ISBN 0-905927-18-4

BOHS
British Occupational Hygiene Society
Technical Handbook Series No. 1

Some Applications of Statistics in Occupational Hygiene

by Peter Dewell

Science Reviews Ltd in association with
H & H Scientific Consultants, Leeds
1989

ISBN 0-905927-18-4

BOHS

British Occupational Hygiene Society
Technical Handbook Series No. 1

Some Applications of Statistics in Occupational Hygiene

by Peter Dewell, B.Sc., F.I.O.H.,
Consultant Occupational Hygienist,
Redditch, U.K.

Series Editor **Dr. D. Hughes, University of Leeds.**

Science Reviews Ltd in association with
H & H Scientific Consultants, Leeds
1989

H H
S C

H & H Scientific Consultants Ltd
in association with **Science Reviews Ltd**
P.O.Box MT27, Leeds LS17 8QP, U.K.; tel.: 0532 687189.

Series Editor: Dr D. Hughes, University of Leeds, U.K.

The statistical methods presented in this Handbook are offered in good faith. Users must satisfy themselves that the methods are appropriate for the applications to which they are put.

The views expressed are those of the Author and are not necessarily those of the the British Occupational Hygiene Society.

British Library Cataloguing in Publication Data

Dewell, Peter
Some applications of statistics in occupational hygiene
(British Occupational Hygiene Society technical handbook no. 1)
1. Industrial health. Hazards. Monitoring. Use of statistical analysis
I. Title II. Series
363.1'163

ISBN 0-905927-18-4

CONTENTS

FOREWORD

When the BOHS Technology Committee invited me to prepare a handbook on statistics for Occupational Hygienists the request seemed, initially, straight forward enough. However, after some consideration the conviction grew that yet another book on basic statistics would serve no useful purpose and this has certainly dictated the contents. The somewhat informal style is intended to involve the reader/user, inviting participation, both in thinking about the problems of occupational hygiene statistics and in tracing the mathematics.

As a result the opening chapters briefly cover the ideas of means, standard deviations, and confidence limits — all available in any statistical text. The remainder of the book is much more a personal account of some statistics which I have found to be useful in testing occupational hygiene data taking account of the distribution from which the sample was drawn, and the outcome which this may have on some common comparative tests.

Many aspects of statistics are omitted, including the χ^2 test and other non-parametric tests. Even the statistics which are included in the book are inevitably far from complete. Some statistical expressions are introduced without explanation, (although their context should allow their import to be grasped), and in a chapter entitled "Sample Sizes" there is no discussion of experimental design. It is perhaps only very fortunate hygienists who have the time to carry out well-designed sampling experiments, and it must be presumed that these few have also had the time to study the statistics of experimental design. But knowing something about the inaccuracies associated with sample sizes can be no bad thing. The important question of handling "zero" and "below detectable level" values is also omitted.

Too many hygienists (no blame attaches to them — it's a fault of the system) are engaged in one-off "fire-fighting" exercises, or at best salvaging what they can from a series of measurements. It is to these that the Handbook is mainly aimed, the author and the BOHS Technology Committee hoping that it will open the way to a better understanding of the nature of occupational hygiene measurements.

Understanding statistics (or any other topic) comes from their application. Applying the statistics is today a matter of equipment — no apology is offered for relying on computers or, at the very least, programmable calculators. The age of manual calculation, with cross-checks, must surely have passed.

I would like to take this opportunity to thank the BOHS Technology Committee (and my critics) for their continued support while I was preparing this book. I also thank colleagues and the Director of BCIRA for allowing me to use illustrative data.

17 Brotherton Avenue,
Redditch,
Worcs.

P. Dewell,
February 1989.

Notes on printing.

The text of the book has been prepared, with the exception of title pages and section titles, on an Amstrad PCW8512® computer using LocoScript 2©. The camera-ready copy was prepared on an Amstrad LQ3500® 24 pin printer, using Locomotive Software's® 24 Pin Printer Driver. Computer outputs were prepared on a Sharp PC1500® computer with printer/plotter interface.

1 STATISTICS, NUMBERS AND HANDLING THEM

1.1 Introductory Remarks

Statistics applied to Occupational Hygiene data are concerned with numbers, frequency of occurrences, concentrations, levels, distributions, tests of various kinds and so on. They are therefore no different from any other statistics, and like them they require facilities to handle numbers and (at the risk of stating the obvious) they require the numbers themselves. Too often occupational hygiene "data" consist of a single observation, perhaps as a gas detector reading. In extreme cases it might not be past imagining to think of thousands of pounds being spent on control measures, based on such slim evidence. Applying statistics to almost any OH data will soon convince you that the (statistical) confidence you can usually place on your data is meagre in the extreme. How sure are you that the new control system *has* reduced the noise? Is the new sampler really any different from the old one and is the sampling strategy you've used because it was set up ten years ago by your predecessor any good at all?

1.2 The Statistics

The statistics described in this book can be thought of as the minimum tool-kit needed by an occupational hygienist to be able to say "I have analysed the results statistically". You will find that, to some extent, the basics of statistics have been assumed – the book is concerned with applying the usually available statistics and extending them into areas not often covered in the more readily available standard texts. Some of the statistics proposed may not be "secure" (to use a legal term) but in the absence of anything better they will at least allow you to move beyond "the average concentration was x mg/m^3" in the interpretation and reporting of your results.

A most important point to remember is to state, when you are reporting some results which have been analysed statistically, which statistic has been used. This may not be an obvious point, since one assumes that the reader of your report will probably know what a "mean" looks like and has at least a feel for the meaning of standard deviation. This may not be the case if you say "applying the t-test to the results from the two instruments it can be seen that XYZ collects more than ABC". Which t-test? What was the significance level? Should you have applied the F- (variance ratio) test first to see if the variances were similar before proceeding to the t-test? If you did why didn't you say so? It becomes even more important to state the statistic used (perhaps even giving references) when you start using some of the more arcane procedures now available for analysing your data.

A final observation is that having yet another book on statistics on the bookshelf, or even reading it, will be of little value unless the statistics are practised on real, or in their absence, artificial data.

1.3 The Numbers

For sure if you don't have enough data there is no way you can carry out any sensible statistical analysis. You may get a mean from two results but you will be hard pressed to say what kind of distribution they come from. At the other end of the scale it is now easier than it has ever been to analyse masses of data which have accumulated in the files for the last n years. Apparently impossible tasks like tracing changes in the the blood lead concentrations of a Company's lead workers over the last ten years is now, or should be, a matter of a few minutes at most.

Once the statistics have been applied to the numbers is this the end of it? Perhaps so, too often. But you should remember that you are *not* a statistician, but an occupational hygienist. You use statistics only as a tool of your trade – like another meter. You use them firstly to understand the nature of your results, and secondly to predict what might happen if the measurements were to be repeated in the same or similar circumstances or location. If you do not *use* your statistical analyses there's not much point in doing them. At the same time, the application of common sense can save you from many embarrassments.

1.4 Handling Numbers

In addition to presuming that today's occupational hygienist has a basic knowledge of statistics it is also presumed that he has a number crunching machine, either a pocket calculator or access to a computer or even one of his own (personally or within his group). Pocket calculators will be programmable, with an adequate number of memories, statistical functions (preferably capable of being used within a program) and the necessary support functions like log and e^x. Calculators with a single memory and only the four arithmetic functions can be kept for working out the best buy in the supermarket.

The bewildering number of computers, operating systems and high level languages available prevents anything other than a very general approach to their use in occupational hygiene statistics. Perhaps BASIC in one form or another is the commonest programming language and is very adequate. H & H Scientific Consultants Ltd have available a number of statistical programs, originally written in BASIC for a Sharp PC1500 pocket computer with printer/plotter. These programs are also available on BBC cassettes in the original Sharp BASIC form and Amstrad PCW 3" CF2 discs translated to Mallard Basic. BOHS has published a BASIC program for microcomputers which tests monitoring data for the underlying nature of the distribution. There are other sources of "stats" and "graphics" software, but you should be sure before you buy that they do what you want. Too many "stats" programs seem to consist of histograms, pie charts and little else – not much use to a hygienist.

The essential difference between using a calculator and a computer in occupational hygiene statistics is that the calculator (usually) only calculates the intermediate tables stepwise, which have to be noted before further processing can be done. Also there may be a risk of errors in loading the program and data into the calculator. On the other hand once a program has been proved for a computer it will always carry out the required operations in

a single pass without any errors once the data have been input and verified (unless you have requested a perverse operation like asking the program to calculate the logarithm of a negative number).

You will find in this book that the approach is calculator/computer biased. This is no accident since the author has his own views on doing statistics with log tables, pencil, paper and mental arithmetic. He's tried this method and a variety of calculators and computers. He concludes that the right calculator is a powerful tool, and while some computers may be faster than others, at the practical level all computers are fast.

In Chapter 11 a brief section on logarithms is included. It has been known for minor puzzlement to arise when first using calculators and computers to find that the log of 0·2 is -0·6990, when by using tables it is $\bar{1}$·3010, read as "bar 1, ·3010". Also in Chapter 11 there are brief explanations of Minimum Variance Unbiased Estimators, and Sturges' Rule which may be of help. More commonly used terms can all be found in introductory statistical text books, and are not given here, which may account for the rather abrupt introduction of some of the statistics used in the book.

At various points you will find reference to "the degradation of data". This reflects the fact that some statistics by their nature group data together for the purpose of carrying out a test, or demonstrating a feature, usually to simplify the arithmetic or reduce the amount of calculation. Such grouping obviously destroys the integrity of the individual data points. There may, however, be a statistic available which does not destroy this integrity which will perform essentially the same test. In such cases the statistic which conforms to the woodwork master's admonition to "keep your wood as long as you can as long as you can" should be chosen.

A final point, applicable to both calculators and computers. Do not become bemused by the precision of which they are capable. An error of one in the last place in a ten-digit display is equivalent to an error of 1 ft in the distance to the moon. None of your results will probably warrant more than four significant figures, and too often occupational hygiene data are little better than ±10%. Now that's a statistic to conjure with!

1.5 Bibliography

This bibliography is by no means extensive or exclusive. There are many good modern books on statistics, as well as classical works which have not been included. New statistics are being developed all the time. It is for this reason that any book on statistics will necessarily be both out-of-date and imperfect (not just because of the misprints - a good enough reason for consulting at least two books to ensure that any errors have been detected and corrected).

A browse through the contents and indices of books on the shelves of your neighbourhood University bookshop or local library will often throw up a chapter of particular interest to you.

Where no publication date is given it can be assumed that the book is in a state of constant reprint or update.

References to additional sources are given at the end of each chapter.

CHOU, Y., (1970) Statistical Analysis, Holt, Rinehart and Winston, New York, N.Y.

COOKE, D., CRAVEN, A.H., and CLARKE, G.M., (1982) Basic Statistical Computing, Edward Arnold, London.

DAVIES, O.L., and GOLDSMITH, P.L., (1984) Statistical Methods in Research and Production, 4th rev. ed., Longman, London.

GUTTMAN, I., and WILKS, S.S., (1965) Introductory Engineering Statistics, John Wiley, N.Y., N.Y.

HAYSLETT, M.S., and MURPHY, P., (1968) Statistics made Simple, W.H. Allen, London. (Useful simple introduction.)

KENNEDY, J.B., and NEVILLE, A.M., (1976) Basic Statistical Methods for Engineers and Scientists, Harper International Edition, Harper & Row Inc., New York, N.Y. (Has much material not commonly found.)

KING, J.R., (1971) Probability Charts for Decision Making, Industrial Press Inc., New York., N.Y. (Out of print. Useful for starting the study of distributions, but some misprints and critical errors in the statistics.)

LEE, J.D., and LEE, T.D., (1982) Statistics and Computer Methods in BASIC, Van Nostrand Reinhold, Wokingham, Berks.

LEE, J.D., and LEE, T.D., (1982) Statistics and Numerical Methods in BASIC for Biologists, Van Nostrand Reinhold, Wokingham, Berks. (Both these books are useful sources of statistics and BASIC routines, but each contains much of the other. The latter book is out of print.)

MORONEY, M.J., Facts from Figures, Penguin Books, Harmondsworth, Mdsx. (Useful simple introduction, with full explanatory text rather than theory.)

SNEDECOR, G.W., and COCHRAN, W.G., (1967) Statistical Methods, Iowa State University Press, Ames, Iowa. (Very full, but readable.)

YULE, G.U., and KENDALL, M.G., (1965) An Introduction to the Theory of Statistics, Griffin, London. (Heavy going.)

Statistical tables can be found in many books on statistics. The major work (which also contains much valuable material on the underlying statistics) in this category is

PEARSON, E.S., and HARTLEY, H.O., (1976) Biometrika Tables for Statisticians, Vols. 1 and 2, Charles Griffin for the Biometrika Trustees, High Wycombe.

Small books of tables are available, one of the more popular being

ROHLF, F.J., and SOKAL, R.S., Statistical Tables, Freeman, San Francisco, Cal. This book of tables contains much which has been superseded by calculators and computers, like tables of five figure logarithms, but it still contains many other useful tables.

The following have been published by NIOSH, Cincinnati, Ohio, and are particularly relevant to occupational hygiene. For some comments on NIOSH and its publications see Section 11.5.

BAR-SHALOM, Y., BUDENAERS, D., SCHAINKER, R., and SEGALL, A., (1975) Handbook of Statistical Tests for Evaluating Employee Exposure to Air Contaminants, DHEW Pubn. No. 75-147.

LEIDEL, N.A., and BUSCH, K.A., (1975) Statistical Methods for the Determination of Noncompliance with Occupational Health Standards, DHEW Pubn. No. 76-159.

LEIDEL, N.A., BUSCH, K.A., and CROUSE, W.E., (1975) Exposure Measurement Action Level and Occupational Environmental Variability, DHEW Pubn. No. 76-131.

LEIDEL, N.A., BUSCH, K.A., and LYNCH, J.R., (1977) Occupational Exposure Sampling Strategy Manual, DHEW Pubn. No. 77-173.

NIOSH have also published material on the statistics of chemical analysis, some of which can be translated from this area and applied to occupational hygiene data.

H & H Scientific Consultants Ltd, P.O. Box MT27, Leeds, LS17 8QP, can supply a list of occupational hygiene programs (including statistical programs mentioned in this book) in BASIC. Abstracts, prepared for the BOHS Technology Committee, of these programs can be obtained on request from the BOHS Office, 1 St Andrew's Place, Regent's Park, London, NW1 4LB.

The BOHS Office can also supply (£2.00) Technical Guide Series No. 1, "Statistical Analysis of Monitoring Data by Microcomputer", August 1983.

2 DISTRIBUTIONS

2.1 Basic Ideas

A collection of data gathered from the same population of values will generally fit some distribution – the "random" numbers which are generated by your pocket calculator or computer will lie between zero and one, with a uniform (or rectangular) distribution. This means that any value between 0 and 1 has an equal chance of occurring.

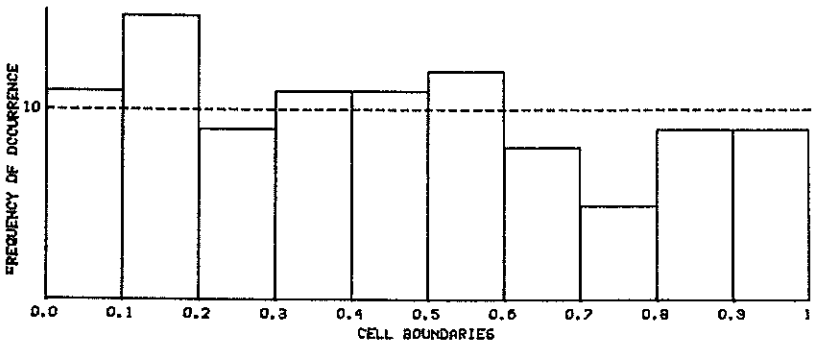


Figure 2.1 Uniform distribution of 100 random numbers.

In Fig. 2.1 the theoretical value for 100 results from a uniform distribution is for ten occurrences per cell, or in other words a probability of 0·1 (10%), but variations can, and will, occur. The variations in Fig. 2.1 may well cause concern, but using another statistic, the χ^2 test (one of the many tests of randomness and not discussed in this book) you would find that $\chi^2 = 6{\cdot}4$ with 9 degrees of freedom, suggesting that using this test at least the observed frequencies of occurrence are in fact quite acceptable for a random sample of size 100 drawn from a uniform distribution, $p = 0{\cdot}7$.

Repeated analytical measurements will be expected to have a mean with a spread of results about this mean. The spread of results will be normally distributed, with a high probability of a result occurring at or close to the mean, and a decreasing probability of a result as the value departs from the mean. The shape of the probability curve is shown in Fig. 2.2.

The x-axis is in standard deviations, the z-axis is the ordinate of the normal curve, and P(x) is the area (probability) to the left of x=1 in this case and Q(x) the area to the right of $x = 1$. It is clear that $P(x) + Q(x) = 1$, and the area (p) to the left and right of $x = 0$ is 0·5.

For a full discussion of the theory of the Normal distribution you should refer to any standard textbook on statistics. You should at least be able to read and interpret a table of Standard Normal Deviates. Referring to Fig. 2.2

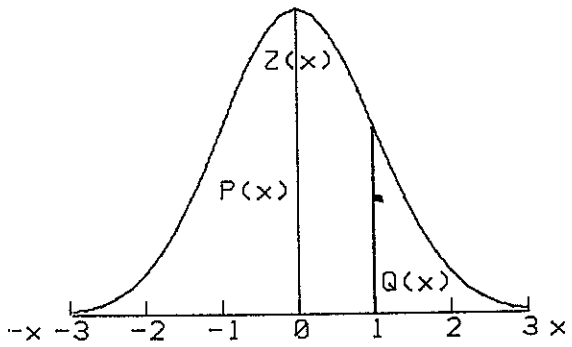


Figure 2.2 The Normal Curve.

you will find that some tables give the value P(x) (i.e. those which start at 0·5) while others give the area from x = 0 to x = +x, i.e. 0·5 - Q(x). The integral equations in Section 10.2.1 should clear up any residual problems you may have, while Fig. 2.3 shows, on a single scale, the relation between cumulative probability and standard deviations.

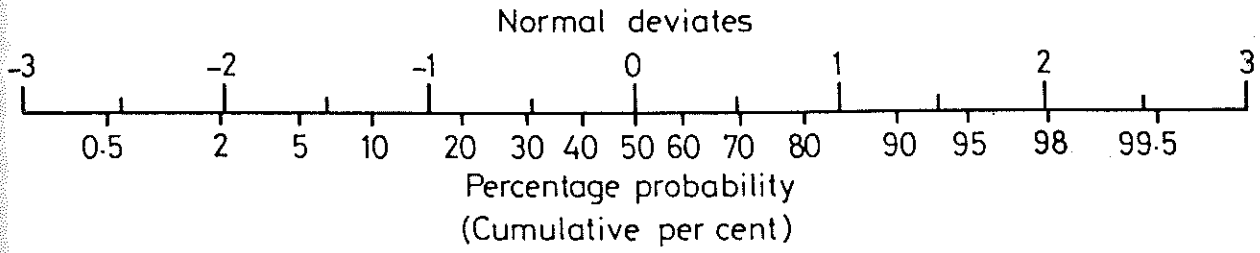


Figure 2.3 The Normal Cumulative Probability Scale.

The percentage scale is unequally divided but symmetrical about 50%. Normal deviates are symmetrical about 0 and are equal. These scales are used in plotting both normal and lognormally distributed data. Other scales will be needed for other distributions.

Similarly the t-distribution is described fully in most statistics texts, and again it is presumed that you can read the t-distribution table, and understand the nature of degrees of freedom and one- and two-tailed values in the table. The χ^2 distribution is not used in this book, but it certainly has a place in occupational hygiene statistics, and has already been mentioned in passing in this Chapter.

2.2 Distributions in Occupational Hygiene

It seems to be an occupational hazard of occupational hygienists to believe that their data (airborne concentrations in particular) are never distributed normally but always lognormally, that is the logarithms of the data are normally distributed. Some concentrations of styrene are shown in Fig. 2.4 which show a "typical" set of data which might be thought to be lognormally distributed, with the peak frequency of occurrence (mode) well to the left of centre, and a long "tail" of concentrations extending to the highest values.

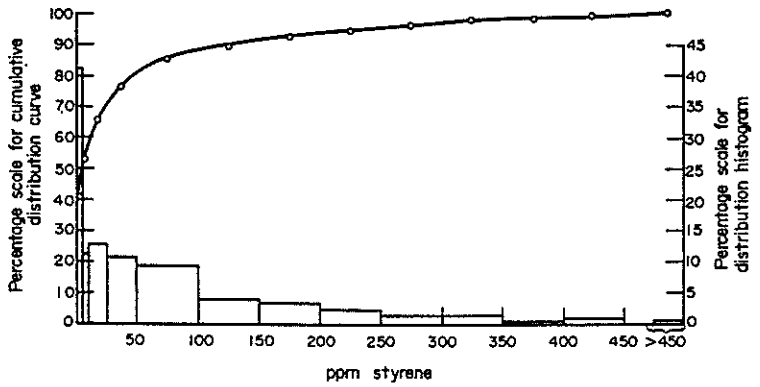


Figure 2.4 Histogram and cumulative curve of styrene concentrations.

It seems to be an occupational hazard of occupational hygienists to believe that their data (airborne concentrations in particular) are never distributed normally but always lognormally, that is the logarithms of the data are normally distributed. Some concentrations of styrene are shown in Fig. 2.4 which show a "typical" set of data which might be thought to be lognormally distributed, with the peak frequency of occurrence (mode) well to the left of centre, and a long "tail" of concentrations extending to the highest values.

At the very least the histogram should be plotted on a logarithmic scale as in Fig. 2.5 if the concentrations are lognormally distributed.

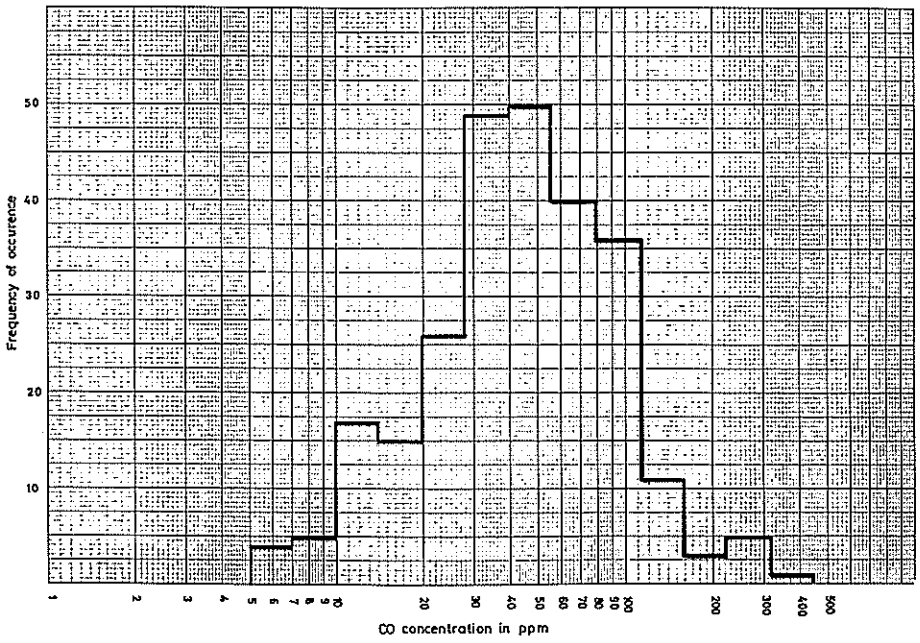


Figure 2.5 Histogram of frequency of occurrence of CO concentrations.

The "typical bell-shaped curve" of the normal curve, Fig. 2.2 (and to which Fig. 2.5 might be thought an approximation), is often invoked to "show" that the data are normally distributed, or, if the data are to be plotted on a logarithmic scale, lognormally distributed. Worse, a plot such as Fig. 2.4,

showing the extended tail of the histogram, is too often said to be "lognormally distributed" without further ado. On some occasions the investigator might take the logarithms of such data and replot these to give the "typical bell-shaped curve" and use this to say that the logged data are normally distributed, and hence the data are lognormally distributed. It would seem only prudent to test for the most likely distribution, since, as we will see, it might be important to get the right one. Chapter 6 describes a method of testing experimental data for goodness-of-fit of individual data points, but for data presented and only available in histogram form (or the derived cumulative curve) an alternative test must be used, such as the χ^2 test or the Kolmogorov-Smirnov test. It must also be clear that the test and choice of distribution should really be made *before* the histogram, or single point plot described in Chapter 6, displaying the results is prepared for the final report. Mage[1] has some cogent comments on the all-pervasive lognormal distribution.

It seems that the most popular, or at least most common, distributions met in occupational hygiene are the normal and lognormal distributions. There may be others, and the most likely are the Type I and II Extreme Value Distributions (EVDs). These are, like the lognormal distribution, usually positively skewed, that is with a tail at the higher values. These tails extend to even higher values than the lognormal distribution, and take an even more exaggerated form than Fig. 2.4. They could arise in the following way. Of the 48 10-minute samples of a compound with a 10-minute STEL which could be taken in an eight-hour shift, the highest in each hour is recorded. These eight highest values may well be from an extreme value distribution. Further daily testing for all the available 10 minute sampling periods would be needed to show convincingly that the hourly highest concentrations *were* from either a Type I or Type II (log) EVD.

The log EVD bears the same relation to the EVD as does the lognormal to the normal distribution – the data are logged before plotting. You can perhaps see that the EVDs could be most useful in working with STELs or the infrequent, but still likely, occurrence of very high concentrations. The EVDs will not be discussed further here, except to note that the probability scale for the Type I and II EVDs are the same as for Rosin-Rammler distribution paper used in particle size analysis. In fact Rosin-Rammler paper can be used for plotting log EVD data, since particle sizes are plotted on a logarithmic scale. Rosin-Rammler paper, like probability papers, can be obtained from the usual suppliers should you like to examine it.

2.3 The Lognormal Distribution

There is nothing especially difficult about handling the lognormal distribution. All the features of the normal distribution are available for computational purposes and all that is required is to work with the logarithms of the data. There are, however, some restrictions, and other peculiarities. Firstly, the lognormal distribution can have no zero or negative data values, since the logarithm of numbers ≤ 0 is undefined. Also the mean and standard deviation are in the exponentiated form, i.e. they are not, in general, presented in their logged form, although it is usually much more convenient to carry out calculations using the logs of the Geometric Mean (GM) and Geometric Standard Deviation (GSD). This is particularly true when using computers, since the same program then does for normally and lognormally distributed data. Although there is a statistic corresponding to the Coefficient of Variation of the normal distribution, it is not used very much (and not at all

in this book), the GSD serving just as well. The dimensions of the GM are the same as those for the AM and arithmetic SD, namely those of the original units, ppm, mg/m³ etc., but the GSD is dimensionless. Also, just as a standard deviation of zero means that there is no spread of values about the mean (because all data values are identical), if the geometric standard deviation is 1, again all the results are identical (although it is difficult to conceive of anyone wishing to attribute such a data set to a lognormal distribution). Perhaps the point to note is that the logarithm of 1, to any base, is zero, and consequently the geometric standard deviation is always greater than one.

Numerical relationships between logarithmic and arithmetic parameters of the lognormal *population* are given in Section 10.2.8 and the calculation of the minimum variance unbiased estimates of the arithmetic mean and standard deviation for a sample drawn from a lognormal distribution is shown in Section 3.5.

Brief accounts of the use of the lognormal distribution can be found in Leidel and Busch (1975), Leidel et al. (1975) and Leidel et al. (1977), with more information in King. References to these sources will be found in the bibliography to Chapter 1. For the fullest discussion, going far beyond the needs of hygienists, you would find the major work by Aitchison and Brown invaluable (Reference 2, Chapter 3).

If you use the normal and lognormal distributions to *predict* possible values you will find that the lognormal distribution always predicts values >0, while the normal distribution predicts values <0 sometimes at comparatively modest probabilities in Fig. 2.3, say in the 2% to 10% region. This is sometimes used as an excuse for saying that all occupational hygiene data must be lognormally distributed, to avoid such predicted negative concentrations. You should always be aware that your predictions will not be exact, and the fewer the number of your data, the less reliable will your predictions be.

2.4 Artificial Samples

It is useful to generate artificial samples for the study of the statistics described in this book, especially if there is a dearth of real OH data upon which to practise. Such data can be generated by using the random number function on a calculator or computer. "Random numbers" on calculators and computers are obtained using a formula which need not concern us here, except to note that it requires a "seed" which governs the series produced, the same seed will generate the same series. Some computers take their seed unseen from the internal clock, but others require the seed to be deliberately sown, otherwise the same series of random numbers is produced each time the program is run – not much fun and even less use. Also, calculators may only give random numbers to only a few, perhaps three, decimal places. They all give numbers in the range $0 \leq x < 1$ or $0 < x < 1$ but the range can obviously be expanded by multiplying by a factor, and moved by adding a constant.

2.4.1 General Method of Generation

Each of the random numbers from the uniform distribution provided by a calculator or computer can be thought of as probability values which must be converted to the appropriate deviate scale. This, for normally and lognormally distributed data, will be the normal deviate scale, or scale of standard

deviations (see Fig. 2.3), and for EVDs another scale. To simulate some random data we also need to specify an appropriate mean and standard deviation.

The random variate is generated by drawing a random number from the range $0<x<1$, converting this to the deviate and then to the random variate, perhaps representing an airborne concentration. The first conversion can be done using the inverse normal integral approximation given in Section 10.2.2 (or using the Normal table). For normal and lognormal distributions the random number 0·5 would, of course, transform to a deviate of zero. Random numbers below 0·5 will give deviates <0 and random numbers >0·5 will give deviates >0. So for a normal distribution and a random number = 0·119 the corresponding deviate, v, would be -1·18. Interestingly the conversion for EVDs is numerically much easier to calculate.

The variate is obtained from this value and the mean and standard deviation by substituting in the linearized form of the cumulative distribution. If, say, the mean = 5 and SD = 2, in the example above the random normal variate will be

$$RNV = v.SD + Mean = -1{\cdot}18 \times 2 + 5 = 2{\cdot}64$$

This procedure is repeated for as many random variates as required, perhaps 5, if daily shift averages for a week are being simulated, or 48 for a day's 10-minute Short Term Exposures. To generate lognormally distributed data the equation above becomes

$$\ln RNV = v.\ln GSD + \ln GM$$

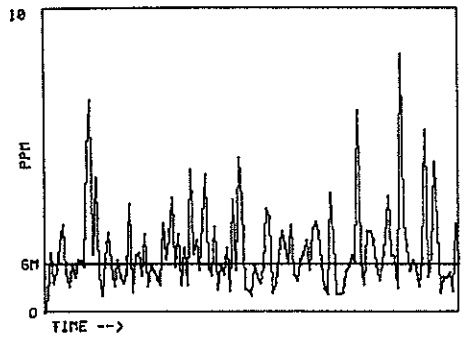
It was stated earlier that the range of random numbers from computers and calculators is $0 \leq x < 1$, and it may be wise to weed out any random numbers = 0, since these will compute as $-\infty$ ND, which is not on. Also three-place numbers from some calculators mean that the random concentrations will be "quantized", but I leave it to you to show that 0·001 and 0·999 will give minimum and maximum concentrations, much closer to the mean than, say, 0·000001 or 0·999999. Clearly there cannot be concentrations corresponding to values between 0·001 and 0·002 and so on, when using uniform random numbers generated on such a calculator followed by conversion to random normal deviates and variates using the polynomial approximation and linear equation.

This approach is quite general for *any* distribution, and by using the appropriate conversion from random number to random deviate a sequence of random values derived from a particular distribution may be generated. An alternative and more rapid method[2,3] of generating simulated normally or lognormally distributed data is given in Section 10.2.9.

2.4.2 A Random Lognormal "Sample"

It will be clear that a computer is ideal for such simulation. Fig 2.1 was derived from a computer generated series of one hundred random numbers (HISTOGRAMS[4]). Fig 2.6, (using RND LNOR CONCS[4]) shows the output from a Sharp PC1500 for 136 random "measurements" (as one-day shift average concentrations? see Section 5.3.2) for a lognormally distributed data set with GM = 1·608 mg/m³ and GSD = 2·0. The similarity to typical outputs from continuous recording instruments can be seen. Also random variates can be treated using any of the statistics described in this book. For example the printout gives the various means and standard deviations for the specified distribution and for the random sample drawn from it. It is not necessary to plead poverty of real occupational hygiene data in order to become proficient

in the application of statistics when artificial data are so easily generated.



136 SAMPLES FROM A LOGNORMAL
DISTRIBUTION, WITH POPn. GM= 1.688
AND POPn. GSD= 2
POPn. ARITH MEAN= 2.845
POPn. VAR. 2.579

SAMPLE GEOM MEAN 1.697
SAMPLE GEOM Std Devn 1.79
SAMPLE ARITH MEAN 2.824
SAMPLE ARITH VARIANCE 1.824
MVU ARITH MEAN 2.808
MVU ARITH VARIANCE 1.689

Figure 2.6 Random Lognormal Concentrations.

2.5 References

1 MAGE, D.T. (1985) "The Procrustean Fit – A Useful Statistical Tool for Decision Making" *Journal of Irreproducible Results*, v. 30, no. 4, p. 32.

2 COOKE, D., CRAVEN, A.H., and CLARKE, G.M. (1982) Basic Statistical Computing, Edward Arnold (Publishers) Ltd, London.

3 COOKE, D. (1985) Private communication, correcting a misprint in Reference 2.

4 OH Programs are available from H & H Scientific Consultants Ltd, P.O. Box MT27, Leeds, LS17 8QP.

3 MEANS AND STANDARD DEVIATIONS

3.1 Means and Spread of Results

This is an area in which it is assumed you already have a working knowledge, and what follows has been included for the sake of completeness and revision.

3.2 Arithmetic Mean

In the equation

$$\bar{x} = \Sigma x/n \qquad [3.1]$$

$\bar{x}$ is the arithmetic mean, AM, of the n items of data. It is also the best estimate we have of the population mean, μ, from which the n items were drawn as a sample, if the population is normally distributed. The case of the best estimate of the arithmetic mean for the lognormal distribution will be covered more fully below in Section 3.5.

3.3 Arithmetic Standard Deviation

The equations

$$s = \sqrt{[\Sigma(x - \bar{x})^2/(n - 1)]} \qquad [3.2a]$$

$$= \sqrt{[(\Sigma x^2 - n\bar{x}^2)/(n - 1)]} \qquad [3.2b]$$

$$= \sqrt{[(\Sigma x^2 - (\Sigma x)^2/n)/(n - 1)]} \qquad [3.2c]$$

are three forms which can be used to calculate the estimate of the population standard deviation s from the sample data. Equation [3.2a] is least prone to the generation of rounding errors in the calculation of s and can easily be incorporated into computer programs by first calculating the arithmetic mean from [3.1]. This means that two passes of the data are used. Cooke et al.[1] give a BASIC routine which calculates s with one pass of the data as it is entered, although it is slower than computing [3.2a] because of the additional computing called for. Pocket calculators use a single entry of the data and you will find memories containing Σx and Σx^2 as well as n, with the AM calculated from [3.1] and the SD from [3.2c], directly from the memory contents, leading to the potential for arithmetic errors. While many calculators and computers have an adequate number of significant figures, some may only work to seven, and the use of routines which obviate the risk of errors is to be preferred.

The standard deviation is the most appropriate estimate of the spread of the population from which the sample data values were drawn. Other measures, such as the range (of the sample), give less information, and the labour involved in calculating the SD is these days little more than finding the extreme values (by inspection). although its expectation increases with

increasing sample size.

A close relative of the standard deviation is the estimate of the population variance, s^2. The standard deviation for the whole population (which we will rarely have) is σ, and the population variance is σ^2 in which n-1 in [3.2] is replaced by n.

As with the mean, these statistics are most easily interpreted for data which are normally distributed, or more strictly, to data drawn from a normally distributed population. Again the best estimator for the (arithmetic) variance of a lognormally distributed population will be given below.

3.4 Geometric Mean and Geometric Standard Deviation

These statistics are derived from using the natural (or common) logarithms in [3.1] and [3.2].

$$y = \ln x \qquad [3.3]$$

$$\bar{y} = \Sigma(\ln x)/n = \Sigma y/n \qquad [3.4]$$

$$g = \sqrt{[\Sigma(y - \bar{y})^2/(n - 1)]} \qquad [3.5a]$$

$$= \sqrt{[(\Sigma y^2 - n\bar{y}^2)/(n - 1)]} \qquad [3.5b]$$

$$= \sqrt{[(\Sigma y^2 - (\Sigma y)^2/n)/(n - 1)]} \qquad [3.5c]$$

$$\text{Geometric Mean (GM)} = \exp(\bar{y}) \qquad [3.6]$$

$$\text{Geometric SD (GSD)} = \exp(g) \qquad [3.7]$$

Thus the method of calculating the GM is similar to that for calculating the AM but for the GM the logarithms of the data are first used to calculate the mean (ln data), which is then exponentiated(†), and similarly for the GSD. The same reasoning applies to σ and g and the variance of the logged(†) data.

(†) See Section 11.2

The arithmetic mean and SD of the data are still = $\bar{x}$ (= $\Sigma x \div n$) and s (from [3.1] and [3.2]), and the best estimates of the ln GM and ln GSD of the parent population (assuming that it is lognormally distributed) are $\bar{y}$ and g. The variance, g^2, is also the best estimator for the logged data. On the other hand $\bar{x}$ and s are not the best estimators of the arithmetic mean and standard deviation of the parent (lognormally distributed) population from which the sample of n was drawn.

In occupational hygiene it is often the case that samples are taken from a lognormally distributed population (although this should be tested as described in Chapter 6). The problem can be visualised in the practical case when ten short term "samples", of 10 minute duration, are taken at random during a nominal length (480 minute) day. From this sample of 10 we want to make the best estimate of the arithmetic mean of the daily exposure. If the data are normally distributed the arithmetic mean is the best estimate for the shift average, but if they are lognormally distributed an alternative measure of the average should be used. The best estimate of the population arithmetic mean of a lognormally distributed sample is α.

So from the data an estimate of, say, the time-weighted average (TWA) or arithmetic mean of the population is made, either by assuming that the sample arithmetic mean is the value required or by evaluating

$$\alpha = \exp(\bar{y} + \tfrac{1}{2}(\ln GSD)^2)$$

although this is only true for the population. It is unsafe to apply this procedure, or to use the sample arithmetic mean, for samples of small size. A third method of calculating the TWA is available, using the minimum variance unbiased estimator.

3.5 Minimum Variance Unbiased Estimators

Minimum variance unbiased estimators are statistics, such as the mean and standard deviation, calculated from the available sample data and which use a form of calculation which ensures that they are estimators of the statistic for the population from which the sample is drawn, and have minimum variance and minimum bias. They are discussed briefly in Section 11.3.

Aitchison and Brown[2], and Bar-Shalom et al[3] give a method of calculating the minimum variance unbiased (MVU) estimator, $\underline{a}$, for the population arithmetic mean α (although there are unfortunate misprints in both). Aitchison and Brown also give a method of calculating $\underline{b}^2$, the MVU estimator for β^2, the population arithmetic variance. In each case these estimators are the best. In occupational hygiene it will be obvious that you are really interested in the AM of the population (the full working day?) from which your n short-term samples were taken. The MVU estimators are what you want, not the AM or SD of the data. Bar-Shalom also gives three nomograms for the solution of $\underline{a}$ for $n = 3,4$, $n = 5,6$ and $n \geq 7$. Omitting the derivation given by Aitchison and Brown, a function $\psi_n(t)$ is defined by the power series

$$\psi_n(u) = 1 + \frac{n-1}{n}u + \frac{(n-1)^3}{n^2(n+1)}\frac{u^2}{2!} + \frac{(n-1)^5}{n^3(n+1)(n+3)}\frac{u^3}{3!} + \ldots\ldots \qquad [3.8]$$

By making $u = \frac{1}{2}g_y^2$ (g_y^2 = variance of logged data) the power series can be evaluated, and is used to calculate $\underline{a}$, the MVU estimate of the population arithmetic mean, α, from

$$\underline{a} = \exp(\bar{y}).\psi_n(\tfrac{1}{2}g_y^2) \qquad [3.9]$$

Perhaps of lesser interest is the estimate of the arithmetic variance, $\underline{b}^2$, of the lognormally distributed population, which is evaluated by generating two $\psi_n(u)$ series, in which the first $u = 2g_y^2$,

and in the second $u = \frac{n-2}{n-1}.g_y^2$ to give

$$\chi_n(u) = [\psi_n(2g_y^2) - \psi_n(\frac{n-2}{n-1}g_y^2)] \qquad [3.10]$$

The values of these two ψ series are first calculated to give $\chi_n(t)$ which is then used to evaluate

$$\underline{b}^2 = \exp(2\bar{y}).\chi_n(u) \qquad [3.11]$$

Equations [3.8] to [3.11] appear formidable, but [3.8] and [3.9] can be evaluated to any desired precision in no more than five lines of BASIC, or on a primitive 72 step programmable pocket calculator. The equations for the variance, [3.10] and [3.11], need an extra line of BASIC, and two passes on the calculator. The mathematics are thus no obstacle to calculating the correct estimates for the population arithmetic mean and variance of lognormally distributed data. That is if you have already shown that the data are indeed from this distribution, by using the methods described in Chapter 6.

In Section 5.3.2 you will find that MVU estimates of the population mean and standard deviation of a set of data are quite close to the sample mean and

standard deviation. This is not always the case. A greater difference was found in 57 data points which were lognormally distributed. The sample mean and standard deviation were 112 and 154 respectively, while the MVU estimates for the mean and standard deviation for the population were 139 and 353. Such large differences are apparently not uncommon, and may have significant effects on consequent decisions or subsequent statistics. It is, on the whole, better to use the MVU estimates of the arithmetic parameters for the population rather than the sample arithmetic mean and standard deviation.

3.6 References

1 COOKE, D., CRAVEN, A.H., and CLARKE, G.M., (1982) Basic Statistical Computing, Edward Arnold, London.

2 AITCHISON, J., and BROWN, J.J., (1957) The Lognormal Distribution, Cambridge University Press, Cambridge.

3 BAR-SHALOM, Y., BUDENAERS, D., SCHAINKER, R., and SEGAL, A., (1975) Handbook of Statistical Tests for Evaluating Employee Exposure to Air Contaminants, U.S. Dept of Health, Education and Welfare, NIOSH, Cincinnati, Ohio. (HEW Pubn. No. (NIOSH) 75-147).

4 CONFIDENCE LIMITS OF MEANS

4.1 How Sure Are You?

In Chapter 3 we have seen how to calculate the mean and spread (standard deviation) of the sample and estimate these parameters for the population from which the sample was drawn. The distinction has to be made that while the calculated mean of the data values is numerically exact, this value is only an estimate of the mean of the population.

The use of the word "estimate" immediately suggests some uncertainty in the calculated results. This uncertainty can be calculated from the data we have, and a knowledge of the t-distribution. The derivation and applications of the t-distribution are well described in any good basic book on statistics. This distribution is similar to the normal distribution, but it invokes a new idea, the "degrees of freedom" (DF) associated with the data. When the degrees of freedom are infinite the distribution is identical to the normal distribution, but as the number of DF decreases the values corresponding to the deviations increase for the same tabulated probabilities. You can think of this effect as a reflection of increasing uncertainty as the number in the sample gets smaller.

4.2 Confidence Limits

The uncertainty of the estimate of the mean for the population can be expressed as "with 90% confidence (or some other level) the mean of the population lies between these two values". This is the same as saying that if we were to sample from the population to give 100 estimates of the mean, 90 of them would lie between the two values, or confidence limits. Sometimes we may need to say that "the mean, with 95% confidence (or some other level) will not exceed (or be less than) one value (or another)".

In essence this means that if we take 100 (repeat) sets of samples from the same population (5 daily personal dust concentration measurements for each of 100 weeks from the possible lifetime of the operative) the means of 90% of these (weekly) sets will lie within the 90% confidence limits either side of the true mean for the population. The means of 5% of the sets will be below the lower limit, and 5% above the higher. It is most unlikely that we could sample daily for 100 weeks. Instead from one week's sampling we must calculate an estimate of the mean dust concentration, and use the same data to estimate the confidence limits within which the true mean will lie.

For this sample of approximately normally distributed data
0·192, 0·401, 0·505, 0·612, 0·645, 0·654, 0·666, 1·132
the (arithmetic) mean is 0·601, which is the estimate of the mean of the population from which the sample was drawn. With 90% confidence the population mean lies between 0·420 and 0·781. Or with 95% confidence the mean will not exceed 0·781, nor be less than 0·420. Clearly there is some lack of confidence

in the estimate, since the mean could be greater than 0·781, or less than 0·420, but it seems fair to suggest this is unlikely.

Confusion can sometimes arise in the interpretation of the "90%" and "95%" used above – how do these two levels give the same results? The explanation lies in the fact that they represent the same confidence limits, but while the first is "two-tailed" the second is "one-tailed". That is the 90% confidence (between) level is looking at the central symmetrical 90% of the area under the t-distribution curve (which it has been said is of similar form to the normal distribution curve). The 95% (not exceeding or not less than) level is the same (90%) area plus the 5% of the area at one or other of the ends, see Fig. 4.1.

In fact you will find the table is laid out not as percentage confidence levels, but as percentage significance levels, and

$$\% \text{ confidence} = 100 - \% \text{ significance}.$$

The significance is commonly denoted by α, which is the sum of the areas below the two tails. The value of t in the table may correspond to the significance α or $\alpha/2$, and you should make sure that you use the correct percentage value. So when looking up the required value of t you will first have to decide whether your confidence level is one- ($\alpha/2$) or two-tailed (α) and then convert the confidence to significance.

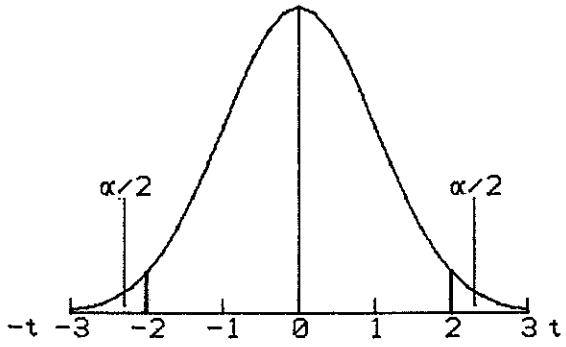


Figure 4.1 The t-distribution curve.

As an alternative to using tables you could calculate t for a given confidence level and DF on a computer using the formula in Section 10.2.4.

4.3 Calculation of Confidence Limits

The mean, standard deviation and sample size n are needed to calculate the confidence limits of the mean, together with the appropriate tabulated value of t.

First the standard error (of the mean), $s_{\bar{x}}$, is calculated from the estimate of the standard deviation and the number in the sample

$$s_{\bar{x}} = s/\sqrt{n} \qquad [4.1]$$

The value of t is then looked up in the t table, for the chosen confidence level and for n-1 degrees of freedom. The standard error is multiplied

by t, and this value is added and subtracted from the mean to give the confidence limits.

$$\text{Confidence limits} = \bar{x} \pm t.s_{\bar{x}} \quad [4.2a]$$

$$= \bar{x} \pm t.s/\sqrt{n} \quad [4.2b]$$

For the data above and choosing a two-tailed confidence level of 90% (giving a one-tailed confidence level of 95%, which translates to a significance of 5%) with 8-1 = 7 DF, we find in the table t = 1·895, and equation [4.2] gives the values 0·601 ± 0·180 = 0·420 and 0·781, accommodating rounding errors to three decimal places.

Other confidence levels may be chosen, such as 80% (two-tailed) or 99% (one-tailed). A lower confidence level merely brings the limits closer to the mean, and higher confidence levels widen the interval. The mean, with 80% confidence, lies between 0·466 and 0·736 (t = 1·415). The mean will, with 99% confidence, not be less than 0·315 nor exceed 0·886 (t = 2·998). Or, if you like, the mean, with 98% confidence, will lie between 0·315 and 0·886.

It is useful to do a few practice runs with dummy data until you are confident (sorry) in the calculation of confidence limits.

4.4 Confidence Limits for Lognormally Distributed Data

Equations [4.1] and [4.2] are applicable to the logarithms of lognormally distributed data.

$$\text{Standard error of the (log) mean } s_{\bar{y}} = g/\sqrt{n} \quad [4.3a]$$

$$\text{Logs of confidence limits} = \bar{y} \pm t.s_{\bar{y}} \quad [4.4a]$$

$$= \bar{y} \pm t.g/\sqrt{n} \quad [4.4b]$$

Translated into Geometric Means and Geometric Standard Deviations these become, by analogy with [4.2a] and [4.2b]

$$\text{Confidence limits} = GM \times \text{or} \div \exp(s_{\bar{y}})\uparrow t \quad [4.5a]$$

$$= GM \times \text{or} \div GSD\uparrow(t/\sqrt{n}) \quad [4.5b]$$

The symbol ↑ is "computerese" for "raised to the power of".

Looking at the sets of equations [4.4] and [4.5] it is clear that in plain number handling terms [4.4b] is easier (and on a computer faster) to compute, followed by a final exponentiation to get the two numerical limits, and is to be preferred to evaluating [4.5]. For the data listed above the geometric mean is 0·543 and the geometric standard deviation is 1·669. The 95% confidence limits are 0·385 and 0·766. You will see that these are not symmetrical (arithmetically) about the GM, whereas the limits about the AM (4.3 above) were symmetrical. The confidence limits of the logged data are, of course, symmetrical about $\bar{y}$.

4.5 Other Confidence Limits

Clearly there are confidence limits which can be applied to the standard (and geometric) deviations. These variables are distributed according to the χ^2 distribution. These confidence limits are generally considered either to be of only marginal interest or numerically too difficult to handle and are only discussed in a few text books (and this isn't one of them). It may be that, with the wide availability of computers to handle the number crunching, these confidence limits might become more accessible, perhaps leading to a better understanding of why measured occupational hygiene data are distributed the way they are.

This illustrates that there are, on the one hand, more statistics available for hygienists if they wish to apply them. On the other hand the mathematics of calculating confidence limits for the MVU estimators for the arithmetic mean and standard deviation of lognormally distributed data (Section 3.6) have not yet been worked out, and you should not be tempted to substitute these MVU values in the equations above; you will obviously get a pair of values, but they will be difficult to justify statistically. There is more on confidence limits in Chapters 6 and 7.

4.6 Another Example — Respirable Quartz Concentrations

This example and the one given above are, in fact, taken from data used in Chapter 7, so that when you come across a computer output there such as

WITH 90% CONF. MEAN LIES BETWEEN 8.45 AND 14.69

you will know where it comes from and to which mean (arithmetic or geometric) the confidence limits apply, although for the purposes of illustration the confidence limits of both distributions will be worked out here. In general only one or other would be recorded, or output, although both would be available within a computer with the appropriate program installed.

The data (in mg/m³) are

	0·036	0·035	0·129	0·079
$\log_{10}$	−1·4437	−1·4559	−0·8894	−1·1024

The problem is to find the 95% confidence limits of the arithmetic and geometric means. There are (n − 1) degrees of freedom, n = 4, so DF = 3, for which the t-distribution table gives t = 3·182, which will be used in the table below.

The quartz concentrations give the following table, for the arithmetic and logarithmic data, using the equations in the last column.

	Arithmetic	Logarithmic	Antilogs	From Eqn
Mean	0·0698	−1·2229	0·0599 (GM)	[3.1]
SD	0·0445	0·2762	1·8887 (GSD)	[3.2a]
Std Error	0·0223	0·1381	1·4436 (10↑SE)	[4.1]
t.SE	0·0708	0·4394		
Mean + t.SE	0·1406	−0·7835	0·1646 (Upper 95% CL)	[4.2a]
Mean − t.SE	−0·0011	−1·6622	0·0218 (Lower 95% CL)	[4.2a]

Again it will be seen that although the confidence limits for the arithmetic and logarithmic data are symmetrical about their respective means, the numerical limits about the geometric mean are not. The only useful thing about equations [4.5a] and [4.5b] is that they warn us that the limits are a multiple or quotient of the geometric mean of a lognormally distributed set of data. It will also be seen that a negative concentration of −0·0011 mg/m³ is predicted as the lower confidence limit for the normal (arithmetic) distribution!

If this should be thought an embarrassment some level of confidence lower than 95% could be chosen, say 80%, for which, with 3 DF, t = 1·638. The values of t.SE for the arithmetic and logarithmic data now become 0·0365 and 0·2262. From these and the means we get the arithmetic 80% confidence limits of 0·0698 ± 0·0365 = 0·1062 and 0·0333 mg/m³. At least there is no negative concentration and the limits are closer to the mean, but this is the price to be paid for selecting a not very high level of confidence. The confidence

limits for the geometric mean are 0·1008 and 0·0356 mg/m³, and similar comments apply.

"The 95% confidence limits of the geometric mean are 0·0218 and 0·1646", is the equivalent of saying that, with 97·5% confidence, the geometric mean of the population from which this sample of four concentrations was drawn will not exceed 0·1646 mg/m³, nor, with the same (97·5%) confidence will it be less than 0·0218, these being the one-tailed distribution of t.

It might be thought that for a geometric mean of 0·0599 these are wide limits within which we can expect to find the true population geometric mean, and indeed they are. But that is the nature of the lognormal distribution (and any lognormally distributed occupational hygiene data). We could make the limits closer to the mean by choosing a smaller confidence level. For a confidence level of 80%, or $\alpha = 0{\cdot}2$, t_3 is found from the t-distribution table to be 1·638. Working through the table above again we find that the 80% confidence limits for the geometrical mean are now 0·0355 and 0·1008, certainly closer to 0·0599, and perhaps of more comfort than the 95% confidence limits, but this has been achieved at the cost of expecting 20 out of 100 sample means to be outside these limits, instead of only 5 out of 100 for the 95% level (were we to have the time to sample four days per week for 100 weeks in order to estimate the operative's lifetime weekly mean exposure in this way). We must accept a trade-off between high confidence levels and wide limits on one hand and low confidence and tighter limits on the other.

The wide confidence limits of these quite typical occupational hygiene data should remind us of how unlike they can be to data derived under laboratory conditions. An analyst faced with such data from repeat measurements would conclude that he did not have an analytical method. But occupational hygienists must live with such results, or choose to use rather lower confidence levels than might be more usual. The one thing they cannot do is to ignore the spread of their results.

5 SAMPLE SIZES

5.1 Some of the Problems

The question "What size of sample do I need?" is easily asked, and not at all easy to answer. In fact without more qualification or additional information it is impossible. One qualification might be "in order to say, with 90% confidence, that the mean of my sample would not be more than 10% in error of the true mean". The additional information might be in the form of a statement concerning the sample itself and the population from which it is to be drawn, or in the form of preliminary data, or accumulated experience from previous similar measurements.

Even with this additional information available it is almost inevitable that the proposed sample size will be so large that the occupational hygienist will be hard pressed to justify the cost of implementing the sampling programme. This is unfortunate but both the hygienist and his management or client must be aware that either the costs must be met or the confidence in the results will be significantly reduced. It seems that in the real world little can be done about this state of affairs, except be aware of it.

The problem arises from the nature of so many occupational hygiene measurements. Even a passing acquaintance with repeat or simultaneous measurements of nominally the same condition will have convinced you that OH measurements are, in general, not very reproducible. If you have not done any such measurements you should, since your previous experience of repeatedly measuring the length of a piece of string and working out the mean and standard deviation (seriously) will be of no help to you in finding out how variable occupational hygiene measurements can be.

5.2 The Single Sample

It is unfortunate that the word "sample" is almost always used for a single measurement of a sound level, concentration of welding fume (further perhaps compounded by calling it the "dose"), or blood lead level. Even if the "sample" is taken over a full shift it is simply a measure of the conditions at the time the measurement was made. It will be obvious from Chapter 3 (standard deviations) that from only one measurement there is no way you can get any idea of what the spread of all possible measurements or of what another measurement might be, since in calculating the estimate of the population standard deviation you will have a denominator of zero ($n - 1$, when $n = 1$) in Equation [3.2]. Division by zero gives ∞ in mathematics, and on computers and calculators results in overflow or errors (or both).

Nor can you calculate a coefficient of variation for the same reason.

5.3 Sample Size for Two Cases

The two methods of finding the required sample size given below are by no means the only ones available. They do, however, answer two properly formulated questions, unlike the incomplete one posed at the start of this Chapter. Sampling size is discussed in many good statistics texts and in some depth by Yates[1].

5.3.1 Coefficient of Variation Known, Error Limit Chosen

If the objective of the planned sampling programme is to estimate the population mean, whether arithmetic or geometric, for the appropriately distributed data, the prior knowledge needed is the estimate of the standard deviation from a preliminary survey or previous experience in similar circumstances.

The statistic t is given by the equation

$$t = \frac{|\mu - \bar{x}|}{s_{\bar{x}}} \qquad [5.1]$$

where μ = population mean

$\bar{x}$ = sample mean and

$s_{\bar{x}}$ = standard error of the mean.

But $s_{\bar{x}} = s/\sqrt{n}$,

where s = estimate of population standard deviation and

n = sample size.

If we express $|\mu - \bar{x}|$ as a percentage of the population mean then

$$E = \frac{|\mu - \bar{x}|}{\mu} \times 100 \qquad [5.2]$$

where E is the percentage error.

We also have the coefficient of variation,

$$V = s/\mu \qquad [5.3]$$

From these equations

$$n = (t.V/E)^2 \qquad [5.4]$$

V is estimated from earlier data, as mentioned above, t is read from the table of t-distribution values, for some chosen level of confidence for ∞ degrees of freedom (DF), and E is a level of (acceptable or chosen) error. The use of DF = ∞ assumes that the population is infinite, that is the number of possible one-day samples which could be taken in the workplace could stretch to a lifetime. Since a year would accumulate about 200 days, or for 10-minute short term samples, a possible month's samples would be nearly 1000 (48 samples/day, 5 days/week and 4 weeks/month) the approximation of ∞ DF is acceptable.

5.3.2 Examples and Discussion

For a *normal* distribution with μ = 100 mg/m^3 and σ = 30 mg/m^3, which have been derived from experience or estimated from a short sampling exercise, and

choosing an error limit of 10%, and 95% confidence we select DF = ∞, for which t = 1·960.

$$n = (1{\cdot}960.30/100/0{\cdot}1)^2$$
$$= 34{\cdot}57, \text{ or } 35 \text{ to next integer.}$$

That is to say, in order to estimate the population mean of the concentrations, so that with 95% confidence the estimate would be within 10% of the "true" mean we would need to take 35 full shift (or 10-minute STEs) "samples".

This applies to normally distributed data. What if you suspect or know that the data are lognormally distributed? The Central Limit Theorem (see any good text book) tells us that the distribution of means is normal even for data samples which are lognormally distributed. As a consequence the same formula can be used to calculate the required sample size for lognormally distributed data.

In Section 4.6 the minimum variance unbiased (MVU) estimators for the population arithmetic mean and standard deviation of lognormally distributed data were discussed. The question arises "should these MVU estimators be used for such data?" since the Central Limit Theorem says we will not be interested in the population GM. The answer cannot be found in text books, but it would probably be valid, or perhaps be a more secure guess to use the MVU estimates, than to use the raw sample arithmetic mean and particularly the *sample* standard deviation from a trial sample (which can be very different from the MVU arithmetic standard deviation), even though the confidence limits for the MVU estimators cannot be calculated exactly.

The following data for shift average respirable dust concentrations (in mg/m³) measured over eight days for a man always doing the same job in a foundry are lognormally distributed,

1·27, 1·33, 1·36, 1·49, 1·67, 1·75, 1·80, 2·48

For how many days should samplers be deployed to estimate the man's true mean exposure within 10% error limits with 90% confidence?

GM of data = 1·608 mg/m³ (Best estimate of GM for population)
GSD of data = 1·244 (Best estimate of GSD for population)
AM of data = 1·644 mg/m³
ASD of data = 0·392 mg/m³
Minimum Variance Unbiased AM = 1·642 mg/m³ (Best estimate of AM for population)
Minimum Variance Unbiased SD = 0·359mg/m³ (Best estimate of ASD for population)
"CV" (from MVU estimators) = 0·219

From [5.4] $n = (1{\cdot}645.0{\cdot}219/0{\cdot}1)^2$
$= 12{\cdot}96 = 13$

If the sample CV were used the required sample number is 16. Generally the MVU estimators, especially the variance, are such that they will give much greater sample sizes than will the raw sample CV.

It should be noted that although this suggests that 13 shift "samples" need to be taken to get an estimate of the man's mean respirable dust exposure with 90% confidence within 10% limits, the real case will probably be even worse. If we were to use the logarithmic data (despite what the Central Limit Theorem tells us) the GSD of 1·2 is not very high, and a value of 2 or even more might be expected in many instances. Even so a GSD of 1·244 suggests a sample size of 58 for us to have 90% confidence that the log of the sample

geometric mean will be within ±10% of the true log GM (since in this case we are looking at the lognormal distribution), which may cause some surprise.

From these two examples it must be clear that the arithmetic mean (for the population) of exposure of an operative can only be estimated with any sensible confidence and with reasonable error limits by making an adequate (large) number of measurements. If you and your management are determined to be serious about making reliable estimates of long term (year to lifetime) occupational hygiene exposures you had best abandon the idea that the single full shift sample is telling you very much, other than allowing you to state in your report that "on the day when the measurement was made the concentration was …". Any interpretation beyond this, such as ascribing this value to the weekly, yearly or lifetime exposure, or even comparing it with an exposure limit (EL), assuming this itself has been derived sensibly, would seem to to be rather more than risky.

The only alternative to large sample sizes is to accept lower confidence and higher error limits.

5.4 The "NIOSH Method"

If the type of distribution and the relevant standard deviation are not known, and it seems imprudent to make assumptions for these, an alternative is to propose that (at least) one result in the sample to be taken from a population should be in the top T%, with C% confidence. In this case the population size N is known, as are T% and C% (or at least they can be decided), and from these data the sample size n can be calculated.

5.4.1 The Association of T%, C%, N and n

Once again you will be surprised at the relatively large number in your sample for reasonable values of T% and C%, particularly if the population size is not large. This approach is given in Leidel et al.[2] and by Crosby[3]. The presumption is that the population is homogeneous, a group of workers exposed to the same conditions, doing the same job, and without any distinguishable differences in the way they work or are exposed. Another example of a homogeneous population would be the 48 ten-minute samples which can be taken during a full 8h working shift using static or personal samplers, again assuming that the exposure is, so far as can be judged, uniform throughout the day.

As an aside, although this (and any other) method will tell you *how many* samples to take during the day, it will not tell you *when* to take them. This, too, needs to be decided, probably best by using a sampling timetable generated using random numbers.

Tables of sample sizes needed to ensure that one result will be in the top 10% or 20% (T%) with 90% or 95% confidence (C%) are given in the references quoted. These values were obtained from the following formulae, but for the sake of a tidy presentation in the reference sources the results have been rounded.

N = group size
n = sample size
$1 - \alpha$ = confidence (C%/100)
τ = proportion of group (τ = top T%/100)
$N_o = N.\tau$

For a population of infinite size ($N = \infty$)

$$n = \frac{\log \alpha}{\log (1-\tau)} \qquad [5.5]$$

This sets an upper sample size limit for given values of T% and C%. For groups of size N ($N<\infty$),

$$\alpha = \frac{(N-N_o)!}{(N-N_o-n)!} \cdot \frac{(N-n)!}{N!} \qquad [5.6]$$

$$= \frac{p_n^{N-N_o}}{p_n^N} \qquad [5.7]$$

In either form the confidence ($1 - \alpha$) is easily calculated on a pocket calculator with factorial or permutation functions, or on a microcomputer using the approximation given in Section 10.2.7. The calculation of sample sizes (n) for given group sizes (N), confidence ($1 - \alpha$) and top fraction (τ) is rather more tedious to set up as an iterative routine but is entirely practicable with care.

Sample sizes have been recalculated for group sizes up to 50 using the program SAMCON[4] and although the new values are very similar to those given by Leidel, there are some minor differences which are usually advantageous (smaller sample sizes) to the hygienist. These recalculated values are shown in Tables 5.1-4, page 28.

An examination of the tables will show some apparent anomalies. For example, in Table 5.1, for a group of size 32 or 33 a sample size of 14 is proposed, but for a group size of 28 a sample size of 15 is needed. The discontinuities and apparent anomalies, which have been smoothed out in the NIOSH tables, are due to the influence of N_o ($= N.\tau$) in the equations, since N_o must be an integer (the top 10% of a group size of 25 men = 3 men, not 2·5 men). SAMCON outputs either the % confidence, the required sample size or the data for Tables 5.1-5.4 when the other parameters in equation [5.6] have been input. Perhaps this is a case where you should report which statistic was used – the "NIOSH tables" or the calculated values of sample size.

5.4.2 Examples and Consequences

It frequently happens that the sample size used by a hygienist falls far short of that which should be used due to shortage of staff, equipment or time (money in all cases), but using equations [5.5]-[5.7] will show you (and perhaps your management) what little confidence there may be in the results obtained from too brief a survey.

For example if only 4 men can be sampled from a homogeneous group of 16, the confidence will be only 45% that the exposure of one of the four will be in the top 10% of the concentrations to which all 16 men are exposed (10% of 16 = 2). An alternative interpretation would be that the confidence is 73% that one of the four would be in the top 20% (20% of 16 = 4). For a group of 16 the sample sizes suggested by NIOSH are 12 and 8 to be 90% confident that one of the 12 or 8 should be in the the top 10% or 20% respectively. The corresponding sample sizes from Tables 5.1 and 5.3 are 11 and 7, both a small, but helpful, advantage over the sample sizes from the NIOSH tables, although still illustrating the need to take an adequate number of samples if any reasonable level of confidence is to be placed in the results.

Even a casual examination of these tables and earlier sections in this Chapter will show that too many sampling exercises in occupational hygiene are probably carried out with sample sizes which are far too small. Although this may be the general practice it is still helpful to understand why this is so, even allowing for the difficulty in identifying, let alone being able to work with, homogeneous groups of workers. The hygienist should be prepared to propose to management or client that it would seem only prudent (in the legal sense) to plan and implement adequate sampling surveys if sufficient data are to be collected in order that sound decisions on the expenditure of considerable sums for control measures can be made. Certainly an inadequate data base can lead to over- or under-specification (and corresponding over- or under-expenditure, with the probable need for additional improvements in the latter case) of control systems. A poor data base due to a poor sampling strategy also makes the comparison of conditions "before" and "after" much more difficult and decreases the confidence one can have in the comparison, if only in the statistical sense.

5.5 Final Remarks on Sample Size

From the two methods of estimating a sample size outlined above which are needed to say anything with any confidence about either the mean of an occupational hygiene data set, or to ensure that the highest of a sample is in the top T% of the group, you will have seen that large sample sizes are the order of the day. As has been explained this is due to the wide variation with time (and space if it applies) of occupational hygiene data. This must be compared with laboratory or workshop measurements when sample sizes will be much smaller – three or perhaps five – for us to be sure we have as reliable a result as we are likely to get. There are other methods of estimating sample size, some with greater precision, but their use hardly seems to be justified – for certain they will not predict *smaller* sample sizes! There is no substitute for a properly designed sampling programme with an adequate number of samples.

5.6 References

1 YATES, F., (1965) Sampling Methods for Censuses and Surveys, 3rd ed., Griffin & Co. Ltd., London.

2 LEIDEL, N.A., BUSCH, K.A., and LYNCH, J.R., (1977) Occupational Exposure Sampling Strategy Manual, NIOSH, US Dept. of Health, Education and Welfare, Cincinnati, Ohio. (DHEW Pubn. No. (NIOSH) 77-173).

3 CROSBY, T., (1982) Statistics of Compliance *in* Sampling, Statistics and Epidemiology, Birmingham October 1980, Institute of Occupational Hygienists.

4 SAMCON, a computer program in BASIC available from H & H Scientific Consultants Ltd. Leeds.

Table 5.1. Sample size n for top 10% ($\tau = 0{\cdot}1$) and 90% confidence ($\alpha = 0{\cdot}1$)

Sample size	8	9	10	11	12	13	14	15	16	17	18
For groups of size N	11	10	14	16	17	19	20	28	30	39	49
	12	13	15	21	18	24	26	29	36	40	50
					22	25	27	34	37	46	
					23	31	32	35	38	47	
							33	41	43	48	
								42	44		
									45		

If N <= 9 then n = N

Table 5.2. Sample size n for top 10% ($\tau = 0{\cdot}1$) and 95% confidence ($\alpha = 0{\cdot}05$)

Sample size	9	10	11	12	13	14	15	16	17	18	19	20	21	22
For groups of size N	11	13	14	15	17	18	19	20	27	28	30	38	40	49
	12			16	21	22	24	25	32	29	36	39	47	50
						23		26	33	34	37	45	48	
								31		35	43	46		
										41	44			
										42				

If N <= 10 then n = N

Table 5.3. Sample size n for top 20% ($\tau = 0{\cdot}2$) and 90% confidence ($\alpha = 0{\cdot}1$)

Sample size	4	5	6	7	8	9	10
For groups of size N	6	7	8	10	15	20,24,25	40
			9	13	18	28-30	45
			11	14	19	32-39	49
			12	16	22	41-44	50
				17	23	46-48	
				21	26		
					27		
					31		

If N <= 5 then n = N

Table 5.4. Sample size n for top 20% ($\tau = 0{\cdot}2$) and 95% confidence ($\alpha = 0{\cdot}05$)

Sample size	5	6	7	8	9	10	11	12
For groups of size N	7	8	9	10	14,15,17	19,20,23	25,29	35,39,40
			11	13	18,21,22	24,26-28	30-34	44,45
			12	16		31	36-38	48-50
							41-43	
							46,47	

If N <= 6 then n = N

6 PROBABILITY PLOTTING

6.1 Plotting Distributions

Usually the first method of plotting the distribution of a sample of results to be considered is the histogram. The idea is simple, the data being grouped into a series of ranges, and the frequency of occurrence of results in these ranges being plotted as in Figs. 2.1, 2.4 and 2.5. Many off-the-shelf computer graphics and statics programs offer these "bar charts" but at best they are not very helpful and at worst can be misleading. An example of how misleading histograms can be was suggested in Chapter 2. In essence the problem lies in the understandable desire to say that the results depicted in a histogram fit a normal (or some other) distribution without actually doing anything further to show that this is so, believing that the effort expended in acquiring the data needed to produce a histogram is evidence enough.

More helpful than the histogram is the cumulative probability plot, derived from the histogram, on graph paper graduated with the appropriate probability and variable scales. If the results fit the nominated distribution the cumulative plot will fall on a straight line. But although the fit to a straight line can be better judged by eye than the fit to a curve imposed upon the histogram, even this needs to be tested numerically.

6.2 Histograms and Probability Plots

In order to be able to plot a frequency histogram or a cumulative curve there must obviously be a sufficient number of data available in order to group them into the cells from which the histogram is to be constructed. Ten values are unlikely to give a very useful histogram. The use of probability plotting positions overcomes this problem and in theory at least as few as three points can be used to plot the estimated cumulative probability curve, in much the same way as a cumulative curve can be drawn from a histogram.

In addition, the use of probability plotting is not limited to small numbers of data points – in fact there is no (upper) limit to the number of points. The author believes that there is no longer any place (other than for purposes of demonstration) for histograms and cumulative probability plots derived from them. He bases this assertion on the following.

1. By collecting the data values into the cells, the original data are degraded. Perhaps 1000 data values would be collected arbitrarily into between 10 and 20 cells, giving only this number upon which to do any statistical tests, such as goodness-of-fit tests, or even calculating the mean and standard deviation. If Sturges' Rule (see Section 11.4) is used to calculate the cell numbers and cell boundaries there would be 11 cells for 1000 points.

2. Probability plotting of individual data points can be done for any number of points greater than 3, and the data are not degraded, since each point appears on the plot. This has a significant advantage over the histogram, or its derived cumulative curve, since the shape of the cumulative

curve is defined over its full length by the individual points, and deviations from a straight line, which can be of interest and importance, are apparent, rather than being hidden within cells. This effect can be very easily missed in the end cells of histograms.

3. Considering 1. and 2. together, the labour of producing a histogram or a cumulative probability plotting curve with modern computing aids will be the same – simply the entry of the original individual data (assuming you have suitable programs).

There thus now seems to be no justification for adhering to the use of histograms and cumulative probability curves derived from them as a basis for statistical analysis. Although they are still of residual interest, too often they are incorrectly constructed, with no heed paid to the underlying distribution and their theory will not be discussed further.

6.3 Probability Plotting Positions

The idea behind probability plotting positions is not hard to understand, nor is their calculation difficult. They are used to locate the individual values of small (or large) samples of *ranked data* on the cumulative distribution curve of the presumed population from which the sample was drawn. The data are first arranged in increasing order (ranked) and then each point is allocated a percentage plotting position on the probability scale. It is for this reason that the plotting positions are frequently called "Rankits" (by analogy to Probits and Logits which are plotting methods used in studying the effects found in toxicology).

6.3.1 Selection of Appropriate Plotting Positions

Statisticians (and users of the technique of probability plotting) seem to have different ideas of what values, in probability or percentage terms, the probability plotting positions should have, and of how they should be calculated. Imagine that the second ranked concentration of a sample of five will, on (some sort of) average, plot at a certain percentage value, say 30·36%, on the cumulative percentage or probability scale, Fig. 6.1. Repeating the sample of five will give some other concentration for this second ranked point, but this second point from the repeat sample will still be plotted, according to the rule for the second point, at 30·36%.

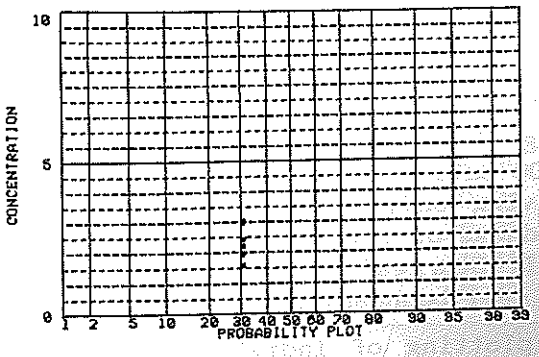


Figure 6.1 Repeated plotting of second point.

This gives a vertical spread of concentrations at 30·36% for each of the second highest values from a series of samples of size five. The process of using plotting positions is thus only an approximation to the full cumulative probability curve, but it does allow useful deductions to be made from a single quite small sample. Also there is always dispute about which "average" plotting position should be used – mean, median or mode (see any introductory text book), and what formula gives which. Filliben[1] and Cunnane[2] are among those who have considered some of the problems.

The differences in plotting position given by choosing the mean or median and in the formulae selected from calculating them are generally not great, particularly for large sample sizes, but it may be of some comfort to use one which has been shown to give the "best" representation of plotting position for a particular distribution. especially when only a few points are available, as so frequently occurs with occupational hygiene data.

6.3.2 Tabulated Values of Rankits

Plotting positions for normally (and lognormally) distributed data are tabulated in Pearson and Hartley[3], Table 28, and also in Leidel et al.[4] Table I- 1 for sample sizes from 2 to 50. In the first reference they are expressed as normal deviates and in the second as percentage plotting positions – they nevertheless correspond one with the other. While the NIOSH table can be reduced to a computable form, the initial percentage values for each sample size still need to be stored in an array in a computer program. For sample sizes >50 NIOSH suggest using (i – 0·5)/N for the ith point in a sample of size N.

6.3.3 "Universal" Plotting Formulae

In addition to the formula above, the simple formula i/(N + 1) is often quoted and used for *all* distributions, e.g. King[5] and Kennedy & Neville[6]. This formula is *exactly* true only for the mean positions of the uniform distribution. The mode of the plotting position for the uniform distribution is (i – 1)/(N – 1). None of these formulae is used in this book but they are included here so that you know what they are if you come across them. Whichever plotting position formula is used the points are distributed symmetrically about 50% on the *x* (probability) axis, or, in terms of standard deviations, about zero. Also the greater the number of points to be plotted the farther from the centre will the outermost ones be.

6.3.4 Filliben's Formula

Almost all investigators agree that probability plotting formulae for different distributions ((log)normal, (log)extreme value and others) have the general form

$$p = (i - a)/(N + b), \text{ where } a = (1 - b)/2 \qquad [6.1]$$

For normally (and by implication lognormally) distributed data Filliben[1] chooses to use the median plotting position, for which the last (Nth) and first *exact* plotting positions are

$$p_N = 0{\cdot}5^{1/N}, \text{ and } p_1 = 1 - p_N \qquad [6.2]$$

Remaining values are given by

$$p_i = (i - 0{\cdot}3175)/(N + 0{\cdot}365) \quad (1<i<N) \qquad [6.3]$$

Percentage probability plotting positions, needed to plot sample data sets on commercial probability paper, are of course given by $P_i = 100 \times p_i$. For computer plotting and statistical calculations the probability positions need to be transformed to the corresponding normal deviates.

Filliben chooses the median plotting position, rather than the mean, in order to develop his concept of using the probability plot correlation coefficient (ppcc), r, to test the goodness-of-fit of a sample to a particular distribution. By using the median, r becomes independent of the mean and variance of the sample for normal populations, and the ppcc r is still the best test (according to Filliben) for other (non-normal) distributions. For formulae to calculate the plotting positions of other distributions you should see, for example, Cunnane[2]. Filliben's formulae, [6.2] and [6.3], are not the best for the (log)extreme value distributions, nor probably is $i/(N + 1)$, which is to be found in the fundamental works by Gumbel on these distributions and in Kennedy and Neville[6].

So for a sample size of 5 the fifth (highest value) will be plotted at 87·06%, the first (smallest value) at 12·94%, and the second, third and fourth at 30·36%, 50%, and 68·64% respectively. The end points for the sample of 1000 mentioned above are at 0·069% and 99·931%, and the second point will be at 0·168%. The 500th and 501st points will plot at 49·95% and 50·05%. This is, as would be expected, a greater spread, with much tighter packing, than is the case for five points.

6.4 Correlation Coefficient

Most data which the hygienist will come across can usually be approximated by either a normal or lognormal distribution. You should be aware that there are many other distributions which may occur, albeit more rarely, in particular the extreme value distributions, but these are not considered further.

From this point there is a heavy computational load. It must be clear that while all this computation can be done, writing down tables of intermediate results, on a suitable programmable calculator, it is no effort for a computer once the program has been written. Nor is there a risk of data input errors using a computer once the program has been proved and the raw data been checked and corrected if necessary.

Having ranked the data in increasing magnitude, and allocated percentage plotting positions (rankits), the plot on normal or lognormal probability paper can be tested for the best fitting distribution.

First, the two means and standard deviations are calculated for the raw data and the logged data. If the data are distributed lognormally the Minimum Variance Unbiased estimates of the population arithmetic mean and variance (see Section 3.6) can be calculated from the geometric mean and standard deviation, or more likely from the corresponding log values, and from N.

Next, the percentage plotting positions must be converted to their corresponding standard deviations. This is most easily done using the polynomial approximation shown in Section 10.2.2. This converts the non-linear percentage scale to a linear one of normal deviates, as shown in Fig. 2.3. The data values are then handled in raw and logged form in parallel while calculating the correlation coefficients and regression equations. In other

words values on the x-axis are now in terms of standard deviations, and on the y-axis as raw ranked data in one case and ranked data logarithms in the other.

In this section the notation follows the usual one of using x for the abscissa (in this case the linear scale of standard deviations) and y for the ordinate, (the data or their logarithms). Do not confuse the use of x and y here with their use as data values and their logarithms elsewhere in this book.

The correlation coefficient, r, is given in its general form by

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{[\Sigma(x_i - \bar{x})^2 . \Sigma(y_i - \bar{y})^2]}} \qquad [6.4]$$

Since $x_i = -x_{n-i+1}$ and $\bar{x} = 0$, the plotting positions being symmetrical about 50% probability (or 0 ND), the formula reduces to

$$r = \frac{\Sigma x_i y_i}{\sqrt{\Sigma x_i^2} . \sqrt{\Sigma(y_i - \bar{y})^2}} \qquad [6.5]$$

Because the data have been ranked in increasing order the correlation coefficient is always >0, and in fact has minimum values, related to the number of points. For these reasons the usual published tables which give the significance of the correlation coefficient given by [6.5] must not be used. The significance table provided by Filliben should be used, or the approximation to it given in a BASIC program by Dhanoa[7].

The distribution which has the higher correlation coefficient is assumed to be the more likely one, and subsequent operations are usually carried out on the data for this distribution. The alternative distribution may be selected if desired, and all the required information should still be available in the computer.

6.5 Regression Equation

The least square regression equations, (y on x in the usual nomenclature), of data and log data on standard deviation are calculated in the usual way, remembering that the symmetry of the plotting positions about zero SD can again simplify the calculation, although why this should be of much concern to a computer might be debated.

$$m\ (\textit{slope}) = (n\Sigma x_i y_i - \Sigma x_i . \Sigma y_i)/(n\Sigma x_i^2 - (\Sigma_i)^2) \qquad [6.6]$$

$$c\ (\textit{intercept}) = (\Sigma x_i^2 . \Sigma y_i^2 - \Sigma x_i . \Sigma x_i y_i)/(n\Sigma x_i^2 - (\Sigma x_i)^2) \qquad [6.7]$$

This is the way calculators handle regression (and correlation) data. You should examine the memories of your calculator to confirm this. You will find registers for Σx, Σy, Σx^2, Σy^2, and Σxy. These equations involve the subtraction of two large positive quantities one from the other, and is potentially inaccurate yet the method is widely taught. A better way which reduces the likelihood (and size) of errors arising from this cause can, like the calculation of the estimate of the standard deviation, be done on computers by first calculating the means of x's and y's, $\bar{x}$, and $\bar{y}$.

The straight line passes through the point $(\bar{x},\bar{y})$, and if we put

$$X_i = (x_i - \bar{x}) \text{ and } Y_i = (y_i - \bar{y})$$

then $$m = \Sigma X_i Y_i / \Sigma X_i^2 \qquad [6.8]$$

and $$c = (\Sigma X_i^2 . \Sigma y_i - \Sigma x_i . \Sigma X_i Y_i)/(n\Sigma X_i^2) \qquad [6.9]$$

This last equation (for c) is slightly better than first calculating the slope m and then substituting in

$$\bar{y} = m\bar{x} + c$$

especially when c is close to zero (although this is *never* the case for data fitting a lognormal distribution). It would seem good computing practice to use the equations [6.8] and [6.9] when writing the program. You are unlikely to notice the extra time to work through two "FOR - NEXT" loops, even on "slow" computers, remembering that no computers are slow (ask anyone who recalls doing statistics on narrow ruled foolscap fly, with pencil and mechanical calculator), although some may be faster than others.

It should be noted that in some books it is suggested that the slope of the regression line (or more properly the slope of the cumulative "curve" drawn by hand on commercial probability paper) can be taken as the standard deviation. This might have been a short cut of adequate accuracy before calculators and computers were as widely available as they are at present. Certainly with computers only one entry of the original data is sufficient to calculate all the parameters needed without resorting to graphical methods. Plotting the results nevertheless is still a useful aid in visualising the results, and perhaps confirming that the regression line *does* intersect the 50% line at the mean, and highlighting deviations from a straight line, as may frequently happen, at either end.

6.6 Confidence Limits

The method of calculating the confidence limits of the means, both arithmetic and geometric, has been discussed in Chapter 3, and it has also been pointed out that the confidence limits of the MVU estimate of the arithmetic mean of lognormally distributed data cannot be calculated.

Having got the regression equation, other confidence limits can be calculated, including those:-

1. for the slope of the line,
2. for the confidence area of the regression line,
3. for the influence of the confidence limits of $\hat{y}$ on the confidence limits of the line,
4. for the confidence limits of an estimate of y.

This last is of particular interest to hygienists, since it allows an estimate to be made of what might have been had another sample of n results been taken. In other words it allows a prediction to be made, at some proposed confidence level, for example, of what the highest or lowest value could be for another set of random measurements taken from the same population.

The steps, following the calculation of the regression equation, are

1. calculate the standard deviations ($S_{y.x}$) of the points from the least square line

$$S_{y.x} = \sqrt{[(\Sigma(y_i - \{m.x_i + c\})^2)/(n - 2)]} \qquad [6.10]$$

This is sometimes known as the "standard deviation from regression" or the "standard error of estimate". Note that y_i are the data values and $(mx_i + c)$ are the values of y ($\hat{y}$) predicted from the x_i (plotting position) values. Note also the n - 2 degrees of freedom.

2. calculate the standard error of the predicted individual $\hat{y}$ values

$$E_{yi} = S_{y.x}\sqrt{[1 + 1/n + (x_i - \bar{x})^2/(\Sigma\{x - \bar{x}\}^2)]} \qquad [6.11]$$

Note that since $\bar{x} = 0$, the origin of the x axis, and that $\Sigma(x - \bar{x})^2$ values are already available from the regression calculations, only x_i has to be entered into the equation for each point, this being the rankit in SD values. Also since x_i is symmetrical about the origin and is squared in the equation, one calculation suffices for symmetrical points. The "1" has a profound effect on E_{yi} and there is nothing to be done about it, but the more points there are contributing to the line the smaller will E_{yi} be as 1/n decreases (sample size n increases).

3. decide on a confidence level (say 90%) and calculate the degrees of freedom = n - 2. Look up the value of t in the t-distribution table, using the two-tailed distribution, or calculate t using the approximation given in Section 10.2.4. For example a week's full shift samples, i.e. five days, DF = 3, and t = 2·353 for a confidence level of 90%.

4. calculate the confidence interval (about the line) for each value of $\hat{y}$, the y value on the line, not the data (or its log) value

$$\hat{y}_L = \hat{y} \pm t_{n-2}.E_{yi} \qquad [6.12]$$

The values of $\hat{y}_L$ are plotted for each rankit point, that is $t_{n-2}.E_{yi}$ for each value are plotted as offsets above and below $\hat{y}$ on the regression line. These are the 90% confidence limits we have in proposing that were we to take (or were we to have taken) another n samples (sampled on another week?) from the population, the results would not have fallen outside these values.

Note the four-fold symmetry of the offset values, since $\pm t_{n-2}.E_{yi}$ is obviously symmetrical above and below the line, and at the same time the ith and (n-i+1)th points are symmetrical about 0 SD or 50%. These symmetries can be used to ease the computational load for both calculators and computers.

6.7 References

1 FILLIBEN, J.J., (1975) The Probability Plot Correlation Coefficient for Normality, *Technometrics*, v. 17, no. 1, 111-117.

2 CUNNANE, C., (1978) Unbiased Plotting Positions - A Review, *J. Hydrology*. v. 37, pp. 205-222.

3 PEARSON, E.S., and HARTLEY, H.O., (1976) Biometrika Tables for Statisticians, v. 1, Table 28, Charles Griffin, High Wycombe.

4 LEIDEL, N.A., BUSCH, K.A., and LYNCH, J.R., (1977) Occupational Hygiene Sampling Strategy Manual, US Dept. of Health, Education and Welfare, NIOSH, Cincinnati, (DHEW (NIOSH) Pubn. No. 77-173).

5 KING, J.R., (1971) Probability Charts for Decision Making, Industrial Press Inc., New York, N.Y.

6 KENNEDY, J.B., and NEVILLE, A.M., (1976) Basic Statistical Methods for Engineers and Scientists, Harper International Edition, Harper & Row Inc., New York, N.Y.

7 DHANOA, M.S., (1982) A BASIC Computer Program which tests for Normality and the Presence of Outliers, *Laboratory Practice*, v. 31, no. 4, 330-331.

7 EXAMPLES OF PROBABILITY PLOTTING

7.1 Initial Comments

Although only three probability plotting cases will be discussed, this Chapter has been set aside for them, since the detail is probably novel and will require rather more explanation than has been given in earlier Chapters. The cases are all real, although you could generate your own data using random normal deviates (see Section 4.4). The purpose of this extended coverage is not to convert you to the belief that all data from occupational hygiene measurements are lognormally distributed, or even to show that very few are lognormally distributed, but simply to show that the best fitting distribution can be selected with minimum effort with the aid of a computer, although if a calculator is used, even a programmable one, the effort will be considerable but still worthwhile.

As we shall see in this Chapter and in Chapters 8 and 9, knowing the distribution can be of profound significance when applying statistics which are more usually applied to normally distributed data, or to any data under the assumption that all data are normally distributed. Some of the data in Chapters 8 and 9 have been tested for distribution. You may like to retrace the calculations to make sure that you get the same results and that the correct distributions have been chosen.

The outline of the necessary steps has already been given, but this Chapter will follow the arithmetic of a simple example in detail, before discussing the outcome of other examples. One program which may be used to determine the most likely distribution is given in the BOHS Technical Guide No.1[1]. The arithmetic used in this Chapter is taken from the OH Program FILPLOT[2] as are the figures which have been output by the Sharp PC1500 computer. FILPLOT has in fact been used as the model, even to the extent of invoking the operations which the program undertakes from time to time. This may be of help if you wish to write your own program, and should not intrude significantly while following the arithmetic.

A comment on the use of programmable calculators and computers was made in the last Chapter, and at this point perhaps it might be appropriate to comment further on their use. Rankits are easy to calculate. The polynomial conversion of percentage plotting points to standard deviations can also be done conveniently on a programmable calculator, although each rankit must be entered in turn. A calculator with statistical functions is convenient for the calculation of means and standard deviations and also regression equations and a few correlation coefficients, but this will call for at least two entries of the ranked data, depending on the calculator, but the labour and risk of input errors increases with increasing sample size. By the time it comes to calculating the confidence limits of another sample the calculator must, in practical terms, give way to the computer whatever the sample size.

Although the statistics are, of course, numerical the output of a graphical representation is always a help, and since the graph will not be used for reading off values, such as percentiles, as one might from commercial probability paper, the skeletal graticule and its reduced size is no disadvantage. Fig. 7.1 shows two typical probability graticules which might be output as part of the program.

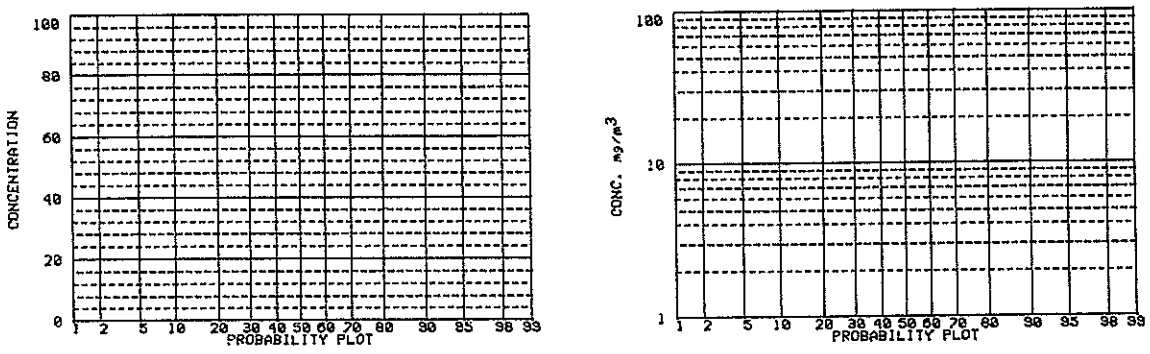


Figure 7.1 Normal and 2-cycle lognormal probability graticules.

7.2 Example 1. Four Full-Shift Dust Concentrations

These data are taken from an extended exercise in an ironfoundry – other results from the series are used in Chapter 9. The data are respirable quartz concentrations in mg/m³ measured on four successive days on a foundry worker. As has already been suggested the arithmetic is *at all times* carried out on the raw or logged data, exponentiating logarithmic values only when a numerical output is required for the benefit of mere mortals. Also only a few places of decimals are used here, although the computer of course runs with the full precision of which it is capable, rounding only for outputs.

7.2.1 The data inputs, then, are

0·036, 0·035, 0·129, 0·079.

7.2.2 After checking and correction if necessary the data are ranked, using a suitable sorting routine, the plotting positions are computed using [6.2] and [6.3], and the logarithms of the data are found (natural logs in the program). All data for the lognormal distribution are now worked through as logarithms only.

Ranked data	0·035	0·036	0·079	0·129
Plotting positions, %	15·91	38·55	61·45	84·09
Logs of data	-3·352	-3·324	-2·538	-2·048

Although the probability plotting percentages are output as shown, it should be remembered that only half are computed, the remainder being derived by the simpler and faster expedient of subtracting the values from 100. As has already been mentioned this symmetry can be, and is, used to advantage in the program.

7.2.3 Means and standard deviations for data and log data are computed, using [3.1], [3.2b], [3.4] and [3.5b]. In the "b" equations the values of $x-\bar{x}$, or deviation, is used.

Mean and SD of arithmetic data	0·0698	0·0445
Mean and SD of logged data	-2·8157	0·6359

7.2.4 Remembering the symmetry of % plotting positions about 50% and of SDs about 0, percentage plotting positions (% pp) are converted to SDs using the polynomial approximation in Section 10.2.2 (or the table of Normal Deviates).

% pp	15·91	38·55	61·45	84·09
SD pp	-0·998	-0·291	+0·291	+0·998

Only the second half (SD>0) values are calculated, the remainder simply being obtained by observing the symmetry and changing the sign.

7.2.5 The next step is to calculate, in parallel, the sums of squares and products needed to find the standard deviation from regression, correlation coefficient, and slope (the intercept is the mean) for the arithmetic and logged data. With the usual usage of x and y in regression analysis

Arithmetic	Sums of (data deviations)^2	0·005943
	Sums of data deviations × SD	0·106326
	Sums of (SD)^2	2·161553
Logarithmic	Sums of (data deviations)^2	1·213061
	Sums of data deviations × SD	1·530523

(Sums of (SD)^2, not recomputed = Sums of (SD)^2 for arithmetic data).

7.2.6 These data are now used in [6.5], [6.8] and [6.10] to get

	Normal	Lognormal	From
r	0·945	0·938	[6.5]
Slope	0·04918	0·70806	[6.8]
SD_{REG} ($S_{x.y}$)	0·01887	0·25431	[6.10]

7.2.7 At this point in the program parameters for both distributions are output and the choice is offered to work from now on with the distribution having the higher correlation coefficient. If this is r_{LNOR} the MVU estimates of AM and ASD will already have been output (Section 3.5).

7.2.8 Depending on the distribution chosen the regression equation

y = mx + c gives intercepts at 1% and 99% (x = ±2·32635 NDs, m and c from 7.2.3 and 7.2.6 above). These intercepts are calculated for the purpose

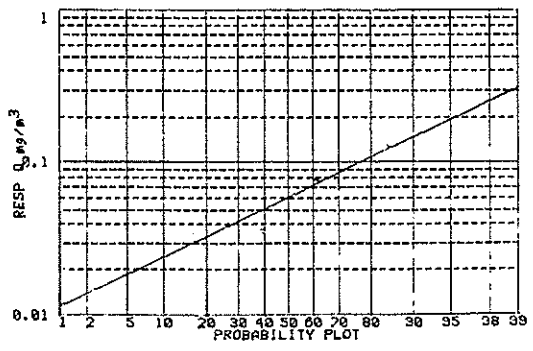


Figure 7.2 Ranked data and their probability plotting positions

of manual plotting on the appropriate commercial normal or lognormal probability paper. At this point there is sufficient data to allow Fig. 7.2 to be drawn.

Also the numerical 95% confidence limits for the mean are output using [3.1]-[3.2b], t being calculated in a subroutine (see Appendix 10A.1) for the required degrees of freedom, 3, (t = 3·1786 in this case) or read from the tabulated values (t = 3.182, the calculated value is <0·1% low).

7.2.9 Next the intercepts on the regression line of the probability plotting positions, $\hat{y}$, can be calculated, entering the SDs from 7.2.4 in the regression equation. Then the standard errors E_{yi} in [6.11] and confidence limits $\bar{y}_L$ in [6.12] for each of these intercept points are calculated. These confidence limits, of course, have + and - values, and can be thought of, for the sake of brevity, as "offsets" above and below the $\hat{y}$ values at the plotting positions. In fact the program combines [6.11] and [6.12]. and also for the printed output merely does this for the higher (+) offset of the highest and lower (-) offset of the lowest points.

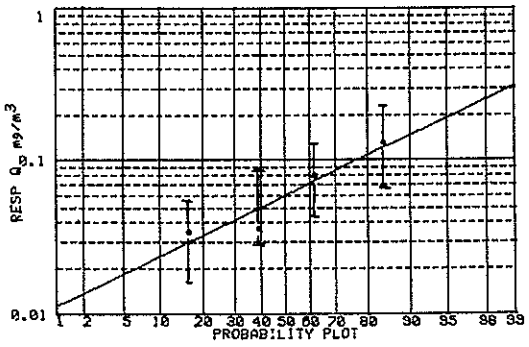


Figure 7.3 Confidence limits for the four plotting positions

For convenience the envelope of the offsets is calculated as a curved line approximated to by the computer as a series of straight segments. If the offsets were actually plotted as vertical lines defining the confidence limits at each plotting point the picture would become very confused for a large sample size. Fig. 7.3 illustrates the reality of the confidence limits for another random sample drawn from the same population.

The value of the confidence level is set at 80% (two tail), which is the equivalent of 90% (one tail) and the value of t is again calculated from the subroutine in the program for the (n-2) degrees of freedom for use in calculating the confidence limits E_{yi} in [6.12]. The calculated value is 1·883, the tabulated value 1·886.

Table for confidence envelope

%	SD	$\hat{y}$	±offset	
99	2·327	-1·168	0·928	($\hat{y}$ and offset are logs!)
95	1·645	-1·651	0·758	
80	0·841	-2·220	0·602	
50	0*	-2·816	0·535	
(20)	(-0·841)	-3·412	(0·602)	
(5)	(-1·645)	-3·981	(0·758)	
(1)	(-2·327)	-4·463	(0·928)	

* The polynomial gives -1E-07, which is near enough to 0.

The exact forms for 1 and 2 DF are not in the program – there is little expectation for precision from small sample numbers and little call for high confidence levels in occupational hygiene data.

The values in parentheses are not calculated using the equations since they are symmetrical about 50% in column 1, and 0 in column 2. The actual values of the offset points are given by ($\hat{y} \pm$ offset), which must be exponentiated to be plotted on commercial log probability paper, so that at 99% the envelope for the highest value expected (notionally from another sample from the same population) is

exp(−1·168 + 0·928) = 0·7866, and for the lowest expected value at 99%

exp(−1·168 − 0·928) = 0·1229. As has been pointed out this envelope has no real existence in fact, it merely defines the confidence limits the n values which $\hat{y}$ would have if plotted individually. It certainly has no meaning at 1% and 99%, except to enable the envelope to be drawn, since these values are beyond the lowest and highest plotting positions.

In the computer program used to work these examples this routine is also used to calculate the highest expected value from another sample of the same size, in this case 4, from the same population, except that the SD used in the regression equation is that already calculated for the highest plotted point at 84·09%. This is followed by calculating the standard error, E_{yi} from [6.11] and the offset value [6.12] $t_{n-2}.E_{yi}$, the one calculation doing for both the upper and lower values, the first being added and the second being subtracted from $\hat{y}_i$, the intercepts of the line at, in this case 84·091% and 15·91%, that is ±0·998 SD, to give, with 90% confidence, the highest and lowest expected values from another sample. The offset is 0·6265 which is added to $\hat{y}_4$ and subtracted from $\hat{y}_1$ to give −1·4825 and −4·1489, which exponentiate to 0·2271 and 0·0159.

7.2.10 The program also offers a graphical output, and a rerun of the same data in case the other distribution should of interest.

Obviously there is a considerable amount of computing involved, not beyond the capabilities of a pocket calculator but tedious. A computer is the ideal tool, not least because it always gets it right if the input data are correct. The computer output obviously does not include all the data above, being limited to that given below, and the optional graphical output, Fig. 7.4, here shown as presented by the Sharp PC1500.

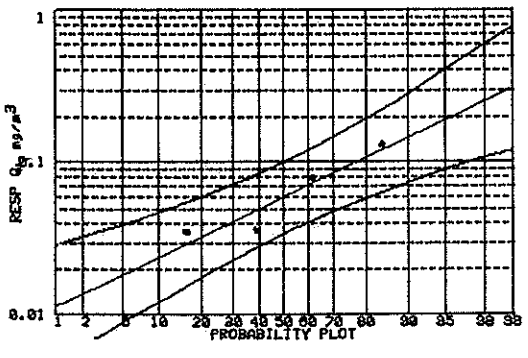


Figure 7.4 Probability plot for Example 1.

A typical text output from a computer program would be the data as input (for checking and if necessary correcting them and as a permanent record) and ranked, together with the percentage plotting positions for the ranked data, followed by

LOGNORMAL DISTRIBUTION R= 0.945, R FOR NORMAL= 0.938

SAMPLE ARITH MEAN= 0.0698, SD= 0.0445
MVU ESTIMATE OF ARITH MEAN= 0.0694, OF SD= 0.0428
GEOM MEAN= 0.0599 GEOM SD= 1.8887
JOIN UP 0.0115 & 0.3108 AT 1% AND 99%

WITH 95% CONFIDENCE THE MEAN LIES BETWEEN 0.0218 and 0.1645

WITH 90% CONFIDENCE THE MAXIMUM VALUE FROM ANOTHER SET OF 4 RESULTS WOULD NOT EXCEED 0.2271, THE MINIMUM WOULD NOT BE LESS THAN 0.0158

80% CONFIDENCE LIMITS ENVELOPE FOR OTHER Y VALUES

PLOT %	UPPER 80% LIMIT	LOWER 80% LIMIT
1	0.0292	0.0046
5	0.0398	0.0088
20	0.0602	0.0181
50	0.1023	0.035
80	0.1982	0.0595
95	0.4093	0.09
99	0.7865	0.1229

Not all the outputs from the program will be of interest each time it is run. If it is only intended to select the best fitting distribution only the correlation coefficients might be wanted. If you are concerned with STELS only highest value from another sample of the same size would be needed or if you are trying to define what sensitivity you really need in your sampling and analysis the lowest from another sample might be of most interest.

7.2.11 As was pointed out at the beginning of this section this data set was only one of a series. In fact there were 48 such sets arising from 8 men × 2 days × 3 concentrations (total dust, respirable dust and respirable quartz). Of these sets 31 were found to be lognormally distributed, at the $p = 0{\cdot}05$ level for Filliben's Probability Plot Correlation Coefficient (PPCC), r. In some cases the normal distribution was also significant at $p = 0{\cdot}05$, but 3 of the sets did not fit either distribution at $p = 0{\cdot}05$. In addition to the personal samplers a GCA 101 (β absorption) portable dust monitor was used to take some 530 readings in 50 data sets. Of these 21 were lognormally distributed, 22 were normal and 7 failed to fit either distribution at $p = 0{\cdot}05$. Again there were cases for which both distributions could be significant at $p = 0{\cdot}05$. In such cases it was usual to use the distribution with the higher PPCC. The effort which went into the work was not directed solely at a study of the form of the distributions but was utilized for occupational hygiene purposes. The determination of the distributions was simply a necessary step in the overall investigation. The frequencies with which the two distributions occurred show that it is necessary to test for both, and that either (and indeed both or neither) may occur.

7.3 Example 2. Short-term Concentrations

Suppose the following 10 personal exposures to formaldehyde were measured at random times for a continuous process, the sampling period being 10 minutes.

0·055, 0·088, 0·192, 0·401, 0·505, 0·612, 0·645, 0·654, 0·666, 1·132.

7.3.1 The procedure for an initial statistical analysis should be familiar by now, and first a somewhat simplistic approach will be followed.

1 Rank the data (they are already ranked), and calculate logs.
2 Calculate means and standard deviations.
3 Allocate percentage probability plotting positions.
4 Convert these to SDs.
5 Calculate the regression equations for arithmetic data against SDs and logs of data against SDs and the correlation coefficients.
6 Depending on requirements calculate the confidence limits for the mean of the selected distribution, and of concentrations for another sample of 10.
7 Plot the distribution.
8 Proceed to *apply* the statistics to your problem – the computer output is of little use on its own.

Working firstly on the simple assumption that all occupational hygiene data are lognormally distributed, we find that the following data can be calculated immediately:-

Arithmetic mean	0·495 ppm	Arithmetic SD	0·326 ppm
Geometric mean	0·360 ppm	Geometric SD	2·674
MVU est. of AM	0·548 ppm		

None of the measurements exceeds the Short Term Exposure Limit (STEL) value of 2 ppm.

Calculating the probability plotting positions and the confidence limits for another sample of 10 from the same population (same working shift of 48 such 10 minute periods) and plotting them in Fig. 7.5, we find that, with 90% confidence, there could be one (the highest) of the sample which would exceed the STEL, with a concentration of 3·05 ppm. This might cause some minor alarm. In an effort to massage the results without doing too much damage it could perhaps be proposed that the highest concentration of 1·132 ppm is a "statistical outlier" and should be omitted, after all it is about twice as high as the next highest (whatever else is done, don't do a statistical test!).

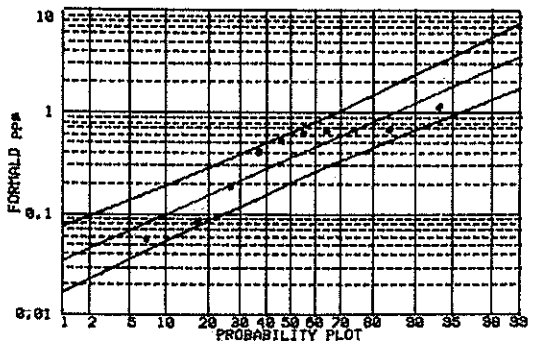


Figure 7.5 Probability plot for 10 concentrations.

This will reduce the means to:-

Arithmetic mean	0·424 ppm
Geometric mean	0·317 ppm
MVU est. of AM	0·467 ppm

And the standard deviations will change to

ASD	0·251 ppm
GSD	2·590

This, together with the reduction of n from 10 to 9 will influence the plotting positions, and the highest value expected from another sample. The result, in Fig. 7.6, is that this highest expected value is now 2·613 ppm, still above the STEL and not really very much of a return for the effort.

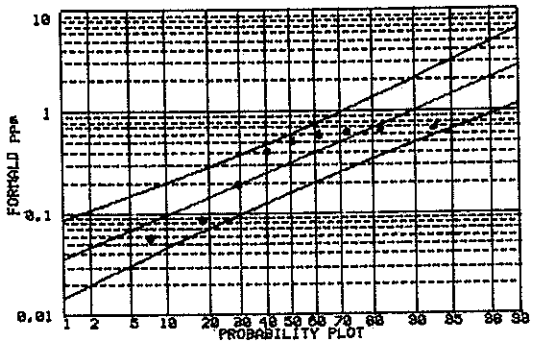


Figure 7.6 Probability plot for 9 lowest concentrations.

There is an alternative, which does have some occupational hygiene merit. From the notes made during the sampling exercise it is found that when the lowest concentrations were measured the man was not at his machine. These concentrations *might* be part of his daily exposure pattern, but not of his possible exposure pattern due to the machine. So they can be removed from consideration, restoring the highest concentration at the same time. This time the means increase to:-

AM	0·601 ppm
GM	0·543 ppm

Also the highest concentration expected from another (still lognormally distributed) sample would be 1·609 ppm, below the STEL at last, as can be seen from Fig. 7.7.

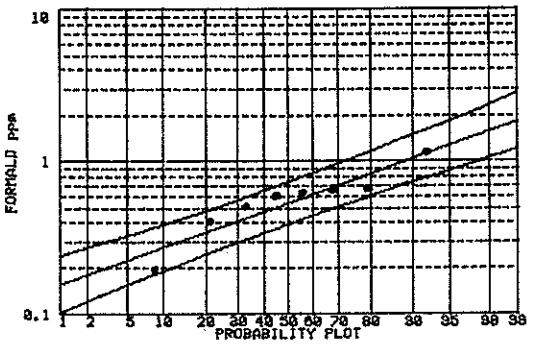


Figure 7.7 Probability plot for 8 highest concentrations.

Again this all seems to have been somewhat laboured – why not start as it should have been analysed?

7.3.2 It should have been decided *before* sampling began whether the purpose of the exercise was to estimate the man's STE under all working conditions if these typically included the absences, or whether it was only intended to consider the man/machine relationship. On this basis the decision would be made either to include or to exclude the two lowest concentrations from further consideration, on the grounds that the conditions under which the samples were collected indicate that these two values do not belong to the exposure population associated with the machine. Both options will be covered here, but only for the purposes of illustrating their effects.

The computer outputs for the analyses are given below, and the first thing to notice is that the distributions are, in both cases, normal and will be treated only as such.

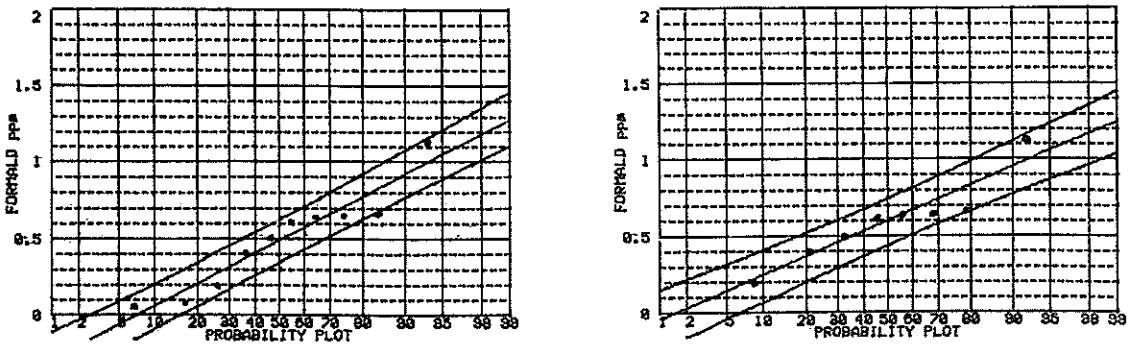
Formaldehyde, all points

NORMAL DISTRIBUTION R= 0.961, R FOR LOGNORMAL= 0.926
ARITH MEAN= 0.495 ARITH SD= 0·326
WITH 95% CONFIDENCE THE MEAN LIES BETWEEN 0.262 and 0.728
WITH 90% CONFIDENCE THE MAXIMUM VALUE FROM ANOTHER SET OF 10 RESULTS
WOULD NOT EXCEED 1.165 AND THE MINIMUM WOULD NOT BE LESS THAN-0.175

Formaldehyde, only top 8 points

NORMAL DISTRIBUTION R= 0.938, R FOR LOGNORMAL= 0.929
ARITH MEAN= 0.601 ARITH SD= 0·269
WITH 95% CONFIDENCE THE MEAN LIES BETWEEN 0.376 and 0.826
WITH 90% CONFIDENCE THE MAXIMUM VALUE FROM ANOTHER SET OF 8 RESULTS
WOULD NOT EXCEED 1.163 AND THE MINIMUM WOULD NOT BE LESS THAN 0.039

In the event the inclusion or omission of the two lowest concentrations does not affect the estimated highest value from another sample, both are below the STEL, although the means and SDs change. Notice, too, the prediction of a negative concentration! The probability plots are shown in Fig. 7.8



All 10 points Highest 8 points

Figure 7.8 Probability plots for formaldehyde concentrations.

7.4 Example 3. Welding Fume

These are some personal (MIG) welding fume concentrations. The only thing known about them is that they are from a group of men doing the same job, the

samples were taken according to BS 6691: Part 1:1986, and the first sample was for a short-term operation not really typical of the general work.

Total particulates, mg/m³: 0·8, 4·2, 6·5, 7·8, 8·8, 9·5, 11·4, 13·4, 14·0, 15·6, 18·0, 18·1.

Since the first value is reported to be not of the "homogeneous" group (that is the same man, the same machine, continuous work rate, and all conditions uniform) it can be immediately discarded, leaving 11 concentrations. The computer analysis, omitting the output of the data and the probability plotting percentages is

NORMAL DISTRIBUTION R= 0.988, R FOR LOGNORMAL= 0.974
ARITH MEAN= 11.57 ARITH SD= 4.64
GEOM MEAN= 10.62 GEOM SD= 1.58

JOIN UP 0 & 23.15 AT 1% AND 99%

WITH 95% CONFIDENCE THE MEAN LIES BETWEEN 8.45 and 14.69

WITH 90% CONFIDENCE THE MAXIMUM VALUE FROM ANOTHER SET OF 11 RESULTS WOULD NOT EXCEED 20.47 AND THE MINIMUM WOULD NOT BE LESS THAN 2.67
80% CONFIDENCE LIMITS ENVELOPE FOR OTHER Y VALUES

PLOT %	UPPER 80% LIMIT	LOWER 80% LIMIT
1	1.35	-1.36
5	4.61	2.16
20	8.5	6.27
50	1.65	10.5
80	16.88	14.64
95	20.98	18.53
99	24.5	21.79

Again, the data are probably from a normal distribution, with more negative predictions. The "zero" intercept of the regression line on the 1 percent probability line is actually -0·001014. Remembering that it is often imprudent to extrapolate beyond the limits of the data this value should not cause alarm, the data extending in this case from 6·11% to 93·89%, using Filliben's formula. Also the negative lower confidence limit of the envelope at 1% is somewhat more than mythical (see 7.2.9 above) again because the plot does not actually go lower than 6·11% (the plotting point of the lowest concentration) and the envelope can only define the ends of the confidence limits above and below the line at the actual probability plotting positions.

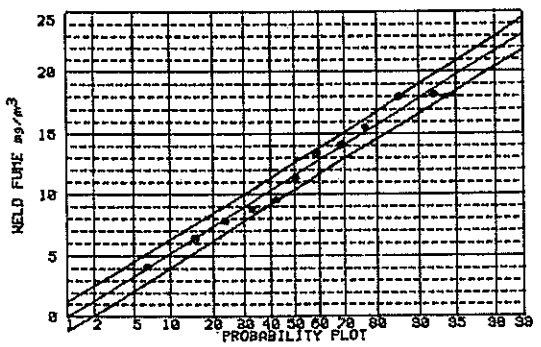


Figure 7.9 Probability plot for welding fume concentrations.

To what use the analysis was put is not known – perhaps it helped in a comparison test with conditions after local exhaust ventilation was installed, or was used to compare conditions in another plant doing the same welding operation, or with a different welding method (remember this was MIG welding).

7.5 Concluding Remarks

Probability plotting is not, of itself, much use. Its value lies in the use to which it is put. You will find that many statistics are applicable only to data which are normally distributed. Sometimes this is stated but often it is only implied. Confirming that the data are normally distributed when they might be expected to be lognormally distributed, or showing they are indeed lognormally distributed will allow you to decide whether the statistic can be used "as is" or whether its form (and interpretation) must be modified before it can be applied, as will be found in Chapters 8 and 9.

7.6 References

1 BOHS Technical Guide Series No. 1, 1983, "Statistical Analysis of Monitoring Data by Microcomputer", (£2.00), from the BOHS Office, 1 St Andrew's Place, Regent's Park, London NW1 4LB August 1983.

2 FILPLOT A program available from H & H Scientific Consultants Ltd, P.O. Box MT27, Leeds LS17 8QP.

8 PAIRED T-TESTS

8.1 Use of Paired t-tests

The use of the word "tests" in the plural is deliberate. In most text books it is treated in the singular, and is applied to data in pairs to test whether there is a statistically significant difference between them, or more correctly "Are the means of the differences between the corresponding pairs significantly different?" This is not the same as "Are the means of the two data sets significantly different?" which is the question considered in the next Chapter.

The occasions when such pairs arise by chance, especially in occupational hygiene, are rare and should be treated with the deepest suspicion. Most usually paired data occur as the result of a deliberate attempt to compare two effects. A common one might be "Does the new sampler give the same result as the old sampler?" The easiest way to test this is to set up an experiment to test the two samplers together under a variety of conditions. We will see how this worked in a real case.

The usually described test is for the mean arithmetic *difference* between the pairs, but some occupational hygiene data are lognormally distributed, and following the arguments proposed earlier, that is to work with the logarithms of the data, the paired t-test then becomes one of testing not for mean (arithmetic) *difference* between the pairs being different from zero but for the mean *ratios* between the pairs being different from unity.

This can be modified to test whether the mean ratio between pairs is different from some specified ratio other than one, just as one can test for an arithmetic mean difference between the paired data not being equal to some value other than the usual zero.

A critical point about the paired t-test is that it is not permitted to rearrange the data from the pairs, for example, by increasing rank or by random choice. The data must *always* retain their original paired association.

8.2 Calculation of Mean Differences

The differences between the two sets of paired data x_1, x_2, . . x_n and z_1, z_2, . . z_n are calculated.

Data	Data	Difference
x_i	z_i	$d_i = x_i - z_i$

for values of i from 1 to n, then the mean, $\bar{d}$, standard deviation, σ_d, and standard error of $\bar{d}$, $SE_{\bar{d}}$, are calculated.

$$\bar{d} = \Sigma d_i / n$$

$$\sigma_d = \sqrt{[\Sigma(d_i - \bar{d})^2/(n - 1)]}$$

$$SE_{\bar{d}} = \sigma_d/\sqrt{n}$$

DF = n - 1, there are n-1 degrees of freedom for n pairs of data.

From this data t is calculated

$t = \bar{d}/SE_{\bar{d}}$ which is then compared with the values in the t-distribution table with the correct degrees of freedom and at some chosen level of significance, say 5%. If $t_{calc} > t_{table}$ then, at the selected level, the mean differences are significant. Alternatively the probability, p, can be calculated directly.

8.3 A Case for The Paired t-test

Allen, Bellinger and Higgins[1] reported on a new sampler designed for use in the lead industry. The sampler was paired with the conventional (UKAEA) sampler and the pairs tested in the field firstly as static samplers with the sampling heads about 50 mm apart. Ten pairs of results were obtained for background lead-in-air concentrations in a battery pasting department. A second test was carried out with the UKAEA sampler mounted on fourteen workers' lapels, and the new sampler mounted very close to the orinasal region, and fixed in relation to it.

The concentrations (mg/m^3) measured for the static tests were

Run	UKAEA	New	d
1	0·043	0·049	-0·006
2	0·062	0·069	-0·007
3	0·072	0·075	-0·003
4	0·120	0·111	0·009
5	0·063	0·035	0·028
6	0·067	0·043	0·024
7	0·086	0·076	0·010
8	0·087	0·118	-0·031
9	0·037	0·029	0·008
10	0·051	0·049	0·002

Applying the paired t-test to these data we get

$\bar{d} = 0{\cdot}0034$ mg/m^3, $\sigma_d = 0{\cdot}0168$ mg/m^3, and calculated $t_9 = 0{\cdot}639$ which is not significant, since t_9 at 5% confidence is 2·262. In fact p is 53·8%. The conclusion is that the mean differences are not significantly different from zero and the lead in air concentrations are equivalent, when the samplers are operating as static samplers.

When sampling with the UKAEA sampler on the lapel and the new one at a true orinasal position between nose and mouth Allen, Bellinger and Higgins found the concentrations were

Worker	UKAEA	New	d
1	0·223	0·124	0·099
2	0·354	0·199	0·155
3	0·292	0·148	0·144
4	0·224	0·142	0·082
5	0·137	0·115	0·022
6	0·115	0·094	0·021
7	0·327	0·360	-0·033
8	0·174	0·208	-0·034
9	0·117	0·104	0·013
10	0·155	0·186	-0·031
11	0·110	0·167	-0·057
12	0·256	0·072	0·184
13	0·170	0·142	0·028
14	0·181	0·119	0·062

From these data we get

$\bar{d} = 0{\cdot}0468$ mg/m³, $\sigma_d = 0{\cdot}0769$ mg/m³, calculated $t_{13} = 2{\cdot}279$ which this time is significant at the 5% (two-tailed) level ($p = 4{\cdot}0\%$). The tabulated value of t_{13} at 5% significance is 2·160. The mean difference between the measured concentrations is not zero. The measured concentrations at lapel and nose/mouth are different. (Note that it is not correct to say that the *samplers* are different, since it has already been shown that they give the same concentrations when used in similar conditions). The paired t-test is one thing, the occupational hygiene is another – you have to take your courage in your hands and say "the lapel is not the best location for a 'breathing zone' sampler", at least in the pasting department.

8.4 Extending the Paired t-test

It may be that in some occupational hygiene data there is some suspicion or even an *a priori* expectation of the difference between pairs not being arithmetic, but a ratio. One method of analysis might be expected to be four times as sensitive as another or the new local exhaust ventilation system is intended to reduce background concentrations to an eighth of what they were. For such cases, although they would obviously give mean differences between pairs which were not equal to zero, a more informative test would be one which showed that the ratios of the pairs were different from one, or even in these two examples were not different from 4 or 0·125. This would be the null hypothesis.

Although the lead-in-air data do not suggest that the differences between pairs will be better expressed as ratios rather than arithmetic differences we can use them to illustrate the technique.

Since it is a Friday afternoon, raining and the computer is free you might feel inclined to apply Filliben's correlation coefficient test to the four data sets. You would find that all four data sets are lognormally distributed. You might also do a simple least square regression and correlation analysis on the static and personal data. In the first case the pairing may be lost, since the Filliben plotting routine rearranges both sets of data into increasing order. This has happened with both the static and personal sampler data sets. On the other hand a standard regression or correlation coefficient analysis of z on x obviously retains the pairing. If you superimpose the Filliben plots for the UKAEA and the new sampler for the static sampling you will find that they are essentially coincident (which is no more than you'd expect), but for the personal sampling they are offset and parallel.

Since the probability plots are done on the logarithmic scale (and knowing what we do about logarithms) it might not be out of order to think that the personal sampler offset represents not an arithmetic offset, but a logarithmic one, i.e. *ratio* differences, rather than the *arithmetic* differences we have tested for already. At this point proceed with caution, since the plots have been "fixed" by the ranking process and are no longer paired, but nevertheless it might make sense to test for ratio differences on the original (unranked) data. This is reinforced by looking at the least square regressions. The slope for the static samplers is close to unity, and the intercept is very small. Not very interesting on its own.

But for the personal sampler data pairs the slope is far from one, 0·39 in fact, again suggesting that there might be more than just a simple arithmetic difference between pairs.

The final column for the static sampling data has been changed to $d_{log} = (\log x_i - \log y_i)$

Run	UKAEA	New	d_{log}
1	0·043	0·049	-0·057
2	0·062	0·069	-0·046
3	0·072	0·075	-0·018
4	0·120	0·111	0·034
5	0·063	0·035	0·255
6	0·067	0·043	0·193
7	0·086	0·076	0·054
8	0·087	0·118	-0·132
9	0·037	0·029	0·106
10	0·051	0·049	0·017

Working with d_{log} instead of d, we find that

$\bar{d}_{log} = 0·0405$, $\sigma_d = 0·1176$, $t_9 = 1·089$ which is not significant ($p = 30·4\%$). The interpretation of this is that $10\uparrow\bar{d}_{log} = 1·0978$, in other words the *ratios* of the pairs are not significantly different from unity when operating as static samplers.

For the personal sampler pairs we can, from the data below, calculate $\bar{d}_{log} = 0·1174$, $\sigma_d = 0·1894$, $t_{13} = 2·320$ which is significant at the 5% (two-tailed) level ($p = 3·7\%$). In this case the log differences between the pairs is significantly different from zero, and in fact the mean of the differences of the logarithms is interpreted as a ratio between pairs of $10\uparrow\bar{d}_{log} = 1·305$. If we had had real evidence *beforehand* for the existence of this ratio between pairs, we could have tested the null hypothesis that the ratio between pairs on the lapel and at the nose/mouth was 1·3, and would then have found $t = 0·069$, which is not at all significant statistically, with $p = 94·6\%$.

The data for the personal sampler pairs were

Worker	UKAEA	New	d_{log}
1	0·223	0·124	0·255
2	0·354	0·199	0·250
3	0·292	0·148	0·295
4	0·224	0·142	0·198
5	0·137	0·115	0·076
6	0·115	0·094	0·088
7	0·327	0·360	-0·042
8	0·174	0·208	-0·078
9	0·117	0·104	0·051
10	0·155	0·186	-0·079
11	0·110	0·167	-0·181
12	0·256	0·072	0·551
13	0·170	0·142	0·078
14	0·181	0·119	0·182

You should beware of attempting to work with *ratios* in the third column of the tables above, and summing the ratios before dividing by n, since you will get the wrong answer. If you insist on working in ratios at this point instead of log differences they must be "summed" according to the arithmetic (rather than using logarithms) method of calculating the geometric mean,

(Geometric) mean ratio = $\sqrt[n]{[\Pi(x_i/y_i)]}$

$= \sqrt[n]{[x_1/y_1 \times x_2/y_2 \times x_3/y_3 \times \ldots \times x_n/y_n]}$.

(Π represents the operation "multiply the terms", just as the more familiar Σ means "sum the terms".)

Working with logarithms is easier, and retains the concept and actuality of working with the "differences between pairs". It also means that the same program can be used on a computer or calculator, substituting logarithms of data for the original data when testing for a ratio different from 1:1 or 1:r, if you have some reason for testing this alternative.

8.5 A Simple Example

Two on-filter samples are to be analysed by direct methods, X-ray diffraction and infra-red absorption. The analysed masses of quartz on the filters by these two methods are 30 and 22 mg for the first filter and 32 and 26 mg for the second. The differences are 8 (30-22) and 6 mg (32-26). The mean of the differences is 7 mg and the SD 1·414 mg. The SE of $\bar{y}$ is 1·414 and $t_{calc} = 7$, with 1 DF. The (two-tailed) tabulated value of t at 5% and 1 DF is 12·706, and on this slender evidence it can be said that the mean differences are not significantly different from zero. It might give us just a little confidence to carry on with a full-scale comparison between the two analytical methods.

8.6 When Statistics are Unnecessary

Over twenty years ago the following personal dust concentrations were measured using personal samplers worn by two fettlers in an ironfoundry. The men operated with the same tools on the same castings and, by mutual agreement, fettled the same number of castings each day. As can be seen from Figs 8.1 and 2 the work was done without local exhaust ventilation, more or less in the open, on trestles and blocks.

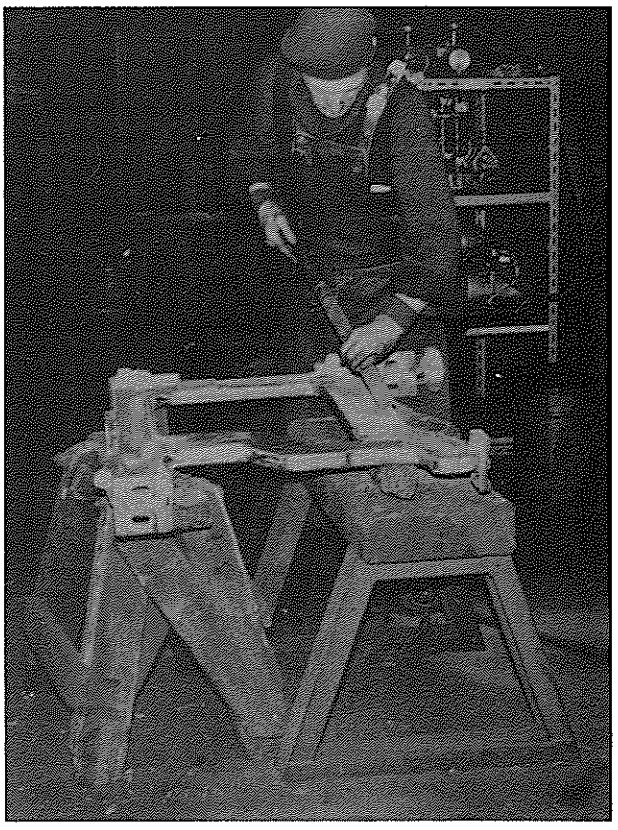
Their dust exposures (in mg/m³) were

	Fettler "A"	Fettler "B"
Respirable dust	4·23	1·45
Respirable quartz	0·694	0·192
Percent quartz in resp. dust	16	13

The concentrations are quite different, although the work, and the way it is done, is the same. The quartz contents are the same, so the cleanliness or otherwise of the castings does not seem to be important. To explain the difference in respirable dust and quartz concentrations the records made on site must be consulted, in this case the photographs. These remind the investigator of what he saw on site, that Mr. "A" was a short man, who was bent so close to the work supported on relatively high blocks that he felt the need for eye protection. On the other hand Mr. "B" was tall and an arthritic, (notice the wrist bands, once thought to be a useful prophylactic) standing upright, working on trestles and much further away from the dust source (and hence in no need of eye protection!). Even had there been a week's results for both men available for statistical analysis they would never have *explained* the reason for the differences in the concentrations.

8.7 Reference

1 ALLEN, J., BELLINGER, E.G., and HIGGINS, R.I., (1981) "A full shift true breathing zone air sampler and its application to lead workers", *Proc. Inst. Mech. Eng.*, v. 195, no. 25, pp. 325-328.



Fettler "A" Fettler "B"

Figure 8.1. Fettlers in an ironfoundry. (See text for discussion)

9 F– AND T– TESTS

9.1 F–test, and t-test on Means for Small Samples

The application of the unpaired t-test to the difference of the means of two sample sets of data is rather less subtle than the paired t-test, since the data differences are not tested in pairs, but the test is only applied to the difference between the means of the two samples. On the other hand, since the data are not paired, the samples do not need to be of the same size. Although it may not be the best statistics, you could still apply the test discussed in this Chapter to a difference between means of two sets of data where one, or possibly more, result had been lost. This could arise due to the caprices of occupational hygiene sampling in a large experiment originally designed for a paired t-test. In such cases you should at least be honest, and state which statistic you have used and why, rather than simply report that "a significant difference was (or was not) found between the two sets of results".

The test is in two parts – the F-test (or Variance Ratio Test) followed by the t-test. The F-test is necessary since the application of the t-test assumes that the two data sets come from populations with the same (or similar) standard deviations, and hence variances. The means of the two populations may, of course, differ, and the t-test is applied to detect whether this difference between the means is significantly different from zero. This is the "null hypothesis", although the test can be modified to test for a difference other than zero, the null hypothesis then being that the difference between means is not different from the proposed one.

In the variance ratio test (F-test) it is implicit that the populations from which the two samples are drawn are normally distributed. This is usually the case for laboratory measurements, such as analytical results by two methods, and in workshops when measuring the diameters of turned components coming from two lathes. It is this usualness which perhaps allows the frequent application of the t-test without first performing the F-test.

An alternative test is used when the sample sizes are large, but the small sample test is the only one considered here, since this is the most likely situation you will meet when handling occupational hygiene data.

9.2 The F–, or Variance Ratio, Test

It is quite common, although not correct, simply to apply the t-test to the difference of the two means. It is always worth while applying the F- (variance ratio) test to the data before the t-test is applied.

$$F, \text{ the variance ratio, } = \frac{s_1^2}{s_2^2} \qquad [9.1]$$

where the groups are, by convention, arranged so that s_1 (the sample, or "(n-1)" standard deviation) for the first group is greater than s_2, so that F is greater than 1. The closer F is to 1 the closer are the variances.

This is the optimum condition for applying the t-test, although statisticians assure us that the t-test is a "robust" test, and can be applied with some degree of success (rather than "confidence") even to sets of data with widely disparate variances, hence the common omission of the F-test. Nevertheless it would seem only prudent to remember to apply this test to data which might be from quite different distributions. If both the F- and t-test are incorporated into the same calculator or computer program this will ease the labour of the calculations and at the same time ensure that "a result" isn't being churned out for the t-test alone without regard to the applicability of the test to the input data.

Tables giving the critical values of the F-distribution are available for deciding the significance of F, using $n_1 - 1$ and $n_2 - 1$ (numerator and denominator) as the degrees of freedom. The significance can also be computed, as in the OH Program "F&T TESTS", based on a routine given by Lee and Lee (see bibliography, Chapter 1).

If no tables are available to determine the significance of F, nor a computer program with the necessary routine, the two values of F from the raw data and the logged data will give an *indication* of the more likely distribution to which both belong (normal or lognormal), since the value of F for the more likely distribution will be the nearer to 1. This will indicate to which distribution the t-test should be more properly applied. This use of the F-test is not considered in any text books known to the author, and is not as good a test as applying probability plotting, correlation and regression to *both* data sets separately for both distributions. Perhaps the F-test is more of an indicator for the distributions of both groups considered together, rather than a test for either group considered separately. Obviously the F-test is applied to both distributions simultaneously. This degrades the information in the data, much as histograms do (see Chapter 6).

Although it is customary to ensure that $F \geq 1$, it is not absolutely necessary, since the significance of F is the same as for 1/F.

9.3 The t-test on Differences of Means

For unpaired data the statistic of interest is $|\bar{x}_1 - \bar{x}_2|$, which is the modulus or absolute value (ABS on a computer) of the difference between the means, calculated as usual from the values of the two series, 1 and 2.

The estimate of the population variance, s_c^2, is also needed

$$s_c^2 = \frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{(n_1 - 1) + (n_2 - 1)} \qquad [2b]$$

The SD of the difference of the means, s_d is also calculated, using s_c^2 from [9.2].

$$s_d = s_c\sqrt{[(n_1 + n_2)/(n_1 n_2)]} \qquad [9.3a]$$

$$= s_c\sqrt{(1/n_1 + 1/n_2)} \qquad [9.3b]$$

The significance of the difference between the means is measured by the ratio of the difference to its SD, i.e.

$$t = \frac{|\bar{x}_1 - \bar{x}_2|}{s_d} \qquad [9.4]$$

This calculated value of t is compared with the value in the t-table for a chosen significance level (two-tailed) with $(n_1 + n_2 - 2)$ degrees of freedom. As with the paired t-test the higher the calculated value of t, the greater is the significance of the difference. For $n_1 = 4$, $n_2 = 5$, DF = 7, and a calculated t value of 1·895 the conclusion would be that the difference (between the means) was not significant, since there is a 10% probability of such a difference arising by chance. A calculated t value greater than 5·4, however would be significant, since such a value (from $|\bar{x}_1 - \bar{x}_2|$), would have less than 0·1% probability of arising by chance.

9.4 Examples of the F– and t–tests

As usual only the various marker data are given to help you to ensure that you are on the right track when you work the examples through.

Example 1. Two methods of injecting analytical samples of 6-chloro-o-cresol into a mass spectrometer were tried to see if there were any difference between them. The analytical results were compared, using the F- and t-tests, since the number of analyses were different in each group, and the analyses are not, in any case, paired. (6-chloro-o-cresol isn't a very common chemical, and is never used in disinfectants, despite its relationship to the chlorinated phenols. Its potential for tainting food is measured in ppb, and it can be absorbed from the air at these levels into food).

The results were as follows:-

Method 1. 7·77, 7·41, 7·84, 7·86, 7·63

Method 2. 7·57, 7·68, 7·87, 7·65, 7·60, 7·67.

The arithmetic means for Methods 1 and 2 are 7·7020 and 7·6733, and the standard deviations are 0·1865 and 0·1052.

The variance ratio, F, from [9.1] is 3·142 (Variance = SD^2). From the F-ratio critical value tables, this is not significantly different from 1 at the 10% level. In fact p = 12·1% so it is safe to proceed with the t-test.

From [9.4] the calculated value of t is 0·322, which again is not significant (p = 75·5%, for 4 + 5 = 9 DF) and we can assume that either method of sample injection into the mass spectrometer is as good as the other. More formally the difference between the means is not significantly different from zero.

The following sections will help you to trace alternative routes of analysis. If the logs of the data are tested in the same way the Geometric Means are 7·7002 and 7·6727, and the GSDs 1·0247 and 1·0137. The variance ratio of the log-variances (not the GSD^2) is 3·220, still not significant, and t is 0·307. You should notice that, like the paired t-test for lognormally distributed data, this is testing for $(\log GM_1 - \log GM_2) = 0$. That is testing for $\log (GM_1/GM_2) = 1$ since $\log 1 = 0$. So in this case the *ratio* between the geometric means is not different from unity.

The two sets of data have been analysed using the probability plot correlation test. Arithmetic and geometric means and standard deviations are, of course, as above. For Method 1 the correlation coefficients are $r_{NOR} = 0·941$, $r_{LNOR} = 0·939$, and for Method 2, $r_{NOR} = 0·912$, $r_{LNOR} = 0·914$. Because there is so little difference between the respective correlation coefficients for each method (even if Method 2 looks more lognormal than

normal, if we are to believe the third place of decimals) it seems hardly worth while making much more of it.

Example 2. Two urea/formaldehyde resins used for binding foundry sand were suspected of evolving different concentrations of formaldehyde during mixing. An experiment was devised to test this. The results for Resin 1 were 11·82, 12·01, 13·17 and 13·93 ppm in the test rig. For Resin 2 the results were 11·26, 10·82, 10·82 and 9·88 ppm. The means are 12·73 and 10·70, but is the difference in means significantly different from zero? The F-test for the arithmetic data gives 2·937, a difference in variances which could occur by chance with 20% probability, so the t-test can be applied to the difference of the means. This gives $t = 3·531$, which is only 1·23% probable if the populations from which the two formaldehyde concentrations were drawn had the same variance and mean. The difference between the means is significant at the 5% level.

The tests on the logged data suggest that they will be just slightly more reliable than for the arithmetic data ($F = 1·973$, $p = 29·5\%$ and $t = 3·629$, $p = 1·1\%$), but the difference between the arithmetic and the logarithmic cases is trivial. Omitting the test on the logged data would hardly embarrass anyone, although in transposing the results to foundry practice it is possible that knowing a ratio of formaldehyde emissions might be more useful than having a simple arithmetic increment, which might allow us to say that we would expect a reduction of x% in formaldehyde emissions during mixing by using one resin rather than another (but hardly worth while in this case).

Example 3. A worker in an ironfoundry laid cores in moulds. Four full-shift personal samples were taken in each of two successive weeks for cores coated with a) a siliceous core wash in the first week, and b) a non-siliceous core wash in the second. The daily concentrations (mg/m³) of respirable dust (RD) were:-

Siliceous core wash (S) 1·11, 2·44, 0·78, 0·66

Non-siliceous core wash (N) 1·28, 1·13, 1·08, 0·99,

and of respirable quartz (RQ):-

Siliceous core wash (S) 0·401, 0·183, 0·190, 0·126

Non-siliceous core wash (N) 0·023, 0·025, 0·039, 0·024

The samples are not paired (except by chance by day of the week). The question is "has the use of non-siliceous core wash improved conditions?" and apart from the answer "Obviously", it will be of some comfort to show this statistically. Not all cases will be as clear-cut.

Knowing what we do of occupational hygiene data ("they are always lognormally distributed") it might be prudent to start by testing for the best fitting distribution.

For the RD S samples $r_{NOR} = 0·900$, and $r_{LNOR} = 0·951$ and

for the RD N samples $r_{NOR} = 0·983$, and $r_{LNOR} = 0·989$.

Since $r_{LNOR} > r_{NOR}$ it looks as if we will be using the logs of the data and testing for the ratio of the geometric means not being different from 1, rather than the difference of the arithmetic means not being different from 0 for the respirable dust concentrations.

The data give $GM_{RD\ S}$ = 1·0866, $GSD_{RD\ S}$ = 1·7882 and
$GM_{RD\ N}$ = 1·1152, $GSD_{RD\ N}$ = 1·1130.

F (for the log variances) = 29·474, a probability of 1% that the difference in variances could have arisen by chance—the t-test can be applied with only limited confidence.

t (for means of log data) = 0·088. If the parent populations have the same mean and variance this difference has a 93·3% probability of occurring. There is no difference between the geometric means by the ratio test.

If you work through the data, this time assuming that they are both normally distributed, you will find that

the data give $AM_{RD\ S}$ = 1·2475, $ASD_{RD\ S}$ = 0·8175 and

$AM_{RD\ N}$ = 1·1200, $ASD_{RD\ N}$ = 0·1214.

F = 45·355, greater than F for the logged data, some confirmation that we were wiser to work with logs. t = 0·309, not quite as good as for the logged data, but, bearing in mind the robustness of the test, enough for us to say that the difference between the means is still not significant. But in both cases F is getting rather far from 1 — both the arithmetic and geometric variances of the two parent populations are becoming uncomfortably different, not perhaps invalidating the t-test but making it at least risky to apply. Even so we are reasonably sure that there isn't a real difference from one week to the other for the mean respirable dust concentrations.

The case for the respirable quartz concentrations is more interesting. Using the same approach

for the RQ S samples r_{NOR} = 0·899, and r_{LNOR} = 0·944 and

for the RQ N samples r_{NOR} = 0·846, and r_{LNOR} = 0·862. Again we probably have two lognormal distributions to deal with,

and $GM_{RQ\ S}$ = 0·2047, $GSD_{RQ\ S}$ = 1·6242

$GM_{RQ\ N}$ = 0·0271, $GSD_{RQ\ N}$ = 1·2782.

Dealing with the logged data, F = 3·906, close to 1, p = 14·6%, the (log) variances are not significantly different and we can go ahead with the t-test with confidence. t = 7·442, (p = 0·03%) with very little chance of the ratio between the geometric means being 1. The geometric means are significantly different.

Working through the arithmetic data F = 256·316 (p = 0·041%). The value of F is so high that the two sets of data are most unlikely to come from parent normally distributed populations with similar variances, and the t-test is probably not valid — it might be robust but even with t = 3·26 (p = 1·73%) the evidence is much stronger for a ratio between the geometric means being different from 1 rather than the difference between the arithmetic means not being zero. The use of a non-siliceous core wash has certainly reduced the airborne respirable quartz concentration, and the reduction is best shown as a ratio, rather than an arithmetic reduction.

Remembering that at the time of the investigation the Exposure Limit for silica-containing dusts was

$$EL = \frac{10}{\%\ \text{quartz} + 2}\ \text{mg/m}^3$$

the ratio RD/EL can be calculated, which for concentrations below the EL should be <1. Again probability plotting and the F- and t-tests were used to show that

1) the distributions for S and N are both lognormal
2) the F-test is satisfactory for lognormal data (F = 11·6, p = 3·7%)
3) the t-test shows that the ratio of the geometric mean of RD_S/EL_S to RD_N/EL_N is different from unity (t = 6·395, p = 0·069%)
4) the value of F for the arithmetic (normal) case is 301·58, making it difficult to justify applying the t-test to the difference between the means of RD_S/EL_S and RD_N/EL_N. In fact t = 3·259, p = 1·73%, much poorer evidence than that provided by testing the difference between the log-means which is, of course, the log of the ratio of the geometric means.

Example 4. An interesting case concerns Cadmium in urine from workers in two departments at the same works. In the first group of 20 men the Cd in urines were

1, 1, 1, 3, 3, 4, 4, 4, 5, 6, 6, 6, 7, 7, 9, 9, 10, 11, 24, and 56.

In the second group of 31 men the results were

1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 7, 10, 12, 13, and 18.

These two sets are so obviously different, with means of 8·85 and 4·61, that the confirmation, using the t-test, is almost superfluous, or is it? Immediate calculation of t gives 1·8, with (20 + 31 − 2) = 49 degrees of freedom, giving p = 7·8%, so the difference is, in fact, not very significantly different from zero, and certainly not at the 5% significance level.

Applying the full F- and t-tests to the arithmetic and the logged data gives

Arithmetic:- F = 9·475, p = <0·01%, extremely unlikely that the variances are the same (the samples coming from very different populations), and t = 1·8, as above, but with very little justification for applying the test.

Logarithmic:- F = 1·409, p = 19·6%, indicating that it is highly likely that the variances of the *logged* data are the same, and t = 1·826 (with GMs of 5·35 and 3·34). Thus t has about the same significance as before, p = 7·4%, (except that there is now much more justification for applying the t-test) and the chance that the ratio of the GMs is different from 1 is not high, with about the same significance as the difference of the means.

Armed with hindsight the data can be tested for the best fitting distribution, using the techniques described in Chapter 6. We find the Filliben probability plotting correlation coefficients for the Cadmium-in-urine to be

Set 1, r_{NOR} = 0·726, r_{LNOR} = 0·964 and
Set 2, r_{NOR} = 0·893, r_{LNOR} = 0·971, confirming that both sets of data are indeed lognormally distributed in this case.

These F- and t-tests are quite interesting (perhaps disappointing), since they show that the arithmetic *difference* of the means is not *significantly* different from zero at the 5% level (which is *not* what intuition suggested) and, at a very similar level of significance, the *ratio* of the (geometric) means is not different from one. But lurking behind it all are the F-tests telling us that the arithmetic variance ratio is significantly different from

1, casting a shadow over the applicability of the t-test to the *difference* of the arithmetic means, while the F-test on the logged data suggests that we would, in any case, be better off, as statisticians, looking to the *ratio* test.

What this means in occupational health or hygiene terms remains to be decided, but there may be more significance in an increased *ratio* of Cd in urine between the departments than a simple arithmetic *difference*. We can always say that although the statistics do not support the proposition that the first set give a statistically significant higher AM or GM than the second set, there is, perhaps, other supporting evidence to be examined such as β_2-microglobulin, Cd in air results, and examination of work practices and processes (never to be under-estimated or ignored). Common sense might help, as would the application of a test for the better fitting distribution *before* embarking on the F- and t-tests, as described in Chapter 6.

9.5 Some Concluding Remarks

It is perhaps fortunate that in the examples discussed above the distributions for the data sets in each pair have been the same—either normal or lognormal (except for a trivial disparity in the 6-chloro-o-cresol case). You should notice that the non-normal case is not discussed in text books (it being assumed that the reader will only meet normally distributed data!), and although the extension of the use of the F- and t-tests to the lognormal distribution in particular is valid, (as it was in the paired t-test if we accept that we are now testing for the ratio of the GMs being different from 1) there will be days when one set of the two results is normally distributed and the other lognormally distributed. This would be shown by applying the probability plotting test before the F-test, and not by appling the F-test to the arithmetic and logged data. It is hard to say what the F-test would suggest in such a case. Should we test the sets for the difference or the ratio between the appropriate means? It would seem best to first test for the best fitting distributions and then test, using both types of F- and t-tests, and report results for both the normal and lognormal t-tests, commenting on any ambiguities, and leaving the problem of interpretation there.

10 SOME FORMULAE AND USEFUL NUMERICAL APPROXIMATIONS

10.1 Purpose and Sources

The purpose of including these formulae and numerical approximations is to assist hygienists to become better acquainted with the numerical manipulation of statistics and to overcome the problem of royalty fees which need to be paid if statistical tables are reproduced. The statistical tables required to carry out all the operations described in this book (except for the significance table for Filliben's probability plot correlation coefficient) are, in any case, available in any respectable statistics book.

The OH Programs which are available from H & H Scientific Consultants Ltd are all supported internally with routines which eliminate the use of statistical tables—that is what computers and calculators are for. The list of formulae is obviously incomplete but covers the main ones which can be applied to the statistics described in this book. The routines are either exact or approximations of sufficient accuracy. You may find others but before using them you should be sure that they are sufficiently accurate. For example alternative routines are available for converting confidence (or significance) and DF to Student's t, but these, and indeed the one quoted become increasingly inaccurate at low probabilities and low DF. Also Dhanoa (see Reference 7, Chapter 6) gives routines for calculating the significance of Filliben's PPCC, and of Grubbs' test for (single) outliers.

Many of the routines given here and elsewhere are based on the use of polynomial approximations.These are best evaluated whenever possible using "Newton's method" (usually known as Horner's rule). This applies to both calculators and computers. It avoids the evaluation of powers and is both quicker and more accurate. For example

$$(at + bt^2 + ct^3 + dt^4 + et^5)$$

is evaluated as

$$((((et + d)t + c)t + b)t + a)t.$$

The principal sources used in this compilation are:-

Abramowitz, M., and Stegun, I.A. (10th printing 1972) Handbook of Mathematical Functions, with Formulas, Graphs and Mathematical Tables. National Bureau of Standards, Applied Mathematical Series 55, U.S. Dept of Commerce, USGPO, Washington DC.

This is rather more than a handbook, and repays deep study by anyone wanting to explore the application of some arcane mathematics to computing. For example you might choose to use a continued fraction to calculate the normal integral, rather than the polynomial approximation given below.

Lee, J.D., and Lee, T.D., Statistics and Numerical Methods in BASIC for Biologists (out of print) and Statistics and Computer Methods in BASIC, both 1982, Van Nostrand Reinhold Co. Ltd., Wokingham. Both books are valuable for statistics and BASIC programs and both contain the routines used in OH Programs under a copyright arrangement. The sources are given here, but not the BASIC routines.

Cooke, D., Craven, A.H., and Clarke, G.M. (1982) Basic Statistical Computing, Edward Arnold, London. This book contains some more useful BASIC routines, although deeper exploration will show that a few of the algorithms are inferior to those in Lee and Lee, or which can be formed from formulae given in Abramowitz and Stegun. If you need the routine to generate random normal deviates you will find that the one given in the book is not correct, but the correct version is given in Section 10.2.9.

Two routines in BASIC are included in Appendix 10A to illustrate the point that there is nothing to fear in using these approximations in programs. Appendix 10A also lists OH Programs which are associated with the statistics described in this book.

10.2 List of Formulae and Approximations

10.2.1 Normal Integral (Area under the Normal Curve)

$Z(x)$ = Ordinate = $1/\sqrt{(2\pi)}.\exp(-\frac{1}{2}x^2)$

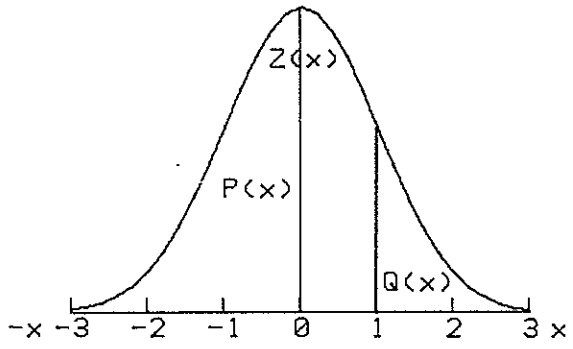
$P(x)$ = Area to left of x

$Q(x)$ = Area to right of x

$P(x) + Q(x) = 1$

$$Q(x) = 0{\cdot}5 - \frac{1}{\sqrt{(2\pi)}} \int_0^x \exp(-\tfrac{1}{2}x^2)dx$$

$$P(x) = \frac{1}{\sqrt{(2\pi)}} \int_{-\infty}^x \exp(-\tfrac{1}{2}x^2)dx$$



To derive P(x) from x use this approximation

a = 0·319381530 b =−0·356563782 c = 1·781477937
d =−1·821255978 e = 1·330274429 r = 0·2316419

$$t = \frac{1}{1 + rx}$$

$P(x) = 1 - Z(x)(at + bt^2 + ct^3 + dt^4 + et^5)$

Z(x) is ordinate at x, see above.

Error less than $7{\cdot}5\times10^{-8}$. Source: Abramowitz and Stegun, 26.2.18.

10.2.2 Inverse Normal Integral (SD from p)

To derive x_p from $Q(x_p) = p$

a = 2·515517 b = 0·802853 c = 0·010328
d = 1·432788 e = 0·189269 f = 0·001308

$t = \sqrt{\ln(1/Q^2)}$

$$x = t - \frac{a + bt + ct^2}{1 + dt + et^2 + ft^3}$$

Error less than $4{\cdot}5\times10^{-4}$. Source: Abramowitz and Stegun, 26.2.23.

10.2.3 Probability from Student's t

To derive p from t and DF use the BASIC routine from Lee and Lee.

10.2.4 Student's t from Probability and Degrees of Freedom

To derive t from p (or confidence or significance)

If $A(t_{DF}|DF) = 1 - 2p$ and $Q(x_p) = p$ then

$$t_{DF} \simeq x_p + \frac{g_1(x_p)}{DF} + \frac{g_2(x_p)}{(DF)^2} + \frac{g_3(x_p)}{(DF)^3} + \cdots\cdots$$

$$g_1(x) = \frac{1}{4}(x^3 + x)$$

$$g_2(x) = \frac{1}{96}(5x^5 + 16x^3 + 3x)$$

$$g_3(x) = \frac{1}{384}(3x^7 + 19x^5 + 17x^3 - 15x)$$

$$g_4(x) = \frac{1}{92160}(79x^9 + 776x^7 + 1482x^5 - 1920x^3 - 945x)$$

Remember to use Newton's method of evaluating these polynomials and watch the signs of the coefficients. This is probably the slowest routine to run on a computer (and too big for most programmable calculators) not least because it must start with the additional routine to convert p to SD, the x_p term in t_{DF} (see Section 10.2.2).

Source: Abramowitz and Stegun, 26.7.5.

A routine in BASIC is given in Appendix 10A with *exact* values of t for DF = 1 and 2.

10.2.5 Probability from F (variance ratio)

To derive p from F and DF use the BASIC routine given by Lee and Lee.

10.2.6 Filliben's Probability Plotting Position

Filliben plotting positions, as percentage probabilities for ranked data points (rankits).

n = number of points, i = order, i.e. 1st, 2nd, ... ith, ... nth.
% plotting position for nth = $0{\cdot}5^{1/n} \times 100$
1st = 100 − nth
remaining ith = $100 \times (i - 0{\cdot}3175)/(n + 0{\cdot}365)$.

Apart from the arithmetic mean, this is perhaps the easiest statistic to program. Source: Filliben, Reference 1, Chapter 6.

10.2.7 Log Factorial (Log Γ Function)

This approximation, in the log or natural form, is much faster than a FOR-NEXT BASIC loop, except for small n!. It is obvious that some computers will overflow sooner than others if n! is calculated using a loop or the approximation, which is very accurate. The log form of the approximation is used in the program SAMCON (see Chapter 5) and in Fisher's Exact Test (not dealt with in this book). Log n! will not cause overflow problems, but you should be decide whether you are going to use natural or common logarithms. The formula works for either. The form below is for natural logarithms.

$n! = \Gamma(n + 1)$, so by making $z = n + 1$ then $\ln n! = \ln \Gamma(z)$ and

$$\ln \Gamma(z) \simeq (z-\tfrac{1}{2}).\ln z - z + \tfrac{1}{2}\ln(2\pi) + 1/(12z) - 1/(360z^3) + 1/(1260z^5) - 1/(1680z^7)$$

Source: Abramowitz and Stegun, 6.1.41.
A routine in BASIC is given in Appendix 10A for common logarithms.

10.2.8 Relationships Between Arithmetic and Logarithmic Parameters for a Lognormal Distribution

This table is taken from Appendix M, Reference (2), Chapter 5. The conversions are for the *population* and they can lead to significant errors if applied indiscriminately to *sample* data which are only approximately lognormally distributed. In particular it is better to use the MVU estimators for the arithmetic mean and standard deviation than to apply the conversions indicated by *, although their judicious use can often be helpful in obtaining an insight into the distribution.

x is the arithmetic value of a point on the distribution, y is its natural logarithm, ln x (base e), *never* common logarithms in this instance.

μ = true AM of x distribution (*not* of sample)
σ = true ASD of x distribution (*not* of sample)
μ_1 = true AM of y distribution
σ_1 = true ASD of y distribution
GM = Geometric Mean of x distribution
GSD = Geometric Standard Deviation = $\exp(\sigma_1)$

Given	To Obtain	Use
μ_1	*GM*	$\exp(\mu_1)$
μ,σ	*GM*	$\mu^2/\sqrt{(\mu^2 + \sigma^2)}$
σ_1	*GSD*	$\exp(\sigma_1)$
μ,σ	*GSD*	$\exp\sqrt{[\ln(1 + \sigma^2/\mu^2)]}$
μ_1,σ_1	μ *	$\exp(\mu_1 + \frac{1}{2}\sigma_1^2)$
GM,σ_1	μ *	$(GM)\exp(\frac{1}{2}\sigma_1^2)$
μ_1,σ_1	σ *	$\sqrt{[[\exp(2\mu_1 + \sigma_1^2)][\exp(\sigma_1^2) - 1]]}$
GM,σ_1	σ *	$\sqrt{[(GM)^2[\exp(\sigma_1^2)][\exp(\sigma_1^2) - 1]]}$
GM	μ_1	$\ln(GM)$
μ,σ_1	μ_1	$\ln \mu - \frac{1}{2}\sigma_1^2$
μ,σ_1	*GM*	$\exp(\ln \mu - \frac{1}{2}\sigma_1^2)$
GSD	σ_1	$\ln(GSD)$
μ,σ	σ_1	$\sqrt{[\ln(1 + \sigma^2/\mu^2)]}$

* To obtain μ and σ it is better to use the MVU estimators on your data than to assume the sample arithmetic parameters = population AM and ASD.

10.2.9 Method of Generating Normally Distributed Random Data

There is an alternative method due to Marsaglia and Bray (see References 2 and 3, Chapter 2) of generating normal deviates which can be used to simulate normally or lognormally distributed data. This method avoids the conversion from probability to deviate, and in fact generates pairs of normal deviates simultaneously. As with all random number generators the independence of successive random numbers is assumed. Be sure that you do not repeatedly generate the same series - use a RANDOMIZEing function to sow new seeds.

The steps of the method are as follows

1. Generate two random variables, p_1 and p_2, from the uniform distribution $0 \leq p_1, p_2 \leq 1$.
2. Calculate $W_1 = 2.p_1 - 1$ and $W_2 = 2.p_2 - 1$.
3. Calculate $W = W_1^2 + W_2^2$. If $W \geq 1$ return to step 1.
4. Calculate $C = \sqrt{[-2.(\ln W)/W]}$.
5. Calculate $V_1 = C.W_1$ and $V_2 = C.W_2$. V_1 and V_2 are the required normal deviates.
6. Return to step 1 for more pairs of random deviates.

These normal deviates can then be entered into the linearized equation for the normal cumulative curve as in Chapter 2 to create as many pairs of random variates as the experiment calls for.

The "quantization" effect discussed in Chapter 2 will not be so apparent, if at all, when using Marsaglia and Bray's method of generating random normal deviates using a calculator which gives uniformly distributed random numbers to only three places of decimals.

Appendix 10A

The routines in 10A.1 and 10A.2 are written in Sharp PC1500 BASIC. You will be able to judge how easy they might be to translate to your machine's dialect or alternative language. They are taken from OH Programs, 10A.3.

10A.1 Conversion of Significance and DF to Student's t

(From Section 10.2.4). The value of PI (π) is assumed to be resident, otherwise it must be assigned numerically or from PI=ATN(1)*4 (in radians). Confidence = 1 - significance.

```
800 INPUT "INPUT CONFIDENCE LEVEL >50% ";P:IF P<50 THEN 800
801 INPUT "DEGREES OF FREEDOM ";I
810 Q=P/100:Y=(1-Q)/2:GOSUB 1020:REM I=DF
820:
830 PRINT "t =";T:STOP
1010:
1020 REM CONF -> t
1030 IF I=1 THEN T=TAN(Q*PI/2):RETURN:REM t FOR 1 DF
1040 IF I=2 THEN T=Q/SQR(1-Q*Q)*SQR(2):RETURN:REM t FOR 2 DF
1050 REM P->SD
1060 Q=SQR(-2*LOG(Y))
1070 V1=(.010328*Q+.802853)*Q+2.515517
1080 V2=((.001308*Q+.189269)*Q+1.432788)*Q+1
1090 Y=Q-V1/V2
1100 Y=Y*Y:REM START OF SD -> t
1110 V1=(Y+1)/4
1120 V2=((5*Y+16)*Y+3)/96
1130 V3=(((3*Y+19)*Y+17)*Y-15)/384
1140 V4=((((79*Q+776)*Y+1482)*Y-1920)*Y-945)/92160
1150 T=SQR(Y)*((((V4/I+V3)/I+V2)/I+V1)/I+1)
1160 RETURN
```

Exact conversions for DF 1 and 2 need only one line each and appear at 1030 and 1040 after which you RETURN to the program. These two lines are based on series expansions given in Abramowitz and Stegun, 26.7.3 and 26.7.4 and will satisfy all but the most fastidious demands for accurate values of t at the remaining lower degrees of freedom and small α (significance). You will notice that the routine for higher degrees of freedom needs to start with the conversion of p to SD, using the approximation in Section 10.2.2 above.

10A.2 Log Factorial Approximation (From Section 10.2.7)

```
1000 Q=Q+1:REM ! -> GAMMA
1010 IF Q=1 OR Q=2 THEN Q=0:RETURN:REM 0! and 1!
1020 IF Q=3 THEN Q=LOG10(2):RETURN:REM 2!
1030 IF Q=4 THEN Q=LOG10(6):RETURN:REM 3!
1040 IF Q=5 THEN Q=LOG10(24):RETURN:REM 4!
1050 S=(Q-.5)*LOG10(Q):T=(LOG10(2*PI))/2
1060 U=(Q-1/(12*Q)+1/(360*Q↑3)-1/(1260*Q↑5)+1/(1680*Q↑7)*LOG10(EXP(1))
1070 Q=S+T-U:REM Q=log x!, base 10
1080 RETURN
```

The subroutine is entered with Q as the integer for which the log factorial is required and is left with Q = $\log_{10}Q!$, the required value. You will notice the quick exits at 1010-1040. You could add a few more if you wish. Or you could use a loop for summing the logarithms of the first few factorials *before* forming the Γ function at 1000 (IF Q<14 THEN FOR I=2 TO Q, say) and then RETURN from the loop, since the summing loop for logs will be faster than the approximation below some factorial value, log(13!) on one computer. If it matters.

10A.3 Occupational Hygiene Statistics Programs

BOHS, 1983, Technical Guide Series No. 1, "Statistical Analysis of Monitoring Data by Microcomputer", (£2.00), 1983, The BOHS Office, 1 St Andrew's Place, Regent's Park, London NW1 4LB. Presented as a listing with commentary this program accepts OH data and applies Filliben's probability plot test to them. Output is limited.

OH Programs

These programs are in BASIC, written for a Sharp PC1500, with printer/plotter. Some have been translated to BBC BASIC and Mallard® BASIC for the AMSTRAD PCW8256/8512. They can be obtained from H & H Scientific Consultants Ltd. P.O. Box MT 27, Leeds LS17 8QP as listings, with cassettes for the Sharp PC1500, either as ASCII files on Sharp or BBC cassettes for loading into BBC machines, or Sharp ASCII or PCW ready-to-run on 3" CF2 discs for the PCW. Some versions have not been translated.

1 MEANSETC For the calculation of arithmetic and geometric means, and from an input confidence level it outputs the confidence limits. Accepts individual or grouped data input.

2 MVU Demonstration program and in-depth discussion of mvu estimators for the arithmetic mean and SD of lognormally distributed data.

3 RNDLNOR (Sharp only) Plots random lognormal output as a pseudo-strip chart recorder output, with various population and sample statistics.

4 SAMCON (NIOSH Sample size) Relates group number (N), sample size (n), one of sample in top T% with confidence C%. From input of N, n, and T% outputs C%, or from N, T% and C% outputs sample size, or will output table of N and n for input T% and C%.

5 FILPLOT Takes individual data points, ranks them, allocates plotting positions, performs correlation and regression for normal and lognormal distributions, and offers choice of which to work with. Outputs arithmetic and geometric means and SDs (and MVU estimates of arithmetic mean and SD if lognormal), values at input percentiles, correlation coefficients, instructions for plotting on commercial probability paper, 95% confidence limits of mean, 90% confidence limits of highest and lowest values from another sample from the same distribution, and table for plotting the 80% confidence envelope for another sample. Option of graphical output on Sharp and PCW (or Hewlett-Packard plotter), with auto-ranging and capability of changing y-axis range on the screen of the PCW before print-out. Rerun of data for other distribution if required.

6 PAIREDT Paired t-test for differences and ratios.

7 FANDTTEST F- and t-test for difference and ratio of means.

8 RNDTTS Random timetables for 10 minute sampling periods within a shift. Accommodates 24h clock, mid-shift break and night shifts. Any (sane) number of samples can be called for, and as many timetables as required.

Also available

CHICUM, a χ^2 best fit test for cumulative normal or lognormal data.

CHISQU, a χ^2 program including continuity correction and Fisher's Exact Test, for observed or observed and expected data.

ZERCON, a form of FILPLOT which will accept zero data for certain applications.

HISTOGRAM, a demonstration program to give random uniform, normal, or lognormal grouped data for histograms, group sizes based on Sturges' rule, but this can be changed. Can also be run to accept real data which can then be tested using CHICUM.

All programs are internally supported by the necessary statistics – no tables are needed. No filing facilities for data storage or input are included in the programs – these you must work out for yourself, since they will be language and system dependent.

11 SOME EXPLANATIONS

11.1 Background

While it is always risky to make assumptions there comes a time when this must be done. Certain assumptions have been made regarding the extent of your knowledge of statistics. No excuse is offered for presuming that you will be able to find out for yourself any topics in the book where this assumption is invalid. It was not the author's intention to go over the ground that is more than adequately covered in many other books, such as those listed in the bibliography to Chapter 1. The intention of the book is to bend the use of statistics towards occupational hygiene and show that hygienists can, and should, apply statistical tests to their data. However a few topics deserve some explanation, and these are addressed below.

11.2 A Reminder on Logarithms

If you are at home with common and natural logarithms, and in particular with their appearance on calculators and computers, as opposed to tabulated values, you can happily omit this section.

The *logarithm* of a number is the *power* by which its *base* must be raised to give the number. In

$$a = b^x$$

a is the number whose logarithm is required, b is the base, and x is the power, or logarithm, to which the base is raised. An alternative expression which rearranges $a = b^x$ is

$$x = \text{logarithm}_b(a)$$

All numbers >0 have logarithms, but the logarithm of zero is $-\infty$. Negative numbers do not have logarithms. Logarithms of numbers greater than one are positive (>0), the logarithm of one is always zero whatever the base, and those of numbers less than one are negative (<0). Any positive number greater than one can be used as the base. The bases most usually met are 10 (for *common* logarithms) and 2·78182818... (represented by the symbol *e*, for *natural* or *Napierian* logarithms, or even *exp*, where it is followed by a complicated exponent, as in $\exp(\ln \mu - \frac{1}{2}\sigma_1^2)$. Historically the "exp x" form is more correct than "e^x" to the extent that "$\exp_{10}$ x" was considered the "correct" way of expressing 10 raised to the power x.

The phrases "taking logarithms" and "raising to the power of" are these days not uncommonly replaced by the words "logged" and "exponentiated", this usage presumably following the adage that "there ain't no noun that can't be verbed", although hygienists with data loggers attached to various sampling instruments should take care to distinguish between the two uses of "logged".

If you are unsure of *mantissas* and *characteristics* perhaps the following will help to clear up any doubts you may have on the presentation of logarithms on a calculator or computer. The $\log_{10}$ of 2 is given in the log tables as 0·3010, and of 0·2 as $\bar{1}$·3010. The $\bar{1}$ is the characteristic, or

exponent of 10, and the ·3010 is the mantissa. Algebraically the logarithm is equivalent to $-1 + ·3010 = -0·6990$, which is what a calculator or computer will output for $\log_{10}$ of 0·2. For base e logarithms the characteristic for numbers less than 1 (but greater than 0) are again negative, but again the algebra is the same and the calculator gives the "right" answer without mixing negative (characteristic) and positive (mantissa) values, so that $\log_e$ 0·2 is −1·6094 from a calculator, rather than $\bar{3}·6974 + 0·6932 = \bar{2}·3906$ from tables, where $\bar{3}·6974 = \log_e 1/10$ and $0·6932 = \log_e 2$.

These various methods are used to get the arithmetic and geometric means discussed above from the following data.

0·03, 0·4, 0·7, 1·2

	Arithmetic	$\log_{10}$ (table)	$\log_{10}$ (calculator)	$\log_e$ (calc)
	0·03	$\bar{2}·4771$	−1·5229	−3·5066
	0·40	$\bar{1}·6021$	−0·3979	−0·9163
	0·70	$\bar{1}·8451$	−0·1549	−0·3569
	1·20	0·0792	0·0792	0·1823
Totals	2·33	$\bar{4}$ (characteristics)	−1·9965	−4·5975
		+2·0035 (mantissas)		
		=$\bar{2}$·0035 (but this is no help in the next line!)		
Divide by n = 4 to get means				
	2·33/4	(−4+2·0035)/4	−1·9965/4	−4·5975/4
		$\bar{1}$·5009	−0·4991*	−1·1494*
Means=	0·5825 (AM)	=0·3169 (GM)	=0·3169 (GM)	=0·3168 (GM)

Note: * are the means of the logarithms.

The final value (geometric mean) in the second column comes from the table of antilogs, and the last two GMs from the inverse log functions on the calculator, 10^x and e^x, or in computer parlance 10↑x and EXP(x). The two means (arithmetic and geometric) and the arithmetic standard deviation have the same dimensions as the original data. So that if the data are in mg/m³, then so are these statistics. It is an error to express the geometric standard deviation in the same dimensions – it is in fact dimensionless, best thought of as a ratio.

Logarithms were commonly used to simplify multiplication and division before calculators and computers became widely available. The logarithms of the numbers to be multiplied or divided are read from a table, added or subtracted respectively and the resultant logarithm converted back to a number using a table of antilogarithms.

Calculators have the logarithmic functions built in. The natural logarithm is often shown as "ln" and (using the "F", "upper case" or second function key) the inverse function (antilogarithm), or e^x, can be regained. A similar double function key "log" and "10^x" serves for the base 10. On computers which you might use, the BASIC logarithmic functions may be LN (for ln or $\log_e$) and LOG (for $\log_{10}$) or perhaps LOG (for ln or $\log_e$) and LOG10 (for $\log_{10}$). Don't blame me for the confusion! The antilog functions are invariably EXP (for antilogs of natural logarithms) and 10↑ or 10^ (for base 10 logs).

As an aside if you have a y^x key on your calculator you can confirm that $2·5^{3·2} = 18·7675...$ Here 3·2 is the logarithm of 18·7675 to the base 2·5. As a check use the same key to see if $10^{0·30103}$ comes very close to 2.

Natural logarithms are invariably used in pure mathematics and the mathematical side of statistics, as in the conversions in Chapter 10. Base 10 or base e logarithms can safely be used in all numerical operations (followed

by a final exponentiation) discussed in this book. The most likely place common (base 10) logarithms will be encountered in occupational hygiene is in sound where the units are decibels. In this case you *must* use common logarithms (or you will be in deep trouble!).

Because of the ease of manipulating data using calculators and computers it is obviously easier to let whatever calculating aid you may be using do the worrying about the logarithms and sign of their mantissas and characteristics.

11.3 Minimum Variance Unbiased Estimators

There are a number of ways one might suggest for estimating the "average" or mid-point value of a set of data. For ease of calculation (apart from the tedium of finding them) one could divide the sum of the extreme values by 2. Experience (and probably mathematical analysis) however shows that this gives very *variable* results when the samples are repeated – the *variance* is large and far from a *minimum*. Other methods might be proposed which could be *biased*, either high or low relative to the real value. All these methods would, in general be aimed at the *estimation* of the mean of the population from which the experimental samples were drawn. The *Minimum Variance Unbiased Estimator* of the arithmetic mean of a normally distributed population is in fact the arithmetic mean of the sample.

Similarly when we want to describe the spread of results we should choose a measure which is unbiased and, if repeated many times, is itself of minimum variance. So the range of the small sample is a poor estimator of the population from which the sample is drawn (although some statistics still rely on it). The *sample standard deviation* (see Section 3.3) is the best (least biased with the least variance) estimator of the spread of the population we can have. If we have the data for the *whole* (large) population then the n – 1 in the denominator of [3.2a-c] is nearly the same as the denominator n mentioned for the population variance in Section 3.2. The truth of the use of n – 1 (Bessel's correction) or n in calculating the standard deviation is better described in other text books (e.g. Kennedy and Neville, Appendix A, see bibliography Chapter 1).

There are other MVU estimators for other parameters used in statistics. The discussion in Section 3.5 considers the fact that the arithmetic mean and standard deviation of the sample values are not the MVU estimators for a lognormally distributed population, and that the MVU estimators are quite easy to calculate with modern calculating aids.

11.4 Histograms and "Sturges' Rule"

It has been pointed out that collecting the data into groups in order to construct histograms causes degradation of the data, since only two values, the group limits, are used to describe all the data in the group. If the cell size or ratio is imprudently selected much valuable information may be irretrievably lost, and this loss of information may conceal interesting and important features, especially at the ends of the distribution.

Before the days of computers it was convenient to collect data into such groups to reduce the amount of calculation to be performed. This is no longer a valid argument for using histograms in statistical analysis.

There are occasions, particularly when presenting results, when histograms may be acceptable, or even desirable, and "Sturges' Rule" can be

used to determine the "best" number of groups into which data should be collected for the construction of histograms. Its practical value is doubtful except as a guide, and is irrelevant if the data are to be plotted individually as described in Chapters 6 and 7. Sturges' Rule defines both the number of groups and the group size. The Rule is generally given as

$$\text{Number of cells} = 1 + 3{\cdot}322 \times \log_{10}N \qquad [11.1a]$$

N being the sample size, or in some similar form, depending on the number of decimal places and the logarithmic base which the user may choose. The general form is

$$\text{Number of cells} = 1 + \ln N/\ln 2 \qquad [11.1b]$$

which is suitable for use on calculators and computers, as in the program "HISTOGRAMS"[1]. The cell size can then be found from

$$\text{Cell size} = \text{Range/Number of cells}$$
$$\text{Cell size} = (x_{max} - x_{min})/\text{Number of cells} \qquad [11.2a]$$

The cell size will rarely be in exact and convenient units, and can be adjusted within reason. These simple formulae [11.1b] and [11.2a] are clearly for normally distributed data, giving equal cell sizes of, say, 10 with boundaries at 10, 20, 30, 40 and so on. If the data are lognormally distributed [11.1a and b] are still valid but [11.2a] becomes

$$\text{Cell } \textit{ratio} = \text{antilog}((\log x_{max} - \log x_{min})/\text{Cell Number}) \qquad [11.2b]$$
$$= \text{antilog}[(\log (x_{max}/x_{min}))/\text{Cell Number}] \qquad [11.2c]$$

Notice that the *cell size* has become the *cell ratio*. Again the cell ratio will probably not be ideal, that is, a convenient number like $\sqrt{2}$ to give cell boundaries at 0·5, 0·7, 1, 1·414, 2, ... Also notice the need to know whether the data are normally or lognormally distributed *before* the table of frequencies of occurrence for the histogram is drawn up. Or put another way, ideally you should construct two histograms, using [11.1b] to find the number of cells and also [11.2a] *and* [11.2c] to get the cell intervals in arithmetic and logarithmic terms, and then test both histograms for goodness of fit (to the respective distributions) before deciding on which is the more appropriate. Two tests are commonly used, the χ^2 test of goodness of fit and one of the various versions of the Kolmogorov-Smirnov test, but neither is described here, since the routines described in Chapters 6 and 7 are superior. Remember that your data need to be entered into a computer only once, but once entered you should apply a statistic which makes optimum use of them.

11.5 NIOSH

NIOSH is not a small town in Wisconsin. It is the National Institute of Occupational Safety and Health, an American organization based in Cincinnati, Ohio. One of its responsibilities is "to insure safe and healthful working environments." It carries out research and development programmes and disseminates the results. NIOSH is a rich source of information, including statistical procedures, for occupational hygienists. Some publications from NIOSH can be obtained direct free from NIOSH at Robert A. Taft Laboratories, 4676 Columbia Parkway, Cincinnati, OH 45226, but stocks held are usually small. Two alternative sources of NIOSH publications are:-

1) Superintendent of Documents, US Government Printing Office, (US GPO), Washington, D.C. 20402. The US GPO requires prepayment or, better, an account maintained in credit. The GPO takes Visa and Mastercharge (Access) cards.

2) National Technical Information Service (NTIS), U.S. Dept. of Commerce, Springfield, Va. 22161. Again suitable payment with order, an account or credit cards as above plus American Express can be used for payment.

NTIS have agents in many countries. The UK agent is Microinfo Ltd, PO Box 3, Alton, Hants., GU34 2PG.

The British Library, Lending Division, Boston Spa, also carry NIOSH books.

11.6 Null Hypothesis

Many statistical tests, such as the F-test and t-test, used in Chapters 8 and 9 are based on the "null hypothesis". In using this hypothesis it is always postulated that there is *no significant difference* between the distributions being compared. The probability of the *actual* difference occurring due to chance alone is calculated, and if this probability is small, the null hypothesis is rejected, and it is inferred that the difference is real. Thus the null hypothesis for the F-test postulates that there is no difference between the variances, i.e. the *variance ratio* = 1. The probability of this occurring by chance can be read from the appropriate statistical tables, or computed. If F is close to 1 then the probability of this occurring by chance is high, and the hypothesis (that the variances are such that the two samples could have come from populations with the same variance) is accepted.

For the t-test the null hypothesis might be that the mean of the sample is not different from the mean of some population mean. The statistic t is calculated from the data. Again the probability of this value of t occurring is looked up in the t-distribution table or it is calculated, and if the the probability is high enough, the hypothesis (that there is no significant difference between the means) is accepted. If t is large and p is small, the null hypothesis is rejected (the means are significantly different), but it should be noted that the null hypothesis can never be formally proved to be correct.

Commonly, but not always, the t-test is used to test the hypothesis that the difference between means is zero. But it follows from the fact that the test can be used to test the mean from the data with some other population (or sample) mean, these means need not be the same. In this case the null hypothesis is that the difference between the means is some value other than zero. This could have been the basis of the t-test to be applied to the means of the logs of the samplers in Section 8.5 when discussing the case of both being used as personal samplers. With *a priori* evidence the null hypothesis that the ratio of geometric means was not significantly different from 1·30 could have been proposed and tested, or in the more directly and computationally convenient form of the program that the difference of the (natural) logarithmic means was not significantly different from 0·2624.

11.7 Reference

1 "HISTOGRAMS", a demonstration program in Basic which generates random uniform, normal or lognormal data, using Sturges' rule to set cell sizes and cell intervals before collecting the data into frequency of occurrence cells. The program can also be used for "real" data. From H & H Scientific Consultants Ltd, P.O. Box MT27, Leeds, LS17 8QP.

12 OCCUPATIONAL HYGIENE STATISTICS GLOSSARY

12.1 Difficult Words in Occupational Hygiene Statistics

Some of the expressions in the book may be somewhat novel, and this glossary has been included to clear up any misunderstandings you might have. It will also cater for the terminology which you will encounter in the second and subsequent books (as yet unwritten) in this series. The Glossary is taken from a fuller document[1].

Data Set : Numerical results from two or more unrelated measurements made at the workplace.

Distribution : The arrangement of a *data set* in ascending order of magnitude, usually done because no one can think of anything better to do with it.

Lognormal Distribution : Statistical *distribution* universally ascribed to any distribution of occupational hygiene data, irrespective of the fact that it can never be shown statistically that the results are lognormally distributed at any sane level of confidence (see[2]). Any high GSDs are best used sparingly to panic management and ensure employment for another year.

Statistically Insignificant : Describes any outcome of human exposure to harmful agents, up to, and including, loss of taste, hearing, sight, sense of smell or any parenchyma (after[3]).

Statistically Significant : Describes the outcome a study was designed to produce[3].

Statistical Artefact : *Data set* showing that all the remedial measures implemented to improve working conditions have in fact made them worse (after[3]).

Null Hypothesis : (1) The presumption that hygienists don't need to apply statistics to their data.

(2) The proposition that hygienists haven't sufficient data for statistical analysis.

(3) Conclusion you do not want to reach, but constantly try to attain[3].

Type I Error : Application of a statistical test without having properly formulated the *Null Hypothesis*[3].

Type II Error : Believing you understand *Type I Error*[3].

Homogeneous Group : Arbitrary or *biased* (any old) collection of *random* individuals.

Random : Biased.

Biased : Distorted—usually arises from neglecting a *factor*.

Factor : Useful to apply to data, like poultices to boils. (Makes them better if they are poorly, unless you pick at them).

Correlate : As in "correlates with"; suspect either a tautology or a *non sequitur* when you see this in a report.

Trend Analysis : Immortalized by one-time US Defense Secretary Harold Brown, when he commented on the failure of two submarine test firings of the Tomahawk cruise missile "Failure in the past increases the probability of success in the future".

12.2 References

1 Clari Fie and Ed U. Kate, (unpublished work – rejected by every Health & Safety publication) "A Glossary of Occupational Hygiene Terms".

2 Mage, D.T., (1985) "The Procrustean Fit – A Useful Statistical Tool for Decision Making", *Journal of Irreproducible Results*, v.30, no.4, p. 32.

3 Iles, R.L., (1984) "A Dictionary of Pharmaceutical Research, Comments and Excerpts", *JIR*, v.29, no.3, pp. 14-15.

Occupational Hygiene Monograph Series: ISSN 0141-7568.
Editor: Dr D.Hughes, University of Leeds.

No.16, Phosphorus-32: Practical Radiation Protection, ISBN 0-905927-67-2, P.E.Ballance, L.R.Day and J.Morgan, AURPO, 1987.

No.15, A Guide to Radiation Protection in the Use of X-Ray Optics Equipment, ISBN 0-905927-52-4, Health and Safety Executive Working Group, 1986.

No.14, The Control of Microorganisms responsible for Legionnaires' Disease and Humidifier Fever, ISBN 0-905927-22-2, B.P.Ager and J.A.Tickner, Health and Safety Executive, 1985.

No.13, Health Physics Aspects of Radioiodines, ISBN 0-905927-76-1, D.Prime, AURPO, 1985.

No.12, Allergy to Chemicals and Organic Substances in the Workplace, ISBN 0-905927-51-6, G.W.Cambridge and B.F.J.Goodwin, Unilever,1984.

No.11, The Disposal of Hazardous Wastes, ISBN 0-905927-26-5, G.E.Chivers, University of Sheffield. **Out of print.**

No.10, Education and Training in Occupational Hygiene, ISBN 0-905927-21-4, Peter J.Hewitt, University of Bradford, 1983.

No.9, The Performance, Installation, Testing and Limitations of Microbiological Safety Cabinets, ISBN 0-905927-16-8, R.P.Clark, M.R.C., 1983, reprint 1989.

No.8, A Guide to the Safe Use of X-ray Diffraction and Spectrometry Equipment, ISBN 0-905927-11-7, E.B.M. Martin, AURPO, 1983.

No.7, Adventitious X Radiation from High Voltage Equipment: Hazards and Precautions, ISBN 0-905927-90-7, E.B.M.Martin, AURPO, 1982. **Out of print.**

No.6, Health Physics Aspects of the Use of Tritium, ISBN 0-905927-85-0, E.B.M.Martin, AURPO, 1982, reprinted 1984.

No.5, Notes on Ionizing Radiations: Quantities, Units, Biological Effects and Permissible Doses, ISBN 0-905927-80-X, D.Hughes, University of Leeds, 1982, reprinted 1983.

No.4, A Literature Survey and Design Study of Fumecupboards and Fume-Dispersal Systems, ISBN 0-905927-50-8, D.Hughes, 1980, reprinted 1987.

No.3, The Toxicity of Ozone, ISBN 0-905927-30-3, D.Hughes, 1979.

No.2, Electrical Safety-Interlock Systems, ISBN 0-905927-45-1, D.Hughes, 1978, reprinted with additions 1985.

No.1, Hazards of Occupational Exposure to Ultraviolet Radiation, ISBN 0-905927-15-X, D. Hughes, University of Leeds, 1978, reprinted 1982. **Out of print.**

HHSC

H & H Scientific Consultants Ltd
in association with **Science Reviews Ltd**
P.O.Box MT27, Leeds LS17 8QP, U.K.; tel.: 0532 687189.

HHSC Handbook Series:
Editor: Dr D.Hughes, University of Leeds.

HHSC Handbook No. 4, Discharging to Atmosphere from Laboratory-Scale Processes, ISBN 0-948237-03-1, Spring 1989, by Dr D.Hughes, University of Leeds..

HHSC Handbook No. 3, The BURL Guide to the Control of Substances Hazardous to Health Regulations, ISBN 0-948237-02-3, 1989, by Peter J.Hewitt, University of Bradford.

HHSC Handbook No. 2, Fumecupboards Revisited, ISBN 0-948237-01-5, 1986, by Dr J.D.Cook, Environmental Safety Group, Harwell Laboratory, and Dr D.Hughes, University of Leeds.

HHSC Handbook No. 1, The Radman Guide to the Ionising Radiations Regulations 1985, ISBN 0-948237-00-7, reprinted 1986, by Radman Associates.

Medical Research Council Video Programs (airflow visualization by Schlieren photography):
Biological Containment (Microbiological Safety Cabinets);
Laminar Downflow Cabinets; **The Open-Bench Environment;**
Surgeon's Clothing; **The Human Micro-Environment.**

Video Training Program: Protection from Ionising Radiation. Five modules: Scientific Background, Units, Principal Hazards, Control of External Radiation, Measuring External Radiation, plus Notes; Commission of the European Communities (Health and Safety Directorate) and University of Sheffield.

B.O.H.S. Technical Guide Series: ISSN 0266-6936.
Editors: Dr D.Hughes, University of Leeds, and Dr T.L.Ogden, Health and Safety Executive.

T.G. No. 7, Controlling Airborne Contaminants in the Workplace, ISBN 0-905927-42-7, M.Piney, R.J.Alesbury, B.Fletcher, J.Folwell, F.S.Gill, G.L.Lee, R.J.Sherwood, J.A.Tickner, 1987, reprinted 1988.

T.G. No. 6, The Sampling and Analysis of Compressed Air to be used for Breathing Purposes, ISBN 0-905927-17-6, G.L.Lee, D.Coker, R.N.J.Barraclough, M.J.Gorham, A.J.Bloom, K.R.Haigh,1985.

T.G. No. 5, The Selection and Use of Personal Sampling Pumps, ISBN 0-905927-86-9, R.M.Wagg, D.T.Coker, J.R.P.Clarke, G.L.Lee, P.Leinster, B.Miller, M.Piney, 1985.

T.G. No. 4, Dustiness Estimation Methods for Dry Materials: Their Uses and Standardization; and **The Dustiness Estimation of Dry Products: Towards a Standard Method,** ISBN 0-905927-71-0, C.M.Hammond, N.R.Heriot, R.W.Higman, A.M.Spivey, J.H.Vincent, A.B.Wells, 1985.

T.G. No. 3, Fugitive Emissions of Vapours from Process Plant Equipment, ISBN 0-905927-66-4, A.L.Jones, M.Devine, P.R.Janes, D.Oakes, N.J.Western, 1984.

continued overleaf