Some theoretical aspects of urban structure

Lawrence M. Ostresh
Southern Illinois University Edwardsville

Follow this and additional works at: https://spark.siue.edu/etd

Recommended Citation
https://spark.siue.edu/etd/65

This Thesis is brought to you for free and open access by the Graduate School at SPARK. It has been accepted for inclusion in Theses, Dissertations, and Culminating Projects by an authorized administrator of SPARK. For more information, please contact magrase@siue.edu,tdvorak@siue.edu.
ACKNOWLEDGMENT

The author wishes to thank the members of his committee, Dr. Harry Kircher, Dr. Carl Lossau, and Mr. Richard Guffy, Chairman, for their kind help and thoughtful suggestions in the writing of this thesis. The author is further indebted to Mr. Samuel J. Meltz, Mr. John Weever, and the late Mr. Richard Piepenberg for a series of stimulating discussions during the spring and summer of 1968 which led to the birth of the ideas herein contained. Finally, the author wishes to thank his beloved wife, Leslie, for long months of paper-strewn rooms, and for many midnight hours spent in typing copy.
DEDICATION

For Mr. Richard Guffy, who teaches one to think.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Background</td>
<td></td>
</tr>
<tr>
<td>Statement of Problem</td>
<td></td>
</tr>
<tr>
<td>II. Methodology</td>
<td>5</td>
</tr>
<tr>
<td>Selection of Cities</td>
<td></td>
</tr>
<tr>
<td>Data Collection</td>
<td></td>
</tr>
<tr>
<td>Estimation of Parameters</td>
<td></td>
</tr>
<tr>
<td>Tests of Significance</td>
<td></td>
</tr>
<tr>
<td>Limitations</td>
<td></td>
</tr>
<tr>
<td>III. Theory</td>
<td>30</td>
</tr>
<tr>
<td>Notation</td>
<td></td>
</tr>
<tr>
<td>Characteristics of the Lognormal Distribution</td>
<td></td>
</tr>
<tr>
<td>The Relationship Between Population Distribution and Population Density</td>
<td></td>
</tr>
<tr>
<td>A Deductive Explanation of the Lognormality of Urban Population Distribution</td>
<td></td>
</tr>
<tr>
<td>Implications of the Postulates</td>
<td></td>
</tr>
<tr>
<td>IV. Conclusion</td>
<td>62</td>
</tr>
<tr>
<td>V. Bibliography</td>
<td>64</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Cities Meeting Minimum Requirements for Testing Lognormal Hypothesis.</td>
<td>7</td>
</tr>
<tr>
<td>2. Estimates of Parameters of Cities Studied.</td>
<td>11</td>
</tr>
<tr>
<td>3. Statistical Test Results.</td>
<td>24</td>
</tr>
<tr>
<td>4. Regression Analysis of SMSA and Central City Population.</td>
<td>57</td>
</tr>
<tr>
<td>5. Regression of ( \mu ) and ( \sigma^2 ) on the Logarithm of Population.</td>
<td>58</td>
</tr>
<tr>
<td>Figure</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1. The Lognormal Distribution, $P = 1, \mu = 0, \sigma = 0.5$</td>
<td>4</td>
</tr>
<tr>
<td>2. Enumeration Sections, South Bend, Indiana</td>
<td>9</td>
</tr>
<tr>
<td>3. Cumulative Frequency Graph: Fort Wayne, Indiana</td>
<td>12</td>
</tr>
<tr>
<td>4. Cumulative Frequency Graph: Grand Rapids, Mich</td>
<td>13</td>
</tr>
<tr>
<td>5. Cumulative Frequency Graph: Lubbock, Texas</td>
<td>14</td>
</tr>
<tr>
<td>6. Cumulative Frequency Graph: Nashville, Tenn</td>
<td>15</td>
</tr>
<tr>
<td>7. Cumulative Frequency Graph: New Orleans, La</td>
<td>16</td>
</tr>
<tr>
<td>8. Cumulative Frequency Graph: Phoenix, Ariz</td>
<td>17</td>
</tr>
<tr>
<td>9. Cumulative Frequency Graph: Rochester, N. Y</td>
<td>18</td>
</tr>
<tr>
<td>10. Cumulative Frequency Graph: Salt Lake City, Utah</td>
<td>19</td>
</tr>
<tr>
<td>11. Cumulative Frequency Graph: San Antonio, Tex</td>
<td>20</td>
</tr>
<tr>
<td>12. Cumulative Frequency Graph: South Bend, Ind</td>
<td>21</td>
</tr>
<tr>
<td>13. Cumulative Frequency Graph: Salt Lake City, Utah, Northern Sector</td>
<td>22</td>
</tr>
<tr>
<td>14. Probability Triangle</td>
<td>40</td>
</tr>
<tr>
<td>15. Cumulative Frequency Graph Showing Convergence of a Theoretical to a Normal Distribution</td>
<td>44</td>
</tr>
<tr>
<td>16. Cumulative Frequency Graph of Density, Population, and Area</td>
<td>51</td>
</tr>
<tr>
<td>17. Cumulative Frequency Graph of Density, Population, and Area, Assuming Truncation at 44th Percentile</td>
<td>53</td>
</tr>
<tr>
<td>18. Scattergram of Size of SMSA and Size of Central City</td>
<td>56</td>
</tr>
<tr>
<td>19. Scattergram of Regression of $\mu$ and $\sigma^2$ on the Logarithm of Population</td>
<td>60</td>
</tr>
</tbody>
</table>
INTRODUCTION

Background

The fundamental urban aspect is people: A city's size is the number of its inhabitants, not its land area. People are distributed unevenly throughout the city, both during the day (the worker-shopper distribution) and during the night (the resident distribution). The distribution of the resident population long ago attracted the attention of Mark Jefferson. In his landmark urban study, "The Anthropography of Some Great Cities," published in 1909,¹ he shows that for the cities studied, population density is low in the center of the city, reaches its peak a short distance outward, and then tapers off gradually as distance increases. Bogue's study of some sixty-seven of the largest United States cities shows that this decline in density continues for many miles beyond the city limits (up to 300 miles in some cases).²

An important advance in urban analysis was made by Colin Clark in 1951. In his paper Urban Population Densities³ he gives evidence that

¹Bulletin of the American Geographical Society, XLI, 537-566.
³Journal of the Royal Statistical Society, Series A, CXIV, 490-496.
the decline in density with increasing distance follows a simple negative exponential rule. Clark writes:

Let $X$ be the distance in miles from the center of the city. Let $Y$ be the density of resident population in thousands per square mile. Then (except in the central business zone) —

$$Y = A e^{-bX}$$

That the falling off density is an exponential function, as in the above equation, appears to be true for all times and all places studied, from 1801 to the present day, and from Los Angeles to Budapest.  

In the equation, $A$ is the central density, and $b$ is the coefficient of rate of decline. Both are constant for any city, but vary between cities.

At the center of the city in the formula, $X$ is equal to zero and $Y$, therefore becomes equal to $A$. It is a hypothetical rather than an actual figure, because in fact the center of the city is occupied by the business zone with few or no resident inhabitants. Nevertheless it remains a useful figure; it shows the point to which densities are tending, if we measure the densities of the inner residential suburbs and continue extrapolating them inwards to reach the center of the city.

In his review of Clark's work, Stewart states

Mr. Clark perhaps overstates the deviation found when relatively few people live in the central business district. The fact is that the high density center of population is not the business district, except in cities smaller than he investigated.

---

4Ibid., pp. 490-491.

5Ibid., p. 491.

Whether a shift of origin would improve the validity of Clark's model is unknown. To the author's knowledge, no empirical studies have tested this hypothesis.

An alternate means of accounting for the low central density is given by Berry, Simmons, and Tennant. After reviewing studies relating to Clark's model, and noting the contributions of Alonzo and Muth in providing it with a theoretical rationale, they hypothesize that the model should hold better for net density than gross density. This appears to be true, at least for Chicago, the only city on which they report.

In this paper the distribution of the resident urban population with respect to distance from the city center is studied and shown to approximate lognormality. A probabilistic generation model is then hypothesized to account for the distribution and some of the implications of the model are examined. One of the implications is that gross density varies lognormally with distance, in conflict with Clark's earlier formulation. If Clark's model is restricted to net densities, however, the conflict disappears. Indeed, the two models in conjunction may have important implications about residential land use distributions.

**Statement of Problem**

**HYPOTHESIS:** The resident urban population is distributed lognormally with respect to distance from the city center.

---

That is:

\[ dP = \frac{P}{\sqrt{2\pi \sigma r}} e^{-\frac{1}{2\sigma^2} (\ln r - \mu)^2} \quad r > 0 \]

where \( dP \) is the number of persons in a ring of radius \( r \) and width \( dr \), \( P \) is the total population of the city, and \( \mu \) and \( \sigma \) are parameters of location and dispersion, respectively. A graph of (1.1) is given in Figure 1. With a change in scale, the same graph can be used to show the variation of gross density with distance. As can be seen, the curve is unimodal and positively skewed, which is in agreement with Jefferson's findings, and with more recent research.

Fig. 1.—The lognormal distribution, \( P = 1 \), \( \mu = 0 \), \( \sigma = 0.5 \).
II

METHODOLOGY

In order to test the lognormal hypothesis, it is necessary to gather data on the distribution of population for selected cities, estimate the parameters \( \mu \) and \( \sigma \), and compare the expected and the actual distributions by means of a statistical test of significance. Each of these points is discussed in turn.

Selection of Cities

As originally conceived, the lognormal hypothesis would have applied exactly to a city developed on an unbounded, featureless plain, with uniform access to the center from all directions. Such a city would be circular in shape with lines of equal population density forming concentric circles around the center of the city, which contains all non-residential land uses. Obviously no existing city fits such a description in either site or structure. Yet some approach it more nearly than others. Accordingly a list of minimum requirements of cities used in testing the lognormal hypothesis was established. Since large cities frequently have radial arteries emanating from the center (an approximation to the uniform access assumption) the first requirement concerns size:

1. The population of the central city is at least 100,000.

2. The center of the city, operationally defined as the peak land value intersection (PLVI), is farther than five miles from a large body of water, large hills, or mountain chains, these being considered large if they distort more than 90° of circular perimeter.
3. The PLVI is farther than five miles from a river or estuary greater than one-half mile average width.

4. The PLVI is farther than five miles from a state or international boundary.

5. The central city is farther than twenty-five miles from any other city containing more than one-third its population.

6. Both city block and census tract publications or their equivalents are available at Lovejoy Library, Southern Illinois University, Edwardsville, Illinois. The study area for each city is confined to its surrounding Standard Metropolitan Statistical Area (SMSA), or equivalent.

Item six is essentially pragmatic in nature, but it removed the possibility of observing the same city over time, since only data for 1960 were available, and these only for the United States.

The above requirements are of course arbitrary. Another researcher would have developed a different list, or possibly none at all. Consideration of regional location, for example, may have led to interesting results. It was not undertaken primarily because the author wished to investigate general rather than specific aspects of urban structure. An obvious task for some future study is to determine the effect that systematic deviations from the above or a similar list of requirements have on the goodness of fit between observed and expected urban residential distributions. This is equally true for the lognormal and the Clark negative exponential hypothesis.

The decision as to whether a city passed items two, three, and four (natural barriers and political boundaries) was made on the basis of Census Tract maps, the location of the PLVI being inferred from the size of the tracts—the PLVI is generally in the central business district, which in turn is generally in the center of the
set of smallest tracts. Fifty-one cities satisfied these requirements, and were tested against item five (distance from other cities) by use of a road atlas. This eliminated nine from consideration bringing the total number of cities qualifying for testing the lognormal hypothesis to forty-two. These are presented in Table 1.

**TABLE 1**

Cities Meeting Minimum Requirements for Testing Lognormal Hypothesis

<table>
<thead>
<tr>
<th>Albuquerque, N.M.</th>
<th>Fresno, Calif.</th>
<th>Rochester, N.Y.*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlanta, Ga.**</td>
<td>Grand Rapids, Mich.*</td>
<td>Rockford, Ill.</td>
</tr>
<tr>
<td>Austin, Tex.**</td>
<td>Houston, Tex.</td>
<td>Sacramento, Calif.</td>
</tr>
<tr>
<td>Baltimore, Md.</td>
<td>Knoxville, Tenn.</td>
<td>Salt Lake City, Utah*</td>
</tr>
<tr>
<td>Columbus, Ohio</td>
<td>Lubbock, Tex.*</td>
<td>South Bend, Ind.*</td>
</tr>
<tr>
<td>Dayton, Ohio</td>
<td>Nashville, Tenn.*</td>
<td>Syracuse, N.Y.</td>
</tr>
<tr>
<td>Denver, Colo.</td>
<td>New Orleans, La.*</td>
<td>Topeka, Kans.</td>
</tr>
<tr>
<td>Des Moines, Iowa</td>
<td>Oklahoma City, Okla.</td>
<td>Tucson, Ariz.</td>
</tr>
<tr>
<td>Fort Wayne, Ind.*</td>
<td>Pittsburgh, Penn.</td>
<td>Wichita Falls, Tex.</td>
</tr>
</tbody>
</table>

The above cities were numbered and a table of random digits used to select ten, against which the lognormal hypothesis would be tested. These ten cities are shown with an asterisk in Table 1. Two of the cities in Table 1, Atlanta, Ga. and Austin, Tex., were not considered for inclusion in this study, since the author had already analyzed them. It was their analysis which led to the lognormal hypothesis.
Data Collection

Data were gathered in the following manner:

1. The PLVI was ascertained by writing a letter to the tax assessor of the subject cities.

2. Census Tract and City Blocks maps and reports (U.S. Census, 1960) were procured.

3. A sheet of transparent mylar was placed on top of the Census Tract map and the geographic center (centroid) of each tract was marked with pencil.

4. The mylar sheet was then placed on top of a grid and the $X, Y$ co-ordinates of each centroid were entered on data sheets, along with the tract identification number.

5. A similar technique was applied to remaining Census Tract maps if there were more than one for the city. Prominent boundary intersections were used to register the various maps.

6. The tracts nearest the PLVI were located on the City Blocks map, and divided into sections which increased in area with increasing distance from the PLVI. (See the South Bend example in Figure 2). The number of tracts thus sectioned varied from city to city rather unsystematically, unfortunately, as did section boundaries. In defense, it may be stated that to develop a rigid method of sectioning would have involved an undue amount of labor. Furthermore, if unsystematic, the sectioning was at least not gerrymandered, since knowledge of the population in each block was withheld until sectioning was complete.

7. Sectioned City Blocks centroids were transferred to the data sheet in the same manner as Census Tract centroids along with register locations.

8. Data sheets were completed with the addition of map scales; tract and section populations; $P$, the total population of the SMSA; and transformation distances. The latter, in conjunction with the map scale and register co-ordinates served to locate each centroid properly with respect to distance and direction from the PLVI, regardless of the map on which the centroid originally appeared.

9. Data from these sheets were coded on General Electric Time Share Terminal paper tape for electronic data processing.
Base Source: South Bend, Indiana, By Census Tracts
And Blocks: 1960, U. S., Bureau of the Census

Fig. 2—Enumeration sections, South Bend, Ind.
Estimation of Parameters

Parameters were estimated by the method of maximum likelihood. In the case of the normal distribution, this involves finding the mean and variance of the sample values and assuming these are the most likely estimates of the mean and variance of the population. Since the lognormal distribution can be transformed into the normal simply by taking the logarithms of the distance, the method is simple and straightforward. The equations for $\mu$ and $\sigma^2$, the sample mean and variance, are:

$$
\mu = m = \frac{1}{P} \sum_{i=1}^{N} P_i \cdot \ln r_i \quad \sigma^2 = s^2 = \frac{1}{P-1} \sum_{i=1}^{N} P_i (\ln r_i - m)^2
$$

where $P_i$ is the population of the $i^{th}$ areal unit; $r_i$ is the distance from its centroid to the PLVI; and $N$ is the number of centroids.

The actual computations were performed by the computer once data were entered. Since $r_i$ was not an input, it was solved for by the well known relationship:

$$
\sqrt{\sum (x_i - x_0)^2 + (y_i - y_0)^2}
$$

where $x_i, y_i$ are the co-ordinates of the $i^{th}$ centroid; $x_0, y_0$ the co-ordinates of the PLVI. Estimates of $\mu$ and $\sigma^2$ for the ten cities studied are given in Table 2.
TABLE 2

Estimates of Parameters of Cities Studied

<table>
<thead>
<tr>
<th>City</th>
<th>Geometric Mean, $\mu$</th>
<th>Standard Deviation, $\sigma$</th>
<th>Variance, $\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fort Wayne, Ind.</td>
<td>0.984</td>
<td>0.848</td>
<td>0.719</td>
</tr>
<tr>
<td>Grand Rapids, Mich.</td>
<td>0.824</td>
<td>0.607</td>
<td>0.369</td>
</tr>
<tr>
<td>Lubbock, Tex.</td>
<td>0.863</td>
<td>0.624</td>
<td>0.390</td>
</tr>
<tr>
<td>Nashville, Tenn.</td>
<td>1.263</td>
<td>0.675</td>
<td>0.456</td>
</tr>
<tr>
<td>New Orleans, La.</td>
<td>1.130</td>
<td>0.751</td>
<td>0.564</td>
</tr>
<tr>
<td>Phoenix, Ariz.</td>
<td>1.675</td>
<td>0.870</td>
<td>0.758</td>
</tr>
<tr>
<td>Rochester, N.Y.</td>
<td>1.123</td>
<td>0.831</td>
<td>0.690</td>
</tr>
<tr>
<td>Salt Lake City, Utah</td>
<td>1.420</td>
<td>0.838</td>
<td>0.702</td>
</tr>
<tr>
<td>San Antonio, Tex.</td>
<td>1.261</td>
<td>0.661</td>
<td>0.437</td>
</tr>
<tr>
<td>South Bend, Ind.</td>
<td>0.984</td>
<td>0.768</td>
<td>0.590</td>
</tr>
</tbody>
</table>

Tests of Significance

One graphical and two statistical tests of lognormality were applied to data from each of the ten cities.

**Graphical Test**

The graphical test simply involved plotting each city's cumulative frequency distribution on log-probability paper. This paper has a logarithmic scale in one direction and a cumulative normal scale in the other direction, so that it produces a straight line for any cumulative lognormal distribution. The cumulative frequency distributions were obtained by ranking the centroids in increasing order by distance from the city center, adding successive centroid populations so that a running summation is formed, and then dividing each summation by the total population and multiplying this by one-hundred to obtain a percentage. Distances were plotted on the logarithmic scale, and cumulative percent of population on the probability scale. Results are shown in Figures 3-13. The straight
Fig. 3.—Cumulative Frequency Graph: Fort Wayne, Ind.
Fig. 4.—Cumulative Frequency Graph: Grand Rapids, Mich.
Fig. 5.—Cumulative Frequency Graph: Lubbock, Tex.
Fig. 6.—Cumulative Frequency Graph: Nashville, Tenn.
Fig. 7.—Cumulative Frequency Graph: New Orleans, La.
Fig. 8.—Cumulative Frequency Graph: Phoenix, Ariz.
Fig. 9.—Cumulative Frequency Graph: Rochester, N. Y.
Fig. 10—Cumulative Frequency Graph: Salt Lake City, Utah.
Fig. 11.—Cumulative Frequency Graph: San Antonio, Tex.
Fig. 12.—Cumulative Frequency Graph: South Bend, Ind.
Fig. 13.—Cumulative Frequency Graph: Salt Lake City, Utah, Northern Sector.
lines indicate the position of the theoretical curves, based on the estimates of $\mu$ and $\sigma$ given in Table 2.

All of the graphs display a slight downward concavity, especially near the origin, which indicates that the lognormal approximation tends to underestimate the number of persons living close to the center. In seven of the cities, this central dip is not particularly pronounced except for the one-half percent of the population living closest to the center. Of the remaining three, Lubbock, San Antonio, and Salt Lake City, more serious departures from linearity are present on the graph. Lubbock (Figure 5) appears to be composed of two straight line segments, with the break in slope occurring at the first quartile. San Antonio (Figure 11) has a rather pronounced bulge at about the ninth decile. No attempt has been made to interpret these departures from lognormality. In the case of Salt Lake City, the break occurring at the eighth decile only occurs in the southern half of the SMSA, towards Provo, Utah. This is deduced from Figure 13, which shows a well defined linear trend for centroids located in the northern half of the SMSA.

**Statistical Tests**

**Kolmogorov-Smirnov:** The statistic is based on $D_n$, defined as the maximum of all deviations of the empirical from the theoretical cumulative distributions. The test is non-parametric and distribution free, but requires that the theoretical distribution be completely specified, that is, no parameters may be estimated from the sample. Since $\mu$ and $\sigma$ were estimated from each of the city samples, the Kolmogorov-Smirnov statistic is not strictly applicable in this context.
study. Professor Clements has suggested, however, that the test may be applied in a conservative sense, in that if a significant difference does indeed occur, one would have high confidence that the empirical and theoretical curves do not describe the same population. Also, if the sample is large, the distribution of $D_n$ is not likely to differ greatly from its tabled values. Professor Clements went on to note that if none of the ten cities showed significant difference, it would be well to test the power of the statistic.  

Values of $D_n$ were derived by comparing the empirical cumulative distribution with a lognormal cumulative distribution having the same mean and variance—the sample size, $N$, is equal to the total number of centroids for each city. These are shown together with the computed chi-square in Table 3.

<table>
<thead>
<tr>
<th>CITY</th>
<th>Number of Observations, $N$</th>
<th>Maximum Deviation, $D_n$</th>
<th>Confidence Level</th>
<th>Chi-Square</th>
<th>Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fort Wayne, Ind.</td>
<td>85</td>
<td>9.7</td>
<td>.80</td>
<td>12.6</td>
<td>.95-.975</td>
</tr>
<tr>
<td>Grand Rapids, Mich.</td>
<td>82</td>
<td>8.4</td>
<td>.80</td>
<td>8.6</td>
<td>.90</td>
</tr>
<tr>
<td>Lubbock, Tex.</td>
<td>36</td>
<td>21.2</td>
<td>.90-.95</td>
<td>16.0</td>
<td>.99-.995</td>
</tr>
<tr>
<td>Nashville, Tenn.</td>
<td>97</td>
<td>7.6</td>
<td>.80</td>
<td>4.3</td>
<td>.90</td>
</tr>
<tr>
<td>New Orleans, La.</td>
<td>189</td>
<td>4.9</td>
<td>.80</td>
<td>3.9</td>
<td>.90</td>
</tr>
<tr>
<td>Phoenix, Ariz.</td>
<td>154</td>
<td>5.3</td>
<td>.80</td>
<td>7.7</td>
<td>.90</td>
</tr>
<tr>
<td>Rochester, N.Y.</td>
<td>143</td>
<td>5.6</td>
<td>.80</td>
<td>3.3</td>
<td>.90</td>
</tr>
<tr>
<td>Salt Lake City, Utah</td>
<td>113</td>
<td>8.4</td>
<td>.80</td>
<td>12.0</td>
<td>.95-.975</td>
</tr>
<tr>
<td>San Antonio, Tex.</td>
<td>138</td>
<td>8.1</td>
<td>.80</td>
<td>15.9</td>
<td>.99-.995</td>
</tr>
<tr>
<td>South Bend, Ind.</td>
<td>81</td>
<td>7.8</td>
<td>.80</td>
<td>7.4</td>
<td>.90</td>
</tr>
</tbody>
</table>

Note: Except when bracketed, the true confidence level is less than the figure given.

8Dr. Kermit Clements, interview held at Southern Illinois University, March and April, 1969.
Of the ten cities, only one, Lubbock, deviates significantly from lognormality at the .90 confidence level according to the Kolmogorov-Smirnov test. Inasmuch as the graphical test indicates strong departures from lognormality for Salt Lake City and San Antonio, in addition to Lubbock, one may conclude that the power of the test has indeed been reduced by having to estimate parameters from the sample. Still, the test is useful, in that we may with great confidence assert that the distribution of the residence population of Lubbock, Texas, by distance from the city center, is not lognormal.

Chi-square: This test is well known to geographers after the spirited Zobler-Mackay exchange on its uses in regional geography. As a consequence of the assumed loss of power of the Kolmogorov-Smirnov test, the author decided to subject his data to the chi-square test of goodness of fit.

For each city, eight class intervals were established, with limits determined by the following formula:

\[ c = \mu \pm \frac{c}{2} \cdot \ell \]

\[ \ell = 1, 2, 3 \]

where \( \mu \) and \( \sigma \) are the mean and variance for the city. The first and last intervals were unbounded on their lower and upper sides, respectively, so that between them, the eight class intervals contained the total city population.

---

Chi-square was computed from the formula

\[ \chi^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} \]

where \( O_i \) is the observed population of the \( i \)th class interval, \( E_i \) is the expected population, and \( K \) is the number of class intervals. This formula assumes that the sum of the \( O_i \)'s equals the number of observations, but for this study, this is clearly not the case, since the sum of the \( O_i \)'s equals the population of the city, whereas the number of observations is equivalent to the number of centroids. Thus the chi-square computed from II.1, is unduly inflated. A simple relationship exists for correcting this, however:

Let

\[ T = \sum_{i=1}^{K} O_i \]
\[ N = \text{Number of observations} \]
\[ \chi^2 = \text{chi-square as computed from II.1} \]
\[ \chi^2_N = \text{corrected chi-square} \]

Then

\[ \chi^2_N = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i \cdot \frac{N}{T}} \cdot \frac{N^2}{T^2} \]

and

\[ \chi^2_N = \frac{N}{T} \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \frac{N}{T} \chi^2 \]

Whether II.2 is valid is unknown. If the number of persons corresponding to each centroid were a constant, say C, then II.2 would seem quite reasonable, for then the probability that \( C \) people would be found in a class interval containing \( n \) centroids would equal the probability that \( n \) centroids were in the class interval. Similarly, if there were only random variation from C, II.2 would seem to be approximately true, although this is based merely on
intuition. A thorough mathematical analysis of the problem is clearly needed, but cannot be undertaken here.

Results of the chi-square test, as modified above, are given in Table 3. In addition to the significant differences indicated for Lubbock, Salt Lake City, and San Antonio, which were expected from the graphical test, Fort Wayne appears to deviate significantly (at the .05 - .025 level) from lognormality. This comes as a surprise, for on lognormal paper, the fit appears quite good (Figure 3).

Conclusions: The conclusions which may be derived from any statistical goodness of fit test are somewhat negative when the aim of the research is to establish correspondence between an empirical and a theoretical curve, since the tests are designed to determine significant difference rather than agreement. Thus, the fact that Rochester shows a neither significant deviation, $D_n$, nor a significant chi-square indicates only that there is insufficient evidence to reject the lognormal hypothesis for that city. It does not give us a degree of certainty that the lognormal hypothesis is valid, because there are an infinite number of curves which would fit the empirical distribution as well or better than the lognormal.

Nevertheless, on the basis of the three modes of testing, one may tentatively accept the lognormal hypothesis for six of the ten cities. Of the remaining four, Fort Wayne shows significant difference on the chi-square test but not on the graphical or deviation tests; Salt Lake City and San Antonio show departures from lognormality on the chi-square and graphical tests, but not on the deviation test; and Lubbock is a uniformly poor fit. It is concluded that the lognormal
distribution is acceptable for describing the distribution of the residence population by distance from the city center for some, but not all, cities. Further research is clearly needed.

**Limitations**

There are strong limitations on the usefulness of this investigation, stemming largely from the small number of cities studied and from lack of spatial and temporal generality. (Only cities located in the United States were studied, and these only for 1960). Further, research was limited to those cities which passed a rather rigorous list of requirements, and although this was thought a necessary step in the study design, it has decreased the generality of the findings.

Also, there is a serious limitation on the interpretation to be given the confidence levels for the statistical tests (Table 3), since the assumptions underlying each test have been violated. In the Kolmogorov-Smirnov deviation test, it is thought that estimation of parameters from the sample leads to "conservative" confidence levels; but the behavior of the "corrected" chi-square is unknown. The confidence levels reported are thought to have heuristic value, however.

Finally there is the problem of random sampling, on which all statistical tests are based. The question is whether, having gathered data from all of the census tracts within an SMSA, one has a random sample. The answer to this rests ultimately on the status to be accorded statistical populations: May they be considered a
"random sample of some hypothetical population of possible values"?\textsuperscript{11}

Thomas and Anderson, quoting Fisher and others, seem to think so,\textsuperscript{12}
as does the author: Thus this is not thought to be a limitation on the study.


\textsuperscript{12}Ibid., pp. 435-436.
III

THEORY

The conclusions of the preceding section are assumed to be valid, at least for some cities. Notation is introduced and some characteristics of lognormal distributions are discussed. The relationship between population distribution and population density is examined and a set of postulates are put forth to explain the former. Consequences of these postulates are then explored and are shown to have important bearing on relationships geographers have found interesting.

Notation

It will be convenient to introduce the following notation, in order to expedite discussion:

- $x, y$ variables
- $x_i, y_i$ particular values, or realizations of $X$ and $Y$
- $\{x_i\}$ a sequence of $X$'s
- $P\{A\}$ the probability of the event $A$
- $E\{x_i\}$ the mean, or expectation, of $\{x_i\}$
- $Var\{x_i\}$ the variance of $\{x_i\}$
- $\gamma : N(\mu, \sigma^2)$ $\gamma$ is normally distributed with mean $\mu$ and variance $\sigma^2$

---

Characteristics of Lognormal Distributions

Definition: Given a positive variate \( X \) such that \( 0 < X < \infty \) such that
\[
Y = \ln X \quad \text{and} \quad Y \sim N(\mu, \sigma^2),
\]
then \( X \) is said to be lognormally distributed with mean \( \mu \) and variance \( \sigma^2 \). That is, \( X \sim \lambda(\mu, \sigma^2) \)

The distribution functions are defined as:
\[
N(y | \mu, \sigma^2) = P \{ Y \leq y \} \\
\lambda(x | \mu, \sigma^2) = P \{ X \leq x \}
\]

Density Function: The density function for \( X \) is the derivative of its distribution function:
\[
\frac{d \lambda(x | \mu, \sigma^2)}{dx} = \frac{1}{x \sigma \sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2 \sigma^2}\right)
\]
from the definition of the normal distribution and the relation
\[
Y = \ln X
\]

Measures of Central Tendency and Quantiles: \(^{15}\) Given \( \mu \) and \( \sigma^2 \), the position on the \( X \)-axis of the mean, median, and mode can be computed

\(^{14}\)Condensed from Aitchison and Brown, The Lognormal Distribution, passim, chap. i, and p. 110.

\(^{15}\)A "quantile" is the point on the \( X \)-axis corresponding to a given fraction of the area under the density curve. A quantile of order .3 for example locates the point at which 30% of the number of observations are less than or equal to \( X \). Special names have been given to certain of the quantiles: Thus the .25 and the .75 quantiles are called the first and third "quartiles," respectively; the .5 quantile is called the "median"; the .1, .2, .3, ..., .9 quantiles, the "deciles"; and the .01, .02, .03, ..., .99 quantiles, the
as follows:

\[ X_{\text{mean}} = e^{\mu + \frac{1}{2} \sigma^2} \]
\[ X_{\text{median}} = e^\mu \]
\[ X_{\text{mode}} = e^{\mu - \sigma^2} \]

Quantiles of any order may be determined from the quantiles of the standardized normal distribution by the relationship:

\[ Q_q^\sigma = e^{\mu + U_q \sigma} \]

where \( Q_q^\sigma \) and \( U_q \) are the quantiles of order \( q \) of \( \chi(\mu, \sigma^2) \) and of \( \mathcal{N}(0,1) \), respectively. For example, the first decile is \( e^{-1.282 \sigma} \), the third quartile, \( e^{\mu + 0.745 \sigma} \).

**Reproductive Properties:** The following theorems and corollaries will be important in the discussion to follow. They are a result principally of the reproductive properties of the normal distribution and the characteristic property of the logarithmic function

\[ \ln X_1 + \ln X_2 = \ln X_1 \cdot X_2. \]

**Theorem III.1**

If \( Y : \mathcal{N}(\mu, \sigma^2) \) and \( a \) and \( b \) are constants, then

\[ a + bY : \mathcal{N}(a + b\mu, b^2\sigma^2). \]


This leads directly to:

Corollary III.1

If \( X \sim \lambda(\mu, \sigma^2) \) and \( b \) and \( c \) are constants, where \( c > 0 \)
(say \( c = e^a \))

then

\[
  cX^b \sim \lambda(a + b\mu, b^2\sigma^2).
\]

The additive property of the normal distribution leads to:

Theorem III.2

If \( X_1 \) and \( X_2 \) are independent \( \lambda \)-variates such that
\( X_1 \sim \lambda(\mu_1, \sigma_1^2) \)
and \( X_2 \sim \lambda(\mu_2, \sigma_2^2) \), then the product
\( X_1X_2 \sim \lambda(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2) \)

Corollary III.2

\[
  \int_{0}^{\infty} \lambda(\mu_1, \sigma_1^2) d \lambda(\mu_2, \sigma_2^2) = \lambda(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)
\]

This corresponds the convolution property for the normal integral. 17

Corollary III.3

Let a distribution be decomposed into a number of sectors and
suppose that each of the sectors is \( \lambda(\mu', \sigma^2) \) or \( \lambda(\ln x, \frac{1}{2} \sigma^2, \sigma^2) \)
where \( \mu' \) is the arithmetic mean of the sector. Then if

1. \( \sigma^2 \) is constant for all sectors

2. the number of sectors is large enough so that the distribution
of \( \mu' \) approximates to a continuous distribution, and this
distribution is lognormal, say \( \lambda(\mu', \sigma^2) \)

Then from Corollary III.2

\[
  \int_{0}^{\infty} \lambda(\ln x, -\frac{1}{2} \sigma^2, \sigma^2) d \lambda(\mu, \mu_0, \sigma^2) = \lambda(\mu, -\frac{1}{2} \sigma^2, \sigma^2)
\]

Mathematical Series, no. 9, (Princeton University Press, 1946),
p. 190, cited by Aitchison and Brown, The Lognormal Distribution,
p. 11.
**Limit Theorems**: These derive directly from the additive form of the central limit theorem.

**Theorem III.3**

If \( \{X_i\} \) is a sequence of independent, positive variates such that

\[
E \left\{ \ln X_i \right\} = \mu_i \\
D^2 \left\{ \ln X_i \right\} = \sigma_i^2 \\
E \left[ (\ln X_i - \mu_i)^3 \right] = \omega_i^3
\]

all exist for every \( i \), then if

\[
\mu(n) = \sum_{i=1}^{n} \mu_i \\
\sigma^2(n) = \sum_{i=1}^{n} \sigma_i^2 \\
\omega^3(n) = \sum_{i=1}^{n} \omega_i^3
\]

and

then the product

\[
\prod_{i=1}^{n} X_i \sim \lambda \left( \mu(n), \sigma^2(n) \right)
\]

approximately, provided

\[
\lim_{n \to \infty} \frac{\omega(n)}{\sigma^2(n)} = 0
\]

In the special case where \( \{X_i\} \) all have the same distribution:

**Theorem III.4**

If

\[
E \left\{ \ln X_i \right\} = \mu \\
D^2 \left\{ \ln X_i \right\} = \sigma^2
\]

both exist, then

\[
\prod_{i=1}^{n} X_i \sim \lambda \left( n \mu, n \sigma^2 \right)
\]

approximately.
The Relationship Between Population Distribution and Population Density

It has been indicated (above, p.3) that a consequence of the lognormality of the residence urban population is that gross population varies lognormally with distance. This will now be demonstrated.

We assume a city developed on an unbounded, featureless plain, with uniform access to the center from all directions. Under such conditions, a circular city will result, and within it, the residence population is assumed to vary lognormally with distance from the center. That is, from I.1:

$$\frac{dP}{dr} = \frac{P}{r \sigma \sqrt{2\pi}} e^{-\frac{1}{2 \sigma^2} (\ln r - \mu)^2}$$

The area, $A$, of a circle of radius $r$ is $A = \pi r^2$

So that

$$dA = 2\pi r \, dr$$

describes the area of the infinitesimal ring of radius $r$ and width $dr$.

Gross population density is defined as the number of persons $dP$

residing in $dA$. Thus

$$\frac{dP}{dA} = \frac{dP}{2\pi r \, dr} = \frac{P}{2\pi \sigma^2 \sqrt{2\pi}} e^{-\frac{1}{2 \sigma^2} (\ln r - \mu)^2}$$

Since

$$\frac{1}{r} = e^{-\ln r}$$

$$e^x \cdot e^y = e^{x+y}$$
we may rewrite III.4 as

\[ \frac{dP}{dA} = \frac{P}{2\pi r \sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (\ln r - \mu)^2 - \ln r} \]

The exponent

\[ -\frac{1}{2\sigma^2} (\ln r - \mu)^2 - \ln r \]

is expanded thus:

\[ -\frac{1}{2\sigma^2} \left[ (\ln r)^2 - 2\mu \ln r + \mu^2 + 2\sigma^2 \ln r \right] \]

\[ = -\frac{1}{2\sigma^2} \left[ (\ln r)^2 - 2\mu \ln r + \mu^2 - 2\mu \sigma^2 + \sigma^2 \right] \]

so that

\[ e^{-\frac{1}{2\sigma^2}(\ln r - \mu)^2} \ln r \]

\[ = \mu + \frac{\sigma^2}{2}, e^{\frac{1}{2\sigma^2} (\ln r - (\mu - \sigma^2))^2} \]

For any given \( \mu \) and \( \sigma \),

\[ C = \mu + \frac{\sigma^2}{2} \]

\[ b = \frac{C}{2\pi} \]

Thus III.4 may be written

\[ \frac{dP}{dA} = \frac{P \cdot b}{r \sigma \sqrt{2\pi}} e^{-\frac{1}{2\sigma^2} (\ln r - (\mu - \sigma^2))^2} \]

which is lognormally distributed as

As a check, we note that

\[ b = \frac{e^{-\mu + \frac{\sigma^2}{2}}}{2\pi} \]

\[ = e^{-\mu}, e^{\frac{\sigma^2}{2}} \]

The mean, \( \mu \), has been derived from the relation

\[ \mu = \frac{1}{N} \sum_{i=1}^{N} \ln r_i \]

and thus is in units of the logarithm of distance: \( e^{-\mu} \) therefore has units of inverse distance.
The variance, derived from

\[ \sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (\ln r_i - \mu)^2 \]

is a pure number, so that the product, \( b \), is in units of inverse distance, and \( III.4 \) is in units of persons (\( P \)) divided by distance squared \( (r \cdot e^{-\mu}) \), or person per unit area, which is correct.

A Deductive Explanation of the Lognormality of Urban Population Distribution

Perhaps it is presumptuous of the author to attempt an explanation of a relationship for which there is so little direct evidence. The postulates which follow, however, have been found to have predictive power beyond their original intent, and are thus thought useful.

The Postulates

1. In a given city, and a given small interval of time, all houses have an equal chance of attracting a neighbor.

2. The distance between a house and its attracted neighbor is a random proportion of the distance from the attractor to the center of the city. That is, if a house at \( r_i \) attracts a neighbor which locates at \( r_{i+1} \), then

\[ r_{i+1} - r_i = \varepsilon_{i+1} r_i \]

3. The random element \( \varepsilon_i \) is such that \( \sum_{i=1}^{\infty} \varepsilon_i \) is a sequence of independent, positive variates having the same probability distribution and such that the mean \( E \sum_{i=1}^{\infty} \ln (1 + \varepsilon_i) = \mu_0 \) and the variance \( D \sum_{i=1}^{\infty} \ln (1 + \varepsilon_i)^2 = \sigma_0^2 \) both exist.

The author agrees with Bunge when he states "...the plausibility or intuitive reality of a theory is not a valid basis
It is maintained that the postulates as given are not entirely unreasonable, however. Postulate (1) for example seems quite plausible when stated in its aggregate form: The number of people deciding to locate in a given large area tends to be proportional to the number of persons already there. Postulate (2) may be thought of as a result of the diminishing value of land with increase in distance from the center. The farther out, the greater the residence area a house buyer with a given income may purchase, and consequently the further he is from his neighbor. An increase proportional to distance from the center is the simplest possible function which describes this relationship. Postulate (3) is thought to be such a weak assumption that it requires no rationale.

Given the following conditions, the three postulates will generate an approximately lognormal distribution:

1. A focal point, or "center of the city," and at least one house at a distance from the focal point both exist.

2. The only mechanism governing the urban residential distribution is given by the three postulates (that is, one assumes urban growth to occur on an unbounded, featureless plain, with uniform access to the center from all directions).

The Lognormality of an Stage Location

Consider urban growth along some small sector, and suppose that in Postulate (1) the probability that a house attracts a neighbor during some small interval of time, , is , and that . Beginning with only one house, at , it is apparent that if

does not attract a neighbor, we may observe a succession of $\Delta T$'s until it does. When this occurs, there will be two houses, one at $r_0$ and one at $r_1 = r_0 + \epsilon, r_0 = r_0 (1 + \epsilon_1)$, from Postulate (2). Designate the house at $r_1$ a "first stage" house, and allow the passage of more $\Delta T$'s until it attracts a neighbor which locates at $r_2 = r_1 + \epsilon, r_1 = r_1 (1 + \epsilon_2) = r_0 (1 + \epsilon_1)(1 + \epsilon_2)$. Designate this a "second stage" house and proceed in like manner until an $n$ stage house results. Then

$$r_n = r_0 \prod_{i=1}^{n-1} (1 + \epsilon_i) : \chi(\xi_n r_0 + n \chi_{\epsilon_1, \epsilon_2}, n \sigma^2_2),$$
approximately, from Postulate (3), Corollary III.1, and Theorem III.4, with

$$\sigma^2_2 = \frac{E \xi_n (1 + \epsilon_1)^2}{\sigma^2_1} = \frac{D^2 \xi_n (1 + \epsilon_1)^2}{\sigma^2_1}.$$

In practice, the distribution of an $n$ stage location will converge to lognormality quite quickly. Hald gives examples of the distribution of the sum of four random numbers where the variable takes on the values $0, 1, ..., 9$ with equal probabilities, and concludes "that even for $n = 4$ the theoretical distribution does not deviate very much from the corresponding normal distribution."\textsuperscript{19} An analogous statement could of course be made about convergence to a lognormal distribution were the product, rather than the sum, computed.

**The Normality of the Stages**

Suppose the probability $\vartheta$, that a house attracts a neighbor during $\Delta T$ is small. Then by Postulate (1), each existing house attracts a neighbor with probability $\vartheta$. The probability that only one house is attracted is $\vartheta$, the probability that two are attracted

\textsuperscript{19}Hald, *Statistical Theory with Engineering Applications*, p. 193.
is $\Theta^2$, and the probability that $n$ are attracted is $\Theta^n$, by the multiplication rule for probabilities. Assume that $\Theta$ is sufficiently small such that the probability that two or more houses are attracted is nil. Then if by "generation" we mean an interval of time $\Delta T$ during which a neighbor is attracted, the probability distribution of the stages may be computed by summing the products of the combinations of probabilities which result in a stage $K$ house after $N$ generations. This is derived in the following manner (see Figure 14).

<table>
<thead>
<tr>
<th>GENERATION</th>
<th>STAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>B C</td>
</tr>
<tr>
<td>3</td>
<td>D E F</td>
</tr>
<tr>
<td>4</td>
<td>G H I J</td>
</tr>
</tbody>
</table>

Fig. 14.—Probability Triangle

During the first generation, there is only one possible event which may occur: A one-stage house, $A$, is attracted by the original house, $O$. This event may be indicated by writing $O \rightarrow A$, and $P_{O \rightarrow A^2} = 1$, by definition of "generation." During the second generation, two events, $O \rightarrow B$ and $A \rightarrow C$ may occur, since now two houses exist; and these events occur with equal probability, by Postulate (1). Thus $P_{O \rightarrow B^2} = P_{A \rightarrow C^2} = \frac{1}{2}$. The probability $P_{C^2}$ that a house actually locates at $C$, however, is conditioned on the event that $O \rightarrow A$ has already occurred. That is, both $O \rightarrow A$ and $A \rightarrow C$ must occur, and by the multiplication rule for probabilities,

$$P_{C^2} = P_{O \rightarrow A^2} \cdot P_{A \rightarrow C^2}.$$  

$P_{C^2}$ is more simply equal to $P_{O \rightarrow B^2}$. 

During the third stage three houses may attract neighbors, each with probability $\frac{1}{3}$. Four events may occur, however: $0 \rightarrow D; A \rightarrow E; B \rightarrow E; C \rightarrow F$. By analogies with the events $A \rightarrow C$ and $O \rightarrow E$, we see that $P \{F\} = P \{O \rightarrow A\} \cdot P \{A \rightarrow C\}; P \{C \rightarrow F\}$, and that $P \{D\} = P \{O \rightarrow D\}$. $P \{E\}$ is slightly more complicated since it can occur if either $A \rightarrow E$ or $B \rightarrow E$. Both cannot occur, however, since the probability that two houses are attracted during one generation is nil. Thus, by the addition rule for probabilities,

$$P \{E\} = P \{A \rightarrow E\} + P \{B \rightarrow E\},$$

provided $A$ and $B$ both exist. Since this is not certain, we write $P \{E\} = P \{O \rightarrow A\} \cdot P \{A \rightarrow E\} + P \{B \rightarrow E\}$.

The events thus far described are summarized below for the first three generations, together with numerical probabilities:

- $P \{O\} = 1$
- $P \{O \rightarrow A\} = 1$
- $P \{A\} = P \{O \rightarrow A\} = 1$
- $P \{O \rightarrow B\} = \frac{1}{3}$
- $P \{B\} = P \{O \rightarrow B\} = \frac{1}{3}$
- $P \{A \rightarrow C\} = \frac{1}{2}$
- $P \{C\} = P \{O \rightarrow A\} \cdot P \{A \rightarrow C\} = 1 \cdot \frac{1}{2}$
- $P \{O \rightarrow D\} = \frac{1}{3}$
- $P \{D\} = \frac{1}{3}$
- $P \{A \rightarrow E\} = \frac{1}{3}$
- $P \{B \rightarrow E\} = \frac{1}{3}$
- $P \{E\} = P \{O \rightarrow A\} \cdot P \{A \rightarrow E\} + P \{O \rightarrow B\} \cdot P \{B \rightarrow E\} = 1 \cdot \frac{1}{2} + \frac{1}{3}$
- $P \{C \rightarrow F\} = \frac{1}{3}$
- $P \{F\} = P \{O \rightarrow A\} \cdot P \{A \rightarrow C\} \cdot P \{C \rightarrow F\} = 1 \cdot \frac{1}{2} \cdot \frac{1}{3}$

As a check on the above probabilities, we note that either $P \{B\}$ or $P \{C\}$ must occur, but cannot both occur. Their sum must therefore equal unity, and this is so, since $P \{B\} = P \{C\} = \frac{1}{2}$ and $\frac{1}{2} + \frac{1}{3} = 1$. Similarly $P \{D\} + P \{E\} + P \{F\}$ must equal unity:

$$\frac{1}{3} + \frac{1}{3} + \frac{1}{2} = 1.$$
To find the probability distribution of the stages, one simply totals the probabilities over all generations that a house has been attracted to a given stage. This is then divided by the number of existing houses, \( N + 1 \), where \( N \) is the number of generations. Thus the probability that a house exists at stage one, after three generations, is \( \left( P \frac{A}{2} + P \frac{B}{2} + P \frac{C}{2} \right) / 4 = (1 + 1 + 1) / 4 = \frac{3}{4} \); at stage two, \( \left( P \frac{C}{2} + P \frac{E}{2} \right) / 4 = (1 + 1 + 1 + 1) / 4 = \frac{4}{4} \); and at stage three, \( P \frac{E}{2} / 4 = (1 + 1) / 4 = \frac{2}{4} \). The probabilities sum to 3/4, but since the probability of finding a house at the origin is certain, then 1/4 must be added to this, so that the sum of the probabilities is unity. Note above that the number of factors in each product is equal to the stage, \( K \), so that for example a two-stage has two factors in each of its products \( (1 \cdot \frac{1}{2}, 1 \cdot \frac{1}{2}, \frac{1}{2} \cdot \frac{1}{2}) \). Note also, that the number of terms equals the number of distinct combinations of three and further that each of the \( \frac{1}{n} \), \( n = 1, 2, 3 \) probabilities is equally represented. This will be true in general, regardless of the number of generations or of the size of \( K \), provided that \( K \leq N \). The number of terms for a \( K \)-stage location may be found by the binomial formula \( \binom{N}{K} \), where \( N \) is the number of generations, \( K \) is the stage, and

\[
\binom{N}{K} = \frac{N!}{K!(N-K)!}
\]

One may make use of this formula to determine the probability of finding a house at the \( K \)-stage after \( N \) generations in the following manner:
Let \( N = 5, K = 3 \)

Then by III.7, there are

\[
\frac{5!}{3!(5-3)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}{(1 \cdot 2 \cdot 3) \cdot (1 \cdot 2)} = 10
\]

distinct products to be summed, each with \( K = 3 \) factors, where the probability factors take on the values \( \frac{1}{n} \), \( n = 1, 2, 3, 4, 5 \), such that all are equally represented. This is demonstrated below:

<table>
<thead>
<tr>
<th>COMBINATIONS</th>
<th>PRODUCT</th>
<th>PROBABILITY</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 ( \cdot \frac{1}{2} \cdot \frac{1}{3} )</td>
<td>( \frac{1}{6} )</td>
<td>( \frac{2}{120} )</td>
</tr>
<tr>
<td>1 ( \cdot \frac{1}{2} \cdot \frac{1}{4} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{15}{120} )</td>
</tr>
<tr>
<td>1 ( \cdot \frac{1}{2} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{8} )</td>
<td>( \frac{12}{120} )</td>
</tr>
<tr>
<td>1 ( \cdot \frac{1}{3} \cdot \frac{1}{4} )</td>
<td>( \frac{1}{12} )</td>
<td>( \frac{10}{120} )</td>
</tr>
<tr>
<td>1 ( \cdot \frac{1}{3} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{15} )</td>
<td>( \frac{8}{120} )</td>
</tr>
<tr>
<td>1 ( \cdot \frac{1}{4} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{20} )</td>
<td>( \frac{6}{120} )</td>
</tr>
<tr>
<td>( \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{4} )</td>
<td>( \frac{1}{24} )</td>
<td>( \frac{5}{120} )</td>
</tr>
<tr>
<td>( \frac{1}{2} \cdot \frac{1}{3} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{30} )</td>
<td>( \frac{4}{120} )</td>
</tr>
<tr>
<td>( \frac{1}{4} \cdot \frac{1}{3} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{40} )</td>
<td>( \frac{3}{120} )</td>
</tr>
<tr>
<td>( \frac{1}{2} \cdot \frac{1}{4} \cdot \frac{1}{5} )</td>
<td>( \frac{1}{40} )</td>
<td>( \frac{2}{120} )</td>
</tr>
</tbody>
</table>

\[ \text{Sum of Products} \]

\[ \frac{17}{24} = \frac{85}{120} \]

The probability that a house exists at stage three after \( N = 5 \) generations is then found by dividing \( 17/24 \) by \( N + 1 = 6 \) to give .1180.

At this writing, the author is not able to demonstrate mathematically convergence of the above distribution to the normal distribution, although this is highly suspected. Nor is he able to give formulae to determine rapidly the distribution of stages for large \( N \). As plotted on normal probability paper, however, even with \( N = 31 \), the cumulative frequencies show a distinct linear trend (Figure 15).
Fig. 15.— Cumulative Frequency Graph Showing Convergence of a Theoretical to a Normal Distribution.
If the probability, $\varphi$, that a house attracts a neighbor during $\Delta T$ is made large, and if concurrently, the probability that the house attracts two or more neighbors is kept small, then the distribution of stages converges to the binomial. (Since each of the probability factors will approximately equal unity, and therefore so will their product.) In such a case, the number of houses existing after the $N$th generations is no longer $N + 1$, but $2^N$, since they will double with each generation. The probability that a house exists at the $K$th stage after $N$ generations is therefore given by:

$$P\{K\} = \binom{N}{K} \left(\frac{1}{2}\right)^N$$

III.8

If $N$ is large, then by De Moivre's theorem, III.8 is approximately normally distributed, with mean $\frac{N}{2}$, and variance $\frac{N}{4}$.

That is,

$$P\{K\} = N\left(\frac{N}{2}, \frac{N}{4}\right)$$

III.9

In general, however, we would not expect every house to attract a neighbor in any given time interval $\Delta T$, so that it seems presumptuous to suppose that the mean and variance of the stages are as given, although they perhaps tend around these values. The value of $N$ is easily determined from the total population of the city, $P$, since they are related as:

Then

$$P = 2^N$$

and

$$N = \frac{\ln P}{\ln 2}$$

III.10
Thus, in the small sector (above, p. 38) over which growth has been assumed to take place, the distribution of urban residences has been shown to approximate lognormality for all stages, taken as a whole. This conclusion has been deduced from a set of three postulates (above, p. 37) and two assumptions. It is hoped that one of the assumptions, that Postulate (1) does indeed imply an approximately normal distribution of stages (above, p. 46) will prove unnecessary when more penetrating mathematical analysis is applied. Similarly, it is hoped that an alternative to Corollary III.3 will be found so that the assumption of small variance (above, p. 46) may be dropped.

Of interest is the distribution of residences for the city as a whole, which the evidence presented in Section II has suggested is lognormal. This may be shown to be a result of the above discussion in either of two ways. First, one may assume that the above process occurs in K sectors independently, with the expectation of similar outcomes (i.e., lognormality, with equal means and variances for all K sectors). The various sectors are then pooled to give a composite distribution for the whole city and this is of course lognormal, since only the total population is changed. Second, one may invoke Corollary III.3, which frees one from having to assume a circular city. As a matter of fact, since Corollary III.3 assumes the arithmetic mean varies lognormally, while the variance remains constant, it implies that the edge of the urbanized area varies lognormally with respect to distance from the city center, which is of some interest.
Of concern also is the distribution of persons, rather than households. The author has assumed household size to remain constant as distance from the center increases, or at least to vary only randomly with distance. Dr. Lossau has suggested, however, that this is probably not the case for large cities. If one may assume that household size increases as distance to the center decreases, then the "central dip" (above, p. 23) may be explained. This of course implies that a better fit will occur if the distribution of households, rather than population, is studied.

Implications of the Postulates

The Rank Size Rule

Postulate (1) states that "during any given small time interval" all houses have an equal chance to attract a neighbor. The event that any one house attracts a neighbor during any specified time interval $\Delta T$ may be considered random, and since by Postulate (1) all houses have an equal chance of attracting a neighbor, the total number of houses attracted in the interval $\Delta T$ is randomly proportional to the number already existing. That this leads to an approximately lognormal distribution of the city sizes has been suggested by Berry and Garrison, and Berry has done research in

---

20 Dr. Carl Lossau, discussion held at Southern Illinois University, Dec., 1968.

this area. Siwon, however, has considered nearly the same process in a slightly different light and concludes that city sizes are best described by the Yule distribution, while Curry, arguing from an information theoretic viewpoint derives a negative exponential expression. Which of these distributions best fits empirical evidence will not be discussed here, as it seems to be an open question, but there are interesting theoretical consequences of assuming that the lognormal distribution exactly fits the data.

The Lognormality of Urban Areas and Densities

There is the fascinating relationship that if \( X \) is exactly lognormally distributed, and if \( X = YZ \), then \( Y \) and \( Z \) are also lognormally distributed, except in the special case where one is a constant and the other lognormal. As a hypothesis, one may assume that at least for larger cities, population is exactly lognormally distributed. Denoting the total population of the \( i \)th city by its area by \( A_i \), and its gross density by \( D_i \), then

\[
D_i = \frac{P_i}{A_i} \quad ; \quad P_i = D_i \cdot A_i
\]


26 Aitchison and Brown, The Lognormal Distribution, p. 12.
Then, since neither $D_i$ nor $A_i$ is constant, they are both lognormally distributed, at least for the larger cities.

To test this hypothesis, the author drew a random sample of 50 urban places in Illinois, 1960 from the *County and City Data Book* (1967) then recorded the areas and population. Population was divided by area for each city to compute density, and population, area, and density were plotted on lognormal paper (Figure 16).

Surprisingly, area and density appear to fall along approximately straight lines, but population does not. Population seemingly increases much too rapidly for even the logarithmic transformation to normalize the city size. Tiedemann took the square root of the logarithms of city sizes to normalize his Michigan data, but a different tactic was attempted here. Areas for urban places of less than 2500 are not listed in the *County and City Data Book* and consequently could not have been in the sample drawn. Samples which suppress a portion of the population are called "censored" and the point where the loss of information occurs is called the "point of truncation," so that the Illinois urban place sample has a point of truncation of 2500 persons. By count, there are 248 urban places

---


29 The count was taken from the *County and City Data Book*, pp. 585-586, for places larger than 2500 persons, and from U.S. Department of Commerce, Bureau of the Census, *Census of Housing, 1960: Illinois: State and Small Area*, HC(1), n. 15, pp. 125-127, for places in the range 1000-2500 persons.
Fig. 16—Cumulative Frequency Graph of Density, Population, and Area
in the size range 1000-2500 persons, and there are 327 urban places with larger than 2500 persons, making a total of 575 urban places. The sample drawn may then be thought to represent the upper 327/575 = 56% of the population of urban places. To expedite analysis, six additional sample cities were drawn randomly. The results, when graphed on lognormal paper beginning at the 45th percentile, show a greatly improved linear trend (Figure 17). Also on Figure 17 are the graphs of density and area, based on the expanded sample. Area shows a clear linear trend, but density does not, for unknown reasons.

The Stewart and Warntz Equation

In their intriguing paper "Physics of Population Distribution,"\(^{30}\) Stewart and Warntz state and give evidence for the relationship

\[
C = \frac{P}{A}^{3/4}\tag{III.12}
\]

where \(P\) is the population of any city, \(A\) is its area, and \(C\) is a constant for all cities. The value of \(C\) is determined by linear regression of the logarithms of population and area since the transformation

\[
\ln C = \frac{3}{4}\ln P - \frac{1}{4}\ln A\tag{III.13}
\]

describes a linear trend. No clear explanation for the relationship in III.12 has ever been given, but a formulation similar to that of III.13 is a direct consequence of the lognormal hypothesis.

Assuming that III.10 and III.11 are valid, we may write

\[
r = \mu \left( \frac{\ln x + \frac{\mu}{2\ln 2}}{\ln 2} \right) + \frac{\mu^2}{2\ln 2} + C\tag{III.14}
\]

Fig. 17—Cumulative frequency graph of density, population, and area, assuming truncation at 44th percentile.
where \( P \) is the total city population, which, from 1,1, apparently ranges to infinity. Obviously \( P \) does not range to infinity, but it is reasonable to assume that it extends beyond the city limits of most large urban areas. It is of interest to ask: What is the relationship between the size of the population of a central city and that of the total urban region?

If they are linearly related, then there exists some constant \( b \) such that

\[
\frac{P_{ci}}{P_{ri}} = b \quad \text{for all } i
\]

where \( P_{ci} \) is the population of the \( i \)th central city, and \( P_{ri} \) is the population of the total urban area. That \( b \) may be considered a quantile (above, p. 31) is obvious, and combining 111.2 with 111.14, we have:

\[
C^b = e^{ln r_0 + \frac{\mu_0 ln P}{2 ln^2}} + \frac{1}{b} \left( \frac{\mu^2}{4 ln^2} + C \right)^\frac{1}{2}
\]

III.15

By rotating III.15 about the origin, one may then estimate the area of the city, given \( r_0 \), \( \mu_0 \), \( b \), and \( C \). There is reason to believe, however, that \( b \) is equal to 50% (below p. 55), so that \( \frac{1}{b} \) equals zero and III.15 becomes

\[
C^b_{50} = e^{ln r_0 + \frac{\mu_0 ln P}{2 ln^2}}
\]

Further, since

\[
A = \pi C^b_{50}^2
\]

\[
ln A = ln \pi + 2 ln r_0 + \frac{\mu_0 ln P}{ln^2}
\]

III.16

which is identical to III.13 provided

\[
ln C' = -ln \pi - 2 ln r_0
\]

III.17
The Stewart and Warntz equation is no longer mystifying, and we may interpret the finding that the power term, $3/4$, has remained constant over time\textsuperscript{31} to mean that
\[
\frac{M_0}{\ln 2} = \frac{3}{4}
\]
(above, p. 39) is a constant over time.

**A City-Region Relationship**

An important assumption of the preceding section is that the relationship between the population size of an urban region (operationally defined as an SMSA) and that of its central city is linear. Furthermore, given evidence that the ratio of city population to SMSA population is 1:2, certain computations may be simplified (above, p. 54).

These points were tested by linear regression (see scattergram, Figure 18) of a random sample of fifty United States SMSA's, 1960, drawn from the *County and City Data Book, 1967*\textsuperscript{32}. Since the selection was random, several of the cities were drawn more than once. When this happened, they were counted more than once. Of the fifty cities, nine had multiple central cities, and were deleted from the analysis, bringing the sample size down to forty-one. There is some question as to which of the two variables is independent, and which dependent (in the statistical sense), and therefore the regression was run twice. The results are

\textsuperscript{31}Ibid.

\textsuperscript{32}pp. 432-573, passim.
Fig. 18.—Scattergram of Size of SMSA and Size of Central City Population (Thousands)
summarized in Table 4.

TABLE 4

Regression Analysis of SMSA and Central City Populations

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>EQUATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central City</td>
<td>Y = 48,921 + .3869 X</td>
</tr>
<tr>
<td>SMSA</td>
<td>y = 25,285 + 1.9605 X</td>
</tr>
</tbody>
</table>

Correlation Coefficient: r = .87; r² = .76

These results, together with Figure 18, tend to confirm the assumption of linearity and give evidence for the simplifying ratio. Indeed, the regression slope of SMSA to central city (1.96) is strikingly close to the desired slope of 1/1.50 = 2.0. There are of course statistical techniques for determining whether a regression slope deviates from a hypothetical slope by more than one would expect from chance alone. However, the technique requires the assumption of normal populations, which is of doubtful validity (above, pp. 43-52).

The Regression of the Means and Variances

If \( \mu_i \) and \( \sigma_i^2 \) are the mean and variance respectively of the \( i \)th city, and \( P_i \) is its total population, then from III.13

\[
\mu_i = \ln r_0 + \frac{\mu_0 \ln P_i}{2 \ln 2}
\]

\[
\sigma_i^2 = \frac{\sigma_0^2 \ln P_i}{4 \ln 2} + C
\]
One may use the regression of $\mu_i$ on $\ln \Pi_i$ of a sample of cities to estimate $\ln r_0$ and $\mu'/\ln 2$; and the regression of $\sigma^2$ on $\ln \Pi_i$ to estimate $\mu_o/4\ln 2$ and $c'$. These estimates can then be solved for $\mu_o$ and compared to the theoretical value of $\mu_o$ derived from III.16. Hypothetically, $\frac{\mu_o}{\ln 2} = 3$ so that $\mu_o = \ln 2 \times .75 = .8799$. Results of the regression, based on the ten cities studied in Section II (above, pp. 5-29) are summarized in Table 5.

**TABLE 5**

Regression of $\mu$ and $\sigma^2$ on the Logarithm of Population

<table>
<thead>
<tr>
<th>VARIABLES</th>
<th>EQUATIONS</th>
<th>CORRELATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dep.</td>
<td>Indep.</td>
<td></td>
</tr>
<tr>
<td>$\mu_i$</td>
<td>$\ln \Pi_i$</td>
<td>$\mu_i = -2.4564 + .2803 \ln \Pi_i$</td>
</tr>
<tr>
<td>$\sigma^2_i$</td>
<td>$\ln \Pi_i$</td>
<td>$\sigma^2_i = -.3527 + .0715$</td>
</tr>
</tbody>
</table>

$^*+$ Significant at .05 level

Since the distribution of cities by size is approximately lognormal, the logarithm of size is normally distributed; also, since $\mu_0$ is the mean of the distribution of logarithms of distance from the city center, and since these distances are lognormally distributed, then $\mu_i$ is normally distributed. If we may assume that the population variance of $\mu_i$ for all values of $\ln \Pi_i$ is constant, then we may use the T-test to determine whether the regression slope of $\mu_i$ versus $\ln \Pi_i$ deviates from the theoretical slope more than we would expect from chance alone.
If \( \hat{\mathcal{M}}_o \) (theoretical) = .5199, then the theoretical slope is computed from III.19 as:

\[
\frac{\hat{\mathcal{M}}_o}{2 \ln 2} = \frac{.5199}{2 \ln 2} = .3750
\]

so that we wish to compare .3750 with .2803. The null hypothesis is that there is no difference between the slope values: With \( T = .560 \), and with \( 10 - 2 = 8 \) degrees of freedom, this is accepted at the .95 confidence level. \( T_{.05} = .546 < .506 < T_{.00} = .889 \) with 8 d.f.

The variances are chi-square over degrees of freedom distributed and thus their slope cannot be meaningfully compared with the theoretical slope by means of the T-test. However, the slopes may be compared by inspection.

From III.20

\[
\frac{\hat{\mathcal{M}}_o}{4 \ln 2} > \frac{(.5199)^2}{4 \ln 2} = .0974
\]

The sample slope (Table 5) is .0715, but unfortunately, this fine correspondence cannot be given much weight, for as the scattergram (Figure 19) and the correlation coefficient (Table 5) both indicate, the relationship between \( \ln P_i \) and \( \sigma_i^2 \) is poor, and nearly any line passing through the dots may be considered a "best fit".

There is admittedly some problem with the constant in the variance regression model, since it is negative. Recall that it represents

\[
P \sigma_i^2 \{ \ln(1 + e_i) \}^2
\]

(above, p. 39) and should thus always be positive, and theoretically close to zero. Insertion of the theoretical variance regression slope into the regression model only compounds the difficulty, since
Fig. 19.—Scattergram of Regression of $\mu$ and $\sigma$ on the Logarithm of Population
the constant is still negative, and of greater magnitude. Four alternate explanations of this discrepancy come to mind:

1. The estimates of the variances for the ten cities are too low.

2. The cities themselves have uncommonly low variances.

3. There is some initial size a city must reach before any of the equations hold. (This is quite possible. The initial size, based on the point at which the variance becomes positive, is then computed to be \( \ln \frac{-352.7}{0.918} \) which corresponds to a city of 1400 people).

4. The equations themselves are wrong or require extensive modification. This is the most probable of the choices.

Much further study is required, notably an analysis of many more cities, in widely scattered locations and from many periods of time, before any firm conclusions may be reached.
CONCLUSION

It is hypothesized that the resident urban population is distributed lognormally with respect to distance from the city center. Ten cities were selected for testing this hypothesis, with mixed results. It is concluded that for some cities, the lognormal hypothesis is valid as presently formulated, but that for others, some revision must be made. The nature of these revisions is not discussed other than to note that the hypothesis should more nearly explain the distribution of households, rather than population, with respect to distance from the center.

This revision is suggested by a hypothetical construct designed to yield a lognormal distribution of household distances from the center. Certain implications of the construct are examined and are found to have bearing on aspects of urban structure geographers have found interesting. Since hitherto, these aspects have not been shown to be mathematically related, it is tentatively suggested that the construct may be useful for urban analysis.

The tentative nature of this suggestion cannot be over-stressed, for the empirical basis of the study is but ten cities, sharply restricted as to time and place. Many more cities must be analysed; many more implications must be studied. It is hoped that others will take an interest in the investigations and the approach taken herein, if only to offer criticism; for it is only through a
dialogue of ideas that knowledge of the world around us may be gained.
SOURCES CONSULTED

**Books**


**Government Sources**


Interviews

Clements, Kermit. Interview held at Southern Illinois University. March and April, 1969.


Journals


VITA
Graduate School
Southern Illinois University

Lawrence M. Ostresh, Jr. September 13, 1942

2108 Waterman Avenue, Granite City, Illinois

University of Illinois 1960-1962
Southern Illinois University 1966-1968 B.A., Geography
Southern Illinois University 1968-1969 Geography

Thesis Title: Some Theoretical Aspects of Urban Structure
Advisor: Mr. Richard Guffy