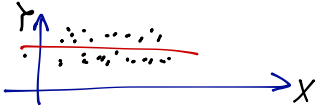
$0 \le \rho^2 \le 1 - \frac{E[Y - E[Y|X]]^2}{E[Y - E[Y]]^2} =: \eta \le 1$ ← quotient



If Y and X independent
top < bot , $\therefore \rho^2 < 1$

Lecture 17

6.1. Measures of association for continuous observations. Tests of independence "near" normal random variables.

It is well known that the **correlation coefficient** for two random variables X and Y is defined as $\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{E[(X-E(X))(Y-E(Y))]}{\sqrt{E(X-E(X))^2 E(Y-E(Y))^2}}$. It is also known that $-1 \le \rho \le 1$ (this is precisely the Cauchy-Schwartz inequality) and that this coefficient indicates how strong the linear stochastic relationship between X and Y is: the closer in absolute value to one the coefficient, the stronger the relationship. Value of $\rho = \pm 1$ indicates a perfect linear relationship, i.e. existence of deterministic coefficients a and b such that $X = aY + b$ holds. In the opposite direction, when $\rho = 0$, we say that X and Y are uncorrelated. This means complete lack of linear relationship (**but does not necessarily mean independence of X and Y**). Note however that if in addition X and Y have a joint multivariate **normal** distribution then $\rho = 0$ also means **independence** of X and Y since for joint multivariate **normal**, the notion of independence and uncorrelatedness coincide. Having a sample of N pairs of observations X_i, Y_i, one can easily **estimate** the correlation coefficient by using the consistent estimator

$$\hat{\rho} = r = \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^N X_i Y_i - N\bar{X}\bar{Y}}{\sqrt{[\sum_{i=1}^N X_i^2 - N(\bar{X})^2][\sum_{i=1}^N Y_i^2 - N(\bar{Y})^2]}}$$

(called also **Pearson's** r.)

The following statistic $T = r\sqrt{\frac{N-2}{1-r^2}}$ can be used for testing significance of ρ. Under $H_0 : \rho = 0$, the distribution of T is Student's T with $df = N - 2$. Hence, rejection region of the test : $H_0 : \rho = 0$ versus $H_1 : \rho \neq 0$ would be $T : |T| > t_{n-2,\alpha/2}$. When the joint distribution of X and Y is multivariate normal, the same test can be used for testing independence of X and Y (because of the mentioned equivalence of independence and uncorrelatedness under normality). More sophisticated reasoning shows that the same test can be used successfully also to test independence (i.e. $H_0 : F_{(X,Y)}(x,y) = F_X(x)F_Y(y)$ for all x, y) also in case of continuous observations that not necessarily come from multivariate normal family. The power of such a test is significantly reduced, however and there arises the question for alternative tests of independence when the observations do not have a joint multivariate normal distribution.

6.2. Rank based correlation coefficients (Spearman's ρ and Kendall's τ).

6.2.1. Spearman's ρ. Spearman introduced his rank-based correlation coefficient (sometimes called **Spearman's** ρ) as a variant of the usual Pearson's r whereby the observations X_i and Y_i are simply replaced by their ranks $R(X_i)$ and $R(Y_i)$, respectively. Hence

$$\rho = \frac{\sum_{i=1}^N R(X_i)R(Y_i) - N(\frac{N+1}{2})^2}{\sqrt{[\sum_{i=1}^N R(X_i)^2 - N(\frac{N+1}{2})^2][\sum_{i=1}^N R(Y_i)^2 - N(\frac{N+1}{2})^2]}}$$

Note however that **if there are no ties**, the denominator is in fact **not a function of the ranks** (WHY (!)) and the calculation can be simplified to a computationally more attractive form:

$$\rho = 1 - \frac{6\sum_{i=1}^N [R(X_i) - R(Y_i)]^2}{N(N^2-1)}$$

$\rho = \frac{12 \sum R(X_i) R(Y_i)}{N(N^2-1)} - \frac{3(N+1)}{N-1}$ $\quad H_0 : \rho = \rho_0$

$(\hat{\rho} - \rho_0) \approx N(0, \frac{1}{N-1})$

$Z = \sqrt{N-1}(\hat{\rho} - \rho_0) \xrightarrow{d} N(0,1)$

$* \frac{N(N^2-1)}{12}$

It should be stressed though that the first formula is the more general one and in case of presence of tied observations, the first, not the second, is the formula that should be applied!

Spearman's rank correlation is often used as a distribution free test statistic to test independence between two random variables X and Y. The exact distribution under the hypothesis of independence is tabulated for $N \leq 30$. For bigger values of N, the approximated $p-value$ are based on a normality approximation. Denote by Z the $N(0,1)$ random variable. Then:

- **Two-tailed Test**. Consider the two-sided alternative H_1 : Either (a) there is a tendency for the larger values of X to be paired with the larger values of Y, or (b) there is a tendency for the smaller values of X to be paired with the larger values of Y. Then p-value$\approx 2P(Z \geq |\rho|\sqrt{N-1})$.
- **Lower-tailed Test for Negative Correlation**. Consider the one-sided alternative H_1 : There is a tendency for the smaller values of X to be paired with the larger values of Y, and vice versa. Then p-value$\approx P(Z \leq \rho\sqrt{N-1})$.
- **Upper-tailed Test for Positive Correlation**. Consider the one-sided alternative H_1 : There is a tendency for the larger values of X and Y to be paired together. Then p-value$\approx P(Z \geq \rho\sqrt{N-1})$.

6.2.2. Kendall's τ. An alternative to Spearman's coefficient has been suggested by Maurice Kendall. For an easier explanation, let us introduce some terminology. Looking at the N pairs (X_i, Y_i) of observations, we shall say that two observations (for example (1.1, 2.3) and (1.8,2.9)) are *concordant* if both members of one observation are larger than their respective members of the other observation. If, on the other hand, the two numbers in one observation differ in opposite directions from the respective members in the other observation (for example (1.3,2.1) and (1.6,1.2)) we shall call such a pair of observations *discordant*. More precisely, this means that a pair of bivariate observations (X_1, Y_1) and (X_2, Y_2) is concordant if $\frac{Y_2-Y_1}{X_2-X_1} > 0$ and discordant if $\frac{Y_2-Y_1}{X_2-X_1} < 0$. Theoretically, for continuous observations there would be no ties and correspondingly there is no chance for the ratio to be zero or undefined. Then if from all $N(N-1)/2$ possible pairs, N_c were concordant, $N_d = N(N-1)/2 - N_c$ would be discordant. Kendall's τ in case of no ties is defined as

$$\tau = \frac{N_c - N_d}{[N(N-1)/2]}$$

If all pairs are concordant it is equal to 1, if all pairs are discordant, it equals -1. There is a clear intuition behind the above definition of τ. You can easily see that it is the empirical estimator of the difference of the probabilities $P_c - P_d$ where

$$P_c = \text{ probability to get a concordant pair;}$$

$P_d =$ probability to get a discordant pair of bivariate observations $(X_i, Y_i), (X_j, Y_j), i \neq j$.

Indeed, if the X and the Y observations are independent then

$$P_c = P((X_i < X_j) \bigcap (Y_i < Y_j)) + P((X_i > X_j) \bigcap (Y_i > Y_j)) =$$

because of the assumed independence $= P(X_i < X_j)P(Y_i < Y_j) + P(X_i > X_j)P(Y_i > Y_j) =$

since X_j and X_i are interchangeable $= P(X_i > X_j)P(Y_i < Y_j) + P(X_i < X_j)P(Y_i > Y_j) =$

again using the independence $= P((X_i > X_j) \bigcap (Y_i < Y_j)) + P((X_i < X_j) \bigcap (Y_i > Y_j) = P_d$

should hold and therefore, the difference $P_c - P_d$ would be equal to zero under the assumption of independence. Therefore, the size of the estimated difference (which, as we have seen, is the meaning of the τ coefficient) could be used to measure deviations from the null hypothesis of independence. If the null hypothesis is rejected, that would imply that the pairs of observations tend to be concordant (if $H_1 : \tau > 0$), or to be discordant (if $H_1 : \tau < 0$), or just to be dependent (if $H_1 : \tau \neq 0$).

Since in practice ties may occur, let us note that if $X_1 = X_2$, no comparison can be made (the denominator would be zero) whereas if $Y_1 = Y_2$ then the ratio is zero and could be considered as 50/50 concordant/discordant. Hence, it makes sense to define τ in all cases (including when there are ties) as:

$$\tau = \frac{N_c - N_d}{N_c + N_d} \qquad -1 \leq \tau \leq 1$$

where all pairs with $X_i \neq X_j$ are compared.

Again, as in Spearman's ρ, tables exist for the distribution of τ under the null hypothesis of independence (and for small values of N.) Asymptotically, the statistic $T = N_c - N_d$ (which is up to a constant the τ statistic) has a limiting normal distribution:

$$T = N_c - N_d \approx N(0, \frac{N(N-1)(2N+5)}{18}) \qquad \left(\frac{N(N-1)}{2}\right)$$

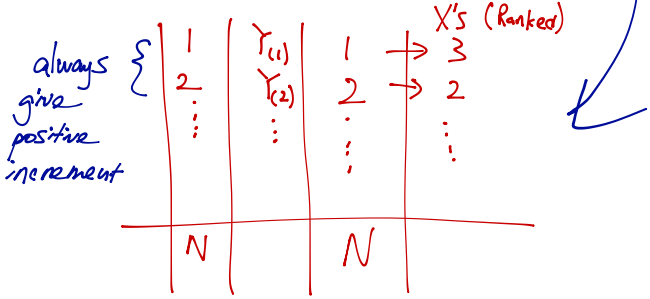
Therefore, testing in the case of each of the three alternatives (Two-tailed, Lower-tailed, and Upper-tailed test) can be worked based on the Normal theory using the fact that under the null hypothesis, $\frac{T\sqrt{18}}{\sqrt{N(N-1)(2N+5)}}$ has a limiting standard normal distribution (details are left for the student to formulate).

tie observation

6.2.3. Some comparisons It should be noted that, as you can easily realize, if the Y_i are ranked in the natural order $1, 2, \ldots, N$ and the ranks of the X_i that correspond to this order, are denoted by R_i then

$$\tau = \frac{1}{N(N-1)} \sum_{i \neq j} \sum sign(i-j)sign(R_i - R_j)$$

holds (WHY(!)). Hence, Kendall's τ is a **nonlinear rank-based** statistic. Spearman's Statistic, on the other hand, can be written as $\sum_{i=1}^{N} a_i R_i + b$ for suitably chosen constants a_i and b (ie., it is a linear rank statistic). (Prove this claim (!)). In a certain sense, it can be shown that Spearman's statistic can be considered to bee the projection of the Kendall coefficient into the family of linear rank statistics. In this sense, some statisticians claim that Kendall's tau is the "more accurate measure" and Spearman's rho is its "simple" approximation; hence they give a preference to Kendall's tau. Experience also shows that



τ is non-linear transformed of the rank based statistic

Spearman's ρ is linear transform of the rank based statistic

the convergence of Kendall's tau to the limiting asymptotic distribution is quicker. This is another reason to prefer Kendall's tau.

6.3. Measures of association. Different definitions of the Contingency coefficient for a contingency table with no ordered categories

As discussed in Lecture XIV, Carl Pearson was the pioneer who introduced the **contingency coefficient** C defined as $C = \sqrt{\frac{Q}{Q+N}}$. Another reason besides intuition to define the contingency coefficient in such a way was that if observations from a two-dimensional multivariate normal distribution with a (usual) correlation coefficient ρ were classified into an $r \times c$ contingency table then it can be shown that $C^2 \to \rho^2$ when both r and c are increased. But since the coefficient C can never be equal to one, some further suggestions have been made later to overcome this difficulty. It was Tschuprov who introduced the coefficient $T = \sqrt{\frac{Q}{N[(r-1)(c-1)]^{1/2}}}$. It can be seen that for a table with $r = c$, this coefficient can be equal to one in the case of perfect dependence. To see this, first note that (as it can be easily seen) the equality $Q = N[\sum_{i,j} \frac{O_{ij}^2}{O_{i.}O_{.j}} - 1]$ holds. Therefore, if for a squared contingency table ($r = c$) we had all frequencies outside the diagonal being zero, we would have $Q = N(r-1)$ and therefore the value of T will be equal to one. Further reasonable definition has been introduced by Cramer in 1946. He defines the coefficient $CR = \sqrt{\frac{Q}{N \min(r-1,c-1)}}$ which **always** can be equal to one. Note that for a squared table, $CR = T$ holds.

Finally, it should be said that hardly any of the coefficients discussed in 6.3, is a very good measure of association. Their main drawback is the lack of a clear probabilistic interpretation (as opposed to the coefficients in 6.1 and 6.2.)

6.4. Measures of association in multiple classifications.

When $k = 2$ sets of paired rank measurements are available and we would like to investigate their association, the methods of Section 6.2 are appropriate. If we assume that $k > 2$ sets of blocked rank measurements are available, the same question might be of interest. We would be interested both in testing the hypothesis that the k sets are independent and in finding a measure that numerically expresses the relationship between the rankings. In the common language of this problem, we are confronted with k **observers**, each of whom is presented with the same set n of objects to be ranked. (Think about a wine tasting experiment in which $k = 4$ experts are asked to taste and rank $n = 7$ different brands of wine each). We would be interested if the rankings of these experts are concordant or discordant. If the experts were only $k = 2$ we could use Kendall's tau to express the concordance quantitatively and could test the significance using the method outlined in section 6.2. How do we proceed if the experts are more than two? One possibility is, of course, to apply Kendall's tau for each paired samples of n ranks in which case we would end up with $k(k-1)/2$ sets of tau-values. However, if $k(k-1)/2$ tests of the null hypothesis of independence are then made, the tests would not be independent and the overall probability of I type error is difficult to keep under control. We need a **single measure** of overall association for the set of **all** k rankings.

We can order the data in a $k \times n$ table with rows and columns designating observers and objects (as opposed to blocks and treatments in Friedman's test of Section 5.3). The table entries R_{ij} represent the rank given by the ith observer to the jth object. The ith row is a permutation of the numbers $1, 2, \ldots, n$ then. The ranks in each column are indicative about the agreement between the observers (the less variable they are, the closer the

agreement between the observes). Under complete agreement of the observers the column totals $R_j, j = 1, 2, \ldots, n$ will be some permutation of the numbers

$$k, 2k, 3k, \ldots, nk.$$

On the other hand, the average column sum in each column is $k(n+1)/2$. It is easy to see that the value

$$\sum_{j=1}^{n} \left[kj - \frac{k(n+1)}{2} \right]^2 = k^2 n \frac{n^2 - 1}{12}$$

is the maximal value that the statistic

$$S = \sum_{j=1}^{n} \left[R_j - \frac{k(n+1)}{2} \right]^2$$

could attain (the minimal is of course, equal to 0) and therefore

$$W = \frac{S}{k^2 n \frac{n^2-1}{12}} = \frac{12S}{k^2 n(n^2-1)}$$

provides a measure of **concordance** between the rankings (or dependence of the rankings). It can be shown that under the null hypothesis, the statistic

$$Q = \frac{12S}{kn(n+1)} = k(n-1)W$$

has asypmotically χ^2_{n-1} distribution. You are now in a position to formulate an asymptotic test for concordance of the rankings of the k observers. Do it!