

REPORT

Distribution of correlation coefficient for samples taken from a bivariate normal distribution

M J Salter, N M Ridler and M G Cox

September 2000

Distribution of correlation coefficient for samples taken from a bivariate normal distribution

M J Salter[†], N M Ridler[†] and M G Cox[‡]

[†]Centre for Electromagnetic and Time Metrology

[‡]Centre for Mathematics and Scientific Computing

ABSTRACT

When measuring a complex S -parameter several repeat measurements are made. The best estimate of the S -parameter and an elliptical uncertainty region are calculated from statistics of the repeat measurements. In particular the correlation coefficient between the real and imaginary parts $r(x,y)$ is used in calculating the uncertainty region. The set of repeat measurements can be considered as a sample from a bivariate normal distribution. In this report the distribution of correlation coefficient $r(x,y)$ calculated for samples from a bivariate normal distribution is investigated by generating a large number of samples using the multivariate normal distribution simulator MULTNORM. The effect of the population correlation coefficient $\rho(x,y)$ and of the sample size n on the distribution is examined. For small samples it is found that the distributions are non-normal, broad and sometimes skew. This has implications for the reliability of confidence regions arrived at based on a small number of repeat measurements. The distribution of Fisher's z (a statistic defined in terms of r) is also investigated and is found to be more normal than the distribution of r . This statistic is useful for estimating 95% confidence intervals for ρ .

©Crown Copyright 2000
Reproduced by permission of the Controller HMSO

ISSN 1467-3932

National Physical Laboratory
Teddington, Middlesex, UK TW11 0LW

Extracts from this report may be reproduced provided the source
is acknowledged and the extract is not taken out of context

Approved on behalf of the Director, NPL, by Dr S Pollitt
Head of the Centre for Electromagnetic and Time Metrology

CONTENTS

1 INTRODUCTION.

2 MEAN, STANDARD DEVIATION AND RANGE OF SAMPLE CORRELATION
COEFFICIENT AS A FUNCTION OF SAMPLE SIZE AND POPULATION
CORRELATION COEFFICIENT 2

3 DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENT FOR DIFFERENT
SAMPLE SIZES AND POPULATION CORRELATION COEFFICIENTS 2

4 DISTRIBUTION OF FISHER’S Z FOR DIFFERENT SAMPLE SIZES AND
POPULATION CORRELATION COEFFICIENTS 3

5 CONCLUSION

6 REFERENCES.....6

APPENDIX GLOSSARY OF STATISTICAL SYMBOLS USED 7

APPENDIX 2 THE CORRELATION COEFFICIENT FOR A SAMPLE OF SIZE 2 IS $\pm 1/10$

APPENDIX 3 A PROCEDURE FOR SAMPLING FROM A BIVARIATE NORMAL
DISTRIBUTION.....

Figures 1 to 39 appear after Appendix 3 on pages 24 to 43

This page is left blank

1 INTRODUCTION

The results of independent measurements of an S -parameter of a device under test (DUT) can be modelled as a complex valued (or 2-dimensional vector valued) random variable associated with some bivariate normal distribution. Here the term “independent” is taken to mean that the DUT is disconnected and reconnected and possibly the VNA (vector network analyser) is recalibrated between each pair of measurements. Thus a set of n independent repeat measurements is regarded as a random sample of size n from a bivariate normal distribution. The parameters of the bivariate distribution (the complex mean $(\mu_x + j\mu_y)$, the variances in real and imaginary parts $(\sigma^2(x)$ and $\sigma^2(y))$ and the correlation coefficient between real and imaginary parts $(\rho(x,y))$) are determined by the DUT, the measurement system and the engineer performing the measurements. The symbols used for the parameters of the distribution and the corresponding statistics of the sample which are used to estimate the population parameters are listed in Appendix 1.

The complex mean $(\bar{x} + j\bar{y})$ of the sample is taken as the best estimate of the S -parameter and an elliptical confidence region is constructed from the variances in the mean of the real and imaginary parts $(s^2(\bar{x})$ and $s^2(\bar{y}))$ and the correlation coefficient between real and imaginary parts $(r(x,y))$ ¹. If further samples of size n were taken then the values of the sample statistics would, in general, vary from sample to sample. In particular the best estimate of an S -parameter and the size and orientation of the elliptical confidence region would vary from sample to sample. The distributions of the statistics can be investigated by generating a very large number of samples and calculating the statistics for each sample.

The multivariate normal distribution simulator MULTNORM (see Appendix 3) can be used to generate a random sample of size n taken from a bivariate normal distribution. The bivariate normal distribution is specified by means of the following parameters: (i) the number of random variables (for a bivariate distribution this is equal to 2), (ii) the complex mean $(\mu_x + j\mu_y)$, (iii) the variances $(\sigma^2(x)$ and $\sigma^2(y))$ and (iv) the correlation coefficient $(\rho(x,y))$. This report describes an investigation using MULTNORM into the distribution of sample correlation coefficient $r(x,y)$. The effect on the distribution of the population correlation coefficient $\rho(x,y)$ and of the sample size n (keeping the population mean $(\mu_x + j\mu_y)$ and variances $(\sigma^2(x)$ and $\sigma^2(y))$ fixed) is examined. Small and medium sized samples are of particular interest because of practical limitations on the number of repeat measurements that can be carried out.

As an alternative to the Monte Carlo simulation technique described here, an analytical treatment of the distribution of the sample correlation coefficient was given by Fisher[#] in 1915² as described by Kendall and Stuart³. A fairly complicated formula for the probability density function is given which depends on the population correlation coefficient and on the sample size.

[#] Sir Ronald A. Fisher (1890-1962) was statistician at Rothamstead Experimental Station from 1913-33 and thereafter professor at University College London.

2 MEAN, STANDARD DEVIATION AND RANGE OF SAMPLE CORRELATION COEFFICIENT AS A FUNCTION OF SAMPLE SIZE AND POPULATION CORRELATION COEFFICIENT

MULTNORM was used to generate samples of size n from a bivariate normal distribution (with mean $(\mu_x, \mu_y) = (0,0)$, variance $(\sigma^2(x), \sigma^2(y)) = (1,1)$ and specified correlation coefficient $\rho(x,y)$) for all integer values of n from 2 to 100. For each value of n , a large number of samples (40,000) of that size were generated and the correlation coefficient $r(x,y)$ of each sample was calculated. The mean, standard deviation, maximum and minimum of the 40,000 correlation coefficients are shown plotted against n in Figs 1 to 3 for three different bivariate normal distributions. For all three distributions the standard deviation and the difference between maximum and minimum get smaller as the sample size increases. Thus as the sample size increases the spread in values of correlation coefficient from one sample to another decreases. For Fig 1 (population correlation coefficient = 0) the mean of the 40,000 correlation coefficients is close to zero for all sample sizes. For Figs 2 and 3 (population correlation coefficient = 0.5 and 0.9 respectively) the mean correlation coefficient is markedly different from the population value for small sample sizes (less than 10 say). Thus, in these cases, the sample correlation coefficient is not an unbiased estimator of the population correlation coefficient for small sample sizes. For samples of size 2 the correlation coefficient is always either +1 or -1 as shown in appendix 2.

3 DISTRIBUTION OF SAMPLE CORRELATION COEFFICIENT FOR DIFFERENT SAMPLE SIZES AND POPULATION CORRELATION COEFFICIENTS

A large number of samples (one million) of each of the following sizes 3, 4, 5, 7, 10, 100 were generated from each of the three bivariate normal distributions discussed above (namely those with mean $(\mu_x, \mu_y) = (0,0)$, variance $(\sigma^2(x), \sigma^2(y)) = (1,1)$ and correlation coefficients $\rho(x,y)$ equal to 0, 0.5 and 0.9). The distribution of these one million values are plotted in the even numbered Figures from 4 to 39 for the different values of population correlation coefficient and sample size.

For the case of population correlation coefficient = 0, all the distributions are symmetrical about 0. For small sample sizes the distributions are clearly non-normal. As already mentioned for a sample of size 2 the only values of sample correlation coefficient obtained are ± 1 and so in this case the distribution is an extreme example of a U-shaped distribution. When the sample size is three the distribution is U-shaped with peaks at ± 1 . For a sample size of four the distribution is uniform. By the time the sample size has increased to 100 the distribution looks fairly normal. For the case of population correlation coefficient = 0.5 or 0.9, the distributions for small sample size are skew being more skew for the larger value of population correlation coefficient. As before the distribution for a sample size of 3 is peaked at ± 1 and the distributions become more normal as the sample size is increased.

For each combination of population correlation coefficient and sample size, the mean, standard deviation and 95% confidence interval of the distribution of one million sample correlation coefficients were calculated. These are listed in Table 1. The 2.5 and 97.5 percentiles were calculated by sorting the one million values into ascending order and taking the 25,001st and 974,999th values. The interval between these two percentiles encompasses 95% of the values and so is referred to as a 95% confidence interval. For population correlation coefficients of 0.5 and 0.9 it can be seen from Table 1 that the mean sample

correlation coefficient departs from the population value for small samples. As previously mentioned this is also evident from Figs 2 and 3.

Table 1 Properties of the distribution of sample correlation coefficient

Correlation coefficient (ρ)	Sample size (n)	Mean	Standard deviation	95% confidence interval
0	3	1.566E-5	0.707	(-0.997, 0.997)
0	4	-2.455E-5	0.577	(-0.949, 0.950)
0	5	-3.444E-4	0.500	(-0.879, 0.878)
0	7	-1.443E-6	0.408	(-0.755, 0.754)
0	10	-7.328E-5	0.333	(-0.631, 0.632)
0	100	4.806E-5	0.101	(-0.197, 0.196)
0.5	3	0.406	0.635	(-0.980, 0.999)
0.5	4	0.436	0.502	(-0.802, 0.987)
0.5	5	0.452	0.424	(-0.609, 0.964)
0.5	7	0.468	0.336	(-0.353, 0.918)
0.5	10	0.478	0.267	(-0.157, 0.867)
0.5	100	0.498	0.076	(0.339, 0.636)
0.9	3	0.821	0.363	(-0.542, 1.000)
0.9	4	0.854	0.236	(0.134, 0.998)
0.9	5	0.869	0.175	(0.375, 0.995)
0.9	7	0.881	0.118	(0.563, 0.987)
0.9	10	0.889	0.083	(0.670, 0.979)
0.9	100	0.899	0.020	(0.856, 0.932)

4 DISTRIBUTION OF FISHER’S Z FOR DIFFERENT SAMPLE SIZES AND POPULATION CORRELATION COEFFICIENTS

In 1921 Fisher^{3, 4} introduced a new statistic *z* which is a function of the sample correlation coefficient *r* and is defined by

$$z = \frac{1}{2} \log_e \frac{1+r}{1-r}$$

It follows that

$$r = \tanh z$$

This statistic tends to normality much faster than *r* as the sample size (*n*) is increased. The distribution of *z* is given by the following approximate rule:

For random samples of size *n* from a bivariate normal distribution with population correlation coefficient ρ the distribution of *z* is a normal distribution with mean

$$\xi = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho}$$

and with variance

$$\frac{1}{n-3}$$

Note that the variance is independent of the population correlation coefficient (ρ).

The distributions of Fisher's z corresponding to the distributions of sample correlation coefficient (r) discussed above (namely $n=3, 4, 5, 7, 10, 100$ for each of $\rho=0, 0.5, 0.9$) are plotted in the odd numbered Figures from 4 to 39. The means and standard deviations of each of these distributions are listed in Table 2 where they are compared with the values predicted by the approximate rule.

Table 2 Properties of the distribution of Fisher's z

Correlation coefficient (ρ)	Sample size (n)	Mean	$\frac{1}{2} \log \frac{1+\rho}{1-\rho}$	Standard deviation	$\sqrt{\frac{1}{n-3}}$
0	3	-1.060E-4	0.000	1.569	-
0	4	3.587E-4		0.906	1.000
0	5	-6.165E-4		0.684	0.707
0	7	-5.977E-5		0.495	0.500
0	10	-4.793E-5		0.377	0.378
0	100	4.703E-5		0.102	0.102
0.5	3	0.844	0.549	1.532	-
0.5	4	0.689		0.890	1.000
0.5	5	0.640		0.674	0.707
0.5	7	0.602		0.489	0.500
0.5	10	0.581		0.374	0.378
0.5	100	0.552		0.102	0.102
0.9	3	2.053	1.472	1.394	-
0.9	4	1.739		0.834	1.000
0.9	5	1.644		0.643	0.707
0.9	7	1.570		0.476	0.500
0.9	10	1.531		0.367	0.378
0.9	100	1.477		0.101	0.102

By inspection, the distributions of z are all fairly symmetrical even in cases where the corresponding distribution of r is strongly skewed. However to determine the extent to which the distributions are normal one would have to investigate more closely their peaks and tails. It can be seen from Table 2 that, for a given ρ , the estimates of the mean and the standard deviation given by the approximate rule improve with the sample size and that useful estimates can be obtained even for quite small samples (larger than 5 say).

The fact that the distribution of z is approximately normal allows the standard theory of the normal distribution to be applied to determine a 95% confidence interval for ξ . This interval can then be transformed to give a 95% confidence interval for ρ .

Example

As an example, suppose that six repeat measurements of a complex S -parameter are made and a value of 0.6 is calculated for the correlation coefficient between the real and imaginary parts. This corresponds to a random sample of size $n = 6$ from a bivariate normal distribution with a sample correlation coefficient $r = 0.6$. The corresponding value of Fisher's z is

$$z = \frac{1}{2} \log_e \frac{1+0.6}{1-0.6} = 0.69$$

Assume that z is distributed according to the approximate simple rule given above (normally distributed with mean $\xi = \frac{1}{2} \log_e \frac{1+\rho}{1-\rho}$ and variance $\frac{1}{n-3}$). The probability that a normal random variable is within 1.96 standard deviations of the mean is 0.95 and so

$$p\left(\xi - \frac{1.96}{\sqrt{n-3}} \leq z \leq \xi + \frac{1.96}{\sqrt{n-3}}\right) = 0.95$$

where $p()$ represents the probability that the given condition holds. It follows that

$$p\left(z - \frac{1.96}{\sqrt{n-3}} \leq \xi \leq z + \frac{1.96}{\sqrt{n-3}}\right) = 0.95$$

Thus $\left(z - \frac{1.96}{\sqrt{n-3}}, z + \frac{1.96}{\sqrt{n-3}}\right)$ is a 95% confidence interval for ξ . Carrying out the calculations gives

$$\xi_+ = z + \frac{1.96}{\sqrt{n-3}} = 1.82$$

$$\xi_- = z - \frac{1.96}{\sqrt{n-3}} = -0.43$$

Performing the inverse transformation gives a 95% confidence interval for ρ (ρ_+, ρ_-) where

$$\rho_+ = \tanh(r_+) = 0.95$$

$$\rho_- = \tanh(r_-) = -0.41$$

Hence $\rho = 0.6 + 0.35, -1.01$

5 CONCLUSION

The above discussion suggests the following. For small sample sizes (less than 10 say) the distribution of sample correlation coefficient from a bivariate normal distribution is non-normal with a wide 95% confidence interval. If the parent population has a non-zero correlation coefficient then the small sample distribution is skew and has a mean value which departs from the correlation coefficient of the parent population. In a practical measurement of an S -parameter a single sample is taken and the correlation coefficient (together with the variances in the real and imaginary parts) is used to determine the elliptical uncertainty region. This suggests that the elliptical uncertainty region based on a small number of repeat measurements is subject to a considerable uncertainty. Fisher's z provides a method of estimating a 95% confidence interval for the population correlation coefficient from a single

sample and from this an indication of the reliability of the orientation of the uncertainty ellipse.

6 REFERENCES

- 1 Paul Young, What is Correlation ?, ANAMET News, Issue 12, Spring 1999
- 2 R. A. Fisher, Frequency-distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, 10, 507 (1915)
- 3 Alan Stuart and J. Keith Ord, Kendall's Advanced Theory of Statistics, Volume Distribution Theory, 5th ed (1987), chapter 16
- 4 R. A. Fisher, On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1, No. 4, 1 (1921)

APPENDIX 1 GLOSSARY OF STATISTICAL SYMBOLS USED

Population Parameters

Let $P(x,y)$ be the bivariate probability density function of the two random variables x and y which represent the real and imaginary parts of the complex random variable $x + jy$.

- Mean (expectation) of real part

$$\mu_x = E(x) = \iint xP(x,y)dxdy$$

- Mean (expectation) of imaginary part

$$\mu_y = E(y) = \iint yP(x,y)dxdy$$

- Variance of real part:

$$\begin{aligned}\sigma^2(x) &= E((x - \mu_x)^2) \\ &= \iint (x - \mu_x)^2 P(x,y)dxdy\end{aligned}$$

- Variance of imaginary part:

$$\begin{aligned}\sigma^2(y) &= E((y - \mu_y)^2) \\ &= \iint (y - \mu_y)^2 P(x,y)dxdy\end{aligned}$$

- Standard deviation of real part

$$\sigma(x) = \sqrt{\sigma^2(x)}$$

- Standard deviation of imaginary part

$$\sigma(y) = \sqrt{\sigma^2(y)}$$

- Covariance between real and imaginary parts:

$$\begin{aligned}\sigma(x,y) &= \sigma(y,x) = E\{[x - \mu_x][y - \mu_y]\} \\ &= \iint (x - \mu_x)(y - \mu_y)P(x,y)dxdy \\ &= \iint xyP(x,y)dxdy - \mu_x\mu_y\end{aligned}$$

- Correlation coefficient between real and imaginary parts:

$$\rho(x,y) = \frac{\sigma(x,y)}{\sqrt{\sigma^2(x)\sigma^2(y)}} = \frac{\sigma(x,y)}{\sigma(x)\sigma(y)}$$

- Covariance matrix

$$\begin{pmatrix} \sigma^2(x) & \sigma(x,y) \\ \sigma(x,y) & \sigma^2(y) \end{pmatrix} = \begin{pmatrix} \sigma^2(x) & \rho(x,y)\sqrt{\sigma^2(x)\sigma^2(y)} \\ \rho(x,y)\sqrt{\sigma^2(x)\sigma^2(y)} & \sigma^2(y) \end{pmatrix}$$

Sample Statistics

If the sample consists of the n complex numbers $x_i + jy_i$ ($i = 1$ to n) then the following statistics calculated from the sample can be used to estimate the corresponding parameters of the parent population:

- Mean of real part:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

- Mean of imaginary part:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- Variance of real part:

$$s^2(x) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Variance of imaginary part:

$$s^2(y) = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

- Standard deviation of real part

$$s(x) = \sqrt{s^2(x)}$$

- Standard deviation of imaginary part

$$s(y) = \sqrt{s^2(y)}$$

- Covariance between real and imaginary parts:

$$s(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Correlation coefficient between real and imaginary parts:

$$r(x, y) = \frac{s(x, y)}{\sqrt{s^2(x)s^2(y)}} = \frac{s(x, y)}{s(x)s(y)}$$

- Covariance matrix

$$\begin{pmatrix} s^2(x) & s(x, y) \\ s(x, y) & s^2(y) \end{pmatrix} = \begin{pmatrix} s^2(x) & r(x, y)\sqrt{s^2(x)s^2(y)} \\ r(x, y)\sqrt{s^2(x)s^2(y)} & s^2(y) \end{pmatrix}$$

The following sample statistics estimate the corresponding parameters of the distribution of sample mean which is also a bivariate normal distribution

- Variance in mean of real part:

$$s^2(\bar{x}) = \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Variance in mean of imaginary part:

$$s^2(\bar{y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (y_i - \bar{y})^2$$

- Standard error in mean of real part

$$s(\bar{x}) = \sqrt{s^2(\bar{x})}$$

- Standard error in mean of imaginary part

$$s(\bar{y}) = \sqrt{s^2(\bar{y})}$$

- Covariance in mean between real and imaginary parts:

$$s(\bar{x}, \bar{y}) = \frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Correlation coefficient between real and imaginary parts:

$$r(\bar{x}, \bar{y}) = \frac{s(\bar{x}, \bar{y})}{\sqrt{s^2(\bar{x})s^2(\bar{y})}} = \frac{s(\bar{x}, \bar{y})}{s(\bar{x})s(\bar{y})} = r(x, y)$$

- Covariance matrix of mean

$$\begin{pmatrix} s^2(\bar{x}) & s(\bar{x}, \bar{y}) \\ s(\bar{x}, \bar{y}) & s^2(\bar{y}) \end{pmatrix} = \begin{pmatrix} s^2(\bar{x}) & r(\bar{x}, \bar{y})\sqrt{s^2(\bar{x})s^2(\bar{y})} \\ r(\bar{x}, \bar{y})\sqrt{s^2(\bar{x})s^2(\bar{y})} & s^2(\bar{y}) \end{pmatrix}$$

APPENDIX 2 THE CORRELATION COEFFICIENT FOR A SAMPLE OF SIZE 2 IS ± 1

For a sample of size 2 from a bivariate distribution consisting of points (x_1, y_1) and (x_2, y_2) the correlation coefficient $r(x, y)$ is given by

$$r(x, y) = \frac{(x_1 - \bar{x})(y_1 - \bar{y}) + (x_2 - \bar{x})(y_2 - \bar{y})}{[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2]^{1/2} [(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2]^{1/2}}$$

where $\bar{x}$ is the mean of the first component and $\bar{y}$ is the mean of the second component. It follows that

$$\begin{aligned}(x_1 - \bar{x}) &= x_1 - \left(\frac{x_1 + x_2}{2} \right) \\ &= \frac{1}{2}(x_1 - x_2)\end{aligned}$$

Similarly

$$(x_2 - \bar{x}) = -\frac{1}{2}(x_1 - x_2)$$

$$(y_1 - \bar{y}) = \frac{1}{2}(y_1 - y_2)$$

$$(y_2 - \bar{y}) = -\frac{1}{2}(y_1 - y_2)$$

Hence

$$r(x, y) = \frac{\frac{1}{2}(x_1 - x_2)(y_1 - y_2)}{\frac{1}{2}|(x_1 - x_2)(y_1 - y_2)|} = \pm 1$$

APPENDIX 3 A PROCEDURE FOR SAMPLING FROM A BIVARIATE NORMAL DISTRIBUTION

A procedure for sampling from a multivariate normal distribution

Maurice Cox, Centre for Information Systems Engineering, NPL

20 May 1998

Abstract

This document describes a procedure for sampling from a multivariate normal (Gaussian) distribution, given the defining parameters of that distribution. The approach uses a decomposition of the covariance matrix of the distribution, together with a conventional Gaussian random number generator.

1 Introduction

Of particular value in electrical impedance work and in a number of other areas of metrology is the ability to sample from a *multivariate* distribution. Often the need is to work with a *bivariate* distribution, where the two variables correspond to the real and imaginary part of a complex variable. Sometimes it is necessary to work with a number of complex variables. Frequently, there exists correlation between these components. This paper describes an approach for such sampling.

In what follows the symbol j is reserved for $\sqrt{-1}$.

2 Problem statement

Let x_i , $i = 1, \dots, n$, be n random variables, in general correlated. Let μ_i denote the (arithmetic) mean of x_i . Let

$$V = \begin{pmatrix} \sigma_1^2 & \sigma_{1,2} & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \sigma_{2,n} \\ \vdots & \vdots & \vdots \\ \sigma_{n,1} & \sigma_{n,2} & \sigma_n^2 \end{pmatrix} \quad (1)$$

denote the covariance matrix of the x_i . Here

$$\sigma_i^2 = E\{(x_i - \mu_i)^2\}$$

is the variance of x_i (i.e., σ_i is the standard deviation of x_i) and

$$\sigma_{i,k} = E\{(x_i - \mu_i)(x_k - \mu_k)\}$$

is the covariance of x_i and x_k , where E denotes mathematical expectation. Since $\sigma_{i,k} = \sigma_{k,i}$, the matrix is symmetric.

Additionally, V is positive definite or at least semi-definite [2, page 322]. In practice, V will be positive definite unless there is perfect correlation ($\sigma_i\sigma_k = \sigma_{i,k}$) between two variables x_i and x_k . It is possible that due to rounding or truncation errors in the way it was formed that V will fail to be positive definite or even semi-definite. In this case, V can be “repaired” such that it satisfies the required property: see Section D.

It is required to sample from the multivariate normal distribution whose mean is $\mu = [\mu_1, \dots, \mu_n]^T$ and whose covariance matrix is V .

The sample will be in the form of a set of points, each of which is an n -dimensional point.

3 Theory

Let $\mathbf{d} = [d_1, \dots, d_n]^T$, where each d_i is an independent sample from the standardized normal distribution (viz., with mean zero and standard deviation unity) $N(0, 1)$.

Let

$$V = M^T M \quad (2)$$

be a factorization of V .

Then

$$\mathbf{e} = M^T \mathbf{d}$$

is the required sample.

The proof of this result is given in Appendix A.

It remains to determine a suitable factorization of (2). Since the above proof holds for any such factorization, we select the Cholesky factorization (see Appendix B) on the grounds of simplicity, efficiency, numerical stability and availability of algorithms.

The Cholesky factor R is an upper triangular matrix of the form

$$R = \begin{pmatrix} r_{1,1} & r_{1,2} & \cdot & r_{1,n} \\ 0 & r_{2,2} & \cdot & r_{2,n} \\ \cdot & \vdots & \cdot & \vdots \\ 0 & 0 & \cdot & r_{n,n} \end{pmatrix} \quad (3)$$

The elements $r_{i,k}$ of R can be found by equating coefficients in the defining equation $R^T R = V$, yielding the *Cholesky decomposition algorithm*. An illustration of the derivation of the algorithm for the case $n = 2$ is given in Appendix C.

4 High-level solution

The high-level solution to the problem can be expressed succinctly in the following step-by-step manner.

Input the parameters of the n -dimensional multivariate normal distribution, i.e., the vector μ of n means and the variance-covariance matrix V of order n .

2. Calculate the Cholesky factor of V , i.e., the upper-triangular matrix R of order n such that $R^T R = V$.
3. Use a Gaussian random number generator to generate a vector $\mathbf{d}$ of n normal variates drawn from a distribution with zero mean and unit standard deviation.
4. Provide the required sample, a vector $\mathbf{e}$ of n values drawn from a multivariate normal distribution with the given parameters by forming $\mathbf{e} = R^T \mathbf{d}$.
5. Repeat Steps 3 and 4 as many times as required.

5 Matlab software

Appendix E contains the listing of a Matlab function MULTNORM.M for determining a sample of size m from a multivariate normal distribution, given the vector μ of means and the covariance matrix V . This function allows V to be indefinite (as might arise in a case where the variables are highly correlated and rounding errors have been committed in forming V). Minimal correction is made to V in order to ensure that it is appropriate numerically for Cholesky factorization. A function CVREPAIR.M is provided in Appendix F for this purpose.

Appendix G contains the listing of a Matlab script MLTNRMDR.M which acts as a driver for MULTNORM. This script can be used in either of two modes:

As a demonstration program, in the case of a bivariate normal distribution, with fixed, pre-specified mean values and covariance matrix,

2. Under user-control, in which the means and covariance matrix can be entered, as well as the size of the required sample.

In the former case, or if $n = 2$ in the second, a plot of the sample is provided.

The plot produced by the demonstration program is shown in Figure 1.

6 Implementation

A Fortran routine callable from Visual Basic, based on the appended Matlab functions and scripts, will be prepared which returns numbers drawn from a multivariate normal distribution. Usage documentation and notes explaining each module will be provided.

The Matlab routines that prototype such an implementation are provided here. The only exceptions are the standard facilities, found in typical software libraries, for

- Cholesky decomposition,
- eigendecomposition,
- generating numbers from a univariate normal distribution,
- the computational precision.

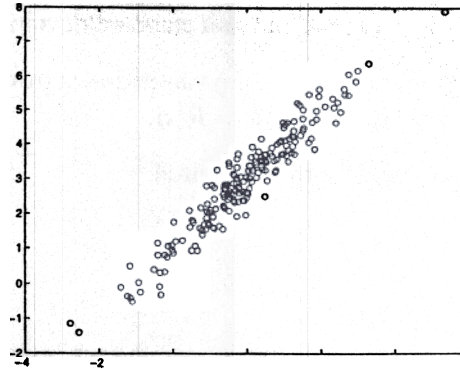


Figure A sample from a bivariate normal distribution having highly correlated variables

7 Examples of use

Examples of use of the Fortran routine will be provided.

8 Repairing a covariance matrix

For the reasons stated in Section 5, a specified covariance matrix V may not be nonnegative definite, in a case where correlation is very high, as a consequence of rounding and other errors made in its formation. Such a matrix can be “repaired”, by replacing it by the matrix that is closest to it in an appropriate sense and has the required properties. An algorithm for this purpose [2, page 322] is based on the eigendecomposition of V . A slight extension of this algorithm, to render it more robust, is given in Appendix D.

9 Concluding remarks

An approach for sampling from multivariate normal distributions has been described. Prototype software in the form of Matlab functions based on the approach have been provided. It is intended that “production” software based on the prototype will be used for a range of simulation problems arising in impedance work.

References

- [1] G. H. Golub and C. F. Van Loan. *Matrix Computations*. John Hopkins University Press, Baltimore, MD, USA, 1996. Third edition.

- [2] G. Strang and K. Borre. *Linear Algebra, Geodasy and GPS*. Wiley, Wellesley-Cambridge Press, 1997.

A Proof of sampling by factorization

Let

$$V = M^T M$$

be a factorization of a covariance matrix V order n .

Then, if $\mathbf{d}$ is a sample of size n from the standardized normal distribution $N(0, 1)$,

$$\mathbf{e} = M^T \mathbf{d}$$

is the required sample.

The proof of this result is straightforward. Since $E\{\mathbf{d}\} = 0$, the expectation of $\mathbf{e}$ is

$$E\{\mathbf{e}\} = E\{M^T \mathbf{d}\} = M^T E\{\mathbf{d}\} = 0.$$

The covariance of $\mathbf{e}$ is therefore

$$\text{cov}(\mathbf{e}) = E\{\mathbf{e}\mathbf{e}^T\} = E\{M^T \mathbf{d} \mathbf{d}^T M\} = M^T E\{\mathbf{d} \mathbf{d}^T\} M.$$

But $E\{d_i d_k\}$ is unity if $i = k$ and zero otherwise. Hence, $E\{\mathbf{d} \mathbf{d}^T\}$ is the identity matrix, and so

$$\text{cov}(\mathbf{e}) = M^T M = V,$$

as required.

B Factorizations of the covariance matrix

There are several possible factorizations of the covariance matrix V of the form $V = M^T M$ (Equation (2)), principally those given by

1. $M = R$, where R is the Cholesky factor of V [1, page 143],
2. $M = D^{1/2} Q^T$, where D is the diagonal matrix containing as its diagonal elements the eigenvalues of V and Q is the matrix of corresponding eigenvectors of V [1, page 393],
3. $M = H$, where H is the matrix square root of V [1, page 149].

Because Option 1 satisfies the criteria of simplicity, efficiency, numerical stability and availability of algorithms, we select it for the current purpose. It can be formed using a simple and stable algorithm and the operation count for determining it is proportional to n^3 . For Options 2 and 3 the corresponding algorithms are also stable but more complicated, and the operation counts are larger multiples of n^3 .

Note that, if required, both factorization and repairing could be carried out in terms of the eigendecomposition.

C Illustrative derivation of the Cholesky decomposition algorithm

Given a positive semi-definite (covariance) matrix V of order n , the defining equation for the (upper-triangular) Cholesky factor (of order n) is

$$V = R^T R. \quad (4)$$

Consider the derivation of the Cholesky decomposition algorithm in the case $n = 2$. In full, for this case, Equation (4), using (1) and (3) and the symmetry of V , becomes

$$\begin{bmatrix} r_{1,1} & 0 \\ r_{1,2} & r_{2,2} \end{bmatrix} \begin{bmatrix} r_{1,1} & r_{1,2} \\ 0 & r_{2,2} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$$

from which, by comparing elements,

$$\begin{aligned} r_{1,1}^2 &= \sigma_1^2, \\ r_{1,1}r_{1,2} &= \sigma_{1,2}, \\ r_{1,2}^2 + r_{2,2}^2 &= \sigma_2^2. \end{aligned}$$

Hence the required elements of R can immediately be determined using

$$\begin{aligned} r_{1,1} &= \sigma_1, \\ r_{1,2} &= \sigma_{1,2}/r_{1,1} \\ r_{2,2} &= (\sigma_2^2 - r_{1,2}^2)^{1/2}. \end{aligned}$$

The generalization to order n is straightforward [1, page 143].

D Repairing a covariance matrix that is not nonnegative definite

An algorithm for repairing a covariance matrix [2, page 322] first determines the eigendecomposition

$$V = Q D Q^T, \quad (5)$$

where D is the diagonal matrix of eigenvalues of V and Q the corresponding matrix of eigenvectors. If one (or more) of the eigenvalues (diagonal elements of D) is negative, V is not strictly a *covariance matrix* and does not possess a Cholesky factor. The closest such acceptable matrix is given by replacing any offending eigenvalue by zero, i.e., replacing diagonal elements of D by zero and re-forming V from Equation (5). This operation is equivalent to expressing the decomposition as

$$V = \sum_{i=1}^n \lambda_i \mathbf{q}_i \mathbf{q}_i^T,$$

and replacing it by

$$V = \sum_{i: \lambda_i \geq 0} \lambda_i \mathbf{q}_i \mathbf{q}_i^T.$$

Mathematically, this approach is fully acceptable. Numerically, there can be a difficulty, however. The Cholesky decomposition itself suffers from (inevitable but mild) rounding errors. Since the resulting

matrix is semidefinite, it may not be possible to form the Cholesky decomposition. The problem will manifest itself as an attempt to take the square root of a number which is negative (but small in magnitude). See Appendix C.

As a consequence of the above discussion, the approach of [2, page 322] is extended slightly by replacing each negative λ_i by

$$C\eta \max_{1 \leq k \leq n} |\lambda_k|,$$

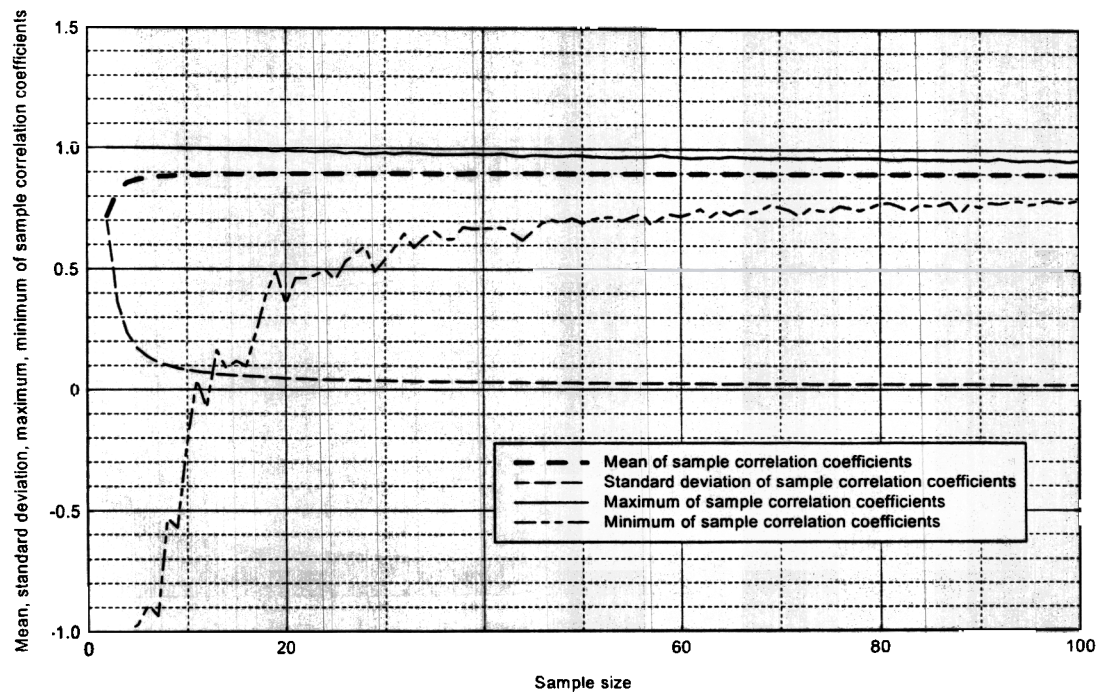
where C is a positive constant, and η the *computational precision*, i.e., the smallest representable floating-point number such that when added to unity in the floating-point arithmetic employed, the result differs from unity. This approach has been found to operate satisfactorily with $C = 1$. Research would be needed to find a value of C which guarantees that in the presence of rounding error the Cholesky decomposition of the repaired V always exists.

```
X = randn(n, m);
% Provide the required sample (the Cholesky factor acts as a transformation
% from the uncorrelated standardized space to that required)
Y = mu*ones(1,m) + R'*X;
% end MULTNORM.M
% -----
% This software has not been fully subjected to NPL's Quality
% Assurance procedures. No warranty or guarantee applies to this
% software, and therefore any users should satisfy themselves
% that it meets their requirements.
% -----
```

% Assurance procedures. No warranty or guarantee applies to this
% software, and therefore any users should satisfy themselves
% that it meets their requirements.
% -----

```
        figure(gcf)
    end
% end MLTNRMDR.M
% -----
% This software has not been fully subjected to NPL's Quality
% Assurance procedures. No warranty or guarantee applies to this
% software, and therefore any users should satisfy themselves
% that it meets their requirements.
% -----
```

Fig 3 Properties of distribution of sample correlation coefficients of 40,000 random samples from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9 as a function of sample size



E Matlab function for generating samples from a multivariate normal distribution

```

function Y = multnorm(mu, V, m, repair)
% MULTNORM.M      Samples from a multivariate normal distribution.
%
% Version 0.1      Created 98-03-20.  Last updated 98-04-23.
% Project file     ISE/C3/46A.
% Author           M G Cox, Scientific Software Section,
%                  Information Systems Engineering, NPL.  Crown Copyright.
%-----
% MULTNORM determines samples from a multivariate normal distribution,
% given the defining parameters of that distribution.
%-----
% Input
%   mu      n x 1  Vector of mean values
%   V       n x n  Covariance matrix (only the upper triangle is used)
%   m       1 x 1  Size of required sample
%   repair  1 x 1  Parameter indicating whether the covariance matrix is
%                  to be repaired in the case that numerically it fails
%                  to be positive definite.  Repair is carried out
%                  according to whether this parameter is present or
%                  absent
% Output
%   Y       n x m  Required sample.  Each column contains an
%                  n-dimensional point drawn from the distribution
%-----
% Modules used
%   cvrepair
%-----
% Method of call
%   Y = multnorm(mu, V, m, repair)
%-----

% Check compatibility of input data
n = length(mu);
if size(V,1) ~= n | size(V,2) ~= n
    error('Vector of means and covariance matrix have incompatible dimensions')
end
% Form Cholesky factor R of V
[R, deficiency] = chol(V);
% Repair if required and requested, and recompute R.  Otherwise, report if
% repair not requested and V is nonnegative definite
if deficiency ~= 0
    if nargin > 3
        R = chol(cvrepair(V));
    else
        error('Covariance matrix not semidefinite and repair not requested')
    end
end
% Form m n-dimensional samples from the n-dimensional standardized normal
% distribution N(0, 1) x N(0, 1) x ... x N(0, 1)

```

F Matlab function for repairing a covariance matrix such that it is non-negative definite

```

function Vrep = cvrepair(V)
% CVREPAIR.M      Repairs a covariance matrix such that it is nonnegative
%                  definite.
%
% Version 0.1      Created 98-04-23.  Last updated 98-04-23.
% Project file     ISE/C3/46A.
% Author           M G Cox, Scientific Software Section,
%                  Information Systems Engineering, NPL.  Crown Copyright.
% -----
% CVREPAIR repairs a covariance matrix V such that it is nonnegative
% definite.  This computation is undertaken by
% (1) determining the eigendecomposition  $V = QDQ'$ ,
% (2) replacing negative elements of the vector D of eigenvalues by the
%     product of the computational precision and the largest eigenvalue,
% (3) forming  $Vrep = QDQ'$ .
% This approach constitutes a simple extension of [1].  In [1], the
% negative eigenvalues of D are replaced by zero rather than the value
% above.  Such zero values can still cause V to be perceived as
% indefinite by the Cholesky factorization algorithm, as a
% consequence of rounding errors.
%
% References
% [1] G. Strange and K. Borre.  Linear Algebra, Geodasy and GPS.
%     Wiley, Wellesley-Cambridge Press, 1997.
% -----
% Input
%   V      n x n  Covariance matrix (only the upper triangle is used)
%
% Output
%   Vrep   n x n  Repaired covariance matrix
% -----
% Method of call
%   Vrep = cvrepair(V)
% -----

n = size(V,1);
% Form full V (under the assumption that only the upper triangle was
% provided as input) and its eigendecomposition
[Q, D] = eig(triu(V,1) + diag(diag(V)) + tril(V,-1));
% Form vector whose elements are the eigenvalues D except that
% negative values are replaced by the product of the computational
% precision and the largest eigenvalue
Drep = max([diag(D)'; eps*max(diag(D))*ones(1,n)]);
% Repair V
Vrep = Q*diag(Drep)*Q';
% end CVREPAIR.M
% -----
% This software has not been fully subjected to NPL's Quality

```

G Matlab driver for MULTNORM, sampling from a multivariate normal distribution

```
% MLTNRMDR.M      Driver for MULTNORM - samples from a multivariate
%                  normal distribution.
% Version 0.1      Created 98-03-24.  Last updated 98-04-23.
% Project file     ISE/C3/46A.
% Author           M G Cox, Scientific Software Section,
%                  Information Systems Engineering, NPL.  Crown Copyright.
%-----
% MLTNRMDR calls MULTNORM to determine samples from a multivariate normal
% distribution, given the defining parameters of that distribution.
%-----

userprob = menu('Run demonstration problem or self-selected problem?', ...
                'Demonstration', ...
                'Self-selected');
if userprob == 1
%   Set particular values for a bivariate problem
    mu = [2  3]';
    V = [2.0  1.9;  1.9  2.0];
    m = 200;
else
%   Input user-prescribed values
    mu = input('Provide column vector of mean values ');
    V = input('Provide covariance matrix column by column ');
    m = input('Provide size of sample required ');
end
% Print the problem data
format compact
disp('Means')
disp(mu)
disp('Covariance matrix')
disp(V)
disp('Sample size requested')
disp(m)
% Ask whether covariance-matrix repair is needed
repair = menu('Covariance matrix V to be repaired or not? ',
              'Repair V (the uncommon case)', ...
              'Do not repair V (the usual case)');
% Determine the required sample
if repair == 1
    Y = multnorm(mu, V, m, repair);
else
    Y = multnorm(mu, V, m);
end
% Plot the result if the problem is two-dimensional
if length(mu) == 2
    plot(Y(1,:), Y(2,:), 'o')
    title('Generated sample')
    xlabel('Variable 1')
    ylabel('Variable 2')
```

Fig 1 Properties of distribution of sample correlation coefficients of 40,000 random samples from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0 as a function of sample size

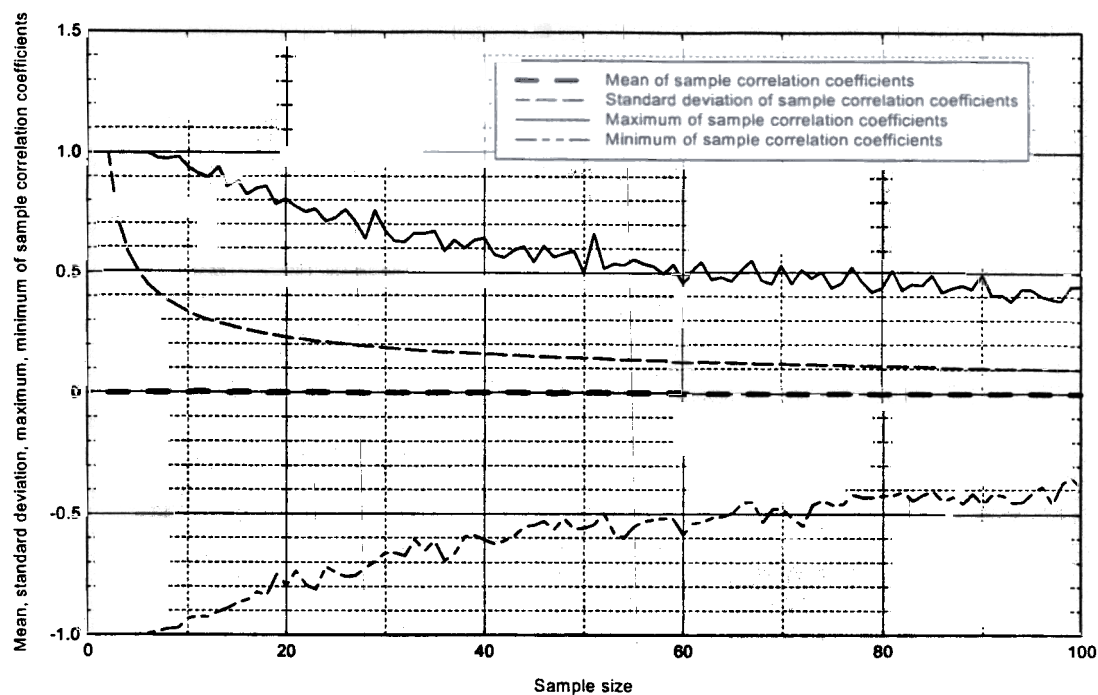


Fig 2 Properties of distribution of sample correlation coefficients of 40,000 random samples from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5 as a function of sample size

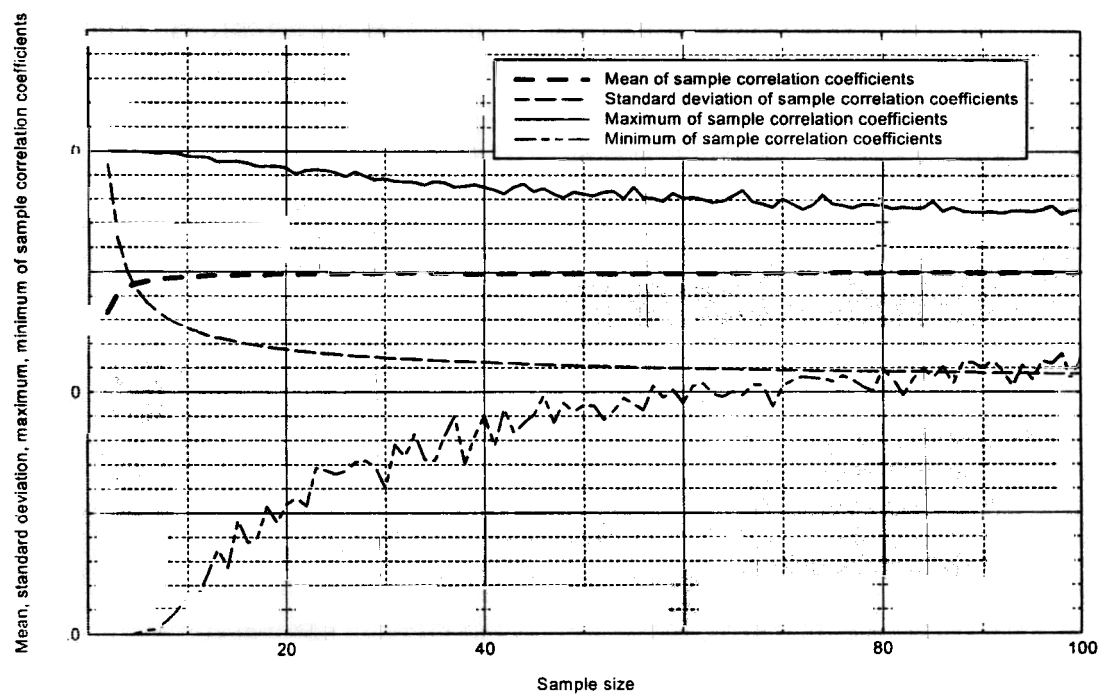


Fig 4 Distribution of sample correlation coefficients of one million random samples of size 3 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

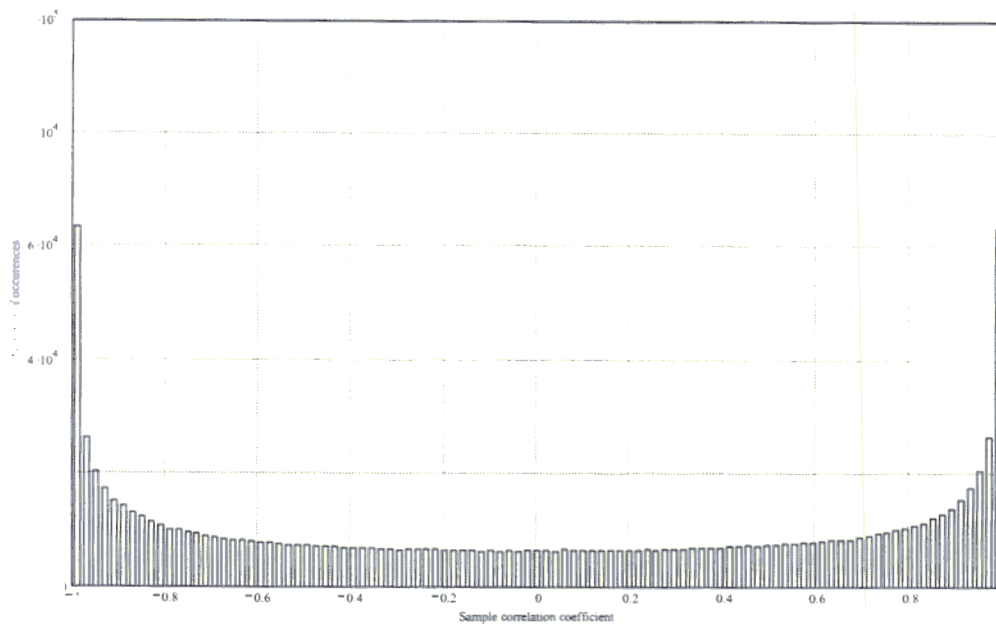


Fig 5 Distribution of Fisher's z corresponding to Fig 4

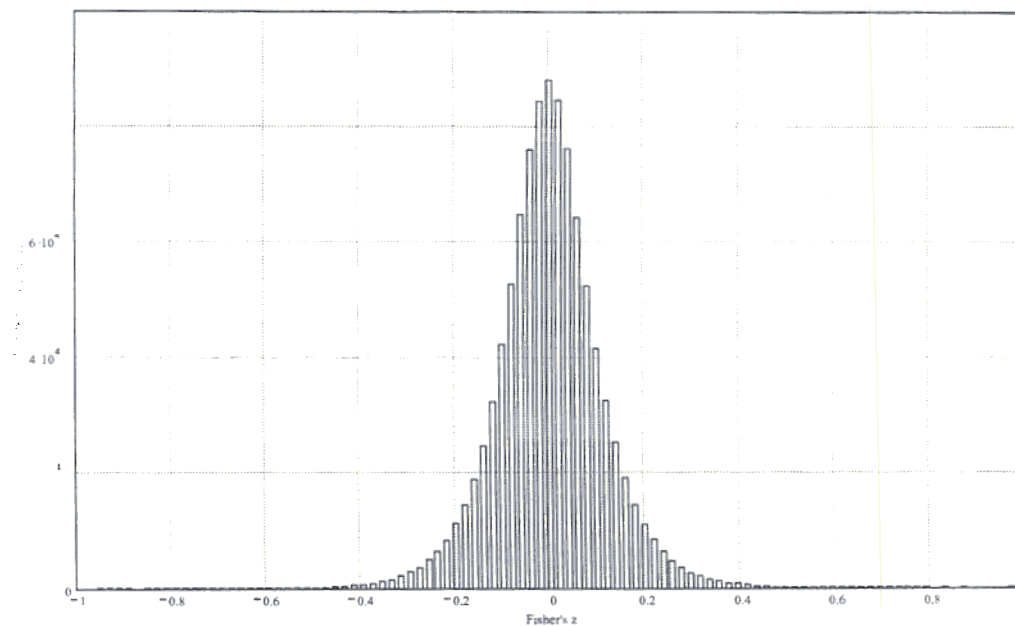


Fig 6 Distribution of sample correlation coefficients of one million random samples of size 4 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

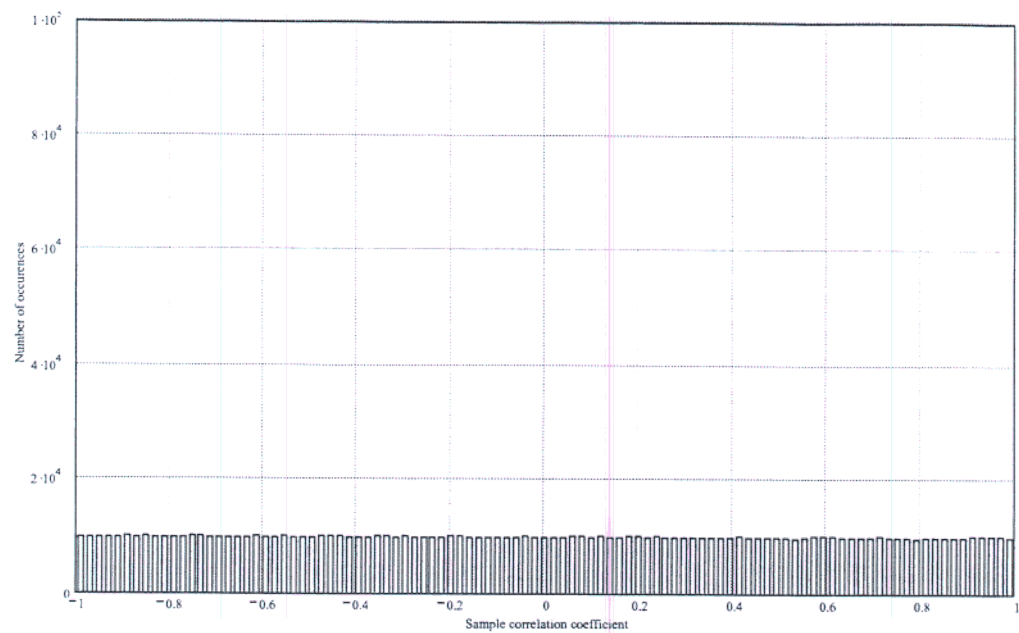


Fig 7 Distribution of Fisher's z corresponding to Fig 6

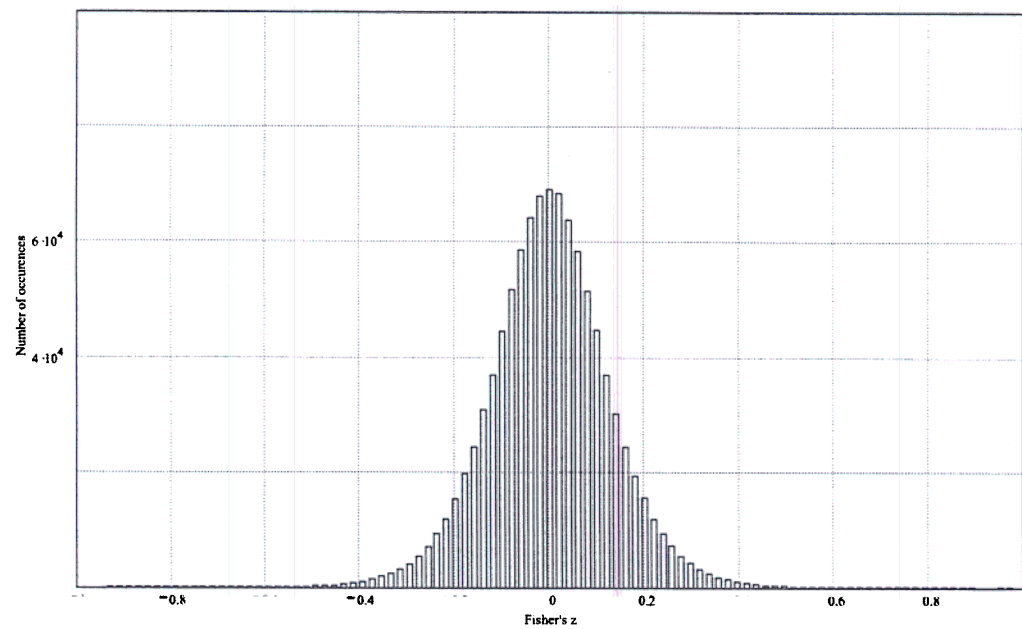


Fig 8 Distribution of sample correlation coefficients of one million random samples of size 5 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

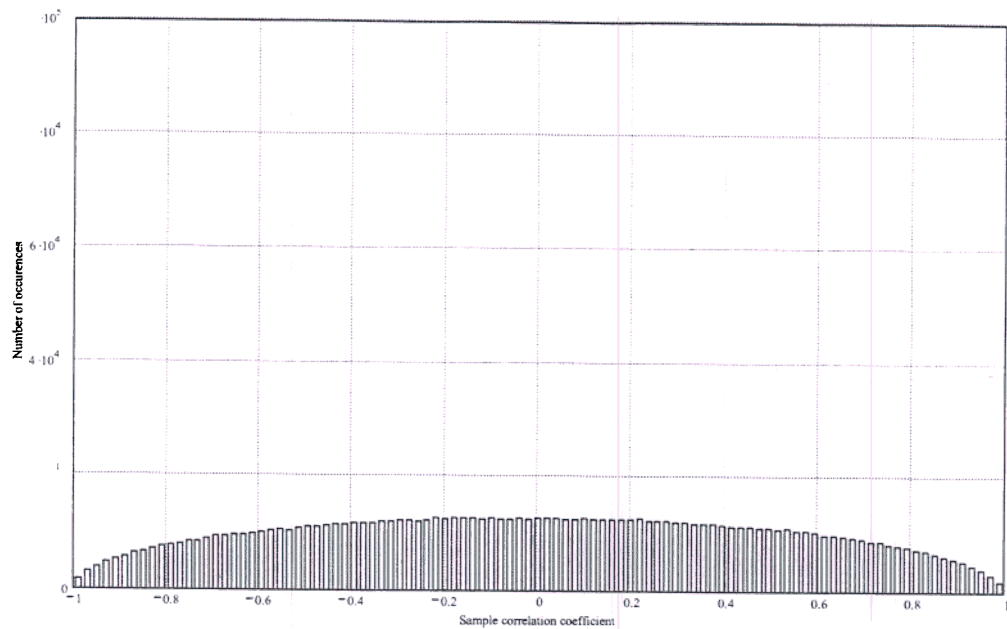


Fig 9 Distribution of Fisher’s z corresponding to Fig 8

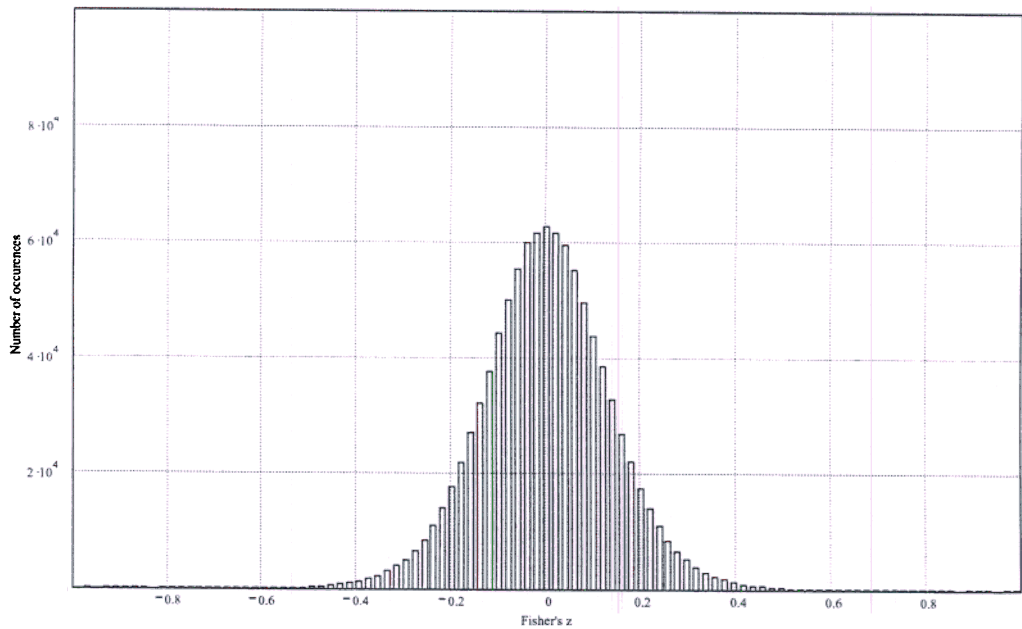


Fig 10 Distribution of sample correlation coefficients of one million random samples of size 7 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

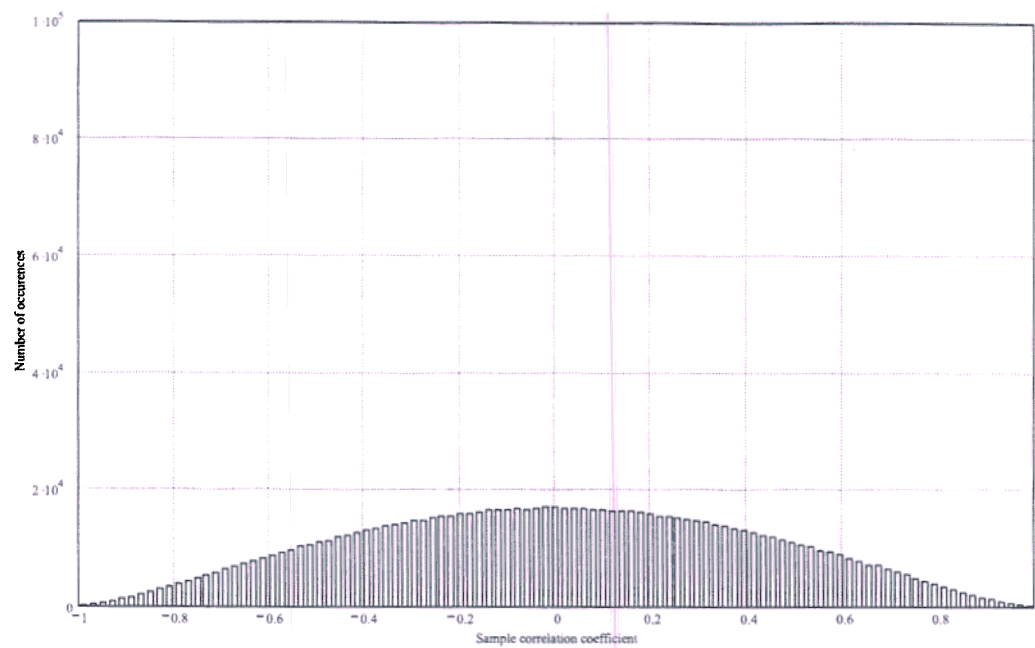


Fig 1 Distribution of Fisher's z corresponding to Fig 10

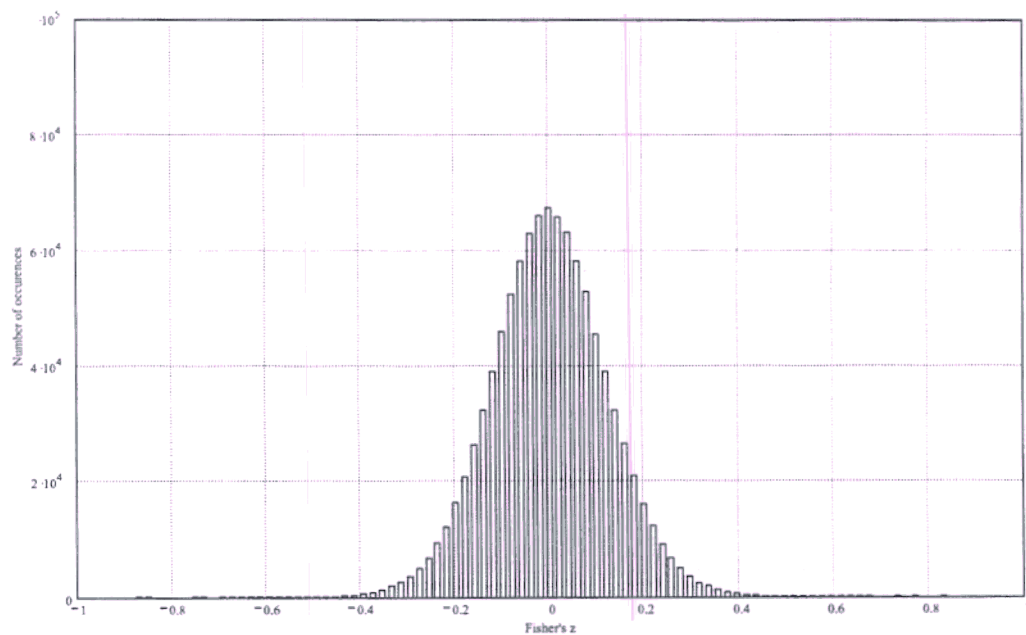


Fig 12 Distribution of sample correlation coefficients of one million random samples of size 10 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

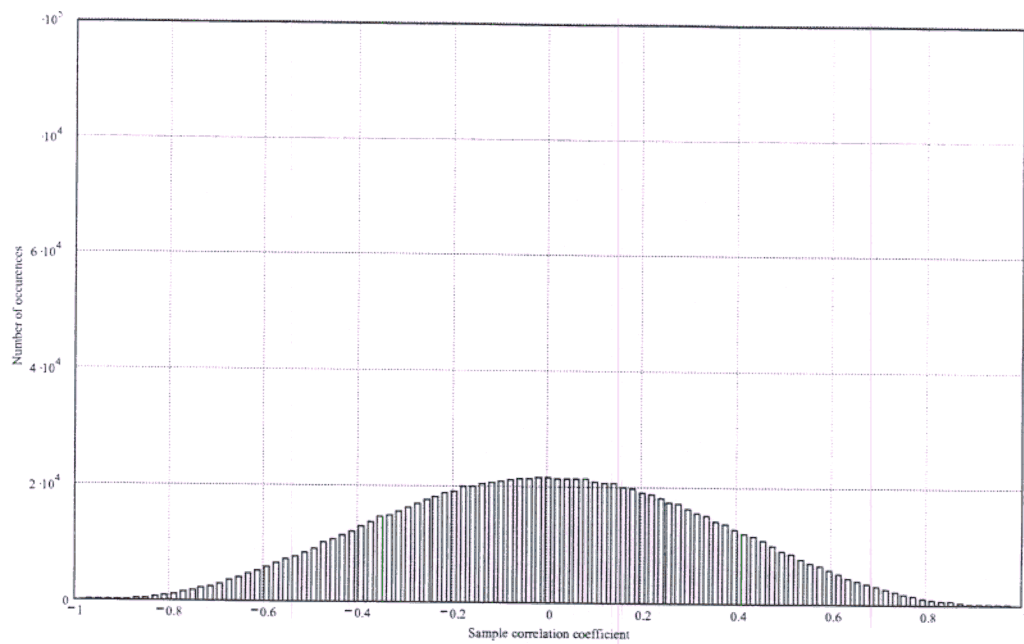


Fig 13 Distribution of Fisher's z corresponding to Fig 12

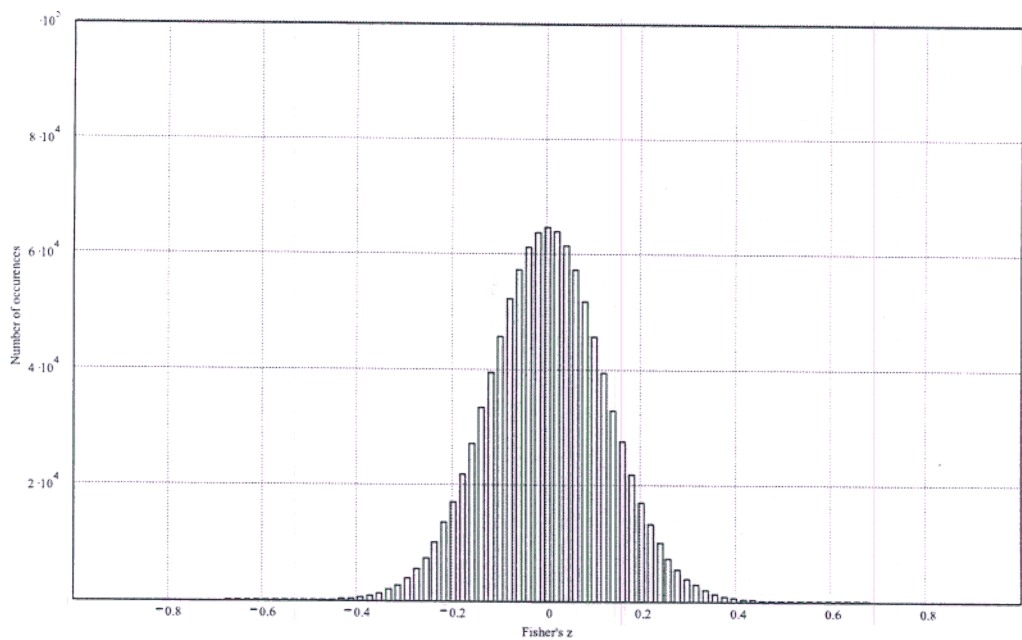


Fig 14 Distribution of sample correlation coefficients of one million random samples of size 100 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0

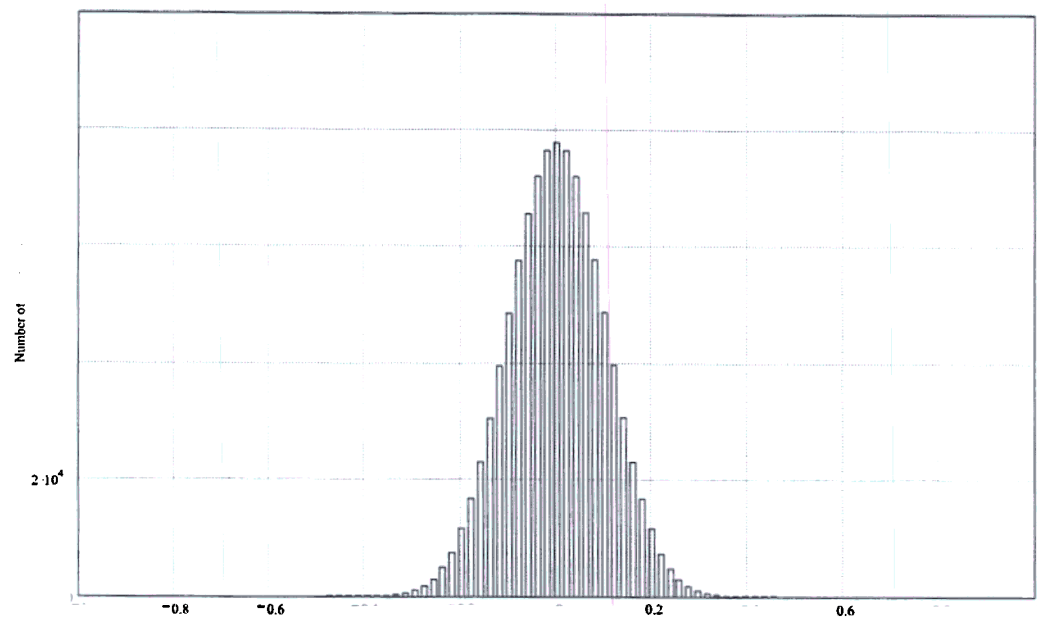


Fig 15 Distribution of Fisher's z corresponding to Fig 14

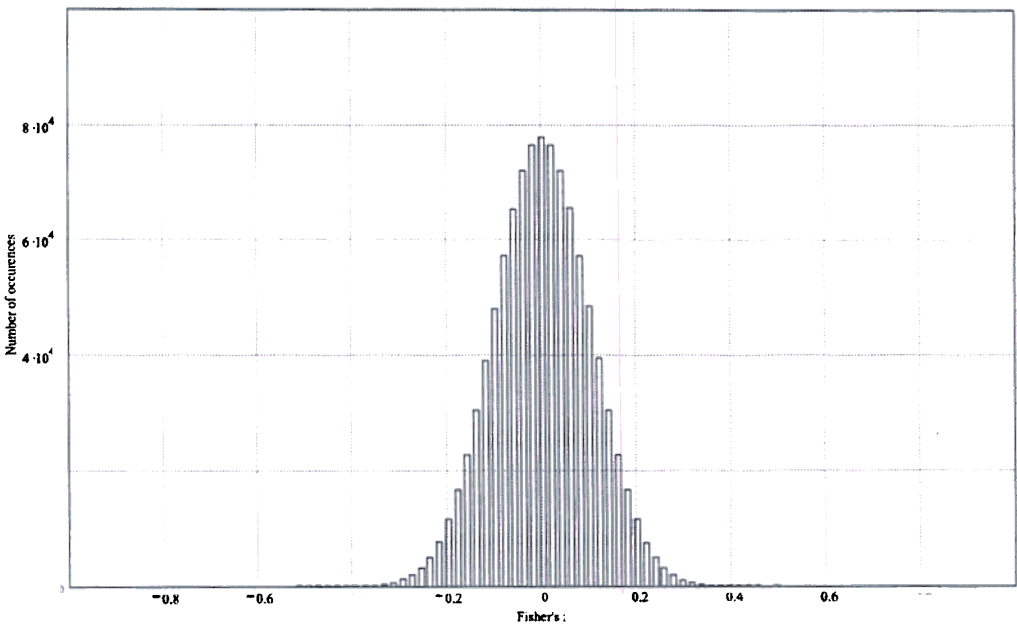


Fig 16 Distribution of sample correlation coefficients of one million random samples of size 3 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

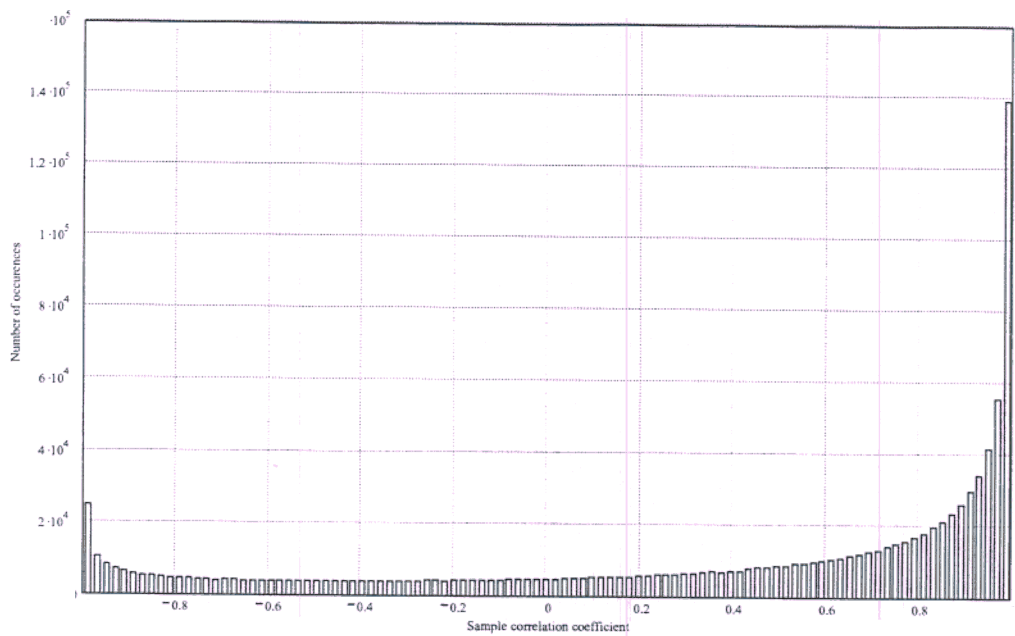


Fig 17 Distribution of Fisher's z corresponding to Fig 16

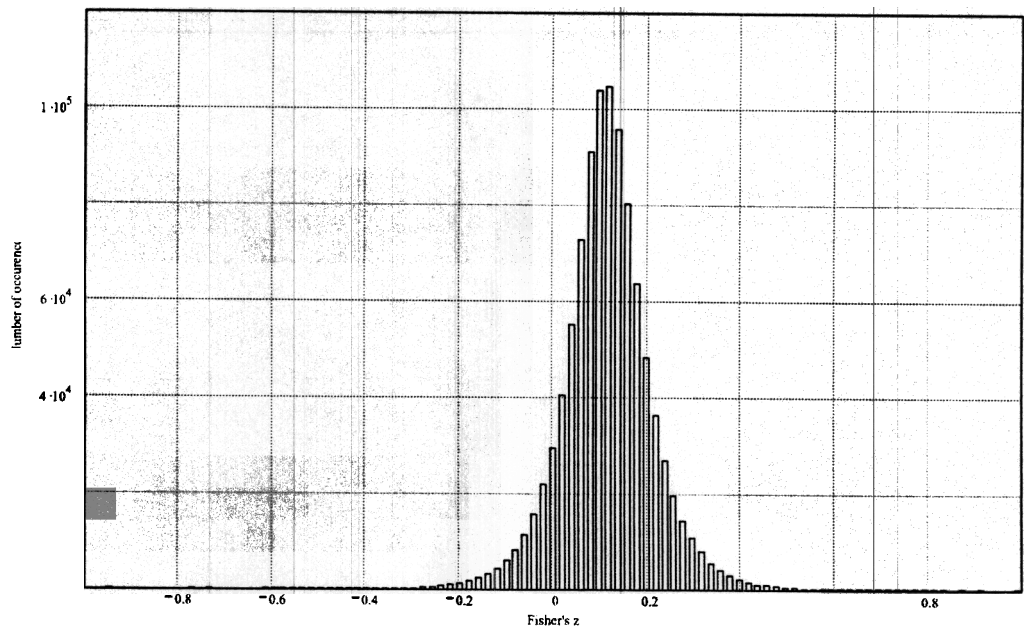


Fig 18 Distribution of sample correlation coefficients of one million random samples of size 4 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

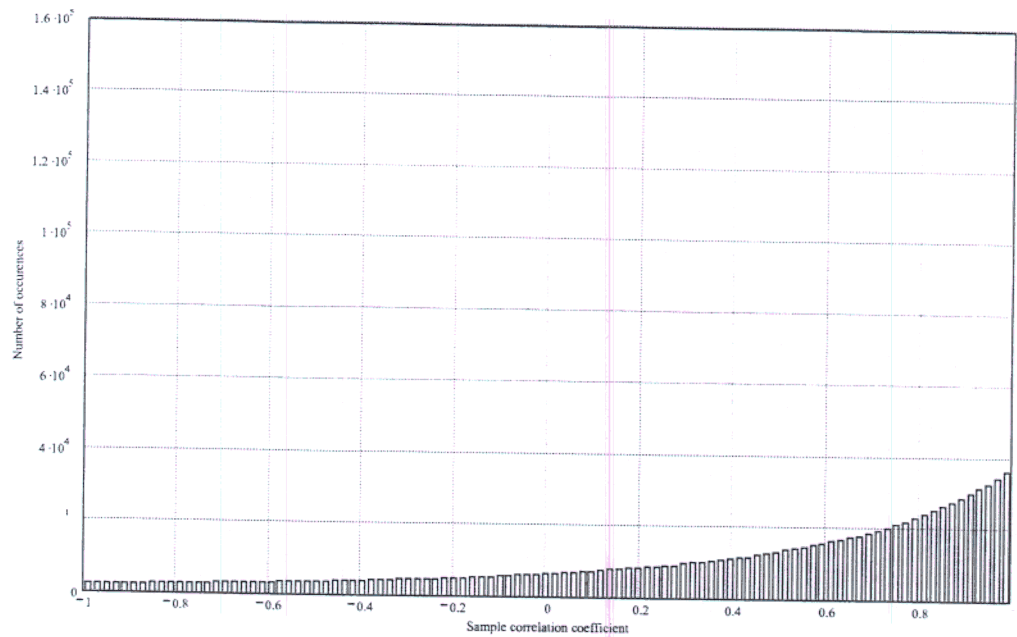


Fig 19 Distribution of Fisher's z corresponding to Fig 18

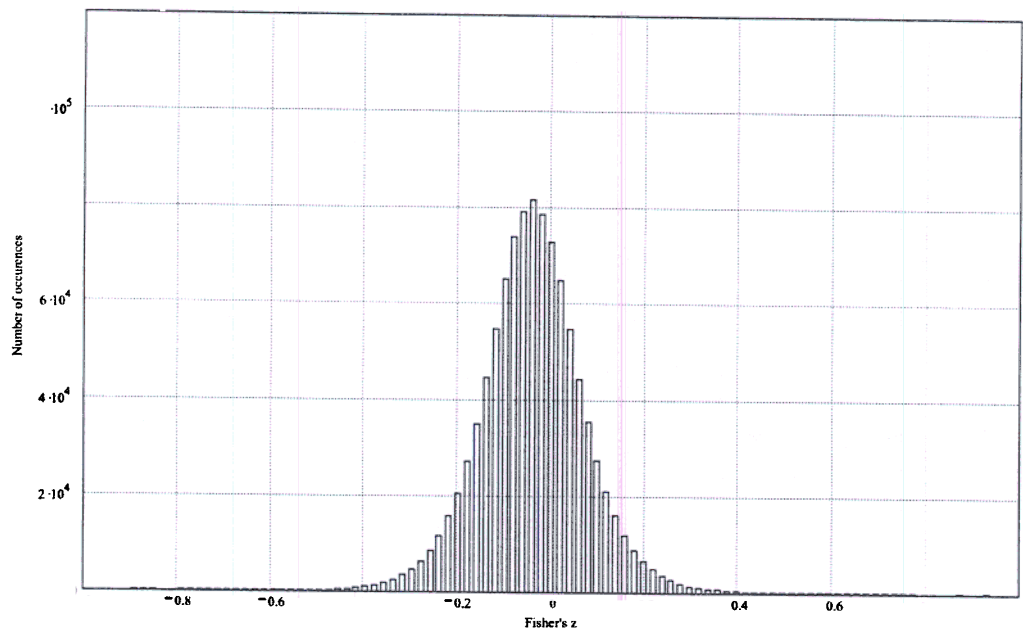


Fig 20 Distribution of sample correlation coefficients of one million random samples of size 5 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

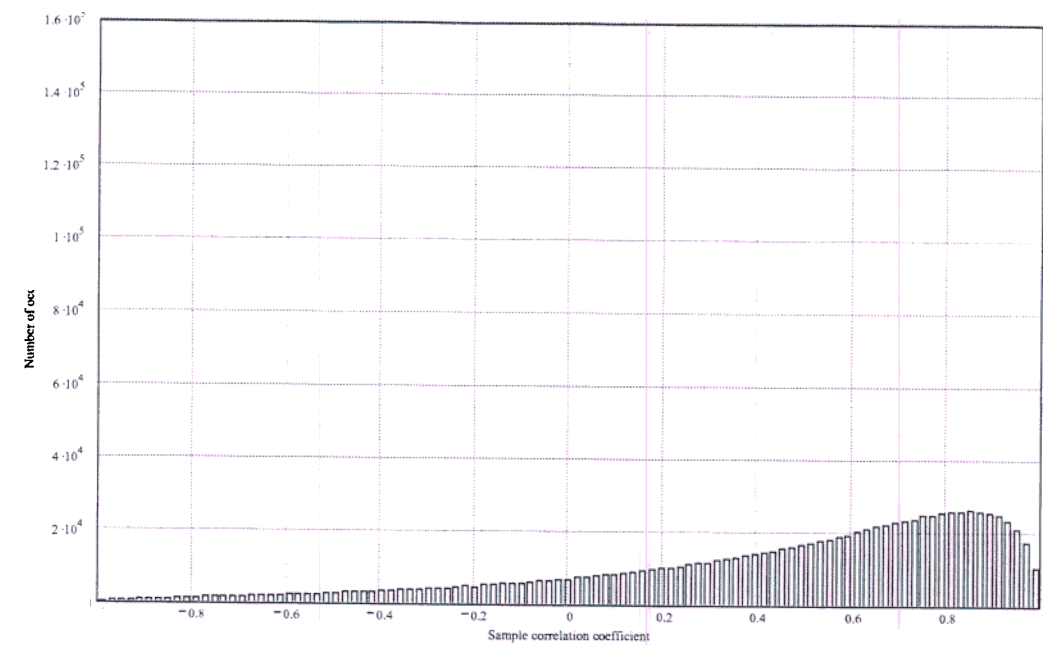


Fig 21 Distribution of Fisher's z corresponding to Fig 20

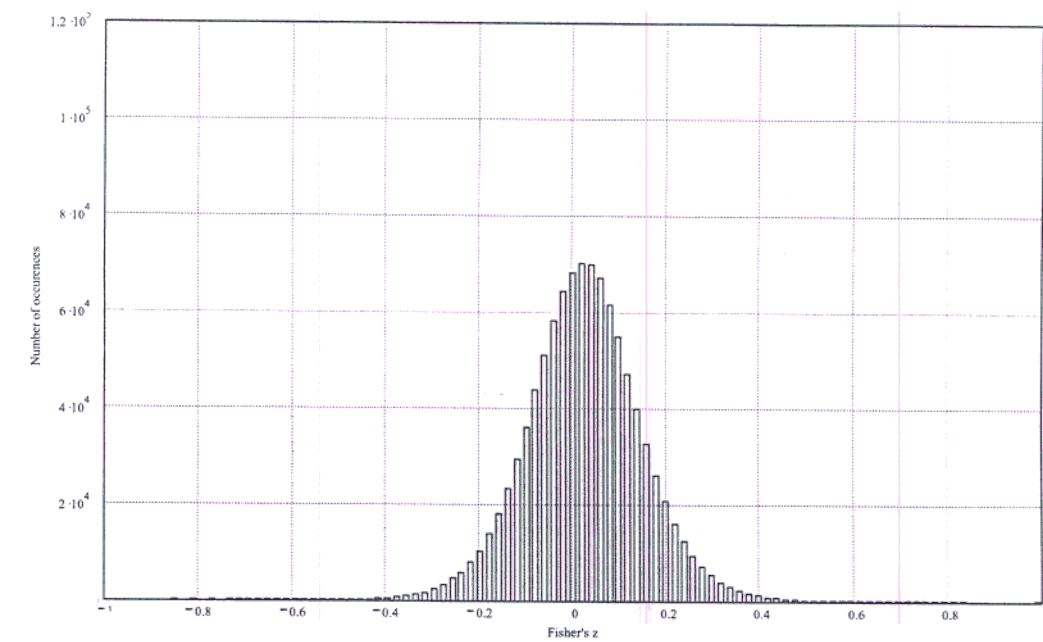


Fig 22 Distribution of sample correlation coefficients of one million random samples of size 7 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

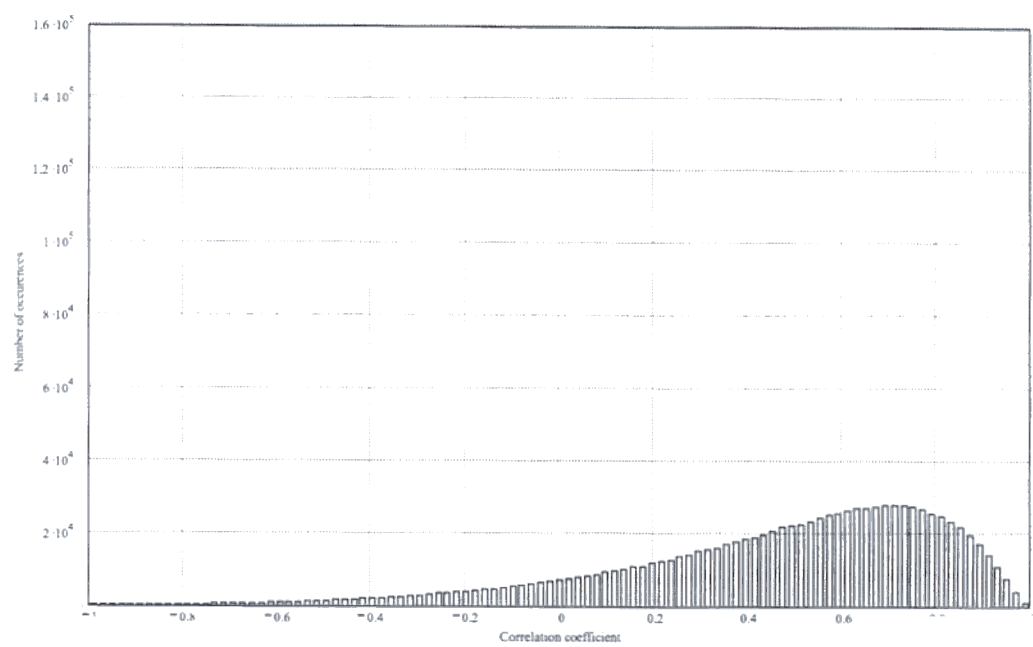


Fig 23 Distribution of Fisher's z corresponding to Fig 22

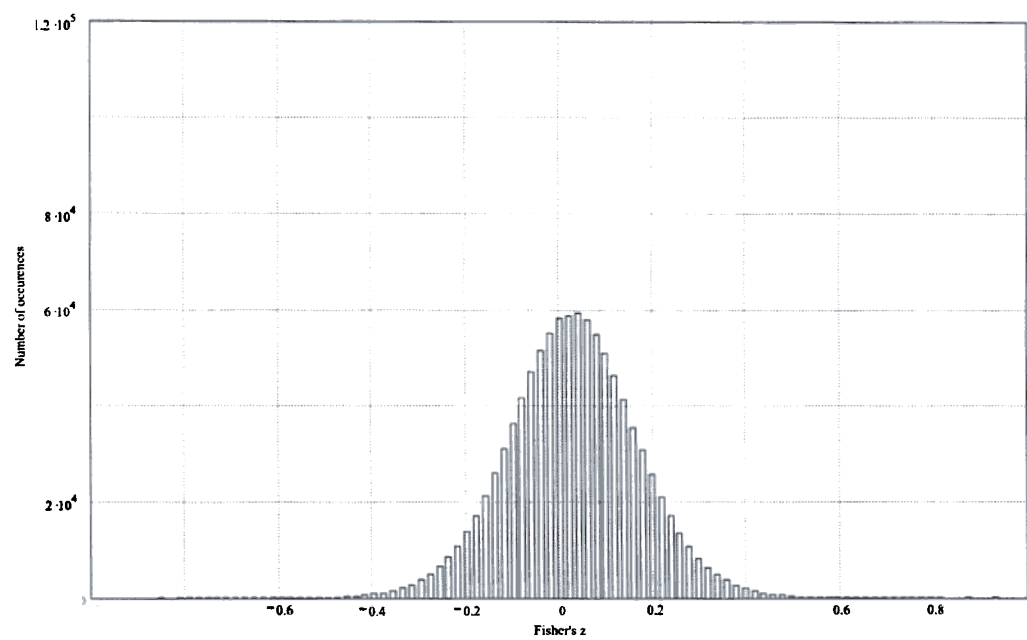


Fig 24 Distribution of sample correlation coefficients of one million random samples of size 10 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

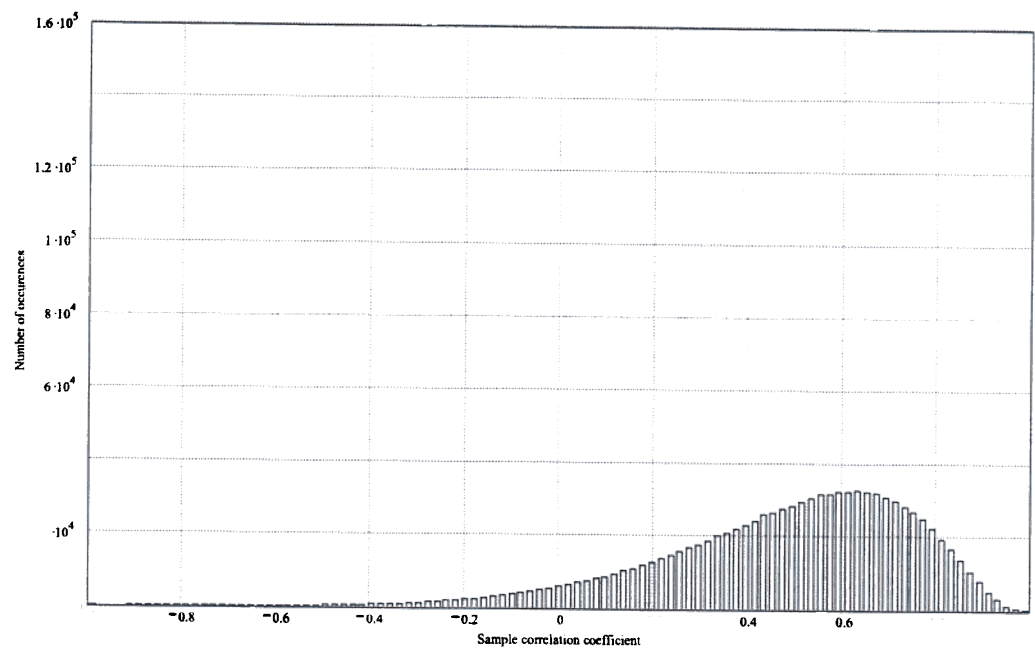


Fig 25 Distribution of Fisher's z corresponding to Fig 24

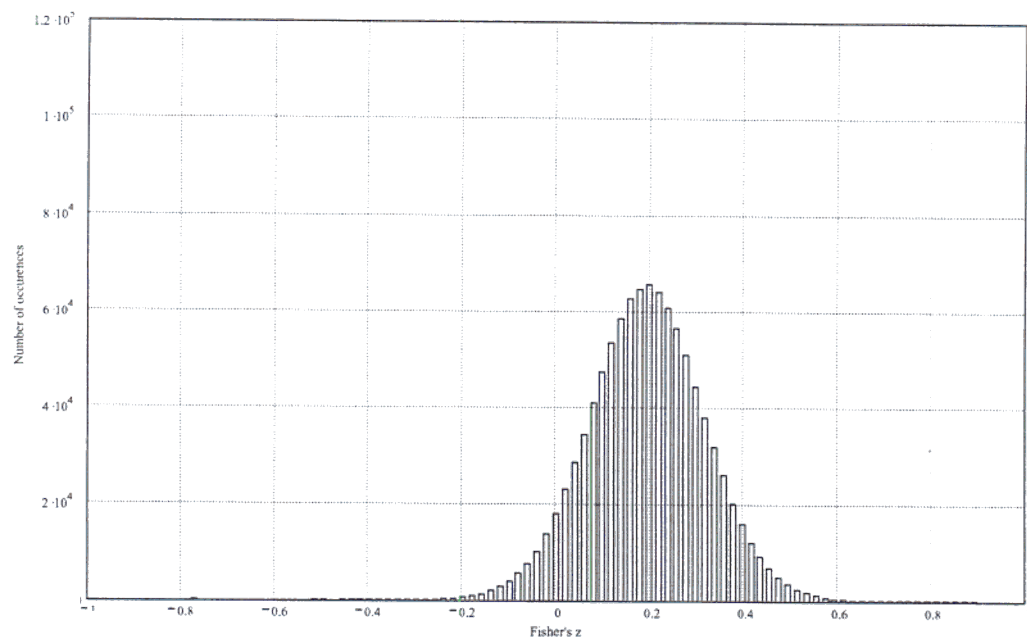


Fig 26 Distribution of sample correlation coefficients of one million random samples of size 100 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.5

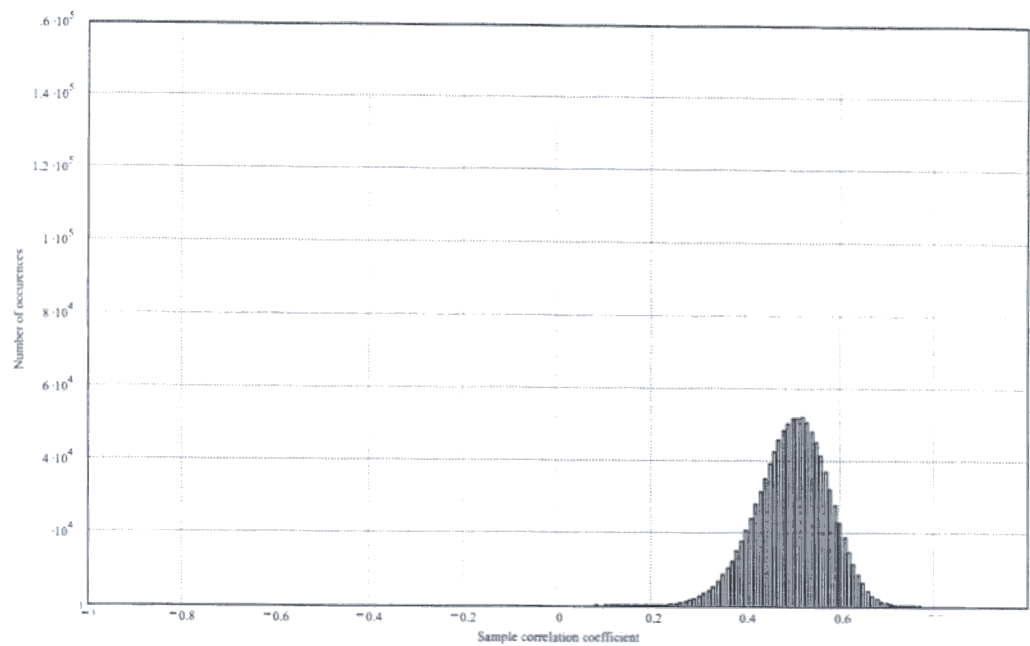


Fig 27 Distribution of Fisher's z corresponding to Fig 26

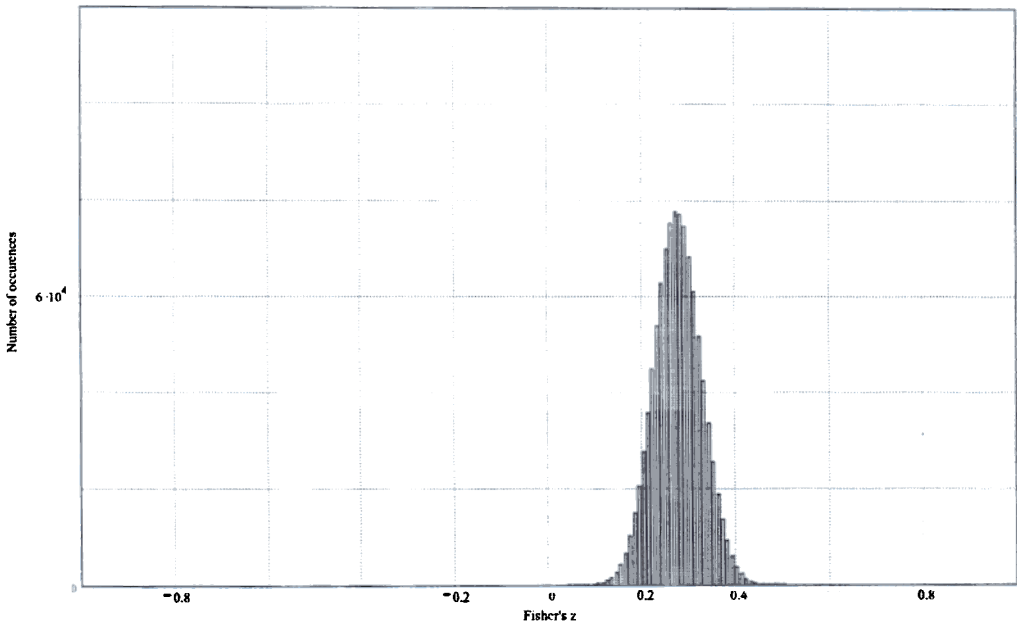


Fig 28 Distribution of sample correlation coefficients of one million random samples of size 3 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

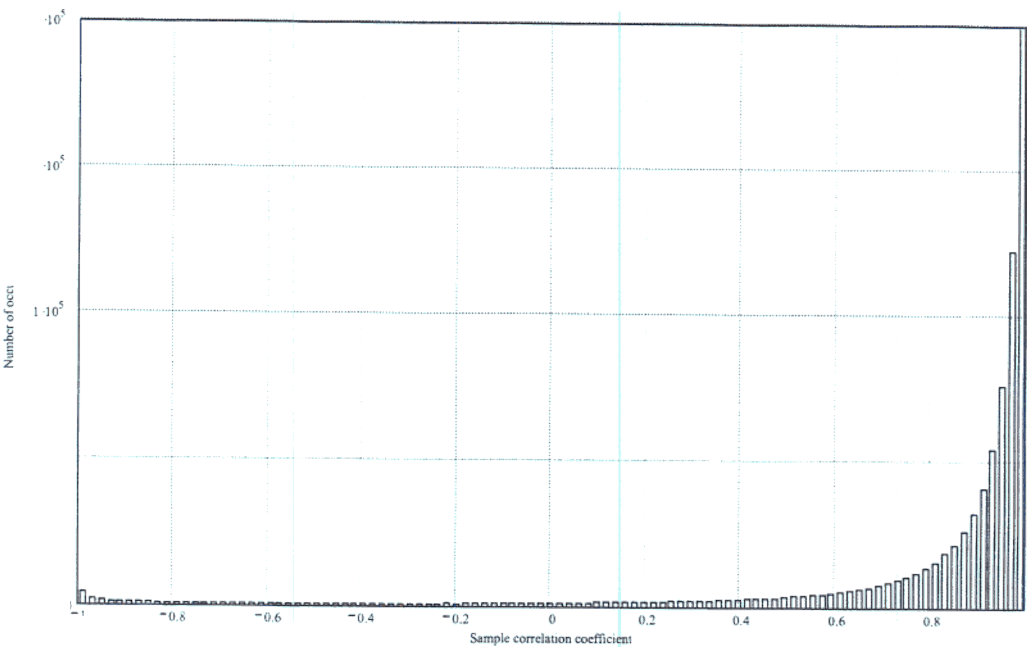


Fig 29 Distribution of Fisher's z corresponding to Fig 28

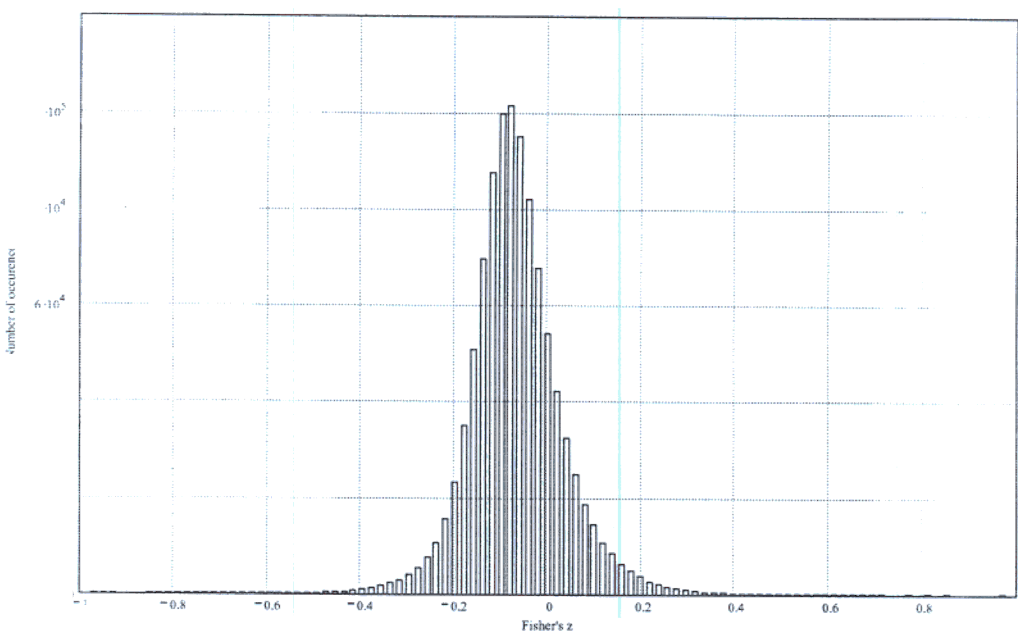


Fig 30 Distribution of sample correlation coefficients of one million random samples of size 4 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

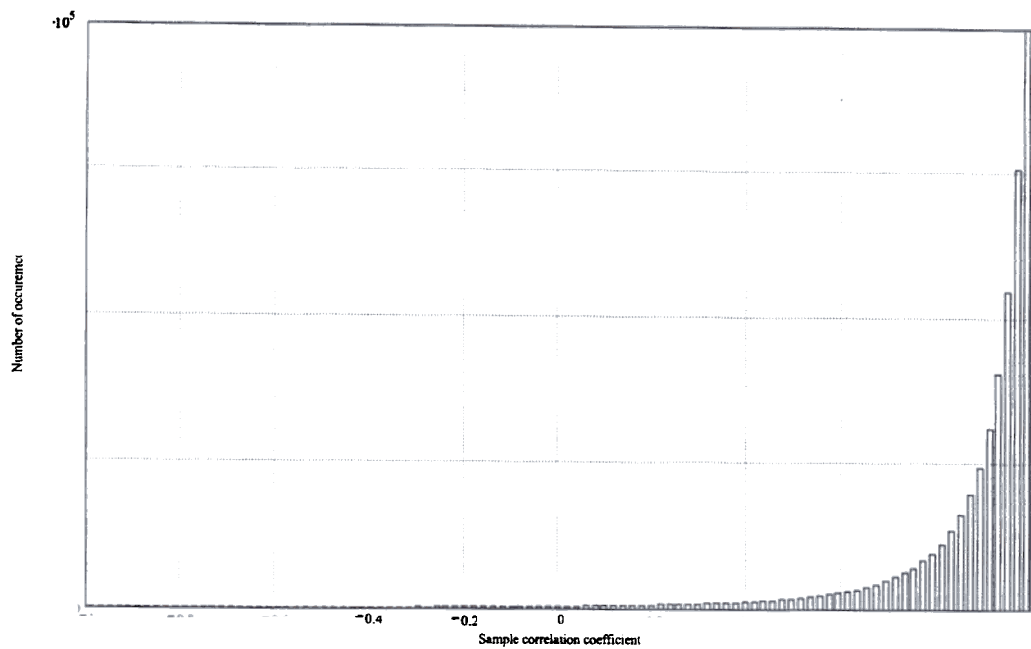


Fig 31 Distribution of Fisher's z corresponding to Fig 30

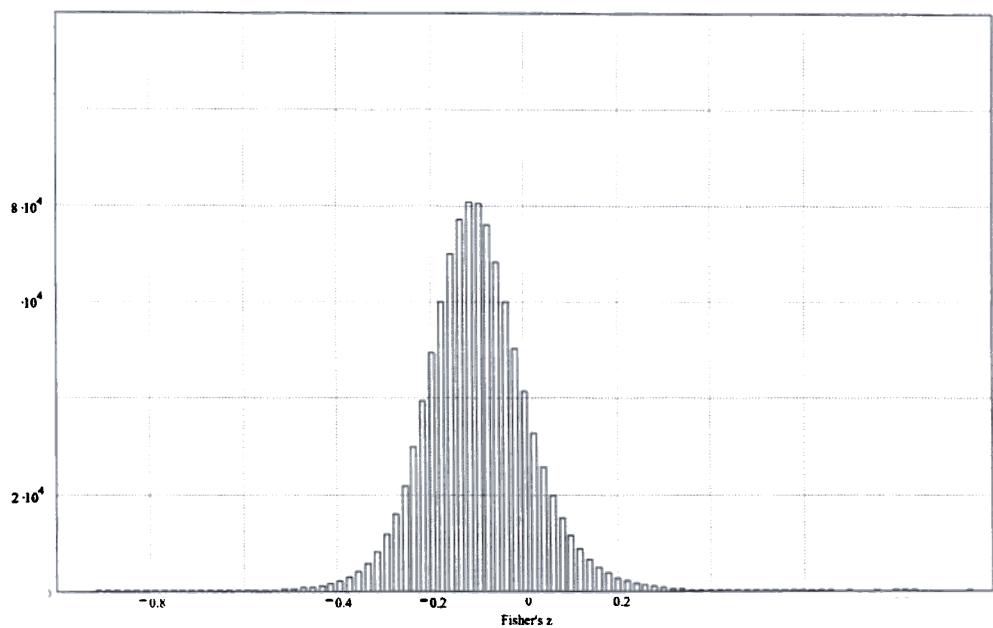


Fig 32 Distribution of sample correlation coefficients of one million random samples of size 5 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

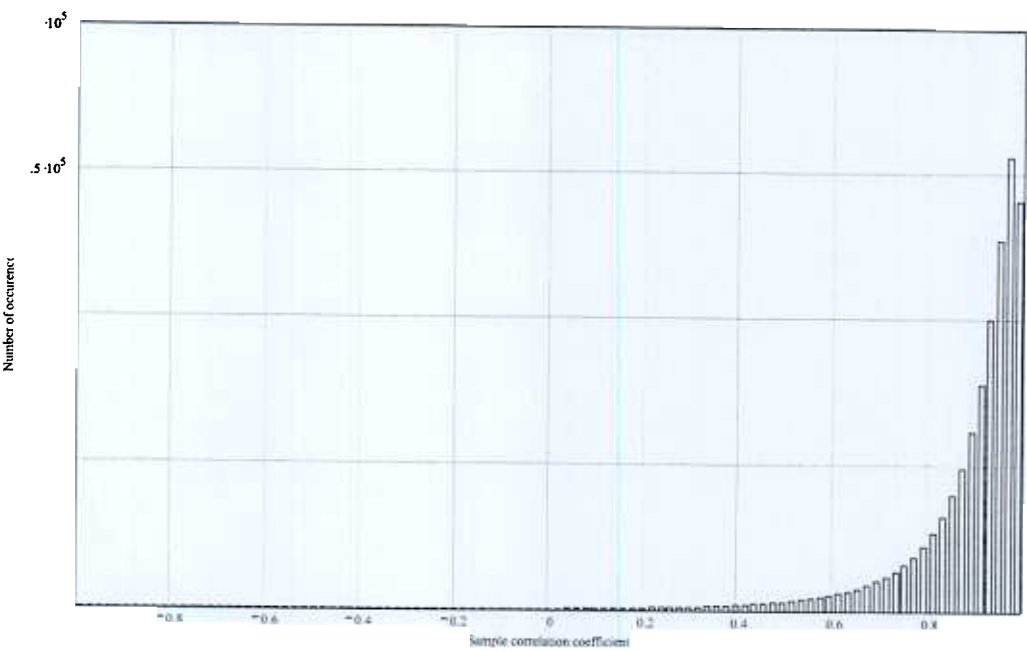


Fig 33 Distribution of Fisher's z corresponding to Fig 32

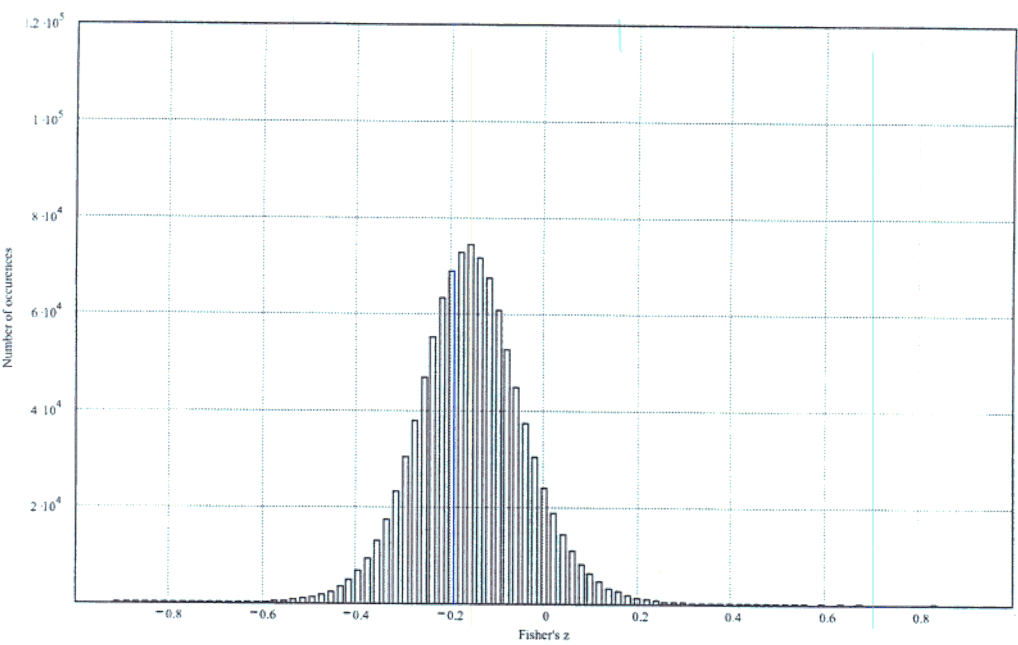


Fig 34 Distribution of sample correlation coefficients of one million random samples of size 7 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

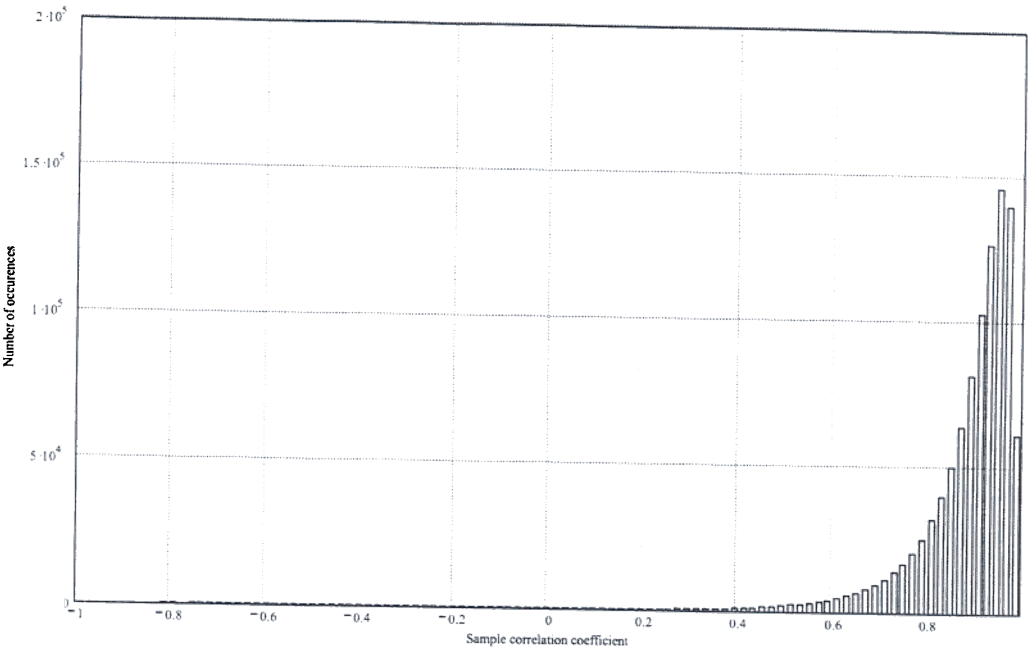


Fig 35 Distribution of Fisher's z corresponding to Fig 34

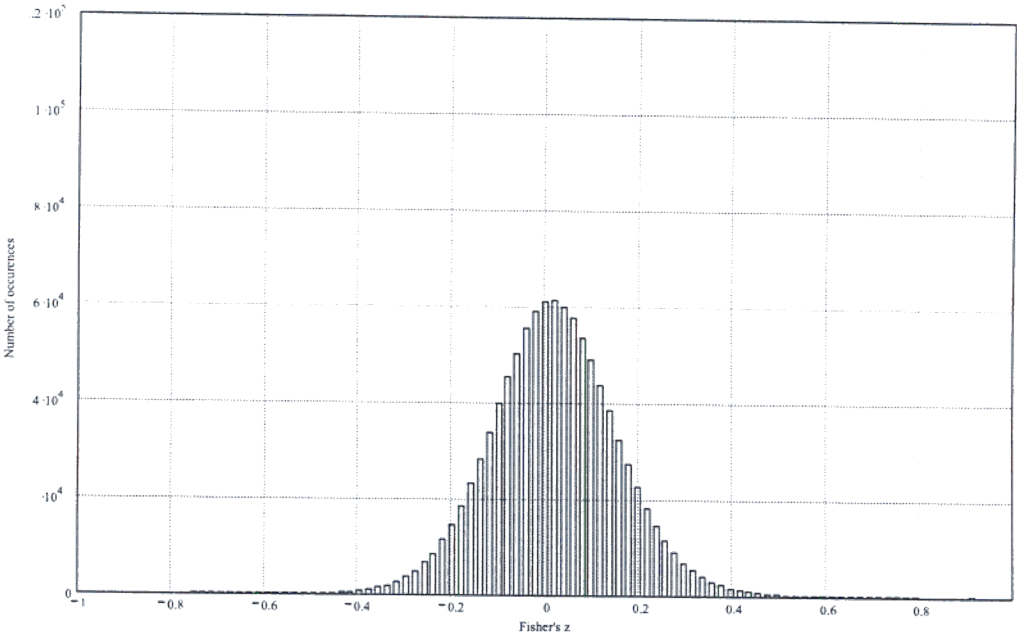


Fig 36 Distribution of sample correlation coefficients of one million random samples of size 10 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

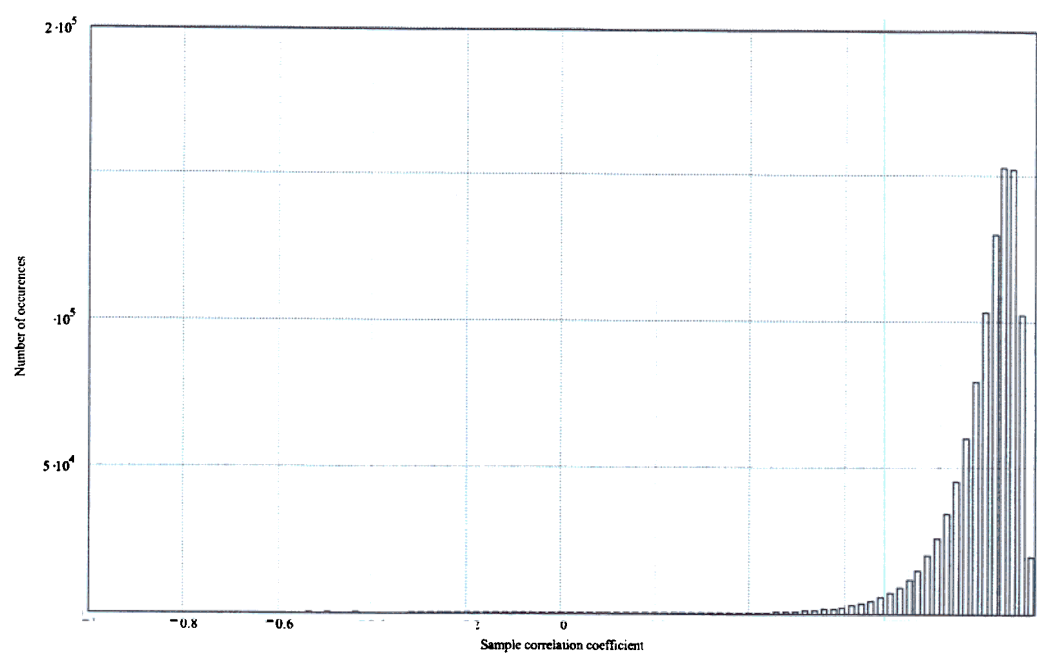


Fig 37 Distribution of Fisher's z corresponding to Fig 36

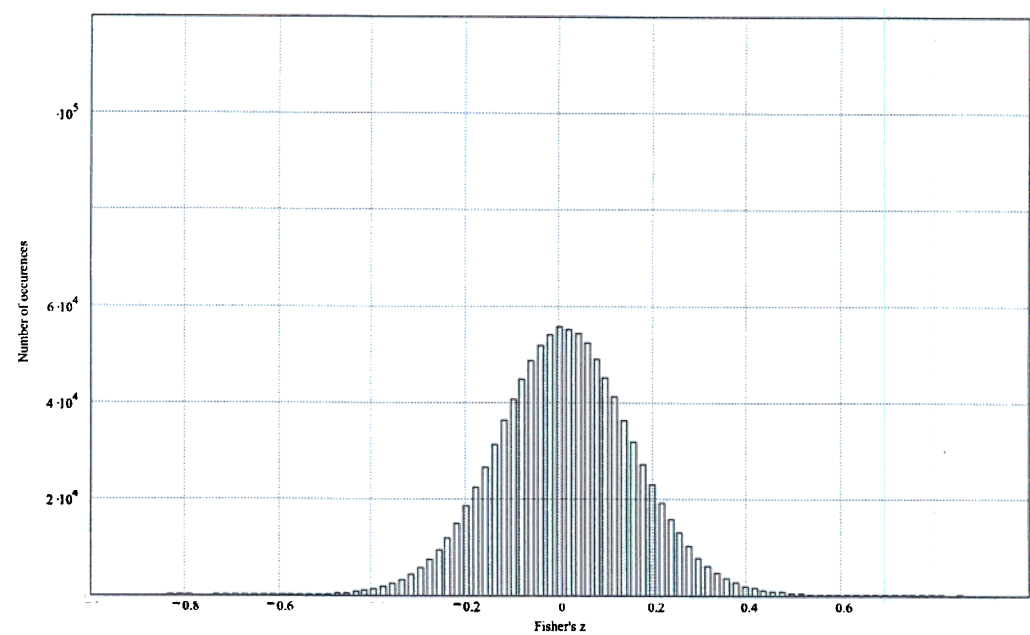


Fig 38 Distribution of sample correlation coefficients of one million random samples of size 100 from a bivariate normal distribution with population mean (0, 0), population variance (1, 1) and population correlation coefficient 0.9

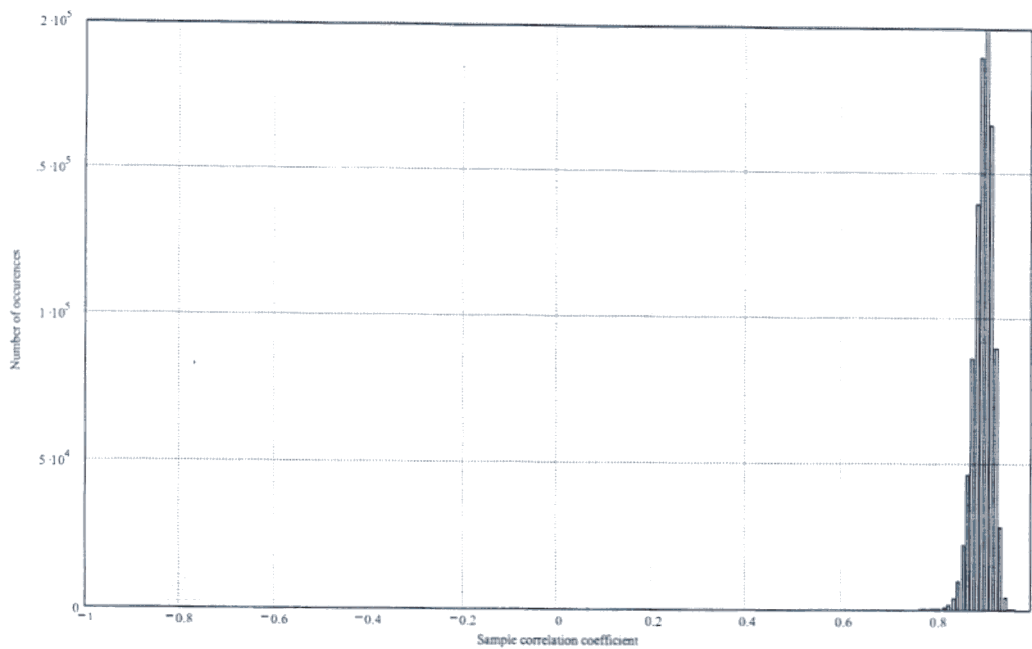


Fig 39 Distribution of Fisher's z corresponding to Fig 38

