STAT 110

CHAPTER # 10 & 13

PREPARED BY: GASIM MUDAWI

YEAR 1431-1432 H

Chapter 10: Correlation and Regression:

الفصل العاشر: الارتباط والانحدار

Correlation is a statistical method used to determine whether a linear relationship between variables exists.

A correlation coefficient is a measure of how variables are related.

Scatter Plots:

A scatter plot is a graph of the ordered pairs (x,y) of numbers consisting of the independent variable, x, and dependent variable, y.

A scatter plot is visual way to describe the nature of the relationship between the independent and dependent variables.

Example: (10-1) page 530

Construct a scatter plot for the following data:

Subject	Age x	Pressure y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

Correlation Coefficient: (Pearson Correlation Coefficient):

معامل ارتباط بيرسون

The correlation coefficient computed from the sample data measures the strength and direction of a linear relationship between two variables.

The symbol for the sample correlation coefficient is r.

The symbol for the population correlation coefficient is *p*.

The range of the correlation coefficient is from 1- to +1.

If there is a strong positive linear relationship between the variables, the value of r will be close to +1.

If there is a strong negative linear relationship between the variables, the value of r will be close to -1.

When there is no linear relationship between the variables or only a weak relationship, the value of r will be close to 0.

Strong negative linear relationship	no linear relationship	strong positive linear relationship

ملاحظات مهمة جدا جدا:

- تنحصر قيمة معامل ارتباط بيرسون بين -1 و+1
- إذا كانت قيمة معامل الارتباط بالسالب (نقول أن هناك ارتباط عكسي)
- إذا كانت قيمة معامل الارتباط بالموجب (نقول أن هناك ارتباط طردي)
- إذا كانت قيمة (r) أقل من 0.5 يقال أن هناك ارتباط ضعيف weak
- إذا كانت قيمة (r) من 0.5 إلى 0.74 يقال أن هناك ارتباط متوسط moderate
- إذا كانت قيمة (r) من 0.75 إلى 0.99 يقال أن الارتباط قوي strong

Formula for the Correlation Coefficient r:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{\left[n(\sum x^2) - (\sum x)^2\right]\left[n(\sum y^2) - (\sum y)^2\right]}}$$

Where n is the number of data pairs.

Example (10-4) – page 534

Compute the value of the correlation coefficient (Pearson Correlation Coefficient) for the data for the data obtained in the study of age and blood pressure given in Example (10-1):

Subject	Age x	Pressure y			
A	43	128			
B	48	120			
C	56	135			
D	61	143			
E	67	141			
F	70	152			

Solution:

Subject	Age x	Pressure y	xy	x^2	y^2
A	43	128	5504	1849	16384
B	48	120	5760	2304	14400
C	56	135	7560	3136	18225
D	61	143	8723	3721	20449
E	67	141	9447	4489	19881
F	70	152	10640	4900	23104
	345	819	47634	20399	112443

$$r=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right)-\left(\sum x\right)^2\right]\left[n\left(\sum y^2\right)-\left(\sum y\right)^2\right]}}=\frac{6(47634)-(345)(819)}{\sqrt{\left[6(20399)-(345)^2\right]\left[6(112443)-(819)^2\right]}}=0.897$$

The correlation coefficient suggests a strong positive relationship between the age (x), and blood pressure (y).

Example (20): Note: (from second exam, first term on 10-1-1430H)
A calculus instructor is interested in finding the relationship between the final exam grades of students enrolled in Calculus I (x) and Calculus II (y) at this college.

$\sum x=400$, $\sum y=380$ $\sum x^2=32656$ $\sum y^2=29546$
$\sum xy=31023$ $n=5$

Find the Pearson correlation coefficient (r):

$$r=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\left(\sum x^2\right)-\left(\sum x\right)^2\right]\left[n\left(\sum y^2\right)-\left(\sum y\right)^2\right]}}=\frac{5(31023)-(400)(380)}{\sqrt{\left[5(32656)-(400)^2\right]\left[5(29546)-(380)^2\right]}}=0.94$$

Interpret the value of the Pearson Correlation coefficient (r)!!!!!!!!!!!!!!

there is strong positive linear correlation.

Regression: الانحـــــــدار

Regression is statistical method used to describe the nature of the relationship between variables, that is, positive or negative, linear or nonlinear.

Line of best fit:

Means that the sum of the squares of the vertical distances from each point to the line is t at a minimum.

Regression line Equation: معادلة خط الانحدار (معادلة الخط المستقيم)

$$y = a + b\,x$$

b = slope الميل a= intercept of y axis الجزء المقطوع من محور الصادات

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2}$$

المطلوب في السؤال: أوجد قيمة من الميل والجزء المقطوع وما علينا إلا التعويض في المعادلات السابقة وبعد إيجاد القيم نعوض بها في معادلة الانحدار السابقة.

From Example (20): find the slope (b)and intercept (a) of the line if you have:

$\sum x = 400$, $\sum y = 380$ $\sum x^2 = 32656$ $\sum y^2 = 29546$

$\sum xy = 31023$ $n = 5$

$$a = \frac{\left(\sum y\right)\left(\sum x^2\right) - \left(\sum x\right)\left(\sum xy\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{(380)(32656) - (400)(31023)}{5(32656) - (400)^2} = 0.024$$

$$b = \frac{n\left(\sum xy\right) - \left(\sum x\right)\left(\sum y\right)}{n\left(\sum x^2\right) - \left(\sum x\right)^2} = \frac{5(31023) - (400)(380)}{5(32656) - (400)^2} = 0.95$$

Then the equation of the regression line is: $y = 0.024 + 0.95\,x$

Predict a Calculus II exam score for a student who receive an 80 in Calculus I.

بعد إيجاد قيمة كل من الميل والجزء المقطوع وبعد التعويض بهما في معادلة الانحدار ، فإنه يعطيني قيمة (x) ويطلب مني إيجاد قيمة (y) ويطلب مني أن أتنبأ بقيمة (y) أي إيجادها.

Note that: Calculus I = x, Calculus II = y

Y = 0.024 + 0.95 (80) = 76

Then Calculus II = 76 when Calculus I = 80

CHAPTER 13: SPEARMAN RANK CORRELATION COEFFICIENT:

معامل ارتباط سبيرمان للرتب

Formula for computing the spearman rank correlation:

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Where: d: differences in ranks, n: number of data pairs

EXAMPLE (13-7):

Two student were asked to rate eight different textbooks for a specific courses on categories, find the Spearman Rank Correlation:

Textbook	Student 1 rating	Student 2 rating			
A	4	4			
B	10	6			
C	18	20			
D	20	14			
E	12	16			
F	2	8			
G	5	11			
H	9	7			

نقوم بترتيب قيم x بإعطاء أصغر رقم 1 ثم الرقم اللي بعده 2 ثم 3 وهكذا إلى آخر رقم من قيم x ، وأكرر نفس العملية بالنسبة لقيم y كما في الجدول التالي:

Textbook	Student 1 rating	Student 2 rating	R1	R2	d	d^2
A	4	4	7	8	-1	1
B	10	6	4	7	-3	9
C	18	20	2	1	1	1
D	20	14	1	3	-2	4
E	12	16	3	2	1	1
F	2	8	8	5	3	9
G	5	11	6	4	2	4
H	9	7	5	6	-1	1
						30

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*30}{8(64-1)} = 0.645$$

Then, there is a moderate relationship between student 1 & student 2 rating.

هناك علاقة متوسطة طردية بين x وy

ملاحظة:

عمود d يمثل الفروق (حاصل طرح) رتب x ورتب y ومجموع d دائما يساوي صفر.

EXAMPLE: (FROM FIRST TERM EXAM 10-1-1430H)

هذا السؤال ورد في اختبار دوري سابق: إذا كانت الفروق بين رتب متغيرين هي كما يلي:

If the differences between the ranks of two variables are:

(-1, -3, 1, -2, 1, 3, 2, -1) then answer the following questions:

1. The sample size is: 8 حجم العينة عبارة عن عدد المشاهدات

2. The value of the correlation coefficient is: قيمة معامل ارتباط سبيرمان

d	-1	-3	1	-2	2	3	2	-1
d^2	1	9	1	4	4	9	4	1

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} = 1 - \frac{6*30}{8(64-1)} = 0.645$$

3. The correlation coefficient value means that there is: Moderate linear relationship.

بعد أوجدنا قيمة r يسألك ما هو نوع الارتباط وبما أن قيمة r تساوي 0.645 إذن فهناك ارتباط طردي متوسط.

EXPERCISE: FIND THE SPEARMAN CORRELATION:

Textbook	CALCULUS1	CALCULUS 2	R1	R2	d	d^2
A	80	70				
B	70	80				
C	90	90				
D	80	70				
E	60	50				
F	80	60				
G	50	40				

^^^ نهاية الفصل ^^^^

هذا العمل مجاني ولا يباع

ونتمنى من الله أن ينفع به أبناءنا في جميع أنحاء الكرة الأرضية ،،،،

ولا تنسونا من صالح دعاءكم ،،،،،،

gasim_hhh@hotmail.com