Correlation & Regression

Outlines

- Correlation
  - Definition
  - Deviation Score Formula, Z score formula
  - Hypothesis Test
- Regression
  - Intercept and Slope
  - Un-standardized Regression Line
  - Standardized Regression Line
  - Hypothesis Tests

1. Direction

- Positive (+)
- Negative (-)

2. Degree of association

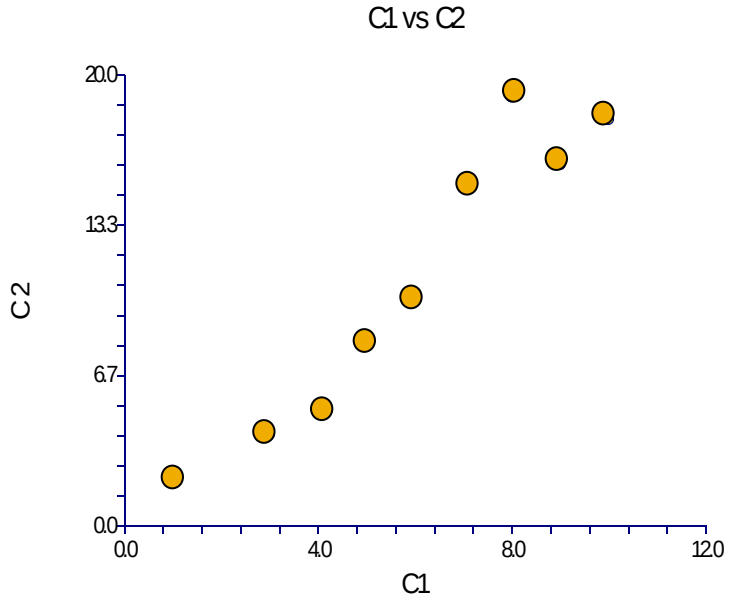
- Between –1 and 1
- Absolute values signify strength

3. Form

- Linear
- Non-linear

Correlation: 1. Direction

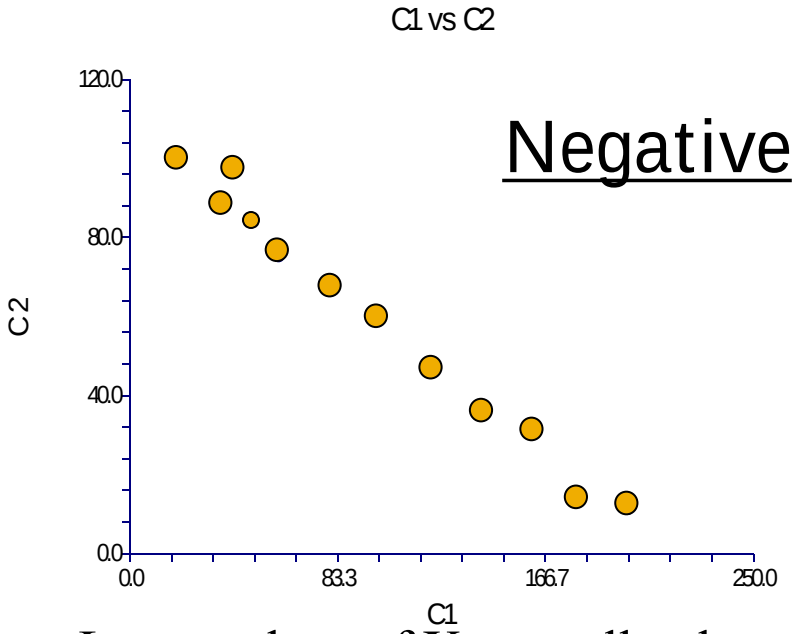
Positive



Large values of X = large values of Y,

Small values of X = small values of Y.

- e.g. IQ and SAT



Negative

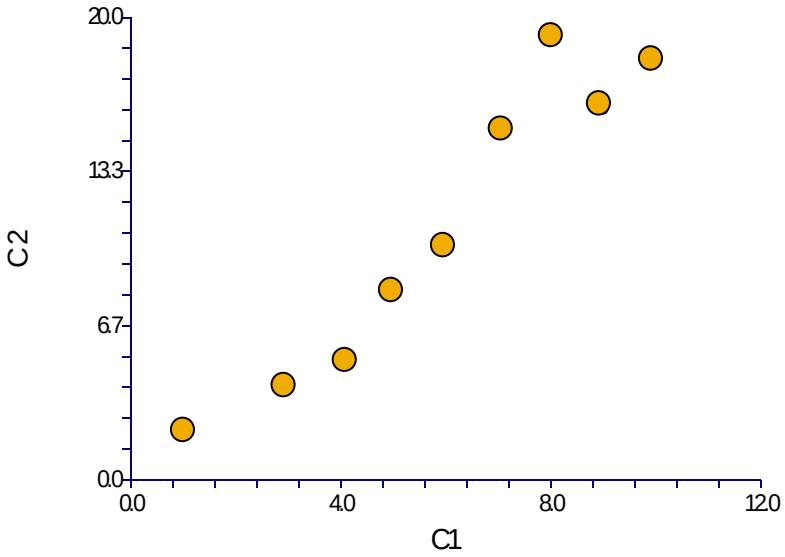
Large values of X = small values of Y

Small values of X = large values of Y

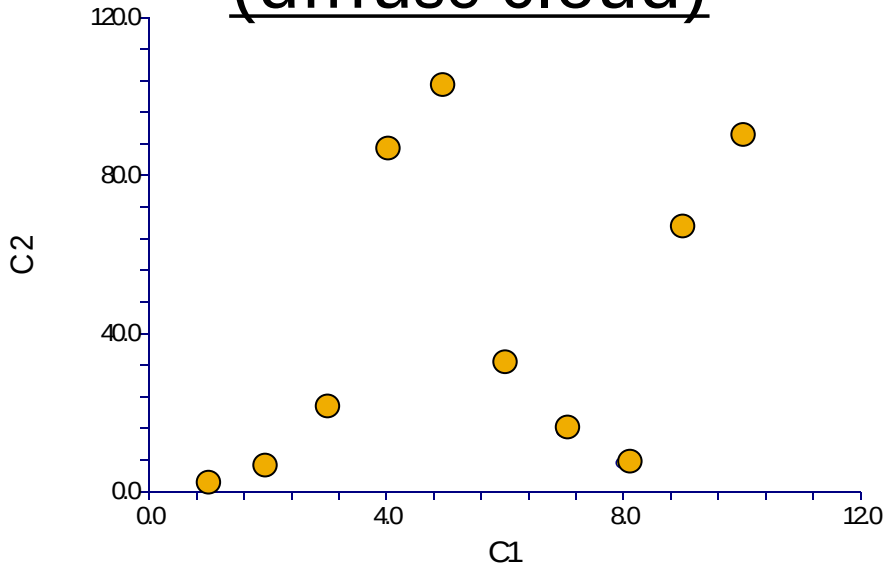
-e.g. SPEED and ACCURACY

Correlation: 2. Degree of association

Strong
(tight cloud)

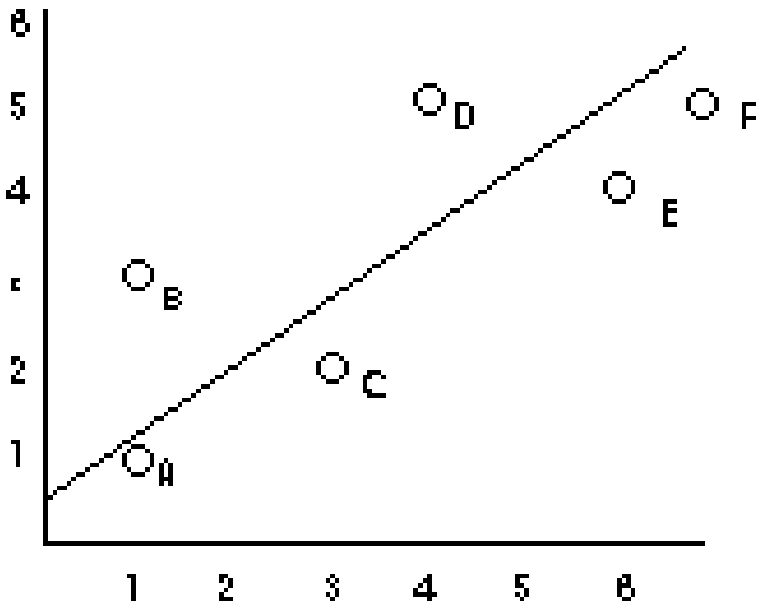


Weak
(diffuse cloud)

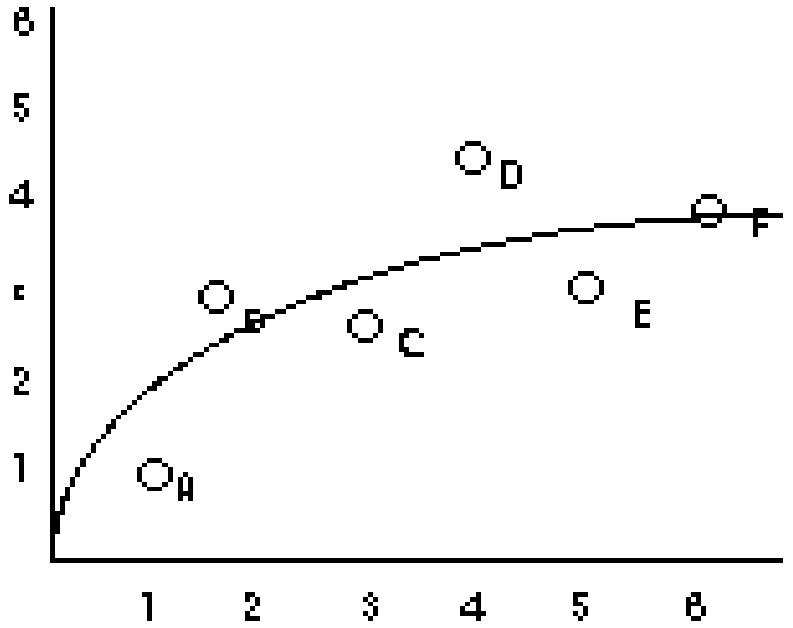


Correlation: 3. Form

Linear



Non- linear



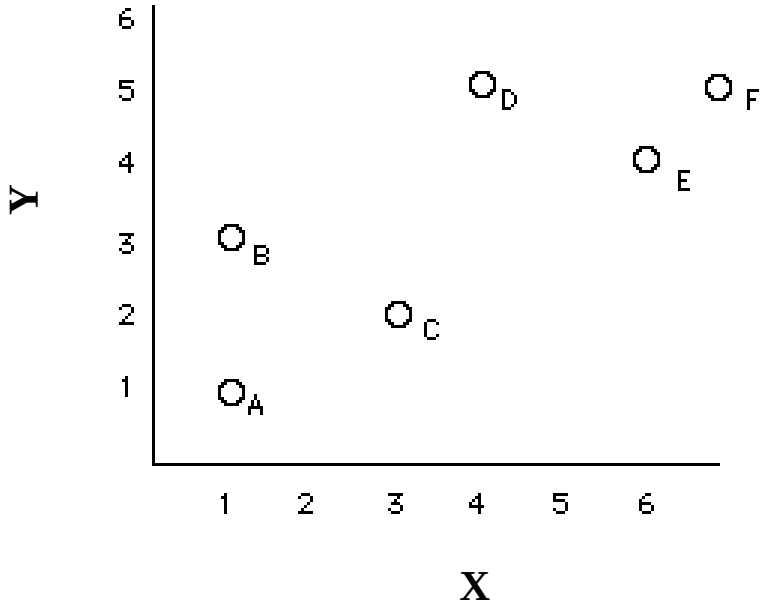
Correlation: Definition

Correlation: a statistical technique that measures and describes the degree of linear relationship between two variables

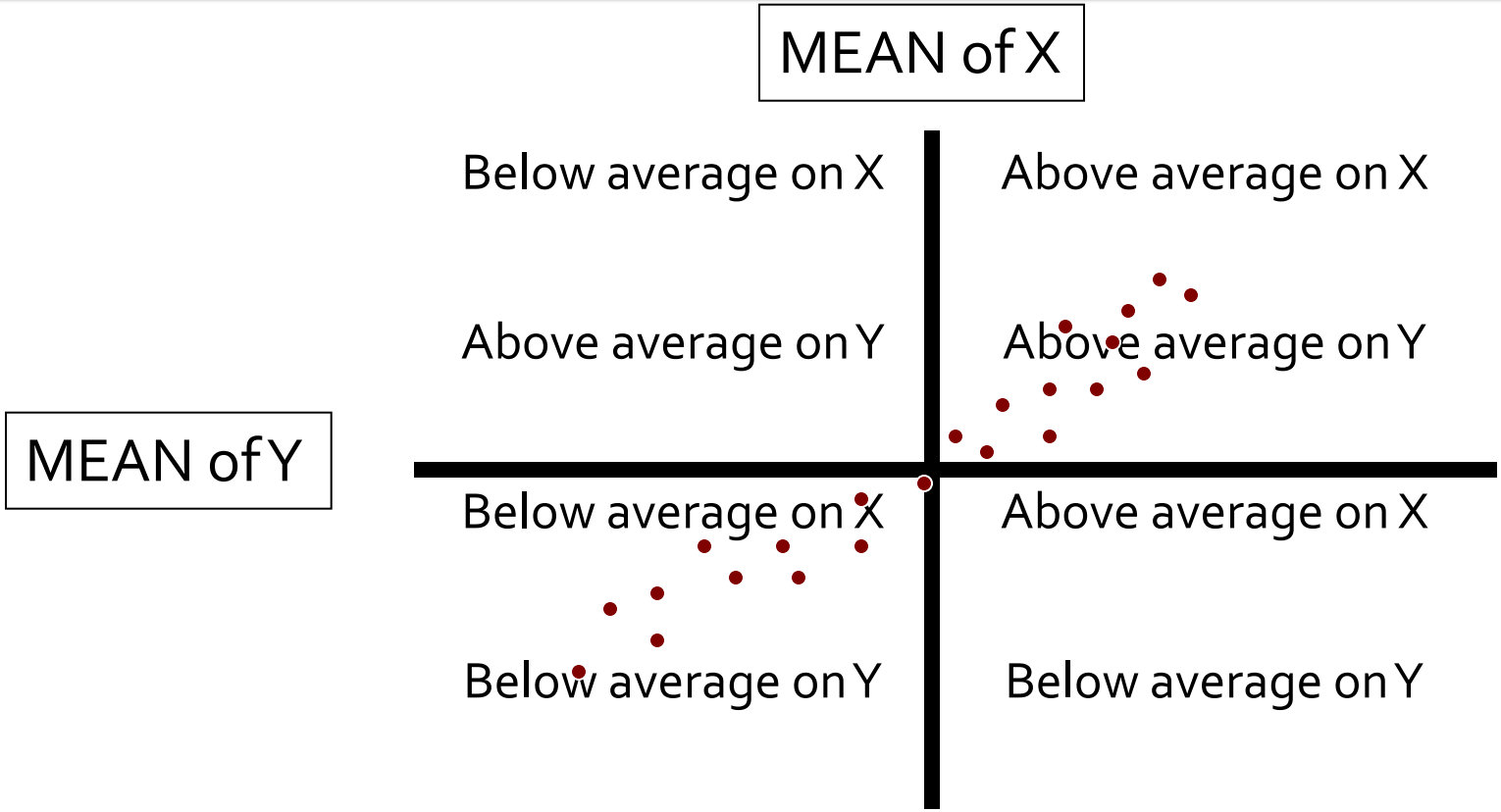
Dataset

Obs	X	Y
A	1	1
B	1	3
C	3	2
D	4	5
E	6	4
F	7	5

Scatterplot



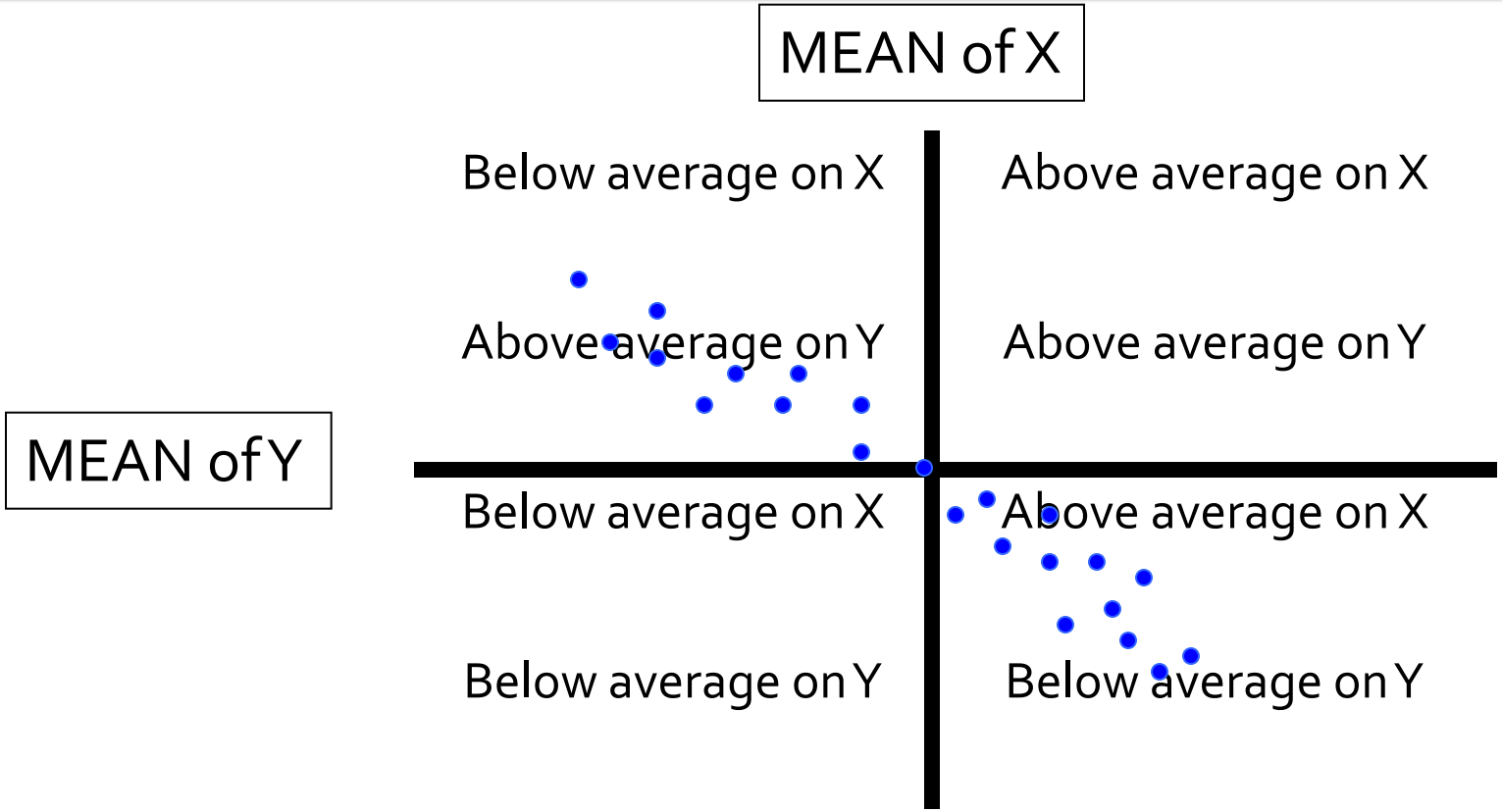
The logic of regression



Cross-Product = $(X - \overline{X})(Y - \overline{Y})$

For a strong positive association, the cross-products will mostly be positive

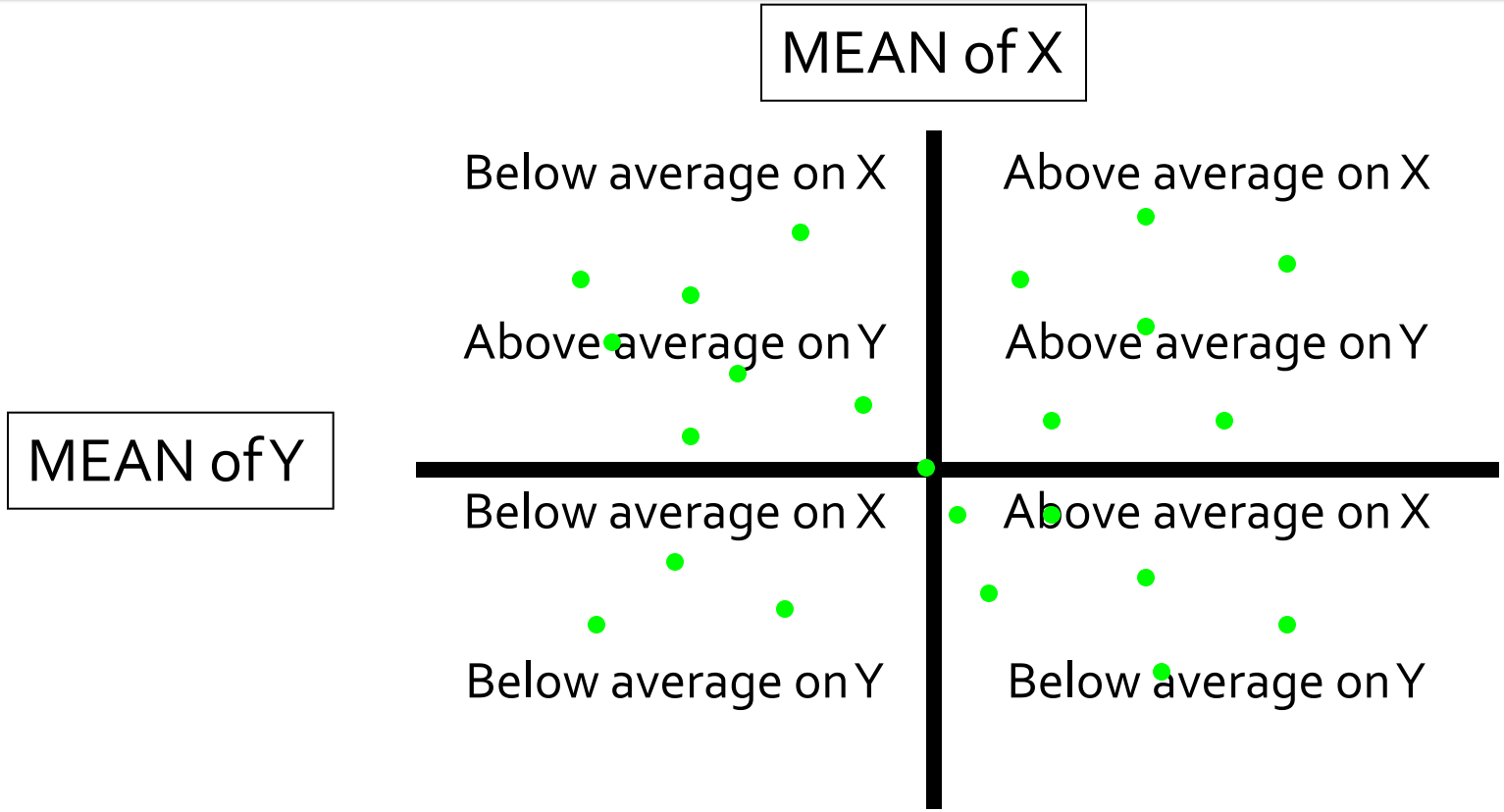
The logic of regression



Cross-Product = $(X - \bar{X})(Y - \bar{Y})$

For a strong negative association, the cross-products will mostly be negative

The logic of regression



Cross-Product = $(X - \bar{X})(Y - \bar{Y})$

For a weak association, the cross-products will be mixed

Pearson Correlation Coefficient

Symbol: r, R

A value ranging from -1.00 to 1.00 indicating the strength (look to the number of correlation coefficient) and direction (look to the sign of the correlation coefficient) of the linear relationship.

- Absolute value indicates strength
- +/- indicates direction

$$r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}}$$

Pearson Correlation Coefficient

- Assumptions:

1. The errors in data values are independent from one another
2. Correlation always requires the assumption of a straight-line relationship
3. The variables are assumed to follow a bivariate normal distribution

Pearson Correlation Coefficient

	Femur	Humerus	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
A	**38**	**41**					
B	**56**	**63**					
C	**59**	**70**					
D	**64**	**72**					
E	**74**	**84**					
Mean	58.2	66.00					
					SS_X	SS_Y	SP

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

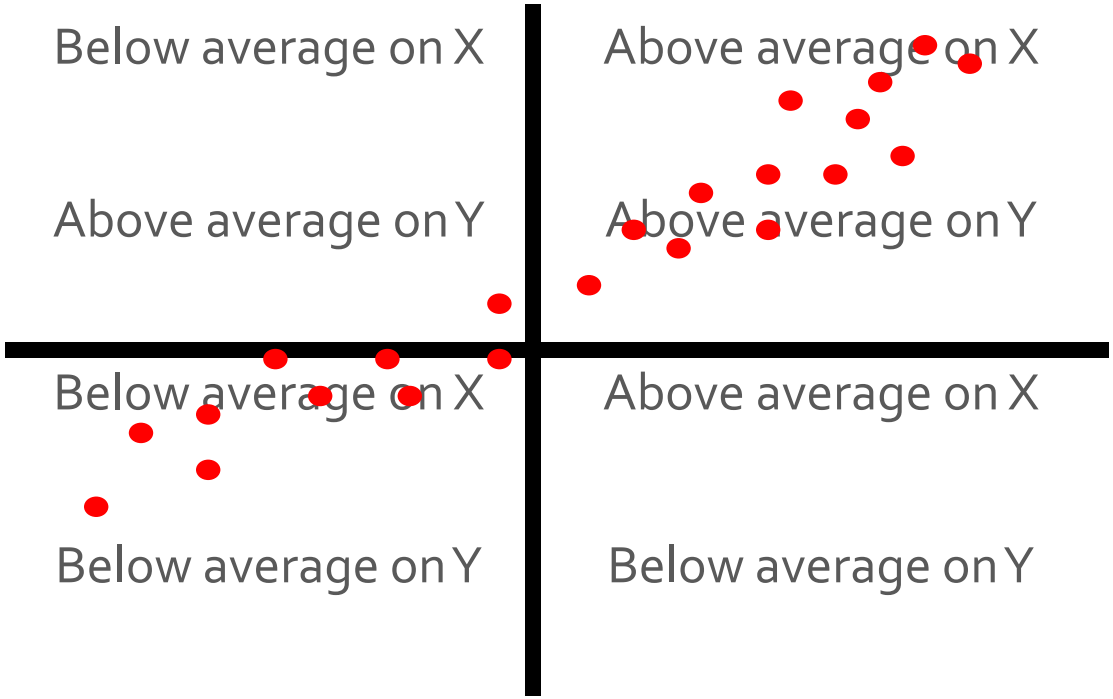
Pearson Correlation Coefficient

	Femur	Humerus	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
A	**38**	**41**	-20.2	-25	408.04	625	505
B	**56**	**63**	-2.2	-3	4.84	9	6.6
C	**59**	**70**	0.8	4	.64	16	3.2
D	**64**	**72**	5.8	6	33.64	36	34.8
E	**74**	**84**	15.8	18	249.64	324	284.4
mean	58.2	66.00			**696.8**	**1010**	**834**
					SS_X	SS_Y	SP

r = 0.99

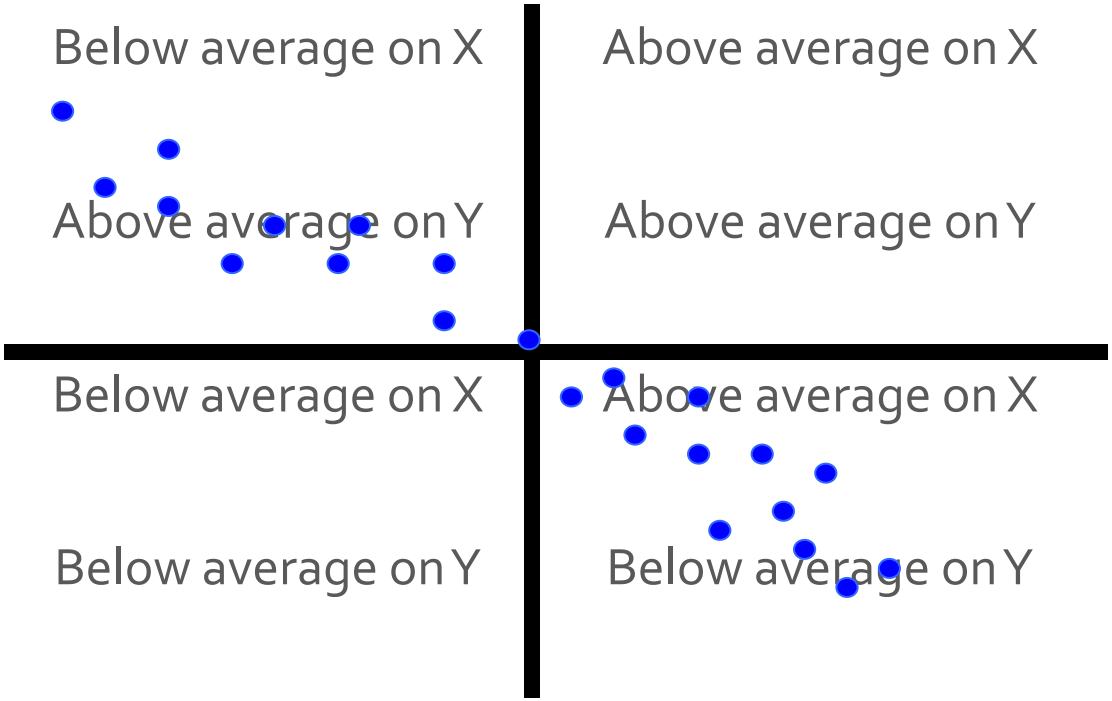
Pearson Correlation Coefficient

- For a strong positive association, the SP (sum of products) will be a big positive number



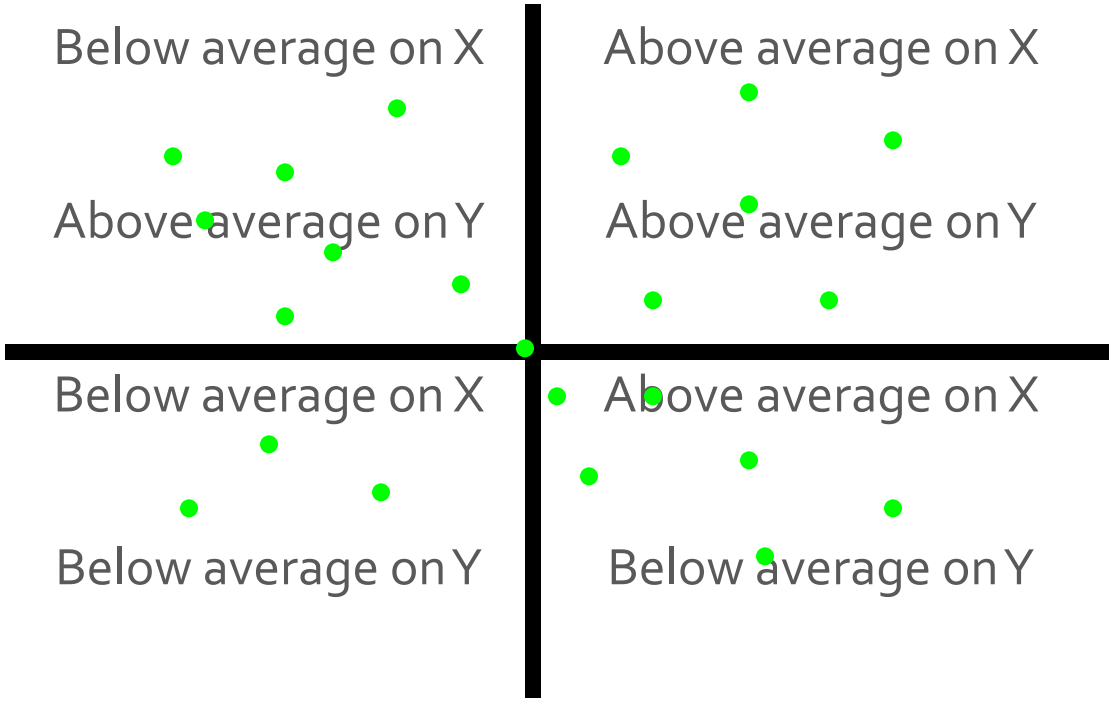
Pearson Correlation Coefficient

- For a strong negative association, the SP will be a big negative number



Pearson Correlation Coefficient

- For a <u>weak</u> association, the SP will be a small number (+ and – will cancel each other out)



Pearson Correlation Coefficient: Interpretation

- A measure of strength of association: how closely do the points cluster around a line?
- A measure of the direction of association: is it positive or negative?
- Colton [Colton T. Statistics in Medicine. Little Brown and Company, New York, NY 1974] rules:
  - $R \subset [-0.25 \text{ to } +0.25] \rightarrow$ No relation
  - $R \subset (0.25 \text{ to } +0.50] \cup (-0.25 \text{ to } -0.50] \rightarrow$ weak relation
  - $R \subset (0.50 \text{ to } +0.75] \cup (-0.50 \text{ to } -0.75] \rightarrow$ moderate relation
  - $R \subset (0.75 \text{ to } +1) \cup (-0.75 \text{ to } -1) \rightarrow$ strong relation

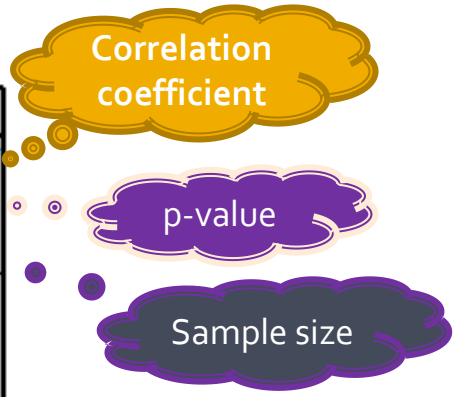
Pearson Correlation Coefficient: Interpretation

- The P-value is the probability that you would have found the current result if the correlation coefficient were in fact zero (null hypothesis).
- If this probability is lower than the conventional significance level (e.g. 5%) ($p < 0.05$) → the correlation coefficient is called statistically significant.

Correlations

		ISET	LogPexp
ISET	Pearson Correlation	1	.653**
	Sig. (2-tailed)		2.178E-016
	N	124	124
LogPexp	Pearson Correlation	.653**	1
	Sig. (2-tailed)	.000	
	N	124	124

**. Correlation is significant at the 0.01 level (2-tailed).



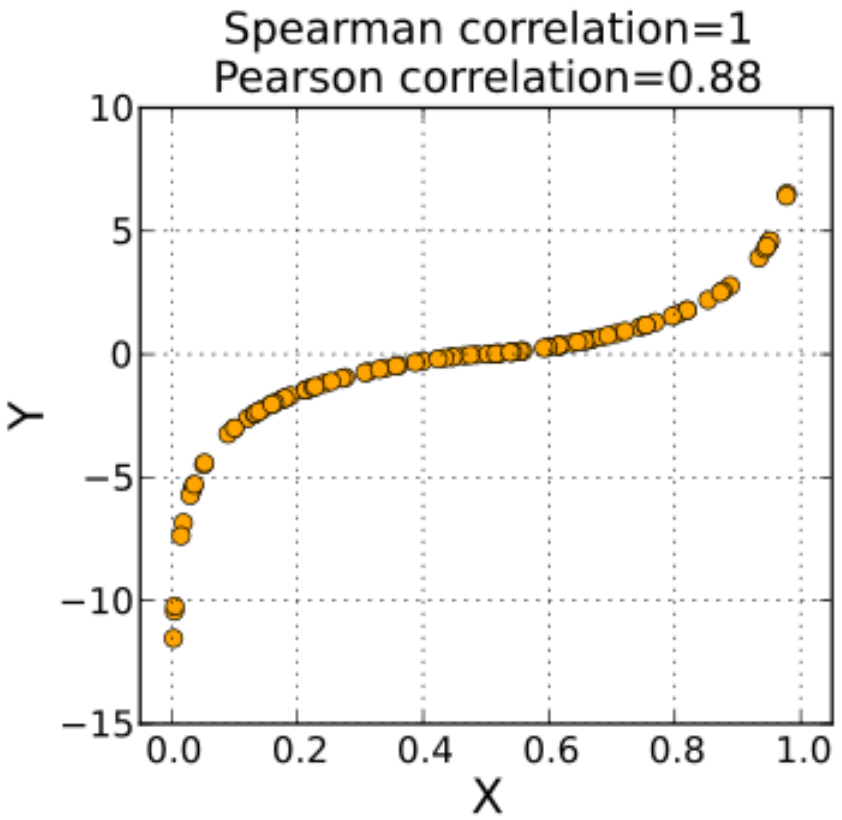
Spearman Rank Correlation Coefficient

- Not continuous measurements
- The assumption of bivariate normal distribution is violated
- Symbol: ρ (Rho Greek Letter)

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \times (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman Rank Correlation Coefficient

- The sign of the Spearman correlation indicates the direction of association between *X* (the independent variable) and *Y* (the dependent variable).
- $\rho = 1 \rightarrow$ the two variables being compared are monotonically related. N.B. This does not give a perfect Pearson correlation.



Interpretation of r-squared (r^2)

- The amount of covariation compared to the amount of total variation
- The percent of total variance that is shared variance
- E.g. If r = 0.80, then X explains 64% of the variability in Y (and vice versa)

Properties of correlation coefficient

- A standardized statistic – will not change if you change the units of X or Y.
- The same whether X is correlated with Y or vice versa
- Fairly unstable with small n
- Vulnerable to outliers
- Has a skewed distribution

Correlation coefficient by example

- Enciu A, Zamfir CZ, Nicolescu A, Ida A. THE ANALYSIS OF CORRELATIONS BETWEEN THE MAIN TRAITS OF WOOL PRODUCTION ON MILK BREED – PALAS. Lucrări Ştiinţifice - Seria Zootehnie ????;57:50-54.

Table 1 Correlation and regression coefficients between wool production and fiber diameter related to the age of the sheep (shearing season)

Sheep category	Breed	Shearing season (age)	r ± sr	b ± sb
Female yearlings	Milk Breed Palas	1	0.187±0.055(***)	0.117±0.022 (***)
Male yearlings	Milk Breed Palas	1	0.204±0.109 (ns)	0. 185±0.098 (*)
Ewes	Milk Breed Palas	2 - 10	-0.043±0.218(ns) 0.361±0.071 (***)	-0.035±0.099(ns) 0.125± 0.025 (***)
Rams	Milk Breed Palas	2 - 4	0.081±0.442(ns) 0.257±0.176 (ns)	0.065± 0.028 (ns) 0.196±0.113 (ns)

Notes: ns – not significant (P > 0,05) * - significant (P < 0,05) ** - distinctly significant (P < 0,01)
*** - very significant (P < 0,001)

Correlation matrix

Table 5

Correlation matrix for the broad set of explanatory variables considered in the analysis of risk factors associated with the SCM among lactating cow in the Savannah region of Nigeria

Variable	A	B	C	D	E	F	G	H	I	J	K	L
A	1											
B	0.9637	1										
C	0.0273	0.0148	1									
D	0.143	0.0945	0.0702	1								
E	0.0397	0.0434	0.5789	0.0182	1							
F	−0.1778	−0.1315	−0.3027	−0.2847	−0.1752	1						
G	−0.0949	−0.0563	−0.0903	−0.208	−0.1284	0.3426	1					
H	0.018	0.0298	0.5884	0.022	−0.027	−0.1781	0.0397	1				
I	−0.018	−0.0298	−0.5884	−0.022	0.027	0.1781	−0.0397	−1	1			
J	−0.019	−0.0751	0.4626	0.1916	0.2678	−0.6544	−0.217	0.2722	−0.2722	1		
K	−0.1964	−0.1503	−0.3469	−0.3396	−0.2008	0.8725	0.373	−0.2041	0.2041	−0.75	1	
L	−0.0015	−0.0016	0.0004	−0.0012	0.0073	0.0016	0.0012	0.0002	−0.0002	0.0009	0.0018	1

A=age, B=parity, C=washing hands before milking, D=breed, E=heifer and cow, F=month, G=cow type, H=pre-striping before milking, I=feeding after milking, J=washing of teats, K=management system, L=quarter.

Shittu *et al. BMC Veterinary Research* 2012 **8**:134 doi:10.1186/1746-6148-8-134

Simple Linear regression
Multiple linear regression

Linear Regression: Assumptions

- The errors in data values (e.g. the deviation from average) are independent from one another
- Regressions depends on the appropriateness of the model used in the fit
- The independent readings (X) are measured as exactly known values (measured without error)
- The variance of Y is the same for all values of X
- The distribution of Y is approximately normal for all values of X

Linear Regression

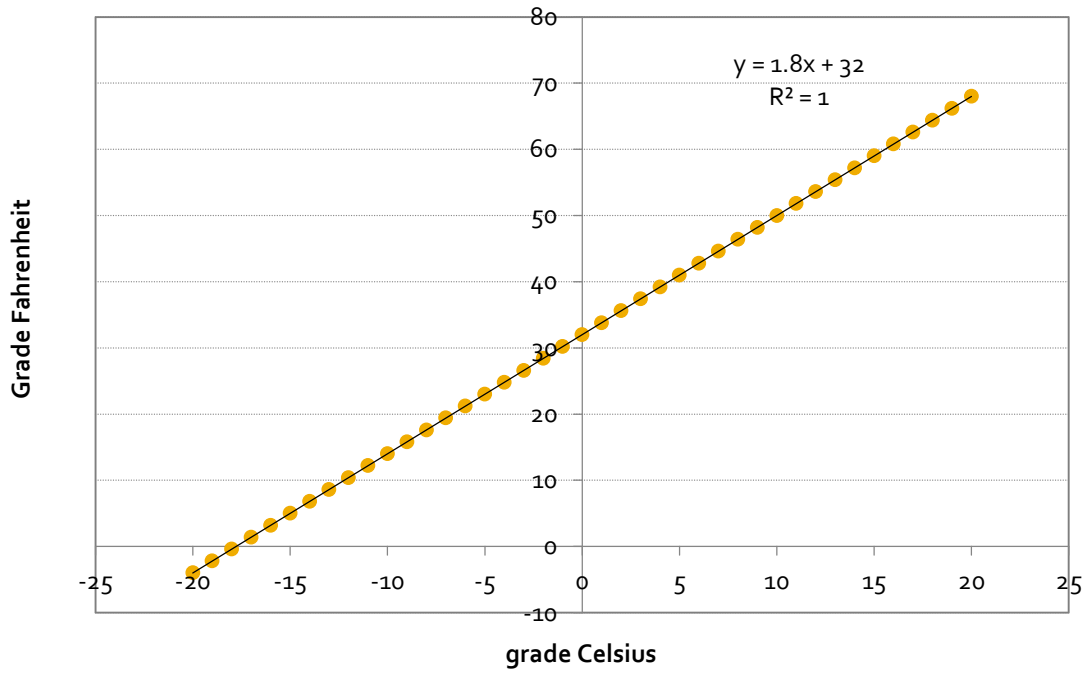
- But how do we describe the line?
- If two variables are linearly related it is possible to develop a simple equation to predict one variable from the other
- The outcome variable is designated the Y variable, and the predictor variable is designated the X variable
- E.g. centigrade to Fahrenheit:

$$F = 32 + 1.8^{o}C$$

this formula gives a specific straight line

Linear Equation

- General form is Y = a + bX
- The prediction equation: $\tilde{Y}$ = a+ bX
  - a = intercept, b = slope, X = the predictor, Y = the criterion
- *a and b are constants in a given line; X and Y change*



Slope and Intercept

- Equation of the line: $\tilde{Y} = a + bX$
- The slope b: the amount of change in Y with one unit change in X

$$b = r\frac{s_y}{s_x} = \frac{SP}{SS_X}$$

- The intercept a: the value of Y when X is zero

$$a = \overline{Y} - b\overline{X}$$

- The slope is influenced by r, but is not the same as r

Table 2

The effect of herd paratuberculosis sero-status (positive, negative or non-negative) on milk, fat and protein yield, somatic cell count score (SCCS) and calving interval [mean (95% CI)]

Variables	Positive	Non-negative	Negative	F-value	Adjusted R^2	P-value
Milk (kg)	6981.44 (6594-7369)	6928.00 (6594-7369)	6601.85 (6408-6795)	1.995	0.447	0.138
Fat (kg)	242.20 (231-253)	238.50 (226-251)	238.03 (232-244)	0.213	0.7	0.809
Protein (kg)	222.09 (216-228)	220.79 (214-228)	215.75 (213-219)	2.031	0.871	0.133
SCCS (score)	3.26 (2.7-3.8)	3.38 (2.7-4.0)	2.76 (2.5-3.0)	2.469	0.021	0.087
Calving interval (day)	386.42 (368-405)	391.63 (364-419)	381.67 (372-391)	0.696	0.063	0.5

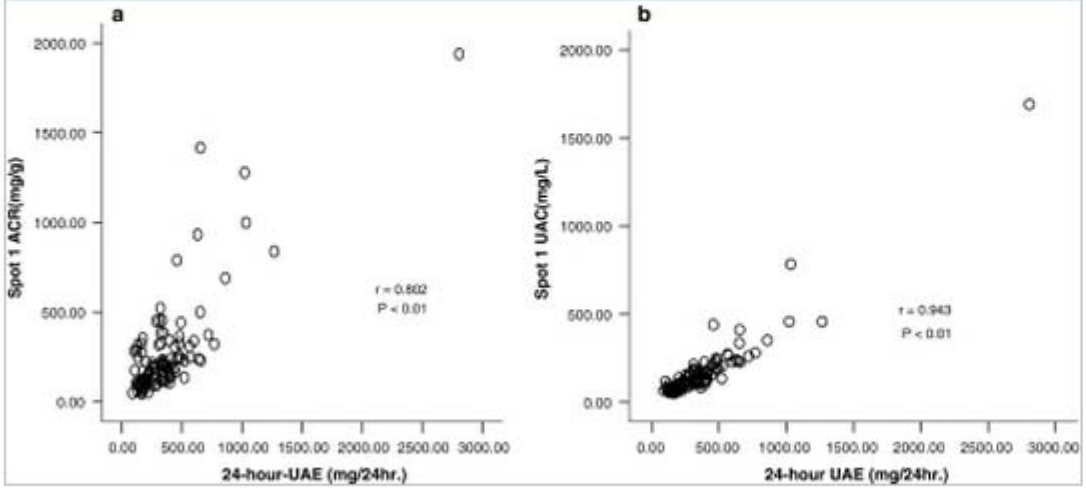
Hoogendam *et al. Irish Veterinary Journal* 2009 **62**(Suppl 4):265 doi:10.1186/2046-0481-62-4-265

From: Indian J Clin Biochem. 2011 July; 26(3): 283–289.
Published online 2011 April 30. doi: 10.1007/s12291-011-0136-0

<< Prev Next >>

Fig. 2

Click on image to zoom



Scatterplot graph of Pearson's correlation between spot 1 (FMV) and 24-h UAE. **a** spot 1 ACR (mg/g) versus 24-h UAE (mg/24 h), **b** spot 1 UAC (mg/l) versus 24-h UAE (mg/24 h)

Summary!

- Assessment of the strenght of association between 2 quantitative variables (normal distributed) —› correlation coefficient
- Prediction of one variable (Y) using another variable (X) —› regression