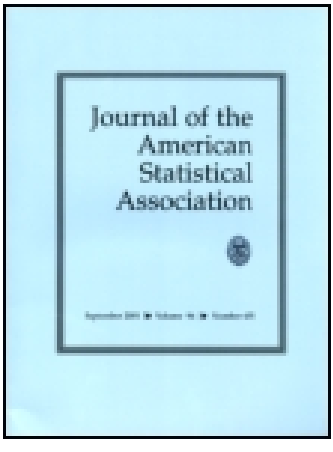

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Some Practical Techniques in Serial Number Analysis

Leo A. Goodman [a]
[a] University of Chicago
Published online: 11 Apr 2012.

To cite this article: Leo A. Goodman (1954) Some Practical Techniques in Serial Number Analysis, Journal of the American Statistical Association, 49:265, 97-112

To link to this article: http://dx.doi.org/10.1080/01621459.1954.10501218

SOME PRACTICAL TECHNIQUES IN SERIAL NUMBER ANALYSIS

LEO A. GOODMAN
University of Chicago

The problem discussed is that of sampling from continuous and discrete uniform distributions. An application of this problem is presented which deals with the analysis of serial numbers on manufactured items in order to estimate the total number of items manufactured. Estimates of bounded relative error are obtained. Some justification for the use of these estimates is presented from the loss (cost) function point of view. Confidence intervals for the parameters are obtained and graphs are presented which may be used to determine the sample size required for confidence intervals of a given expected relative length. Tests of hypotheses are discussed. A method is presented for determining whether the serial numbers obtained are a random sample from a population of consecutive serial numbers.[1]

CONTENTS

CHART

[1] The author is indebted to Mr. William A. Aronow, New Holland Machine Company, and to Professor Harry V. Roberts, University of Chicgo, for very helpful comments.

1. INTRODUCTION

THE analysis of serial numbers has several practical applications. We shall describe two such uses. The interested reader will no no doubt think of still other applications.

a) A commercial company could use the methods of serial number analysis in order to estimate the production and capacity of its competitors. Representatives from the company could obtain the serial numbers of showroom equipment as well as equipment in use which has been produced by the competitors. Many of the basic methods have been developed for analyzing the serial numbers obtained by the company representatives (see [3]).

b) An organization has been using equipment which was purchased many years ago. The question was raised as to how many pieces of equipment had been purchased. No records were immediately available to determine the total purchase, since the purchase had been made years ago. Since serial numbers had been placed on each piece of equipment at the time of purchase, the serial numbers obtained from a sample of the equipment could be used to estimate the total purchase. Section 5 describes how this method was used to estimate the total number of pieces of equipment (desks, bookcases, etc.) which were purchased for the Division of the Social Sciences, The University of Chicago.

2. SUMMARY

Some of the practical problems which are of importance to organizations using "serial number analysis" will be considered here.

The arithmetic involved in the analysis of serial numbers seems to be simpler if the unknown production p is "assumed so large that variation is continuous" (see [3], p. 629). Some results for the "continuous variation" case will be presented which will serve as an approximation to the exact results. Some exact results will then be discussed.

The problem of obtaining confidence intervals for the total production p is studied. The sample size necessary to obtain confidence intervals of a given average relative length is determined. The power of tests of hypotheses concerning the true value of the production is also examined.

Rather than use an estimate of the production p which is unbiased or which minimizes the average of the squared error (see [3]) it might be desirable to have an estimate of which we are "almost certain" that it will be no more than, say, 1.2 times p and no less than, say, $0.8p$. The estimate which maximizes the probability of being included in the

desired interval may be determined. For example, if d is the difference between the largest and smallest serial number in a sample of thirty-one serial numbers, then we can be "99.99% confident" that the estimate $1.20d$ will be between $0.8p$ and $1.2p$. In other words, we can be "99.99% confident" that the relative error of the estimate $1.20d$ of p is less than .2. Justification of the use of such estimates of "bounded relative error" is presented within the framework of the theory of statistical decisions.

A method is also presented for testing the basic assumptions made in serial number analysis by examining the serial numbers which have been obtained. It is possible to test the hypothesis that the serial numbers obtained are a random sample. This method may also be used to detect whether there is a change in the procedure of serial numbering.

An application of the methods described herein is discussed in the final section.

3. CONTINUOUS VARIATION

In this section we shall assume that the serial numbers have a continuous uniform distribution between the initial serial number s and the final serial number $s+p$, where the total production p is unknown. Both the case when the initial serial number s is known and also when it is unknown will be considered.

3.1. *Initial Number Known*

When the initial number s is known, we might subtract s from each serial number obtained. The serial numbers (after the subtraction has been made) will then be uniformly distributed between 0 and p. The production p will be estimated using a sample of n serial numbers.

3.1.1. *Confidence intervals.* Let us first consider the problem of obtaining confidence intervals for p. If g is the largest serial number observed, suppose we state that "the total production p is between g and ag," where a is some constant greater than 1. Then the probability that this statement will be incorrect is $1/a^n$. That is, such a statement will be incorrect if and only if $ag<p(g<p/a)$. If $n=1$, the probability that $g<p/a$ is $\int_0^{p/a} dx/p = 1/a$. Since each observation is independent, the probability that *all* observations, and therefore g in particular will be less than p/a is $1/a^n$. This probability $1/a^n=\alpha$ of making an incorrect statement may be made small by choosing a large value for the constant a, or by obtaining a large sample of n serial numbers. We might first determine how small the probability α of making an incorrect statement should be, and then determine a or n from the relation

$\alpha = 1/a^n$. The interval "g to ag" in which it is stated that p lies is called the "$(1-\alpha)\cdot 100\%$ confidence interval" since the probability is $1-\alpha$ that the statement will be correct.

The length of the confidence interval in which it is stated that p lies is $ag-g=g(a-1)$. Since the expected value of g is $pn/(n+1)$, the expected length of the interval is $pn(a-1)/(n+1)$. The expected relative length of the interval is $n(a-1)/(n+1)=\lambda$. We might first determine how small the probability α of making an incorrect statement should be and also how small the expected relative length λ of the confidence interval should be. The sample size n of the serial numbers may then be determined by the relations

$$n(a-1)/(n+1) = \lambda \quad \text{and} \quad \alpha = 1/a^n \quad \text{or}$$

$$a = \lambda + 1 + \lambda x \quad \text{and} \quad a = \alpha^{-x}, \quad \text{where} \quad x = 1/n.$$

For any given values of α and λ, graphs of the functions $\lambda+1+\lambda x$ and α^{-x} can be drawn. The value x_0 of x where the two graphs intersect is then the desired solution of the last two equations. The reciprocal $1/x_0=n_0$ of this solution is the desired sample size. If then n_0 serial numbers are obtained, we will have $(1-\alpha)\cdot 100\%$ confidence in the statement that "p lies between g and ag." The expected relative length of this confidence interval is the desired value λ.

It is interesting to note that among all $(1-\alpha)\cdot 100\%$ confidence intervals of the form "a_1g to a_2g," where $1 \leqq a_1 < a_2$, the confidence interval with the smallest average length is obtained by taking $a_1=1$, which is what we have done.

3.1.2. *Testing hypotheses.* Let us now consider the problem of testing the hypothesis that the total production is a given value p_0. This hypothesis will be rejected when the given value p_0 does not lie within the confidence interval. In other words, having observed a sample of serial numbers, we make a confidence statement that "p is between g and ag," and reject the "null" hypothesis that the total production is a given value p_0 if this value lies outside the confidence interval. The probability is $\alpha=1/a^n$ of rejecting this hypothesis when it is in fact true. We should like the probability of rejecting the null hypothesis (that the total production is p_0) to be large, when the hypothesis is in fact false (i.e., when the total production is a value p different from p_0). This probability $1-\beta$ of correctly rejecting the null hypothesis, when in fact the true production is p, may be determined by the following formula:

$$1 - \beta(p) = \begin{cases} 1, & \text{when } p < p_0/a \\ \alpha(p_0/p)^n, & \text{when } p_0/a \leqq p \leqq p_0 \\ 1 - (1 - \alpha)(p_0/p)^n, & \text{when } p > p_0. \end{cases}$$

We call $1-\beta(p)$ the power function of the test.

The formula for the power function $1-\beta(p)$ follows directly from the following considerations. The null hypothesis that the total production is a given value p_0 will be rejected whenever $p_0<g$ or $p_0>ag$. But $g<p_0/a$ if and only if all observations are less than p_0/a. The probability that an observation will be less than p_0/a is p_0/ap, when in fact the true production is $p>p_0/a$. Hence the probability that all observations will be less than p_0/a (i.e., $g<p_0/a$), is $(p_0/ap)^n=(p_0/p)^n\alpha$ if $p>p_0/a$. If $p<p_0/a$, rejection of the null hypothesis is certain since $g\leqq p<p_0/a$. The probability that at least one observation will be greater than p_0 (i.e., $g>p_0$) is zero for $p<p_0$, and it is $1-(p_0/p)^n$ for $p>p_0$. From these conclusions the formula for the power function follows directly.

We might first determine how small the probability α of incorrectly rejecting the null hypothesis should be and also how large the probability $1-\beta$ should be of correctly rejecting the null hypothesis when a particular alternative hypothesis $p=p_1$ (different from p_0) is true. If the alternate hypothesis $p=p_1$ has been specified the appropriate sample size of the serial numbers required can be determined by solving the equation

$$1 - \beta = 1 - \beta(p_1)$$

for the value of n. For example, if $p_0\alpha\leqq p_1\leqq p_0$, then

$$1 - \beta = \alpha(p_0/p_1)^n$$

$$(1 - \beta)/\alpha = (p_0/p_1)^n$$

or

$$n = \log\,[(1 - \beta)/\alpha]/\log\,[p_0/p_1].$$

3.1.3. *Estimates of bounded relative error.* In [3], the problem of point estimation of p was considered and the unbiased estimate of p which had the smallest variance was given. The relation between this unbiased estimate and various other point estimates of p was examined. The problem of point estimation will now be considered from a somewhat different point of view. We might want to be "almost certain" that the estimate of production p obtained from the sample of n serial numbers will not be more than 1.2 times as large as the true production p, and

will not be smaller than $0.8p$. If the estimate is of the form cg, where $c \geqq 1$ is a constant and g is the largest among the n serial numbers, then the probability that the estimate cg will lie between $0.8p$ and $1.2p$ is

$$(1.2/c)^n - (0.8/c)^n, \quad \text{when} \quad c \geqq 1.2$$

and

$$1 - (0.8/c)^n, \quad \text{when} \quad c \leqq 1.2.$$

Hence the probability that cg will lie between $0.8p$ and $1.2p$ is maximized when $c = 1.2$ and, in that case, the probability is

$$1 - (0.8/1.2)^n.$$

The sample size n necessary in order that we can be "$(1-\alpha)\cdot 100\%$ confident" that $1.2g$ lies between $0.8p$ and $1.2p$ is determined by the relation

$$1 - \alpha = 1 - (0.8/1.2)^n$$

or

$$n = \log \alpha / \log (0.8/1.2).$$

It may be desirable to determine an interval $c_1 g$ to $c_2 g (1 \leqq c_1 \leqq c_2)$ of which we can be at least "$(1-\alpha)\cdot 100\%$ confident" that any given estimate of the form cg in that interval ($c_1 \leqq c \leqq c_2$) will lie between $0.8p$ and $1.2p$. In order to obtain such an interval, it is clear that the sample size n must be greater than $\log \alpha / \log (0.8/1.2)$. In that case, the values of c_1 and c_2 are determined by

$$1 - \alpha = 1 - (0.8/c_1)^n$$

and

$$1 - \alpha = [(1.2)^n - (0.8)^n]/c_2^n, \quad \text{since} \quad c_1 \leqq c \leqq c_2.$$

We might wish to determine an interval $c_3 g$ to $c_4 g$ of which we can be "$(1-\alpha)\cdot 100\%$ confident" that the entire interval will lie between $0.8p$ and $1.2p$. If $n > \log \alpha / \log (0.8/1.2)$, appropriate values of $c_3 < 1.2$ and $c_4 > 1.2$ can be determined by the relation

$$(1.2/c_4)^n - (0.8/c_3)^n = 1 - \alpha.$$

More generally, if an estimate cg is desired which maximizes the probability of being included between $k_1 p$ and $k_2 p$ (where the k's are

given constants such that $k_1 < k_2$) then the estimate should be $k_2 p$. If the sample size n is greater than $\log \alpha / \log (k_1/k_2)$, then the probability is at least $1-\alpha$ that any given estimate of the form g times a given constant in the interval $c_1 g$ and $c_2 g$ will lie between $k_1 p$ and $k_2 p$, where

$$c_1^n = k_1^n/\alpha$$

and

$$c_2^n = [k_2^n - k_1^n]/(1 - \alpha).$$

Also, the probability is $1-\alpha$ that the entire interval $c_3 g$ to $c_4 g$ will lie between $k_1 p$ and $k_2 p$ where

$$(k_2/c_4)^n - (k_1/c_3)^n = 1 - \alpha.$$

In practice it may sometimes be possible to determine the constants k_1 and k_2 so that if the estimate $\hat{p}$ of p is between $k_1 p$ and $k_2 p$ it will be "close enough." By "close enough" we mean that no loss is incurred when an estimate $\hat{p}$ of p is made which is between $k_1 p$ and $k_2 p$. When the estimate $\hat{p}$ is not between $k_1 p$ and $k_2 p$, then the loss incurred in using an estimate which is not "close enough" may be some given constant, say, 1. If the loss incurred in estimating p by $\hat{p}$ may in fact be described by the function

$$L(\hat{p}, p) = \begin{cases} 0 & \text{when } k_1 p < \hat{p} < k_2 p, \\ 1 & \text{otherwise,} \end{cases}$$

then the estimate which maximizes the chance of being included between $k_1 p$ and $k_2 p$ also minimizes the expected loss. Hence the estimate $k_2 g$ which maximizes the chance of being included between $k_1 p$ and $k_2 p$ may be justified within the framework of the theory of statistical decisions. For a more general discussion of the problem treated in this paragraph the reader is referred to [2].

3.1.4. *Tests of randomness and consecutive serial numbering.* It has been assumed herein that the n serial numbers obtained are a random sample from all the serial numbers which are distributed uniformly (numbered consecutively) between the initial serial number s and the final serial number $s+p$, where s or $s+p$ (or both) may be unknown. Before applying the statistical methods which have been based on this assumption, it is desirable to examine the sample of n serial numbers and test whether this assumption is justified. That is, the hypothesis that the serial numbers were obtained from a random sample of n observations from a uniform distribution between s and $s+p$ should

be tested. The question "Are the serial numbers a random sample?" will be studied.

When the initial serial number s is known, it has been assumed that the serial numbers (after s has been subtracted from each serial number) are uniformly distributed between 0 and p, where p is unknown. The n serial numbers have been assumed to be a random sample of serial numbers. Let us now consider the problem of testing the hypothesis that the n serial numbers are a random sample. We note that the hypothesis to be tested is not concerned with determining the unknown true value of the production p. Several tests are available for the hypothesis that the n serial numbers are a random sample from all the serial numbers uniformly distributed between 0 and p, where p is not specified. Consider all n serial numbers obtained except the largest serial number g. If the hypothesis to be tested is true, then this sample of the $n-1$ smallest serial numbers will be uniformly distributed between 0 and g, when g is given. Hence, dividing these $n-1$ serial numbers by g, the numbers obtained will be uniformly distributed between 0 and 1, when the hypothesis to be tested is true. In order to test the hypothesis of randomness, we might test whether these $n-1$ serial numbers (divided by g) are uniformly distributed between 0 and 1. This can be done using the Kolmogorov statistic or one of the other statistics (e.g., chi-square, maximum difference, etc.) described in [1]. For example, if $n=31$, a graph of the sample cumulative distribution of the $n-1=30$ smallest serial numbers obtained (when divided by the largest serial number obtained can be drawn). The maximum absolute difference between this sample cumulative and the cumulative of the uniform distribution (the diagonal line) is then determined. From Table 1 ($N=30$), on page 428 of [1], we find that the probability is .97745 that this maximum absolute difference between the cumulatives will be less than 8/30. Hence, if a test is to be performed at the .02255 level of significance, we will accept the hypothesis of randomness whenever the maximum absolute difference between the cumulatives is less than 8/30.

If the hypothesis of randomness is accepted, the analysis described in the preceding sections herein and in [3] could then be used. If the hypothesis is rejected, the sample of serial numbers should be examined to determine what is nonrandom about it. On the basis of such an inquiry *ad hoc* methods for estimating the true production p could be determined.

This approach may also be used to see whether there are changes in

the procedure of serial numbering. If the procedure changes (i.e., if the serial numbers are not uniformly distributed between the initial serial number and the final serial number), then a random sample of the serial numbers might indicate a nonuniform distribution. The test proposed in this section may be considered as a test of the hypothesis that serial numbering was done consecutively, as well as a "test of randomness."

3.2. *Initial Number Unknown*

3.2.1. *Confidence intervals.* Let us first consider the problem of obtaining confidence intervals for p.

The probability that the difference d between the largest and smallest among the n serial numbers is greater than $p/b(b \geqq 1)$ may be determined by the following relation (see [4], page 386):

$$\begin{aligned}\Pr\{p \geqq d \geqq p/b\} &= \Pr\{1 \geqq d/p \geqq 1/b\} \\ &= \int_{1/b}^{1} n(n-1)z^{n-2}(1-z)dz \\ &= 1 - nb^{1-n} + (n-1)b^{-n} \\ &= \Pr\{d \leqq p \leqq bd\}.\end{aligned}$$

Suppose the statement is made that "the total production p is between d and bd," where b is some constant. Then the probability α that this statement will be incorrect is $nb^{1-n}+(1-n)b^{-n}=\alpha$. This probability α of making an incorrect statement may be made small by choosing a large value for the constant b, or by obtaining a large sample of n serial numbers. We might first determine how small the probability α of making an incorrect statement should be, and then determine b or n from the relation $\alpha=nb^{1-n}+(1-n)b^{-n}$. Tables are available which will simplify the computations (see [5], [6]). A reprint of [6] may be purchased from *Biometrika*.

Let us illustrate the methods just described by a numerical example. If α is chosen equal to 0.05, the value of $1/b$ can be determined from the entries in column $4=v_1$ on p. 174 of [6] where $2(n-1)=v_2$. If $n=31$ serial numbers have been obtained, then $1/b$ is determined by the entry in the fourth column ($v_1=4$) and third row from the bottom ($v_2=60$) of the table on page 174 in [6]. Hence $1/b=.85591$ and $b=1.17$. Upon observing 31 serial numbers, we will be 95% confident in the statement that "the total production p lies between d and $1.17d$."

The length of the 95% confidence interval for $n=31$ serial numbers is $d(1.17-1)=0.17d$. Since the expected value of d is $p(n-1)/(n+1)$,[2] the expected length of the interval is $(0.17)\ p(n-1)/(n+1)=0.16p$. The expected relative length of the interval is $\lambda=0.16$. We might first determine how small the probability α of making an incorrect statement should be and also how small the expected relative length λ of the confidence interval should be. Then the relations

$$(b-1)(n-1)/(n+1)=\lambda \quad \text{and} \quad nb^{1-n}+(1-n)b^{-n}=\alpha$$

can be used to determine b and the necessary sample size n. Writing $1/b=y$ and $1/(n-1)=x$, the first relation can be replaced by $1/y=\lambda+1+2\lambda x$.

Other methods for determining n may also be used; e.g., successive approximation procedures.

3.2.2. *Testing hypotheses.* The problem of testing the hypothesis that the total production is a given value p_0 may be studied in the same way as was done in Section 3.1.2. Direct computations may be made for any test at a given level α of significance in order to determine the power function of the test. The tables in [5] and [6] may be used to simplify computation.

3.2.3. *Estimates of bounded relative error.* Let us now consider the problem of point estimation of p from the same point of view as in Section 3.1.3. We might want to be "almost certain" that the estimate of p obtained from the sample of n serial numbers "will not be more than $1.2p$ nor smaller than $0.8p$." If the estimate is of the form cd, where $c>1$ is a constant and d is the difference between the largest and smallest among the n observed serial numbers, then the probability that the estimate will be between $0.8p$ and $1.2p$ is maximized when

$$(1.2)^{n-1}(1-1.2/c)-(0.8)^{n-1}(1-0.8/c)=0$$

or when

$$(1) \qquad c=[(1.2)^n-(0.8)^n]/[(1.2)^{n-1}-(0.8)^{n-1}].$$

The sample size n necessary in order that we can be "$(1-\alpha)\cdot 100\%$ confident," that cg lies between $0.8p$ and $1.2p$ is determined by the relation

[2] The reader will notice that the expected value of d presented on page 627 of [3] is $(p+1)(n-1)/(n+1)$. The formula in [3] was derived for the exact model whereas the formula in this text is for the continuous variation model. Hence, $d(n+1)/(n-1)-1$ is the unbiased estimate of p in the exact model (see [3]) whereas the unbiased estimate of p for the continuous variation model is $d(n+1)/(n-1)$.

$$1 - \alpha = \int_{0.8/c}^{1.2/c} n(n-1)z^{n-2}(1-z)dz$$
$$= \Pr\{0.8/c \leqq d/p \leqq 1.2/c\} \tag{2}$$

and relation (1).

If the sample size is larger than the sample required by the preceding relations (1) and (2), two constants $c_1 \leqq c$ and $c_2 > c$ can be determined, where c is defined by relation (1), such that we can be at least "$(1-\alpha)\cdot 100\%$ confident" that any given estimate of the form d times a given constant in the interval $c_1 d$ to $c_2 d$ will lie between $0.8p$ and $1.2p$. The values of c_1 and c_2 are determined by the relations

$$1 - \alpha = \Pr\{0.8/c_2 \leqq d/p \leqq 1.2/c_2\}$$

and

$$1 - \alpha = \Pr\{0.8/c_1 \leqq d/p\}.$$

It may be desirable to determine an interval $c_3 d$ to $c_4 d$ of which we can be "$(1-\alpha)\cdot 100\%$ confident" that the entire interval will lie between $0.8p$ and $1.2p$. When the sample size n is larger than the sample required by relations (1) and (2), appropriate values of $c_3 < c$ and $c_4 > c$ may be determined by the relation

$$1 - \alpha = \Pr\{0.8/c_3 \leqq d/p \leqq 1.2/c_4\}.$$

The numbers 0.8 and 1.2 can be replaced by k_1 and k_2 respectively in the preceding discussion to obtain more general results. A justification of estimates of bounded relative error may be presented, as was done in Section 3.1.3, within the framework of the theory of statistical decisions. The estimate cd which maximizes the chance of being included within $k_1 p$ and $k_2 p$ is also the estimate which minimizes the expected loss if no loss is incurred when the estimate is within $k_1 p$ and $k_2 p$ and a constant loss is incurred otherwise.

Let us illustrate the computations required in the preceding discussion by considering a sample of $n = 31$ serial numbers. The value of c as defined by relation (1) is equal to 1.20 (to three significant digits), when $n = 31$. Hence, the estimate $1.20\,d$ maximizes the chance of being included between $0.8p$ and $1.2p$. From the tables on page 54 of [5] we find that the chance is .9999 that $1.20d$ will lie between $0.8p$ and $1.2p$.

Suppose we wish to be 95% confident of all statements made, i.e., $\alpha = .05$. The second column ($p = 30$) of the table ($q = 2$) on page 54 of [5] presents the distribution of d. Using this information together with

the entry in the eighth column ($v_1=60$) and the fourth row ($v_2=4$) on page 175 of [6], we see that c_2 is about $1.2/(1-.011585)=1.21$ (to three significant digits). Hence if the estimate of production p based on 31 serial numbers is $1.21d$, then the probability is 0.95 that this estimate will be between $0.8p$ and $1.2p$. From the table on page 174 of [6] ($v_1=4$, $v_2=60$) we see that

$$\Pr\{0.8 \leqq d/p\} > .95.$$

Hence, we are at least 95% confident that any given estimate (of the form d times a given constant) in the interval d and $1.21d$ will lie between $0.8p$ and $1.2p$. We also find from the tables that the probability is about .95 that the entire interval d to $1.21d$ will lie between $0.8p$ and $1.2p$.

3.2.4. *Tests of randomness and consecutive serial numbering.* Let us consider the hypothesis that the n serial numbers obtained are a random sample from the population of uniformly distributed serial numbers. In the case where the initial number is unknown, we consider all n serial numbers obtained except the largest serial number g and the smallest serial number f. If the hypothesis to be tested is true, then this sample of $n-2$ serial numbers (all except g and f) will be uniformly distributed between f and g, when f and g are given. Hence, subtracting f from these $n-2$ serial numbers and then dividing the numbers obtained by $g-f$, the adjusted numbers will be uniformly distributed between 0 and 1, when the hypothesis to be tested is true. In order to test the hypothesis of randomness, we might test whether these $n-2$ adjusted serial numbers (when f is subtracted from the serial numbers and the numbers obtained are then divided by $g-f$) are uniformly distributed between 0 and 1. This can be done using the Kolmogorov statistic or one of the other statistics (e.g., chi-square, maximum difference, etc.) as mentioned in Section 3.14. For example if $n=31$, the sample cumulative distribution of the $n-2=29$ adjusted serial numbers obtained can be graphed. The maximum absolute difference between this sample cumulative and the cumulative of the uniform distribution (the diagonal line) can then be determined. From Table 1 ($N=29$) on page 428 of [1], we note that the probability is .98076 that this maximum absolute difference between the cumulatives will be less than 8/29. Hence, if a test is to be performed at the .01924 level of significance, the hypothesis of randomness and consecutive (uniformly distributed) serial numbers will be accepted whenever the maximum absolute difference between the cumulatives is less than 8/29.

4. THE EXACT MODEL

In the preceding sections we have assumed that the serial numbers have a continuous uniform distribution between the initial serial number s and the final serial number $s+p$. This was done in order to simplify the problem and because for practical problems (when the value of p is large) the results obtained will serve as an approximation to results for the exact model of a discrete, finite, uniform population (see [3]).

On page 624 of [3], the exact confidence intervals and tests of hypotheses are obtained for the case where the initial serial number is known. Since exact confidence intervals and tests of hypotheses were not discussed in [3] for the case where the initial serial number is unknown, we shall now consider that problem.

From [3], we see that the probability that the difference d between the largest and smallest among n serial numbers will be less than or equal to a given constant c may be determined from the relation

$$\Pr\{d \leqq c \mid n, p\} = \sum_{d=n-1}^{c} n^{(2)}(d-1)^{(n-2)}(p-d)/p^{(n)}$$
$$= nc^{(n-1)}/(p-1)^{(n-1)} - (n-1)(c+1)^{(n)}/p^{(n)},$$

where $c^{(m)} = c!/(c-m)!$. As a first approximation to this probability we might replace the exact model by the model of a continuous uniform distribution and obtain $Pr^1\{d \leqq c \mid n, p\} = n(c/p)^{n-1} - (n-1)(c/p)^n$ for which convenient tables are available (see [5] and [6]).

Suppose we wish to test the null hypothesis that $p = p_0$ against the alternative hypothesis $p > p_0$. Then the rejection region for a significance test at level α is obviously $d > c_1 + 1$ where c_1 is the largest integer satisfying

$$\Pr\{d \leqq c_1 \mid n, p_0\} < 1 - \alpha.$$

If we wish to test the null hypothesis $p = p_0$ against the alternative hypothesis $p \leqq p_0$, then the rejection region for a significance test at level α is $d \leqq c_2$ where c_2 is the smallest integer satisfying

$$\Pr\{d \leqq c_2 \mid n, p_0\} > \alpha.$$

A two-sided test at level α of the null hypothesis $p = p_0$ against the two-sided alternative $p \neq p_0$ is defined by the acceptance region $c_2 \leqq d \leqq p_0 - 1$. A two-sided test at the 2α level might be based on the acceptance region $c_2 \leqq d \leqq c_1 + 1$.

The results of the preceding paragraph may now be used to obtain confidence intervals. That is, the left-sided $1-\alpha$ confidence interval is $p \geqq k_1$, where k_1 is the smallest integer satisfying

$$\Pr \{d < d_0 | \ n, k_1\} < 1 - \alpha,$$

and d_0 is the actual difference between the largest and smallest among the n serial numbers observed. The right-sided $1-\alpha$ confidence interval is $p \leqq k_2$, where k_2 is the largest integer satisfying

$$\Pr \{d \leqq d_0 | \ n, k_2\} > \alpha.$$

A two-sided $1-\alpha$ confidence interval is $d+1 \leqq p \leqq k_2$ and a two-sided $1-2\alpha$ confidence interval is $k_1 \leqq p \leqq k_2$.

5. AN APPLICATION

The Division of the Social Sciences of the University of Chicago has been using equipment (desks, bookcases, etc.) upon which serial numbers had been placed. The question was raised as to how many such pieces of equipment were there.

The serial numbers on thirty-one pieces of equipment were observed. The 31 serial numbers obtained were:

> 83, 135, 274, 380, 668, 895, 955, 964, 1113, 1174, 1210, 1344, 1387, 1414, 1610, 1668, 1689, 1756, 1865, 1874, 1880, 1936, 2005, 2006, 2065, 2157, 2220, 2224, 2396, 2543, 2787.

The serial numbers range from 83 to 2787. The sample cumulative distribution of the 29 serial numbers obtained between the smallest and largest serial numbers is graphed in Figure 5.1. The diagonal line in Figure 5.1 represents the uniform cumulative distribution between the smallest serial number 83 and the largest serial number 2787. From Figure 5.1 we see that the maximum absolute difference between the two cumulative distributions is $(9.65-5)/29 = .16$. If the serial numbers obtained are a random sample from a population of uniformly distributed serial numbers, then there is more than a $1-.68280 = .3172$ probability of obtaining a maximum absolute difference of .16 or larger (see page 428, Table 1, $N=29$, in [1]). Hence the null hypothesis that the serial numbers obtained are a random sample from a population of consecutive serial numbers is accepted.

From Section 3.2.1 we see that the unbiased estimate of the total number p of pieces of equipment is $d\ 32/30 = (2787-83)32/30 = (2704)32/30 = 86528/30 = 2884.3$ for the continuous variation model (2883.3 for the exact model). Also, the 95% confidence interval for p is "$2704 \leqq p \leqq 1.17(2704)$" or "$2704 \leqq p \leqq 3163.7$."

From Section 3.2.3 we see that the chance is .9999 that the estimate

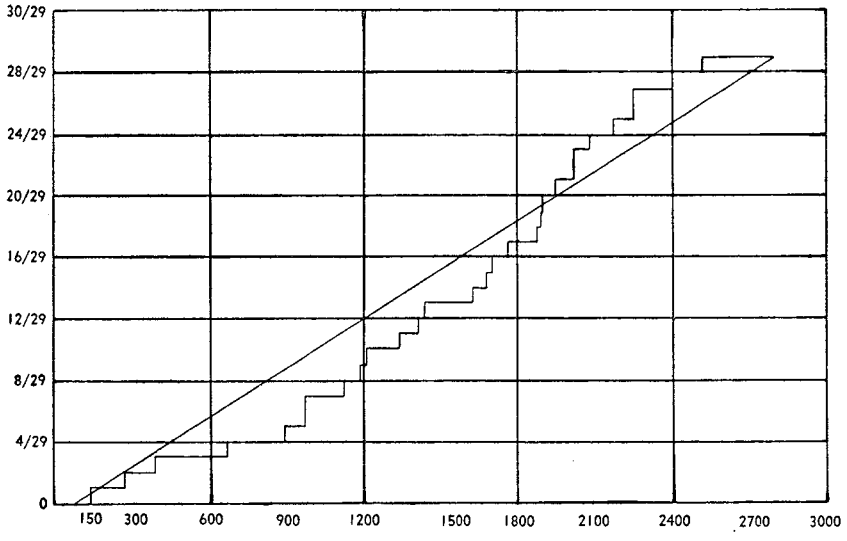

FIG. 5.1. Sample Cumulative Distribution of the 29 Observed Serial Numbers Between the Smallest and Largest.

$1.20d = 1.20(2704) = 3244.8$ will be within 20 per cent of p. This estimate minimizes the expected loss if no loss is incurred when the estimate is within 20 per cent of p and a constant loss is incurred otherwise. The probability is .95 that the estimate $1.21d = 1.21(2704) = 3271.8$ will be within 20 per cent of p. In fact the probability is appropriately .95 chance that the entire interval d to $1.21d$, or 2704 to 3271.8 will lie within 20 per cent of p.

It was a relatively simple task to obtain the serial numbers of 31 pieces of equipment and then to estimate p in the manner described herein. Determining the true value of p (the total number of pieces of equipment) was much more time consuming. These pieces of equipment had been purchased in the period between 1928 and 1934 and no records were immediately available to determine the total purchase. We are indebted to Mrs. Ruth Denney, Aministrative Assistant to the Dean of the Social Sciences. After several days and many inquiries, Mrs. Denny was able to locate the records and found that the total number p of pieces of equipment was *2885*.

6. REFERENCES

[1] Birnbaum, Z. W., "Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size," *Journal of the American Statistical Association*, 47 (1952), 425–41.

[2] Blackwell, D., and M. A. Girshick, *Theory of Games and Statistical Decisions*, in press.

[3] Goodman, Leo A., "Serial number analysis," *Journal of the American Statistical Association*, 47 (1952), 622–34.
[4] Mood, A. M., *Introduction to the Theory of Statistics*, McGraw-Hill Book Company, New York, 1950.
[5] Pearson, Karl, *Tables of the Incomplete Beta Function*, Cambridge University Press, London, 1932.
[6] Thompson, Catherine M., "Tables of percentage points of the incomplete beta function," *Biometrika*, 32 (1941), 151–81.