C. Describe relation between 2 qualitative variables

Cross table with levels of one variable in rows, levels of the second variable in columns:

	Y(1)	**Y(2)**	**. . .**	**Y(m)**	
X(1)	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$	M_1
X(2)	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$	M_2
			. . .		
X(k)	$n_{k,1}$	$n_{k,2}$		$n_{k,m}$	M_k
	K_1	K_2		K_m	**N**

Usually,

- The first variable (rows) is a independent (describing, cause, input) variable

- The second variable (columns) is a dependent (descriptive, result, output) variable

	Y(1)	Y(2)	. . .	Y(m)	
X(1)	$n_{1,1}$	$n_{1,2}$		$n_{1,m}$	M_1
X(2)	$n_{2,1}$	$n_{2,2}$		$n_{2,m}$	M_2
			. . .		
X(k)	$n_{k,1}$	$n_{k,2}$		$n_{k,m}$	M_k
	K_1	K_2		K_m	**N**

In cell ij of table, $n_{i,j}$: number of observations belonging simmutaneously to the level i of the first variable and to the the level j of the second variable,

M_i : sum of all numbers in the row i (number of observations in *i-th* level of the first variable),

K_j : sum of all numbers in the column j (number of observations in *j-th* level of the second variable)

N : total number of all observations in the sample

Table with percentages % across rows: $n_{i,j} / M_i$ gives information about distrribution of "output" variable Y in each level of "input" variable X

	Y(1)	**Y(2)**	**. . .**	**Y(m)**
X(1)	$n_{1,1} / M_1$	$n_{1,2} / M_1$		$n_{1,m} / M_1$
X(2)	$n_{2,1} / M_2$	$n_{2,2} / M_2$		$n_{2,m} / M_2$
			. . .	
X(k)	$n_{k,1} / M_k$	$n_{k,2} / M_k$		$n_{k,m} / M_k$

Table with percentages % across columns: $n_{i,j} / K_j$ gives information about distrribution of "input" variable X in each level of "output" variable Y

	Y(1)	**Y(2)**	**. . .**	**Y(m)**
X(1)	$n_{1,1} / K_1$	$n_{1,2} / K_2$		$n_{1,m} / K_m$
X(2)	$n_{2,1} / K_1$	$n_{2,2} / K_2$		$n_{2,m} / K_m$
			. . .	
X(k)	$n_{k,1} / K_1$	$n_{k,2} / K_2$		$n_{k,m} / K_m$

Table with percentages % in whole samle: $n_{i,j} / N$ gives information about total distrribution in sample

	Y(1)	**Y(2)**	**. . .**	**Y(m)**
X(1)	$n_{1,1} / N$	$n_{1,2} / N$		$n_{1,m} / N$
X(2)	$n_{2,1} / N$	$n_{2,2} / N$		$n_{2,m} / N$
			. . .	
X(k)	$n_{k,1} / N$	$n_{k,2} / N$		$n_{k,m} / N$

Using SPSS

SPSS : Commands

Analyze

Descriptive Statistics

Crosstabs …

For calculations percentages across rows or across columns we use in addition the **Cells ...** command box and choose **Row** or **Column** respectively.

D. Describe relation between qualitative variable and quantitative variable

Using cross table with columns present the groups determined by qualitative variable and Mean value of quantitative variable taken separately in each group:

1-st Group $Y = y_1$	*2-nd Group* $Y = y_2$	. . .	*k-th Group* $Y = y_k$
$Mean/_{Y=y_1}(X)$	$Mean/_{Y=y_2}(X)$		$Mean/_{Y=y_k}(X)$

Information from the table:

- *Mean value of variable is highest, lowest in which group of variable Y*
- *Difference between mean values of variable X take in different levels of variable Y , etc.*

Examle:

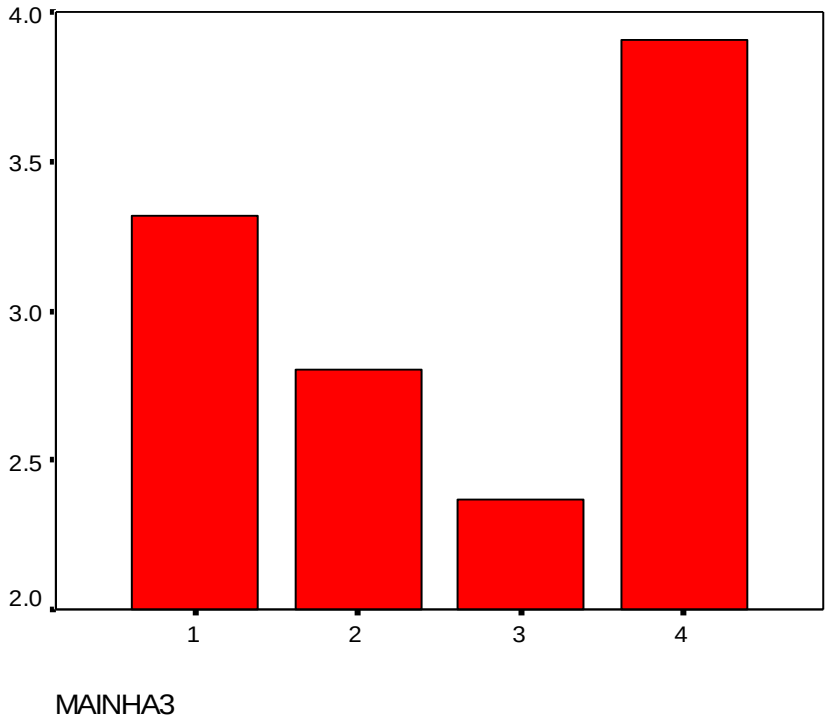
- Qualitative variable Y : "*Economic status of household*"
- Quantitative variable X: "*Food expenditure per capita in household*",

Lowest	Lower	Average	Higher	Highest
3.511	**4.808**	**7.105**	**8.455**	**9.650**

***Remark**: In the table, instead of Mean value we can use other statistical parameters of quantitative variable: Median, Min, Max, Standard Deviantion, Range, etc.*

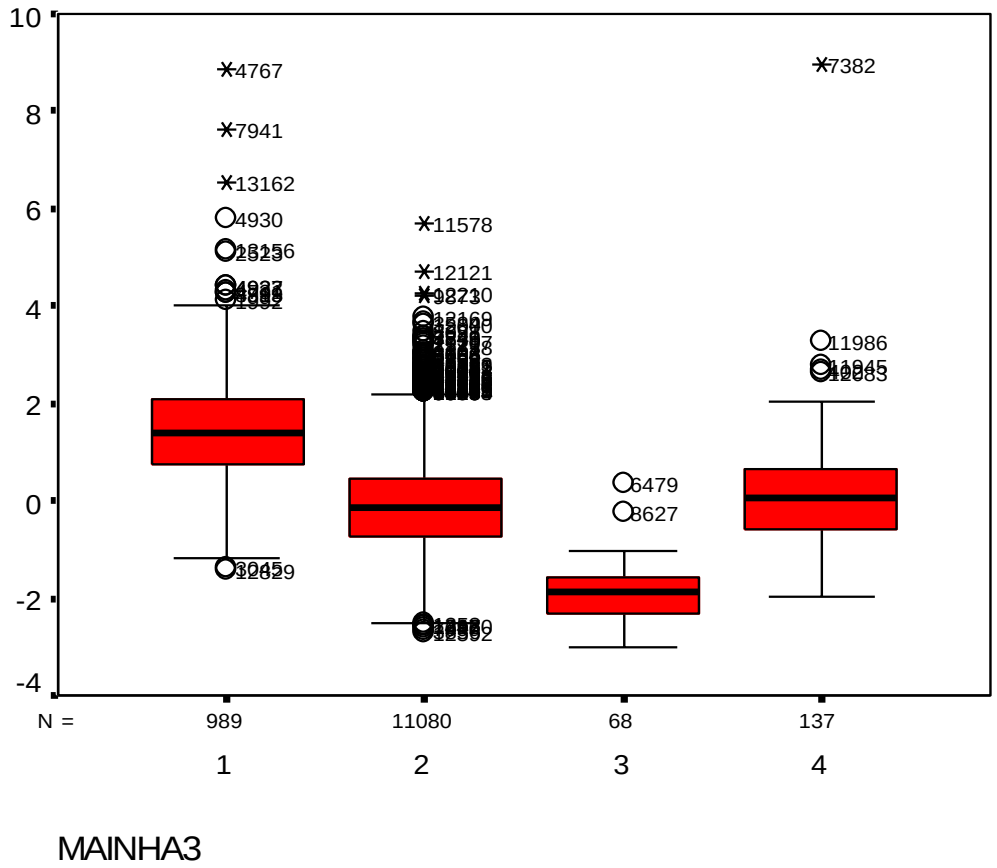
Using bar chart

Colums of bar chart present statistical parameters of quantitative variable X (Mean(X), Med(X), Min(X), etc.) in groups of qualitative variable Y:



Using box plot

Box plot is using to compare distributions of quantitative variable in different groups of qualitative variable:



Using SPSS to describe relation between qualitative and quantitative variables

SPSS : Command

Analyze

Descriptive Statistics

Explore …

Then choose **Simple**, **Summaries for groups of cases** and put quantitative variable int **Dependent list**, put qualitative variable into **Factor list**

Char-graph-plot

SPSS : Use command

Graph

Bar ...

(or **Pie ...** or **Boxplot …**)

Then choose **Other summary function,** and a suitable statistical perameter function (mean, median, min, max, etc.)

E. Describe relation between 2 quantitative variables

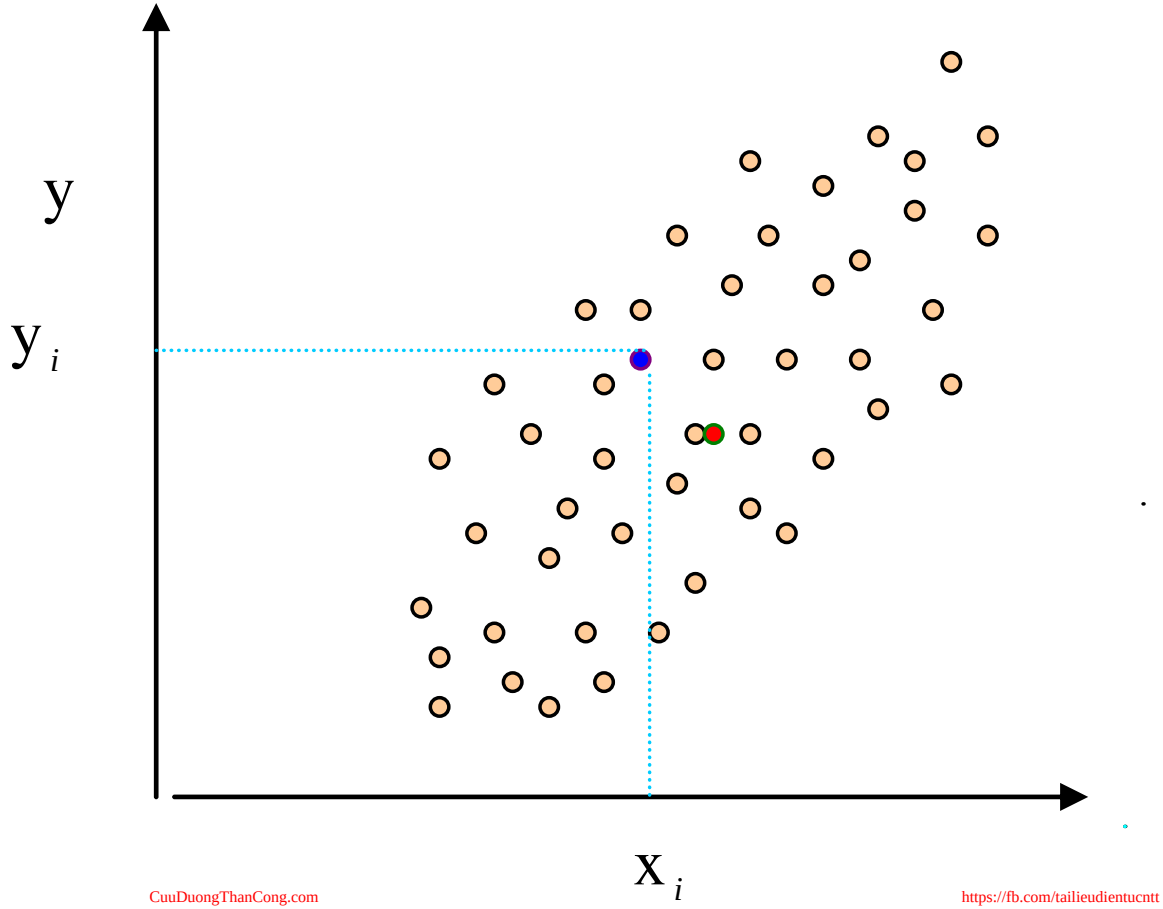
For primary describing relation between 2 quantitative variables we can use dot (scatter) plot, covariance và linear correlation coefficient of two quantitative variables.

(a) Scatter plot

For quantitative variables X, Y with sample

$$E = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Dot plot of sample E is performed by drawing n points with cordinates (x_i, y_i) :



Notes

(a) Scatter plot providing two-dimentional picture of data represents distribution of data. In that plot we can see concentration area of data, see if there are somes outliers, abnormal points, etc.

(b) Scatter plot can be used to compare several populations: draw several samples (differently coloured) on a common plot

b) Covariance and linear correlation coefficient

(i) For presentation relationship between two quantitative variables we can use ***covariance*** between those variables

$$Cov(X,Y) = \frac{1}{n}\sum_{k=1}^{n} (x_k - Mean(X)).(y_k - Mean(Y)) \quad .$$

Property of covariance

1) Symmetric: $Cov(X,Y) = Cov(Y,X)$

2) Depends on measure scale of X and Y :

$$Cov(a.X,b.Y) = a.b.Cov(Y,X)$$

3) Does not depend on origin of co-ordinates: for all numbers a and b the following equality always valids:

$$Cov(X+a,Y+b) = Cov(X,Y) \quad ,$$

b) Covariance and linear correlation coefficient

ii) *Linear correlation coefficient:*

$$r(X,Y) = Cov(X,Y) / (\sigma(\boldsymbol{X}).\sigma(\boldsymbol{Y})) .$$

Property:

1) Symmetric: $r(X,Y) = r(Y,X)$

2) Not depent on measure scale of variables: For all numberr a , b different from 0 and a.b>0 we have

$$r(aX,bY) = r(X,Y)$$

3) Not depend on origin of co-ordinates: for all numbers a and b , the following is true:

$$r(X+a,Y+b) = r(X,Y) ,$$

Linear correlation coefficient measures the linear dependence between two variables:

$$-1 \le r(X,Y) \le 1$$

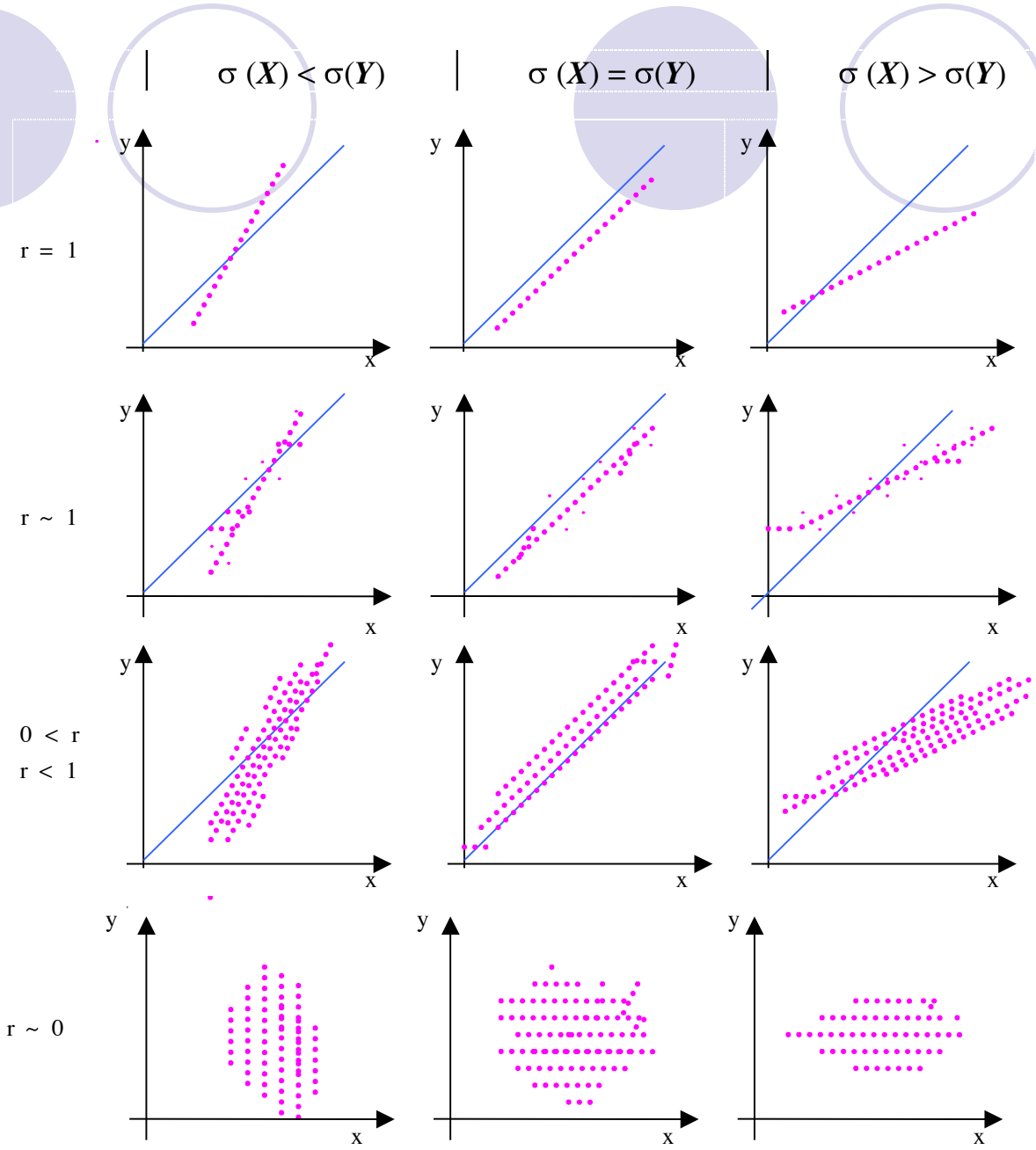
a) X and Y are completely linearly dependent:

$r(X,Y) = 1$ if and only if $Y = aX + b$ with $a > 0$,

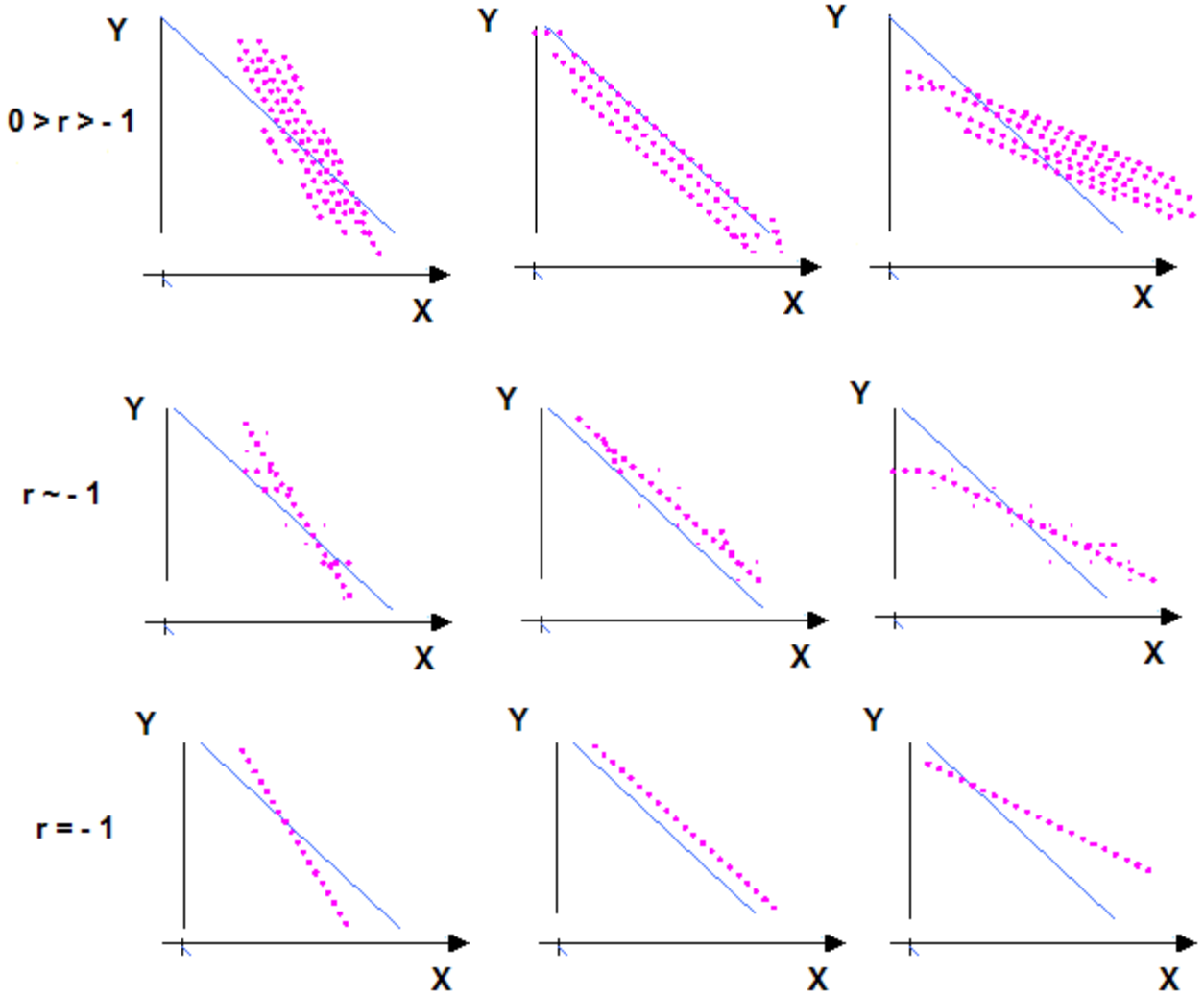
$r(X,Y) = -1$ if and only if $Y = aX + b$ with $a < 0$.

b) If $r(X,Y)$ close to 1 (or -1) then X and Y are very strongly related, can have some linear correlation,

c) If $r(X,Y)$ close to 0 then X and Y are linearly independent, there is not linear relation between them.



| $\sigma(X) < \sigma(Y)$ | $\sigma(X) = \sigma(Y)$ | $\sigma(X) > \sigma(Y)$ |



Remarks

1) If $E(1)$ and $E(2)$ are two samples then

From $r(E(1)) \sim 1$ and $r(E(2)) \sim 1$ does not imply $r(E(1) \cup E(2)) \sim 1$.

2) From $r(E(1)) \sim 0$ and $r(E(2)) \sim 0$ does not imply $r(E(1) \cup e(2)) \sim 0$.

3) From $r(E(1)) < 0$ and $r(E(2)) < 0$ does not imply $r(E(1) \cup e(2)) < 0$.

4) - Between X and Y may be some non-linear relation although

$$r(X,Y) = 0 ,$$

Example: $y = x^2$

- Between X and Y may be some non-linear relation in spite of

$$r(X,Y) \sim 1 ,$$

Example: For $y = \sqrt{x}$ we have $r = 0.981$.

5) From $r(X,Y) \sim 1$ and $r(Y,Z) \sim$ does not imply $r(X,Z) \sim 1$.

6) From $r(X,Y) \sim 0$ and $r(Y,Z) \sim 0$ does not imply $r(X,Z) \sim 0$.

APPLETS

Ch□ơng 8.1