DEPENDENCE ANALYSIS OF RANDOM VARIABLES, WITH HYDROLOGIC APPLICATIONS

By

J. Reimann

Department of Civil Engineering Mathematics, Technical University, Budapest, H-1521.

(Received: July 14, 1981)

Summary

Instead of the usual correlation coefficient for determining the closeness and monotony of stochastic relationship between random variables, the indicator correlation is suggested, of numerical values easy to calculate at quantile curve points.

The quantile curve is suggested for approximating the functional relationship between the two random variables. Certain optimum properties of the quantile curve are pointed out.

For a monotonous functional relationship between two random variables of exponential distribution, the quantile curve representing the relationship will be shown to coincide with the line of orthogonal regression.

Hydrology is often facing the need of easy and fast information on the stochastic connection of random variables in order to foresee and forecast certain phenomena and to make correct decisions.

Quantity and quality characteristics of rivers and lakes are random variables. Some pairs of random variables are in functional relation, while other pairs have some tendencies in common, are in stochastic relationship, so-called correlation (in a wide sense). The theoretical and practical importance of the analysis of stochastic dependence of pairs of random variables is seen by the great many relevant publications in the last decades.

The examination of the independence—dependence conditions between two or more random variables has a particular importance, when

1. the values of one variable are easier to measure than those of the other;

2. the values of one variable can be observed earlier in time than those of the others.

Case 1 could be exemplified by the water stage/discharge relation. Obviously, stage is much simpler to measure than is discharge.

Another example for case 1 is to examine the relationship between water quality characteristics. Certain quality characteristics are easy and fast to measure instrumentally, others being accessible to laboratory chemical analysis.

For instance, an eventual close relation between electrical conductivity and the quantity of solved mineral salts permits to assess the total quantity of solved mineral salts from conductometry indication prior to laboratory analysis, etc.

The reader is thought to be able to give as many similar examples as can the author.

As an example for case 2, an eventual close relationship between flood levels and runoff times of a given river permits to conclude from the peak value on the approximate duration of the runoff.

The majority of sequences of basic hydrologic variables are discrete sequences, obtained by observations at discrete times. These time series consist of dependent random variables. With increasing time interval the dependence between the members of time series generally much decreases. It is of interest to determine the greatest interval where data are still dependent, decisive for the time range of reliable forecasting from the time series. In other words the memory of the time series is as long as the elements of the time series are dependent, as far a forecast is possible. In simulating a given hydrological time series by means of a Markov-chain, the order of the chain is determined by the memory of the time series. The memory of the time series is just as important for the application of other models e.g. autoregressive processes.

Now let us see how to measure the closeness of the relation between two random variables X and Y.

B. Schweitzer and E. F. Wolf [5] specify the following conditions for a reasonable set of desiderata for a symmetric, nonparametric measure of dependence $R(X, Y)$ for two continuously distributed random variables X and Y.

(A) $R(X, Y)$ is defined for any X and Y;

(B) $R(X, Y) = R(Y, X)$;

(C) $0 \leq R(X, Y) \leq 1$;

(D) $R(X, Y) = 0$ if and only if X and Y are independent;

(E) $R(X, Y) = 1$ if and only if each of X, Y is a strictly monotonous function of the other;

(F) if f and g are strictly monotonous in ranges X and Y, respectively, then $R[f(X), g(Y)] = R(X, Y)$;

(G) if the distribution of X and Y is bivariate normal, with correlation coefficient r, then $R(X, Y)$ is a strictly increasing function of $|r|$;

(H) if (X, Y) and (X_n, Y_n) $(n = 1, 2, \ldots)$ are pairs of random variables with joint distributions H and H_n, respectively, and if the sequence H_n converges weakly to H, then $\lim\limits_{n \to \infty} R(X_n Y_n) = R\,(X, Y)$.

These conditions are some modifications of the axiom of Rényi [4] which seems to be too strong at least for nonparametric measures of dependence.

E. L. LEHMAN [2] pointed out certain new directions in the domain of stochastic dependence, introducing the quadrant dependence between random variables X and Y in the following way.

Let us compare the probability of any quadrant $X < x$, $Y < y$ under the distribution $H(x, y)$ of (X, Y) with the corresponding probability in the case of independence. The pair (X, Y) is positively quadrant dependent if

$$P(X < x, Y < y) \geq P(X < x) P(Y < y) \tag{1}$$

for all x, y.

The dependence is strict if inequality holds for at least some pair (x, y).

Similarly (X, Y) is negatively quadrant dependent if (1) holds with the inequality sign reversed.

Have a closer look at the meaning of quadrant dependence.

On the basis of (1)

$$\frac{P(X < x, Y < y)}{P(X < x)} = P(Y < y \mid X < x) \geq P(Y < y). \tag{2}$$

This means that the conditional probability of $(Y < y)$ provided $(X < x)$ exceeds the unconditional one of event $(Y < y)$, i.e. small values of X tend to be associated with small values of Y and large values of X tend to be associated with large values of Y. In case of negative dependence, large values of one variable tend to be associated with small values of the other.

The set of inequalities (1) is equivalent to each of the following:

$$\begin{aligned} P(X < x, Y \geq y) &\leq P(X < x) \cdot P(Y \geq y) \\ P(X \geq x, Y < y) &\leq P(X \geq x) \cdot P(Y < y) \\ P(X \geq x, Y \geq y) &\geq P(X \geq x) \cdot P(Y \geq y). \end{aligned} \tag{3}$$

If a statistical sample is available for the related values (X_1, Y_1), $(X_2, Y_2), \ldots$ $\ldots (X_n, Y_n)$ of random variables X and Y to be plotted as a set of points in the plane, then in case of positive quadrant dependence the following pattern of points is obtained (Fig. 1). Let the continuous distribution functions of X and Y be $F(x)$ and $G(y)$, resp., and the joint distribution function $H(x, y)$

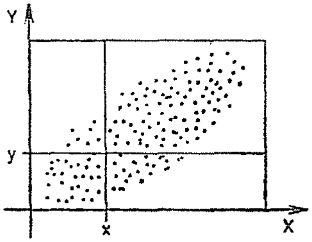

Fig. 1. Typical shape of the point pattern in the case of positive quadrant dependence

then the positive quadrant dependence — according to (1) — between the distribution functions is expressed by

$$H(x, y) \geq F(x)G(y). \tag{4}$$

Similarly, the negative quadrant dependence:

$$H(x, y) \leq F(x)G(y). \tag{5}$$

Knowledge of a positive (or negative) quadrant dependence between random variables X and Y offers some information on the connection between the two variables. In practice, the question arises whether there is a positive or negative quadrant dependence between X and Y or not, and if the answer is yes, how this dependence can be measured and how close it is.

Practical analyses start from the two-dimensional sample (X_1, Y_1), (X_2, Y_2), . . ., (X_n, Y_n) to be plotted as a set of points in the plane, then medians (m_1, m_2) of the distributions of X and Y are calculated by drawing the following figure:

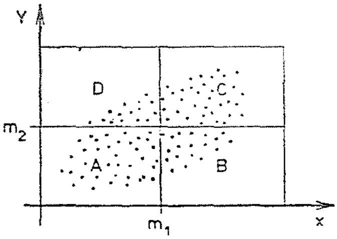

Fig. 2. Division into four fields of the value range of the two-dimensional sample (X_1, X_1), (X_2, X_2), . . ., (X_n, X_n) by medians m_1 and m_2

In case of positive quadrant dependence, more points are found in quadrants A and C than in B and D. In case of independence, about equal numbers of points fall into each quadrant, for $P(X < M_1) = \frac{1}{2}$, $P(Y < M_2) = \frac{1}{2}$, and thus $P(A) = P(B) = P(C) = P(D) = \frac{1}{4}$, where M_1 and M_2 are theoretical medians.

To measure the closeness of positive association, the following measure is formed:

$$\delta(M_1, M_2) = P(A + C) - P(B + C) = P(A + C) - 1 - P(A + C) =$$
$$= 2P(A + C) - 1.$$

Since $P(A) + P(B) = \frac{1}{2}$ (by definition of M_2), furthermore: $P(B) + P(C) = \frac{1}{2}$ (by definition of M_1) therefore $P(A) = P(C)$, i.e.

$$\delta(M_1, M_2) = 4P(A) - 1. \tag{6}$$

The practical approximate determination of the measure (M_1, M_2) is rather simple in possession of empirical medians M_1, M_2. All to be done is to count the points falling in set A. For a number k, the statistical estimate of $\delta(M_1, M_2)$ is:

$$\hat{\delta}(M_1, M_2) = 4\frac{k}{n} - 1.$$

Mosteller was the first to measure $\delta(M_1, M_2)$ defined by Eq. (6) statistical features of which were investigated by Blomqvist [1].

It certainly did not escape the attentive reader that the measure $\delta(M_1, M_2)$ does not measure the quadrant dependence in general but only a special case, the median dependence. For practical applications it is generally sufficient to get a fast look at the dependence conditions, namely the quadrant dependence cannot be tested at every point (x, y) of the plane. It is, however, more satisfactory to check the dependence in several, rather than in a single point.

To this end the measure defined by Eq. (6) will be generalized and the derived mesure will be seen to be easy to calculate along points of a properly selected plane curve.

Another measure for the relationship closeness between random variables

As a measure of the stochastic dependence between two random variables X, Y the concept of indicator correlation will be introduced. Let the distribution functions of random variables X and Y be $F(x)$, and $G(y)$, respectively. Both are assumed to be strictly monotonic, continuous functions. The joint distribution function of X and Y is denoted $H(x, y)$.

Let us introduce the indicator variables ξ_x and η_y for the fixed pair of values (x, y).

$$\xi_x = \begin{cases} 1 & \text{if} \quad X < x \\ 0 & \text{if} \quad X \geq x \end{cases}$$

$$\eta_y = \begin{cases} 1 & \text{if} \quad Y < y \\ 0 & \text{if} \quad Y \geq y. \end{cases}$$

Correlation coefficient of indicator variables ξ_x and η_y is obtained as

$$\varrho(\xi_x, \eta_y) = \frac{E(\xi_x, \eta_y) - E(\xi_x)E(\eta_y)}{D(\xi_x)D(\eta_y)} = \\ = \frac{H(x, y) - F(x)G(y)}{\sqrt{F(x)[1 - F(x)]G(y)[1 - G(y)]}} = \tilde{\varrho}(x, y). \tag{8}$$

Correlation coefficient $\tilde\varrho(x, y)$ is called the indicator correlation of random variables X and Y with respect to the pair of values (x, y). As $\tilde\varrho(x, y)$ may have a different value in every point (x, y) of the plane, it cannot be directly applied to measure the closeness of relationship between variables X and Y.

Calculating, however, the $\tilde\varrho(x, y)$ values at certain points along a properly selected plane curve, a rather informative measure for the relationship between the two variables will be seen to result.

Be $\tilde x_\alpha$ the α-quantile of distribution of variable X such that:

$$P(X < \tilde x_\alpha) = F(\tilde x_\alpha) = \alpha.$$

Again, be $\tilde y_\alpha$ the α-quantile of distribution of variable Y such that:

$$P(Y < \tilde y_\alpha) = G(\tilde y_\alpha) = \alpha.$$

In particular $\tilde x_{\frac{1}{4}}$, and $\tilde x_{\frac{3}{4}}$ are lower, and upper quantiles of the distribution of X, respectively, and $\tilde x_{\frac{1}{2}}$ is the median. Remind that knowledge of the quantiles is rather informative of the distribution.

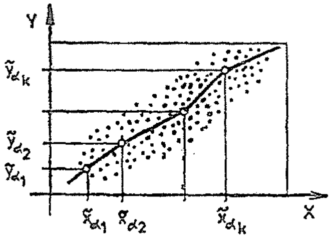

Fig. 3. Empirical quantile curve connecting the identical quantile values $(x_{\alpha_i}, y_{\alpha_i})$

The concept to be introduced will often be referred to in the following. Let us consider the set of points defined by quantile point pairs $(\tilde x_\alpha, \tilde y_\alpha)$ travelling α through interval $(0, 1)$ to be called quantile curve. In practice this curve can be approximately imaged by assuming to have a two-dimensional sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ for associated values of random variables X and Y, to be represented by a cluster of points.

Theoretical quantile values $\tilde x_\alpha$ and $\tilde y_\alpha$ are approximated by empirical quantiles from the sample and points $(\tilde x_{\alpha_i} \tilde y_{\alpha_i})$ belonging to the same α_i value connected. (The resulting curve is only a statistical approximation of the quantile curve, but in practical analysis it is current and generally sufficient.) By the time, distributions $F(x)$ and $G(y)$ of variables X and Y are assumed to be known, and so are quantiles $\tilde x_\alpha, \tilde y_\alpha$. In this case, equation of the quantile curve is easy to establish using distribution functions $F(x)$ and $G(y)$, namely:

$$F(\tilde x_\alpha) = G(\tilde y_\alpha) = \alpha$$

$$\tilde y_\alpha = G^{-1}(\alpha) = G^{-1}[F(\tilde x_\alpha)]$$

yielding for the quantile curve

$$y = G^{-1}[F(x)]. \tag{9}$$

Let us point out a rather important feature of the quantile curve. If there exists a strictly monotonous increasing continuous functional relationship between X and $Y : Y = \varphi(X)$, then $\tilde{y}_\alpha = \varphi(\tilde{x}_\alpha)$. Thus, with the given stipulations for a monotonous increasing functional relationship between two random variables the curve of the function and the quantile curve are identical. To prove this statement is rather simple, namely

$$\alpha = P(Y < \tilde{y}_\alpha) = P(\varphi(X) < \tilde{y}_\alpha) = P(X < \varphi^{-1}(\tilde{y}_\alpha))$$

as $P(X < \tilde{x}_\alpha) = \alpha$ and thus, with the stipulations:

$$\varphi^{-1}(\tilde{y}_\alpha) = \tilde{x}_\alpha$$

$$\tilde{y}_\alpha = \varphi(\tilde{x}_\alpha). \tag{10}$$

(10) is of importance for finding the functional relationship between two random variables.

Let the value of the indicator correlation $\tilde{\varrho}(x, y)$ be calculated at points of the quantile curve:

$$\tilde{\varrho}_\alpha = \tilde{\varrho}(\tilde{x}_\alpha, \tilde{y}_\alpha) = \frac{H(\tilde{x}_\alpha, \tilde{y}_\alpha) - F(\tilde{x}_\alpha)G(\tilde{y}_\alpha)}{\sqrt{F(\tilde{x}_\alpha)[1 - F(\tilde{x}_\alpha)]G(\tilde{y}_\alpha)[1 - G(\tilde{y}_\alpha)]}} = \frac{H(\tilde{x}_\alpha, \tilde{y}_\alpha) - \alpha^2}{\alpha - \alpha^2}. \tag{11}$$

Note that the quantile curve in the form above holds for a positive association. In case of a negative association, quantile values $(\tilde{x}_\alpha, \tilde{y}_{1-\alpha})$ are connected, and then

$$\tilde{\varrho}_\alpha = \frac{H(\tilde{x}_\alpha, \tilde{y}_{1-\alpha}) - \alpha(1 - \alpha)}{\alpha - \alpha^2}. \tag{12}$$

Practical utility of indicator correlation $\tilde{\varrho}_\alpha$ can be appreciated by examining some characteristics.

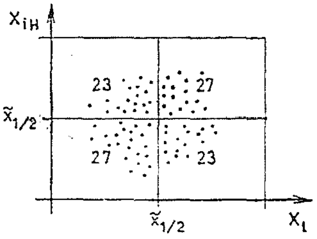

Fig. 4. Independence test by indicator correlation for the annual highest stages on the Tisza River

a) Evidently, $-1 \leq \tilde{\varrho}_\alpha \leq 1$ as it is a special correlation coefficient. Thus $|\tilde{\varrho}_\alpha| \leq 1$;

b) $\tilde{\varrho}_\alpha = 0$ if and only if $H(\tilde{x}_\alpha, \tilde{y}_\alpha) = F(\tilde{x}_\alpha)G(\tilde{y}_\alpha) = \alpha^2$, that is, iff X and Y are independent along the quantile curve. Thereby the indicator correlation is superior to the usual correlation coefficient;

c) for an arbitrary, monotonous functional relationship between X and Y, $|\tilde{\varrho}_\alpha| = 1$. Namely, let $Y = \varphi(X)$, then Eq. (10) leads to:

$$H(\tilde{x}_\alpha, \tilde{y}_\alpha) = P(X < \tilde{x}_\alpha, Y < \tilde{y}_\alpha) = P\big(X < \tilde{x}_\alpha, \varphi(X) < \varphi(\tilde{x}_\alpha)\big) = P(X < \tilde{x}_\alpha) = \alpha.$$

Also the inverse of the statement is valid, as the value is 1 only if $H(\tilde{x}_\alpha, \tilde{y}_\alpha) = \alpha$.

d) The indicator correlation is invariant against the concordant monotonous transformation (either both monotonous increasing or both monotonous decreasing) of variables.

Let $U = f(X)$, $V = g(Y)$, where f and g are strict monotonous functions. Then $\tilde{u}_\alpha = f(\tilde{x}_\alpha)$, $\tilde{v}_\alpha = g(\tilde{y}_\alpha)$

$$P(U < \tilde{u}_\alpha, V < \tilde{v}_\alpha) = P\big(F(x) < f(\tilde{x}_\alpha)g(Y) < g(\tilde{y}_\alpha)\big) = P(X < \tilde{x}_\alpha, Y < \tilde{y}_\alpha)$$

$$P(U < \tilde{u}_\alpha) = P\big(f(X) < f(\tilde{x}_\alpha)\big) =$$

$$= P(X < \tilde{x}_\alpha);\ P(V < \tilde{v}_\alpha) = P\big(g(Y) < g(\tilde{y}_\alpha)\big) = P(Y < \tilde{y}_\alpha).$$

The values in Eq. (11) can only be calculated if joint distribution function $H(x, y)$ of variables X and Y is known. In practice it is seldom known, therefore it should be estimated by statistic means

$$\hat{\varrho}_\alpha = \frac{\dfrac{k}{n} - \alpha^2}{\alpha - \alpha^2} \tag{13}$$

where $\dfrac{k}{n}$ is the relative frequency of points within quadrant $(\tilde{x}_\alpha, \tilde{y}_\alpha)$.

A statistical value of about zero hints to the independence of variables X and Y, while a value close to 1 predicts some monotonous functional relationship. The question arises, at what numerical value of $\hat{\varrho}_\alpha$ can X and Y be considered to be independent. Remind that in the case of independence

$$E\left(\frac{k}{n}\right) = \alpha^2$$

$$D\left(\frac{k}{n}\right) = \sqrt{\frac{\alpha^2(1-\alpha^2)}{n}} = \frac{\alpha}{\sqrt{n}}\sqrt{1-\alpha^2}.$$

From the *Moivre—Laplace* limit distribution theorem, in the case of independence:

$$P\left(\left|\frac{k}{n} - \alpha^2\right| \geq \frac{2\alpha}{\sqrt{n}}\sqrt{1-\alpha^2}\right) \simeq 0.05. \tag{14}$$

Hence from Eq. (13):

$$P\left(\left|\frac{\frac{k}{n}-\alpha^2}{\alpha-\alpha^2}\right| \geq \frac{2\alpha}{\sqrt{n}}\,\frac{\sqrt{(1+\alpha)(1-\alpha)}}{\alpha(1-\alpha)}\right) = P\left(|\hat{\varrho}_\alpha| \geq \frac{2}{\sqrt{n}}\sqrt{\frac{1+\alpha}{1-\alpha}}\right). \tag{15}$$

For $\alpha = \frac{1}{2}$

$$\hat{\varrho}_{\frac{1}{2}} = \frac{\frac{k}{n}-\frac{1}{4}}{\frac{1}{2}-\frac{1}{4}} = 4\frac{k}{n}-1 \tag{16}$$

the same as measure (M_1, M_2) in (7) examined by Blomqvist for statistical features. In case of $\alpha = \frac{1}{2}$ according to (15):

$$P\left(|\hat{\varrho}_{\frac{1}{2}}| < 2\sqrt{\frac{3}{n}}\right) = 0.95.$$

As a practical application of the above, let the independence of stage maxima of the River *Tisza* measured at *Szeged* in the period 1876 to 1975 be examined by means of the empirical indicator correlation, Eq. (13).

Be the sequence of yearly maxima at *Szeged*:

$$X_1, X_2, \ldots, X_{100}.$$

Forming continuous pairs (X_1, X_2); (X_2, X_3); (X_3, X_4); . . . ; (X_{99}, X_{100}) and plotting them as a set of in-plane points yields in final account:

The median of the time sequence $\tilde{x}_{\frac{1}{2}} = 648$ cm. The numbers of points in each quadrant are seen in Fig. 5. Based on Eq. (16):

$$\hat{\varrho}_{\frac{1}{2}} = 4\frac{27}{100} - 1 = 0.08.$$

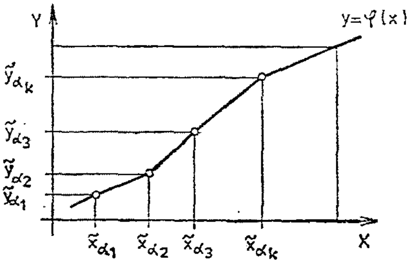

Fig. 5. Approximate determination of the monotonous functional relationship between random variables with the help of the quantile curve

The critical value in Eq. (14) — about corresponding to the level 0.05 — is:

$$\frac{2}{\sqrt{n}}\sqrt{\frac{1+\alpha}{1-\alpha}}=\frac{2\sqrt{3}}{10}\approx 0.346.$$

Since the observed value of $\hat{\varrho}_{\frac{1}{2}}$ is far below the critical level, there is no reason to reject the hypothesis of independence.

	cm		cm		cm		cm		cm
1876	786	1900	525	1924	870	1948	714	1972	606
1877	795	1901	680	1925	681	1949	495	1973	475
1878	720	1902	668	1926	759	1950	517	1974	807
1879	806	1903	508	1927	477	1951	550	1975	692
1880	627	1904	450	1928	542	1952	648		
1881	845	1905	518	1929	458	1953	706		
1882	691	1906	550	1930	496	1954	454		
1883	728	1907	758	1931	603	1955	657		
1884	613	1908	595	1932	923	1956	689		
1885	565	1909	642	1933	660	1957	604		
1886	534	1910	496	1934	526	1958	730		
1887	660	1911	563	1935	594	1959	436		
1888	847	1912	753	1936	472	1960	582		
1889	805	1913	802	1937	703	1961	394		
1890	566	1914	778	1938	638	1962	820		
1891	668	1915	791	1939	579	1963	587		
1892	630	1916	791	1940	847	1964	764		
1893	726	1917	514	1941	855	1965	748		
1894	568	1918	349	1942	780	1966	799		
1895	884	1919	916	1943	366	1967	790		
1896	525	1920	708	1944	654	1968	600		
1897	730	1921	325	1945	560	1969	626		
1898	604	1922	774	1946	525	1970	961		
1899	460	1923	637	1947	602	1971	521		

Derivation of functional relationship from measurements

The relation between two random variables X and Y is usually characterized by the conditional expected value curves $E(X|Y)$ or $E(Y|X)$, the so-called regression curves. In practice often linear regression is made up with, that is, finding the straight line best fitting the two-dimensional set of points

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

by means of the least squares' method.

To discover the relevant structural relationship, WALD suggests to draw a straight across the centroids of two half parts of the two-dimensional sample.

The method presented below for examining regression in a wider sense (termed by *Wald* "structural relationship") has the advantages over *Wald*'s method, partly of being applicable in case of other than linear regression and even to give the approximate shape of the regression curve, and partly, of requiring little computing work.

Our procedure gives the structural relationship between random variables X and Y by the curve connecting the quantile points $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ and $(\tilde{x}_\alpha, y_{1-\alpha})$ in case of positive, and negative association, respectively (quantile curve). As we have seen, if there is a strictly monotonous functional relationship $Y = \varphi(X)$ between X and Y, then the curve $y = \varphi(x)$ coincides with the quantile curve, provided distribution functions of the random variables are strictly monotonous continuous functions.

Knowledge of the marginal distribution functions $F(x)$ and $G(y)$ of the pair (X, Y) permits to compute quantile values $\tilde{x}_{\alpha_1}, \tilde{x}_{\alpha_2}, \ldots, \tilde{x}_{\alpha_k}$ and $\tilde{y}_{\alpha_1}, \tilde{y}_{\alpha_2}, \ldots$ $\ldots, \tilde{y}_{\alpha_k}$ for variables X and Y, respectively.

In case of positive association forming pairs $(x_{\alpha_1}, \tilde{y}_{\alpha_1}), (\tilde{x}_{\alpha_2}, \tilde{y}_{\alpha_2}), \ldots, (\tilde{x}_{\alpha_k}, \tilde{y}_{\alpha_k})$ and connecting them by straight sections results in the approximation of the quantile curve (Fig. 5).

The procedure is the same for negative association where the quantile curve crosses points $(\tilde{x}_{\alpha_i}, \tilde{y}_{1-\alpha_i})$. Provided neither single-variable distribution functions $F(x)$ and $G(y)$ of random variables X and Y are known, the corresponding sample quantiles, that is, the empirical quantiles of ordered samples are:

$$X_1^* < X_2^* < \ldots < X_n^*; \quad Y_1^* < Y_2^* < \ldots < Y_n^*.$$

As the sample quantiles are unbiased estimates of theoretical quantiles, the curve constructed as above from empirical quantiles of a sufficiently high sample number is a good approximation of the theoretical quantile curve.

One may wonder how the defined quantile curve can fit the two-dimensional set of sample points. Is there some optimal property of the quantile curve similar to that of the regression curve? Regression curve $E(Y|X = x)$ of variable Y on X is known to be a conditional expected value curve with the optimal property of minimum variance of corresponding Y values in case of $X = x$ with the same abscissa. The regression curve $E(X|Y = y)$ has a similar property. Also the quantile curve has a certain optimal property for both random variables X and Y simultaneously, resulting from its construction principle. The quantile curve, set of points $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ where α is passing through the interval $(0, 1)$ is a curve where points of the two-dimensional sample exhibit a minimum weighted average of absolute deviation. For a given point $(\tilde{x}_\alpha, \tilde{y}_\alpha)$, relationships

$$(1-\alpha)\int_{-\infty}^{\tilde{x}_\alpha} |x - \tilde{x}_\alpha| f(x)\,dx + \alpha \int_{\tilde{x}_\alpha}^{\infty} |x - \tilde{x}_\alpha| f(x)\,dx = \min$$

and

$$(1-\alpha)\int_{-\infty}^{y_\alpha}|y-\tilde{y}_\alpha|g(y)\,dy+\alpha\int_{\tilde{y}_\alpha}^{\infty}|y-\tilde{y}_\alpha|g(y)\,dy=\min \tag{17}$$

are simultaneously met.

To prove (17) let us find the number z such that

$$(1-\alpha)\int_{-\infty}^{z}(z-x)f(x)\,dx+\alpha\int_{z}^{\infty}(x-z)f(x)\,dx=\min.$$

Hence

$$\int_{-\infty}^{z}zf(x)\,dx-\int_{-\infty}^{Z}xf(x)\,dx-\alpha\int_{-\infty}^{z}zf(x)\,dx+\alpha\int_{z}^{\infty}xf(x)\,dx+$$
$$+\alpha\int_{-\infty}^{z}xf(x)\,dx-\alpha\int_{z}^{\infty}zf(x)\,dx=\min$$

that is

$$zF(z)-\alpha zF(z)+\alpha m-\int_{-\infty}^{z}xf(x)\,dx-\alpha\int_{z}^{\infty}zf(x)\,dx=$$
$$=zF(z)-\alpha zF(z)-\alpha z+\alpha zF(z)+\alpha m-$$
$$-\int_{-\infty}^{z}fx(x)\,dx=\min\quad\big(\text{where } m=E(X)\big).$$

Introducing function

$$\varphi(z)=zF(z)-\alpha(z-m)-\int_{-\infty}^{z}xf(x)\,dx$$

there may be a minimum only where

$$\varphi'(z)=F(z)+zf(z)-\alpha-zf(z)=0$$

end

$$F(z)=\alpha,\ z=F^{-1}(\alpha)=\tilde{x}_\alpha.$$

Of course, an analog relationship results for $\tilde{y}_\alpha$, α-quantile of random variable Y.

Monotonous relationship between random variables of exponential distribution

For certain hydrological applications functional relationship between exponentially distributed random variables is of special interest.

For example, in flood hydrology the excess over a given (sufficiently high) difference between peak and c-level is exponentially distributed, and so is the runoff time (Fig. 6).

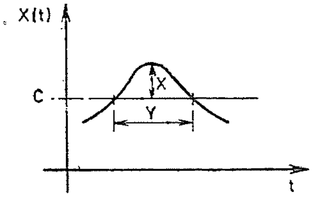

Fig. 6. An (approximately) linear correlation exists within a given flood wave between the magnitude X of exceeding the warning level c and the length X of the flood wave

Be X and Y random variables of exponential distribution, with distribution functions:

$$X : F(x) = 1 - \bar{e}^{\alpha x}$$

$$Y : G(y) = 1 - \bar{e}^{\beta y}.$$

Assume a monotonous (increasing) relationship $Y = \varphi(X)$ between X and Y, where $\varphi(x)$ is a continuous function. In this case function $y = \varphi(x)$ can only be linear:

$$y = \varphi(x) = \frac{E(Y)}{E(X)} x = \frac{\alpha}{\beta} x. \tag{18}$$

This statement, however surprising at the first glance, is easy to understand because in case of exponentially distributed random variables a monotonous relationship is necessarily rectilinear represented by a line passing through the origin with a slope equal to the ratio of the expected values.

In conformity with (10), a monotonous increasing relationship existing between two random variables is coincident with the quantile curve, hence from Eq. (9):

$$\varphi(x) = G^{-1}[F(x)].$$

If

$$z = G(y) = 1 - \bar{e}^{\beta y}$$

then

$$y = G^{-1}(z) = -\frac{1}{\beta} \ln (1 - z)$$

$$y = G^{-1}[F(x)] = -\frac{1}{\beta} \ln (1 - 1 + \bar{e}^{\alpha x}) = \frac{\alpha}{\beta} X = \frac{E(Y)}{E(X)} x.$$

(Note: if distribution medians of X and Y are $\tilde{x}_{\frac{1}{2}}$ and $\tilde{y}_{\frac{1}{2}}$, resp., then

$$y = \frac{E(Y)}{E(X)} x = \frac{\tilde{y}_{\frac{1}{2}}}{\tilde{x}_{\frac{1}{2}}} x.$$

Namely for $F(x) = 1 - \bar{e}^{\alpha x}$ the $\tilde{x}_{\frac{1}{2}}$ value becomes:

$$\bar{e}^{\alpha x} = \frac{1}{2}$$

$$\tilde{x}_{\frac{1}{2}} = \frac{\ln 2}{\alpha}, \text{ similarly } \tilde{y}_{\frac{1}{2}} = \frac{\ln 2}{\beta}$$

i.e.

$$\frac{\tilde{y}_{\frac{1}{2}}}{\tilde{x}_{\frac{1}{2}}} = \frac{\frac{\ln 2}{\beta}}{\frac{\ln 2}{\alpha}} = \frac{\frac{1}{\beta}}{\frac{1}{\alpha}} = \frac{\alpha}{\beta}.$$

Orthogonal regression for exponentially distributed random variables

Be (X, Y) an arbitrary in-plane point of joint values of random variables X and Y.

Let us find a line L for which the expected value of the square normal distance d of random point (X, Y) is minimum (Fig. 7).

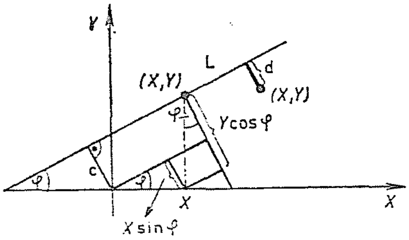

Fig. 7. Determination of the orthogonal regression straight line L

For points (X, Y) on a line L at a distance c from the origin including angle φ with the positive X-axis:

$$Y \cos \varphi - X \sin \varphi = c. \tag{19}$$

In general, an arbitrary point (X, Y) is not on line L, so values φ and c are wanted for solving the problem

$$E(d^2) = E(Y \cos \varphi - X \sin \varphi - c)^2 = \min.$$

Let

$$f(\varphi, c) = E(Y^2) \cos^2 \varphi - E(XY) \sin 2\varphi + E(X^2) \sin^2 \varphi -$$
$$- 2E(Y)c \cos \varphi + 2E(X)c \sin \varphi + c^2$$

$$\frac{\partial f}{\partial c} = -2E(Y)\cos\varphi + 2E(X)\sin\varphi + 2c = 0$$

that is

$$E(Y)\cos\varphi - E(X)\sin\varphi = c.$$

Line L passes through point $E(x) = m_1$, $E(y) = m_2$, that is, from (19):

$$L = (Y - m_2^2)\cos\varphi - (X - m_1)\sin\varphi = 0.$$

The minimum problem to be solved is:

$$f(\varphi) = E(d^2) = E[(Y - m_2)\cos\varphi - (X - m_1)\sin\varphi]^2 = \min$$

i.e.,

$$f(\varphi) = \sigma_1^2\sin^2\varphi - 2\varrho\sigma_1\sigma_2\sin\varphi\cos\varphi + \sigma_2^2\cos^2\varphi = 0$$

$$\frac{\partial f}{\partial\varphi} = \sigma_1^2\sin 2\varphi - 2\varrho\sigma\sigma_2\cos 2\varphi - 2\sigma_2^2\sin 2\varphi = 0$$

that is:

$$\operatorname{tg} 2\varphi = \frac{2\varrho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2}$$

yielding the slope; hence the equation for line L is:

$$y - m_2 = \frac{2\varrho\sigma_1\sigma_2}{\sigma_1^2 - \sigma_2^2 + \sqrt{(\sigma_1^2 - d^2)^2 + 4\varrho^2\sigma_1\sigma_2}}\,(X - m_1). \tag{20}$$

If random variables X and Y are exponentially distributed with distribution functions $F(x) = 1 - \bar{e}^{\alpha x}$ and $G(y) = 1 - \bar{e}^{\beta y}$, resp., furthermore there is a monotonic functional relationship between them, then $Y = \varphi(X)$ is linear i.e. Eq. (20) becomes:

$$y - \frac{1}{\beta} = \frac{2\,\frac{1}{\alpha}\,\frac{1}{\beta}}{\frac{1}{\alpha^2} - \frac{1}{\beta^2} + \sqrt{\left(\frac{1}{\alpha^2} + \frac{1}{\beta^2}\right)^2}}\left(x - \frac{1}{\alpha}\right) = \frac{\alpha}{\beta}x - \frac{1}{\beta}$$

that is,

$$y = \frac{\alpha}{\beta}x,$$

exactly the quantile curve.

As a conclusion, the line of orthogonal regression between exponentially distributed random variables needs not be computed since the quantile curve equation can be directly obtained in knowledge of expected values, or the empirical quantile curve is simply produced from the sample.

References

1. Blomqvist, N.: On a Measure of Dependence Between two Random Variables. Ann. Math. Statist. Vol. 21. (1950).
2. Lehmann, E. L.: Some Concept of Dependence. The Annals of Math. Statist. Vol 37. (1966)
3. Reimann, J.: Mathematical-Statistical Analysis of Flood Characteristics. (In Hungarian). Hidrológiai Közlöny, 4. 1974.
4. Rényi, A.: On Measures of Dependence. Acta Mathematica X/3—4 (1960)
5. Schweizer, B.—Wolff, E. F.: On Nonparametric Measures of Dependence for Random Variables. The Annals of Statistics Vol. 9. No. 4. 1981.
6. Wald, A.: The Fitting of Straight Lines if both Variables are Subject to Error. Ann. Math. Statist. Vol. 11. (1940).

Prof. Dr. József Reimann, H-1521, Budapest