the non-robust mean and very robust median.

Another example are the *quantiles*. A pth sample quantile is roughly a point θ such that pn observations are less than θ and $(1-p)n$ observations are greater than θ. The precise definition has to take into account, that the value pn may not be an integer. One possibility is to call a *pth sample quantile* any $\hat\theta$ that solves the inequalities

$$-1<\sum_{i=1}^{n}\Big((1-p)1\{X_i<\theta\}-p1\{X_i>\theta\}\Big)<1. \tag{4.3}$$

† The *sign-function* is defined as $\operatorname{sign}(x)=-1,0,1$ if $x<0$, $x=0$ or $x>0$, respectively. In the case of the median the equation is valid provided the middle observation is not tied to other observations, in particular when all observations are different.

This is an approximate M-estimator for $\psi(x) = 1-p, 0, -p$ when $x < 0$, $x = 0$ or $x > 0$, respectively. The "approximate" refers to the inequalities: it is required that the value of the estimating equation be inside the interval $(-1, 1)$, rather than exactly zero. This may seem a rather wide tolerance interval for a zero. However, all solutions turn out to have the same asymptotic behaviour. In any case, except for special combinations of p and n, there is no hope of finding an exact zero, because the criterion function is discontinuous with jumps at the observations. If no observations are tied, then all jumps are of size one and at least one solution $\hat{\theta}$ to the inequalities exists.‡ When tied observations are present, it may be necessary to increase the interval $(-1, 1)$ to ensure the existence of solutions. Note that the present ψ function is monotone, as in the previous examples, but not symmetric about zero (for $p \neq 1/2$). □

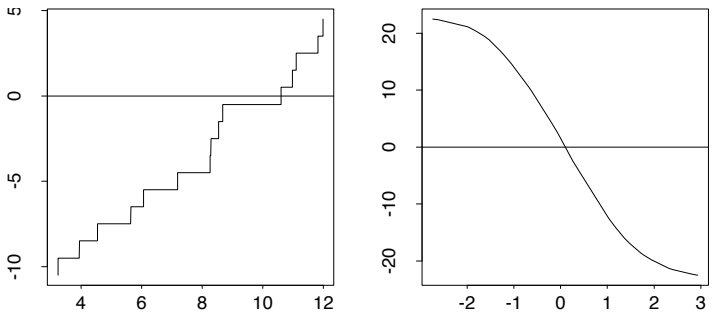

Figure 4.1. The functions $\theta \to \Psi_n(\theta)$ for the 80%-quantile and the Huber estimator for samples of size 15 from the gamma(8,1) and standard normal distribution, respectively.

4.3 Example (Regression). Consider a random sample of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the regression model

$$Y_i = f_\theta(X_i) + e_i,$$

for i.i.d. errors $e_1, \ldots, e_n$ that are independent of $X_1, \ldots, X_n$, and a function f_θ that is known up to a parameter θ. The special case $f_\theta(x) = \theta^T x$ corresponds to linear regression. A popular estimator for θ is the least squares estimator, which minimizes $\sum_{i=1}^n \big(Y_i - f_\theta(X_i)\big)^2$. An alternative is the least absolute deviation estimator, which minimizes $\sum_{i=1}^n \big|Y_i - f_\theta(X_i)\big|$. □

4.4 Example (Weighted linear regression). Consider a random sample of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ following the linear regression model

$$Y_i = \theta^T X_i + e_i,$$

‡ We could use the value of θ where the function $\theta \to \Psi_n(\theta)$ jumps across zero. This is often unique. It is certainly one of the values allowed in (4.3).

for i.i.d. errors $e_1, \ldots, e_n$ that are independent of $X_1, \ldots, X_n$. The usual estimator for the regression parameter θ is the least squares estimator, which minimizes $\sum_{i=1}^n (Y_i - \theta^T X_i)^2$. Outlying values of X_i ("leverage points") or extreme values of (X_i, Y_i) jointly ("influence points") can have an arbitrarily large influence on the value of the least squares estimator, which therefore is "nonrobust". As in the case of location estimators, a more robust estimator for θ can be obtained by replacing the square by a function $m(x)$ that grows less rapidly as $x \to \infty$, for instance $m(x) = |x|$ or $m(x)$ the primitive function of Huber's ψ. Usually, minimizing an expression of the type $\sum_{i=1}^n m(Y_i - \theta X_i)$ will be equivalent to solving a system of equations

$$\sum_{i=1}^n \psi(Y_i - \theta^T X_i) X_i = 0.$$

The estimators obtained in this way are protected against influence points, but may still suffer from leverage points, and hence are only partly robust. To obtain fully robust estimators, we can change the estimating equations to, for given weight functions $w(x)$,

$$\sum_{i=1}^n \psi(Y_i - \theta^T X_i) w(X_i) = 0.$$

Here we protect against leverage points by choosing w bounded. The choices $\psi(x) = x$ and $w(x) = x$ correspond to the (nonrobust) least squares estimator. □

4.1 Consistency

If the estimator $\hat{\theta}_n$ is used to estimate the parameter θ, then it is certainly desirable that the sequence $\hat{\theta}_n$ converges in probability to θ. If this is the case for every possible value of the parameter, then the sequence of estimators is called *asymptotically consistent*. For instance, the sample mean $\bar{X}$ is asymptotically consistent for the population mean $\mathrm{E}X$, provided the population mean exists. This follows from the law of large numbers. Not surprisingly this extends to many other sample characteristics. For instance, the sample median is consistent for the population median, whenever this is well-defined. What can be said about M-estimators in general?

We are now implicitly assuming that the set of possible parameters is a metric space. Write d for the metric. Suppose the estimator $\hat{\theta}_n$ maximizes a random criterion function

$$\theta \to M_n(\theta).$$

Clearly, the "asymptotic value" of $\hat{\theta}_n$ depends on the asymptotic behaviour of the functions M_n. Under suitable normalization it is typically the case that

$$M_n(\theta) \xrightarrow{\mathrm{P}} M(\theta), \qquad \text{every } \theta \tag{4.4}$$

for some nonrandom function $\theta \to M(\theta)$. For instance, if $M_n(\theta)$ is an average of the form $\mathbb{P}_n m_\theta$ as in (4.1), then the law of large numbers gives this result with $M(\theta) = Pm_\theta$, provided this expectation exists.

It seems reasonable to expect that the maximizer $\hat{\theta}_n$ of M_n converges to the maximizing value of M. This is what we wish to prove in this section. However, the convergence (4.4) is too weak to ensure the convergence of $\hat{\theta}_n$. Since the value $\hat{\theta}_n$ depends on the whole function $\theta \to M_n(\theta)$ an appropriate form of "functional convergence" of M_n to M is needed, strengthening the pointwise convergence (4.4). There are several possibilities. In this section we discuss an approach based on uniform convergence of the criterion functions. The assumption of uniform convergence is too strong for some applications and it is not easy to verify by elementary methods, but the approach illustrates the general idea.

Given an arbitrary random function $\theta \to M_n(\theta)$ consider estimators $\hat{\theta}_n$ that *nearly maximize* M_n, i.e.

$$M_n(\hat{\theta}_n) \geq \sup_\theta M_n(\theta) - o_P(1).$$

(Whether M_n achieves its maximum is of no importance for the consistency result.) It is assumed that the sequence M_n converges to a non-random map $M\colon \Theta \to \bar{\mathbb{R}}$. Condition (4.5) of the following theorem requires that this attains its maximum at a unique point θ_0, and only parameters close to θ_0 may yield a value of $M(\theta)$ close to the maximum values $M(\theta_0)$.

4.5 Theorem. *Let M_n be random functions and let M be a fixed function of θ such that for every $\varepsilon > 0$*

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{\mathrm{P}} 0,$$

$$\sup_{\theta : d(\theta, \theta_0) \geq \varepsilon} M(\theta) < M(\theta_0). \tag{4.5}$$

Then any sequence of estimators $\hat{\theta}_n$ that nearly maximize M_n converges in probability to θ_0.

Proof. By the property of $\hat{\theta}_n$, we have $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$. Since the uniform convergence of M_n to M implies the convergence of

$M_n(\theta_0) \xrightarrow{\mathrm{P}} M(\theta_0)$, the right side equals $M(\theta_0) - o_P(1)$. It follows that $M_n(\hat\theta_n) \geq M(\theta_0) - o_P(1)$, whence

$$\begin{aligned} M(\theta_0) - M(\hat\theta_n) &\leq M_n(\hat\theta_n) - M(\hat\theta_n) + o_P(1) \\ &\leq \sup_\theta |M_n - M|(\theta) + o_P(1) \xrightarrow{\mathrm{P}} 0. \end{aligned}$$

by the first part of assumption(4.5). By the second part of assumption (4.5), there exists for every $\varepsilon > 0$ a number $\eta > 0$ such that $M(\theta) < M(\theta_0) - \eta$ for every θ with $d(\theta, \theta_0) \geq \varepsilon$. Thus $d(\theta, \theta_0) \geq \varepsilon$ implies that $M(\theta_0) - M(\theta) > \eta$ and hence

$$\mathrm{P}\big(d(\hat\theta_n, \theta_0) \geq \varepsilon\big) \leq \mathrm{P}\big(M(\theta_0) - M(\hat\theta_n) > \eta\big) \to 0,$$

in view of the preceding display. ∎

Instead of through maximization, an M-estimator may be defined as a zero of a criterion function $\theta \to \Psi_n(\theta)$. It is again reasonable to assume that the sequence of criterion functions converges to a fixed limit:

$$\Psi_n(\theta) \xrightarrow{\mathrm{P}} \Psi(\theta).$$

Then it may be expected that a sequence of (approximate) zeros of Ψ_n converges in probability to a zero of Ψ. This is true under similar restrictions as in the case of maximizing M-estimators. In fact, this can be deduced from the preceding theorem by noting that a zero of Ψ_n maximizes the function $\theta \to -\|\Psi_n(\theta)\|$.

4.6 Theorem. *Let Ψ_n be random functions and let Ψ be a fixed function of θ such that for every $\varepsilon > 0$*

$$\begin{aligned} \sup_{\theta \in \Theta} \big\|\Psi_n(\theta) - \Psi(\theta)\big\| &\xrightarrow{\mathrm{P}} 0, \\ \inf_{\theta : d(\theta, \theta_0) \geq \varepsilon} \big\|\Psi(\theta)\big\| > 0 &= \big\|\Psi(\theta_0)\big\|. \end{aligned}$$

Then any sequence of estimators $\hat\theta_n$ such that $\Psi_n(\hat\theta_n) = o_P(1)$ converges in probability to θ_0.

Proof. This follows from the preceding theorem, on applying it to the functions $M_n(\theta) = -\|\Psi_n(\theta)\|$ and $M(\theta) = -\big\|\Psi(\theta)\big\|$. (Remember that $\big|\|x\| - \|y\|\big| \leq \|x - y\|$ for any vectors x and y.) ∎

The main difficulty with applying the preceding theorems in concrete examples is to verify the assumptions of uniform convergence of the criterion functions. There are many methods to do this. For the situation that $M_n(\theta)$ or $\Psi_n(\theta)$ takes the form of an average $\mathbb{P}_n m_\theta$ or $\mathbb{P}_n \psi_\theta$ there is even a name for the uniform convergence: a set of functions $x \to f_\theta(x)$ indexed by a

parameter θ running through a parameter set Θ is called a *Glivenko-Cantelli class* if

$$\sup_{\theta\in\Theta}\bigl|\mathbb{P}_n f_\theta - Pf_\theta\bigr| \xrightarrow{\mathrm{P}} 0.$$

This type of result is also called a uniform law of large numbers, because it asserts that $\mathbb{P}_n f_\theta \xrightarrow{\mathrm{P}} Pf_\theta$ uniformly in θ. We include, without proof, one sufficient condition for a set of functions to be Glivenko-Cantelli.

4.7 Lemma. *For every θ in a compact metric space Θ let $x \to f_\theta(x)$ be a given measurable function. Suppose that $\theta \to f_\theta(x)$ is continuous for every x and suppose that there exists a function F such that $|f_\theta| \le F$ for every θ and $PF < \infty$. Then $\sup_{\theta\in\Theta} |\mathbb{P}_n f_\theta - Pf_\theta| \xrightarrow{\mathrm{P}} 0$.*

4.8 Example (Cauchy likelihood). Suppose that we define an estimator $\hat\theta_n$ as the point of maximum of the function $\theta \to M_n(\theta) = -1/n\sum_{i=1}^n \log\bigl(1+(X_i-\theta)^2\bigr)$. (This is the log likelihood of a sample from the Cauchy distribution.) Then we can apply the preceding lemma to the functions

$$m_\theta(x) = -\log\bigl(1+(x-\theta)^2\bigr).$$

These are continuous with respect to θ, for every x. For θ ranging over a compact interval, say $\theta \in [-K, K]$, we have that $(x-\theta)^2 = x^2 - 2\theta x + \theta^2 \le x^2 + 2K|x| + K^2$ and hence

$$\bigl|m_\theta(x)\bigr| \le \log(1 + x^2 + 2K|x| + K^2).$$

We can take the function F as in the preceding lemma equal to the right side of this display. If the distribution P of the observations possesses a density p, then we shall have $PF = \int F(x)\,p(x)\,dx < \infty$ if, for instance, $p(x)$ decreases fast enough as $|x| \to \infty$. This includes for instance the Cauchy density $p(x) = (1+x^2)^{-1}\pi^{-1}$.

The restriction to parameters θ in a compact interval is mathematically somewhat unnatural, but not unpractical, because in practice we can never maximize over infinite sets.[þ] □

Even though uniform convergence of the criterion functions as in the preceding theorems can often be verified, it is stronger than needed for consistency. The following lemma is one of the many possibilities to replace the uniformity by other assumptions.

[þ] By more complicated arguments the restriction can also be shown to be mathematically unnecessary.

4.9 Lemma. *Let Θ be a subset of the real line and let Ψ_n be random functions and Ψ a fixed function of θ such that $\Psi_n(\theta) \to \Psi(\theta)$ in probability for every θ. Assume that each map $\theta \to \Psi_n(\theta)$ is nondecreasing and that $\Psi_n(\hat\theta_n) = o_P(1)$. Let θ_0 be a point such that $\Psi(\theta_0 - \varepsilon) < 0 < \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. Then $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$.*

Proof. If the map $\theta \to \Psi_n(\theta)$ is nondecreasing and has a unique zero at $\hat\theta_n$, then this zero must be strictly between every pair of points $\theta_1 < \theta_2$ with $\Psi_n(\theta_1) < 0 < \Psi_n(\theta_2)$. Therefore, for any $\varepsilon > 0$,

$$\big\{\Psi_n(\theta_0 - \varepsilon) < 0, \Psi_n(\theta_0 + \varepsilon) > 0\big\} \subset \big\{\theta_0 - \varepsilon < \hat\theta_n < \theta_0 + \varepsilon\big\}.$$

Because $\Psi_n(\theta_0 - \varepsilon) \xrightarrow{\mathrm{P}} \Psi(\theta_0 - \varepsilon) < 0$, we have that $\mathrm{P}\big(\Psi_n(\theta_0 - \varepsilon) < 0\big) \to 1$. (If $Y_n \xrightarrow{\mathrm{P}} \mu$, then $\mathrm{P}(Y_n < \mu') \to 1$ for every $\mu' > \mu$.) Combined with a similar argument for $\Psi_n(\theta_0 + \varepsilon)$, this shows that the probability of the event on the left side converges to one. Thus the probability of the event on the right side converges to one as well, and hence $\hat\theta_n$ is consistent.

If $\hat\theta_n$ is only a near zero, then this argument is not quite right. However, we still have that, for every $\varepsilon, \eta > 0$,

$$\big\{\Psi_n(\theta_0 - \varepsilon) < -\eta, \Psi_n(\theta_0 + \varepsilon) > \eta\big\} \subset \big\{\theta_0 - \varepsilon < \hat\theta_n < \theta_0 + \varepsilon\big\} \cup \big\{\Psi_n(\hat\theta_n) \notin [-\eta, \eta]\big\}.$$

For sufficiently small $\eta > 0$ (for instance 2η equal to the smallest of $-\Psi(\theta_0 - \varepsilon)$ and $\Psi(\theta_0 + \varepsilon)$), the probability of the left side still converges to one. The probability of the second event on the right side converges to zero if $\Psi_n(\hat\theta_n) = o_P(1)$. Then the probability of the first event on the right side converges to one, as desired. ∎

4.10 Example (Median). The sample median $\hat\theta_n$ is a (near) zero of the map $\theta \to \Psi_n(\theta) = n^{-1}\sum_{i=1}^n \operatorname{sign}(X_i - \theta)$. By the law of large numbers,

$$\Psi_n(\theta) \xrightarrow{\mathrm{P}} \Psi(\theta) = \mathrm{E}\operatorname{sign}(X - \theta) = \mathrm{P}(X > \theta) - \mathrm{P}(X < \theta),$$

for every fixed θ. Thus, we expect that the sample median converges in probability to a point θ_0 such that $\mathrm{P}(X > \theta_0) = \mathrm{P}(X < \theta_0)$: a population median.

This can be proved rigorously by applying Theorem 4.5. However, even though the conditions of the theorem are satisfied, they are not entirely trivial to verify. (The uniform convergence of Ψ_n to Ψ is proved essentially in Theorem 5.1 in the next chapter.) In this case it is easier to apply Lemma 4.9. Since the functions $\theta \to \Psi_n(\theta)$ are nonincreasing, it follows that $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$ provided that $\Psi(\theta_0 - \varepsilon) > 0 > \Psi(\theta_0 + \varepsilon)$ for every $\varepsilon > 0$. This is the case, for instance, if the observations possess a density that is positive in a neighbourhood of its median. (Draw a pricture to see this.) More generally, it suffices that the population median is unique: $\mathrm{P}(X < \theta_0 - \varepsilon) < \frac{1}{2} < \mathrm{P}(X < \theta_0 + \varepsilon)$ for all $\varepsilon > 0$. □

4.2 Asymptotic Normality

Suppose a sequence of estimators $\hat{\theta}_n$ is consistent for a parameter θ, which ranges over an open subset of a Euclidean space. The next question of interest concerns the order at which the discrepancy $\hat{\theta}_n - \theta$ converges to zero. The answer depends on the specific situation, but for estimators based on n replications of an experiment the order is often $n^{-1/2}$. Then multiplication with the inverse of this rate creates a proper balance, and the sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ converges in distribution, most often a normal distribution. This is interesting from a theoretical point of view. It also makes it possible to obtain approximate confidence sets. In this section we derive the asymptotic normality of M- and Z-estimators.

We can use a characterization of M-estimators either by maximization or by solving estimating equations. Consider the second possibility. Let $X_1, \ldots, X_n$ be a sample from some distribution P, and let a random and a "true" criterion function be of the form:

$$\Psi_n(\theta) \equiv \frac{1}{n}\sum_{i=1}^{n} \psi_\theta(X_i) = \mathbb{P}_n \psi_\theta; \qquad \Psi(\theta) = P\psi_\theta.$$

Assume that the estimator $\hat{\theta}_n$ is a zero of Ψ_n and converges in probability to a zero θ_0 of Ψ. For simplicity, first assume that θ is one-dimensional.

4.11 Theorem. *Assume that the map $\theta \to \psi_\theta(x)$ is twice continuously differentiable in a neighbourhood B of θ_0, for every fixed x, with derivatives $\dot{\psi}_\theta(x)$ and $\ddot{\psi}_\theta(x)$ such that $\left|\ddot{\psi}_\theta(x)\right| \le \ddot{\psi}(x)$ for a function $\ddot{\psi}$ with $P\ddot{\psi} < \infty$ and every $\theta \in B$. Furthermore, suppose that $P\psi_{\theta_0}^2 < \infty$, $P|\dot{\psi}_{\theta_0}| < \infty$ and $P\dot{\psi}_{\theta_0} \neq 0$. If $\hat{\theta}_n$ are zeros of $\theta \to \Psi_n(\theta)$ that are consistent for a zero θ_0 of $\theta \to \Psi(\theta)$, then the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a normal distribution with mean zero and variance $P\psi_{\theta_0}^2/(P\dot{\psi}_{\theta_0})^2$.*

Proof. Since $\hat{\theta}_n \to \theta_0$, it makes sense to expand $\Psi_n(\hat{\theta}_n)$ in a Taylor series around θ_0. This takes the form

$$0 = \Psi_n(\hat{\theta}_n) = \Psi_n(\theta_0) + (\hat{\theta}_n - \theta_0)\dot{\Psi}_n(\theta_0) + \tfrac{1}{2}(\hat{\theta}_n - \theta_0)^2 \ddot{\Psi}_n(\tilde{\theta}_n),$$

where $\tilde{\theta}_n$ is a point between $\hat{\theta}_n$ and θ_0. This can be rewritten as

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\sqrt{n}\Psi_n(\theta_0)}{\dot{\Psi}_n(\theta_0) + \frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)}.$$

If we can show that $\ddot{\Psi}_n(\tilde{\theta}_n) = O_P(1)$, then, since $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ by assumption, the term $\frac{1}{2}(\hat{\theta}_n - \theta_0)\ddot{\Psi}_n(\tilde{\theta}_n)$ is equal to $o_P(1)O_P(1) = o_P(1)$, and does not play a role in the asymptotics. Therefore, the desired result follows from several applications of Slutsky's lemma, once it has been proved that

$$\sqrt{n}\Psi_n(\theta_0) \rightsquigarrow N(0, P\psi_{\theta_0}^2), \quad \dot{\Psi}_n(\theta_0) \xrightarrow{\mathrm{P}} P\dot{\psi}_{\theta_0}, \quad \ddot{\Psi}_n(\tilde{\theta}_n) = O_P(1).$$

Since $\sqrt{n}\Psi_n(\theta_0) = n^{-1/2}\sum \psi_{\theta_0}(X_i)$ and $\mathrm{E}\psi_{\theta_0}(X_i) = P\psi_{\theta_0} = 0$, the first follows by the central limit theorem. Since $\dot{\Psi}_n(\theta_0) = n^{-1}\sum_{i=1}^n \dot{\psi}_{\theta_0}(X_i)$ is an average, the second follows by the law of large numbers.

It remains to prove the third statement: $\ddot{\Psi}_n(\tilde{\theta}_n) = O_P(1)$. This also concerns an average and therefore should follow by the law of large numbers. However, since the terms $\ddot{\psi}_{\tilde{\theta}_n}(X_i)$ in the average $\ddot{\Psi}_n(\tilde{\theta}_n)$ are stochastically dependent through $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \ldots, X_n)$, we must be more careful. On the event $A_n = \{\tilde{\theta}_n \in B\}$, which happens with probability tending to one by assumption, we have, by the triangle inequality,

$$\bigl|\ddot{\Psi}_n(\tilde{\theta}_n)\bigr| \leq \frac{1}{n}\sum_{i=1}^n \bigl|\ddot{\psi}_{\tilde{\theta}_n}(X_i)\bigr| \leq \frac{1}{n}\sum_{i=1}^n \ddot{\psi}(X_i).$$

The right side converges in probability to $P\ddot{\psi}$, by the law of large numbers, and hence is bounded in probability. Now, for any constant M,

$$\mathrm{P}\Bigl(\bigl|\ddot{\Psi}_n(\tilde{\theta}_n)\bigr| > M\Bigr) \leq \mathrm{P}\Bigl(\frac{1}{n}\sum_{i=1}^n \ddot{\psi}(X_i) > M\Bigr) + \mathrm{P}(A_n^c).$$

The first term of the right can be made arbitrarily small by choice of M, uniformly in n, while the second term converges to zero as $n \to \infty$. This gives the desired result. ∎

The preceding theorem can be extended to higher dimensional parameters. For a k-dimensional parameter, we use k estimating equations. Then the criterion functions are maps $\Psi_n\colon \mathbb{R}^k \to \mathbb{R}^k$ and the derivatives $\dot{\Psi}_n(\theta_0)$ are $(k \times k)$-matrices that converge to the $(k \times k)$-matrix $P\dot{\psi}_{\theta_0}$ with entries $P\partial/\partial\theta_j \psi_{\theta_0,i}$. The final statement becomes

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N_k\left(0, \bigl(P\dot{\psi}_{\theta_0}\bigr)^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T \bigl(P\dot{\psi}_{\theta_0}^T\bigr)^{-1}\right). \tag{4.6}$$

This can be proved in the same way, taking care that ordinary multiplications become matrix multiplications and divisions multiplication by an inverse.

The most complicated assumption in the preceding theorem is the *domination condition* that requires that $\ddot{\psi}_\theta(x)$ can be bounded above by an integrable function $\ddot{\psi}$, uniformly for θ in a neighbourhood of θ_0. The "smallest" candidate for $\ddot{\psi}$ is $\sup_{\theta\in B}|\ddot{\psi}_\theta|$. (The function $\ddot{\psi}$ may be arbitrary, but we are using the double-dot notation to indicate that it is related to the second order partial derivatives.) We note that

$$\sup_{\theta\in B} \mathrm{E}\bigl|\ddot{\psi}_\theta(X)\bigr| \leq \mathrm{E}\sup_{\theta\in B}\bigl|\ddot{\psi}_\theta(X)\bigr|,$$

and the inequality is almost always strict. It is rarely possible to calculate the right side explicitly and hence we need to upper bound it by a simpler expression, such as $\mathrm{E}\ddot{\psi}(X)$ for some $\ddot{\psi}$, to see that it is finite, as required.

4.12 Example (Cauchy likelihood). The log likelihood function corresponding to a random sample from the Cauchy-distribution with location θ takes the form $\theta \to M_n(\theta) = -1/n \sum_{i=1}^n \log\bigl(1 + (X_i - \theta)^2\bigr)$. The maximum likelihood estimator is a zero of $\Psi_n(\theta) = \mathbb{P}_n \psi_\theta$ for

$$\psi_\theta(x) = \frac{(x-\theta)}{1 + (x-\theta)^2}.$$

The partial derivatives with respect to θ are given by $\dot\psi_\theta(x) = -\psi'(x-\theta)$ and $\ddot\psi_\theta(x) = \psi''(x-\theta)$ for $\psi(x) = x/(1+x^2)$. The functions ψ, ψ' and ψ'' are continuous and have limit zero at $\pm\infty$ and hence are uniformly bounded. Therefore, the conditions of the preceding theorem are satisfied for every P such that $P\dot\psi_{\theta_0} \neq 0$. In Section 4.3 we shall see that the latter is true in particular for P equal to the Cauchy distribution with location θ_0. We can also directly verify that in this case $P\dot\psi_{\theta_0} = -1/4$.

It follows that every sequence $\hat\theta_n$ of zeros of Ψ_n that is consistent for a zero θ_0 of Ψ is asymptotically normal. In general, the function Ψ_n may have many zeros, corresponding to local maxima and minima of M_n. By the results of the preceding section the zero corresponding to the absolute maximum can be shown to be consistent for the point of absolute maximum of the limit criterion function M, but we omit a proof. □

The preceding theorem requires that the function $\theta \to \psi_\theta(x)$ possesses two continuous derivatives with respect to the parameter, for every x. This is true in many examples, but fails, for instance, for the Huber function, and the function $\psi_\theta(x) = \operatorname{sign}(x-\theta)$, which yields the median. Nevertheless, both the Huber estimator and the median are asymptotically normal. That such simple, but statistically important, examples cannot be treated by the preceding approach has motivated much effort to derive the asymptotic normality of M-estimators by more refined methods. One result is the following theorem, which assumes less than one derivative instead of two derivatives.

4.13 Theorem. *For each θ in an open subset of Euclidean space let $x \to \psi_\theta(x)$ be a vector-valued measurable function such that for every θ_1 and θ_2 in a neighbourhood of θ_0*

$$\bigl\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\bigr\| \le \dot\psi(x)\, \|\theta_1 - \theta_2\|,$$

for some measurable function $\dot\psi$ with $P\dot\psi^2 < \infty$. Assume that $P\|\psi_{\theta_0}\|^2 < \infty$ and that the map $\theta \to P\psi_\theta$ is differentiable at a zero θ_0, with nonsingular derivative matrix V_{θ_0}. If $\Psi_n(\hat\theta_n) = \Psi_n(\theta_0) - o_P(n^{-1})$, and $\hat\theta_n \xrightarrow{\mathrm{P}} \theta_0$, then the sequence $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1} P\psi_{\theta_0}\psi_{\theta_0}^T (V_{\theta_0}^{-1})^T$.

The first condition of the theorem, involving a function $\dot\psi(x)$ requires that the map $\theta \to \psi_\theta(x)$ be *Lipschitz* in θ with *Lipschitz constant* $\dot\psi(x)$,

for every fixed x. The function $\dot\psi$ may be arbitrary, but will be related to the partial derivatives of ψ_θ with respect to θ. If $\theta \to \psi_\theta(x)$ is continuously differentiable in a neighbourhood of θ with derivative $\dot\psi_\theta(x)$, then

$$\begin{aligned}\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| &= \Big\| \int_0^1 \dot\psi_{\theta_1 + t(\theta_2-\theta_1)}(x)\, dt(\theta_1 - \theta_2)\Big\| \\ &\le \sup_{0\le t\le 1} \|\dot\psi_{\theta_1+t(\theta_2-\theta_1)}(x)\|\, \|\theta_1 - \theta_2\|.\end{aligned}$$

Then the natural candidate for $\dot\psi$ is the supremum over $\dot\psi_\theta$ for θ ranging over a neighbourhood B of θ_0, and the condition reduces to

$$P \sup_{\theta\in B} \|\dot\psi_\theta\|^2 < \infty.$$

This is similar to the "domination condition" of Theorem 4.11. However, this time we are concerned with a first derivative rather than a second derivative, and this may be relaxed to a Lipschitz condition.

The final assertion of Theorem 4.13 is in agreement with (4.6), provided that we can identify

$$V_\theta = \frac{\partial}{\partial\theta} P\psi_\theta, \qquad \text{and} \quad P\dot\psi_\theta = P\frac{\partial}{\partial\theta}\psi_\theta.$$

Under the conditions of the Theorem 4.11 this "changing of the order of expectation and differentiation" is permitted. However, in general the derivative V_θ after integration may well exist, even if the pointwise derivative $\dot\psi_\theta(x)$ and hence $P\dot\psi_\theta$ do not.

The proof of the preceding theorem is too complicated to be given here. It is not the best theorem of its type, but it is a reasonable compromise between simplicity and general applicability.

4.14 Example (Huber estimator). For a given x, the function $\theta \to \psi_\theta(x) = \psi(x-\theta)$ with ψ the Huber function is differentiable except at the two points $\theta = x \pm k$, where it has different left and right derivatives. The derivative takes the values 0 and 1. From this, or from a picture, we can see that $\big|\psi(x-\theta_1) - \psi(x-\theta_2)\big| \le |\theta_1 - \theta_2|$ for every pair θ_1, θ_2. Thus the function is Lipschitz with Lipschitz constant $\dot\psi(x) = 1$.

If the probability measure P has a density p, then

$$P\psi_\theta = \int \psi(x-\theta)p(x)\, dx = \int \psi(x)p(x+\theta)\, dx.$$

For a sufficiently regular density p this function is differentiable with respect to θ with derivative $V_\theta = \int \psi(x)p'(x+\theta)\, dx$. Then the conditions of Theorem 4.13 are satisfied provided that $V_{\theta_0} \neq 0$. □

4.15 Example (Median). The sample median $\hat{\theta}_n$ is a (near) zero of the map $\theta \to P\psi_\theta$ for $\psi_\theta(x) = \operatorname{sign}(x-\theta)$. It is consistent for the population median θ_0. If the observations possess a differentiable distribution function F, then $P\psi_\theta = 1-2F(\theta)$ is differentiable at θ_0 with derivative $V_{\theta_0} = 2f(\theta_0)$. Since $P\psi_{\theta_0}^2 = P1 = 1$, we find the asymptotic variance for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ to be equal to $1/\big(4f^2(\theta_0)\big)$.

Unfortunately, the function ψ_θ does not satisfy the Lipschitz condition of the preceding theorem, as it is not even continuous. To prove the asymptotic normality of the sample median we must either refine this theorem or apply another argument. The conclusion that the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance $1/\big(4f^2(\theta_0)\big)$ is correct. □

4.16 Example (Misspecified model). Suppose that we postulate a model $\{p_\theta : \theta \in \Theta\}$ for a sample of observations $X_1, \ldots, X_n$, but the model is misspecified in that the true underlying distribution does not belong to the model. If we decide to use the postulated model anyway, and obtain an estimate $\hat{\theta}_n$ from maximizing the likelihood $\sum \log p_\theta(X_i)$, what is the asymptotic behaviour of $\hat{\theta}_n$?

At first sight, it might appear that $\hat{\theta}_n$ would behave erratically due to the use of the wrong model. However, this is not the case. First, we expect that $\hat{\theta}_n$ is asymptotically consistent for a value θ_0 that maximizes the function $\theta \to P\log p_\theta$, where the expectation is taken under the true underlying distribution P. The density p_{θ_0} could be viewed as the "projection" of the true underlying distribution P on the model using the *Kullback-Leibler divergence*, which is defined as $P\log(p_\theta/p)$, as a "distance" measure: p_{θ_0} minimizes this quantity over all densities in the model. Second, we expect that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix

$$V_{\theta_0}^{-1} P\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T V_{\theta_0}^{-1T}.$$

Here $\ell_\theta = \log p_\theta$ and V_θ is the derivative matrix of the map $\theta \to P\dot{\ell}_\theta$. The preceding theorem with $\psi_\theta = \dot{\ell}_\theta$ gives sufficient conditions for this to be true.

The asymptotics give insight in the question whether estimate $\hat{\theta}_n$ has practical value. The answer depends on the specific situation. However, if the model is not too far off from the truth, then the estimated density $p_{\hat{\theta}_n}$ may be a reasonable approximation for the true density. □

4.2.1 Asymptotic Relative Efficiency

The asymptotic distribution of Z- and M-estimators can be used in two ways. First, they allow the construction of asymptotic confidence regions.

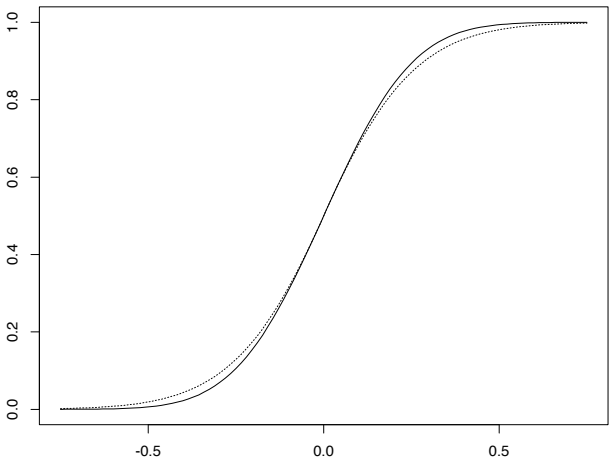

Figure 4.2. The distribution function of the sample median (dotted curve) and its normal approximation for a sample of size 25 from the Laplace distribution.

For instance, in the setting of Theorem 4.11 a natural estimator for the asymptotic variance is given by

$$\hat{\sigma}_n^2 := \frac{\mathbb{P}_n \psi_{\hat{\theta}_n}^2}{(\mathbb{P}_n \dot{\psi}_{\hat{\theta}_n})^2}.$$

Under some regularity conditions this should be consistent for the true asymptotic variance and then the confidence interval $\theta = \hat{\theta}_n \pm \hat{\sigma}_n / \sqrt{n}\, \xi_\alpha$ is of asymptotic confidence level $1 - 2\alpha$.

Second, the limit results can be used to compare the quality of different M-estimators. Suppose that we aim at estimating a parameter θ and can choose between two estimator sequences $T_{n,1}$ and $T_{n,2}$ such that, for $i = 1, 2$ and every value of θ,

$$\sqrt{n}(T_{n,i} - \theta) \overset{\theta}{\rightsquigarrow} N\big(0, \sigma_i^2(\theta)\big).$$

Then, on asymptotic grounds, we would prefer the sequence with the smallest asymptotic variance. This sequence would give better precision, at least for large n, as can be seen, for instance, from the fact that the asymptotic confidence interval obtained in the preceding paragraph would be shorter. A good quantitative measure of comparison is the quotient

$$\frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}$$

of the two asymptotic variances. This number, called the "asymptotic relative efficiency" of the two estimator sequences, has an attractive interpretation in terms of the numbers of observations needed to attain the same goal with each of two sequences of estimators.

Let $\nu \to \infty$ be a "time" index, and suppose that it is required that, as $\nu \to \infty$, our estimator sequence attains mean zero and variance $1/\nu$. More precisely, if we use the estimator sequence $T_{n,i}$, then the requirement is to use at time ν an appropriate number $n_{\nu,i}$ of observations such that, as $\nu \to \infty$,

$$\sqrt{\nu}(T_{n_{\nu,i},i} - \theta) \overset{\theta}{\rightsquigarrow} N(0,1).$$

Thus $n_{\nu,1}$ and $n_{\nu,2}$ are the numbers of observations needed to meet the requirement with each of the two estimator sequences. Then, if it exists, the limit

$$\lim_{\nu\to\infty} \frac{n_{\nu,2}}{n_{\nu,1}}$$

is called the *relative efficiency* of the estimators. (In general, it depends on the parameter θ.)

Since $\sqrt{\nu}(T_{n_\nu} - \theta)$ can be written as $\sqrt{\nu/n_\nu}\,\sqrt{n}_\nu(T_{n_\nu} - \theta)$, it follows that necessarily $n_{\nu,i} \to \infty$, and also that $n_{\nu,i}/\nu \to \sigma_i^2(\theta)$. Thus, the relative efficiency is just

$$\lim_{\nu\to\infty} \frac{n_{\nu,2}/\nu}{n_{\nu,1}/\nu} = \frac{\sigma_2^2(\theta)}{\sigma_1^2(\theta)}.$$

If the value of this quotient is bigger than 1, than the second estimator sequence needs proportionally that many observations more than the first to achieve the same (asymptotic) precision.

4.17 Example. Suppose that we wish to estimate the symmetry point of a distribution F that is symmetric about some point θ, based on a random sample from this distribution. Since the symmetry point is both the median and the mean of F, two possible estimators are the sample median and the sample mean. Their relative efficiency is equal to

$$\frac{1/(4f^2(\theta))}{\int x^2\, f(x)\, dx - \theta^2}.$$

The comparison depends on the shape of the underlying distribution. If the underlying distribution is standard normal, then the relative efficiency is equal to $\big(4/(2\pi)\big)^{-1}/1 = \pi/2$. That this is bigger than 1 should not be surprising, since the sample mean is the "best" estimator for every finite n. If the underlying distribution is Laplace (density $f(x) = \frac{1}{2}e^{-|x|}$), then the relative efficiency is equal to $(1/1)/2 = 1/2$. In this case the median is a better estimator, at least in an asymptotic sense. Using the mean instead of the median would be the same as "throwing half of the observations away". □

4.3 Maximum Likelihood Estimators

Maximum likelihood estimators are examples of M-estimators. In this section we specialize the consistency and the asymptotic normality results of the preceding sections to this important special case. (This reverses the historical order. Maximum likelihood estimators were shown to be asymptotically normal first, by Fisher in the 1920s and rigorously by Cramér in the 1940s. General M-estimators were not introduced and studied until the 1960s.)

If $X_1, \ldots, X_n$ are a random sample from a density[♯] p_θ, then the maximum likelihood estimator $\hat{\theta}_n$ maximizes the function $\theta \to \sum \log p_\theta(X_i)$, or equivalently, the function

$$M_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\theta}{p_{\theta_0}}(X_i) = \mathbb{P}_n \log \frac{p_\theta}{p_{\theta_0}}.$$

(Subtraction of the "constant" $\sum \log p_{\theta_0}(X_i)$ does not change the location of the maximum and is mathematically convenient.) If we agree to let $\log 0 = -\infty$, then this expression is with probability one well-defined if p_{θ_0} is the true density. The asymptotic function corresponding to M_n is

$$M(\theta) = \mathrm{E}_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}}(X) = P_{\theta_0} \log \frac{p_\theta}{p_{\theta_0}}.$$

The number $M(\theta)$ is called the *Kullback-Leibler divergence* of p_θ and p_{θ_0}; it is often considered a measure of "distance" between p_θ and p_{θ_0}, although it does not have the properties of a mathematical distance. Based on the results of the previous sections we may expect the maximum likelihood estimator to converge to a point of maximum of $M(\theta)$. Is the true value θ_0 always a point of maximum? The answer is affirmative and, moreover, the true value is a unique point of maximum if the true density is *identifiable*:

$$P_\theta \neq P_{\theta_0}, \qquad \text{every } \theta \neq \theta_0. \tag{4.7}$$

This requires that the model for the observations is not the same under the parameters θ and θ_0. Identifiability is a natural and even a necessary condition: if the parameter is not identifiable, then consistent estimators cannot exist.

4.18 Lemma. *Let p_θ be a collection of probability densities such that (4.7) holds. Then $M(\theta) = P_{\theta_0} \log p_\theta / p_{\theta_0}$ attains its maximum uniquely at θ_0.*

Proof. First note that $M(\theta_0) = P_{\theta_0} \log 1 = 0$. Hence we wish to show that $M(\theta)$ is strictly negative for $\theta \neq \theta_0$.

[♯] The results of this section are also valid for discrete distributions. In that case use $P_\theta(X = x)$ as the "density" $p_\theta(x)$, and replace integrals by sums.

Since $\log x \le 2\bigl(\sqrt{x}-1\bigr)$ for every positive x,

$$P_{\theta_0}\log\frac{p_\theta}{p_{\theta_0}} \le 2P_{\theta_0}\Bigl(\sqrt{\frac{p_\theta}{p_{\theta_0}}}-1\Bigr) = 2\Bigl(\int\sqrt{p_\theta p_{\theta_0}}\,dx - 1\Bigr)$$
$$= -\int\bigl(\sqrt{p_\theta}-\sqrt{p_{\theta_0}}\bigr)^2\,dx,$$

since $\int\sqrt{p_\theta}^2\,dx = 1$ for every θ. This is always nonpositive, and zero only if p_θ and p_{θ_0} define the same probability measure. By assumption the latter happens only if $\theta = \theta_0$. ∎

Thus, under regularity conditions and identifiability the sequence of maximum likelihood estimators is consistent for the true parameter. Next, consider asymptotic normality. The maximum likelihood estimator solves the likelihood equations

$$\frac{\partial}{\partial\theta}\sum_{i=1}^n \log p_\theta(X_i) = 0.$$

Hence it is an M-estimator for ψ_θ equal to the *score function* $\dot\ell_\theta = \partial/\partial\theta \log p_\theta$ of the model. In view of the preceding section, we expect that the sequence $\sqrt{n}(\hat\theta_n - \theta)$ is under θ asymptotically normal with mean zero and covariance matrix

$$\bigl(P_\theta\ddot\ell_\theta\bigr)^{-1}P_\theta\dot\ell_\theta\dot\ell_\theta^T\bigl(P_\theta\ddot\ell_\theta^T\bigr)^{-1}.$$

Under regularity conditions, this reduces to the inverse of the Fisher information matrix[†]

$$I_\theta = P_\theta\dot\ell_\theta\dot\ell_\theta^T.$$

To see this in the case of a one-dimensional parameter, differentiate the identity $\int p_\theta\,dx \equiv 1$ twice with respect to θ. Assuming that the order of differentiation and integration can be reversed, we obtain $\int\dot p_\theta\,dx \equiv \int\ddot p_\theta\,dx \equiv 0$. Together with the identities

$$\dot\ell_\theta = \frac{\dot p_\theta}{p_\theta}; \qquad \ddot\ell_\theta = \frac{\ddot p_\theta}{p_\theta} - \left(\frac{\dot p_\theta}{p_\theta}\right)^2,$$

this implies that

$$P_\theta\dot\ell_\theta = 0; \qquad P_\theta\ddot\ell_\theta = -I_\theta.$$

Thus, $\bigl(P_\theta\ddot\ell_\theta\bigr)^{-2}P_\theta\dot\ell_\theta^2 = I_\theta^{-1}$. The higher dimensional case follows in the same manner.

[†] Note that presently we take the expectation P_θ under the parameter θ, whereas the derivation in preceding section is valid for a generic underlying probability structure, and does not conceptually require that the set of parameters θ indexes a set of underlying distributions.

We conclude that maximum likelihood estimators satisfy

$$\sqrt{n}(\hat{\theta}_n - \theta) \overset{\theta}{\rightsquigarrow} N(0, I_\theta^{-1}).$$

This is a very important result, as it implies that maximum likelihood estimators are asymptotically optimal. The convergence in distribution means roughly that the maximum likelihood estimator $\hat{\theta}_n$ is $N\big(\theta, (nI_\theta)^{-1}\big)$-distributed for every θ, for large n. Hence, it is "asymptotically unbiased" and "asymptotically of variance" $(nI_\theta)^{-1}$. According to the Cramér-Rao theorem, the variance of an unbiased estimator is at least $(nI_\theta)^{-1}$. Thus, we could infer that the maximum likelihood estimator is asymptotically uniformly minimum variance unbiased and in this sense optimal. We wrote "could", because the preceding reasoning is informal and unsatisfying. Note, for instance, that the asymptotic normality does not warrant any conclusion about convergence of the moments $\mathrm{E}_\theta \hat{\theta}_n$ and $\mathrm{var}_\theta\, \hat{\theta}_n$; nor did we introduce an asymptotic version of the Cramér-Rao theorem.

However, the message that maximum likelihood estimators are *asymptotically efficient* is correct. The justification through asymptotics appears to be the only general justification of the method of maximum likelihood. In some form this result was found by Fisher in the 1920s, though a better insight was only obtained in the 1950s, 1960s and early 1970s.

4.19 Example (Logistic regression). Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent, identically distributed random vectors with Y_i taking values in $\{0, 1\}$ with conditional probabilities determined by

$$\mathrm{P}_{\alpha,\beta}(Y_i = 1 | \, X_i = x) = \frac{1}{1 + e^{-\alpha - \beta x}}.$$

The distribution of X_i is unknown, but assumed not to depend on the unknown parameters (α, β). We can estimate the parameters α and β by the maximum likelihood estimators. Even though the point of maximum of the likelihood cannot be calculated explicitly, this can be calculated numerically by an iterative scheme to solve the likelihood equations $\sum_{i=1}^n \dot{\ell}_{\alpha,\beta}(X_i) = 0$. Here the score function of the model is given by, with $\Psi(u) = (1 + e^{-u})^{-1}$,

$$\dot{\ell}_{\alpha,\beta}(x, y) = \frac{y - \Psi(\alpha + \beta x)}{\Psi(\alpha + \beta x)\big(1 - \Psi(\alpha + \beta x)\big)} \Psi'(\alpha + \beta x) \begin{pmatrix} 1 \\ x \end{pmatrix}.$$

Thus the Fisher information matrix is given by, with X distributed as $X_1, \ldots, X_n$,

$$I_{\alpha,\beta} = \mathrm{E} \frac{\Psi'(\alpha + \beta X)^2}{\Psi(\alpha + \beta X)\big(1 - \Psi(\alpha + \beta X)\big)} \begin{pmatrix} 1 & X \\ X & X^2 \end{pmatrix}.$$

This matrix is nonsingular under the (very reasonable) condition that the distribution of X is non-degenerate. □

In the preceding informal derivations and discussion, it is implicitly understood that the density p_θ possesses at least two derivatives with respect to the parameter. While this can be relaxed considerably, a certain amount of smoothness of the dependence $\theta \to p_\theta$ is essential for the asymptotic normality. Compare the behaviour of the maximum likelihood estimator in the case of uniformly distributed observations: it is neither asymptotically normal, nor asymptotically efficient.

4.20 Example. Let $X_1, \ldots, X_n$ be a sample from the uniform distribution on $[0, \theta]$. Then the maximum likelihood estimator is the maximum $X_{(n)}$ of the observations. Since the variance of $X_{(n)}$ is of the order $O(n^{-2})$ we expect that a suitable norming rate in this case is not $\sqrt{n}$, but n. Indeed, for each $x < 0$

$$\mathrm{P}_\theta\big(n(X_{(n)} - \theta) \le x\big) = \mathrm{P}_\theta\big(X_1 \le \theta + x/n\big)^n = \left(\frac{\theta + x/n}{\theta}\right)^n \to e^{x/\theta}.$$

Thus, the sequence $-n(X_{(n)} - \theta)$ converges in distribution to an exponential distribution with mean θ. Consequently, the sequence $\sqrt{n}(X_{(n)} - \theta)$ converges to zero in probability.

Note that most of the informal operations in the preceding introduction are illegal or not even defined for the uniform distribution, starting with the definition of the likelihood equations. It can be shown that the informal conclusion that the maximum likelihood estimator is asymptotically efficient is also wrong in this case. □

We conclude this section with a theorem that establishes the asymptotic normality of maximum likelihood estimators rigorously. Clearly, the asymptotic normality follows from Theorem 4.13 applied to ψ_θ equal to the score function $\dot\ell_\theta$ of the model. The following theorem applies directly to maximum likelihood estimators. It is a very clever theorem in that its conditions somehow ensure the relationship $P_\theta \ddot\ell_\theta = -I_\theta$, even though existence of a second derivative $\ddot\ell_\theta$ is not required. The proof of the theorem is omitted.

4.21 Theorem. *For each θ in an open subset of Euclidean space, let $x \to p_\theta(x)$ be a probability density such that $\theta \to \log p_\theta(x)$ is continuously differentiable for every x and such that, for every θ_1 and θ_2 in a neighbourhood of θ_0,*

$$\big|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\big| \le \dot\ell(x)\,\|\theta_1 - \theta_2\|,$$

for a measurable function $\dot\ell$ such that $P_{\theta_0}\dot\ell^2 < \infty$. Assume that the Fisher information matrix $I_\theta = P_\theta \dot\ell_\theta \dot\ell_\theta^T$ is continuous in θ and nonsingular. Then the maximum likelihood estimator $\hat\theta_n$ based on a sample of size n from p_{θ_0} satisfies that $\sqrt{n}(\hat\theta_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $I_{\theta_0}^{-1}$ provided that $\hat\theta_n$ is consistent.

Problems

1. Let $p_{\mu,\sigma^2}(x)$ be the density of the $N(\mu,\sigma^2)$-distribution and let P an arbitrary probability distribution (not necessarily normal).
(i) Calculate $M(\mu,\sigma^2) = P\log p_{\mu,\sigma^2}$;
(ii) For which (μ,σ^2) does $M(\mu,\sigma^2)$ attain its maximal value?
(iii) What does this suggest about the behaviour of the estimators $(\hat{\mu}_n,\hat{\sigma}_n^2)$ that maximize $(\mu,\sigma^2) \to \mathbb{P}_n \log p_{\mu,\sigma^2}$? Is this true?

2. Let $X_1,\ldots,X_n$ be a sample from a strictly positive density that is symmetric about some point. Show that the Huber M-estimator for location is consistent for the symmetry point.

3. Define $\psi(x) = 1-p, 0, -p$ when $x < 0, 0, > 0$.
(i) Show that $\mathrm{E}\psi(X-\theta) = 0$ implies that $\mathrm{P}(X<\theta) \le p \le \mathrm{P}(X \le \theta)$.
(ii) Derive a condition for the consistency of the corresponding Z-estimator.

4. Let $X_1,\ldots,X_n$ be a random sample from a strictly positive density. Let $\psi(x) = 2(1+e^{-x})^{-1}-1$ and let $\hat{\theta}_n$ be the solution of the equation $\sum_{i=1}^n \psi(X_i-\theta) = 0$.
(i) Show that $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ for some parameter θ_0. Express θ_0 in the density of the observations;
(ii) Prove that $\sqrt{n}(\hat{\theta}_n-\theta_0)$ converges in distribution and find an expression for the variance of the limit distribution.

5. Suppose that Θ is compact and suppose that $M\colon\Theta \to \mathbb{R}$ is continuous and attains a unique absolute maximum at θ_0. Show that (4.5) holds.

6. Suppose that Θ is a set with finitely many elements and the map $\theta \to Pm_\theta$ possesses a unique maximum at $\theta_0 \in \Theta$. Show that every sequence $\hat{\theta}_n$ such that $\mathbb{P}_n m_{\hat{\theta}_n} \ge \max_\theta \mathbb{P}_n m_\theta$ satisfies $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$. (Assume that $P|m_\theta| < \infty$ for every θ.)
(i) using a direct proof;
(ii) using Theorem 4.5.

7. Suppose that $M_n\colon\mathbb{R} \to \mathbb{R}$ and $M\colon\mathbb{R} \to \mathbb{R}$ are (deterministic) functions such that M_n has a unique absolute maximum in θ_n and M has a unique absolute maximum in θ_0. Suppose that $\sup_{\theta\in\mathbb{R}}\big|M_n(\theta)-M(\theta)\big| \to 0$ and for every $\varepsilon > 0$ there exists $\delta > 0$ such that $M(\theta) < M(\theta_0)-\delta$ for every $|\theta-\theta_0| \ge \varepsilon$. Show by a direct argument that $\theta_n \to \theta_0$.

8. Find a sequence of fixed (nonrandom) functions $M_n\colon\mathbb{R} \to \mathbb{R}$ that converges pointwise to a limit M and such that each M_n has a unique maximum at a point θ_n and M at θ_0, but the sequence θ_n does not converge to θ_0. Can you also find a sequence M_n that converges uniformly?

9. Find an expression for the asymptotic variance of the Huber estimator for location if the observations are normally distributed.

10. Define $\psi(x) = 1-p, 0, -p$ when $x < 0, 0, > 0$. Show by an informal argument that, under appropriate conditions, the corresponding Z-estimator is asymptotically normal with zero mean and asymptotic variance $p(1-p)/f\big(F^{-1}(p)\big)^2$.

11. Let $X_1, \ldots, X_n$ be independent $N(\mu, 1)$-distributed random variables. Define $\hat{\theta}_n$ as the point of minimum of the function $\theta \to \sum_{i=1}^n (X_i - \theta)^4$.

(i) Show that $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta_0$ for some θ_0. Which θ_0?

(ii) Show that $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges to a normal distribution.

(iii) Determine the asymptotic relative efficiency of $\hat{\theta}_n$ and the sample mean as an estimator of μ.

12. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from the distribution of a vector (X, Y) satisfying

$$Y = f_{\theta_0}(X) + e,$$

for e a random variable that is independent of X and a function f_θ that is known up to a parameter $\theta \in \mathbb{R}^k$. (For instance, $f_\theta(x) = \theta_1 + \theta_2 x$ or $f_\theta(x) = \theta_1 \log x + \theta_2 e^{\theta_3 x}$.) Let $\hat{\theta}_n$ be the point of minimum of $M_n(\theta) = n^{-1} \sum_{i=1}^n \big(Y_i - f_\theta(X_i)\big)^2$.

(i) Find the limit in probability M of the function M_n.

(ii) Which condition on Ee ensures that M has a point of minimum at θ_0? Which implication does this have for the consistency of $\hat{\theta}_n$?

(iii) Find, informally, an expression for the asymptotic covariance matrix of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

13. In the setting of the preceding problem, but with $k = 1$, let $\hat{\theta}_n$ be a zero of the function $\Psi_n(\theta) = n^{-1} \sum_{i=1}^n w(X_i)\big(Y_i - f_\theta(X_i)\big)$.

(i) Find the limit in probability Ψ of the function Ψ_n.

(ii) Which condition on Ee ensures that Ψ has a zero at θ_0? Which implication does this have for the consistency of $\hat{\theta}_n$?

(iii) Find, informally, an expression for the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

14. In the setting of the preceding problem, assume that the errors possess a normal distribution with mean zero and that $f_\theta(x) = \theta x$. Which of the two weight functions $w(x) = x$ and $w(x) = 1$ yields the smallest asymptotic variance? (Assume that $\mathrm{E}X_i \neq 0$.)

15. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ are a random sample from the distribution of a vector (X, Y) satisfying

$$Y = \alpha_0 + \beta_0 X + e,$$

for e a random variable that is independent of X. Let $(\hat{\alpha}_n, \hat{\beta}_n)$ be the point of minimum of $M_n(\alpha, \beta) = n^{-1} \sum_{i=1}^n \big|Y_i - \alpha - \beta X_i\big|$.

(i) Find the limit in probability M of the function M_n.

(ii) Which condition on the distribution of e ensures that M has a point of minimum at (α_0, β_0)? Which implication does this have for the consistency of $(\hat{\alpha}_n, \hat{\beta}_n)$?

(iii) Find, informally, an expression for the asymptotic covariance matrix of $\sqrt{n}(\hat{\alpha}_n - \alpha_0, \hat{\beta}_n - \beta_0)$.

16. Determine the relative efficiency of the sample median and the sample mean based on a random sample from the uniform distribution on $[0, 1]$.

17. Give an expression for the relative efficiency of the Huber estimator and the sample mean based on a random sample from the normal distribution with variance 1.

18. Calculate the Kullback-Leibler divergence between two exponential distributions with different scale parameters. When is it maximal?

19. Calculate the Kullback-Leibler divergence between two normal distributions with different location and scale parameters. When is it maximal?

20. Calculate the Kullback-Leibler divergence between two Laplace distributions with different locations. When is it maximal?

21. Let $X_1, \ldots, X_n$ be i.i.d. Poisson$(1/\theta)$-distributed.
(i) Calculate the Fisher information I_θ in one observation;
(ii) Derive the maximum likelihood estimator for θ and show by a direct argument that it is asymptotically normal with variance I_θ^{-1}.

22. Let $X_1, \ldots, X_n$ be i.i.d. $N(\theta, \theta)$-distributed.
(i) Calculate the Fisher information I_θ in one observation;
(ii) Derive the maximum likelihood estimator for θ and show by a direct argument that it is asymptotically normal with variance I_θ^{-1}.

23. Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$-distributed.
(i) Calculate the Fisher information I_{μ,σ^2} in one observation for the parameter (μ, σ^2) and its inverse;
(ii) Verify by direct calculations that the maximum likelihood estimator for (μ, σ^2) is asymptotically normal with mean zero and covariance matrix I_{μ,σ^2}^{-1}.

24. Let X be Poisson distributed with density $p_\theta(x) = \theta^x\, e^{-\theta}/x!$. Show by direct calculation that $\mathrm{E}_\theta \dot{\ell}_\theta(X) = 0$ and $\mathrm{E}_\theta \ddot{\ell}_\theta(X) = -\mathrm{E}_\theta \dot{\ell}_\theta^2(X)$. In Section 4.3 these identities are obtained informally by differentiation under the integral (sum). Is it obvious from results from analysis that this is permitted in this case?

25. Let $X_1, \ldots, X_n$ be a sample from the $N(\theta, 1)$-distribution, where it is known that $\theta \geq 0$. Show that the maximum likelihood estimator is not asymptotically normal under $\theta = 0$. Why does this not contradict the theorems of this chapter?

5
Nonparametric Estimation

Statistical models are called *parametric models* if they are described by a Euclidean parameter (in a nice way). For instance, the binomial model is described by a single parameter p, and the normal model is given through two unknowns: the mean and the variance of the observations. In many situations there is insufficient motivation for using a particular parametric model, such as a normal model. An alternative at the other end of the scale is a *nonparametric model*, which leaves the underlying distribution of the observations essentially free. In this chapter we discuss two problems of nonparametric estimation: estimating a distribution function or a density of the observations if nothing is known a-priori.

5.1 Estimating Distributions

Let $X_1, \ldots, X_n$ be a random sample from a distribution function F on the real line. The *empirical distribution function* is defined as

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\{X_i \le t\}.$$

It is the natural estimator for the underlying distribution F if this is completely unknown. Since $n\mathbb{F}_n(t)$ is binomially distributed with mean $nF(t)$, the estimator $\mathbb{F}_n(t)$ is unbiased for estimating $F(t)$ for every fixed t. By the law of large numbers it is also consistent,

$$\mathbb{F}_n(t) \xrightarrow{\text{as}} F(t), \qquad \text{every } t.$$

Furthermore, by the central limit theorem it is asymptotically normal,

$$\sqrt{n}\big(\mathbb{F}_n(t) - F(t)\big) \rightsquigarrow N\Big(0, F(t)\big(1 - F(t)\big)\Big).$$

These properties indicate that the "functional estimator" $t \to \mathbb{F}_n(t)$ is a reasonable estimator for the function $t \to F(t)$. This can also be underscored by studying the properties of $\mathbb{F}_n$ as a function, rather than the properties of $\mathbb{F}_n(t)$ for each t separately.

The *Glivenko-Cantelli theorem* extends the law of large numbers, and gives uniform convergence of the random function $t \to \mathbb{F}_n(t)$. The uniform distance

$$\|\mathbb{F}_n - F\|_\infty = \sup_t \bigl|\mathbb{F}_n(t) - F(t)\bigr|$$

is known as the *Kolmogorov-Smirnov* statistic.

5.1 Theorem (Glivenko-Cantelli). *If $X_1, X_2, \ldots$ are i.i.d. random variables with distribution function F, then $\|\mathbb{F}_n - F\|_\infty \stackrel{\text{as}}{\to} 0$.*

Proof. By the strong law of large numbers, both $\mathbb{F}_n(t) \stackrel{\text{as}}{\to} F(t)$ and $\mathbb{F}_n(t-) \stackrel{\text{as}}{\to} F(t-)$ for every t. Given a fixed $\varepsilon > 0$, there exists a partition $-\infty = t_0 < t_1 < \cdots < t_k = \infty$ such that $F(t_i-) - F(t_{i-1}) \le \varepsilon$ for every i. This is easiest seen by drawing a picture of F. (More formally, we can define $t_0 = -\infty$ and next, recursively, $t_{i+1} = \inf\bigl\{t > t_i \colon F(t) \ge F(t_i) + \varepsilon\bigr\}$. Then $F(t_{i+1}) - F(t_i) \ge \varepsilon$ by the right continuity of F and $F(t_{i+1}-) - F(t_i) \le \varepsilon$.) Now, for $t_{i-1} \le t < t_i$,

$$\begin{aligned}
\mathbb{F}_n(t) - F(t) &\le \mathbb{F}_n(t_i-) - F(t_{i-1}) \le \mathbb{F}_n(t_i-) - F(t_i-) + \varepsilon,\\
\mathbb{F}_n(t) - F(t) &\ge \mathbb{F}_n(t_{i-1}) - F(t_i-) \ge \mathbb{F}_n(t_{i-1}) - F(t_{i-1}) - \varepsilon.
\end{aligned}$$

Together these bounds yield the inequality

$$\|\mathbb{F}_n - F\|_\infty \le \sup_i \bigl|\mathbb{F}_n(t_i) - F(t_i)\bigr| \vee \sup_i \bigl|\mathbb{F}_n(t_i-) - F(t_i-)\bigr| + \varepsilon.$$

The convergence of $\mathbb{F}_n(t)$ and $\mathbb{F}_n(t-)$ for every fixed t is certainly uniform for t in the finite set $\{t_1, \ldots, t_{k-1}\}$. Conclude that $\limsup \|\mathbb{F}_n - F\|_\infty \le \varepsilon$, almost surely. This is true for every $\varepsilon > 0$, whence the lim sup is zero. ∎

The central limit theorem can also be extended from a central limit theorem for every fixed t to a "uniform" or "functional" central limit theorem, but this is more involved. A first step is to prove the joint weak convergence of finitely many coordinates. By the multivariate central limit theorem, for every $t_1, \ldots, t_k$,

$$\sqrt{n}\bigl(\mathbb{F}_n(t_i) - F(t_i), \ldots, \mathbb{F}_n(t_k) - F(t_k)\bigr) \rightsquigarrow \bigl(\mathbb{G}_F(t_1), \ldots, \mathbb{G}_F(t_k)\bigr),$$

where the vector on the right has a multivariate-normal distribution, with mean zero and covariances

$$\mathrm{E}\mathbb{G}_F(t_i)\mathbb{G}_F(t_j) = F(t_i \wedge t_j) - F(t_i)F(t_j). \tag{5.1}$$

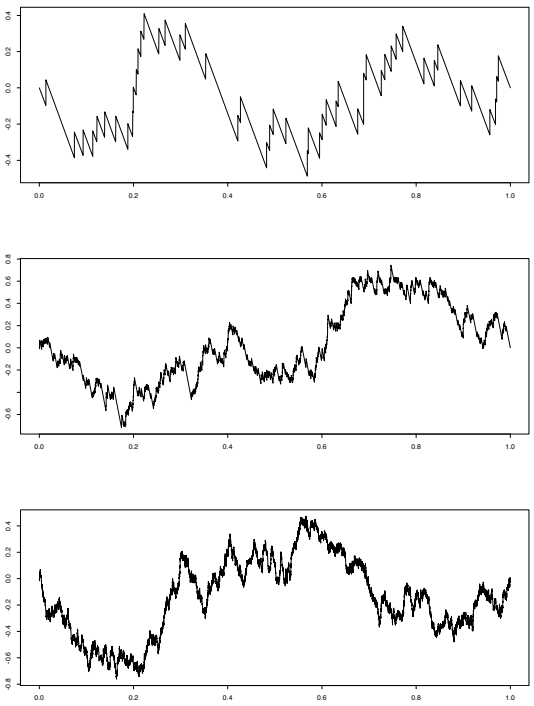

Figure 5.1. Three realizations of the uniform empirical process, of 50, 500 and 5000 observations, respectively.

This suggests that the sequence of *empirical processes* $\sqrt{n}(\mathbb{F}_n - F)$, viewed as random functions, converges in distribution to a "Gaussian stochastic process" $\mathbb{G}_F$. This can be made precise in a mathematical theorem, which is known as *Donsker's theorem*. The Gaussian process $\mathbb{G}_F$ is known as a *Brownian bridge* process.

Figure 5.1 shows some realizations of the empirical process for a sample from the uniform distribution on $[0, 1]$. The roughness of the sample path for $n = 5000$ is remarkable, and typical. It is carried over onto the limit process $\mathbb{G}_F$, for it can be shown that, with probability one, the function $t \to \mathbb{G}_F(t)$ is continuous, but nowhere differentiable. This is how the process $\mathbb{G}_F$ received its name: its sample path resemble the displacement connected to the physical "Brownian" movement of particles in a gas.

Some popular global measures of discrepancy for real-valued observations are

$$\sqrt{n}\|\mathbb{F}_n - F\|_\infty, \qquad \text{(Kolmogorov-Smirnov)},$$

$$n \int (\mathbb{F}_n - F)^2 \, dF, \qquad \text{(Cramér–von-Mises)}.$$

These are used to test whether the true distribution of the observations is F. Large values of these statistics indicate that the null hypothesis that the true distribution is F is false. To carry out a formal test of this null hypothesis an appropriate quantile is chosen from the null distribution of these statistics. For large n these null distributions do not depend much on n any more, because the statistics can be shown to converge in distribution. The critical value of the test is usually chosen equal to the upper α-quantile of the limit distribution.

It is probably practically more relevant to test the goodness-of-fit of composite null hypotheses, for instance the hypothesis that the underlying distribution F of a random sample is normal, i.e. belongs to the normal location-scale family. To test the null hypothesis that F belongs to a certain family $\{F_\theta : \theta \in \Theta\}$, it is natural to use a measure of the discrepancy between $\mathbb{F}_n$ and $F_{\hat{\theta}}$, for a reasonable estimator $\hat{\theta}$ of θ. For instance, a modified Kolmogorov-Smirnov statistic for testing normality is

$$\sup_t \sqrt{n}\Big|\mathbb{F}_n(t) - \Phi\Big(\frac{t-\bar{X}}{S}\Big)\Big|.$$

Many goodness-of-fit statistics of this type also converge to limit distributions, but these are different due to the "extra randomness" introduced by the estimator $\hat{\theta}$.

5.2 Estimating Densities

Let $X_1, \ldots, X_n$ be a random sample from a density f on the real line. If we would know that f belongs to the normal family of densities, then the natural estimate of f would be the normal density with mean $\bar{X}_n$ and variance S_n^2, i.e. the function

$$x \to \frac{1}{S_n\sqrt{2\pi}} e^{-\frac{1}{2}(x-\bar{X}_n)^2/S_n^2}.$$

In this section we suppose that we have no prior knowledge of the form of f, and want to "let the data speak as much as possible for itself".

There are several possibilities for contructing estimators $\hat{f}$ for f. In this section we discuss the *kernel method*.

Let K be a probability density with mean 0 and variance 1, for instance the standard normal density. A *kernel estimator* with *kernel* or *window* K is defined as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h} K\Big(\frac{x - X_i}{h}\Big).$$

Here h is a positive number, still to be chosen, called the *bandwidth* of the estimator. It turns out that the choice of the kernel K is not crucial, but

the quality of $\hat{f}$ as an estimator of f depends strongly on the choice of the bandwidth.

A kernel estimator is an example of a *smoothing method*. The construction of a density estimator could be viewed as smoothing out the total mass 1 over the real line. Given a random sample of n observations it is reasonable to start with allocating the total mass in packages of size $1/n$ to the observations. Next a kernel estimator distributes the mass that was allocated to X_i smoothly around X_i, not homogenously, but according to the kernel and bandwidth.

More formally, we can view a kernel estimator as the sum of n small "mountains" given by the functions

$$x \to \frac{1}{nh} K\Big(\frac{x - X_i}{h}\Big).$$

Every small mountain is centred around an observation X_i and has surface area $1/n$, for any bandwidth h. For small bandwidth the mountain is very concentrated (a peak), while for large bandwidth the mountain is low and flat. Figure 5.2 shows how the mountains can add up to a single estimator. If the bandwidth is small, then the mountains remain separated and their sum is peaky. On the other hand, if the bandwidth is large, then the sum of the individual mountains is too flat. Intermediate values of the bandwidth should give the best results.

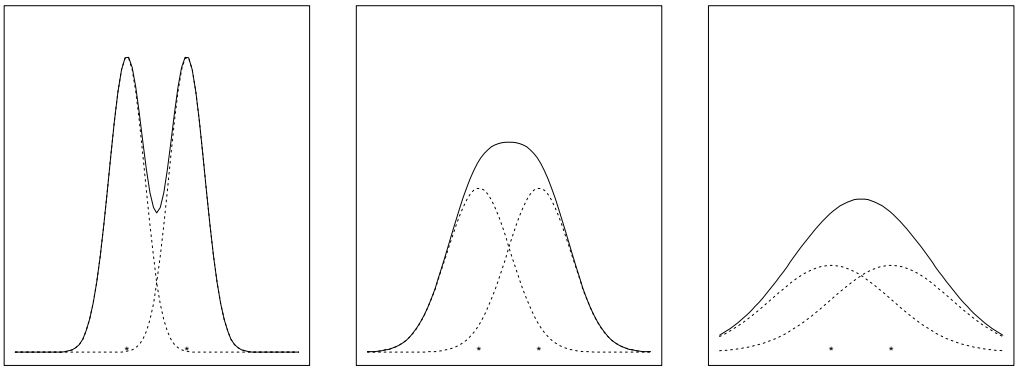

Figure 5.2. The kernel estimator with normal kernel and two observations for three bandwidths: small (left), intermediate (middle) and large (right). The figures shows both the contributions of the two observations separately (dotted lines) and the kernel estimator (solid lines), which is the sum of the two dotted lines.

Figure 5.3 shows the kernel method in action on a sample from the normal distribution. The solid and dotted lines are the estimator and the true density respectively. The three pictures give the kernel estimates using three different bandwidths, small, intermediate and large, each time with the standard normal kernel.

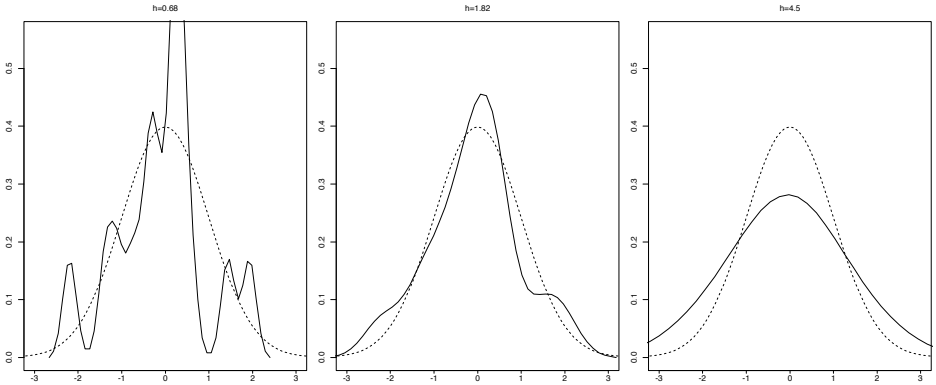

Figure 5.3. Kernel estimates for the density of a sample of size 15 from the standard normal density for three different bandwidths, using a normal kernel. The dotted line gives the true density.

A popular criterion to judge the quality of density estimators is the *mean integrated square error*, which is defined as

$$\begin{aligned}\mathrm{MISE}_f(\hat{f}) &= \int \mathrm{E}_f\big(\hat{f}(x) - f(x)\big)^2\,dx \\ &= \int \mathrm{var}_f\,\hat{f}(x)\,dx + \int \big(\mathrm{E}_f\hat{f}(x) - f(x)\big)^2\,dx.\end{aligned}$$

This is the mean square error $\mathrm{E}_f\big(\hat{f}(x) - f(x)\big)^2$ of $\hat{f}(x)$ as an estimator of $f(x)$ integrated over the argument x. If the mean integrated square error is small, then the function $\hat{f}$ is close to the function f.

As can be seen from the second representation, the mean integrated square error is the sum of an integrated "variance term" and a"bias term". The mean integrated square error can be small only if both terms are small. We shall show that the two terms are of the orders

$$\frac{1}{nh}, \qquad \text{and} \qquad h^4,$$

respectively. Then it follows that the variance and the bias terms are balanced for $(nh)^{-1} \sim h^4$, which means an optimal choice of bandwidth equal to $h \sim n^{-1/5}$.

Informally, these orders follow from simple Taylor expansions. For instance, the bias of $\hat{f}(x)$ can be written

$$\begin{aligned}\mathrm{E}_f\hat{f}(x) - f(x) &= \int \frac{1}{h}K\Big(\frac{x-t}{h}\Big)\,f(t)\,dt - f(x) \\ &= \int K(y)\big(f(x-hy) - f(x)\big)\,dy,\end{aligned}$$

by the substitution $(x-t)/h = y$. Developing f in a Taylor series around x and remembering that $\int yK(y)\,dy = 0$, we see, informally, that this is

equal to

$$\int K(y)\bigl(-hyf'(x) + \tfrac{1}{2}(-hy)^2 f''(x) + \cdots\bigr)\,dy$$
$$= \int y^2 K(y)\,dy \tfrac{1}{2}h^2 f''(x) + \cdots.$$

Thus, the square bias is of the order h^4. The variance term can be handled similarly. A precise theorem is as follows.

5.2 Theorem. *Suppose that f is twice continuously differentiable with $\int \bigl|f''(x)\bigr|^2\,dx < \infty$. Furthermore, suppose that $\int |y|K^2(y)\,dy$ is finite. Then there exists a number C_f such that for small $h > 0$*

$$\mathrm{MISE}_f(\hat{f}) \le C_f\bigl(\frac{1}{nh} + h^4\bigr).$$

Consequently, for $h_n \sim n^{-1/5}$, we have $\mathrm{MISE}_f(\hat{f}) = O(n^{-4/5})$.

Proof. Since a kernel estimator is an average of n independent random variables, the variance of $\hat{f}(x)$ is $(1/n)$ times the variance of one term. Hence

$$\mathrm{var}_f\,\hat{f}(x) = \frac{1}{n}\,\mathrm{var}_f\,\frac{1}{h}K\Bigl(\frac{x - X_1}{h}\Bigr) \le \frac{1}{nh^2}\mathrm{E}_f K^2\Bigl(\frac{x - X_1}{h}\Bigr)$$
$$= \frac{1}{nh}\int K^2(y) f(x - hy)\,dy.$$

Take the integral with repect to x on both left and right sides. Since $\int f(x - hy)\,dx = 1$ is the same for every value of hy, the right side reduces to $(nh)^{-1}\int K^2(y)\,dy$, by Fubini's theorem. (This asserts, among others, that repeated integrals of a nonnegative, measurable function of several variables does not depend on the order in which the integrals are computed.) This concludes the proof for the variance term.

To upper bound the bias term we first write the bias $\mathrm{E}_f\hat{f}(x) - f(x)$ in the form as given preceding the statement of the theorem. Next we insert the formula

$$f(x + h) - f(x) = hf'(x) + h^2\int_0^1 f''(x + sh)(1 - s)\,ds.$$

This is a Taylor expansion with the Laplacian representation of the remainder. We obtain

$$\mathrm{E}_f\hat{f}(x) - f(x) = \int\!\!\int_0^1 K(y)\left[-hyf'(x) - (hy)^2 f''(x - shy)(1 - s)\right]ds\,dy.$$

Since the kernel K has mean zero by asumption, the first term inside the square brackets can be deleted. Using the Cauchy-Schwarz inequality $(\mathrm{E}UV)^2 \le \mathrm{E}U^2\mathrm{E}V^2$ on the variables $U = Y$ and $V = Yf''(x - ShY)$

for Y distributed with density K and S uniformly distributed on $[0,1]$ independent of Y, we see that the square of the bias is bounded above by

$$h^4 \int K(y)y^2\,dy \int\!\!\int_0^1 K(y)y^2 f''(x-shy)^2\,(1-s)^2\,ds\,dy.$$

The integral of this with respect to x is bounded above by

$$h^4 \Big(\int K(y)y^2\,dy\Big)^2 \int f''(x)^2\,dx\,\frac{1}{3}.$$

This concludes the derivation for the bias term.

The last assertion of the theorem is trivial. ∎

The rate $O(n^{-4/5})$ for the mean integrated square error is not impressive if we compare it to the rate that could be achieved if we knew a-priori that f belonged to some parametric family of densities f_θ. Then, likely, we would be able to estimate θ by an estimator such that $\hat\theta = \theta + O_P(n^{-1/2})$, and we would expect

$$\mathrm{MISE}_{f_\theta}(f_{\hat\theta}) = \int \mathrm{E}_\theta\big(f_{\hat\theta}(x) - f_\theta(x)\big)^2\,dx \sim \mathrm{E}_\theta(\hat\theta-\theta)^2 = O(n^{-1}).$$

This is a factor $n^{-1/5}$ smaller than the mean square error of a kernel estimator.

At second thought this loss in efficiency is only a modest price. After all the kernel estimator works for every density that is twice continuously differentiable, while the parametric estimator will presumably fail miserably when the true density does not belong to the postulated parametric model.

Moreover, the lost factor $n^{-1/5}$ can be (almost) regained if we assume that f has sufficiently many derivatives. Suppose that f is m times continuously differentiable. Drop the condition that the kernel K is a probability density, but use a kernel K such that

$$\int K(y)\,dy = 1, \quad \int yK(y)\,dy = 0, \ \ldots\ldots\ , \int y^{m-1}K(y)\,dy = 0,$$

$$\int |y|^m K(y)\,dy < \infty, \quad \int |y|K^2(y)\,dy < \infty.$$

Then, by the same arguments as before, the bias term can be expanded in the form

$$\begin{aligned}\mathrm{E}_f\hat f(x) - f(x) &= \int K(y)\big(f(x-hy)-f(x)\big)\,dy \\ &= \int K(y)\frac{1}{m!}h^m y^m f^{(m)}(x)\,dy + \cdots\end{aligned}$$

Thus the square bias is of the order h^{2m} and the bias-variance trade-off $(nh)^{-1} \sim h^{2m}$ is solved for $h \sim n^{1/(2m+1)}$. This leads to a mean square

error of the order $n^{-2m/(2m+1)}$, which approaches the "parametric rate" n^{-1} as $m \to \infty$. This argument is made precise in the following theorem, whose proof proceeds as before.

5.3 Theorem. *Suppose that f is m times continuously differentiable with $\int \bigl|f^{(m)}(x)\bigr|^2\,dx < \infty$. Then there exists a number C_f such that for small $h > 0$*

$$\mathrm{MISE}_f(\hat{f}) \le C_f\bigl(\frac{1}{nh} + h^{2m}\bigr).$$

Consequently, for $h_n \sim n^{-1/(2m+1)}$, we have $\mathrm{MISE}_f(\hat{f}) = O(n^{-2m/(2m+1)})$.

In principle it might be possible that the slower rate of convergence of a kernel estimator could be improved by using another type of estimator. This is not the case. The rate $n^{-4/5}$ or $n^{-2m/(2m+1)}$ is best possible as soon as we require that the estimator should work for a nonparametric model. We shall make this precise in the following theorem, which we state without proof.

First we note that the constants C_f in the preceding theorems are uniformly bounded in f such that $\int \bigl|f^{(m)}(x)\bigr|^2\,dx$ is uniformly bounded. We can see this from inspecting the proof more closely. Thus, letting $\mathcal{F}_{m,M}$ be the class of all probability densities such that this quantity is bounded by M, there exists a constant $C_{m,M}$, depending on m and M only, such that the kernel estimator with bandwidth $h_n = n^{-1/(2m+1)}$ satisfies

$$\sup_{f \in \mathcal{F}_{m,M}} \mathrm{E}_f \int \bigl(\hat{f}_n(x) - f(x)\bigr)^2\,dx \le C_{m,M}\Bigl(\frac{1}{n}\Bigr)^{2m/(2m+1)}.$$

The following theorem shows that this upper bound is sharp, and the kernel estimator optimal, in that the maximum risk on the left side is bounded below by a similar expression for *every* density estimator $\hat{f}_n$, for every fixed m and M. The proof is omitted.

5.4 Theorem. *There exists a constant $D_{m,M}$ such that for any density estimator $\hat{f}_n$*

$$\sup_{f \in \mathcal{F}_{m,M}} \mathrm{E}_f \int \bigl(\hat{f}_n(x) - f(x)\bigr)^2\,dx \ge D_{m,M}\Bigl(\frac{1}{n}\Bigr)^{2m/(2m+1)}.$$

Problems

1. Show that the distribution of the Kolmogorov-Smirnov statistic is the same for every continuous distribution function F.

2. Let $X_1, \ldots, X_n$ be a random sample from some arbitrary distribution P. What is the natural nonparametric estimate for $P(B)$ for some fixed set B?

3. Suppose that $X_1, \ldots, X_n$ is a random sample from the normal distribution with variance 1. Compare the asymptotic variance of the estimators $\mathbb{F}_n(t)$ and $\Phi(t - \bar{X}_n)$ of $\mathrm{P}(X_1 \leq t)$.

4. Suppose that $X_n \rightsquigarrow X$ for a limit X with a continuous distribution function. Show that $\sup_x \bigl|\mathrm{P}(X_n \leq x) - \mathrm{P}(X \leq x)\bigr| \to 0$.

5. Show, informally, that under sufficient regularity conditions, the mean integrated square error of a kernel estimator $\hat{f}_n$ with bandwidth h satisfies

$$\mathrm{MISE}_f(\hat{f}) \sim \frac{1}{nh} \int K^2(y)\, dy + \tfrac{1}{4} h^4 \int f''(x)^2\, dx \left(\int y^2\, K(y)\, dy \right)^2.$$

What does this imply for an optimal choice of the bandwidth?

6. Let $X_1, \ldots, X_n$ be a random sample from the normal distribution with variance 1. Calculate the mean square error of the estimator $\phi(x - \bar{X}_n)$ of the common density.

6
Appendix:
Some Probability Theory

A *probability space* $(\Omega, \mathcal{U}, \mathrm{P})$ consists of an arbitrary set Ω, a σ-field $\mathcal{U}$ and a probability measure P. A *σ-field* is a collection of subsets (called 'events') that is closed under taking countable unions and taking complements, and contains the empty set. A *probability measure* is a map $\mathrm{P}\colon \mathcal{U} \to [0,1]$ with the properties:

(i) $\mathrm{P}(\emptyset) = 0$ and $\mathrm{P}(\Omega) = 1$;

(ii) if $A_1, A_2, \ldots$ is a sequence of pairwise disjoint sets, then $\mathrm{P}(\cup_i A_i) = \sum_i \mathrm{P}(A_i)$.

Two important consequences of the preceding axiomas for probabilities are the monotone convergence theorems for probabilities.

6.1 Lemma.
(i) if $A_1 \subset A_2 \subset \cdots$, then $\mathrm{P}(A_i) \uparrow \mathrm{P}(\cup_i A_i)$;
(ii) if $A_1 \supset A_2 \supset \cdots$, then $\mathrm{P}(A_i) \downarrow \mathrm{P}(\cap_i A_i)$.

In statistics we rarely explicitly mention or construct an underlying probability space, but we formulate results in terms of random variables and random vectors. A *random variable* X is a map $X\colon \Omega \to \mathbb{R}$ such that the set $\{X \le x\}$ (which is shorthand for $\{\omega\colon X(\omega) \le x\}$) is contained in $\mathcal{U}$ for every $x \in \mathbb{R}$. A *random vector* with values in $\mathbb{R}^k$ is a vector $X = (X_1, \ldots, X_k)$ of random variables. The condition that the sets $\{X \le x\}$ are contained in $\mathcal{U}$ is referred as the (Borel-) *measurability* of X. It allows us to talk about the probabilities $\mathrm{P}(X \le x)$.

The *distribution function* of X is the map $x \to \mathrm{P}(X \le x)$. We use this definition for vectors and variables alike, where in the case of a vector the inequality $X \le x$ is understood coordinate-wise: $X \le x$ if and only if $X_i \le x_i$ for every i.

6.2 Lemma. *A function $F\colon \mathbb{R} \to [0,1]$ is a distribution function of some random variable iff both*

(i) F *is nondecreasing and right continuous;*
(ii) $\lim_{x\to-\infty} F(x) = 0$ *and* $\lim_{x\to\infty} F(x) = 1$.

The necessity of (i)–(ii) can easily be derived from the axiomas for probabilities. That every function F that satisfies (i)–(ii) is the distribution function of some random variable is harder to see, but is proved in measure theory.

The properties of a distribution function in higher dimensions are somewhat more complicated.

We call two random vectors X and Y *equal in distribution* if they have the same distribution function. In other words, X and Y are equal in distribution if and only if the probabilities of the events $\{X \in B\}$ and $\{Y \in B\}$ are equal for every set of the form $B = (-\infty, b]$. The special role given here to the "lower corners" $(-\infty, b]$ is odd and not necessary: we automatically have equality for almost all sets B. This is recorded in the following lemma (i)–(ii), together with two other characterizations.

The collection of *Borel sets* in $\mathbb{R}^k$ is defined as the smallest σ-field that contains all sets of the form $(-\infty, b]$. This is a very large collection of sets and $\{X \in B\}$ can be shown to be an event for every random vector X and every Borel set B. It is difficult to construct sets that are not Borel sets.

6.3 Lemma. *The following statements are equivalent for every pair of random vectors* X *and* Y*:*
(i) $\mathrm{P}(X \le b) = \mathrm{P}(Y \le b)$ *for every* $b \in \mathbb{R}^k$*;*
(ii) $\mathrm{P}(X \in B) = \mathrm{P}(Y \in B)$ *for every Borel set* B*;*
(iii) $\mathrm{P}(a^T X \le b) = \mathrm{P}(a^T Y \le b)$ *for every* $a \in \mathbb{R}^k$ *and* $b \in \mathbb{R}$.

In measure theory the *expectation* is defined for every nonnegative random variable, and for every random variable whose absolute value has finite expectation. Usually it is sufficient to know the definition for discrete and continuous variables. If X is discrete, or continuous with density f, then

$$\mathrm{E}X = \sum_x x\mathrm{P}(X = x), \qquad \text{or} \qquad \mathrm{E}X = \int x f(x)\,dx.$$

Some important properties of expectations are given in the following lemma.

6.4 Lemma. *For any random variables* X *and* Y *whose expectations exist,*
(i) $\mathrm{E}(aX + bY) = a\mathrm{E}X + b\mathrm{E}Y$*;*
(ii) if $X \le Y$*, then* $\mathrm{E}X \le \mathrm{E}Y$*;*
(iii) $|\mathrm{E}X| \le \mathrm{E}|X|$.

Proof. Part (i) is a known result. For (ii) we first infer directly from the definition of expectation that $Y - X \ge 0$ implies that $\mathrm{E}(Y - X) \ge 0$. This implies $\mathrm{E}Y \ge \mathrm{E}X$ by (i). For (iii) first note that $X \le |X|$ and $-X \le |X|$. Next apply (ii). ∎

If $X: \Omega \to \mathbb{R}^k$ is a random vector and $g: \mathbb{R}^k \to \mathbb{R}^m$ a given function, then $g(X): \Omega \to \mathbb{R}^m$ is, by the definition given previously, a random vector if $\{g(X) \leq y\} \in \mathcal{U}$ for every y. This is true for most maps g, but not necessarily true for every map g. A map such that $g(X)$ is a random vector whenever X is a random vector, is called *measurable*.

7
Woordenlijst

σ-field	sigma-algebra
almost surely	bijna zeker
asymptotically consistent	asymptotisch consistent
asymptotically efficient	asymptotisch efficiënt
bandwidth	bandbreedte
Borel sets	Borel verzameling
bounded in probability	begrensd in kans
Brownian bridge	Brownse brug
Cauchy likelihood	Cauchy aannemelijkheidsfunctie
central limit theorem	centrale limietstelling
central moments	centrale momenten
chisquare distribution	chikwadraat verdeling
confidence interval	betrouwbaarheidsinterval
continuous mapping	continue afbeelding
converge almost surely	convergeert bijna zeker
converge in distribution	convergeert in verdeling
converge in probability	convergeert in kans
convergence in law	convergeert in verdeling
correlation	correlatie
covariance matrix	covaraintie matrix
defective distribution function	defectieve verdelingsfunctie
differentiable	differentieerbaar
distribution function	verdelingsfunctie
domination condition	dominerings voorwaarde
empirical distribution	empirische verdeling
empirical distribution function	empirische verdelingsfunctie
empirical process	empirisch proces
estimator	schatter
equal in distribution	gelijk verdeeld

estimating equations	schattingsvergelijking
expectation	verwachting
exponential family	exponentiële familie
independence	onafhankelijkheid
identifiable	identificeerbaar
joint convergence	simultane convergentie
joint distribution	simultane verdeling
kernel	kern
Kullback-Leibler divergence	Kullback-Leibler divergentie
kurtosis	kurtosis
Laplace transform	Laplace getransformeerde
law of large numbers	wet van de grote aantallen
level	onbetrouwbaarheidsdrempel
location estimators	locatie schatter
logistic regression	logistische regressie
marginal convergence	marginale convergentie
maximum likelihood estimator	meest aannemelijke schatter
mean integrated square error	verwachte geintegreerde kwadratische fout
measurability	meetbaarheid
median	mediaan
nonparametric model	niet-parametrisch model
parametric model	parametrisch model
probability measure	kansverdeling
probability space	kansruimte
quantile	kwantiel
random variable	stochastische grootheid
random vector	stochastische vector
regression	regressie
relative efficiency	relatieve efficiëntie
robust statistics	robuuste statistiek
sample correlation coefficient	steekproefcorrelatie coefficiënt
sample variance	steekproef variantie
score function	score functie
sign-function	tekenfunctie
skewness	scheefheid
smoothing method	gladstrijk methode
statistic	statistiek
strong law of large numbers	sterke wet van de grote aantallen
tight	beperkt
uniformly tight	uniform beperkt
weak convergence	zwakke convergentie, convergentie in verdeling
weak law of large numbers	zwakke wet van de grote aantallen
weighted linear regression	gewogen lineaire regressie
window	window