المحاضرة الثانية

الإحصاء الحيوي

د. وديع علي

البيانات الإحصائية ومؤشراتها الوصفية (تتمة)

مقاييس النزعة المركزية:

تعريف النزعة المركزية:

يقصد بالنزعة المركزية ميل البيانات إلى التمركز حول قيمة معينة في منتصف هذه البيانات.

قياس النزعة المركزية:

يمكن قياس النزعة المركزية للبيانات (أي التمركز حول القيمة الوسطى للبيانات) بعدة مقاييس:

1- المتوسط الحسابي Arithmetic Mean

2- الوسط الهندسي Geometric Mean

3- الوسط التوافقي Harmonic Mean

4- الوسيط Median

5- المنوال Mod

1 - المتوسط الحسابي

تعريف: يعرف المتوسط الحسابي لمجموعة من القيم بأنه عبارة عن مجموع هذه القيم على عددها.

طرق حساب المتوسط الحسابي:

أولاً: في حالة البيانات غير المبوبة

يرمز للمتوسط الحسابي لعينة $x_1, x_2, \dots, x_n$ بـ $\bar{x}$ ويعرف بالعلاقة:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

ثانياً:

في حالة البيانات المبوبة

إذا كانت البيانات معطاة وفق جداول تكرارية فئوية فإن مركز الفئة يمثل جميع القيم الواقعة في الفئة وبالتالي تكرار مركز الفئة يساوي تكرار الفئة، عندئذٍ:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} n_i x_i$$

حيث x_i مركز الفئة i $(i = 1,2,\ldots,k)$ و k عدد الفئات.

مثال (1):

تمثل البيانات التالية عدد المصابين بمرض أنفلونزا الخنازير في يوم واحد والمسجلة في 30 بلداً من العالم وذلك حسب ترتيب الحصول عليها:

6	16	9	8	16	10	19	9	18	27
28	26	7	8	18	33	40	28	37	41
10	11	18	8	23	40	33	3	20	30

والمطلوب حساب المتوسط الحسابي لعدد المصابين.

الحل:

1- من أجل البيانات المفردة نطبق العلاقة $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ فنحصل على:

$$\bar{x} = \frac{1}{30}(6 + 16 + 9 + \cdots + 20 + 30) = \frac{600}{30} = 20$$

2- نرتب البيان السابق في جدول يضم خمس فئات.

لدينا طول الفئة يساوي

$$L = \frac{41 - 3}{5} \cong 8$$

i	I_i	x_i	n_i	$n_i x_i$
1	[3-11[	7	10	70
2	[11-19[	15	6	90
3	[19-27[	23	4	92
4	[27-35[	31	6	186
5	[35-43[	39	4	156

وبالتالي:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i = \frac{1}{30}\sum_{i=1}^{5} n_i x_i = \frac{594}{30} = 19.8$$

معايير استخدام المتوسط الحسابي:

1- لحساب المتوسط الحسابي يجب أن تكون البيانات متجانسة

مثال:

بفرض أن: $X : 50,49,51$ عندئذ $\bar{x} = \frac{50+49+51}{3} = 50$

أما إذا كان: $X : 1,0,49$ فإن المتوسط لا يمكن أن يعبر عن البيانات لأن القيم مشتتة.

2- لاستخدام المتوسط الحسابي يجب أن تكون: أ) البيانات كمية ، ب) التوزيع طبيعي.

قاعدة: إذا كان $2\,SD < \bar{x}$ يكون التوزيع طبيعياً وإذا كان $2\,SD > \bar{x}$ يكون التوزيع غير طبيعي.

مزايا وعيوب المتوسط الحسابي

المزايا:

1- سهولة حسابه والتعامل معه جبرياً

2- دخول جميع القيم في حسابه (لا يمكن إهمال أي قيمة عند حسابه)

3- يعتبر المتوسط الحسابي الأساس في معظم عمليات الاستدلال الإحصائي.

العيوب:

1- يتأثر تأثراً كبيراً بالقيم المتطرفة أو الشاذة مما يفقده معناه وأهميته أحياناً.

2- لا يمكن حسابه من البيانات الوصفية

3- لا يمكن حسابه من الجداول التكرارية المفتوحة من أحد الطرفين أو من كليهما

4- لا يمكن إيجاده بيانياً أو بالرسم.

مثال (2):

إذا كانت النسب المئوية لاستجابة عينة مؤلفة من 6 أشخاص مصابين بمرض عضال لعقار جديد هي:

10% ,15%, 6%, 25%, 20%, 50% فإن متوسط استجابة هؤلاء المرضى هي:

$$\bar{x}=\frac{0.15+0.06+0.25+0.20+0.50+0.10}{6}=0.21$$

والنسبة المئوية لمتوسط استجابة المرضى 21% بعيدة عن النسب المئوية لاستجابة عناصر العينة.

2- الوسيط Med

أولاً: الوسيط لمعلومات غير مبوبة:

يرمز لوسيط العينة $x_1, x_2, \cdots, x_n$ المرتبة تصاعدياً بـ Med ويعرف بأنه القيمة التي تقع في وسط العينة، أي أن عدد نقاط العينة التي تسبقها يساوي عدد نقاط التي تليها. فإذا كان حجم العينة فردياً فإن الوسط هو القيمة التي ترتيبها $\frac{n+1}{2}$ وإذا كان حجم العينة زوجياً فإن الوسيط هو المتوسط الحسابي للقيمتين اللتين ترتيبهما $\frac{n}{2}$ و $\frac{n}{2}+1$.

ثانياً: الوسيط لمعلومات مبوبة :

إذا كانت $x_1', x_2', \ldots, x_m'$ النقاط المتباينة في العينة التي حجمها n $(m \le n)$ وكانت تكرارات هذه النقاط هي $n_1, n_2, \ldots, n_m$ حيث $n_1+n_2+\ldots+n_m=n$ وإذا رمزنا بـ K_i لمجموع تكرارات العناصر الأصغر من العنصر x_i' عندئذ يكون الوسط هو القيمة $Med = x_i$ التي يكون من أجلها K_i محققاً للمتراجحة:

$$\frac{n}{2} \le K_i \le \frac{n}{2}+n_i$$

وإذا كانت العينة مرتبة في جدول تكراري فان الوسط Me هو النقطة من محور الفئات التي إذا أقمنا منها عموداً قسم هذا العمود المدرج التكراري النسبي للعينة إلى قسمين متساويين في المساحة ولذلك يحسب الوسط من العلاقة الآتية:

$$Med = L + \frac{\frac{n}{2}-K_{i-1}}{n_i} \times I$$

حيث I طول الفئة الوسيطية و L الحد الأدنى للفئة الوسيطية و K_{i-1} التكرار التجميعي الصاعد للفئة التي تسبق الفئة الوسيطية و n_i تكرار الفئة الوسيطية.

ملاحظة:

تحدد الفئة الوسيطية التي تحوي الوسيط وذلك بحساب $\frac{n}{2}$ ومقارنتها بقيم K_i وعندئذ تكون الفئة الوسيطية هي الفئة التي تقابل أول قيمة لـ K_i تحقق العلاقة $K_i \ge \frac{n}{2}$.

مثال (3):

لنأخذ سلسلة درجات124 طالباً مبوبة مع تكراراتها المطلقة والتجميعية كما في الجدول التالي

الرقم المتسلسل i	1	2	3	4	5	6	المجموع
مجالات التبويب $[x_i , x_{i+1}[$	50-55	55-60	60-65	65-70	70-75	75-80	×
التكرارات المطلقة n_i	20	40	30	20	10	4	124
التكرارات التجميعية k_i	20	60	90	110	120	124	×

نلاحظ أن عدد القيم n = 124 وأن نصفها هو $\frac{n}{2} = 62$ أي أن رقم (ترتيب) الوسيط هو حوالي 62 . لذلك نتابع الأرقام في سطر التكرارات التجميعية k_i حتى نصل ونتجاوز العدد 62 فنجد أنه $k_3 = 90$, أي أن الوسيط يقع في الفئة الثالثة وإن قيمته Med محصورة بين طرفيها , أي أن $60 \leq Med \leq 65$ وتسمى هذه الفئة بالفئة الوسيطية أو المجال الوسيطي ويكون:

$$Med = L + \frac{\frac{n}{2} - K_{i-1}}{n_i} \times I = 60 + \frac{62-60}{30} \times 10 = 60.33$$

وهي قيمة واقعة في المجال الوسيطي وقريبة جداً من طرفه الأيسر.

معايير استخدام الوسيط:

لاستخدام الوسيط يجب أن يكون:

أ) البيانات كمية

ب) البيانات مرتبة

ت) التوزيع غير طبيعي.

مثال: بفرض أن $X : 1,0,49$ عندئذ لحساب الوسيط نرتب البيانات أولاً: $X : 0,1,49$ ويكون الوسيط يساوي 1 أي أن $Me = 1 \ (0 \ to \ 49)$ نلاحظ أن المتوسط الحسابي $\bar{x} = 50$ لا يعبر عن البيانات لأن تشتتها كبير.

ملاحظة: الوسيط لا يتأثر بالقيم المتطرفة فعلى سبيل المثال إذا كان $X : 0,0,0,0,10000$ فإن:

$$Me = 0 \ (0 \ to 10000)$$

ولا يمكننا وصف البيانات من خلال المتوسط الحسابي في هذه الحالة مع أن $\bar{x} = \frac{10000}{5} = 2000$.

3- المنوال Mod:

تعريف المنوال : هو القيمة من قيم X التي تقابل التكرار الأكبر , أي هي الأكثر تكراراً وتسمى (الموضة) , وهو لا يعرف إلا على المعلومات المرتبة أو المبوبة .

طرق حساب المنوال:

أولاً: في حالة البيانات غير المبوبة:

لحساب المنوال نستخدم التعريف مباشرة أي أن قيمة المنوال هي القيمة الأكثر تكراراً.

مثال (4) :

لنأخذ المثال التالي عن درجات الطلاب في العملي (لـ 50 طالباً) :

الرقم المتسلسل i	1	2	3	4	5	6	المجموع
قيم x المرتبة : x_i	2	4	8	16	32	40	×
التكرارات المطلقة	3	5	10	15	10	7	50

نلاحظ أن القيمة التي تقابل التكرار الأكبر (15) هي قيمة X الرابعة أي أن المنوال هو:

$$Mod = x_4 = 16$$

أي أن الدرجة 16 هي الأكثر تكراراً بين درجات الطلاب في العملي .

ثانياً:المنوال لمعلومات مبوبة :

لحساب المنوال في حالة البيانات المبوبة نتبع الخطوات الآتية:

1- تحديد فئة المنوال (وهي الفئة المناظرة لأكبر تكرار في الجدول التكراري)

2- حساب المنوال من المعادلة:

$$Mod = L + \frac{d_1}{d_1 + d_2} \times I$$

حيث:

L بداية فئة المنوال

d_1 _ أكبر تكرار - التكرار السابق له

d_2 _ أكبر تكرار - التكرار اللاحق له

I _ طول الفئة المنوالية.

مثال (5):

لنأخذ المعلومات التالية عن درجات 220 طالباً المبوبة كما في الجدول التالي :

i	1	2	3	4	5	6	7	المجموع
$[x_i , x_{i+1}[$	[20 – 30[	[30 – 40[	[40 – 50[	[50 – 60[	[60 – 70[	[70 – 80[	[80 – 90[	×
n_i	32	40	45	60	25	15	3	220

نلاحظ أن التكرار الأكبر هو (60) وهو يقابل الفئة الرابعة المحددة بـ **[50 – 60[** وتسمى هذه الفئة بالفئة المنوالية .

أي أن المنوال Mod يقع في هذا المجال وأن قيمته تحقق العلاقة: $60 \leq Mod \leq 50$ ولحساب قيمة المنوال نلاحظ أن $I = 60 - 50 = 10$ وإن التكرار الذي يسبقه يساوي 54 و التكرار الذي يليه يساوي 25 وبذلك نجد أنه يمكننا حساب قيمة المنوال Mod من العلاقة:

واعتماداً على معطيات الجدول السابق يمكننا حساب قيمة المنوال كما يلي :

$$Mod = 50 + \frac{60-45}{(60-45)+(60-25)} \times 10 = 50 + 10\,\frac{15}{50}$$

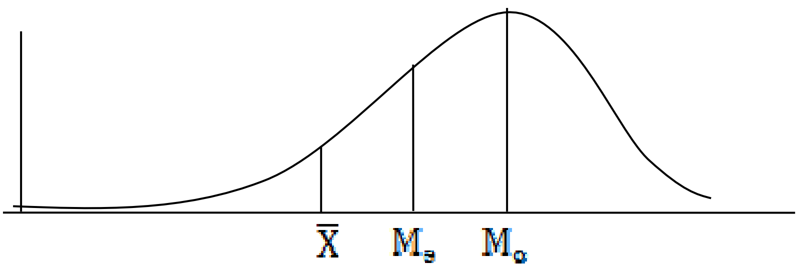
$$Mod = 50 + 3 = 53$$

تطبيقات مشتركة للمتوسط الحسابي والوسيط والمنوال :

إن قيم المتوسط الحسابي والوسيط والمنوال قد تكون مختلفة أو متقاربة أو متساوية . وهنا نميز الحالات الأساسية التالية :

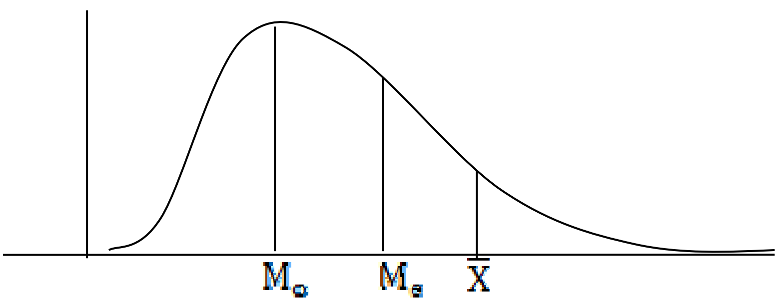
1- إذا كانت العلاقة بينهم كما يلي : $\bar{x} < Med < Mod$

فإن التوزيع التكراري يكون ملتوياً نحو اليسار (أي أنه يمتد نحو اليسار) كما يلي :



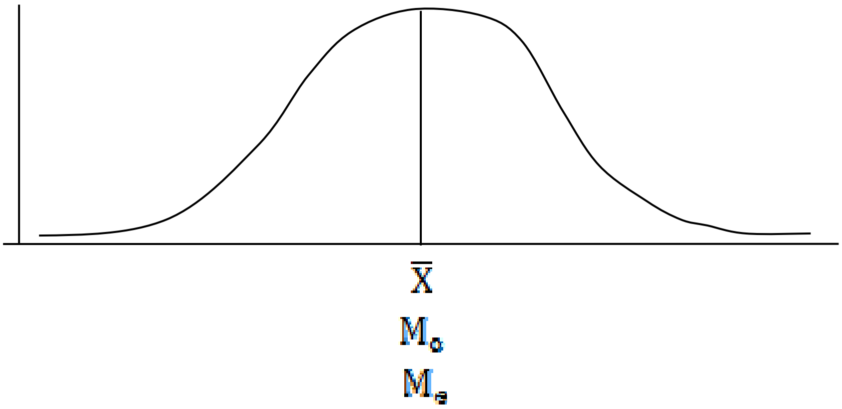
2- إذا كانت العلاقة بينهم كما يلي : $Mod < Med < \bar{x}$

فإن التوزيع التكراري يكون ملتوياً نحو اليمين (أي أنه يمتد نحو اليمين) كما يلي :



3- إذا كانت العلاقة بينهم كما يلي : $\bar{x} = Med = Mod$

فإن التوزيع التكراري يكون متناظراً تماماً كما يلي :



ملاحظة:

إذا كانت العينة وحيدة المنوال فإن العلاقة التقريبية بين متوسطها الحسابي وأوسطها ومنوالها هي:

$$Mod = 3Med - 2\bar{x}$$

مزايا وعيوب المنوال:

المزايا:

1- لا يتأثر في حسابه بالقيم المتطرفة أو الشاذة

2- سهولة حسابه أو إيجاده

3- يمكن حسابه من الجداول التكرارية المفتوحة من أحد الطرفين أو من كليهما

4- يمكن إيجاده بالرسم

5- يمكن حسابه من البيانات الوصفية

العيوب:

1- عدم دخول جميع القيم في حسابه

2- أقل مقاييس النزعة المركزية استخداماً

3- يعاب على المنوال عدم وجوده في بعض البيانات كما يمكن وجود أكثر من منوال في بعض البيانات

4- عديم الفائدة في البيانات قليلة العدد .

مقاييس التشتت:

من مقاييس التشتت (أو التبعثر) سندرس المدى، الانحراف المتوسط، التشتت (أو التباين) والانحراف المعياري. هذه المقاييس تعطي درجة تباعد البيان الإحصائي عن المتوسط الحسابي. فعلى سبيل المثال لدينا ثلاث مجموعات من القياسات على النحو التالي:

a)	4	5	6	7	8
b)	0	3	6	9	12
c)	-4	1	6	11	16

نلاحظ أن المتوسط الحسابي لكل مجموعة هو 6 كما أن الوسط هو 6 لكل من المجموعات الثلاث رغم الفرق بين قيم المجموعات الثلاث.

1- المدى:

إذا كانت $x_1, x_2, ..., x_n$ عينة حجمها n فيسمى الفرق بين أكبر قيمة max وأصغر قيمة min في العينة بالمدى ويرمز له بـ R ويكون:

$$R = \max(x) - \min(x)$$

أما في الجدول التكراري فيحسب المدى على أساس الفرق بين قيمة الحد الأعلى للفئة الأخيرة وقيمة الحد الأدنى للفئة الأولى. إذاً المدى هو أبسط مقاييس التشتت وأسهلها. وبالتالي إذا كان للعينتين A و B نفس المتوسط الحسابي فنقول أن العينة A أكثر تجانساً مع متوسطها $\overline{x_A}$ من العينة B مع متوسطها $\overline{x_B}$ إذا كان مدى العينة A أصغر من مدى العينة

2- التباين (أو التشتت)

وهو المتوسط الحسابي لمربعات انحرافات نقاط العينة عن متوسطها الحسابي. ويرمز لتشتت العينة $x_1, x_2, ..., x_n$ بـ s^2 ويعطى:

آ- في حالة البيانات غير المبوبة:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (I)$$

ب- في حالة البيانات المبوبة في جدول تكراري يضم k فئة بالعلاقة:

$$s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \quad (II)$$

حيث x_i يمثل مركز الفئة i و n_i تكرارها.

يمكن حساب التشتت من العلاقتين المكافئتين للعلاقتين السابقتين التاليتين:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \quad (I')$$

بالنسبة للبيانات غير المبوبة

$$s^2 = \frac{1}{n}\sum_{i=1}^{k} n_i x_i^2 - \bar{x}^2 \quad (II')$$

بالنسبة للبيانات المبوبة.

يمكن بسهولة البرهان أن تباين القيم المتساوية يساوي الصفر وأن تباين المعطيات عن متوسطها الحسابي أصغر من تباين هذه المعطيات عن أي قيمة أخرى مختلفة عنه.

3- الانحراف المعياري:

تعريف:

الانحراف المعياري لعينة هو الجذر التربيعي الموجب لتشتتها ويرمز له بالرمز s ويقيس كيفية توزيع المشاهدات حول المتوسط.

ملاحظة: يرمز للتباين المعدل للعينة $x_1, x_2, \ldots, x_n$ بـ $\hat{s}^2$ ويعطى بالعلاقة:

$$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

بالمقارنة مع تباين العينة نجد:

$$\hat{s}^2 = \frac{n}{n-1}s^2$$

بما أن لكل عينة قيمتها العددية المميزة لها فهذا يعني أن للمجتمع الإحصائي الذي استخلصت منه العينة قيم عددية مميزة خاصة به مقابلة للقيم المميزة للعينة والتي يمكن تعيينها إذا كانت غير معلومة، وتمكننا من إيجاد العلاقة التي تربطها مع مثيلاتها من العينة ويتم ذلك بالاعتماد على نظرية الاحتمالات.

مثال (6): لنأخذ المعلومات المبوبة في الجدول التالي عن درجات الطلاب في مقرر الإحصاء لـ 60 طالباً.

j	1	2	3	4	5	6	7
i	مجالات الدرجات [[	التكرارات n_i	مراكز المجالات x_i'	$n_i.x_i'$	$x_i'-\bar{x}'$	$(x_i'-\bar{x}')^2$	$n_i.(x_i'-\bar{x}')^2$
1	30-40	8	35	280	-22.6	510,76	4086.08
2	40-50	11	45	495	-12.6	158.76	1746.36
3	50-60	15	55	825	-2.6	6.76	101.4
4	60-70	13	65	845	7.4	54.76	711.88
5	70-80	9	75	675	17.4	302.76	2724.84
6	80-90	4	85	340	27.4	750.76	3003.04
	المجموع	60		3460			12373.6

ونظراً لعدم وجود قيم محددة ضمن مجال من مجالات الدرجات لاستخدامها في حساب المتوسط. نقوم بإيجاد وحساب مراكز هذه المجالات ونعتبر كل مركز ممثلاً لمجاله. ولحساب المتوسط الحسابي نقوم بضرب مراكز المجالات x_i' بالتكرارات المقابلة n_i فنحصل على عناصر العمود الرابع $n_i.x_i'$. وبذلك نجد أن المتوسط الحسابي لدرجات الطلاب يساوي:

$$\bar{x}' = \frac{\sum_{i=1}^{n} n_i x_i'}{n} = \frac{3460}{60} = 57.6$$

ولحساب التباين نحسب الفر وقات $x_i'-\bar{x}'$ ثم نأخذ مربعاتها $(x_i'-\bar{x}')^2$ فنحصل على عناصر العمود السادس، ثم نضرب المربعات بالتكرارات المقابلة n_i ، فنحصل على عناصر العمود السابع، ومنه نجد أن التباين يساوي:

$$s^2 = \frac{\sum_{i=1}^{n} n_i (x_i'-\bar{x}')^2}{n} = \frac{12373.6}{60} = 206.22$$

وبذلك نجد أن الانحراف المعياري لدرجات الطلاب يساوي:

$$s = \sqrt{s^2} = \sqrt{206.22} = 14.36$$

4- معامل الاختلاف (Coefficient of Variation):

هو أحد مقاييس التشتت ويستخدم للمقارنة بين تشتت ظاهرتين مختلفتين أو أكثر , الظاهرة (المتغير) التي معامل اختلافها أقل تكون متجانسة لأنها أقل تشتتاً ويرمز له بالرمز $C.V(x)$ ويعطى بالعلاقة:

$$C.V(x)=\left(\frac{S}{\overline{x}}\right)\cdot 100\%$$

حيث S الانحراف المعياري و $\overline{x}$ المتوسط الحسابي .

مثال (7): قام أحد مراقبي السيطرة النوعية على الإنتاج بأخذ عينة عشوائية من الإنتاج خلال يومين كما هو مبين في الجدول الآتي:

الانحراف المعياري	المتوسط الحسابي	n	اليوم
100	1000	200	الأول
150	96	200	الثاني

المطلوب: في أي يوم أعطت الآلة إنتاجا أقل تشتتاً أو أكثر تجانساً؟

الحل:

$$C.V(1)=\frac{S}{\overline{x}}\cdot 100\%=\frac{100}{1000}\cdot 100\%=10\%$$
$$C.V(2)=\frac{150}{960}\cdot 100\%=15.62\%$$

أي أن الآلة في اليوم الأول معامل اختلافها أقل من اليوم الثاني أي أنها تكون أكثر تجانساً لأنها أقل تشتتاً .

تمرين1:

أجريت دراسة على 200 مريض بسرطان الكبد مع التهاب الكبد B ، وتم تقسيم المرضى إلى مجموعتين المجموعة 1 تم إعطاؤها علاج واحد والمجموعة 2 تم إعطاؤها أكثر من علاج في نفس الوقت وتم عرض البيانات في الجدول الآتي:

المتغير	المجموعة 1	المجموعة 2
العمر $(\bar{x} \pm S)$	34 ± 10	33 ± 11
النوع		
ذكر	60	64
أنثى	40	36
درجة الورم		
1	30	20
2	34	35
3	36	45
الإصابة بالتهاب الكبد B		
نعم	19	17
لا	18	83
الإقامة		
مدينة	64	40
ريف	36	60

عدد نوع المتغيرات الآتية:

1- العمر:

a) مستمر

b) رتبي

c) أسمي

2- النوع:

a) رتبي

b) أسمي

c) مستمر

3- درجة الورم:

a) رتبي

b) مستمر

c) أسمي

4- الإصابة بالتهاب الكبد B :

a) مستمر

b) أسمي

c) رتبي

5- الإقامة:

a) أسمي

b) رتبي

c) مستمر

تمرين 2:

ضع صح أو خطأ أمام كل من العبارات الآتية وصحح الخاطئة منها:

1) $\bar{x} \pm S$ لمستويات $IL-2$ (إنترلوكين) في المصل للحالات كان $164.7 \pm 183.8\ ng/ml$

الجواب: خطأ التوزيع غير طبيعي لا يمكن أن نعبر عن البيانات باستخدام $\bar{x} \pm S$ ويجب استخدام الوسيط والمدى

2) $\bar{x} \pm S$ لمستويات $IL-2$ في المصل كان $164.7 \pm 183.8\ ng/ml$ والوسيط كان 86.6 والمدى $(-881\ to\ 55)$

الجواب: خطأ لأن البيانات تم التعبير عنها باستخدام $\bar{x} \pm S$ و الوسيط والمدى

3) عدد الذكور المتبرعين بالدم في هذه الدراسة كان 235. تراوحت أعمارهم بين 20 و 48 عام وكان المتوسط 29.1 عام

الجواب: خطأ لأن البيانات تم التعبير عنها باستخدام المتوسط والمدى

4) كان متوسط عمر المرضى 28.8 عام والمدى $(11\ to\ 57)$ عام وكانت الإصابة بالذئبة الحمامية (SLE) عند الإناث أكبر منها عند الذكور.

الجواب: خطأ لأن البيانات تم التعبير عنها باستخدام المتوسط والمدى.