Selected Tables in Mathematical Statistics

Volume 6

Edited by the Institute of Mathematical Statistics

SELECTED TABLES IN MATHEMATICAL STATISTICS

Volume VI

THE DISTRIBUTION OF THE SIZE OF THE MAXIMUM CLUSTER OF POINTS ON A LINE

by

NORMAN D. NEFF and JOSEPH I. NAUS

Edited by the Institute of Mathematical Statistics

Coeditors

W. J. Kennedy
Iowa State University

and

R. E. Odeh
University of Victoria

Managing Editor

J. M. Davenport
Texas Tech University

AMERICAN MATHEMATICAL SOCIETY
PROVIDENCE, RHODE ISLAND

This volume was prepared with the aid of:

D. R. Cook, University of Minnesota
J. E. Gentle, International Mathematical & Statistical Libraries, Inc.
W. J. Hemmerle, University of Rhode Island
K. Hinklemann, Virginia Polytechnic Institute and State University
R. L. Iman, Sandia Laboratories
A. Londhe, Iowa State University
D. Meeter, Florida State University
D. B. Owen, Southern Methodist University
S. Pearson, Bell Laboratories
N. S. Urquhart, New Mexico State University
R. H. Wampler, National Bureau of Standards

1980 *Mathematics Subject Classification*
Primary 62Q05; Secondary 62E15.

International Standard Serial Number 0094-8837
International Standard Book Number 0-8218-1906-2
Library of Congress Card Number 74-6283

PREFACE

This volume of mathematical tables has been prepared under the aegis of the Institute of Mathematical Statistics. The Institute of Mathematical Statistics is a professional society for mathematically oriented statisticians. The purpose of the Institute is to encourage the development, dissemination, and application of mathematical statistics. The Committee on Mathematical Tables of the Institute of Mathematical Statistics is responsible for preparing and editing this series of tables. The Institute of Mathematical Statistics has entered into an agreement with the American Mathematical Society to jointly publish this series of volumes. At the time of this writing, submissions for future volumes are being solicited. No set number of volumes has been established for this series. The editors will consider publishing as many volumes as are necessary to disseminate meritorious material.

Potential authors should consider the following rules when submitting material.

1. The manuscript must be prepared by the author in a form acceptable for photo-offset. This includes both the tables and introductory material. The author should assume that nothing will be set in type although the editors reserve the right to make editorial changes.

2. While there are no fixed upper and lower limits on the length of tables, authors should be aware that the purpose of this series is to provide an outlet for tables of high quality and utility which are too long to be accepted by a technical journal but too short for separate publication in book form.

3. The author must, wherever applicable, include in his introduction the following:

   (a) He should give the formula used in the calculation, and the computational procedure (or algorithm) used to generate his tables. Generally speaking, FORTRAN or ALGOL programs will not be included but the description of the algorithm used should be complete enough that such programs can be easily prepared.

   (b) A recommendation for interpolation in the tables should be given. The author should give the number of figures of accuracy which can be obtained with linear (and higher degree) interpolation.

   (c) Adequate references must be given.

   (d) The author should give the accuracy of the table and his method of rounding.

(e) In considering possible formats for his tables, the author should attempt to give as much information as possible in as little space as possible. Generally speaking, critical values of a distribution convey more information than the distribution itself, but each case must be judged on its own merits. The text portion of the tables (including column headings, titles, etc.) must be proportional to the size 5–1/4″ by 8–1/4″. Tables may be printed proportional to the size 8–1/4″ by 5–1/4″ (i.e., turned sideways on the page) when absolutely necessary; but this should be avoided and every attempt made to orient the tables in a vertical manner.

(f) The table should adequately cover the entire function. Asymptotic results should be given and tabulated if informative.

(g) An example or examples of the use of the tables should be included.

4. The author should submit as accurate a tabulation as he can. The table will be checked before publication, and any excess of errors will be considered grounds for rejection. The manuscript introduction will be subjected to refereeing and an inadequate introduction may also lead to rejection.

5. Authors having tables they wish to submit should send two copies to:

Dr. Robert E. Odeh, Coeditor
Department of Mathematics
University of Victoria
Victoria, B. C., Canada V8W 2Y2

At the same time, a third copy should be sent to:

Dr. William J. Kennedy, Coeditor
117 Snedecor Hall
Statistical Laboratory
Iowa State University
Ames, Iowa 50011

Additional copies may be required, as needed for the editorial process. After the editorial process is complete, a camera-ready copy must be prepared for the publisher.

Authors should check several current issues of *The Institute of Mathematical Statistics Bulletin* and *The AMSTAT News* for any up-to-date announcements about submissions to this series.

ACKNOWLEDGMENTS

The tables included in the present volume were checked at the University of Victoria. Dr. R. E. Odeh arranged for, and directed this checking with the assistance of Mr. Bruce Wilson. The editors and the Institute of Mathematical Statistics wish to express their great appreciation for this invaluable assistance. So many other people have contributed to the instigation and preparation of this volume that it would be impossible to record their names here. To all these people, who will remain anonymous, the editors and the Institute also wish to express their thanks.

To: **PATRICIA and SARAH**

Contents of VOLUMES I, II, III, IV and V of this Series

TABLE OF CONTENTS

THE DISTRIBUTION OF THE SIZE OF THE MAXIMUM CLUSTER OF POINTS ON A LINE

Norman D. Neff Trenton State College

Joseph I. Naus Rutgers University

ABSTRACT

Given N points randomly drawn from the unit line, let $\tilde{n}_p$ be the size of the largest number of points clustered within an interval of length p. Let $\tilde{p}_n$ be the size of the smallest interval that contains n out of the N points. The distributions of $\tilde{n}_p$ and $\tilde{p}_n$ are related: $\Pr(\tilde{p}_n \leq p) = \Pr(\tilde{n}_p \geq n)$. We denote the common probability $P(n;N,p)$. Tables are given for $P(n;N,p)$ and for the expectation over N of $P(n;N,p)$ where N is a Poisson random variate.

1. INTRODUCTION

Researchers in many fields deal with the clustering of events in time and space. A quality control expert investigates clusters of defectives. A communications engineer designs system capacity to accommodate clusters. An educational psychologist sets success quotas to gauge learning. Experts in accident prevention, reliability, traffic control, ecology, epidemiology and many other fields focus on unusually large clusters. The probabilities of large clusters under various models are tools of the natural, physical, and social sciences.

The present tables provide probabilities for the size of the largest cluster of random points on the line. Given N points independently drawn from the uniform distribution on (0,1), let $\tilde{n}_p$ be the largest number of points to be found in any subinterval of (0,1) of length p. Let $P(n;N,p)$ denote the probability $\Pr(\tilde{n}_p \geq n)$.

Received by the editors December 1978 and in revised form June 1979.
AMS (MOS) Subject Classifications (1970): Primary 62Q05; Secondary 62E15.

Certain applications deal with points generated from a Poisson process, where the total number of points in the unit interval is a Poisson random variate with expectation λ. Let $P'(n;\lambda,p)$ denote the expectation taken over N of $P(n;N,p)$, where N is a Poisson random variate with expectation λ. Tables are given for $P(n;N,p)$ and $P'(n;\lambda,p)$. Section two describes applications of these probabilities to a wide variety of scientific and statistical areas. Section three summarizes known results on the probabilities and details approaches that were used to compute the tables.

2. APPLICATIONS

This section gives a variety of applications for $P(n;N,p)$ and for $P'(n;\lambda,p)$.

2.1 Applications of P(n;N,p)

(a) Developability of silver specks. The photographic signal recorded on a film depends not only on the total number of photons during exposure, but also on their time sequence. Hamilton, Lawton, and Trabka (1972, p. 855) note that this dependency is called reciprocity failure, a key aspect of photographic science. They note that $P(n;N,p)$ was studied early in photographic history by Silberstein (1939) and others and plays a role in low intensity reciprocity failure (a decrease in sensitivity at less than optimum intensities). A classical model for the developability of silver specks is that a speck will develop if k or more photons are absorbed within the decay time, t_o, of the nucleus.

Silberstein (1945), Berg (1945), and Mack (1948, 1950) find the expected number of n-aggregates, or clusters, where n points form an aggregate if "they are all contained within a subinterval, p, of (0,1), no matter how placed in (0,1)."[1] Berg (1945, p. 340) and Mack (1948, p. 784) indicate their interest in $P(n;N,p)$ and derive an approximate formula for it in the case of sufficiently rare aggregates. Silberstein (1945, p. 319) states that

> The rigorous determination of the probability P of any given number of aggregates, n-ets, among N points is exceedingly complicated and becomes to all purposes impracticable when N exceeds a few units.

[1] Silberstein (1945, p. 319).

Silberstein makes two exceptions, namely, the probability of no aggregate, 1-P(2;N,p), and the probability of the largest possible aggregate, P(N;N,p).

(b) Clustering of leukemia cases. Childhood leukemia is a relatively rare disease. Scientists seeking clues as to common causative agents investigate unusual clusters of cases. They ask whether clusters of a given size within a given time period are likely to occur by chance. Ederer, Myers, and Mantel (1962, p. 9), who develop approaches to this problem, state:

> In considering temporal clusters we chose the calendar year as the unit of time, and in this way obtained 5 non-overlapping years in a 5-year period. The reader will recognize that a 5-year period in fact contains a continuum of overlapping periods one year in length. However, the distribution of the maximum number of cases in a year under the null hypothesis [of randomness] cannot be readily determined unless the number of periods is restricted in some way.

It is natural to look at clusters where they fall, be it in a calendar year, or in a year period that overlaps two calendar years. The above authors note that Pinkel and Nefzeger (1959) did look at the unrestricted continuum. However, if one does this, then one must use the appropriate null distribution in checking for significance. For the above example, P(n;N,0.2) is the distribution of the maximum number of cases in a year, under the null hypothesis that the N cases are distributed at random over the five-year period, for the unrestricted continuum of 1-year periods. For example, given N = 15 cases over the five year period, the chance that there is any calendar year that contains as many as seven cases is 0.07 (from Appendix - Table 1 in Ederer, Myers, and Mantel). This is much less likely than the probability that there is a one-year period (calendar or not) that contains at least 7 cases. From Table 1a of our appendix, P(7;15,0.2) = 0.30.

(c) Dialing calls. Dialing is started for fifteen phone calls at times that are distributed at random over a one-minute period. The dialing time for a call is ten seconds. Find the probability that eight or more phone calls are being dialed at the same time.

Feller (1958, p. 397) notes that compound events such as "seven calls within a minute on a certain day" involve complicated sample spaces and goes on to state,

> We cannot deal here with such complicated sample spaces and must defer the study of the more delicate aspects of the theory.

Given 15 phone calls initiated during a minute, the event "eight or more calls initiated during a ten-second interval" is equivalent to the event that for some $i=1,2,\ldots,8$, dialing for the $(7+i)$th phone call started less than ten seconds after dialing started for the ith phone call. Interpolating in Table 1 we find the desired probability $P(8;15,1/6) = 0.037$.

2.2 Applications of $P'(n;\lambda,p)$

(d) Visual perception. Photons arrive at a receptor in the eye according to a Poisson process. Under conditions of low illumination it is observed that the retinal neurones do not discharge for each photon, but rather it is conjectured that there is a triggering effect from several photons. One of the classical theories of perception is that if n or more photons arrive within an integration time t, then this triggers an impulse and the neurone discharges. What is frequently of interest is the distribution of waiting times till discharge. Leslie (1969, p. 379) states that the distribution of the waiting time till discharge under the preceding model is intractable and gives an alternative model. Van de Grind et al. (1971) note that an analytic solution is lacking and use a Monte Carlo approach to estimate the distribution under the classical model. The probability that the waiting time till first discharge is less than or equal to T is $P'(n;\lambda,t/T)$, where

$$P'(n;\lambda,t/T) = \Sigma_{N=n}^{\infty} e^{-\lambda T}(\lambda T)^N P(n;N,t/T)/N!. \qquad (2.1)$$

Tables 2 and 2a provide values for $P'(n;\lambda,p)$. Outside the range of these tables we can use approximations of the form

$$P'(n;\lambda,1/2L) \doteq 1 - [1 - P'(n;2\lambda/L,\tfrac{1}{4})]^{L-1}[1 - P'(n;\lambda/L,\tfrac{1}{2})]^{-L+2}.$$

The reasoning behind this approximation and its generalization for other p values is detailed in the appendix under the discussion of the use of Tables 2 and 2a.

Van de Grind et al. (1971) use a Monte Carlo approach to chart the values of λ such that $P'(k;\lambda,t/T) = 0.60$. Based on a simulation of 10^4 flashes they find $P'(4;\lambda,1/200) = 0.60$ for λ somewhere between 90 and 100 (given the roughness of the plot). From the above approximation and using values from Table 2 we find

$$P'(4;100,1/200) \doteq 1 - [1 - P'(4;2,\tfrac{1}{4})]^{99}[1 - P'(4;1,\tfrac{1}{2})]^{-98}$$
$$= 1 - [1 - 0.018456295]^{99}[1 - 0.007452004]^{-98} \doteq 0.67 .$$

Similarly, $P'(4;90,1/200) \doteq 0.54$.

(e) Pedestrians on a street. Furth (1920) did an empirical study of the number of people on a certain segment of a street. Feller (1958, pp. 370-372) discusses this study and shows how it is related to a moving average process, an example of a nonmarkovian process. Feller points out how this process is related to our problem.

If the pedestrians are walking at a fixed rate, then the time each pedestrian spends on the segment is a constant, C. If the arrival times are exponential, then the probability that the street has had at any time before T as many as n pedestrians on it is $P'(n;\lambda,C/T)$.

(f) A counter problem. A Poisson process with mean rate λ generates impulses that are received by a counter. The counter registers whenever n impulses have occurred together in an interval of length less than t. $P'(n;\lambda,t/T)$ gives the c.d.f. of the waiting time until the first registration of the counter. Janossy (1944) and Schroedinger (1944) consider the rate of n-fold accidental coincidences in counters, and find asymptotic expressions for the probability. Domb (1950) studies various counter problems and derives an implicit formula for the probability that k "n-clusters" occur in (0,T), where he defines an n-cluster as a "recorded event whose dead period contains n-1 other nonrecorded events." Domb's solution for this type of cluster is closely related to Mack's (1948, p. 784) solution for a different type of cluster (see application (a)). Both solutions assume that overlapping effects are negligible. Domb's cluster differs in the way an earlier cluster can prevent a later cluster.

(g) A queueing problem. Customers arrive at an n-server system according to a Poisson process with mean λ. Customers are served by any free server on a first come first served basis. Service time is a constant, t_o. If k or more customers arrive within a time t_0, then service will stagnate (with the result that some customers will have to wait, or some impatient customers will leave). Solov'ev (1966) deals with this problem and gives some results for the expected waiting time till stagnation and some approximate results for $P'(n;\lambda,p)$, the c.d.f. of the waiting time. Newell (1963) also mentions the queueing applications of $P'(n;\lambda,p)$ and various generalizations and derives asymptotic results. Our experience with various examples suggests that the approximation of example (d) together with values from Tables 2 and 2a give more accurate approximations than these asymptotic results. For example, from Table 2 find $P'(4;10,0.1) = 0.3741$. This is the exact result to the number of places we are able to compute it by averaging $P(4;N,0.1)$ for N up to 19, and then treating the remaining probabilities as 1. The number of significant places results from averaging over rapidly shrinking Poisson terms. Applying the approximation of example (d) together with the values from Table 2 gives

$$P'(4;10,0.1) = 1 - [1 - P'(4;4,\tfrac{1}{4})]^4[1 - P'(4;2,\tfrac{1}{2})]^{-3}$$
$$= 1 - [1 - 0.156709254]^4[1 - 0.068630099]^{-3} = 0.3740456.$$

The best previous approximation, that of Newell (1963) and Ikeda (1965) is $P'(n;\lambda,p) \doteq 1 - \exp(-\lambda^n p^{n-1}/(n-1)!)$, which for this example gives $P'(4;10,0.1) \doteq 0.811$.

(h) Breaking strength of materials. Several models postulate the breaking of an object whenever there are several flaws within a certain distance: the breakaway due to several close dislocations in crystals; a multistrand thread breaking when several strands have weaknesses within a common length; earthquakes triggered by shift flaws within a certain distance; nervous breakdowns resulting from several severe personal problems occurring within a short period of time.

2.3 $P(n;N,p)$ and $P'(n;\lambda,p)$ as approximations to the probability of a generalized run.

Let N one's and M-N zeros be randomly arranged in a row. Let the random variable $\tilde{k}_m$ be the maximum number of one's within any m consecutive positions in the arrangement. An n:m quota is defined as at least n one's within some m consecutive trials. $\Pr(\tilde{k}_m \geq n)$ is the probability of a quota, and we denote this probability as $P(n \text{ in } m|N \text{ in } M)$. Saperstein (1969, 1972, 1973), Naus (1974), Greenberg (1970), Huntington (1974, 1976) and others derive various results for $P(n \text{ in } m|N \text{ in } M)$.

Saperstein (1972) gives bounds on the upper 5% points for $\tilde{k}_m$, for $N = 5(1)15$, $m/M = 0.2(0.1)0.7$, and certain values of M. The bounds are based on the formula for the case $n > N/2$. Huntington (1974, pp. 204-285) gives tables for the distribution of $P(n \text{ in } m|N \text{ in } M)$ for two classes of cases.

Case 1: $m/M = 1/L$, $L = 2(1)5$, $n > [N/2] + 1$.

Case 2: $m/M = 0.35(.05)0.50$, and full range of n.

Naus (1974) gives a simple expression for the case $M/m = L$, L an integer, where $n > N/2$:

$$\Pr(\tilde{k}_m \geq n) = 2\Sigma_{s=n}^{N} \binom{m}{s}\binom{M-m}{N-s}/\binom{M}{N} + (Ln-N-1)\binom{m}{n}\binom{M-m}{N-n}/\binom{M}{N}. \quad (2.2)$$

The formula for the other case $n \leq N/2$ is not computationally simple and is in terms of sums of LxL determinants. For this other case, and the cases where $m/M \neq 1/L$, L an integer, one requires either a high speed computer and substantial programming effort, or an approximation. For the case where $P(n \text{ in } m|N \text{ in } M)$ is small (less than 0.1), equation (2.2) provides a good approximation for the case $n \leq N/2$. (It is an approximation for this case because equation (2.2) fails to count the possibility that there might be several non-overlapping sets of n successes each within m trials. Of course, for $n > N/2$, this cannot happen. It is a good approximation for $n \leq N/2$ and small $P(n \text{ in } m|N \text{ in } M)$, because the chance of two non-overlapping quotas is of smaller order of magnitude. Equation (2.2) does take account of the nonnegligible chance that there might be overlapping sets of n successes each within m trials.)

Another approximation is available when m and M are large relative to n and N. The limit of $P(n \text{ in } m|N \text{ in } M)$ as $m \to \infty$, $M \to \infty$, such that $m/M = p$, is $P(n;N,p)$. We can use $P(n;N,p)$ to approximate $P(n \text{ in } m|N \text{ in } M)$ in much the same way that the normal distribution is used to approximate the binomial. Table 2 in Saperstein (1972) is based on this approximation. For example, in a manufacturing process there were eleven defectives in 500 items coming off an assembly line. After reviewing the defectives it is found that six of the defectives occurred within 100 consecutive trials. We approximate P(6 in 100|11 in 500) by P(6;11, 100/500) which from Table 1 equals 0.198. Equation (2.2) gives the exact probability P(6 in 100|11 in 500) = 0.185. In a similar way $P'(n;\lambda,p)$ approximates the expectation over N of P(n in m|N in M), where N is a binomial random variable. We now give applications of P(n in m|N in M).

(i) <u>A learning criterion</u>. Psychologists studying transfer and learning sometimes use a generalized run criterion to decide when to terminate a particular treatment. At each trial, the psychologist counts the number of successes in the last m trials. If at any trial this number exceeds a critical number n (the criterion), this signals a change in the underlying process. For an experiment with $M(> m)$ trials, the critical number n is chosen to give a specified experiment wide level of significance. To set the experiment wide criterion, we require the probability that within M trials there exists a subsequence of m consecutive trials with at least n successes. This is the probability P(n in m|N in M), or the related probability where N is a binomial random variate.

The special case n = m deals with the event of a run of m successes. Bogartz (1965) derives probabilities for the special case of m-2 and m-1 correct responses out of m trials. Bogartz notes that an approach based on Markov chains is practical only for very simple cases, and gives an approximation. Runnels, Thompson, and Runnels (1968) reexamine Bogartz's approaches and give additional calculations for the case m-1 out of m trials.

(j) <u>Target detection systems</u>. Various target detection systems react to a "quota" of responses where a quota is a set of m consecutive trials in which there are at least k successes. Goldman and Bender (1962) consider events where a run immediately follows a quota, and derive the distribution of waiting times till the event. The derivation is done for the case of where the run size is at least equal to the number of successes in the quota, thereby allowing treatment of the event as a recurrent event.

For the case of run size less than the number of successes in the quota, Brookner (1966) describes a general procedure. The procedure is to transform the original sequence of overlapping states to an m-state first-order Markov chain. The method is fairly general in that the original trials need not be independent. However, the method becomes unwieldy for m and n of moderate size because of the resulting large number of states in the transformed matrix. The method is feasible for quotas of a few successes in a few trials.

(k) Quality control. In quality control, a point outside the control chart limits for the mean signals that the process may be out-of-control. In addition, runs of observations above or below the mean, even if within the control chart limits, are additional warning signs. Roberts (1958) develops a series of zone tests that combine these two ideas. Roberts sets up zones within the usual control limits such that if too many observations within a consecutive group of observations fall within the zone, the process is called out-of-control. (For example, "too many" might be 4 out of 7 consecutive observations falling between 1.28 and 3 sigma limits above the mean). To study the operating characteristics of such plans, one needs the probabilities of quotas. Saperstein (1976) gives a variety of other zone tests.

(l) Acceptance sampling. Troxell (1972) develops acceptance sampling plans based on a quota of batches being unacceptable. The special cases considered are 2 out of m, and 3 out of m batches. Given present results for $P(n \text{ in } m | N \text{ in } M)$, the acceptance sampling plans can be generalized to the case of n out of m batches.

(m) Faults in a sequence of trials. Leslie (1967) describes some ingenious uses of a type of generalized run different than but related to quotas. The occurrence of many leaky pipe joints within a given distance may justify replacement. Faulty sleepers in railway tracks can cause problems if too many occur within a given spacing. Several genetic mutations within a given distance on a chromosome may lead to defects.

(n) Other applications. Ecologists study clustering of diseased plants in transects through a field. Meteorologists investigate the alternating of