Correlation

Thomas Struppeck, FCAS, MAAA

Correlation

Thomas Struppeck, FCAS, MAAA

Abstract

Actuaries frequently are called upon to estimate sums of random variables. Such sums arise in a variety of contexts, as aggregate loss distributions, as losses including loss adjustment expense, as losses to a particular layer in stop loss reinsurance. If the quantities being summed were independent, things would be easy, however this is seldom the case. Generally, there will be some amount of "correlation" between the summands.

This paper examines the Pearson product moment correlation coefficient's strengths and weaknesses and discusses two non-parametric alternatives: the Spearman rank correlation coefficient and Kendall's tau statistic.

Correlation

Cor-re-la-tion n.

1. A causal, complementary, parallel, or reciprocal relationship, especially a structural, functional, or qualitative correspondence between two comparable entities.
2. Statistics. The simultaneous change in value of two numerically valued random variables.[1]

What is correlation? As can be seen from the above dictionary definition, correlation is a difficult concept to precisely define. Complicating the matter is the fact that the word "correlation" has a common usage (definition 1) and a technical usage. Upon hearing the word "correlation", actuaries may think of the (Pearson product moment) correlation coefficient, this preconceived notion can lead to further confusion.

In this paper, we will consider several notions of correlation and examine their relative strengths and weaknesses. After this brief introductory section, we will develop the Pearson product moment correlation coefficient and see why it specifically is central to the problem of estimating aggregate loss distributions. In the third section, we examine two other possible measures of correlation, Spearman rank correlation and Kendall's tau. These measures are in some ways more natural than product moment correlation, and they have appeared at various places in the actuarial literature.

In the fourth section, we examine some ways of estimating correlation coefficients. As we shall see, this can be a very tricky business. Sometimes pairs of quantities that at first glance would appear to be highly correlated, turn out not to be. We will see an example in which two independent random variables appear to be correlated!

Oftentimes the goal of studying some correlated random variables is to obtain an aggregate distribution. A common method for estimating the aggregate distribution is to use Monte Carlo methods. Our fifth section is devoted to examining different ways of simulating correlated random variables along with some cautions for the various methods.

---

[1] *The American Heritage® Dictionary of the English Language, Fourth Edition, 2000, Houghton Mifflin Company.*

Finally, before the Conclusions section, we consider a hypothetical aggregate stop loss and examine how our choice of correlation model can vastly change the estimated value of the different layers.

Independence and Correlation

In the normal course of business insurers commonly extend vastly more in total limit outstanding than they have in total assets. Consider a homeowners writer that insures 10,000 houses in each of fifty states for $100,000 each. The total sum insured is $50 billion. Assuming a premium of $1,000 per house, total annual premiums would be $500 million, so policyholder surplus might be $200 million. In this example, the total sum insured is 250 times the available surplus. How can such leverage work? The answer, of course, is that the insurer does not expect claims from all of its insureds at once. There is an assumption that losses occur somewhat independently from one another. This spreading of the risk is the essence of insurance.

What is Correlation?

First, let us review some probability. Suppose that X and Y are random variables with finite means, then:

$$E(X+Y) = E(X) + E(Y).$$

That is, the expected value of the sum is the sum of the expected values. Since expected value is the amount that the insurer will have to pay on average, it is naturally of considerable interest to insurers.

Insurers are also concerned about the variation around this expected value. This variability is commonly measured by using the standard deviation or its square, the variance. The variance is the expected squared deviation from the mean, namely:

$$VAR(X) = E((X - E(X))^2)$$

Or equivalently:

$$VAR(X) = E(X^2) - E(X)^2$$

Unlike the rule for expected value, the variance of a sum is not sum of the variances, except under specific conditions. Instead there are other terms:

$$\begin{aligned} VAR(X+Y) &= E((X+Y)^2) - E(X+Y)^2 \\ &= E(X^2 + 2XY + Y^2) - E(X)^2 - 2E(X)E(Y) - E(Y)^2 \\ &= E(X^2) - E(X)^2 + E(Y^2) - E(Y)^2 + 2E(XY) - 2E(X)E(Y) \end{aligned}$$

Denoting $E(XY) - E(X)E(Y)$ by $COV(X,Y)$, we obtain:

$$VAR(X+Y) = VAR(X) + VAR(Y) + 2\,COV(X,Y)$$

As we see above, computing the variance of the sum of X and Y leads us to consider the expected value of the product of X and Y and the product of their expected values. This product term in turn motivates the definition of covariance. It is worth noting that the covariance of X with itself is the variance of X.

Covariance is one step away from our goal, the correlation coefficient. If X is expressed in one unit of measure (say, meters) and Y is expressed in another unit (say, seconds), then the unit for the covariance will be the product (in this case meter-seconds). It is possible to normalize the covariance (assuming finite, non-zero second moments for X and Y), by dividing by the product of the standard deviations of the two variables. In his delightful book [F], Feller[2] suggests that a physicist might call this quantity "dimensionless covariance"; this ratio is the Pearson product-moment correlation coefficient. Since we will be looking at several possible measures of correlation, we will henceforth refer to this measure as "dimensionless covariance".

This quantity measures the extent to which X and Y are linearly related. It ranges between minus one and one. Being a measure of the linear relationship between two quantities might lead one to suspect that there is some connection between the correlation coefficient and linear regression. Indeed there is such a relationship, and it highlights a subtle point. Select a sample from the jointly distributed population (X, Y). From this sample, first compute the sample correlation coefficient. Then do two regressions: regress Y on X, and X on Y. In each of these regressions, X and Y have asymmetric roles: X is the independent variable and Y is the dependent

[2] Every actuary should read, at a minimum, section I.5 of Volume Two, "The Persistence of Bad Luck".

variable in the first and vice versa for the second. From each of the two regression lines, a slope is obtained. The sample correlation coefficient is the geometric mean of these two slopes (see for example, [CS].) The sample correlation coefficient treats X and Y equally; neither is thought of as being a function of the other, they are simply related somehow.

There are several ways that the correlation coefficient could be zero. The most important case is when the variables are independent. Independence, however, is a much stronger condition than correlation zero. Recall that two or more random variables are called independent if their joint density function factors into the product of their respective density functions. If two random variables are independent, knowledge about the value of one of them tells you nothing new about the value of the other one; i.e. the conditional distribution is with probability one the same as the unconditional distribution. On the other hand, if two random variables have a non-zero correlation, then knowing the value of one of them might give you information about the other one.[3]

Here is an example. Consider two lines of insurance, A and B, which have losses jointly distributed as follows:

	1	2
1	1/6	1/6
2	1/4	1/12
3	1/12	1/4

So the losses for line A can be either 1 or 2 and the losses for line B can be 1, 2, or 3 with the probabilities for the 6 possible pairs of events given by the above table. This table is the joint distribution function and it contains all of the information about A and B. For example, by summing the columns we get the marginal distribution for A, namely 50% of the time it is 1 and 50% of the time it is 2. Similarly by summing the rows, we get the marginal distribution for B, which is also uniform, this time on the three values 1, 2, and 3.

If A and B were independent, the probability of any given pairing would be 1/6. Evidently A and B are not independent. If we compute the Pearson correlation coefficient, we find that it is about 20.4%. It is

[3] More exactly, the probability that the conditional distribution is different than the unconditional distribution is positive.

interesting to note that the conditional distribution of A given that B=1 is the same as the unconditional distribution of A; however, if we are given that B=2 instead, then we have a different conditional distribution for A. (A in this case has a 75% chance of being 1 and a 25% chance of being 2, as the reader may check.)

Correlation and covariance are central concepts in modern portfolio theory. Consider a simple portfolio consisting of N securities, X_i, each of which has a value that is randomly distributed with mean M and variance V. Further, suppose that each pair of security prices has covariance equal to C, then:

$$E(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} E(X_i) = N \cdot M$$

and:

$$VAR(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} VAR(X_i) + 2\sum_{j<i} Cov(X_i, X_j)$$

$$= N \cdot V + 2\binom{N}{2} C = N \cdot V + 2\frac{N(N-1)}{2} C$$

The first sum in the variance line has N terms, but the second sum has N(N-1)/2 terms. As N gets larger, the second term, which is quadratic in N, starts to dominate. What happens if we examine the average instead of the sum of the X_i?

$$VAR(\frac{1}{N}\sum_{i=1}^{N} X_i) = \frac{1}{N^2} VAR(\sum_{i=1}^{N} X_i)$$

$$= \frac{N \cdot V}{N^2} + 2\frac{N \cdot (N-1) \cdot C}{N^2 \cdot 2} = \frac{V}{N} + \frac{N^2 - N}{N^2} C$$

The first term goes to zero as N goes to infinity, but the second term is bounded away from zero by C, the pairwise covariance. This last term, called "non-diversifiable risk" (because it doesn't go away as N gets large), is a cornerstone of the capital asset pricing model (CAPM) and of other factor models such as arbitrage pricing theory (APT).

Knowing the correlation coefficient is not enough

There are many ways that two random variables can co-vary. Even if we know the marginal distribution of X and Y and we know the correlation of X and Y, there generally will be many possible joint distribution functions, and hence, different ways that they might co-vary. A good way to visualize the relationship between two variables is to examine a scatterplot of them.

Consider the two data sets in Figures I and II. Both have standard normal marginal distributions, and the each pair of variables has correlation coefficient $\rho = 80\%$. The pair of normal random variables shown in Figure I is multivariate normal. We will use the term "binormal" for a pair of multivariate normal random variables. Each of the random variables shown in Figure II is normal, but their joint distribution function is not the multivariate normal. The pair in Figure II has been engineered so that the random variables exhibit extreme behavior simultaneously much more often than the pair in Figure I, yet the pairs in Figures I and II both have $\rho = 80\%$. Notice how the second and fourth quadrants have fewer observations in Figure II than in Figure I. These two quadrants represent pairs of random variables that have opposite signs. Similarly, the first and third quadrants appear to have more observations in Figure II than in Figure I, especially away from the origin. These points represent observations with both random variables being exceptionally far away from average.[4] This comparison illustrates a major problem with using any single number to attempt to describe how two things co-vary: one number cannot do the job.

Remarkably, for continuous random variables, there is an object that does capture the exact way that two or more random variables co-vary. Furthermore, it captures only the interaction between the variables, being independent of their respective marginal distributions. This object is called a copula and is described briefly in a following section. The next section serves to give further motivation to the introduction of copulas.

[4] The probability of an observation being in quadrants one or three minus the probability of an observation being in quadrants two or four is one way to define Kendall's tau, which we will discuss later.

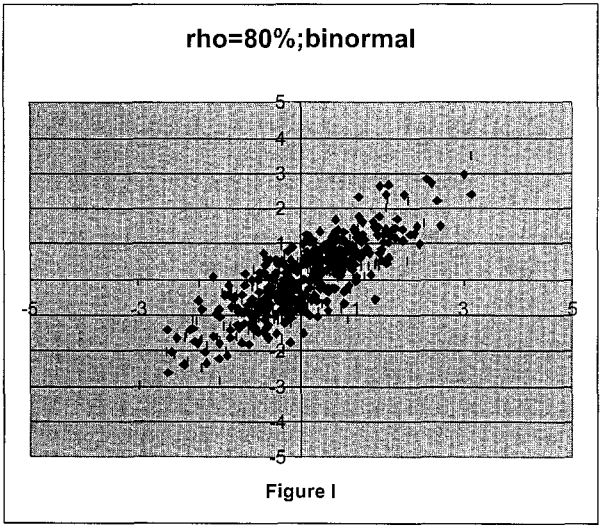

Figure I

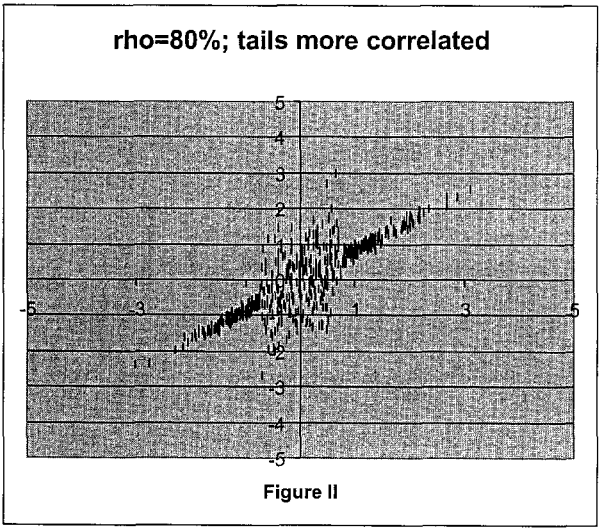

Figure II

Mango and Sandor's experiment

In a recent paper [MS], Mango and Sandor report on an experiment they performed to test a simulation procedure for generating correlated excess losses. They were attempting to estimate how much capital a reinsurer needs to support its book. They began by generating a binormally distributed pair of random variables with mean 0, variance 1, and specified correlation. As is well-known, a multinormal random variable is uniquely defined by its covariance matrix and its vector of means.

They take the pair (X, Y) of normal random variables and compute from it a pair (U, V) of lognormal random variables by setting U = exp(X) and V = exp(Y). They then censor these observations from below retaining only the pairs (U, V) where both exceed some threshold. At this point, they compute the correlation coefficient of the censored pair and observe that it is considerably smaller than the original correlation used when generating the pair (X, Y). Since the amount of correlation strongly influences how much capital is needed, they found this result to be alarming.

It will be instructive to examine these steps again, inserting one intermediate step.

Step 1: Generate a pair (X, Y) of normal random variables, binormally distributed with mean 0, variance 1, and specified dimensionless covariance.

Step 2: Denote by F() the cumulative distribution function for a normal random variable with mean 0 and variance 1. Determine the point in the unit square given by F(X), F(Y). In other words, determine the joint cumulative distribution function for this pair. (This step we have inserted into Mango and Sandor's process.)

Step 3: Compute U = exp(NORM(X)) and V = exp(NORM(Y)), where exp is the exponential function and NORM is the inverse of the cumulative distribution function that we introduced into the process in Step 2.

Step 4: Censor the results from below deleting all observations that have either U or V less than some threshold.

Step 1 is the generation of a pair random numbers and in this case the imposition of some sort of correlation structure on them.

Step 2 takes these random numbers and transforms them into a distribution function over the unit square. This particular choice of random number generation and transformation creates what is called a normal copula. The marginal distribution for each of the two coordinates is uniform on (0,1).

Step 3 takes these uniform (0,1) marginals and converts them to specified (in this case lognormal) marginals.

Step 4 takes the pair of lognormally distributed random variables and censors them from below.

Mango and Sandor noticed that the dimensionless covariance dropped. In which steps did they lose it? Some was lost in step 2 and some more in step 3 (they combined these steps). Also, some was lost in Step 4.

The loss in steps 2 and 3 comes about because dimensionless covariance is not preserved under non-linear transformations. A binormal pair of random variables is created in Step 1 and two nonlinear transformations are applied to it. There are measures which we will discuss later that are preserved under these transformations.

The loss in step 4 is more fundamental. The process of censoring is not a transformation --- some observations are discarded. Paraphrasing Embrechts, et al [EMS], the normal copula exhibits independence in the tails. What this means is that as we focus our attention on the observations in the extreme upper right portion of the first quadrant the pairs of observations that we find there are distributed as though they were (nearly) independent. This happens even though if we look at all of the observations they will exhibit some correlation.

Copulas

Recently the notion of a copula has begun to become more widely known by North American actuaries. Venter's paper, Tails of Copulas, is a particularly nice source [V]. (Other good sources are [W], [EMS], and Nelson's book [N].)

A copula is a dependence structure between two or more random variables. Let X and Y be (real-valued) random variables with joint distribution function F(x,y). That is:

$$F(x,y) = \Pr(X \leq x \text{ and } Y \leq y) \text{ for all real x and y}$$

If we just consider the variable X, we can ask about its distribution. It is given by the marginal distribution function, $F_X(x)$. Similarly, the marginal distribution of Y is given by $F_Y(y)$.

If we can find a function C: $[0,1]^2 \rightarrow [0,1]$ with $F(x,y) = C(F_X(x), F_Y(y))$, then we call that function a copula for X and Y.

Copulas have many nice features. First, they always exist. Secondly, in the case when X and Y are continuous, they are unique. This feature follows from what is known as Sklar's Theorem. It would take us too far afield to discuss this important result further. (The interested reader should see, for example, [ELM] or [N].)

One corollary of Sklar's Theorem is that, given a pair of continuous marginal distributions, every possible joint distribution with those marginals is given by a copula. The copula contains all of the information about the dependence structure of the joint distribution function; the marginal distributions contain the information about the individual components.

Considering the above, it seems that a good measure of the dependence of two (continuous) random variables would depend only on their copula since the copula captures all of the dependence information. Alas, dimensionless covariance is not a function of just the copula (in general it will depend of the marginals also), but there are measures that depend only on the copula. Since the copula captures all of the dependence information, this would suggest that these other measures are more natural. We discuss two of these measures later, but first let us see some other problems with dimensionless covariance.

Problems with dimensionless covariance

As we have seen, dimensionless covariance has several nice properties. It is easy to compute an estimate of it from a sample. It is related to linear regression, a familiar topic. In the case of a standard multivariate normal, knowing the pairwise correlations (equivalently, the covariance matrix) completely determines the joint distribution.

Ironically, these same properties highlight major weaknesses with it. The fact that it is easy to compute may lead practitioners to use it when it is not appropriate. The relationship to linear regression may make management feel that this is the right measure to consider, even if better

measures are offered. And finally, the fact that the covariance matrix contains all of the pertinent information in the multivariate case has led many to believe that it is true in general, when in fact, this relationship is seldom true.

The dimensionless covariance is invariant under positive affine transforms. I.e. if you multiply one of the variables by a positive constant and add another constant, the transformed variables have the same correlation coefficient. However, it is not preserved under more general transforms even if these transforms are monotone increasing. This noninvariance is a distinct weakness which seems to still not be fully understood by all practicing actuaries. Wang and Mango-Sandor touch on this issue.

Another problem with dimensionless covariance was alluded to earlier, namely that it is a single number. There are other problems too. While the dimensionless covariance is always between minus one and one, it is not always possible to achieve those bounds. Wang [W] gives explicit bounds for some lognormal random variables and shows that the range of achievable correlations can be made arbitrarily small. Wang shows that the largest possible dimensionless covariance between a lognormal random variable with parameters μ and 1 and another one with parameters $\mu \cdot \sigma$ and σ^2 is:

$$\frac{e^{\sigma} - 1}{\sqrt{e^{\sigma^2} - 1}\sqrt{e - 1}}$$

As the reader may check, this quantity tends to zero as sigma tends to infinity. So, in particular it is possible to have a pair of comonotonic lognormal random variables with dimensionless covariance arbitrarily close to zero.

Here is another easy example. Let X and Y be Bernoulli random variables with success probabilities of 10% and 20%, respectively. To maximize the correlation coefficient, we make X and Y equal to one simultaneously as often as possible. To do this we select a uniform (0,1) random variable, U and define:

X = 1, if U<0.1 X = 0, otherwise

Y = 1, if U<0.2 Y = 0, otherwise

We now compute the covariance of X and Y.

$$COV(X,Y) = E(XY) - E(X)E(Y) = 0.1 - 0.02 = 0.08$$

The standard deviations of X and Y are 0.3 and 0.4, respectively, so their correlation coefficient is 2/3. These two variables move together as much as possible, but their correlation coefficient is not 1. The reader can check that the smallest possible correlation between such a pair of Bernoulli random variables is $-1/6$.

In the case of multivariate normal random variables (or more generally elliptical distributions), the correlation matrix is the canonical measure, i.e. specifying the correlation matrix uniquely defines the distribution. Therefore, it may be surprising to some that the correlation lives near the center of the distribution. What I mean by that is that, in the tails, there is little correlation. As mentioned earlier, this is what Embrechts means when he says that the normal copula exhibits tail independence. Mango and Sandor observe this property in their study in which they consider lognormal random variables [MS].[5]

It is easy to produce a spreadsheet that generates, say, 500 pairs of standard binormal random variables with ρ=60% and then computes the sample correlation when these are truncated at 1 from below or censored at 1 from below. If you make such a spreadsheet, you will observe that in both cases, the resulting population correlation is smaller than 60% and the observed sample correlation is usually smaller. These examples illustrate two potential missteps for modelers:

Mistake 1: Assuming that by modeling the ground-up losses and their correlations accurately, you will automatically get the excess loss correlations correct.

Mistake 2: Assuming that because you know that the excess layers do not exhibit significant correlation, the ground-up losses will also not exhibit any.

Finally, correlation is very hard to measure. While the sample correlation is easily computed, the confidence intervals around it are quite wide, even under the assumption that the variables are binormally distributed. The author's spreadsheet exhibits considerable instability in the

[5] Since the tail independence is a property of the copula, one expects that they would have observed something similar for any marginal distributions.

sample correlations --- even in the sample correlations of the binormal random variables themselves. Brehm [B] alludes to some of the problems of measuring correlations from data.

Here is one suggestion: when analyzing data, looking at a scatterplot of the data is almost always a good idea.

Consider this pair of jointly distributed random variables:

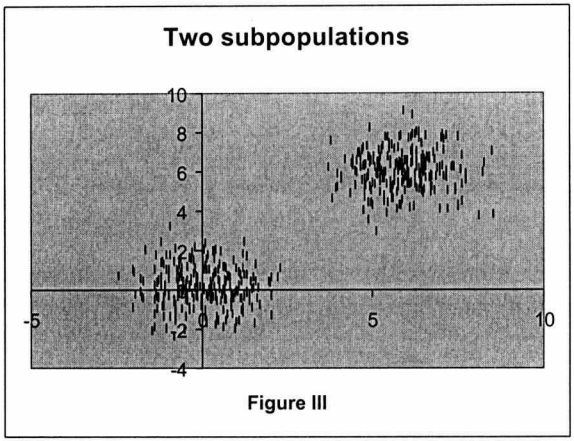

Figure III

A computation of the correlation coefficient will show a significant correlation, however it is misleading. As we see from examining the plot, we really have two classes. Once the two classes are identified and split, the apparent correlation goes away. We will see a second example of this type when we look at simulating correlated random variables. The variables here could be profitability for two lines of business, perhaps auto and general liability. One cluster might be urban risks and the other cluster rural risks. Given that a risk is rural, the auto and general liability are independent, similarly, given that the risk is urban, the auto and general liability are independent. However, naively computing the correlation of these two lines without partitioning the data, we will be misled into thinking that the lines are highly correlated.

Care should also be taken when adjusting data. Commonly, one adjusts for changing exposure bases over time by examining loss ratios. A danger with this is that loss ratios depend on premium and premium depends on the underwriting cycle. It is possible that apparent correlation between

losses from various lines is actually simply simultaneous fluctuations in premium adequacy caused by the underwriting cycle. If correlation between loss ratios is to be computed one should adjust for any premium redundancy or inadequacy first.

Another computational danger arises from binning of data. Insurance data frequently has a large degree of "lumpiness". Sometimes data is binned (perhaps all values between 100 and 300 are coded as 200). In addition, policy limits, deductibles, and case reserving methodologies can cause observations to be clustered around certain values. These data quirks can play havoc with canned correlation calculating software. Even if a more detailed analysis is attempted, the loss of detail in the data may lead to spurious results. The non-parametric methods described in the next section are more robust in this regard.

Since the computation of dimensionless covariance is essentially fitting a least squares line through the data, it is very sensitive to outliers. Observations that appear to be outliers should not just be ignored. Their existence should immediately call into question the assumption that the data is binormal (insurance data almost never is). There are non-parametric measures that are not as sensitive to the exact population distribution. We will examine two of these in the next section.

So, in summary, there are several problems with dimensionless covariance as a measure of correlation:

1) Confidence interval tests are usually predicated on the underlying data being binormal;
2) Insurance data frequently is binned or dirty in some other way, when reviewing this type of data, computing correlation is especially dubious;
3) Seldom can every possible value between -1 and 1 be achieved;
4) It is not preserved under increasing transformations; and finally,
5) It can be overly sensitive to outliers and to the distribution of the population which is almost never know.

Other Measures

As central to the study of aggregate losses as correlation and covariance are, as we shall see later, they really are not well behaved with respect to transformations of the data. There are other notions of correlation that are better behaved in that regard and are in some ways more intuitive than correlation.

The first of these measures is Spearman rank correlation. The idea behind rank correlation is very simple. Suppose that we have N pairs of observations of X and Y. (Remember that we are trying to understand how they are jointly distributed, so we must always think of the pair.) We look at the set of observed X values and the set of observed Y values (for the moment, assume no repeated values in either set). We take the N values of X and rank them from smallest (which gets assigned the number 1) up to largest (which gets assigned the number N). We similarly rank the Y values. We now look at the pairs of rankings, which we show sorted by X ranking:

Pairs (X, Y)	Pairs (Rankings)
(2.1, 7.4)	(1, 3)
(4.9, 9.8)	(2, 4)
(7.7, 7.3)	(3, 2)
(8.0, 5.2)	(4, 1)

Notice that all of the information about X and Y is gone, all that is left is information about how they are paired together. I.e. it is a function of their copula. With all of the information about X and Y abstracted away, we have reduced the problem to one involving only finite sets. Once we know N and the pairs of rankings, we have extracted all of the information.

Having sorted the pairs in order by their X rankings, we now count the number of "inversions" in the Y rankings, in this example we have: 3 appears before 1 and 2, 4 appears before 1 and 2, and 2 appears before 1: a total of 5 inversions.

The maximum possible number of inversions is N(N-1)/2 and the minimum possible is 0. The number of inversions is normalized to lie between -1 and 1 by subtracting N(N-1)/4 and dividing by N(N-1)/4. If N=4, as in this case, we get 2/3.

The result is the Spearman rank correlation. Like the dimensionless covariance, it ranges from -1 to 1 with negative scores indicating random variables that move opposite one another while positive scores tend to indicate random variables that move together.

An alternative way to define Spearman rank correlation is to compute the dimensionless covariance of the rankings, it turns out that this is equivalent to the above definition.

Another closely related measure is the Kendall rank correlation coefficient tau ("Kendall's tau"). As before, we take pairs (X, Y) and rank them. Then we count the number of times that the first variable is greater than the second variable (in rank). This count yields a number between 0

and N which we subtract from N/2 and normalize by dividing by N/2 to produce a final result in the range of −1 to 1 as for the other two measures. The result is Kendall's tau. Using the example with four observations above, the only pair with the first coordinate's rank larger than the second's is the pair (3, 2), so we subtract 1 from 4/2 and divide by 4/2. This tells us that Kendall's tau for this data is 0.5.

There is an equivalent way to think of Kendall's tau that gives it a nice intuitive interpretation. Select two pairs of observations (X_1, Y_1) and (X_2, Y_2) from the jointly distributed population. Call this pair concordant if either $X_1 < X_2$ AND $Y_1 < Y_2$ or $X_1 > X_2$ AND $Y_1 > Y_2$; otherwise call the pair discordant (if ties are permitted this definition needs to be adjusted). Let P(C) and P(D) be the probabilities of concordant and discordant observations, respectively. Then Kendall's tau is P(C) – P(D).

In a scatterplot, Kendall's tau can be thought of as the probability that an observation is in quadrants one or three (concordant) minus the probability that it is in quadrants two or four (discordant).

So, Kendall's tau measures the likelihood that two random variables move in the same direction. It is a statement about probability; it does not try to relate the size of the changes, only their directions. I suspect that when people say that two things are correlated, they generally have something like this relationship in mind.

A third non-parametric measure, which we mention in passing is the Gamma statistic. It is similar to Kendall's tau in that it also is a probability. It is useful when there are many ties (such as you might find in frequency data). It is the probability of concordance minus the probability of discordance divided by one minus the probability of ties. (See for example, Siegel & Castellan [SC])

All three of these non-parametric measures have the nice feature that they are equal to one in the case on concordance and minus one in the case of discordance. Also, they are invariant under increasing transformation, so they are actually properties of the copula.

Kendall's tau (τ) and Spearman's rank statistic (R) cannot differ by very much, as illustrated by the following inequality due to Daniels [see N]:

$$-1 \leq 3\tau - 2R \leq 1$$

What this inequality tells us is that if either Kendall's tau or the Spearman R is near 1 (or −1) then the other must be also, and when one is

near 0 the other cannot be too far away. For example, if the Spearman coefficient is zero then Kendall's tau is at most 1/3 in absolute value.

An alternative development of Kendall's tau, Spearman rank correlation, and dimensionless covariance can be found in Wang's paper on aggregation of risk portfolios [W].

Simulation

One of the more common methods for estimating aggregate distributions is to run a Monte Carlo simulation. Reinsurance may have features such as drop-downs or shared retentions, or any number of other features that make analytic modeling practically impossible. In these cases, Monte Carlo simulation may be the only available tool.

Determining how to incorporate correlation information into a simulation is an important problem. It is quite common to use the normal copula in these cases, as we have seen earlier there are some pitfalls associated with this.

Some canned simulation packages can induce a given Spearman rank correlation. One possible method for simulating a given rank correlation structure is described below.

Parameter uncertainty is commonly dealt with through use of a simulation. The incorporation of this uncertainty can produce some unexpected results. Let X and Y be two independent identically distributed random variables that are both normal with variance 1 and mean θ, where θ is either 1 or 2, but is not known which. Suppose that we run a simulation and assign θ the value 1 half of the time and the value 2 the other half of the time, using the same value of θ for both X and Y. We then compute the correlation of X and Y across the simulation. We will show a positive correlation even though X and Y are independent! The scatterplot of the result will look similar to plot number III. (The scatterplot of the two lines of insurance with an urban cluster and a rural cluster.)

In the previous example, θ was constant in each scenario. Frequently, parameter uncertainty arises from uncertainty about the parameter for each individual. This uncertainty can manifest also itself as observed correlation. Meyers [M] has written extensively about this topic.

Even if it is believed that the correlation is known, it may be best to use Kendall's tau or rank correlation anyway. There are several reasons for this. In the case of binormal random variables, there is a formula for converting from dimensionless covariance to rank correlation and vice versa.

So, in at least one important case they are almost the same.[6] Secondly, the confidence intervals that arise in the calculation of dimensionless covariance from samples are generally very wide. If one uses rank correlation, it is possible to achieve the specified rank correlation when performing a Monte Carlo simulation, a goal which in general in unachievable in the case of dimensionless covariance as we have seen.

One possible method to generate random variables with a given rank correlation in a simulation is to do the following:

Assume that the variables are X and Y and that a method exists for generating instances of each of these variables. We will also assume that a simulation with N iterations is to be run.

Step 1: Generate N instances of X and Y. Sort each of these from smallest to largest.

Step 2: Notice that paired this way (sorted) they are fully concordant so the rank correlation is one.

Step 3: (Loop) While the rank correlation is larger than desired, interchange two of the Y's moving a smaller one up and a larger one down.

The choice of pair to swap in step 3 can be tailored to make the process appear to have more or less tail dependence as the modeler desires.

An Example of an Aggregate Stop Loss

To illustrate some of these concepts, let us examine a simplified example of an aggregate stop loss. Our model consists of 100 risks, each of which has a 5% chance of suffering a loss of (exactly) 1. For simplicity, we will assume a zero percent discount rate.

We will divide these losses into three layers: a primary layer that covers losses in aggregate from 0 up to 6, a "working layer" that covers losses from 7 up to 10 and a catastrophe (CAT) layer that covers losses from 11 to 100.

The expected loss to the entire program is 5 (since there are 100 risks each with a 5% chance of loss). Of course, this expectation does not mean

[6] This is the classic case for dimensionless covariance where everything works as expected.

that we expect no loss in the working layer. If the losses were independent Bernoulli trials, the probability of 7 or more losses would be about 23.4% and the probability of 11 or more losses would be less than 0.5%.

What happens if the losses are not independent? The answer is that things can be very different. How they are different depends on the joint distribution function of these 100 variables. We will look at three different models.

First, is it possible for the pairwise correlation to be negative for each pair of insureds? At first blush, ones intuition might lead one to think that you cannot have 100 things which are all pairwise negatively correlated. Positive correlation intuitively means that when one variable goes up the other tends to also, so negative correlation should mean that they move in opposite directions. If we have 100 such variables, then when one of them goes up we expect the other 99 go down ... that means that most of them are moving together! As compelling as this argument is, it is wrong. To see this fallacy, consider our example of 100 risks each with a 5% chance of loss where we are given that the total number of losses is exactly 5. Now, knowledge that one given risk had a loss decreases the probability that any other risk had a loss (from 5 in 100 to 4 in 99). Similarly, knowledge that a given risk did not have a loss increases every other risk's chance of loss (from 5 in 100 to 5 in 99). This information shows pairwise negative correlation. Of course, in this case we could have done a simple calculation to explicitly compute it also (it is equal to $-1/95$).

When we look at the effect of this particular correlation structure on our excess losses we see that it results in no excess losses at all! Since there are always exactly five losses, all of the loss occurs in the primary layer. In this example, negative correlation helps the excess layers and hurts the primary layers. Of course, this example is somewhat unrealistic.

In the second model, we will set the pairwise correlation to a specified value using a normal copula. One way to achieve this pairwise correlation is to generate 101 independent standard normals. One of these variables, say the first one, we will call the "market factor". Each of our one hundred correlated normal random variables will be obtained by taking a multiple of the "market factor" and adding to it one of the other standard normals. The resulting variable will be normal (the sum of normals is normal), but it will require rescaling to be standard (variance 1). If the multipliers are selected appropriately, any specified non-negative correlation can be achieved in this manner.

Here is a simple example. Suppose that we wish to achieve a pairwise correlation of 36%. We create standard normals which are correlated 60%

with the market factor (60% being the square root of 36%). If we call the market factor M, the independent normals X_i, and the correlated normals Y_i, we would set:

$$Y_i = 0.8X_i + 0.6M$$

The Y_i have mean 0 (because both X_i and M do) and variance 1 (because $0.8^2 + 0.6^2 = 1^2$). An easy calculation left to the reader shows that the correlation between distinct Y_i and Y_j is indeed 36%.

These normals are then used to create a copula (i.e. the inverse transformation is used to obtain 100 uniformly distributed (0,1) random variables) the components of which, in turn, are used to determine if the corresponding risk has a loss. If the uniform random variable is less than 0.05, then there is a loss on that risk.

The overall expected losses remain 5, however the sharing of these losses amongst the layers depends heavily on the correlation (however it is measured). The higher the correlation, the more expected loss there is above the primary layer.

Strangely, beyond a certain point additional correlation stops hurting the working layer and actually begins to lower its expected losses. This reduction happens when the additional expected loss coming in from the primary layer is exceeded by the expected loss passed up to the CAT layer. Global reinsurers have observed similar phenomenon when dealing with currencies that have extreme inflation.

In the third model, we will consider the following method for generating losses with a specified pairwise correlation that includes a catastrophe component. Select 101 standard independent normals as before and generate uniform random variables as in the prior example. Now generate one more independent standard normal. If it is smaller than some specified amount, set all of the one hundred risks equal to the first one. It is easy to do the parameter selection so that a given pairwise correlation is exactly achieved.

This model has the exact same correlation as the prior model, but it is vastly more dangerous for the CAT layer and safer for the primary (again the working layer might either benefit or lose). Effectively, our 100 separate risks get replaced by one giant risk a certain percentage of the time. When this giant risk has a loss, the CAT layer is totally wiped out. This is a form of common shock model.

Spectacularly large losses to the CAT layer will almost never be observed in the prior model because of the tail independence of the normal

copula, but they will be observed with this modification. Certain risks (in particular credit losses) may be better modeled by something with a very fat tail, such as a mixture of this last model with the prior one.

Conclusions

Determining how the correlation of risks changes the aggregate loss distribution is a very difficult problem. One number cannot describe the relationship between two jointly distributed random variables. It is possible to make two models with identical correlation coefficients that exhibit wildly different excess aggregate losses. The difference lies in the copula.

The copula (which captures the structure of joint distribution) and the marginals (which capture the individual random variable) can be used to specify pairs (or more generally tuples) of jointly distributed random variables. Casualty actuaries are beginning to use these tools to understand the correlation structure of insurance loss random variables. This use will no doubt increase in the future.

Because of its relationship to the second moment of the aggregate loss distribution, Pearson product moment correlation (which we have referred to as dimensionless covariance) is a fundamental quantity. However, it lacks some of the nice features of the non-parametric measures, Kendall's tau and Spearman rank correlation.

The non-parametric measures may be somewhat more intuitive and may better capture what management is thinking when they say that two things are correlated.

References:

[B] Brehm, P.; "Correlation and the Aggregation of Unpaid Loss Distributions"

[CS] Clark, C. and Schkade, L.; "Statistical Methods for Business Decisions", South-western Publishing Co., 1969

[ELM] Embrechts, P.; Lindskog, F.; McNeil, A.; "Modeling Dependence with Copulas and Applications to Risk Management"

[EMS] Embrechts, P.; McNeil, A.; and Straumann, D; "Correlation and Dependence in risk Management: Properties and Pitfalls"

[F] Feller, W.; "An Introduction to Probability Theory and Its Applications", Wiley, 1968

[M] Meyers, G.; "Estimating Between Line Correlations Generated by Parameter Uncertainty"

[MS] Mango, D. and Sandor, J.; "Dependence Models and the Portfolio Effect"

[N] Nelson, Roger B.; An Introduction to Copulas, Springer, 1999

[SC] Siegel, S. and Castellan, N. J.; Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill Book Company, 1988.

[V] Venter, G.; "Tails of Copulas"

[W] Wang, S.; "Aggregation of Correlated Risk Portfolios: Models and Algorithms", 1999 PCAS LXXXV