المحاضرة السابعة عشرة

5-6 الانحدار (Ragression) :

يعرف الانحدار بأنه العلاقة التي تربط بين متغيرين كميين y , x و الغرض من دراسة هذا الموضوع هو تحديد هذه العلاقة ثم استخدامها للتنبؤ بقيم المتغير التابع y إذا علمت قيمة المتغير المستقل x . و بعبارة أخرى فإن الهدف هو تقدير متوسط المتغير y عند معرفة قيمة المتغير x . على سبيل المثال إذا افترضنا أن x المتغير الذي يمثل طول الشخص ، و y المتغير الذي يمثل وزن الشخص وسجلنا قياسات أطوال مجموعة من الأشخاص وأوزانهم فإننا نتوقع علاقة تربط طول الشخص (x) بوزنه (y) . فإذا اكتشفنا هذه العلاقة يمكننا تقدير متوسط وزن مجموعة الأشخاص الذين لهم طول مفروض سلفاً. و يعود مفهوم الانحدار Regression للعالم غالتون عام 1886 عندما درس العلاقة بين أطوال الآباء و أطوال أبنائهم فقد أخذ عينة من الآباء و قاس أطوالهم و أطوال أبنائهم ، فوجد أن متوسط أطوال الأبناء ينزع ليكون قريباً من متوسط أطوال الأبناء للمجتمع . و الدليل على ذلك فقد لاحظ أن أطوال أبناء الآباء طوال القامة أقصر من أطوال آبائهم و أطوال أبناء الآباء قصار القامة أطول من آبائهم . لذلك أطوال الأبناء ترتبط بأطوال الآباء لكنها دائماً تنزع باتجاه معدل أطوال المجتمع . أو معدل أطوال الأبناء تنحدر من معدل طول المجتمع و تدعى هذه الظاهرة بالانحدار نحو المعدل .

و لهذا الموضوع تطبيقات كثيرة في العلوم الصحيحة كما في العلوم الاجتماعية والاقتصادية و الزراعية ... الخ. فمثلاً قد نريد التنبؤ بوزن الشخص إذا علمنا طوله أو عمره أو التنبؤ بشدة الإصابة بسرطان إذا عُلمت كمية التدخين أو التنبؤ

بعدد ضربات القلب إذا علمنا الكمية الكبرى للأكسجين المستنشق أو التنبؤ بإمكانية الإصابة بنوبة قلبية إذا علمنا كمية الكوليسترول.

لدراسة العلاقة بين متغيرين نأخذ عينة عشوائية ونسجل

$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ مشاهدات المتغيرين (x, y) ثم نرسم المخطط الانتشاري ، فإذا تبين أن نقاط العينة تتوضع حول منحى خطي (مستقيم في المستوي OXY) نكون قد أوجدنا دليلاً أولياً على وجود علاقة خطية بين المتغيرين x ,y . في الخطوة التالية نبدأ بتحليل الانحدار أي الكشف عن العلاقة الرياضية $y = ax + b + e$ التي تربط x بـ y . و أخيراً يمكننا الاستفادة من هذه العلاقة في تقدير قيم المتغير التابع y بدلالة قيم معطاة للمتغير المستقل x .

ندعو المعادلة السابقة بمعادلة الانحدار و ندعو الثوابت a, b معالم (وسطاء) العلاقة، أما الحد e فهو الخطأ المرتكب وهو الفرق بين القيم المشاهدة y_i ومقدراتها $ax_i + b$ حيث $1 \le i \le n$ و ندعو المستقيم $y = ax + b$ بخط الانحدار.

1–5–6 تقدير المعالم بطريقة المربعات الصغرى :

عند رسم مخطط الانتشار للبيانات سنجد كما ذكرنا مجموعة من النقاط المنتشرة على شكل منحى خطي ، و لن تكون واقعة على استقامة واحدة ، و ذلك لأن المتغير y يتأثر بعوامل و متغيرات أخرى عدا المتغير x ، و سنحاول تعيين ذلك المستقيم الذي يمر وسط هذه النقاط ، و بمعنى آخر سنبحث عن ذلك الخط المحدد بـ a و b و الذي يكون الأقرب إلى تلك النقاط و بأسلوب تحليلي إذا بدلنا قيم المشاهدات

في المعادلة المفترضة فسنحصل على n $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ علاقة :

$$y_1 = ax_1 + b + e_1$$

$$y_2 = ax_2 + b + e_2 , \dots$$

$$y_n = ax_n + b + e_n$$

حسب طريقة المربعات الصغرى سنبحث عن القيمتين a وb اللتين تجعلان $\sum_{i=1}^{n} {e_i}^2$ مجموع مربعات الفروق بين القيم المشاهدة y_i و المحسوبة $ax_i + b$ أصغر ما يمكن فإذا رمزنا بـ $\hat{a}$, $\hat{b}$ لتلك القيمتين اللتين تجعل المقدار الآتي أصغر ما يمكن بالنسبة لكل من a , b

$$Q\ (\ a\ , b\ ) = \sum_{i=1}^{n} \left(y_i - \hat{a}x_i - \hat{b}\right)^2$$

حيث إن المقدار $Q\ (\ a\ , b\ )$ يمثل مجموع الفروقات بين القيم الفعلية والقيم المقدّرة. نجعل مشتقي هذه المعادلة بالنسبة لكل من α و b مساوياً للصفر :

$$\sum_{i=1}^{n} x_i \left(y_i - \hat{a}x_i - \hat{b}\right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \hat{a}x_i - \hat{b}\right) = 0$$

وبحل هاتين المعادلتين بالنسبة لـ $\hat{a}$ و $\hat{b}$ نجد :

$$\hat{a} = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (11\text{–}6)$$

$$\hat{b} = \bar{y} - a\bar{x} \qquad (12\text{–}6)$$

حيث $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

و يمكن استخدام صيغة حسابية مكافئة لحساب $\hat{a}$ تنتج من تقسيم كل من البسط والمقام في (6–11) على n حيث يحصل :

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} {x_i}^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \qquad (13-6)$$

أو

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{y}\bar{x}}{\sum_{i=1}^{n} {x_i}^2 - n(\bar{x})^2} \qquad (14-6)$$

<u>ملاحظة</u> : ندعو المعادلة $\hat{y} = \hat{a}\,x + \hat{b}$ بمقدر المربعات الصغرى لمعادلة الانحدار للمجتمع $y = a\,x + b + e$.

مثال (6–7) :

بالعودة للمثال (6–2) الذي درسنا فيه ارتباط الكمية الكبرى للأكسجين المستنشق (x) مع معدل ضربات القلب (y) و بالنظر إلى الشكل (6–1) الذي يوضح وجود علاقة خطية بين مشاهدات x و مشاهدات y . لنستخدم طريقة المربعات الصغرى لإيجاد هذه العلاقة $\hat{y} = \hat{a}\,x + \hat{b}$

الحل :

في الجدول (6–2) حسبنا القيم الآتية :

$$\sum_{i=1}^{10} \boldsymbol{x}_{\mathbf{i}} = 333 \quad , \quad \bar{\boldsymbol{x}} = \frac{333}{10} = 33.3 \quad , \quad \sum_{i=1}^{10} \boldsymbol{y}_{\boldsymbol{i}} = 1386$$

$$\bar{\boldsymbol{y}} = \frac{1386}{10} = 138.6 \ , \sum_{i=1}^{10} \boldsymbol{x}_{\boldsymbol{i}} \boldsymbol{y}_{\boldsymbol{i}} = 51940 \ , \ \sum_{i=1}^{n} {\boldsymbol{x}_{\boldsymbol{i}}}^{\mathbf{2}} = 13237$$

نعوض في العلاقة (6–14) :

$$\hat{a} = \frac{\sum_{i=1}^{n} x_i y_i - n\,\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2} = \frac{51940-10\,(33.3)(138.6)}{13237\,-10\,(33.3)^2} = 2.694$$

ثم في العلاقة (6−12) :

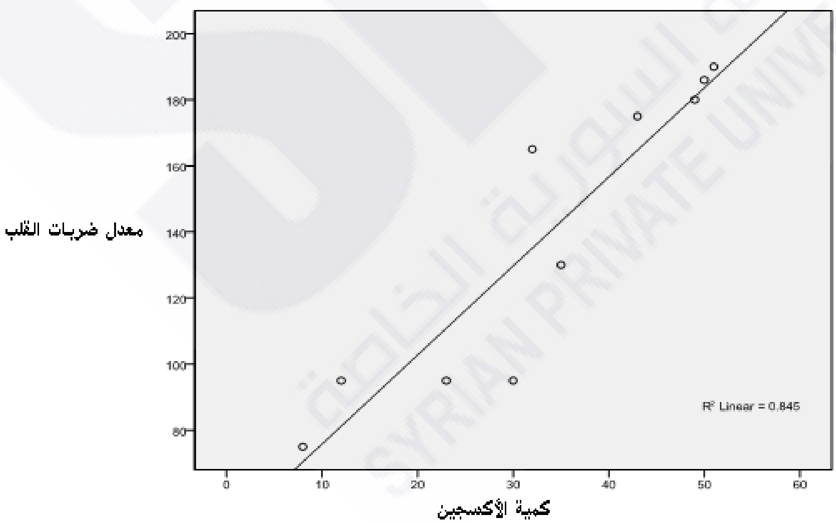
$$\hat{b} = \bar{y} - a\,\bar{x} = 138.6 - (2.694)(33.3) \cong 48.9$$

فيكون مقدر المربعات الصغرى لمعادلة انحدار معدل ضربات القلب y على الكمية الكبرى للأكسجين المستنشق x هو :

$$\hat{y} = 2.694\ x + 48.9$$

يمكن من هذه المعادلة التنبؤ بمعدل ضربات القلب لشخص إذا كانت كمية الأكسجين المستنشقة على سبيل المثال تساوي $x = 30$ و ذلك بأن نعوض قيمة 30 بـ x لنحصل على $\hat{y} = 2.694\ (30) + 48.9 = 129.7$

و لرسم خط الانحدار و هو المستقيم الذي معادلته $y = 2.694\ x + 48.9$ يكفي تحديد نقطتين يمر منهما $M_1(0\ ,48.9)\ ,M_2(30\ ,129.7)$



الشكل (6−6) مستقيم الانحدار لعلاقة كمية الاكسجين بمعدل ضربات القلب.

6-5-2 اختبار الفرضيات حول ميل خط الانحدار الخطي a :

وجدنا أن مقدر معادلة الانحدار المتغير y على المتغير x هي :

$$\hat{y} = \hat{a}\,x + \hat{b} \qquad (15\text{–}6)$$

و عندما لا يكون هناك علاقة خطية بين y و x فإن ميل المستقيم في تلك المعادلة يكون صفراً $a = 0$ و هذا يكافئ إن كان معامل الارتباط الخطي ρ بين المتغيرين x و y معدوماً ، أي إن : $\rho = 0$ و من ثَمَّ يمكن استبدال فرضية العدم و الفرضية البديلة لانعدام a أي $H_0 : a = 0$ (لا يوجد علاقة خطية مقابل الفرضية) و $H_1 : a \neq 0$ (هناك علاقة خطية) بفرضية العدم و الفرضية البديلة المتعلقة بـ ρ و المكافئة لها $H_0 : \rho = 0$ مقابل $H_1 : \rho \neq 0$.

فإذا فرضنا أنه لا علاقة خطية بين x و y (a=0) فهذا يكافئ $\rho = 0$ فعندئذ من أجل مستوى دلالة α تكون الفرضية صحيحة إذا وقعت قيمة الإحصاء

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}}$$

في المجال $\left[-t_{\alpha/2}(n-2)\ ,\ t_{\alpha/2}(n-2)\right]$ حيث n حجم العينة و R مقدر معامل الارتباط الخطي ρ .

مثال(6–8):

هل البيانات الواردة في المثال (6–2) تشير لوجود علاقة خطية بين معدل ضربات القلب y و الكمية الكبرى للأكسجين المستنشق . وذلك من أجل $\alpha = 0.01$.

الحل : نقوم بالخطوات الآتية :

1– نضع فرضية العدم و الفرضية البديلة ، و ذلك من أجل $\alpha = 0.01$

$H_1 : a \neq 0$ لا يوجد علاقة خطية مقابل الفرضية $H_0 : a = 0$
هناك علاقة خطية

2- حسبنا في المثال (6-2) مقدر معامل الارتباط r فوجدنا $r \cong 0.92$
وحسبنا في المثال (6-3) قيمة إحصائية التقدير $t_0 \cong 6.7$.

3- نعين منطقة القبول و ذلك من أجل $\alpha = 0.01$. نحسب $\alpha/_2 = 0.005$
نعين القيمة $t_{0.005}(8)$ والواقعة في جدول ستودنت في سطر درجة الحرية 8
و عمود 0.005 فنجد $t_{0.005}(8) = 3.355$.

4- بما أن $t_0 = 6.7$ تقع خارج $[-3.355 \quad 3.355]$ نرفض H_0 و
نقبل H_1 أي يوجد علاقة خطية بين معدل ضربات القلب و كمية الأكسجين .

6-5-3 معامل التحديد Coefficient Of Determination

بفرض أن x , y متغيران مرتبطان خطياً و بفرض أن علاقة الانحدار للعينة $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ هي $\hat{y} = \hat{a}\,x + \hat{b}$ حيث $\hat{a}$ و $\hat{b}$ مقدرا الوسيطين a و b نعرف معامل التحديد لهذه العلاقة و الذي سنرمز له بـ R^2 بأنه نسبة التباين (أو التغير) في بيانات المتغير y المفسرة بالتباين (أو بالتغير) في قيم المتغير المستقل x. و بعبارة أخرى هي نسبة التباين المفسر بالنموذج إلى التباين الكلي في بيانات المتغير y

$$R^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \qquad (16-6)$$

حيث $\sum_{i=1}^{n}(y_i - \bar{y})^2$ يقيس حجم تباين المشاهدات عن متوسطها بينما $\sum_{i=1}^{n}(\hat{y} - \bar{y})^2$ فيقيس حجم تباين $\widehat{y_i}$ بيانات المتغير y المحسوبة من النموذج (6−15) و يمكن أن نثبت أن

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \qquad (17-6)$$

حيث $y_i = \hat{a}\,x_i + \hat{b}$.

إن قيمة معامل التحديد تحقق $0 \le R^2 \le 1$. فكلما اقتربت قيمة R^2 من الواحد زادت جودة النموذج في تفسير التغير . و كلما كانت R^2 أقرب إلى الصفر كانت النموذج أقل جودة في تفسير التغير في البيانات. وإذا $R^2=1$ فإن النقاط $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ تقع جميعها على خط معادلة الانحدار .

نستنتج من أجل معادلة الانحدار $\hat{y} = \hat{a}\,x + \hat{b}$.

1− معامل التحديد R^2 ليس الا مربع معامل الارتباط الخطي R .

2− ومعامل الارتباط الخطي R ليس إلا الجذر التربيعي لمعامل التحديد مضروباً بإشارة $\hat{a}$ حيث :

$$Sig(\hat{a}) = +1 \quad \text{عندما} \quad 0 < \hat{a}$$

$$\text{و} \quad Sig(\hat{a}) = -1 \quad \text{عندما} \quad \hat{a} < 0$$

مثال (6−9):

بالعودة لمثال كمية الأكسجين و معدل ضربات القلب حيث كانت معادلة الانحدار الخطي $\hat{y} = 2.694\,x + 48.9$

و بما أن $\hat{a} > 0$ فإن $R^2 = \rho^2 = (0.92)^2 \cong 0.85$ و نفسر ذلك بقولنا إن: 85% تقريباً من التغير في قيم y (معدل ضربات القلب) تفسر بمعادلة خط الانحدار المقدرة أي تفسر بتغير كمية الأكسجين المستنشق .

تمارين ومسائل

1- أجريت تجربة لدراسة تأثير درجة الحرارة x على نتائج إحدى العمليات الكيميائية وتم الحصول على البيانات الآتية:

x	−5	−4	−3	−1	−1	0	1	2	3	4	5
y	1	5	4	7	10	8	9	13	14	13	18

أ- احسب معامل الارتباط الخطي لبيرسون .

ب- اختبر الفرضية $H_0: \rho = 0$ ضد الفرضية البديلة $H_1: \rho \neq 0$ عند مستوى معنوية α=0.01 .

ت- أوجد نموذج الانحدار الخطي المقدر.

ث- احسب معامل التحديد وفسره .

2- إذا كان معروف أن هناك علاقة بين فترة الإصابة بمرض معين وعدد البكتريا في العضو المصاب. تم اختيار 10 مصابين بهذا المرض وسجلت أطول فترات إصابتهم بالمرض عند دخولهم المستشفى وحصلنا على البيانات الآتية:

X عدد البكتريا (بالألف)	9	10	5	7	10	6	7	4	8	6
Y فترة الإصابة (باليوم)	12	11	8	9	13	10	14	8	11	7

أ- احسب معامل الارتباط الخطي لبيرسون .

ب- احسب معامل سبيرمان لارتباط الرتب ، ثُمَّ قارن بين النتيجتين.

3- أجريت تجربة لتقدير العلاقة بين درجة الحرارة ومعدل دقات القلب في الضفدعة المسماة Rana pipiens سجلت البيانات في الجدول الآتي:

الحيوان	1	2	3	4	5	6	7	8	9
X درجة الحرارة	2	4	6	8	10	12	14	16	18
Y دقات القلب بالدقيقة	5	11	11	14	22	23	32	29	32

أ- احسب معامل سبيرمان لارتباط الرتب

ب- أوجد معادلة الانحدار المقدرة؟

ت- اختبر فرض العدم $H_0: a = 0$ ضد الفرض البديل $H_1: a \neq 0$ عند مستوى معنوية α=0.05.

ث- احسب معامل التحديد للعلاقة المقدرة وفسر النتيجة.

4- أجريت تجربة لتقدير العلاقة بين العمر ودقات القلب (في الدقيقة) عند الإناث اللواتي أعمارهنّ تتراوح من واحد إلى 13 سنة . استخدم البيانات المعطاة في

الجدول الآتي في إيجاد:

الأنثى	1	2	3	4	5	6	7	8	9	10	11	12	13
X العمر	1	2	3	4	5	6	7	8	9	10	11	12	13

83	84	88	90	90	88	93	92	99	102	108	108	111	y دقات القلب

أ- أوجد معادلة الانحدار المقدرة بين العمر وعدد دقات القلب؟

ب- احسب معامل التحديد .

ت- اختبر معنوية معادلة الانحدار من أجل $\alpha = 0.02$.

5- الجدول الآتي يبين طول الجمجمة x وعرضها y بالمليمتر والمطلوب إيجاد معامل الارتباط الخطي ، ومعامل سبيرمان لارتباط الرتب .

71	58	72	73	79	66	76	70	80	63	X الطول
35	39	37	42	46	39	38	45	42	40	Y العرض

6- الجدول الآتي يعطي قياسات تمثل طول الأب وطوال الابن الأصغر.

(البيانات مقيسه لأقرب بوصة).

74	69	72	70	64	68	x
70	66	73	69	68	67	y

أوجد معامل الارتباط الخطي البسيط.

7- لدراسة العلاقة بين عدد الكيلوجرامات التي يفقدها شخص في برنامج لإنقاص الوزن وعدد الأسابيع التي يقضيها لإنقاص الوزن ، اختيرت عينة عشوائية من 5 أشخاص ممن يتبعون هذا البرنامج الغذائي وتم الحصول على البيانات الآتية :

عدد الأسابيع x	6	5	4	9	11
الوزن المخفض y	3	2	1	4	5

والمطلوب : قدر قيمة معامل الارتباط الخطي . واختبر معنوية الارتباط عند مستوى معنوية α=0.01 .

8- فيما يأتي أوزان وأطوال مجموعة من الذكور البالغين والمطلوب :

الطول x	159	180	175	150	170	171	165	176
الوزن y	68	88	79	65	70	73	63	74

أ- أوجد معادلة خط الانحدار المقدرة لانحدار الوزن على الطول ثم استخدمها في تقدير وزن شخص طوله 183سم.

ب- أوجد معادلة خط الانحدار المقدرة لانحدار الطول على الوزن .

9- قام باحث بالحصول على بيانات من 1000 مريض نفسي ، وذلك لمدة 5 سنوات، فإذا كان x يمثل الدرجة التي حصل عليها الشخص في بداية العلاج ، Y الدرجة التي حصل عليها بعد تلقي العلاج باستخدام البيانات الآتية:

$\sum x_i^2 = 14000$ ، $\sum x_i y_i = 3000$ ، $\sum y_i = 5000$ ،$\sum x_i = 3000$

والمطلوب:

أ- أوجد معادلة الانحدار المقدرة.

ب- أوجد قيمة yعندما x=4.

ت- إذا كانت قيمة s_y=10 فأوجد معامل الارتباط r.

10- يعطي الجدول التالي أعمار الزوج والزوجة بالسنوات لعينة من 6أزواج:

عمر الزوجة X	35	25	51	25	53	42
عمر الزوج Y	38	25	49	31	55	44

أ- أوجد معامل الارتباط؟

ب-اختبر معنوية معامل الارتباط من أجل α=0.04 .

11- في دراسة أجريت على إحدى أنواع الثدييات وجد أن حجم المخ يتغير مع وزن الجسم من فرد لآخر وأن العلاقة بين حجم المخ ووزن الجسم على الشكل:

وزن الجسم X	30	35	37	40	41	44	46	47	49	52	54
حجم المخY	360	379	380	390	409	408	412	419	425	435	439

أ- أوجد معامل الارتباط الخطي لبيرسون .

ب- استخدم البيانات الآتية لإيجاد معادلة الانحدار المقدرة.

ت- اختبر معنوية العلاقة عند مستوى معنوية α=0.05 .

12- لدراسة العلاقة بين الهيموجلوبين x (مقيساً mg/100ml) وعدد كرات الدم الحمراء y بالمليون لكل ملليمتر مكعب ، اختيرت عينة عشوائية من 12 ذكراً بالغاً من مجتمع ما وتم قياس تركيزات الهيموجلوبين وعدد كرات الدم الحمراء لكل مفردة والبيانات معطاة في الجدول الآتي :

x	15.2	16.4	14.2	13.0	14.5	16.1	15.2	14.8	15.7	14.9	15.6	14.7
y	5.1	5.4	4.5	4.2	4.3	6.1	5.2	4.3	4.7	4.8	4.6	4.8

والمطلوب :احسب معامل سبيرمان لارتباط الرتب.

13-عينة عشوائية من 200 رجل متزوج ، وتم تصنيفهم في الجدول الآتي تبعاً للتعليم وعدد الأطفال :

التعليم	عدد الأطفال		
	0-1	2-3	أكثر من 3
بسيط	14	37	32
متوسط	19	42	17
جامعي	12	17	10

والمطلوب : احسب معامل التوافق بين عدد الأطفال ومستوى التعليم . هل تعتقد هناك علاقة بين عدد الأطفال ومستوى التعليم؟

14- يعطى الجدول الآتي تصنيف لعينة عشوائية من 2764 شخصاً حسب الدخل بالدولار والفترة منذ آخر زيارة لاستشارة طبيب .

الدخل	منذ 6 شهور	من 7 شهور لسنة	أكثر من سنة
أقل من 3000	186	38	35
3000-4999	227	54	45
5000-6999	219	78	78
7000-9999	355	112	140
أكثر من 10.000	653	285	259

والمطلوب : هل تعتقد هناك علاقة ارتباط بين الدخل ومراجعة الطبيب؟

15- لدراسة العلاقة بين حساسية الجلد من ضوء الشمس ولون العين حصل طبيب متخصص في الأمراض الجلدية على البيانات التالية ، وذلك من عينة عشوائية من 100 شخص .

لون العين	تأثير الأشعة		
	قوي	متوسط	ضعيف
أزرق	19	27	4
رمادي أو أخضر	7	8	5
بني	1	13	16

والمطلوب : هل يمكن القول : إنَّ هناك علاقة بين لون العين وحساسية الجلد ؟