

BU-149-M

A MATHEMATICAL MODEL FOR THE DISTRIBUTION OF
LENGTHS OF CHROMOSOMAL DEFICIENCIES
INVOLVING A SPECIFIC LOCUS

Technical Report No. 10

Department of Navy
Office of Naval Research

Contract No. Nonr-409(39)
Project No. (NR 042-212)

K. H. Thompson, B. Wallace, and W. T. Federer
Biometrics Unit and Plant Breeding Department
New York State College of Agriculture
Cornell University
Ithaca, New York

This work was supported in part by the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the United States Government.

A MATHEMATICAL MODEL FOR THE DISTRIBUTION OF LENGTHS OF
CHROMOSOMAL DEFICIENCIES INVOLVING A SPECIFIC LOCUS*

K. H. Thompson, B. Wallace, and W. T. Federer

Abstract

A mathematical model of random two-hit breakage and deficiency is developed for the condition that the chromosomal segment lost includes an arbitrary but specific locus. Both a discrete band model and a continuous chromosome model are considered in presenting theoretical distributions for deficiency lengths, deficiency midpoints, and other related characteristics. The problem of estimating the location of essential loci is discussed for the continuous model.

*Biometrics Unit Mimeo No. BU-149-M, Cornell University.

A MATHEMATICAL MODEL FOR THE DISTRIBUTION OF LENGTHS OF
CHROMOSOMAL DEFICIENCIES INVOLVING A SPECIFIC LOCUS*

K. H. Thompson, B. Wallace, and W. T. Federer

Experimental distributions of breaks and chromosomal recombinations in Drosophila melanogaster have been studied by several authors in attempts to throw light on the mechanism of the origin of chromosomal rearrangements. Stadler (1932), Bauer, et al. (1938), Catcheside (1938), Fano (1943), Kaufmann (1946) and many others have contributed to this literature. While there is evidence (Bauer, Demerec, and Kaufmann, 1938) that breakage, either spontaneous or X-ray induced, occurs largely at random throughout the length of a chromosome, the accumulated evidence of observed chromosomal alterations indicates that recombination may not be random.

In obtaining alterations to test the hypothesis of random breakage and recombination, inversions and deficiencies are particularly useful because they admit fairly standard cytological or morphological techniques of identification. Inversion loops may be detected and measured with reasonable facility in the salivary chromosomes of D. melanogaster, and a series of 49 induced inversions of the X chromosome was obtained by Bauer et al. (1938). This distribution was compared by Federer, Steel, and Wallace (1961) to several mathematical models.

The cytological detection of chromosomal deficiencies, however, is impractical unless the deficiency includes a mutant marker. This technique was used by Demerec and Fano (1941) to obtain a series of 37 Notch deficiencies of the X chromosome. These authors observed that the expected frequency of deficiencies involving n bands out of a total number of N bands is equal to $2(N-1-n)/(N-1)(N-2)$ and that the a priori probability that any specific band is included in a deficiency of length n is approximately equal to $n/(N-2)$.

These results may be obtained by assuming that any two random and independent breaks may lead with equal probability to a deficiency or to an inversion of the chromosomal segment involved. This assumption probably does not obtain, however, for large chromosomal segments which might be expected to be more viable

*Biometrics Unit Mimeo No. BU-149-M, Cornell University

as inversions than as deficiencies, nor for very short segments which might be lost as deficiencies more often than recombining as inversions. This model also does not take into consideration the shorter deficiencies which may arise as a result of a single break or those deficiencies associated with more complicated rearrangements. However, in spite of the fact that we know chromosomal alteration to be a complex mechanism, the two break model proposed affords a useful hypothesis and perhaps a reasonable approximation.

The calculations of Demerec and Fano (1941) are easily extended to obtain the exact distribution of deficiency lengths, given that a specific locus is included. It is convenient to assume that breakage is equally likely between any two bands of a chromosome whose length is divided into N distinct bands. Hence, excluding the two terminal bands, there are $(N-1)(N-2)/2$ different and equally likely deficiencies, of which $(N-1-n)$ are of length n ; $n=1,2,\dots,N-2$. It may further be determined that there are $(a-1)(N-a)$ different and equally likely deficiencies which include a specific band a ; $a=2,3,\dots,N-1$. Simple counting techniques suffice to show that the frequency of deficiencies of length n which include band a ; $a \leq (N+1)/2$, is equal to

$$\begin{array}{ll} \frac{n}{(a-1)(N-a)} & \text{if } 1 \leq n < (a-1) , \\ \frac{1}{(N-a)} & \text{if } (a-1) \leq n \leq (N-a) , \text{ and} \\ \frac{N-n-1}{(a-1)(N-a)} & \text{if } (N-a) < n \leq (N-2) . \end{array}$$

The frequency of deficiencies of length n which include band a is symmetrical in a . For N odd or even, the distribution for a is equal to the distribution for $(N-a+1)$; hence, the frequency of deficiencies of length n which include band a ; $a > (N+1)/2$, is equal to

$$\begin{array}{ll} \frac{n}{(N-a)(a-1)} & \text{if } 1 \leq n < (N-a) , \\ \frac{1}{(a-1)} & \text{if } (N-a) \leq n \leq (a-1) , \text{ and} \\ \frac{N-n-1}{(N-a)(a-1)} & \text{if } (a-1) < n \leq (N-2) . \end{array}$$

The calculations of Demerec and Fano (1941) may be extended further to include the exact proportion of deficiencies of fixed length n which include a specific band a . This frequency may be seen to depend upon the location of band a . For fixed $n < N/2$, the proportion of deficiencies of length n which include a specific band a is equal to

$$\begin{aligned} \frac{a-1}{N-n-1} & \quad \text{if} \quad 2 \leq a \leq n, \\ \frac{n}{N-n-1} & \quad \text{if} \quad n < a < (N-n+1), \quad \text{and} \\ \frac{N-a}{N-n-1} & \quad \text{if} \quad (N-n+1) \leq a \leq (N-1). \end{aligned}$$

The number of distinct deficiencies which include band a is symmetrical in n ; that is, the number of deficiencies of fixed length n which include any given band a is equal to the number of deficiencies of fixed length $(N-n-1)$ which include the same band a . Hence, for fixed $n \geq N/2$, the proportion of deficiencies of length n which include a specific band a is equal to

$$\begin{aligned} \frac{a-1}{N-n-1} & \quad \text{if} \quad 2 \leq a \leq (N-1-1), \\ 1 & \quad \text{if} \quad (N-n-1) < a < (n+2), \quad \text{and} \\ \frac{N-a}{N-n-1} & \quad \text{if} \quad (n+2) < a < (N-1). \end{aligned}$$

The above frequencies and probabilities are based on a discrete band concept which is commensurate with the techniques of cytological identification and measurement. However, the frequency distribution of deficiency lengths, given that a specific locus a is included, may also be obtained by assuming that breakage may occur at any point of a chromosome of length c . If we assume as before that exactly two independent and random breaks induce a single deficiency, we may let the independent random variables X and Y denote the position of the first and second breaks, respectively. Thus, the joint distribution of X and Y now represent our mathematical model for random two-hit breakage and deficiency and is the bivariate uniform distribution

$$f(x,y) = \frac{1}{c^2} \quad \text{for } 0 < x < c, \quad 0 < y < c,$$
$$= 0 \quad \text{otherwise.}$$

By applying the appropriate linear transformations to X and Y , the frequency distributions of deficiency lengths, midpoints, and other characteristics of interest may be obtained. In particular, the distributions of $V=|Z|=|X-Y|$ and $W=(X+Y)/2$ over the region $(0 < x < a < y < c, \quad 0 < y < a < x < c)$ represent the distributions of deficiency lengths and midpoints, given that the deficiency includes a given locus a .

We now proceed to compute the distributions of V and W . Denoting the region of definition by R , by which we mean that (x,y) for fixed $0 < a < c$ satisfies either $0 < x < a < y < c$ or $0 < y < a < x < c$, we begin by observing that

$$g(x,y|R) = \frac{1}{2a(c-a)} \quad \text{for } (x,y) \in R,$$
$$= 0 \quad \text{otherwise.}$$

The region R over which this joint density is defined is represented graphically in Figure 1.

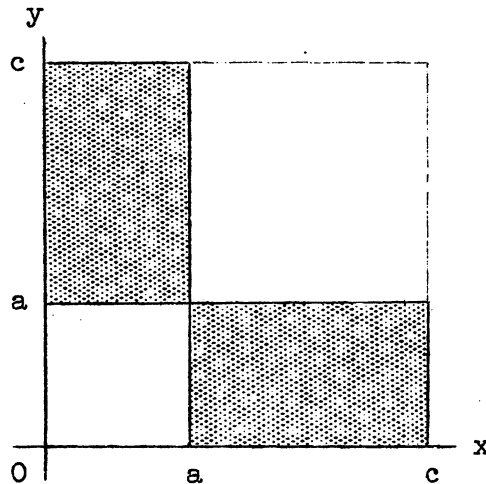


Figure 1. Region for which $g(x,y|R) \neq 0$.

It is now convenient to compute the joint distribution of $Z=(X-Y)$ and $W=(X+Y)/2$ over the same region R . Since the Jacobean of the transformation is equal to 1, $g(w,z|R)$ is uniformly distributed as

$$\begin{aligned}
 g(w,z|R) &= \frac{1}{2a(c-a)} \quad \text{for } 0 < 2a < c, \quad a < 2w < 2a, \quad 2(a-w) < |z| < 2w ; \\
 & \quad \text{for } 0 < 2a < c, \quad 2a < 2w < c, \quad 2(w-a) < |z| < 2w ; \\
 & \quad \text{for } 0 < 2a < c, \quad c < 2w < c+a, \quad 2(w-a) < |z| < 2(c-w) ; \\
 & \quad \text{for } c < 2a < 2c, \quad a < 2w < c, \quad 2(a-w) < |z| < 2w ; \\
 & \quad \text{for } c < 2a < 2c, \quad c < 2w < 2a, \quad 2(a-w) < |z| < 2(c-w) ; \\
 & \quad \text{for } c < 2a < 2c, \quad 2a < 2w < c+a, \quad 2(w-a) < |z| < 2(c-w) ; \\
 & = 0 \quad \text{otherwise.}
 \end{aligned}$$

The region of definition is represented graphically in Figure 2 below.

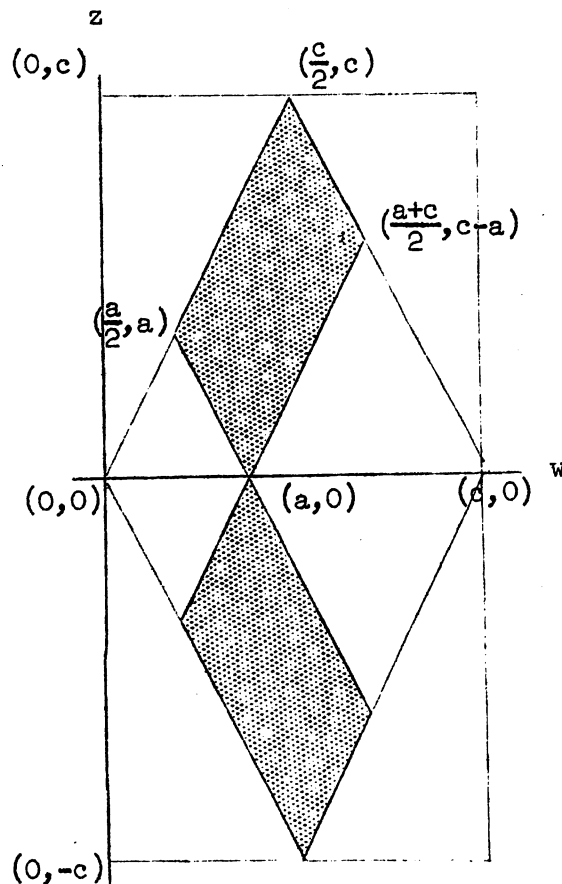


Figure 2. Region for which $g(w,z|R) \neq 0$.

The joint distribution of $V=|Z|=|X-Y|$ and $W=(X+Y)/2$ is now easily obtained because of the symmetry of $g(w,z|R)$ in z and becomes

$$\begin{aligned}
 g(w,v|R) &= \frac{1}{a(c-a)} && \text{for } 0 < 2a < c, && a < 2w < 2a, && 2(a-w) < v < 2w ; \\
 & && \text{for } 0 < 2a < c, && 2a < 2w < c, && 2(w-a) < v < 2w ; \\
 & && \text{for } 0 < 2a < c, && c < 2w < c+a, && 2(w-a) < v < 2(c-w) ; \\
 & && \text{for } c < 2a < 2c, && a < 2w < c, && 2(a-w) < v < 2w ; \\
 & && \text{for } c < 2a < 2c, && c < 2w < 2a, && 2(a-w) < v < 2(c-w) ; \\
 & && \text{for } c < 2a < 2c, && 2a < 2w < c+a, && 2(w-a) < v < 2(c-w) ; \\
 & && = 0 && \text{otherwise.}
 \end{aligned}$$

The region of definition may be represented as the shaded area of the first quadrant of Figure 2 where the z axis is replaced by v .

Finally, the distribution of the lengths of deficiencies $V=|X-Y|$ which cover the locus a is computed as $g(v|R) = \int g(w,v|R)dw$. Hence, for fixed $0 < 2a < c$,

$$\begin{aligned}
 g(v|R) &= \frac{1}{a(c-a)} \int_{\frac{2a-v}{2}}^{\frac{2a+v}{2}} dw = \frac{v}{a(c-a)} && \text{for } 0 < v < a, \\
 &= \frac{1}{a(c-a)} \int_{\frac{v}{2}}^{\frac{2a+v}{2}} dw = \frac{1}{c-a} && \text{for } a < v < c-a, \\
 &= \frac{1}{a(c-a)} \int_{\frac{v}{2}}^{\frac{2c-v}{2}} dw = \frac{c-v}{a(c-a)} && \text{for } c-a < v < c, \text{ and} \\
 &= 0 && \text{otherwise.}
 \end{aligned}$$

For fixed $c < 2a < 2c$,

$$\begin{aligned}
 g(v|R) &= \frac{1}{a(c-a)} \int_{\frac{2a-v}{2}}^{\frac{2a+v}{2}} dw = \frac{v}{a(c-a)} && \text{for } 0 < v < c-a, \\
 &= \frac{1}{a(c-a)} \int_{\frac{2a-v}{2}}^{\frac{2c-v}{2}} dw = \frac{1}{a} && \text{for } c-a < v < a, \\
 &= \frac{1}{a(c-a)} \int_{\frac{v}{2}}^{\frac{2v-c}{2}} dw = \frac{c-v}{a(c-a)} && \text{for } a < v < c, \text{ and} \\
 &= 0 && \text{otherwise.}
 \end{aligned}$$

It is now quite clear that without loss of generality we need consider only $0 < 2a < c$; that is, the end of the chromosome from which we measure the locus a is unimportant. The distribution of V for $0 < 2a < c$ is shown graphically in Figure 3 below and shows the symmetry in $a=c-a$.

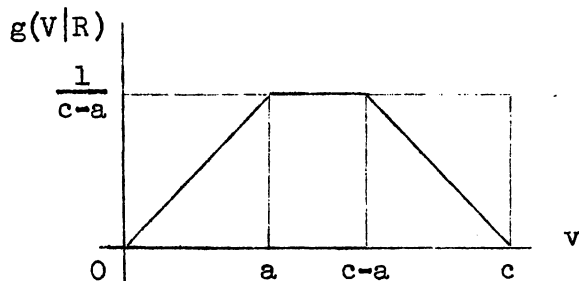


Figure 3. Distribution of $V=|Z|=|X-Y|$ = deficiency length given that the locus $0 < 2a < c$ is included.

The mean and variance of V are $c/2$ and $[(c-a)^2 + a^2]/12$, respectively.

The distribution of the midpoints of deficiencies which cover the locus a may be computed similarly as $g(w|R) = \int g(w,v|R) dv$. For fixed $0 < 2a < c$,

$$g(w|R) = \frac{1}{a(c-a)} \int_{2(a-w)}^{2w} dv = \frac{2(2w-a)}{a(c-a)} \quad \text{for } a < 2w < 2a,$$

$$= \frac{1}{a(c-a)} \int_{2(w-a)}^{2w} dv = \frac{2}{c-a} \quad \text{for } 2a < 2w < c,$$

$$= \frac{1}{a(c-a)} \int_{2(w-a)}^{2(c-w)} dv = \frac{2(c-2w+a)}{a(c-a)} \quad \text{for } c < 2w < a+c, \text{ and}$$

$$= 0 \quad \text{otherwise.}$$

For fixed $c < 2a < 2c$,

$$g(w|R) = \frac{1}{a(c-a)} \int_{2(a-w)}^{2w} dv = \frac{2(2w-a)}{a(c-a)} \quad \text{for } a < 2w < c,$$

$$= \frac{1}{a(c-a)} \int_{2(a-w)}^{2(c-w)} dv = \frac{2}{a} \quad \text{for } c < 2w < 2a,$$

$$= \frac{1}{a(c-a)} \int_{2(w-a)}^{2(c-w)} dv = \frac{2(c-2w+a)}{a(c-a)} \quad \text{for } 2a < 2w < a+c, \text{ and}$$

$$= 0 \quad \text{otherwise.}$$

This distribution is shown graphically in Figures 4 and 5 below to show the nature of the symmetry in $a=c-a$.

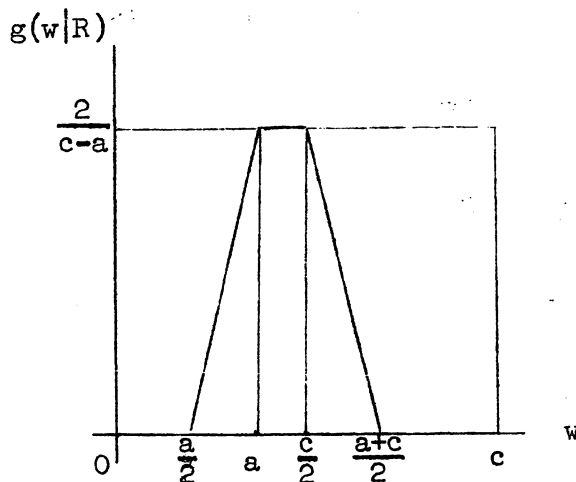


Figure 4. Distribution of $W=(X+Y)/2$ = deficiency midpoints given that the locus $0 < 2a < c$ is included.

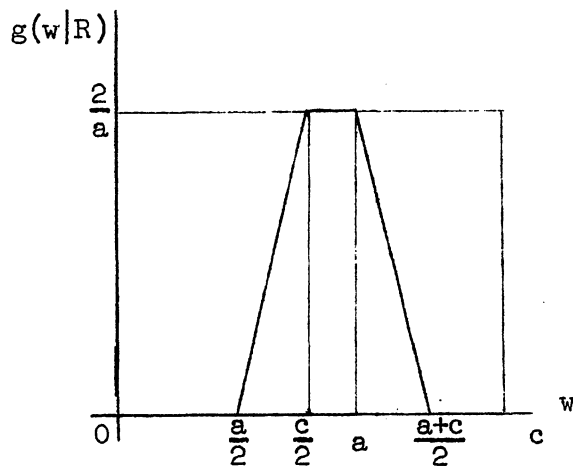


Figure 5. Distribution of $W=(X+Y)/2$ = deficiency midpoints given that the locus $c < 2a < 2c$ is included.

The mean and variance of W are $(c+2a)/4$ and $[(c-a)^2 + a^2]/48$, respectively.

The conditional distributions may now be computed as $g(w|v,R) = g(w,v|R)/g(v|R)$ and $g(v|w,R) = g(w,v|R)/g(w|R)$. The conditional distribution of W , the midpoint of the deficiencies, for any fixed value of V , the deficiency length, given that the deficiency includes the locus a , is given as

$$\begin{aligned}
 g(w|v,R) &= \frac{1}{v} && \text{for } 0 < 2a < c, && 0 < v < a, && 2a-v < 2w < v+2a; \\
 &= \frac{1}{a} && \text{for } 0 < 2a < c, && a < v < c-a, && v < 2w < v+2a; \\
 &= \frac{1}{c-v} && \text{for } 0 < 2a < c, && c-a < v < c, && v < 2w < 2c-v; \\
 &= \frac{1}{v} && \text{for } c < 2a < 2c, && 0 < v < c-a, && 2a-v < 2w < v+2a; \\
 &= \frac{1}{c-a} && \text{for } c < 2a < 2c, && c-a < v < a, && 2a-v < 2w < 2c-v; \\
 &= \frac{1}{c-v} && \text{for } c < 2a < 2c, && a < v < c, && v < 2w < 2c-v; \\
 &= 0 && \text{otherwise.}
 \end{aligned}$$

The conditional distribution of V, the length of the deficiency, for any fixed value of W, the deficiency midpoint, given that the deficiency includes the locus a, is given as

$$\begin{aligned}
 g(v|w,R) &= \frac{1}{2(2w-a)} && \text{for } 0 < 2a < c, && a < 2w < 2a, && 2(a-w) < v < 2w; \\
 &= \frac{1}{2a} && \text{for } 0 < 2a < c, && 2a < 2w < c, && 2(w-a) < v < 2w; \\
 &= \frac{1}{2(c-2w+a)} && \text{for } 0 < 2a < c, && c < 2w < c+a, && 2(w-a) < v < 2(c-w); \\
 &= \frac{1}{2(2w-a)} && \text{for } c < 2a < 2c, && a < 2w < c, && 2(a-w) < v < 2w; \\
 &= \frac{1}{2(c-a)} && \text{for } c < 2a < 2c, && c < 2w < 2a, && 2(a-w) < v < 2(c-w); \\
 &= \frac{1}{2(c-2w+a)} && \text{for } c < 2a < 2c, && 2a < 2w < c+a, && 2(w-a) < v < 2(c-w); \\
 &= 0 && \text{otherwise.}
 \end{aligned}$$

The mathematical model discussed to this point has assumed that the loss of a chromosomal segment is based solely upon the location of two chance breaks in the chromosome and not upon the distance between the two breaks. In fact,

chromosomal deficiencies of large lengths involving a significant portion of a chromosome are rarely observed. One reason may be that viability is in some way proportional to the length of the deleted portion of the chromosome. It is also likely that there exist several loci distributed along the length of every chromosome which are essential to viability. Deficiencies including these loci would therefore never be observed. However, without loss of generality, the mathematical model that we have considered is applicable to this situation where the locus a is bounded at an unknown distance on each side by an essential locus; say a_s and a_t where $0 < a_s < a_t < c$. a_s and a_t now become the effective ends of the chromosome insofar as deficiencies which include the locus a are concerned although their locations must be estimated.

If we let $S=W-(V/2)$ and $T=W+(V/2)$ represent respectively the proximal and distal breaks of a deficiency which includes the locus a ; that is, such that $0 < a_s < s < a < t < a_t < c$, it is easy to show from previous considerations of the model that

$$g(s|R) = \frac{1}{a-a_s} \quad \text{for } 0 < a_s < s < a,$$

$$= 0 \quad \text{otherwise, and}$$

$$g(t|R) = \frac{1}{a_t-a} \quad \text{for } a < t < a_t < c,$$

$$= 0 \quad \text{otherwise.}$$

In fact, S and T are independent, and their distributions are useful in estimating a_s and a_t . If we have n independent deficiencies which include the locus a , then

$$\xi = \left(\frac{n+1}{n}\right)[\max(T_1, \dots, T_n)] - \frac{a}{n}$$

is an unbiased estimator of a_t , and

$$\zeta = \left(\frac{n+1}{n}\right)[\min(S_1, \dots, S_n)] - \frac{a}{n}$$

is an unbiased estimator of a_s . Since $\max(T_1, \dots, T_n)$ and $\min(S_1, \dots, S_n)$ are

complete and sufficient for a_t and a_s respectively, ξ and ζ have minimum variance among all unbiased estimators of a_t and a_s ; namely,

$$\text{Var}(\xi) = \frac{(a_t - a)^2}{n(n+2)}, \quad \text{and}$$

$$\text{Var}(\zeta) = \frac{(a - a_s)^2}{n(n+2)} .$$

In particular, such unbiased estimators of a_t and a_s as $(2\bar{w} - a + \bar{v})$ and $(2\bar{w} - a - \bar{v})$ respectively, where $\bar{w} = \Sigma W_1/n$ and $\bar{v} = \Sigma V_1/n$, are inadmissible under this model.

Literature Cited

- Bauer, H., Demerec, M., and Kaufmann, B. P. 1938. X-Ray induced chromosomal alterations in Drosophila melanogaster. Genetics 23: 610-630.
- Catcheside, D. G. 1938. The effect of X-ray dosage upon the frequency of induced structural changes in the chromosomes of Drosophila melanogaster. J. Genet. 36: 307-320.
- Demerec, M. and Fano, U. 1941. Mechanism of the origin of X-ray induced notch deficiencies in Drosophila melanogaster. Proc. Nat. Acad. Sci. 27: 24-31.
- Fano, U. 1943. Mechanism of induction of gross chromosomal rearrangements in Drosophila sperms. Proc. Nat. Acad. Sci. 29: 12-18.
- Federer, W. T., Steel, R. G. D., and Wallace, B. 1961. A mathematical model for lengths and mid-points of inversions in chromosomes. Biometrics Unit Bulletin 129-M, Cornell University.
- Kaufmann, B. P. 1946. Organization of the chromosome. I. Break distribution and chromosome recombination in Drosophila melanogaster. J. Exp. Zool. 102: 293-320.
- Stadler, L. J. 1932. On the genetic nature of induced mutations in plants. Proc. Sixth Int. Cong. Genetics I: 274-294.

CORNELL UNIVERSITY - BIOMETRICS UNIT

Distribution List for Unclassified Technical Reports

Contract Nonr-401(39)

Project (NR 042-212)

Head, Logistics and Mathematical Statistics Branch Office of Naval Research Washington 25, D. C.	3	Professor W. G. Cochran Department of Statistics Harvard University Cambridge, Massachusetts	1
Commanding Officer Office of Naval Research 346 Broadway New York 13, New York	1	Dr. C. Clark Cockerham Institute of Statistics North Carolina State College Raleigh, North Carolina	1
ASTIA Document Service Center Arlington Hall Station Arlington 12, Virginia	10	Professor Cyrus Derman Dept. of Industrial Engineering Columbia University New York 27, New York	1
Institute for Defense Analyses Communications Research Div. von Neumann Hall Princeton, New Jersey	1	Professor Benjamin Epstein Applied Mathematics and Statistics Laboratory Stanford University Stanford, California	1
Technical Information Officer Naval Research Laboratory Washington 25, D. C.	6	Dr. W. T. Federer Biometrics Unit Plant Breeding Department Cornell University Ithaca, New York	1
Professor T. W. Anderson Department of Mathematical Statistics Columbia University New York 27, New York	1	Dr. R. J. Freund Department of Statistics and Statistical Laboratory Virginia Polytechnic Institute Blacksburg, Virginia	1
Professor Z. W. Birnbaum Laboratory of Statistical Research Department of Mathematics University of Washington Seattle 5, Washington	1	Professor H. P. Goode Dept. of Industrial and Engineering Administration Cornell University Ithaca, New York	1
Professor A. H. Bowker Applied Mathematics and Statistics Laboratory Stanford University Stanford, California	1	Professor W. Hirsch Institute of Mathematical Sciences New York 3, New York	1
Professor Ralph A. Bradley Department of Statistics Florida State University Tallahassee, Florida	1	Professor Harold Hotelling, Associate Director Institute of Statistics University of North Carolina Chapel Hill, North Carolina	1
Dr. John W. Cell Department of Mathematics North Carolina State College Raleigh, North Carolina	1	Professor Oscar Kempthorne Statistics Laboratory Iowa State University Ames, Iowa	1
Professor Herman Chernoff Applied Mathematics and Statistics Laboratory Stanford University Stanford, California	1		

Professor Gerald J. Lieberman Applied Mathematics and Statistics Laboratory Stanford University Stanford, California	1	Professor L. J. Savage Statistical Research Laboratory Chicago University Chicago 37, Illinois	1
Dr. Arthur E. Mace Battelle Memorial Institute 505 King Avenue Columbus 1, Ohio	1	Professor W. L. Smith Statistics Department University of North Carolina Chapel Hill, North Carolina	1
Professor J. Neyman Department of Statistics University of California Berkeley 4, California	1	Professor Frank Spitzer Department of Mathematics University of Minnesota Minneapolis, Minnesota	1
Dr. F. Oberhettinger Department of Mathematics Oregon State College Corvallis, Oregon	1	Dr. H. Teicher Statistical Laboratory Engineering Administration Bldg. Purdue University Lafayette, Indiana	1
Professor Herbert Robbins Mathematical Statistics Dept. Fayerweather Hall Columbia University New York 27, New York	1	Professor M. B. Wilk Statistics Center Rutgers-The State University New Brunswick, New Jersey	1
Professor Murray Rosenblatt Department of Mathematics Brown University Providence 12, Rhode Island	1	Professor S. S. Wilks Department of Mathematics Princeton University Princeton, New Jersey	1
Professor H. Rubin Department of Statistics Michigan State University East Lansing, Michigan	1	Professor J. Wolfowitz Department of Mathematics Lincoln Hall Cornell University Ithaca, New York	1
Professor I. R. Savage School of Business Admin. University of Minnesota Minneapolis, Minnesota	1		