الفصل 3: سلاسل إحصائية ذات متغيرين (Bivariate statistical series)

أثناء دراسة مجتمع إحصائي ما يمكن أحيانا ملاحظة أكثر من متغير في نفس الوقت. ومن ثم نتعامل مع متغيرين (متغيرين كميين، نوعيين أو مختلط)، نتناول في هذا الفصل دراسة العلاقة بين متغيرين في مجتمع إحصائي واحد وذلك عمومًا لأننا نحاول معرفة ما إذا كان هناك رابط بينهما وما هي شدة الارتباط، نعرّف عندئذ متغيرين إحصائيين X و Y، وتسمى الثنائية (X, Y) متغيرة إحصائية ذات بعدين.

مثالا: دراسة نقطة الفرض ونقطة الامتحان لمقياس الرياضيات لطلبة سنة أولى.

I.1.3- جدول التوزيع المشترك و سحابة النقط (Contingency table or crosstab and scatter plot)

ليكن X و Y المتغيرين المدروسين، p عدد قيم المؤشر بالنسبة للمتغير X، q عدد قيم المؤشر بالنسبة للمتغير Y و n العدد الإجمالي للمتغيرات. مجموعة الثنائيات $(X_i, Y_j)_{1\leq i\leq p, 1\leq j\leq q}$ تشكل سلسلة إحصائية ذات متغيرين.

ملاحظة: إذا كان المتغير الأول X عبارة عن زمن تسمى السلسلة الإحصائية بسلسلة زمنية.

أ- جدول التوزيع المشترك (Contingency table or crosstab):

إذا كان المتغيرين كميين والبيانات متقطعة فإن توزيعهما عادة يعرض على شكل جدول ذومدخلين يسمى جدول التوزيع المشترك، أو الجدول ذي المدخلين.

$X \therefore Y$	Y_1	Y_2	…	…	Y_q	المجموع
X_1	n_{11}	n_{12}	…	…	n_{1q}	$n_{1.}$
X_2	n_{21}	…	…	…	…	…
…	…	…	…	…	…	…
X_p	n_{p1}	…	…	…	n_{pq}	$n_{p.}$
المجموع	$n_{.1}$	…	…	…	$n_{.q}$	$n_{..} = n$

ملاحظة: في حالة بيانات مستمرة فإننا نعوض X_i بمراكز الفئات x_i في الجدول السابق و بالتالي في جميع العلاقات، ونفس الشئ مع Y_i .

ونعرف التكرارات التالية:

- التكرار الهامشي (marginal frequency):

$n_{i.} = \sum_{j=1}^{q} n_{ij}$ تمثل مجموع مكونات السطر i

$n_{.j} = \sum_{i=1}^{p} n_{ij}$ تمثل مجموع مكونات العمود j

من الواضح أن لدينا $\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij} = \sum_{i=1}^{p} n_{i.} = \sum_{j=1}^{q} n_{.j} =$n

إذن: التوزيع الهامشي $n_{i.}$ هو عدد المشاهدات X_i للمتغير X مهما يكن توزيع المتغير Y.

التوزيع الهامشي $n_{.j}$ هو عدد المشاهدات Y_i للمتغير Y مهما يكن توزيع المتغير X.

مثال: كل الطلبة المتحصلين على العلامة 12 في الامتحان مهما كانت علامتهم في الفرض.

– التكرار النسبي الهامشي (Marginal relative frequency): و يعرف بالتكرار النسبي المشترك أو التكرارات النسبية الهامشية f_{ij}، $f_{i.}$ و $f_{.j}$:

$$f_{ij} = \frac{n_{ij}}{n} \quad و \quad f_{i.} = \frac{n_{i.}}{n} \text{ و } f_{.j} = \frac{n_{.j}}{n}$$

ملاحظة: التكرار النسبي المشترك يحقق مايلي:

$$\sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij} = 1 \text{ و } f_{ij} \geq 0; \forall i,j \text{ و } \sum_{j=1}^{q} f_{i.} = 1 \text{ و } \sum_{i=1}^{p} f_{.j} = 1$$

– التكرار النسبي الشرطي (conditional relative frequency): ويعرف من أجل كل قيمة لـ i و j :

$$f_{i/j} = \frac{n_{ij}}{n_{.j}} \quad و \quad f_{j/i} = \frac{n_{ij}}{n_{i.}}$$

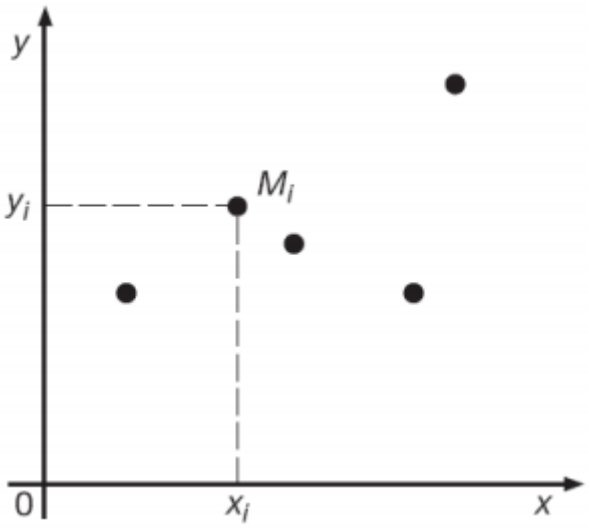
حيث $f_{j/i}$ هي توزيع المتغير X *بتثبيت المؤشر* i للمتغير Y

مثال: توزيع علامات الامتحان للطلبة الذين لهم نفس نقطة الفرض.

مثال 1: لتكن السلسلة الاحصائية ثنائية الأبعاد للثنائية (X , Y) التالية:

$X \therefore Y$	0	1	2	3	المجموع
2	3	4	0	6	13
3	4	3	3	2	12
4	2	3	3	2	10
المجموع	9	10	6	10	35

ب– سحابة النقط (scatter plot): يستخدم هذا النوع من الرسم البياني لتمثيل توزيع البيانات الثنائية حيث يتم رسم النقاط $M_i(X_i; Y_j)$ في معلم متعامد ومناسب، مجموعة النقط M_i تمثل القيم لكل ثنائية من المتغيرات وتسمى **سحابة النقط** في $\mathbb{R}^2$ لسلسلة ذات متغيرين كميين، و يعطي شكل سحابة النقط معلومات حول نوع الارتباط المحتمل.

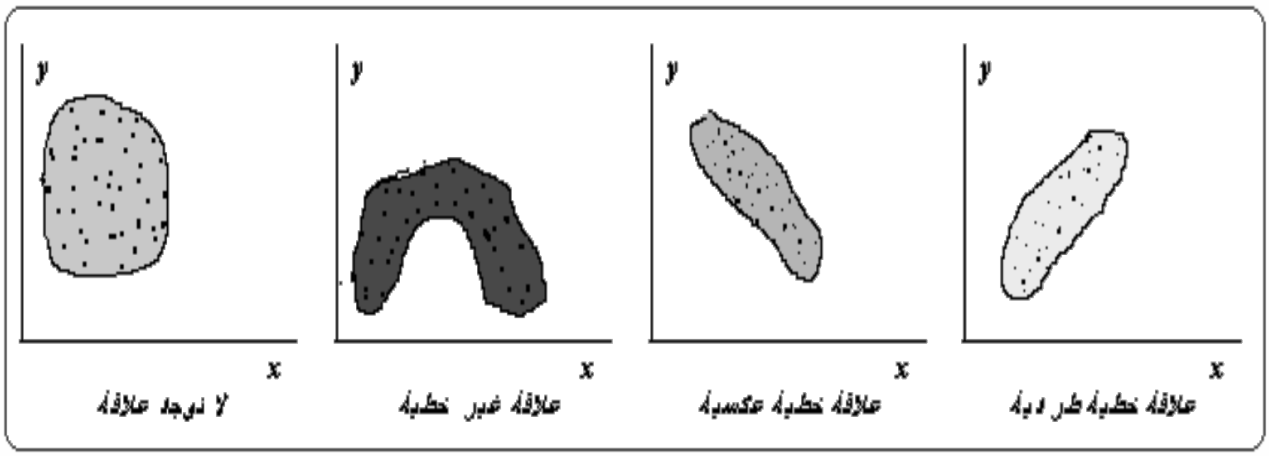


الشكل- 1- سحابة النقط

ملاحظة: حتى يكون لدراسة سحابة النقط معنى ينبغي ان يكون هناك عدد كبير من النقط.

– النقطة المتوسطة (The middle point): ونرمز لها بالرمز Γ حيث $G(\overline{X},\overline{Y})$ حيث $\overline{X}$ و $\overline{Y}$ يمثلان المتوسطات الهامشية للمتغيرين X و Y على الترتيب .

– أشكال الإنتشار:



2.3.I– التوزيعات الهامشية و الشرطية (Marginal and conditional distributions)

أ- التوزيعات الهامشية: يتم تحديد التوزيع الهامشي عن طريق عزل الاعمدة الأولى والأخيرة في جدول الإرتباط، العمود الاول يحتوي على قيم المتغير X والعمود الأخير يحتوي على التكرارات المقابلة لهذه القيم. بمعنى آخر على حافة جدول التوزيع المشترك يمكن استخراج فقط البيانات الخاصة بالمتغير X وفقط الخاصة بالمتغير Y. هذه التوزيعات يمكن تمثيلها على شكل جداول إحصائية:

- **التوزيع الهامشي لـ X:**

X	التكرار الهامشي $n_{i.}$	التكرار النسبي الهامشي $f_{i.}$
X_1	$n_{1.}$	$f_{1.}$
X_2	$n_{2.}$	$f_{2.}$
.	.	.
.	.	.
.	.	.
X_p	$n_{p.}$	$f_{q.}$
المجموع	n	**1**

- **التوزيع الهامشي لـ Y:**

Y	التكرار الهامشي $n_{.j}$	التكرار النسبي الهامشي $f_{.j}$
Y_1	$n_{.1}$	$f_{.1}$
Y_2	$n_{.2}$	$f_{.2}$
.	.	.
.	.	.
.	.	.
X_q	$n_{.q}$	$f_{.q}$
المجموع	n	**1**

- **خصائص السلاسل الهامشية:**

– المتوسطات الهامشية للمتغيرين X و Y :

$$\bar{X}_M = \frac{1}{n}\sum_i n_{i.}X_i = \sum_i f_{i.}X_i$$

و

$$\bar{Y}_M = \frac{1}{n}\sum_j n_{.j}Y_j = \sum_j f_{.j}Y_j$$

– التباينات الهامشية للمتغيرين X و Y :

$$V_M(X) = \frac{1}{n}\sum_{i=1}^{p} X_i^{\,2} - (\bar{\boldsymbol{X}})^{\mathbf{2}}$$

و

$$V_M(Y) = \frac{1}{n}\sum_{j=1}^{q} Y_j^{\,2} - (\bar{\boldsymbol{Y}})^{\mathbf{2}}$$

– الإنحرافات المعيارية الهامشية للمتغيرين X و Y :

$$\boldsymbol{\sigma_X} = \sqrt{V_M(X)} \quad \text{و} \quad \boldsymbol{\sigma_Y} = \sqrt{V_M(Y)}$$

مثال: بالرجوع للمثال 1 في الفصل 3، أحسب كل من $\boldsymbol{\sigma_X}$ و $\boldsymbol{\sigma_Y}$.

ب- التوزيعات الشرطية:

توزيع المتغير Y لما المتغير X يساوي X_i، يسمى التوزيع الشرطي لـ Y من أجل $X=X_i$:

$Y/X=X_i$	Y_1	$\dots\ Y_j\ \dots.$	Y_q	المجموع
التكرارات	n_{i1}	$\dots\ n_{ij}\ \dots$	n_{iq}	$n_{i.}$

هذا التوزيع لـ n_i مشاهدات الموافقة للشرط $X=X_i$ يمكن عرضه في شكل تكرارات نسبية شرطية :

$$\sum_{j=1}^{q} f_{j/i}=1 \quad \text{حيث} \quad f_{j/i}=\frac{n_{ij}}{n_{i.}}$$

$Y/X=X_i$	Y_1	$\dots\ Y_j\ \dots.$	Y_q	المجموع
التكرارات النسبية الشرطية	$f_{1/i}$	$\dots\ f_{j/i}\ \dots$	$f_{q/i}$	1

$f_{j/i}$: يعني التكرار النسبي لـ Y_j لما $X=X_i$ ولدينا p توزيع شرطي لـ Y من أجل (i=1.....p).

من أجل كل قيمة X_i يمكننا حساب المتوسط الشرطي $\overline{Y_i}$ وكذلك التباين الشرطي V_i :

$$V_i=\sum_{j=1}^{q} f_{j/i}\left(Y_j-\overline{Y_i}\right) \quad \text{و} \quad \overline{Y_i}=\sum_{j=1}^{q} f_{j/i}Y_j$$

وبالتالي

$$\overline{Y}=\sum_{i=1}^{p} f_{i.}\overline{Y_i}$$

بشكل متناظر يمكننا الحصول على التوزيع الشرطي لـ X من أجل $Y=Y_i$:

$X/Y=Y_j$	X_1	$\dots\ X_i\ \dots.$	X_p	المجموع
التكرارات	n_{1j}	$\dots\ n_{ij}\ \dots$	n_{pj}	$n_{.j}$

هذا التوزيع لـ n_i مشاهدات الموافقة للشرط $X=X_i$ يمكن عرضه في شكل تكرارات نسبية شرطية :

$$\sum_{i=1}^{p} f_{i/j}=1 \quad \text{حيث} \quad f_{i/j}=\frac{n_{ij}}{n_{.j}}$$

$X/Y=Y_j$	Y_1	$\dots\ Y_j\ \dots.$	Y_q	المجموع
التكرارات النسبية الشرطية	$f_{1/j}$	$\dots\ f_{i/j}\ \dots$	$f_{p/j}$	1

$f_{i/j}$: يعني التكرار النسبي لـ X_i لما $Y=Y_j$ ولدينا q توزيع شرطي لـ X من أجل (j=1.....q).

من أجل كل قيمة Y_j يمكننا حساب المتوسط الشرطي $\overline{X_J}$ وكذلك التباين الشرطي V_j :

$$V_j=\sum_{i=1}^{p} f_{i/j}\left(X_i-\overline{X_J}\right) \quad \text{و} \quad \overline{X_J}=\sum_{i=1}^{p} f_{i/j}X_i$$

وبالتالي

$$\overline{X}=\sum_{j=1}^{q} f_{.j}\overline{X_J} \ .$$

2.3.I– مقاييس الارتباط (correlation coefficients): يستخدم تحليل الارتباط لتحديد نوع و قوة العلاقة بين المتغيرين، وتقاس طبيعة، قوة و أثر العلاقة بين المتغيرين بعدة مقاييس منها:

أ– التغاير(covariance) : و يسمى ايضا التباين المشترك وهو عدد حقيقي يقيس ارتباط و كيفية انتشار النقاط حول النقطة المتوسطة.

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}(X_i - \bar{X})(Y_j - \bar{Y})$$

أو

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}X_iY_j - \overline{X}\overline{Y}$$

يشير التغاير إلى اتجاه العلاقة بين المتغيرين X و Y ، وهكذا يمكن تمييز الحالات التالية:

– إذا كانت $cov(X,Y) > 0$؛ فيمكننا القول أن العلاقة بين المتغيرين إيجابية. وفي هذه الحالة، يتغير هذان المتغيران في نفس الاتجاه.

– إذا $cov(X,Y) <0$؛ ومن ثم يمكننا القول أن العلاقة بين المتغيرين سلبية. وفي هذه الحالة، يتغير هذين المتغيرين في اتجاهين متعاكسين.

– إذا كانت $cov(X,Y)= 0$ فيمكننا القول أنه لا توجد علاقة بين المتغيرين. وفي هذه الحالة، فإن اختلاف أحدهما لا يؤدي إلى اختلاف الآخر.

- **بعض خواص التغاير:**

أ – $cov(X,Y) = cov(Y,X)$ ب– $cov(X,X) = V_M(X)$

ب– مقياس الارتباط الخطي (Linear correlation coefficient): ويعتبر مقياسا للارتباط الخطي بين متغيرين كميين وهو نسبة دون تمييز (بدون وحدة) ونرمز له بـ ρ_{XY}

$$\rho_{XY} = \frac{cov(X,Y)}{\sigma_X\sigma_Y}$$

– خواص:

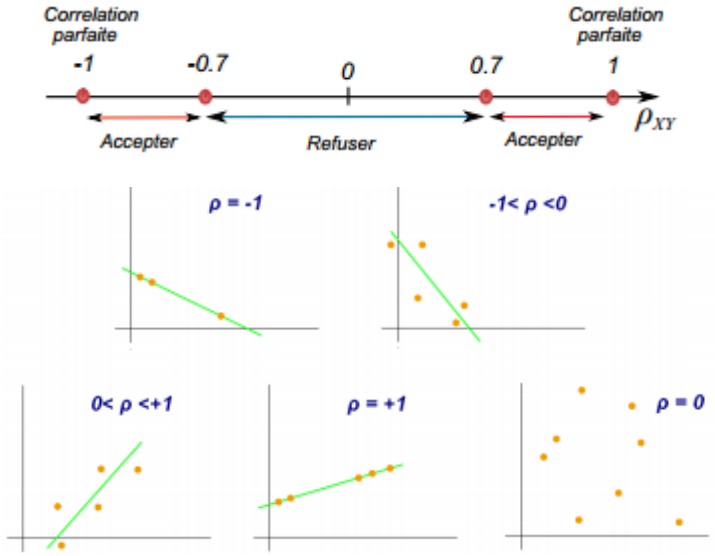
أ– $\rho(X,Y) = \rho(Y,X)$ ب– $-1 \leq \rho \leq 1$.

ت- $\rho(X,X)=1$ لان: $cov(X,X)=V_M(X)$ د- $\rho(X,-X)=-1$.

هـ- $\rho(aX+b,cY+d)=\rho(X,Y);\ a,c\neq 0$

ملاحظة:

- كلما اقتربت القيمة المطلقة ل ρ_{XY} من 1، فهذا يشير إلى وجود علاقة خطية قوية جدًا بين المتغيرين X و Y.
- لما القيمة المطلقة ل ρ_{XY} تساوي 1فإن هناك ارتباط أعظمي بين المتغيرين X و Y.
- كلما اقتربت القيمة المطلقة ل ρ_{XY} من 0، زاد غياب الاتصال الخطي بين X و Y.
- إذا كان $|\rho_{XY}| < 0.7$، وبالتالي يتم رفض التعديل الخطي (يرفض الخط المستقيم).
- اذا كان $|\rho_{XY}| \geq 0.7$، وبالتالي يتم قبول التعديل الخطي (يقبل الخط المستقيم).



الشكل- 3- أمثلة على مخططات التشتت ذات قيم معامل الارتباط المختلفة.

ملاحظة: توجد ثنائيات من المتغيرات لها نفس التوزيعات الهامشية لكنها تختلف في التغاير و الارتباط لهذا فإن التغاير و معامل الارتباط الخطي يقيسان التداخل بين المتغيرين.

مثال: بالرجوع للمثال 1 في الفصل 3، أحسب ρ_{XY} . ماذا تستنتج؟

I.3.3– مستقيم الانحدار ومستقيم مايير (Regression line and Mayer line)

أ– مستقيم الانحدار:

الفكرة هي تحويل سحابة من النقط إلى مستقيم، يجب أن يكون هذا المستقيم أقرب ما يمكن إلى كل نقطة، ولذلك سنسعى إلى تقليل الاختلافات بين النقاط والمستقيم، ولهذا نستخدم طريقة المربعات الصغرى. تهدف هذه الطريقة إلى شرح سحابة النقاط بمستقيم يربط بين المتغيرين X و Y ، وهذا يعني: $Y = aX + b$

بحيث تكون المسافة بين سحابة النقط والمستقيم في حدها الأدنى. ولذلك فإن مستقيم الانحدار الذي يجعل المسافة بينه وبين النقاط في حدها الأدنى و الذي يشمل **النقطة المتوسطة** لسحابة النقط يعطى بواسطة:

$$D(Y/X): Y = aX + b$$

حيث: $a = \frac{cov(X,Y)}{V_M(X)}$ و $b = \bar{Y} - a\bar{X}$

أو

$$D(X/Y): X = a'Y + b'$$

حيث: $a = \frac{cov(X,Y)}{V_M(Y)}$ و $b' = \bar{Y} - a'\bar{X}$

ب– مستقيم مايير: ويتم بكتابة معادلة المستقيم (G_1G_2) حيث G_1 النقطة المتوسطة لنصف النقاط الأولى و G_2 النقطة المتوسطة للنقاط المتبقية.

ملاحظة:

- يسمح مستقيم التعديل المرسوم من سحابة النقاط بيانيًا بتقدير Y بمعرفة X (تقدير لـ X بمعرفة Y).
- معادلة مستقيم التعديل تسمح لنا بحساب تقدير لـ Y بمعرفة X (تقدير لـ X بمعرفة Y).

مثال: من أجل السلسلة الإحصائية ذات المتغيرين X و Y التالية:

X(نفقات الإعلان بالآلاف دج)	10	15	20	30	35
Y (حجم المبيعات بالآلاف دج)	400	700	800	900	950

1- مستقيم الانحدار:

$$a = \frac{cov(X,Y)}{V_M(X)} = \frac{1650}{86} = 19.19 \quad \text{و} \quad b = \bar{Y} - a\bar{X} = 750 - 19.19 \times 22 = 327.82$$

ومنه مستقيم الإنحدار هو

$$Y = 19.19\,X + 327.82$$

ومقياس الارتباط الخطي هو:

$$\rho_{XY} = \frac{\text{cov(X,Y)}}{\sigma_X \sigma_Y} = \frac{1650}{9.27 \times 194.93} = 0.91$$

نلاحظ أن: $\rho_{XY} \approx 1$

وبذلك هناك علاقة خطية موجبة بين X و Y.

حجم المبيعات لمدة شهر حيث يبلغ الإنفاق الإعلاني 40.000 دج (X=40) هو:

$$Y = 19.19 \times 40 + 327.82 = 1095.42$$

أي 1095420 دج.

ب- مستقيم ماير:

$\overline{X_1}$ متوسط نصف النقاط الأولى: $\overline{X_1} = 12.5$ ، $\overline{Y_1}$ متوسط نصف النقاط الأولى: $\overline{Y_1} = 550$

$\overline{X_2}$ متوسط نصف النقاط الأولى: $\overline{X_2} = 28.33$ ، $\overline{Y_2}$ متوسط نصف النقاط الأولى: $\overline{Y_2} = 833.33$

ومنه $G_1(12.5; 550)$ و $G_2(28.33\ ;\ 883.33)$

$$a = \frac{883.33 - 550}{28.33 - 12.5} = 21.06$$

ولدينا النقطة G_1 تنتمي الى المستقيم $Y = 21.06\,X + b$ أي:

$$550 = 21.06 \times 12.5 + b$$

ومنه $b = 286.75$

وبالتالي: $Y = 21.06\ X + 286.75$.