المحاضرة السادسة عشرة

الفصل السادس

الترابط و التنبؤ

Correlation and Prediction

6-1 المقدمة Introduction :

في كثير من التطبيقات نكون مهتمين بدراسة عدة ظواهر (عدة متغيرات) معاً، و ذلك بهدف معرفة إن كان هناك أيّ علاقة أو ارتباط بينهما ، لذلك سوف ننظر إلى مجموعتين مختلفتين من المشاهدات (القياسات) و نرى إن كان هناك علاقة ما بينها . هل التغير في إحدى المجموعتين يصاحبه تغير في المجموعة الأخرى و في أي اتجاه؟ و ما قوة تلك العلاقة ؟ ففي مجال الصحة كما في الاقتصاد و علم الاجتماع و غيرهما تطرح مثل التساؤلات الآتية : هل هذان المتغيران مرتبطان و ما طبيعة العلاقة بينهما و هل أحدهما ينبئنا عن الآخر؟ فمثلاً نتساءل هل هناك علاقة بين الطول و الوزن لمجموعة من الأشخاص. هل هناك علاقة للتعرض للإشعاع و الإصابة بمرض السرطان؟ وهل المدخن أكثر عرضه للإصابة بسرطان الرئة ؟ وهل هناك علاقة بين الوزن و ارتفاع ضغط الدم ؟.... إلخ من التساؤلات .

و للإجابة عن هذه الأسئلة نحتاج إلى دراسة العلاقة بين مجموعتين من القراءات (المشاهدات) مرتبة على شكل أزواج (X, Y) حيث X تمثل المتغير الأول و

Y تمثل المتغير الثاني و المعنى من كلمة مرتبة هو أن نجعل المكان الأول لمشاهدات المتغير الأول و المكان الثاني لمشاهدات المتغير الثاني. فإذا أخذنا عينة من الأشخاص البالغين فإن X ستكون طول الشخص ، و Y ستكون وزنه، وندعو قيم X مشاهدات المتغير المستقل (independent variable) وقيمY مشاهدات المتغير التابع (variable dependent) .

قراءة المتغير الأول X أو المتغير المستقل أحياناً يكون مسيطراً عليه أو متحكماً فيه و قراءة المتغير Y تكون نتيجة التجربة.

فإذا كان لدينا متغيران الأول كمية الدواء والثاني مدة الشفاء فإن مشاهدات المتغير الأول x يمكن التحكم بها ، بينما مدة الشفاء Y فتسجل فيما بعد. و بعد معرفة طبيعة العلاقة يمكننا التنبؤ بقيم Y عند معرفتنا بكمية الدواء المعطاة .

6-1-1 مخطط الانتشار Scatter Diagram :

قبل إجراء تحليل رياضي على البيانات المشاهدة لغرض معرفة إذا كان هناك أيّ علاقة بين أي متغيرين فإنه من الأفضل رسم ما يسمى بمخطط الانتشار لتلك البيانات إذ نمثل مشاهدات المتغير المستقل على المحور الأفقي و مشاهدات المتغير التابع على المحور العمودي ، لنجد أزواج المشاهدات ممثلة بنقاط مبعثرة في المستوى Oxy . إن توزيع تلك النقاط يعطينا صورة أولية تساعد في كشف العلاقة بين المتغيرين إن كانت موجودة .

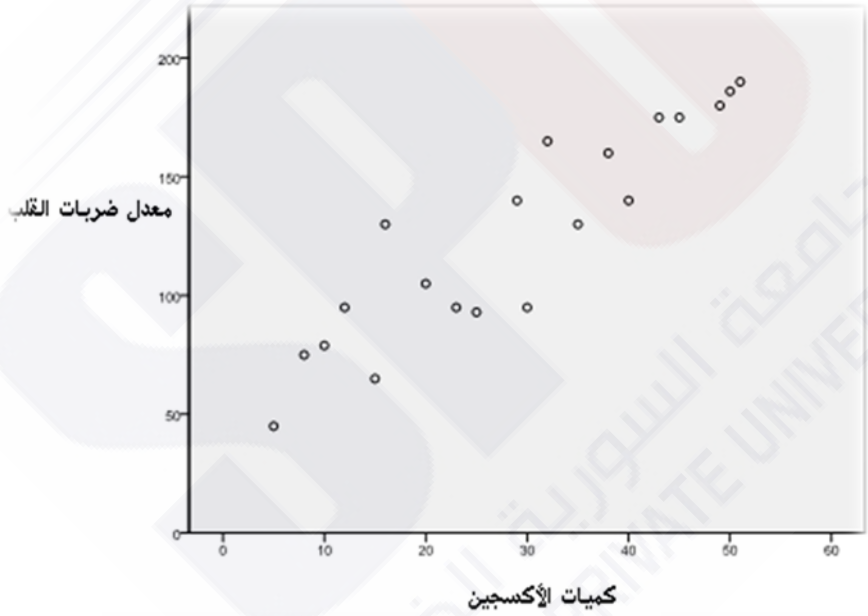
مثال (6-1):

في أحد البحوث اخترنا عشوائياً عشرين شخصاً لمعرفة العلاقة بين X المتغير المستقل وهو الكمية الكبرى للأوكسجين المستنشق و Y المتغير التابع وهو معدل ضربات القلب و سجلنا النتائج في الجدول الآتي **(6-1)**

x	43	49	50	12	8	32	51	30	35	23
y	175	180	186	95	75	165	190	95	130	95
x	25	16	38	40	29	15	10	20	5	45
y	93	130	160	140	140	65	79	105	45	175

الجدول (6–1)

و الشكل (6–1) الآتي يمثل شكل الانتشار بين المتغيرين X و Y لمشاهدات العينة



الشكل (6–1) مخطط الانتشار لكمية الأوكسجين ومعدل ضربات القلب.

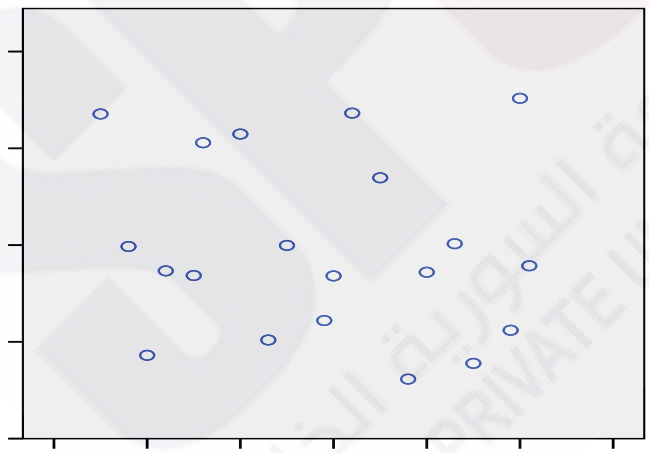
من مخطط الانتشار السابق نرى بوضوح أن هناك نزعة متمثلة في زيادة قيم y بشكل خطي مع تزايد قيم x ، و هذه العلاقة الخطية ليست تامة بمعنى أن هناك تغيراً عشوائياً ضمن مجموعة الأشخاص الذين استنشقوا نفس كمية الأكسجين .

فمثلاً في هذه العينة شخصان استنشقا الكميتين المتقاربتين 16 , 15 ، من الأكسجين في حين كانت معدلات ضربات القلب متباعدة 130 , 65 وهذا الفارق الكبير قد يعود لعوامل أخرى مثل الوزن – العمر .

6-1-2 أشكال الانتشار و الارتباط الخطي :

الغرض الأساسي من تحليل الارتباط هو قياس قوة العلاقة الخطية بين متغيرين. سوف نستعرض الآن بعض العلاقات الممكنة بين المتغيرين المستقل X و المتغير التابع Y.

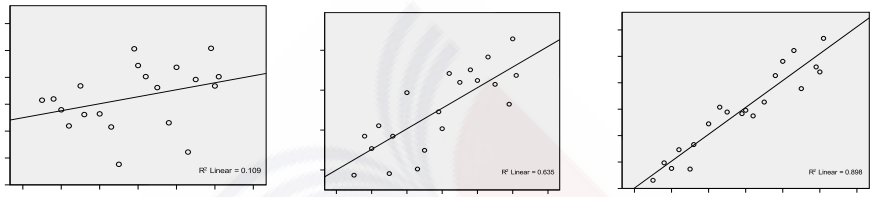
أولاً: إذا كانت النقاط منتشرة عشوائيا (مبعثرة) في المستوي كما في الشكل (6-2) فلا يوجد في هذه الحالة علاقة خطية بين x و y



الشكل (6-2) شكل مخطط الانتشار لمتغيرين لا علاقة بينهما

ثانياً: إذا تزايدت قيم المتغير y مع ازدياد قيم المتغير المستقل x بدرجات مختلفة، فيكون هناك ارتباط موجب بين المتغيرين x و y وقد تكون هذه

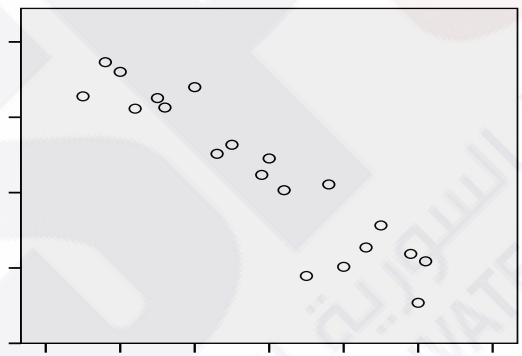
العلاقة قوية أو متوسطة أو ضعيفة و الأشكال (6–3 – a , b , c) توضح ذلك :



a– علاقة قوية **b– علاقة متوسطة** **c – علاقة ضعيفة**

الشكل (6–3) أشكال الانتشار لثلاث علاقات قوية ومتوسطة وضعيفة.

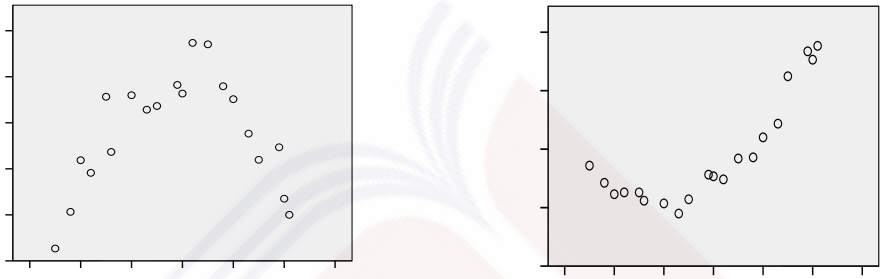
ثالثاً: إذا تناقصت قيم المتغير y بدرجات مختلفة مع تزايد مشاهدات x يكون هناك علاقة ارتباط عكسية بين x و y و نمثل ذلك في الشكل (6–4) :



الشكل (6–4) شكل الانتشار لعلاقة ارتباط خطية عكسية

رابعاً : قد تنتشر النقاط الممثلة لأزواج المشاهدات (x, y) بأشكال مختلفة ، و قد لا تكون خطية كما يوضح ذلك الشكلين التاليين (6–5) ففي الأول ترتبط المشاهدات y مع المشاهدات x بشكل غير خطي و لها شكل تابع من الدرجة

الثانية تقعره للأسفل ، إما في الشكل الثاني فهي علاقة تابع من الدرجة الثانية تقعره للأعلى.



التقعر نحو الأعلى **التقعر نحو الأسفل**

الشكل (6–5) شكلا الانتشار لعلاقتين من الدرجة الثانية

وسنهتم فقط بالعلاقات الخطية أو التي يمكن تحويلها إلى خطية .

6–2 معامل الارتباط الخطي لبيرسون :

:Pearson Linear Correlation coefficient

معامل الارتباط الخطي لبيرسون و يرمز له بـ R هو عبارة عن مقياس لقوة العلاقة الخطية بين متغيرين ، و هو يعكس مدى تماسك التأثير الناتج عن التغير في قيم المتغير x على التغير في قيم المتغير y و قيمة معامل الارتباط الخطي تكون دائماً بين 1- و 1+ فقيمته الموجبة دلالة على أن العلاقة الخطية بين x و y طردية أي تزايد قيم الأول يؤدي لتزايد قيم الثاني (إذا كانت R قريبة من الواحد : فالعلاقة قوية ، وإذا كانت قريبة من $\frac{1}{2}$ فتكون متوسطة ، وإذا كانت قريبة من الصفر فهي ضعيفة). الشكل(6–3) .

أما إذا كانت قيمته سالبة R<0 فإن العلاقة بين المتغيرين تكون سالبة أو عكسية. الشكل (6–4). فمثلاً نتوقع أن تكون R لمشاهدات x (الطول) مع مشاهدات y (الوزن) لمجموعة من الأشخاص موجبة أو طردية كما أننا نتوقع أن تكون R لمشاهدات y (سعر السيارة) مع x (عمر السيارة) قوية سالبة أو عكسية أي ينقص سعر السيارة مع زيادة عمرها .

و لمعامل ارتباط بيرسون عدة صيغ متكافئة فإذا فرضنا أزواج المشاهدات لمتغيرين x,y هي :

$$(x_1 , y_1),\ (x_2 , y_2)\ , . . .\ , (x_n , y_n)$$

فنعرف معامل الارتباط لبيرسون R بأنه متوسط جداءات القيم المعيارية للمتغيرين x,y أي:

$$R = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right).\left(\frac{y_i - \bar{y}}{S_y}\right) \qquad (1–6)$$

حيث S_y, S_x القيم المعيارية لمشاهدات y,x على الترتيب

$$S_x{}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i{}^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right)$$

$$S_y{}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i{}^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}\right)$$

وللعلاقة السابقة عدة صيغ متكافئة يمكن استخدامها في حساب معامل الارتباط منها :

$$R = \frac{\sum_{i=1}^{n} x_i .y_i - n.\bar{x}.\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i{}^2 - n.\bar{y}^2}} \qquad 2–6$$

وتكتب أيضا بالصيغة الآتية:

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad 3-6$$

مثال (6–2)

لدراسة العلاقة بين الكمية الكبرى للأكسجين المستنشق X ومعدل ضربات القلب Y ومن أجل عشرة أشخاص أخذنا أول عشر أزواج من القيم في الجدول (6–1) وجدنا من الشكل (6–1) أنه هناك علاقة إيجابية قوية بالاعتماد على الصيغة (6–3) لعلاقة الارتباط نلخص عملية إيجاد معامل الارتباط في الجدول الآتي :

الجدول (6–2)

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
43	175	9.7	36.4	353.08	94.09	1324.96
49	180	15.7	41.4	649.98	246.49	1713.96
50	186	16.7	47.4	791.58	278.89	2246.76
12	95	-12.3	-43.6	928.68	453.69	1900.96
8	75	-25.3	-63.6	1609.08	640.09	4044.96
32	165	-1.3	26.4	-34.32	1.69	696.96
51	190	17.7	51.4	904.47	313.29	2641.96
30	95	-3.3	-43.6	143.88	10.89	1900.96
35	130	1.7	-8.6	-14.62	2.89	73.96
23	95	-10.3	-43.6	449.08	106.09	1900.96

بجمع قيم الاعمدة نجد :

$$\sum_{i=1}^{n} \boldsymbol{x_i} = 333 \quad , \quad \sum_{i=1}^{n} \boldsymbol{y_i} = 1386$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}}) = 0 \quad , \quad \sum_{i=1}^{n} (\boldsymbol{y_i} - \bar{\boldsymbol{y}}) = 0$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})(\boldsymbol{y_i} - \bar{\boldsymbol{y}}) = 5786.2$$

$$\sum_{i=1}^{n} (\boldsymbol{x_i} - \bar{\boldsymbol{x}})^2 = 2148.1 \ , \ \sum_{i=1}^{n} (\boldsymbol{y_i} - \bar{\boldsymbol{y}})^2 = 18446.4$$

و منه :

$$\bar{\boldsymbol{x}} = \frac{333}{10} = 33.3 \quad , \ \bar{\boldsymbol{y}} = \frac{1386}{10} = 138.6$$

بالتبديل في الصيغة (6–3)

$$R = \frac{5786.2}{\sqrt{2148.1}\sqrt{18446.4}} = \frac{5786.2}{6301.1} = 0.92$$

ملاحظة :

إن استخدام الصيغة (6–2) أسهل في التطبيقات العملية . و لنقوم بحساب R مرة ثانية و باستخدام الجدول المساعد :

x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2
43	175	7525	1849	30625
49	180	8820	2401	32400
50	186	9300	2500	34596
12	95	1140	144	9025
8	75	600	46	5625
32	165	5280	1024	27225
51	190	9690	2601	36100
30	95	2850	900	9025
35	130	4550	1225	16900
23	95	2185	520	9025
$\sum_{i=1}^{n} x_i$ =333	$\sum_{i=1}^{n} y_i = 1386$	$\sum_{i=1}^{n} x_i \cdot y_i = 51940$	$\sum_{i=1}^{n} x_i^2 = 13237$	$\sum_{i=1}^{n} y_i^2 = 201546$

الجدول (6–3)

$$R = \frac{51940 - 10\,(33.3)(138.6)}{\sqrt{13237 - 10\,(33.3)^2}\sqrt{210546 - 10\,(138.6)^2}}$$

$$= \frac{5786.2}{\sqrt{2148.2}\sqrt{18446.4}} = 0.92$$

تدل الإشارة الموجبة والقيمة القريبة من الواحد على أن علاقة الارتباط بين كمية الأكسجين ومعدل ضربات القلب طردية وقوية أي زيادة كمية الأكسجين المستنشق يؤدي لزيادة ضربات القلب في الواقع العلاقة هنا ليست سببية، ويجب التركيز أن الارتباط لا يعني السببية.

اختبار الفروض حول معامل الارتباط الخطي :

:Testing of Hypotheses for Linear Correlation Coefficient

بعد حساب معامل الارتباط الخطي للعينة المعطاة نطرح السؤال الآتي هل قيمة R المحسوبة من الصيغ السابقة تدل على أن هناك علاقة بين المتغيرين في المجتمع الذي سحبت منه العينة ، بمعنى آخر ما هي القيمة التي إذا كانت قيمة R أكبر منها يكون هناك علاقة ارتباط وإذا كانت قيمة R أصغر منها تكون العلاقة ضعيفة ، ومن ثَمَّ لا يوجد ارتباط خطي بين قيم X وقيم Y للإجابة عن هذا السؤال نقوم باختبار فرض العدم H_0 كما يأتي:

المتغيران غير مرتبطين خطياً H_0 ضد الفرض البديل H_1 و هو ذو طرفين كمايلي (المتغيران مرتبطان خطياً) : H_1.

في الحقيقة الفرض البديل H_1 يمكن أن يكون على الصيغة

المتغيران مرتبطان خطياً بشكل موجب : H_1

أوالمتغيران مرتبطان خطياً بشكل سالب : H_1

فإذا استخدمنا الرمز ρ للدلالة على معامل الارتباط الخطي للمجتمع ، فنكتب

فرضية العدم $H_0 : \rho = 0$

مقابل الفرض البديل ذي طرفين $H_1 : \rho \neq 0$

أو مقابل الفرض البديل بطرف واحد إمّا $H_1 : \rho > 0$

وإمّا $H_1 : \rho < 0$

لاختبار الفروض السابقة H_0 , H_1 لا بد من الأمور الآتية:

1- حساب الاحصاء t_0 لمعامل الارتباط R للعينة حيث t_0 تحسب تحت صحة فرضية العدم H_0 كما يأتي :

$$t_0 = \frac{R-\rho}{S_R} = \frac{R-0}{\sqrt{\frac{1-R^2}{n-2}}} = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \qquad (4-6)$$

و لهذا الإحصاء توزيع ستودنت بدرجات حرية تساوي $\gamma = n-2$ حيث $S_R = \sqrt{\frac{1-R^2}{n-2}}$ هو الخطأ المعياري عند اعتبار R مقدر لـ ρ .

2- تحديد مستوى المعنوية α . قد تكون 0.01 , 0.02 , 0.05 ,....حسب أهمية البحث.

3- نستخدم جدول توزيع ستودنت بدرجة حرية $\gamma = n-2$ لتعيين القيمة الحرجة $(\gamma)\ t_{1-\alpha/2}$ التي تحصر على يسارها و تحت منحني الكثافة لستودنت بدرجة γ مساحة مقدارها $1-\alpha/2$.

4- ثم نتخذ القرار المناسب وفق الآتي:

a) إذا كانت القيمة المطلقة للقيمة المحسوبة t_0 أكبر من القيمة الحرجة ($|t_0| > t_{1-\alpha/2}(\gamma)$) نرفض الفرضية H_0 و نقبل H_1 ذات الطرفين ويكون الارتباط معنوياً.

b) إذا كانت القيمة المطلقة للقيمة المحسوبة t_0 أصغر من القيمة الحرجة ($|t_0| < t_{1-\alpha/2}(\gamma)$) نقبل الفرضية H_0 ، و نعتبر أن قيمة R ليست معنوية ، أي ليست هناك علاقة ارتباط بين المتغيرين.

مثال (6-3)

بالعودة للمثال السابق (6-2) حيث حسبنا معامل ارتباط بيرسون للعينة ، و كان $R = 0.92$ وهو قيمة تقديرية لـ ρ معامل ارتباط x و y نقوم باختبار معنوية معامل الارتباط R عند مستوى معنوية $\alpha = 0.05$ مثلا.

الحل:

1- نصيغ الفرض الإحصائي

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

2- نحسب قيمة t_0 من العلاقة (6-4) و تحت فرضية H_0

$$t_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} = \frac{0.92\sqrt{10-2}}{\sqrt{1-(0.92)^2}} \cong \frac{2.6}{\sqrt{0.15}} \cong 6.7$$

3- نعين منطقة قبول H_0 و هي المجال المعطى بالشكل:

$$\left[-t_{1-\alpha/2}(n-2)\,, t_{1-\alpha/2}(n-2)\right] = [-t_{0.025}(8), t_{0.975}(8)]$$

و القيمة $t_{0.975}(8) = 0.23$ نجدها في سطر 8 و عمود 0.025 في جدول توزيع ستودنت.

إذن $t_0 = 6.7$ لا تقع في المجال $[-2.306 , 2.306]$ أو $|t_0| > 2.306$ أي t_0 المحسوبة تقع في منطقة رفض H_0 ، ومن ثَمَّ الفرضية $\rho = 0$ غير صحيحة، و ذلك بدرجة ثقة أكبر من 95% ، و هذا يعني أننا على ثقة مقدارها 95% بأن هناك علاقة خطية بين الكمية الكبرى للأكسجين المستنشق و معدل ضربات القلب .

3-6 معامل سبيرمان لارتباط الرتب

Spearman's Rank Correlation Coefficient :

إن معامل الارتباط الخطي لبيرسون الذي سبق ذكره يمكن أن يستخدم لقياس الارتباط الخطي بين متغيرين كميين ، عندما تكون ملاحظات كُلٍّ من x و هي (x_i) و y وهي (y_i) مقادير كمية. فلا يمكن استخدامه إذا كانت المشاهدات ليست كمية (اسمية أو رتبية) ، لذلك دعت الحاجة لإيجاد مقياس للارتباط في حال كانت قيم أحد المتغيرين أو كليهما غير كمية. و من هذه المقاييس معامل سبيرمان لارتباط الرتب الذي يمكن استخدامه لقياس الارتباط بين المتغيرات التي يمكن ترتيب قيمها أي إذا كانت المتغيرات رتبية أو فئوية .

ليكن لدينا مجموعة مكونة من n من الأزواج المرتبة لمشاهدات العينة

$$(x_1 , y_1) , (x_2 , y_2) \ , \dots \dots \dots , (x_n , y_n)$$

و لنرمز بـ $r(x_i)$ لرتبة المشاهدة x_i في العينة $x_1 , x_2 , \dots , x_n$ (أي ترتيبها في العينة بعد ترتيب العينة تصاعديا) وبـ $r(y_i)$ لرتبة المشاهدة y_i في العينة $y_1 , y_2 , \dots , y_n$.

يتم استخدام رتب قيم المتغير x و رتب قيم المتغير y بشكل تصاعدي أو تنازلي معاً بشرط تطبيق نفس الطريقة لكلا المتغيرين. و في حال تساوي مشاهدتين $x_i = x_{i+1}$ نعطي لكل منها نفس الرتبة و تساوي متوسط رتبتي القيمتين .

وبعد استخراج رتب قيم المتغيرات نحسب معامل سبيرمان لارتباط الرتب كما يأتي:

a) معامل سبيرمان لارتباط الرتب هو معامل الارتباط الخطي لبيرسون لرتب المتغيرين أي معامل سبيرمان لارتباط الرتب يعطى من العلاقات المتكافئة الآتية:

$$r_s = \frac{\sum_{i=1}^{n}(\gamma_i - \overline{\gamma})(w_i - \overline{w})}{\sqrt{\sum_{i=1}^{n}(\gamma_i - \overline{\gamma})^2}\sqrt{\sum_{i=1}^{n}(w_i - \overline{w})^2}} \qquad (5-6)$$

$$r_s = \frac{\sum_{i=1}^{n}\gamma_i w_i - n\overline{\gamma}\overline{w}}{\sqrt{\sum_{i=1}^{n}\gamma_i{}^2 - n\,(\overline{\gamma})^2}\sqrt{\sum_{i=1}^{n}w_i{}^2 - n\,(\overline{w})^2}} \qquad (6-6)$$

$$r_s = \frac{n\sum_{i=1}^{n}\gamma_i w_i - \sum_{i=1}^{n}\gamma_i \sum_{i=1}^{n} w_i}{\sqrt{n\sum_{i=1}^{n}\gamma_i{}^2 - (\sum_{i=1}^{n}\gamma_i)^2}\sqrt{n\sum_{i=1}^{n}w_i{}^2 - (\sum_{i=1}^{n}w_i)^2}} \qquad (7-6)$$

حيث

$$\overline{w} = \frac{1}{n}\sum_{i=1}^{n} w_i \;،\; \overline{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i \;،\; w_i = r\,(y_i) \;،\gamma_i = r\,(x_i)$$

b) اذا لم يكن هناك تشابه بين رتب قيم كُلٍّ من المتغيرين x و y فإننا نحسب معامل سبيرمان لارتباط الرتب من الصيغة البسيطة المكافئة الآتية :

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i{}^2}{n(n^2-1)} \qquad (8-6)$$

حيث: $d_i = r\,(x_i) - r\,(y_i)$ هو الفرق بين رتبتي x_i و y_i. و يجوز استخدام العلاقة البسيطة هذه و الحصول على قيمة تقريبية في حال وجود عدد

قليل من الرتب المتساوية . أما إذا كانت عدد الرتب المتساوية في أحد المتغيرين أو كليهما كبيراً فيجب استخدام العلاقة الأولى فقط .

مثال (6–4) :

لدراسة العلاقة بين كمية التدخين x والمقيسة بمتوسط عدد السجائر اليومية و شدة الإصابة بسرطان الرئة y أخذنا عينة عشوائية من عشرة أشخاص من المدخنين الذين أصيبوا بمرض سرطان الرئة ، و سجلنا مشاهدات كل من المتغيرين x و y في الجدول الآتي :

x	5	10	15	15	20	25	30	30	30	35
y	A	B	A	B	C	E	D	C	E	E

حيث مشاهدات المتغير Y هي :

A = خفيفة جداً ، B= خفيفة ، C = إصابة متوسطة ، D = إصابة شديدة ، E = شديدة جداً .

أوجد معامل ارتباط الرتب لسبيرمان .

الحل : أولاً نوجد رتب قيم كل المتغيرات

x_i	5	10	15	15	20	25	30	30	30	35
$r(x_i)$	1	2	3.5	3.5	5	6	8	8	8	10

y_i	A	A	B	B	C	C	D	E	E	E
$r(y_i)$	1.5	1.5	3.5	3.5	5.5	5.5	7	9	9	9

نلخص العمليات الحسابية في الجدول (6–3) لحساب r_s من الصيغة البسيطة (6–8):

x	Y	رتبة x $r(x)=\gamma$	رتبة y $r(y)=w$	$d=r(x)-r(y)$	d^2
5	A	1	1.5	- 0.5	0.25
10	B	2	3.5	- 1.5	2.25
15	A	3.5	1.5	2	4
15	B	3.5	3.5	0	0
20	C	5	5.5	- 0.5	0.25
25	E	6	9	- 3	9
30	D	8	7	1	1
30	C	8	5.5	2.5	6.25
30	E	8	9	-1	1
35	E	10	9	1	1
					$\sum_{i=1}^{10} {d_i}^2 = 25$

الجدول (6–3)

$$r_s = 1 - \frac{6\sum_{i=1}^{n} {d_i}^2}{n(n^2-1)} = 1 - \frac{6\,(25)}{10\,(100-1)} = 1 - \frac{150}{990} = +0.848$$

نحسب معامل ارتباط الرتب لسبيرمان باستخدام الصيغة (6–5) أي نطبق معامل الارتباط الخطي لبيرسون على بيانات الرتب في العمودين الثالث و الرابع من الجدول(6–2) السابق :

$$r_s = \frac{\sum_{i=1}^{n}(\gamma_i-\overline{\gamma})(w_i-\overline{w})}{\sqrt{\sum_{i=1}^{n}(\gamma_i-\overline{\gamma})^2}\sqrt{\sum_{i=1}^{n}(w_i-\overline{w})^2}} = \frac{67}{\sqrt{80}\sqrt{79}} = +0.843$$

بمقارنة النتيجتين نلاحظ هناك فرقاً (بسيطاً) و الاختلاف بين النتيجتين يُعزى لوجود عدد كبير من الرتب المتشابهة ولاسيّما رتب قيم المتغير y و من قيمة معامل الارتباط هذه $r_s = 0.843$ يتضح أن هناك علاقة طردية قوية نوعاً ما

بين المتغيرين x و y ، و تلك العلاقة تعني أن شدة الإصابة بسرطان الرئة تعود لزيادة كمية التدخين.

6-4 معامل الاقتران و معامل التوافق :

:Coefficient Of Contingency Coefficient Of Association

يستخدم معامل الاقتران و معامل التوافق لقياس قوة الارتباط بين متغيرين اسميين (وصفيين) (nominal variables) حيث لا نستطيع استخدام مقياس الارتباط لبيرسون و مقياس ارتباط الرتب لسبيرمان لتلك البيانات . فعندما كلّ من المتغيرين x,y يأخذ فقط حالتين 1 , 0 (مدخن و غير مدخن أو مريض و سليم) .نستخدم معامل الاقتران. أما عندما أي من المتغيرين x و y أو كليهما يأخذ عدة قيم أو عدة حالات مثل 0,1,2 أو لون العينين (أسود– أزرق – بني) أو لون البشرة (أبيض – أسمر – أشقر) فنستخدم معامل التوافق لقياس شدة الارتباط بين المتغيرين.

أولاً : معامل الاقتران :

بفرض أننا نرغب في دراسة العلاقة بين صفتين لأفراد مجتمع ما و كل صفة تأخذ حالتين فقط مثل التدخين و الجنس ، فأي فرد من أفراد مجتمع ما سيكون مدخناً أو غير مدخن ، و كذلك سيكون إما ذكراً وإمّا أنثى ، أي إن الصفة الأولى x تقسم المجتمع إلى مدخنين و غير مدخنين و الصفة الثانية y تقسم المجتمع كذلك إلى فئتين ذكور و إناث ، فإذا رمزنا بـ A: لعدد المدخنين الذكور، B: لعدد المدخنين الإناث ، C: لعدد غير المدخنين الذكور ، D: لعدد غير المدخنين الإناث ، أو وفق الجدول الآتي:

الصفة X / الصفةY	الحالة الأولى (مدخنون)	الحالة الثانية (غير مدخنين)
الحالة الأولى (ذكور)	A	C
الحالة الثانية (إناث)	B	D

نعرف معامل الاقتران r_c بالعلاقة الآتية:

$$r_c = \frac{A\,D - B\,C}{A\,D + BC} \qquad (9-6)$$

مثال (6–5) :

عند دراسة علاقة التدخين بالتعليم في إحدى المؤسسات أخذت عينة عشوائية مكونة من 50 موظفاً ، و كانت النتائج :

التدخين / التعليم	لا يدخن	يدخن
متعلم	25	5
غير متعلم	10	10

احسب معامل الاقتران r_c بين التدخين و التعليم .

الحل :

باستخدام العلاقة (6–9):

$$r_c = \frac{25\,(10) - 5\,(10)}{25\,(10) + 5\,(10)} = \frac{200}{300} = 0.67$$

نجد شدة الارتباط متوسطة. أي نسبة المدخنين في مجتمع المتعلمين أقل من نسبة المدخنين في مجتمع غير المتعلمين.

ثانياً : معامل التوافق :

أوجد كرامر Cramer (1946) مقياساً للارتباط يستخدم عندما يكون للمتغيرين الوصفين أكثر من حالتين أو عندما يكون متغير وصفي له أكثر من حالتين و الثاني كمي ، ويدعى معامل التوافق . فإذا فرضنا أن للمتغير (الصفة) x الحالات الآتية ($x_1 , x_2 , \dots , x_r$) و للمتغير (الصفة) y الحالات ($y_1 , y_2 , \dots , y_s$) حيث إحداها على الأقل اسمية (وصفية) و رمزنا بـ f_{ij} لتكرارات العينة التي لها الحالة i للصفة الأولى و لها الحالة j للصفة الثانية y ورتبنا الجدول:

الصفة y / الصفة x	y_1 y_2 .... y_s	المجموع
x_1	f_{11} f_{12} .... f_{1s}	$f_{1.}$
x_2	f_{21} f_{22} .... f_{2s}	$f_{2.}$
.	.	.
.	.	.
x_r	f_{r1} f_{r2} .... f_{rs}	$f_{r.}$
المجموع	$f_{.1}$ $f_{.2}$ .... $f_{.s}$	$n = f_{..}$

$f_{i.}$:عدد التكرارات في العينة التي لها الحالة i للصفة الأولى x.

$f_{.j}$:عدد التكرارات في العينة التي لها الحالة j للصفة الثانية y.

من الجدول السابق نحسب المقدار B

$$B = \frac{{f_{11}}^2}{f_1.f_1} + \frac{{f_{12}}^2}{f_1.f_2} + \cdots + \frac{{f_{rs}}^2}{f_r.f_s}$$

و نعرف معامل التوافق r_a :

$$r_a = \sqrt{\frac{B-1}{B}} \qquad \textbf{(10–6)}$$

مثال (6–6):

بهدف دراسة علاقة لون البشرة لمجموعة من الأمهات مع لون بشرة المولود الأول لكل منهن . اخترنا بشكل عشوائي عينة من مئة أم و عرفنا المتغير x لون بشرة الأمهات (أبيض – حنطي – أسمر) و كذلك المتغير y لون بشرة المولود الأول و يأخذ نفس الحالات و سجلنا النتائج في الجدول الآتي :

الأمهات / المولود الأول	أبيض	حنطي	أسمر	المجموع
أبيض	27	6	7	40
حنطي	8	17	5	30
أسمر	5	7	18	30
المجموع	40	30	30	100

بين إذا كان هناك توافق بين لون بشرة الطفل الأول و لون بشرة الأم .

الحل :

نحسب $B = \frac{{f_{11}}^2}{f_1.f_1} + \frac{{f_{12}}^2}{f_1.f_2} + \cdots + \frac{{f_{33}}^2}{f_3.f_3}$

$$= \frac{(27)^2}{(40)(40)} + \frac{(6)^2}{(30)(40)} + \frac{(7)^2}{(30)(40)} + \frac{(8)^2}{(40)(30)} + \frac{(17)^2}{(30)(30)} + \frac{(5)^2}{(40)(30)} + \frac{(5)^2}{(30)(30)} + \frac{(7)^2}{(30)(30)} + \frac{(18)^2}{(30)(30)}$$

$$\cong 0.46 + 0.03 + 0.041 + 0.05 + 0.32 + 0.03 + 0.021 + 0.05 + 0.36 = 1.36$$

نبدل في (6- 10):

$$r_a = \sqrt{\frac{B-1}{B}} = \sqrt{\frac{0.36}{1.36}} = 0.51$$

إن معامل التوافق $r_a \cong 0.51$ يبين أن قوة الارتباط بين لون البشرة للأمهات و للأبناء متوسطة ليست قوية.