The Annals of Mathematical Statistics
1970, Vol. 41, No. 2, 427–439

THE ASYMPTOTIC DISTRIBUTION OF THE MEASURE OF RANDOM SETS WITH APPLICATION TO THE CLASSICAL OCCUPANCY PROBLEM AND SUGGESTIONS FOR CURVE FITTING

By Gedalia Ailam

Michigan State University and
The Israeli Institute for Biological Research

1. Introduction. Probability properties of the measure of the union of random sets have theoretical as well as practical importance [4], [5]; e.g., the probability distribution of the total area hit by bombs or shells is of much interest in gunnery [6]. The area of a slide covered by particles falling on it has a considerable importance for particle sampling. Similarly, some properties of distribution functions [2] as well as the classical occupancy problem of Statistical Mechanics may be formulated as properties of measures of the union of random sets. In the present paper we derive an asymptotic distribution for these measures and apply the results to the classical occupancy problem as a special case.

Large sample nonparametric tests for the multivariate case of curve fitting are also suggested as application of the asymptotic distribution.

2. Preliminaries. As the present paper is inherently based on a previous one [1] it may be of advantage to summarize the former's necessary essentials. In [1] it has been shown that measures of the union of random sets may be treated with the aid of "Coverage Spaces" which were defined as triplets $(\mathscr{A}, \mathscr{B}, M)$ formed by two probability spaces $\mathscr{A} = (X, S, P)$ and $\mathscr{B} = (Y, T, Q)$ and a measurable set M in their product probability space. An experiment of size n was defined as a point $\mathbf{y}^n = (y_1, \cdots, y_n)$ in $\mathscr{B}^n$. The covered part of X due to the experiment $\mathbf{y}^n$ was $U_1{}^n M(y_i)$ where $M(y_i)$ denotes the section of M determined by y_i. The "vacancy", $p(\mathbf{y}^n)$ (the measure of that part of X which is not covered due to the experiment $\mathbf{y}^n$) was, thus, equal to $P\{\bigcap_1{}^n M'(y_i)\}$ where M' denotes the complement of M. The coverage space $(\mathscr{B}, \mathscr{A}, W)$ where W denotes the image of M under the natural transformation of $\mathscr{A} \times \mathscr{B}$ on $\mathscr{B} \times \mathscr{A}$ was called "The Conjugate Coverage Space to $(\mathscr{A}, \mathscr{B}, M)$".

It was proved that the kth moment of vacancy for an experiment of size n in any coverage space is equal to the nth moment of vacancy for an experiment of size k in the conjugate coverage space. This property of duality will be used in the sequel.

Concerning the choice of a coverage space model for a specific coverage problem, it is obvious that many coverage spaces may represent the same coverage problem, and even when $\mathscr{A}$ is already fixed there still remains a vast class of probability spaces for possible $\mathscr{B}$ spaces. And generally even when $\mathscr{B}$ is also fixed, there still remain many possibilities for M. It may also readily be observed that a class of coverage problems having the space $\mathscr{A}$ in common may always be represented by

Received November 10, 1967; revised July 14, 1969.

a class of coverage spaces having in common both $\mathscr{A}$ and $\mathscr{B}$, so that the coverage spaces of the class may differ by M only.

3. Theorem.

Let $\mathscr{B} = (Y, T, Q)$ be a probability space such that (Y, T) is the Euclidean N-space with Lebesgue σ-field.

Let $\mathscr{A} = (K, S, P)$ be a probability space such that K is a subset of Y of finite positive Lebesgue measure, S the sub σ-field of T relative to K and P the Lebesgue measure on Y normalized to K.

Denote by $q(x)$ the density of the absolute continuous component of Q relative to P and let Q satisfy the following regularity condition on K:

(1) *There exists an $R \in S$ with P(boundary of R) $= 0$, $q(x) = 0$ a.e. on $K-R$, for which either $P(R) = 0$, or there exists a sequence of partitions of R,*

$$\{\Pi_m\} = \{(\pi_{1m}, \pi_{2m}, \cdots, \pi_{mm})\}, \qquad m = 1, 2, \cdots$$

such that $P(\pi_{im}) > 0$, P(boundary of π_{im}) $= 0$ for every i and m and such that

$$Q^{\frac{1}{2}}(\pi_{im})/P(\pi_{im}) \to \infty \quad \text{uniformly in } i \text{ when } \quad m \to \infty. \tag{1.1}$$

Let the sets M_n satisfy the following conditions:

$$M_n \subset \{(x, y) \,|\, |x-y| \leqq c/n^{1/N}\}, \qquad n = 1, 2, \cdots, \qquad \text{for some constant } c. \tag{2}$$

$$nQ\{M_n(x)\} \to_{\text{a.e.}} a(x) \qquad x \in R \tag{3}$$

$$\int_K n[\exp(nQ\{M_n(x_1) \cap M_n(x_2)\}) - 1]\, dP(x_2) = b_n(x_1) < b \quad \text{and} \quad b_n(x_1) \to_{\text{a.e.}} b(x_1). \tag{4}$$

Then, letting an experiment of size n correspond to the nth coverage space and denoting

$$Z_n(\mathbf{y}_n) = n^{\frac{1}{2}}[p(\mathbf{y}_n) - Ep(\mathbf{y}_n)], \tag{5}$$

the distribution of Z_n tends to the normal distribution with mean zero and variance σ^2 where

$$\sigma^2 = \int_K e^{-2a(x)}\, b(x)\, dP(x) - [\int_K a(x)\, e^{-a(x)}\, dP(x)]^2. \tag{6}$$

REMARK. It is worthwhile to mention that when $\{M_n(x)\}$ is a regular sequence of closed sets containing the point x ([9] page 106) and $nP_n\{M_n(x)\} \to t(x)$ then $a(x) = t(x)f(x)$ a.e. where $f(x)$ is the density of the absolute continuous component of Q relative to P ([2] page 198).

PROOF. In order to carry out the proof, estimations of the variance and the fourth central moment of vacancy as well as an estimation of the limit of a sequence of tails of multinomial distributions will be required.

LEMMA (a). *The variance of vacancy. Let $\{(\mathscr{A}, \mathscr{B}, M)\}$ be a coverage space where $\mathscr{A} = (X, S, P)$, $\mathscr{B} = (Y, T, Q)$ and assume*

$$P\{x_2 \in X \,|\, Q\{M(x_1) \cap M(x_2)\} \neq 0\} \leqq h \quad \text{for a.e.} \quad x_1 \in X. \tag{7}$$

Denote,

$$u_i = nQ\{M(x_i)\} \qquad i = 1, 2, \cdots$$
$$u_{ik} = nQ\{M(x_i) \cap M(x_k)\} \qquad i, k = 1, 2, \cdots.$$

Then for an experiment of size n

$$\sigma^2 p(\mathbf{y}_n) = (1/n)\int_{X^2} \exp(-u_1 - u_2)\, n(e^{u_{12}} - 1)\, dP^2 - (1/n)(\int_X u_1\, e^{-u_1}\, dP)^2 + \theta_1/n^2 + \theta_2\, h/n \tag{8}$$

where $|\theta_1| \leqq (3e+1+4/e^2)/e^2$ *and* $|\theta_2| \leqq 2/e^2$.

PROOF OF (a). Let

$$f(a, n) = [e^{-a}(1 - a^2/2n) - (1 - a/n)^n].$$

Then for any $n = 1, 2, \cdots$ and any a, $o \leqq a \leqq n$

$$|f(a, n)| \leqq n^{-2}/e. \tag{9}$$

This may be shown as follows: For $a = o$ and $a = n$ (9) is valid. Now,

$$\partial f/\partial a = e^{-a}(1 - a^2/2n + a/n) + (1 - a/n)^{n-1}$$
$$\partial f/\partial a = 0 \Rightarrow (1 - a/n)^n = e^{-a}(1 - a^2/2n - a^2/n^2 + a^3/2n^2).$$

Therefore,

$$|f(a, n)| \leqq \max |e^{-a}(1 - a^2/2n) - e^{-a}(1 - a^2/2n - a^2/n^2 + a^3/2n^2)|$$
$$= \max |a^2 - a^3/2|\, e^{-a} n^{-2} \leqq n^{-2}/e.$$

As shown in [2], it may be proved that under the same conditions also,

$$|e^{-a} - (1 - a/n)^n| \leqq 1/en. \tag{10}$$

Now applying the theorem cited in the preliminaries we obtain,

$$\sigma^2 p(\mathbf{y}_n) = E[p(\mathbf{y}_n)^2] - [Ep(\mathbf{y}_n)]^2$$
$$= \int_{X^2} \left[1 - \frac{u_1 + u_2 - u_{12}}{n}\right]^n dP^2 - \int_{X^2} \left(1 - \frac{u_1}{n}\right)^n \left(1 - \frac{u_2}{n}\right)^n dP^2$$
$$= \int_{X^2} \exp[-(u_1 + u_2 - u_{12})] \left[1 - \frac{(u_1 + u_2 - u_{12})^2}{2n}\right] dP^2 - \int_{X^2} \exp(-u_1 - u_2) \left(1 - \frac{u_1^2}{2n}\right)\left(1 - \frac{u_2^2}{2n}\right) dP^2 + \theta_1'/n^2$$

where $|\theta_1'| \leqq 3/e + 1/e^2$.

The latter expression for $\sigma^2 p(\mathbf{y}_n)$ may also be written as,

$$\sigma^2 p(\mathbf{y}_n) = \int_{X^2} \exp(-u_1 - u_2)(e^{u_{12}} - 1)\, dP^2 - [(\int_X u_1\, e^{-u_1}\, dP)^2]/n + (\int_X u_1^2\, e^{-u_1}\, dP)^2/(4n^2) + (1/2n)\int_{X^2} \{(u_1 + u_2)^2 \exp[-(u_1 + u_2)] - (u_1 + u_2 - u_{12})^2 \exp[-(u_1 + u_2 - u_{12})]\}\, dP^2 + \theta_1'/n^2.$$

The last two integrands are bounded by $4/e^2$ and the last one vanishes for a.e. x_1 outside the set defined in (7) the measure of which is bounded by h; therefore the integral is bounded by $4h/e^2$. Hence,

$$\sigma_n^2 = (1/n)\{\int_{X^2} \exp[-(u_1+u_2)]\, n(e^{u_{12}}-1)\, dP^2 - (\int_X u_1\, e^{-u_1}\, dP)^2\} + \theta_1/n^2 + \theta_2\, h/n$$

where $|\theta_1| \leqq (3e+1+4/e^2)/e^2$ and $|\theta_2| \leqq 2/e^2$.

LEMMA (b). *The fourth central moment of vacancy. Let* $\{(\mathscr{A}, \mathscr{B}, M)\}$ *be a coverage space where* $\mathscr{A} = (X, S, P)$, $\mathscr{B} = (Y, T, Q)$, *and for which* (7) *holds, and let correspond to the coverage space an experiment of size n, then*

$$E[p(\mathbf{y}_n) - Ep(\mathbf{y}_n)]^4 \leqq a_1 h^2 + a_2 h/n + a_3/n^2 \tag{11}$$

where a_1, a_2 *and* a_3 *are absolute constants.*

PROOF. Applying (9), the previous notations, and denoting

$$A = \{(x_1, x_2) \in X^2 \mid u_{12} \neq 0\},$$

$$A' = \{(x_1, x_2, x_3) \in X^3 \mid \text{at least two out of } u_{12}, u_{13}, u_{23} \text{ are different from zero}\},$$

$$A'' = \{(x_1, x_2, x_3, x_4) \in X^4 \mid \text{at least two out of } u_{12}, u_{13}, u_{14}, u_{23}, u_{24}, u_{34} \text{ are different from zero}\},$$

we obtain as special cases of the theorem cited in the introduction,

$$Ep(\mathbf{y}_n) = \int_X (1-u_1/n)^n\, dP = \int_X e^{-u_1}(1-(2n)^{-1}u_1^2)\, dP + \theta_1/n^2 \qquad |\theta_1| \leqq 1, \tag{12}$$

$$\begin{aligned} E[p(\mathbf{y}_n)]^2 &= \int_{X^2} [1-(u_1+u_2-u_{12})/n]^n\, dP^2 \\ &= \int_{X^2} \exp(-u_1-u_2)[1-(u_1+u_2)^2/2n]\, dP^2 \\ &\quad + \int_A [\exp(-u_1-u_2+u_{12}) - \exp(-u_1-u_2)]\, dP^2 \\ &\quad + \theta_2\, h/n + \theta_3/n^2, \qquad |\theta_2| \leqq 1, |\theta_3| \leqq 1. \end{aligned} \tag{13}$$

$$\begin{aligned} E[p(\mathbf{y}_n)]^3 &= \int_{X^3} [1-Q\{M(x_1) \cup M(x_2) \cup M(x_3)\}]^n\, dP^3 \\ &= \int_{X^3} [1-(u_1+u_2+u_3)/n]^n\, dP^3 + 3\int_{A\times X} \{[1-(u_1+u_2-u_{12}+u_3)/n]^n \\ &\qquad - [1-(u_1+u_2+u_3)/n]^n\}\, dP^3 + d_1 \int_{A'} R(x_1, x_2, x_3)\, dP^3. \end{aligned}$$

Where d_1 is a combinatorial constant, $|R(x_1, x_2, x_3)| \leqq 1$ and $A \times X$ stands for the Cartesian product of A and X. But according to the definition of A', $P^3 A' \leqq 3h^2$, thus

$$\begin{aligned} E[p(\mathbf{y}_n)]^3 &= \int_{X^3} \exp(-u_1-u_2-u_3)[1-(u_1+u_2+u_3)^2/(2n)]\, dP^3 \\ &\quad + 3\int_{A\times X} [\exp(-u_1-u_2+u_{12}) - \exp(-u_1-u_2)]\, e^{-u_3}\, dP^3 \\ &\quad + 3d_1 h^2 + 3h/(en). \end{aligned}$$

Denoting

$$E_{12} = \int_A \exp(-u_1-u_2)(e^{u_{12}}-1)\, dP^2,$$

we get for the third moment

$$E[p(\mathbf{y}_n)]^3 = \int_{X^3} \exp(-u_1-u_2-u_3)[1-(u_1+u_2+u_3)^2/(2n)]\,dP^3 + 3E_{12}\int_X e^{-u_1}\,dP + 3d_1 h^2 + 3h/(en). \tag{14}$$

In the same manner we obtain for the fourth moment,

$$\begin{aligned} E[p(\mathbf{y}_n)]^4 &= \int_{X^4} \exp[-u_1-u_2-u_3-u_4)[1-(u_1+u_2+u_3+u_4)^2/(2n)]\,dP^4 \\ &\quad + 6\int_{A\times X^2}(\exp(-u_1-u_2+u_{12})-\exp(-u_1-u_2)]\,e^{-u_3-u_4}\,dP^4 \\ &\quad + 6d_2 h^2 + 6h/(en) \\ &= \int_{X^4} \exp(-u_1-u_2-u_3-u_4)[1-(u_1+u_2+u_3+u_4)^2/(2n)]\,dP^4 \\ &\quad + 6E_{12}(\int_X e^{-u_1}\,dP)^2 + 6d_2 h^2 + 6h/(en) \end{aligned} \tag{15}$$

where d_2 is a combinatorial constant. For the fourth central moment μ_4 we have

$$\mu_4 = m_4 - 4m_3 m_1 + 6m_2 {m_1}^2 - 3{m_1}^4$$

where m_1, m_2, m_3, m_4 are the corresponding moments, so that by combining (12)–(15) and as a consequence of the moments being bounded by 1, we obtain,

$$\begin{aligned} E[p(\mathbf{y}_n)-Ep(\mathbf{y}_n)]^4 &= \int_{X^4} \exp(-u_1-u_2-u_3-u_4)\{[1-(u_1+u_2+u_3+u_4)^2/(2n)] \\ &\quad -4[1-(u_1+u_2+u_3)^2/(2n)][1-{u_4}^2/(2n)] \\ &\quad +6[1-(u_1+u_2)^2/(2n)][1-{u_3}^2/(2n)][1-{u_4}^2/(2n)] \\ &\quad -3[1-{u_1}^2/(2n)][1-{u_2}^2/(2n)][1-{u_3}^2/(2n)][1-{u_4}^2/(2n)]\}\,dP^4 \\ &\quad +E_{12}\int_{X^2}\exp(-u_3-u_4)(6-4.3+6.1)\,dP^2 \\ &\quad +a_1 h^2 + a_2 h/n + a_3/n^2 \end{aligned}$$

where a_1, a_2 and a_3 are absolute constants. As a consequence of the symmetry of the expression in the u_i-s and of the boundedness of the function $x^2 e^{-x}$ for non-negative $x-s$, we obtain

$$E[p(\mathbf{y}_n)-Ep(\mathbf{y}_n)]^4 = a_1 h^2 + a_2 h/n + a_3/n^2.$$

LEMMA (c). *Tails of multinomial distributions. Let P_n denote the multinomial probability*:

$$P_n = P_n(n_1,\cdots,n_{k(n)}) = (n!/\textstyle\prod_{i=1}^{k(n)} n_i!)\prod_{i=1}^{k(n)} p_{in}^{n_i}; \qquad \sum_{i=1}^{k(n)} n_i = n$$

where $p_{in} \geqq 0$; $\sum_{i=1}^{k(n)} p_{in} = 1$; *then*

$$P_n \bigcup_{i=1}^{k(n)} \{(n_1,\cdots,n_{k(n)}) \,|\, (|np_{in}-n_i| > n_n')\} \leqq n/{n_n'}^2. \tag{16}$$

PROOF. According to the Tchebyshev inequality, the following inequalities hold for every i,

$$\begin{aligned} P_n\{(n_1,\cdots,n_{k(n)}) \,|\, |np_{in}-n_i| > n_n'\} &\leqq np_{in}(1-p_{in})/{n_n'}^2 \\ &= (n/{n_n'}^2)p_{in}(1-p_{in}) \leqq (n/{n_n'}^2)p_{in}. \end{aligned}$$

Therefore,

$$P_n \bigcup_1^{k(n)} \{(n_1, \cdots, n_{k(n)}) \mid |np_{in} - n_i| > n_n'\} \leqq \sum_1^{k(n)} P_n\{n_1, \cdots, n_{k(n)}) \mid |np_{in} - n_i| > n_n'\}$$
$$\leqq \sum_1^{k(n)} (n/n_n'^2) p_{in} = n/n_n'^2.$$

PROOF OF THE THEOREM. Each of the sets M_n may be partitioned into a sum of three disjoint sets $M_{\alpha n}$, $M_{\beta n}$ and $M_{\gamma n}$ where,

$$M_{\alpha n} = \{(m, m') \in M_n \mid m \in R, m' \notin K - R\},$$
$$M_{\beta n} = \{(m, m') \in M_n \mid m \in K - R, m' \notin K - R\},$$
$$M_{\gamma n} = M_n - M_{\alpha n} - M_{\beta n}.$$

According to this partition we obtain out of the original coverage space $(\mathscr{A}, \mathscr{B}, M_n)$ the three coverage spaces $(\mathscr{A}, \mathscr{B}, M_{\alpha n})$, $(\mathscr{A}, \mathscr{B}, M_{\beta n})$ and $(\mathscr{A}, \mathscr{B}, M_{\gamma n})$. The coverage $q(\mathbf{y}_n)$ in the original space may be written for any experiment $\mathbf{y}_n$ as

$$q(\mathbf{y}_n) = q_\alpha(\mathbf{y}_n) + q_\beta(\mathbf{y}_n) + \theta q_\gamma(\mathbf{y}_n)$$

where q_α, q_β, q_γ are the coverages in the corresponding spaces and θ is a non-negative number not exceeding one which may be dependent on $\mathbf{y}_n$. Denoting by $Z_{\alpha n}$, $Z_{\beta n}$ and $Z_{\gamma n}$ the random variables defined by (5) corresponding to the three latter coverage spaces, we shall show that the distributions of $Z_{\beta n}$ and $Z_{\gamma n}$ tend to the degenerate distribution with zero as its point of increase. As $Z(\mathbf{y}_n) = Z_\alpha(\mathbf{y}_n) + Z_\beta(\mathbf{y}_n) + \theta Z_\gamma(\mathbf{y}_n)$ it will prove that $Z(\mathbf{y}_n)$ and $Z_\alpha(\mathbf{y}_n)$ have the same limiting distribution if any.

For proving the degeneracy it is sufficient to prove that the variances in case tend to zero when $n \to \infty$.

Denoting by H_n the set of all points of K with distances from the boundary of R not exceeding $2c/n^{1/N}$, we have $M_{\beta n} \subset H_n \times H_n$ and $M_{\beta n}(x_1)$ as well as $M_{\beta n}(x_2)$ being subsets of the set H_n. Now H_n tends to the boundary of K when $n \to \infty$ as a consequence of the boundary being closed, $Q\{M_{\beta n}(x_1) \cap M_{\beta n}(x_2)\}$ vanishes outside H_n, and (7) holds with $h = c/n$. So that according to (8) and using the previous notations, we have

$$\sigma^2 Z_{\beta n} \leqq \int_{H_n \times H_n} \exp(-u_1 - u_2)\, n(e^{u_{12}} - 1)\, dP^2 + O(1/n).$$

As a consequence of (4) we get

$$\sigma^2 Z_{\beta n} \leqq bc/n^{1/N} + O(1/n) \to 0 \quad \text{as} \quad n \to \infty.$$

For proving the assertion concerning $Z_{\gamma n}$, we observe that according to (2), $M_n(x)$ is included in a sphere centered at x such that for every x the volume of the sphere is some constant k divided by n. Therefore, we get according to Lebesgue derivation theorem ([9] page 115) that

$$\limsup nQM_n(x) \leqq kf(x) \qquad \text{a.e.}$$

where $f(x)$ is the derivative of Q at x (which exists a.e.). But in the case of $Z_{\gamma n}$,

$M_{\gamma n}(x) \subset K\text{–}R$ and according to Assumption (1) $f(x) = 0$ a.e. on $K\text{–}R$; therefore, $\limsup nQM_{\gamma n}(x) = 0$ a.e. So that

$$\sigma^2 Z_{\gamma n} = \int_{X^2} \exp(-u_1 - u_2)\, n(e^{u_{12}} - 1)\, dP^2 - (\int_X u_1\, e^{-u_1}\, dP)^2 + \theta_1/n + \theta_2/n$$
$$\leqq \int_{(K-R)^2} n(e^{u_{12}} - 1)\, dP^2 + \theta/n \leqq \int_{K-R} (k/n) n(e^{u_1} - 1)\, dP + \theta/n$$
$$= \int_{K-R} k(e^{u_1} - 1)\, dP + \theta/n \qquad |\theta| \text{ bounded.}$$

As $n \to \infty$, $u_1 \to 0$ a.e. on $K\text{–}R$, therefore we get as a consequence of conditions (2) and (4), the Egorov theorem and the Lebesgue dominated convergence theorem [7] that $\sigma^2 Z_{\gamma n} \to 0$ as $n \to \infty$. Now, in case $P(R) = 0$, R may be chosen to be empty and according to the previous discussion $Z_n(\mathbf{y}_n)$ tend to the degenerate distribution which is compatible with the assertion.

In the case where $P(R) > 0$ and K is replaced by R we obtain a coverage space for which the random variable $Z_n(\mathbf{y}_n)$ has a constant ratio $P(K)/P(R)$ to the corresponding random variable in the original space ($n = 1, 2, \cdots$). Therefore it is sufficient to prove the theorem for the case $R = K$.

Now, assuming condition (1.1) holds for $R = K$, let $\{\Pi_m\}$ be a sequence of partitions for which (1.1) holds.

Let m be the number of elements in Π_m and let $K_{1m}, K_{2m}, \cdots, K_{mm}$ denote these elements. Then decomposing M_n into the sets $M_{1n}, M_{2n}, \cdots, M_{mn}, M_{rn}$ where

$$M_{in} = M_n \cap (K_{im} \times K_{im}) \qquad i = 1, 2, \cdots, m$$

and

$$M_{rn} = M_n - \bigcup_{i=1}^{m} M_{in} \qquad \text{then,}$$

$$q(\mathbf{y}_n) = q_1(\mathbf{y}_n) + q_2(\mathbf{y}_n) + \cdots + q_m(\mathbf{y}_n) + \theta q_r(\mathbf{y}_n)$$

where $q_i(\mathbf{y}_n)$ denotes the coverage for the experiment $\mathbf{y}_n$ in the coverage space $(\mathscr{A}, \mathscr{B}, M_{in})$, and $0 \leqq \theta \leqq 1$. Similarly

$$Z(\mathbf{y}_n) = Z_1(\mathbf{y}_n) + Z_2(\mathbf{y}_n) + \cdots + Z_m(\mathbf{y}_n) + \theta Z_r(\mathbf{y}_n).$$

Now, M_{rn} is the union of m sets of the type $M_{\beta n}$ of the previous discussion, so following the same arguments dealing with $Z_{\beta n}$ we get the conclusion that for m fixed as well as for $m = m(n)$ tending slowly enough to infinity when $n \to \infty$, the distribution of Z_r tends to the degenerate distribution with zero as its point of increase.

Therefore, without loss of generality M_n may be replaced by $\bigcup_{i=1}^{m(n)} M_{in}$ where $m(n)$ tends slowly enough to infinity when $n \to \infty$.

Thus the sequence of coverage spaces $\{(\mathscr{A}, \mathscr{B}, M_n)\}$ was modified by replacing the sets M_n with the corresponding sets $\bigcup_{i=1}^{m(n)} M_{in}$ and leaving the spaces $\mathscr{A}$ and $\mathscr{B}$ unchanged. Let $\mathscr{A}_{in} = (K_{im}, S_{in}, P_{(in)})$ and $\mathscr{B}_{in} = (K_{im}, S_{in}, Q_{(in)})$ where S_{in} denotes the sub σ-field of S relative to K_{im}, $P_{(in)}$ is the Lebesgue measure normalized to K_{im}, and $Q_{(in)}$ denotes the conditional probability of Q given K_{im}. Denoting $C_{in} = (\mathscr{A}_{in}, \mathscr{B}_{in}, M_{in})$ and replacing M_n by $\bigcup_{i=1}^{m(n)} M_{in}$ we have,

$$p(\mathbf{y}_n) = \sum_{i=1}^{m(n)} p(\mathbf{y}_{in}) P(K_{im})$$

where $\mathbf{y}_{in}$ denotes the vector composed of the components of $\mathbf{y}_n$ which are in K_{im}, and $p(\mathbf{y}_{in})$ denotes the corresponding vacancy in C_{in}.

Let $n(i)$ denote the number of components of $\mathbf{y}_{in}$ and let

$$Z_{in}(\mathbf{y}_{in}) = (n(i))^{\frac{1}{2}}[p(\mathbf{y}_{in}) - Ep(\mathbf{y}_{in})],$$

i.e. Z_{in} denotes the analogue of Z_n for the coverage space C_{in}; then we have

$$Z_n(\mathbf{y}_n) = \sum_{i=1}^{m} [n/n(i)]^{\frac{1}{2}} Z_{in}(\mathbf{y}_{in}) P(K_{im})$$

where $n(i)$ is random and $P(K_{im})$ arises from the normalization. Now,

$$\Pr(Z_n \in S) = \sum \Pr\{\textstyle\sum_{i=1}^{m} [n/n(i)]^{\frac{1}{2}} Z_{in} P(K_{im}) \in S \mid n(1), n(2), \cdots, n(m)\}$$
$$\cdot \Pr\{n(1), n(2), \cdots, n(m)\}; \qquad \textstyle\sum_1^m n(i) = n.$$

But according to the construction of the sets M_{in} $(i = 1, \cdots, m)$, the conditional probabilities of Z_{in} $(i = 1, \cdots, m)$ are independent and we have therefore for the characteristic functions the following relations:

$$\phi_n(\theta) = \sum_{\sum n(i) = n} \Pr\{n(1), \cdots, n(m)\} \cdot \prod_{i=1}^{m} \phi_{in(i)}[(n/n(i))^{\frac{1}{2}} P(K_{im}) \cdot \theta]$$

where $\phi_n(\theta)$ and $\phi_{in}(\theta)$ denote the characteristic functions of Z_n and of Z_{in} correspondingly. Inserting the values for $\Pr\{n(1), n(2), \cdots, n(m)\}$ and replacing $Q(K_{im})$ with Q_{im} and $P(K_{im})$ with P_{im} we obtain

$$\phi_n(\theta) = \sum_{\sum n(i) = n} n! \prod_{i=1}^{m} (Q_{im}^{n(i)}/n(i)!) \phi_{in}[(n/n(i))^{\frac{1}{2}} P_{im}\theta]$$

or

$$\phi_n(\theta) = \sum{}' n! \prod_{i=1}^{m} (Q_{im}^{n(i)}/n(i)!) \phi_{in}[(n/n(i))^{\frac{1}{2}} P_{im}\theta] + R_n$$

where the $'$ denotes that summation is carried out over all complexes $(n(1), \cdots, n(m))$ for which $|n(i) - nQ_{im}| \leqq n^{\frac{3}{4}}$, and R_n stands for the sum of the remaining terms. But ϕ_{in} being characteristic functions are uniformly bounded; therefore, $R_n \to 0$ when $n \to \infty$ as a result of (16). So that in order to prove the theorem it is sufficient to prove that

$$\phi_n^*(\theta) = \sum{}' n! \prod_{i=1}^{m} (Q_{im}^{n(i)}/n(i)!) \phi_{in}[(n/n(i))^{\frac{1}{2}} P_{im} \cdot \theta]$$

tends to the characteristic function of the normal distribution.

Now, according to the differentiability properties of the characteristic functions ([7] page 144) $\phi(\theta) = 1 + im_1\theta - m_2\theta^2/2 + \eta\mu_3(\theta^3/3!)$ where m_1, m_2 and μ_3 denote the first moment, second moment and third absolute moment of the random variable, while $|\eta| \leqq 1$. But according to (11), condition (2) and the construction of the spaces $\mathscr{A}_{im}$, the fourth central moment of $p(\mathbf{y}_{in})$ is bounded by $a_1 h_{in}^2 + a_2 h_{in}/n(i) + a_3/[n(i)]^2$ with $h_{in} = c/(nP_{im})$. The third absolute central moment of $p(\mathbf{y}_{in})$ is therefore bounded by $c_1/(nP_{im})^{\frac{3}{2}} + c_2/(n(i)nP_{im})^{\frac{3}{4}} + c_3/(n(i))^{\frac{3}{2}}$ where c_1, c_2 and c_3 are constants. The third absolute moment of Z_{in} is therefore bounded by, $c_1[(n(i)/n)/P_{im}]^{\frac{3}{2}} + c_2[(n(i)/n)/P_{im}]^{\frac{3}{4}} + c_3$. So that we obtain for $\phi_n^*(\theta)$,

$$\phi^*(\theta) = \sum{}' (n!/\textstyle\prod_1^m n_i!) \prod_1^m Q_{im}^{n(i)} \prod_1^m \{1 - p_{im}^2[n/(nQ_{im} + n'(i)]\sigma_{inn'(i)}^2 \theta^2/2$$
$$+ \eta_1 p_{im}^{\frac{3}{2}} + \eta_2 p_{im}^{\frac{9}{4}}[n/(nQ_{im} + n'(i))]^{\frac{3}{4}} + \eta_3 p_{im}^3/[n/(nQ_{.m} + n'(i))]^{\frac{3}{2}}$$

where $n'(i)$ stands for $n(i)-nQ_{im}$, $\sigma^2_{inn'(i)}$ denotes the corresponding variance of Z_{in}, and η_1, η_2 and η_3 are bounded.

But as $|n'(i)| \leqq n^{\frac{3}{4}}$ we have for $m = m(n)$ chosen such that Q_{im} decreases slowly enough that

$$\phi_n^*(\theta) = \sum'(n!/\prod_1^m n_i!)\prod_1^m Q_{im}^{n(i)}\prod_1^m\{1-P_{im}^2/[(1+n(i)/(nQ_{im}))Q_{im}]\sigma^2_{inn'(i)}\theta^2/2 + O(P_{im}^{\frac{3}{2}}) + O(P_{im}^{\frac{9}{4}}/Q_{im}^{\frac{3}{4}}) + O(P_{im}^3/Q_{im}^{\frac{3}{2}})\} \tag{17}$$

where the boundedness is uniform in i and in all the sets $\{n_i\}$ under the summation sign.

Now the right-hand product in (17) is of the form $\prod_{j=1}^{m(i)}(1-a_{ij})$ and it may be easily verified by taking logarithms and expanding into a power series that when $i\to\infty$, $\sum_{j=1}^{m(i)} a_{ij}\to a$ and $\max_j|a_{ij}|\to 0$, then $\prod_1^{m(i)}(1-a_{ij})\to e^{-a}$. But according to the assumptions of the theorem and the estimate (8) of the variance of coverage, the variance σ_n^2 of Z_n tends to a limit when $n\to\infty$. On the other hand,

$$\sigma_n^2 = \sum_{\sum n(i)=n}(n!/\prod_1^m n(i)!)\prod_1^m Q_{im}^{n(i)}\sum_{i=1}^m[P_{im}^2/(Q_{im}+n'(i)/n)]\sigma^2_{inn(i)} \tag{18}$$

and because of (1.1):

$$\sum_i P_{im}^{\frac{3}{2}} \leqq \max_i P_{im}\sum_i P_{im} = \max P_{im}\to_{m\to\infty} 0,$$

$$\sum_i P_{im}^{\frac{9}{4}}/Q_{im}^{\frac{3}{4}} = \sum P_{im}(P_{im}^{\frac{5}{4}}/Q_{im}^{\frac{3}{4}}) \leqq \max_i(P_{im}^{\frac{5}{4}}/Q_{im}^{\frac{3}{4}}) \leqq \max P_{im}^{\frac{1}{4}}[P_{im}^{\frac{4}{3}}/Q_{im}]^{\frac{3}{4}}\to 0,$$

$$\sum_i P_{im}^3/Q_{im}^{\frac{3}{2}} \leqq \max_i(P_{im}^{\frac{4}{3}}/Q_{im})^{\frac{3}{2}}\to 0.$$

Therefore, if it will be shown that

$$\max_i\left|P_{im}^2\sigma^2_{inn'(i)}/(Q_{im}+n'(i)/n)\right|\to 0 \tag{18a}$$

and

$$\sum_{i=1}^m P_{im}^2\sigma^2_{inn'(i)}/(Q_{im}+n'(i)/n)\to\lim\sigma_n^2 \quad \text{uniformly.} \tag{18b}$$

It will follow that the right-hand product in (17) tends to $\exp(-\lim\sigma_n^2\theta^2/2)$ and consequently $\phi_n^*(\theta)$ tends to the same limit function, which proves the theorem.

In order to prove (18a) we use the following notations,

$$v_{ij} = (n+n'(i)/Q_{im})Q\{M_n(x_j)\} \qquad j=1,2$$

$$v_{i12} = (n+n'(i)/Q_{im})Q\{M_n(x_1)\cap M_n(x_2)\}.$$

According to (8), we have for $\sigma^2_{inn'(i)}$

$$\sigma^2_{inn'(i)} = (1/P_{im}^2)\int_{K_{in}^2}\exp(-v_{i1}-v_{i2})(nQ_{im}+n'(i))(e^{v_{i12}}-1)\,dx_1\,dx_2 - (1/P_{im}^2)\left[\int_{K_{in}} v_{i1}\,e^{-v_{i1}}\,dx_1\right]^2+\eta/(nQ_{im}+n'(i)) \tag{19}$$

where η is bounded uniformly in m, n, i and $n'(i)$. Therefore, for n sufficiently large

$$P_{im}^2\sigma_{inn'(i)}^2/(Q_{im}+n'(i)/n) \leqq \int_{K_{in}^2} n\exp(-v_{i1}-v_{i2})(e^{u_{i12}}-1)\,dx_1\,dx_2 + \eta P_{im}^2/[n(Q_{im}+n'(i)/n)^2]. \tag{20}$$

Now choosing $m = m(n)$ such that both $\max_i P_{im}$ and $\max_i Q_{im}$ tend slowly enough to zero we get as a result of the condition (1.1) that the factor multiplying η,

$$P_{im}^2/[n(Q_{im}+n'(i)/n)^2] \sim P_{im}^2/nQ_{im}^2 = O(1/(nP_{im}^3)) \to 0 \quad \text{uniformly.}$$

The integral on the right-hand side of the inequality (20) may be written as

$$\int_{K_{in}} e^{-v_{i1}}\,dx_1 \int_{K_{in}} e^{-v_{i2}}\,n(e^{v_{i12}}-1)\,dx_2,$$

but $v_{i1} \geqq 0$, and according to (4) the right-hand side integrand is uniformly bounded by b; therefore the whole expression is bounded by bP_{im} and tends uniformly to zero, which proves assertion (18a). In order to prove (18b) it has to be observed that

$$\begin{aligned} P_{im}^2\sigma_{inn'(i)}^2/(Q_{im}+n'(i)/n) &= (P_{im}^2\sigma_{ino}^2/Q_{im})[1+O(n(i)/nQ_{im})] \\ &= (P_{im}^2\sigma_{ino}^2/Q_{im})[1+o(1)] \end{aligned}$$

where the boundedness is uniform in i and $n'(i)$.

So that according to (19) we have

$$\begin{aligned} \sum_{i=1}^m P_{im}\sigma_{ino}^2/Q_{im} = \sum_{i=1}^m \int_{K_{im}^2} &\exp[-nQ_n M_n(x_1) - nQ_n M_n(x_2)] \\ &\times n\{\exp[nQ_n(M_n(x_1)\cap M_n(x_2))]-1\}\,dx_1\,dx_2 \\ &-\sum_{i=1}^m \left(\int_{K_{im}} nQ_n M_n(x)\exp(-nQ_n M_n(x))\,dx\right)^2 + o(1) \end{aligned}$$

and as a result of the special structure of M_n, the right-hand side of the equation is equal to σ_n^2 up to $o(1)$. We get, therefore, that for every set $\{n(i)\}$ over which the summation $\sum'$ is carried over we have uniformly

$$\sum_1^m P_{im}^2\sigma_{inn'(i)}^2/(Q_{im}+n'(i)/n) = \sigma_n^2 + o(1),$$

which proves assertion (18b) and concludes the proof of the theorem.

DISCUSSION. The conditions of the theorem are obviously not necessary ones, as it is readily seen from the relation of coverage space-coverage problem, also from the proof of the theorem. For instance, as the measures of the random sets and their intersections are preserved under measure preserving autotransformations of Y, it is evident that condition (2) may be replaced by the weaker condition that (2) should be satisfied for some measure preserving autotransformation of Y. But while, in this case, there is no real loss of generality, because we may choose for any coverage problem any suitable coverage space, the situation is different concerning conditions (3) and (4). These two conditions may be weakened by imposing the appropriate conditions on the corresponding integrals rather than on the integrands. But as the necessary conditions are still not attained it was

preferred to sacrifice generality for more simplicity in formulation. Concerning condition 1, the (1.1) part of it is a Lipschitz-type condition which makes sure that $Q(S)$ should not tend too fast to zero relative to $P(S)$ when S contracts to points for which the density of Q is equal to 0. This condition is certainly satisfied in case Q has on K a density bounded from below by a positive number. It is also readily seen that the theorem remains valid when (1.1) is weakened by demanding it to hold only approximately, i.e. that there should exist a sequence of positive h_n's with $h_n \to 0$ such that all the boundaries of $R_m = \{x \in R \mid Q'(x) \geqq h_m\}$ are of measure zero, where $Q'(x)$ denotes the density of Q at x. This result is observed by breaking down the coverage spaces into two parts, such that R_m is the space X of the first part, and letting $m = m(n)$ tend slowly enough to infinity. Then the distribution of coverages in the sequence of the first parts tends to the appropriate normal distribution, and that of the second parts tends to the degenerate distribution with zero as its point of increase. The asymptotic normality follows from the theorem while the asymptotic degeneracy follows by the same arguments used for proving the limiting degeneracy of Z_β and Z_γ. Also, as commented by the referee, the Lebesgue σ-field may be replaced in the theorem by the Borel σ-field.

4. Applications.

1. *The classical occupancy problem.* This problem which has special importance in mechanical statistics deals with the distribution of the number of empty cells when balls are thrown independently into a set of cells such that each of the cells is equally likely to catch each of the balls. Irving Weiss [9] proved that a sufficient condition for the fraction of empty cells to be asymptotically normally distributed is one in which the ratio of the number of balls to that of the cells should be kept constant. A. Rényi [8] has generalized Weiss' theorem by proving that a necessary and sufficient condition for a limiting normal distribution is:

$$n \exp[-(N+1)/n]\{1-[1+(N+1)/n]\exp[-(N+1)/n]\} \to \infty$$

where n and N denote the number of cells and the number of balls respectively. V. P. Čistiakov and I. I. Viktorova [3] gave sufficient conditions for normal limiting distribution in cases where the probability of catching a ball may be different for different cells.

Taking as $\mathscr{A}$ and as $\mathscr{B}$ the Euclidean segment [0, 1) with Lebesgue σ-field and measure; denoting by A_{in} and by A'_{in} the interval $[i/n, (i+1)/n)$ in K and in Y respectively, and taking $M_n = \sum_{i=0}^{n-1} A_{in} \times A'_{in}$, then the coverage space $(\mathscr{A}, \mathscr{B}, M_n)$ is a proper coverage space for the Classical Occupancy Problem and $p(\mathbf{y}_n)$ is equal to the fraction of empty cells for the cases of Weiss and Rényi. Thus the Classical Occupancy problem may be treated as a special case of a coverage problem. Similarly, by taking $A'_{i,n} = [a_{i,n}, a_{i+1,n})$, $0 = a_{0,n} \leqq a_{1,n} \leqq \cdots \leqq a_{n-1,n} = 1$, we obtain a coverage space corresponding to the case of Čistiakov and Viktorova (evidently any permutation of the A'_{in}'s may be taken for the same problem). Now the theorem of Weiss as well as the empty cells case of Čistiakov and

Viktorova are special cases of the theorem in the present paper. For instance, we have for Weiss' case,

(1.1)W $R = K; P = Q$; thus $P^{\frac{1}{2}}(S)/Q(S) = P^{\frac{1}{2}}(S) \to 0$ uniformly for equi-partitions of R

(2)W $$M_n \subset \{(x, y) \mid |x-y| \leqq 1/n\}$$

(3)W $$nQ\{M_n(x)\} = n \cdot 1/n = 1$$

(4)W $$\int_K n[\exp(nQ\{M_n(x_1) \cap M_n(x_2)\}) - 1]\, dP(x_2) = \int_{A_{in}} n[\exp n(1/n) - 1]\, dx_2 = e - 1$$

where A_{in} is that interval which includes x_2. Similarly, all the conditions of the theorem are satisfied for the case of Čistiakov and Viktorova.

2. *Multivariate curve fitting*. F. N. David has proposed a univariate curve fitting test based on the Classical Occupancy [4]. As the classical occupancy is a special case of a coverage problem it is reasonable to assume that better tests may possibly be obtained when the class of coverages is not confined to the classical occupancy only. Unfortunately, computations of distributions of coverages are involved and the difficulties are practically insurmountable, except in a few simple cases.

The theorem proved in the present paper asserts that for large samples the distribution of coverage is approximately normal for the class of coverage spaces satisfying some regularity conditions, a property which is also valid for the multivariate case. This property may be used for construction of multivariate curve fitting tests for large samples. Various tests based on coverages may be suggested which have the advantage of being easily carried out by the aid of a computer and have also the advantage that the power of the test against any alternative may be easily computed. The following is an example of such a test.

Let the null hypothesis be that $F(\mathbf{x})$ is the distribution function of a vector $\mathbf{x}$ in the Euclidean N-space where $F(\mathbf{x})$ satisfies the regularity condition (1.1) on some Lebesgue measurable and bounded set K. Now, given a large sample $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, feed into the memory of a computer the n N-cubes parallel to the coordinates centered at the $\mathbf{x}_i$'s and having sides of length s[1]. Measure (count) the "volume" covered by the union of the cubes. Now the asymptotic volume of that union is normally distributed with mean given by [2] $\int_K \exp[-nvf(\mathbf{x})]\, dP(\mathbf{x})$ and variance obtained from formula (8), where $f(\mathbf{x})$ denotes the density function of $\mathbf{x}$ (which exists a.e. and is equal a.e. to the Radon–Nikodym derivative of the Lebesgue absolute continuous part of P), v denotes the volume of the covering

[1] The probability of discriminating between the hypothesis and the alternative becomes small for too large as well as for too small values of s. In the first case the probability of the vacancy being $1 - nv$ will be close to 1 for both the hypothesis and alternative, while in the second case the same is true for vacancies equal to zero. There is a value of s which optimalizes the discrimination between hypothesis and alternative. Evidently it depends on n. The larger the n the smaller this optimal value is. s has therefore to be chosen accordingly.

cube and P denotes the Lebesgue measure. A test of the null hypothesis may, therefore, be carried out by comparing the volume covered with the expected volume according to the corresponding normal distribution scale. Similarly, the power of the test against any alternative may be computed.

Evidently the cubes may be replaced by "boxes" with sides which may be different from each other and which may also be functions of $\mathbf{x}$. The test in that case is carried out in the same manner as the previous one after the proper modifications of the formulas for the asymptotic mean and variance were made [2].

By a proper choice of the side lengths as functions of $\mathbf{x}$ the test may be made "best". Similarly, in case the set K is not bounded the test may still be carried out by suitably transforming K into the unit cube and appropriately transforming $F(\mathbf{x})$.

REFERENCES

[1] AILAM, G. (1966). Moments of coverage and coverage spaces. *J. Appl. Probability* **3** 550–555.

[2] AILAM, G. (1968). On probability properties of measures of random sets and the asymptotic behaviour of empirical distribution functions. *J. Appl. Probability* **5** 196–202.

[3] CISTIAKOV, V. P., and VIKTOROVA, I. I. (1965). Asymptotic normality in a problem of balls falling into different boxes when the probabilities of falling into different boxes are different. *Teor. Verojatnost. i Primenen.* **10** 162–167.

[4] DAVID, F. N. (1950). Two combinatorial tests of whether a sample has come from a given population. *Biometrika* **37** 97–110.

[5] HEMMER, P. Chr. (1959). A problem of geometrical probabilities. *Norske Vid. Selsk. Forh. Trondheim* **32** 117–120.

[6] GARWOOD, F. (1947). The variance of overlap of geometrical figures with reference to a bombing problem. *Biometrika* **34** 1–17.

[7] LOÈVE, M. (1955). *Probability Theory*. Van Nostrand, New York.

[8] RÉNYI, A. (1962). Three new proofs and a generalization of a theorem of Irving Weiss. *Magyar. Tud. Akad. Mat. Fiz. Oszt. Közl.* **7** 203–215.

[9] SAKS, S. (1937). *Theory of the Integral*, (2nd ed.) Hafner, New York.

[10] WEISS, I. (1958). Limiting distributions of some occupancy problems. *Ann. Math. Statist.* **29** 878–884.