A STATISTICAL APPROACH TO SOME MINE VALUATION AND ALLIED PROBLEMS ON THE WITWATERSRAND.

A STATISTICAL APPROACH TO SOME MINE VALUATION AND ALLIED PROBLEMS ON THE WITWATERSRAND

By

D. G. Krige, B.Sc.(Eng.)(Rand)

Thesis presented for the
Degree of M.Sc. in Engineering,
University of the Witwatersrand

JOHANNESBURG,
15th March, 1951.

INTRODUCTION

The determination of the probable tonnage and grade of payable ore remaining from time to time in a mine and the correct policy of selective mining based on such determinations, is of vital importance to the mining engineer and the investor of capital. It is surprising, therefore, that more attention has not been devoted on the Witwatersrand to the scientific improvement of mine valuation methods, which at present consist almost entirely of the application of simple arithmetic and empirical formulae based on practical experience.

Experience based on intelligent observation and practical experimentation has, no doubt, throughout the history of mankind provided the basis for the advancement of all the sciences as well as the necessary confidence in approaching the multitude of problems scientific and otherwise which have had to be faced from time to time. It is also evident that without the foundation stone of elementary arithmetic the so called exact sciences could not have attained their present degree of development. Practical experience and elementary arithmetic have, therefore, naturally also been indispensable in providing the background for present mine valuation methods on the Rand.

These methods, however, ignore the additional information and experience which can be gained from a careful statistical analysis of the behaviour of gold values both individually and collectively. In the writer's opinion what is called for in improving the present methods is, therefore, not the discarding of the valuable experience already gained, but the widening of such experience by approaching the subject on a statistical basis, an approach which will in turn

inevitably/...

inevitably lead to the adoption of improved methods of valuation.

The science of statistics has expanded rapidly during the last two decades and its value as an indispensable tool is now recognised not only by research workers and scientists but also, ever increasingly, by the commercial and industrial world. This being the case it is noteworthy that in a mining field such as the Rand with its highly developed and advanced mining methods, singularly little attention has been paid to the analysis of mine valuation problems on a modern statistical basis. This omission is even more striking when cognisance is taken of the wealth of sampling data concerning the gold ore which is available and of the far-reaching decisions and deductions constantly being based on such data. Various contributions have been made from time to time towards the application of statistics to mine valuation on the Rand* but a systematic practical approach on clearly defined fundamental concepts still appears to be lacking.

The object of this paper is, therefore, to attempt to indicate how the mine valuator can gain practical experience in the statistical study of gold values, and how such experience and specialised statistical methods can be applied profitably in solving many of the existing problems and in improving the general standard of mine valuation on the Rand. For this purpose, digression into the somewhat specialised field of mathematical statistics will be necessary, but it is hoped that the mine valuator who lacks the mathematical background to grasp the detailed statistical reasoning fully, will be able to appreciate the fundamental concepts and if convinced, will be able to apply the suggested methods

intelligently/...

---

*Refs. 13, 14, 15, 3, 5 & 2 - see Bibliography.

intelligently. It is for this reason that the writer has attempted to explain certain basic statistical concepts in more detail than may appear necessary.

The writer makes no claim that any of his suggested methods are necessarily unique nor the final word in statistical application, but it is his earnest hope that the thoughts presented may arouse the interest of those who have the welfare of mining at heart, and in so doing, assist in the already overdue closing of the present gap in the mine valuation branch of the science of mining.

TABLE OF CONTENTS

Page

CHAPTER IV.

CHAPTER V.

CHAPTER VI.

CHAPTER I

DEFINITION OF FUNDAMENTAL CONCEPTS

The intelligent observer has no doubt often been amazed at the regularity and order behind what at first glance, appears to be a chaotic variation in the attributes of an object, event or condition. The individual heights of the people forming the population of a town, for example, appear from a casual investigation to vary haphazardly, and yet when such height measurements are grouped according to the frequency of occurrence of individual sizes over the full range of sizes a surprisingly uniform and regular trend in such frequencies will be found. Thus intelligent observation and analysis will generally disclose the regular pattern and definite law behind the apparent chaos, i.e. the method behind the apparent madness. Statistics is the branch of applied mathematics which suitably provides the scientific aid required for such observation and analysis.

Even an experienced mine valuator on the Rand may believe that the variation between gold values along a stretch of drive, raise or stope face is haphazard. This is not the case, however, and it follows naturally that the establishment of the regular pattern and laws followed by such values, and the correct interpretation thereof, must open up new avenues of approach to the benefit of mine valuation in general.

It is as well to stress at this stage that the basic problem of mine valuation is that the <u>actual</u> gold content of a block of ore to be stoped is unknown and that it can never be determined exactly until the ore has been mined and the gold extracted. Even in the latter event the content can only be inferred since it is impossible to measure the gold lost in

mining/...

mining exactly, and from a practical point of view the ore from a single block cannot generally be kept separate underground and in the reduction works. The actual gold value of an intact block of ore can, therefore, only be estimated from the limited number of values available round its periphery, the orthodox estimate being based on the arithmetic mean of such a set of available values, i.e. the mean of such values is accepted as being the indicated mean value of the block of ore. The object of a statistical approach to mine valuation is to determine the reliability of such existing methods of estimation and to develop, where possible, methods which will on average yield closer and more reliable estimates of the actual mean value of the ore from the limited available sampling information.

Before this can be done, however, the following fundamental statistical terms and their application to mine valuation on the South African gold fields have to be defined.

1. Population.

The common concept of a "population" is that of a large group of persons, each "member" of the population being identified by his or her own particular attributes such as height, weight, age, wealth, etc. In the statistical sense, however, the measurements of any one attribute of the individual persons in such a group constitute a population (of measurements) and each such measurement is regarded as a member of the population.

In the case of a gold mine, the ore body can be regarded as a single ore parcel which can be subdivided into a large number of small parcels of ore, each of these smaller parcels having its own attributes, the vital one naturally being its gold content. The aim in framing the ideal policy of selective mining is to select for stoping purposes only

those/...

those parcels of ore which contain sufficient gold to pay for all expenditure incurred up to and including the extraction of this gold, and to leave intact all parcels with an insufficient gold content to cover such costs. In practice, except in the case of unusually wide auriferous reef bodies, this process of selection is effected in respect of reef "parcels" which in each case occupy the entire width of the reef body (or economic band of reef), and the "payable" and "unpayable" parcels can consequently be depicted on the plane of the reef by the areas covered by these parcels. For practical purposes, therefore, a reef body in a particular mine can be regarded as a large "area" of reef consisting of smaller individual reef "areas", each "area" being identified in particular by the gold content of the ore parcel (or "volume" or tonnage of ore) it represents. The gold contents of such individual small reef "areas" within a large reef "area" can from a statistical angle, consequently be regarded as the "members" of a "population".

The smallest "area" of reef the gold content of which is measured in practice, is that represented by the cross sectional area of the standard size channel cut in the process of sampling across the width of the reef body at a sampling section, and on average measures approximately six square inches. For mine valuation purposes, therefore, the measurements of the gold contents of all the standard size (6 sq. in.) reef "areas" which constitute a larger reef "area" will be regarded as a population and every such individual measurement will be a member of the population. The basic population is comprised of the actual gold contents of these "areas" but these can in practice only be measured by underground sampling, and hence the observed population consists of a number of measurements of the actual gold contents concerned.

In/...

In the practical case of a block of ore measuring, say 200 ft. x 200 ft., which has been sampled at 5 foot intervals round its periphery, the measured gold contents of the 160 odd standard size reef "areas" at the corresponding number of sample sections will constitute the only known members of the population of measurements of the gold contents of the odd million 6 sq. in. reef "areas" constituting the entire block.

The gold content of any one such standard size reef "area" (6 sq. ins.) will be measured by the assayed gold content of the sample(s) obtained from the channel cut at the corresponding sampling section, i.e. by the (weighted average) dwt/ton of the sample(s) x tonnage of sample(s). Now, since the tonnage of the sample(s) is directly proportional to the volume of the sample(s), and the volume is in turn directly proportional to the overall sampled width (when the cross sectional area of the channel cut for every sample is identical), it follows that the measurement of the gold content of a standard size reef "area" is directly proportional to the average dwt/ton over the sampled width x the sampled width

= total inch dwts for the sample section corresponding to the area concerned.

The inch dwts of a sample section can, therefore, be accepted as a measurement (requiring only multiplication by some constant factor to yield the actual number of dwts) of the gold content of a standard size of reef "area" (6 sq. ins.) corresponding to this sampling section.

Where the reef width is relatively narrow the stoping width is determined entirely by practical mining considerations and is fairly constant. In such a case the inch dwt value at a sampling section divided by the more or less constant factor of the stoping width so as to yield the dwt/ton value over the stoping width will also, therefore, provide a measurement of the gold content of the relevant standard size reef area.

Similarly/...

Similarly, in the case of a wide variable reef width having a definite influence on the stoping width, but where neither of these widths appears on average to be related to the corresponding inch-dwt values* the dwt/ton value over the stoping width at a sampling section will on average also provide a measure of the gold content of the corresponding standard size reef area.

In the unusual case where there appears to be a definite relationship between the stoping widths and corresponding inch-dwt values at the various sample sections, the problem is more complicated and will not be considered in this paper.

From a practical point of view, therefore, the use of either the inch-dwt value or the dwt/ton value over the stoping width at a sampling section can be justified and should yield the same eventual answer, since the average dwt/ton value for the tonnage of ore in a block is the quotient of the average inch-dwt value and the average stoping width.

For the purpose of this thesis the inch-dwt measure will be used almost invariably and a population will therefore be considered as being comprised of a number of inch-dwt values of sample sections corresponding to "standard" size reef areas. In the case of a block of ore, for example, the population will consist of all the theoretically possible inch-dwt values which could be obtained if the block were to be extracted by a process of continuous sampling.

Similarly the sample values obtained from a stretch of drive, raise or stope face can be considered to be equivalent to that obtained from a relatively narrow and elongated "area" of reef containing a population of sample section values.

A case where the area concept is departed from is in the analysis of the distribution of calculated ore reserve values. In this case the population in effect comprises the indicated mean values of a number of blocks of ore. In order,

however/...

---

* i.e. where the full range of stoping width variations is likely to be associated with every category of inch-dwt values.

however, to allow for the fact that these block areas are usually not only very divergent in size, but also insufficient in number to reflect the proper distribution of the ore reserve values, the tonnages of the various value categories provide a better frequency measure. The population will therefore in this case consist of all the individual tons of ore in the ore reserves each at the indicated average value of the ore block of which it forms part.

2. Sampling from a Population.

In the statistical sense "sampling" implies the selection, at random, of a limited number of members of a population, the group of selected members constituting the so called "sample". To the mine valuator "sampling" implies the physical act of chisselling out a few pounds of reef (and waste) material for assay purposes, and "samples" imply the separate packages of reef (and waste) material obtained in "sampling". It is, therefore, obvious that in the application of statistics to mine valuation a clear distinction is required between the above dual meanings of both "sampling" and "sample". Since this thesis is primarily intended for the benefit of mine valuators, the valuation interpretation of these two terms will be maintained and the corresponding statistical terms will be referred to in the following manner:-

Statistical Term	Valuation Equivalent
"Sample"	A set of sample values drawn from a population of sample values.
"Sampling"	The act of drawing a set of sample values from a population of such values.

The term sample where used in this thesis, therefore, unless qualified is used in the mine valuation sense and an individual sample value will be the inch-dwt value at a

sampling/...

sampling section, i.e. a member of a population of individual sample values.

1. Random Sampling.

A considerable part of statistical theory has been built up round the basic concept of "random" sampling (statistical sense), i.e. the concept of the drawing of a set of sample values (valuation sense) from a population of such values in a purely random and unbiassed manner. Briefly this means that every individual member of a population must have an equal chance of selection.

Consider now an area of a reef body from which a set of "random" sample values is required. It is common sense to any mine valuator that if, for instance, 10 samples are taken in, say, the confined space of one corner of this reef "area," the values of such samples will, in all probability, not be representative of the values in the "area" as a whole, and since the theoretical sample values in the remainder of the "area" had no chance of selection at all, the 10 sample values will certainly not be "random." To ensure, therefore, that all sample values have an equal chance of selection the ideal practical method would apparently be to divide the "area" into ten equal portions and to select without bias one sample per portion, such a method being virtually equivalent to grid sampling on a square pattern.

In practice, however, samples can only be taken round the periphery of an ore block. The theoretical sample values in the interior of the block, therefore, have no chance of selection, and "random" sampling in the ideal sense becomes impossible. Where, however, the selection of the locations of the drives and raises bounding a block of ore has not been influenced in any way by sampling values previously known or inferred, and where sampling round the periphery is carried out

without/...

without bias, it is contended that the results will in general conform to those which would be obtained from ideal "random" sampling of the block on a grid pattern. Details of an experimental attempt to confirm this contention will be found in Chapter V, paragraph 1.

4. Homogeneous and Non-homogeneous Populations.

The sample values along a well defined reef horizon (or sedimentation unit) where the original pattern of gold deposition has not subsequently been upset by factors such as leaching, can be regarded from a practical point of view as constituting a homogeneous population. Where however two reefs merge or where the basic gold distribution has been upset by e.g., the hydrothermal addition of gold or the leaching out of a proportion of the gold, the resultant population will no longer be homogeneous and may disclose characteristics foreign to those of the constituent or of the original population(s) respectively. In cases where non-homogeneity is suspected, therefore, the problems should be approached either from the angle of the constituent homogeneous populations (where a mixture of populations is suspected), or of the reconstruction of the original homogeneous population. Such problems however are specialised and fall outside the basic concepts which require consideration at this stage.

5. Frequency Histograms and Curves.

The statistical analysis of a population of values consists primarily of the segregation of such values into a range of selected value categories. The population is then represented graphically by plotting the limits of the range of values within each value category as abscissae and on each such range of values as base, a rectangle with area in direct proportion to the frequency of occurrence of the values in the

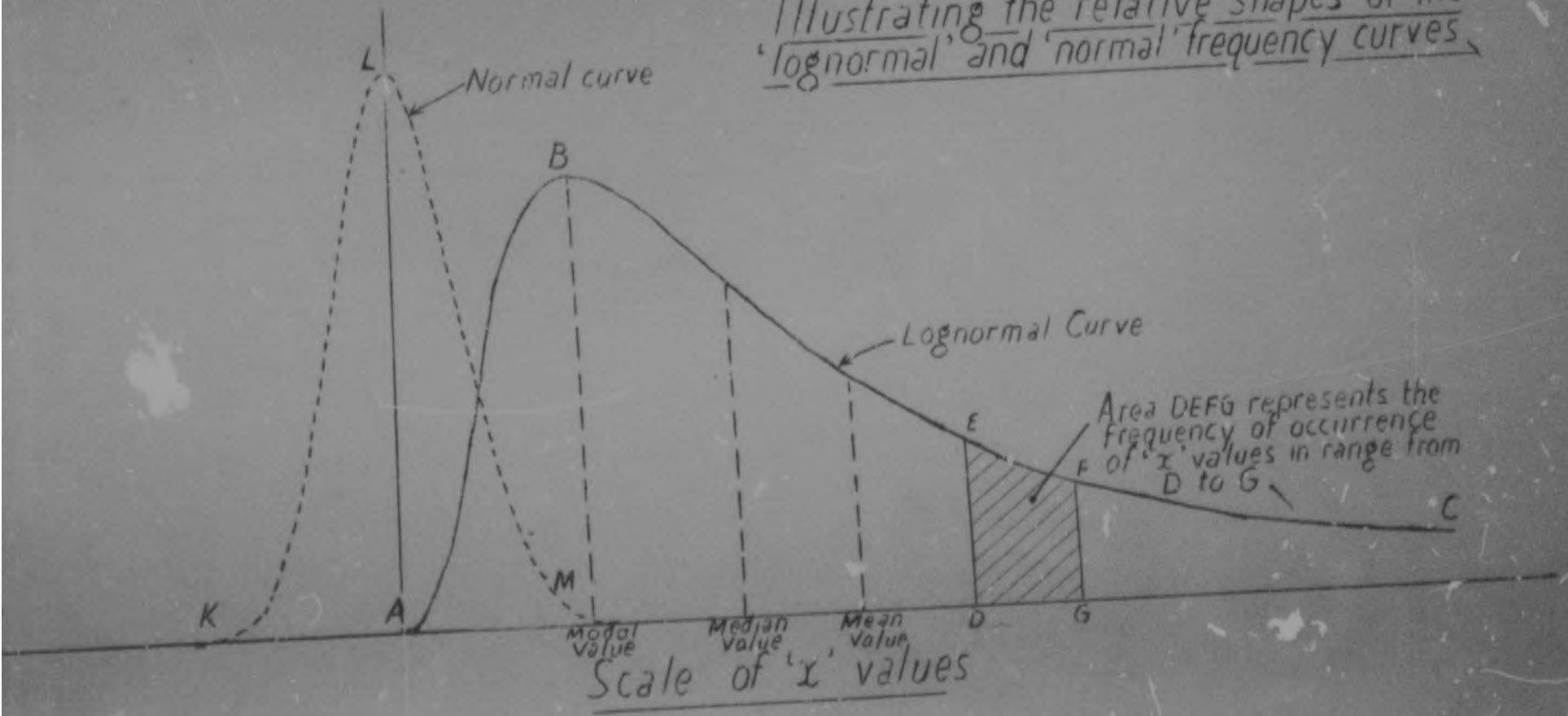
value/...

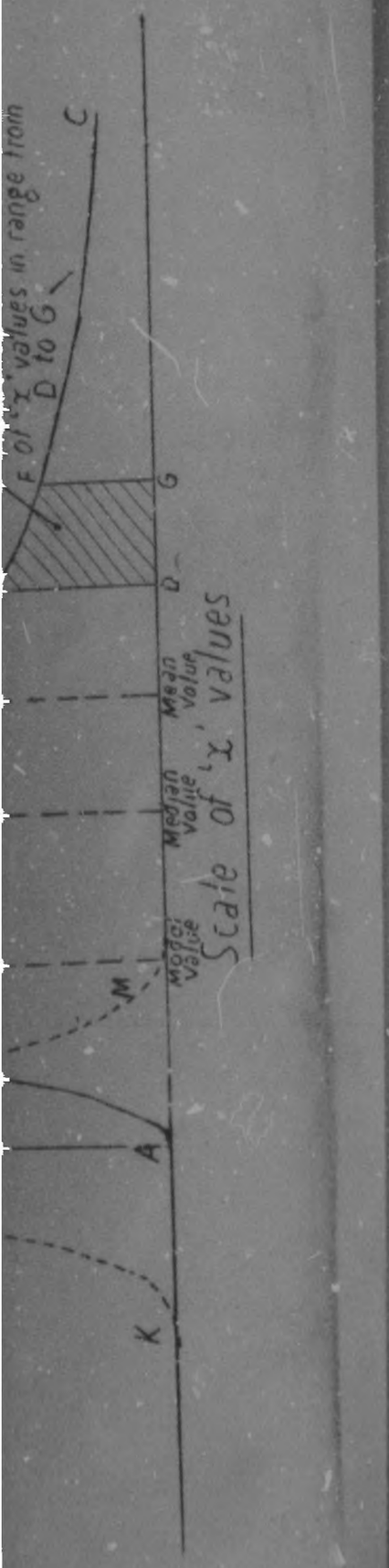
value category concorned. The resultant step diagram is called a frequency histogram, and where the value ranges are made sufficiently small, this step diagram will in the limiting case, merge into a smooth curve called a frequency curve.

CHAPTER II/...

DIAGRAM No. 1

Illustrating the relative shapes of the 'lognormal' and 'normal' frequency curves





CHAPTER II

THE GENERAL CHARACTERISTICS AND APPLICATIONS OF THE LOGNORMAL FREQUENCY CURVE

1. The Shape of the Lognormal Curve.

The fact that the gold values obtained in sampling a reef area could be represented by a frequency histogram of definite shape was known as far back as 1919* but it was not until recent years that the type of frequency curve which could be fitted satisfactorily to such a histogram was recognised as the lognormal frequency curve.** A typical lognormal curve is illustrated by the curve ABC on Diagram No. 1.*** The frequency of occurrence of the values falling into the value category DG are represented by the area under the curve between the two ordinates corresponding to the values D and G, i.e. area DEFG.

As suggested by the name "lognormal," this curve is related to the well known "normal" curve of error, and can be transformed into the latter by plotting the abscissae on a logarithmic scale. If, for example, the abscissae of the lognormal curve ABC are plotted on a logarithmic scale, with the retention of the frequencies in the corresponding value categories, this curve will be transformed into a "normal" curve of the type illustrated by the dotted curve KLM on the diagram. For practical reasons arising out of the plotting, these two curves have been represented in a purely illustrative manner and do not indicate the relative positions or shapes of these two curves in respect of the same frequency distribution.

2. The/...

*Ref. 13. **Refs. 3, 5 & 2. ***See opposite.

2. The Application of the Lognormal Curve in Various Fields.

The lognormal frequency curve is not peculiar to the distribution of gold values and has been found to be applicable in a large number of widely different fields as the following brief list will indicate:-

The incomes of individuals in a nation.*

The sizes of grains in samples from sedimentary deposits.**

The sizes of sandgrains in samples from windblown sand.***

The sizes of particles of silver in a photographic emulsion.****

Sensitivities of animals of same species to drugs.****

Numbers of plankton caught in different hauls with a net.****

Amounts of electricity used in medium class homes in the U.S.A.****

Reaction times of human beings in a word test.****

Number of words in sentences from works of G. B. Shaw.****

Diameters of particles of airborne dust in coal mines.****

As far as the Witwatersrand gold field and its extensions are concerned, evidence from a number of the chain of mines stretching for more than a hundred miles from Heidelberg in the east to the West-Wits line in the west as well as from mines in the Klerksdorp Sector, indicates that it is highly probable that the lognormal curve can be applied throughout to all the economic reef horizons either directly or indirectly (where the population of gold values is not homogeneous). An analysis of the borehole values for the Basal Reef in the Orange Free State field confirm the natural expectation that these values are also lognormally distributed.

3. Suggested/...

---

*Ref.5. **Refs. 11 & 18. ***Ref. 19. ****Ref. 20.

3. Suggested Reason for the Lognormal Distribution of Gold Values.

The above indications of the general applicability of the lognormal frequency curve suggest that lognormal distributions result from definite natural laws which cover at least all the fields referred to in paragraph 2 above.

A reef body can be regarded as a mixture of gold and waste particles, the relative concentration of the former per unit reef area being measured as previously explained by the inch-dwt value of a sample section. Disregarding the variation in particle sizes, the mixture of gold and waste particles can be considered analogous to a mixture of, say, black and white balls respectively, and the reef in a block of ore analogous to a layer of such a mixture of balls covering a corresponding area. If, now an even layer of balls was formed by the spreading of the mixture of balls in an entirely random manner over the area of the block concerned, it can be shown by the application of the basic theories of probability that the concentrations of black balls per unit small area, (say, 6 sq. ins.) will vary according to the "normal" frequency distribution law, i.e. the frequency distribution of such concentrations will conform to the "normal" frequency curve.

If the gold particles were deposited in a random manner, one would, therefore, naturally expect the gold concentrations per standard size area as measured by the sample section inch-dwts also to be distributed "normally," whereas, in fact, the logarithms of such inch-dwt values are "normally" distributed.

The following references to Nature's use of the linear and logarithmic scales are, however, of particular interest in this connection and may provide a partial explanation:-

"The/...

"The linear scale, since it was first cut on the wall of an Egyptian temple, has come to be accepted by man almost as if it were the one unique scale with which Nature works and builds, whereas it is nothing of the sort. Its sole value lies in giving due prominence to the differences and sums of quantities when these are what we want to display. But Nature, if she has any preference, probably takes more interest in the ratios between quantities; she is rarely concerned with size for the sake of size."*

"Linear scales are seldom acceptable to Nature. A millimetre difference between the diameters of two boulders is insignificant but a millimetre difference between one sand grain and another is a large and important inequality. The natural scale for size classification, therefore, is logarithmic."**

It is, therefore, not surprising that the distributions of e.g., grain sizes in a sample from a sedimentary horizon and the sizes of sand grains in naturally deposited windblown sand deposits, tend to be "normal" when the sizes are measured on a logarithmic scale. Without enlarging in any way on the above quotations or endeavouring to reconcile the observed facts with the laws controlling the settling of (gold) particles in liquids, it appears that the fact that gold values are distributed "normally" only when measured on a logarithmic scale could be explained from the natural laws already known, and that research in this direction will prove profitable, bearing in mind particularly the remaining doubts as to the origin of the gold in the Witwatersrand reefs.

4. General/...

---

*Ref. 19, p. 2. **Ref. 18, p. 15.

4. General Basic Conclusions to be drawn from the Lognormal Distribution of Gold Values.

Since the gold values in a homogeneous reef body can be expected to be distributed lognormally, the question naturally arises whether this will be the case irrespective of the size of the area concerned. The selection of the boundaries of a mine in respect of which the lognormal distribution of gold values has been observed, is generally arbitrary and it is, therefore, natural to expect the distributions of gold values within portions of such a mine also to be lognormal irrespective of the size of such portions. This has been confirmed by practical experiment for various sizes of reef areas down to a size smaller than that of the average ore reserve block on a mine.*

Further, since the decision to select samples corresponding to reef areas of 6 sq. ins. each is also arbitrary, the size of the sample should also not affect the typically lognormal distribution of gold values. This is confirmed by the observed fact that in an ideal case** the average values of ore reserve blocks within a mine are also distributed lognormally.

It is immediately evident from the illustration of a typical lognormal curve (Diagram No. 1), that the inch-dwt values comprising a lognormal frequency curve cover the entire theoretical range from zero to infinity. Bearing in mind that the area under the curve corresponding to any particular value category is a direct measure of the frequency of occurrance of values in this value category, it is evident that since the curve approaches the x-axis asymptotically in the range of the higher value categories, the frequency of

occurrence/...

---

*See Chapter V.

**Where the mine is 100% payable and the ore reserves, therefore, include all ore blocks; and where the natural distribution has not been upset by previous mining operations.

occurrence of extremely high values is relatively small but that it can only become zero for infinitely large values.

In drawing a set of values at random from a population of values, the probability of drawing a value in any particular value category is measured by the relative frequency of occurrence of values in this category, e.g., if 10% of the total values occur in a particular value category, the probability of drawing a value from this category will naturally be 1 in 10. It is, therefore, evident that in the case of a lognormal distribution of values, the probability of striking an extremely high value is slight but will never become non-existent except for infinitely large values.*

It is also evident from Diagram No. 1 that every lognormal distribution, no matter what its mean value may be, must comprise a mixture of values ranging theoretically from zero to infinity. A distribution with a low average value will, therefore, always contain a proportion (even if infinitely small) of relatively high values, and vice versa a distribution with a high average value will contain a proportion of relatively low values. In mine valuation, therefore, the occurrence of relatively high values (even if only occasionally) in a low-grade block of ore is quite natural, and similarly also the occurrence of low values in a high-grade block of ore.

Considering now the practical aspect of, say, a block of ore or a stope face in respect of which only a limited number of all the possible values is available, it is evident that the possible combinations of, say, 10 sample values each, which can be drawn from the complete distribution formed by all

the/...

---

*In practice the maximum gold value possible will be that corresponding to pure gold, i.e. some 583,000 dwts/ton, or say, 29 million inch-dwts for a 50-inch stoping width.

the sample values, will be infinite and, therefore, that the probability of striking two identical sets of 10 values each is slight. Further in view of the wide range of values covered by the parent population, the striking of a set in which all 10 values are identical, is virtually impossible. If, therefore, in sampling, say, 10 sections along a stope face, the inch-dwt values are found to be identical or to lie within a very close range of values, the result is either that of a highly improbable event or must be suspected of not being genuine.

A further basic conclusion to be drawn from the knowledge of the lognormal frequency distribution of gold values is that the individual sample values available in respect of a block of ore or a stope face represent only a few known values out of a virtually infinite number of values which can be obtained by repeated sampling. Where the few known sample values are distributed over the range of values in approximately the same proportions as the total number of possible sample values, the mean value of these few samples will naturally correspond closely to the true mean value of all the possible samples. In practice, however, some of the relatively few extremely high values in the parent population of values, must at one time or another be struck in taking a set of samples, and will in such an event appear to be out of accord with the rest of the sample values in the set, and will raise the average value of the set to an abnormally high figure. Such values are generally regarded as "anomalous," "freak," or the result of bad sampling, and are in practice usually "cut" or "adjusted" by arbitrary methods in order to yield what, at any rate, appears to be a more reliable average result. Such apparently anomalous values are, however, genuine members of the population of values along the stope face or in the block

of/...

of ore, and are, therefore, in no sense truly anomalous or freak. The correct approach to the problem of estimating the true mean value of the unknown population of values, (i.e. of the stope face or ore block) from the few known members of this population, (i.e. from the few available sample values) is, therefore, to fill in the gaps between these known values in such a way as to result in the best estimate of the parent distribution of values, i.e. the population, without discarding or "cutting" any one value which may appear to be anomalous. This is basically the aim in approaching the problems of mine valuation from a statistical angle.

It is also evident that since even adjoining sample values cannot be expected to be identical, a fact which has in a practical way often been observed from the results of check sampling in the same groove, <u>any sample value cannot be regarded as having a so called "area or distance of influence,"</u> except in so far as it is related to the actual reef area previously occupied by the sampled material, i.e. approximately 6 sq. ins. Where sampling is done at, say, 5 foot intervals, the "influence" of a sample value cannot, therefore, on any logical grounds be extended for a distance of $2\frac{1}{2}$ ft. on either side of the relevant sample section. It appears, therefore, that sampling need not be carried out at rigidly determined regular intervals, and that where such intervals are in practice, irregular (but not such as to introduce any obvious bias), weighting of individual values by their so called "distances of influence" cannot be upheld on scientific grounds. It also follows that <u>the conception that an occasional high value encountered in sampling successive stope faces in a low-grade block of ore is indicative of a patch of high-grade ore</u> (extending halfway from the relevant sample section to surrounding sampling sections), <u>is entirely erroneous</u>.

CHAPTER III/...

CHAPTER III

A MATHEMATICAL ANALYSIS OF CERTAIN PROPERTIES OF THE LOGNORMAL DISTRIBUTION AND LOGNORMAL CORRELATION SURFACE

Note: In this and subsequent chapters the following references will apply:-

Normal curve	- the normal curve of error on which the major part of the theory of statistics is based.*
Mean	- arithmetic mean.
Sample	- sample in the mining valuation sense as distinct from "sample" in the statistical sense, the latter being referred to in this thesis as a "set of sample values."
Logarithms	- Naperian logs unless otherwise stated, i.e. logs to the base "e."

1. General Form.

The most general mathematical expression for the lognormal curve is based on the assumption of some lower finite value limit for the variable "x" and no upper value limit and involves three parameters.** In the case of the distribution of gold values, which can range from nil to a theoretical value of infinity,*** a somewhat simpler expression, involving only two parameters, can be employed, viz:-

$$y = Ke^{-a^2(\log x - b)^2} \quad \text{****} \qquad \ldots\ldots\ldots (1)$$

where $K = \dfrac{a(\text{total area under curve})}{\sqrt{\pi}\ .\ e^{\,b + \frac{1}{4a^2}}}$

and where x = the variable, e.g. gold value

a/...

---

*Ref. 4, p.114. **Ref. 1 & Ref. 9, p.236

***The maximum possible value is naturally that for pure gold which is still not equivalent to an infinite number of dwts/ton, but can for practical purposes be regarded as such.

****Refs. 2 & 3.

a and b are parameters,

and $\int_{\bar{x}}^{x+dx} y.dx$ = area under the curve between the ordinates corresponding to the abscissae x and (x + dx)

= frequency of occurrence of x values lying between these ordinates

It must be stressed that whereas the variable x, e.g. gold value, can be plotted directly as the abscissae in graphing this curve, the frequency of occurrence of values being the other variable, is related to specific ranges of x values, and cannot, therefore, be plotted directly as the ordinates. The frequency of occurrence of values within a specific range of values or value category is consequently represented by the area under the curve between the values of x forming the outer limits of the range concerned.

In statistics the total area under the curve is usually taken as unity, i.e. total frequency = 1 or 100%, and frequencies are then required to be expressed as fractions of the total, and K then becomes

$$\frac{a}{\sqrt{\pi}\ .\ e^{b + \frac{1}{4a^2}}}$$

2. Arithmetic Mean.

The arithmetic mean of the lognormal population

$$m = \int_{0}^{\infty} xydx$$

$$= e^{b + \frac{3}{4a^2}} * \qquad \ldots\ldots\ldots (2)$$

from which $b = \log m - \frac{3}{4a^2}$

3. Transposed/...

---

* Ref. 2.

3. Transposed Form.

Substituting for b and K in (1) above

$$y = \frac{a}{\sqrt{\pi}} e^{-\log x - a^2(\log \frac{x}{m} + \frac{1}{4a^2})^2} \qquad \ldots\ldots\ldots (3)$$

In this expression "a" and "m" are the two parameters, and it is consequently evident that for a specific value of "m," i.e. the mean value of the population, the shape of the curve is determined entirely by the other parameter "a." This aspect was investigated in some detail by Ross.*

4. Median and Mode.

The position of the median, i.e. the x value at which the area under the curve is bisected and on either side of which 50% of the total frequency of x values will lie, is determined from

$$\int_0^x y.dx = .5$$

from which*

$$x = m.e^{b + \frac{1}{2a^2}}$$

$$= m.e^{-\frac{1}{4a^2}} \qquad \ldots\ldots\ldots (4)$$

The position of the mode, i.e. the x value corresponding to the maximum frequency per unit "dx" interval and thus to the peak of the curve is defined by*

$$x = e^{b}$$

$$= m.e^{-\frac{3}{4a^2}} \qquad \ldots\ldots\ldots (5)$$

and the height of the mode by

$$y = K$$

$$= \frac{a}{\sqrt{\pi}.e^{b + \frac{1}{4a^2}}} = \frac{ae^{\frac{1}{2a^2}}}{m.\sqrt{\pi}} \qquad \ldots\ldots\ldots (6)$$

The/...

*Ref. 2.

The relative graphical positions of the mode, median and mean of the lognormal distribution are indicated on Diagram No. 1,(see Chapter II).

5. Moments and Standard Deviation.

The 2nd moment v_2 of the distribution about the origin is defined as

$$v_2 = \int_{x=0}^{x=\infty} y.x^2.dx$$ which from (1) above with the total area under the curve equal to unity

$$= \frac{a}{\sqrt{\pi}} \int_0^{\infty} x^2 e^{-b - \frac{1}{4e^2} - a^2(\log x - b)^2} .dx$$

substituting $\begin{cases} w = a \log x - ab - \frac{3}{2a} \\ \text{from which } dx = \frac{x}{a}dw \\ \text{and } x = e^{\frac{w}{a} + b + \frac{3}{2a^2}} \end{cases}$.

this reduces to

$$(e^{2b + \frac{2}{a^2}}) \frac{1}{\sqrt{\pi}} \int_{w=-\infty}^{w=+\infty} e^{-w^2}.dw \qquad \ldots\ldots\ldots (7)$$

$= e^{2b + \frac{2}{a^2}}$ since the last factor in (7) is the integral of a form of the normal curve of error = unity since the total area under the log-normal curve has also been taken as unity

and from (2) above

$$v_2 = m^2 e^{\frac{1}{2a^2}} \qquad \ldots\ldots\ldots (8)$$

Variance: The second moment about the arithmetic mean "m;" i.e. the variance, is given by

μ_2 = second moment about the origin - (mean)^2*

$$= m^2 e^{\frac{1}{2a^2}} - m^2$$

$$= m^2(e^{\frac{1}{2a^2}} - 1) \qquad \ldots\ldots\ldots (9)$$

The/...

* Ref. 4, p.65.

The standard deviation, i.e. $\sqrt{\text{variance}}$

$$= \sqrt{\mu_2} = m\sqrt{e^{\frac{1}{2a^2}} - 1} \qquad \ldots\ldots\ldots (10)$$

The coefficient of variation, i.e. the standard deviation divided by the arithmetic mean*

$$= \sqrt{e^{\frac{1}{2a^2}} - 1} \qquad \ldots\ldots\ldots (10a)$$

The 3rd moment, v_3, of the distribution about the origin is defined as

$v_3 = \int_0^\infty y.x^3.dx$, which can in a manner similar to that used for the 2nd moment be reduced to

$$v_3 = m^3 e^{\frac{3}{2a^2}} \qquad \ldots\ldots\ldots (11)$$

But the 3rd moment about the mean

$$= \mu_3 = \int_0^\infty y(x-m)^3 dx = v_3 - 3mv_2 + 2m^3 **$$

$$= m^3 e^{\frac{3}{2a^2}} - 3m^3 e^{\frac{1}{2a^2}} + 2m^3$$

$$= m^3 (e^{\frac{1}{2a^2}} - 1)^2 (e^{\frac{1}{2a^2}} + 2) \qquad \ldots\ldots\ldots (12)$$

And the 3rd moment about the mean in standard units, i.e. in units of (standard deviation)^3 in order to reduce the result to a pure number***

$$= \alpha_3 = (e^{\frac{1}{2a^2}} - 1)^{\frac{1}{2}} (e^{\frac{1}{2a^2}} + 2) \qquad \ldots\ldots\ldots (13)$$

$$= (e^{\frac{1}{2a^2}} + 2)\ (\text{coeff. of variation}) \qquad \ldots\ldots\ldots (13a)$$

The/...

---

*Ref. 4, p.90. **Ref. 4, p. 65. ***Ref. 4, p. 72.

The 4th moment.

By a similar procedure it can be shown that

$$v_4 = m^4 e^{\frac{3}{2a^2}}$$

$$\mu_4 = m^4 (e^{\frac{1}{2a^2}} - 1)^2 (e^{\frac{2}{a^2}} + 2e^{\frac{3}{2a^2}} + 3e^{\frac{1}{a^2}} - 3)$$

$$\text{and } \alpha_4 = 3 + (e^{\frac{1}{2a^2}} - 1)(e^{\frac{3}{2a^2}} + 3e^{\frac{1}{a^2}} + 6e^{\frac{1}{2a^2}} + 6) \quad \ldots\ldots (14)$$

$$= \underline{e^{\frac{2}{a^2}} + 2e^{\frac{3}{2a^2}} + 3e^{\frac{1}{a^2}} - 3} \quad \ldots\ldots (15)$$

6. Skewness and Kurtosis.

The skewness* of the curve is measured by α_3, (No. (13) above), and as "$\frac{1}{2a^2}$" can never be negative for real values of "a," α_3 will always be positive.** This means that the mode or peak of the curve will always be to the left of the mean value and the curve will always have its longer tail on the right-hand side of the mean value. It is also evident that as "a" approaches infinity, α_3 approaches zero and the curve thus loses its skewness, the mode then approaching coincidence with the mean.

The kurtosis*** of the curve is measured by α_4, and is related to whether the curve is flat-topped with filled out shoulders or sharply peaked, the peak of the normal curve being accepted as the criterion. The kurtosis for the normal curve is equal to 3, for a flat-topped curve it is less than 3, and for a sharply peaked curve it exceeds 3. From No. (15) above it is evident that α_4 will always exceed 3 and, therefore, the lognormal curve is always more sharply peaked than the normal curve and is therefore said to be leptokurtic. It is also

clear/...

---

*Ref. 4, pp.73 & 111, & Ref. 8, p.11.
**In the limiting case when $a = \infty$, $\alpha_3 = 0$.
***Ref. 4, pp.73 & 111, & Ref. 8, p.11.

clear that as "a" approaches infinity, α_4 approaches 3, i.e. the kurtosis of the normal curve.

The values of "a" commonly encountered on the Rand in dealing with the values of individual gold sampling sections range from approximately 0.5 to 0.8, giving the following range of values for the skewness and kurtosis.

TABLE 1

a =	.5	.6	.7	.8
Skewness, α_3 =	21.2	9.5	6.4	4.6
Kurtosis, α_4 =	3,949	335	122	55

These values give some idea of the extreme skewness and peakedness of most of the curves encountered in dealing with the distribution of gold values.

7. Curve of log x: Relation between Arithmetic and Geometric Means.

As stated in Chapter II, the name "lognormal" curve implies that the logarithm of the variable x is distributed "normally," i.e. according to the "normal curve". The lognormal curve (No. (3) above), can be "normalised" as follows:-

From (3)

$$ydx = f(x)dx = \frac{a}{\sqrt{\pi}} e^{-\log x - a^2(\log \frac{x}{m} + \frac{1}{4a^2})^2} .dx$$

Substituting $(z = \log x$
(from which $dx = xdz = e^z.d$

this reduces to

$$F(z)dz = \frac{1}{\sqrt{2\pi} \cdot \sqrt{\frac{1}{2a^2}}} e^{\frac{1}{-2(\frac{1}{2a^2})}\left[z - (\log m - \frac{1}{4a^2})\right]^2} .dz \quad ..... \quad (16)$$

Thus/...

Thus the variable z, i.e. log x is normally distributed* with

$$\text{mean} = \log m - \frac{1}{4a^2} \qquad \ldots\ldots\ldots (17)$$

$$\text{standard deviation} = \sqrt{\frac{1}{2a^2}}$$

$$\text{and variance} = \frac{1}{2a^2} \qquad \ldots\ldots\ldots (18)$$

But from No. (4) above the median of the lognormal curve

$$= me^{-\frac{1}{4a^2}}$$

and hence the log of the median $= \log m - \frac{1}{4a^2}$

But, from No. (17) this is the mean of the distribution of the logs of the variable "x." Further, as the mean of the logs of a number of values equals the log of the geometric mean of such values, it is evident that <u>the Normal Curve (No.(16)) is symmetrical about a value equivalent to the log of the geometric mean of the parent lognormal population,</u> ..... (19)

and further, that the <u>median and geometric mean of the log-normal population are coincident</u>. ..... (20)

<u>Note</u>: The normal distribution (No. (16)) above, can be reduced to the common form, for which tables are printed in nearly every book on statistics,

$$\text{i.e. } \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} .dw$$

$$\text{by substituting } w = \frac{z - (\log m - \frac{1}{4a^2})}{(\frac{1}{2a^2})^{\frac{1}{2}}}$$

= deviations of "z" from its mean in standard units, i.e. units of its standard deviation.

<u>Relation</u>/...

---

* Ref. 4, p. 115; also Ref. 5.

Relation between arithmetic and geometric mean.

From Nos. (4) and (20) above

$$\text{Geometric mean} = me^{-\frac{1}{4a^2}}$$

$$= (\text{arithmetic} \quad)e^{-\frac{1}{4a^2}} \quad \ldots\ldots\ldots (21)$$

$$\text{or log of geometric mean} = \text{log of arithmetic mean} - \frac{1}{4a^2} \quad \ldots\ldots\ldots (21a)$$

8. The Distribution of a Product of "x", i.e. of "x" multiplied by a constant.

Let $q = kx$ (i.e. $dx = \frac{dq}{k}$)

where k = a constant

i.e. $x = \frac{q}{k}$

Substitution in No. (3) yields

$$f(q)dq = \frac{a}{\sqrt{\pi}} e^{-\log(\frac{q}{k}) - a^2(\log\frac{q}{km} + \frac{1}{4a^2})^2} .dq(\tfrac{1}{k})$$

$$= \frac{a}{\sqrt{\pi}} e^{-\log q - a^2(\log\frac{q}{km} + \frac{1}{4a^2})^2} .dq \quad \ldots\ldots\ldots (22)$$

By comparison with No. (3), the distribution of "q" is, therefore, also lognormal with identical parameter "a" and mean = mk. Multiplication of the variable x by a constant factor, therefore, has no effect on the relative shape of the curve and merely changes the mean of the distribution in the same proportion as the change in the individual "x" values.

It follows, therefore, that the unit in which the lognormal variable "x" is expressed has no effect on the parameter "a" of the distribution.

9. Area under the Curve, (i.e. Frequency), and the Average of all "x" Values above any "x" Value

The area under the curve, i.e. the frequency of "x" values above an "x" value of, say, "x_1" will be

$$\int_{x=x_1}^{x=\infty} ydx = \int_{x_1}^{\infty} \frac{a}{\sqrt{\pi}} e^{-\log x - a^2(\log\frac{x}{m} + \frac{1}{4a^2})^2} .dx$$

Substituting/...

Substituting $w = a\sqrt{2}(\log\frac{x}{m} + \frac{1}{4a^2})$

and $w_1 = a\sqrt{2}(\log\frac{x_1}{m} + \frac{1}{4a^2})$

$$\text{the area} = \frac{1}{\sqrt{2\pi}} \int_{w_1}^{\infty} e^{-\frac{w^2}{2}} . dw \qquad \ldots\ldots\ldots (23)$$

The average value of all "x" values above an "x" value of "x_1"

$$= m_{x_1} = \frac{\int_{x_1}^{\infty} xydx}{\int_{x_1}^{\infty} ydx} \quad *$$

which with the same substitution as above reduces to

$$m_{xi} = m. \frac{\int_{w_1 - \frac{1}{a\sqrt{2}}}^{\infty} e^{-\frac{w^2}{2}} . dw}{\int_{w_i}^{\infty} e^{-\frac{w^2}{2}} . dw} \qquad \ldots\ldots\ldots (23a)$$

This can be solved by the use of standard tables of the integral of $(\frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}})$ which are available in almost any textbook on statistics.

10. Sampling (in the statistical sense) from a Lognormal Population.

(a) Distribution of arithmetic means of sets of sample values: Now consider the process of drawing an infinite number of sets of N random sample values each from a lognormal population. The means of such sets of values will in turn yield/...

---

* Ref. 2.

yield a new frequency distribution with an overall mean equal to that of the parent population. In the case of an arbitrary population this distribution of means will have the same general form as that of the parent population, but its variance will be less, its skewness much less and its kurtosis very much less than that of the parent population.* Further, as the number of samples per set, N, is increased, the skewness and kurtosis will approach the corresponding values for the normal curve. The position in the case of the lognormal parent population can now be examined from the knowledge of the formulae** applicable to arbitrary populations:-

$$\text{Variance} = \frac{\text{Variance of parent population}}{\text{Number of values per set}}$$

which from No. (9)

$$= \frac{m^2(e^{2a^2} - 1)}{N} \qquad \ldots\ldots\ldots (24)$$

where "m" and "a" are the parameters of the parent population and N = number of samples per set.

The 3rd moment (about the mean in standard units)

$$= \frac{\text{3rd moment of parent population}}{\sqrt{N}}$$

which from No. (13)

$$= \frac{(e^{2a^2} - 1)^{\frac{1}{2}}(e^{2a^2} + 2)}{\sqrt{N}} \qquad \ldots\ldots\ldots (25)$$

The 4th moment (about the mean in standard units)

$$= 3 + \frac{1}{N}(\text{4th moment of parent population} - 3)$$

which from No. (14)

$$= 3 + \frac{1}{N}(e^{2a^2} - 1)(e^{2a^2} + 3e^{a^2} + 6e^{2a^2} + 6) \qquad \ldots\ldots\ldots (26)$$

$$= 3 + \frac{1}{N}(e^{a^2} + 2e^{2a^2} + 3e^{a^2} - 6) \qquad \ldots\ldots\ldots (27)$$

An analysis of formulae Nos. (24) to (27) indicates that the distribution of the means of sets of sample values

from/...

---

* Ref. 8, p. 108. ** Ref. 6, pp. 102/3

from a lognormal parent population is not itself truly lognormal, but more skew and peaked for the same variance. Tests carried out with observed lognormal populations have, however, indicated that for practical purposes, the distribution of the means of sets of sample values selected at random can be regarded as lognormal and that No. (24) above can be employed to arrive at its theoretical variance,

$$\text{i.e. } m(e^{2a_{x_1}^{\frac{1}{2}}} - 1) = \frac{m(e^{2a^{\frac{1}{2}}} - 1)}{N}$$

The parameter "a_{x_1}" can, therefore, be calculated and will specify the relative shape of the lognormal frequency curve which closely approximates the distribution of means of sets of N random samples drawn from the parent lognormal population with parameter "a."

(b) Distribution of geometric means of sets of sample values and of an "improved" estimate of the arithmetic mean: From No. (16) it was seen that the distribution of the logarithm of the lognormal variable "x" is "normal" and defined by

$$f(z).dz = \frac{1}{\sqrt{2\pi}.\sqrt{\frac{1}{2a^2}}} e^{-\frac{1}{2(\frac{1}{2a^2})}\left[z - (\log m - \frac{1}{4a^2})\right]^2} .dz \quad .... (28)$$

where $z = \log x$

Now the frequency distribution of the means ($\bar{z}_1$) of sets of N z values each, drawn from this normal population of the logs of x will also be normal,* with variance = $\frac{1}{2Na^2}$, and the same overall mean z value, and will thus be defined by

$$F(\bar{z}_1)d\bar{z}_1 = \frac{1}{\sqrt{2\pi}.\sqrt{\frac{1}{2Na^2}}} e^{-\frac{1}{2(\frac{1}{2Na^2})}\left[\bar{z}_1 - (\log m - \frac{1}{4a^2})\right]^2} .d\bar{z}_1 \quad ..(29)$$

where/...

---

* Ref. 6, p. 103, & Ref. 8, p. 231.

where $\bar{z}_i$ = mean of the logs of N x values in the i^{th} set of sample values

= logarithm of the geometric mean of these N "x" values

= log $\bar{g}_i$, where $\bar{g}_i$ = geometric mean of i^{th} set of sample values

and No. (29) then, by substitution, reduces to

$$Q(\bar{g}_i)d\bar{g}_i = \frac{1}{\sqrt{2\pi}.\sqrt{\frac{1}{2Na^2}}} e^{-Na^2\left[\log \bar{g}_i - \left(\log m - \frac{1}{4a^2}\right)\right]^2} - \log \bar{g}_i \, .d\bar{g}_i \quad \ldots\ldots\ldots (30)$$

Reference to Nos. (3) and (9) will indicate, therefore, that the geometric means of sets of N (lognormally distributed) sample values each, are also lognormally distributed with variance $= m^2(e^{\frac{1}{2Na^2}} - 1)$

and mean = $\log m - \frac{1}{4a^2}$

But in Nos. (21) and (21a) it was shown that the geometric and arithmetic means of a lognormal population are directly related through the parameter "a" of the population. Thus if the "a" of the parent population is known, it is possible to arrive at an estimate of the true population mean "m" from the calculated geometric mean "$\bar{g}_i$" of a set of sample values from the formula

$$\text{geometric mean} = (\text{arithmetic mean}).e^{-\frac{1}{4a^2}}$$

Therefore taking

$$\bar{g}_i = h_i e^{-\frac{1}{4a^2}} \quad \ldots\ldots\ldots (31)$$

where h_i = an estimate of the true arithmetic mean of the population derived from the i^{th} set of sample values.

Then/...

Then from (30)

$$f(h_i)dh_i = \sqrt{\frac{Na^2}{\pi}} e^{-Na^2(\log h_i - \log m)^2 - \log h_i} .dh_i \quad \ldots (32)$$

The mean of this distribution of h_i

$$= \int_0^{\infty} f(h_i).h_i.dh_i$$

$$= \sqrt{\frac{Na^2}{\pi}} \int_0^{\infty} e^{-Na^2(\log h_i - \log m)^2} .dh_i$$

Substituting $w = a\sqrt{N}(\log h_i - \log m) - \frac{1}{2a\sqrt{N}}$

the mean of the distribution

$$= me^{\frac{1}{4Na^2}} . \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{w^2}{2}} .dw$$

$$= me^{\frac{1}{4Na^2}} \quad \ldots\ldots (33)$$

The mean of the distribution of h_i will, therefore, always exceed the true mean "m" of the parent population and consequently "h_i," as an estimate of "m" will be biassed, the bias factor $e^{\frac{1}{4Na^2}}$ only disappearing when N approaches infinity. The bias is very evident in the case of N = 1 when the observed geometric mean must be the same as the observed arithmetic mean, i.e. "$\bar{g}_i$" must equal "h_i," whereas "h_i" then becomes equal to

$$\bar{g}_i e^{\frac{1}{4a^2}}$$

This suggests that "h_i" must be corrected by a factor dependent on N and equivalent to $e^{-\frac{1}{4a^2}}$ when N = 1 and

to/...

to 1 when $N = \infty$. Similar corrections are required in most statistical estimates based on a limited number of sample values, e.g. Bessel's correction.*

Division by the bias factor in No. (33), i.e. $e^{\frac{1}{4Na^2}}$, provides this correcting term and the corrected estimate of the population mean now becomes

$$\bar{m}_1 = \frac{\bar{g}_1 . e^{\frac{1}{4a^2}}}{e^{\frac{1}{4Na^2}}} = \bar{g}_1 e^{\frac{1}{4a^2}(1 - \frac{1}{N})} \qquad \ldots\ldots\ldots (34)$$

from which it is evident that when

$N = 1$, $\bar{m}_1 = \bar{g}_1$

and when $N \to \infty$, $\bar{m}_1 \to \bar{g}_1 . e^{\frac{1}{4a^2}}$

As a check on the unbiassed nature of the statistic "$\bar{m}_1$" as an estimate of "m," substitution in No. (30) yields

$$F(\bar{m}_1)d\bar{m}_1 = \frac{a\sqrt{N}}{\sqrt{\pi}} \int_0^{\infty} e^{-\log \bar{m}_1 - Na^2(\log \bar{m}_1 - \log m + \frac{1}{4Na^2})^2} . d\bar{m}_1 \qquad \ldots\ldots\ldots (35)$$

From No. (3) it will be seen that the distribution of "$\bar{m}_1$" is lognormal with mean = "m," i.e. the true mean of the parent population, and the other parameter = $a\sqrt{N}$ (corresponding to "a" of the parent population). The statistic "$\bar{m}_1$" is, therefore, an unbiassed estimate of the population mean and from No. (9) the variance of the distribution of "$\bar{m}_1$" will be

$$= m^2(e^{\frac{1}{2Na^2}} - 1) \qquad \ldots\ldots\ldots (36)$$

Efficiency of "$\bar{m}_1$:"

When the variance of a normal population is known, the maximum likelihood estimate of the population mean is the mean of the set of sample values drawn from the population**

and/...

*Ref. 6, p. 125. **Ref. 8, p.273.

and, therefore, since the derivation of "$\bar{m}_1$" is based on the means ($\bar{g}_1$) of sets of sample values from the normally distributed variable log x = z, (see No. (28)), "$\bar{g}_1$" and "$\bar{m}_1$" will be maximum likelihood estimates of the means of the "z" and "x" populations, respectively.

The efficiency of "$\bar{m}_1$" relative to that of the straight mean "$\bar{x}_1$" of a set of sample values from a lognormal distribution can now be gauged as follows:-

$$\text{Variance of } \bar{x}_1 = \frac{m^2(e^{\frac{1}{2a^2}} - 1)}{N} \quad \text{... from No. (24)}$$

$$\text{Variance of } \bar{m}_1 = m^2(e^{\frac{1}{2Na^2}} - 1) \quad \text{... from No. (36)}$$

where "m" = mean of parent population

and "a" = parameter of parent population

Therefore when the above two variances are equal the arithmetic mean "$\bar{x}_1$" of a set of sample values will be as "reliable" as the statistic "$\bar{m}_1$," and this will occur when

N_a = the number of samples employed on the arithmetic mean basis

$$= \frac{m^2(e^{\frac{1}{2a^2}} - 1)}{m^2(e^{\frac{1}{2N_g a^2}} - 1)}$$

$$= \frac{e^{\frac{1}{2a^2}} - 1}{e^{\frac{1}{2N_g a^2}} - 1} \quad \ldots\ldots (37)$$

(N_g being = number of samples employed in estimating the true mean from "$\bar{m}_1$.")

From No. (37) it is obvious that when $N_a = 1$, $N_g = 1$ and the two bases will thus yield the same result; and that when $N_a > 1$, the "$\bar{m}_1$" basis will always require a smaller number of samples to yield results as reliable as the straight arithmetic mean basis, i.e. $N_g < N_a$. The extent of the

improvement/...

improvement of the former basis on the latter is illustrated in the following table for a range of "a's" commonly met in practice.

TABLE 2

N_g	a = .5	a = .6	a = .7	a = .8	a = .9	a = 1.0
	N_a	N_a	N_a	N_a	N_a	N_a
2	4	3	3	2	2	2
5	13	9	8	7	6	6
10	29	19	17	15	13	12
20	61	40	34	30	27	25
50	157	102	86	75	69	65
100	316	206	173	151	138	129
500	1,593	1,037	870	759	694	642
1,000	3,195	2,066	1,740	1,518	1,377	1,297

Taking an example from the above table, sets of 10 sample values each drawn from a population having an "a" of .5 will on the "$\bar{m}_1$" basis yield results as "reliable"as sets of 29 sample values each on the usual arithmetic mean basis.

The values of "a" normally encountered in the distribution of individual underground values range from about .5 to .8, and thus the "$\bar{m}_1$" basis will roughly yield results equivalent to those obtained from 2½ to 3 times as many samples on the customary arithmetic mean basis.

The effort involved in estimating the true mean value on the "$\bar{m}_1$" basis is relatively small and the method straight forward, involving only the calculation of the geometric mean and its multiplication by $e^{\frac{1}{4a^2}(1 - \frac{1}{N})}$, where N is known and "a" can in many cases be determined by experiment as shown in Chapter V.

(c) Distribution/...

(c) Distribution of the variances of sets of sample values drawn from a lognormal population: For this purpose only "large" sets of samples will be considered, (say $\geq$100 per set) and the problem will be approached from the angle of the "normal" distribution formed by the logs of the variable. In the case of a "normal" parent population the individual variances of large sets of samples drawn from it, can be regarded as normally distributed with standard deviation

$$= \sqrt{\frac{1}{2a_{x_1}^2}} = \sqrt{\frac{2}{N}} \cdot (\text{variance of parent population})^*$$

$$= \sqrt{\frac{2}{N}} \cdot \left(\frac{1}{2A^2}\right) \qquad \ldots\ldots \text{(see No. (18))}$$

where N = number of samples per set

A = parameter "a" of parent population

and a_{x_1} = estimate of parameter "A" from the calculated variance $\frac{1}{2a_{x_1}^2}$ of the logs of the sample values in the i^{th} set of samples.

For confidence limits of 0.95, i.e. if a 1 in 20 chance of an extremely low or high value of $\frac{1}{2a_{x_1}^2}$ is disregarded,

$$\frac{\frac{1}{2a_{x_1}^2} - \frac{1}{2A^2}}{\frac{1}{2A^2} \cdot \sqrt{\frac{2}{N}}} = \pm 1.96^{**}$$

where $\frac{1}{2a_{x_1}^2}$ = variance of logs of values in the i^{th} set of samples

from which

$$a_{x_i} = A \cdot \sqrt{\frac{1}{1 \pm 1.96\sqrt{\frac{2}{N}}}} \qquad \ldots\ldots\ldots (38)$$

There/...

---

*Ref. 8, p. 289. **Ref. 8, p.290.

There is therefore a 95%, or 19 in 20, chance of the value of the parameter "a_{x_i}" as determined (on the logarithmic basis) from a set of N sample values drawn from a lognormal population, lying between the limits defined by No. (38). The following is a table of the percentage maximum variation in the parameter "a_{x_i}" for 95% confidence limits and specific values of N.

TABLE 3

N	$\sqrt{\frac{1}{1 + 1.96\sqrt{\frac{2}{N}}}}$	$\sqrt{\frac{1}{1 - 1.96\sqrt{\frac{2}{N}}}}$
30,000	99.2%	100.8%
20,000	99.0%	101.0%
15,000	98.9%	101.2%
10,000	98.6%	101.4%
7,500	98.4%	101.6%
5,000	98.1%	102.0%
4,000	97.9%	102.3%
3,000	97.6%	102.6%
2,000	97.0%	103.3%
1,500	96.6%	103.8%
1,000	95.9%	104.7%
900	95.7%	105.0%
800	95.4%	105.3%
700	95.1%	105.7%
600	94.8%	106.2%
500	94.3%	106.8%
400	93.7%	107.7%
300	92.8%	109.1%
200	91.4%	111.5%
100	88.5%	117.6%

e. g. /...

e.g., there is a 19 in 20 chance of the value of "a_{x_1}" determined from a set of 500 samples lying within the range of 94.3% to 106.8% of "A," the true parameter of the parent population.

11. Combination of Lognormal Subpopulations with Identical Parameters "a," and Lognormally Distributed Means.

The hypothesis that the combination of such subpopulations will itself constitute a lognormal population, at any rate for all practical purposes, has been confirmed by test calculations which need not be reproduced here.

Consider, therefore, a parent population with parameter "A" and mean "M," consisting of "k" subpopulations each with parameter "a" and with means m_1, m_2, m_3 ... m_k, which are lognormally distributed with parameter "a_m" (and mean M). Then, since

Variance of parent population

= weighted average variance within subpopulations + variance between means of subpopulations* ..... (39)

$$M^2(e^{\frac{1}{2A^2}} - 1) = \frac{m_1^2(e^{\frac{1}{2a^2}} - 1) + m_2^2(e^{\frac{1}{2a^2}} - 1) + \dots}{k} + M^2(e^{\frac{1}{2a_m^2}} - 1)$$

$$= (e^{\frac{1}{2a^2}} - 1)\left(\frac{m_1^2 + m_2^2 + \dots m_k^2}{k}\right) + M^2(e^{\frac{1}{2a_m^2}} - 1)$$

But $\frac{m_1^2 + m_2^2 + \dots m_k^2}{k}$ = 2nd moment of "m_1" about origin

$$= M^2 e^{\frac{1}{2a_m^2}} \quad \text{... from No. (8)}$$

$$\therefore M^2(e^{\frac{1}{2A^2}} - 1) = M^2 e^{\frac{1}{2a_m^2}}(e^{\frac{1}{2a^2}} - 1) + M^2(e^{\frac{1}{2a_m^2}} - 1)$$

which reduces to

$$\frac{1}{A^2} = \frac{1}{a_m^2} + \frac{1}{a^2} \qquad \dots\dots (40)$$

12. The/...

*Ref. 4, p. 101.

12. <u>The Relationship between a Parent Lognormal Population and the Means of Sets of Sample Values drawn from the Lognormal Subpopulations stipulated in No. 11 above</u>.

Consider now the case where an infinite number of sets of N samples each are drawn from one of these subpopulations. The means of these sets of sample values can, from paragraph 10(a) above, be regarded as a lognormal distribution with

$$\text{Variance} = \frac{m_1^2(e^{\frac{1}{2a^2}} - 1)}{N}$$ where m_1 = mean of 1^{th} subpopulation,

and the weighted average variance of the series of distributions of means if the procedure is applied to each of the above subpopulations will be

$$\frac{(e^{\frac{1}{2a^2}} - 1)(m_1^2 + m_2^2 + \ldots m_k^2)}{N \quad k}$$

$$= \frac{(e^{\frac{1}{2a^2}} - 1)}{N} \cdot M^2 e^{\frac{1}{2a_m^2}}$$

Now, regarding for a moment each distribution of means as a subset and the distribution of all these means as the parent population with parameter "a_x," then from N (39)

Variance of all the means = Variance between means of subsets + weighted average variance of the means within the subsets,

and since the means of the subsets are identical with the means of the subpopulations,

$$M^2(e^{\frac{1}{2ax^2}} - 1) = M^2(e^{\frac{1}{2a_m^2}} - 1) + M^2 e^{\frac{1}{2a_m^2}} \cdot (\frac{e^{\frac{1}{2a^2}} - 1}{N})$$

and $$e^{\frac{1}{2a^2}} - 1 = N(e^{\frac{1}{2a_x^2}} - e^{\frac{1}{2a_m^2}} - 1)$$

But/...

But from (40) above

$$\frac{1}{2a_m^2} = \frac{1}{2A^2} - \frac{1}{2a^2}$$

$$\therefore e^{\frac{1}{2a^2}} - 1 = N(e^{\frac{1}{2a_x^2} - \frac{1}{2A^2} + \frac{1}{2a}} - 1)$$

whence $$e^{\frac{1}{2a^2}} = \frac{N - 1}{Ne^{\frac{1}{2a_x^2} - \frac{1}{2A^2}} - 1} \qquad \ldots\ldots\ldots (41)$$

$$= \frac{1 - \frac{1}{N}}{e^{\frac{1}{2a_x^2} - \frac{1}{2A^2}} - \frac{1}{N}} \qquad \ldots\ldots\ldots (41a)$$

The following serves as a practical check on the derivation of (41a):-

When N becomes large, a set of samples drawn from a subpopulation will approximate the subpopulation itself, "a_x" will then approximate "a_m," and (41a) should merge into the identity (40). This is evident since $\frac{1}{N}$ approaches zero as N approaches ∞, and then

$$e^{\frac{1}{2a^2}} = \frac{1}{e^{\frac{1}{2a_x^2} - \frac{1}{2A^2}}}$$

and hence $\frac{1}{A^2} = \frac{1}{a^2} + \frac{1}{a_x^2}$

which is identical with (40) when "a_m" is substituted for "a_x."

13. Fitting the Lognormal Curve: Improvements in the Estimate of the Population Mean.

(a) Fitting by moments of observed values: The general type of lognormal curve with three parameters can be fitted by solving the equations for the mean, variance and 3rd moment

simultaneously/...

simultaneously.* In the case of the lognormal curve as applied to gold values, there are, however, only two parameters, i.e. the mean "m," and "a." Fitting by moments would consequently only necessitate the substitution of the observed mean for "m" in the formula, and solving for "a" in the expression for the variance,

$$\text{i. e. } m^2(e^{2a^2} - 1) = \text{observed variance.}$$

This method will, obviously, not yield any "improvement" in the estimate of the population mean since the mean of the observed values is accepted as being equivalent to the population mean.

(b) Fitting by moments of logs of values: Since the distribution of the logs of a lognormal variable is normal,** it is possible to fit a normal curve to the logs of the observed values by the customary method of equating the mean of the set of log values, (i.e., in this case the mean of logs of observed values), to the log population mean and equating the calculated variance of the logs of the observed values to the variance of the log population.*** The parameters of the lognormal population can then be obtained by substitution in formulae (17) and (18). It can be shown in the following manner, that this method yields a considerable improvement in the estimate of the mean of the lognormal parent population.

The means of the logs of sets of N values each from a lognormal population are distributed normally with variance

$= \frac{1}{2NA^2}$**** where $\frac{1}{2A^2}$ = variance of parent normal population.

Where N is large, (say, > 30), the variances of the logs/...

---

*Ref. 1. **Par. 7 above. ***With Bessel's correction where required.

****Par. 10(b) above.

logs of such sets of values will also be distributed approximately normally with variance

$$= (\frac{1}{2A^2})^2 (\frac{2}{N})^*$$

and therefore the estimates of the variance of the parent normal population (i.e. $\frac{1}{2a_{x_i}^2}$) after applying Bessel's

correction, i.e. $(\frac{1}{2a_{x_i}^2})(\frac{N}{N-1})$ ..... (42)

(where $\frac{1}{2a_{x_i}^2}$ = calculated variance of logs of set of values)

will also be distributed normally with variance

$$= (\frac{2}{4NA^4}) \cdot \frac{N^2}{(N-1)^2}^{**} \quad \text{..... (42a)}$$

Referring now to No. (34) the factor

$$\frac{1}{4A^2}(1 - \frac{1}{N}) = \frac{1}{2A^2}(\frac{N-1}{2N}).$$

$$= (\frac{N-1}{2N}) \cdot \text{normal population variance}$$

Substituting now the <u>estimate</u> of the normal population variance provided by (42a), the <u>estimate</u> of the factor $\frac{1}{4A^2}(1 - \frac{1}{N})$ becomes

$$(\frac{1}{2a_{x_i}^2})(\frac{N}{N-1})(\frac{N-1}{2N}) = \frac{1}{4a_{x_i}^2} \quad \text{..... (43)}$$

and will be distributed normally with variance

$$= \frac{2N^2}{4NA^4(N-1)^2} \cdot \frac{(N-1)^{2**}}{4N^2}$$

$$= \frac{1}{8NA^4} \quad \text{..... (44)}$$

Now/...

---

*Ref. 8, p. 289. ** Based on Ref. 6, p.125 & Ref. 9, p.25.

Now, the estimate of the mean of the lognormal parent population provided by No. (34)

$$= \bar{m}_1 = \bar{g}_1 e^{\frac{1}{4A^2}(1 - \frac{1}{N})}$$

is lognormally distributed and therefore

$$\log \bar{m}_1 = \log \bar{g}_1 + \frac{1}{4A^2}(1 - \frac{1}{N})$$

$$= \left.\begin{matrix} \text{mean of logs of set of} \\ \text{observed values} \end{matrix}\right) + \frac{1}{4A^2}(1 - \frac{1}{N}) \quad \ldots\ldots \quad (45)$$

is normally distributed. Now, where the "A" of the parent population is unknown, and is estimated from the logs of the set of observed values, the estimate of the factor $\frac{1}{4A^2}(1 - \frac{1}{N})$ will be distributed normally* with variance defined by No. (44) above.

Since the first factor in No. (45) above (mean of logs of values) is also distributed normally with variance $= \frac{1}{2NA^2}$** the new estimate of the log of the lognormal population mean "m," i.e. $\log p_i = \log \bar{g}_i + \frac{1}{4ax_i^2}$ (from Nos. (45) and (43) above) ..... (46)

will be normally distributed with

$$\text{Variance} = \frac{1}{2NA^2} + \frac{1}{8NA^4}$$

and the corresponding estimate "p_i" of the lognormal population mean will be distributed lognormally with variance***

$$= m^2(e^{\frac{1}{2NA^2} + \frac{1}{8NA^4}} - 1) \quad \ldots\ldots \quad (47)$$

The statistic "p_i" will then (from No. (46))

$$= (\text{geometric mean of sample values})(e^{\frac{1}{4ax_i^2}}) \quad \ldots\ldots \quad (48)$$

where $\frac{1}{2ax_i^2}$ = calculated variance of logs (to the base e) of the set of sample values.

The/...

*Provided N = large. **Par. 10(b) above. ***No. (9) above.

The relative efficiency of "p_1" can now be calculated on the same basis as that of "$\bar{m}_1$" in paragraph 10(b) above, by calculating the equivalent number of sample values required on the orthodox arithmetic mean method to yield on average as reliable a result as the statistic "p_1." In the following table these numbers of equivalent samples are listed for various values of the parameter "A" of the parent population.

TABLE 4

N_p	A = .5	A = .6	A = .7	A = .8	A = .9	A = 1.0
	N_a	N_a	N_a	N_a	N_a	N_a
50	77	62	57	53	52	52
100	157	126	114	108	105	103
500	799	641	574	541	527	519
1,000	1,597	1,281	1,152	1,086	1,054	1,038

Where N_a = Number required on arithmetic mean basis

and N_p = number required using statistic "p_1"

e.g., where "A" = .5, 50 samples on the "p_1" basis will on average yield as reliable a result as 77 samples on the orthodox arithmetic mean basis.

The improvement indicated above is by no means as marked as that obtained from "$\bar{m}_1$" (paragraph 10(b) above) when "A" is known, but is still sufficient over the range of "A" usually encountered, (i.e. .5 to .8) to warrant its employment.

The above method of fitting the lognormal curve and of obtaining an improvement in the estimate of the population mean can naturally be applied for all values of "N" but Table 4 is only applicable for large values of N exceeding, say, 30 since the derivation of (42a) is based on this

assumption ...

assumption. The determination of the improvement provided by "p_i" for smaller values of "N," falls beyond the scope of this thesis.

Practical examples of the application of this method are provided in Chapter VI.

(c) Fitting by the Theory of Maximum Likelihood: This method has been developed for the lognormal curve by Sichel* and Finney** and is based, as in the case of the method under (b) above, on the calculated mean and variance of the logs of the set of sample values concerned. The solution provided by the Theory of Maximum Likelihood, however, maximises the combined probability of obtaining the observed mean and variance in a set of logs of sample values, and thus provides a "better" estimate than that under (b) above.

This method requires the solution of an infinite series which involves a fair amount of calculation to arrive at a sufficiently close result, or alternatively, the use of tables of Bessel's functions, and the overall theoretical improvement obtained is for all practical purposes, the same as that under (b) above. Three examples quoted by Sichel for sets of 10 samples each have been solved using the method under (b) above, and the results obtained differed from his to the maximum extent of only a few percent.

The use of this more complicated method appears, therefore, not to be justified from a practical point of view, except possibly in particular cases, e.g. borehole values, where the time factor in calculation is of little consequence.

14. Lognormal Correlation.

As shown in paragraph 7 above, the distribution of the logarithms of a lognormal variable is normal and it is, therefore/...

---

*Ref. 5. **Ref. 7.

therefore, evident that a normal correlation surface of the logs of two joint lognormal variables can be transposed into a lognormal correlation surface of these two variables. The problem of the lognormal correlation surface will consequently be approached from the angle of the relatively simple normal correlation surface formed by the logs of the variables, and only the ideal case of the normal correlation surface with homoscedastic regression system* and linear regression will be considered.

Now let:-

x and y = two joint lognormal variables

z and t = the corresponding joint normal variables, i.e. log x and log y, respectively

$\bar{x}$ and $\bar{y}$ = means of the two lognormal populations

$\bar{z}$ and $\bar{t}$ = means of the two normal populations of "z" and "t"
= logs of the geometric means of the two lognormal populations of "x" and "y" (see paragraph 7)

a_y = parameter of the lognormal population of "y" and of the normal population of "t"

a_x = parameter of the lognormal population of "x" and of the normal population of "z"

a_{yx} = parameter of the lognormal distribution formed by each of the various arrays of "y's" in the x categories
= parameter of the normal distribution formed by each array of "t's"

a_{xy} = parameter of the lognormal distribution formed by each array of "x's"
= parameter of the normal distribution formed by each array of "z's"

From/...

*Ref. 4, p. 209.

From paragraph 7 above it is, therefore, evident that:-

The mean of all the "z's" = $\bar{z}$ = log of mean of the x's - $\frac{1}{4a_x^2}$

$$= \log \bar{x} - \frac{1}{4a_x^2} \quad \ldots\ldots\ldots (49)$$

The mean of all the "t's" = $\bar{t}$ = $\log \bar{y} - \frac{1}{4a_y^2}$ ..... (49a)

The mean of each array of "z's" = log of mean of corresponding array of x's - $\frac{1}{4a_{xy}^2}$ ...(50)

The mean of each array of "t's" = log of mean of corresponding array of y's - $\frac{1}{4a_{yx}^2}$ ...(50a)

Further, as the <u>line of regression of "t" on "z"</u> is formed by the series of means of arrays of "t's" it will correspond to the series of (logs of the means of arrays of y's - $\frac{1}{4a_{yx}^2}$). It is, therefore, evident that <u>the line (or curve) of regression of "y" on "x"</u> will on the logarithmic graph of "t" on "z," plot as a straight line parallel to the line of regression of "t" on "z" at a "vertical" distance, (i.e. parallel to the y axis) of + $\frac{1}{4a_{yx}^2}$, from the latter. ..... (51)

Similarly, the <u>line or curve of regression of "x" on "y"</u> will on the logarithmic graph of "z" on "t" plot as a straight line parallel to the line of regression of "z" on "t" at a distance parallel to the x axis of + $\frac{1}{4a_{xy}^2}$ from the latter. ..... (51a)

It should be noted at this stage that only a 45^o line on double logarithmic graph paper will convert into a straight line on ordinary graph paper, and that the latter line will always pass through the origin, e.g.,

log y/...

$\log y = \log x + \log S$ is a 45° line on double logarithmic paper converting into

$y = Sx$ on ordinary graph paper

where S = a constant

Any other straight line on double logarithmic paper will convert into a curve on ordinary graph paper, e.g.,

$\log y = p \log x + \log S$ (where p = slope $\neq 45^\circ$)

converts to

$$y = {}_e p \log x + \log S$$

$$= {}_e \log S \, {}_e p \log x$$

$$= K {}_e p \log x \quad \text{(i.e. a type of exponential curve)}$$

Line of regression $y = Sx$:

Now consider a line of regression of "y" on "x" of

$y = Sx$

converting to

$\log y = \log x + \log S$

$t = z + \log S$

∴ from the conclusions reached above (No. (51)) the line of regression of "t" on "z" will be

$$t = z + \log S - \frac{1}{4a_{yx}^2} \qquad \ldots\ldots (53)$$

with slope = 1

But the slope of the line of regression of "t" on "z" in general

$$= \frac{\text{Standard deviation of } t}{\text{Standard deviation of } z} \text{(Correlation Coefficient)*}$$

$$= \frac{\sqrt{\frac{1}{2a_y^2}}}{\sqrt{\frac{1}{2a_x^2}}} \cdot r \qquad \ldots \text{ (from Nos. (17) and (18))}$$

$$= \frac{a_x}{a_y} \cdot r$$

when/...

*Ref. 4, p. 177.

when the slope = 1, therefore, the coefficient of correlation

$$r = \frac{a_y}{a_x} \qquad \ldots\ldots\ldots (54)$$

Line of regression z on t:

Further, the general formula for the slope of the line of regression of z on t is

$$\frac{\text{Standard deviation of } z}{\text{Standard deviation of } t}.r \text{ *}$$

$$= \frac{a_y}{a_x}.r \text{ and from No. (54) this}$$

$$= \frac{a_y^2}{a_x^2} \qquad \ldots\ldots\ldots (55)$$

This line of regression will, therefore, be

$$z = \frac{a_y^2}{a_x^2}t + K \qquad \ldots\ldots\ldots (55)$$

and it must pass through the point $\bar{z},\bar{t}$ which, from Nos. (49) and (49a) can be stated as

$$\log \bar{x} - \frac{1}{4a_x^2} \text{ , } \log \bar{y} - \frac{1}{4a_y^2}$$

Solving for K in No. (56) yields

$$K = \log \bar{x} - \frac{a_y^2}{a_x^2}.\log \bar{y}$$

and No. (56) becomes

$$z = \frac{a_y^2}{a_x^2}t + \log \bar{x} - \frac{a_y^2}{a_x^2}.\log \bar{y} \qquad \ldots\ldots\ldots (57)$$

Now, the variance of each array of z's

$$= \frac{1}{2a_{xy}^2} = \text{Variance of z's}(1 - r^2) \text{ **} \quad \text{which from No. (54)}$$

$$= \frac{1}{2a_x^2}(1 - \frac{a_y^2}{a_x^2}) \qquad \ldots\ldots\ldots (58)$$

Consequently, from No. (51a) above, the line or curve of regression of x on y can be obtained from

$$z = \frac{a_y^2}{a_x^2}t + \log \bar{x} - \frac{a_y^2}{a_x^2}.\log \bar{y} + \frac{1}{4a_{xy}^2} \qquad \ldots\ldots\ldots (59)$$

i.e. $\log x'$...

---

*Ref. 4, p. 177. **Ref. 4, p. 180.

i.e. $\log x = \frac{a_y^2}{a_x^2}\log y + \log \bar{x} - \frac{a_y^2}{a_x^2}\log \bar{y} + \frac{1}{4a_x^2}(1 - \frac{a_y^2}{a_x^2})$ .... (59a)

and where in a special case $\bar{x} = \bar{y}$

$$\log x = \frac{a_y^2}{a_x^2}\log y + (\log \bar{x} + \frac{1}{4a_x^2})(1 - \frac{a_y^2}{a_x^2}) \quad \text{.....} \quad (59b)$$

It is evident, therefore, that in the case of a log-normal correlation surface, a straight line of regression of y on x of the type $y = Sx$ will have a corresponding curve of regression of x on y as defined by No. (59). This property finds particular application in the subsequent examination of bias errors in mine valuation, (Chapter VI, paragraph 3).

CHAPTER IV/...

CHAPTER IV

A PRACTICAL GRAPHICAL METHOD OF CURVE FITTING FOR LOGNORMAL DISTRIBUTIONS

1. Transposition of the Lognormal Curve into a Straight Line.

From No. (3) above, the lognormal frequency function is

$$F(x)dx = \frac{a}{\sqrt{\pi}} e^{-\log x - a^2(\log\frac{x}{m} + \frac{1}{4a^2})^2} .dx$$

substituting $w = a\sqrt{2}(\log\frac{x}{m} + \frac{1}{4a^2})$ ..... (60)

$$\text{and } dx = \frac{x}{a\sqrt{2}}dw$$

$$f(w)dw = \frac{a}{\sqrt{\pi}} e^{-\log x - \frac{w^2}{2}} \cdot \frac{x}{a\sqrt{2}}dw$$

$$= \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} .dw \qquad \text{..... (61)}$$

which is the expression for the normal curve symmetrical about the line w = o.

$$\text{Also } \int_0^{\infty} F(x)dx = \int_{-\infty}^{+\infty} f(w)dw$$

(since when x = o, w = $-\infty$ and when x = ∞, w = ∞)

$$= 2\int_0^{\infty} f(w)dw \qquad \text{..... (62)}$$

Now $\int_0^{\infty} f(w)dw$ is usually listed in tables for the normal curve as equivalent to 0.5 since it only represents a half/...

half of the total area under the normal curve. Also as $\int_0^{w_1} f(w)dw$, i.e. the area under the curve betwsen values for "w" of zero and of "w_1," is measured (in these tables) from the median (= mode = mean) of the curve, whereas the cumulative frequency on the lognormal curve is measured from zero "x," i.e. the left-hand extremity of the curve, the following frequencies will correspond:-

Cumulative Frequency Normal Curve $\int_0^{w} f(w)dw$	Cumulative Frequency Lognormal Curve $\int_0^{x} F(x)dx$	Value of "w"
-0.5	0%	$-\infty$
0	50%	0
+0.5	100%	$+\infty$
-0.3 (i.e. 0 - .3)	20% (i.e. 50% - 30%)	.842*
+0.4 (i.e. 0 + .4)	90% (i.e. 50% + 40%)	1,281*

The fact that when "w" = o, the cumulative frequency on the lognormal curve = 50%, can also be illustrated as follows:-

When "w" = o, from No. (60)

$$\log x = \log u - \frac{1}{4a^2}$$

$$x = me^{-\frac{1}{4a^2}} = \text{median of lognormal curve (see No.(4))}$$

= value corresponding to a 50% cumulative frequency.

Thus/...

*Explanation given in subsequent discussions.

TABLE 5

Cum. F. %	"$\frac{w}{\sqrt{2}}$"	Cum. F. %	"$\frac{w}{\sqrt{2}}$"	Cum. F. %	"$\frac{w}{\sqrt{2}}$"	Cum. F. %	"$\frac{w}{\sqrt{2}}$"
0	$-\infty$	25	-0.4770	51	0.0177	77	0.5225
.5	-1.8214	26	-0.4549	52	0.0355	78	0.5461
1.0	-1.6450	27	-0.4333	53	0.0532	79	0.5702
2.0	-1.4522	28	-0.4121	54	0.0710	80	0.5951
3.0	-1.3299	29	-0.3913	55	0.0889	81	0.6208
4.0	-1.2379	30	-0.3708	56	0.1067	82	0.6472
5.0	-1.1631	31	-0.3506	57	0.1247	83	0.6747
6.0	-1.0994	32	-0.3307	58	0.1428	84	0.7032
7.0	-1.0436	33	-0.3111	59	0.1609	85	0.7329
8.0	-0.9936	34	-0.2917	60	0.1791	86	0.7641
9.0	-0.9481	35	-0.2725	61	0.1975	87	0.7965
10.0	-0.9062	36	-0.2535	62	0.2160	88	0.8308
11	-0.8673	37	-0.2347	63	0.2347	89	0.8673
12	-0.8308	38	-0.2160	64	0.2535	90	0.9062
13	-0.7965	39	-0.1975	65	0.2725	91	0.9481
14	-0.7641	40	-0.1791	66	0.2917	92	0.9936
15	-0.7329	41	-0.1609	67	0.3111	93	1.0436
16	-0.7032	42	-0.1428	68	0.3307	94	1.0994
17	-0.6747	43	-0.1247	69	0.3506	95	1.1631
18	-0.6472	44	-0.1067	70	0.3708	96	1.2379
19	-0.6208	45	-0.0889	71	0.3913	97	1.3299
20	-0.5951	46	-0.0710	72	0.4121	98	1.4522
21	-0.5702	47	-0.0532	73	0.4333	99	1.6450
22	-0.5461	48	-0.0355	74	0.4549	99.5	1.8214
23	-0.5225	49	-0.0177	75	0.4770	100	$+\infty$
24	-0.4994	50	$\pm$0.0000	76	0.4994		

Thus
indirect measu
curve, and the
established di
Normal Curve.*
Thus w = funct
$= a\sqrt{2}$
or $\frac{w}{\sqrt{2}} = a$ log
This equation
$Y = KX +$
where $Y = \frac{w}{\sqrt{2}}$, X
K = "a,"
and C = $(\frac{1}{4a}$ –
As w
graphical strai

2. Logarithmic
Table
frequencies co
of areas of the
above.
Thus
ordinary graph
tive frequency
as illustrated
rithmic scale f
of the gold valu
frequencies as
lognormal distri

*E.g., Ref. 4,
***See map pocke