Cost functions for sample surveys

Garnet Ernest McCreary

Iowa State College

Follow this and additional works at: https://lib.dr.iastate.edu/rtd

Part of the Statistics and Probability Commons

Recommended Citation
McCreary, Garnet Ernest, "Cost functions for sample surveys " (1950). Retrospective Theses and Dissertations. 13525.
https://lib.dr.iastate.edu/rtd/13525
COST FUNCTIONS FOR SAMPLE SURVEYS

by

Garnet Ernest McCrea

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Statistics

Approved:

In Charge of Major Work

Head of Major Department

Dean of Graduate College

Iowa State College
1950
THE QUALITY OF THIS REPRODUCTION IS DEPENDENT UPON THE QUALITY OF THE COPY SUBMITTED. BROKEN OR INDISTINCT PRINT, COLORED OR POOR QUALITY ILLUSTRATIONS AND PHOTOGRAPHS, PRINT BLEED-THROUGH, SUBSTANDARD MARGINS, AND IMPROPER ALIGNMENT CAN ADVERSELY AFFECT REPRODUCTION.

IN THE UNLIKELY EVENT THAT THE AUTHOR DID NOT SEND A COMPLETE MANUSCRIPT AND THERE ARE MISSING PAGES, THESE WILL BE NOTED. ALSO, IF UNAUTHORIZED COPYRIGHT MATERIAL HAD TO BE REMOVED, A NOTE WILL INDICATE THE DELETION.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>PART</th>
<th>CONTENTS</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I.</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>II.</td>
<td>REVIEW OF LITERATURE</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>A. Travel Functions and Field Costs</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>B. Decision Functions</td>
<td>5</td>
</tr>
<tr>
<td>III.</td>
<td>EXPECTED MINIMUM DISTANCE</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>A. Points and Sampling Units</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>B. The Theoretical Problem of Minimum Distance</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>C. Grid and Airline Distance When n = 2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>D. Mean Square Distance and Expected Grid Distance for n = 2 When the Points Have any Space Distribution over any Plane Area</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>E. Minimum Grid Distance Among n = 3 Points in a Square</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>F. Empirical Investigations of Minimum Average Distances</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>G. An Upper Bound to Minimum Grid Distance in a Rectangle</td>
<td>23</td>
</tr>
<tr>
<td>IV.</td>
<td>RELATIONSHIP OF GRID TO AIRLINE DISTANCE</td>
<td>26</td>
</tr>
<tr>
<td>V.</td>
<td>EXPECTED DISTANCE OVER A REGION IN A PREDESIGNATED DIRECTION</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>A. Grid Distance Among n Points in a Rectangle</td>
<td>35</td>
</tr>
<tr>
<td></td>
<td>B. Grid Distance Among n Points When a Rectangle is Stratified</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>C. Mean Square Distance Among n Points in a Circle, Semicircle and Quadrant</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>D. Grid Distance from a Fixed Point to a Random Point</td>
<td>68</td>
</tr>
<tr>
<td>VI.</td>
<td>ANALYSIS OF TIME AND MILEAGE IN THE FIELD FOR THREE IOWA FARM SURVEYS</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>A. Description of the Surveys</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>B. Between Sampling Unit Mileage</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>C. Home Mileage</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>D. Least Squares Estimation of the Remaining Mileage</td>
<td>97</td>
</tr>
</tbody>
</table>
PART I. INTRODUCTION

In planning sample surveys administrators are interested in both the accuracy and the cost of various designs. Usually there are many possible sampling designs and sizes of samples which will give the required accuracy but the corresponding costs are unknown.

Any overall cost function should include office costs as well as field costs. This study concerns itself with field costs which are by far the more difficult to estimate. The field costs considered are those for surveys in which the investigators must travel to certain random locations in a geographical area. Under these circumstances, if the total distance travelled and its various components can be estimated from a travel function, if the speed in m.p.h. for each component of travel can be prognosticated, and if the average time to make an observation or interview is known, it is possible to arrive at a cost function for the field work.

In this thesis, theory is presented that will enable one to make an estimate of the total distance travelled in a survey for a wide variety of sample designs where the population has a uniform space distribution over any stratum. Interviewers kept mileage and time records for three different Iowa farm surveys. The data derived from these records are presented and compared with theoretical results.

The overall problem of deciding what is the accuracy required is much more difficult and brings up the question of the loss that is sustained from errors in estimates, as well as the cost of the survey.
In Part VII some of the elementary notions of Wald's decision function are applied to determining sample size, where one does not know the parameters of his distribution of error.
PART II. REVIEW OF LITERATURE

A. Travel Functions and Field Costs

Interest in cost functions began in 1932 when Neyman (16) showed there was an optimum rate of sampling for each stratum depending on the variation within and the size of each stratum. Using Lagrangian multipliers, it is possible to determine the optimum rates within a particular design if one minimizes cost for a fixed accuracy or alternately maximizes the accuracy for a fixed cost. Yates and Zaccopany (31) in 1935 described a method of determining the optimal percentage of sampling in field experiments taking into account the cost of harvesting various size samples. They also investigated the optimum size of the sampling unit in a particular instance.

P. C. Mahalanobis (14) in 1940 considered a cost function which depended partially on the distance travelled. He stated that it was easy to see that the expected length of a route connecting $n$ points scattered at random within any given area was

$$\frac{n - 1}{\sqrt{n}}$$

In 1942 Jessen (11) used this as a basis for setting up a field cost function. For a survey in Iowa, Mahalanobis and Jessen in their respective papers set up a variance function showing how the sampling

...Mahalanobis should have stated that it equals $\sqrt{n} \frac{n - 1}{\sqrt{n}}$ for a systematic sampling in a plane (see Part V.A),
error varied with the size of the sampling unit (or with mean square
distance within an s.u.). It was possible then to arrive at two equations
using a Lagrange multiplier and thus to obtain formally, for a given
total cost, the optimum size and number of sampling units for maximum
accuracy.

Several research workers (including Mahalanobis) realized that
expression (1) is not the exact expected minimum distance among n random
points. B. Ghosh (7) in 1943 obtained the distribution of the distance
between two points where each point had a uniform distribution in the
rectangle ab. Previous to this, Williamson's "Integral Calculus" (30)
had given the mean (the expected distance) of this distribution and
Jessen (11) had shown the uncorrected second moment. Independent of
Ghosh and independent of each other, A. M. Mood\textsuperscript{1} and H. C. S. Thom\textsuperscript{1} in
1945 also found the distribution of the distance between two random
points in a rectangle.

Marks (16) in 1948 obtained a lower bound\textsuperscript{2} for the expected travel
in any bounded two dimensional Borel set of area A to be

\[
\frac{1}{n} \sqrt[n]{A} \cdot \frac{n - 1}{\sqrt{n}}. \quad (2)
\]

An upper bound of the expected minimum distance was found by M. N. Ghosh

\textsuperscript{1}Both solutions are unpublished but Mood's derivation may be found
in the July 1949 Progress Report to B.A.E. Ames, Iowa, Statistical

\textsuperscript{2}This was misprinted at the beginning of Marks' note, but was given
correctly at the end of the note. M. N. Ghosh (7) used the incorrect
expression.
(9) to be

\[ 1.27 \sqrt{\frac{1}{n}} \]

for large \( n \).

**B. Decision Functions**

In 1944 Von Neumann and Morgenstern (24) showed, among other things, that a zero sum two person game is generally strictly determined if a finite number of mixed strategies are admitted. Wald, a year later, extended the solution to the case of a denumerably infinite number of strategies. Previous to that, in 1939 Wald (25) had given an argument for using risk functions for the problems of statistical inference. So when Von Neumann's theory of games was published, Wald (26) regarded statistical inference as a zero sum two person game between the statistician and Nature as the two players. Wald extended his ideas on decision functions to sequential decision functions (27) and in a later paper (29) further generalized his conditions on the spaces of strategies, the weight function, the space of all admissible distribution functions and the cost function of experimentation.

Cochran (6) has pointed out the Neyman criterion assumes that either cost or accuracy is fixed in advance. Actually one should determine optimum cost and optimum accuracy simultaneously. Nordin (19), Blythe (4) and Yates (32) have given what Wald would call a Bayes solution to the problem (for example see Arrow, Blackwell and Girshick (2)). Their solutions differ because they assume different cost function and/or different distributions of the error. The assumption that Nordin, Blythe
and Yates make is that one knows at least one parameter of the distribution, usually the variance or standard error, and that it enters into the expected loss function. Berkson (3) in 1947 used a risk function in which he knew the parameters of the distributions.
1. Points, if the intensity distribution function is known, can be
found from a map which is the shortest route (in time) among
the points. There are no possible paths from these an investigator
scheme only once (unless it is interesting to test one of the random
points, randomly drawn), then choose the shortest path which touches
a given a uniform space distribution and any condition on the
given a uniform space distribution of a given

2. The theoretical problem of minimum distance

is slightly less than the distance between the center of such units.

If the intensity of the center of the sample unit is not a random point, then the distance between the center of the sample unit is not a great deal of harm to done by regrading the center of the sample unit as an area of the region. If the sample units are fairly large
areas part of the region, then of the sample units are equal
or the region and if the density of e., is constant over all
sample units are guide mainly in area which respect to the total area
region. In sample that is achieved exactly if the
be chosen, i.e., there exists a uniform space distribution of the

are strict, assume that every point in a region is equally likely to
Pare III. EXPECTED MINIMUM DISTANCE
In some studies, it is not feasible to travel in a straight line from one point to the next, i.e. airline distance. In the midwest where there is one set of roads a mile apart running east and west and another set of roads a mile apart running north and south, rectangular or grid distance between the points seems to be more appropriate. In other parts of the country, the roads form no systematic pattern and the road distance among n random points will be at least as great as straight line or airline distance and may be even greater than grid distance.

C. Grid and Airline Distance When n = 2

It is not too difficult to obtain the expected distance between two random points for a uniform space distribution over a regular area such as a rectangle or ellipse.

The expected grid distance between two random points in a rectangle \( ab, b \leq a \), is
\[ E(r_e) = \frac{1}{2} D_g = \frac{a + b}{3}. \] (1)

The corresponding expected airline distance has been found by several researchers (see Part II) to be

\[ E(r_a) = \frac{1}{2} D_a = \frac{1}{6} \left[ \frac{b^2}{a^2} \log_e \left( \frac{a}{b} + \sqrt{1 + \left( \frac{b}{a} \right)^2} \right) + \frac{a^2}{b^2} \log_e \left( \frac{b}{a} + \sqrt{1 + \left( \frac{a}{b} \right)^2} \right) \right] + \frac{a^2 + b^2}{15^2} \left( 3 - \frac{a^2}{b^2} - \frac{b^2}{a^2} \right) + \frac{1}{15} \left( \frac{a^3}{b^2} + \frac{b^3}{a^2} \right). \] (2)

For a square i.e. \( b = a \) the variance of the airline distance is

\[ 2 \, \mathbb{V} \, (r_a) = a^2 \left[ \frac{1}{3} - \frac{1}{9} \left( \log_e (1 + \sqrt{2}) + \frac{\sqrt{2} + 2}{3} \right) \right]^2 \] (3)

\[ \approx 0.0614 \, a^2. \]

The variance of grid distance in a rectangle is

\[ 2 \, \mathbb{V} \, (r_g) = \frac{a^2 + b^2}{18}. \] (4)

It is easy to see that the coefficient of variation for airline or grid distance is fairly high (more than 15 percent in all cases). In fact for grid distance it may be shown by the Schwarz Inequality that the C. V. is always \( \geq 50 \) percent.

The expected grid distance between two random points in an ellipse with semiaxes \( a \) and \( b \) is

\[ 2 \, D_g = \frac{256}{45^2 \pi a} (a + b). \] (5)

\(^1\) The subscript before \( D \) is the number of random points and the subscript \( g \) after \( D \) indicates grid distance. The subscript \( a \) after \( D \) indicates airline distance.
Williamson's "Integral Calculus" (30) on page 319 gives an ingenious method for finding the airline distance in a circle to be

$$2 D_a = \frac{125}{45^n} a$$

where the \(a\) on the right is the radius of the circle.

A rectangle with sides \(a\) and \(b\) and an ellipse with axes \(a/\sqrt{m}\) and \(b/\sqrt{m}\) have the same area. In this case the grid distance for the rectangle is approximately \((a + b)(.3333)\) and for the ellipse is approximately \((a + b)(.3252)\). A square and circle whose area is \(a^2\) have airline distance approximately .5214\(a\) and .5108\(a\) respectively. Thus the average distance does not change much (relative to the coefficient of variations) between rectangles and ellipses of equal area. Hence, if a rectangle of the same area as any fairly regular plane figure is taken and the longer direction of the rectangle used to approximate the longer direction of the figure, then the formulas for rectangles should give a fairly good estimate for this plane figure.

Figure 3.
D. Mean Square Distance and Expected Grid Distance for \( n = 2 \) 
When the Points Have any Space Distribution over any Plane Area

Unfortunately the square root of the mean square distance does not provide the answer that is sought. In fact it is biased upward with respect to airline distance.

The mean square distance may be found between two random points distributed over any plane area \( \Omega \) over which there exists any space distribution \( f(x, y) \).

Then

\[
E(r^2) = \iint_{\Omega} \left[ (x - x_1)^2 + (y - y_1)^2 \right] f(x_1, y_1) f(x_2, y_2) \, dy_2 \, dy_1 \, dx_2 \, dx_1
\]

\[
= \int_{\Omega} \int_{\Omega} \left[ (x - m_x)^2 + (y - m_y)^2 - 2(x - m_x)(x_1 - m_x) \right.
\]

\[
+ (y_2 - m_y)^2 + (y_1 - m_y)^2 - 2(y_2 - m_y)(y_1 - m_y) \left. \right] f(x_1, y_1) f(x_2, y_2) \, dy_2 \, dy_1 \, dx_2 \, dx_1
\]

\[
= \int_{\Omega} \left[ (x - m_x)^2 + (y - m_y)^2 \right] f(x_2, y_2) \, dy_2 \, dx_2
\]

\[
+ \int_{\Omega} \left[ (x_1 - m_x)^2 + (y_1 - m_y)^2 \right] f(x_1, y_1) \, dy_1 \, dx_1
\]

\[- 2 \int_{\Omega} \int_{\Omega} \left[ (x_2 - m_x)(x_1 - m_x) + (y_2 - m_y)(y_1 - m_y) \right] f(x_1, y_1) f(x_2, y_2) \, dy_2 \, dy_1 \, dx_2 \, dx_1
\]
\[ = 2 \int \frac{1}{\sigma^2} \left[ (x - \mu_x)^2 + (y - \mu_y)^2 \right] f(x, y) \, dy \, dx \]

\[ = 2(\mu_{20} + \mu_{02}). \]

That is, the mean square distance is twice the second moment of the distribution in the \( x \) direction plus twice the second moment of the distribution in the \( y \) direction.

In theory it is also possible to find the grid distance between two points for any distribution over an irregular plane area \( \cdots \). We have

\[
E(r_g) = \int \int_{\Omega} \left[ |x_2 - x_1| + |y_2 - y_1| \right] f(x_1, y_1) f(x_2, y_2) \, dy \, dx \\
= \int \int_{\Omega} |x_2 - x_1| f(x_1, y_1) f(x_2, y_2) \, dy \, dx \\
+ \int \int_{\Omega} |y_2 - y_1| f(x_1, y_1) f(x_2, y_2) \, dy \, dx .
\]

If \( y \) can be expressed in terms of \( x \) and if the integrations are feasible then

\[
E(r_g) = \int \int |x_2 - x_1| p_2(x_2)p_1(x_1) \, dx \, dx \\
+ \int \int |y_2 - y_1| q_2(y_2)q_1(y_1) \, dy \, dy,
\]

where the integrations are over the range of the \( x \)'s in the first part and over the range of the \( y \)'s in the second part. The \( p \) and \( q \) functions
are not exactly marginal distributions but are fairly closely related to them.

E. Minimum Grid Distance Among $n = 3$ Points in a Square

Define a reasonable path to be such that we move from one point to the next in two directions at most. In practice this path may not be reasonable because road conditions may be such that for a saving of time, money and car one should move in more than two directions in going from one point to another, (i.e. one may go out of his way in order to reach highways).

For a certain configuration of points there are three possible orders in which the points may be taken, each having a different total length. For a minimum one follows the path among the three points which minimizes the total grid distance.

Let $x_1 < x_2 < x_3$ and $y_1 < y_2 < y_3$.

![Figure 1](image)

Represent the three points by

$$(x_{a_1}, y_{b_1}), (x_{a_2}, y_{b_2}), (x_{a_3}, y_{b_3})$$
and suppose the points are joined in the order \( \alpha_1, \alpha_2, \alpha_3 \). Assume first that \( \alpha_1 = 1, \alpha_2 = 2 \) and \( \alpha_3 = 3 \). There are 6 permutations of the \( \beta \)'s and each can occur with probability \( \frac{1}{6} \).

<table>
<thead>
<tr>
<th>( \beta_1 )</th>
<th>( \beta_2 )</th>
<th>( \beta_3 )</th>
<th>Min. Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_1 &lt; \beta_2 &lt; \beta_3 ) then</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>( \beta_1 &lt; \beta_3 &lt; \beta_2 )</td>
<td>1</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>( \beta_2 &lt; \beta_1 &lt; \beta_3 )</td>
<td>2</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>( \beta_3 &lt; \beta_1 &lt; \beta_2 )</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>( \beta_2 &lt; \beta_3 &lt; \beta_1 )</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>( \beta_3 &lt; \beta_2 &lt; \beta_1 )</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

For \( (x_1,y_1), (x_2,y_2), (x_3,y_3) \) and \( (x_1,y_3), (x_2,y_3), (x_3,y_1) \) the configuration of points is such that it is obvious that the minimum "reasonable" path is \( R = x_3 - x_1 + y_3 - y_1 \).

![Figure 5.](image)

For the other four configurations there are three possible paths. For \( \beta_1 = 1, \beta_2 = 3, \beta_3 = 2 \) the \( \alpha \)'s may be taken in the following orders...
(without duplication of paths):

\[
\begin{array}{ccc}
\alpha_1 & \alpha_2 & \alpha_3 \\
1 & 2 & 3 \\
1 & 3 & 2 \\
2 & 1 & 3 \\
\end{array}
\]

Path

\[x_3 - x_1 + y_3 - y_1 + y_3 - y_2 = R + y_3 - y_2\]
\[x_3 - x_1 + y_3 - y_1 + x_3 - x_2 = R + x_3 - x_2\]
\[x_3 - x_1 + y_3 - y_1 + x_2 - x_1 + y_2 - y_1 = R + x_2 - x_1 + y_2 - y_1\]

![Diagram](attachment:image.png)

*Figure 6.*

It can easily be verified that similar relations hold for \(\beta_1 = 2\), \(\beta_2 = 1\), \(\beta_3 = 3\), etc. For all permutations of the \(\beta\)'s the distance includes \(R\) and \(\frac{2}{3}\) of the permutations have an added element which can be any one of three components. Hence

\[
3 \mathbb{E} \frac{D}{\xi} = \mathbb{E} \text{(Minimum path length)} = \mathbb{E}(R) + \frac{2}{3} \mathbb{E} \left[ \min (y_3 - y_2, x_3 - x_2, x_3 - x_1 + y_3 - y_1) \right].
\]

(1)

Now assuming a uniform distribution of the \(x\)'s and \(y\)'s then

\[
\mathbb{E}(R) = \frac{a + b}{2}.
\]

(2)

Let

\[
\begin{align*}
u_1 &= x_2 - x_1 & u_1 &> 0 \\
u_2 &= x_3 - x_2 & u_2 &> 0 \\
u_1 + u_2 &< a
\end{align*}
\]
\[
\left\{ \frac{\bar{t}}{x} + \frac{x(x-\bar{t})}{x} + x(x-\bar{t}) \right\} \left[ \frac{x}{x-\bar{t}} + \frac{x}{x-\bar{t}} \right] + \frac{w}{x-\bar{t}} \frac{w}{x-\bar{t}}
\]

\[ a \leq x \leq b \]

\[ \left[ \frac{w^w}{w^w - \frac{w^w}{w^w}} \right] = \frac{(x)b - 1}{1} \]

By

(9)

\[ \left[ a \leq e_A + t_A, x < e_A, x < e_A \right] d + \left[ x > e_A, x < e_A \right] d + \left[ x > e_A \right] d = \left[ x > e_A, x > e_A \right] \frac{d}{d} \left( A_{n^A} \cdot e_n, a \right) \frac{d}{d} = \left[ x > e_A, x > e_A \right] \frac{d}{d} = \frac{(x)b}{1} \]

Then, \( (A_{n^A} \cdot e_n, a \cdot e_n) \frac{d}{d} \) for the c.d.f. \( t_A = \frac{1}{d} \) and \( A = t_A + t_A \)

The problem is to find

(10)

\[ (e_A - t_A - q) \frac{d^q}{d^q} = \frac{e_A}{d^e} \frac{d^e}{d^e} (e_A + t_A) \]

Then

\[ q > e_A + t_A \]

(11)

\[ 0 < e_A \]

\[ e_A - e_A = e_A \]

\[ 0 < t_A \]

\[ t_A - e_A = t_A \]

Similarly let

(12)

\[ (e_n - t_n - q) \frac{d^q}{d^q} = (e_n \cdot t_n) \]

Then the distribution of \( t_n \) and \( e_n \) is
\[- \frac{6x^3}{a^6} \left\{ \frac{(a - 2x)^3}{2} + \frac{3(a - 2x)^2x}{3} + \frac{3(a - 2x)x^2}{4} + \frac{x^3}{5} \right\} \]

\[+ \frac{3x^3}{a^6} \left\{ \frac{(a - 2x)^3}{3} + \frac{3(a - 2x)^2x}{4} + \frac{3(a - 2x)x^2}{5} + \frac{x^3}{6} \right\} \quad 0 \leq x < \frac{a}{2} \]

\[+ \left[ \frac{3}{40} \frac{(a - 2x)^6}{a^6} + \frac{(a - x)^3}{a^6} \left\{ \frac{(a - 2x)^3}{4} + \frac{3(a - x)(a - 2x)x^2}{10} + \frac{3(a - x)^2(a - 2x)}{26} \right\} \right. \]
\[+ \left. \frac{6(a - 2x)^4}{a^6} \left\{ \frac{(3a - 4x)^2}{5} - \frac{5(a - 7x)^2}{12} + \frac{(7a - 10x)^2}{32} - \frac{(9a - 13x)^2}{200} \right\} \right] \quad \frac{a}{2} \leq x < \frac{2}{3} a. \] (7)

Now it may be shown that

\[
\int_0^a x g(x) \, dx = \int_0^a \left\{ 1 - g(x) \right\} \, dx. \] (8)

Hence

\[
\int_0^a x g(x) \, dx = a \left( \frac{1}{20} + \frac{1}{27} + \frac{17.555}{160000} \right). \] (9)

From (1) and (2)

\[
3 \, D = a \left( 1 + \frac{1}{10} + \frac{1}{48} \right) + \frac{17.555}{(61,236)10^3} \approx a(1.12112491). \] (10)
F. Empirical Investigations of Minimum Average Distances

No exact formula for the shortest average distance between n points exists. C. J. Maloney started some empirical studies on the distance, both airline and grid, among random points. In Appendix B is an example of an Average Travel Path Sheet for n = 5, side ratio 1:1. Maloney took a 5" x 5" square and divided it into 100 smaller squares. Hence each smaller square is 1/2" to the side. These were numbered in an ordered sequence from 00 to 99. Then n numbers between 00 and 99 were drawn at random without replacement. Therefore n randomly selected small squares within the larger 5 x 5 have been selected. The lower left hand corner of each small square selected might be regarded as being a random point in the 5 x 5 square (i.e. a rectangle with side ratio 1:1). For purposes of this report the difference between these points and purely random points is small. Both airline and grid distance were measured from these lower left hand corner points.

For 2:1 side ratio two 5 x 5 squares were used with a side of one coinciding with a side of the other. Thus we have a 5" x 10" rectangular area in which there are 200 smaller squares. Using the same scheme and the same numbers on the small squares as before random squares, from which approximately random points were obtained, were drawn. The only difference was that it was necessary to determine which 5 x 5 square had been selected for every random number drawn. This was done by making a further random drawing of numbers 0---9. If the number was even the left hand 5 x 5 square was selected, and if the number was odd the right hand 5 x 5 was selected.
For rectangles with sides in the ratio $\frac{1}{4} : 1$ two $5 \times 5$ squares were laid side by side and a line which bisected each square did so in such a way that all the small squares below the line have numbers $< 50$ and all numbers above the line are $\geq 50$. In this way 2 rectangles with side ratio $\frac{1}{4} : 1$ were formed. In one of these random numbers between 00 and 49 were drawn and then it was determined which $5 \times 5$ square it was in by random odd or even numbers as above. For the other random numbers between 50 and 99 were drawn and then it was determined which $5 \times 5$ square had been selected, as before. The random numbers determined the small squares and hence the approximate random points in a rectangle.

The following tables show the values of $n$ considered, the average minimum distance among the $n$ points where the average is determined from $k$ replications, and the standard error of this distance. In Table I, III and V we have airline distance, $D_a$, and in Table II, IV and VI we have grid distance, $D_g$. Also in Tables III and IV the last two lines give the average minimum distance and its standard error when the area of the rectangle is 25 as in the other tables.
### TABLE I

**Squares 5 x 5, Side Ratio 1 : 1**

<table>
<thead>
<tr>
<th>Airline Distance</th>
</tr>
</thead>
</table>
| No. of replications (k) | 100 | 100 | 100 | 50 | 25 | 25 | 25 | 10  
| No. of random points (n) | 2 | 3 | 4 | 5 | 9 | 12 | 16 | 25  
| Average distance ($D_a$) | 2.697 | 4.268 | 5.431 | 6.734 | 10.312 | 12.568 | 15.124 | 19.92  
| Standard error ($\frac{s}{\sqrt{k}}$) | .1211 | .1244 | .1638 | .2335 | .2498 | .3049 | .3298 | .3292  
| Pop. st. error ($\frac{c}{\sqrt{k}}$) | .1240 |  

### TABLE II

**Squares 5 x 5, Side Ratio 1 : 1**

<table>
<thead>
<tr>
<th>Grid Distance</th>
</tr>
</thead>
</table>
| No. of replications (k) | 100 | 100 | 100 | 50 | 25 | 25 | 25 | 10  
| No. of random points (n) | 2 | 3 | 4 | 5 | 9 | 12 | 16 | 25  
| Average distance ($D_g$) | 3.445 | 5.438 | 6.850 | 8.360 | 12.880 | 15.040 | 18.140 | 24.200  
| Standard error ($\frac{s}{\sqrt{k}}$) | .1676 | .1589 | .1993 | .2567 | .2891 | .4396 | .4159 | .4955  
| Pop. st. error ($\frac{c}{\sqrt{k}}$) | .1667 |  

### TABLE III
**Rectangle 5 x 10 = 50, Side Ratio 2 : 1**

<table>
<thead>
<tr>
<th>No. of replications (k)</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>25</th>
<th>25</th>
<th>25</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of random points (n)</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>Average distance (D_a)</td>
<td>4.355</td>
<td>6.366</td>
<td>8.525</td>
<td>10.071</td>
<td>14.900</td>
<td>16.594</td>
<td>20.964</td>
<td>27.320</td>
</tr>
<tr>
<td>Standard error (\frac{s}{\sqrt{k}})</td>
<td>.2250</td>
<td>.2305</td>
<td>.2072</td>
<td>.1770</td>
<td>.3902</td>
<td>.4614</td>
<td>.4417</td>
<td>.9062</td>
</tr>
<tr>
<td>St. error on area 25</td>
<td>.1591</td>
<td>.1631</td>
<td>.1465</td>
<td>.1252</td>
<td>.2759</td>
<td>.3263</td>
<td>.3123</td>
<td>.6408</td>
</tr>
<tr>
<td>Pop. st. error (\sigma/\sqrt{k})</td>
<td>.2161</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE IV
**Rectangles 5 x 10, Side Ratio 2 : 1**

<table>
<thead>
<tr>
<th>No. of replications (k)</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>25</th>
<th>25</th>
<th>25</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of random points (n)</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>12</td>
<td>16</td>
<td>25</td>
</tr>
<tr>
<td>Average distance (D_g)</td>
<td>5.47</td>
<td>7.890</td>
<td>10.625</td>
<td>12.715</td>
<td>18.18</td>
<td>20.2</td>
<td>25.46</td>
<td>33.5</td>
</tr>
<tr>
<td>Standard error (\frac{s}{\sqrt{k}})</td>
<td>.2685</td>
<td>.2894</td>
<td>.2760</td>
<td>.2517</td>
<td>.4948</td>
<td>.5664</td>
<td>.4974</td>
<td>1.2042</td>
</tr>
<tr>
<td>Average distance (D_g) on area 25</td>
<td>3.868</td>
<td>5.579</td>
<td>7.513</td>
<td>8.991</td>
<td>12.855</td>
<td>14.284</td>
<td>18.003</td>
<td>23.688</td>
</tr>
<tr>
<td>St. error on area 25</td>
<td>.2040</td>
<td>.2047</td>
<td>.1952</td>
<td>.1760</td>
<td>.3499</td>
<td>.4005</td>
<td>.3517</td>
<td>.6515</td>
</tr>
<tr>
<td>Pop. st. error (\sigma/\sqrt{k})</td>
<td>.1863</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### TABLE V

<table>
<thead>
<tr>
<th>No. of replications (k)</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>50</th>
<th>25</th>
<th>25</th>
<th>25</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error (s/√k)</td>
<td>.2261</td>
<td>.2091</td>
<td>.1975</td>
<td>.2133</td>
<td>.2904</td>
<td>.2458</td>
<td>.2140</td>
<td>.4728</td>
</tr>
<tr>
<td>Pop. st. error (s/√k)</td>
<td>.2234</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### TABLE VI

<table>
<thead>
<tr>
<th>No. of replications (k)</th>
<th>100</th>
<th>100</th>
<th>100</th>
<th>50</th>
<th>25</th>
<th>25</th>
<th>25</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. distance (Dg)</td>
<td>4.37</td>
<td>6.90</td>
<td>8.19</td>
<td>9.95</td>
<td>14.0</td>
<td>16.16</td>
<td>18.56</td>
<td>23.1</td>
</tr>
<tr>
<td>Standard error (s/√k)</td>
<td>.2409</td>
<td>.2376</td>
<td>.2273</td>
<td>.2708</td>
<td>.3663</td>
<td>.3509</td>
<td>.2522</td>
<td>.7589</td>
</tr>
<tr>
<td>Pop. st. error (s/√k)</td>
<td>.2430</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
In the above tables as the number of points increase the distance increases along a curve which looks somewhat like a parabola. It is to be noted that \( \frac{a}{\sqrt{n}} \) while it is subject to sampling fluctuations should however give a fair idea of the variation of the distances. For side ratio 1 : 1 it is to be noted that in the main the sampling errors of the mean increase as \( D \) increases. However, in the case of side ratios 2 : 1 and 4 : 1 this trend does not seem to hold and suggests that shape is a limiting factor in the variation. There are then two factors affecting the variation: (1) the increase due to increase in \( D \) as \( n \) increases; (2) the decrease due to elongation of a rectangle of fixed area.

In the tables the population sampling errors are calculated both for airline and grid distance for \( n = 2 \). It is to be noted that the sampling standard error is quite close to the population standard error.

G. An Upper Bound to Minimum Grid Distance in a Rectangle

The method followed was that used by M. N. Ghosh (9) for the upper bound of the minimum airline distance. The number of points \( n \) is assumed to be large. There is only one path out of the \( n! \) possible paths which gives the minimum grid distance. We specify a path which may be the minimum but in general it is not and hence its length is an upper bound to the minimum. The expected length of the \( s \)-path (specified path) is then equal to or greater than the expected length of the minimum path. The rectangular area \( A \) is divided into \( m \) squares.

For grid distance the expected within block distance is from section G

\[
(l_1) = \frac{2}{3} \sqrt{\frac{A}{m}}
\]
and the expected between block distance is

\[(l_3) = \sqrt{\frac{A}{m}} + \frac{1}{3} \sqrt{\frac{A}{m}} = \frac{4}{3} \sqrt{\frac{A}{m}}.\]

Hence, following M. N. Ghosh, assuming \( m \) is large the expected length of the \( s \)-path is

\[E(L) = (n - m)(l_1) + m(l_2), \quad (1)\]

To ascertain the value of \( m \) (in relation to \( n \)) which will make the expression (1) a minimum, differentiate with respect to \( m \) and equate to zero.

This leads to

\[\frac{m}{n} = \frac{1 - 2e^{-\frac{n}{m}}}{1 + e^{-\frac{n}{m}}} = \frac{1}{2}. \quad (2)\]

By iterative methods one may solve (2) and obtain \( m = 0.5 \) \( n \) approximately.

Then \( E(L) \) for the optimum average number of points per square is \( 1.48\sqrt{A} \) \( n \).

If one does not have the optimum average number of points per square for example if \( m = n \) or \( m = \frac{1}{3} n \) then \( E(L) \) is respectively \( 1.56\sqrt{A} n \) and \( 1.56\sqrt{A} n \).

Hence \( 1.48\sqrt{A} n \) is an upper bound for the expected grid travel among \( n \) points and it is the minimum grid length of the \( s \)-path. In many farm surveys the \( s \)-path is the one that is approximately followed for first calls when repeated callbacks are made. The interviewer has a few sampling units per county and he intends to make all his calls while in the county. He cannot minimize his path because he is continually going back and forth between s.u.'s to pick up those not-at-
home on the first visit. In fact his path for first calls to the s.u.'s may be like a path in which he chooses, at random within a county, the order in which he makes first calls to the sampling units. Furthermore the sampling unit where he finishes his interviewing may be regarded as a random point in the county. Presumably he would choose the nearest point in the next county to start interviewing and in this way he would travel a shorter distance than the x-path. But if he goes back and forth across county lines to make callbacks then again the between county path resembles the distance between two random points one in each county.
PART IV. RELATIONSHIP OF GRID TO AIRLINE DISTANCE

In graphs I, II and III the empirical average grid distance is plotted against the empirical average airline distance for all n that have been considered in Part III.F. For each shape the points fall very close to their regression line. The following table shows values of Pearson's r and the regression coefficient b for the shapes considered.

<table>
<thead>
<tr>
<th>TABLE VII</th>
</tr>
</thead>
<tbody>
<tr>
<td>Empirical Relationship of Grid to Airline Distance</td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>r</td>
</tr>
<tr>
<td>b</td>
</tr>
</tbody>
</table>

The regression line must pass through the origin since for \( n = 1 \),
\[
D_a = D_g = C.
\]

The grid distance and airline distance were in these empirical investigations measured using the same points in the same order. It would seem (if measurement and rounding errors are neglected) that expected airline distance is perfectly linearly correlated with expected grid distance. Since the regression line passes through the origin, then it appears from the data that expected grid distance varies as expected airline distance where the ratio varies with the shape of the rectangle.

A heuristic proof of why this should be so is presented below. Consider, at first, two points in a rectangle ab, \( b < a \). Take the point
Graph I. Regression of Grid Distance on Airline Distance
Rectangles 5x5, side ratio 1:1
Graph II. Regression of Grid Distance on Airline Distance
Rectangles 10x5, side ratio 2:1
### Graph III. Regression of Grid Distance on Airline Distance

Rectangles 10x2.5, side ratio 4:1
on the left to be fixed. Then the second point may occur anywhere on
the semi-circle of radius $r$ with equal probability. This statement is
not true when the semicircle passes outside the rectangle ab. However
it will be shown later that for $b$ moderately less than $a$ this exception
does not hurt the argument much.

\[ \gamma = \frac{\text{Grid distance}}{\text{Airline distance}} = \frac{x + |y|}{\sqrt{x^2 + y^2}} = |\sin \theta| + \cos \theta. \]

Thus the ratio depends only on a function of the random angle $\theta$. It
is necessary to find the expected value of the function above. Now the
probability density function is independent of $r$ and is in fact $f(\theta) = \frac{1}{\pi}$. Hence

\[ E \left( \frac{\text{Grid distance}}{\text{Airline distance}} \right) = \frac{1}{\pi} \int_{0}^{\pi} (\sin \theta + \cos \theta) \, d\theta + \frac{1}{\pi} \int_{\pi}^{0} (-\sin \theta + \cos \theta) \, d\theta \]

\[ = \frac{4}{\pi} \approx 1.273 \]
Let to that this is
value of distance. Now for in which one moves from
is the same as the expected value of sample Grid distance to the expected
average distance for any number of points, the expected value
then a, that one can approximate the expected value of observed Grid to
from the above discussion one can show, when p is not too much less
is not very good for a > 1; one would expect
for a relatively wide 1: 1, 2: 1 and 1: 1 respectively. The assumption
expected average distance. The empirical ratio is 1.277, 1.269, 1.160
the same as the expected value of the ratio of Grid to any
Grid distance becomes the expected Grid distance and the expected
average Grid distance becomes the average Grid distance and average
number of repetitions increases the average Grid distance to the average average distance. As the
Grid distance to the average distance is the same as the expected value
Hence in this case the expected value of the ratio of any partition

\[
\left( \begin{array}{c}
X_1 \\
\vdots \\
X_i \\
\vdots \\
X_n
\end{array} \right)
\frac{X_1 + \cdots + X_i}{X_1 + \cdots + X_n + X_i}
\]

written as

of the mean Grid distance of the to the mean Grid distance may be
for the empirical work k repetitions were taken, the expected value

or \( n = 2.73 \)
\[ E \left( \frac{\gamma_1 r_1 + \gamma_2 r_2 + \ldots + \gamma_n r_n}{r_1 + r_2 + \ldots + r_n} \right) \]

\[ = 1.273. \]  

(3)

As \( n \) gets larger the empirical work indicates that the ratio gets smaller. This seems reasonable since the assumption about the limits on \( \theta \) seem to be further from the truth (i.e., more of the semicircles will be partly outside the rectangle). Also for \( n \) points the empirical work was done with the criterion of minimizing overall distance, hence formula (3) would be more appropriate for Part V.A.

It may be noted that this approximate solution is absolutely true in the case of circular regions for \( n = 2 \). The ratio of expected grid distance to expected airline distance (see Part III.C) is \( \frac{h}{w} \) and this section gives the expected value of the ratio of grid to airline distance as \( \frac{h}{w} \). Also from Part V.C the expected airline distance from the center of a circle to a random point is \( \frac{2}{3} a \) and the corresponding expected distance is \( \frac{2a}{3w} \) (Part V.D).

For \( n = 2 \) and any shape of rectangle one can find the exact expected value of the ratio. If one lets \( x_2 - x_1 = u, y_2 - y_1 = v \) then the exact expectation is

\[ E(\gamma) = \frac{h}{a^2 b^2} \int_0^b \int_0^a \frac{u + v}{\sqrt{u^2 + v^2}} \ (a - u)(b - v) \, du \, dv. \]  

(4)

This integrates to
\[ H(\gamma) = \frac{1}{3k^2} \left[ \frac{\ln k + 1}{2} \log (k + \sqrt{k^2 + 1}) + \left( \frac{3k - 2}{2} \right) \sqrt{k^2 + 1} - 2k + 1 \right] \]
\[ + \frac{1}{3} \left[ \frac{\ln k + k^2}{2} \log \left( \frac{1 + \sqrt{k^2 + 1}}{k} \right) + \left( \frac{3 - 2k}{2} \right) \sqrt{k^2 + 1} - (2-k)k \right] \]

where \( k = \frac{b}{a} \), \( 0 \leq k \leq 1 \).

Now for \( k = 1, 1/2, 1/4 \) then \( H(\gamma) \) is 1.274, 1.254 and 1.205 respectively as compared with the empirical ratios 1.277, 1.256 and 1.160. The exact value for \( n = 2 \) of the ratio of expected grid distance to expected airline distance is 1.278, 1.242 and 1.167 for \( k = 1, \frac{1}{2} \) and \( \frac{1}{4} \) respectively. The approximate value (1) is very close to the actual value given by (5) when we have a square.

Now as \( k \to 0 \) it is easy to show that \( H(\gamma) \to 1 \) as one would expect since the rectangle becomes a straight line. It may also be shown that as \( k \) varies between 0 and 1 that \( H(\gamma) \) has a maximum when \( k = 1 \). Despite several attempts\(^1\) it has not been possible to show that \( H(\gamma) \) is monotonically decreasing as \( k \) decreases from 1 to 0. We would expect this would be the case on intuitive grounds and from a comparison of the values of \( H(\gamma) \) for the shapes considered.

From the above discussion the ratio of expected grid distance to expected airline distance has been shown to be approximately a constant for a particular shape of rectangle and any number of random points. Furthermore this constant is fairly stable over the shapes considered. So if one knows either expected grid distance or expected airline distance it is then a simple matter to estimate the other. For example if the expected airline distance is known for rectangles where \( b \) is not too much

\(^1\)See Appendix A
less than a one can estimate expected grid distance by multiplying by 1.2.
PART V.  EXPECTED DISTANCE OVER A REGION 
IN A PREDESIGNATED DIRECTION

A. Grid Distance Among n Points in a Rectangle

Because the complexity of finding the expected minimum grid distance 
becomes very great for \( n > 3 \) and because minimizing distance is sometimes 
not practical the order in which the points are taken is fixed in this part. 
For example if our rectangle is oriented East and West one can 
move North, South and West but never East. For \( n = 2 \) the distance is 
given in Part III.C since the order in which the points are to be taken 
does not matter.

For \( n = 3 \) the length of the path depends on the configuration of the 
points. By using the notation and theory developed in Part III.E one 
always travels a distance \( R = x_3 - x_1 + y_3 - y_1 \).

\[
\begin{align*}
(x_2, y_3) & \quad (x_1, y_2) & \quad (x_3, y_1) \\
(x_1, y_1) & \quad (x_2, y_3) & \quad (x_3, y_1)
\end{align*}
\]

\[D = (x_3-x_1) + (y_3-y_1) + (y_2-y_3) \quad D = (x_3-x_1) + (y_3-y_1) + (y_2-y_3) \quad D = (x_3-x_1) + (y_3-y_1) + (y_2-y_3)\]

\text{Figure 3.}

There are six possible reasonable paths, all equally likely, which corre-
spond to
<table>
<thead>
<tr>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>$\beta_3$</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>$x_3 - x_1 + y_3 - y_1$</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>the same average distance</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>3</td>
<td>(see figure 8)</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>$x_3 - x_1 + y_3 - y_1$</td>
</tr>
</tbody>
</table>

However in moving from left to right three of these six are the same distance on the average as the other three. Hence

$$E(D) = E(R) + \frac{2}{3} E(y_3 - y_2)$$  \hspace{1cm} (1)

or

$$E(D) = E(R) + \frac{2}{3} E(y_2 - y_1)$$

$$= \frac{a + b}{2} + \frac{2}{3} \left( \frac{b}{4} \right)$$

$$= \frac{3a + bb}{6}.$$  \hspace{1cm} (2)

This value may also be derived as follows. For 3 points the expected distance between contiguous points in the $x$ direction is $\frac{a}{4}$ and the expected distance between contiguous points in the $y$ direction is $\frac{b}{4}$. Put in the nine expected points and number them as in Figure 9. Then the length of the path $(1,5,9)$ is the expected value of the path when $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 3$; the length of the path $(6,2,9)$ is the expected value of the path when $\beta_1 = 2$, $\beta_2 = 1$, $\beta_3 = 3$, etc. Each of these permutations is equally likely so the
average of these expected lengths gives the average grid distance for 3 points viz.

\[ \frac{1}{6} (3a + 4b) . \]

It can easily be shown that for \( r \) points on a straight line of length \( x \) the expected distance between two contiguous points is \( \frac{x}{r + 1} \).

Hence in the case of four points, as above, all possible paths consist of all possible ways of joining four points, where each point is in a different row and different column of the expected points from any of the other three points selected. The design of the expected points would be as below. The average grid path is

\[ \frac{1}{24} \left( \frac{72a + 120b}{5} \right) = \frac{3a + 5b}{5} . \]  

\((3)\)

Figure 10.

Similarly for five random points average grid path is

\[ \frac{1}{120} \left( \frac{480a + 960b}{6} \right) = \frac{3a + 6b}{4.5} . \]

\((4)\)
Thus by induction one has that

\[ 2^n \, D_g = \frac{a}{3} + \frac{b}{3} \]

\[ 3 \, D_g = 2 \left( \frac{a}{4} \right) + \frac{2b}{3} \]

\[ 4 \, D_g = \frac{3}{5} a + \frac{3b}{5} \]

\[ 5 \, D_g = \frac{4}{6} a + \frac{4b}{6} \]

\[ \cdots \cdots \cdots \cdots \cdots \]

\[ n \, D_g = \frac{n-1}{n+1} a + \frac{(n-1)b}{n+1} = \left( \frac{a}{n+1} + \frac{b}{n} \right) \] (5)

where \( n \, D_g \) is the formula for expected grid distance for \( n \) random points when one travels in the principal direction of the rectangle.

However if \( b < a \) and travelling always in the 'y' direction (5) can be easily modified to

\[ n \, D_g = (n - 1) \left( \frac{a}{3} + \frac{b}{n+1} \right) \] (6)

In many surveys either (5) or (6) (or both) is the appropriate one to use. In a one stage survey in which the interviewer is assigned a group of counties, each containing a few sampling units, the interviewer has to plan to move from one county into the next and visit the s.u.'s as he goes. Assume no callbacks are made. An interviewer cannot always minimize the distance between sampling units in a county and still end up by being near the next county he wishes to enter. Minimizing the distance within a county increases the distance between sets of s.u.'s.
It hardly seems reasonable to adopt this criterion. A plan in which the interviewer moves from east to west or north to south over the county seems more reasonable. Maximizing the distance between points within a county would seem to result in more travel than this East-West or North-South plan.

Less attention may be paid to callbacks now since one can use the estimation scheme suggested by Politz and Simmons\textsuperscript{1}, or one can subsample as outlined by Hansen and Hurwitz\textsuperscript{2}.

It is interesting to note that for a square region of area $A$ with 3 random points moving over the region in a definite direction

$$ 3D_g \approx 1.167 \sqrt{A} \text{ while the minimum grid distance among 3 points in a square is } 1.121 \sqrt{A}. $$

For $n$ points there are $n - 1$ line segments joining them. Therefore from equation (5) each line segment averages $\frac{a}{n+1} + \frac{b}{3}$. Now as the number of points become infinite the average length of each line segment goes to $\frac{b}{3}$. In the limiting case on any line in the rectangle parallel to the side $b$, there are an infinite number of random points. There is no criterion for selecting among the infinite points on the line as to what order the points are to be visited in. Hence the problem is the same as the average distance between two random points on a line of length $b$. This average distance is, of course, $\frac{b}{3}$. The total expected distance (equation (5)) for an infinite number of points is infinite.

\textsuperscript{1}Politz, A. and Simmons, W. An attempt to get the "Not at Home" into the sample without callbacks. 44:9-31. 1949.

\textsuperscript{2}Hansen, M. and Hurwitz, W. The problem of nonresponse in sample surveys. 41:517-529. 1946.
It may be shown for \( n > 16 \) that the expected distance under the
criterion of moving in one direction exceeds an upper bound to minimum
grid distance given in Part III.G.

B. Grid Distance Among \( n \) Points When a
Rectangle is Stratified

Suppose one has a rectangular area \( A \) over which there is a uniform
space distribution. Suppose one draws a sample which is systematic and
aligned\(^1\) in both directions (see H. H. Quenouille (20)). If the size
of the sample is \( m \) this would mean that \( m \) square equal-sized strata are
created with one point in each stratum. The theory is not applicable
to the case where one has an approximately uniform space distribution
and consequently the strata are approximately equal in size. The
distance travelled is easily seen to be

\[
D_n = (n - 1) \frac{\sqrt{\frac{A}{m}}}{n}.
\]

This distance is evidently either
airline or grid distance. It is
rather remarkable that it agrees
very closely with the empirical
minimum grid distance of Part III.F
for \( n > 3 \) (see reference (17)).
Mahalanobis first stated this
formula for random samples in (12)
without specifying the constant.

\(^1\)It is to be noted that one can draw a sample systematic in both
directions but not aligned in either direction. The expected distance
would be greater for this case than that given in expression (1).
Now suppose that instead of a systematic sample one has a random point in each stratum.\(^1\) The grid distance is

\[
D_g = m l_1 + (m - 1) l_2
\]

where \(l_1\) is the within stratum distance and \(l_2\) is the between stratum distance. In this case \(l_1 = 0\) and \(l_2 = \frac{h}{3} \sqrt{\frac{A}{m}}\). So

\[
D_g = \frac{h}{3} \sqrt{\frac{A}{m}} \frac{n - 1}{\sqrt{n}} - \frac{h}{3} \sqrt{\frac{A}{n}}\text{ for } n \text{ large.}
\]

In general suppose there are \(k\) points per stratum and \(m\) square strata where \(m\) is large, \(km = n\). Then

\[
l_1 = \frac{(k - 1)(k + 1)}{3(k + 1)} \sqrt{\frac{A}{m}}
\]

and

\[
l_2 = \frac{k + 1}{3(k + 1)} \sqrt{\frac{A}{m}}
\]

Hence

\[
D_g = \frac{k + 3}{3 \sqrt{k}} \sqrt{\frac{A}{n}} = \frac{n + 3m}{3 \sqrt{m}} \sqrt{\frac{A}{n}}
\]

One can find the minimum of the above function with respect to \(m\). It turns out that

\[
m = \frac{1}{3} n \text{ or } k = 3
\]

is the optimum stratification for a fixed sample size \(n\) since this gives the shortest average distance, where the path is in one general direction.

---

\(^1\)This case is also approximated in a sample drawn systematically from a list of s.u.'s where the following conditions occur: (1) the s.u.'s are only approximately the same area or shape. (2) consecutively numbered s.u.'s in the list are nearly always contiguous geographically. (3) the sampling rate is not such that the sampling units chosen are systematic in a geographic sense. (4) the space distribution of s.u.'s is only approximately uniform. (5) only one s.u. is picked in each systematic interval of the list.
from one stratum into the next. At the optimum stratification

\[ n_D = \frac{2}{\sqrt{3}} \sqrt{A} n \approx 1.155 \sqrt{A} n. \]  

(5)

It has been shown in Part V.A that with 3 points in a square travelling in one direction among the points is not much greater than the expected minimum grid path. Hence the overall distance might be greater for a stratified sample when one minimizes the path within a stratum because the expected between stratum path would be longer than when one moves in a certain direction across strata.

Even for \( n \) and \( m \) not large, in which case

\[ n_D = \frac{k + 3}{3k} \sqrt{A} n \quad \text{or} \quad \frac{(k + 7) \sqrt{k}}{3(k + 1)} \sqrt{A} n, \]  

(6)

it may be shown that the integral value of \( k \) which makes this a minimum is 3, as before, for all \( n \geq 3 \). In fact on minimizing (6) with respect to \( k \) one obtains

\[ n(k - 3)(k^2 + 2k + 1) = k(k^2 - 4k + 7) \]

or

\[ \frac{(n - 1)k^3 + (4 - n)k^2 - (5n + 7)k - 3n}{3} = 0. \]  

(7)

A root of such an equation is a function of \( n \). But \( k = 3 \) gives a value +12 to the R.H.S. and this is independent of \( n \). No other integral value of \( k \) for \( n \geq 3 \) comes anywhere as near as close to being a root of the above equation (7).

\[ ^{1} \text{For 4 and 5 points per stratum } n_D \text{ is respectively } 1.167 \sqrt{A} n \text{ } \] 

and \( 1.192 \sqrt{A} n \) respectively. For 2 points \( n_D \) is \( \frac{2}{3} \sqrt{A} n + \frac{A}{n} \)

\[ = \frac{5}{3 \sqrt{2}} \sqrt{A} n \approx 1.179 \sqrt{A} n. \]
It would seem obvious that the expected minimum distance for a random sample would be less than the expected minimum distance for a stratified sample of the same size. This seems rather difficult to prove. However one can prove that for a certain shape of region A that the grid distance moving in a certain direction for a stratified sample is greater than the grid distance moving in the same direction for a random sample of the same size. Specify that the area be a rectangle with dimensions \( m \sqrt{\frac{A}{m}} \times \sqrt{\frac{A}{m}} \). The m strata then are squares with sides \( \sqrt{\frac{A}{m}} \). Hence the distance for the stratified sample of size n and k per strata is

\[
\frac{D^S}{g} = n \frac{m(k - 1)(k + 1)}{3(k + 1)} \sqrt{\frac{A}{m}} + (m - 1) \left( \frac{2}{k + 1} + \frac{1}{3} \right) \sqrt{\frac{A}{m}}.
\]

For a random sample

\[
\frac{D^R}{g} = (n - 1) \left( \frac{\sqrt{\frac{A}{m}}}{3} + \frac{m \sqrt{\frac{A}{m}}}{n + 1} \right).
\]

Simplifying one can show

\[
\frac{D^S}{g} > \frac{D^R}{g}
\]

for all \( m > 1 \). The difference between the stratified and random distance is not large on the average. It may be found to be

\[
\frac{2(m - 1)}{m(k + 1)(n + 1)}.
\]

The restriction on the shape of the area is not so drastic as one might think if the interviewers move along a tier of strata in this fashion or along a tier of counties in the random case. There would be an added
travel component for going from one tier to the next but in general it would be about the same for the stratified and the random case.

On the other hand, using a square region and under the assumption that one always moves over the region or over rows of strata in a definite direction (even though this is unrealistic when one has a large number of purely random points), in order to minimize travel one should stratify when \( n > 3 \). Use the optimum stratification and since \( m \) and \( n \) are small in this case use the exact distance, viz.

\[
m_1 + (m - 1) l_2 = \left( \frac{2}{\sqrt{3}} - \frac{5 \sqrt{3}}{6n} \right) \sqrt{A} \cdot n.
\]

Hence from equation (6) of Part V.A and the above equation one should stratify if

\[
\left( \frac{2}{\sqrt{3}} - \frac{5 \sqrt{3}}{6n} \right) \sqrt{A} \cdot n < \frac{(n - 1)(n + 1)}{3(n + 1)} \sqrt{A}
\]

or on reducing when \( n > 3 \).

It may be shown for \( n \) large and \( k > 13 \) per stratum that a stratified sample in which one moves in a certain direction at all times has a longer path than the \( n \)-path of a purely random sample of the same size.

There is an element of unreality in moving over a stratified area in one direction because for most areas there are several tiers of strata and one must change direction to move to a different tier and then proceed in one direction along that tier. The path is lengthened by this moving from one tier to the next under the above assumptions. How much is it lengthened each time one turns a corner of this kind? This increase may be easily shown for a right-angle turn to be

\[
\frac{k - 1}{k + 1} \frac{\sqrt{k} \sqrt{A}}{\sqrt{\frac{k}{n}}}
\]
where \( k \) is the number of points per stratum and \( m \) is number of square strata.

For the expected overall distance, where the interviewer moves from one stratum to the next, it is not necessary to have the same probability density over all strata as long as it is uniform within each stratum. One could have varying numbers of sample units in each stratum and using the formulas developed one could find the expected length of the path. The strata must be approximately square and have about the same area. The sampling might be proportionate stratified, but it need not be.

One should be able to divide the region into the appropriate number of squares. Rarely will one be able to do this exactly. However if the area is sufficiently large and the number of points fairly large, it is possible to approximate square strata.

C. **Mean Square Distance Among \( n \) Points in a Circle, Semicircle and Quadrant**

For \( n > 3 \) it becomes very difficult to find either expected airline or expected grid distance in circular regions. Having found mean square distance it is possible to estimate mean airline distance and mean grid distance approximately.

1. **Airline distance between a random point and a fixed point at the center**

In this section assume the headquarters is at the center of the arc and that the \( n \) points are randomly chosen from the uniform space distribution over a circle of radius \( a \). Hence the distribution of any random point \((r,\theta)\) is given in polar coordinates by
\[ f(r, \theta) \, dr \, d\theta = \frac{r \, dr \, d\theta}{n \, a^2} \]

or in Cartesian coordinates by
\[ f(x, y) \, dx \, dy = \frac{dx \, dy}{n \, a^2} \]

when the center of the circle is taken as origin.

Now the mean distance, \( r \), from the headquarters to any random point is given by
\[
E(r) = \frac{1}{n a^2} \int_0^a \int_0^{2\pi} r^2 dr \, d\theta = \frac{2}{3} a. \quad (1)
\]

2. **Distance among random points for \( n = 2, 3 \) in a circle**

Now suppose there are two random points and that the interviewer's load is such that he can visit both points before returning to his headquarters. What is the mean distance travelled? The mean distance from the headquarters to either of the points is \( 2/3 \, a \) and from Part III.C the mean distance between two random points in a circle is \( \frac{128}{45\pi} a \). Hence the question is answered. However it is of interest to compute the mean square distance between two random points \( (r_1, \theta_1) \) and \( (r_2, \theta_2) \). Denote the distance by \( z_1 \). Then
\[
aE(z_1^2) = \frac{1}{(n a^2)^2} \int_0^a \int_0^a \int_0^{2\pi} \int_0^{2\pi} \left[ r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_2 - \theta_1) \right] r_1 r_2 \, d\theta_2 d\theta_1 dr_1 dr_2 \]
\[
= a^2. \quad (2)
\]
Now
\[ 2E(s_1) = \frac{12a}{\pi^2} \approx 0.905a \]
while
\[ \sqrt{2E(s_1^2)} = a. \]
Hence the mean distance is 9.5 percent less than the square root of the mean square distance. The square root of the mean square distance can always be regarded as an upper bound to the corresponding mean distance.

In the case of three random points \((r_1, \theta_1)(r_2, \theta_2)\) and \((r_3, \theta_3)\) the problem is to find the mean distance travelled if the interviewer can visit all three before returning to his headquarters. There are 3! ways of choosing the order in which the points are taken and three of these, in general, will result in a different length of path. (In the case of 2 points there were 2! ways of choosing the order in which the points were taken but both of these resulted in the same path). Of these three different paths one is a minimum. For three points it is desirable to always take the minimum path. However the finding of the expected length of the minimum path
would involve reasoning similar to Part III.E, only a good deal more complex. Furthermore it would be a difficult problem\(^1\) to even find the mean distance between any pair of points. For instance the first point of the pair is picked at random from the three points and other point of the pair is the point next to it in a counter-clockwise direction. If one represents this distance by \(z_1\), the first point by \((r_1, \theta_1)\) and the second by \((r_2, \theta_2)\) then

\[
3^E(z_1) = \frac{3!}{(\pi a^2)^3} \int_0^a \int_0^a \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{r_1 r_2^2 - 2r_1 r_2 \cos(\theta_2 - \theta_1)}{r_1 r_2 r_3^2 d\theta_3 d\theta_2 d\theta_1 dr_1 dr_2 dr_3}.
\]

However it is possible to find the mean square distance between any pair of points which are located adjacent to each other in a counter-clockwise direction. This is given by

\[
3^E(z_1^2) = \frac{3!}{(\pi a^2)^3} \int_0^a \int_0^a \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \frac{(r_1^2 + r_2^2 - 2r_1 r_2 \cos(\theta_2 - \theta_1))}{r_1 r_2 r_3^2 d\theta_3 d\theta_2 d\theta_1 dr_1 dr_2 dr_3} = a^2 \left(1 - \frac{k}{3\pi a}\right).
\]

Hence if the interviewer picks the first point to visit at random from the three points then his mean distance to this point is given by equation (1). If from this point he moves always in either a clockwise or counter-clockwise direction (determined in advance) in order to visit the

\(^{1}\)With 2 points in a circle Williamson (30) arrived at the mean distance between these random points by a trick method.
remaining points, then the mean square distance between the first and second points is given by equation (3). Between the second and third points formula (3) is valid also. In addition we can regard the third point as random since the first point was random and we moved in a direction chosen in advance. Hence formula (1) will give the mean distance.

3. Distance between adjacent pairs of points in a circle

In general for n points if the interviewer picks the first point to visit at random from the n points that are given and then proceeds always in a certain direction chosen in advance (say counter-clockwise), the distance between any pair of adjacent points \((r_k, \theta_k)\) and \((r_{k+1}, \theta_{k+1})\) is given by

\[
E(s_k^2) = \frac{n!}{(n^2)^n} \int_0^{2\pi} \cdots \int_0^{2\pi} r_k^{2n} r_{k+1}^{2n} \cos^{2n}(\theta_{k+1} - \theta_k) \, dr_k \cdots dr_k
\]

If one relabels the n points and then the above may be written as

\[
E(s_l^2) = \frac{n!}{(n^2)^n} \int_0^{2\pi} \cdots \int_0^{2\pi} r_1^2 r_2^2 \cos^2(\theta_2 - \theta_1) \, dr_1 \cdots dr_k
\]

If the interviewer proceeds in a pre-chosen direction from first to last point there would be \((n - 1)\) such mean square distances as given
above. Furthermore the first and last points were chosen at random
from the n points so that we know the mean distance from headquarters
to these points is given by equation (1). It is not necessary that
the interviewer visit the n points before returning to his headquarters
in order to obtain mean square distances or mean distances as the case
may be. If the interviewer, say, starts at a random point of the
configuration and returns to his headquarters when a point on his route
happens to be near the headquarters then, of course, equation (1) would
not give the mean distance to the headquarters. If however the inter-
viewer breaks off his circuit for any other reason, then the mean
distance to his headquarters will be given by (1). When he resumes
his circuit he must again proceed from point to point in a definite
direction. The interviewer may start again at the next point and main-
tain his counter-clockwise course or he may pick a point at random, not
yet visited, and proceed in either direction (chosen without reference
to the configuration of the points). He may also start at the last
point and proceed in a clockwise direction.

Now to evaluate the integral of equation (4) one can reduce it to

\[ E(a^2) = \frac{n!}{(na^2)^n} \int_0^{2\pi} \int_0^{2\pi} \left[ \frac{(a^n)_{n+1}}{2^{n+1}} + \frac{2(a^2)_{n+1}}{3!} + \frac{2(a^2)^{n+1}}{3.2^n} \cos(\theta_2 - \theta_1) \right] \frac{(2\pi - \theta_2)^n}{(n-2)!} d\theta_2 d\theta_1 \]

\[ = \frac{n!}{(na^2)^n} \int_0^{2\pi} \int_0^{2\pi} \left[ \frac{(a^n)_{n+1}}{2^n} - \frac{g}{9} \frac{(a^2)_{n+1}}{2^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2\pi - \theta_2)^n}{(n-2)!} \cos(\theta_2 - \theta_1) d\theta_2 d\theta_1 \right] \]

\[ = a^2 \left[ 1 - \frac{g}{9} \frac{n!}{(2\pi)^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2\pi - \theta_2)^n}{(n-2)!} \cos(\theta_2 - \theta_1) d\theta_2 d\theta_1 \right]. \quad (5) \]
In order to evaluate the above integral it is necessary to expand
\((2\pi - \theta_2)^{n-2}\) by the binomial theorem. When one does so one is interested
in evaluating integrals of the kind

\[
\int_0^{2\pi} \int_0^{2\pi} \theta_2^m \cos((\theta_2 - \theta_1))d\theta_2 d\theta_1 \quad m = 0, 1, \ldots, n - 2
\]

\[
= \int_0^{2\pi} \int_0^{2\pi} \theta_2^m (\cos \theta_2 \cos \theta_1 + \sin \theta_2 \sin \theta_1) d\theta_2 d\theta_1 .
\]

(6)

Now if \(m\) is even

\[
\int_0^{2\pi} \int_0^{2\pi} \theta_2^m \cos \theta_2 \cos \theta_1 d\theta_2 d\theta_1 + \int_0^{2\pi} \int_0^{2\pi} \theta_2^m \sin \theta_2 \sin \theta_1 d\theta_2 d\theta_1
\]

\[
= m! \int_0^{2\pi} \left( \frac{\theta_2^m}{m!} - \frac{\theta_2^{m-2}}{(m-2)!} + \ldots + \frac{(-1)^{m/2}}{\frac{m}{2}} \right) \sin \theta_2 \cos \theta_1 d\theta_1
\]

\[
+ m! \int_0^{2\pi} \left( \frac{\theta_2^{m-1}}{(m-1)!} - \frac{\theta_2^{m-3}}{(m-3)!} + \ldots + \frac{(-1)^{m/2}}{\frac{m}{2}} \frac{\theta_2^{m-2}}{(m-2)!} \right) \cos \theta_2 \cos \theta_1 d\theta_1
\]

\[
+ m! \int_0^{2\pi} \left( -\frac{\theta_2^m}{m!} + \frac{\theta_2^{m-2}}{(m-2)!} + \ldots - \frac{(-1)^{m/2}}{\frac{m}{2}} \right) \cos \theta_2 \sin \theta_1 d\theta_1
\]
\[ + m! \int_0^{2\pi} \left( \frac{\theta_{m-1}}{(m-1)!} - \frac{\theta_{m-3}}{(m-3)!} + \ldots + (-1)^{\frac{m}{2}} \frac{\theta_{m}}{(m)!} \right) \sin \theta d\theta \]

\[ = m! \int_0^{2\pi} \left[ \sum_{k=1}^{\frac{m+2}{2}} (-1)^k \frac{\theta_{m-2k+2}}{(m-2k+2)!} \right] \sin \theta \cos \theta d\theta \]

\[ + m! \int_0^{2\pi} \left[ \sum_{k=1}^{\frac{m}{2}} (-1)^{k-1} \frac{(2\pi)^{m-2k+1}}{(m-2k+1)!} + \sum_{k=1}^{\frac{m+2}{2}} (-1)^{k+1} \frac{(2\pi)^{m-2k+2}}{(m-2k+2)!} \right] \cos \theta d\theta \]

Collecting the terms and using the fact that \( \int_0^{2\pi} \cos \theta d\theta = 0 \) the above reduces to

\[ = \int_0^{2\pi} \sin \theta d\theta = 0 \] the above reduces to

\[ m! \int_0^{2\pi} \left[ \sum_{k=1}^{\frac{m}{2}} (-1)^k \frac{\theta_{m-2k+1}}{(m-2k+1)!} \right] d\theta \]

This on integration becomes

\[ m! \left[ \sum_{k=1}^{\frac{m}{2}} (-1)^{k} \frac{(2\pi)^{m-2k+2}}{(m-2k+2)!} \right] \] (7)
Similarly if \( m \) is odd (6) becomes

\[
m! \int_0^{2\pi} \sum_{k=1}^{m+1} \frac{(-1)^k \theta_1^{m-2k+2}}{(m-2k+2)!} \sin \theta_1 \cos \theta_1 \, d\theta_1
\]

\[
+ m! \int_0^{2\pi} \left[ \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \right] \left[ \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \cos \theta_1 \right] \, d\theta_1
\]

\[
+ m! \int_0^{2\pi} \left[ \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \right] \left[ \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \cos \theta_1 \right] \, d\theta_1
\]

\[
+ m! \int_0^{2\pi} \left[ \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \right] \left[ \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \sin \theta_1 \right] \, d\theta_1
\]

\[
= m! \left[ \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \right] . \quad (7)
\]

Now

\[
\int_0^{2\pi} \int_0^{2\pi} \frac{(2\theta - \theta_1)^{n-2}}{(n-2)!} \cos \left( \theta_2 - \theta_1 \right) \, d\theta_2 \, d\theta_1
\]

\[
= \int_0^{2\pi} \int_0^{2\pi} \left[ \sum_{m=0}^{n-2} \frac{(-1)^m \theta_2^{n-m-2} \theta_2^m}{(n-m-2)! m!} \right] \cos \left( \theta_2 - \theta_1 \right) \, d\theta_2 \, d\theta_1.
\]
By (7) the above equals

\[
\sum_{m=0}^{n-2} (-1)^m \frac{(2m)_{n-m-2}}{(n-m-2)!} \left[ \sum_{k=1}^{\frac{m}{2}} \text{or} \frac{m+1}{2} \right] \frac{(2n)^m-2k+2}{(m-2k+2)!} \]

with the last term of the second summation \( \frac{m}{2} \) if \( m \) is even, or \( \frac{m+1}{2} \) if \( m \) is odd. However (8) is not in a very concise form. Fortunately it can be reduced to a somewhat neater form. If one expands (8) and finds the coefficients of the powers of \( 2^n \) one will find (8) becomes for \( n \) even

\[
\sum_{j=1}^{n-2} \frac{(2^n)_{n-2j}}{(n-2j)!} (-1)^j \left[ \sum_{r=0}^{n-2j-1} (-1)^r \right] \frac{(n-2j)_{n-2j}}{(n-2j)!} \left[ (1-1)^{n-2j} \frac{(n-2j)!}{(n-2j)!} \right]
\]

\[
= \sum_{j=1}^{n-2} (-1)^{j-1} \frac{(2^n)_{n-2j}}{(n-2j)!} \cdot (9)
\]

Similarly for \( n \) odd

\[
\sum_{j=1}^{n-1} \frac{(2^n)_{n-2j}}{(n-2j)!} (-1)^{j-1} \left[ \sum_{r=0}^{n-2j-1} (-1)^r \right] \frac{(n-2j)_{n-2j}}{(n-2j)!} \left[ (1-1)^{n-2j} \frac{(n-2j)!}{(n-2j)!} \right]
\]

\[
= \sum_{j=1}^{n-1} (-1)^{j-1} \frac{(2^n)_{n-2j}}{(n-2j)!} \cdot (9)
\]

The formula (8) may also be reduced to a different form, viz.

\[
\sum_{j=1}^{n-2} \frac{(2^n)^{2j}}{(2j)!} (-1)^{\frac{n-2j-2}{2}} \cdot (10)
\]
for n even and

\[
\frac{n-1}{2} \sum_{j=1}^{\frac{n-2}{2}} \frac{(2^n)^{2j-1} \cdot \frac{n-2j-1}{2}}{(2j-1)!} (-1)^j \quad (10)
\]

for n odd.

Using formula (9) in (5) one finds that

\[
E(z_1^2) = a^2 \left[ 1 + \frac{8}{9} \sum_{j=1}^{n-2 \text{ or } \frac{n-1}{2}} (-1)^j \frac{n!}{(2^n)^{2j} (n-2j)!} \right] \quad (11)
\]

This is the mean square distance between any pair of adjacent points (in a counter-clockwise direction say) provided the first point of the n was chosen at random. If one takes the square root of (11) he would obtain an upper bound for the mean distance between pairs of points taken in the given order. For n points it is not known how much \( n E(z_1^2) \) overestimates \( E(z_1^2) \). It has been shown earlier that for \( n = 2 \) the mean distance is 9.5 percent less than the mean square distance. A little later on it will be shown for \( n = \infty \) that the mean distance is 6.6 percent less than the mean square distance. For n between 2 and \( \infty \), on general grounds, it would seem that the amount of bias is between 9.5 percent and 6.6 percent. In fact the bias may decrease monotonically as n increases.

It is perhaps easiest to find \( \lim_{n \to \infty} E(z_1^2) \) by going back to equation (5) and considering the integral

\[
\frac{n!}{(2^n)^n} \int_{0}^{2\pi} \int_{0}^{2\pi} \frac{(2^n - 2) \cdot n-2}{(n-2)!} \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 . \quad (12)
\]
Now the maximum value of $\cos(\theta_2 - \theta_1)$ is 1. If we put 1 in for $\cos(\theta_2 - \theta_1)$ then (12) is

$$\leq \frac{n!}{(2n)^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2n - \theta_2)^n}{(n-2)!} \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 \leq 1 \text{ for all } n.$$  

Now integrating (12) by parts twice one can show that the expression (12) is equal to

$$1 - \frac{n!}{(2n)^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2n - \theta_2)^n}{n!} \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 . \quad (13)$$

Now in a similar fashion to the above one can show that the right member of the above expression is

$$\leq \frac{n!}{(2n)^n} \frac{(2n)^{n+2}}{(n+2)!} = \frac{(2n)^n}{n(n+1)} \text{ for all } n.$$  

Hence

$$\lim_{n \to \infty} \frac{n!}{(2n)^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2n - \theta_2)^n}{n!} \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 \leq 0$$

and hence

$$\lim_{n \to \infty} \frac{n!}{(2n)^n} \int_0^{2\pi} \int_0^{2\pi} \frac{(2n - \theta_2)^{n-2}}{(n-2)!} \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 \geq 1.$$  

But (12) has been proven $\leq 1$ for all $n$ so for $n = \infty$ it must be equal to 1.
Since this is the case then from (5)

\[ E(s_1^2) = \frac{a^2}{9}. \]  

(14)

This result may be arrived at in a slightly different way by considering the right member of (13) in the form

\[
\int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \cos (\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1
\]

which is

\[
\int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \, d\theta_2 \, d\theta_1 + \int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \, d\theta_2 \, d\theta_1
\]

\[
\int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \, d\theta_2 \, d\theta_1
\]

Figure 15.

Now for an \( \varepsilon \) however small one can find an \( n_0 \) large enough and an \( \gamma_0 \) small enough so that for all \( n > n_0, \gamma < \gamma_0 \)

\[
\int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \, d\theta_2 \, d\theta_1 < \frac{\varepsilon}{3},
\]

\[
\int_0^{2\pi} \int_0^{2\pi} (1 - \frac{\theta_2}{2\pi})^n \, d\theta_2 \, d\theta_1 < \frac{\varepsilon}{3}
\]
and
\[ \int_{0}^{\frac{\pi}{2}} \int_{\theta_1}^{\pi} \left(1 - \frac{\theta_2}{\pi}\right)^n \, d\theta_2 \, d\theta_1 < \frac{\epsilon}{3}. \]

Hence
\[ \int_{0}^{2\pi} \int_{0}^{2\pi} \left(1 - \frac{\theta_2}{2\pi}\right)^n \cos(\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1 \leq \epsilon. \]

and this can be made as small as one pleases by making n large enough.

Hence (12) can be made arbitrarily close to 1. Putting this in equation (5) one has as before
\[ E(x^n) = \frac{\alpha^n}{a^2}. \]  

(14)

The third way this result is arrived at is to notice what happens as n becomes infinite in the circle. If one draws any radius then, of course, there are a denumerable infinity of points on that line. However there is no rule for taking them in order, since in this case one is not minimizing the distance. The only rule is that the points be taken in order in a counter-clockwise direction. Hence one would select two points at random from the infinite number on the radius. One would want to know what is the mean square distance between them. This is given by

![Figure 16.](image-url)
\[ E(z_1)^2 = \frac{2}{(na)^2} \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \int_0^a (r_1 - r_2)^2 r_1 r_2 d\theta d\phi d\theta d\phi = \frac{a^2}{9}. \] (14)

But \( E(z_1) \) can also be obtained using the above reasoning to be

\[ E(z_1) = \frac{2}{(na)^2} \int_0^{2\pi} \int_0^{2\pi} \int_0^{2\pi} \int_0^a (r_1 - r_2) r_1 r_2 d\theta d\phi d\theta d\phi = \frac{2}{15}. \]

Hence one can find how much \( \sqrt{\infty E(z_1)} = \frac{a}{3} \) overestimates \( \infty E(z_1) = \frac{2}{15} \) as we did earlier.

While one can estimate approximately how much the square root of the mean square distance overestimates the mean distance we do this under the assumption of moving in a definite direction. However in a certain percentage of all possible choices of the locations of the \( n \) points it will be possible to find a path among them that would be shorter than moving in a counter-clockwise direction. Hence the mean distance overestimates the mean minimum distance. It may be difficult to tell in many configurations what is the minimum path and this difficulty increases very rapidly as \( n \) increases. An obvious configuration where going in a certain direction would not minimize the total path is

![Figure 17.](image)
illustrated in Figure 17.

All the above development is for airline distance. However Part IV enables one to transform to mean grid distance by multiplying by 1.2 approximately.

4. Distance between adjacent pairs of points in a semicircle

Similarly for a semicircle with n random points one can find the mean square distance between any two points, where the points are taken in a definite direction. Then

\[
\begin{align*}
E(r^2) & = \left( \frac{2}{na} \right)^n n! \int_0^\infty \int_0^{\pi} \int_0^{\theta_1} \int_0^{\theta_2} \cdots \int_0^{\theta_{n-1}} \left[ r_1^2 + r_2^2 - 2r_1r_2 \cos(\theta_2 - \theta_1) \right] \\
& = \left( \frac{2}{na^2} \right)^n n! \left[ \int_0^\infty \left( \frac{a^2}{n!} \right)^{n+1} + \left( \frac{a^2}{2n+1} \right)^{n+1} \right. \\
& \quad \left. + \frac{2(a^2)^{n+1}}{2n+1} \int_0^\infty \int_0^{\pi} \int_0^{\theta_1} \frac{(n-\theta_2)^{n-2}}{(n-2)!} \cos(\theta_2-\theta_1) d\theta_2 d\theta_1 \right] \\
& = a^2 \left[ 1 - \frac{6}{5} \frac{n!}{n^n} \int_0^{\pi} \int_0^{\theta_1} \frac{(n-\theta_2)^{n-2}}{(n-2)!} \cos(\theta_2-\theta_1) d\theta_2 d\theta_1 \right].
\end{align*}
\]

Take \( n \) even and consider

\[
\int_0^{\theta_1} \int_0^{\pi} \gamma_n (\cos \theta_2 \cos \theta_1 + \sin \theta_2 \sin \theta_1) d\theta_2 d\theta_1
\]

(16)
\[ + m! \int_0^\pi \left[ \sum_{k=1}^{\frac{m}{2}} \frac{\eta^{m-2k+1}}{(m-2k+1)!} + \sum_{k=1}^{\frac{m}{2}} (-1)^k \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \cos \theta_1 \cos \theta_1 d\theta_1 \right] \]

\[ + m! \int_0^\pi \left[ \sum_{k=1}^{\frac{m+2}{2}} (-1)^{k+1} \frac{\eta^{m-2k+2}}{(m-2k+2)!} + \sum_{k=1}^{\frac{m+2}{2}} (-1)^{k+1} \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \cos \theta_1 \sin \theta_1 d\theta_1 \right] \]

\[ + m! \int_0^\pi \sum_{k=1}^{\frac{m}{2}} (-1)^k \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \sin^2 \theta_1 d\theta_1 \]

Now collecting terms and using the fact that

\[ \int_0^\pi \sin \theta_1 d\theta_1 = 2 \text{ and } \int_0^\pi \cos \theta_1 d\theta_1 = 0 \text{ the above becomes} \]

\[ = \frac{m+2}{2} \sum_{k=1}^{\frac{m+2}{2}} (-1)^{k+1} \frac{\eta^{m-2k+2}}{(m-2k+2)!} + m! \sum_{k=1}^{\frac{m+2}{2}} (-1)^k \frac{\eta^{m-2k+2}}{(m-2k+2)!} \]

\[ = m! \sum_{k=1}^{\frac{m+2}{2}} (-1)^{k+1} \frac{\eta^{m-2k+2}}{(m-2k+2)!} + 2m! (-1)^m \]

For \( m \) odd (16) reduces to

\[ 2m! \sum_{k=1}^{\frac{m+1}{2}} (-1)^{k+1} \frac{\eta^{m-2k+2}}{(m-2k+2)!} + m! \sum_{k=1}^{\frac{m+1}{2}} (-1)^k \frac{\eta^{m-2k+2}}{(m-2k+2)!} \]
\[
\begin{align*}
&= m! \sum_{k=1}^{\frac{m+1}{2}} (-1)^{k+1} \frac{v^{m-2k+2}}{(m-2k+2)!} \\
&= m! \sum_{k=1}^{\frac{m+1}{2}} (-1)^{k+1} \frac{v^{m-2k+2}}{(m-2k+2)!}.
\end{align*}
\]

Hence

\[
\int_{\theta_1}^{\theta_2} \int_{\theta_1}^{\theta_2} \sum_{m=0}^{n-2} (-1)^{m} \frac{v^{n-m-2}}{(n-m-2)!} m^2 \cos (\theta_2 - \theta_1) \, d\theta_2 \, d\theta_1
\]

\[
= \sum_{m=0}^{n-2} (-1)^{m} \frac{v^{n-m-2}}{(n-m-2)!} \sum_{k=1}^{\frac{m+1}{2}} (-1)^{k+1} \frac{v^{m-2k+2}}{(m-2k+2)!} + 2 \sum_{m=0}^{n-3} (-1)^{m} \frac{v^{n-m-2}}{(n-m-2)!}.
\]

The first term above may be reduced in the same manner as (8) was reduced to (9). The last term exists only for even values of \( m \) and hence one can set \( m = 2k \) in it. One then obtains for (17)

\[
\sum_{j=1}^{\frac{n-2}{2} \text{ or } \frac{n-1}{2}} (-1)^{j} \frac{v^{n-2j}}{(n-2j)!} + 2 \sum_{k=0}^{\frac{n-2}{2} \text{ or } \frac{n-1}{2}} (-1)^{k} \frac{v^{n-2k-2}}{(n-2k-2)!}.
\]

This may be reduced further by setting \( k + 1 = j \) and thus obtaining for the second term

\[
2 \sum_{j=1}^{\frac{n}{2} \text{ or } \frac{n-1}{2}} (-1)^{j-1} \frac{v^{n-2j}}{(n-2j)!}.
\]

Hence (17) reduces to

\[
\sum_{j=1}^{\frac{n}{2} \text{ or } \frac{n-1}{2}} (-1)^{j-1} \frac{v^{n-2j}}{(n-2j)!} + 2 (-1)^{\frac{n-2}{2}}
\]

median
where the last term is present for \( m \) even. Otherwise it is zero. Hence we have for (15)

\[
\begin{align*}
W(n^2) &= a^2 \left[ 1 - \frac{2}{9} \sum_{j=1}^{n-2} \frac{(-1)^{j-1} n!}{(n-2j)!} \frac{n^2}{n^2} \frac{2n^4}{2^2} \frac{n-2}{2} \right]
\end{align*}
\]

(18)

where the last term in \( \{ \} \) brackets is added only when \( n \) is even.

5. Distance between adjacent pairs of points in a quadrant

One may also derive the mean square distance between adjacent points in a quadrant when there are \( n \) points drawn at random from a uniform space distribution over the quadrant. This is given by

\[
\begin{align*}
W(n^2) &= n! \left( \frac{r_1^2 + r_2^2 - 2r_1r_2 \cos (\theta_2 - \theta_1)}{n^2} \right)
\end{align*}
\]

For \( n \) even the integral

\[
\int_0^\frac{\pi}{2} \int_0^\frac{\pi}{2} \omega \left( \cos \theta_2 \cos \theta_1 + \sin \theta_2 \sin \theta_1 \right) d\theta_2 d\theta_1
\]

(20)
\[ m! \int_0^\pi \left[ \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k-1} \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} + \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^k \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \right] \sin \theta_1 \cos \theta_1 \theta_1 \, d\theta_1 \]

\[ + m! \int_0^\pi \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^k \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \cos^2 \theta_1 \, d\theta_1 \]

\[ + m! \int_0^\pi \left[ \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k+1} \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} \cos \theta_1 \right] \sin \theta_1 \, d\theta_1 \]

\[ + m! \int_0^\pi \left[ \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k-1} \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} + \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^k \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \right] \sin \theta_1 \, d\theta_1 \]

Collecting terms and using the fact that \( \int_0^{\pi/2} \sin \theta_1 \, d\theta_1 = \int_0^{\pi/2} \cos \theta_1 \, d\theta_1 = 1 \) results in

\[ \frac{m+2}{2} \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k-1} \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} + m! \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^k \frac{\theta_1^{m-2k+2}}{(m-2k+2)!} + m! \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k-1} \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \]

\[ = m! (-1)^{\lfloor \frac{m}{2} \rfloor} + m! \sum_{k=1}^{\lfloor \frac{m}{2} \rfloor} (-1)^{k-1} \frac{\theta_1^{m-2k+1}}{(m-2k+1)!} \times \]

For \( m \) odd (20) becomes
\[
\frac{m^2}{2} \sum_{k=1}^{m} (-1)^{k-1} \left( \frac{n}{2} \right)^{m-2k+1} = \frac{m^2}{2} \sum_{k=1}^{m} (-1)^{k-1} \frac{\left( \frac{n}{2} \right)^{m-2k+1}}{(m-2k+1)!}
\]

Hence

\[
\int \int \int_{\theta_1}^{\theta_2} (\theta_2 - \theta_1)^{n-2} \cos (\theta_2 - \theta_1) d\theta_2 d\theta_1
\]

where \( k = 2m \) since the last term exists only for even \( m \). The first term may be reduced by finding the coefficients of the powers of \( \frac{n}{2} \) and using the fact that

\[
-\frac{(n-3)}{(n-3)!} = 0.
\]

For \( n \) even the first term disappears. For \( n \) odd it equals \((-1)^{(n-1)/2}\). Hence (21) becomes for \( n \) odd after setting \( k + 1 = j \)

\[
\frac{n-1}{2} \left( \frac{n}{2} \right)^{n-2j} \sum_{j=1}^{n-1} (-1)^{j-1} \frac{\left( \frac{n}{2} \right)^{n-2j}}{(n-2j)!}
\]
converted to grid, the trial for the theory developed in Part I. If
how much the plane is. In this estimated mean distance have been
distance has been converted to mean distance by 

\[ d = r \sqrt{\frac{1 - \frac{Z^2}{N}}{1 - \frac{Z^2}{M}}} \]

In the formula below the square root of the mean square

and in the same direction, the mean square distance between any two points of

where the shape of the region has very little effect on mean distance. One can

It was pointed out earlier in Part III. 0. that for a = 2. and 0. the mean distance between the

In the later an 

where the second term of the odd

Now putting these in (19), one has

\[ \frac{d}{\sqrt{\frac{1 - \frac{Z^2}{M}}{1 - \frac{Z^2}{N}}}} \]

and for a even becomes
Square and a circular region where tested in the above table.

The effect of the difference in the deflection of "circular" for a circular region and corrected to mean Erd distance, as in the section and corrected for a square as well as the formulae developed and the WHO gives an accurate result. As such, the formulae developed much simpler to use the formulae for a square for the circular region.

Mean square distance in the circular region, mean for n equal to the square. 

For a large square number it is almost imperceptible to calculate the

| 1.20 | 6° | 9° | 9° |
| 1.59 | 25° | 56° | 56° |
| 1.97 | 25° | 56° | 56° |
| 2.50 | 36° | 74° | 71° |
| 2.85 | 36° | 74° | 71° |

<table>
<thead>
<tr>
<th>Conversion Factors</th>
<th>Conversion of approximate exact distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1.18m 1.36m 1.50m 1.72m</td>
</tr>
</tbody>
</table>

(All regions have the same area)

Comparison of approximate exact distance

Table III

Obtain the exact mean distance from Part III.

From Part A: If 

\( \frac{y}{1} + (\frac{2}{1}) \) then

the Erd distance between adjacent points in a square of area \( a \) is given

known exactly from the theory developed earlier in this section. The

approximation of the conversion factor, for \( n = \infty \) the Erd distance is
D. Grid Distance From a Fixed Point to a Random Point

First consider a circle, semicircle or quadrant in which the fixed point is the centre for the arc described. If one lets this point be the origin then the absolute distance from this fixed point to the random point is the same for every quadrant. If one chooses the first quadrant then

\[
E(r_g) = \frac{\mu}{\pi a^2} \int_0^{\pi/2} \int_0^a (r \sin \theta + r \cos \theta) r \, dr \, d\theta
\]

\[
= \frac{2a}{\pi^2}
\]

where \(a\) is the radius of the arc. In Part V.C(1) the corresponding expected airline distance was found to be \(\frac{2}{3} a\) and hence the ratio of expected grid distance to airline distance is \(\frac{\mu}{\pi}\). In

Part IV we showed the expected value of the ratio under some circumstances was also \(\frac{\mu}{\pi}\).

Next, in a rectangle in which the fixed point is at one of the angles of the rectangle find the expected grid distance to a random point in the rectangle. Again let the fixed point be the origin and let the sides of the rectangle be \(a\) and \(b\) then
\[ E(r_g) = \frac{1}{ab} \int_{0}^{b} \int_{0}^{a} (x+y) \, dx \, dy \]

\[ = \frac{a + b}{2}. \quad \text{(2)} \]

The corresponding airline distance for a square was found by Williamson (30) p. 370 ex. 71 to be

\[ E(r_a) = \frac{1}{3} \left[ a \sqrt{2} + a \log 1 + \sqrt{2} \right] \approx 0.765a. \]

The ratio of expected grid distance to expected airline distance is approximately 1.307.

Now for a rectangle one wants to find the expected absolute grid distance to a random point from a point fixed anywhere within the rectangle. Suppose furthermore that this point, with respect to the origin at one angle, has coordinates \((ka, jb)\) where \(0 \leq k, j \leq 1\). Now divide the original rectangle into four smaller rectangles such that an angle of each one is at \((ka, jb)\). From (2) for each of these small rectangles one can calculate the expected absolute grid distance from \((ka, jb)\) to a random point in the rectangles, given that there is a random point in each one. If there is only one random point in the rectangle, \(ab\), the probability of its falling in any of the smaller rectangles would be proportionate to
its size. (The assumption throughout this section is that the random point(s) has a uniform distribution). Hence the distance from \((ka, jb)\) to a random point would be the expected distance in each of the smaller rectangles multiplied by the probability of the random point being there, i.e.

\[
\mathbb{E}(r_g) = \frac{ak + bj}{2} z_j + \frac{a(1-k) + bj}{2} (1-k)j + \frac{ak + b(1-j)}{2} k(1-j)
\]

\[
+ \frac{a(1-k) + b(1-j)}{2} (1-k)(1-j) = \frac{a(1-2k+2k^2)+b(1-2j+2j^2)}{2}.
\]  

As a special case of this the fixed point is taken at the center of the rectangle. In this case \(k = j = \frac{1}{2}\) and

\[
\mathbb{E}(r_g) = \frac{a+b}{4}.
\]

Suppose now the fixed point is outside the rectangle ab. Make this point the origin with axes parallel to the sides of the rectangle. Put in the absolute values of the coordinates for the random point and for the vertices of the rectangle as shown in Figure 21. Then

\[
\mathbb{E}(r_g) = \frac{1}{ab} \int_{d}^{e} \int_{c}^{d} (x+y) dx dy = \frac{e + c}{2} + \frac{f + d}{2}.
\]  

\(^1\)This is really the average expected distance from the center of the rectangle to four points, one in each of four equisized strata. However we can show that the expected distance in this case is the same as the expected distance from the center of a rectangle to a random point. The variance of the distance for the case considered in the text would be much less than when there is but one point at random in the rectangle.
Actually $e - c = a$ and $f - d = b$. However if the point outside the rectangle is such that one axis cuts the rectangle into two smaller rectangles the formula must be modified somewhat. In fact, one can apply formula (4) to each of the smaller rectangles, under the assumption that a random point is in each, and then multiply by the probability of the random point being in one or the other. For example in Figure 22

$$E(r_e) = \left[ \frac{0+e}{2} + \frac{f+d}{2} \right] \frac{e(f-d)}{(d+e)(f-d)} + \left[ \frac{0+e}{2} + \frac{f+d}{2} \right] \frac{c(f-d)}{(d+e)(f-d)}$$

$$= \frac{(e+f+d)}{2(d+e)} e + \frac{(e+f+d)}{2(d+e)} c.$$

**Figure 22.**

The theory of the distance from a fixed point to a random point can be extended to the distance from a fixed point to any number of independently chosen random points. The expected distance from a fixed point to a random point should estimate the average distance of the fixed point from the randomly chosen points.
PART VI. ANALYSIS OF TIME AND MILEAGE IN THE FIELD FOR THREE IOWA FARM SURVEYS

A. Description of the Surveys

For the purposes of this report the following descriptions of these surveys should be sufficient for understanding the analysis. All three inquiries were designed to obtain estimates for the whole state.

The "Sources of Information" survey in October 1947 used 351 half-size Open Country Master Sample segments. The counties were strata. Within counties from a list, segments were drawn systematically at the same rate for each county. In the geographic sense the segments chosen may be regarded as random. This will be called a random systematic selection. Each segment was cruised, i.e. the households were numbered and all tracts of land within the designated area mapped and identified as part of some farm. For those farms over 30 acres the operator and homemaker were to be both interviewed, while for farms under 30 acres a shortened schedule was used with either the homemaker or operator as a respondent. Revisitation of the operators and homemakers of farms over 30 acres who were not-at-home on first call\(^1\) was required. For example, the instructions to the interviewers were: (a) Never leave the county without trying to contact the not-at-home on two different days. (b) If the interviewer is at any time (after leaving the county) within 15 miles of the non-interviewed farm, make another try to contact the respondent.

\(^{1}\)In this thesis call and callback are used interchangeably.
The "April Survey of Agriculture, 1948" was based on interviews with farm operators on 403 farm segments in Open Country, Rural Places and Incorporated Places. The counties were ordered in a contiguous manner and a random systematic draw was made from the cumulated list of Master Sample segments. The farms on these sampling units were subsampled at the rate of 1/2 and the information obtained at the first farm which fell into the subsample could be used in drawing the rest of the subsample. (The subsampling was systematic). The farm headquarters were numbered on the Farm Identification Sheet in a clockwise direction around the segment from the corner of the segment closest to where the interviewer's route intersects the segment boundary. All tracts within the sample segment were mapped. A small amount of information on all farms in the segment, not interviewed, for whatever reason, was obtained from those farm operators who were interviewed. The interviewer was to obtain at least one interview in each segment, provided the segment contained a farm headquarters of the subsample. If the interviewer did not accomplish this he was to make a return trip later, provided the round trip mileage from any point in his route was less than 30 miles. If there were other not-at-homes, in a segment, then he was to return to the segment after leaving it only if his route were within five miles of the segment.

The "Media Study" of 1949 used 200 sampling units in the Open Country selected in a systematic random manner. The sampling unit consisted of two contiguous master sample segments, one designated on the map as red, the other as green. Interviewers were instructed to drive around the red segment and sketch locations of all dwellings and tracts of land and
later to check this information with the first operator interviewed in the sample. A "fixed take" of three farm operators was required in each sampling unit. If there were exactly 3 operators in the red segment then all these were interviewed. If there were less than 3 operators in the red segment, then all there were in the red were interviewed and these were complemented by a subsample of operators in the green segment. In this case the interviewer started either at the north or east corner of the green segment and sketched in farms and tracts of land as he went. But he needed only go around the green segment until he had found enough farms to complete his "fixed take" of three operators. If there were more than three farms in the red segment then rules were applied for subsampling to obtain three farm operators for interview. From those selected a small amount of information was obtained for the other farms in the red segment. Landlords (if any) of the three farms chosen for interview were to be visited also. If the landlords lived in the segment or in the territory assigned to the interviewer then he was to interview them. At the end of the survey a team of interviewers went out and found as many as possible of the previously noninterviewed landlords that had fallen into the sample. Only part of this mileage appeared in the mileage records. Interviewers were to make at least one callback to get those farm non-interview households (except in the case of refusal) before taking a substitute farm in the green segment.

In the "Sources of Information" survey the interviewers kept mileage records on a county basis. Each time they entered or left a county, arrived at or left a s.u., at lunch, arrived at or left night's head-quarters, arrived at or left home, they recorded their odometer readings
in the space provided. In the "April Survey" and the "Media Study"
both mileage and time records were kept on a log basis, that is when-
ever an interviewer arrived at or left a s.u., arrived at or left lunch,
arrived at or left night's headquarters, arrived at or left home and
(in the case of the Media Study) arrived at or left a landlord's house,
he recorded the time and odometer reading and checked what event was
happening. From these records it was possible to obtain the mileage
and time spent for "Between S.U.", the mileage and time spent for "Within
S.U.", the mileage and time spent in going "Home", mileage and time spent
in going to and from and at "Lunch", the mileage and time spent in going
to "Night's Headquarters" and in the case of the "Media Study" the
mileage and time spent in finding landlord's outside the s.u. and the
time to interview the landlord.

It should be possible to present each of these components by call-
back. However, actually it is very difficult to say for example that
travel and time spent in going "Home" or to "Night's Headquarters", etc.
should be put under say Third Callback because the last s.u. visited
before these events had been a Third Callback. Undoubtedly some of the
time and mileage for these components are due to the fact that callbacks
were made. How much, will be estimated later on.

One should be able to obtain the time and mileage for each callback
for the "Within S.U." component. However some subjectiveness came in
even here. Often an interviewer began to work in a s.u. just before
lunch or before he quit for the day or before he went to a landlord's
house outside the s.u. ("Media"). After lunch, the next day or after

1 See Appendix B for the form used.
visiting a landlord’s house, as the case may be, he returned to finish his interviewing. This was not defined to be a callback. In these cases the mileage and time spent within an s.u. was all assigned to First Call. The mileage and time spent in going to and coming from lunch place or night’s headquarters or landlord’s house was assigned to the appropriate component (“Lunch”, “Night’s Headquarters” or “Landlord”) on First Call. On the other hand if the interviewer, after having spent considerable time in the s.u. left for lunch, for night’s headquarters or landlord’s house and returned to the same s.u. then all the mileage and time was assigned as callback mileage and time.

In the Media study “considerable time” was defined as three hours because the total time spent within a s.u. was long (average of 5.23 hours) and because some interviewers did their mapping and prelisting first and then returned to the s.u. to do their interviewing. In the other surveys this situation did not arise very often because the interviewers could visit all sample farms in a s.u. in a shorter time. Hence there was a tendency for interviewers to finish up a s.u. as far as possible before going to lunch or night’s headquarter. Thus somewhat arbitrary decisions were made from the records regarding the call to which the mileage and time was assigned. Notwithstanding these precautions the data in Tables IX – XVII show probably more callbacks than there really were and more mileage and time assigned to callbacks than was actually the case.

For the “Between S.U.” mileage and time it was necessary to set up certain rules so that one would be sure to obtain this component as
accurately as possible. If lunch, night's headquarters or landlord's house (Media Study) were on the direct route between s.u.'s then all the mileage and time spent in travelling was assigned to "Between S.U." component. If, however, the lunch place, night's headquarters or landlord's house were not on the direct route of the interviewer as he proceeded from one s.u. to the next, then, from the total mileage and time between the two s.u.'s the author subtracted the amount of mileage and time he estimated it would take to go from the direct route to the lunch place, night's headquarters or landlord's house. Or the author took 1/2 the total mileage and time between the two s.u.'s for the "Between S.U." component and the other 1/2 for the appropriate event that happened between the two s.u.'s.

If an interviewer left a s.u. to revisit another then his time and mileage for this trip would be callback mileage. If from this revisited segment he went to a third s.u., not previously visited, then this mileage and time would be "Between S.U." on first call. If, however, after revisiting this second s.u., he went back to the first s.u., then this time and mileage would go into Second Callback. It is easy to extend this rule to cover all practical situations in which one has to determine which callback the mileage and time "Between S.U." is to be assigned to. Even where there was an intervening event in going between s.u.'s one can combine the rule of the preceding paragraph with the rule in this one. The rule is that the mileage and time in going from s.u. y to make the k-th call on s.u. x is assigned to "Between S.U." on the k-th callback.
**TABLE IX**

Total Milesage: Sources of Information Survey

(Based on a stratified random sample of 351 sampling units, counties as strata)

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Home</td>
<td>4546</td>
<td>370</td>
<td>783</td>
<td>7551</td>
<td>1103</td>
<td>11353</td>
</tr>
<tr>
<td>Lunch</td>
<td>927</td>
<td>66</td>
<td>187</td>
<td>2420</td>
<td>358</td>
<td>3958</td>
</tr>
<tr>
<td>Within S.U.</td>
<td>154</td>
<td>6</td>
<td>60</td>
<td>491</td>
<td>64</td>
<td>775</td>
</tr>
<tr>
<td>Between S.U.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Night's H.Q.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>5627</td>
<td>442</td>
<td>1030</td>
<td>10462</td>
<td>1525</td>
<td>19086</td>
</tr>
</tbody>
</table>

78
<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>T</td>
<td>M</td>
<td>T</td>
<td>M</td>
<td>T</td>
<td>M</td>
</tr>
<tr>
<td>Home</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Call</td>
<td>198:55&lt;sup&gt;a&lt;/sup&gt;</td>
<td>6520</td>
<td>119:30</td>
<td>313</td>
<td>835:35</td>
<td>1837</td>
</tr>
<tr>
<td>Third Call or</td>
<td>6:20</td>
<td>117:1</td>
<td>15:1</td>
<td>_</td>
<td>23:10</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Time in hours and minutes
### TABLE XI

**Total Time and Mileage: Media Survey**

*(Based on a one stage sample of 200 Sampling Units)*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Home</td>
<td>Lunch</td>
<td>Within</td>
<td>Between</td>
<td>Night's</td>
<td>To or From</td>
<td>At</td>
<td>Total</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S.U.</td>
<td>S.U.</td>
<td>H.Q.</td>
<td>Landlords</td>
<td>Landlords</td>
<td></td>
</tr>
<tr>
<td>T M</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
<td>M T</td>
</tr>
<tr>
<td>-----</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>---------</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **First Call**
  - 306:2 T 10374
  - 14:12 783
  - 37:54 753:41 1604
  - 114:25 3023
  - 212:17

- **Second Call**
  - 36:45
  - 14:10 1018
  - 222
  - 232:38

- **Third or More Calls**
  - 2:30
  - 19:12 35
  - 4:15 46
  - 2:16 153
  - 3:34 5:35
  - 177
  - 108:13

- **Total**
  - 322:16
  - 11026
  - 160:53
  - 967
  - 1041:35
  - 2133
  - 187:19
  - 51:47
  - 80:01
  - 204:6
  - 114:25
  - 3023
  - 131:55
  - 2041:39
  - 24:32

*Time is in hours and minutes*
### TABLE XII

**Mileage per S.U.**

<table>
<thead>
<tr>
<th>Source of Information</th>
<th>First Call Between S.U.</th>
<th>First Call Within S.U.</th>
<th>Second Call Between S.U.</th>
<th>Second Call Within S.U.</th>
<th>Third Call or More Between S.U.</th>
<th>Third Call or More Within S.U.</th>
<th>Number of S.U.'s</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sources of Information</td>
<td>21.51</td>
<td>2.23</td>
<td>6.89</td>
<td>.53</td>
<td>1.40</td>
<td>.17</td>
<td>351</td>
</tr>
<tr>
<td>April Survey</td>
<td>17.59</td>
<td>4.30</td>
<td>1.28</td>
<td>.53</td>
<td>.39</td>
<td>.13</td>
<td>383</td>
</tr>
<tr>
<td>Media Outside</td>
<td>25.91</td>
<td>8.52</td>
<td>7.13</td>
<td>2.44</td>
<td>2.39</td>
<td>1.27</td>
<td>96</td>
</tr>
<tr>
<td>Media Estimate Used</td>
<td>25.91</td>
<td>8.02</td>
<td>7.13</td>
<td>1.88</td>
<td>2.39</td>
<td>.77</td>
<td>200</td>
</tr>
<tr>
<td>Media Inside</td>
<td>12.49</td>
<td>7.56</td>
<td>3.21</td>
<td>1.36</td>
<td>1.10</td>
<td>.30</td>
<td>104</td>
</tr>
</tbody>
</table>
### TABLE XIII

<table>
<thead>
<tr>
<th>Sources of Information</th>
<th>Home</th>
<th>Lunch</th>
<th>Night's Headquarters</th>
<th>Between S.U.</th>
<th>Within S.U.</th>
<th>To and From Number of Landlords</th>
<th>Total</th>
<th>Rate of Callbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>April Survey</td>
<td>17.45</td>
<td>.93</td>
<td>6.25</td>
<td>19.27</td>
<td>5.46</td>
<td>383</td>
<td></td>
<td>.21</td>
</tr>
<tr>
<td>Media Outside</td>
<td>46.46</td>
<td>5.50</td>
<td>11.55</td>
<td>35.42</td>
<td>12.23</td>
<td>19.27</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>Media Estimate Used</td>
<td>4.84</td>
<td>11.55</td>
<td>35.42</td>
<td>10.67</td>
<td>200</td>
<td></td>
<td></td>
<td>.70</td>
</tr>
<tr>
<td>Media Inside</td>
<td>63.13</td>
<td>4.21</td>
<td>9.01</td>
<td>16.80</td>
<td>9.22</td>
<td>11.64</td>
<td>104</td>
<td></td>
</tr>
<tr>
<td>Sources of Information</td>
<td>First Call</td>
<td>Second Call</td>
<td>Third Call or More</td>
<td>Number of S.U.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------------</td>
<td>------------</td>
<td>-------------</td>
<td>--------------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>April Survey</td>
<td>.80&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2.18</td>
<td>.06</td>
<td>.28</td>
<td>.02</td>
<td>.06</td>
<td>351</td>
<td></td>
</tr>
<tr>
<td>Media Outside</td>
<td>.98</td>
<td>3.41</td>
<td>.22</td>
<td>1.09</td>
<td>.10</td>
<td>.63</td>
<td>96</td>
<td></td>
</tr>
<tr>
<td>Media Estimated Used</td>
<td>3.77</td>
<td></td>
<td></td>
<td></td>
<td>.41</td>
<td></td>
<td>200</td>
<td></td>
</tr>
<tr>
<td>Media Inside</td>
<td>.48</td>
<td>4.10</td>
<td>.12</td>
<td>.97</td>
<td>.03</td>
<td>.21</td>
<td>104</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Time is in hours.
<table>
<thead>
<tr>
<th>Sources of Information</th>
<th>Home</th>
<th>Lunch</th>
<th>Night's Headquarters</th>
<th>Between S.U.</th>
<th>Within S.U.</th>
<th>At Landlords</th>
<th>To and from Landlords</th>
<th>Total No. of S.U.'s</th>
</tr>
</thead>
<tbody>
<tr>
<td>April Survey</td>
<td>.54(^a)</td>
<td>.34</td>
<td>.30</td>
<td>.83</td>
<td>2.52</td>
<td></td>
<td></td>
<td>383</td>
</tr>
<tr>
<td>Media Outside</td>
<td>1.35</td>
<td>.61</td>
<td>.47</td>
<td>1.26</td>
<td>5.14</td>
<td>.75</td>
<td>.71</td>
<td>96</td>
</tr>
<tr>
<td>Media Estimate Used</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.21</td>
<td></td>
<td>200</td>
</tr>
<tr>
<td>Media Inside</td>
<td>1.91</td>
<td>.80</td>
<td>.33</td>
<td>.63</td>
<td>5.27</td>
<td>.58</td>
<td>.47</td>
<td>104</td>
</tr>
</tbody>
</table>

\(^a\)Time is in Hours
<table>
<thead>
<tr>
<th>Sources of Information</th>
<th>Home</th>
<th>Lunch</th>
<th>Knight's Headquarters</th>
<th>Between S.U.</th>
<th>To and from Landlords</th>
</tr>
</thead>
<tbody>
<tr>
<td>April Survey</td>
<td>32.3</td>
<td>2.76</td>
<td>20.7</td>
<td>21.9</td>
<td></td>
</tr>
<tr>
<td>Media Outside</td>
<td>34.3</td>
<td>6.78</td>
<td>24.5</td>
<td>28.0</td>
<td>27.0</td>
</tr>
<tr>
<td>Media Inside</td>
<td>33.1</td>
<td>5.28</td>
<td>26.9</td>
<td>26.5</td>
<td>24.9</td>
</tr>
</tbody>
</table>
In the "Sources of Information" study the supervisor's mileage was included because in addition to his supervisory duties he interviewed in a certain region. Most of his supervisory mileage went into the "Home" component and this tended to inflate this total in comparison with the other surveys where the supervisor's time and mileage were not recorded. In the "April Survey" one interviewer, who worked in the state statistician's office, did not receive the instructions on, or forms for keeping mileage and time records that those who attended interviewer school in Ames did. Consequently the totals for the "April Survey" are based on 383 segments rather than 403 segments.

The total time and mileage for the "April Survey" and the "Media Study" and the total mileage for the "Sources of Information" are given in Tables IX - XVI where the breakdown is Callbacks x Components. In the "Media Study" no attempt was made to separate the time and mileage for landlords by call. Many of the landlords were interviewed afterwards by the supervisors if they were not obtained by the regular interviewers.

When these tables are put on a per sampling unit basis one can more clearly see the relations among these different surveys. In the "Media Study" there were two groups of interviewers, viz. (1) those who lived in Ames and came home only on weekends and (2) those who lived in the region in which they were interviewing. From those interviewers who lived outside the area in which they were interviewing (abbreviated to Media Outside) one can obtain estimates of the "Between S.U." and the "Night's Headquarters" components. But for those interviewers
living inside the region (shortened to Media Inside) the "Between S.U." travel is not comparable to the "Between S.U." travel of other surveys or the Media Outside interviewers. These interviewers probably left home in the morning, interviewed in one, two or three segments and then returned home. The next day they would start out in some other direction from home and pick up one, two or three s.u.'s, etc.

The mileage within a s.u. can be averaged over the Media Outside interviewers and the Media Inside interviewers. The Media Outside interviewers visited 96 sampling units while the Media Inside interviewers visited the remaining 104 s.u.'s. The mileage per s.u. for "Between S.U." and "Within S.U." components by First, Second and Third Calls are given in Table XII. The "Between S.U." on First Call for these three surveys will be compared with theoretical formula in Section B. In section D a least squares method for obtaining estimates of constants is shown. This can be used for predicting the average within s.u. mileage if the relative size of the s.u. is given.

In Table XIII are given the mileage per s.u. for "Home", "Lunch", "Night's Headquarters", "Between S.U." and "Within S.U." and "Landlord's House" when all calls are added together. Also given is the rate of callback for each survey. This is arrived at by counting the total number of times sampling units are entered for purposes of making second or third or more calls and dividing this number by the total number of sampling units in the survey. In the "Sources of Information" survey and the "April Survey" one or two interviewers generally made their home their night's headquarters. Their mileages were assigned to "Home" when appropriate rather than to Night's Headquarters. In the "Sources for
of Information" survey 15 s.u.'s and in the "April Survey" 65 s.u.'s were visited by interviewers who made home their night's headquarters. Hence for mileage per s.u. for Night's Headquarters the author divided the total for that component by 336 for the "Sources of Information" survey and by 316 for the "April Survey". The Media Inside interviewers did establish night's headquarters other than their home and hence there is a component for them but it is fairly difficult to ascertain what s.u.'s were covered from these night's headquarters. The mileage for "Home" depends on the number of times the interviewer comes home and the location of his home. This will be covered in more detail in Section C.

B. Between Sampling Unit Mileage

In the "Sources of Information" survey the interviewers in general made all necessary callbacks while in the county, although there was some travelling back and forth across county lines. It would seem that the path for "Between S.U." on First Call was approximately the s-path of Part III.G, where the square blocks are the counties and the rectangle is the state of Iowa. One may assume that there are 100 square counties in Iowa. The expected value of the s-path is $1.61 \sqrt{A_n} = 7392$ miles as compared with the actual "Between S.U." on First Call of 7551 miles. It is fairly difficult to compute what the sampling error of the s-path is. At least some of the difference may be attributed to sampling error.

1There are 98 counties of approximately the same size and one large county Pottawattamie can be conveniently divided into two nearly square parts.
In the "April Survey of Agriculture 1948", because of the low rate of callbacks, the interviewers' path should not resemble an s-path as closely as in the "Sources of Information" survey. In this case the expected length of the s-path is 5270 miles while the actual distance travelled "Between S.U." on First Call was 7092 miles.\(^1\) One may also regard the counties as strata and use the formulae of Part V.B where the interviewer moves from one county to the next without reversing a direction. If one adds a correction for the number of times the interviewer has to turn a corner to go from one tier of counties to another tier, the result is an expected distance of 6107 miles. Even in this case, where the rate of callback is fairly low, the formulae of Part V.B underestimates the actual mileage. Of course it is not always possible, nor practical, to follow a grid pattern of travel if a paved highway will take the interviewer close to his next sampling unit. The highway route may be longer.

In the "Media Study" the expected length of the s-path was 4937 miles. In this survey the "Between S.U." mileage records of only five interviewers could be used for comparison with other surveys. These five interviewers visited 96 out of 200 sampling units. The other nine interviewers lived in the region in which they were interviewing and consequently the "Between S.U." mileage is not the same for the two sets of interviewers. Hence the author assumed the "Between S.U." mileage on First Call was the same for the 10\(^1\) s.u.'s as for the 96 and he expanded on the basis of the 96 to estimate the mileage to be 5161 miles.

---

\(^1\)Information on mileage for five counties was missing but this is the expanded total assuming the "Between S.U." mileage was the same in these five counties as in the remaining 95.
200 s.u.'s.

It would seem from the above examples that when a fair rate of callbacks is made then the interviewers tend to follow the grid s-path and that the corresponding expected value provides a fairly good estimate of the mileage "Between S.U." on First Call. It is well to remember that the theory only covers a uniform space distribution over a region or a stratum. In Iowa the population of farm sampling units are fairly uniformly distributed over the whole state.

C. Home Mileage

In the three surveys for which mileage records are presented the "Home" component might be explained by the formulas of Part V.D. If one can assume the interviewers always went to a random point in the region in which they were interviewing and always returned home from a random point in the region in which they were interviewing, then the formulas would apply. We must also assume the frequency of home travel is the same whether the interviewers are working in a region near home or not.

In the "Sources of Information" survey there were only a few interviewers and hence the region they worked in was fairly large. Usually the interviewers could arrange to start interviewing at a s.u. nearer their home than the majority of s.u.'s assigned to them and then arrange their route so that, when they went home, the last s.u. before going home was also nearer home than the majority of s.u.'s assigned to them. Hence the theoretical formulas in this case would overestimate the
distance. Also interviewers who were working in a region fairly close to home came home often than those whose region was farther from home. In fact, there were two interviewers who were able to come home every night because their home was either in the region or close to the region in which they were interviewing. The total number of trips to and from home was 80 and the number of miles travelled for this component was 5627. One can assume that Iowa is a rectangle and that Ames or Des Moines are at the centre of this rectangle. In addition one may compute separately the number of miles for the interviewer who lived in Ames and interviewed in the surrounding district. Actually four interviewers lived in Des Moines and the other three in Ames. The formula for "Home" distance would give

\[
55\left(\frac{200 + 280}{4}\right) + 25\left(\frac{\log\left(\frac{1}{2}\right) + \log\left(\frac{5}{8}\right)}{2}\right) + \frac{4}{18}\left(\frac{40 + 12 + 36}{2}\right) + \frac{5}{18}\left(\frac{20 + 12 + 36}{2}\right)
\]

= 7522 miles

if we assume Iowa is a rectangle 280 by 200 miles. This overestimates the actual miles by about 2000 miles. Knowing that the assumptions of the formula are not met one would expect to overestimate the actual home mileage considerably.

In the "April Survey of Agriculture" five out of the nineteen interviewers lived in the region in which they were interviewing. For these five one can approximate their home mileage by use of formula (3) of V.D, taking into account the relative position of home in the region in which the interviewer worked. The accompanying map shows the information needed to prognosticate travel home. The prediction of home travel for these four interviewers is given below as
Figure 23. Location of Interviewers Who Lived in the Region in Which They Interviewed April Survey of Agriculture, 1948
In the experiment, only those intercepts were used in the regression on which there were intercepts.

If the intercepts of part A & B in 17.75 were

and home is 17.75 and the same expected length for those two sustain

It is interesting to note the actual mistake for "between S.U." home

more mistakes.

author estimated the between S. U. mistakes and substituted it in the

or between these two S. U.'s when the intercepts were not the same home in the course

successive whether or not the intercepts had gone home in the course

between S. U." mistake was obtained for every pair of S. U.'s presented in the

between the actual mistake in the survey a

hence the total expected distance is 759 whereas the actual distance

\[ H = 7590 \]

is to

covered these points. The expected distance to and from home or home

one can estimate the home mistake for the intercepts who

If now one assumes that the remaining S. U.'s are at random over the

\[
\left( \frac{11}{10} \times \frac{2}{x + (\frac{7}{1}) y} + \frac{11}{10} \times \frac{2}{y + (\frac{7}{1}) z} \right) g + \left( \frac{2}{(\frac{7}{1}) t + (\frac{7}{1}) s} \right)^2 + \\
\left( \frac{2}{x + (\frac{7}{1}) y} \right)^2 + \left( \frac{2}{y + (\frac{7}{1}) z} \right)^2 + \left( \frac{2}{s + (\frac{7}{1}) t} \right)^2
\]
"Between S.U." mileage was estimated for the case in which an interviewer left a s.u., went home and then came back to a different s.u. For these Media Inside interviewers the estimate of home mileage from the formulas should be fairly good. These Media Inside people made 149 trips to and from home. The expected distance for these interviewers was 6336 miles as compared with 6566 miles, the actual mileage travelled to and from home. For one interviewer, Carley, it was possible (by making a transformation of one county which did not affect the distance) to regard the region as semi-circular with the fixed point at the origin. In this instance the author employed the formula of Part V.C, transformed for grid distance. For the other interviewers the author employed the formulas of Part V.D.

In the case of Carley's mileage it was possible to estimate the "Between S.U." mileage from the formulas of Part V.D, where the semi-circular region contained 13 s.u.'s and altogether there were 13 trips between s.u.'s. The expected grid distance was 205 miles and the actual grid distance was 225 miles. The shape of the districts of the other interviewers were such that one could not test the "Between S.U." component on the basis of Part V.D. The best one could do with these other interviewers was to find the number of times they travelled between s.u.'s and consider the between s.u. as between two s.u.'s in the same county. There were on the average two s.u.'s per county. Hence the mileage for "Between S.U." is estimated as

\[ 70 \left[ \frac{2h}{3} + \frac{2h}{3} \right] = 1120. \]
Figure 24. Location of Homes of Interviewers and the Boundaries of the Interviewers' Regions, Media Survey.
The total "Between S.U." for the Media Inside then is estimated as 1325 whereas the actual mileage was 1747.

The home mileage for the Media Outside interviewers can be estimated separately for each one if the number of trips is known, the location of their homes with respect to the district, and the size and shape of the district in which they are interviewing. The expected distance in this case is 5057 whereas the actual distance travelled home was 4460. The districts in which these interviewers worked were fairly large and they usually arranged their travel so that the beginning s.u. and the ending s.u. (i.e., before interviewing and before going home) were located to make the "Home" mileage as short as possible. It may be noted that the number of trips to and from home was 49.

In all of these surveys there was some sort of geographic stratification either deliberate or in the systematic drawing of s.u.'s from a list ordered in a geographic manner. The effect of this on the expected distance from a fixed point to a random point is for all practical purposes nil. It is easy to demonstrate this. If one has a uniform distribution over a rectangle and he creates equal strata with one random point in each, then the average expected distance from any fixed point in the rectangle to these stratified points is the same as the expected distance from the fixed point to a random point anywhere in the rectangle.

The variation expected in our estimates of "Home" mileage has not been mentioned. This is because the variance, when averaged over many trips, is negligible as compared with biases due to the fact that the assumptions about our theory are met only partially, in practice. For
instance, if one has a hundred strata in a rectangle with a random point in each, and if the interviewer makes 100 trips (i.e. one to each strata) from a fixed point in the centre of the rectangle, then the standard error of the total "Home" mileage is approximately 100 miles. This is relatively small compared with the bias introduced because the assumptions of the theory are not met.

D. Least Squares Estimation of the Remaining Mileage

Now one can estimate constants for the components for which no theoretical formulas have been developed. These constants can, in turn, be used to predict these same components of mileage for a fairly wide variety of proposed designs of sampling a uniform space distribution. Include as part of the knowledge of the given design the total amount of time spent on the average within each segment and the rate of callback.

Until the present, most cost studies have consisted mainly of breaking down the costs of a particular survey. No attempt has been made to predict the cost of another survey from this breakdown, except implicitly, another survey of the same population of s.u.'s using the same design. It would seem to be very valuable if one could predict the cost of any particular design in advance and compare that cost with those for other feasible designs. One can obtain the mileages for "Between S.U." on First Call and "Home" from our theoretical formulas. But the other components may be approached in a different manner by least squares estimates. Suppose one has n observations on y and one is fitting k constants \( a_1, \ldots, a_k \), \( n < k \). Then there are n equations which are expressed
in matrix notation as \[ Y = Ua + \epsilon \]

where \[ Y \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad a \equiv \begin{pmatrix} a_1 \\ \vdots \\ a_k \end{pmatrix}, \quad \epsilon \equiv \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} \quad \text{and} \quad U \equiv \begin{pmatrix} u_{11} & \cdots & u_{1k} \\ \vdots & \ddots & \vdots \\ u_{n1} & \cdots & u_{nk} \end{pmatrix}. \]

\( U \) is the coefficient matrix of the \( a \)'s. The least squares estimate of \( a \) is

\[ \hat{a} = S^{-1}U^TY \]

where

\[ S = U^TU. \]

Now data is presented for only three surveys but they vary considerably in design. By fitting two constants only one degree of freedom for error remains and consequently one cannot say whether the constants explain most of the variation or not. One really needs to have the data of several more surveys, in which he has a uniform distribution, before he can establish whether the fit is good enough. The following linear hypotheses are given as tentative relations that might exist from a priori considerations. It may be the relations are not of the first degree in the constants, but of higher degree. This could be tested if more data were available. One could also test what constants are, or are not, necessary to explain the variation.

To explain "Night's Headquarter's" mileage use the following linear hypothesis

\[ y = \sqrt{\frac{A}{n}} \tau + Ts + \epsilon \]

where \( A \) is the average area of an Iowa county, \( n \) the average number of
points within a county and \( T \) the total time spent within a s.u. The reason for using \( \sqrt{\frac{A}{n}} \) is that there is usually a county seat near the center of the county, where an interviewer can spend the night. The more sampling units there are in a county the less distance, on the average, the interviewer will have to travel to his night's headquarters and it has been shown earlier that distance depends on the square root of the area and the \( \sqrt{n} \). The average time spent within a s.u. gives a measure of how often, on the average, the interviewer will go to night's headquarters. From these three equations

\[
\begin{align*}
13.12 & \quad r + 1.61 \quad s + e_1 = 4.54 \\
12.24 & \quad r + 2.52 \quad s + e_2 = 6.25 \\
17.38 & \quad r + 5.21 \quad s + e_3 = 11.55
\end{align*}
\]

one estimates \( r = 0.0966 \) and \( s = 1.914 \). For the "Sources of Information" survey no time records existed but one could find the total time spent in the field, estimate the total interviewing time, estimate the speed travelled outside the segments in m.p.h. and, using these, one could estimate the average time within a s.u. to be 1.61 hours.

For estimating the mileage "Within S.U." on First Call one can scale the estimates of the sizes of the s.u. relative to say a Master Sample size segment. In the "Sources of Information" survey the size was .5 of a m.s.s. For the "April Survey" it would be about the same size as a m.s.s. and for the "Media Study" in view of the instructions it would be 1.5. Hence the linear estimate is

\[
y = 5f + c.
\]
Note that $S$ is the measure of size of the s.u.

Hence

\[
\begin{align*}
.5f + \epsilon_1 &= 2.23 & \hat{y} \\
1.0f + \epsilon_2 &= 4.50 & 4.53 \\
1.8f + \epsilon_3 &= 8.02 & 8.16
\end{align*}
\]

and $f = 4.71$.

For the total mileage on all calls within a s.u. one may use the linear relationship

\[
y = Sf + Cg + \epsilon
\]

where $C$ is the rate of callback. This results in

\[
\begin{align*}
.5f + .40g + \epsilon_1 &= 2.93 & 3.11 \\
1.0f + .21g + \epsilon_2 &= 5.46 & 5.67 \\
1.8f + .70g + \epsilon_3 &= 10.67 & 10.50
\end{align*}
\]

which leads to $\hat{f} = 5.63$ and $\hat{g} = 1.05$.

In the above three cases the amount of variation removed by the constants is significant at the 5 percent level. However, because there are only one or two d.f. for error, this means the test is not too reliable.

For the total mileage per s.u. between s.u.'s one may set up the following linear hypothesis

\[
y = \hat{B} + Cc + \sqrt{\frac{A}{n}} d + \epsilon
\]
where \( B \) is the between s.u. estimate arrived at from the theory in Section B. Hence one has the following three equations

\[
\begin{align*}
\hat{y} &= 21.06 + 0.40c + 13.21d + \epsilon_1 = 29.81 \\
21.15 + 0.21c + 12.24d + \epsilon_2 = 19.27 \\
24.69 + 0.70c + 17.36d + \epsilon_3 = 35.42
\end{align*}
\]

from which one obtains \( \hat{c} = 11.311 \) and \( \hat{d} = 0.2004 \). This is not a particularly good fit. Therefore it suggests that perhaps other constants should be added to or some of the present constants dropped from the function or the degree of the function changed. The author assumed that the "Between S.U." variation on callbacks depended on the rate of callback and how far, on the average, the s.u.'s were apart. It might be that the rate of callback is too heterogeneous and should be split up into two parts, one for second callbacks and the other for third or more callbacks.

With the hypothesis concerning mileage for "Lunch" the author was not completely successful. The linear hypothesis used was

\[
y = Cg + T_h + \epsilon.
\]

However in the "April Survey" lunches were carried in the car more often than in the other surveys and hence there would be no mileage for this. One should really have a term for the number of times lunches were eaten in the car. Also in the "Media" survey, because of the long time spent within an s.u., the interviewer quite often left the segment for lunch and returned immediately afterward to continue his interviewing. This
mileage and time were assigned to "Lunch". In the other surveys, with a shorter time spent within a segment, the interviewers did not do this so often. Usually they had lunch while going between s.u.'s and this mileage was assigned to "Between S.U." rather than "Lunch" unless they went out of their way for lunch. Usually the lunch place was on the direct route between s.u.'s and, if not, only the amount of mileage they went out of their way was assigned to "Lunch". The equations used were

\[ y \]
\[ .40g + 1.45h + c_1 = 1.26 \]
\[ .21g + 2.18h + c_2 = .92 \]
\[ .70g + 3.77h + c_3 = 4.84 \]

and from these \( g = 4.42 \) and \( h = .2761 \). The coefficient of \( h \) is \( T_1 \), the time spent "Within S.U." on First Call. One probably should also have a term which contains \( \sqrt{A \over M} \) since the more s.u.'s there are, the less the distance one would expect to travel to a lunch place.

Tables XIV and XV contain time on a s.u. basis. It probably could be partially analysed by least squares in much the same manner as the mileage was. However there exists only two surveys for which there is a breakdown of time into these various components. It may be noted here that the Media Inside interviewers spent much less time travelling between s.u.'s and less time on going to night's headquarters than the Media Outside interviewers. On the other hand they spent more time going home.

However one can obtain an estimate of the time spent in various
activities if one has estimated the mileage for these activities.
Table XVI reveals the speed for each activity (except "Lunch") is fairly
constant and a rough guess could be made of the m.p.h. for each event.
This, when multiplied by the appropriate mileage for that event, would
give the time for that event. The miles per hour for lunch is not
constant because the "Lunch" component of time contains the actual time
spent at lunch as well as the time going to and from lunch. For the
"April Survey", since the interviewers carried their lunch a good deal
of the time, there was a time component for "Lunch" but no mileage and
this tended to depress the ratio of miles to hours.
PART VII. DETERMINING SAMPLE SIZE BY USE OF THE MINIMAX PRINCIPLE

As Cochran (6) has pointed out the Neyman criterion assumes that either cost or accuracy is fixed in advance. Actually one should determine optimum cost and optimum accuracy simultaneously.

Nordin (19), Blythe (4) and Yates (32) have given what Wald would call a Bayes solution to the problem. Their solutions differ because they assume different cost functions and/or different distributions of the error. The assumption that they make is that one knows at least one parameter of the distribution, usually the variance or standard error, and that it enters into the expected loss function.

However it is not necessary to know the parameter of the binomial distribution in order to reach a decision on the size of the sample to be taken from that population. Assume the statistician is playing a game with nature. The decision is made under the assumption that nature is trying to maximize the loss function and the statistician is trying to minimize the loss function.

For definiteness assume \( Y \) is a linear estimate of the population total \( (\nu_y) \) of those who answer "yes" to a certain question in a sample survey. Assume the loss resulting from the deviation of the estimate from the population total increases proportionate to the square of the deviation, i.e.

\[
l(s) = a^2
\]

where \( s = Y - \nu_y \), \( Y = \nu_y \) and \( y = 1 \) if "yes" is recorded on the schedule. Otherwise \( y = 0 \).
Assume the cost function is $C(n) = b \sqrt{n} + cn + d$. Then the expected loss is

$$L(n,p) = a \int_{-\infty}^{+\infty} s^2 f(s,n) ds = a \sqrt{\frac{a}{n}} = a \sqrt{\frac{a}{n}} \frac{p(1-p)}{n}$$

if one assumes no finite correction term. It is to be noted for this weight function that one does not need to know the distribution function of the errors.

Hence one wants to find the $\text{Min Max} = \text{Max Min}$

\[ \alpha \beta \gamma \delta \epsilon \zeta \eta \theta \iota \kappa \lambda \mu \nu \xi \omicron \pi \rho \sigma \tau \upsilon \phi \chi \psi \omega \]

\[ \alpha \beta \gamma \delta \epsilon \zeta \eta \theta \iota \kappa \lambda \mu \nu \xi \omicron \pi \rho \sigma \tau \upsilon \phi \chi \psi \omega \]

solution of

$$u = L(n,p) + G(n)$$

where $0 \leq p \leq 1, 0 \leq n \leq N$. Suppose $a = 1, b = 20, c = 2, d = 100$ and $N = 100$. Then the maximum with regard to $p$ of (1) is

$$\frac{10,000}{a} (1 - 2p) = 0 \text{ or } p = .5.$$

Putting this in (1) results in

$$\frac{10,000 (0.5)(0.5)}{a} + 20 \sqrt{n} + 2n + 100.$$

If one minimizes the above with respect to $n$ he has

$$-\frac{2500}{n^2} + \frac{10}{\sqrt{n}} + 2 = 0$$

or

$$n = 25.$$

Now $\frac{\partial u}{\partial p} = \frac{\partial n}{\partial n} = 0$ for $p = .5$ and $n = 25$. Also $\frac{\partial^2 u}{\partial p^2}$ is

negative, $\frac{\partial^2 u}{\partial n^2}$ is positive and $\frac{\partial^2 u}{\partial p \partial n} = 0$ for $p = .5$ and $n = 25$ hence
one has a saddle point or minimax at this point (see reference (21)).
In this case the worst that nature can do is to make \( p = .5 \) and under
the assumptions regarding the constants and the weight function the
statistician should take a sample of 25.

The question arises does one always have a minimax solution when
\( u \) is of the above form. It is obvious that \( p = .5 \) gives a maximum in
that direction for any value of \( n \). One has a minimum \( n \) for \( p = .5 \) if

\[
\frac{\partial u}{\partial n} = - \frac{a \, n^2 \, (.25)}{n^4} + \frac{b}{2(n^3)} + c = 0
\]

(2)

and

\[
\frac{\partial^2 u}{\partial n^2} = \frac{a \, n^3 \, (.5)}{n^4} - \frac{b}{4n^3} > 0
\]

(3)

That is, one has a minimax solution if the root of (2) is less than

\[
\left( \frac{2 \, a \, n^2}{b} \right)^{\frac{2}{3}}
\]

and of course less than \( N \). There will be only one positive real root of
(2). The others will be either negative or imaginary.

Take another example in which all the conditions of the previous
example are met and where the cost varies with \( p \). This might happen in
an interview in which a "yes" answer to the most important question would
entail further interviewing whereas a "no" answer would not. Specifically
let

\[
C(a, p) = b \, \sqrt{n} + \frac{c \, a}{1 - p} + d
\]
then
\[ u = an^2 \frac{p(1 - p)}{n} + b \sqrt{n} + \frac{an^m}{1 - p} + d. \]

If \( b = 0, \ c = 5, \ a = 1, \ N = 100 \) and \( d = 100 \) then a minimax solution exists for \( p \approx .60 \) and \( n \approx .21. \)

In general, suppose
\[ u = L(n, \theta) + o(n, \theta) \]

where \( u \) is a bounded function in \( n \) and \( \theta, \theta \) being a parameter of the distribution of the errors of estimate. There exists a minimax solution to the problem provided one can find roots of
\[ \frac{\partial u}{\partial \theta} = 0, \quad \frac{\partial u}{\partial n} = 0 \]

such that on substituting these roots
\[ \frac{\partial^2 u}{\partial \theta \partial n} < 0, \quad \frac{\partial^2 u}{\partial n^2} > 0 \text{ and } D = \frac{\partial^2 u}{\partial \theta \partial n} - \frac{\partial^2 u}{\partial \theta^2} \frac{\partial^2 u}{\partial n^2} > 0. \]

They will be such that
\[ \min_{n} \max_{\theta} \min_{\theta} = \max_{n} \min_{\theta} \max_{n}. \]

There might be more than one point within the range on \( n \) and \( p \) that satisfies the above relation. Then take the one where \( n \) is least. Of course there may be no saddle point for \( u \) over the range of \( n \) and \( p \) that is being considered. In that case one may build up a randomized decision function for the strategies of nature and the statistician. This in general would be fairly difficult to do but would give a minimax solution.

If randomized decision functions are admitted one may apply the
minimax principle to other populations in addition to the binomial population. A minimax solution (Wald (27)) would exist because \( n \) could be regarded as denumerably infinite even though nature's parameter is not bounded.
PART VIII. DISCUSSION AND SUMMARY

Assume that one is planning a survey over a rectangular area in which there is a fairly uniform space distribution of s.u.'s. Also assume that the interviewers all live in the center of the rectangle and go home equally often whether the region in which they are interviewing is close to home or not. Also assume that the s.u.'s are drawn in the random systematic manner that is usually followed. Hence one may regard some political unit or multiple thereof as a stratum provided these political units are all approximately square and of equal size. For convenience call these counties, although they may well be pseudo-counties, built up from a certain number of smaller political units (for example M.C.D's) so as to be of approximately equal size. These counties are of such a size that they contain on the average from one to six s.u.'s, i.e. $1 \leq k \leq 6$. Assume that the number of callbacks is fairly small so that the interviewer can move across the county picking up the s.u.'s as they occur and moving on into the next county. It will be assumed that a pattern of roads runs north and south and east and west at a distance of a mile apart. Hence one assumes any s.u. can be reached in a grid fashion. Also take $n$, the number of s.u.'s and $m$ the number of counties to be fairly large. Neglect the increment in mileage due to turning a corner in going from one county to the next and also neglect the fact that there will be no "Between S.U." mileage when the interviewer goes home and comes back to a different s.u. Under the above
assumptions a tentative travel function which gives the total mileage for a sample survey can be set up. It should be fairly accurate within a particular range of certain constants. It is difficult to say what that range is for these constants, but it would be at least as great as the range for the three surveys described previously. The travel function for a one stage sample is

\[ y = H \left[ \frac{a + b}{4} \right] + \frac{n + \frac{3m}{3}}{\sqrt{n}} \sqrt{A} + \left[ \frac{.0966}{\sqrt{A}} n + 1.914 n T \right] + \left[ n(5.63 S + 1.080) \right] + \left[ 11.31 n C + .2004 \sqrt{A} n \right] + \left[ n(4.42 C + .2761 T_1) \right] \]

This can be simplified to

\[ y = H \left[ \frac{a + b}{4} \right] + \frac{n + \frac{3m}{3}}{\sqrt{n}} \sqrt{A} + .2970 \sqrt{A} n + n(1.914 T + .2761 T_1) + 5.63 S + 16,810 \]

where

- \( H \) - number of trips to and from home
- \( A \) - area of the rectangle
- \( a, b \) - sides of the rectangle, \( ab = A \)
- \( n \) - number of s.u.'s
- \( m \) - number of counties
- \( k \) - number of s.u.'s within a county
- \( T \) - total time spent within a s.u.
- \( T_1 \) - time spent on first call to an s.u.
- \( S \) - size of s.u. using the m.s.s. as base = 1
- \( C \) - number of callbacks as a ratio to total number of s.u.'s.
One may also estimate the total time roughly from (1) and Table XVI. That is one can guess the m.p.h. for the various activities and divide this into the total number of miles in the corresponding activities as given by certain terms of (1).

It is to be noted that \( k = 3 \) would make (1) a minimum in the same way as this value for \( k \) made equation (4) of Part V.B a minimum. Of course, we may put in other values for \( k \) in equation (1) and find the total distance travelled. It may be from the point of view of minimizing sampling error that \( k = 3 \) is not desirable.

One may have more alternatives than merely varying the number of p.s.u.'s per stratum. The statistician may want to consider the possibilities of quite different designs. Also within these distance functions he can vary the sampling rates.

For comparison with the previous travel function suppose the following conditions hold: (1) It is feasible to take a two stage sample in which the p.s.u.'s are of approximately equal area and are chosen with approximately equal probability. (This would have to be the case under the assumption of a nearly uniform space distribution over the rectangle). (2) The p.s.u.'s are approximately circular with the night's headquarters at the center of the p.s.u. (The p.s.u.'s may not be circular at all but may be approximately square, for example, see Table VIII). (3) The interviewers proceed from s.u. to s.u. visiting them in a counter-clockwise order (say). (4) The number of days (4) in the field in each p.s.u. is known. (5) The number of trips (N) the
interviewers will make between p.s.u.'s is also known. Assume that each interviewer travels between p.s.u.'s in such a way that the p.s.u.'s are "adjacent" in a counter-clockwise direction (say). (6) There is no stratification of the rectangle. (7) All the other assumptions of the previous travel function (which are not in conflict with the above) are met. Then the travel function for this case would be

\[
y = N \left[ \frac{a + b}{n} \right] + N \sqrt{\frac{A}{w}} \left[ 1 + \frac{g}{2} \sum_{j=1}^{\text{even}} \frac{2}{2} \left( \frac{p-1}{2} \right)^{j} \left( -1 \right)^{j} \frac{p^j}{(2^n)^j (n-2j)^j!} \right]
\]

\[
+ p(k - 1) \sqrt{\frac{A}{w}} \left[ 1 + \frac{g}{2} \sum_{j=1}^{\text{odd}} \frac{2}{2} \left( \frac{p-1}{2} \right)^{j} \left( -1 \right)^{j} \frac{k^j}{(2^n)^j (n-2j)^j!} \right] + \frac{c}{3^n}
\]

\[
+ \left[ n(5.63 s + 1.08 c) \right] + \left[ n(11.31 c + .2004 \sqrt{\frac{A}{w}}) \right]
\]

\[
+ \left[ n(4.42 c + .2761 t) \right]
\]

where

\[
k - \text{number of s.u.'s chosen per p.s.u.}
\]

\[
p - \text{number of p.s.u.'s chosen}
\]

\[
u - \text{number of p.s.u.'s in the rectangle.}
\]

In the above formula it was assumed that the conversion factor for
changing the square root of mean square distance to airline distance
times the conversion factor for changing airline distance to grid
distance was equal to one.

Under different assumptions functions similar to equations (1) and
(3) could be constructed. Hence within certain fairly broad limits one
can change the design of any proposed survey and estimate for each change
how the field costs are changed. Presumably one knows something of the
changes in sampling errors created by these changes in designs (see for
example references (1), (5), (10), (11), (13) and (22)). It is then
possible to strike an optimum considering both costs and variances as in
references (11) and (31). If one cannot do this in a formal manner he
at least can do some empirical work and use intuitive judgments, for
example, on whether the extra cost is justified in order to decrease
sampling error a certain amount. A formal method of making this decision
is presented in Part VII for a population in which one does not know the
parameters that enter the loss function. However the solution is given
under fairly restrictive conditions. In general, it would be necessary
to construct a randomized decision function in order to find the minimax
solution to the problem.

In the cost function presented in this part the effect of the design
of the survey on costs of operations other than field costs has been
ignored. These other variable costs should be incorporated in a cost
function, but in general they are not as important, nor as difficult to
estimate, as field costs. The fixed costs (i.e., costs independent of
design) do not enter into the design of surveys but of course should be
included in any contractual agreement.
A very general cost function for an interview survey would include:

Overhead, i.e., Cost of Negotiations, Research, etc. \( K_0 \)

Questionnaire Design, Pretest and Instructions \( K_1(s_1,s_2) \)

Drawing of Sample \( K_2(d) \)

Printing of Questionnaires \( K_3(d) \)

Cost of Materials \( K_4(d) \)

Field Costs \( K_5(d,s_1) \)

Coding, Punching and Editing \( K_6(d,s_1,s_2) \)

Tables \( K_7(s_1,s_2) \)

Error Tables \( K_8(d) \)

where

\( s_1 \) - length of questionnaire

\( s_2 \) - type of questions

\( d \) - the design of the survey.

As far as achieving the optimum design of survey one can consider

\( K_0 + K_1 + K_7 = k \) as a fixed cost. For a one stage sample, with counties as strata, consider one has \( h \) households within each s.u., \( n \) s.u.'s and \( m \) counties. Take \( K_2 = k_3n \), \( K_3 = k_3nh \), \( K_4 = k_4m \). Also \( K_7 \) will be given by the methods described earlier in this section. Assume \( K_6(s_1,s_2) = nh k_6(s_1,s_2) \). Then in this case

Total Cost = \( k + k_3n + k_3nh + k_4m + k_5y + wy' + k_6nh + K_8(d) \)

where

\( k_5 \) - the rate of pay per mile,

\( w \) - the rate of pay per hour in the field plus the rate of expenses per hour, other than mileage expenses,
$y$ - the estimate of total number of miles,

$y'$ - the estimate of the number of hours in the field.

The total cost function would be more complicated if one considered a multistage sample. However in a particular survey it could be estimated fairly accurately from the theory and the data presented here, if there is a uniform space distribution of the elements over each stratum.
PART II. CONCLUSIONS

The problem of estimating the total travel for any sample design where the sampling units may have any space distribution has not been solved completely. However some very useful formulae for a uniform space distribution have been developed which can be used for a wide variety of designs.

1. The expected grid distance between two points in a rectangle $ab$ is $\frac{a + b}{2}$ and between two points in an ellipse with semiaxes $a$ and $b$ is $\frac{256}{\pi} (a + b)$. It is shown then, that for equal area regions moderate changes in shape do not influence distance greatly. It is easy to show that the expected coefficient of variation for grid distance is 50 percent for a square and that it is always less than 71 percent for any shape of rectangle. In fact as $\rho = \frac{b}{a}$ decreases in the interval from one to zero the expected coefficient of variation increases monotonically from 50 percent to 71 percent.

2. The expected minimum grid distance for $n = 3$ in a square of side $a$ is approximately 1.121 $a$.

3. An upper bound to minimum grid distance for $n$ points in a convex region (and some non-convex regions) of area $A$ is $1.48 \sqrt{A} n$ where $n$ is large.

4. The ratio of expected grid distance to expected airline distance seems to be fairly constant over changes in $n$ and changes
slightly with changes in shape of the region being sampled. For rectangles of side ratio 1 : 1, 2 : 1 and 4 : 1 the expected value of the ratio of grid distance to airline distance is 1.27, 1.25 and 1.21 respectively, if we take the points in the order of their occurrence from left to right. It is to be noted that the ratio of expected grid distance to expected airline distance is almost the same as the expected value of the ratio of grid distance to airline distance for most common shapes of regions. For a circular region, \( n = 2 \), the ratio of the expected values equals the expected value of the ratio.

5. The expected grid distance among \( n \) points in a rectangle, if one takes the points in the order in which they occur from left to right, is \((n - 1) \left( \frac{a}{a + 1} + \frac{b}{3} \right)\) where \( a \) is the length of the rectangle in the direction left to right.

6. If the number of strata, \( m \), is large in a rectangular region of area \( A \) then the expected grid distance proceeding always in one direction is \( \frac{n + 3m}{3\sqrt{m}} \sqrt{A} \). Under the above assumptions the expected distance is a minimum when we have three points per stratum. For this case the total distance is \( 1.155 \sqrt{A} n \).

7. For a certain shape of rectangle and all \( n \) it can be proven that the expected distance for a stratified sample is greater than the expected distance for a random sample if one follows in both cases a path that seems reasonable.

8. The expected distance for a stratified sample will exceed an upper limit of the minimum grid distance of a random sample of
the same size over the same area if \( k > 13 \), where \( k \) is number of points per stratum.

9. The expected distance for a stratified sample is increased by
\[
\frac{k - 1}{k + 1} \sqrt{\frac{k}{2}} \sqrt{\frac{A}{a}}
\]
each time a corner is turned.

10. It is only necessary that over each stratum there be a uniform density function in order to apply and extend the theory of the expected distance. That is, the densities may vary from stratum to stratum.

11. The expected airline distance of a random point from the center of the arc which forms a circle, semicircle or quadrant is \( \frac{2}{3} \) the radius.

12. The expected grid distance of a random point from the center of the arc which forms a circle, semicircle or quadrant is \( \frac{5}{3w} \) the radius.

13. The expected grid distance of a random point from the point \((ka, jb)\), where \( 0 \leq k, j \leq 1 \) and the rectangle \( ab \) has one vertex at the origin, is
\[
\frac{a(1 - 2k + 2k^2) + b(1 - 2j + 2j^2)}{2}
\]

14. Suppose the fixed point is the origin and the rectangle is entirely in one quadrant with respect to the axes. The coordinates of the vertices nearest and farthest from the origin are \((c,d)\) and \((e,f)\). Then the expected distance of a random point in the rectangle from the fixed point is
\[
\frac{c + d + e + f}{2}
\]
15. The mean square distance in a circle between any pair of random points adjacent in either clockwise or counter-clockwise direction is

\[
a^2 \left[ \begin{array}{cc}
even & odd \\
\frac{n-2}{2} & \frac{n-1}{2} \\
\end{array} \right] \frac{(-1)^{j-1}}{(2n)^2j(n-2j)!}. 
\]

16. The mean square distance in a semicircle between any pair of random points adjacent in either clockwise or counter-clockwise direction

\[
a^2 \left[ \begin{array}{cc}
even & odd \\
\frac{n-2}{2} & \frac{n-1}{2} \\
\end{array} \right] \frac{(-1)^{j-1}}{n^2j(n-2j)!} + \frac{2n!}{n^2} (-1)^{\frac{n-2}{2}).
\]

The last term in ( ) brackets is added only when \( n \) is even.

17. The mean square distance in a quadrant between any pair of random points adjacent in either clockwise or counter-clockwise direction is

\[
a^2 \left[ \begin{array}{cc}
even & odd \\
\frac{n}{2} & \frac{n-1}{2} \\
\end{array} \right] (-1)^{j-1} \frac{n!}{(\frac{n}{2})^2j(n-2j)!} + \frac{n!}{\frac{n}{2}^n} (-1)^{\frac{n-1}{2}}.
\]

The second term in ( ) brackets exists only if \( n \) is odd.

18. The author is fairly confident that the square root of the above mean square distances are from 6.6 percent to 9.5 percent greater than the corresponding mean airline distances.
19. In general, to use Wald's minimax principle in deciding on
size of sample one would have to find the appropriate random-
ized decision functions for both nature and the statistician.
PART X. LITERATURE CITED


PART XI. ACKNOWLEDGMENT

The author wishes to thank Dr. R. J. Jessen for suggesting this problem and guiding the course of the work.

In addition Dr. A. M. Mood, Dr. G. W. Brown, Dr. G. R. Seth and Dr. T. A. Bancroft have been very helpful at various stages of the development. The author is also grateful for the interest of Dr. H. F. Thielman and Dr. J. A. Nordin in the study.

The investigation was financed by the Iowa Agricultural Experiment Station, Project 1005, and the Bureau of Agricultural Economics' Bankhead-Jones project SRF 2-63 entitled "Research in the statistics of agriculture and the associated statistical theory".
PART XII. APPENDICES
\[ \frac{r}{I+e^{r/2} + 1} \] 

Also, for \( n \) large,

- If \( n \) is positive at \( n = 1 \), negative at \( n = 2 \), and \( n \) approaches \( n \) on \( n \), then \( n \) is between 1 and 2. Hence two roots are imaginary and one is real.

Hence \( \lambda \) is a root of \( \zeta = 0 \) in \( \lambda \). The derivative of \( \lambda \) at the critical point is

\[ \frac{\lambda}{\lambda + 1} \]

Now \( \lambda = 0 \) and \( \lambda \) can

study \( \lambda \) between 1 and 1. Hence one can

consider \( \lambda \) for \( \lambda > 1 \) and \( \lambda > 0 \) when \( \lambda > 0 \) then \( \lambda = 0 \) and \( \lambda = 1 \). Let \( \lambda > 1 \). \( \lambda \) is

\[ \lambda = \frac{\lambda}{\lambda + 1} \]

and

\[ \lambda = \frac{\lambda}{\lambda + 1} \]

Where

\[ \lambda = \frac{\lambda}{\lambda + 1} \]

From equation (5) in part A:

- Investigation of the shape of \( \lambda \) for \( \lambda > 0 \)

- Part XII. Appendices
One can see that $f''(0) = -\infty$, $f''(1)$ is positive and $f''(k)$ approaches zero positively as $k$ gets large. Hence it is known that $f''(k)$ has a maximum at $k_1$ and that for some value $k_2$, $0 \leq k_2 \leq 1$, $f''(k_2) = 0$, since $f''(k)$ is continuous.

Furthermore

$$f'(k) = \frac{k-2k^3-3}{\sqrt{k^2+1}} + 2k - 2 + (2+k) \log \frac{1+\sqrt{k^2+1}}{k}. \quad (6)$$

Hence $f'(0) = +\infty$. Furthermore $f'(1)$ is negative and $f'(k)$ approaches zero negatively as $k \to \infty$. Hence $f'(k)$ has a minimum at $k_3$ and a point of inflection at $k_1$. Furthermore at some point $k_3$, $0 \leq k_3 < k_2$, $f'(k)$ has a root, since $f'(k)$ is continuous. Also $f'(k)$ is monotonically increasing between $k = 1$ and $k = \infty$.

![Figure 25.](image-url)
Now from equation (3) $f(0) = 1/2$, $f(1) \approx 0.6369$ and $f(k)$ approaches $1/2$ as $k$ approaches infinity. Hence $f(k)$ has a maximum at $k_3$ and a point of inflection at $k_4$. But for $k$ between $1$ and $\infty$ $f(k)$ is monotonically decreasing. Hence $\phi(k)$ is monotonically increasing from $k = 0$ to $k = 1$.

Now the author is interested in showing that $E(\gamma) = f(k) + \phi(k)$, $0 \leq k \leq 1$ is monotonically increasing. We know that $\phi(k)$ is increasing over the whole interval and that $f(k)$ is increasing in the interval $0$ to $k_3$. If $E(\gamma)$ is increasing monotonically from $k_3$ to $k_4$ then it must be increasing from $k_4$ to $1$ because $f'(k)$ is negative and increasing from $k_4$ to $1$ while it is negative and decreasing from $k_3$ to $k_4$. At the same time $\phi(k)$ is positive and increasing over the whole range $k_3$ to $1$.

Now $f'(1) + \phi'(1) = 0$ and also one can show that

$$f(0.95) + \phi(0.95) < f(1) + \phi(1).$$

Furthermore $f''(0.95)$ is positive hence $k_4$ must be to the left of $k = 0.95$.

Now it may be easily shown that the maximum of $f(k)$ is to right of $k = 1/2$, so that $k_3$ is on the right of $1/2$. Hence one needs only consider $E(\gamma)$ between $k = 1/2$ and $k = 0.95$ to show monotony.

One wants to show between $k_3$ and $k_4$ that

$$f'(k) + \phi'(k) \geq 0.$$  

Since $f(\frac{1}{2}) = \phi(k)$ then $\phi'(k) = -\frac{1}{k^2} f'(k)$.

Hence one has to show

$$f'(k) \geq \frac{1}{k^2} f'(\frac{1}{2}).$$

Further investigations are necessary before it can be shown that $E(\gamma)$ is monotonically increasing over the range $0 \leq k \leq 1$. 
E. Forms Used

Form I is a photostatic copy of an Average Travel Path sheet used in the empirical investigations of the minimum distance in a rectangle among a randomly chosen points.

Form II is a photostatic copy of a Daily Mileage and Time Record sheet. This particular sheet was the one that was used in the interviewer school ("Media Study") for illustrating the method for making entries on the blank Daily Mileage and Time Record sheets.
**Sample First Second**

<table>
<thead>
<tr>
<th>Grid</th>
<th>8.0</th>
<th>7.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Airline</td>
<td>6.0</td>
<td>5.4</td>
</tr>
</tbody>
</table>

**Form I. An Average Travel Path Sheet**
<table>
<thead>
<tr>
<th>Time</th>
<th>AM/PM</th>
<th>Quarter Reading</th>
<th>Segment Number</th>
<th>Lunch Hour Location</th>
<th>Lunch Hour Place</th>
<th>Landlord Location</th>
<th>Landlord Place</th>
</tr>
</thead>
<tbody>
<tr>
<td>8:15</td>
<td>X</td>
<td>153</td>
<td></td>
<td></td>
<td></td>
<td>Malvern</td>
<td>X</td>
</tr>
<tr>
<td>8:45</td>
<td>X</td>
<td>166</td>
<td>126 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:25</td>
<td>X</td>
<td>172</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>10:40</td>
<td>X</td>
<td>181</td>
<td>127 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11:40</td>
<td>X</td>
<td>195</td>
<td>127 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:10</td>
<td>X</td>
<td>203</td>
<td>Silver Springs X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1:10</td>
<td>X</td>
<td>203</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>1:30</td>
<td>X</td>
<td>201</td>
<td>127 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2:05</td>
<td>X</td>
<td>203</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>2:25</td>
<td>X</td>
<td>209</td>
<td>126 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4:05</td>
<td>X</td>
<td>225</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>4:30</td>
<td>X</td>
<td>225</td>
<td>126 X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5:00</td>
<td>X</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:10</td>
<td>X</td>
<td>239</td>
<td>Ames X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6:10</td>
<td>X</td>
<td>239</td>
<td>*</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>6:30</td>
<td>X</td>
<td>251</td>
<td>Ames X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Households to be Interviewed

<table>
<thead>
<tr>
<th>Segment Number</th>
<th>in the Segment</th>
<th>Outside the Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>126</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>127</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>128</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Form II. A Hypothetical Mileage and Time Record Sheet