CONTROL SYSTEMS LABORATORY

APPROXIMATE DISTRIBUTIONS OF SAMPLE INFORMATION FOR USE IN ESTIMATING TRUE INFORMATION BY CONFIDENCE INTERVALS

Report Number R-76

1956

Contract DA-36-039-SC-56695
Project 8-103A, D/A Project 3-99-10-101

UNIVERSITY OF ILLINOIS · URBANA · ILLINOIS

"The research reported in this document was made possible by support extended to the University of Illinois, Control Systems Laboratory, jointly by the Department of the Army (Signal Corps and Ordnance Corps), Department of the Navy (Office of Naval Research), and the Department of the Air Force (Office of Scientific Research, Air Research, and Development Command), under Signal Corps Contract DA-36-039-SC-56695, Project 8-103A, D/A Project 3-99-10-101."

Report Number R-76

APPROXIMATE DISTRIBUTIONS OF SAMPLE INFORMATION FOR USE IN ESTIMATING TRUE INFORMATION BY CONFIDENCE INTERVALS

Prepared by:

H.T. David and W. H. Kruskal
Statistical Research Center
University of Chicago
Chicago, Illinois

and

L. Augenstine and H. Quastler
Control Systems Laboratory
University of Illinois
Urbana, Illinois

1956

CONTROL SYSTEMS LABORATORY
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS
Contract DA-36-039-SC-56695

Numbered Pages: 37

TABLE OF CONTENTS

I. INTRODUCTION

Information measures are defined in terms of certain probability distributions. In practice, we can only take samples from such distributions, and estimate the true information measures from the samples. In telecommunication samples of virtually infinite size can be obtained in a msec. or so. But in many other situations where one wishes to apply information measures it is either impracticable or impossible to obtain very large samples from a constant source. In such situations, one must make the best use of a limited number of observations when evaluating the information measures. Thus, the application of information theory to small-sample situations depends upon the development of an appropriate small-sample distribution theory. One problem that has been attacked by several authors[1,2,3,4] is that of finding unbiased or nearly unbiased estimates of the true information functions. The stochastic model generally used in this context is one in which sample frequencies are generated by one or several multinominal samplings. It can be shown that no truly unbiased estimates of information functions exist in this situation. Accordingly, the search had to be restricted to nearly unbiased estimates.

The population functions we considered below are exemplified by H, T, and A:

$$(1.1)\quad H = -\sum_i p_i \log p_i$$

H is the uncertainty associated with the set of probabilities, p_i, representing the states of a single variable.

$$(1.2)\quad T = \sum_{ij} p_{ij} \log_2 \frac{p_{ij}}{p_i \cdot p_j}$$

where $p_i = \sum_j p_{ij}$. This can be written as

$$(1.2a)\quad T = H(i) + H(j) - H(i,j)$$

where H(i) is the uncertainty about the ith variate (input), H_j the uncertainty about the jth variate (output) and H(i,j) the joint uncertainty about both the variates i and j. T is the measure of transmission in a single channel where the input-output relationships can be represented by a 2-dimensional matrix of probabilities, p_{ij}.

$$(1.3)\quad A = \sum_{ijk} p_{ijk} \log_2 \frac{p_{ijk} \cdot p_i \cdot p_j \cdot p_k}{p_{ij} \cdot p_{ik} \cdot p_{jk}}$$

where $p_i = \sum_{jk} p_{ijk}$ and $p_{ij} = \sum_k p_{ijk}$

$$(1.3a)\quad A = - H(i) - H(j) - H(k) + H(i,j) + H(i,k) + H(j,k) - H(i,j,k).$$

A, which is an interaction term, is the difference in transmission between two variates due to having or not having knowledge of a third variate. A is defined by a set of probabilities, p_{ijk}, which populate a 3-dimensional matrix representing the input-output probabilities for multi-channel transmission.

The most natural way of estimating information functions is to use the observed sample frequencies as estimates of the population frequencies and evaluate the functions from them. These estimates will be biased, i.e., their average values for a large number of repititions of the sampling procedure are not the same as the population value. In particular, H will be negatively biased, (i.e., it will be too small) and T will be positively biased, (too large). The size of the bias can be approximated according to

Miller and Madow[1] by:

$$(1.4)\quad \text{Bias} \doteq \pm \frac{\text{degrees of freedom}}{1.3863\ n}$$

where "degrees of freedom" is a quantity depending on the number of categories, n is the sample size, and the sign depends on the function to be estimated. It turns out, incidentally, that this bias term is very effective in locating the mean and the principal mode of the sample distribution.

However, relation (1.4) does not apply to the situation where all probabilities are equal. Rogers and Green[2] have developed an expression which gives the exact value of the bias of H for the equiprobable case. Good[3] has presented an alternative estimation of H which is almost unbiased and is valuable in the case where there are many classes sparsely populated. His estimator does not use the sample frequencies as estimators of the population frequencies. Another method also based on a different estimation of population frequencies is being developed by Blyth[4].

All of these methods deal with the problem of point estimation, i.e., of obtaining a single number to represent the "best" estimate of an information function from a single sample. However, at the time we started our work, a satisfactory theory of confidence intervals did not exist. In most experimental work, the estimation of confidence intervals is more important than point estimation. This is true because the problem of whether the value of a functional derived from a given set of observations is or is not compatible with some theoretical value is the one most frequently encountered. To improve this situation we pooled the talents of three statisticians (from the Statistical Research Center, University of Chicago) two experimenters (from the Bio-Systems Group,

of the Control Systems Laboratory, University of Illinois) and one high speed digital computer (ILLIAC, University of Illinois).

It is possible to obtain a general idea of the variability of information functions in a given situation by using Monte Carlo methods. With the help of a high speed digital computer a sampling distribution for a given sample size and set of probabilities is readily constructed. We seriously considered using this method to construct a catalog of sampling distributions of information functions for all situations of interest. However, such a catalog would need to be prohibitive in size even if one were satisfied with a low precision. Therefore, we have concentrated on developing analytic methods for constructing confidence intervals. This means essentially finding a function of the sample frequencies which has the following properties: i) it can be inverted to yield a confidence statement and ii) the tails of its distribution have a predictable behavior.

The general method we used was to investigate a likely function with the help of Monte Carlo methods and used the results to suggest a more suitable function. To begin with, we oriented ourselves by generating distributions of the maximum likelihood estimators, $\hat{H}$, $\hat{T}$, $\hat{A}$, constructed by substituting the sample (i.e., Monte Carlo generated) frequencies, $\hat{p}_i$, for the population frequencies, p_i, in relations (1.1, 1.2, 1.3). The means of the distributions so generated confirmed the Miller-Madow estimate of the bias (relation 1.4). The tails of the distributions were usually right skewed, i.e., the right tail was larger than the left one. $\hat{H}$ was usually more skewed than $\hat{T}$, and $\hat{T}$ more than $\hat{A}$.

We next investigated a normalization of $\hat{H}$ which is distributed asymptotically unit normal in the sense of convergence in distribution.

$$(1.5)\quad K = \frac{\sqrt{n}\,(\sum_i \hat{p}_i \log_2 \hat{p}_i - \sum_i p_i \log_2 p_i)}{\left(\sum_i \hat{p}_i \log_2^2 \hat{p}_i - \left[\sum_i \hat{p}_i \log_2 \hat{p}_i\right]^2\right)^{1/2}}$$

With small samples this function is also usually skewed to the right. The skewness of K is linked to the skewness of $\hat{H}$ itself. Therefore, the next reasonable step seemed to be to work not with the normalized $\hat{H}$ itself, but with a normalized function of $\hat{H}$ designed such that the skewness is minimized. Such a function is the exponential: it turned out to be very successful except for cases with nearly equal probabilities.

In order to avoid an indiscriminate and time-consuming exploration of various functions of the e^x type, it was decided to look for a principle which could be used to pinpoint a desirable transformation. One such principle is variance-stabilization. This happily leads to an arcsine transformation which has the exponential property of a monotonically increasing slope.

Following these introductory remarks are three sections. Section II is a thorough exposition of certain considerations pertaining to the arcsine transformation and contains graphical summaries of representative results*. Section III is practical and contains the recipe for constructing confidence intervals using the arcsine transformation. Section IV is a collection of representative distributions. Section V is a summary.

---

* This section is based on a lecture given by H.T. David at a meeting of the American Institute of Mathematical Statistics, Ann Arbor, Michigan (1955).

II. THEORETICAL

We have been concerned with establishing computationally simple confidence intervals for a large class of information functions. Besides simplicity, our criterion has been that nominal and actual confidence levels be in good agreement. Relations (1.1, 1.2, 1.3) are some examples of information functions, as we will understand them here. The discussion will not concern Fisherian information or other related concepts. At the risk of perhaps being more abstruse than is necessary, we define exactly the type of function that will be dealt with; these functions are precisely those which, under certain assumptions discussed below, lead to an asymptotic variance term in eqn. (1) amenable to the arcsine transformation proposed here.

a) Let $(x_1, .., x_k)$ be a set of k arguments, and let π^α $[\alpha: 1, 2, .., A]$ be A partitions of this set, with elements π^α_β, all distinct. A function f of $(x_1 .., x_k)$ is of the information type if $f(x_1, .., x_k) = \sum_{i=1}^{k} x_i \phi_i(x_1, .., x_k)$, where $\phi_i(x_1, .., x_k) = \sum_{\alpha=1}^{A} C^\alpha_{\beta(i)} \log \left(\Sigma x_\nu / x_\nu \in \pi^\alpha_{\beta(i)} \right)$, with $\sum_{\alpha=1}^{A} C^\alpha_{\beta(i)}$ constant over i - $\pi^\alpha_{\beta(i)}$ stands here for the element of π^α containing x_i, and $C^\alpha_{\beta(i)}$ is a constant associated with $\pi^\alpha_{\beta(i)}$.

b) An information function $I(x_1, .., x_k)$ is a function of the information type whose arguments are restricted by $\Sigma x_i = 1$, $x_i \geqslant 0$.

c) Note that, for the information functions H, T and A defined above, the number of partitions π^α is, respectively 1, 3 and 7, and the $C^\alpha_{\beta(i)}$ are all equal to +1 or -1 (for the logarithmic base 2).

The arguments of the information function I (or, specifically, H, T, A, etc.) can be population probabilities p, in which case I is a population characteristic. Or they may be sampling fractions $\hat{p}$, in which case I is denoted by $\hat{I}$, (or, specifically, $\hat{H}$, $\hat{T}$, $\hat{A}$, etc.) the maximum likelihood estimate and a natural statistic for inference on its population counterpart.

This discussion is based on $\hat{I}$. Other statistics might also be useful, depending perhaps on the nature of the sampling process giving rise to the $\hat{p}$. Our assumption regarding this process has been that cells are filled by independent repetitions of a single multinominal trial encompassing the entire array. Some very useful prior work has already been done on the distribution problem for $\hat{I}$ under our assumption of simple or single multinomiality. First, Miller[5] has written $E(\hat{H})$ in the form

$$E(\hat{H}) = H + \frac{B_1}{n} + \frac{B_2}{n^2} + O\left(\frac{1}{n^3}\right)$$

where B_1 depends only on the number of categories and <u>not</u> on the population p's, and correspondingly for $E(\hat{T})$.

B_1 is equal to

$$B_{1,H} = -\left(\frac{k-1}{2n}\right)(\log_2 e)$$

for H, and is equal to

$$B_{1,T} = \left(\frac{(r-1)(c-1)}{2n}\right)(\log_2 e)$$

for T. B_1 generalizes immediately for general I, by the linearity of expectation.

Miller also pointed out that $\hat{T}$ is a multiple of the log of the

likelihood ratio for testing independence in a bi-viriate array, hence is distributed assymptotically as a multiple of χ^2 under independence, which is equivalent to T = 0. Further, Madow and Miller[1] wrote Var $(\hat{H})$ in the form

$$\text{Var}\,(\hat{H}) = \frac{\Sigma\, p_i \left[\log_2 p_i\right]^2 - H^2}{n} + \frac{A_2}{n^2} + \frac{A_3}{n^3} + O\left(\frac{1}{n^4}\right)$$

and pointed out that $\hat{H}$, suitably normed, tends in distribution to the normal when the population p's are not all equal, and to central χ^2 when they are. Other very useful work has been done by I.J. Good[3], who developed an almost unbiased estimate of H.

Concerning the distribution of the general $\hat{I}$, we can draw on the easily verified fact that, for any well-behaved function G of (k-1) k-nomial probabilities p_i,

$$(2.1)\quad \sqrt{n}\left[G(\hat{p}_i) - G(p_i)\right] \xrightarrow{D} N\left[0,\ \sum_{i=1}^{k-1} p_i\,(G^{(i)})^2 - \left(\sum_{i=1}^{k-1} p_i\, G^{(i)}\right)^2\right]$$

where the symbol $\xrightarrow{D}$ means "tends in distribution", and $G^{(i)}$ is the partial derivative of G with respect to p_i. This is the central fact on which the following discussion is based. (Another possibly very useful fact, mentioned to us by S. Kullback, but not explored here, is that $I(\hat{p})$, at least for all its usual specializations, tends to normality vis non-central χ^2-ness).

For information functions, the asymptotic variance term in (2.1) becomes

$$\sum_{i=1}^{k-1} p_i (I^{(i)})^2 - \left(\sum_{i=1}^{k-1} p_i\, I^{(i)}\right)^2 = \sum_{i=1}^{k} p_i\, \phi_i^2 - \left(\sum_{i=1}^{k} p_i\, \phi_i\right)^2 = \sum_{i=1}^{k} p_i\, \phi_i^2 - I^2 = V(p_i).$$

Though it doesn't constitute an essential link in the present arguemnt, it seems of some interest to locate the information functions in the larger domain of functions $G(p_1, .., p_k)$, $\Sigma p_i = 1$, $= \sum_{i=1}^{k} p_i g_i (p_1, .., p_k)$ which admit a reduction of the asymptotic variance expression similar to that possible for information functions. To this end we can make three remarks:

1) Let $\lambda_i = \sum_{\alpha=1}^{k} p_\alpha g_\alpha(i)$, where the superscript denotes differentiation with respect to the indicated argument of $g_\alpha (p_1, .., p_k)$.

Then $\sum_{i=1}^{k-1} p_i (G^{(i)})^2 - (\sum_{i=1}^{k-1} p_i G^{(i)})^2 = \sum_{i=1}^{k} p_i g_i^2 - G^2$ if and only if

$$\sum_{i=1}^{k-1} p_i \left[(g_i-g_k) + (\lambda_i-\lambda_k)\right]^2 - \left\{\sum_{i=1}^{k-1} p_i \left[(g_i-g_k) + (\nu_i-\nu_k)\right]\right\}^2 = \sum_{i=1}^{k-1} p_i (g_i-g_k)^2 -$$

$$\left[\sum_{i=1}^{k-1} p_i (g_i - g_k)\right]^2$$

2)It is clearly sufficient for the condition of 1) that

$$\sum_{i=1}^{k} x_i \frac{\partial}{\partial x_j} g_i(x_1, .., x_k) = \sum_{i=1}^{k} x_i \frac{\partial}{\partial x_m} g_i(x_1, .., x_k) \text{ identically,}$$

for j, m: 1, 2, ..., k and arbitrary unrestricted arguments x_i.

3) Let $\pi^\alpha \alpha$:1, 2, .., A be defined as in the above definition of a function of the information type, and let $g_i (x_1, .., x_k) =$

$\sum_{\alpha=1}^{A} {}^{\alpha}_{\beta(i)}(x_1,..,x_k)$, where the arguments of ${}^{\alpha}_{\beta(i)}$ are exactly the x's in the partition element $\pi^{\alpha}_{\beta(i)}$ containing x_i. Then, if $\psi^{\alpha}_{\beta(i)}$ depends on its arguments only through their sum, the condition of 2) is equivalent to "G is an I".

Returning now to the main discussion, we will focus attention on two attractive and available first order terms: Miller's bias correction B_1/n, and the first order variance term $V(p_1)$. B_1/n deserves to be called attractive, if accurate, because it does not depend on the population p's. The variance term if accurate, also has several virtues, not the least of which is simplicity; some of the others are given below.

How accurate are these two terms? B_1/n has been tested by Madow and Miller, and also by us, intensively for H, and also for T and A. It appears to be surprisingly good; it is almost more effective as an estimate of Med $(\hat{I})$ - I than as an estimate of $E(\hat{I})$ - I, which seems not at all undesirable for our purposes, since a confidence interval procedure pertains to quantile rather than moment estimation. Also, it appears to be literally better than the supposedly finer estimate $B_1/n + B_2/n^2$.

The variance term has been less extensively investigated, but the results we do have are again gratifying. We considered the case of H for n = 31 and three categories, with class probabilities belonging to the one-parameter set

$$p,\ 1/2 - p/2,\ 1/2 - p/2$$

We then plotted against p both the actual stand. dev. of $\hat{H}$, as estimated by Illiac on the basis of 200 samples, and also the approximation $V(p)/n$. The two curves show excellent agreement.

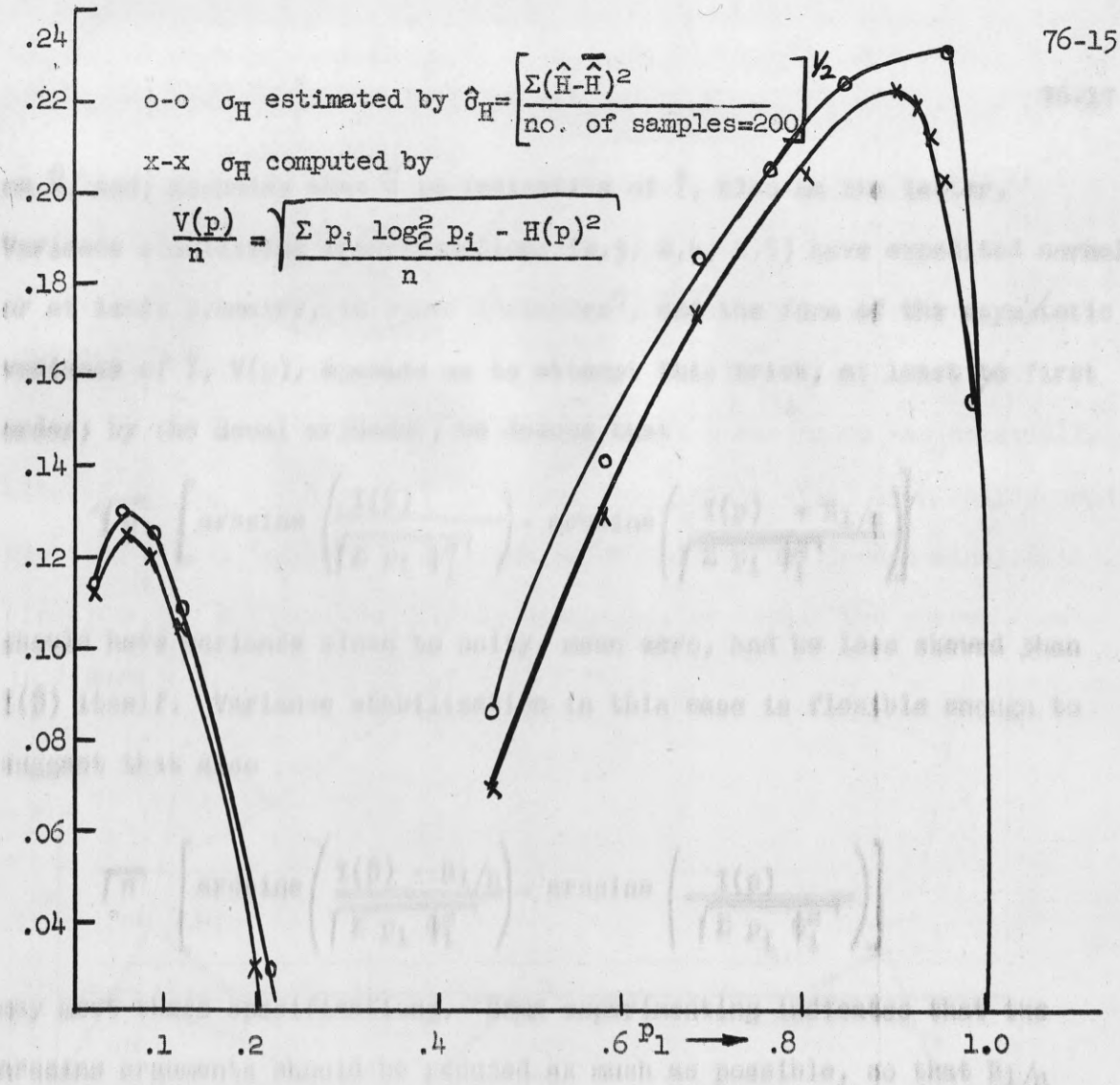


Fig. 2.1 Standard deviation of H for a trinomial distribution ($p_2=p_3=\frac{1-p_1}{2}$)

It seems reasonable, therefore, inducing from these specific investigations to the general I, to begin by looking at confidence intervals based on $\frac{B1}{n}$ and V(p). The first statistic that suggests itself is

$$(2.2)\quad \frac{\sqrt{n}\,(I(\hat{p}) - \left[I(p) + \frac{B_1}{n}\right]}{\left[\Sigma\, \hat{p}_i\, \phi_i^2 - I^2(\hat{p})\right]^{1/2}}$$

If one computes by Illiac (200 sample runs) the actual size of the nominal upper and lower 5 and 10% tails of this statistic, for H with the probability set (p_1 1/2-p/2, 1/2-p/2) given above, one obtains (again plotting against p).

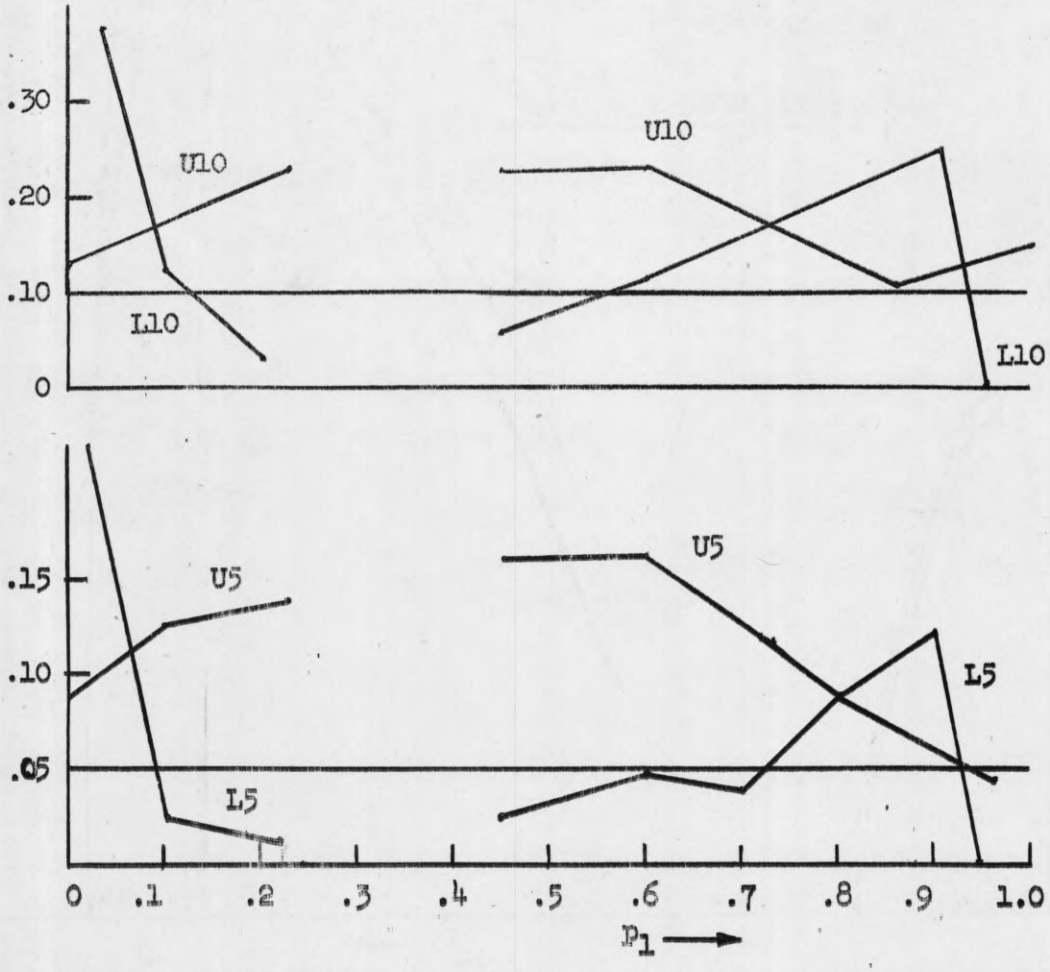


Fig. 2.2 5 and 10% limits of the distribution of expression (2.2) for a trinomial distribution where $p_2 = p_3 = \frac{1-p}{2}$

The results are disappointing, partly because the distribution of the statistic is skewed. This skewness seems to be due primarily to the numerator, that is $\hat{p}$ itself, when p is small or large, and to the denominator, that is $V(\hat{p})^{1/2}$, in the middle region. In consequence, on might expect that replacing $H(\hat{p})$ by $H(p)$ in the denominator would improve the picture in the center. This substitution yields a statistic almost as useful for interval construction as the original. However, things are not improved.

Since the skewness of $\hat{p}$ appears as partly responsible for the mediocre behavior of these first expressions, one is led to consider transformations

on $\hat{H}$, and, assuming that $\hat{H}$ is indicative of $\hat{I}$, also on the latter. Variance stabilizing transformations (2.3, 2.4, 2.5) have expedited normality, or at least symmetry, in other instances[6], and the form of the asymptotic variance of $\hat{I}$, $V(p)$, enables us to attempt this trick, at least to first order; by the usual argument, we deduce that

$$\sqrt{n}\left[\arcsin\left(\frac{I(\hat{p})}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right) - \arcsin\left(\frac{I(p) + B_1/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right)\right]$$

should have variance close to unity, mean zero, and be less skewed than $I(\hat{p})$ itself. Variance stabilization in this case is flexible enough to suggest that also

$$\sqrt{n}\left[\arcsin\left(\frac{I(\hat{p}) - B_1/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right) - \arcsin\left(\frac{I(p)}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right)\right]$$

may meet these specifications. Some experimenting indicates that the arcsine arguments should be reduced as much as possible, so that B_1/n should appear in the first term when it is positive, and in the second when it is negative. This empirical fact may be introduced into the expression by writing

$$(2.3)\quad \sqrt{n}\left[\arcsin\left(\frac{I(\hat{p}) - B_1^+/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right) - \arcsin\left(\frac{I(p) + B_1^-/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right)\right]$$

Expression (2.4) will denote expression (2.3) with the sample estimate $\sqrt{\Sigma\, \hat{p}_i\, \phi_i^2}$ introduced in the first arcsine term

$$(2.4)\quad \sqrt{n}\left[\arcsin\left(\frac{I(\hat{p}) - B_1^+/n}{\sqrt{\Sigma\, \hat{p}_i\, \phi_i^2}}\right) - \arcsin\left(\frac{I(p) + B_1^-/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right)\right].$$

Expression (2.5) will be expression (2.4) with the sample estimate $\sqrt{\Sigma\, p_i\, \phi_i^2}$ also introduced in the second arcsine term

$$(2.5)\quad \sqrt{n}\left[\text{arcsine}\left(\frac{I(\hat{p}) - B_1^{+}/n}{\sqrt{\Sigma\, \hat{p}_i\, \hat{\phi}_i^2}}\right) - \text{arcsine}\left(\frac{I(p) + B_1^{-}/n}{\sqrt{\Sigma\, \hat{p}_i\, \hat{\phi}_i^2}}\right)\right]$$

Expressions (2.3), (2.4) and (2.5) all are asymptotically normal, since they are well-behaved functions of multinomial probabilities. They further should have mean approximately zero, and variance approximately 1. Expression (2.3) will conform to these specifications best. However, it is impossible to construct simple confidence intervals on the basis of it. It does so happen that, for H with three categories, the equicontours of

$$\text{arcsine}\left(\frac{C - B_1^{+}/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right) - \text{arcsine}\left(\frac{H(p) + B_1^{-}/n}{\sqrt{\Sigma\, p_i\, \phi_i^2}}\right)$$

do seem to follow fairly closely the equicontours of H(p) itself, on the triangle in which the population p's take values, so that a fairly tight confidence interval might be obtainable for H by way of these contours. However, one of our goals was simplicity, and it seems doubtful whether it could be achieved here; hence we proceed to expression (2.4).

Expression (2.4) is especially suitable for establishing confidence intervals for $I(p)/\sqrt{\Sigma\, p\ \phi_i^2}$ and hence for the coefficient of variation, since these two population parameters are monotonically related, at least to first order. However, as regards intervals for I(p), the only possibility seems again to lie in contour matching. Quite unlike the case of expression (2.3) this turns out to be unfeasible. For example, in the case of H, $H/\sqrt{\Sigma}$ is equal to unity in the middle of all faces and edges of the simplex in

which the p's take values, whereas the value of H itself drops off monotonically from the center along every ray.

Our real hope for a simple confidence interval is therefore expression (2.5). The question is whether the insertion of the two sample quantitives in place of their population analogous has materially altered the distribution; it turns out that things still look fairly good. Plotting actual (computed by Illiac - 100 sample runs) and nominal tail sizes for the H situation already discussed, we obtain the curves shown in figure 2.3.

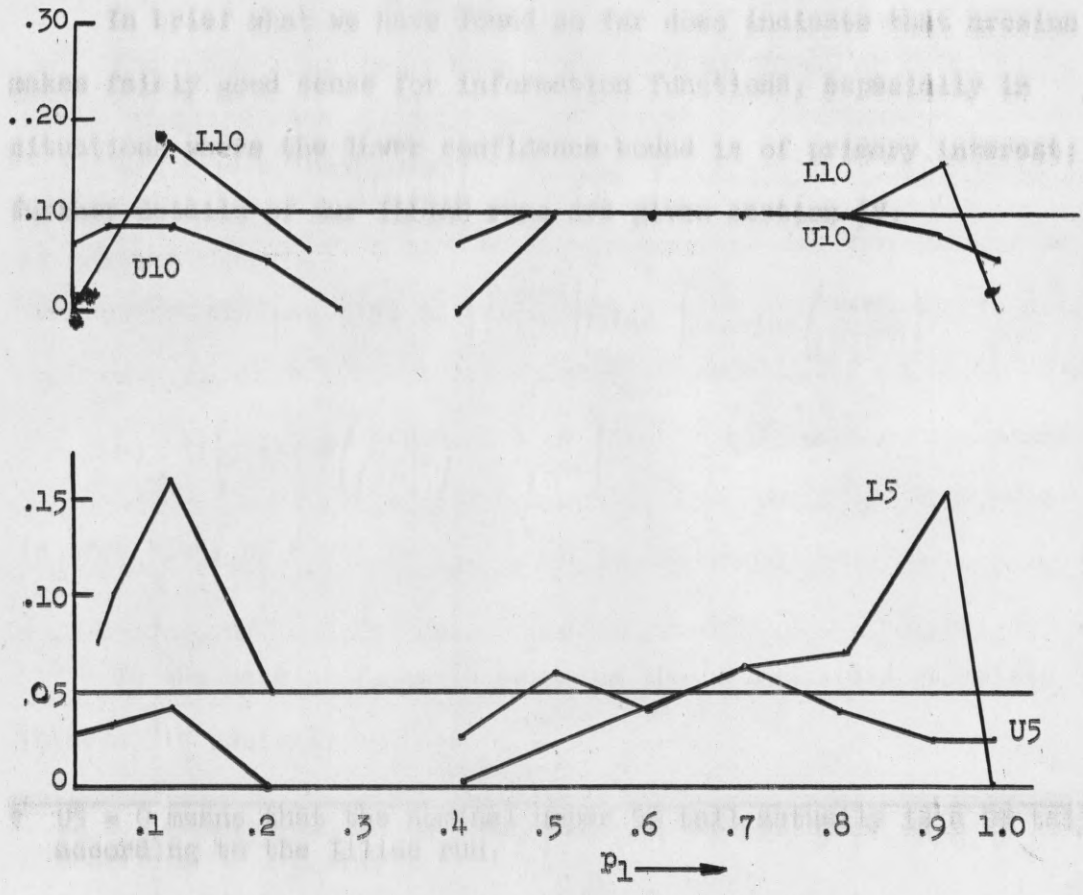


Fig. 2.3 5 and 10% limits of the distribution of expression (2.5) for a trinomial distribution where $p_2 = p_3 = \frac{1-p_1}{2}$

What these graphs show first is that the arcsine may be expected to give good lower confidence bounds (corresponding to the U curves) for H, which, if not exact, are at least conservative. The L curves are also pretty good except for the two symmetrically placed peaks which, both for .05 and .10, seem to reach to approximately .17, although we haven't checked the entire neighborhood. These peaks are quite reminiscent of what has been found by Eisenhart in[9] for the Binomial. If they could be eliminated, we would be in good shape, at least as far as H is concerned. We have tried to cut them down in several ways, for example, by a device analogous to the Freeman-Tukey correction for the binomial arcsine and Poisson square root[7]; this did not work out. We have had somewhat more success reassigning the sample fraction zero to rare classes (quite analogously to what has been suggested for the Binomial in[8] and[9]) but this hasn't been entirely satisfactory either.

We've been curious, of course, about the performance of the arcsine in situations other than H with three categories. We've applied expression (2.5) to H with five categories and n = 31, avoiding very low probabilities. The same pattern manifests itself: the upper tail is conservative, and the lower tail is not far from exact.

Our studies of $\hat{T}$ are fragmentary at the moment; the following are typical results for a low, intermediate and high value of population T, all with n = 31.

Figure IV

.133	.100	.100
.100	.133	.100
.100	.100	.133

T = .01

U5 = 0*
U10 = .04
L10 = .13
L5 = .07

.267	.033	.033
.033	.267	.033
.033	.033	.267

T = .67

U5 = .05
U10 = .09
L10 = .16
L5 = .13

.301	.016	.016
.016	.301	.016
.016	.016	.301

T = 1.03

U5 = .07
U10 = .16
L10 = .07
L5 = .04

Things are quite hopeful especially in view of the good behavior of arcsine in the presence of low population probabilities.

In brief what we have found so far does indicate that arcsine makes fairly good sense for information functions, especially in situations where the lower confidence bound is of primary interest; further details of our ILLIAC runs are given section IV.

* U5 = 0 means that the nominal upper 5% tail actually is a 0% tail, according to the Illiac run.

III. CONSTRUCTION OF CONFIDENCE INTERVALS

We now give a "recipe" for the computation of a symmetric two-sided confidence interval for H and for T. This should illustrate how one-sided or two-sided confidence intervals can be computed in general.

First, fix a level of confidence (95%, say) and determine the corresponding symmetric unit normal deviates (+1.96 and -1.96 when the level is 95%).

For H, compute

$$H(\hat{p}) = -\sum_{i=1}^{k} \hat{p}_i \log_2 \hat{p}_i$$

and

$$S(\hat{p}) = \sum_{i=1}^{k} \hat{p}_i (\log_2 \hat{p}_i)^2 \quad ,$$

where

k : the number of classes (categories)

$\hat{p}_i$: the observed class frequencies ($\sum_i \hat{p}_i = 1$).

Now compute

$$\left[\text{arcsine } \frac{H(\hat{p})}{S(\hat{p})}\right] + \frac{1.96}{\sqrt{n}}$$

and

$$\left[\text{arcsine } \frac{H(\hat{p})}{S(\hat{p})}\right] - \frac{1.96}{\sqrt{n}}$$

(We are now assuming a 95% level of confidence), where arcsine is measured in radians and n is the number of trials (items classified).

The upper confidence bound, $\bar{H}$, for H is now computed as follows:

1) If $\left[\text{arcsine}\left(\frac{H(p)}{S(\hat{p})}\right) + \frac{1.96}{\sqrt{n}}\right]$

is less than $\pi/2$, compute

$$\frac{C(k)}{n} = \frac{\left[(k-1)/2\right]\left[\log_2 e\right]}{n}$$

and

$$F = \frac{C(k)}{n} + \left\{\sqrt{S(\hat{p})}\right\}\left\{\text{sine}\left[\arcsin\left(\frac{H(\hat{p})}{S(\hat{p})}\right) + \frac{1.96}{\sqrt{n}}\right]\right\}.$$

Then $\bar{H}$ is either F or $\log_2 k$, whichever is smaller,

2) If $\left[\arcsin\left(\frac{H(\hat{p})}{\sqrt{S(\hat{p})}}\right) + \frac{1.96}{\sqrt{n}}\right]$ is

equal to or greater than $\pi/2$, then

$$\bar{H} = \log_2 k$$

The lower confidence bound, $\underline{H}$, for H is computed as follows:

1) If $\left[\arcsin\left(\frac{H(\hat{p})}{\sqrt{S(\hat{p})}}\right) - \frac{1.96}{\sqrt{n}}\right]$

is greater than 0,

$$\underline{H} = \frac{C(k)}{n} + \left\{\sqrt{S(\hat{p})}\right\}\left\{\text{sine}\left[\arcsin\left(\frac{H(\hat{p})}{\sqrt{S(\hat{p})}}\right) - \frac{1.96}{\sqrt{n}}\right]\right\}$$

2) If $\left[\arcsin\left(\frac{H(\hat{p})}{\sqrt{S(\hat{p})}}\right) - \frac{1.96}{\sqrt{n}}\right]$

is less than, or equal to 0,

$$\underline{H} = 0$$

In the case of T, again assuming that a two-sided symmetric 95% interval is desired, compute

$$T(\hat{p}) = -\sum_{i=1}^{r}\sum_{j=1}^{c} \hat{p}_{ij}\left[\log_2 \hat{p}_{i.} + \log_2 \hat{p}_{.j} - \log_2 p_{ij}\right]$$

and

$$S_T(\hat{p}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \hat{p}_{ij} \left[\log_2 \hat{p}_{i.} + \log_2 \hat{p}_{.j} - \log_2 \hat{p}_{ij}\right]^2$$

where

r : the number of rows ("inputs")

c : the number of columns ("outputs")

$\hat{p}_{ij}$: the observed frequency of the simultaneous occurence of input i and output j ($\sum_{i=1}^{r} \sum_{j=1}^{c} \hat{p}_{ij} = 1$).

$\hat{p}_{i.}$: $\sum_{j=1}^{c} \hat{p}_{ij}$

$\hat{p}_{.j}$: $\sum_{i=1}^{r} \hat{p}_{ij}$.

Now compute

$$A = \left[\text{arcsine}\left(\frac{T(\hat{p}) - \frac{D(r,c)}{n}}{\sqrt{S_T(\hat{p})}}\right)\right] + \frac{1.96}{\sqrt{n}}$$

and

$$B = \left[\text{arcsine}\left(\frac{T(\hat{p}) - \frac{D(r,c)}{n}}{\sqrt{S_T(\hat{p})}}\right)\right] - \frac{1.96}{\sqrt{n}}$$

where

$$D(r,c) = \left[(r-1)(c-1)/2\right]\left[\log_2 e\right] ,$$

n is the number of items (input-output combinations) classified, and where arcsine, measured in radians, is defined as zero for negative arguments.

The upper confidence found, $\bar{T}$, for T is now computed as follows:

1) If A is less than $\pi/2$, compute

$$G = \left(\sqrt{S_T/\hat{p}}\right)\left(\sin A\right)$$

Then $\overline{T}$ is either G, or H(i) or H(j), whichever is smallest.

2) If $A \geq \frac{\pi}{2}$,

$\overline{T}$ = H(i) or H(j), whichever is smaller.

The lower confidence bound, $\underline{T}$, for T is computed in the following way:

1) If B is greater than zero,

$$\underline{T} = \left(\sqrt{S_T(\hat{p})}\right)\left(\sin B\right)$$

2) If $B \leq 0$,

$\underline{T} = 0$

Note that the computation of unsymmetric or one-sided intervals afford no additional difficulties of any kind. For example, the upper bounds $\overline{H}$ and $\overline{T}$ considered here are equally well interpreted as 97.5% upper bounds of one-sided intervals (i.e. intervals with lower bound identically zero).

IV. RESULTS

This section contains examples of the sampling distributions (cumulative distributions) of the various functions we have investigated. Since we were interested in normal approximation all distributions were plotted on normal probability paper. These distributions were constructed by a Monte Carlo method using ILLIAC. All of the functions tested had one feature in common; they involved a probability matrix. Thus, in all cases the Monte Carlo operations consisted of the generation of a large number of probability matrices from which values of the particular function being tested were then calculated.

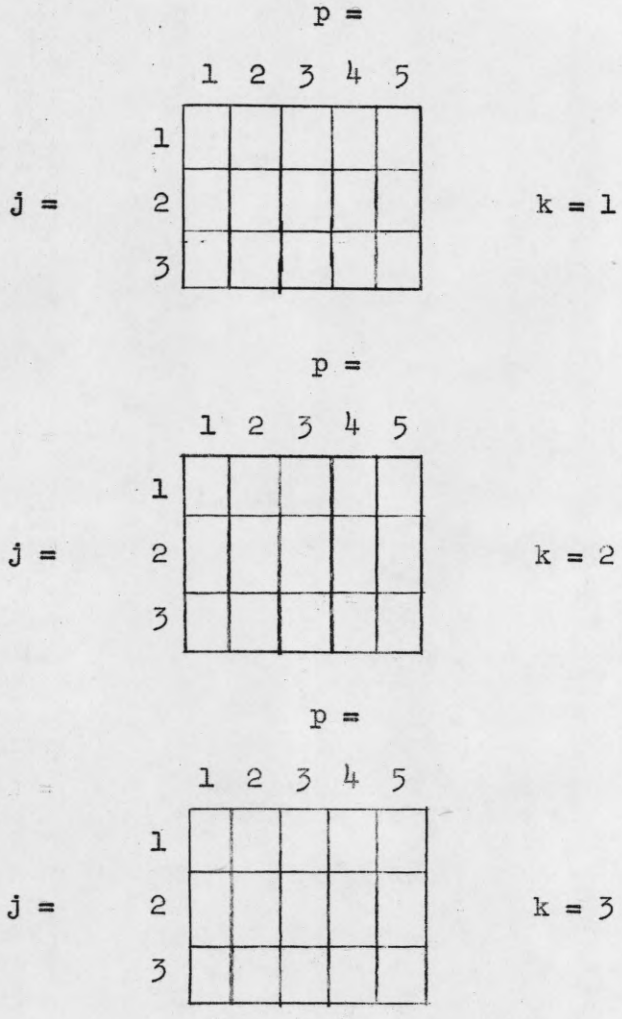
In practice a probability matrix, concerning the variates associated with the particular function, was read into ILLIAC. A set of n random numbers was generated and each random number was assigned to one of the matrix cells according to the probabilities which were specified in the input.* The resulting matrix cell population was converted into a set of relative frequencies, $\hat{p}_1, \hat{p}_2, \ldots\ldots \hat{p}_k$. This set formed a maximum likelihood estimate of the "true", i.e., input, probabilities, $p_1, p_2, \ldots . p_k$.

Values for the function being tested were then computed on the basis of the generated matrix, $\hat{p}_i$. Those values were estimators $f(\hat{p}_i)$ of the "true" values of the function, $f(p_i)$, computed from the input probabilities, p_i. This process was repeated s times and a distribution of the s estimates formed. The resulting distribution is one which could have been obtained with s experiments of n trials each, given a set of true probabilities, p_i. From these distributions it was possible to determine the mean, variance, normalcy, etc. for the input parameters specified.

---

* Any cell having an input probability of zero remained empty, i.e., that outcome was treated as impossible.

The computational procedures adopted are as follows: when a 3-dimensional matrix was input into ILLIAC the columns were given i-designations, the rows j-designations and the planes k-designations.



Distributions involving $\hat{H}$ were generated by filling the j = k = 1 row of the matrix with h random numbers according to the P_{ill} probability set; distributions involving $\hat{T}$ by filling the k = 1 plane with t random numbers according to the P_{ijl}'s; and $\hat{A}$ distributions by filling all of the matrix with a random numbers according to the P_{ijk}'s.

The distribution of information functions, for samples of limited size, are necessarily discontinuous. This accounts for the steps in the cumulative distributions observed in the figures in this section. The arrows in the figures indicate that increment which yields cumulative frequencies of 0 or 100.

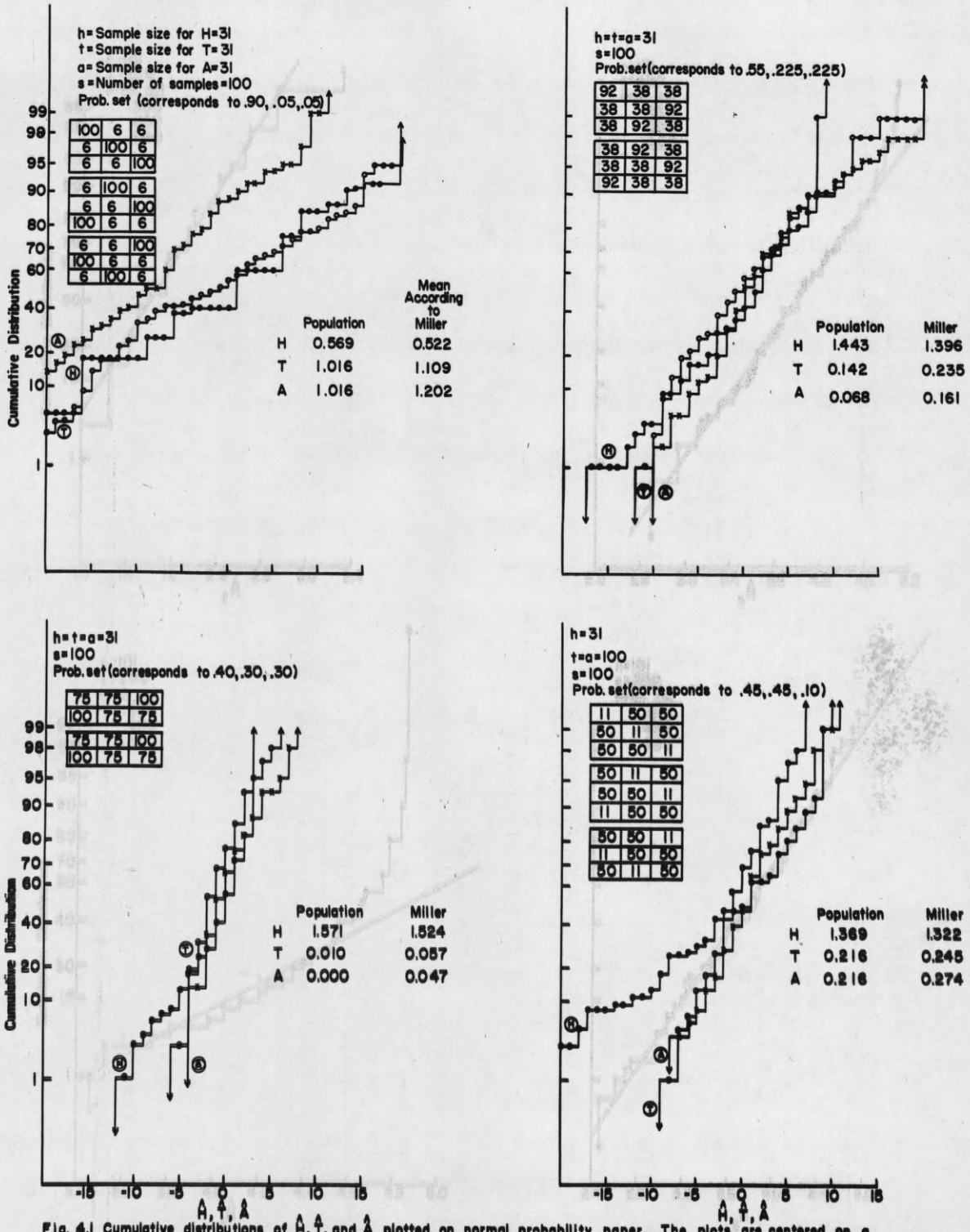


Fig. 4.1 Cumulative distributions of $\hat{H}$, $\hat{T}$, and $\hat{A}$ plotted on normal probability paper. The plots are centered on a mean which was estimated according to relation (1.4). The abscissal increments are 0.02 bits.

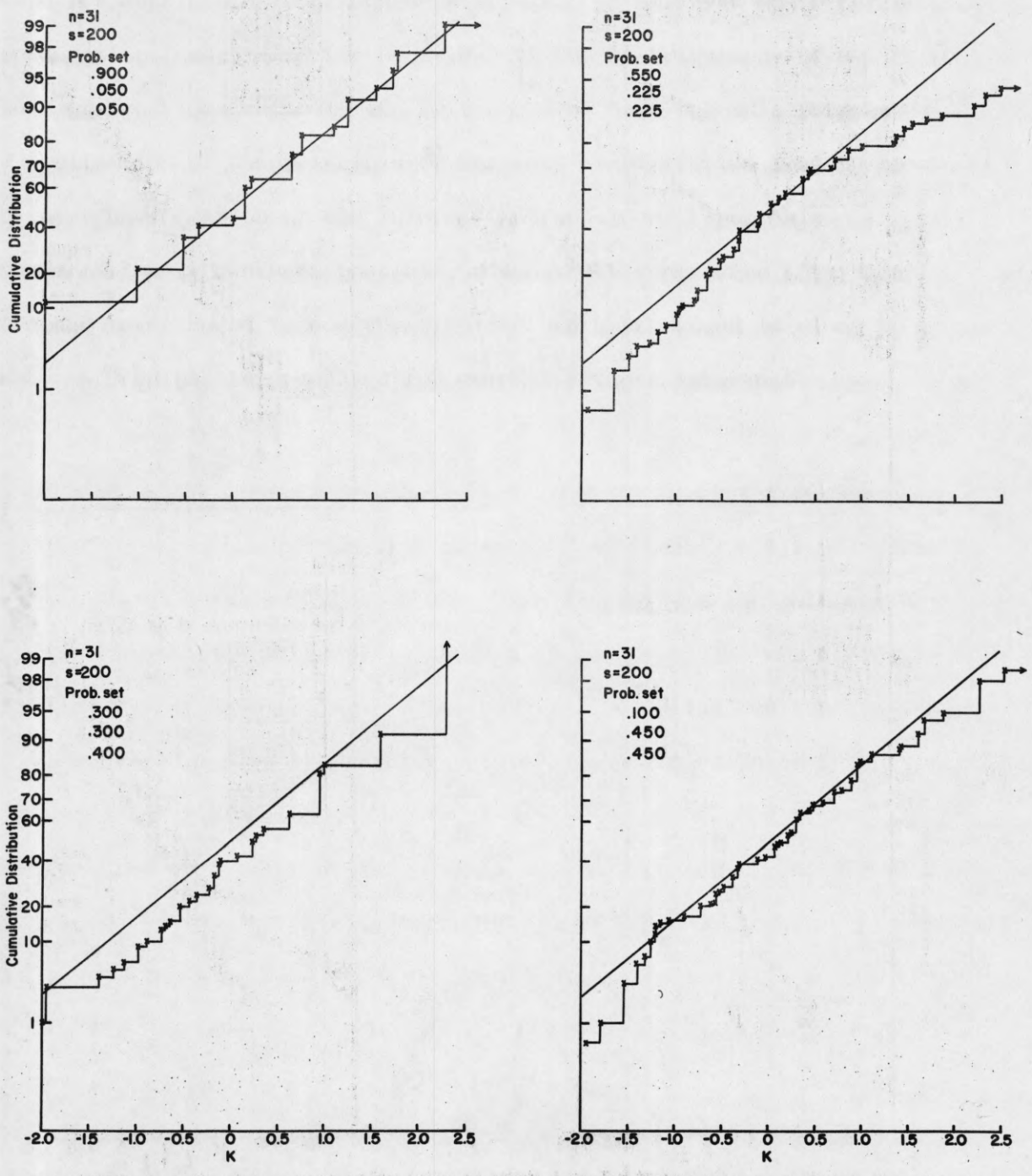


Fig. 4.2. Plots of the cumulative distribution for the function K defined in relation (1.5). The straight lines represent the unit normal distribution which K approaches asymptotically.

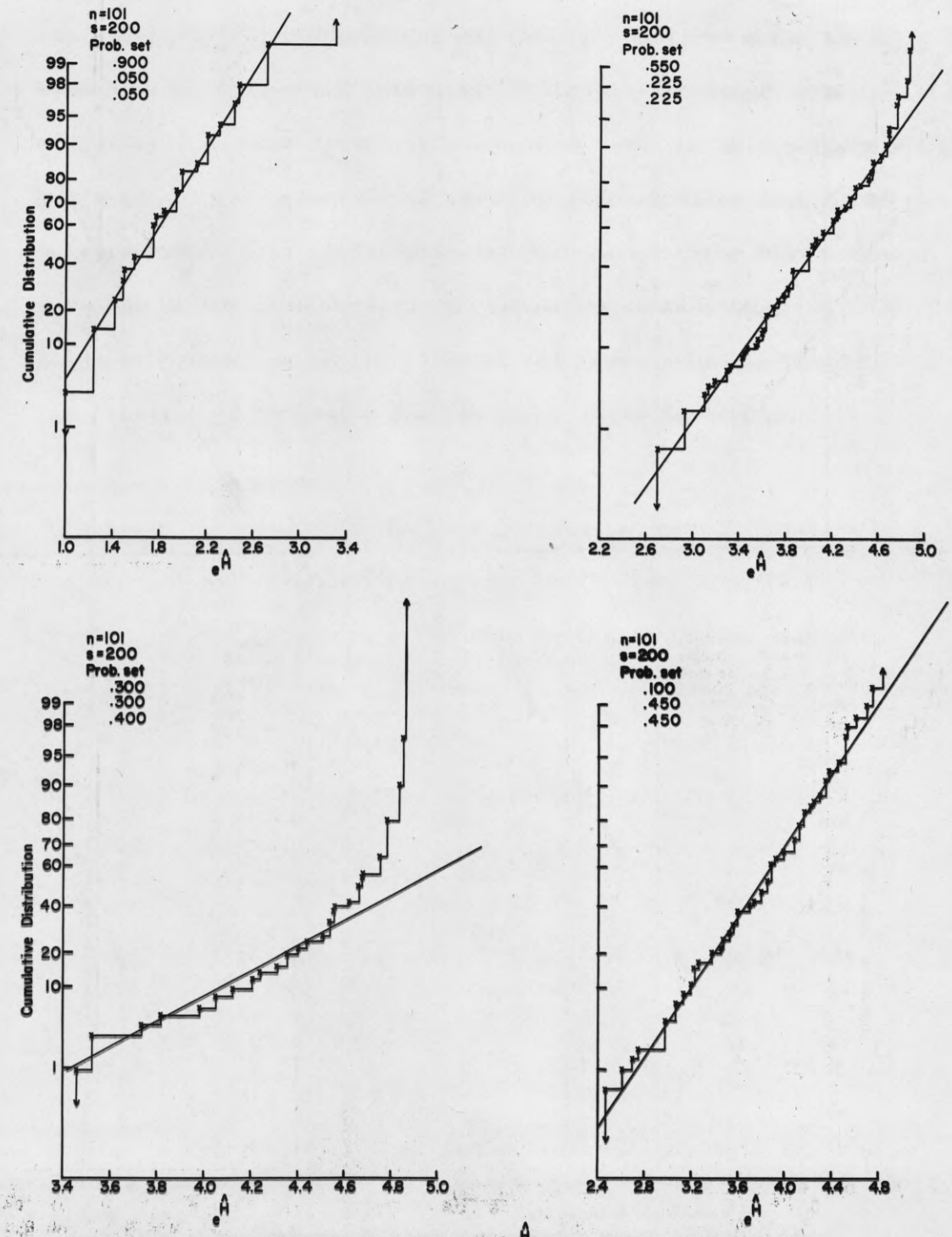


Fig. 4.3. Plots of the cumulative distribution of the function e^A. The straight lines are fitted by eye.

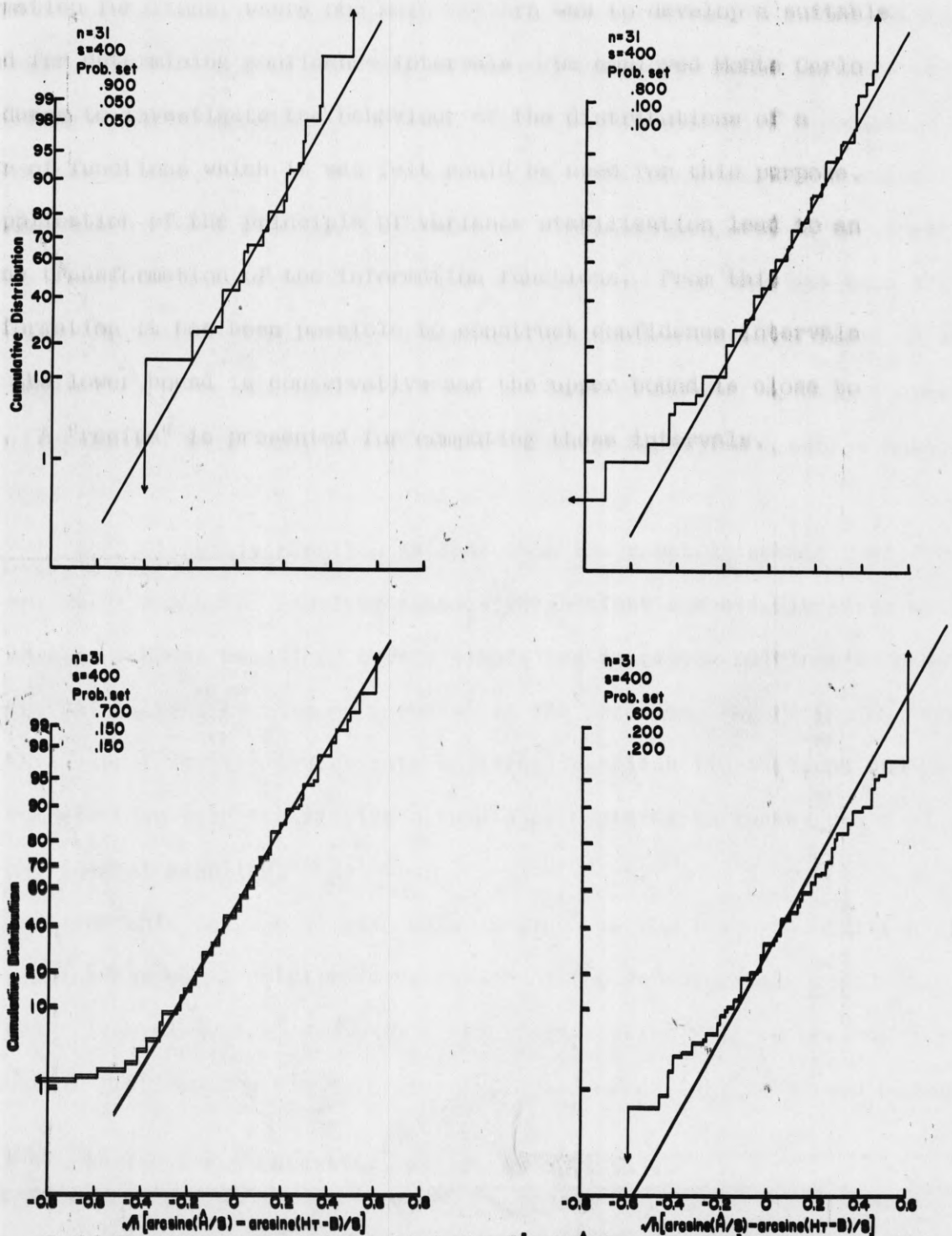


Fig.4.4. Plots of the cumulative distribution of the function $\sqrt{n}[\arcsin(\hat{A}/S) - \arcsin(H_T - B)/S]$ defined in relation (2.5). The straight lines represent the normal distribution which the function asymptotically approaches.

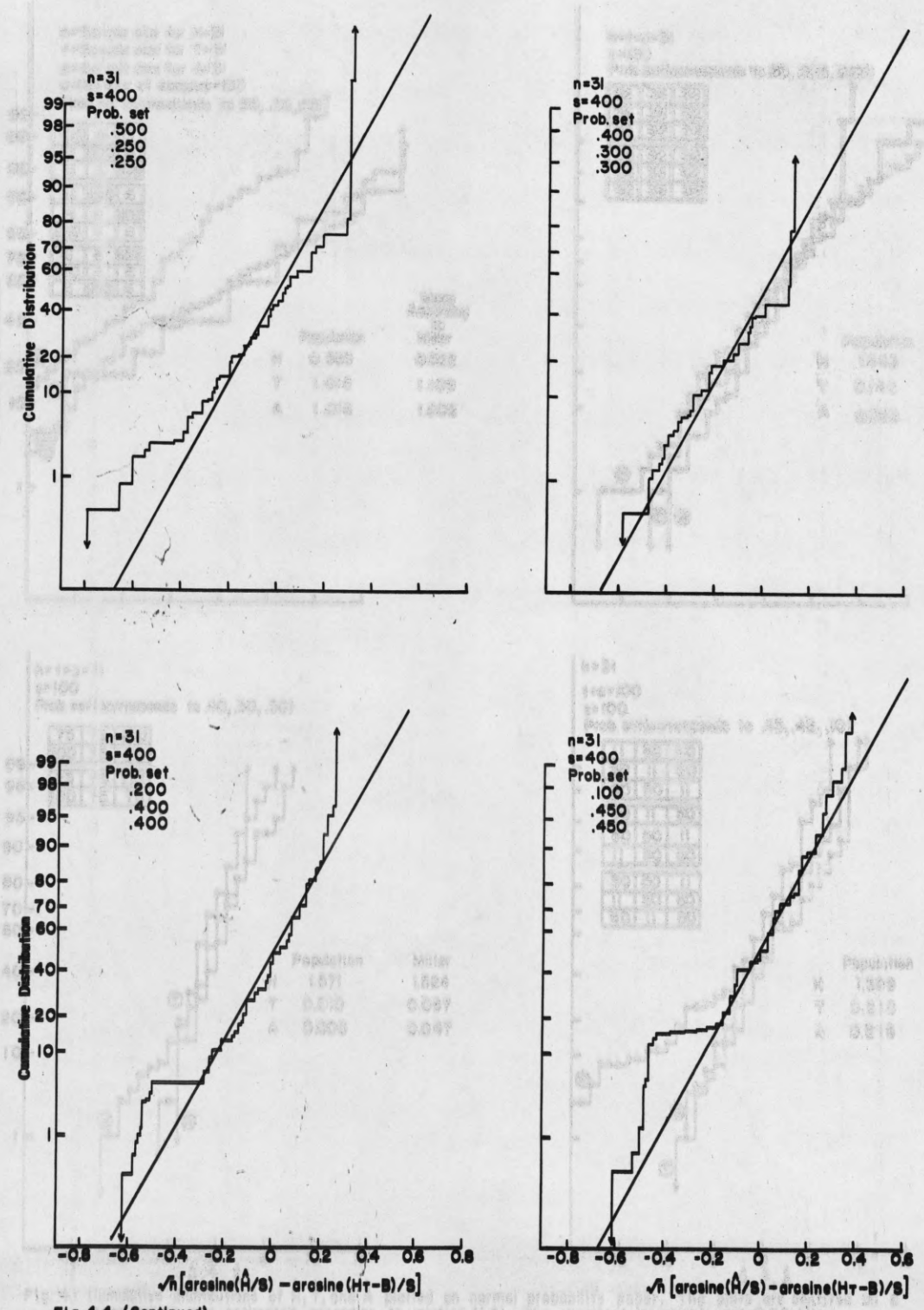
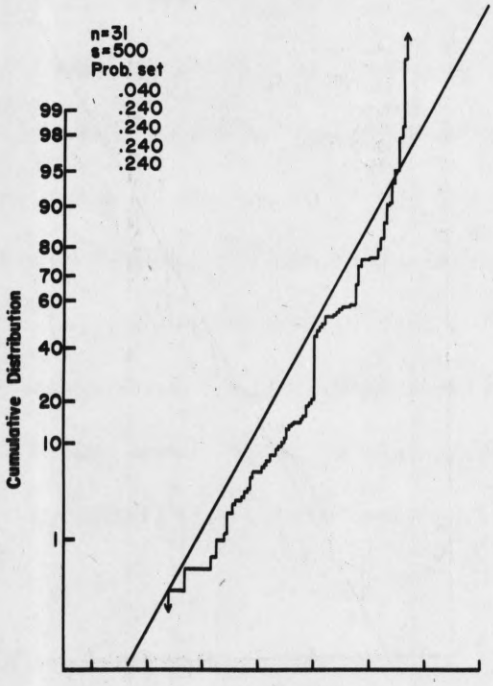
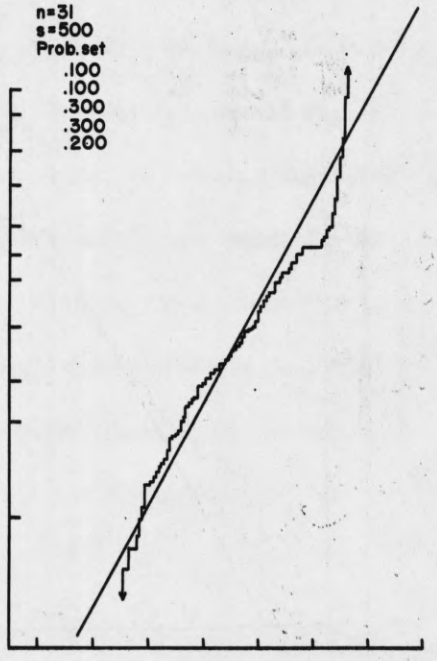
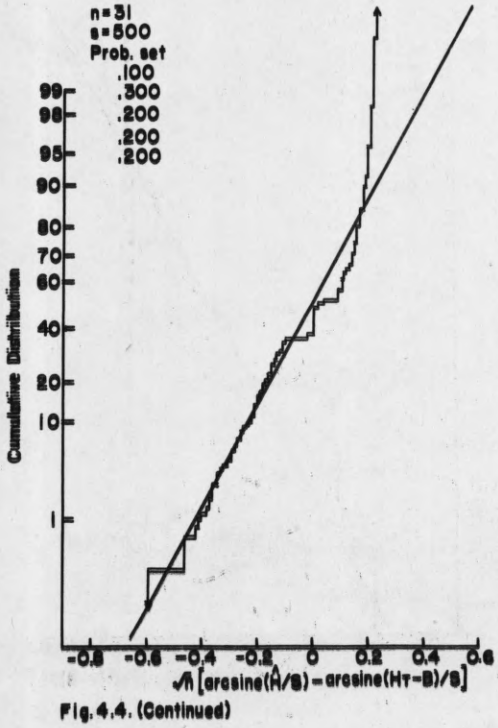


Fig. 4.4. (Continued)







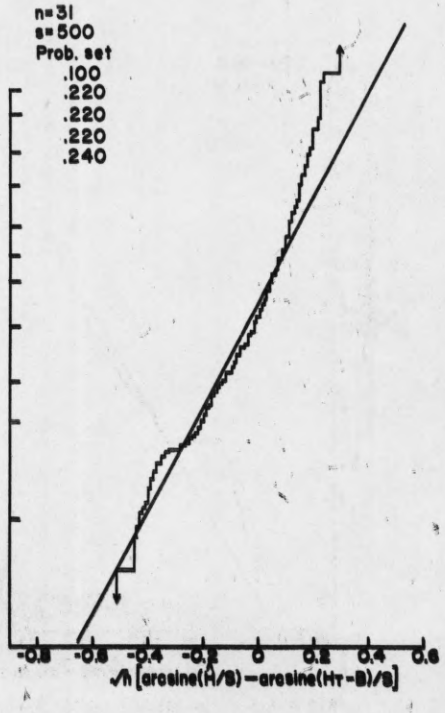


Fig. 4.4. (Continued)

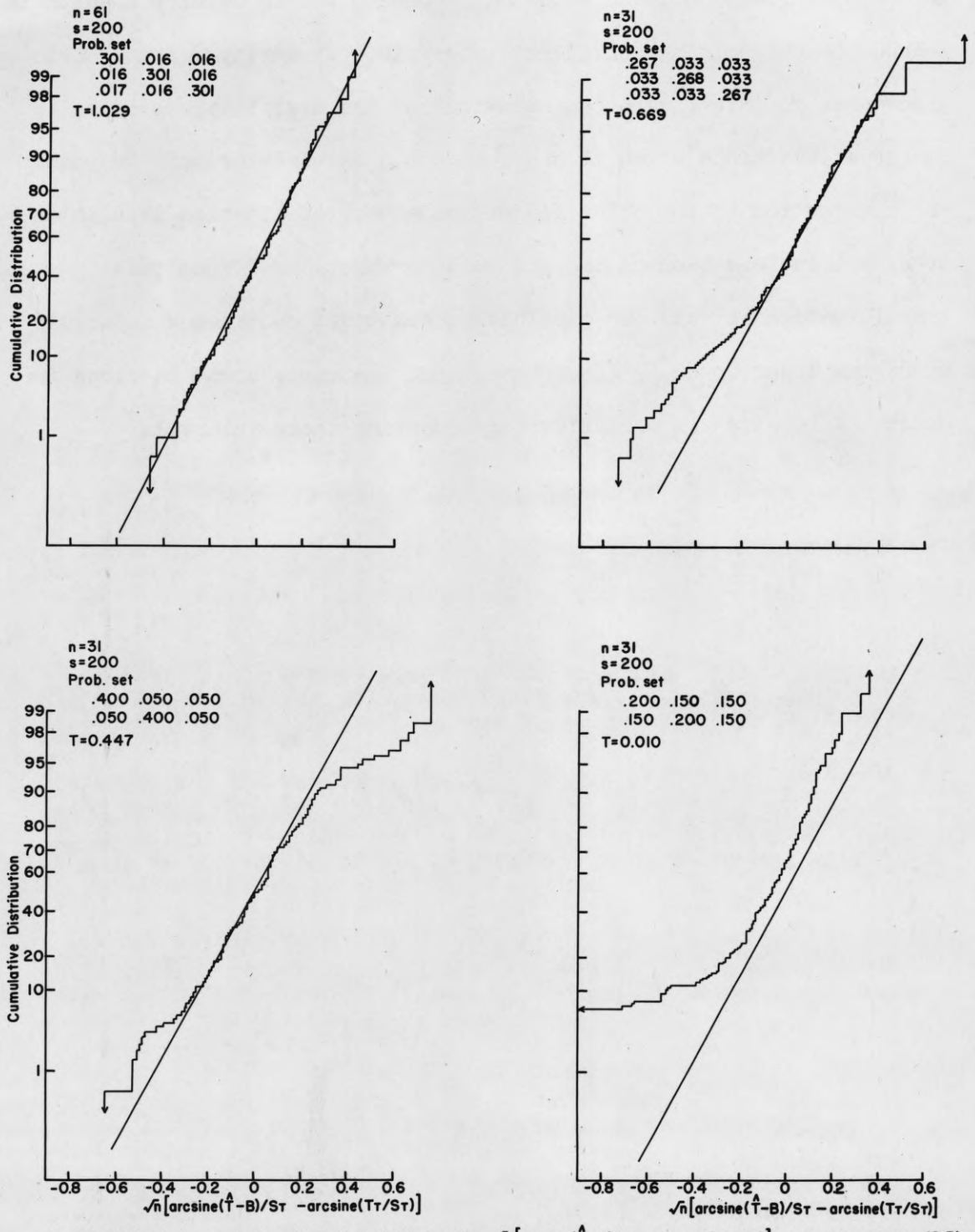


Fig.4.5. Plots of the cumulative distribution of the function $\sqrt{n}\left[\arcsin(\hat{T}-B)/S_T - \arcsin(T_T/S_T)\right]$ defined in relation (2.5). The straight lines represent the normal distribution which the function asymptotically approaches.

V. SUMMARY

We have described a study concerning the distribution theory of information functions, where our main concern was to develop a suitable method for determining confidence intervals. We employed Monte Carlo procedures to investigate the behaviour of the distributions of a number of functions which it was felt could be used for this purpose. The application of the principle of variance stabilization lead to an arcsine transformation of the information functions. From this transformation it has been possible to construct confidence intervals where the lower bound is conservative and the upper bound is close to exact. A "recipe" is presented for computing these intervals.

REFERENCES

1. Miller, G.A., and Madow, W.G., On the Maximum Likelihood Estimate of the Shannon Wiener Measure of Information, Report No. AFCRC-TR-54-75, Air Force Cambridge Research Center, 1954.

2. Rogers and Green

3. Good, I.J., "The Population Frequencies of Species", Biometrika, Vol. 40 (1953) pp. 237-264.

4. Blyth, C; Personal Communication.

5. Miller, G.A., "Note on the bias of information estimates" in Quastler, H. (ed.), Information Theory in Psychology: Problems and Methods, Glencoe, Illinois, The Free Press (1956).

6. Blom, G., "Transformations of the Binomial, Negative Binomial, Poisson and χ^2 Distributions", Biometrika, Vol. 41 (1954) pp. 302-316.

7. Freeman, F., and Tukey, J.W., "Transformations related to the angular and the square root", Ann. of Math. Stat., Vol. 21, No. r (1950), pp. 607-611.

8. Bartlett, M.S., "Square Root Transformations in the Analysis of Variance", Supplement to the Journal of the Royal Statistical Society, Vol. 3 (1936) pp. 68-78.

9. Eisenhart, C., "Inverse Sine Transformations of Proportions", in Eisenhart, C., Hastay, M., and Wallis, W.A., (ed). Techniques of Statistical Analysis, New York, New York, McGraw-Hill, 1947, pp. 395-416.

10. Holloway, J., and Woodbury, M., Application of Information Theory and Discriminant Function Analysis to Weather Forecasting and Forecast Verification. Technical Report No. 1 of the Meteorological Statistics Project of the Institute for Cooperative Research of the University of Pennsylvania, 1955.

11. Holloway, J., and Woodbury, M., Techniques for Summarizing Information in the Sea Level Pressure Pattern. Technical Report No. 2 of the Meteorological Statistics Project of the Institute for Cooperative Research of the University of Pennsylvania, 1955.

12. Holloway, J., Holt, A., Manchly, J., Woodbury, M., Topics in Statistical Meteorology. Final Report of the Meteorological Statistics Project of the Institute for Cooperative Research of the University of Pennsylvania, 1955.