ON A GENERAL CONCEPT OF "IN PROBABILITY"[1]

BY JOHN W. PRATT

Harvard University

1. Summary. Chernoff [1] has called attention to a paper of Mann and Wald [5], which provides a general theory and a convenient notation for the derivation of theorems concerning stochastic limits and limit distributions. The present paper attempts to clarify the first of these topics, stochastic limits, by applying one form of the definition of convergence in probability to any event rather than just to convergence. As in the convergence case, most of the reasoning one is intuitively disposed to do in this connection is valid. Its justification is made more transparent but no more difficult by broadening the applicability of the definition. Thus broadened, "in probability" neither implies nor follows from "with probability one."

2. Introduction. This section leads up to a general concept of "in probability."

Suppose $\{x_n\}$, $\{r_n\}$ are sequences of points of the extended real line $[-\infty, \infty]$. It is customary to write $x_n = o(r_n)$ if $x_n/r_n \to 0$ as $n \to \infty$, and $x_n = O(r_n)$ if x_n/r_n is bounded for large n. (Saying "for large n" allows a finite number of x_n/r_n to be infinite or undefined.) Writing out the definitions fully, we have:

$x_n = o(r_n)$ if, for every positive η, for some N, for every $n > N$, $|x_n/r_n| \leqq \eta$;

$x_n = O(r_n)$ if, for some η and N, for every $n > N$, $|x_n/r_n| \leqq \eta$.

("For some" and "there exist(s) . . . such that" are equivalent.)

Suppose now that $\{X_n\}$ is a sequence of random variables on $[-\infty, \infty]$. It is customary to define o_p and O_p by adding probability requirements to the definitions of o and O as follows.

DEFINITION 1. $X_n = o_p(r_n)$ if, for every positive ϵ and η, for some N, for every $n > N$, $P_n\{|X_n/r_n| \leqq \eta\} \geqq 1 - \epsilon$.

$X_n = O_p(r_n)$ if, for every positive ϵ, for some η and N, for every $n > N$, $P_n\{|X_n/r_n| \leqq \eta\} \geqq 1 - \epsilon$.

$X_n = o_p(1)$ is also written: $X_n \underset{p}{\to} 0$. Note that only the distributions of the individual X_n are referred to. This is emphasized by the use of P_n to denote probabilities. The presence of any joint distribution is irrelevant, and indeed at least one common use of the definition is in connection with asymptotic distributions, where n is related to the sample size and there is not naturally a joint distribution at all.

If we fix η first in the above definition of o_p, we find immediately that it is equivalent to another common definition:

Received September 3, 1957; revised December 28, 1958.

[1] Research carried out at the Statistical Research Center, University of Chicago, and at Harvard University, under sponsorship of the Statistics Branch, Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

DEFINITION 1′. $X_n = o_p(r_n)$ if, for every positive η, $P_n\{|X_n/r_n| \leqq \eta\} \to 1$ as $n \to \infty$.

We cannot "fix η first" in the definition of O_p.

If we fix ϵ first, we are led, less directly, to less familiar variations:

DEFINITION 1″. $X_n = o_p(r_n)$ if, for every positive ϵ, there exist c_n with $P_n\{|X_n| \leqq c_n\} \geqq 1 - \epsilon$ such that $c_n = o(r_n)$.

DEFINITION 1‴. $X_n = o_p(r_n)$ if, for every positive ϵ, there exist S_n with $P_n\{X_n \varepsilon S_n\} \geqq 1 - \epsilon$ such that $x_n \varepsilon S_n$ for all n implies $x_n = o(r_n)$.

Definitions 1″ and 1‴ remain equivalent to Definition 1 if O is substituted for o. For both o and O, this equivalence can be proved directly by letting S_n^ϵ be $[-c_n^\epsilon, c_n^\epsilon]$ and c_n^ϵ be the smallest number such that $P_n\{|X_n| \leqq c_n^\epsilon\} \geqq 1 - \epsilon$, i.e., the upper-tail ϵ-probability point of $|X_n|$. In Section 5, the equivalence of Definition 1‴ to Definition 1 will be proved for o as Corollary 2 and for O as Corollary 3 of Theorem 6. As far as I know, Definitions 1″ and 1‴ were first stated explicitly by Chernoff. It is proved in [5] that the condition of 1‴ implies the condition of 1.

Definitions 1″ and 1‴ are easier to work with than the original definition for many purposes because of the way they separate the stochastic and limiting aspects of the situation. They also seem to me to have at least as much intuitive meaning and reasonableness. Definition 1‴ suggests immediately the generalization introduced in the next section.

3. Definition and fundamental properties of a general concept of "in probability." Suppose, for $n = 1, 2, \cdots$, P_n is the distribution of the random variable X_n in the set $\mathfrak{X}_n$, that is $P_n(X_n \varepsilon S_n) = P_n(S_n)$ is a probability measure on the measurable subsets S_n of $\mathfrak{X}_n$. If S_n is a measurable subset of $\mathfrak{X}_n$, the event $X_n \varepsilon S_n$ will be called an "X_n-event" E_n. If S is any subset of the product space $\mathfrak{X} = \times_{n=1}^{\infty} \mathfrak{X}_n$, the event $(X_1, X_2, \cdots) \varepsilon S$ will be called an "$(X_1, X_2, \cdots)$-event" E. (S need not be measurable, and indeed measurability need not be defined for subsets of $\mathfrak{X}$.)

DEFINITION 2. The $(X_1, X_2, \cdots)$-event E will be said to "occur in probability," written $\mathcal{P}(E)$, if, for every positive ϵ, there exist X_n-events E_n of probability at least $1 - \epsilon$ such that E occurs whenever all E_n occur.

A more formal version of this, in terms of sets, is

DEFINITION 2′. $\mathcal{P}(S)$ if, for every positive ϵ, there is a sequence $\{S_n\}$ such that (1) S_n is a measurable subset of $\mathfrak{X}_n$, (2) $P_n(S_n) \geqq 1 - \epsilon$, and (3) $\times_n S_n \subset S$.

The concept defined here is, as one would hope, independent of the choice of underlying random variables from among those possible. More precisely, suppose $Y_1, Y_2, \cdots$ have given distributions. Then $\mathcal{P}\{(f_1(Y_1), f_2(Y_2), \cdots) \varepsilon T\}$ has the same meaning whether $X_n = Y_n$ or $X_n = f_n(Y_n)$ in Definition 2, that is whether the event $(f_1(Y_1), f_2(Y_2), \cdots) \varepsilon T$ is regarded as a $(Y_1, Y_2, \cdots)$-event or as an $(f_1(Y_1), f_2(Y_2), \cdots)$-event. This amounts to the trivial fact that there are sets S_n with $P_n(Y_n \varepsilon S_n) \geqq 1 - \epsilon$ such that $(f_1(y_1), f_2(y_2), \cdots) \varepsilon T$ whenever $y_n \varepsilon S_n$ for all n if and only if there are sets S'_n with $P_n\{f_n(Y_n) \varepsilon S'_n\} \geqq 1 - \epsilon$ such that $(f_1(y_1), f_2(y_2), \cdots) \varepsilon T$ whenever $f_n(y_n) \varepsilon S'_n$ for all n.

We now investigate the behavior of $\mathcal{P}$ in connection with some elementary relations of events such as $E \Rightarrow F$ (which means the event F occurs whenever the event E occurs.) The theorems are also stated in terms of sets. Thus, for instance, $E \Rightarrow F$ is equivalent to $S \subset T$ if S and T are sets corresponding to events E and F. Proofs given in terms of events can, of course, be directly translated into (perhaps more formal) proofs in terms of sets.

THEOREM 1. *If* $E \Rightarrow F$, *then* $\mathcal{P}(E) \Rightarrow \mathcal{P}(F)$.

THEOREM 1′. *If* $S \subset T$, *then* $\mathcal{P}(S) \Rightarrow \mathcal{P}(T)$.

This is an immediate consequence of the definition of $\mathcal{P}$.

THEOREM 2. $\mathcal{P}$(*for all* α, E^α) $\Leftrightarrow$ *for all* α, $\mathcal{P}(E^\alpha)$, *provided the range of* α *is countable.*

THEOREM 2′. $\mathcal{P}(\bigcap_\alpha S^\alpha) \Leftrightarrow$ *for all* α, $\mathcal{P}(S^\alpha)$, *provided the range of* α *is countable.*

PROOF. $\Rightarrow$ is an immediate consequence of Theorem 1, whatever the range of α.

To prove $\Leftarrow$, suppose $\alpha = 1, 2, \cdots$, and $\mathcal{P}(E^\alpha)$ for all α, and $\epsilon > 0$. For each α, by Definition 2, there are X_n-events E_n^α of probability at least $1 - 2^{-\alpha}\epsilon$ such that E_n^α for all n implies E^α. Let $E_n \Leftrightarrow$ for all α, E_n^α. Then E_n for all n implies for all α, E^α. Furthermore,

$$P_n\{E_n\} \geqq 1 - \sum_{\alpha=1}^{\infty} P_n\{\text{not } E_n^\alpha\} \geqq 1 - \sum_{\alpha=1}^{\infty} 2^{-\alpha}\epsilon = 1 - \epsilon.$$

ϵ was arbitrary. Therefore, by Definition 2, $\mathcal{P}$(for all α, E^α).

THEOREM 3. $\mathcal{P}$ (*for some* α, E^α) *if, for some* α, $\mathcal{P}(E^\alpha)$.

THEOREM 3′. $\mathcal{P}(\bigcup_\alpha S^\alpha)$ *if, for some* α, $\mathcal{P}(S^\alpha)$.

Here the range of α is arbitrary. The proof is trivial. (Technically, Theorem 1 is involved.)

The converse of Theorem 3 is false. If it were true, since we have $\mathcal{P}(E$ or not $E)$ for every E, we would have either $\mathcal{P}(E)$ or $\mathcal{P}$(not E) for every E, which is clearly absurd. For instance, let $P_1(X_1 = 0) = \frac{1}{2}$ and let E be the event $X_1 = 0$.

$\mathcal{P}$(not E) $\Rightarrow$ not $\mathcal{P}(E)$, for otherwise $\mathcal{P}(E$ and not $E)$ by Theorem 2. The converse, not $\mathcal{P}(E) \Rightarrow \mathcal{P}$(not E), is false, for otherwise we would again have either $\mathcal{P}(E)$ or $\mathcal{P}$(not E) for every E.

The following theorem covers Theorems 1–3 and summarizes what can be obtained from them.

THEOREM 4. *If* $\phi(E^\alpha) \Rightarrow F$, *then* $\phi(\mathcal{P}(E^\alpha)) \Rightarrow \mathcal{P}(F)$, *for any logical combination or formula* ϕ, *provided only that* ϕ *involves at most a countable number of and's and no not's. More precisely, we suppose that* ϕ *consists of a finite number of phrases "for all" and "for some," each with an index set, which must be countable in the case of "for all."*

For instance, we might have $\phi(E^\alpha)$ = "for some $A \varepsilon \mathcal{A}$, for all $\alpha \varepsilon A$, E^α," where each $A \varepsilon \mathcal{A}$ is countable, though $\mathcal{A}$ need not be.

THEOREM 4′. *If* $\phi'(S^\alpha) \subset T$, *then* $\phi(\mathcal{P}(S^\alpha)) \Rightarrow \mathcal{P}(T)$, *where* ϕ' *is a finite* *series of the set operations of intersection (with countable index) and union (with arbitrary index), and ϕ is obtained from ϕ' by replacing intersection by "for all" and union by "for some."*

PROOF. $\phi(\mathcal{P}(E^\alpha)) \Rightarrow \mathcal{P}(\phi(E^\alpha))$ by successive application of Theorems 2 and 3. $\mathcal{P}(\phi(E^\alpha)) \Rightarrow \mathcal{P}(F)$ by Theorem 1.

These theorems may seem like poorly disguised trivialities, as indeed they are. My main purpose has been to remove the disguise from some useful trivialities. For example,

THEOREM 5. *Suppose that*

$$f_n^{(j)}(X_n) = O_p(r_n^{(j)}), \qquad j = 1, \cdots, J,$$
$$g_n^{(k)}(X_n) = o_p(s_n^{(k)}), \qquad k = 1, \cdots, K,$$

and that $h_n(x_n) = O(t_n)$ whenever

$$f_n^{(j)}(x_n) = O(r_n^{(j)}), \qquad j = 1, \cdots, J$$
$$g_n^{(k)}(x_n) = o(s_n^{(k)}), \qquad k = 1, \cdots, K$$

Then it follows that $h_n(X_n) = O_p(t_n)$. Furthermore, if $O(t_n)$ is replaced by $o(t_n)$ in the hypothesis, the conclusion is $h_n(X_n) = o_p(t_n)$.

This looks formidable, as does Corollary 1 of [5], of which it is a paraphrase. However, Chernoff reports in [1] that he has found Theorem 5 very useful, and it seems to be about the least general theorem which covers the cases that occur in practice. Both Theorem 5 above and Corollary 1 of [5] are weaker in several respects than Theorem 4, and are covered directly by it and proved in the same way once the equivalence of Definitions 1 and 1‴ is established. To realize that Definition 1 will be tractable when put in the form 1‴ is the nontrivial part of the reasoning.

4. Examples. It will be proved (Corollaries 2 and 3 of Theorem 6 below), and a direct proof has already been indicated in the next to last paragraph of Section 2, that $f_n(X_n) = O_p(r_n)$ if and only if $\mathcal{P}(f_n(X_n) = O(r_n))$ and the same for o_p and o, that is, Definition 2 is actually a generalization of Definition 1. This fact permits the theorems of the last section to be used to carry out a certain common type of argument. Chernoff [1] gives some examples, and others follow here, as well as a case where the argument is tempting but cannot be used.

4.1. If $Y_n - Y_n' \underset{p}{\to} 0$, $Z_n - Z_n' \underset{p}{\to} 0$, $Y_n' = O_p(1)$, $Z_n' = O_p(1)$, and $f(y, z)$ is (jointly) continuous, then $f(Y_n, Z_n) - f(Y_n', Z_n') \underset{p}{\to} 0$.

To prove this, apply Theorem 4 with $X_n = (Y_n, Y_n', Z_n, Z_n')$; E^1 the event $Y_n - Y_n' \to 0$; E^2 the event $Z_n - Z_n' \to 0$; E^3 the event $\{Y_n'\}$ bounded; E^4 the event $\{Z_n'\}$ bounded; F the event $f(Y_n, Z_n) - f(Y_n', Z_n') \to 0$; and $\phi(E^1, E^2, E^3, E^4)$ the simultaneous event E^1 and E^2 and E^3 and E^4. The fact is used that a continuous function is uniformly continuous on bounded sets.

The treatment of special cases of this example in the literature indicates the value of the approach codified here. For instance, Cramér ([2], p. 255), attributing the result to Slutsky [6], uses the relation between convergence in probability and convergence in distribution to prove that if $Y_n \xrightarrow[p]{} y, \cdots, Z_n \xrightarrow[p]{} z$ where $y, \cdots, z$ are constants, and if f is a power of a rational function, then $f(Y_n, \cdots, Z_n) \xrightarrow[p]{} f(y, \cdots, z)$. (A finite number of arguments $y, \cdots, z$ is no more difficult than two, of course.) Halmos ([3], p. 94, Problem 1) outlines an ingenious proof that $Y_n Z_n - Y'_n Z'_n \xrightarrow[p]{} 0$ if $Y_n - Y'_n \xrightarrow[p]{} 0$, $Z_n - Z'_n \xrightarrow[p]{} 0$, $Y'_n = Y$, and $Z'_n = Z$. (An easy extension to the case $Y'_n = O_p(1)$ and $Z'_n = O_p(1)$ and a slight further argument would then prove the first sentence of this subsection (4.1) for f a power of a rational function.)

4.2. It is easy to see that $y_n/n \to 0 \Rightarrow \max(y_1, \cdots, y_n)/n \to 0$. A hasty application of Theorem 4 (or Theorem 1) would then lead to the conclusion that $Y_n/n \xrightarrow[p]{} 0 \Rightarrow \max(Y_1, \cdots, Y_n)/n \xrightarrow[p]{} 0$.

To see that this conclusion is false, let the Y_n be independently exponentially distributed with means μ_n, that is, for $t > 0$, $P_n\{Y_n > t\} = \exp(-t/\mu_n)$. Then $Y_n/n \xrightarrow[p]{} 0$ if (and only if) $\mu_n/n \to 0$. Let μ_1 be arbitrary and $\mu_n = n/\log n$, $n > 1$. Then $Y_n/n \xrightarrow[p]{} 0$. However, $P_n\{\max(Y_1, \cdots, Y_{2n}) \leqq 2n\epsilon\} \leqq \prod_n^{2n} P_n(Y_j \leqq 2n\epsilon) \leqq [1 - (2n)^{-2\epsilon}]^n \to 0$ for $\epsilon < \frac{1}{2}$, so $\max(Y_1, \cdots, Y_n)/n$ does not $\xrightarrow[p]{} 0$.

The difficulty is that $\max(Y_1, \cdots, Y_n)/n \xrightarrow[p]{} 0$ is not equivalent to $\mathcal{P}(\max(Y_1, \cdots, Y_n)/n \to 0)$ for any X_n for which $Y_n/n \xrightarrow[p]{} 0$ is equivalent to $\mathcal{P}(Y_n/n \to 0)$ so that Theorem 4 cannot be applied. If $X_n = Y_n$, then $\max(Y_1, \cdots, Y_n)$ is not a function of X_n alone, so Corollary 2 of Theorem 6 cannot be applied to show $\max(Y_1, \cdots, Y_n)/n \xrightarrow[p]{} 0$ equivalent to $\mathcal{P}(\max(Y_1, \cdots, Y_n)/n \to 0)$. If $X_n = \max(Y_1, \cdots, Y_n)$, then $Y_n/n \xrightarrow[p]{} 0$ is similarly not equivalent to $\mathcal{P}(Y_n/n \to 0)$. If $X_n = (Y_1, \cdots, Y_n)$, $\max(Y_1, \cdots, Y_n)/n \to 0$ is not an event on the whole of the product space of the ranges of the X_n, but only on the subspace with points $(y_1; y_1, y_2; y_1, y_2, y_3; \cdots)$.

4.3. Suppose $Y_n - Y \xrightarrow[p]{} 0$, $Z_n - Z \xrightarrow[p]{} 0$, and $f(y, z)$ is (jointly) continuous. Then $f(Y_n, Z_n) - f(Y, Z) \xrightarrow[p]{} 0$.

As in 4.2, we cannot apply Theorem 4 directly. However, letting (Y_n, Y'_n, Z_n, Z'_n) have the distribution of (Y_n, Y, Z_n, Z) for each n, the hypothesis is equivalent to $Y_n - Y'_n \xrightarrow[p]{} 0$, $Z_n - Z'_n \xrightarrow[p]{} 0$ and the conclusion to $f(Y_n, Z_n) - f(Y'_n, Z'_n) \xrightarrow[p]{} 0$. Thus the desired result follows from 4.1.

Having to restate $Y_n - Y \xrightarrow[p]{} 0$ as $Y_n - Y'_n \xrightarrow[p]{} 0$, etc., is no loss mathematically, since the statements are equivalent. Indeed this very equivalence gives some insight into the problem. It shows that we will have to make use of the uniform continuity of f on bounded intervals. Further, it emphasizes that when $Y_n - Y \xrightarrow[p]{} 0$, the fact that the value of Y is the same for each n contributes nothing essential. Of course, the fact that the distribution of Y is the same for each n may contribute; indeed, it does, since it gives $Y'_n = O_p(1)$.

4.4. Suppose $Y_n - Y \xrightarrow[p]{} 0$. Suppose also F_n is a random function and, for every M, $\sup_{|y| \leqq M} | F_n(y) - f(y) | \xrightarrow[p]{} 0$. Then $F_n(Y_n) - f(Y) \xrightarrow[p]{} 0$.

The usefulness of results of this kind is that the limiting distributions of Y_n and Y'_n are the same if $Y_n - Y'_n \xrightarrow[p]{} 0$. Thus, for example, in 4.4 the problem of finding the limiting distribution of $F_n(Y_n)$ has been reduced to finding that of $f(Y)$ which is ordinarily much easier.

4.5. Suppose X_n is the proportion of successes in n independent binomial trials with probability ρ of success. $X_n - \rho = O_p(1/\sqrt{n})$. If f is once differentiable at ρ, $f(X_n) = f(\rho) + f'(\rho)(X_n - \rho) + o_p(1/\sqrt{n})$, and

$$\sqrt{n}[f(X_n) - f(\rho)] = f'(\rho)\sqrt{n}(X_n - \rho) + o_p(1).$$

Now the asymptotic distribution of $\sqrt{n}(X_n - \rho)$ has variance $\rho(1 - \rho)$, so the asymptotic distribution of $\sqrt{n}[f(X_n) - f(\rho)]$ will have variance independent of ρ if and only if $f'(\rho)$ is a multiple of $1/(\rho(1 - \rho))^{\frac{1}{2}}$. This gives $f(\rho) =$ arcsin $\sqrt{\rho}$, up to affine transformation. Note that we have not proved anything, except heuristically, about the limit of the variance of $\sqrt{n}[f(X_n) - f(\rho)]$, even though $\sqrt{n}(X_n - \rho)$ has variance $\rho(1 - \rho)$ for every n. This is because $E(o_p(1))$ is not necessarily $o(1)$. On the other hand, what we really want, often, when we ask for asymptotically constant variance is that the variance of the asymptotic distribution shall be constant, since in this case an F statistic based on the transformed variates will have an F distribution asymptotically. This justification of the arcsin transformation has no bearing, of course, if the purpose of the transformation is to make some effects additive.

4.6. If X_n-events $E_n(\epsilon)$ are given with $P_n\{E_n(\epsilon)\} \geqq 1 - \epsilon$, then $\mathcal{P}$ {for some ϵ, for all n, $E_n(\epsilon)$}. This remark is trivial, but it is technically involved, along with Corollaries 2 and 3 below, in showing that Theorem 4 of this paper includes Theorem 1 and Corollary 1 of the Mann-Wald paper [5].

5. A theorem showing the equivalence of the definitions. It will be proved from Theorem 6 that $X_n \underset{p}{\rightarrow} 0$ is equivalent to $\mathcal{P}(X_n \rightarrow 0)$ and $X_n = O_p(1)$ to $\mathcal{P}(X_n = O(1))$. These facts are a little more easily proved directly. However, Theorem 6 and Corollary 1 may be of interest in themselves.

THEOREM 6. *Suppose that, for every n and α, E_n^α is an X_n-event and $E_n^\alpha \Rightarrow E_n^\beta$ if $\beta > \alpha$. Then*

$$P_n(E_n^\alpha) \rightarrow 1 \quad \textit{uniformly in} \quad n \quad \textit{as} \quad \alpha \rightarrow \infty \Leftrightarrow$$

$$\inf_n P_n(E_n^\alpha) \rightarrow 1 \quad \textit{as} \quad \alpha \rightarrow \infty \Leftrightarrow$$

$$\mathcal{P}(\textit{for some } \alpha, \textit{for all } n, E_n^\alpha).$$

(We have in mind that the range of α is the positive integers, although the proof applies to more general partially ordered sets.)

PROOF. The first $\Leftrightarrow$ is immediate.

To prove the second $\Rightarrow$, suppose $\inf_n P_n(E_n^\alpha) \rightarrow 1$ as $\alpha \rightarrow \infty$. Let ϵ be an arbitrary positive number. There exists σ such that, for all $\alpha \geqq \sigma$ and all n, $P_n(E_n^\alpha) \geqq 1 - \epsilon$. Let $E_n \Leftrightarrow E_n^\sigma$. Then $P_n(E_n) \geqq 1 - \epsilon$. Furthermore, E_n for all $n \Rightarrow$ for some α (namely σ), for all n, E_n^α. Therefore, by Definition 2, $\mathcal{P}$ (for some α, for all n, E_n^α).

To prove the second $\Leftarrow$, suppose $\mathcal{P}$(for some α, for all n, E_n^α), and let ϵ be an arbitrary positive number. By Definition 2, there exist E_n with $P_n(E_n) \geqq 1 - \epsilon$ such that E_n for all $n \Rightarrow$ for some α, for all n, E_n^α.

Fix m arbitrarily. E_n for all $n \Rightarrow$ for some α, E_m^α. Therefore, the other E_n being irrelevant, $E_m \Rightarrow$ for some α, E_m^α. Since $E_m^\alpha \Rightarrow E_m^\beta$ for $\beta > \alpha$, it follows the event (E_m but not E_m^α) decreases to the null event as $\alpha \to \infty$. Then there is an x_m where E_m occurs but E_m^α occurs only for α so large that $P_m(E_m$ but not $E_m^\alpha) \leqq \epsilon$.[2]

Suppose x_m has been so chosen for every m. At $(x_1, x_2, \cdots)$, E_m occurs for all m, and, consequently, there exists σ such that, for all m, E_m^σ occurs. For $\alpha \geqq \sigma$, E_m^α occurs at x_m, whence $P_m(E_m$ but not $E_m^\alpha) \leqq \epsilon$. Therefore, for all m, for all $\alpha \geqq \sigma$, $P_m(E_m^\alpha) \geqq P_m(E_m) - \epsilon \geqq 1 - 2\epsilon$. But ϵ was arbitrary and σ didn't depend on m. Therefore $P_m(E_m^\alpha) \to 1$ uniformly in m as $\alpha \to \infty$.

COROLLARY 1. $P_n(E_n) \to 1$ as $n \to \infty \Leftrightarrow \mathcal{P}($ for some α, for all $n \geqq \alpha$, $E_n)$.

PROOF. Let $E_n^\alpha = E_n$ for $n \geqq \alpha$, let E_n^α be the universal event $X_n \varepsilon \mathfrak{X}_n$ otherwise, and apply Theorem 6.

COROLLARY 2. $f_n(X_n) = o_p(r_n) \Leftrightarrow \mathcal{P}(f_n(X_n) = o(r_n))$.

PROOF. Let $Y_n = |f_n(X_n)/r_n|$. $f_n(X_n) = o_p(r_n) \Leftrightarrow$ for every positive η, $P_n\{Y_n \leqq \eta\} \to 1$ as $n \to \infty \Leftrightarrow$ for every α, $P_n\{Y_n \leqq 1/\alpha\} \to 1$ as $n \to \infty \Leftrightarrow$ for every α, $\mathcal{P}(\lim\sup\ Y_n \leqq 1/\alpha) \Leftrightarrow \mathcal{P}($for every α, $\lim\sup Y_n \leqq 1/\alpha) \Leftrightarrow \mathcal{P}(f_n(X_n) = o(r_n))$. (The range of α is the positive integers.) The first $\Leftrightarrow$ is Definition 2, the second is immediate, the third follows from Corollary 1, the fourth follows from Theorem 2, and the last is virtually definition.

COROLLARY 3. $f_n(X_n) = O_p(r_n) \Leftrightarrow \mathcal{P}(f_n(X_n) = O(r_n))$.

PROOF. Let $Y_n = |f_n(X_n)/r_n|$. $f_n(X_n) = O_p(r_n) \Leftrightarrow$ for every positive ϵ, for some η and N, for every $n > N$, $P_n\{Y_n \leqq \eta\} \geqq 1 - \epsilon \Leftrightarrow$

$$\inf_{n>N} P_n\{Y_n \leqq N\} \to 1 \quad \text{as} \quad N \to \infty \Leftrightarrow$$

$\mathcal{P}($for some N, for all $n > N$, $Y_n \leqq N) \Leftrightarrow \mathcal{P}(f_n(X_n) = 0(r_n))$. The first $\Leftrightarrow$ is Definition 1, the second is immediate, the third follows from Theorem 6, and the last is virtually definition.

6. Relation to Probability One. One might ask what relation there is between $\mathcal{P}(E)$ and $\Pr(E) = 1$. This question is not entirely natural, in that $\Pr(E) = 1$ refers to the joint distribution of the X_n, while $\mathcal{P}(E)$ refers only to their individual (marginal) distributions. The joint distribution may sometimes be changed so that $\Pr(E)$ is changed without changing the marginals. For example, suppose $X_1, X_2, \cdots$ have the standard normal distribution. If they are independent, $\Pr(X_n$ diverges as $n \to \infty) = 1$. If $X_1 = X_2 = \cdots$, $\Pr(X_n$ diverges as $n \to \infty) = 0$. In any case, $\mathcal{P}(X_n$ diverges as $n \to \infty)$. To prove the latter, let E_n in Definition 2 be the event X_n not between a_n and a_{n+2}, where a_n takes on, in rotation, the $k\epsilon/2$-points of the standard normal distribution.

[2] The range of α must be a directed system for "$\alpha \to \infty$" to have meaning. Further restriction of the range of α, and the hypothesis that for every n, $E_n^\alpha \Rightarrow E_n^\beta$ if $\beta > \alpha$, are required only to establish this statement, and hence only to prove the second $\Leftarrow$, which is the least trivial part of the theorem.

As another example of what can happen, let X_n be normal with mean 0 and variance ϵ_n and let $X_1, X_2, \cdots$ be independent. Then

$$X_n \underset{p}{\rightarrow} 0 \quad \text{if and only if} \quad \epsilon_n \rightarrow 0.$$

$$\Pr(X_n \rightarrow 0) = 1 \quad \text{if, for every positive} \quad \epsilon, \Sigma\Phi(\epsilon/\epsilon_n) < \infty,$$

$$\Pr(X_n \text{ diverges}) = 1 \quad \text{if, for some positive} \quad \epsilon, \Sigma\Phi(\epsilon/\epsilon_n) = \infty,$$

where $\Phi(t)$ is the tail area above t of the standard normal distribution. Letting E be the event X_n diverges, we see that we may have, even for independent $X_1, X_2, \cdots, \Pr(E) = 1$ yet $\mathcal{P}(\text{not } E)$.

There are some events for which this is not the case. For example,

THEOREM 7. *If, for every* α, $\Pr(E^\alpha) = 1 \Rightarrow \mathcal{P}(E^\alpha)$, *then* $\Pr($*for every* $\alpha, E^\alpha) = 1 \Rightarrow \mathcal{P}($*for every* $\alpha, E^\alpha)$, *provided the range of* α *is denumerable.*

This is an immediate consequence of Theorem 2.

THEOREM 8. *Suppose that, for every* n, $E_n^\alpha \Rightarrow E_n^\beta$ *if* $\beta > \alpha$. *Then* $\Pr($*for some* α, *for every* $n, E_n^\alpha) = 1 \Rightarrow \mathcal{P}($*for some* α, *for every* $n, E_n^\alpha)$.

(The same range of α is possible here as in Theorem 6.)

PROOF. The hypotheses imply that $\Pr(\text{for every } n, E_n^\alpha) \rightarrow 1$ as $\alpha \rightarrow \infty$. It follows that $\inf_n P_n(E_n^\alpha) \rightarrow 1$ as $\alpha \rightarrow \infty$. Theorem 6 completes the proof.

These two theorems may be applied successively to prove $\Pr(E) = 1 \Rightarrow \mathcal{P}(E)$ for many events E. For instance, they cover the facts that $\Pr\{X_n = o(r_n)\} = 1 \Rightarrow \mathcal{P}(X_n = o(r_n))$ and that $\Pr\{X_n = O(r_n)\} = 1 \Rightarrow \mathcal{P}(X_n = O(r_n))$.

Theorems 7 and 8 do not restrict the joint distribution of the X_n. Thus the second sentence of Theorem 8 could be changed to: If the X_n have given distributions and these distributions are the marginals of some joint distribution of the X_n under which $\Pr(\text{for some } \alpha, \text{ for all } n, E_n^\alpha) = 1$, then $\mathcal{P}(\text{for some } \alpha, \text{ for all } n, E_n^\alpha)$. A similar rewording of Theorem 7 is possible.

7. Further generalizations. Definition 2 and Theorems 1–4 generalize immediately to the case that n ranges over some entirely arbitrary set.

A (perhaps more interesting) generalization is to do away with product spaces. This permits the consideration of events not defined on the product space, both those depending (say) on a further variable as well, and those not defined on the whole space, such as $\max(Y_1, \cdots, Y_n) \rightarrow 0$ with $X_n = (Y_1, \cdots, Y_n)$. Suppose, then, for each n, P_n is a probability measure on a Borel field $\mathcal{B}_n$ of subsets of a set $\underline{X}$.

DEFINITION 3. $\mathcal{P}(S)$ for a subset S of $\underline{X}$ if, for every $\epsilon > 0$, there exist $S_n \in \mathcal{B}_n$ such that $P_n(S_n) \geqq 1 - \epsilon$ for all n and $\bigcap S_n \subset S$.

Theorems 1′–4′ continue to hold. However, some curious things can nevertheless happen. For instance, $\mathcal{P}(S)$ for every S in the following case, essentially the case that P_n is the joint distribution of (X_1, X_n).

THEOREM 9. *Let* $\mathcal{A}_n$ *be a Borel field of subsets of* $\underline{X}_n$, $n = 1, 2, \cdots$; *let* $\underline{X} = \times_{n=1}^\infty \underline{X}_n$; *let* $\mathcal{B}_n$ *be the smallest Borel field including all subsets of* $\underline{X}$ *of the form* $\{x: x_1 \in A_1, x_n \in A_n\}$, $A_1 \in \mathcal{A}_1$, $A_n \in \mathcal{A}_n$; *let* Q *be a probability measure on*

$\mathfrak{A}_1$; *let* P_n *be a probability measure on* $\mathfrak{B}_n$ *such that* $P_n\{x:x_1 \varepsilon A_1\} = Q(A_1)$ *for* $A_1 \varepsilon \mathfrak{A}_1$. *Then* $\mathcal{P}(S)$ *for every* S *provided only* Q *is non-atomic.*

Q is essentially the distribution of X_1 .

PROOF. Choose T_n such that $Q(T_n) \leqq \epsilon$ and $\bigcup T_n = \mathfrak{X}_1$. Let $x \varepsilon S_n$ if and only if $x_1 \not\varepsilon T_n$. Then $P_n(S_n) = 1 - P_n\{x:x_1 \varepsilon T_n\} \geqq 1 - \epsilon$ for all n yet $\bigcap S_n$ is empty. Hence, by Definition 3, $\mathcal{P}(S)$ for every S.

Thus we can now have both $\mathcal{P}(S)$ and $\mathcal{P}(\mathfrak{X} - S)$, but by Theorem 2′, only if we have $\mathcal{P}$(null set), or equivalently, by Theorem 1′, $\mathcal{P}(S)$ for all S.

Theorem 9 makes it amply clear that its hypothesis does not provide a situation in which $X_n - X_1 \underset{p}{\rightarrow} 0$ is equivalent to $\mathcal{P}(X_n - X_1 \rightarrow 0)$, and the generality of Definition 3 does not seem, as one might hope, to permit cases like that of Example 4.3 to be handled by Theorem 4 without introduction of new random variables like Y_n' .

The following question seems to me natural and interesting, but I have not pursued it. Suppose X_n is a function from $\mathfrak{X}$ to the real line and $\mathfrak{B}_n$ is the Borel field of inverse images under X_n of Borel sets. Suppose P_n is a probability measure on $\mathfrak{B}_n$. What restrictions are required to make $X_n \underset{p}{\rightarrow} 0$ and $\mathcal{P}(X_n \rightarrow 0)$ equivalent? We have seen that it is certainly sufficient that X_n be the nth coordinate of X.

8. Miscellaneous remarks.

8.1. A stochastic process X_t is called continuous in probability if, for every u,

$$X_t - X_u \underset{p}{\rightarrow} 0, \quad \text{as} \quad t \rightarrow u.$$

The statement $\mathcal{P}(X_t$ continuous), which might also be read "X_t continuous in probability," has no meaning as it stands. If t is to take the place of n in Definition 2, then for each t, X_t must be a random function of another variable. If "continuous" means "continuous in t," then a collection of stochastic processes is required, one for each n in Definition 2. In both cases X_t refers to something more complicated than a simple stochastic process, and the statement $\mathcal{P}(X_t$ continuous) is therefore, however interpreted, quite different from the statement that (the stochastic process) X_t is continuous in probability.

8.2. I have not been able to see that the point of view explored in this paper throws any light on certain problems which concern convergence in probability and involve pairwise or joint distributions. For example, a stochastic process which is continuous in probability on a finite closed interval is uniformly continuous in probability thereon, that is for every positive ϵ and η, for some δ, $\Pr\{|X_t - X_u| \leqq \eta\} \geqq 1 - \epsilon$ if $|t - u| \leqq \delta$ (see Lévy, [4], pp. 36–37). Again, if $X_n - X_m \underset{p}{\rightarrow} 0$ as $n, m \rightarrow \infty$, then there is a random variable Y such that $X_n \underset{p}{\rightarrow} Y$ (see Halmos, [3], p. 93). Again, $X_n \underset{p}{\rightarrow} 0$ if and only if every subsequence has a subsequence which $\rightarrow 0$ with probability one. Pairwise distributions are involved in the first two cases. The last is interesting in that a condition which involves a joint distribution is equivalent to one which does not. All three are most easily proved directly from Definition 1.

Acknowledgements. I am indebted to Herman Chernoff who aroused my interest in this field when, as a student at Stanford, at his suggestion, I wrote up lectures in which he discussed and amplified [5].

I am also most grateful to Sudhish Ghurye for very stimulating and helpful discussions and suggestions.

REFERENCES

[1] H. Chernoff, "Large sample theory: Parametric case," *Ann. Math. Stat.*, Vol. 27 (1956), pp. 1–22.

[2] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

[3] P. R. Halmos, *Measure Theory*, D. Van Nostrand, New York, 1950.

[4] P. Lévy, *Processus Stochastiques et Mouvement Brownien*, Gauthier-Villars, Paris, 1948.

[5] H. B. Mann, and A. Wald, "On stochastic limit and order relationships," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 217–226.

[6] E. Slutsky, "Über stochastische Asymptoten und Grenzwerte," *Metron*, Vol. 5 (1925), p. 3.