A Notebook of Statistical and Probabilistic Formulae, Methods and Derivations

This is a somewhat idiosyncratic collection based on quantities and methods which I happened to have used over the years. In my personal experience the probabilistic material derives mostly from engineering applications whereas the statistical quantities and methods relate to the analysis of psychology data. Both are of more general applicability.

Notation: Matrices are denoted by round brackets, e.g., (M), and column matrices ("vectors") are denoted by a bar over the symbol, e.g., $\bar{v}$. Transpose is denoted by ^{T}, hence $\bar{v}^T$ is a row matrix (row "vector").

Contents

1. Definitions of Basic Terms

1.1 Probability Density Function (pdf)

The probability that a continuous variable, x, lies within the small range x to $x + dx$ is $P(x)dx$. This defines the probability density function (pdf), $P(x)$, for the variable x. A pdf necessarily integrates to unity,

$$\int_{-\infty}^{+\infty} P(x)dx = 1 \tag{1.1}$$

i.e., it is certain that the variable has some value. (The lower limit of $-\infty$ is replaced by 0 for variables which cannot be negative).

In practice, it is rarely the case that the exact underlying pdf is known. Instead a functional-form of pdf which appears qualitatively suitable is assumed, invariably including unknown parameters (mean, variance, etc.). Typically the pdf must then be estimated from a finite database of x values, $\{x_i, i\epsilon[1, N]\}$. A number of ways of "best fitting" are then available, but it is generally best to fit to the cumulative probability function (see §1.2) rather than directly to the pdf. It is important to appreciate that this procedure involves a number of different assumptions and approximations,

- The database $\{x_i, i\epsilon[1, N]\}$ must be an unbiased (representative) sample of the underlying population;
- Approximation is inherent in the assumed functional form of pdf (e.g., a normal distribution) if the underlying pdf is actually different;
- For any finite N the estimated pdf will be subject to random statistical errors (a second database of N samples will produce a different result, fractional differences typically being of order $1/\sqrt{N}$).

1.2 Cumulative Probability Function (cpf)

The probability that a continuous variable, x, takes a value less than or equal to X defines the cumulative probability function (cpf), $P_{cum}(X)$. The cpf is found in terms of the pdf from,

$$P_{cum}(X) = \int_{-\infty}^{X} P(x)dx \tag{1.2}$$

The lower limit of $-\infty$ is replaced by 0 for variables which cannot be negative. The reverse-cumulative distribution, $\tilde{P}_{cum}(X)$, is the probability that x takes a value greater than or equal to X,

$$\tilde{P}_{cum}(X) = \int_{X}^{+\infty} P(x)dx \tag{1.3}$$

so that $P_{cum}(X) + \tilde{P}_{cum}(X) = 1$.

1.3 Mean

In this notebook the word "mean" is shorthand for the arithmetic mean, and is synonymous with "average" in common parlance. It is sometimes called the "expected value" or the "expectation value". The algebraic definition of the mean of the variable x, written $\bar{x}$, whose pdf is $P(x)$, is,

$$\bar{x} = \int_{-\infty}^{+\infty} xP(x)dx \tag{1.4}$$

From here on it is to be understood that the lower integration limit of $-\infty$ is replaced by 0 for variables which cannot be negative.

For a finite sample, $\{x_i, i\epsilon[1,N]\}$, the mean, or average, is defined by,

$$\langle x\rangle = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.5}$$

Note that $\langle x\rangle$ is only an estimate of the underlying $\bar{x}$ if the database is unbiased (representative). The fractional error (i.e., $\langle x\rangle - \bar{x}$) will typically be proportional to $1/\sqrt{N}$.

1.4 Median

The median of a variable x with pdf $P(x)$ is the value x_{med} which splits the pdf into two regions whose probability is 0.5, i.e., is such that,

$$P_{cum}(x_{med}) = \tilde{P}_{cum}(x_{med}) = 0.5 \tag{1.6}$$

For a discrete set of sampled values, $\{x_i, i\epsilon[1,N]\}$, the median x_{med} is such that an equal number of sampled values are less than x_{med} as are greater than x_{med}. Strictly this definition fails, or fails to be unique, in some cases. For example the median of 1,2,3,3,4 is usually taken to be 3, whilst the median of 1,2,8,9 is usually taken to be 5.

1.5 Mode

The mode of a pdf $P(x)$ is the value of x at the maximum of $P(x)$. Hence the mode is the most probable value of x.

1.6 Standard Deviation

The standard deviation, σ_x, of a continuous random variable, x, is a measure of the spread of its pdf about its mean. It is defined as the root-mean-square (rms) of the deviation from the mean, thus,

$$\sigma_x = \sqrt{\int_{-\infty}^{+\infty}(x-\bar{x})^2 P(x)dx} \tag{1.7}$$

The standard deviation may be divergent for some pdfs. For these pdfs, referred to here as "non-σ" distributions (§2.4, §2.5 and §2.14 are examples), the standard deviation is not a suitable measure of their spread.

The standard deviation of a finite sample $\{x_i, i\epsilon[1,N]\}$ is defined by,

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \langle x\rangle)^2} \tag{1.8}$$

The same notation, σ_x, is used in both (1.7) and (1.8) where no confusion will arise, but it is important to note that (1.8) is, at best, only an approximation for (1.7). If the finite database $\{x_i, i\epsilon[1,N]\}$ is the whole population then (1.8) would be replaced by,

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \langle x\rangle)^2} \tag{1.9}$$

1.7 Variance

The variance is the square of the standard deviation, σ_x^2.

1.8 Coefficient of Variation

The coefficient of variation (CoV) is the standard deviation normalised by the mean,

$$\text{CoV} = \frac{\sigma_x}{\bar{x}} \qquad \text{or} \qquad \text{CoV} = \frac{\sigma_x}{\langle x \rangle} \tag{1.12}$$

1.9 Pearson Correlation Coefficient

The Pearson correlation coefficient between two random variables x and y, denoted C_{xy}, expresses the extent to which they are linearly related. Variables with non-zero correlation cannot be described by independent pdfs like $P_x(x)$ and $P_y(y)$. Instead they have a joint probability density function, $P(x, y)$, in which the probability of the variables being within the ranges x to $x + dx$ and y to $y + dy$ is $P(x, y)dxdy$. The correlation coefficient is defined by,

$$C_{xy} = \frac{1}{\sigma_x \sigma_y} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \bar{x})(y - \bar{y}) P(x, y)\, dxdy \tag{1.13}$$

The generalisation of (1.7) for the variance is then $\sigma_x^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x - \bar{x})^2 P(x, y) dx\, dy$.

Note that if the joint pdf is separable, i.e., if we can write $P(x, y) = P_x(x) P_y(y)$, then it follows immediately from (1.4) and (1.13) that $C_{xy} = 0$.

In practice the Pearson correlation coefficient might be estimated from a random set of pairs of data $\{(x_i, y_i), i\epsilon[1, N]\}$ in which case an estimate of the underlying correlation is,

$$C_{xy} \approx \frac{\sum_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\sum_i (x_i - \langle x \rangle)^2 \sum_i (y_i - \langle y \rangle)^2}} \tag{1.14}$$

If there is a strict deterministic linear relationship between x and y then C_{xy} will be ± 1 (the negative sign applying if the x, y graph has negative slope). If the variables have no underlying linear relationship then $C_{xy} = 0$. Intermediate values, $0 < |C_{xy}| < 1$, indicate an imperfect linear relationship.

It is important to recognise that the Pearson correlation coefficient will only identify ***linear*** relationships. For example, if the variables are deterministically related by $x - \langle x \rangle \propto (y - \langle y \rangle)^2$ they will nevertheless have $C_{xy} = 0$ (assuming a y distribution symmetrical about its mean).

1.10 Rank

If a set of values of a variable, $\{x_i, i\epsilon[1, N]\}$, are put in descending order, so that $x_1 \geq x_2 \geq x_3 \geq \cdots$ then the subscript i is the rank of the value x_i.

1.11 Spearman Correlation Coefficient

A set of pairs of values for two variables x and y can be replaced by a set of pairs of their ranks. The Spearman correlation coefficient between x and y is defined as the Pearson correlation coefficient between their ranks. The Spearman correlation coefficient is therefore a more general means of examining if the two variables have an underlying relationship which is not necessarily linear. The Spearman correlation coefficient will identify any underlying relationship in which y varies monotonically with x. ("Monotonic" means that if x increases then y increases, or stays the same).

1.12 Standard Error

The term "standard error" relates to the accuracy of estimating the mean of an underlying distribution from a small sample. Thus, if n samples are drawn from the population, the "standard error" is the error involved in estimating the true mean from the sample mean. The standard error is given by $\sigma/\sqrt{n}$ where, strictly, σ is the true (whole population) standard deviation. In practice, as this is generally unknown it is replaced by the standard deviation of the n samples.

1.13 Covariance

The covariance between two variables, say x and y, given a set of pairs of their values, $\{(x_i, y_i), i\epsilon[1, N]\}$ is $Cov(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \langle x\rangle)(y_i - \langle y\rangle)$. From (1.11) and (1.14) it follows that the Pearson correlation coefficient is $C_{xy} = \frac{Cov(x,y)}{\sigma_x\sigma_y}$. Given a set of m variables, and N data points at which all these variables are specified, both the pair-wise Pearson correlation coefficients and the pair-wise covariances form $m \times m$ matrices, the correlation or covariance matrices.

2. Some Common PDFs

2.1 The Normal (or Gaussian) PDF

$$P(x) = \frac{1}{\sigma_x\sqrt{2\pi}} exp\left\{-\frac{(x-\bar{x})^2}{2\sigma_x^2}\right\} \tag{2.1}$$

The median and mode are the same as the mean, $\bar{x}$. There is no closed-form expression for the cumulative probability function (which is also known as the error function). Table 1 lists confidence levels derived from the normal distribution, i.e., the probability of being within the stated number of standard deviations above the mean (z).

Table 1: Confidence Levels for the Normal Distribution

z	$P_{cum}(\langle x\rangle - z\sigma_x) = \tilde{P}_{cum}(\langle x\rangle + z\sigma_x)$	Confidence Level*
0	0.5	50%
1	0.1587	84.13%
1.2816	0.1	90%
1.6449	0.05	95%
2	0.0227	97.73%
2.3263	0.01	99%
3	0.00135	99.865%
4	0.000032	99.9968%

*The confidence level is defined by the single-sided value of $1 - P_{cum}(\langle x\rangle - z\sigma_x) = 1 - \tilde{P}_{cum}(\langle x\rangle + z\sigma_x)$. It differs from the confidence interval which is double-sided. For example, the 95% confidence interval would have $P_{cum}(\langle x\rangle - z\sigma_x) = \tilde{P}_{cum}(\langle x\rangle + z\sigma_x) = 0.025$and hence $z = 1.96$.

2.2 The Lognormal PDF

The lognormal pdf is (almost) obtained by replacing x in (2.1) with $ln(x)$. But this would only be normalised if we also integrated (1.1) wrt $ln(x)$, i.e., the integration measure would be $d\big(ln(x)\big) = \frac{dx}{x}$. Consequently we get a new factor of x in the denominator. Also, the lower

limit of the integral would be $ln(x) \to -\infty$, i.e., $x \to 0$, and lognormal distributions are appropriate only when the variable cannot be negative. Hence, finally,

$$P(x) = \frac{1}{sx\sqrt{2\pi}} exp\left\{-\frac{(ln(x)-\mu)^2}{2s^2}\right\} \tag{2.2}$$

where the parameters s, μ are not the standard deviation and mean of x but rather of $ln(x)$. The standard deviation, mean, median, mode and CoV of x are,

$$\sigma_x = \sqrt{exp(s^2) - 1} \cdot exp\left(\mu + \frac{1}{2}s^2\right) \tag{2.2a}$$

$$\bar{x} = exp\left(\mu + \frac{1}{2}s^2\right) \tag{2.2b}$$

$$mode_x = exp(\mu - s^2) \tag{2.2c}$$

$$x_{med} = exp(\mu) \tag{2.2d}$$

$$CoV = \sqrt{exp(s^2) - 1} \tag{2.2e}$$

2.3 Weibull PDF

The Weibull pdf is a favourite with many statistical modellers. It applies to variables which cannot be negative and is given by,

pdf: $$P(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\} \tag{2.3}$$

cpf: $$P_{cum}(x) = 1 - exp\left\{-\left(\frac{x}{\lambda}\right)^k\right\} \tag{2.3a}$$

The parameters k and λ must be non-negative, and the median, mode, mean and standard deviation are given by,

$$x_{med} = \lambda\big(ln(2)\big)^{1/k} \tag{2.3b}$$

$$mode = \lambda\left(\frac{k-1}{k}\right)^{1/k} \text{ for } k > 1 \text{ else } mode = 0 \tag{2.3c}$$

$$\bar{x} = \lambda\Gamma\left(1 + \frac{1}{k}\right) \tag{2.3d}$$

$$\sigma_x = \lambda\sqrt{\Gamma\left(1 + \frac{2}{k}\right) - \left(\Gamma\left(1 + \frac{1}{k}\right)\right)^2} \tag{2.3e}$$

where Γ is the gamma function. Special cases are,

$k = 1$ gives the simple exponential (or Boltzmann) distribution.

$k = 2$ gives the Rayleigh distribution

2.4 Student's t PDF

For any real t,

pdf: $$P(t) = A_\nu\left(1 + \frac{t^2}{\nu}\right)^{-(1+\nu)/2} \tag{2.4}$$

where the normalising factor is given in terms of the gamma function by,

$$A_\nu = \frac{\Gamma\left(\frac{1+\nu}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \tag{2.4a}$$

When ν is a positive integer it is known as the "number of degrees of freedom" as it has that interpretation in the famous "t test" for statistical significance (see §5.2).

The median and mode are zero. The mean is also zero for $\nu > 1$ but is otherwise undefined.

For $\nu > 2$ the variance is $\frac{\nu}{\nu-2}$ but is divergent/undefined for $\nu \leq 2$. Hence the Student's t-distribution is non-σ for $\nu \leq 2$ and doesn't have a finite mean either for $\nu \leq 1$.

A special case is $\nu = 1$ which gives the Cauchy distribution (see §2.5).

The more important special case is $\nu \to \infty$ for which the Student t pdf tends to the normal distribution (with zero mean and unit variance). Hence, the Student t distribution can be viewed as a generalisation of the normal distribution for a finite number of degrees of freedom.

The Student t pdf arises as the distribution of the difference between the mean of ν samples of a population and the underlying (whole population) mean, normalised by the sample standard deviation, in the case that the underlying (whole population) is normally distributed. In standard regression analysis, the Student t distribution is assumed in evaluating the p value and therefore takes into account the finite (and possibly small) number of data points being fitted. It is also the basis of the famous t-test for statistical significance of the regression coefficients (see §5.2).

2.5 Cauchy PDF

This is a non-σ pdf. In fact it does not possess a finite mean either.

$$P(x) = \frac{1}{\pi(1+x^2)} \qquad P_{cum}(x) = \frac{1}{2} + \frac{1}{\pi} tan^{-1}(x) \tag{2.5}$$

The median and the mode are both zero. Readers are urged to digest the lesson from this simple example pdf: it has neither mean nor variance and yet it is a perfectly sensible pdf.

Such pdfs are not usually used in practice simply because any finite sample, $\{x_i, i\epsilon[1,N]\}$, will always have a finite mean and finite variance – and this gives users the false impression that non-σ (and perhaps non-mean) pdfs are ruled out. But this is false logic. A larger dataset, i.e., for larger N, might give an ever increasing value for the mean and variance, so that their limit as $N \to \infty$ does not exist.

2.6 Poisson Distribution

The Poisson distribution applies to discrete variables assumed to take positive indefinite integer values, $x \in [0, +\infty]$. Strictly this means that it does not have a pdf, which applies only for continuous variables, but instead has a "probability mass function" (pmf), which is simply the probability of a given value, x. The Poisson pmf is,

$$P(x) = \frac{e^{-\bar{x}}\bar{x}^x}{x!} \tag{2.6}$$

The variance equals the mean, $\bar{x}$ (so the standard deviation is $\sqrt{\bar{x}}$, noting that this makes sense only because we are dealing with pure, dimensionless, numbers). There is no exact expression for the median but it is given approximately by,

$$x_{med} \approx \bar{x} + \frac{1}{3} - \frac{1}{50\bar{x}} \tag{2.6a}$$

2.7 Chi-Squared PDF

The χ^2-distribution is defined for positive values of the continuous variable, x, with pdf,

$$P(x) = A_{\bar{x}} x^{\frac{\bar{x}}{2}-1} e^{-\frac{x}{2}} \tag{2.7}$$

where the normalisation constant is,

$$A_{\bar{x}} = \left[2^{\frac{\bar{x}}{2}} \Gamma\left(\frac{\bar{x}}{2}\right)\right]^{-1} \tag{2.7a}$$

The variance is twice the mean, $2\bar{x}$. The mode is $max(\bar{x} - 2, 0)$. There is no exact closed-form solution for the median but an approximate expression is,

$$x_{med} = \bar{x}\left(1 - \frac{2}{9\bar{x}}\right)^3 \tag{2.7b}$$

2.8 Beta Distribution

The beta distribution is defined for a continuous variable taking values in the interval $[0,1]$. It is defined in terms of two parameters, α, β, which may take any positive values. Its pdf is,

$$P(x) = A_{\alpha,\beta} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.8}$$

The normalisation constant is a reciprocal beta function,

$$A_{\alpha,\beta} = \frac{1}{B(\alpha,\beta)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \tag{2.8a}$$

The mean, mode and standard deviation are,

$$\bar{x} = \frac{\alpha}{\alpha+\beta} \tag{2.8b}$$

$$\sigma_x = \frac{1}{\alpha+\beta}\sqrt{\frac{\alpha\beta}{1+\alpha+\beta}} \tag{2.8c}$$

For $\alpha, \beta > 1$

$$mode = \frac{\alpha-1}{\alpha+\beta-2} \tag{2.8d}$$

There is an exact expression for the median but only in terms of a generalised form of beta function. For $\alpha, \beta > 1$ an approximate expression is,

For $\alpha, \beta > 1$

$$x_{med} \approx \frac{\alpha-\frac{1}{3}}{\alpha+\beta-\frac{2}{3}} \tag{2.8e}$$

2.9 Binomial Distribution

The binomial distribution applies for a discrete variable, r, which takes positive integer values in the interval $[1, N]$. The probability of a value r occurring (i.e., the probability mass function, pmf) is,

$$P(r) = \binom{N}{r} p^r (1-p)^{N-r} \tag{2.9}$$

where $\binom{N}{r} = \frac{N!}{r!(N-r)!}$ is the binomial coefficient and $0 < p < 1$ is some parameter. It may be interpreted as the probability of r events out of N independent trials when the probability of each event is p. The mean is $\bar{r} = Np$ and the standard deviation is $\sigma_r = \sqrt{Np(1-p)}$.

For sufficiently large N the binomial distribution may be approximated by a normal distribution with the same mean and standard deviation. As the normal distribution extends to minus infinity but the binomial distribution stops at zero, this approximation is only appropriate if the estimated probability at zero is very small.

2.10 Power Law Distribution

A power law distribution has a pdf given by,

$$P(x) = Ax^k \text{ for } x \geq 0 \tag{2.10a}$$

However, this is a bad pdf to be used over the whole range $x \in [0, \infty]$ because it is not normalisable, i.e., the integral (1.1) diverges at infinity if $k > -1$ and alternatively diverges at zero if $k < -1$ (and for $k = -1$ it diverges at both). Consequently, a power law pdf makes sense only for restricted ranges, namely, for some finite X,

$$\text{For } k > -1,\ x \in [0, X] \tag{2.10b}$$

$$\text{For } k < -1,\ x \in [X, \infty] \tag{2.10c}$$

2.11 Exponential (or Boltzmann) Distribution

This is a special case of the Weibull distribution (§2.3) with $k = 1$. Its pdf is thus $\frac{1}{\bar{x}} e^{-\frac{x}{\bar{x}}}$ where $0 \leq x < \infty$.

2.12 Gamma Distribution

The gamma distribution is defined for $x > 0$. Its pdf is,

$$P(x) = A_{\alpha\beta} x^{\alpha-1} exp\{-\beta x\} \tag{2.11a}$$

for real positive parameters α, β where the normalising constant is,

$$A_{\alpha\beta} = \frac{\beta^\alpha}{\Gamma(\alpha)} \tag{2.11b}$$

and Γ is the gamma function. Its mean is $\frac{\alpha}{\beta}$, variance $\frac{\alpha}{\beta^2}$. The mode is zero for $\alpha < 1$ otherwise the mode is $\frac{\alpha-1}{\beta}$.

2.14 Lévy distribution

The Lévy distribution, like the Cauchy distribution, is non-σ, both the mean and the variance being undefined (divergent). Its pdf is defined for $x \geq 0$ by,

$$P(x) = \sqrt{\frac{c}{2\pi}} \cdot \frac{1}{(x-\mu)^{\frac{3}{2}}} exp\left\{-\frac{c}{2(x-\mu)}\right\} \tag{2.12}$$

2.15 Bilinear Distribution

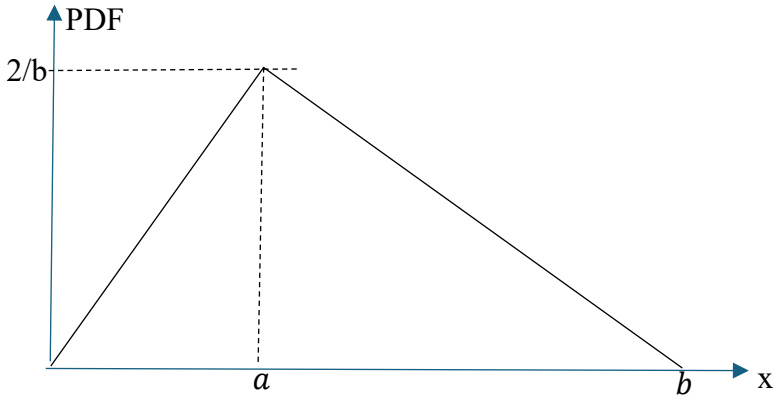
This is a pdf of my own invention, and so is not available in any proprietary software. It has the property of being precisely zero at and below $x = 0$ and also precisely zero above a

certain maximum x (i.e., it has no tail). The pdf is triangular, as shown in Figure 1. Algebraically,

$$P(x) = \frac{2}{ab}x \text{ for } 0 \le x \le a, \text{ or } P(x) = \frac{2}{(b-a)b}(b-x) \text{ for } a \le x \le b \qquad (2.13a)$$

$$\bar{x} = \frac{a+b}{3} \qquad \sigma_x^2 = \frac{1}{18}(a^2 + b^2 - ab) \qquad (2.13b)$$

Figure 1: The Bilinear Distribution



2.16 Availability of the PDFs within Standard Platforms

As of April 2024, Excel / Visual Basic (VB) includes all the above standard distributions except my bilinear distribution, the improper power law distribution and the Lévy distribution. The Cauchy distribution is a special case of the Student-t distribution. VB also includes the F distribution and the hypergeometric distribution. However I leave you to determined, for each distribution, whether the pdf is available or only the cumulative distribution (often both). Also the inverse cumulative distribution P_{cum}^{-1} is not available for all these pdfs, which makes random sampling harder work (see §3).

I expect you can find most if not all the above pdfs within Python or its ancillary facilities, e.g., Numpy Random, plus a number of others (e.g., Pareto, Zipf and Logistic distributions).

3. How to Sample a PDF

You will often wish to write code which randomly samples a given pdf. I explain here how this may be done, assuming you have available a facility to assign to a variable p a random number from a flat (uniform) distribution in the interval $[0,1)$, that is $0 \le p < 1$. In Visual Basic this may be provided by $p =$ RND().

WARNING: All random number generators are really only pseudo-random. Some are better than others. Visual Basic's RND() is not a very good one, though it will be adequate for many purposes. You may need to research better random number generators, which are certainly available in (for example) Python.

Suppose, according to the desired pdf, the probability of the distributed quantity (which we shall call x) taking a value less than or equal to X is p. Then $P_{cum}(X) = p$, by definition of the cumulative distribution. It follows that $X = P_{cum}^{-1}(p)$. This provides the desired random sample, X, of the pdf in terms of the inverse cumulative function, P_{cum}^{-1}, and a random sample of a flat distribution to give p.

Three situations then arise,

(i) If the inverse cumulative function, P_{cum}^{-1}, is known as an analytic expression, then this can be used. However, this is unusual (e.g., there is no such analytic expression for the normal pdf);

(ii) However, the inverse cumulative function, P_{cum}^{-1}, may be provided numerically by the software platform you are using. For example, Visual Basic (usually used via Excel) includes P_{cum}^{-1} for the normal distribution (NORM.S.INV), the lognormal distribution (LOGNORM.INV), the Beta distribution (BETA.INV), the Binomial distribution (BINOM.INV), the Chi-Squared distribution (CHISQ.INV), the F-distribution (F.INV), the Gamma distribution (GAMMA.INV) and the Student-t distribution (T.INV). I suspect no inverse cumulative function is provided in Visual Basic for the Weibull or Hypergeometric functions.

(iii) Finally, if the inverse cumulative function, P_{cum}^{-1}, is not available in either of the above forms then it is necessary to code it numerically yourself. Two situations arise, depending upon whether the cumulative distribution itself, P_{cum}, is available or not. If it is available, for example as an Excel worksheet function CumulativeDistribution, then possible coding in Visual Basic is as follows,

```
Dim Pcum(1400), zarray(1400)
Call StoreCumDist(zarray, Pcum)      'Needs to be run only once
x_sample = Sample(zarray, Pcum)      'Can be called repeatedly, whenever needed
(snip, other main code)
Sub StoreCumDist(zarray, Pcum)
    'Returns an array of the cumulative probability at equal z spacings (z in units of
    standard deviations)
    zmin = -7
    delz = 0.01
    z = zmin
    i = 0
    While z < 7
    Pcum(i) = Application.WorksheetFunction. CumulativeDistribution (z)
    zarray(i) = z
    z = z + delz
    i = i + 1
    Wend
End Sub
Function Sample(zarray, Pcum)
    'Returns a random sample of the distribution
    x = Rnd()       'you may want to use a better random number generator than RND
    'Find which Pcum entry is closest to x,
    i = 0
    Prob = Pcum(i)
    While Prob < x
    i = i + 1
    Prob = Pcum(i)
    Wend
    i1 = i - 1
    Prob1 = Pcum(i1)
    ic = i
```

```
        If Abs(Prob - x) > Abs(x - Prob1) Then ic = i1
        Sample = zarray(ic)
End Function
```

This code avoids having to evaluate the cumulative distribution repeatedly, in seeking its inverse, every time a sample is required. Instead 1400 values of the cumulative distribution are stored. This is likely to be more efficient if more than 1400 samples are called in the main code – and this will usually be the case in Monte Carlo simulations.

If only the pdf is available, not the cumulative distribution, then the Worksheet Function "CumulativeDistribution" must be replaced by a coded routine to evaluate the cumulative distribution (at 1400 points).

4. Monte Carlo Simulation (with Deterministic Model)

Monte Carlo simulation is an umbrella term for any code which simulates a system by random sampling. In this section we shall consider one particular class of Monte Carlo simulations which are common in engineering. For these it will be assumed that given the values of some set of independent variables the value(s) of the desired dependent variable(s) can be found by some deterministic procedure. The independent variables will, in general, be subject to uncertainty but it is assumed that pdfs for each independent variable can be estimated. The dependent, or outcome, variable(s) cannot then be determined with certainty but only their probability distributions determined. Monte Carlo simulation aims to estimate these outcome probabilities.

In engineering applications, for example, the dependent variable may be a binary outcome: failure or avoidance of failure (where "failure" may mean gross structural failure or the occurrence of cracking or some other criterion). In this case, Monte Carlo simulation provides an estimate of the probability of failure.

There are many different Monte Carlo methodologies and it is not the intention here to attempt any summary of them. My advice is to deploy only methodologies which are based upon ***equally likely*** random samples of each of the independent variables. In this case, if N_{trials} such samples are analysed and N_{fails} of these result in failure, then the estimate of failure probability is simply $P_{failure} = N_{fails}/N_{trials}$. If this simple approach is used it will be immediately clear that extremely small failure probabilities, e.g., 10^{-6}, will require a correspondingly large number of trials to be simulated. As a rule-of-thumb, N_{trials} will need to be at least $10/P_{failure}$.

To find equally likely samples, the pdf of each independent variable is divided into N_{bins} "bins", or ranges of the variable. These bins are then assigned a specific value of the variable corresponding to the centroid of the pdf over that bin. The process is illustrated for the normal distribution and the bilinear distribution below.

4.1 Equal Probability Bins for the Normal Distribution

The algorithm is formulated here in terms of a dimensionless, normalised error parameter, z. This is the error in the physical parameter (e.g., temperature, yield stress, etc.) divided by its standard deviation. Hence z is the number of standard deviations by which the quantity deviates from its mean (greater than the mean when z is positive, less than the mean when z

is negative). Hence the relevant normal probability density function (pdf) becomes the standard normal distribution (i.e., with zero mean and unit variance),

$$P(z) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{z^2}{2}\right\} \tag{4.1}$$

The cumulative probability is defined by (1.2).

The algorithm addresses N_v distributed variables, x_i, where $i \in [1, N_v]$. Each parameter takes one of N_{bins} possible values, each of which is defined by the mean of the parameter and the value taken by its error variable, z_i, for the particular random sample in question. Thus,

$$x = \langle x \rangle + \langle z \rangle * \sigma \tag{4.2}$$

where $\langle x \rangle$ is the mean of x, σ is the standard deviation of x, and $\langle z \rangle$ is one of the N_{bins} possible values of the dimensionless error parameter, z. This notation refers to the fact that $\langle z \rangle$ is the mean, or centroidal, value for one of the N_{bins} 'bins' into which the pdf has been divided. The bin ranges are as follows,

$$\text{J}^{\text{th}} \text{ Bin: } \; z \in \left[\xi_{J-1}, \xi_J\right] \tag{4.3}$$

Capital subscripts such as $_J$ will be used to denote bin numbers. The bins $z \in \left[\xi_{J-1}, \xi_J\right]$ are defined so as to represent equal probabilities. Since there are N_{bins} bins this probability must be $1/N_{bins}$. This means that,

$$P_{cum}\left(\xi_J\right) - P_{cum}(\xi_{-1}) = 1/N_{bins} \tag{4.4}$$

The left-most boundary is chosen to be $\xi_0 = -\infty$ so that $P_{cum}(\xi_0) = 0$, and hence (4.4) implies that $P_{cum}\left(\xi_J\right) = J/N_{bins}$ which allows all the bins to be found from,

$$1 \leq J \leq N_b \qquad\qquad \xi_J = P_{cum}^{-1}\left(\frac{J}{N_{bins}}\right) \tag{4.5}$$

From this it follows that $\xi_{N_{bins}} = +\infty$, so the entire parameter space from $-\infty$ to $+\infty$ is spanned by unequal sized bins with equal probabilities.

For each bin, the representative value of z, i.e., $\langle z \rangle$, must be determined. This is taken to be the mean value of z within the bin, i.e.,

Bin I:
$$\left\langle z_J \right\rangle = \frac{\int_{\xi_{J-1}}^{\xi_J} zP(z)dz}{\int_{\xi_{J-1}}^{\xi_J} P(z)dz} \tag{4.6}$$

The denominator is just equal to (4.4), whereas the numerator can be evaluated explicitly for a normal distribution, (8.1), to give,

$$\left\langle z_J \right\rangle = \frac{N_{bins}}{\sqrt{2\pi}}\left(exp\left\{-\frac{\xi_{J-1}^2}{2}\right\} - exp\left\{-\frac{\xi_J^2}{2}\right\}\right) \tag{4.7}$$

Note that the use of (4.7) is particularly important for the first and last bins since it assigns a finite mean $\langle z \rangle$ to a bin of theoretically infinite width. The values of $\langle z \rangle$ for the first and last bins define the extremes of the sampling, i.e., the minimum and maximum values. Explicitly,

$$\langle z \rangle_{min/max} = \mp \frac{N_{bins}}{\sqrt{2\pi}} exp\left\{-\frac{1}{2}\left(P_{cum}^{-1}\left(\frac{1}{N_{bins}}\right)\right)^2\right\} \tag{4.8}$$

Table 2 lists the resulting $-\langle z\rangle_{min} = \langle z\rangle_{max}$ values for a range of N_{bins} values. For example, if you wish to include samples between minus and plus five standard deviations then you need to use just over one million bins. Alternatively, if you wish to include samples between minus and plus four standard deviations you would need to use just over 10,000 million bins. Thus the extra standard deviation ($\pm 5\sigma$ cf $\pm 4\sigma$) requires one hundred times more bins.

Table 2: Numbers of Bins and Corresponding Number of Standard Deviations

Number of Bins	Number of Standard Deviations (± this number about the mean)	Probability of being outside the simulated range per variable*
10	1.75	0.079
30	2.23	0.026
100	2.67	0.0077
300	3.02	0.0025
1000	3.37	7.6×10^{-4}
3,000	3.66	2.5×10^{-4}
10,000	3.96	7.5×10^{-5}
30,000	4.21	2.5×10^{-5}
100,000	4.48	7.5×10^{-6}
300,000	4.71	2.5×10^{-6}
1,000,000	4.95	7.5×10^{-7}
10,000,000	5.38	7.5×10^{-8}

*i.e., 1 – the confidence interval

4.2 Equal Probability Bins for the Bilinear Distribution

The bilinear distribution is defined in §2.15. By randomly sampling an integer r from a flat distribution between 1 and N_{bins} a random sample of the bilinearly distributed variable, x, is provided by setting,

$$\text{If } r \le \frac{a}{b}N_{bins}: x = \frac{1}{2}\left(\sqrt{r} + \sqrt{r-1}\right)\sqrt{\frac{ab}{N_{bins}}} \qquad (4.9a)$$

$$\text{If } r > \frac{a}{b}N_{bins}: x = b - \frac{1}{2}\left(\sqrt{s} + \sqrt{s+1}\right)\sqrt{\frac{b(b-a)}{N_{bins}}} \text{ where } s = N_{bins} - r \qquad (4.9b)$$

The parameters a, b which define the pdf may be fractional. But even if they are integers, in general (4.9ab) will provide a real (fractional) sample for x. The set of N_{bins} values for x resulting from (4.9a,b) as r takes bin numbers 1 to N_{bins} are the corresponding characteristic values for that bin.

4.3 Latin Hypercube Sampling

To avoid sample bias one must ensure that every bin is sampled the same number of times, e.g., once each, and this must be the case for all the distributed independent variables. If there are N_v distributed independent variables then there are ${N_{bins}}^{N_v}$ ways of picking one bin for each variable. In practice it is wildly impracticable to calculate every one of these possible permutations. For example, using $N_{bins} = 10{,}000$ with a very modest 5 distributed independent variables would need 10^{20} trials to cover every permutation of bin choices for

all the variables. In practice it is common to have 20 or 30 distributed variables, pushing the number of trials up to 10^{80} or 10^{120}. But even 10^{20} trials is not a practical possibility. Even if the deterministic core of the code could run in as little as a nanosecond, 10^{20} trials would require 10^{11} seconds, or over 3,000 years.

Instead of aiming to cover every possible permutation we settle for a sampling algorithm which ensures that every bin of every variable is sampled exactly once – not in every combination (which would require ${N_{bins}}^{N_v}$ trials) but in the minimum number of trials for which this is possible, namely N_{bins} trials. This is accomplished using the concept of the Latin hypercube.

Consider an N_v dimensional cube, each side of which is divided into N_{bins} divisions. This hypercube thus contains ${N_{bins}}^{N_v}$ cells, one for each permutation of choice for the N_v variables. A Latin hypercube is defined as a choice of N_{bins} cells out of the possible ${N_{bins}}^{N_v}$ none of which share any row/column/rank/…. etc., covering all N_v directions. In chess terms, this corresponds to placing N_{bins} queens onto a N_v dimensional chess board in such a way that none are *en prise*.

The Latin hypercube algorithm consists of randomly selecting a Latin hypercube and then using all N_{bins} trials which the Latin hypercube represents. This approach ensures that every bin of every variable is used in just N_{bins} trials (albeit in only a very small sub-set, N_{bins}, of possible combinations). Note that this means that the number of trials equals the number of bins, N_{bins}.

The value chosen for N_{bins} determines the greatest number of standard deviations away from the mean which is sampled, according to (4.8) and Table 2. It is not possible with the Latin hypercube algorithm to sample a large number of standard deviations using only a small number of trials – because the number of trials equals the number of bins, and this would conflict with the requirement for bins of equal probability.

Note that the Latin Hypercube methodology constrains the number of bins, N_{bins} , to be the same for all of the N_v independent variables.

There is nothing to stop a simulation using two or more Latin hypercubes, say n hypercubes, so that the total number of trials would then be nN_{bins}. This will produce a better estimate than using just a single Latin hypercube. However, there is no point in doing this because it would be better – for the same total number of trials and hence computer time – to increase the number of bins to N'_{bins} where $N'_{bins} = nN_{bins}$ and use a single Latin hypercube with this larger number of bins as this would increase the range of $\langle z \rangle$ in the simulation and hence be a more reliable estimate.

4.3.1 Specimen Code to Generate a Latin Hypercube (VB)

Let Nb be the number of bins and Nv the number of distributed variables. Hence, the hypercube is of dimension Nv and each of its Nv sides is divided into Nb bins. The array LHC(c,v) defines the Latin hypercube. The possible values taken by LHC are the bin numbers, [1, Nb]. The first index, c, in LHC(c,v) is a sequential identifier, from 1 to Nb, of the occupied cells (which represent the Nb trials). The second index, v, is the variable number, from 1 to Nv. Thus, occupied cell No.1 (i.e., trial No.1) puts variable 1 in bin LHC(1,1), variable 2 in bin LHC(1,2), variable 3 in bin LHC(1,3), etc.

For every variable, v, the Nb numbers LHC(c,v), for c from 1 to Nb, is a permutation of the integers from 1 to Nb, i.e., they are all different. Hence, LHC is defined by setting each of its Nv columns to a permutation of [1,Nb]. Note that it is acceptable for different variables, v, to be assigned the same bin value in a given occupied cell. Different variables might even have the same permutation of bin values (i.e., the same bin for every occupied cell). This is an unlikely but acceptable occurrence. If all variables have the same permutation then the occupied cells of the Latin hypercube are the principal diagonal (regardless of the permutation), though this will not happen by chance for sensible numbers of bins.

It is assumed that there is available a routine "RandomPerm" which will provide a random permutation (array "perm") of the integers 1 to Nb. The following Visual Basic code will generate a random Latin hypercube,

```
Dim LHC(Nb, Nv), perm(Nb)
For v = 1 To Nv
Call RandomPerm(perm, Nb)
For c = 1 To Nb
LHC(c, v) = perm(c)
Next c
Next v
```

If not available within the code platform, "RandomPerm" can be coded in Visual Basic as follows,

```
Sub RandomPerm(perm, Nb)
Randomize
For i = 1 To Nb
perm(i) = i
Next i
For i = 1 To 3 * Nb
j = Int(1 + Nb * Rnd())
10 k = Int(1 + Nb * Rnd())
If k = j Then GoTo 10
keep = perm(j)
perm(j) = perm(k)
perm(k) = keep
Next i
End Sub
```

This coding for "RandomPerm" may not be terribly efficient and you may devise better. However, note that the efficiency of the coding for generating the Latin hypercube is quite unimportant since it is done only once.

4.4 Implementing Correlations

In general one will find that some of the independent variables are correlated. If so, it is essential to include this correlation in the simulation. Failure to do so will result in serious errors in the outcome probability. This section explains how correlations are implemented.

4.4.1 Algorithm for Two Variables

Suppose you have two distributed variables, x and y. They are assumed to have been put in standard form, with zero mean and normalised to unit variance. How can correlation between x and y with a given (Pearson) correlation coefficient, C_{xy}, be implemented?

The contention is that correlation between the two variables can be imposed by using a third variable, ξ, also with zero mean and unit variance, then sampling x and ξ independently (i.e., uncorrelated) and setting y to be,

$$y = C_{xy}x + \sqrt{1 - C_{xy}^2} \cdot \xi \tag{4.10}$$

Thus, if correlation were perfect ($C_{xy} = 1$) then (4.10) would reduce to $y = x$ as required, and perfect inverse correlation ($C_{xy} = -1$) would reduce to $y = -x$ as required. Conversely, if there were no correlation between x and y ($C_{xy} \to 0$) then (4.10) becomes $y = \xi$, i.e., a random variable completely independent of x, again as required. The proof that Equ.(4.10) imposes the desired correlation C_{xy} between *x* and *y* in the general case is given below.

Proof of the Two Variable Algorithm

The pdf of x is written $P(x)$ and the pdf of ξ is denoted $\tilde{P}(\xi)$. We must first check that (A.1) is consistent with y having a mean of zero and a variance of unity. This is proved as follows…

$$\langle y \rangle = \int yP(x)dx\,\tilde{P}(\xi)d\xi = \int \left(C_{xy}x + \sqrt{1 - C_{xy}^2} \cdot \xi\right) P(x)dx\,\tilde{P}(\xi)d\xi$$
$$= \left[C_{xy} \int xP(x)dx + \sqrt{1 - C_{xy}^2} \int \xi\tilde{P}(\xi)d\xi\right] = 0 \tag{4.11a}$$

$$\sigma_y^2 = \int (y - \langle y \rangle)^2 P(x)dx\,\tilde{P}(\xi)d\xi = \int \left(C_{xy}x + \sqrt{1 - C_{xy}^2} \cdot \xi\right)^2 P(x)dx\,\tilde{P}(\xi)d\xi$$
$$= \left[C_{xy}^2 \int x^2 P(x)dx + \left(1 - C_{xy}^2\right) \int \xi^2 \tilde{P}(\xi)d\xi + 2C_{xy}\sqrt{1 - C_{xy}^2} \int x\xi P(x)\tilde{P}(\xi)dxd\xi\right]$$
$$= \left[C_{xy}^2\sigma_x^2 + \left(1 - C_{xy}^2\right)\sigma_\xi^2 + 0\right] = 1 \tag{4.11b}$$

where we have used $\int xP(x)dx = \int \xi\tilde{P}(\xi)d\xi = 0$ (i.e., zero means), and $\int P(x)dx = \int \tilde{P}(\xi)d\xi = 1$ (total probability is unity), and $\int x^2\,P(x)dx = \sigma_x^2 = 1$ and $\int \xi^2\,P(\xi)d\xi = \sigma_\xi^2 = 1$ (unit variance), and $\int x\xi P(x)\tilde{P}(\xi)dxd\xi = 0$ (i.e., x and ξ are uncorrelated). The correlation coefficient between *x* and *y* is,

$$\int \frac{xyP(x)dx\tilde{P}(\xi)d\xi}{\sigma_x\sigma_y} = \int x\left(C_{xy}x + \sqrt{1 - C_{xy}^2} \cdot \xi\right)P(x)dx\tilde{P}(\xi)d\xi$$
$$= C_{xy} \int x^2 P(x)dx\tilde{P}(\xi)d\xi = C_{xy} \tag{4.11c}$$

as required. Note that this proof applies for any probability density functions, $P(x)$ and $\tilde{P}(\xi)$, provided only that the standard deviation exists, i.e., any "with-σ" pdfs (see §6). However the nature of the pdf's assumed for x and ξ will impose a particular pdf on y. If $P(x)$ and $\tilde{P}(\xi)$ are both normal distributions then y will also be normally distributed. Consult specialist texts for the implications of this method of imposing correlation with non-normal distributions.

4.4.2 Algorithm for Multiple Variables

More generally you may have a number of variables, $x, y, z, w. ..$, which are all mutually correlated. How is this multivariable correlation imposed in a Monte Carlo simulation? The imposition of multivariable correlation requires Cholesky decomposition of the correlation matrix. Coding Cholesky decomposition 'from scratch' is roughly as difficult as matrix inversion, so is feasible but not recommended. Users wishing to impose multivariable correlations may opt to use proprietary software which have such facilities in-built. Alternatively, using the more mathematical software platforms, such as Matlab or Python, which include Cholesky decomposition facilities, may provide the best of both worlds in terms of flexibility and ease of use.

If there are N variables which are all correlated, the pair-wise correlation coefficients can be arranged in a matrix called the "correlation matrix", (C). The element C_{ij} of this matrix is the correlation coefficient between the ith and the jth variables. Hence, the diagonal elements of (C) are all unity and the matrix is real and symmetric. A restricted form of the Cholesky decomposition theorem states that any Hermitian, positive-definite matrix, H, can be written in a unique way as the absolute matrix square of a lower triangular matrix with real, positive diagonal elements, i.e., $H = LL^{+}$ where $+$ denotes the complex conjugate transpose in the general case. Real symmetric matrices are a special case of Hermitian positive-definite matrices, and hence any correlation matrix can be written in the form LL^{T} where T denotes the transpose and L has zeros above the main diagonal. A 3×3 example illustrates this,

$$(C) = \begin{pmatrix} 1 & 0.5 & 0.3 \\ 0.5 & 1 & 0.7 \\ 0.3 & 0.7 & 1 \end{pmatrix} = LL^{T} \qquad \text{where, } L = \begin{pmatrix} 1 & 0 & 0 \\ 0.5 & 0.866 & 0 \\ 0.3 & 0.635 & 0.712 \end{pmatrix} \tag{4.12a}$$

Again we assume that all variables have been put in standard form, with zero means and unit variances. The variables which are to be correlated are $x, y, z, \ldots$. To impose this correlation, start with uncorrelated variables $\xi_1, \xi_2, \xi_3, \ldots \xi_N$, also in standard form. If L is the Cholesky "square root" of the correlation matrix, i.e., $(C) = LL^{T}$, then the desired correlated variables are obtained from the independently randomly sampled $\xi_1, \xi_2, \xi_3, \ldots \xi_N$, written as a column vector $\bar{\xi}$, as follows,

$$\begin{pmatrix} x \\ y \\ z \\ etc \end{pmatrix} = L\bar{\xi} \tag{4.12b}$$

The two-variable method of §4.4.1 is just (4.12b) in the case of a 2×2 correlation matrix because in that case we have,

$$(C) = \begin{pmatrix} 1 & C_{xy} \\ C_{xy} & 1 \end{pmatrix} = LL^{T} \qquad \text{where, } L = \begin{pmatrix} 1 & 0 \\ C_{xy} & \sqrt{1 - C_{xy}^{2}} \end{pmatrix} \tag{4.12c}$$

so that (4.12b) gives $x = \xi_1$ and $y = C_{xy}\xi_1 + \sqrt{1 - C_{xy}^{2}} \cdot \xi_2$, in agreement with (4.10). For the above 3×3 example, (4.12a), we get,

$$\begin{aligned} x &= \xi_1 \\ y &= 0.5\xi_1 + 0.866\xi_2 \\ z &= 0.3\xi_1 + 0.635\xi_2 + 0.712\xi_3 \end{aligned} \tag{4.12d}$$

The general case is now clear and very simple to implement providing that the Cholesky decomposition can be carried out. Note that to use this method in practice the variables $x, y, z\ldots$ resulting from (4.12b) will be in standard form and must be converted to the desired physical variables by multiplying by the standard deviation and adding the mean. For this reason the method can only be used for "with-σ" pdfs. Another practical issue is that the User must specify a correlation matrix,(C), which is mathematically possible, a significant constraint on the possibilities for (C). Provided that (C) is real symmetric and positive definite, the above construction guarantees that it is a valid correlation matrix, because L exists. Consequently, the test of a real symmetric matrix being a valid correlation matrix is that it is also positive definite. This is equivalent to all its eigenvalues being positive. The User should therefore test that this is the case before proceeding with the analysis

5. Multivariate Regression

Regression is essentially least-sum-of-squares fitting of a model. Suppose a quantity y is hypothesised to depend upon some set of m independent variables, $\bar{x} = \{x_j, j \in [1, m]\}$. A specific model for this dependence is hypothesised in which y is expressed as a sum over n specified functions of $\bar{x}$, each function being factored by an unknown coefficient, A_i where the index i takes values from 1 to n. Hence,

$$y \approx F(\bar{x}) = \sum_{i=1}^{n} A_i f_i(\bar{x}) \qquad (5.1)$$

Hence, y is assumed to be linearly related to the n unknown coefficients, A_i, but its dependence on the m independent variables, $\bar{x}$, can be as non-linear and as complicated as you wish. An example of a possible model is thus,

$$F(\bar{x}) = A_1 cos(x_1 + x_2) e^{2x_2} + A_2 x_1^3 log x_2$$

In contrast, if the parameter A_3 is to be found by optimising, then the following model is not suitable for regression,

$$F(\bar{x}) = A_1 cos(x_1 + x_2) e^{A_3 x_2} + A_2 x_1^3 log x_2$$

because the dependence on A_3 is not linear.

The model is to be fitted (regressed against) N data points comprising pairs of measured data for y_k corresponding to given $\bar{x}_k$, noting that the latter is m numbers for each k. The best fit is usually defined by minimising χ^2 where,

$$\chi^2 = \frac{1}{N-n} \sum_{k=1}^{N} \left(y_k - F(\bar{x}_k)\right)^2 \qquad (5.2)$$

(5.2) is (almost) the average over all data points of the squared difference between the model prediction of y and the measured y_k (called the residuals). Hence χ^2 is also (almost) the variance of the residuals. "Almost" refers to the use of a denominator of $N - n$ rather than just N, where n is the number of coefficients to be found. (The difference $N - n$ is known as the "degrees of freedom").

"Minimising" here means the smallest χ^2 that can be achieved by varying the n unknown coefficients, A_i. Hence, setting the partial derivatives of χ^2 wrt A_i to zero gives,

$$\frac{\partial \chi^2}{\partial A_i} = -\frac{2}{N} \sum_{k=1}^{N} \left(y_k - F(\bar{x}_k)\right) \frac{\partial F(\bar{x}_k)}{\partial A_i} = -\frac{2}{N} \sum_{k=1}^{N} \left(y_k - F(\bar{x}_k)\right) f_i(\bar{x}_k) = 0 \qquad (5.3)$$

Substituting (5.1) gives a matrix equation for the unknown coefficients, expressed as a column matrix, $\bar{A}$,

$$(M)\bar{A} = \bar{h} \tag{5.4a}$$

or,
$$\sum_{j=1}^{n} M_{ij}A_j = h_i \tag{5.4b}$$

where,
$$h_i = \sum_{k=1}^{N} f_i(\bar{x}_k)y_k \quad \text{and} \quad M_{ij} = \sum_{k=1}^{N} f_i(\bar{x}_k)f_j(\bar{x}_k) \tag{5.4c}$$

Hence the square matrix (M) and the column matrix $\bar{h}$ are both known in terms of the model and the data, and hence (5.4a) is solved for the unknown coefficients by inverting (M),

$$\bar{A} = (M)^{-1}\bar{h} \tag{5.5}$$

5.1 Calculating the Standard Error of the Regression Coefficients

Generally you would use a statistical package to perform the regression, and so can take it on trust that the standard error in the resulting coefficients is correctly calculated. However, you should be aware of how this is done. Here I give only the result not the proof. The standard error in the regression coefficient A_i is the square-root of χ^2 times the i^{th} diagonal element in the matrix $(M)^{-1}$, i.e.,

$$\text{Error in } A_i = \sqrt{\chi^2((M)^{-1})_{ii}} \tag{5.6}$$

The requirement for matrix inversion makes this messy to do by hand when there are more than two regression coefficients. In the special case of linear regression against a single independent variable, i.e., when the model is $y \approx F(x) \equiv A_1 + A_2x$, the errors become,

$$\text{Error in } A_2 = \sqrt{\frac{1}{N-n}\frac{\sum_{i=1}^{N}(y_i - F(x_i))^2}{\sum_{i=1}^{N}(x_i - \langle x \rangle)^2}} \tag{5.7a}$$

$$\text{Error in } A_1 = \sqrt{\sum_{i=1}^{N} x_i^2} \text{ times the error in } A_2 \tag{5.7b}$$

5.2 Significance of Regression Coefficients

Whilst (5.5) will provide the best-fit coefficients in (almost) all circumstances, this does not mean that the apparent dependence of y on (say) variable x_1 is statistically significant. Standard regression software will provide p values for each fitted coefficient, A_i. The p value is (roughly) the probability of the fitted value of A_i having arisen merely by chance. Custom and practice sets the definition of statistical significance at a p value of 5%. Hence the relationship between y and x_i implied by the fitted value of A_i is taken as statistically significant if $p < 0.05$. More generally, the coefficient is said to be "significant at the 95% level" if $p < 0.05$. Alternatively, the coefficient is said to be "significant at the 99.9% level" if $p < 0.001$.

The p value is derived by entering the two-tailed t-distribution (see §2.4) at a t statistic equal to the best estimate coefficient divided by its standard error, the latter being given by (5.5) and (5.6) respectively, and for a number of degrees of freedom set to $N - n$. There are many p value calculators based on the t-test available online, e.g., T Score to P Value Calculator - Statology.

5.3 Relationship Between a Linear Regression Coefficient and the Pearson Correlation

For linear regression against a model, $y = A_1 + A_2 x$, (5.5) gives,

$$A_2 = \frac{Cov(x,y)}{\sigma_x^2} \tag{5.8}$$

But the Pearson correlation coefficient is $C_{xy} = \frac{Cov(x,y)}{\sigma_x \sigma_y}$ and so we have,

$$C_{xy} = \frac{\sigma_x}{\sigma_y} A_2 \tag{5.9}$$

Note that this relationship between the Pearson correlation coefficient and a regression coefficient only applies when the latter is the coefficient of the linear term in a linear model. This reinforces the point made above that the Pearson correlation is specific to linear relationship.

6. Effect Size: Cohen's "d"

Consider two groups distinguished in some way by some parameter (the predictor variable). For example, two sets of patients distinguished by their receiving different treatments. Or, for my BPM Cymru charity data, two sets of non-resident fathers distinguished by having low or high domestic abuse scores (or low or high income, etc.).

Some outcome measure is performed on all subjects, e.g., whether they survive, or their Warwick-Edinburgh mental well-being score, etc. We wish to determine if the difference in the predictor variable between the two groups results in a small or large "effect", i.e., a small or large change in the outcome measure.

Let n_1 be the number of people in the first group, let $\langle x_1 \rangle$ be the mean of the measured outcome measures for group 1, and let s_1^2 be the variance of this set of outcome measures. Ditto for group 2. Cohen's d is defined as,

$$d = \frac{\langle x_1 \rangle - \langle x_2 \rangle}{s} \tag{6.1}$$

where,

$$s = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2}} \tag{6.2}$$

The usual interpretation of Cohen's d is that,

d = 0.2 indicates a small effect

d = 0.5 indicates a medium sized effect

d ≥ 0.8 indicates a large effect.

7. Significance of the Effect: Independent Samples t-Test

Whilst Cohen's d measures the size of an effect (normalised by the joint variance), the Independent Samples t-Test provides a 'p' value for the effect, i.e., it provides the probability for the null hypothesis that there is, in reality, no effect produced by the change in the predictor variable and the apparent effect is just statistical random chance.

This p is derived from the Student t-distribution by entering it at a certain t statistic. (There are many online facilities to calculate p for a given t, e.g., https://www.statology.org/t-score-

p-value-calculator/). Using the same notation as for Cohen's d, the required t statistic is defined by,

$$t = \frac{\langle x_1 \rangle - \langle x_2 \rangle}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{7.1}$$

8. Cronbach's Alpha

Cronbach's Alpha is often said to be a test of the validity of a proposed measure of a single factor "construct", such as mental well-being. Strictly it is only a test of internal consistency, which is one component of validity. (A full demonstration of validity also requires examination of the claim that the proposed measurement procedure does indeed measure what it is purported to measure).

Cronbach's Alpha is limited to constructs consisting of a single factor (so it could not be applied to "personality", since personality is a multi-factorial construct). Whether domestic abuse has been shown to be a single construct I don't know, but it seems unlikely as even coercive control is likely to be multifactorial (I'm guessing).

Envisage a measure which consists of asking k questions in a standard survey questionnaire (e.g., the 24 questions of the DV RIC). Suppose you collect this data from N subjects. Call the response to the I^{th} question by the j^{th} subject X_{jI} where $I \in [1, k]$, $j \in [1, N]$. X_{jI} is assumed to be a numerical measure. It may be binary (0 or 1 only) or take a range of numerical values.

Define σ_{IJ} as the covariance between question I and question J evaluated over all the subjects, i.e.,

$$\sigma_{IJ} = \frac{1}{N-1} \sum_{m=1}^{N} (X_{mI} - \langle X_I \rangle)\big(X_{mJ} - \langle X_I \rangle\big) \tag{8.1}$$

where the mean response to question I is $\langle X_I \rangle = \frac{1}{N} \sum_{m=1}^{N} X_{mI}$. The variance of the different subjects' answers to question I is thus,

$$\sigma_{II} = \frac{1}{N-1} \sum_{m=1}^{N} (X_{mI} - \langle X_I \rangle)^2 \tag{8.2}$$

We denote by $\overline{\sigma_{IJ}}$ the mean of the covariances (i.e., the mean of the off-diagonal elements of the covariance matrix alone, not including the variances), i.e.,

$$\overline{\sigma_{IJ}} = \frac{1}{k(k-1)} \sum_{I=1}^{k} \sum_{J \neq I}^{k} \sigma_{IJ} \tag{8.3}$$

Similarly, the mean variance is,

$$\overline{\sigma_{II}} = \frac{1}{k} \sum_{I=1}^{k} \sigma_{II} \tag{8.4}$$

Those statistics relate to the responses to individual questions and their distributions amongst the different subjects. However, the set of k questions is intended to be a measure of a single construct-factor. This single score for the whole measure (i.e., a single score for each subject) is taken to be the sum of the responses to the individual questions. Thus we define the proposed factor-measure for subject j to be,

$$Y_j = \sum_{I=1}^{k} X_{jI} \tag{8.5}$$

The variance of this total measure across the subjects is thus,

$$s_Y = \frac{1}{N-1}\sum_{j=1}^{N}\big(Y_j - \langle Y\rangle\big)^2 \tag{8.6}$$

where the mean of the Y_j is $\langle Y\rangle = \frac{1}{N}\sum_{j=1}^{N} Y_j$. An exercise is to show that s_Y can be written,

$$s_Y = \sum_{I=1}^{k}\sum_{J=1}^{k}\sigma_{IJ} \equiv \sum_{I=1}^{k}\sigma_{II} + \sum_{I=1}^{k}\sum_{J\neq I}^{k}\sigma_{IJ} \tag{8.7}$$

We may also write this as,

$$s_Y = k\overline{\sigma_{II}} + k(k-1)\overline{\sigma_{IJ}} \tag{8.8}$$

The so-called "systematic" formula for the non-tau-equivalent ρ_T, which I take here to be equivalent to – or a good enough approximation to – Cronbach's alpha, is,

$$\alpha \approx \rho_T = \frac{k^2\overline{\sigma_{IJ}}}{s_Y} \tag{8.9}$$

Using the above expression for s_Y we get, finally,

$$\alpha \approx \left[1 - \frac{1}{k} + \frac{\overline{\sigma_{II}}}{k\overline{\sigma_{IJ}}}\right]^{-1} \tag{8.10}$$

where (8.1-4) are used to evaluate (8.10). A value of α greater than 0.9 indicates excellent internal consistency (validity), whilst a value between 0.8 and 0.9 is considered good, and a value between 0.7 and 0.8 is generally regarded as adequate. Values less than 0.7 suggest the validity or internal consistency are questionable or unsubstantiated by the data analysed.

9. Principal Component Analysis (PCA)

9.1 Introduction to Principal Component Analysis (PCA) and Factor Analysis (FA)

Principal Component Analysis (PCA) and Factor Analysis (FA) are closely related ways of extracting the key features from a set of data. One considers a set of p distributed variables, each of which are been observed (or measured) n times. The archetypal situation envisaged in psychology applications is the analysis of the results of a survey. The survey consists of p questions and n different people have completed the survey. The questions are assumed to have numerical answers, which might be binary (so that yes/no becomes 0 or 1), or a Likert scale (e.g., a positive integer, 1, 2,… to some maximum like 5 or 7 or 24), or a continuous real variable. Only real data is considered here but the following analysis could easily be generalised to complex data.

Hence, the data to be analysed is given by a rectangular matrix, denoted (x), whose component x_{im} where $i \in [1,p]$ and $m \in [1,n]$ is the answer given to question i by the m^{th} person.

9.2 PCA: The Model

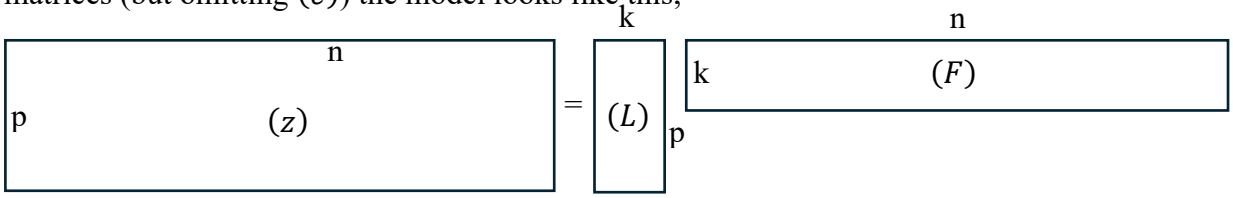
We start by replacing the raw data, i.e., the matrix (x), with a standardised form in which the new matrix (z) has the average answer to each question subtracted. Hence we define,

$$z_{im} = x_{im} - \langle x_i\rangle \qquad \text{where,} \quad \langle x_i\rangle = \frac{1}{n}\sum_{m=1}^{n} x_{im} \tag{9.1}$$

In common with most statistical techniques, the limitation of PCA (and FA) is that they are based on the assumption of an underlying ***linear*** model. In the case of PCA the model is,

$$(z) = (L)(F) + (e) \qquad (9.2)$$

Where (e) is an error matrix and $(L)(F)$ is the best fit for a model of this form where both (L) and (F) are rectangular matrices. In components, and displaying the shapes of the matrices (but omitting (e)) the model looks like this,



The labels on the sides indicate the dimensions, and the schematic illustrates the expected relative size of these dimensions, namely $n > p > k$.

We will also consider the matrix (L) to consist of k column "vectors", denoted $\bar{L}_q$ where $1 \leq q \leq k$. The i[th] component of $\bar{L}_q$ is L_{iq}.

There is an obvious redundancy in (1) in that rescaling $\bar{L}_1$ is equivalent to rescaling the first row of (F). Consequently, we are free to assume the vectors $\bar{L}_q$ are normalised to unity without loss of generality. Hence we put, for all q,

$$\left|\bar{L}_q\right|^2 = \sum_{i=1}^{p}\left(L_{iq}\right)^2 = 1 \qquad (9.3)$$

Hence, the model, Equ.(1), will reduce the np degrees of freedom of the data to $k(n + p - 1)$ parameters to be fitted.

The interpretation of the model is clarified by considering the meaning of the vectors $\bar{L}_q$ if the first row of (F) consists of all 1s, whilst the rest of (F) are zeros. In this case the RHS of (9.2), ignoring (e), is simply n copies of the column vector $\bar{L}_1$, which implies that all the people answer all the questions identically, and the answer to question i is L_{i1}. If the true situation approximated this, then we would conclude that there was one strongly dominant "factor" that determined the answers to all question, by most people, and that the set of answers which defines this "factor" is L_{i1}.

To approximate this situation we would be looking for an (F) matrix whose top row had large components but had small components elsewhere. Alternatively, if the top two rows of (F) had large components, but the rest were small, then we would conclude there were two dominant factors controlling the bulk of people's answers, and that these factors were defined by the sets of answers given by L_{i1} and L_{i2}. Etc.

A best fit to (9.2) therefore provides a set of factors which are effectively defined by the vectors $\bar{L}_q$. Thus, the components of $\bar{L}_1$ provide the "direction in question-space" of the first factor, etc. The rows of matrix (F) provide the relative importance of each of these factors.

9.3 PCA: Fitting the Model

We wish to minimise the sum-of-squared- errors, i.e., this,

$$\chi^2 = \frac{1}{np}\sum_{i=1}^{p}\sum_{m=1}^{n}\left(z_{im} - \sum_{q=1}^{k} L_{iq}F_{qm}\right)^2 \qquad (9.4)$$

This must be minimised wrt variations in both L_{iq} and F_{qm} but subject to the constraint (9.3). Hence, the object function to minimise can be taken to be,

$$f = \sum_{i=1}^{p}\sum_{m=1}^{n}\left(z_{im} - \sum_{q=1}^{k} L_{iq}F_{qm}\right)^2 + \sum_{q=1}^{k}\mu_q \sum_{i=1}^{p}\left(L_{iq}\right)^2 \qquad (9.5)$$

where μ_q are unknown Lagrange multipliers. Hence we require,

$$\frac{1}{2}\frac{\partial f}{\partial L_{iq}} = -\sum_{m=1}^{n}\left(z_{im} - L_{iq}F_{qm}\right)F_{qm} + \mu_q L_{iq} = 0 \qquad (9.6a)$$

and,

$$\frac{1}{2}\frac{\partial f}{\partial F_{qm}} = -\sum_{i=1}^{p}\left(z_{im} - L_{iq}F_{qm}\right)L_{iq} = 0 \qquad (9.6b)$$

where (9.6a) must hold for all i and q, and (9.6b) must hold for all q and m. But (9.3) means that (9.6b) gives,

$$F_{qm} = \sum_{i=1}^{p} z_{im}L_{iq} \qquad (9.7)$$

(5a) can be re-written,

$$L_{iq}\left(\sum_{m=1}^{n}\left(F_{qm}\right)^2 + \mu_q\right) = \sum_{m=1}^{n} z_{im}F_{qm} \qquad (9.8)$$

Inserting (9.7) into the RHS of (9.8) gives,

$$\sum_{m=1}^{n} z_{im}F_{qm} = \sum_{m=1}^{n} z_{im}\sum_{j=1}^{p} z_{jm}L_{jq} = \sum_{j=1}^{p} Cov_{ij}L_{jq} \qquad (9.9)$$

where,

$$Cov_{ij} = \sum_{m=1}^{n} z_{im}z_{jm} = \sum_{m=1}^{n}(x_{im} - \bar{x}_i)\left(x_{jm} - \bar{x}_j\right) \qquad (9.10)$$

is the covariance matrix of the original dataset. Using (9.7) again we find,

$$\lambda_q \equiv \sum_{m=1}^{n}\left(F_{qm}\right)^2 = \sum_{m=1}^{n}\left(\sum_{i=1}^{p} z_{im}L_{iq}\right)\left(\sum_{j=1}^{p} z_{jm}L_{jq}\right) = \sum_{i=1}^{p}\sum_{j=1}^{p} Cov_{ij}L_{iq}L_{jq} \qquad (9.11)$$

which is the qq diagonal component of the matrix $(L)^T(Cov)(L)$. Finally, then (9.8) gives,

$$L_{iq} = \frac{\left((Cov)(L)\right)_{iq}}{\mu_q + \left((L)^T(Cov)(L)\right)_{qq}} = \frac{\left((Cov)(L)\right)_{iq}}{\mu_q + \lambda_q} \qquad (9.12)$$

In terms of the column vectors, $\bar{L}_q$, this can be written,

$$(Cov)\bar{L}_q = \left(\mu_q + \lambda_q\right)\bar{L}_q \qquad (9.13)$$

Recall that this is the condition for the best fit, i.e., minimisation of χ^2, and we see now that the condition reduces to the columns of the matrix (L), i.e., the vectors $\bar{L}_q$, being the eigenvectors of the covariance matrix.

But substitution of (9.13) in $\lambda_q = \left((L)^T(Cov)(L)\right)_{qq}$ gives $\lambda_q = \left(\mu_q + \lambda_q\right)$, using (9.3).

Hence, we conclude that at the minimum the Lagrange multipliers are all zero, $\mu_q = 0$. So we replace (9.13) with,

$$(Cov)\bar{L}_q = \lambda_q\bar{L}_q \qquad (9.14)$$

This uniquely defines the matrix (L), given also the normalisation, (9.3), and ignoring the irrelevance of the order of the columns. The matrix (F) is then given by (9.7).

QED? Not quite…

(Cov) is a $p \times p$ matrix and hence has p eigenvalues/eigenvectors – but we wanted only k 'vectors' $\bar{L}_q$.

9.4 PCA: How Many Factors, k, to Include?

So far we have not specified the number of factors, i.e., the number of columns, k, in matrix (L). Indeed, this is arbitrary except that it must not exceed p. We will see that using $k = p$ will provide a perfect fit with zero errors, but is hardly of any use as it fails to "compactify" the original set of questions. The fewer the factors included, the less good the fit will be, but the more compact the description of the data. An optimal number may be sought in terms of the diminishing returns of including any additional factors. To aid judgment we need to know the contribution to the reduction of χ^2 each additional factor makes. It turns out that the answer to that is simply the eigenvalue, λ_q. Hence, the dominant factors are determined by the magnitude of their corresponding eigenvalues.

To see this, note that,

$$\chi^2 = \sum_{i=1}^{p} \sum_{m=1}^{n} \left[\left(\sum_{q=1}^{k} L_{iq} F_{qm} \right)^2 - 2 z_{im} \sum_{q=1}^{k} L_{iq} F_{qm} + z_{im}^2 \right] \quad (9.15)$$

But $$\sum_{i=1}^{p} \sum_{m=1}^{n} \left(L_{iq} F_{qm} \right)^2 = \sum_{m=1}^{n} \left(F_{qm} \right)^2 \text{ by (2)}$$

And $$\sum_{m=1}^{n} \left(F_{qm} \right)^2 = \sum_{m=1}^{n} \sum_{i=1}^{p} z_{im} L_{iq} \sum_{j=1}^{p} z_{jm} L_{jq} = \sum_{i=1}^{p} \sum_{j=1}^{p} Cov_{ij} L_{iq} L_{jq} = \lambda_q \quad (9.16)$$

But also,

$$\sum_{i=1}^{p} \sum_{m=1}^{n} z_{im} L_{iq} F_{qm} = \sum_{i=1}^{p} \sum_{m=1}^{n} z_{im} L_{iq} \sum_{j=1}^{p} z_{jm} L_{jq} = \sum_{i=1}^{p} \sum_{j=1}^{p} Cov_{ij} L_{iq} L_{jq} = \lambda_q \quad (9.17)$$

And also $\sum_{i=1}^{p} \sum_{m=1}^{n} z_{im}^2 = Tr(Cov)$, the trace of the covariance matrix. So (9.15) becomes,

$$\chi^2 = Tr(Cov) - \sum_{q=1}^{k} \lambda_q \quad (9.18)$$

But as the λ_q are the eigenvalues of the covariance matrix, and because the covariance matrix, being real symmetric, can be diagonalised with the diagonal elements being its eigenvalues, it follows that $Tr(Cov)$ is just the sum of ***all*** p of its eigenvalues. Hence,

$$\chi^2 = \sum_{q=1}^{p} \lambda_q - \sum_{q=1}^{k} \lambda_q \quad (9.19)$$

Hence, the error is zero if we allow $k = p$. Moreover, as claimed above, the reduction in the sum-over-the-squared-errors, χ^2, due to adding a further factor is just the eigenvalue of that new factor. A pragmatic optimum number of factors, k, to include might be decided on the basis of the ratio,

$$\xi = \frac{\sum_{q=1}^{k} \lambda_q}{\sum_{q=1}^{p} \lambda_q} \quad (9.20)$$

being sufficiently close to 1, e.g., 0.95, say.

10. Factor Analysis (FA)

10.1 FA: The Formulation

Factor Analysis uses essentially the same model as (9.2) except that the data is normalised by the standard deviations across people (and so now comprises the coefficients of variation),

$$\hat{z}_{im} = \frac{x_{im} - \bar{x}_i}{\sigma_i} \quad \text{where,} \quad \sigma_i = \sqrt{\frac{1}{n}\sum_{m=1}^{n}(x_{im} - \bar{x}_i)^2} \tag{10.1}$$

and the model becomes,

$$(\hat{z}) = (L)(F) + (e) \tag{10.2}$$

i.e., $$\hat{z}_{im} = \sum_{q=1}^{k} L_{iq} F_{qm} + e_{im}$$

Unlike PCA the vectors $\bar{L}_q$ are not normalised. Instead the rows of the (F) matrix are normalised. These row vectors are denoted $\bar{F}_q$, the m^{th} component of which is F_{qm}. Factor analysis makes two assumptions,

(i) The row vectors $\bar{F}_q$ are orthogonal as well as normalised, i.e., $\bar{F}_q \cdot \bar{F}_{q'} = \delta_{qq'}$. Note that this can be written $\sum_{m=1}^{n} F_{qm} F_{q'm} = \delta_{qq'}$ which also means the $\bar{F}_q$ are uncorrelated.

(ii) The row vectors $\bar{F}_q$ are also uncorrelated with the errors, i.e., $\sum_{m=1}^{n} F_{qm} e_{im} = 0$.

Theorem: $$\sum_{m=1}^{n} \hat{z}_{im}\hat{z}_{jm} = \sum_{q=1}^{k} L_{iq} L_{jq} + \sum_{m=1}^{n} e_{im} e_{jm} \tag{10.3}$$

Proof: $$\sum_{m=1}^{n} \hat{z}_{im}\hat{z}_{jm} = \sum_{m=1}^{n} \left(\sum_{q=1}^{k} L_{iq} F_{qm} + e_{im}\right)\left(\sum_{q=1}^{k} L_{jq} F_{qm} + e_{jm}\right)$$

But, using (i) above,

$$\sum_{m=1}^{n} \sum_{q=1}^{k} L_{iq} F_{qm} \sum_{q'=1}^{k} L_{jq'} F_{q'm} = \sum_{q=1}^{k} \sum_{q=1}^{k} L_{iq} L_{jq'} \delta_{qq'} = \sum_{q=1}^{k} L_{iq} L_{jq} \tag{10.4}$$

whilst the cross-terms are zero due to (ii), i.e.,

$$\sum_{m=1}^{n} \left(\sum_{q=1}^{k} L_{iq} F_{qm}\, e_{jm}\right) = 0$$

Hence we are left with (10.3). QED.

Now the covariance matrix of the errors is,

$$(Cov^e)_{ij} = \sum_{m=1}^{n} e_{im} e_{jm} \tag{10.5}$$

Whilst the correlation matrix of the data is,

$$(Cor)_{ij} = \sum_{m=1}^{n} \hat{z}_{im}\hat{z}_{jm} \tag{10.6}$$

Hence (10.3) can be written,

$$(Cov^e)_{ij} = (Cor)_{ij} - \sum_{q=1}^{k} L_{iq} L_{jq} \tag{10.7}$$

Finally, in Factor Analysis the minimum is sought of the sum-of-squares of the **off-diagonal** error covariances,

$$\varepsilon^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left((Cov^e)_{ij}\right)^2 \tag{10.8}$$

This differs from minimising χ^2, (9.4), because the latter includes the on-diagonal terms in the sum. Purists regard the minimisation of ε^2, (10.8), as being more valid for reasons I shall not attempt to reproduce. Hence, for a given correlation matrix over the data we seek to minimise, with respect to unconstrained variations in L_{iq}, the function,

$$\varepsilon^2 = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} \left((Cor)_{ij} - \sum_{q=1}^{k} L_{iq} L_{jq}\right)^2 \tag{10.9}$$

The minimisation is unconstrained because the L_{iq} are not required to be normalised. Note that, unlike the minimisation of χ^2, (9.4), variations in F_{qm} do not feature and the (F) matrix is not determined by the minimisation of ε^2, (10.9). Only the "factors", i.e., the $\bar{L}_q$, are found.

Moreover, even the factors are not uniquely determined because, if (Λ) is any $k \times k$ real orthogonal matrix and we replace $(L) \to (L)(\Lambda)$ and $(F) \to (\Lambda)^T(F)$ then $(L)(F)$ is invariant and the orthonormal condition, (i), above, is also preserved. Hence Factor Analysis only determines the factors up to an arbitrary rotation in the k-dimensional factor space. For this reason I prefer PCA to FA because the latter gives unique factors.

10.2 FA: How Many Factors to Include?

There is no standard answer to this in the literature as far as I am aware. What I suggest is judging the matter based upon the quantities,

$$\psi_q = \sum_{\substack{i,j=1 \\ i \neq j}}^{p} L_{iq} L_{jq} \tag{10.10}$$

Larger ψ_q will reduce ε^2 more and so are more significant. So, analogous to (9.20), and assuming the ψ_q are ordered in terms of decreasing magnitude, one might adopt a criterion to include k factors where k is the smallest integer such that,

$$\xi = \frac{\sum_{q=1}^{k} \psi_{q(k)}}{\sum_{q=1}^{p} \psi_{q(p)}} > 0.95 \text{ (say)} \tag{10.11}$$

where $\psi_{q(k)}$ are the values of (10.10) resulting from optimising based on k factors and $\psi_{q(p)}$ the result of optimising based on p factors. The latter would be expected to be an exact fit, so that $\varepsilon^2 = 0$ and so (27) gives the criterion as,

$$\xi = \frac{\sum_{q=1}^{k} \psi_{q(k)}}{\sum_{\substack{i,j=1 \\ i \neq j}}^{p} (Cor)_{ij}} > 0.95 \text{ (say)} \tag{10.12}$$

10.3 FA: A Solution Algorithm

Methods for solving the minimisation of ε^2, (10.9), abound in the literature. I have used my own method, based on evaluating partial derivatives numerically and then using the method of steepest descents. It appeared robust and gave exact fits when k was set to p. This is it,

[1] Choose some starting $\bar{L}_q$ (e.g., the eigenvectors of the covariance matrix, the solution for PCA);

[2] Estimate the $p \times k$ partial derivatives $\frac{\partial \varepsilon^2}{\partial L_{iq}}$ numerically using the finite differences for some small increments ΔL_{iq};

[3] Set the next iteration for the $\bar{L}_q$ using $L_{iq} \to L_{iq} - \tau \frac{\partial \varepsilon^2}{\partial L_{iq}}$ for some small τ;

[4] Repeat and expect convergence to unchanging L_{iq} as the minimum is approached (because then all the $\frac{\partial \varepsilon^2}{\partial L_{iq}}$ will be zero).

[5] Confirm robustness of result by sensitivity to change in ΔL_{iq} and τ.

11. Singular Value Decomposition (SVD)

11.1 What is SVD?

Singular Value Decomposition is not a statistical method but a method for diagonalising a matrix. A description of it is included here in order to bring out its relationship with Principal Component Analysis.

Singular Value Decomposition refers to the representation of an arbitrary complex matrix, (M), which will in general be rectangular $(m \times n)$, in the form,

$$(M) = (U)(\Sigma)(V)^{+} \tag{11.1}$$

where $^{+}$ denotes the complex transpose (Hermitian conjugate), and where (U) is an $m \times m$ unitary matrix and (V) is an $n \times n$ unitary matrix, and (Σ) is a diagonal matrix (in general rectangular, $m \times n$). A rectangular diagonal matrix is such that,

$$\Sigma_{ij} = \sigma_i \delta_{ij} \tag{11.2}$$

which applies despite the range of the two subscripts being different in general. Providing we agree on the order of the diagonal terms in (Σ), say in descending order of magnitude, then (Σ) is uniquely determined for a given (M).

(U) and (V) will, in general, be complex when (M) is complex. But if (M) is real then (U) and (V) will be real orthogonal matrices.

The decomposition, (11.1), always exists for arbitrary complex (M).

The key observation is that, for (M) an arbitrary complex matrix, $(M)(M)^{+}$ and $(M)^{+}(M)$ are both Hermitian and hence both are diagonalised by unitary matrices whose columns are their eigenvectors and whose diagonalised components are their (real) eigenvalues. But using (11.1) we have,

$$(M)(M)^{+} = (U)(\Sigma)(V)^{+}(V)(\Sigma)^{+}(U) = (U)(\Sigma)(\Sigma)^{+}(U)^{+} \tag{11.3}$$

This establishes that $(\Sigma)(\Sigma)^{+}$, which is diagonal and whose components are $|\sigma_i|^2$ (other than possible additional zeros) is also the diagonalisation of $(M)(M)^{+}$, i.e., the eigenvalues of $(M)(M)^{+}$ are $|\sigma_i|^2$. Note that $(M)(M)^{+}$ is $m \times m$ and so, if $m \leq n$, the $|\sigma_i|^2$ are its m eigenvalues. If $m > n$ then $|\sigma_i|^2$ are its first n eigenvalues, the rest being zeros.

(11.3) also establishes that the unitary matrix (U) consists of columns which are the eigenvectors of $(M)(M)^{+}$. [Note that it is $(U)^{+}(M)(M)^{+}(U)$ which is diagonal].

In the same way, we have,

$$(M)^{+}(M) = (V)(\Sigma)^{+}(U)^{+}(U)(\Sigma)(V)^{+} = (V)(\Sigma)^{+}(\Sigma)(V)^{+} \tag{11.4}$$

This establishes that $(\Sigma)^{+}(\Sigma)$, which is diagonal and whose components are $|\sigma_i|^2$ (other than possible additional zeros) is also the diagonalisation of $(M)^{+}(M)$, i.e., the eigenvalues of $(M)^{+}(M)$ are $|\sigma_i|^2$. Note that $(M)^{+}(M)$ is $n \times n$ and so, if $n \leq m$, the $|\sigma_i|^2$ are its n eigenvalues. If $n > m$ then $|\sigma_i|^2$ are its first m eigenvalues, the rest being zeros.

(4) also establishes that the unitary matrix (V) consists of columns which are the eigenvectors of $(M)^{+}(M)$.

The alert will spot that I have not proved the existence of the decomposition, (11.1), in the general case – I have merely assumed it in deriving (11.3, 11.4). For rigorous existence proofs see standard sources. However, assuming existence, the required unitary matrices, (U) and (V), are found via solving for the eigenvectors of $(M)(M)^+$ and $(M)^+(M)$.

The σ_i are known as the singular values of the matrix (M) and are uniquely determined by the matrix (other than arbitrary order, which, by convention, it is usual to put in descending order, with trailing additional zeros).

Note that solving for the eigenvalues of $(M)(M)^+$ or $(M)^+(M)$ only provides $|\sigma_i|^2$ and so does not provide the complex phase of the singular values themselves, σ_i. Even in the case of real matrices the sign of the singular values is not determined. However, as (U) and (V) are given by the eigenvector solution, the singular values follow straight away by inverting (1), i.e., $(\Sigma) = (U)^+(M)(V)$.

11.2 Relevance in Principal Component Analysis (PCA)

The factor vectors, $\bar{L}_q$, in PCA are the eigenvectors of the covariance matrix of the data. Now the data is given by the rectangular matrix (z) with components z_{im} where (typically) i runs from 1 to the number of observations per person (p), and m runs from 1 to the number of people (n). This (z) is assumed to have the means-over-people already subtracted. So the covariance matrix is,

$$(z)(z)^+ = (z)(z)^T \text{ for real data} \tag{11.5}$$

with components,

$$(Cov)_{ij} = \sum_{m=1}^{n} z_{im} z_{jm} \tag{11.6}$$

Hence, considering the singular value decomposition of the data matrix itself,

$$(z) = (U)(\Sigma)(V)^+ \tag{11.7}$$

It follows that the factor vectors, $\bar{L}_q$, of PCA are just the columns of matrix (U), and the absolute squares of the components of (Σ) are the eigenvalues of the covariance matrix whose magnitudes indicate the relative importance of the contributions of the factors to the observations.

12. Inference (Prediction from Models)

"Inference" is an umbrella term for any process of judging unobserved quantities from available observations or knowledge, including making predictions.

We have already considered two sorts of inference in sections 4 and 5. In the type of Monte Carlo simulation considered in section 4 it was assumed that a sound theoretical framework was available to calculate the outcome variable(s), e.g., time to structural failure, given specific values for certain independent variables (e.g., yield strength, UTS, operating temperatures, etc etc). The uncertainty in the outcome variable(s) was then a result purely of the uncertainties in the independent variables. Monte Carlo simulations of this type aimed to estimate the distribution of the outcome (e.g., probability of failure in a given time) given the distributions of the independent variables. But this is not the only sort of Monte Carlo simulation.

In section 5 we considered the case where the dependence of the outcome variable(s) on the independent variables is not known via some theoretical framework. Instead it is hypothesised that the outcome depends upon certain functions, $f_i(\bar{x})$, of the independent variables, $\bar{x}$, but the relative weighting of these different functions in determining the outcome is unknown. The outcome is then expressed as $\sum_{i=1}^{n} A_i f_i(\bar{x})$ and hence is linearly dependent on the unknown coefficients A_i which are to be determined by the regression procedure. The particular algebraic form of $\sum_{i=1}^{n} A_i f_i(\bar{x})$ is often referred to as "the model". A special case is the assumption of linear dependence on all the independent variables, $\sum_{j=1}^{m} A_j x_j$.

A generalisation of regression is to consider models which have some non-linear dependence on the unknown coefficients, as well as on the independent variables, $F(\bar{A}, \bar{x})$. There is no numerical algorithm that guarantees to find the optimum fit to such a non-linear function of the unknown coefficients, $\bar{A}$. However, various software platforms offer solution facilities which may (or may not) be effective. The generic problem is that algorithms tend to get stuck in local minima rather than finding the global minimum. An example of an algorithm with this shortcoming was given in §10.3, essentially the method of steepest descents. Although this appears to work well in the case of Factor Analysis there is no guarantee that it will perform well on other problems.

12.1 Maximum Likelihood

There is, however, another situation in which inference is required. This arises when we do not even have any knowledge of any independent variables upon which the outcome variables may depend, and perhaps not even a theoretical framework that would be a guide to what factors might determine the outcome. In such cases we may only have knowledge of the history of previous outcomes. This is best illustrated by a specific example.

Consider a set of N nominal similar structures or components. Over a number of years of operation there have been a few failures. These failures are known to result by first initiating a crack in service which then grows in service to failure. Periodic inspections are carried out and these have revealed some initiated cracks from time to time, these being caught before they could grow to failure. The challenge is to estimate the probability of a failure in some future operating period given only the history of failures and discovered cracks but without any quantifiable structural mechanism.

There may be a temptation to attempt fitting a model using just the data of failures and discovered cracks. But this would be to overlook that the most significant constraint on the problem is that the bulk of components have operated successfully without cracking or failure. For example, suppose there were 100 components, each of which has been inspected, on average, three times over the life of the plant to date (say 30 years). In that time there have been 2 failures and 4 discovered cracks. This compares with circa 300 inspections, 296 of which were crack-free. To produce a Monte Carlo simulation in order to predict the probability of future failures, one approach is to assume some pdf for the service time required to initiate a crack, and another pdf for the additional service period to grow a just-detectable crack to failure. These pdfs will include some unknown parameters (see the examples in section 2). Call these A_i. The challenge is to find the values of these parameters which best fits the history, where the history comprises,

(i) In each year, the components which failed;
(ii) In each year, the components which were discovered to be cracked;
(iii) In each year, the components which were inspected and shown to be uncracked;

A Monte Carlo code can be written which, for any given values of the coefficients, A_i, will, by random sampling of the pdfs, "predict" (i.e., postdict) (i), (ii) and (iii). Unless you are lucky the chances are that the "predicted" outcomes for (i), (ii) and (iii) will not agree with the actual history. However, some will and these successful trials are what are sought. Running, say, 10,000 trials may only produce a handful of successful trials for a given set of coefficients, A_i. However, a different set of coefficients, A_i, may produce a larger number of successful trials. The larger the proportion of trials which are successful, the greater the likelihood that the chosen set of coefficients, A_i, are optimal. This defines the term "likelihood".

Suppose the actual historical record is denoted symbolically by H. This denotes the observed facts. In Bayesian terminology the likelihood is the conditional probability that the history is reproduced given the set of coefficients, A_i. This is denoted $P(H|A_i)$, the probability of H given A_i.

The Maximum Likelihood method of inference seeks to determine the coefficients, A_i, by maximising $P(H|A_i)$, i.e., maximising the proportion of successful trials.

In many situations this is a perfectly decent method. However, there are situations in which this might be misleading.

12.2 Bayesian Inference

A moment's thought shows that, in principle, $P(H|A_i)$ is not what we want to maximise. The reason is simple: it is not actually the coefficients, A_i, that we are given but the history, H. We really want to maximise the probability of the coefficients, A_i, given the history, H. That is, we should ideally be maximising $P(A_i|H)$. The two conditional probabilities are related by Bayes Theorem, which tells us that,

$$P(A_i|H) = \frac{P(H|A_i)P(A_i)}{P(H)} \tag{12.1}$$

In Bayesian speak, whilst $P(H|A_i)$ is the "likelihood", $P(A_i|H)$ is known as the "posterior probability". In these general terms, H would be called the "evidence" whilst the A_i constitute an "hypothesis". The denominator in (12.1), $P(H)$, just acts as a normalising factor. This can be seen because $P(H) = \int P(H|A_i)P(A_i)dA_i$ so that (12.1) becomes,

$$P(A_i|H) = \frac{P(H|A_i)P(A_i)}{\int P(H|A_i)P(A_i)dA_i} \tag{12.2}$$

which is now independent of $P(H)$.

The key difference between Bayesian inference using (12.1) and the method of Maximum Likelihood is the presence of $P(A_i)$, known as "the prior" – that is, the prior probability of the set of coefficients, A_i, before the evidence (data), H, is taken into account.

The importance of the prior in practical applications may be illustrated by a medical example. The evidence, H, may be a positive test result whilst the unknown, represented by A_i, is actually having the medical condition. The "likelihood", $P(H|A_i)$, is therefore the probability of a positive test result given that you have the condition (called the "sensitivity", or true

positive rate, of the test). In contrast, $P(A_i|H)$ is the probability of having the medical condition given a positive test result. This is what you, as the patient, really want to know if you have just had a positive test result. The public will tend to confuse $P(A_i|H)$ with $P(H|A_i)$. If the sensitivity of the test is, say, 80% this leads people to conclude, falsely, that they have an 80% chance of having the condition.

But what if the prevalence of the disease among the public is such that the probability of you having the condition before the test was performed, i.e., the prior, $P(A_i)$, was only 1%? If we also assume that the specificity of the test, i.e., the true negative rate, is (say) 90%, the all-important posterior probability (that you really have the condition) is actually only 7.5%, a rather more comforting conclusion.

Warning: In real medical circumstances the above is likely to be wrong! This is because, if you have been given the test, it is likely to be because you have some symptoms and perhaps some familial background or lifestyle issue which your doctor has judged may indicate the condition. So the "prior" for you will be greater than the 1% applicable to a random member of the public. But what is it? In practice, the devil is in the prior when applying Bayesian inference.

How does the Maximum Likelihood method of §12.1 get around this? In truth it doesn't – it just glosses over it. One might be tempted to consider Maximum Likelihood to result from the more rigorous Bayesian approach when the prior is a flat distribution, i.e., when there is no reason before the data was collected (or the history known) to prefer one set of coefficients, A_i, to another. This is partly valid but does not really avoid an issue of principle: there is no such thing as a flat distribution over an infinite domain. A distribution can be flat only if confined between some finite minimum and some finite maximum. And so, even for a flat distribution, the "prior" still exists as the specification of these finite bounds. This was glossed over in the presentation of Maximum Likelihood but in practice the numerical determination of the likelihood can only be done for a finite range of A_i. Credibility is achieved by showing that the likelihood that results at the extremities of the range of A_i considered is small compared with the likelihood identified at the candidate optimal solution. However, there is then an act of faith that an improved solution (a greater likelihood) could not be found outside the range of A_i considered.

Bayesian inference, and more sophisticated inference techniques, have a large and sophisticated literature. But here we would enter the territory of machine learning, a discipline which has undergone a revolution in the last few years but is beyond the scope of this short note.