A Correlation Model to Analyze Dependent Variables

by

H. Dezfuli & M. Modarres

A CORRELATION MODEL TO ANALYZE DEPENDENT VARIABLES

H. Dezfuli

M. Modarres*

Department of Chemical and Nuclear Engineering
University of Maryland
College Park, MD 20742

Submitted for Publication to IEEE Transactions
On Reliability - March 1986

*To whom correspondence should be sent

A Correlation Model to Analyze Dependent Variables

Key Words- Statistical correlation, Dependent analysis, Statistical uncertainty, Bootstrap method, Correlation coefficient

Readers Aides-

Purpose: Extend state of the art

Special math needed for explanation: Mathematical statistics

Special math needed to use results: None

Result useful to: Reliability engineers and theoreticians

ABSTRACT

In this paper a methodology is formulated to study the dependency between two variables. Statistical correlation coefficient between two variables is used as a measure of the degree of dependency. The method of bootstrap technique is employed to account for the statistical uncertainty in the correlation coefficient that is estimated from available data. A computer program entitled correlation coefficient generator (CCG) is developed to perform the analysis. An example is also presented to demonstrate the methodology. The objective of this example is to determine the dependency between specific actions of control room operator of a nuclear power plant. The methodology and the CCG code are very effective and easy to use.

1. Introduction

Through a close observation of dependent events, dependent failures, or dependent human actions in engineering systems, one can envision the magnitude of the complexity of the problems associated with the analyses of these dependencies (1,2). Causes of dependencies are often numerous and present models can not consider and treat them individually. It is important to consider two factors in the development of any dependent analysis model:

a. Complexity in mechanisms and causes from which the dependency originates.

b. Inadequacy of the data base.

Statistical models provide a means of determining the dependency. An advantage of models based on statistical analysis is that in these models, identification and enumeration of the causes of dependencies are not required. Therefore, in this respect, a statistical model is concerned only with the available data.

Each statistical model is bound to estimate a number of parameters. Again, the major consideration is the amount of data available. With a great deal of data, one can, in principle, be fairly elaborate. It should be recognized that any fancy model would require an estimate of a number of parameters. With only a small amount of data available, it is necessary to use simple and very direct models. Having this consideration in mind, the correlation model

which involves the estimate of only one parameter is developed in this study. The idea here is to obtain statistics which lump many dependencies together. This estimate is the correlation coefficient (ie, r). The value of r represents the degree of dependency between two variables. In the following sections, a discussion of the correlation model, the developed computer program (CCG code), and an application of the model is presented.

2. Notation And Nomenclature

Notation

r	linear correlation coefficient
X	a random variable which can assume values of $x_1, x_2, \ldots, x_n$
Y	a random variable which can assume values of $y_1, y_2, \ldots, y_n$
XY	X times Y
n	sample size
$f_{x_i}, f_{y_j}, f_{X_i\ Y_j}$	frequency that X, Y, or XY take values of x_i, y_j and $x_i y_j$ in the sample
f_{ij}	bivariate frequency of a case in which $X=x_i$ and $Y=y_i$

W_{ij}	weighting factor of a case in which $X=x_i$ and $Y=y_i$
$[W_{ij}]$	an i-by-j square matrix containing weighting factors W_{ij}
$<x_i,y_j>$	a doublet showing a case in which $X=x_i$, and $Y=y_i$
m	number of bivariate frequency table cells with nonzero entries $m \leq n$
N	number of r's to be ordered

Nomenclature

Statistical correlation:	Two measureable statistical characteristics are correlated when a mathematical relationship can be found to correspond these characteristics.
Linear correlation coefficient:	A standard measure of the linear relationship between two variates.
Dependent variables:	Whenever two random variables have a nonzero correlation coefficient.

3. Correlation Theory

A method of determing the relation between one variable

and another which is commonly used in statistics is the regression method. A standardized measure which is commonly used in regression method is a dimensionless parameter called "the correlation coefficient". The correlation coefficient carries information about two aspects of a relationship: (1) Its strength - measured on a scale from 0 to unity. (2) Its direction - indicated by the presence or absence of a minus sign. The important point is to recognize that the strength of a coefficient is entirely independent of its direction (positive or negative). A correlation coefficient of 1.0 is not stronger than a correlation coefficient of -1.0. Mathematically the linear correlation coefficient between X and Y is given by [2]:

$$r = \frac{n \sum_{i,j} f_{X_i Y_j} - \left(\sum_i f_{x_i} X\right)\left(\sum_j f_{y_j} Y\right)}{\sqrt{\left[n \sum_i f_{x_i} x^2 - \left(\sum_i f_{x_i} x\right)^2\right]\left[n \sum_j f_{y_j} y^2 - \left(\sum_j f_{y_j} y\right)^2\right]}} \qquad (1)$$

For example, if a data sample for variables X and Y take the values listed in Table 1, then r from (1) would be 0.402. This indicates that there is some positive correlation between X and Y (ie, there exists a moderate tendency for Y to go up when X goes up, and vice versa).

4. A Computer Based Method to Assess Confidence Limits of Estimated r.

The correlation analysis will be improved by inclusion

of the uncertainty calculation of the estimated r. An analytical approach is presented and explained in [3] to determine confidence limits of r. In this analytical method one should assume a known distribution for example, Gaussian [4]. Obviously, this assumption may not be valid; additionally, a large amount of data may be needed to reach any conclusive results.

It is proposed to use the bootstrap approach [5] as a means to perform the uncertainty calculation associated with the estimated r. The bootstrap method is a computer based technique for the performance of uncertainty calculation and determination of confidence limits. The bootstrap method has recently gained widespread applications [6,7]. The idea in bootstrap method is to mimic the process of selecting many samples of size n from an existing sample in order to find a probability distribution for r's obtained for each sample. The bootstrap samples are generated from the data in the original sample. The name bootstrap, which is derived from the old saying about pulling yourself up by your own bootstrap, reflects the fact that one available sample will give rise to many others .

In a computer approach, the bootstrap samples are generated as follows. The data for the first data set $\langle x_1, y_1 \rangle$ are copied a large number of times and the data for each of the other sets $\{\langle x_j, y_j \rangle \; j=2,3,\ldots,n\}$ are copied an equal number of times. The resulting copies

are thoroughly mixed. Samples of size n are then selected at random, and r is calculated for each sample; variation among r's provides the basis of estimating confidence limits of the true r.

The distribution of r calculated from the bootstrap samples can be treated as if it were a distribution constructed from real samples: it gives an estimate of the statistical accuracy of r that was calculated for the original sample. On a computer, the steps of copying, mixing, and selecting a new set of data are all carried out by a procedure that is much faster but are the mathematical equivalent of hand calculations.

5. Computer Modeling of Bootstrap Method

Let us examine the model which is developed to perform the task of copying, mixing, and subsequent sampling of data. In this model as part of copying and mixing procedure there is a need to first make use of a uniform random number generator, and second, make use of a weighting factor.

Lehmer [8] proposes the use of an efficient uniform random number generator for computers with an available word length of less than or equal to 36 bits. This efficient uniform random number generation is used to produce random numbers between 0 and 1.

The weighting factors are obtained from:

$$W_{ij} = \frac{f_{ij}}{n}. \qquad \begin{matrix} i=1,2,\ldots,n \\ j=1,2,\ldots,n \end{matrix} \qquad (2)$$

For example, $[W_{ij}]$ for the data shown in Table 1 is presented below:

$$[W_{ij}] = \begin{bmatrix} \frac{4}{14} & 0 & 0 & 0 & \frac{1}{14} \\ \frac{1}{14} & \frac{2}{14} & 0 & 0 & 0 \\ \frac{1}{14} & \frac{1}{14} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{14} & 0 \\ \frac{1}{14} & 0 & \frac{1}{14} & 0 & \frac{1}{14} \end{bmatrix}.$$

Assignment of W_{ij} to each cell along with the use of the uniform random number generator allows one to perform the bootstrap procedure in a simple fashion. The basic steps are summarized below:

1. Form the bivariate frequency table from actual observed data.
2. Determine W_{ij} for each cell.
3. Count the number m. For example for Table 1, m is 10.
4. Store the location of each nonzero cell and its ij attribute along with its corresponding

W_{ij}. For example the first three columns of Table 2 shows these values for the data in Table 1.

5. For each nonzero cell, calculate left and right boundary values by summing up all W_{ij} before the cell, and including the W_{ij} for the cell, respectively. Last two colums of Table 2 illustrates the left and right boundary values.
6. Generate a uniform random number between 0 and 1, and select a nonzero cell such that:

$$\text{left boundary of the cell} < \text{generated random number} \leq \text{right boundary of the cell}$$

Register one count for each cell when the above condition is satisfied. Since the length of a nonzero cell is proportional to the frequency of the cell, the larger cells tend to capture proportionally more counts. Repeat this step n times in order to generate one bootstrap sample. At this point the value of r is calculated for each bootstrap sample and recorded. For example, Table 3 shows one typical bootstrap sample generated. For this sample r=0.586.

7. Step 6 is normally repeated many times (10,000 to 100,000 times) depending on the accuracy of the desired confidence intervals of estimated r.

6. Structure of CCG Computer Code

In order to accomplish the formation of bootstrap samples and calculation of confidence limits of the estimated r, the CCG code is developed. This program is written in FORTRAN-V for UNIVAC 1108 computers and consists of three routines:

1. bootstrap sample generator,
2. correlation coefficient estimator,
3. probability interval evaluator.

The basic modeling principles related to the first two routines are outlined in section 5. The third routine basically deals with ordering of r's generated from the bootstrap samples. This task is performed in the CCG via a sorting scheme suggested by Shell [9]. Ordering of r is the most time consuming part of this algorithm. However, the shell ordering routine has an empirical computer time requirement directly proportional to $N^{1.226}$. Following the ordering process statistical confidence limits are determined for various confidence intervals.

7. Application of the CCG Code

As an application of this method lets consider the auxiliary feedwater system (AUXFEED) which is used in pressurized water reactors to cool their steam generators during certain emergencies. This system has two different trains, each capable of independently cooling the steam generators.

In all reported complete failures of AUXFEED system (ie, loss of both trains) in the 10 years period of 1969-1979 as reported to the Nuclear Regulatory Commission [10], some were potentially recoverable by the operator. Under the stated condition, and for all these reported failures, the likelihood that each (AUXFEED) train could have recovered by the operator is determined in [11]. As reported in [11], the likelihood of operator success (or failure) to recover each train is determined to have discrete values; these discrete values are proportional to the severity of the train failures and/or the time available to the operator to act. These discrete values and the frequency of AUXFEED train failures corresponding to each discrete value for the period of 1969-1979 are tabulated in Table 4. If the likelihood of operator failure for each train of AUXFEED is taken as variables X and Y, then r and its confidence limits can be determined by using the CCG. In this case the value of r shows the statistical correlation between an operator's likelihood of failure to recover the two trains of an AUXFEED System.

The output from CCG for 20,000 bootstrap samples shows a point estimate of r=0.84 and 80% confidence interval of $0.68<r< 0.93$. This clearly suggests a strong dependency between operator's likelihood of failure to recover each AUXFEED train. In other words as the likelihood of operator's failure to recover one of the AUXFEED trains goes up, the likelihood of operator's failure to recover the other train would also goes up.

REFERENCES

[1] G. R. Burdick and R. B. Worrell, "Qualitative Analysis in Reliability and Safety Studies", IEEE Transactions on Reliability, vol. R-25, No. 3, 1976, ppl69-181.

[2] H. Dezfuli, "A Correlation Model to Analyze Dependent Failures For Probabilistic Risk Assessment", Ph.D. dissertation, Department of Chemical and Nuclear Engineering, Univ. of Maryland, 1985.

[3] A. Hald, "Statistical Theory with Engineering Applications", John Wiley & Sons, Inc., New York, 1951.

[4] F. N. David, "Tables of the Ordinates and Probability Integral of the Distribution of the Correlation Coefficient in Small Samples", London, 1938.

[5] B. A. Efron, "Bootstrap Methods: Another Look at Jacknife", Ann. Statistics, vol.19, 1979, ppl-26.

[6] P. Diaconis and B. Efron, "Computer-Intensive Methods in Statistics", Scientific American, Vol. 248, No.5, 1983 May, ppll6-130.

[7] M. Modarres and T. Cadman, "Statistical Uncertainty Analysis in Reactor Risk Estimation", Nucl. Eng. Des. J., Vol. 85, 1985, pp385-399.

[8] D. H. Lehmer, "Proceedings of 2nd Symposium on Large-Scale Digital Calculating Machinery", Harvard University Press, Cambrdige, 1951.

[9] Y.T. Lee and S. L. Salem, "Probability Intervals for the Reliability of Complex Systems Using Monte-Carlo Simulation", UCLA-ENG-7758, 1977.

[10] D. L. Shell, "A High-Speed Sorting Procedure", Communi-cations of the Association for Computing Machinery", vol. 2, No. 7, 1959 July, pp30-37.

[11] "Licensee Event Report (LER) Compilation", NUREG/CR-2000, Published monthly by the U.S. Nuclear Regulatory Commission.

[12] J.W. Minarick and C.A. Kukielka, "Precursors to Potential Severe Core Damage Accidents: 1969-1979, A Status Report", NUREG/CR-2497, 1982.

Reference [2] is available from: Department of Chemical and Nuclear Engineering; Attn: Dr. M. Modarres, University of Maryland; College Park, Maryland 20742 USA.

Reference [9] is available from: School of Engineering and Applied Science, University of California; Los Angeles, California 90024 USA.

References [11,12] are available from: National Technical Information Service; Springfield, Virginia 22161 USA.

Dr. H. Dezfuli; Department of Chemical and Nuclear Egineering; College Park, Maryland 20742, USA.

H. Dezfuli: For biography see vol. R-33, 1984 Oct, P328.

Dr. M. Modarres; Department of Chemical and Nuclear Engineering; College Park, Maryland 20742, USA.

M. Modarres (M'83): For biography see vol. R-33, 1984 Oct, P328.

TABLE 1

An Example of A Bivariate Frequency Table
(A 25 Cell Table)

		X					
		x_1	x_2	x_3	x_4	x_5	Total
Y	y_1	4	0	0	0	1	5
	y_2	1	2	0	0	0	3
	y_3	1	1	0	0	0	2
	y_4	0	0	0	1	0	1
	y_5	1	0	1	0	1	3
TOTAL		7	3	1	1	2	14

TABLE 2

Determination of Boundary Values For Each Nonzero Cell

Nonzero Cell No.	$\langle x_i, y_j \rangle$	Weighting Factor (Wij)	Left Boundary Value	Right Boundary Value
1	$\langle X_1, y_1 \rangle$	4/14	0	4/14
2	$\langle x_1, y_5 \rangle$	1/14	4/14	5/14
3	$\langle x_2, y_1 \rangle$	2/14	5/14	7/14
4	$\langle x_2, y_2 \rangle$	1/14	7/14	8/14
5	$\langle x_3, y_1 \rangle$	1/14	8/14	9/14
6	$\langle x_3, y_2 \rangle$	1/14	9/14	10/14
7	$\langle x_4, y_4 \rangle$	1/14	10/14	11/14
8	$\langle x_5, y_1 \rangle$	1/14	11/14	12/14
9	$\langle x_5, y_3 \rangle$	1/14	12/14	13/14
10	$\langle x_5, y_5 \rangle$	1/14	13/14	14/14

Table 3

Assignment of Random Numbers To Various Cells

Nonzero Cell No. → / Random Numbers ↓	1	2	3	4	5	6	7	8	9	10
.4856			✓							
.1366	✓									
.4486			✓							
.8751									✓	
.1459	✓									
.6508						✓				
.7308							✓			
.2966		✓								
.9738										✓
.8602									✓	
.1376	✓									
.7049						✓				
.8909									✓	
.04084	✓									
Total Counts	4	1	2	0	0	2	1	0	3	1

TABLE 4

The Bivariate Frequency Table For the AUXFEED Example

(A 36 Cell Table)

		X						
		X_1=1.0	X_2=0.75	X_3=0.50	X_4=0.25	X_5=0.05	X_6=0.0	TOTAL
Y	Y_1=1.0	2	6	2	0	0	0	10
	Y_2=0.75	0	0	0	3	0	0	3
	Y_3=0.5	0	0	0	0	0	0	0
	Y_4=0.25	0	0	0	0	0	0	0
	Y_5=0.05	0	0	0	0	0	0	0
	Y_6=0.0	0	0	0	0	0	0	0
	TOTAL	2	6	2	3	0	0	13