Correlations

What They Mean – And More Importantly What They Don't Mean

Colin Priest BEc FIAA
Colin_Priest@sigmaplus.com.au

Sigma Plus Consulting
www.sigmaplus.com.au

Introduction

Australian actuaries are increasingly using correlations to measure the relationships between different risks. General insurance actuaries use correlation matrices to allow for diversification of risk. Investment practitioners use autocorrelation models and correlation matrices to describe financial time series, in order to set strategic asset allocations and to price options. Life insurance actuaries allow for correlations between joint lives. But do actuaries understand the implicit assumptions behind the use of correlation coefficients? When is a correlation coefficient significant? When is a correlation measure appropriate? This paper looks at the historical catastrophe event experience in Australia and uses correlation techniques to analyse the relationship between different catastrophe types, and between consecutive years. It extends the correlation concept to introduce copulas and considers the usefulness of tail correlation measures for actuaries. On the way it looks at ways to relax the implicit assumptions that underlie correlation coefficients.

Catastrophe Events in Australia

The Insurance Council of Australia (ICA) publishes a list of large events, and the estimated cost to insurers of each event. Some of these events are large and well known, such as Cyclone Tracy costing $837m in 2001 dollars, while others can cost the industry less than $2m. Each event is advised with an event date, a short description, an original cost, and a cost in 2001 dollars.

From the descriptions I have chosen to categorise the events into one (or more) of the following:

- Fire
- Hail
- Earthquake
- Cyclone
- Flood
- Storm

Unfortunately these categories are fuzzy – some events do not fit into a single category. There is some overlap between hail and storm, between cyclone and flood and between flood and storm. For the purposes of this paper I have spread some events over more than one category.

I have summarised the data to annual totals. These are shown in Appendix A at the end of this paper.

# No Frills Correlation – The Multivariate Normal

Consider two unit normal distributions with a correlation coefficient of 80%. This is a simple example of a multivariate normal distribution. This distribution has a three dimensional density function as shown in Figures 1 and 2.

Figure 1: Multivariate Normal Density with 0% Correlation
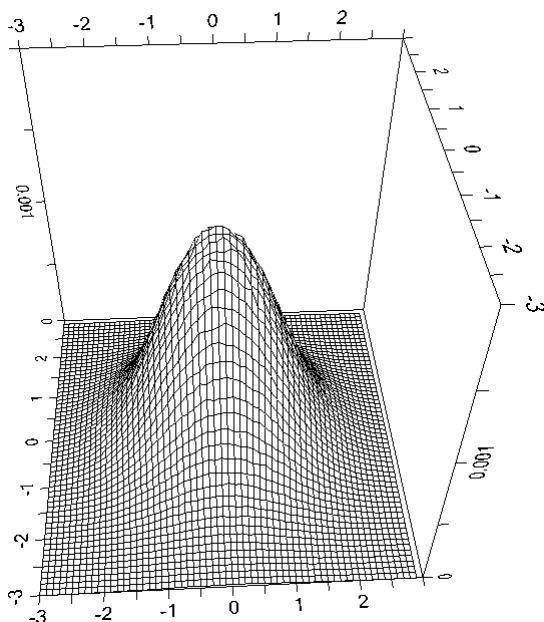
A multivariate normal distribution with 0% correlation is no more than two normal distributions that are statistically independent.

Figure 2: Multivariate Normal Density with 80% Correlation
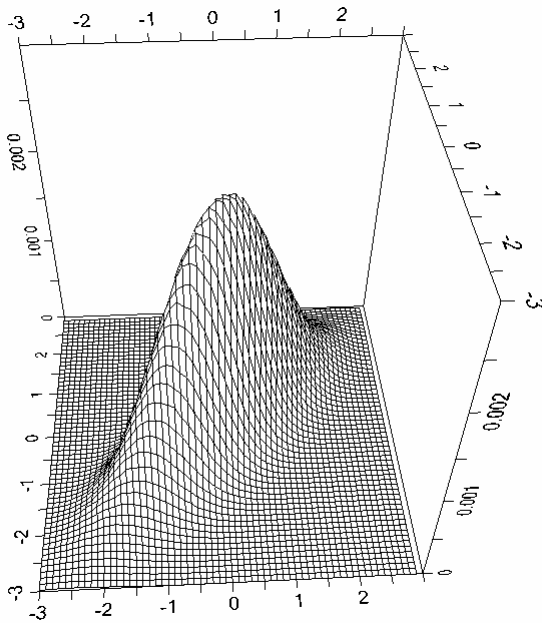
A multivariate unit normal distribution of two values (M_1 and M_2) with 80% correlation can be thought of as the linear combination of two statistically independent normal distributions (N_1 and N_2) with the relationship described in Equation 1.

Equation 1: Multivariate Normal with Correlation Coefficient of 0.8

$M_1 = N_1$
$M_2 = 0.8 * N_1 + Sqrt(1.0 - 0.8 * 0.8) * N_2$

Equation 1 can be generalised to any multivariate unit normal distribution with two values and a correlation coefficient of ρ

Equation 2: Multivariate Normal with Correlation Coefficient of ρ

$M_1 = N_1$
$M_2 = \rho * N_1 + Sqrt(1.0 - \rho * \rho) * N_2$

# The Linearity Assumption

Equation 2 can be rephrased to a more recognisable form:

$Y = a * X + e$ where $e \sim N(0, 1)$

In other words – multivariate normal correlations assume a linear relationship between the variables, with some sample variability around the best fit line. The amount of variability around the straight line is directly related to the correlation coefficient.
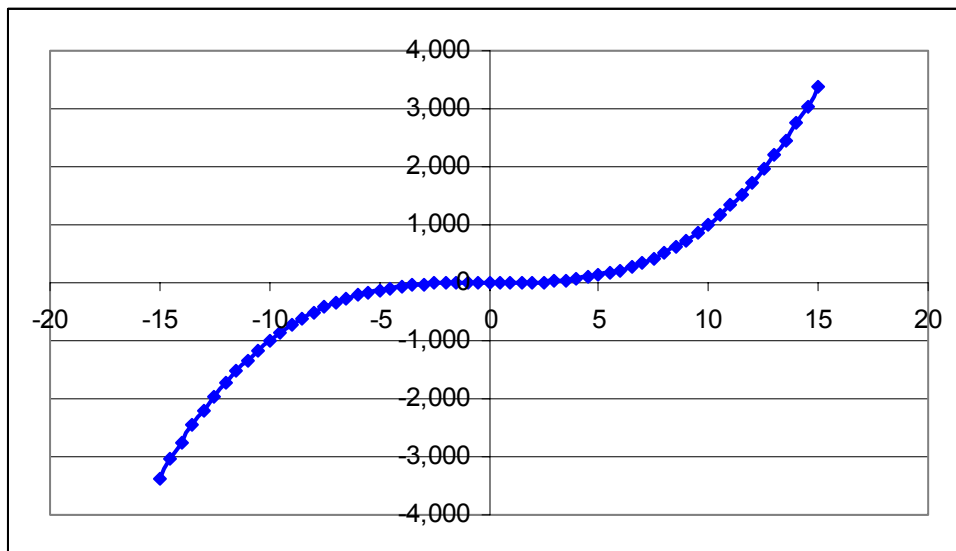
Graph 1: Multivariate Normal with 80% Correlation
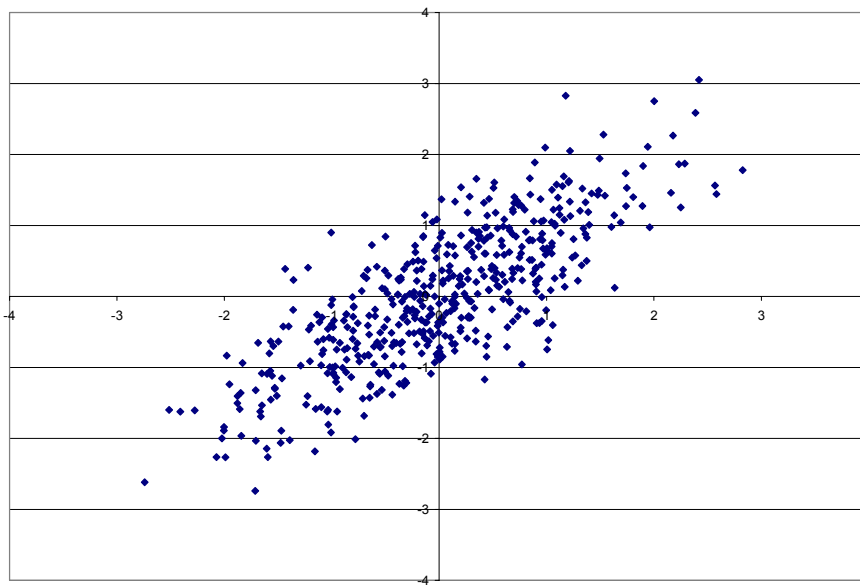
In fact, the square of the correlation coefficient is the same as the r^2 measure in linear regression.
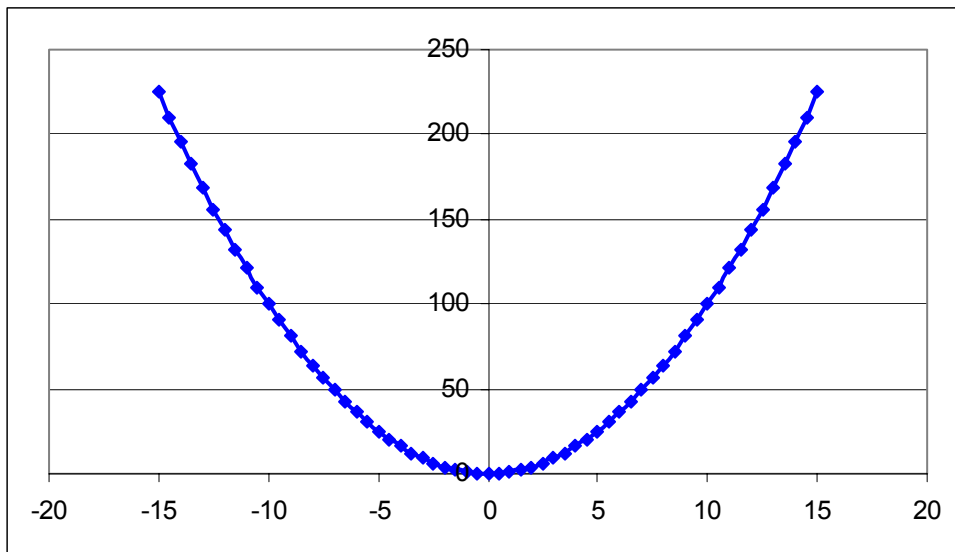
This means that a non-linear relationship will give lower correlations.

Graph 2: A Cubic Relationship

The line in Graph 2 represents the relationship $Y = X^3$, so for every value of X one knows the exact value of Y, and vice-versa. The two numbers have an exact relationship. Yet the correlation coefficient for this range is only 91%.

Graph 3: A Parabola

The relationship in Graph 3 is that $Y = X^2$, so every Y value is the square of the X value. The two numbers have an exact relationship. Yet the correlation coefficient for this range is 0%. How can the correlation coefficient be zero if there is a real relationship? The correlation is zero because the relationship cannot be approximated by a straight line.

The relationship in Graph 2 is just as exact as the relationship in Graph 3, yet the Graph 2 data has 91% correlation and the Graph 3 data has 0% correlation. Such a discrepancy occurs because the correlation coefficient is measuring how well a straight line can approximate the relationship.

RULE: In the same way that a straight line cannot fully describe a non-linear relationship, a correlation coefficient cannot fully describe a non-linear relationship. Always look at the relationship shape before using a correlation coefficient to describe it, and if the relationship is non-linear, then try to find a transformation that makes the relationship linear.

Graph 4: Annual Storm Costs vs. Annual Cyclone Costs
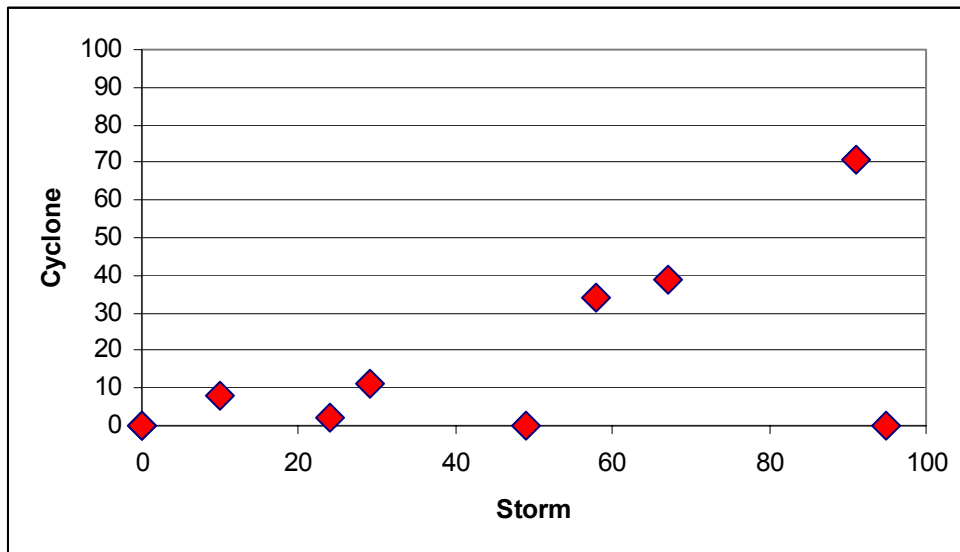
Graph 4 shows the annual cost to insurers of storms versus cyclones. The relationship can be approximated by a straight line. So we do not need to transform the data.

The Normality Assumption

Graph 5: A Normal Distribution
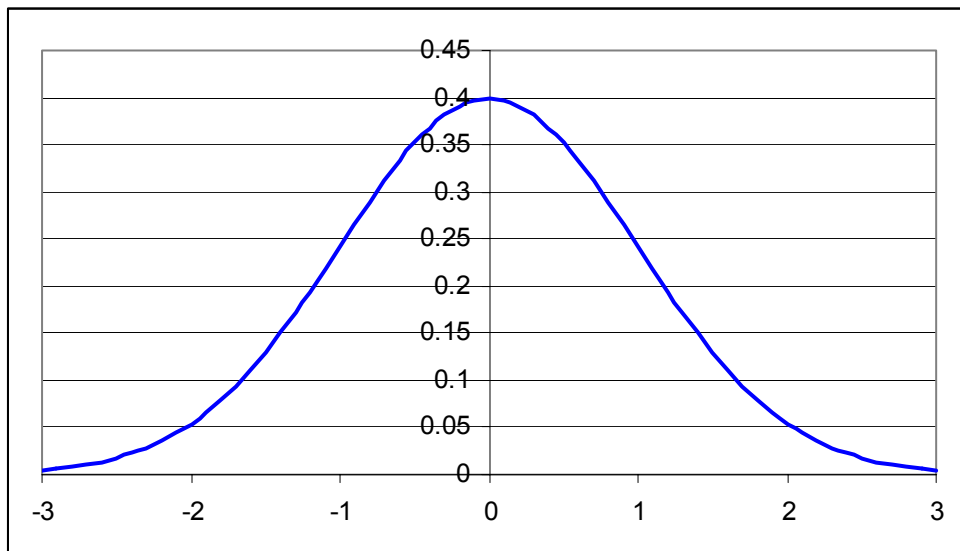
A normal distribution is:

- Symmetrical: the same shape either side of the mean
- Continuous: all values (from the real number set) are possible
- Infinite: all values (from the real number set) are possible

It is these characteristics that make a correlation coefficient a full descriptor of the relationship within a multivariate normal. Any normal distribution can be expressed as a linear function of any other normal distribution.

Asymmetry
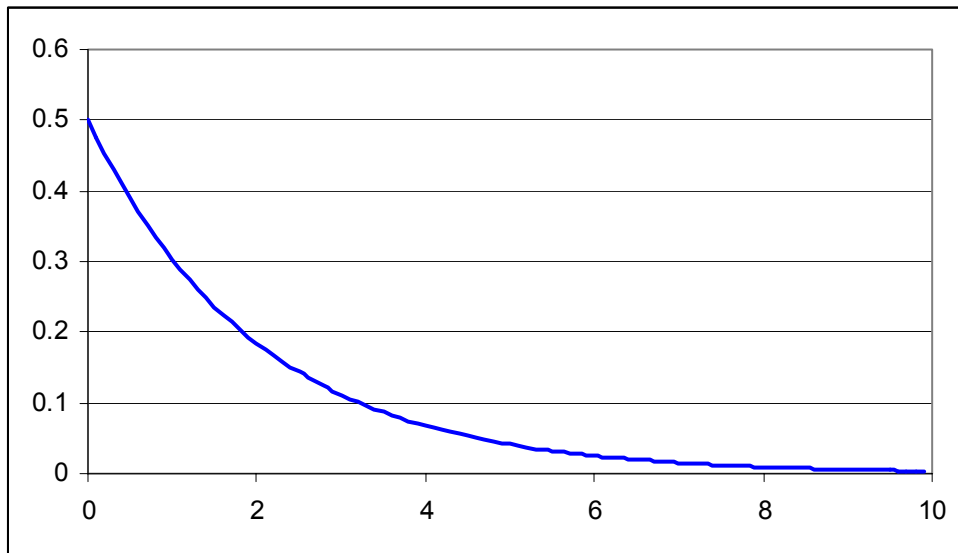
Graph 6: Exponential Distribution

Most statistical distributions are not symmetrical. For example, the exponential distribution is always decreasing. One can measure the level of asymmetry using the skewness measure of the data.

Graph 7: Distribution of Annual Cost of Storms ($2001)
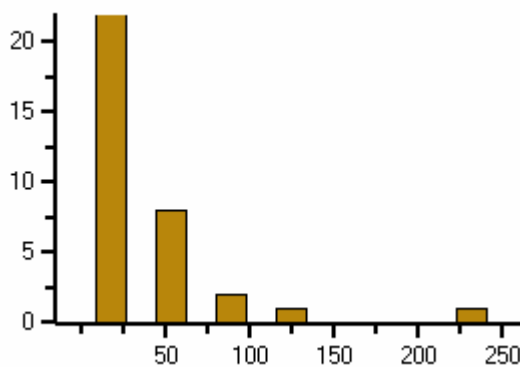
For each catastrophe type the annual cost distribution is highly skewed. For storms the distribution has a skewness measure of 2.4 which is fairly high. For earthquakes the situation is quite extreme. In most years no earthquakes occur, and the cost for those years is zero. But when an earthquake does occur, it can be quite destructive. This can be seen in Graph 8, the annual distribution of earthquake costs, which has a skewness of 5.3 and kurtosis of 27.1

Graph 8: Distribution of Annual Cost of Earthquakes ($2001)
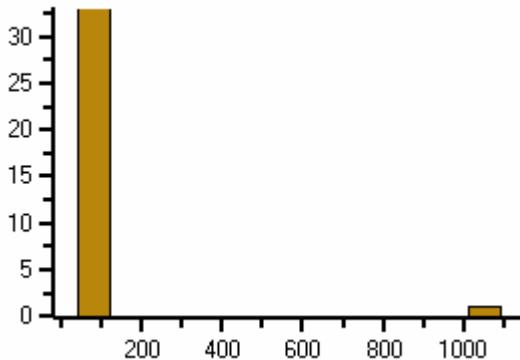
When one correlates two skewed distributions, one is assigning variable strengths of relationship at different parts of the distribution. The linear relationship at the extreme of the distribution will be stronger than the linear relationship at the central, compressed area of the distribution. In a normal distribution the linear relationship is equal throughout the ranges of the distribution.

RULE: When dealing with skewed relationships, consider where you want the strongest relationship to be – in the central values or at the extremes. If you want a strong relationship between central values, then you will need to consider more complex relationships than are described by just a correlation coefficient.

The situation becomes more difficult when relating two distributions with different amounts of skewness i.e. different shapes. It is impossible to have a 100% correlation coefficient between two distributions with different shapes, because if one cannot achieve a linear transformation from one shape to another, then one cannot achieve 100% correlation. So depending upon the level of difference in distribution shapes, some ranges of correlation are impossible.

A more meaningful measure in circumstances of skewed and different shaped distributions may be to compare the measured correlation against the strongest possible correlation for the two distributions.

Another option is to transform the data to give it a symmetrical shape.

Discrete Distributions

Many statistical distributions have discrete values. For example, the number of catastrophe events per annum must be a counting number (a non-negative integer). Distributions with discrete values are rarely able to be expressed as a linear relationship.

Graph 9: Annual Number of Cyclones vs. Storms

For example, one cannot express the number of storms as a linear relationship of the number of cyclones because each number must be a whole number.

Since an exact linear relationship is usually impossible with discrete variables, a correlation coefficient of 100% is usually impossible with discrete variables.
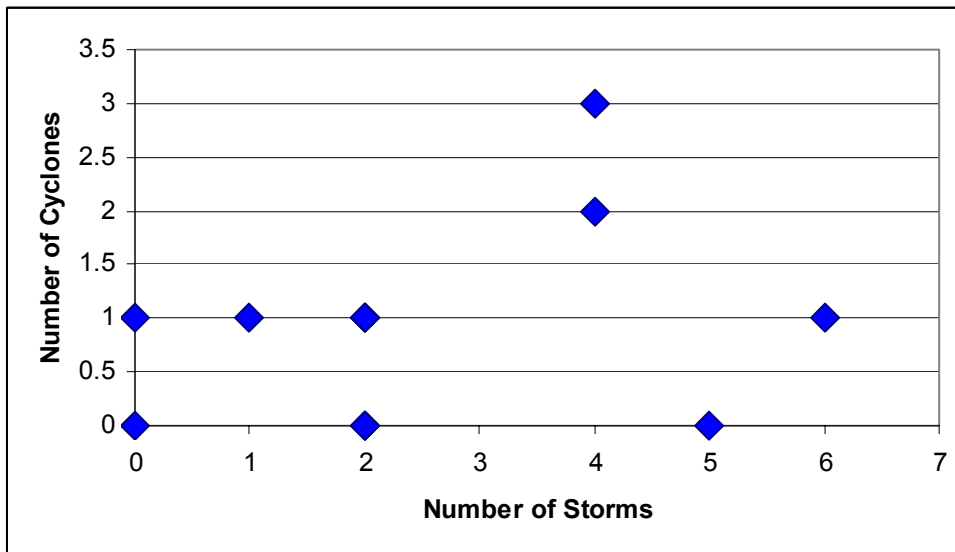
A more meaningful measure in circumstances of discrete distributions may be to compare the measured correlation against the strongest possible correlation for the two distributions.

Whatever transformation one makes to a discrete distribution, it will always remain discrete.

RULE: Always consider how discrete values make the correlation coefficient lower, and hide the strength of a relationship.

Finite Ranges

Normal distributions can range from negative infinity to positive infinity. So any linear combination of normal distributions produces a value that lies on another normal distribution.

Graph 10: A Log Normal Distribution Cannot Take Negative Values
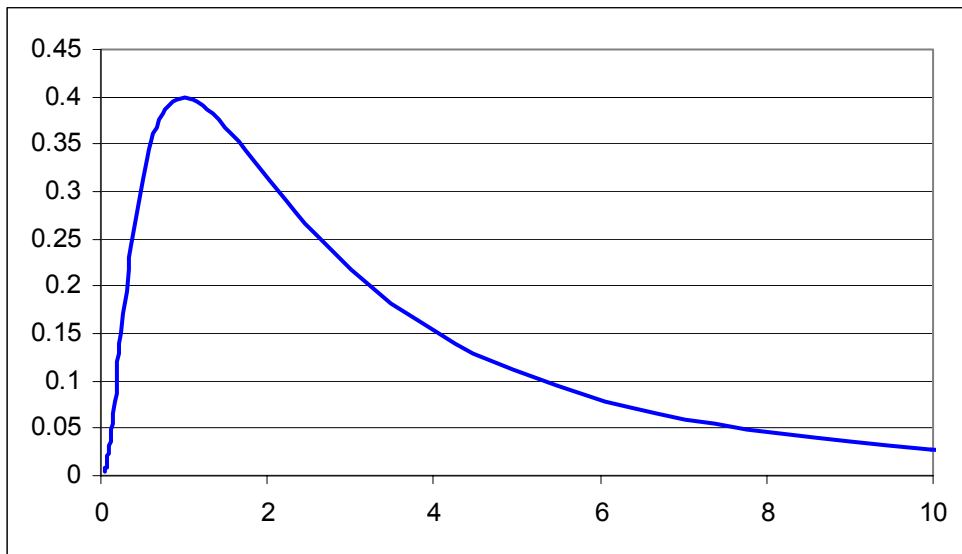
Some distributions have finite ranges. For example, the cost of catastrophes per year must be a non-negative number. But some linear combinations of statistical distributions can result in a negative number. Both the assumptions of normality and linearity fail in such circumstances. A correlation coefficient can describe the strength of the linear relationship for these situations, but an additional measure may be required to capture the non-negative behaviour.

If the boundary conditions are very rarely reached (e.g. a Log Normal with mean 100,000 and standard deviation of 100 will find it almost impossible to approach its minimum possible value of zero), then the normality assumption

RULE: Consider any boundaries upon the values that may be taken. If those boundaries are likely to be reached within a standard sample size, then the distribution is not very well described using linear relationships and normal distributions. In such cases a correlation coefficient can partly only describe a relationship.

Statistical Significance

Now that we know that the correlation coefficient is a measure of the strength of a linear relationship, we can look at some trivial cases to consider the reliability of a correlation coefficient in small samples.

Correlations for Data Sets with a Single Value

It is impossible to measure the correlation when a data set contains only one pair of numbers. This equates to fitting a straight line to a single point.

Graph 11: Correlation for One Data Pair
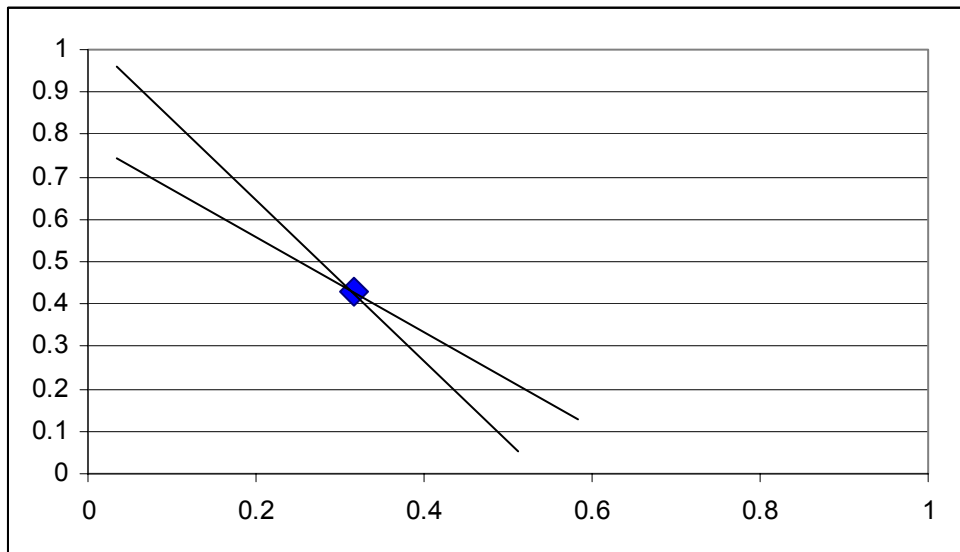
As can be seen in Graph 11, there are many possible lines that pass through the single data point. One does not know which line to use, so one cannot measure a correlation coefficient.

Correlations for Data Sets with a Two Values

As we know for geometry, it is always possible to fit a straight line through two points. A data set with two pairs of values will always produce a correlation coefficient of 100% or -100% because the two points can always be exactly described by a straight line.

Graph 12: Correlation for Two Data Pairs
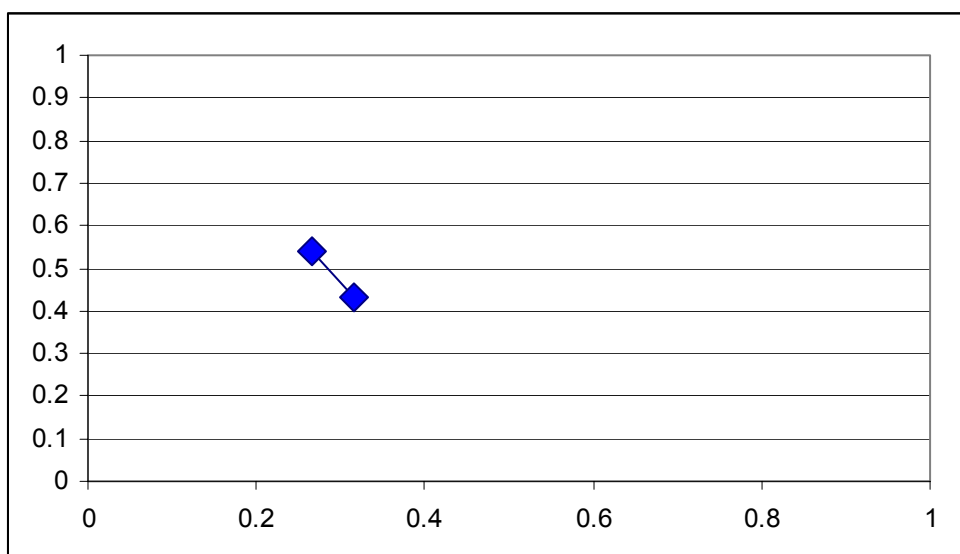
So for two pairs of data, the correlation coefficient will always be statistically insignificant.

Bootstrapping for Statistical Significance

There are some statistical formulae for the significance of correlation coefficients, but these rely upon an assumption of normality. In many real life situations actuaries are working with skewed, discrete and / or curtate data. In such cases we need to look elsewhere for a measure.

I have found bootstrapping to be a practical way of measuring the statistical significance of correlations within non-normal data samples. All one has to do is to generate a number of simulated samples independently drawing from the actual samples (with or without replacement). Since the two data series within the simulation are drawn independently, the simulated samples will have an underlying correlation of zero. One then measures the percentage of simulated samples that have a sample correlation measure greater in magnitude than the real sample. If that percentage is low, then it is unlikely that the underlying correlation coefficient is zero.

Bootstrapping automatically copes with skewed and discrete distributions.

The number of earthquakes in each year is only zero or one. There have been only two years in which earthquakes caused significant losses. So it is quite possible for a spurious correlation measure to occur. The sample correlation for the annual number earthquakes and hail events is 33%. One would not expect earthquakes to be related to hail storms in any physical way. When one does a bootstrap test of statistical significance it validates our prior assumption that there is no relationship.

RULE: The statistical significance is a function of the sample size. Don't just accept large correlation coefficients as significant, and don't just reject low correlation coefficients as insignificant.

Does Bad Luck Come In Threes?

When pricing catastrophe reinsurance actuaries often implicitly assume a process with no memory. They do this by using a Poisson distribution for the number of catastrophe events. We can test the validity of this assumption by testing the significance of the single period lag autocorrelation of the number of catastrophe events (the correlation between the number of catastrophe events in consecutive years).

Figure 3: Autocorrelations for Catastrophe Events

	Fire	Hail	Quake	Cyclone	Flood	Storm
Correlation	-0.1	-0.09	-0.03	0.23	-0.02	0.28
Significant	No	No	No	Yes	No	Yes

I have used the bootstrap technique to test the significance of the correlation measures because the sample distributions are skewed, discrete and non-negative. For cyclone and storm events there is a statistically significant correlation between consecutive years.

This means that bad years for cyclone and storm may come in groups of three!

The significant autocorrelation of cyclones and storms has important implications for the pricing of aggregate deductible catastrophe reinsurance covers and catastrophe reinsurance covers that span more than one year.

Copula Is Not a Dirty Word!

Throughout this paper I have mentioned the possibility of doing transformations upon data in order to allow simple correlation processes to operate more naturally. There is a branch of statistics that deals with the interrelationship of individual distributions into multivariate distributions, transforming their values using their survival functions and then formulating the relationship between those survival functions. The relationship between the survival functions is called a *copula*.

A multivariate normal distribution follows a *normal copula*. The normal copula is one of the simplest, and serves as a good example for understanding how a copula works. Say we wanted to generate some random numbers from two Gamma distributions using a normal copula. The steps to follow would be:

1. generate random numbers from a multivariate normal with the required correlations
2. transform the generated values to their survival function values (using the inverse cumulative density function for the Normal distribution)
3. transform the survival function values to their Gamma distribution values (using the cumulative density function for the Gamma distribution)

The sample values from the Gamma distribution would be correlated, although the correlation coefficient would be lower than that of the Normal copula because the Gamma distribution is skewed. Using an 80% correlation in the Normal copula on a Gamma(1, 5) and a Gamma(5, 1) distribution causes a correlation coefficient of 75%.

Graph 13: Gamma Distribution Samples Using Normal Copula
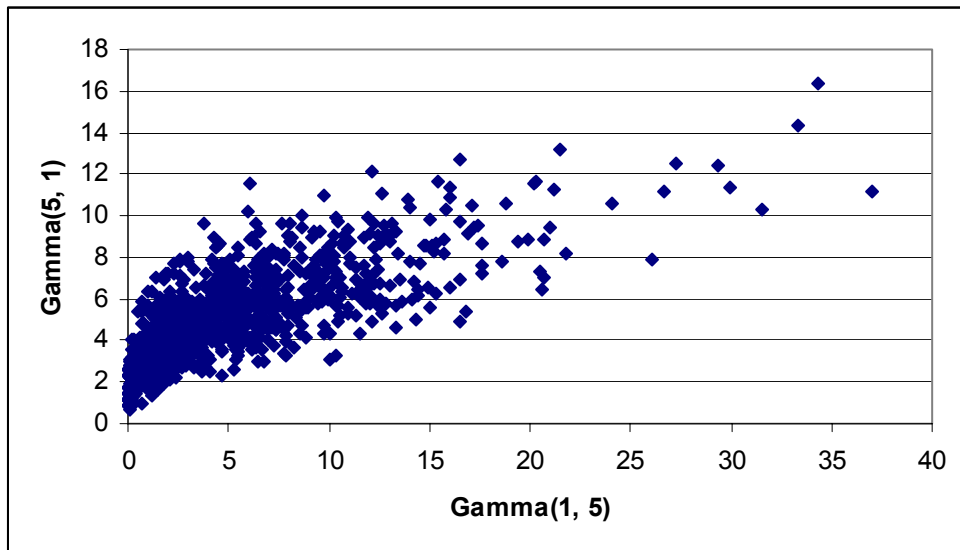
There are many different copulas. For a comprehensive description of different copula I recommend the books in the references section at the end of the paper. Different copula cause different relationships between the survival functions. Some copula cause stronger relationships in extreme values. Other copula

Tail Dependence

There are situations in real life where day to day variability is not correlated, but some large shocks cause a strongly correlated effect between data sets. Some examples are:

- share prices between different stocks
- insurance losses versus share prices
- commodity prices between different commodities
- road accidents between different areas

In many cases we are more concerned with the correlation of severe events than we are with the correlation of everyday events.

Graph 14 shows some real life data (with the labels removed for confidentiality purposes) where the strength of the relationship between the two variables is not constant.

Graph 14: Real Life Correlation Varying at Tails
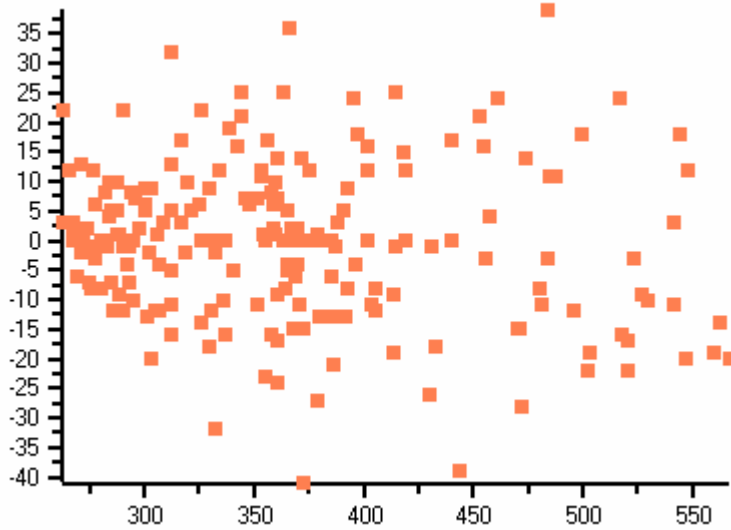
I wanted to determine whether the data in Graph 14 could be adequately described by a Normal copula, or whether the extreme values had a stronger relationship. There is a tail dependence measure defined in Joe (1997) which he defines as below.

Equation 3: Upper and Lower Tail Dependence (Joe 1997)

Definition. If a bivariate copula C is such that

$$\lim_{u \to 1} \overline{C}(u, u)/(1 - u) = \lambda_U$$

exists, then C has **upper tail dependence** if $\lambda_U \in (0, 1]$ and no upper tail dependence if $\lambda_U = 0$. Similarly, if

$$\lim_{u \to 0} C(u, u)/u = \lambda_L$$

exists, C has **lower tail dependence** if $\lambda_L \in (0, 1]$ and no lower tail dependence if $\lambda_L = 0$.

The reasoning behind these definitions is as follows. Suppose $(U_1, U_2) \sim C$. Then

$$\lambda_U = \lim_{u \to 1} \Pr(U_1 > u \mid U_2 > u) = \lim_{u \to 1} \Pr(U_2 > u \mid U_1 > u).$$

Unfortunately the definition in Equation 3 requires that one knows the underlying copula before one can measure the tail dependence. There does not seem to be any broadly accepted sample measure of tail dependence.

So I propose an empirical variant of the tail dependence measures. My proposal is:

Equation 4: Proposed Sample Tail Dependence Measures

$\lambda_U = Pr(U_1 > u \mid U_2 > u)$
$\lambda_L = Pr(U_1 < u \mid U_2 < u)$
where the probability u is equal to 0.25

There are practical problems with taking the limit (as required by Equation 3) on a sample set. So one must choose a value of u at which to measure the probabilities, and use these probabilities as a sample estimate of the tail dependence. My recommendation of 0.25 is based upon some experimentation with different real life data sets, and is a compromise between stability of estimate, and the requirement that the measure resemble the limit in Equation 3.

The data in Graph 14 has an upper tail dependence coefficient of 0.30 and a lower tail dependence coefficient of 0.09.

We have to be able to interpret the upper and lower tail coefficients in order to get meaning from them. So I have used a normal copula with the required correlation, and then measured the probability of obtaining a sample tail dependence measure exceeding the sample values. The upper tail dependence coefficient lies well outside the 90% confidence interval of 0 to 0.20. So we can reject the null hypothesis of a normal copula relationship. The upper tail is more strongly related than a Normal copula would cause. However the lower tail dependence measure is within the confidence interval from a Normal copula.

RULE: If the relationship between extreme values is important, then don't rely on correlation coefficients of the entire sample set. Measure the tail dependencies.

Australian Catastrophe Dependence Measures

Correlation Coefficients of Annual Catastrophe Costs

Figure 4: Correlation Matrix for Annual Catastrophe Costs (1985 onwards, > $30m)

	FinancialYear	Fire	Hail	Earthquake	Cyclone	Flood	Storm
FinancialYear	1	0.32	0.12	-0.16	-0.26	-0.29	0.04
Fire	0.32	1	-0.15	-0.1	-0.13	-0.31	-0.21
Hail	0.12	-0.15	1	0.14	0.13	0.14	-0.04
Earthquake	-0.16	-0.1	0.14	1	0.19	0.09	-0.17
Cyclone	-0.26	-0.13	0.13	0.19	1	-0.03	0.34
Flood	-0.29	-0.31	0.14	0.09	-0.03	1	-0.23
Storm	0.04	-0.21	-0.04	-0.17	0.34	-0.23	1

Using the bootstrap method, none of the sample correlation coefficients are significantly different to zero. We cannot reject the null hypothesis that the annual costs of losses from different catastrophe types are unrelated.

We may be surprised by this result. Some catastrophes are related to how wet the weather has been. It is difficult to imagine a flood putting out a bushfire because floods occur after lots of rain, and bushfires occur during droughts.


Many years ago, when I was studying to be an actuary at university, my professor would remind us that averages were sometimes insufficient to

describe a situation. He would tell us that a man, with his head in a furnace and his feet in ice, has a comfortable average temperature – but the variability of the temperature was of great concern to him! The situation is similar with correlations. A correlation coefficient is like an average – it does not tell us about the extremes.

Tail Dependence of Australian Catastrophes

In this section I will focus upon the relationship between cyclone losses and storm losses. These two catastrophe types have the strongest correlation coefficient in the period, and one would expect that they might be related.

Graph 15: Annual Storm Losses vs. Cyclone Losses
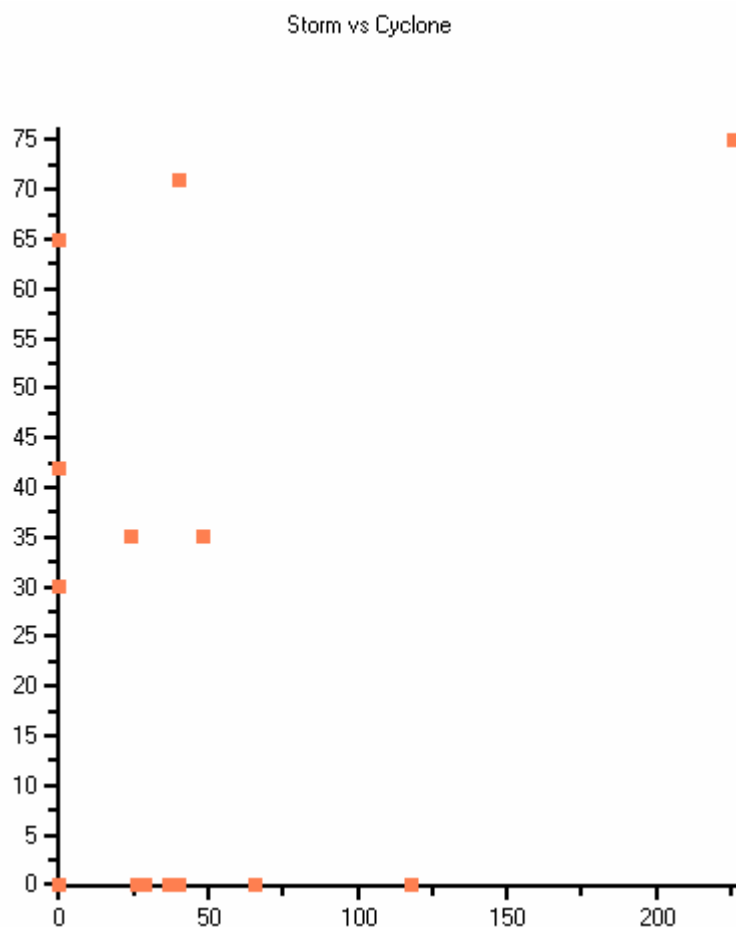
The sample correlation coefficient is 0.34 but this is statistically insignificant. Most of the linear correlation comes from the outlier in the top right hand corner of the graph. The lower tail dependence coefficient is zero – indicating that there is little or no relationship between low values. However the upper tail dependence coefficient is 0.25, which is significantly greater than one would expect from a normal copula.

These sample statistics may indicate that annual storm losses and annual cyclone losses are only related when severe losses occur. This may be due to el-nino climatic cycles.

An insurer will be concerned with the possibility of these two different catastrophe types being strongly positively related. If an insurer had only looked at the correlation coefficient, then it would have incorrectly concluded the level of diversification benefit that it is receiving. Actuaries need to understand more than simple correlations in order to manage risk.

Recommended Reading

Chen, P Y. Popovich, P M. (2002) "Correlation: Parametric and Nonparametric Measures". Sage University Papers Series on Quantitative Applications in the Social Sciences. Sage Publications, California, USA

Joe, H. (1997) "Multivariate Models and Dependence Concepts". Chapman and Hall, Suffolk, UK

Mari, D D. Kotz, S. (2001) "Correlation and Dependence". Imperial College Press, London, UK

Nelson, R B. (1999) "An Introduction to Copulas". Springer-Verlag, New York, USA

Web Resources

http://gro.creditlyonnais.fr/content/en/home_presentation.htm then click on "Applied Mathematics"

http://papers.ssrn.com/sol3/papers.cfm?abstract_id=291140

http://www.soa.org/library/naaj/1997-09.naaj9801_1.pdf

Appendix A: Annual Catastrophe Event Counts

Financial Year	Fire	Hail	Quake	Cyclone	Flood	Storm
1967	1	1	0	0	0	0
1969	0	0	1	0	0	0
1970	0	0	0	1	0	0
1971	0	0	0	0	2	0
1972	0	0	0	2	0	0
1973	0	0	0	1	0	0
1974	0	1	0	2	2	0
1975	0	0	0	1	1	0
1976	0	1	0	2	0	0
1977	1	1	0	1	1	2
1978	0	0	0	1	1	2
1979	0	0	1	1	0	0
1980	1	1	0	2	0	0
1981	0	0	0	1	0	2
1983	1	0	0	0	0	1
1984	0	0	0	1	0	0
1985	0	1	0	0	1	0
1986	1	2	0	1	0	0
1987	1	1	0	0	0	2
1988	0	0	0	1	4	0
1989	0	0	0	1	0	3
1990	0	4	1	1	1	0
1991	1	2	0	1	0	2
1992	1	0	0	0	2	2
1994	1	0	0	0	1	2
1995	0	0	1	1	0	1
1996	0	1	0	1	1	2
1997	1	4	0	1	1	2
1998	1	0	0	1	2	6
1999	0	2	0	2	2	4
2000	0	0	0	3	2	4
2001	0	0	0	0	2	5
2002	1	0	0	0	0	2
2003	1	0	0	0	0	0

Appendix B: Annual Catastrophe Event Costs

Financial Year	Fire	Hail	Quake	Cyclone	Flood	Storm
1967	101	36	0	0	0	0
1968	0	0	0	0	0	0
1969	0	0	12	0	0	0
1970	0	0	0	79	0	0
1971	0	0	0	0	43	0
1972	0	0	0	159	0	0
1973	0	0	0	150	0	0
1974	0	98	0	176	282	0
1975	0	0	0	837	63	0
1976	0	49	0	86	0	0
1977	30	131	0	49	23	62
1978	0	0	0	39	21	59
1979	0	0	12	41	0	0
1980	34	24	0	23	0	0
1981	0	0	0	17	24	60
1982	0	0	0	0	0	0
1983	324	0	0	0	0	19
1984	0	0	0	12	0	0
1985	0	299	0	0	132	0
1986	45	58	0	65	0	0
1987	12	161	0	0	26	42
1988	0	0	0	30	64	2
1989	0	0	0	35	0	51
1990	0	433	1124	42	38	0
1991	12	42	0	75	8	248
1992	12	0	0	0	26	120
1994	58	0	0	0	12	49
1995	0	0	36	11	0	29
1996	0	40	0	2	31	24
1997	10	173	0	8	20	10
1998	3	0	0	71	71	91
1999	0	1776	0	39	39	67
2000	0	0	0	34	18	58
2001	0	0	0	0	36	95
2002	0	0	0	0	0	0
2003	153	0	0	0	0	0