WROCLAW UNIVERSITY

APPLIED STATISTICAL METHODS – Fall 2015

LECTURE NOTES - 5th SET - week 2

ASSOCIATION AND CORRELATION

The framework and problem. Suppose we have a set of iid observations of two continuous variables X and Y. Such observations come in pairs: $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$. We are interested in any association between X and Y. To determine certain types of association, we use measures of association, also called sometimes measures of correlation. The main three measures of association are:

- **Pearson's** r (sample correlation coefficient),
- **Kendall's** τ, and
- **Spearman's** ρ.

The area of statistics that works on association between variables is called **Correlation Analysis**.

CORRELATION ANALYSIS

STEP 1. Plot the data on a scatter plot and decide if you see any pattern in the data.

STEP 2. For some patterns we can compute measures of association. These are **monotonic** patterns/relations.

Monotonic patterns between X and Y:

- Increasing pattern: as X increases, Y increases.
- Decreasing pattern: as X increases, Y decreases.

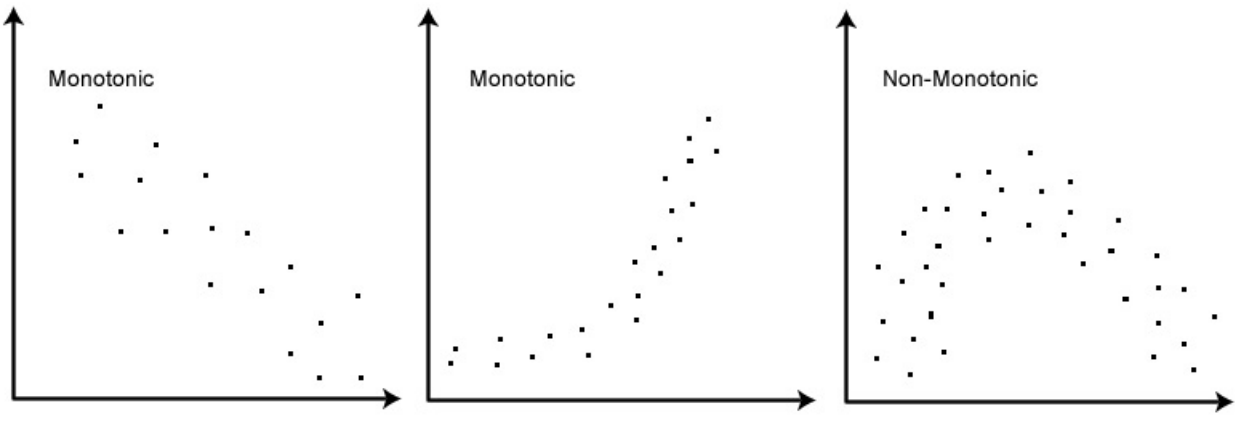

FIGURE 1. Types of association: left panel: monotonic linear; center panel: monotonic nonlinear; right panel: not monotonic.

Measures of correlation for monotonic relationships

- **Kendall's** τ and **Spearman's** ρ
  - Detects any monotonic relation;
  - Nonparametric-based on ranks, so resistant to outliers.
- **Pearson's** r
  - Detects only linear relations,
  - Influenced by outliers.

Common features of all measures of correlation:

- All have values between -1 and 1,
- Correlation=0 means no monotonic relation (τ, ρ), no linear relation r;
- Corr $>$): as X increases, then Y increases;
- Corr < 0: as X increases, then Y decreases.

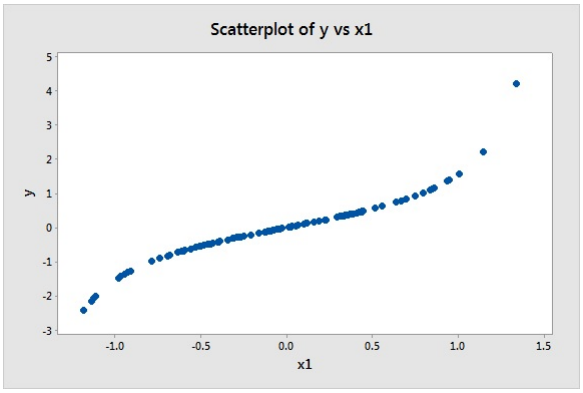

FIGURE 2. Types of association: Nonlinear but perfect monotonic association.

The scatterplot shows a nonlinear monotonic relation. For this data Pearson r= 0.948, and Spearman's $\rho = 1$.

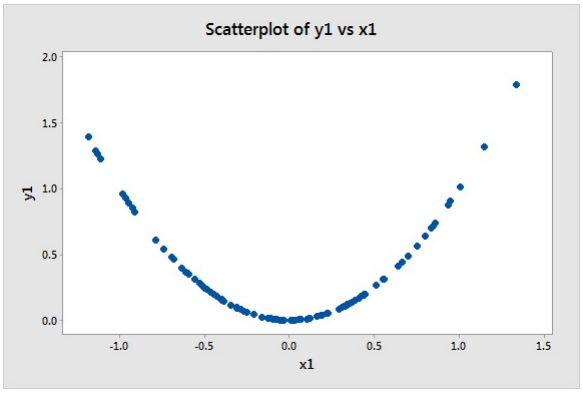

FIGURE 3. Types of association: Nonlinear and non monotonic but perfect relationship.

The scatterplot shows a nonlinear and non monotonic relation. For this data Pearson r= -0.053, and Spearman's $\rho = -0.032$.

TESTING STATISTICAL SIGNIFICANCE OF CORRELATION

The tests will have the same type of null and alternative hypotheses: Ho: corr=0 and Ha: corr $\neq 0$.
The tests differ for different measures of association.
NOTE: When one variable measures time or location, correlation provides information about temporal or spatial trends.

Pearson's correlation r. This is the "usual" correlation which measures the strength of **linear** association between two variables. For a data set $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n))$, the sample Pearson correlation coefficient r is given by

$$r = \frac{1}{n-1}\sum_{i=1}^{n}(\frac{x_i - \bar{x}}{s_x})(\frac{y_i - \bar{y}}{s_y}),$$

where $\bar{x}$ and $\bar{y}$ are sample means of the $X's$ and $Y's$, and s_x and s_y are sample standard deviations of $X's$ and $Y's$.

To test hypotheses about significance of Pearson correlation we assume that the data comes from a bivariate normal distribution.
Hypotheses: Ho: corr=0 and Ha: corr $\neq$ 0. Here by correlation we understand the true population correlation, which we estimate using sample correlation r.
The test statistic for testing above hypotheses is

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}.$$

Under Ho: t_r has a t-distribution with n-2 degrees of freedom. To find the p-value for the test, we use t-tables or software. We can also do the test directly in MINITAB.

Spearman's correlation ρ. Spearman's ρ is the Pearson correlation computed on the ranks of the $X's$ and $Y's$. We start with computing the ranks of $X's$ and $Y's$ independently! Ranks tell us where a given data value stands in the sorted data set. Rank of the smallest data value is 1, rank of the second smallest data value is 2, etc. Rank of the largest value is n. We obtain: $(R_{X_1}, R_{X_2}, \ldots, R_{X_n})$ and $(R_{Y_1}, R_{Y_2}, \ldots, R_{Y_n})$. Once we have ranks, we can compute Spearman's ρ as Pearson correlation on the ranks.
However, when we use the same test statistic as for testing about Pearson's correlation to test about Spearman's correlation, we only get **approximate** p-values. This is because the distribution of the test statistic in

this case is not exactly, only approximately t-distributed. Using the exact distribution of the test statistic we get critical regions given below. If the number of data pairs is over 20, the approximation of the test statistic with t-distribution works very well.

n\α	0.2	0.1	0.05	0.02	0.01	0.002
4	1.000	1.000	–	–	–	–
5	0.800	0.900	1.000	1.000	–	–
6	0.657	0.829	0.886	0.943	1.000	–
7	0.571	0.714	0.786	0.893	0.929	1.000
8	0.524	0.643	0.738	0.833	0.881	0.952
9	0.483	0.600	0.700	0.783	0.833	0.917
10	0.455	0.564	0.648	0.745	0.794	0.879
11	0.427	0.536	0.618	0.709	0.755	0.845
12	0.406	0.503	0.587	0.678	0.727	0.818
13	0.385	0.484	0.560	0.648	0.703	0.791
14	0.367	0.464	0.538	0.626	0.679	0.771
15	0.354	0.446	0.521	0.604	0.654	0.750
16	0.341	0.429	0.503	0.582	0.635	0.729
17	0.328	0.414	0.488	0.566	0.618	0.711
18	0.317	0.401	0.472	0.550	0.600	0.692
19	0.309	0.391	0.460	0.535	0.584	0.675
20	0.299	0.380	0.447	0.522	0.570	0.662
21	0.292	0.370	0.436	0.509	0.556	0.647
22	0.284	0.361	0.425	0.497	0.544	0.633
23	0.278	0.353	0.416	0.486	0.532	0.621
24	0.271	0.344	0.407	0.476	0.521	0.609
25	0.265	0.337	0.398	0.466	0.511	0.597
26	0.259	0.331	0.390	0.457	0.501	0.586
27	0.255	0.324	0.383	0.449	0.492	0.576
28	0.250	0.318	0.375	0.441	0.483	0.567
29	0.245	0.312	0.368	0.433	0.475	0.558
30	0.240	0.306	0.362	0.425	0.467	0.549

FIGURE 4. Critical values for the two-tailed test of association for Spearman's ρ.

Kendall's correlation τ. Kendall's correlation measures strength of a monotonic association and it is based on ranks. In order to compute Kendall's correlation from a sample we first classify the data pairs as **concordant** or **discordant**.

Concordant pairs. A pair of data points (X_1, Y_1) and (X_2, Y_2) is called concordant iff when $X_2 > X_1$ then $Y_2 > Y_1$, i.e. as X increases, Y also increases.

Discordant pairs. A pair of data points (X_1, Y_1) and (X_2, Y_2) is called discordant iff when $X_2 > X_1$ then $Y_2 < Y_1$, i.e. as X increases, Y decreases.

Example: Pairs $(1, 2)$ and $(3, 4)$ are concordant. Pairs (1, 2) and (3,1) are discordant.

In a data set we have n pairs (data points). We can do n(n-1)/2 comparisons. That is we can check n(n-1)/2 pairs of data points for concordance/discordance.
Let P = number of concordant pairs ("plusses"), and M = number of discordant pairs ("minuses"). Then, Kendall's τ is given by

$$\tau = \frac{P - M}{n(n-1)/2}.$$

If all pairs are concordant, then P= n(n-1)/2 and M=0, so $\tau = 1$ (as X increases Y increases).
When all pairs are discordant, then M=n(n-1)/2 and P=0, so $\tau = -1$ (as X increases then Y decreases).
When $\tau = \pm 1$ we have a perfect monotonic relationship between X and Y.

Testing Kendall's correlation for significance.

Test statistic is S= P - M.
Distribution (usually quantiles and/or p-values) of the test statistic is in tables for small samples. Such table is at the end of this set of lecture notes. For large samples we have a large sample approximation of the distribution of the test statistic. The approximating distribution is normal. The approximation works well for 10 or more observations.

$$Z_s = \begin{cases} \frac{s-1}{\sigma_s} & if \quad s > 0 \\ 0 & if \quad s = 0 \\ \frac{s+1}{\sigma_s} & if \quad s < 0, \end{cases} \tag{1}$$

where $\sigma_s = \sqrt{(n/18)(n-1)(2n+5)}$, and statistic Z_s has approximately standard normal distribution. That means that when we use approximate distribution of Z_s, we use standard normal distribution for computing p-values.

Correction for ties for Kendall's τ. When computing τ, note that tied values of X or Y produce 0, rather than a plus or minus to denote concordant or discordant pairs. Ties do not contribute to either P or M, and S and τ are computed exactly as before. An adjustment is required for the large sample approximation of the distribution of S. Namely, we need to change the standard deviation σ_S for the approximating normal.

Adjustment of σ_S. We have to compute the number of ties and the number of values involved in every tie (number of tied values). Consider the following data set:

data values	1	1	1	1	1	2	2	2	3	5	5	7	9	10	10	14	18

TABLE 1. Data for the ties example. The data was presorted.

There are a total of 4 tied groups in the data set. The largest tied group of the data set is of 5 values (five 1's). There is no tied group of 4 values, but there are tied groups of 3 values (three 2s) and two tied groups of 2 values (two 5s and two 10s). Finally, for completeness, lets think of non-repeating values as ties of extent 1. We have five ties of extent 1 (3, 9, 9, 14, and 18). Let t_i denote the number of ties of extent i. For this data we have: $t_1 = 5$, $t_2 = 2$, $t_3 = 1$, $t_4 = 0$, and $t_5 = 1$. For any $i > 5$, $t_i = 0$. the corrected σ_S is given by:

$$\sigma_S = \sqrt{\frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i(i)(i-1)(2i+5)}{18}}.$$

For our data, the corrected standard deviation is

$$\sigma_S = \sqrt{\frac{17 \cdot 16 \cdot 39 - 5 \cdot 1 \cdot 0 \cdot 7 - 2 \cdot 2 \cdot 1 \cdot 9 - 1 \cdot 3 \cdot 2 \cdot 11 - 1 \cdot 5 \cdot 4 \cdot 15}{18}} = \sqrt{567} = 23.81.$$

Table B8 -- Quantiles (p-values) for Kendall's S statistic and tau correlation coefficient

For N>10 use the normal approximation

One-sided p = Prob [S ≥ x] = Prob [S ≤ −x]

For two-sided test (Ha: tau not = 0) use p-value=2(one-sided p)

		N = Number of data pairs					N = Number of data pairs		
x	4	5	8	9	**x**	3	6	7	10
0	0.625	0.592	0.548	0.540	**1**	0.500	0.500	0.500	0.500
2	0.375	0.408	0.452	0.460	**3**	0.167	0.360	0.386	0.431
4	0.167	0.242	0.360	0.381	**5**		0.235	0.281	0.364
6	0.042	0.117	0.274	0.306	**7**		0.136	0.191	0.300
8		0.042	0.199	0.238	**9**		0.068	0.119	0.242
10		0.0083	0.138	0.179	**11**		0.028	0.068	0.190
12			0.089	0.130	**13**		0.0083	0.035	0.146
14			0.054	0.090	**15**		0.0014	0.015	0.108
16			0.031	0.060	**17**			0.0054	0.078
18			0.0156	0.038	**19**			0.0014	0.054
20			0.0071	0.022	**21**			0.0002	0.036
22			0.0028	0.0124	**23**				0.023
24			0.0009	0.0063	**25**				0.0143
26			0.0002	0.0029	**27**				0.0083
28			<0.0001	0.0012	**29**				0.0046
30				0.0004	**31**				0.0023
32				0.0001	**33**				0.0011
34				<0.0001	**35**				0.0005
36				<0.0001	**37**				0.0002
					39				<0.0001
					41				<0.0001
					43				<0.0001
					45				<0.0001

FIGURE 5. Critical values for the two-tailed test of association for Kendall's τ.