A MATHEMATICAL MODEL FOR LENGTHS AND MID-POINTS OF INVERSIONS IN CHROMOSOMES

Technical Report No. 3

Department of Navy

Office of Naval Research

Contract No. Nonr-401(39)
Project No. (NR 042-212)

W. T. Federer, R. G. D. Steel*, and Bruce Wallace
Department of Plant Breeding
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.

* Now at North Carolina State, Raleigh, North Carolina.

A MATHEMATICAL MODEL FOR LENGTHS AND MID-POINTS OF INVERSIONS IN CHROMOSOMES

W. T. Federer,[1] R. G. D. Steel,[2] and Bruce Wallace[3]

BU-129-M January, 1961

In many ways the karyotype of a species is remarkably constant. The number of chromosomes carried by each individual member is, with few exceptions, the same as that of all others; even the shape of individual chromosomes as determined by the position of the centromere is usually invariable.

The arrangement of gene loci within a chromosome on the other hand, often varies considerably from individual to individual. In Dipterous flies, whose giant chromosomes offer an excellent opportunity for study, species after species has been found to possess a wealth of contrasting gene arrangements within the same chromosome. The origin of this variation lies almost exclusively in the inversion of chromosomal segments -- a phenomenon which requires that a chromosome be broken in two places, that the segment between the breaks reverse its position within the chromosome, and that the broken ends heal with the segment in its new position.

Unless the wealth of inversions in a species is overwhelmingly large, a phylogeny of gene arrangements can be constructed such that each gene arrangement differs from another by a single two-break inversion. The logic underlying such phylogenies is supported by the frequent discovery of gene arrangements which were previously postulated as hypothetical ones. We must re-emphasize that phylogenies of this sort can be constructed only in the case of certain Dipteran species. It is a well known fact, nevertheless, that inversions are found in a great many plant and animal species; their origin in these species is undoubtedly comparable to that in Diptera.

Why do wild populations of so many species retain a variety of chromosomal rearrangements? Again a definitive answer is available only in the case of some Drosophila species. Dobzhansky and his students have shown repeatedly that individuals carrying two different gene arrangements are superior to individuals homozygous for one or the other of the various inversions in many, if not all, components of fitness. Although experimental evidence is lacking for other

---

[1] Biometrics Unit, Plant Breeding Department, Cornell University

[2] Now at North Carolina State, Raleigh, North Carolina

[3] Plant Breeding Department, Cornell University

organisms, the simplest explanation for the existence of any polymorphic system is that based on the selective superiority of heterozygous individuals.

It is highly unlikely that the retention of an inversion in any population depends upon the inversion itself; that is, upon the chromosomal breaks, upon the new arrangement, or upon position effects. Indeed, Paget has shown that radiation-induced inversions have an average deleterious effect on fitness. The advantage of inversion heterozygotes which underlies the retention of two or more inversions in a population appear to result from the interaction of blocks of genes; blocks which are held intact by the suppression of normal gene recombination.

We propose to undertake an analysis of the genetic basis of heterosis by utilizing the above facts. Indeed, Sprague and Chao, independently, have used two inversions in analyzing heterosis in corn. We feel, however, that even greater opportunities for analysis are present in Drosophila. The radiation genetics of Drosophila is a well-developed field. These flies breed rapidly; their giant chromosomes can be analyzed with an accuracy unsurpassed in any other group. There already exists an enormous literature (recently reviewed by daCunha) on the inversions which are to be found in a large number of species. Finally, racial and strain hybrids are known to exhibit heterosis.

In brief, our procedure calls for an analysis of the distribution of sizes and position of newly induced inversions retained in populations of hybrid origin but which are, with the exception of the induced inversions, structurally homozygous. This will tell us the location of genes which confer heterosis in these populations as well as the size of the gene blocks needed to confer heterosis. Hybrid populations started by crossing strains from a number of localities can be compared. Material accumulated during these studies can be utilized in attacking many additional problems.

The remainder of this report consists of a mathematical model, in which chromosomes are treated as homogeneous strings subject to breakage at any point, which predicts the theoretical distribution of lengths of inverted chromosomal segments as well as their positions. The distributions of naturally occurring inversions can be compared with those predicted by this model; this comparison will entail library research only. The distribution of newly induced inversions must also be compared with theoretical expectations. This comparison is an essential test of the validity of the present model; it is important that the distribution of newly induced inversions be known so that distortions resulting

from selection within populations can be recognized. Although some information is available in the literature for this study, it is quite possible that the available data will have to be augmented by a new cytological study.

For the purpose of making the theory simple, first assume that there is one and only one inversion per chromosome. That is, only the chromosomes with one inversion will be considered. Secondly, assume that a break is equally likely along any part of the chromosome and that the position of the first break is independent of the position of the second break on the chromosome. This means that the first break (x) and the second break (y) each follow the uniform distribution and that the joint distribution of the two breaks is the product of two uniform distributions.

In mathematical terms, for a chromosome of length c,

$$f(x) = \frac{1}{c} \qquad 0 < x < c$$

$$= 0 \qquad \text{otherwise}$$

$$f(y) = \frac{1}{c} \qquad 0 < y < c$$

$$= 0 \qquad \text{otherwise}$$

(The probability that x (or y) falls in any given interval is $\frac{1}{c}$ times the length of the given interval.)

The joint density function of x and y is:

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

and the joint probability function is

$$P\left\{0 < x < x_0, \ 0 < y < y_0\right\}$$
$$= \int_0^{y_0} \int_0^{x_0} f(x,y)\,dx\,dy = \frac{1}{c^2} \int_0^{y_0} \int_0^{x_0} dx\,dy$$

The region over which the joint density function is defined may be represented graphically as Figure 1.

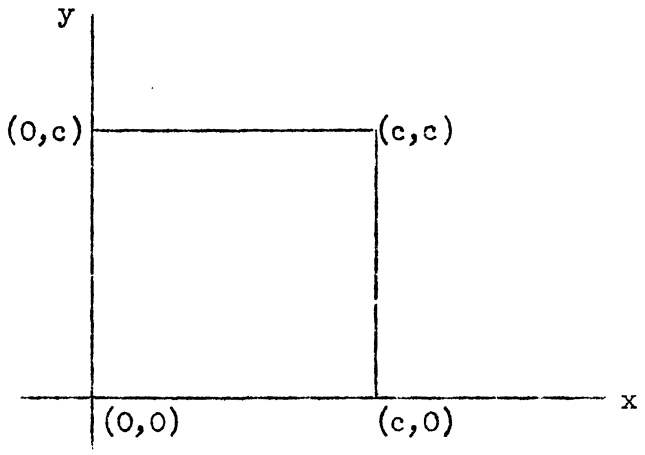

Figure 1. Region for which $f(x,y) \neq 0$.

The problems are to find the distributions of (1) the lengths of the inverted part of the chromosome, (2) the mid-points of the inversions regardless of length, and (3) the mid-points of inversions of fixed length. That is, it is desired to obtain the distributions of $|x-y|$, of $\frac{x+y}{2}$ for all lengths and of $(x+y)/2$ for any fixed value of $|x-y|$.

First, set $z=x-y$ and $w=(x+y)/2$. Then the region of definition of the function $g(w,z)$ is given in Figure 2.

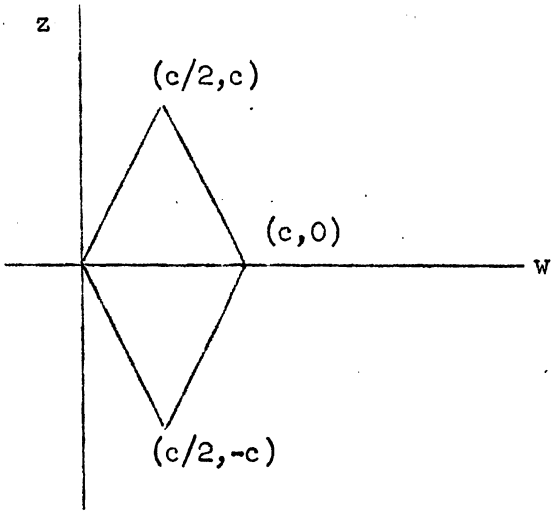

Figure 2. Region for which $g(w,z) \neq 0$.

To find the joint distribution of w and z, we first find

$$J^{-1} = \pm \begin{vmatrix} \frac{\delta w}{\delta x} & \frac{\delta z}{\delta x} \\ \frac{\delta w}{\delta y} & \frac{\delta z}{\delta y} \end{vmatrix} = - \begin{vmatrix} 1/2 & 1 \\ 1/2 & -1 \end{vmatrix} = 1$$

Now

$$f(x,y) = \frac{1}{c^2} J = \frac{1}{c^2}$$

$$= g(w,z), \quad \begin{cases} -2w < z < 2w < c \\ -2c+2w < z < 2c-2w, \quad c/2 < w < c \end{cases}$$

$$= 0 \qquad \text{otherwise}$$

Let $v=|z|=|x-y|$. Then because of the symmetry and uniformity of $g(w,z)$, we have

$$h(w,v) = \frac{2}{c^2}, \quad \begin{cases} 0 < v < 2w < c \\ 0 < v < 2c-2w, \quad c/2 < w < c \end{cases}$$

$$= 0 \qquad \text{otherwise}$$

This is the joint distribution of the two variables of interest and is defined over the region shown in Figure 3.

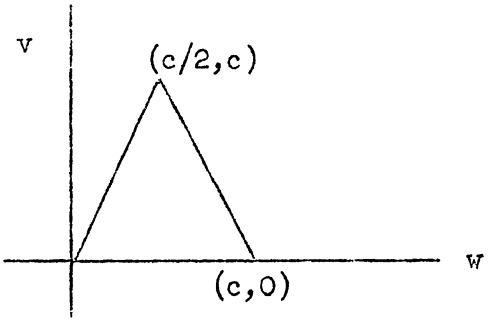

Figure 3. Region for which $h(w,v) \neq 0$

From the distribution of h(w,v),we now find the three distributions of interest, together with the means, variances and other moments of the variables.

First, we find the distribution of lengths of inversions, namely, $h_1(v)$.

$$h_1(v) = \int h(w,v)dw$$

$$= \begin{cases} \frac{2}{c^2} \int_{v/2}^{c/2} dw , & 0<v<c \\ \frac{2}{c^2} \int_{c/2}^{c-v/2} dw , & 0<v<c \end{cases}$$

$$= \frac{2(c-v)}{c^2} , \quad 0<v<c$$

This function is shown graphically in Figure 4.

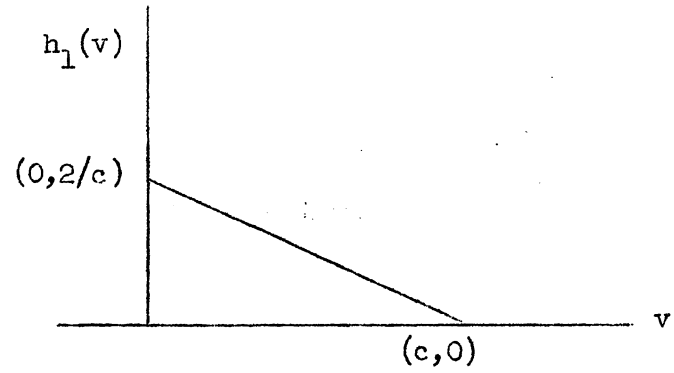

Figure 4. Distribution of $v=|z|=|x-y|$ = length of break.

The mean of v is given by

$$\mu = E(v) = \frac{2}{c^2} \int_0^c (c-v)v \, dv$$

$$= c/3$$

The moments about the mean may be obtained from the following moment generating function.

$$m(t) = E(e^{t(v-c/3)})$$

$$= \frac{2}{c^2}\int_0^c (c-v)e^{t(v-c/3)}dv$$

$$= \frac{2}{c^2}\left[\frac{2}{3}\frac{ce^{t(v-c/3)}}{t}\Big|_0^c - \frac{e^{t(v-c/3)}}{t^2}\left(t\left\{v-c/3\right\}-1\right)\Big|_0^c\right]$$

$$= \frac{2}{c^2}\frac{e^{\frac{2}{3}ct}}{t^2} - \frac{2e^{-\frac{ct}{3}}}{ct} - \frac{2e^{-\frac{ct}{3}}}{c^2t^2}$$

$$= 1 + 2c\frac{(2^3-3(3)+1)}{3^3(3)(2)}\cdot\frac{t}{1} + 2c^2\frac{(2^4+3(4)-1)}{3^4(4)(3)}\cdot\frac{t^2}{2!}$$

$$+ \cdots + 2c^n\frac{\{2^{n+2}+(-1)^n 3(n+2)+(-1)^{n+1}\}}{3^{n+2}(n+2)(n+1)}\cdot\frac{t^n}{n!}$$

$$+ \cdots$$

The coefficient of $t^2/2!$ is the variance,

$$\sigma^2 = \frac{2c^2(2^4+3(4)-1)}{3^4(4)(3)}$$

$$= c^2/18$$

Secondly, we required the distributions of midpoints of inversions. This is simply $h_2(w)$.

$$h_2(w) = \int h(w,v)dv$$

$$= \begin{cases} \frac{2}{c^2}\int_0^{2w} dv, & 0<w<c/2 \\ \frac{2}{c^2}\int_0^{2c-2w} dv, & c/2<w<c \end{cases}$$

$$= \begin{cases} \frac{4w}{c^2}, & 0<w<c/2 \\ \frac{4(c-w)}{c^2}, & c/2<w<c \end{cases}$$

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

This function is shown in Figure 5.

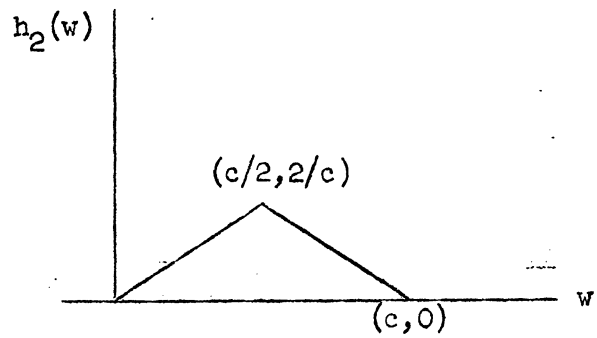

Figure 5. Distribution of $w=(x+y)/2$ = midpoint.

Clearly this is a symmetric function with mean,

$$\mu = c/2$$

Moments about the mean may be obtained from the following moment generating function.

$$m(t) = E(e^{t(w-c/2)})$$

$$= \int_0^{c/2} e^{t(w-c/2)} \frac{4w}{c^2}\, dw + \int_{c/2}^{c} e^{t(w-c/2)} \frac{4(c-w)}{c^2}\, dw$$

$$= \frac{4}{c^2}\left[\left\{ \frac{e^{t(w-c/2)}}{t^2}\,[t(w-c/2)-1] + \frac{ce^{t(w-c/2)}}{2t} \right\} \Bigg|_0^{c/2} \right.$$

$$\left. + \left\{ \frac{ce^{t(w-c/2)}}{2t} - \frac{e^{t(w-c/2)}}{t^2}\,[\,t(w-c/2)-1\,] \right\} \Bigg|_{c/2}^{c} \right]$$

$$= \frac{8}{c^2t^2}\left[-1 + \cosh\frac{ct}{2}\right]$$

$$= \frac{8}{c^2t^2}\left[-1+\left(1-\frac{i^2c^2t^2}{2^2(2!)}+\frac{i^4c^4t^4}{2^4(4!)}-\frac{i^6c^6t^6}{2^6(6!)}+\cdots\right)\right]$$

$$= 1+\frac{4c^2}{2^3(4)(3)}\frac{t^2}{2!}+\frac{4c^4}{2^5(6)(5)}\left(\frac{t^4}{4!}\right)+\cdots$$

$$+\frac{4c^{2n}}{2^{2n+1}(2n+2)(2n+1)}\frac{t^{2n}}{(2n)!}+\cdots$$

This clearly shows that all odd moments about the mean are zero, as they must be for a symmetric distribution. The variance is the coefficient of $t^2/2!$.

$$\sigma^2 = \frac{4c^2}{2^3(4)(3)}$$

$$= \frac{c^2}{24}$$

Finally, for the distribution of the midpoints of inversions of fixed length, we require the conditional distribution of w given v. This is denoted by $h(w|v)$.

$$h(w|v) = \frac{h(w,v)}{h_1(v)}$$

$$= \left(\frac{2}{c^2}\right)\left(\frac{c^2}{2(c-v)}\right)$$

$$= \frac{1}{c-v} \quad , \quad \frac{v}{2} < w < c-\frac{v}{2}$$

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

This distribution is shown graphically in Figure 6.

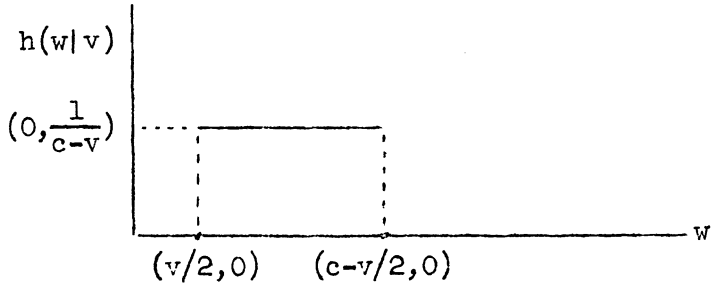

Figure 6. Distribution of $w=(x+y)/2$ for fixed $v=|x-y|$.

The mean of the variable in this conditional distribution is

$$E(w|v) = \int h(w|v)dw$$

$$= \int_{v/2}^{c-v/2} \frac{1}{c-v}\, dw$$

$$= c/2$$

For moments about the mean, we compute the appropriate moment generating function. This is

$$E(e^{t(w-c/2)}\big|\, v) = \int e^{t(w-c/2)} h(w|v)\, dw$$

$$= \frac{1}{c-v} \int_{v/2}^{c-v/2} e^{t(w-c/2)}\, dw$$

$$= \frac{1}{c-v}\left[\frac{2 \sinh \frac{t(c-v)}{2}}{t}\right]$$

$$= 1 + \frac{(c-v)^2(t^2)}{2^2(3)(2!)} + \frac{(c-v)^4(t^4)}{2^4(5)(4!)} + \cdots$$

$$+ \frac{(c-v)^{2n}\ t^{2n}}{2^{2n}(2n+1)\ (2n)!} + \cdots$$

Again, all odd moments about the mean are zero as they should be for a symmetric, in this case the uniform, distribution. It is also clear that

$$\sigma^2 = \frac{(c-v)^2}{12} \quad .$$

This shows that the variance is zero for v=c and increases as the length of the fixed interval decreases.

The distribution of points of proximal and distal breaks is also of importance. These may be found by transforming the variables in $h(w,v)$. We have

$$h(w,v) = \frac{2}{c^2} \quad , \quad \begin{cases} 0<v<2w<c \\ 0<v<2c-2w, \ c/2<w<c \end{cases}$$

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

The proximal point and, at the same time, the distance to the proximal point is given by

$$s = w - v/2$$

Also, the distal point and, at the same time, the distance to the distal point is given by

$$r = w + v/2$$

For this transformation

$$J = \begin{vmatrix} \frac{\delta s}{\delta w} & \frac{\delta r}{\delta w} \\ \frac{\delta s}{\delta v} & \frac{\delta r}{\delta v} \end{vmatrix}$$

$$= \begin{vmatrix} 1 & 1 \\ -1/2 & 1/2 \end{vmatrix} = 1$$

and

$$k(s,r) = \frac{2}{c^2} \quad , \quad 0<s, \ r<c$$

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

In turn,

$$k_1(s) = \frac{2}{c^2} \int_s^c dr$$

$$= \frac{2}{c^2}(c-s) \quad , \quad 0<s<c$$

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

This distribution is the same as that for lengths, viz. $h_1(v)$.

Again,

$$k_2(r) = \frac{2}{c^2} \int_0^r ds$$

$$= \frac{2r}{c^2} , \quad 0<r<c$$

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

This distribution is essentially the same as $k_1(r)$. It differs only in being a reflection. Its graph is given in Figure 7.

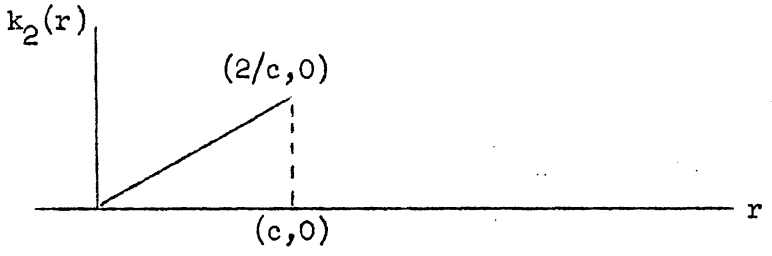

Figure 7. Distribution of r = distal point of an inversion

For $k_1(s)$, the mean is $c/3$ and moments about the mean are given by the generating function obtained for $h_1(v)$.

For $k_2(r)$, the mean is $2c/3$ and moments about the mean may be found from the generating function obtained for $h_1(v)$ by multiplication by $(-1)^n$ where the n-th moment is desired. Thus, the even moments remain the same while the odd moments change sign. The unexpanded form of the generating function may be written as

$$m(t) = \frac{2e^{\frac{2}{3}ct}}{c^2t^2} + \frac{2e^{\frac{ct}{3}}}{ct} - \frac{2e^{\frac{ct}{3}}}{c^2t^2}$$

An example

Bauer et al.* obtained 49 inversions on chromosomes. The chromosome was divided into 20 equal units and the lengths of the inversions were recorded in terms of these units. These workers obtained the frequency of inversions less than one unit, those between one and two units in length, etc. The results obtained are given below in the first two lines:

Length of inversions (in units)	0-1	1-2	2-3	3-4	4-5	5-6	6-7	7-8	8-9	9-10
Frequency of inversions	3	5	6	6	4	7	4	1	4	4
Expected frequency	4.8	4.5	4.3	4.0	3.8	3.6	3.3	3.1	2.8	2.6
Expected percent	9.75	9.25	8.75	8.25	7.75	7.25	6.75	6.25	5.75	5.25

Length of inversions (in units)	10-11	11-12	12-13	13-14	14-15	15-16	16-17	17-18	18-19	19-20
Frequency of inversions	0	0	1	0	2	0	1	0	1	0
Expected frequency	2.3	2.1	1.8	1.6	1.3	1.1	0.8	0.6	0.4	0.1
Expected percent	4.75	4.25	3.75	3.25	2.75	2.25	1.75	1.25	.75	.25

The third line above is obtained as N=49 times the proportion expected (the fourth line) in each unit. For the first unit, the expected proportion is

$$\frac{2}{c^2}\int_0^{c/20}(c-v)dv = \frac{39}{400} = .0975$$

and the expected number is .0975(49)=4.8. The expected proportion of inversions longer than one unit or shorter than two units is

$$\frac{2}{c^2}\int_0^{2c/20}(c-v)dv - .0975 = .19 - .0975 = .0925$$

* Bauer, H., Demerec, M. and Kaufman, B. P. "X-ray induced chromosomal alterations in Drosophila melanogaster." Genetics 23:610-630, 1938.

and the expected number is .0925(49)=4.5, etc. This is obviously a poor fit for these data. Consequently, an alternative model will be presented. Before proceeding, however, the following table has been prepared for situations for which data do fit this model. In order to divide the area for $h_1(v)$ into equal parts consider the following:

$$\frac{2}{c^2}\int_0^x (c-v)dv=\alpha$$

Upon integrating we obtain

$$x^2-2cx+c^2\alpha=0 \ .$$

The two roots of x are

$$x=c(1 \pm (1-\alpha)^{1/2}) \ .$$

The only root usable here is $x=c(1-(1-\alpha)^{1/2})$ since $0 \leq x \leq c$. Setting α=.10,.20, .30, etc. we obtain:

α(%)	10	20	30	40	50	60	70	80	90	100
x/c	0.051	0.106	0.163	0.225	0.293	0.368	0.452	0.553	0.684	1.000

An alternative model

Two possible reasons for an overabundance of shorter inversions relative to longer inversions are:

i) The non-heterochromatic material shrinks under fixation more rapidly than does heterochromatic material.

ii) Following two breaks, the ends of the broken chromosome heal with the remaining parts of the chromosome. Because of the radius of turning it is possible that for longer distances between breaks, the ends tend to heal more frequently with the part from which they were broken than do segments with a shorter distance between breaks.

Therefore, some modification for $h_1(v)$ in figure 4 is required which will take account of either or both of the above situations. (It should be noted here that the deficiency of very small inversions (less than one unit in length) may be due to the inability of the experimenter to observe all such inversions.)

Consider the following alternative to $h_1(v)$ in figure 4:

$$h_1'(v) = \frac{1}{c} + \beta(v - \frac{c}{2}) + \gamma(v^2 - \frac{c^2}{3}) , \qquad 0 < v < c$$

$$= \text{zero} \qquad \text{otherwise}$$

when $\beta = -\frac{2}{c^2}$ and $\gamma = 0$, $h_1'(v)$ becomes $h_1(v)$. This model may be extended as follows:

$$h_1''(v) = \frac{1}{c} + \beta(v - \frac{c}{2}) + \gamma(v^2 - \frac{c^2}{3}) + \delta(v^3 - \frac{c^3}{4}) + \cdots + \rho(v^k - \frac{c^k}{k+1}), \qquad 0 < v < c$$

$$= \text{zero}, \qquad \text{otherwise}$$

The mean and variance for $h_1'(v)$ are:

$$E(v) = \int_0^c v h_1'(v) dv = \frac{c}{12} [6 + \beta c^2 + \gamma c^3] .$$

and

$$E(v^2) - [E(v)]^2 = \int_0^c v^2 h_1'(v) dv - [E(v)]^2$$

$$= c^2 \left\{ \frac{1}{12} + \frac{\gamma c^3}{180} - \frac{(\beta c^2 + \gamma c^3)^2}{144} \right\} .$$

It is interesting to note that the mean equals $c/3$ and the variance equals $c^2/18$ for $\gamma = $ zero and $\beta = -2/c^2$; this is the maan and variance for $h_1(v)$. The function $h_1'(v)$ is shown graphically in figure 7.

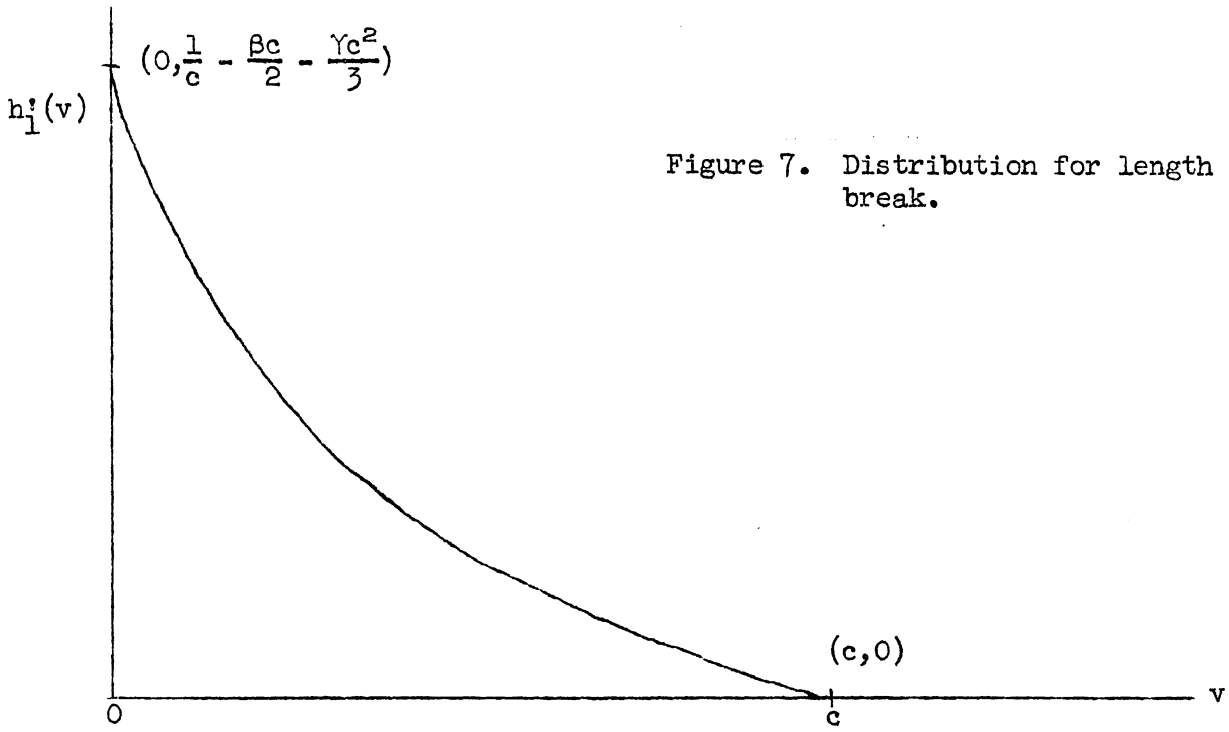

Figure 7. Distribution for length of break.

Dividing the length c into 20 equal units the area under the curve for the ith unit (i=1,2,$\cdots$,20) is computed as:

$$p_i = \int_{\frac{(i-1)c}{20}}^{\frac{ic}{20}} h_1'(v)dv = \frac{1}{32000}\left\{2398+6i-6i^2+\beta c^2(-441+83i-3i^2)\right\}$$

for

$$\gamma = -\frac{3}{2c^3} - \frac{3\beta}{4c} \quad ,$$

which results from the fact that

$$h_1'(v=c) = \frac{1}{c} + \frac{\beta c}{2} + \frac{2\gamma c^2}{3} = 0 \quad .$$

When β is known the proportion of the area between two points for the curve in figure 7, $h_1'(v)$, may be computed directly. For $\beta=-2/c^2$, and therefore $\gamma=0$, the various expected frequencies on page 13 may be computed directly. For example, let i=2 and $\beta=-2/c^2$, then

$$\frac{1}{32000}\left\{2398+12-24-2(-441+166-12)\right\} = \frac{2960}{32000} = .0925 \quad .$$

To obtain the α percent point we note that

$$\int_0^x \left[\frac{1}{c} + \beta(v-\frac{c}{2}) + \gamma(v^2-\frac{c^2}{3})\right] dv = \alpha$$

Solving, the following equation is obtained:

$$\frac{x^3\gamma}{3} + \frac{x^2\beta}{2} + x(\frac{1}{c} - \frac{\beta c}{2} - \frac{\gamma c^2}{3}) - \alpha = 0 \quad .$$

We note that β must be negative and γ must be positive and that

$$\frac{1}{c} + \frac{\beta c}{2} + \frac{2\gamma c^2}{3} = 0 \quad .$$

Therefore, the above equation becomes

$$x^3 - \frac{2x^2c^3\beta}{\beta c^2+2} + xc^2\,\frac{\beta c^2-6}{\beta c^2+2} + \frac{4c^3\alpha}{\beta c^2+2} = 0$$

Letting $y + \frac{2c^3\beta}{3(\beta c^2+2)} = x$ we obtain $y^3 - \frac{yc^2}{3}\left(\frac{\beta c^2+6}{\beta c^2+2}\right) + \frac{c^3}{27(\beta c^2+2)^3}\left\{2\beta c^2(\beta^2c^4 - 36\beta c^2 - 108) + 108\alpha(\beta c^2+2)^2\right\} = 0$. For $p = -\frac{2c^3\beta}{\beta c^2+2}$, $q = \frac{c^2(\beta c^2-6)}{\beta c^2+2}$, $r = \frac{4c^3\alpha}{\beta c^2+2}$, $a = q - p^2/3 = -\frac{c^2}{3}\left(\frac{\beta c^2+6}{\beta c^2+2}\right)^2$, $b = \frac{1}{27}(2p^3 - 9pq + 27r) = \left(\frac{c}{3(\beta c^2+2)}\right)^3\left\{2\beta c^2(\beta^2c^4 - 36\beta c^2 - 108) + 108\alpha(\beta c^2+2)^2\right\}$, $A = \sqrt[3]{-\frac{b}{2} + \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$, and $B = \sqrt[3]{-\frac{b}{2} - \sqrt{\frac{b^2}{4} + \frac{a^3}{27}}}$ the three roots for $x = y - p/3$ are $-p/3 + A + B$, $-\frac{A+B}{2} + \frac{A-B}{2}\sqrt{-3} - p/3$, and $-\frac{A+B}{2} - \frac{A-B}{2}\sqrt{-3} - p/3$. The usable root is the one between zero and c.

Estimators for β and γ from $h_1'(v)$

It should be noted that $h_1'(v)$ in figure 7 goes through the coordinates $(c,0)$ for $v=c$. Therefore, the equation $\frac{1}{c} + \frac{\beta c}{2} + \frac{2\gamma c^2}{3} = 0$ must be satisfied, and $\gamma = -\frac{3}{4c^3}(\beta c^2+2)$. Hence, we may write $h_1'(v)$ in terms of c and β:

$$\int_{\frac{(i-1)c}{20}}^{\frac{ic}{20}} h_1'(v)dx = \int_{\frac{(i-1)c}{20}}^{\frac{ic}{20}} \left(\frac{1}{c} + \beta\left(v - \frac{c}{2}\right) - \frac{3}{4c^3}(\beta c^2+2)\left(v_i^2 - \frac{c^2}{3}\right)\right)dv = \frac{1}{32000}\left\{2398 + 6i - 6i^2 + \beta c^2(-441 + 83i - 3i^2)\right\}$$

Since Y_i/N = proportion of inversions falling in ith interval, minimization of the following sums of squares,

$$\sum_{i=1}^{c}\left[Y_i - \frac{N}{32000}\left\{2398 + 6i - 6i^2 + \beta c^2(-441 + 83i - 3i^2)\right\}\right]^2,$$

results in the following:

$$\hat{\beta} = \frac{32000}{Nc^2}\,\frac{\sum_{i=1}^{20}\left\{Y_i - \frac{N}{32000}(2398 + 6i - 6i^2)\right\}\left\{-441 + 83i - 3i^2\right\}}{\sum_{i=1}^{20}(-441 + 83i - 3i^2)^2}$$

(It should be noted that $\hat{\beta}$ can be obtained if, for example, no observations are possible (or are unreliable) in the first (or other) interval. In this case, $i = 2,3,\cdots,20$ and $N = 46$.)

The variances and covariance for $\hat{\beta}$ and $\hat{\gamma} = -\frac{3}{2c^3} - \frac{3\hat{\beta}}{4c}$ are:

$$V(\hat{\beta}) = 32000^2\, \sigma^2/N^2c^4 \sum_{i=1}^{20}(-441+83i-3i^2)^2$$

$$V(\hat{\gamma}) = \frac{9}{16c^2}\, V(\hat{\beta})$$

$$\operatorname{cov}(\hat{\beta},\hat{\gamma}) = -\frac{3}{4c}\, V(\hat{\beta})$$

An estimator for σ^2 is:

$$\hat{\sigma}^2 = \sum_{i=1}^{20}\left\{Y_i - \frac{N}{32000}\,(2398+6i-6i^2+\hat{\beta}c^2(-441+83i-3i^2)\right\}^2/(20-1)$$

There is a loss of only one degree of freedom since only β is estimated. The restriction on β and γ is provided by the requirement that the distribution goes through the coordinates $(c,0)$, i.e., $\frac{1}{c} + \frac{\beta c}{2} + \frac{2\gamma c^2}{3} = 0 = h_1'(c)$.

Since the Y_i are correlated and have unequal variances, least squares estimators for β and γ are probably inefficient. Therefore, the maximum likelihood estimators are presented below. They are somewhat more difficult computationally.

Now, the Y_i have a multinomial distribution as follows:

$$L = N! \prod_{i=1}^{20} p_i^{Y_i}/Y_i! \quad ,$$

where

$$p_i = \int_{\frac{(i-1)c}{20}}^{\frac{ic}{20}} h_1'(v)dv = \frac{1}{32000}\left\{2398+6i-6i^2+\beta c^2(-441+83i-3i^2)\right\}$$

The log L is equal to

$$\log N! - \sum_{i=1}^{20} \log Y_i! + \sum_{i=1}^{20} Y_i \log p_i$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{20} \frac{Y_i}{p_i}\left(\frac{c^2}{32000}\right)(-441+83i-3i^2)$$

Therefore, the maximum likelihood estimate of β is obtained from the equation:

$$\sum_{i=1}^{20} \frac{Y_i(-441+83i-3i^2)}{2398+6i-6i^2+\beta^* c^2(-441+83i-3i^2)} = 0$$

The $V(\beta^*)$ is obtained as

$$V(\beta^*) = 32000/Nc^4 \sum_{i=1}^{20} \frac{(-441+83i-3i^2)^2}{2398+6i-6i^2+\beta c^2(-441+83i-3i^2)}$$

For the above example we compute (see Table 1 for c=1):

$$\hat{\beta} = -3.73 \qquad \hat{\gamma} = 1.29$$

$$V(\hat{\beta}) = 2.2584 \qquad V(\hat{\gamma}) = 1.2704$$

From Table 2 and the above formulae we compute (for c=1):

$$\beta^* = -4.17 \qquad \gamma^* = 1.63$$

$$V(\beta^*) = 2.1159 \qquad V(\gamma^*) = 1.1902$$

Although the difference between $\hat{\beta}$ and β^* is sizeable, there is little difference between the variances for the two estimates. Also, the variances are quite large; this is due, in part, to the low frequencies of inversions observed in the first class. As stated previously, the experimenter may not detect very small inversions and this would account for the small number observed. Because of this, $\hat{\beta}$ and β^* probably should have beem computed from the observations on the remaining 19 classes.

Table 1. Computations for $\hat{\beta}$ and $\hat{\sigma}^2$.

i	i^2	Y_i	I	$Y_i - \frac{N}{32000}(I)$	II	$Y_i - \frac{N}{32000}\{I-3.7258(II)\}$
1	1	3	2398	-0.6719	-361	-2.73
2	4	5	2386	1.3464	-287	-0.29
3	9	6	2362	2.3832	-219	1.13
4	16	6	2326	2.4383	-157	1.54
5	25	4	2278	0.5118	-101	-0.06
6	36	7	2218	3.6037	- 51	3.31
7	49	4	2146	0.7139	- 7	0.67
8	64	1	2062	-2.1574	31	-1.98
9	81	4	1966	0.9896	63	1.35
10	100	4	1858	1.1549	89	1.66
11	121	0	1738	-2.6613	109	-2.04
12	144	0	1606	-2.4592	123	-1.76
13	169	1	1462	-1.2387	131	-0.49
14	196	0	1306	-1.9998	133	-1.24
15	225	2	1138	0.2574	129	0.99
16	256	0	958	-1.4669	119	-0.79
17	289	1	766	-0.1729	103	0.41
18	324	0	562	-0.8606	81	-0.40
19	361	1	346	0.4702	53	0.77
20	400	0	118	-0.1807	19	-0.07

$N=49$ $\qquad\qquad$ $\hat{\sigma}^2 = \frac{42.6596}{19} = 2.2452$

$I = 2398 + 6i - 6i^2$; $\qquad\qquad$ $II = -441 + 83i - 3i^2$

$\hat{\beta}c^2 = \frac{32000}{49}\left(-\frac{2419.0140}{424004}\right) = -3.725824$

$\hat{\gamma}c^3 = 1.294368$

Table 2. Computations for β^* and $\widehat{V(\beta^*)}$.

i	Y_i	III= I+$\hat{\beta}c^2$II	Y_iII/III	III'= I+$\beta' c^2$II	Y_iII/III'	III"= I+$\beta" c^2$II	Y_iII/III"	III'"= I+$\beta"' c^2$II	Y_iII/III'"	II^2/III'"
1	3	3743.0	-0.2893	3842	-0.2819	3914.2	-0.2767	3903.4	-0.2775	33.39
2	5	3455.3	-0.4153	3534	-0.4061	3591.4	-0.3996	3582.8	-0.4005	22.99
3	6	3178.0	-0.4135	3238	-0.4058	3281.8	-0.4004	3275.2	-0.4012	14.64
4	6	2911.0	-0.3236	2954	-0.3189	2985.4	-0.3155	2980.7	-0.3160	8.27
5	4	2654.3	-0.1522	2682	-0.1506	2702.3	-0.1495	2699.2	-0.1497	3.78
6	7	2408.0	-0.1483	2422	-0.1474	2432.2	-0.1468	2430.7	-0.1469	1.07
7	4	2172.1	-0.0129	2174	-0.0129	2175.4	-0.0129	2175.2	-0.0129	0.02
8	1	1946.5	0.0159	1938	0.0160	1931.8	0.0160	1932.7	0.0160	0.50
9	4	1731.3	0.1456	1714	0.1470	1701.4	0.1481	1703.3	0.1479	2.33
10	4	1526.4	0.2332	1502	0.2370	1484.2	0.2399	1486.9	0.2394	5.33
11	0	1331.9	0	-	0	-	0	1283.5	0	9.26
12	0	1147.7	0	-	0	-	0	1093.1	0	13.84
13	1	973.92	0.1345	938	0.1397	911.8	0.1437	915.7	0.1431	18.74
14	0	810.47	0	-	0	-	0	751.4	0	23.54
15	2	657.37	0.3925	622	0.4148	596.2	0.4327	600.1	0.4299	27.73
16	0	514.63	0	-	0	-	0	461.8	0	30.66
17	1	382.24	0.2695	354	0.2910	333.4	0.3089	336.5	0.3061	31.53
18	0	260.21	0	-	0	-	0	224.2	0	29.26
19	1	148.53	0.3568	134	0.3955	123.4	0.4295	125.0	0.4240	22.47
20	0	47.210	0	-	0	-	0	38.8	0	9.30
Sum	49	-	-0.2071	-	- .0826	-	+ .0174	-	+ .0017	308.65

I, II and $\hat{\beta}c^2 = -3.72582$ from table 1.

$\beta' c^2 = -4.0$ $\qquad$ $\beta^* = B'''c^2 = -4.17$ $\qquad$ $V(\beta^* c^2) = \frac{32000}{49(308.65)} = 2.1159$

$\beta" c^2 = -4.2$