ON THE LOGARITHMIC FREQUENCY DISTRIBUTION AND THE SEMI-LOGARITHMIC CORRELATION SURFACE*

By

PAE-TSI YUAN

INTRODUCTION**

The method of treating frequency curves as developed chiefly by Edgeworth, Kapteyn, Van Uven and Wicksell occupies an important place in both theoretical and applied statistics. The essence of this method may be briefly summarized as follows:

Suppose a function of the variable z is distributed according to the normal law of error. Then, z certainly cannot be also normally distributed, unless the function is a linear function of z. Without losing generality, we shall write the normally distributed function in standard units as $x = f(z)$. Thus the origin of x is its mean and the unit of x is its standard deviation. The relative frequency of values of x between x and $x + dx$ is, therefore

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

and the relative frequency of values of z between z and $z + dz$ is

$$\frac{1}{\sqrt{2\pi}} f'(z) e^{-\frac{1}{2}\left[f(z)\right]^2} dz .$$

Thus if we have an observed frequency distribution of z and we know a normally distributed function of z, then we can

* A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in the University of Michigan.

** Papers written by the writers mentioned in this introduction are listed under the writers' names in the Bibliography at the end of this paper.

graduate the distribution of z by using this formula. Edgeworth calls this method of graduating a frequency distribution the method of translation. In two papers on "Skew Frequency Curves in Biology and Statistics" published in 1903 and 1916, J. C. Kapteyn elegantly set forth a theoretical foundation of this method. Later Wicksell gave a similar justification. Both of them based their "genetic theory of frequency", to use Wicksell's terminology, upon a generalized hypothesis of elementary errors.

In the present paper, we are interested only in the important special case where $x = \frac{1}{c} \log \frac{z-a}{b}$. The frequency function of z, then, becomes:

$$\frac{1}{\sqrt{2\pi}\, c(z-a)}\, e^{-\frac{1}{2c^2}\left(\log \frac{z-a}{b}\right)^2}$$

which is called the logarithmic frequency function.*

Numerous papers have been written on this frequency curve. Among the early writers were Francis Galton and McAllister. But a systematic treatment on the properties of this curve from the standpoint of mathematical statistics is still lacking. Hence, in the first part of this paper, such a treatment will be given, thus leading to some interesting relationships among the characteristics of this curve.

Various methods of determining the parameters of this frequency function have been proposed by writers on this subject. Pearson is the first writer to make use of the method of moments. Later this method was also applied by Jørgensen and Wicksell. In this paper, the method of moments will be considered and a table will be provided to facilitate the computation of the constants by this method.

Edgeworth, Wicksell and Van Uven all have contributed in

---

* For a justification of this frequency function based on Weber-Fechner's Psychophysical Law see the "Calculus of Observations" by E. T. Whittaker and G. Robinson, pp. 217-218. (Blackie & Son Ltd., London and Glasgow, 1929)

extending the method of translation to correlation surfaces. Wicksell's logarithmic correlation surface is particularly noteworthy. In the last part of this paper, a semi-logarithmic correlation surface of two variables will be developed and its properties studied.

The writer wishes to express his appreciation for the assistance Professor Cecil C. Craig has given him in making this study.

PART I

THE LOGARITHMIC FREQUENCY DISTRIBUTION

For the sake of clarity, it is desirable to state at the outset that the logarithmic frequency distribution represented by

$$F(z) = \frac{1}{\sqrt{2\pi}\, c(z-a)}\, e^{-\frac{1}{2c^2}\left(\log \frac{z-a}{b}\right)^2} \tag{1}$$

is unimodal and has three parameters. The parameter a is the finite lower or upper limit of z according to whether b is positive or negative. In the following discussions, unless the sign of b plays an important rôle, we shall take b to be positive and a to be the finite lower limit of z. However, the results of our discussions can be easily modified to cover the case where b is negative and a is the finite upper limit of z.

In the first eight sections of Part I the properties of the logarithmic frequency distribution will be treated from the standpoint of mathematical statistics,* and in section 9 the numerical application of this distribution will be discussed.

1. *AVERAGES*

We shall first give the analytic expressions of four different averages of z and then observe their relative magnitudes.

* Some topics under consideration here in regard to the properties of the logarithmic frequency distribution have also been discussed by many writers, among whom we may particularly mention McAllister, Kapteyn, Pearson and Pretorius. See the references under these writers' names in the Bibliography of this paper.

By definition, the arithmetic mean of z is

$$m = \int_a^\infty zF(z)dz = be^{\frac{c^2}{2}} + a .$$

The logarithm of the geometric mean of z about the point $z = a$ is given by

$$\int_a^\infty \log(z-a)F(z)dz = \log b.$$

Hence, the geometric mean of z about $z = a$ measured from $z = 0$ is

$$m_g = b + a .$$

Since the median of z corresponds to $x = \frac{1}{c}\log\frac{z-a}{b} = 0$, it is equal to

$$m_d = b + a .$$

Setting the derivative $\frac{dF(z)}{dz} = -\frac{\log\frac{z-a}{b} + c^2}{c^2(z-a)}F(z)$

equal to zero, we obtain the mode of z as

$$m_o = be^{-c^2} + a.$$

Thus, the geometric mean and the median are equal. Moreover,

$$m_o < m_d = m_g < m$$

2. *POINTS OF INFLECTION*

The second derivative of $F(z)$ is

$$\frac{d^2F(z)}{dz^2} = \frac{(\log\frac{z-a}{b})^2 + 3c^2 + \log\frac{z-a}{b} - c}{c^4(z-a)^2}F(z).$$

The roots of the equation

$$\left(\log \frac{z-a}{b}\right)^2 + 3c^2 \log \frac{z-a}{b} - c^2 = 0$$

are the points of inflection of the logarithmic frequency curve. We shall denote them by $\check{z}_1$ and $\check{z}_2$,

$$\check{z}_1 = b e^{-\frac{3}{2}c^2\left[1+\sqrt{1+\frac{4}{9c^2}}\right]} + a$$

$$\check{z}_2 = b e^{-\frac{3}{2}c^2\left[1-\sqrt{1+\frac{4}{9c^2}}\right]} + a$$

Note that the quantity under the radical sign is always positive and greater than one. Its square root is, therefore, greater than one in absolute value. Hence, $\check{z}_1 < b + a < \check{z}_2$. That is, the geometric mean and the median of z lie between the points of inflection.

Furthermore, if we observe that the points of inflection may be written in relation to the mode as

$$\check{z}_1 - a = (m_0 - a) e^{-\frac{c^2}{2}\left(1+3\sqrt{1+\frac{4}{9c^2}}\right)}$$

$$\check{z}_2 - a = (m_0 - a) e^{-\frac{c^2}{2}\left(1-3\sqrt{1+\frac{4}{9c^2}}\right)}$$

we see that $\check{z}_1 < m_0 < \check{z}_2$.

But the mean does not always lie between the two inflection points, since

$$\check{z}_1 - a = (m - a) e^{-\frac{c^2}{2}\left(4+3\sqrt{1+\frac{4}{9c^2}}\right)}$$

$$\check{z}_2 - a = (m - a) e^{-\frac{c^2}{2}\left(4-3\sqrt{1+\frac{4}{9c^2}}\right)} .$$

Obviously, $\breve{z}_1$ is always less than the mean. But when $c^2 > 4/7$, the mean is situated above both points of inflection.

Now, the relation of the averages and the points of inflection, when $c^2 < 4/7$, may be expressed by the inequality

$$\breve{z}_1 < m_0 < m_d = m_g < m < \breve{z}_2$$

which holds for almost all practical cases, since c^2 rarely exceeds $4/7$ in practice.

3. *HIGH CONTACT*

A frequency function is said to have high contact, if the function and all its derivatives vanish at the upper and the lower limits of the variable z. We know that the logarithmic frequency function vanishes at both the finite and the infinite limits of z. It can be easily seen that all its derivatives also vanish at these points, if we make the substitution $-z' = \log \frac{z-a}{b}$, which will throw every derivative of the logarithmic frequency function into a product of two factors, one being a polynomial in z' and another being $e^{-\frac{1}{2c^2}z'^2 + kz'}$ where k is a positive integer. Thus, it is obvious that all the derivatives become zero, as z' approaches $\pm\infty$, which correspond to the finite and the infinite limits of z. For instance, this substitution will put the first derivative of the logarithmic frequency function $F(z)$,

$$\frac{dF(z)}{dz} = -\frac{\log \frac{z-a}{c} + c^2}{c^2(z-a)} F(z)$$

into the form

$$\frac{z' - c^2}{\sqrt{2\pi}\, c^3 b^2}\, e^{-\frac{1}{2c^2}z'^2 + 2z'}$$

which clearly goes to zero as z' approaches $\pm\infty$, that is, as z

approaches "a" and infinity.

The logarithmic frequency function, therefore, has high contact.

4. *MOMENTS*

We shall study the practical application of the method of moments to determine the parameters of the logarithmic frequency distribution in section 9. But at present we must know the relationships between the parameters and the moments in order to discuss the properties of dispersion, skewness and kurtosis.

First, we shall express the moments in terms of the parameters:

The s-th moment of z about the point $z=a$ is given by

$$\mu_s' = \int_a^{\infty} (z-a)^s F(z)\,dz = b^s e^{\frac{s^2c^2}{2}}.$$

And we also have the recurring relation $\mu_s' = b e^{\frac{(2s-1)c^2}{2}} \mu_{s-1}'$.

The s-th moment of z about the mean is

$$\mu_s = \int_a^{\infty} (z-m)^s F(z)\,dz = b^s e^{\frac{sc^2}{2}} \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} e^{\frac{k(k-1)c^2}{2}}.$$

Consequently, the s-th standard moment of z, $\alpha_s = \dfrac{\mu_s}{\mu_2^{s/2}}$ is

$$\alpha_s = (e^{c^2}-1)^{-\frac{s}{2}} \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} e^{\frac{k(k-1)c^2}{2}}.$$

Setting s equal to 3 and 4, we have

$$\alpha_3 = \pm(e^{c^2}-1)^{\frac{1}{2}}(e^{c^2}+2)$$

$$\alpha_4 = 3+(e^{c^2}-1)(e^{3c^2}+3e^{2c^2}+6e^{c^2}+6)$$

which will be discussed in connection with skewness and kurtosis. Note that the sign of α_3 follows that of b, because the sign of the third moment of z about the mean is determined by b^3.

Now, we want to express the parameters in terms of the moments. It is clear that there is an infinite number of ways to accomplish this, since there is an infinitude of moments. But we are particularly interested in the expressions of the parameters in terms of the mean and the second and third moments about the mean. Letting $\omega = e^{c^2}$, we have

$$
\begin{aligned}
m &= b\omega^{\frac{1}{2}} + a \\
\mu_2 &= b^2\omega(\omega - 1) \\
\mu_3 &= b^3\omega^{3/2}(\omega-1)^2(\omega+2).
\end{aligned}
\tag{2}
$$

Solving these equations for the parameters, we find ω is the only real root of the cubic:

$$\omega^3 + 3\omega^2 - (4 + \alpha_3^2) = 0 \tag{3}$$

Hence, the parameters c, b and a may be expressed as

$$
\begin{aligned}
c &= (\log \omega)^{\frac{1}{2}} \\
b &= \left(\frac{1}{\omega-1}\right)^{\frac{1}{2}}\left(\frac{1}{\omega}\right)^{\frac{1}{2}}\sigma = \left(\frac{\omega+2}{\alpha_3}\right)\left(\frac{1}{\omega}\right)^{\frac{1}{2}}\sigma \\
a &= m - \left(\frac{1}{\omega-1}\right)^{\frac{1}{2}} = m - \left(\frac{\omega+2}{\alpha_3}\right)\sigma
\end{aligned}
\tag{4}
$$

where the sign of b follows that of α_3, and $\sigma = \sqrt{\mu_2}$. The practical application of (3) and (4) will be discussed in section 9. We shall now turn our attention to some other properties of the logarithmic frequency distribution.

5. *DISPERSION*

The dispersion of z about the mean may be measured by the standard deviation, $\sigma = \sqrt{\mu_2} = be^{\frac{c^2}{2}}(e^{c^2} - 1)^{\frac{1}{2}}$. Denote the deviation of z from the mean in terms of the standard deviation by $t=(z-m)/\sigma$. Then, with the aid of formulae (1), (2) and (4), we obtain the distribution of t as

$$\frac{(e^{c^2}-1)^{\frac{1}{2}}}{\sqrt{2\pi}\,c\left[1+(e^{c^2}-1)^{\frac{1}{2}}t\right]}\, e^{-\frac{1}{2c^2}\left\{\log\left[1+(e^{c^2}-1)^{\frac{1}{2}}t\right]+\frac{c^2}{2}\right\}^2} dt \quad (5)$$

where $(e^{c^2}-1)^{\frac{1}{2}} = \alpha_3/(e^{c^2}+2)$ takes the same sign as α_3.

We know that for the normal distribution 50% of the total frequency lies between the limits $t=-.6745$ and $t=+.6745$. Now, we want to know the similar limits of t for the logarithmic distribution. For that reason, we write t directly in terms of the normally distributed function $x=\frac{1}{c}\log\frac{z-a}{b}$

$$t=\frac{z-m}{\sigma}=\frac{(e^{xc-\frac{c^2}{2}}-1)}{(e^{c^2}-1)^{\frac{1}{2}}}\,. \quad (6)$$

Placing x equal to $-.6745$ and $+.6745$ we have at once the limits

$$t_1=\frac{(e^{-.6745c-\frac{c^2}{2}}-1)}{(e^{c^2}-1)^{\frac{1}{2}}}$$

$$t_2=\frac{(e^{.6745c-\frac{c^2}{2}}-1)}{(e^{c^2}-1)^{\frac{1}{2}}}$$

between which 50% of the total frequency is included. These limits are two quartiles and obviously depend on c. It is clear that one can also locate other deciles and percentiles of t by using (6).

An abstract measure of the dispersion is the coefficient of variability which expresses the standard deviation in terms of the mean. For the logarithmic distribution, it is

$$D=\left|\frac{\sigma}{m-a}\right|=\left|(e^{c^2}-1)^{\frac{1}{2}}\right| \tag{7}$$

which shows that in a logarithmic distribution the larger c^2 is, the greater is the variability.

It is interesting to note that if we also express the deviation of z from the mean in terms of $(m-a)$ and denote it by

$$t'=\frac{z-m}{m-a}=(e^{c^2}-1)^{\frac{1}{2}}t$$

we have by (5) the distribution of t' in this simple form:

$$\frac{1}{\sqrt{2\pi}c(t'+1)}e^{-\frac{1}{2c^2}\left[\log(1+t')+\frac{c^2}{2}\right]^2}dt'. \tag{8}$$

6. *SKEWNESS*

It has been proposed to use $\alpha_3/2$ or α_3 as a measure of skewness of a frequency distribution. For the logarithmic curve, we have shown that

$$\alpha_3=(e^{c^2}-1)^{\frac{1}{2}}(e^{c^2}+2)$$

$$\text{or} \qquad \alpha_3=(\omega-1)^{\frac{1}{2}}(\omega+2). \tag{9}$$

Hence, the absolute value of α_3 increases with c. Since c can take on any finite value whatever, the skewness of the logarithmic curve as measured by α_3 can also have any finite value. Moreover, as we have seen, α_3 of the logarithmic distribution can be positive as well as negative.

In Figure I are shown four logarithmic curves with $m=0$, $\sigma=1$ and with varying α_3's. Various parameters calculated from formulae (4) and important characteristics of these curves are exhibited in Table I.

When $c=0$, α_3 also vanishes. In fact, the logarithmic curve approaches the normal curve of error, as c goes to zero. This can be demonstrated as follows: With the aid of formulae (4) we can write the normally distributed function $x=\frac{1}{c}\log\frac{z-a}{b}$ as

$$x=\frac{1}{c}\log\frac{z-a}{b}$$

$$=\frac{1}{c}\left\{\frac{c^2}{2}+\log\left[1+\frac{z-m}{\sigma}(e^{c^2}-1)^{\frac{1}{2}}\right]\right\}$$

$$=\frac{c}{2}\ \frac{z-m}{\sigma}\ \frac{(e^{c^2}-1)^{\frac{1}{2}}}{c}-\frac{(z-m)^2}{2\sigma^2}\ \frac{(e^{c^2}-1)}{c}+\cdots$$

Now, it can be easily seen that

$$\lim_{c\to 0} x = \frac{z-m}{\sigma}$$

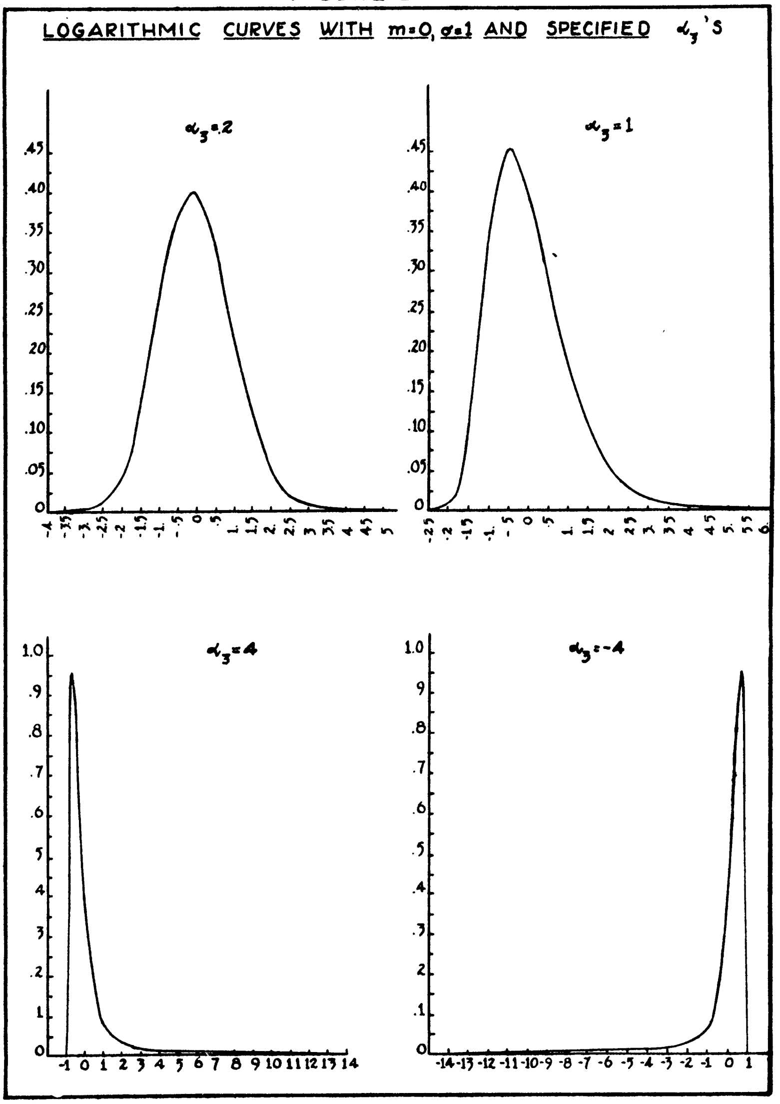
which is a linear function of z. Hence, the logarithmic distribution of z approaches the normal distribution as c approaches zero.

TABLE I

Parameters and Important Characteristics of the Logarithmic Curves with $m=0$, $\sigma=1$ and Specified α_3's

α_3	.2	1	4	-4
ω	1.0044	1.1038	2.0000	2.0000
c	.0665	.3143	.8326	.8326
a	-15.0222	-3.1038	-1.0000	1.0000
b	14.9890	2.9543	.7071	- .7071
$m_g=m_d$	-.0332	-.1495	- .2929	.2929
m_o	- .0991	-.4274	- .6465	.6465
$\check{z}_1$	- 1.09	-1.30	- .9341	.9341
$\check{z}_2$	.90	.50	- .0532	.0532
D	.0666	.3221	1.0000	1.0000
α_4-3	.0712	1.8295	35.0000	35.0000

FIGURE I


LOGARITHMIC CURVES WITH m=0, σ=1 AND SPECIFIED α3's
α3=.2
α3=1
α3=4
α3=-4

Another measure of skewness is defined by Pearson as

$$\chi = \frac{m - m_0}{\sigma} .$$

For the logarithmic curve, it becomes

$$\chi = \frac{1 - \omega^{-\frac{3}{2}}}{(\omega - 1)^{\frac{1}{2}}}$$

which has a maximum value equal to $.6561$, when $\omega = 1.7202$ and $\alpha_3 = 3.1573$. This, however, does not indicate that the skewness of the logarithmic curve is limited. Rather it shows that χ is not a satisfactory measure of skewness, so far as the logarithmic curve is concerned. For any measure of skewness should characterize the skewness of a curve without ambiguity, and χ fails to do so in case of the logarithmic curve. For instance, when we say that a certain logarithmic curve has $\chi = .32$, we may mean either a logarithmic curve with $\alpha_3 = .68$ or one with $\alpha_3 = 36.00$.

When the logarithmic curve is only moderately skew, χ approximately equals $\alpha_3/2$. This can be shown as follows: Letting $h^2 = \omega - 1$, we have

$$\chi = \frac{1-(1+h^2)^{-\frac{3}{2}}}{h} = \frac{3}{2}h - \frac{15}{8}h^3 + \frac{35}{16}h^5 \ldots .$$

and $\alpha_3 = 3h + h^3$.

Hence, for small $|h|$ and hence small $|\alpha_3|$ χ approximately equals $\alpha_3/2$. For instance, when $\alpha_3 = .2$, $\chi = .0991$ which is approximately $\alpha_3/2 = .1$

We may mention here that for the Pearsonian type III curve, the relation $\chi = \alpha_3/2$ always holds. In fact, it appears from Table II that the type III curve and the logarithmic curve are very similar for small $|\alpha_3|$. But the differences between them are already pronounced for $\alpha_3 = 1$, as we can see from Table III.

TABLE II

Ordinates and Areas of the Logarithmic Curve and the Pearsonian Type III Curve

$m = 0 \quad \sigma = 1 \quad \alpha_3 = .2$

z	Ordinate at z		Area from the Lower Limit to z	
	Log. Curve	Type III	Log. Curve	Type III
- 3.5	.0003	.0002	.0000	.0000
- 3.0	.0020	.0020	.0005	.0004
- 2.5	.0124	.0123	.0034	.0034
- 2.0	.0491	.0492	.0172	.0171
- 1.5	.1337	.1341	.0607	.0607
- 1.0	.2587	.2591	.1579	.1582
- .5	.3692	.3687	.3178	.3172
0	.3991	.3986	.5132	.5133
.5	.3366	.3364	.7006	.7002
1.0	.2267	.2267	.8418	.8417
1.5	.1242	1245	.9285	.9284
2.0	.0568	.0567	.9720	.9721
2.5	.0217	.0217	.9906	.9906
3.0	.0072	.0071	.9972	.9973
3.5	.0020	.0020	.9993	.9993
4.0	.0006	.0005	.9998	.9998
4.5	.0002	.0001		

TABLE III

Ordinates and Areas of the Logarithmic Curve and the Pearsonian Type III Curve

$m = 0 \quad \sigma = 1 \quad \alpha_3 = 1$

x	Ordinate at x		Area from the Lower Limit to x	
	Log. Curve	Type III	Log. Curve	Type III
- 2.0	.0084	0	.0009	0
- 1.5	.1196	.1226	.0259	.0190
- 1.0	.3364	.3609	.1398	.1429
- .5	.4498	.4481	.3442	.3528
0	.4040	.3907	.5624	.5665
.5	.2883	.2807	.7363	.7345
1.0	.1791	.1785	.8520	.8488
1.5	.1017	.1043	.9210	.9182
2.0	.0548	.0573	.9590	.9576
2.5	.0295	.0300	.9783	.9788
3.0	.0144	.0151	.9895	.9897
3.5	.0073	.0074	.9948	.9951
4.0	.0036	.0035	.9977	.9977
4.5	.0017	.0017	.9987	.9990
5.0	.0009	.0008	.9993	.9995
5.5	.0004	.0003	.9997	.9998
6.0	.0002	.0002	.9998	.9999
6.5	.0001	.0001	.9999	

7. KURTOSIS

Another important characteristic of a frequency curve is kurtosis measured by $\frac{1}{8}(\alpha_4-3)$ or simply by $\eta=\alpha_4-3$, which equals zero for the normal law of error. If the mean and the standard deviation are taken to be the origin and the unit, respectively, then usually the frequency of a curve in the vicinity of the mean is in excess or in defect to that of a normal curve according to whether η is positive or negative. A curve is said to be platykurtic, if $\eta>0$. It is leptokurtic, if $\eta<0$. Thus, the logarithmic curve is always platykurtic, for its η is

$$\eta=(\omega-1)(\omega^3+3\omega^2+6\omega+6) \tag{10}$$

or

$$\eta=\omega^4+2\omega^3+3\omega^2-6$$

and $\omega>1$ Since the logarithmic curve has only three parameters, there exists a functional relationship between its skewness and kurtosis. This relationship is given through the parameter ω by (9) and (10). We may further deduce the following relations from these two equations:

η is always greater than $\frac{3}{2}\alpha_3^2$. This follows from the fact that $2\eta-3\alpha_3^2=(\omega-1)(2\omega^3+3\omega^2)>0$.

For $|\alpha_3|<6.44$, we have $3\alpha_3^2>\eta$, since $3\alpha_3^2-\eta=(\omega-1)(-\omega^3+6\omega+6)>0$ holds, provided $\omega<2.8$, which corresponds to $|\alpha_3|<6.44$.

For $|\alpha_3|<2.15$, we have $2\alpha_3^2>\eta$, since $2\alpha_3^2-\eta=(\omega-1)(-\omega^3-\omega^2+2\omega+2)>0$ holds provided $\omega<1.4$, which corresponds to $|\alpha_3|<2.15$.

Since practically the value of $|\alpha_3|$ can hardly reach 6.44 or even 2.15, the relations just stated hold for all practical instances.

The relationship existing between η and α_3 is sometimes used as a criterion for applying the logarithmic curve to observed data. We shall discuss this point in section 9.

8. POWERS, ROOTS AND PRODUCTS OF THE LOGARITHMICALLY DISTRIBUTED VARIABLES

If z is logarithmically distributed and has 'a' as its lower limit, $W=(z-a)^k$ is also so distributed, k being any constant. This follows from the fact that if $x=\frac{1}{c}\log\frac{z-a}{b}$ is normally distributed, so is $kx=\frac{k}{c}\log\frac{z-a}{b}$. From the frequency function of z, $F(z)$, given by (1), we find at once the analytic expression of the frequency distribution of W to be

$$\frac{1}{\sqrt{2\pi}\,ckW}\,e^{-\frac{1}{2c^2k^2}\left[\log\frac{W}{b^k}\right]^2}dW. \tag{11}$$

We have learned from the preceding sections that a logarithmic distribution represented by (1) with larger c has greater variability, skewness, and kurtosis. Thus, if $k^2>1$, the variability, skewness, and kurtosis are greater for W than for z. On the other hand, if $k^2<1$, the distribution of z' has greater variability, skewness, and kurtosis.

If the logarithmically distributed variables $z_1, z_2, \cdots, z_n$ are independent and have for their lower limits, $a_1, a_2, \cdots, a_n$, then the product

$$Y=(z_1-a_1)(z_2-a_2)\cdots(z_n-a_n)$$

is also so distributed. This follows from the fact that if

$$x_1=\frac{1}{c_1}\log_1\frac{z_1-a_1}{b_1},\ x_2=\frac{1}{c_2}\log\frac{z_2-a_2}{b_2},\ldots,x_n=\frac{1}{c_n}\log\frac{z_n-a_n}{b_n}$$

are each normally distributed and are independent, their sum also obeys the normal law of error.

Since the variables are independent, the frequency distribution of these n variables is represented by

$$F_1(z_1)F_2(z_2)\cdots F_n(z_n)\,dz_1\,dz_2\cdots dz_n \qquad (12)$$

where $F_i(z_i) = \dfrac{1}{\sqrt{2\pi}\,c_i(z_i - a_i)}\, e^{-\frac{1}{2c_i^2}\left[\log\frac{z_i - a_i}{b_i}\right]^2}$.

Substituting $z_1 - a_1 = Y/(z_2 - a_2)\cdots(z_n - a_n)$ in (12) and integrating the resulting expression with respect to $z_2, \cdots, z_n$ successively over the respective ranges, we have the distribution of Y as

$$\frac{1}{\sqrt{2\pi}\sqrt{c_1^2+c_2^2+\cdots+c_n^2}\,Y}\, e^{-\frac{1}{2(c_1^2+c_2^2+\cdots+c_n^2)}\left[\log\frac{Y}{b_1 b_2\cdots b_n}\right]^2} dY. \qquad (13)$$

Since the sum, $c_1^2 + c_2^2 + \cdots + c_n^2$, is greater than any c_i^2, the distribution of Y has greater variability, skewness, and kurtosis than that of each individual variable.

9. *NUMERICAL APPLICATIONS*

Many methods of fitting a logarithmic frequency curve to observed data have been proposed. But only the method of moments will be considered below.*

The method of moments is very simple to apply. It consists of placing the computed moments in equations (2) and then determining the parameters by solving these equations by formulae (3) and (4).† The only step of computation which requires some time and care to obtain accurate results is the solution of

* Among other methods of graduating the logarithmic frequency distribution, the graphical method proposed by Kapteyn and Van Uven is especially useful. For a description of this method, refer to their paper on "Skew Frequency Curves in Biology and Statistics, 2nd Paper".

† In his paper, "On the Genetic Theory of Frequency", Wicksell also showed the application of the method of moments to the logarithmic frequency distribution. However, he found the parameter "a" first and then proceeded to obtain "log b" and "c".

the cubic,

$$\psi(\omega) = \omega^3 + 2\omega^2 - (\alpha_3^2 + 4) = 0.$$

Hence, it is desirable to have a table which will provide an approximation of the required root of this cubic for a given α_3 Then, the root can be approximated to as great a degree of accuracy as we wish by applying, for instance, Newton's method. That is why Table IV is constructed. Practically, after we obtain an approximate value of ω from Table IV, one single application of Newton's method will almost invariably suffice to give us a value of ω accurate to four decimal places. In Table IV, values of c corresponding to given values of ω are also provided to serve as a check to our computation of c by formulae (4).

TABLE IV

Table Facilitating the Solution of the Cubic

$$\omega^3 + 3\omega^2 - (\alpha_3^2 + 4) = 0$$

ω	α_3	c	ω	α_3	c
1.	0	0			
1.01	.3010	.1000	1.26	1.6623	.4807
1.02	.4271	.1407	1.27	1.6991	.4889
1.03	.5248	.1720	1.28	1.7356	.4969
1.04	.6080	.1980	1.29	1.7717	.5046
1.05	.6820	.2209	1.30	1.8075-	.5122
1.06	.7495+	.2415-	1.31	1.8429	.5196
1.07	.8122	.2602	1.32	1.8781	.5269
1.08	.8712	.2775-	1.33	1.9129	.5340
1.09	.9270	.2936	1.34	1.9475+	.5410
1.10	.9803	.3087	1.35	1.9819	.5478
1.11	1.0315-	.3231	1.36	2.0160	.5545+
1.12	1.0808	.3366	1.37	2.0499	.5611
1.13	1.1285+	.3496	1.38	2.0836	.5675+
1.14	1.1749	.3619	1.39	2.1171	.5738
1.15	1.2200	.3739	1.40	2.1503	.5801
1.16	1.2640	.3852	1.41	2.1835-	.5862
1.17	1.3070	.3962	1.42	2.2164	.5922
1.18	1.3492	.4068	1.43	2.2492	.5981
1.19	1.3905-	.4171	1.44	2.2818	.6038
1.20	1.4311	.4270	1.45	2.3143	.6096
1.21	1.4710	.4366	1.46	2.3467	.6151
1.22	1.5103	.4460	1.47	2.3789	.6207
1.23	1.5491	.4550	1.48	2.4110	.6261
1.24	1.587	.4638	1.49	2.4430	.6315+
1.25	1.6250	.4723	1.50	2.4749	.6368

To illustrate the use of Table IV and to help in studying the application of the logarithmic frequency curve, we take the distribution of the weights of 1,000 female students from the "Synopsis of Elementary Mathematical Statistics"* by Miss B. L. Shook. (See Table V.)

The mean, standard deviation, and skewness for this distribution† are

$$m = 118.74 \text{ lbs.}$$
$$\sigma = 16.91752 \text{ lbs.}$$
$$\alpha_3 = .976424$$

To compute ω, we find from Table IV that for $\alpha_3 = .976424$ ω is approximately $\omega_0 = 1.10$. For a better approximation, we apply Newton's method:

$$\omega = \omega_0 - \frac{\psi(\omega_0)}{\psi'(\omega_0)} = \omega_0 - \frac{\omega_0^3 + 3\omega_0^2 - (\alpha_3^2 + 4)}{3\omega_0^2 + 6\omega_0}$$

$$= 1.10 - \frac{.007596}{10.23} = 1.10 - .000743$$

$$= 1.099\,257$$

By formulae (4), the parameters c, b and a are found to be

$$c = .307\,627$$
$$b = 51.2160 \text{ lbs.}$$
$$a = 65.0423 \text{ lbs.}$$

* Annals of Mathematical Statistics, Vol. I, No. 1 (1930), p. 39.
† Sheppard's corrections have been duly applied.

TABLE V

Observed and Theoretical Distributions of the Weights of 1,000 Female Students
(Original Measurements Made to Nearest 1/10 lb.)

Class Limits (Pounds)	Observed Frequency	Theoretical Logarithmic Distribution		Theoretical Type III Distribution
		By Areas	By Ordinates	By Areas
70- 79.9	2	0	0	0
80- 89.9	16	10	6	4
90- 99.9	82	97	94	102
100-109.9	231	228	234	238
110-119.9	248	255	259	250
120-129.9	196	190	190	184
130-139.9	122	114	111	111
140-149.9	63	57	57	59
150-159.9	23	27	27	29
160-169.9	5	12	12	13
170-179.9	7	6	6	6
180-189.9	1	2	2	3
190-199.9	2	1	1	1
200-209.9	1	1	1	0
210-219-9	1	0	0	0
Total	1,000	1,000	1,000	1,000

Knowing c , b and a , we obtain the geometric mean and the mode:

$$m_g = m_d = 116.2583 \text{ lbs.}$$
$$m_0 = 111.6326 \text{ lbs.}$$

Using these parameters, the theoretical distribution of the weights of 1,000 female students has been computed and is shown in Table V and Figure II. The fit of the logarithmic distribution to the observed data is, indeed, excellent.* The lowest possible weight of female students, according to the theoretical distribution, is 65.04 pounds, which is just about what one would expect after examining the observed data.

Miss Shook† used the type III distribution to fit the same set of observed data and gave the result as shown in the last col-

* Grouping the first three classes into one class and the last six classes into one class, we apply the χ^2 test for goodness of fit and find that the probability to get a worse fit is .70.

† Annals of Mathematical Statistics, Vol. I, No. 3 (1930), p. 242.

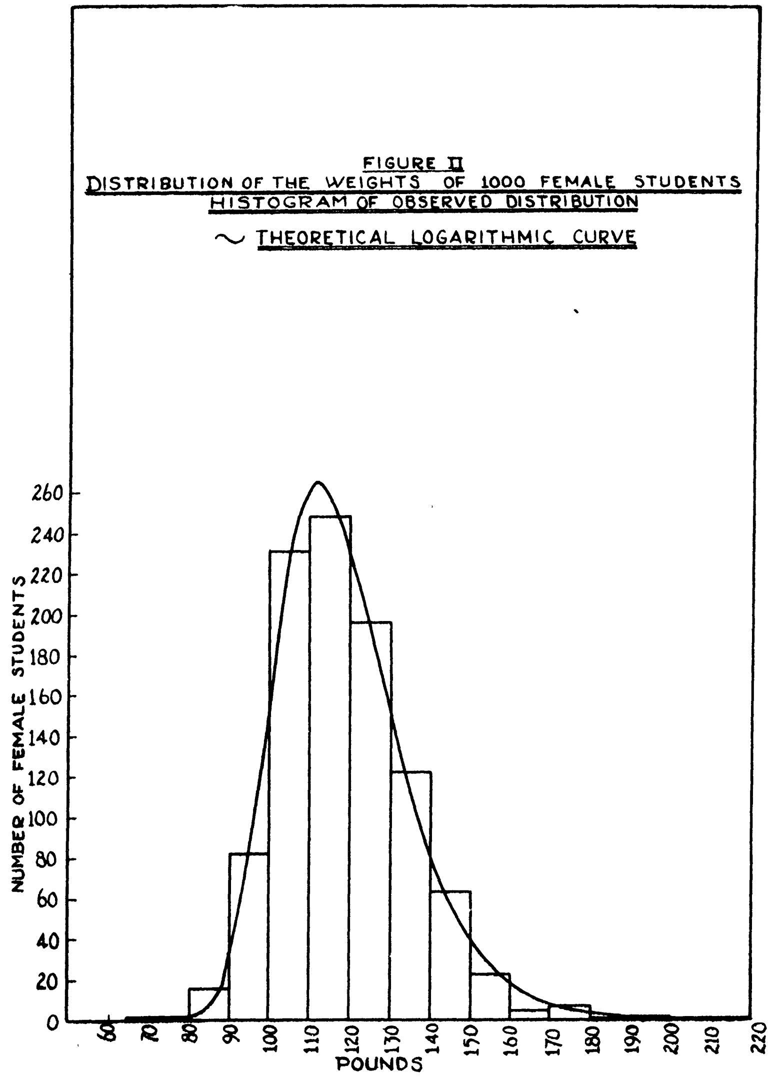


FIGURE II
DISTRIBUTION OF THE WEIGHTS OF 1000 FEMALE STUDENTS
HISTOGRAM OF OBSERVED DISTRIBUTION
THEORETICAL LOGARITHMIC CURVE
NUMBER OF FEMALE STUDENTS
0
20
40
60
80
100
120
140
160
180
200
220
240
260
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
POUNDS

umn of Table V. The fit is not as good as that given by the logarithmic distribution, especially in view of the fact that the type III curve fixes the least possible weight at 84.09 pounds, while as a matter of fact there are two students whose weights are below that limit.‡

From the standpoint of the method of moments, a criterion for the logarithmic distribution to fit a set of observed data is that $\eta = \alpha_4 - 3$ computed directly from the observed data must be approximately the same as the theoretical η computed from formula (10). This criterion, however, does not seem to work in practice. For instance, for the distribution of the weights of 1,000 female students, the theoretical η is 1.7419, while the observed η is 2.4536. But in spite of this fact, the observed distribution, as we have seen, is very satisfactorily fitted by a logarithmic distribution.

Another criterion is to require the observed moments about the lower limit "a" to satisfy approximately the recurring relation

$$\mu'_s = be^{\frac{2s-1}{2}c^2}\mu'_{s-1}$$

for $s = 4$. This criterion is approximately fulfilled by the distribution of the weights of 1,000 female students, for which we have

$$\mu'_4 = 147279 \cdot 10^2$$

$$be^{\frac{7}{2}c^2}\mu'_3 = 146696 \cdot 10^2$$

and

$$\mu'_4 / be^{\frac{7}{2}c^2}\mu'_3 = 1.0040.$$

The fact that a set of observed data may be satisfactorily graduated by the logarithmic distribution but fulfills only the second criterion may be explained on the ground that the com-

‡In fact, since the finite limit of the variable for type III curve is $m - \frac{2}{\alpha_3}\sigma$ and for the logarithmic curves $m - \frac{2+\omega}{\alpha_3}\sigma$, the finite limit is always greater in absolute value for the logarithmic curve than for the type III curve.

paratively wide discrepancy between the observed and theoretical frequency in the classes near the lower limit maks a great difference in the fourth moment about the mean but does not make much difference in the fourth moment about the point "a".

PART II

THE SEMI-LOGARITHMIC CORRELATION SURFACE

Suppose that the correlation surface of the functions, $x = f(u, v)$ and $y = g(u, v)$, is a normal correlation surface and each has its mean as the origin and its standard deviation as the unit. Then, the probability that values of x will lie between x and $x+dx$ and values of y between y and $y+dy$ is

$$\phi(x,y)dxdy = \frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left[x^2-2rxy+y^2\right]} dxdy. \quad (1)$$

It follows that the probability that values of u will lie between u and $u+du$ and values of v between v and $v+dv$ is $F(u,v)\, dudv$ given by

$$\frac{1}{2\pi\sqrt{1-r^2}} e^{-\frac{1}{2(1-r^2)}\left[f^2-2rfg+g^2\right]} \begin{vmatrix} \frac{\partial f}{\partial u} & \frac{\partial f}{\partial v} \\ \frac{\partial g}{\partial u} & \frac{\partial g}{\partial v} \end{vmatrix} du\, dv. \quad (2)$$

$F(u, v)$ is, therefore, a generalized correlation surface of two variables, deduced by extending the method of translation for treating frequency distributions of one variable.

It is clear that in this general form the correlation surface represented by $F(u, v)$ is of little practical use, on account of its complexity. Now a natural simplification suggests itself. That is to take x as a function of u only and y as a function of v

only. By virtue of this simplification, $F(u, v)$ becomes

$$F(u,v)=\frac{1}{2\pi\sqrt{1-r^2}}\,e^{-\frac{1}{2(1-r^2)}\left[f^2-2rfg+g^2\right]}\frac{df}{du}\frac{dg}{dv} \tag{3}$$

which is a great deal easier to handle than before.

Professor Wicksell has made use of (3) for the special case where, in our notations, x and y are

$$x=\frac{1}{c_1}\log\frac{u-a_1}{b_1}$$

$$y=\frac{1}{c_2}\log\frac{v-a_2}{b_2}$$

which leads to the so-called "logarithmic correlation surface".* The surface possesses the property that its marginal distributions as well as the distributions of u for given values of v and distributions of v for given values of u are all logarithmic frequency distributions.

Presently we shall study another case for which

$$x=\frac{u-\gamma}{\lambda}$$

$$y=\frac{1}{c}\log\frac{v-a}{b}\,.$$

The correlation surface $F(u,v)$ given by (3) then becomes:

$$F(u,v)=\frac{e^{-\frac{1}{2(1-r^2)}\left[\left(\frac{u-\gamma}{\lambda}\right)^2-2r\frac{u-\gamma}{\lambda c}\log\frac{v-a}{b}+\left(\frac{1}{c}\log\frac{v-a}{b}\right)^2\right]}}{2\pi\lambda c(v-a)\sqrt{1-r^2}} \tag{4}$$

* In Wicksell's paper, "On the Genetic Theory of Frequency", the theory of the logarithmic correlation function is developed. In his two successive papers quoted in the Bibliography of this paper, the original theory is extended and the application of the extended results illustrated.

which may be appropriately called a semi-logarithmic correlation surface. We shall investigate its marginal distributions, moments and regression curves of the characteristics.

1. *MARGINAL DISTRIBUTIONS*

Now, we shall first find the distribution of the marginal totals of u . This can be, of course, accomplished very easily by integrating $F(u,v)$ with respect to v over the range from a to infinity. The result is:

$$\int_a^\infty F(u,v)dv = \frac{1}{\sqrt{2\pi}\lambda} e^{-\frac{1}{2\lambda^2}(u-\gamma)^2} \tag{5}$$

Thus, the marginal distribution of u obeys the normal laws of error.

Similarly, if we integrate $F(u,v)$ with respect to u over the range from $-\infty$ to ∞ , we find at once the marginal distribution of v as follows:

$$\int_{-\infty}^\infty F(u,v)du = \frac{1}{\sqrt{2\pi}\,c(v-a)} e^{-\frac{1}{2c^2}\left(\log\frac{v-a}{b}\right)^2} \tag{6}$$

which is, clearly a logarithmic distribution and, therefore, has all the properties and characteristics discussed in Part I. Hence, the semi-logarithmic correlation surface is characterized by the fact that one marginal distribution is normal, while the other is logarithmic. It is needless to mention that this does not constitute a sufficient condition for a correlation surface to be a semi-logarithmic correlation surface defined by (4).

2. *MOMENTS*

The moment, μ'_{ij} , of the semi-logarithmic correlation surface about the point $u=\gamma$ and $v=a$ is given by

$$\begin{aligned}\mu'_{ij} &= \int_{-\infty}^\infty\int_a^\infty (u-\gamma)^i(v-a)^j F(u,v)\,du\,dv \\ &= \lambda^i b^j e^{\frac{j^2c^2}{2}} \sum_{k=0}^{i} \binom{i}{k}(jcr)^{i-k} \int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}} t^k e^{-\frac{t^2}{2}}dt\end{aligned} \tag{7}$$

where $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} t^k e^{-\frac{t^2}{2}} dt = \frac{k!}{2^{k/2}\left(\frac{k}{2}\right)!}$, **if k is even**

$= 0,$ if k is odd.

Using relation (7), we can easily calculate the following six moments about the mean of u, m_u, and the mean of v, m_v:

$$\begin{aligned}
\mu_{10} &= m_u - \gamma = 0 \\
\mu_{20} &= \lambda^2 \\
\mu_{01} &= m_v - (be^{\frac{c^2}{2}} + a) = 0 \\
\mu_{02} &= b^2 e^{c^2}(e^{c^2} - 1) \\
\mu_{03} &= b^3 e^{\frac{3}{2}c^2}(e^{c^2} - 1)^2 (e^{c^2} + 2) \\
\mu_{11} &= r\lambda c b e^{\frac{c^2}{2}} .
\end{aligned} \tag{8}$$

Now, we want to solve these equations for the six parameters. As before, we let $\omega = e^{c^2}$ and write $\alpha_{03} = \mu_{03}/\mu_{02}^{\frac{3}{2}}$. Again, we have ω as the only real root of the cubic:

$$\omega^3 + 3^2 - (\alpha_{03}^2 + 4) = 0 \; . \tag{9}$$

The six parameters of the semi-logarithmic correlation surface can be written as:

$$\begin{aligned}
\gamma &= m_u \\
\lambda &= \sqrt{\mu_{20}} = \sigma_u \\
c &= (\log \omega)^{\frac{1}{2}} \\
b &= \left(\frac{1}{\omega}\right)^{\frac{1}{2}} \left(\frac{1}{\omega - 1}\right)^{\frac{1}{2}} \sigma_v = \left(\frac{1}{\omega}\right)^{\frac{1}{2}} \left(\frac{\omega + 2}{\alpha_{03}}\right) \sigma_v \\
a &= m_v - \left(\frac{1}{\omega - 1}\right)^{\frac{1}{2}} \sigma_v = m_v - \left(\frac{\omega + 2}{\alpha_{03}}\right) \sigma_v \\
r &= \frac{\mu_{11} \quad (\omega - 1)^{\frac{1}{2}}}{\sigma_u \, \sigma_v \, (\log \omega)^{\frac{1}{2}}}
\end{aligned} \tag{10}$$

which furnish us a simple practical method for determining the parameters of the semi-logarithmic correlation surface for observed data.

3. *REGRESSION OF THE MEAN*

First, let us observe that the function $F(u,v)$ may be put into the following forms:

$$F(u,v)=\frac{1}{2\pi\sqrt{1-r^2}\,c(v-a)\lambda}\,e^{-\frac{1}{2c^2}\left(\log\frac{v-a}{b}\right)^2}\,e^{-\frac{1}{2(1-r^2)}\left[\frac{u-\gamma}{\lambda}-\frac{r}{c}\log\frac{v-a}{b}\right]^2}$$

$$=\frac{1}{2\pi\sqrt{1-r^2}\,c(v-a)\lambda}\,e^{-\frac{1}{2\lambda^2}(u-\gamma)^2}\,e^{-\frac{1}{2(1-r^2)}\left[\frac{1}{c}\log\frac{v-a}{b}-r\frac{u-\gamma}{\lambda}\right]^2}$$

Hence, the distribution of u for a particular array of v is normal:

$$\theta_1(u,v)=\frac{1}{\sqrt{2\pi}\,\lambda\sqrt{1-r^2}}\,e^{-\frac{1}{2(1-r^2)}\left[\frac{u-\gamma}{\lambda}-\frac{r}{c}\log\frac{v-a}{b}\right]^2} \tag{11}$$

and the distribution of v for a particular array of u is logarithmic:

$$\theta_2(u,v)=\frac{1}{\sqrt{2\pi}\sqrt{1-r^2}\,c(v-a)}\,e^{-\frac{1}{2(1-r^2)}\left[\frac{1}{c}\log\frac{v-a}{b}-r\frac{u-\gamma}{\lambda}\right]^2}$$

$$=\frac{1}{\sqrt{2\pi}\sqrt{1-r^2}\,c(v-a)}\,e^{-\frac{1}{2c^2(1-r^2)}\left[\log\frac{v-a}{be^{cr\frac{u-\gamma}{\lambda}}}\right]^2}. \tag{12}$$

To find the mean of u for a particular value of v, we multiply $\theta_1(u,v)$ by u and integrate the resulting expression with respect to u over the range from $-\infty$ to ∞.

$$\bar{u}=\int_{-\infty}^{\infty}u\,\theta_1(u,v)\,du=\frac{\lambda r}{c}\log\frac{v-a}{b}+\gamma \tag{13}$$

which is the regression equation of the mean of u on v and may be called the logarithmic regression equation.

Similarly, the regression equation of the mean of v on u is found to be:

$$\bar{v} = \int_a^\infty v\theta_2(u,v)dv = be^{cr\frac{u-\gamma}{\lambda} + \frac{c^2}{2}(1-r^2)} + a \tag{14}$$

which may be named exponential regression equation.

Observe the following points:

(a) The regression curves (13) and (14) intersect at the point

$$u = \gamma + \frac{\lambda cr}{2}$$

$$v = be^{\frac{c^2}{2}} + a = m_v .$$

(b) When $r=0$, the curves become two straight lines:

$$\bar{u} = \gamma = m_u$$

$$\bar{v} = be^{\frac{c^2}{2}} + a = m_v$$

which show that $\bar{u}$ is independent of v and $\bar{v}$ is independent of u. We can also see this from the expression $F(u,v)$, which becomes

$$F(u,v) = \frac{1}{\sqrt{2\pi}\lambda} e^{-\frac{1}{2}\left(\frac{u-\gamma}{\lambda}\right)^2} \frac{1}{\sqrt{2\pi}c(v-a)} e^{-\frac{1}{2}\left(\frac{1}{c}\log\frac{v-a}{b}\right)^2}$$

when $r=0$. This is the condition for independence of u and v in a probability sense.

(c) When $r=1$, these two regression curves coincide. This signifies that there exists a complete functional relationship between u and v, namely:

$$\frac{u-\gamma}{\lambda} = \frac{1}{c}\log\frac{v-a}{b} .$$

(d) As we have learned from the studies on the normal correlation surface, r is the coefficient of correlation measuring the linear relationship between $x = \frac{u-\gamma}{\lambda}$ and $y = \frac{1}{c}\log(v-a)$. Thus, it is also a measure of relationships (13) and (14) existing between u and v. If we note that r may be written as

$$r = \frac{\mu_{11}}{\sigma_u \sigma_v} \frac{(e^{c^2}-1)^{\frac{1}{2}}}{c}$$

$$= \frac{\mu_{11}}{\sigma_u \sigma_v} \left(1+\frac{c^2}{2!}+\frac{c^4}{3!}+\frac{c^6}{4!}+\dots\right)^{\frac{1}{2}} \tag{15}$$

we see that r is always greater than $\mu_{11}/\sigma_u \sigma_v$, which would be the coefficient of correlation measuring the linear relationship between u and v, if we treated the correlation surface of u and v as being normal.

The smaller the value of c and α_{03}, the smaller the difference between r and $\mu_{11}/\sigma_u \sigma_v$. In fact, we can show, as we did for one variable case, that as c goes to zero the semi-logarithmic correlation surface approaches the normal correlation surface.

Incidentally, we may remark that the expression (15) is convenient for computing r.

4. *REGRESSION OF THE MOMENTS*

Using the well-known formulae for the moments of the normal curve of error about the mean, we can find at once the s-th moment of $\theta_1(u,v)$ about its mean:

$$M_{s \cdot u} = \int_{-\infty}^{\infty} (u-\bar{u})^s \theta_1(u,v)\, du \tag{16}$$

$$= \frac{s!}{2^{s/2}\left(\frac{s}{2}\right)!} \lambda^s (1-r^2)^{s/2}, \qquad \text{if } s \text{ is even}$$

$$= 0, \qquad \text{if } s \text{ is odd.}$$

This is the regression equation of the s-th moment of u about the mean on v. It follows that the s-th standard moment of u for a given value of v is:

$$\alpha_{s:u} = \frac{M_{s:u}}{M_{2:u}^{s/2}}$$

$$= \frac{s!}{2^{s/2}\left(\frac{s}{2}\right)!}, \qquad \text{if } s \text{ is even} \qquad (17)$$

$$= 0, \qquad \text{if } s \text{ is odd.}$$

Again, by the formulae given in Part I for the moments of the logarithmic distribution, we calculate the s^{th} moment of $\theta_2(u,v)$ about the point "a":

$$M'_{s:v} = \int_a^{\infty} (v-a)^s \theta_2(u,v)\,dv$$
$$= b^s e^{scr\frac{u-\gamma}{\lambda} + \frac{s^2c^2(1-r^2)}{2}} \ . \qquad (18)$$

And the regression equation of the s^{th} moment of v about the mean on u is:

$$M_{s:v} = \int_a^{\infty} (v-\bar{v})^s \theta_2(u,v)\,dv$$

$$= b^s e^{scr\frac{u-\gamma}{\lambda} + \frac{sc^2(1-r^2)}{2}} \sum_{k=0}^{s} (-1)^{s-k} \binom{s}{k} e^{\frac{k(k-1)c^2(1-r^2)}{2}} \ . \qquad (19)$$

The s^{th} standard moment of v for a particular value of u is, therefore,

$$\alpha_{s:v} = \frac{M_{s:v}}{M_{2:v}^{s/2}}$$
$$= \frac{\sum_{k=0}^{s} (-1)^{s-k}\binom{s}{k} e^{\frac{k(k-1)c^2(1-r^2)}{2}}}{\left(e^{c^2(1-r^2)} - 1\right)^{s/2}} \ . \qquad (20)$$

Having obtained the expressions for the regressions of the moments of one variable on the other, we shall now proceed to discuss the scedasticity, clisy and synagic* of the semi-logarithmic correlation surface.

5. *SCEDASTICITY*

From formula (16), we have the regression of the second moment of u about the mean on v:

$$M_{2:u} = \lambda^2(1-r^2) \tag{21}$$

which is the same as in the case of the normal correlation surface, except that r now does not measure the linear relationship between u and v. Since (21) is free of v, the semi-logarithmic correlation surface is homoscedastic, so far as the variable u is concerned.

From the standpoint of estimation, we may also interpret expression (21) to mean that when we estimate the mean value of u for a particular value of v, the error of estimation will be reduced if we use formula (13) instead of the mean of the marginal distribution of u. The standard deviation of the marginal distribution of u is λ, while that of (13) is only $\sqrt{M_{2:u}} = \lambda\sqrt{1-r^2}$ as shown by (21).

The second moment of v for a particular value of u is given by (19):

$$M_{2:v} = b^2 e^{2cr\frac{u-t}{\lambda} + c^2(1-r^2)}\left[e^{c^2(1-r^2)} - 1\right] \tag{22}$$

which is not independent of u. So, the semi-logarithmic correlation surface is not homoscedastic for v. Actually $\sqrt{M_{2:v}}$, the standard deviation of the distribution of v for a given u, increases with u.

However, the relative dispersion or relative error for the

* The term "synagic" was used by S. D. Wicksell to mean the regression of the kurtosis. ("The Correlation Function of Type A, and the Regression of its Characteristics", Kungl. Svenska Vetenskapsakademiens Handlingar, Band 58, Nr. 3; Meddelanden fran Lunds Observatorium, Ser. II, Nr. 17, 1917)

distributions of v for different values of u is a constant, namely:

$$D_v = \frac{M}{\bar{v}-a}$$

$$= \left[e^{c^2(1-r^2)} - 1\right]^{\frac{1}{2}} . \tag{23}$$

Thus, by using formula (14) **to estimate the mean value** of v for a given value of u instead of employing the mean of the marginal distribution of v, we reduce the relative error of estimation, for the relative error of the marginal distribution is $(e^{c^2}-1)^{\frac{1}{2}}$. The reduction of relative error is much pronounced when r is large. In fact, the greater r is, the greater the reduction of relative error and the better the estimation. Hence, r measures the degree of relationships (13) and (14) between u and v.

6. CLISY AND SYNAGIC

Now, we shall study the clisy and synagic of the semi-logarithmic correlation surface or the regression of the skewness and kurtosis of one variable on the other.

The skewness and kurtosis of any distribution represented by $\theta_1(u,v)$ as measured by $\alpha_{3:4}$ and $\eta_u = \alpha_{4:u} - 3$ are, of course, equal to zero, since it is a normal distribution. But the skewness and kurtosis of any distribution of v for particular values of u, according to formula (20), are given by:

$$\alpha_{3:v} = (e^{c^2(1-r^2)} - 1)^{\frac{1}{2}}(e^{c^2(1-r^2)} + 2) \tag{24}$$

$$\eta_v = (e^{c^2(1-r^2)} - 1)(e^{3c^2(1-r^2)} + 3e^{2c^2(1-r^2)} + 6e^{c^2(1-r^2)} + 6) \tag{25}$$

which are two constants. Since the skewness and kurtosis of the marginal distribution of v are given by $(e^{c^2}-1)^{\frac{1}{2}}(e^{c^2}+2)$ and $(e^{c^2}-1)(e^{3c^2}+3e^{2c^2}+6e^{c^2}+6)$, respectively, we may say that the distribution of v for each array of u has smaller skewness and kurtosis, and is, therefore, closer to the normal distribution than the marginal distribution of v. And it is more so, when r is near unity.

7. *REGRESSION OF OTHER CHARACTERISTICS*

In this section, we shall give the regression of other characteristics, such as the median, the geometric mean, the mode, the points of inflection and the finite limit.

The regression equation of the median and the mode of u on v are, of course, the same as that of the mean of u on v, because $\theta_1(u,v)$ is normal. The points of inflection of $\theta_1(u,v)$ are points one standard deviation, i.e., $\sqrt{M_{2:u}}$, to the left and the right of the mean, as this is again a well-known property of the normal distribution.

The regression equation of the median and the geometric mean of v on u is given by

$$m_{d:v} = m_{g:v} = be^{cr\frac{u-\gamma}{\lambda}} + a$$

$$\text{or} \quad \frac{1}{c}\log\left(\frac{m_{d:v}-a}{b}\right) = r\frac{u-\gamma}{\lambda} \tag{26}$$

which differs from the regression equation of the mean or the median of u on v, only in that the constant factor r is on the left member of equation (13) but is on the right member of (26).

The mode of v for special values of u is

$$m_{o:v} = be^{cr\left(\frac{u-\gamma}{\lambda}\right)-c^2(1-r^2)} + a. \tag{27}$$

The regression equations of the points of inflection of v on u are given by

$$\breve{z}_{1:v} = be^{cr\frac{u-\gamma}{\lambda}-\frac{3}{2}c^2(1-r^2)\left[1+\sqrt{1+\frac{4}{9c^2(1-r^2)}}\right]} + a$$

$$\breve{z}_{2:v} = be^{cr\frac{u-\gamma}{\lambda}-\frac{3}{2}c^2(1-r^2)\left[1-\sqrt{1+\frac{4}{9c^2(1-r^2)}}\right]} + a$$

which are not free of u.

Finally, we may add that the finite limit of any distribution of v for a particular array of u is the same as that of the marginal distribution of v.

8. *AN ILLUSTRATION*

For illustrating the application of the semi-logarithmic correlation surface, we take the correlation table of heights and weights of 11,382 school boys between 5 and 14 years of age in Glasgow from L. Isserlis's paper, "On the Partial Correlation Ratio".* We shall treat the height as the variable u and the weight as the variable v. Thus, the marginal distribution of heights is supposed to be normal, while that of weights is supposed to be logarithmic.

Letting the class marks, 49 inches and 56 pounds, be the origins of u and v, respectively, and the class intervals be the respective units, we calculate the moments of this correlation surface:**

$m_u = -.511861$ class intervals
$\sigma_u = 1.7631$ class intervals
$\alpha_{30} = .0177$
$\alpha_{40} = 2.5093$
$m_v = -.205412$ class intervals
$\sigma_v = 2.5781$ class intervals
$\alpha_{03} = .5915$
$\alpha_{04} = 3.1221$
$\mu_{11} = 4.205875$

from which we deduce the following parameters by formulae (10):

$\gamma = -.511861$ class intervals
$\lambda = 1.7631$ class intervals
$\omega = 1.0379$
$c = .1929$
$a = -13.45$ class intervals
$b = 13.00$ class intervals
$r = .9340$

* Biometrika. Vol. XI, 1915, p. 65.

** Note that these numerical results differ somewhat from those given by L. Isserlis, because we have applied Sheppard's corrections to the raw moments.

TABLE VI

Correlation Table of Heights and Weights of 11,382 School Boys between 5 and 14 Years of Age in Glasgow

(Original Measurements of Heights Made to Nearest Inch; Original Measurements of Weights Made to Nearest Pound)

Weight (Pounds) Class Limit	Height (Inches) 30-32	33-35	36-38	39-41	42-44	45-47	48-50	51-53	54-56	57-59	60-62	63-65	Total
24- 28	4	9	2			1							16
29- 33	3	42	62	25	3	1							136
34- 38		16	220	414	72	6							728
39- 43	1	3	51	617	697	95	11	1					1476
44- 48		1	7	122	875	603	38	8	1				1655
49- 53			4	12	249	988	411	33	5	4			1706
54- 58		1	3	1	17	436	905	171	11	4	3		1552
59- 63			1		1	39	630	568	51	6	1		1297
64- 68				1		8	161	621	206	3	2	2	1004
69- 73				1			35	374	340	24	2		776
74- 78							3	106	335	76	5		525
79- 83							2	22	120	93	4	1	242
84- 88						1		8	32	87	8	2	138
89- 93								1	10	36	18	1	66
94- 98									3	23	9	2	37
99-103										5	11	3	19
104-108									1		5	1	7
109-113											1		1
114-118													0
119-123												1	1
Total	8	72	350	1193	1914	2178	2196	1913	1115	361	69	13	11,382

With these parameters, the correlation surface of heights and weights is determined. Now, we shall examine the regression curves of this correlation surface.

Inserting the computed parameters in formulae (13) and (14), we obtain the regression equations of the mean height on weight and the mean weight on height. In Tables VII and VIII, we have the mean heights for specified weights and the mean weights for specified heights. We see, from these tables and from figures III and IV, the agreement the theoretical and observed results is very excellent. In some extreme classes the deviations of the observed values from the theoretical values are more pronounced. But these classes comprise only a small fraction of the total number of cases.

Now, we go further to investigate the scedasticity of the correlation surface of heights and weights. According to the theory, for any particular weight the standard deviation of heights should be a constant and equal to $\sigma(1-r^2)^{\frac{1}{2}} = 1.8893$ inches. This is much less than the standard deviation of the marginal distribution of the heights, which is 5.2893 inches. That 1.8893 inches is quite close to the observed standard deviations is shown by Table IX and Figure V.

The theory asserts that the dispersion of weights is not the same for different heights. But for all arrays of heights the relative dispersion or relative error of weights is independent of heights.

TABLE VII

The Mean Heights for Specified Weights

Weight (Pounds)	Mean Height (Inches) Observed	Theoretical
24- 28	34.4	33.2
29- 33	36.5	36.4
34- 38	39.3	39.3
39- 43	41.8	41.9
44- 48	44.0	44.2
49- 53	46.4	46.4
54- 58	48.5	48.3
59- 63	50.5	50.2
64- 68	52.1	51.9
69- 73	53.2	53.5
74- 78	54.9	55.0
79- 83	56.0	56.4
84- 88	57.1	57.8
89- 93	58.4	59.1
94- 98	58.8	60.3
99-103	60.7	61.5
104-108	60.6	62.6
109-113	61.0	63.6
114-118	....	64.7
119-123	63.0	65.7

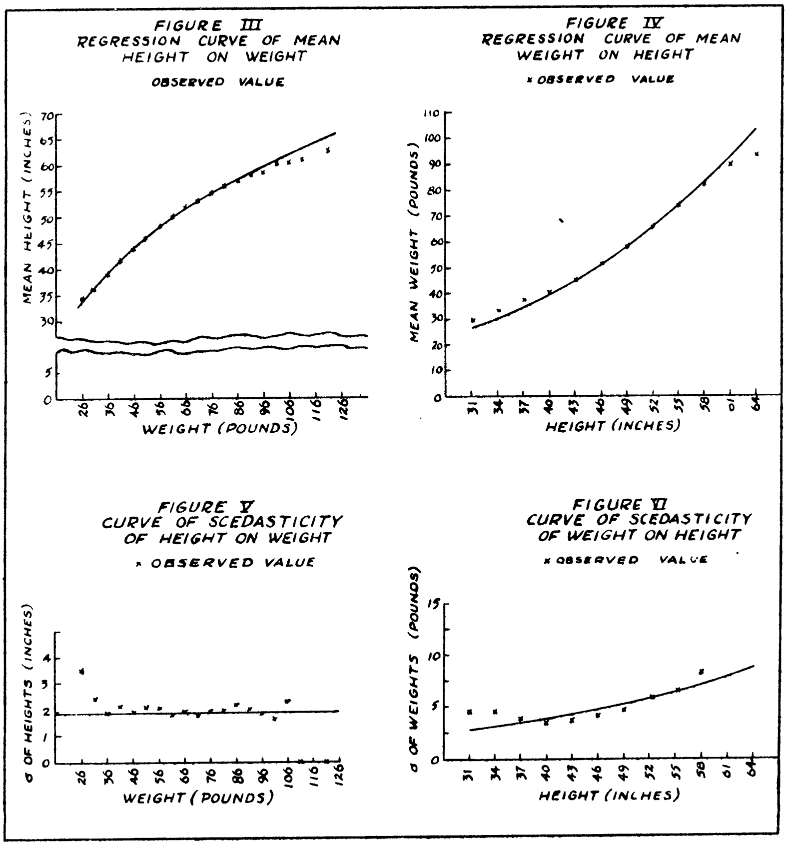


FIGURE III
REGRESSION CURVE OF MEAN HEIGHT ON WEIGHT
OBSERVED VALUE
MEAN HEIGHT (INCHES)
WEIGHT (POUNDS)
FIGURE IV
REGRESSION CURVE OF MEAN WEIGHT ON HEIGHT
× OBSERVED VALUE
MEAN WEIGHT (POUNDS)
HEIGHT (INCHES)
FIGURE V
CURVE OF SCEDASTICITY OF HEIGHT ON WEIGHT
× OBSERVED VALUE
σ OF HEIGHTS (INCHES)
WEIGHT (POUNDS)
FIGURE VI
CURVE OF SCEDASTICITY OF WEIGHT ON HEIGHT
× OBSERVED VALUE
σ OF WEIGHTS (POUNDS)
HEIGHT (INCHES)

TABLE VIII

The Mean Weights for Specified Heights

Height (Inches)	Mean Weight (Pounds) Observed	Theoretical
30-32	29.8	26.0
33-35	32.5	30.0
36-38	36.4	34.4
39-41	39.7	39.3
42-44	44.6	44.7
45-47	50.4	50.8
48-50	57.3	57.5
51-53	65.1	64.8
54-56	72.6	73.1
57-59	81.7	82.1
60-62	89.1	92.2
63-65	92.2	103.3

According to formula (23), for any specified height, the relative error of weights is 7.6%, which is much smaller than the relative error of the marginal distribution of weights, which is

$$(e^{c^2}-1)^{\frac{1}{2}} = 19.5\%.$$

Both the theoretical and observed absolute errors or standard deviations of weights for specified heights have been calculated and are shown in Table X and Figure VI. The agreement between the theoretical and observed dispersions is not as good as for the regression of the mean weight on height. It should be noted here that theoretically the standard deviations of weights for heights over 76 inches are greater than the standard deviation of the marginal distribution of weights, which is 12.8905 pounds.

In interpreting the standard deviations of weights for particular heights, we must bear in mind that the distribution of weights for any given height is not normal, but logarithmic. Hence, a proper interpretation of the dispersion of weights for a given height can be made only with reference to the skewness, masured by the third standard moment of weights, which, according to the theory, is a constant for all different heights. The theoretical third standard moment of the distribution of weights for any given height, as we shall see later, is approximately .2.

TABLE IX

The Standard Deviations of Heights for Specified Weights

	Standard Deviation of Heights (Inches)	
Weight (Pounds)	Observed	Theoretical
24- 28	3.52	1.89
29- 33	2.40	1.89
34- 38	1.91	1.89
39- 43	2.12	1.89
44- 48	1.91	1.89
49- 53	2.07	1.89
54- 58	2.04	1.89
59- 63	1.81	1.89
64- 68	1.87	1.89
69- 73	1.79	1.89
74- 78	1.92	1.89
79- 83	1.95	1.89
84- 88	2.18	1.89
89- 93	2.01	1.89
94- 98	1.86	1.89
99-103	1.62	1.89
104-108	2.34	1.89
109-113	0	1.89
114-118	....	1.89
119-123	0	1.89

TABLE X

The Standard Deviations of Weights for Specified Heights

	Standard Deviation of Weights (Pounds)	
Height (Inches)	Observed	Theoretical
30-32	4.6	2.8
33-35	4.5	3.1
36-38	4.0	3.5
39-41	3.5	3.8
42-44	3.6	4.3
45-47	4.2	4.7
48-50	4.8	5.2
51-53	5.9	5.8
54-56	6.3	6.4
57-59	8.4	7.1
60-62	12.5	7.9
63-65	14.8	8.7

Thus, from Table II in Part I, we find that the probability that any weight will be at most one standard deviation above or below the mean weight for a given height is .6839 instead of .6826, as in the case of the normal distribution. The difference between .6839 and .6826 is slight but should not be overlooked. Moreover, the difference would not be so small, if the skewness were larger.

Another thing we must observe is that since the standard deviation of weights for a given height increases with height, the probability that for a given height the weight will differ from the mean weight for that height by, say, at most one pound is not the same for all different heights, although the probability that for a given height the weight will differ from the mean weight for that height by at most one standard deviation is the same for all different heights. The former probability is greater for smaller heights.

The agreement between the theoretical and observed clisy and synagic is, of course, not expected to be close. Theoretically, the distributions of weights for specified heights should all have $\alpha_{3:v} = .23$ and $\eta_v = \alpha_{4:v} - 3 = .09$. Five observed values of $\alpha_{3:v}$ and η_v are shown below:

Height (Inches)	Observed Skewness of Weights $\alpha_{3:v}$	Observed Kurtosis of Weights η_v
36-38	.22	8.72
42-44	.19	.18
48-50	.29	.79
54-56	.12	1.54
60-62	-.93	.50

The rather large deviations of the observed η_v in the first class from the theoretical η_v and the observed $\alpha_{3:v}$ in the last class from the theoretical $\alpha_{3:v}$ may be accounted for by the fact that only 350 and 69 observations are included in the first and the last classes, respectively.

The observed marginal distribution of heights is very symmetric but is markedly leptokurtic, since its α_4 is about 2.5093. Hence, the fit given by a normal curve is not quite satisfactory, as we can see from Table XI.

The observed marginal distribution of weights is quite skew and platykurtic. As shown by Table XII, the agreement between the observed distribution and the theoretical logarithmic distribution is not very close.

TABLE XI

Relative Frequency Distribution of Heights of 11,382 School Boys between 4 and 15 Years of Age in Glasgow

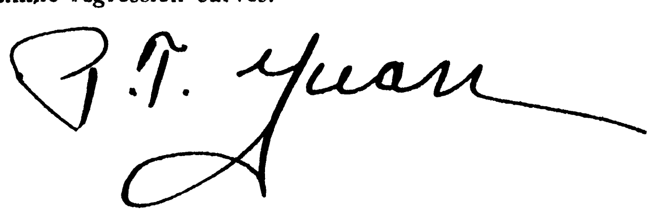
Class Limits (Inches)	Observed Relative Frequency	Theoretical Relative Frequency (Normal Curve)
27-29		.0003
30-32	.0007	.0020
33-35	.0063	.0095
36-38	.0308	.0332
39-41	.1048	.0846
42-44	.1682	.1577
45-47	.1913	.2154
48-50	.1929	.2143
51-53	.1681	.1561
54-56	.0980	.0831
57-59	.0317	.0324
60-62	.0061	.0092
63-65	.0011	.0019
66-69		.0003
Total	1.0000	1.0000

TABLE XII

Relative Frequency Distribution of Weights of 11,382 School Boys between 4 and 15 Years of Age in Glasgow

Class Limits (Pounds)	Observed Relative Frequency	Theoretical Relative Frequency (Logarithmic Curve)
19- 23		.0006
24- 28	.0014	.0048
29- 33	.0119	.0211
34- 38	.0640	.0564
39- 43	.1297	.1041
44- 48	.1454	.1441
49- 53	.1499	.1609
54- 58	.1363	.1504
59- 63	.1139	.1234
64- 68	.0882	.0897
69- 73	.0682	.0602
74- 78	.0461	.0373
79- 83	.0213	.0212
84- 88	.0121	.0127
89- 93	.0058	.0065
94- 98	.0033	.0033
99-103	.0017	.0018
104-108	.0006	.0008
109-113	.0001	.0004
114-118		.0002
119-123	.0001	.0001
Total	1.0000	1.0000

In closing, we may say that the semi-logarithmic correlation surface is not at all uncommon in practice, and the method developed here for treating it should prove rather useful. In fact, our investigation opens up a new way for determining exponential and logarithmic regression curves.



BIBLIOGRAPHY

This Bibliography contains the names of the books and articles which deal with the logarithmic frequency distribution, as well as those which are related to the general problem of the transformation of frequency distributions. But no attempt has been made to include all the literature concerning the transformation of frequency distributions.

ANDERSON, WALTER—Researches into the Theory of Regression, pp. 126-148 and pp. 153-154 (Meddelande från Lunds Astronomiska Observatorium, Ser. II, Nr. 64, 1932)

BAKER, G. A.—Transformations of Bimodal Distributions (Annals of Mathematical Statistics, Vol. I, No. 4, 1930, pp. 334-44)

BERNSTEIN, S.—Sur les Courbes de Distribution des Probabilités (Mathematische Zeitschrift, Vol. 24, 1925, pp. 199-211)

BRUNS, H.—Wahrscheinlichkeitsrechnung und Kollektivmasslehre, pp. 126-139 (B. G. Teubner, Leipzig and Berlin, 1906)

CHARLIER, C. V. L.—Das Strahlungsgesetz (Meddelande från Lunds Astronomiska Observatorium, No. 55, 1913 or Arkiv för Matematik, Astronomi och Fysik, Band 9, No. 11)
Contributions to the Mathematical Theory of Statistics, 7. Frequency Curves of Compound Functions (Meddelande från Lunds Astronomiska Observatorium, No. 61, 1914)

DAVIES, GEORGE R.—The Logarithmic Curve of Distribution (Journal of the American Statistical Association, Vol. XX, 1925, pp. 467-480)
The Analysis of Frequency Distributions (Journal of the American Statistical Association, Vol. XXIV, 1929, pp. 349-366)

DODD, EDWARD L.—The Frequency Law of a Function of Variables with Given Frequency Laws (Annals of Mathematics, Second Series, Vol. 27, 1926, pp. 12-20)

EDGEWORTH, E. F.—On Representation of Statistics by Mathematics (J. R. S. S., Vol. LXI, 1898, pp. 670-700)
A Method of Representing Statistics by Analytic Geometry (Proceedings of the Fifth International Congress of Mathematicians, Vol. II, 1912, pp. 427-40)
On the Use of Analytic Geometry to Represent Certain Kinds of Statistics (J. R. S. S., Vol. LXXVII, 1914, pp. 300-312, pp. 415-432, pp. 653-671, pp. 838-852)
On Mathematical Representation of Statistical Data (J. R. S. S., Vol. LXXIX, 1916, pp. 455-500; Vol. LXXX, 1917, pp. 65-83, pp. 266-288, pp. 411-437)

FISHER, ARNE*—The Mathematical Theory of Probabilities, pp. 235-260 (The MacMillan Co., New York, 1928)

GALTON, FRANCIS—The Geometric Mean in Vital and Social Statistics. (Proc. of the Royal Society, Vol. 29, 1879, pp. 365-367)

JENKINS, THOMAS N.—A Short Method and Tables for the Calculation of the Average and the Standard Deviation of Logarithmic Distribution (Annals of Mathematical Statistics, Vol. III, No. 1, 1932, pp. 45-55)

*Mistakes have been found in the results given by Arne Fisher and Jorgensen.

JØRGENSEN, N. R.*—Undersøgelser over Frequensflader og Korrelation. pp. 45-51 (Arnold Busck København, 1916)

KAPTEYN, J. C.—Skew Frequency Curves in Biology and Statistics. (Noordhoff, Groningen, 1903)
Reply to Professor Pearson's Criticisms (Recueil des Travaux Botanique Néarlandais, Vol. 2, 1906, pp. 216-222)
Skew Frequency Curves in Biology and Statistics. 2nd paper. Ch. I (Hoitsema Brothers, Groningen, 1916)

MCALLISTER, DONALD—The Law of the Geometric Mean (Proc. of the Royal Society, Vol. 29, 1879, pp. 367-376)

NYDELL, STURE—The Mean Errors of the Characteristics in the Logarithmic-Normal Distribution (Skandinavisk Aktuarietidskrift, 1919, pp 134-144)

PEARSON, KARL—"Das Fehlergesetz und Seine Verallgemeinerunger durch Fechner und Pearson", a Rejoinder (Biometrika, Vol. IV, 1905, pp. 170-212)
Skew Frequency Curves, A Rejoinder to Professor Kapteyn (Biometrika, Vol. V, 1906, pp. 168-171)

Pretorius, S. J.—Skew Bivariate Frequency Surfaces Examined in the Light of Numerical Illustrations (Biometrika, Vol. XXII, 1930, pp. 109-223)

RIETZ, H. L.—Frequency Distributions Obtained by Certain Transformations of Normally Distributed Variates (Annals of Mathematics, Second Series, 1923, pp. 292-300)
On Certain Properties of Frequency Distributions of the Powers and Roots of Variates of a Given Distribution (Proc. of the National Academy, Vol. 13, No. 12, 1927, pp. 817-820)
On Certain Properties of Frequency Distributions Obtained by a Linear Fractional Transformation of the Variates of a Given Distribution (Annals of Math. Statistics, Vol. II, No. 1, 1931, pp. 38-47)

VAN UVEN, M. J.—Skew Frequency Curves in Biology and Statistics, 2nd Paper, Chs. II and III (Hoitsema Brothers, Groningen, 1916)
Logarithmic Frequency Distribution (Proc. Koninklijke Akademie van Wetenschappen, Amsterdam, Vol. XIX, No. 3, 1917, pp. 533-546)
Skew Frequency Curves (Proc. Kon. Ak. v. Wet., Amsterdam, Vol. XIX, No. 3, 1917, pp. 670-684)
On Treating Skew Correlation (Proc. Kon. Ak. v. Wet., Amsterdam, Vol. XXVIII, Nos. 8-9, 1925, pp. 797-811; No. 10, pp. 919-935: Vol. XXIX, No. 4, 1926, pp. 580-590; Vol. XXXII, No. 4, 1929, pp. 408-413)
Skew Correlation between Three and More Variables (Proc. Kon. Ak. v. Wet., Amsterdam, Vol. XXXII, No. 6, 1929, pp. 793-807, No. 7, pp. 995-1007, No. 8, pp. 1085-1103)

WICKSELL, S. D.—On the Genetic Theory of Frequency (Arkiv för Matematik, Astronomi och Fysik, Band 12, No. 20, 1917)
On Logarithmic Correlation with an Application to the Distribution of Ages at First Marriage (Meddelande från Lunds Astronomiska Observatorium, No. 84, 1917)
Das Heiratsalter in Schweden 1891-1910, Ein Korrelationsstatistische Untersuchung (Lunds Universitets Årsskrift, N. F. Avd. 2, Bd. 14, Nr. 18, 1918 or Kungl. Fysiografiska Sallskpets Handlingar, N. F. Bd. 29, Nr. 18)

*Mistakes have been found in the results given by Arne Fisher and Jørgensen.