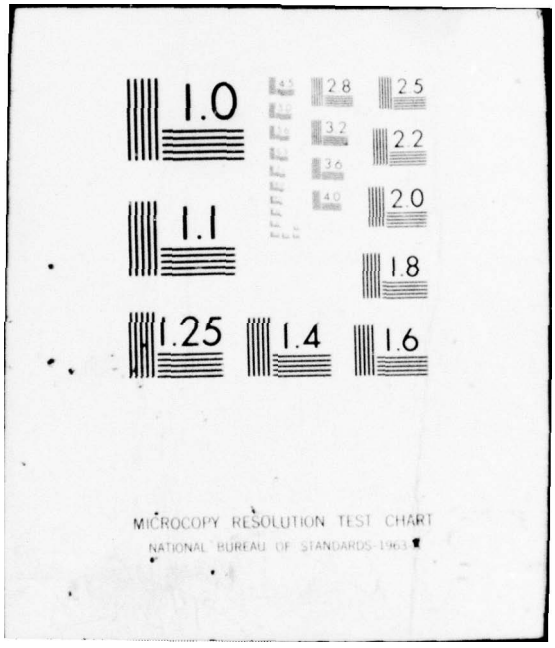
AD-A031 957 WISCONSIN UNIV MADISON MATHEMATICS RESEARCH CENTER F/G 12/1
SOME LIMIT THEOREMS FOR THE INDISTINGUISHABLE BALL PROBLEM WITH--ETC(U)
AUG 76 L HOLST DAAG29-75-C-0024
UNCLASSIFIED MRC-TSR-1660 NL

1 OF 1
AD
A031957

END
DATE
FILMED
1-77


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963

FG

MRC Technical Summary Report #1660

SOME LIMIT THEOREMS FOR THE INDISTINGUISHABLE BALL PROBLEM WITH APPLICATIONS IN NONPARAMETRICS

Lars Holst

Mathematics Research Center
University of Wisconsin—Madison
610 Walnut Street
Madison, Wisconsin 53706

August 1976

Received March 22, 1976

Approved for public release
Distribution unlimited

Sponsored by

U. S. Army Research Office
P.O. Box 12211
Research Triangle Park
North Carolina 27709

UNIVERSITY OF WISCONSIN - MADISON
MATHEMATICS RESEARCH CENTER

SOME LIMIT THEOREMS FOR THE INDISTINGUISHABLE BALL PROBLEM WITH APPLICATIONS IN NONPARAMETRICS

Lars Holst

Technical Summary Report #1660
August 1976

ABSTRACT

In Feller (1968), An Introduction to Probability Theory and Its Applications, Vol. 1, the following urn model is discussed. Consider n m urns and distribute n indistinguishable balls among the urns such that the distinguishable distributions of the balls all have the same probability $1/\binom{n+m-1}{m-1}$.

Let S_k denote the number of balls in the k^{th} urn. Clearly $S_1 + \ldots + S_m = n$. In this paper, random variables of the type $Z = h(S_1, \ldots, S_m)$, especially $h(S_1, \ldots, S_m) = h_1(S_1) + \ldots + h_m(S_m)$, are studied when $m, n \to \infty$ in such a way $m/n \to \rho$, $0 < \rho < \infty$. Some applications of the results in nonparametric statistics are briefly discussed and the limit distribution of $\max(S_1, \ldots, S_m)$ is derived.

AMS(MOS) Classification: primary 60C05, secondary 62G99.

Key words: Indistinguishable ball problem; Bose-Einstein statistics; urn models; occupancy; spacings; limit theorems; combinatorics; geometric distribution; nonparametrics.

Work Unit No. 4 (Probability, Statistics and Combinatorics)

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

SOME LIMIT THEOREMS FOR THE INDISTINGUISHABLE BALL PROBLEM WITH APPLICATIONS IN NONPARAMETRICS

Lars Holst

1. Introduction

In Feller (1968) the following problem is discussed. Consider m urns and distribute n indistinguishable balls into the urns in such a manner that all distinguishable distributions of the balls have the same probability. What is the probability for each distinguishable outcome? A solution is: place $n+m-1$ balls into a row and pick out $m-1$ balls at random (without replacement) and think of the "gaps" as the "walls" of the urns. In this way we obtain a distribution of the balls over the urns which is of the type above. As there are $\binom{n+m-1}{m-1}$ ways of picking out the $m-1$ balls the probability for each outcome is $1/\binom{n+m-1}{m-1}$. If we let $S_1,\ldots,S_m$ denote the number of balls in the urns, we have thus found the joint distribution of $(S_1,\ldots,S_m)$. Note that $S_1 + \ldots + S_m = n$ so the S's are dependent random variables. In this paper we will consider random variables of the type $Z = h(S_1,\ldots,S_m)$, especially $h(S_1,\ldots,S_m) = h_1(S_1) + \ldots + h_m(S_m)$, and limit theorems for such random variables when $m,n \to \infty$ such that $m/n \to \rho$, $0 < \rho < \infty$.

The above urnmodel has been used in physics in connection with Bose-Einstein statistics, see Feller (1968), p. 39.

The urnmodel is also of relevance for the two sample problem in statistics. Let $X_1,\ldots,X_{m-1}(Y_1,\ldots,Y_n)$ be $m-1$ (n) observations from a continuous distribution F_X (F_Y). The problem is to test the hypothesis $H_0 : F_X = F_Y$, i.e. the samples have come from the same parent distribution.

Sponsored by the United States Army under Contract No. DAAG29-75-C-0024.

Let the X's be ordered $X'_1 < X'_2 < \ldots < X'_{m-1}$ (they are different with probability one) and let S_k be the number of Y's in $[X'_{k-1}, X'_k)$, $k = 1, \ldots, m$, with $X'_0 = -\infty$, $X'_m = +\infty$. If both samples have come from the same continuous distribution then $(S_1, \ldots, S_m)$ has the same distribution as in the urn-model. Tests based on statistics of the form $h_1(S_1)+\ldots+h_m(S_m)$ are investigated in Holst and Rao (1976). Such statistics are connected with many non-parametric tests.

We also note that the random variable $(S_1, \ldots, S_m)$ can be constructed in the following way. Take m-1 points at random in (0,1), order them, and let D_k be the distance between the k-1:th and k:th point (the k:th spacing from uniform (0,1)), $k = 1, \ldots, m$ ($k = 0(m)$ corresponds to $0(1)$). Given the D's consider m urns and throw n balls independently into the urns such that the probability of hitting the k:th urn is D_k. If S_k is the number of balls in the k-th urn, then the (unconditional) distribution of $(S_1, \ldots, S_m)$ is as above.

In Park (1973) another formulation is given and a limit theorem is derived for the case $h_k(j) = h(j)$, where $h(j) = 0$ for $j > K$.

In order to properly state limit results sequences of functions $h_{k\nu}$'s etc. are considered. But in order to facilitate the notation we suppress the index ν. We use the notation $\mathcal{L}(U_n) \to \mathcal{L}(U)$ for convergence in distribution. $N(m, \Sigma)$ stands for the normal distribution with mean m and covariance Σ. $Po(\lambda)$ denotes the Poisson distribution with mean λ.

We will always suppose in the following that $m, n \to \infty$ such that $r = m/n \to \rho$, $0 < \rho < \infty$.

2. The characteristic function.

Let $\eta, \eta_1, \eta_2, \ldots$ be i.i.d. random variables with the geometric distribution

$$P(\eta = j) = r/(r+1)^{j+1}, \qquad j = 0,1,2,\ldots ,$$

where $r = m/n$. It is well known that

$$E(\eta) = 1/r , \quad Var(\eta) = (r+1)/r^2 .$$

Let $M \leq m$ and consider for a given function $h(\cdot)$ the random variable

$$Z_M = h(S_1, \ldots, S_M) .$$

Theorem 1. The following representation of the characteristic function of Z_M holds:

$$E(e^{itZ_M}) = (2\pi)^{-1} \binom{n+m-1}{n}^{-1} (r+1)^{n+m} \cdot r^{-m} \cdot$$
$$\int_{-\pi}^{\pi} E(\exp(ith(\eta_1, \ldots, \eta_M) + i\theta \sum_{1}^{M} (\eta_k - 1/r))) \cdot ((r+1-e^{i\theta})e^{i\theta/r}/r)^{M-m} d\theta .$$

Proof. It is well-known that $\eta_1 + \ldots + \eta_m$ has a negative-binomial distribution, i.e.

$$P(\eta_1 + \ldots + \eta_m = n) = \binom{n+m-1}{n} r^m (r+1)^{-(n+m)} ,$$

and that

$$E(\exp(it \sum_{M+1}^{m} (\eta_k - 1/r)) = ((r+1-e^{i\theta})e^{i\theta/r}/r)^{M-m} .$$

Therefore the assertion can be written

$$E(e^{itZ_M}) = (2\pi P(\eta_1 + \ldots + \eta_m = n))^{-1} \cdot$$
$$\cdot \int_{-\pi}^{\pi} E(\exp(ith(\eta_1, \ldots, \eta_M) + i\theta \sum_{1}^{m} (\eta_k - 1/r)))d\theta .$$

Hence it is sufficient to consider the case $M = m$.

We also have for $j_1 + \ldots + j_m = n$, $j_k = 0, 1, 2, \ldots,$

$$P(\eta_1 = j_1, \ldots, \eta_m = j_m \mid \eta_1 + \ldots + \eta_m = n) = 1/\binom{n + m - 1}{n},$$

or the conditional distribution of $(\eta_1, \ldots, \eta_m)$ given $\eta_1 + \ldots + \eta_m = n$ is the same as the distribution of $(S_1, \ldots, S_m)$. Therefore we have

$$\sum_{j_1 + \ldots + j_m = n} P(\eta_1 = j_1, \ldots, \eta_m = j_m) e^{ith(j_1, \ldots, j_m)} / P(\eta_1 + \ldots + \eta_m = n) =$$
$$= E(e^{ith(S_1, \ldots, S_m)}).$$

Now we can write

$$\int_{-\pi}^{\pi} E(\exp(ith(\eta_1, \ldots, \eta_m) + i\theta \sum_1^m (\eta_k - 1/r))d\theta =$$

$$= \int_{-\pi}^{\pi} \sum_{j_1, \ldots, j_m = 0}^{\infty} \exp(ith(j_1, \ldots, j_m) + i\theta \sum_1^m j_k) \cdot e^{-i\theta n} \cdot P(\eta_1 = j_1, \ldots, \eta_m = j_m)d\theta$$

$$= \sum_{j_1, \ldots, j_m = 0} e^{ith(j_1, \ldots, j_m)} P(\eta_1 = j_1, \ldots, \eta_m = j_m) \cdot \int_{-\pi}^{\pi} \exp(i\theta(\sum_1^m j_k - n))d\theta =$$

$$= \sum_{j_1 + \ldots + j_m = n} e^{ith(j_1, \ldots, j_m)} \cdot P(\eta_1 = j_1, \ldots, \eta_m = j_m) \cdot 2\pi =$$

$$= E(e^{ith(S_1, \ldots, S_m)}) \cdot P(\eta_1 + \ldots + \eta_m = n) \cdot 2\pi .$$

The above interchange of summation and integration is allowed because of the absolute convergence of the series.

Combining the above results proves the theorem. ∎

In the following lemmas we consider the "non-random parts" of the representation in Theorem 1. Recall $r = m/n \to \rho$, $0 < \rho < \infty$, when $m,n \to \infty$.

Lemma 1. When $m,n \to \infty$ we have

$$(2\pi)^{-1}\binom{n+m-1}{n}^{-1}(r+1)^{n+m}\, r^{-m}/m^{\frac{1}{2}} = ((r+1)/r^2 2\pi)^{\frac{1}{2}} \cdot (1 + O(1/m)) .$$

Proof. By Stirling's formula

$$\binom{n+m-1}{n}^{-1} = n!(m-1)!/(n+m-1)! =$$

$$= n^n e^{-n}(2\pi n)^{\frac{1}{2}}\, m^m e^{-m}(2\pi m)^{\frac{1}{2}} \cdot m^{-1} \cdot$$

$$(n+m) \cdot (n+m)^{-(n+m)} e^{n+m} (2\pi(n+m))^{-\frac{1}{2}} \cdot e^{O(1/m)} =$$

$$= ((1+r)2\pi)/r^2)^{\frac{1}{2}} \cdot r^m (1+r)^{-n-m} \cdot m^{\frac{1}{2}} \cdot e^{O(1/m)} ,$$

from which the assertion follows. ∎

Lemma 2. Let $M, m \to \infty$ so that $M/m \to \alpha$, $0 < \alpha < 1$, and set

$$f_n(\theta) = ((r+1 - e^{i\theta})\, e^{i\theta/r}/r)^{M-m} .$$

Then for each fixed real number ψ

$$f_n(\psi/m^{\frac{1}{2}}) \to f(\psi) = \exp(-(1-\alpha)\psi^2(\rho+1)/2\rho^2) ,$$

and

$$m^{\frac{1}{2}} \int_{-\pi}^{\pi} |f_n(\theta)|\, d\theta \to \int_{-\infty}^{\infty} |f(\psi)|\, d\psi .$$

Proof. By expanding into Taylor series we find

$$f_n(\psi/m^{\frac{1}{2}}) = (1 + \psi^2(r+1)/2r^2 m + O(m^{-3/2}))^{-m(1-M/m)}$$

$$\to f(\psi) = \exp(-(1-\alpha)\psi^2(\rho+1)/2\rho^2),\ n \to \infty ,$$

which proves the first statement.

Let $d > 0$ be fixed, then

$$m^{\frac{1}{2}} \int_{\pi \geq |\theta| \geq d} |f_n(\theta)| d\theta = m^{\frac{1}{2}} \int_{\pi \geq |\theta| \geq d} (1+2(r+1)(1-\cos\theta)/r^2)^{-(m-M)/2} d\theta$$

$$\leq 2\pi\, m^{\frac{1}{2}} (1+2(r+1)(1-\cos d)/r^2)^{-(m-M)/2} \to 0 .$$

For d sufficiently small and m sufficiently large there exists $K_d > 0$ and $C > 0$ such that for $|\theta| < d$

$$|f_n(\theta)| \leq (1 + K_d \theta^2)^{-C \cdot m} .$$

Therefore for $A = \{\theta ; m^{-1/2 + 1/7} < |\theta| < d\}$

$$|m^{\frac{1}{2}} \int_A f_n(\theta) d\theta| \leq \int_{m^{\frac{1}{2}}A} (1+K_d\psi^2/m)^{-C \cdot m} d\psi \to 0 .$$

In $B = \{\theta ; |\theta| \leq m^{-1/2+1/7}\}$ we can use Taylor expansion and find

$$m^{\frac{1}{2}} \int_B |f_n(\theta)| d\theta \to \int_{-\infty}^{\infty} |f(\psi)| d\psi < \infty ,$$

which completes the proof. ∎

3. Some limit theorems.

First we give a general theorem.

Theorem 2. Suppose that $M, m \to \infty$ such that $M/m \to \alpha$, $0 < \alpha < 1$, and that the sequence $\{h_M(\cdot)\}$, for each fixed real numbers t and s, satisfies

$$E(\exp(ith_M(\eta_1, \ldots, \eta_M) + is \sum_{k=1}^{M} (\eta_k - 1/r)r/((r+1)m)^{\frac{1}{2}})) \to H(t,s)\, e^{-\alpha s^2/2}, \quad m \to \infty,$$

where $H(t,s)$ is continuous and $H(0,s) \equiv 1$. Then with $Z_M = h_M(S_1, \ldots, S_M)$ we have

$$\mathcal{L}(Z_M) \to \mathcal{L}(Z_\alpha)$$

where Z_α's characteristic function is

$$E(e^{itZ_\alpha}) = (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} H(t,s) e^{-s^2/2}\, ds.$$

Proof. By Theorem 1 and Lemma 1 we have

$$E(e^{itZ_M}) = e^{o(1)} \cdot (2\pi)^{-\frac{1}{2}} \cdot \int_{-\pi m^{\frac{1}{2}}\sigma}^{\pi m^{\frac{1}{2}}\sigma} E(\exp(ith_M(\eta_1, \ldots, \eta_M) + is \sum_{1}^{M} (\eta_k - 1/r)/\sigma m^{\frac{1}{2}}))f_n(\psi/m^{\frac{1}{2}}\sigma)d\psi$$

using the notation of Lemma 2 and $\sigma^2 = \mathrm{Var}(\eta) = (r+1)/r^2$. By the extended Lebesgue dominated convergence theorem (see Rao (1973), p. 136) and Lemma 2 it follows from the assumptions that

$$E(e^{itZ_M}) \to (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} H(t,s) \cdot e^{-\alpha s^2/2} \cdot e^{-(1-\alpha)s^2/2}\, ds =$$

$$= (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} H(t,s)\, e^{-s^2/2}\, ds.$$

As $H(t,s)$ is continuous the last integral is a continuous function of t. Thus the assertion follows by the continuity theorem for characteristic functions. ∎

Of particular importance is the special case

$$h(S_1,\dots,S_m) = \sum_1^m h_k(S_k) ,$$

where $h_1(\cdot),\dots,h_m(\cdot)$ are given functions.

Theorem 3. Suppose that $M,m \to \infty$ such that $M/m \to \alpha$, $0 < \alpha \le 1$, and for some $\alpha_0 < 1$ we have for $\alpha_0 \le \alpha \le 1$

$$\mathfrak{L}\begin{pmatrix} \sum_{k=1}^{M} h_k(\eta_k) \\ \sum_{k=1}^{M} (\eta_k - 1/r)r/((r+1)m)^{\frac{1}{2}} \end{pmatrix} \to N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} A_\alpha & B_\alpha \\ B_\alpha & \alpha \end{pmatrix}\right) ,$$

where as $\alpha \to 1-$

$$A_\alpha \to A_1 \quad \text{and} \quad B_\alpha \to B_1 .$$

Then

$$\mathfrak{L}\Big(\sum_1^m h_k(S_k)\Big) \to N(0, A_1 - B_1^2) .$$

Proof. For $\alpha_0 \le \alpha < 1$ the assumptions of Theorem 2 are satisfied and therefore

$$E(e^{itZ_M}) \to (2\pi)^{-\frac{1}{2}} \int_{-\infty}^{\infty} \exp(-(A_\alpha t^2 + 2\alpha B_\alpha ts + s^2)/2)ds$$

$$= \exp(-(A_\alpha - \alpha^2 B_\alpha^2)s^2/2) .$$

So $\mathfrak{L}(Z_\alpha)$ is $N(0, A_\alpha - \alpha^2 B_\alpha^2)$. In the same way we find

$$\mathfrak{L}(Z'_M) = \mathfrak{L}\Big(\sum_{M+1}^m h_k(S_k)\Big) \to N(0, A_1 - A_\alpha - (1-\alpha)^2 (B_1 - B_\alpha)^2) .$$

Now when $\alpha \to 1$ -

$$A_\alpha = \alpha^2 B_\alpha^2 \to A_1 - B_1^2 ,$$

$$A_1 - A_\alpha - (1 - \alpha)^2(B_1 - B_\alpha^2) \to 0 .$$

Using an argument by Le Cam (1958) p. 13-14 it follows that

$$\mathfrak{L}(Z_M + Z'_M) = \mathfrak{L}\left(\sum_{k=1}^{m} h_k(S_k)\right) \to N(0, A_1 - B_1^2) . \qquad \blacksquare$$

Particularly simple is the symmetric case $h_k(\cdot) = h_n(\cdot)$, $k = 1, 2, \ldots, m$.

Theorem 4. Suppose that

$$\mathfrak{L}\begin{pmatrix} \sum_{k=1}^{m} h_n(\eta_k) \\ \sum_{k=1}^{m} (\eta_k - 1/r)r/((r+1)m)^{\frac{1}{2}} \end{pmatrix} \to \mathfrak{L}\begin{pmatrix} U \\ V \end{pmatrix}$$

for some random vector (U,V). Then

$$E(e^{itU + isV}) = G(t) \cdot e^{-(At^2+2Bts+s^2)/2}$$

with $G(t)$ having no normal component and

$$\mathfrak{L}\left(\sum_{1}^{m} h_n(S_k)\right) \to \mathfrak{L}(Z)$$

with

$$E(e^{itZ}) = G(t)\, e^{-(A - B^2)t^2/2} .$$

Proof. By classical limit theorems for independent identically distributed random variables the first assertion follows (cf. Le Cam (1958), p. 8). It is also easily seen that the function $H(t,s)$ of Theorem 2 is given by

$$H(t,s) = (G(t))^{\alpha}\, e^{-\alpha(At^2+2Bts)/2} .$$

Therefore

$$E(e^{itZ_\alpha}) = (2\pi)^{-\frac{1}{2}} (G(t))^\alpha \int_{-\infty}^{\infty} e^{-(\alpha At^2+2\alpha Bts+s^2)/2}\, ds =$$

$$= (G(t))^\alpha e^{-(\alpha A - \alpha^2 B^2)t^2/2} .$$

By the same argument as in Theorem 3 the assertion follows. ∎

If the limit distribution of $\sum_1^m h_n(\eta_k)$ has no normal component then we get:

Theorem 5. Suppose that

$$\mathfrak{L}(\sum_{k=1}^m h_n(\eta_k)) \to \mathfrak{L}(U)$$

where the infinitely divisible distribution $\mathfrak{L}(U)$ has no normal component then

$$\mathfrak{L}(\sum_{k=1}^m h_n(S_k)) \to \mathfrak{L}(U) .$$

Proof. Set

$$U_n = \sum_{k=1}^m h_n(\eta_k)$$

and

$$V_n = \sum_{k=1}^m (\eta_k - 1/r)r/((r+1)m)^{1/2} .$$

As $\mathfrak{L}(U_n) \to \mathfrak{L}(U)$ and $\mathfrak{L}(V_n) \to N(0,1)$, we can select from any subsequence of $\mathfrak{L}(U_n, V_n)$ a convergent subsequence $\mathfrak{L}(U_{n'}, V_{n'})$ using Helly's theorem. Thus by Theorem 4

$$E(e^{itU_{n'}+isV_{n'}}) \to G(t) \cdot e^{-(At^2+2Bts+s^2)/2} .$$

But U has no normal component so $A = B = 0$ and $E(e^{itU}) = G(t)$. As the limit is the same for any subsequence it follows that $\mathcal{L}(U_n, V_n)$ converges. By Theorem 4 the assertion follows. ∎

In the remaining part of this section we consider the statistic

$$h(S_1, \ldots, S_m) = \sum_{k=1}^{m} h(S_k) ,$$

for a fixed function $h(\cdot)$, which does not depend on n. We also suppose that $m, n \to \infty$ so that $m/n = r = \rho$ is fixed. Let η as in Section 2 be a geometric random variable with mean $1/\rho$. We use the notation

$$\mu = Eh(\eta)$$

$$\sigma_{11} = \mathrm{Var}(h(\eta))$$

$$\sigma_{12} = \mathrm{Cov}(h(\eta), \eta)$$

$$\sigma_{22} = \mathrm{Var}(\eta) = (\rho + 1)/\rho^2$$

$$\sigma_{1\cdot 2} = (\sigma_{11} - \sigma_{11}\sigma_{12}^2/\sigma_{22})^{1/2} .$$

The following local limit theorem holds.

Theorem 6. If $h(\eta)$ is an integer valued 1-lattice random variable with finite variance and $\sigma_{1\cdot 2} > 0$ then

$$m^{\frac{1}{2}} P(\sum_{1}^{m} h(S_k) = \nu) - (2\pi)^{-\frac{1}{2}} \cdot \sigma_{1\cdot 2}^{-1} \cdot \exp(-\frac{1}{2}(\nu - m\mu)^2/m\sigma_{1\cdot 2}^2) \to 0 ,$$

uniformly in ν when $m \to \infty$.

Proof. Using Theorem 1 with $M = m$ and Lemma 1 we get

$$E(e^{itZ_m}) = E(\exp(it\sum_1^m h(S_k))) = (m\sigma_{22}/2\pi)^{\frac{1}{2}} \cdot (1 + O(1/m)) \cdot$$

$$\cdot \int_{-\pi}^{\pi} E(\exp(it\sum_1^m h(\eta_k) + i\theta\sum_1^m \eta_k)) \cdot \exp(-in\theta)d\theta .$$

It is well-known that

$$P(Z_m = \nu) = (1/2\pi)\int_{-\pi}^{\pi} E(e^{itZ_m})e^{-i\nu t}dt .$$

Hence we obtain

$$P(Z_m = \nu) = (m\sigma_{22}2\pi)^{\frac{1}{2}}(1 + O(1/m)) \cdot$$

$$(1/2\pi)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} E(\exp(it\sum_1^m h(\eta_k) + i\theta\sum_1^m \eta_k)) \cdot$$

$$\cdot \exp(-i\nu t - in\theta)dt\, d\theta .$$

As above we have

$$P_m(\nu, n) = P(\sum_1^m h(\eta_k) = \nu, \sum_1^m \eta_k = n) =$$

$$= (1/2\pi)^2 \int_{-\pi}^{\pi}\int_{-\pi}^{\pi} E(\exp(it\sum_1^m h(\eta_k) + i\theta\sum_1^m \eta_k)) \cdot \exp(-i\nu t - in\theta)dt\, d\theta .$$

By a multi-dimensional local limit theorem for lattice distributions by Rvačeva, see Rvačeva (1954) theorem 6.1, we have uniformly in ν when $m \to \infty$,

$$mP_m(\nu, n) - (2\pi)^{-1}(\sigma_{11}\sigma_{22} - \sigma_{12}^2)^{-\frac{1}{2}} \cdot$$

$$\cdot \exp(-((\nu - m\mu)^2/\sigma_{11}m + 0 + 0)/2(1 - \sigma_{12}^2/\sigma_{11}\sigma_{22})) \to 0 ,$$

or using the results above

$$m^{\frac{1}{2}}P(Z_m = \nu) = (1 + O(1/m))(2\pi)^{-\frac{1}{2}} \cdot \sigma_{1\cdot 2}^{-1} \cdot$$

$$\cdot((\exp(-(\nu - m\mu)^2/2\sigma_{1\cdot 2}^2 m) + o(1)) =$$

$$= (2\pi)^{-\frac{1}{2}}\sigma_{1\cdot 2}^{-1}\exp(-(\nu - m\mu)^2/2\sigma_{1\cdot 2}^2 m) + o(1) ,$$

with $o(1)$ uniformly in ν. ∎

Remark. The above approach to limit theorems for $h(S_1,\ldots,S_m)$ transforms the problem to study essentially $h(\eta_1,\ldots,\eta_m)$, a function of independent random variables. Of course other theorems than those stated above could be derived in an analogous manner, e.g. using limit theorems for 2-dependence to study $\sum_k h(S_k, S_{k+1})$.

4. Some applications in nonparametrics.

Using Theorem 3 limit distributions, under the null hypothesis, for several nonparametric statistics for the two sample problem can be obtained; e.g. the Wilcoxon-test and the run-test is of the above type. Such applications are considered in Holst and Rao (1976). In that paper "close alternatives" are also studied. To obtain limit distributions under such alternatives seems to require a different method of proof than that used above. Applications of Theorems 4, 5 and 6 are given in the following examples.

Example 1. For the geometric distribution

$$P(\eta = j) = r/(r+1)^{j+1}, \qquad j = 0,1,2,\ldots$$

we have

$$P(\eta > a) = 1/(r+1)^{a}, \qquad a = 0,1,2,\ldots .$$

Letting $I(\cdot)$ be the indicator function we have

$$E \sum_{k=1}^{m} I(\eta_k > a_m) = m\, P(\eta > a_m) = m/(r+1)^{a_m} = \lambda_m$$

where

$$a_m = \log(m/\lambda_m)/\log(r+1) .$$

If $\lambda_m \to \lambda$, $0 < \lambda < \infty$, then it follows by the poisson approximation of the binomial distribution that

$$\mathcal{L}\Big(\sum_{k=1}^{m} I(\eta_k > a_m) \Big) \to Po(\lambda) .$$

By Theorem 5 it follows that

$$\mathcal{L}\Big(\sum_{k=1}^{m} I(S_k > a_m) \Big) \to Po(\lambda) .$$

From this we have

$$P\Big(\sum_{k=1}^{m} I(S_k > a_m) = 0\Big) \to e^{-\lambda} ,$$

which is the same as

$$P(\max(S_1,\ldots,S_m) \le a_m) \to e^{-\lambda} .$$

With $\lambda = e^{-x}$ this can be formulated as

$$P(\max(S_1,\ldots,S_m) \le (\log(m\, e^{x}))/\log(r+1)) =$$

$$= P(\log(r+1)\max(S_1,\ldots,S_m) - \log m \le x) \to e^{-e^{-x}} .$$

Thus we have proved that the random variable $\log((m+n)/n)\max(S_1,\ldots,S_m) - \log m$ converges in distribution to the usual extreme value distribution, cf. David and Barton (1962), p. 231, and Hill (1974), Theorem 1.

Example 2. Using Theorem 4 and similar calculations as in Example 1 we find that the statistics $\max(S_1,\ldots,S_m)$ and $2\sum_{k=1}^{m} I(S_k \neq 0)$ (i.e. essentially the number of runs in the combined sample) are asymptotically independent. The asymptotic distribution of $\max(S_1,\ldots,S_m)$ is given above. The asymptotic distribution of $2\sum_{k=1}^{m} I(S_k \neq 0)$ is the same as that for the run statistic or $N(2mn/(m+n),\ 4m^2n^2/(m+n)^3)$.

Example 3. In Holst and Rao (1976) it is proved that among statistics of the form considered in Section 3 the Dixon-statistic

$$Z_m = \sum_{1}^{m} \binom{S_k}{2}$$

is in a certain sense optimal. After some elementary calculations we find

$$\mu = E\left(\left(\begin{matrix}\eta\\2\end{matrix}\right)\right) = 1/\rho^2$$

$$\sigma^2_{1\cdot 2} = \mathrm{Var}\left(\left(\begin{matrix}\eta\\2\end{matrix}\right)\right) - \left(\mathrm{Cov}\left(\left(\begin{matrix}\eta\\2\end{matrix}\right), \eta\right)\right)^2 / \mathrm{Var}(\eta) = (1+\rho)/\rho^2 .$$

As $\left(\begin{matrix}\eta\\2\end{matrix}\right)$ is integer valued and takes the values 0 and 1 with positive probability, the assumptions of Theorem 6 are fulfilled and therefore

$$m^{\frac{1}{2}} P\left\{\sum_1^m \left(\begin{matrix}S_k\\2\end{matrix}\right) = \nu\right\} - (\rho^2/(1+\rho)2\pi)^{\frac{1}{2}} \cdot \exp(-(\nu - m/\rho^2)^2 \rho^2/2m(1+\rho)) \to 0 ,$$

uniformly in ν when $m \to \infty$.

REFERENCES

[1] David, F. N. and Barton, D. E. (1962). Combinatorial Chance. Griffin, London.

[2] Feller, W. (1968). An Introduction to Probability Theory and its Applications, 1, 3rd. ed., John Wiley and Sons, New York.

[3] Hill, B. M. (1974). The rank-frequency form of Zipf's law. J. Amer. Statist. Assoc. 69, 1017-1026.

[4] Holst, L. and Rao, J. S. (1976). Asymptotic theory for some families of two-sample nonparametric statistics, unpublished.

[5] Le Cam, L. (1958). Un Théorème sur la division d'un intervalle par des points pris au hasard. Publ. Inst. Statist. Univ. Paris 7, 7-16.

[6] Park, C. J. (1973). The distribution of frequency counts of the geometric distribution. Sankhya Ser. A. 35, 106-111.

[7] Rao, C. R. (1973). Linear Statistical Inference and its Applications, 2nd ed., John Wiley and Sons, New York.

[8] Rvačeva, E. L. (1954). On domains of attraction of multi-dimensional distributions. Select. Transl. Math. Statist. and Probability 2, 183-205.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: 1660	2. GOVT ACCESSION NO. (9)
3. RECIPIENT'S CATALOG NUMBER: Technical
4. TITLE (and Subtitle): SOME LIMIT THEOREMS FOR THE INDISTINGUISHABLE BALL PROBLEM WITH APPLICATIONS IN NONPARAMETRICS.	5. TYPE OF REPORT & PERIOD COVERED: Summary Report - no specific reporting period
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s): (10) Lars Holst	8. CONTRACT OR GRANT NUMBER(s): (15) DAAG29-75-C-0024
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Mathematics Research Center, University of Wisconsin, 610 Walnut Street, Madison, Wisconsin 53706	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: U. S. Army Research Office, P.O. Box 12211, Research Triangle Park, North Carolina 27709 (11)	12. REPORT DATE: August 1976
	13. NUMBER OF PAGES: 17
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office): (14) MRC-TSR-1660	15. SECURITY CLASS. (of this report): UNCLASSIFIED (12) 20 p.
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Indistinguishable ball problem; Bose-Einstein statistics; urn models; occupancy; spacings; limit theorems; combinatorics; geometric distribution; nonparametrics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In Feller (1968), An Introduction to Probability Theory and Its Applications, Vol. 1, the following urn model is discussed. Consider m urns and distribute n indistinguishable balls among the urns such that the distinguishable distributions of the balls all have the same probability, $1/\binom{n+m-1}{m-1}$. Let S_k denote the number of balls in the k^{th} urn. Clearly $S_1 + \ldots + S_m = n$. In this paper random variables of the type $Z = h(S_1, \ldots, S_m)$, especially $h(S_1, \ldots, S_m) = h_1(S_1) + \ldots + h_m(S_m)$, are studied when $m, n \to \infty$ in such a way $m/n \to \rho$, $0 < \rho < \infty$. Some applications of the results in nonparametric statistics are briefly discussed and the limit distribution of $\max(S_1, \ldots, S_m)$ is derived.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

221200

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)