Measures of dispersion

In statistics, the measures of dispersion help to interpret the variability of data i.e. to know how much homogenous or heterogeneous the data is. In simple terms, it shows how squeezed or scattered the variable is.

في الإحصاء ، تساعد مقاييس التشتت في تفسير تباين البيانات ، أي معرفة مقدار البيانات المتجانسة أو غير المتجانسة. بعبارات بسيطة ، يوضح مدى ضغط أو تشتيت المتغير.

Types of measures of dispersion:

1. Range. (المدى)

2. Quartile deviation. (الانحراف الربيعي)

3. Average of the absolute deviations. (متوسط الانحرافات المطلقة)

4. Variance . (التباين)

5. Standard deviation. (الانحراف المعياري)

In addition to **the coefficient of variation** that is used to compare two or more sets of data in terms of Dispersion.

بالإضافة إلى **معامل الاختلاف** الذي يستخدم لمقارنة مجموعتين أو أكثر من البيانات من حيث التشتت.

1. Range. (المدى)

The range is the difference between the largest and smallest value in the data, it is the period that

In whichType equation here. the variable under study changes and is denoted by the symbol (R).

A. Non- classified data:

ويحسب المدى في حالة البيانات غير المبوبة كما يلي :

R = largest value - smallest value

R = اصغر قيمة – اكبر قيمة

B. Classified data:

أما في حالة البيانات المبوبة نعتبر أكبر قيمة هي الحد الأعلى للفئة الأخيرة وأصغر قيمة هي الحد الأدنى للفئة الأولى مع مراعاة أن تكون فئات الجدول مرتبة ترتيباً تصاعدياً ، أي يحسب المدى في حالة البيانات المبوبة كما يلي :

R = The upper limit for the last class - The minimum limit for the first class

R = الحد الادنى للفئة الاولى – الحد الاعلى للفئة الاخيرة

إذا كان المدى صغيراً فيعنى ذلك أن البيانات منتشرة في فترة قصيرة أي قريبة من بعضها وتشتتها صغير ، أما إذا كان المدى كبيراً فيعني ذلك أن البيانات منتشرة في فترة طويلة أي متباعدة عن بعضها وتشتتها كبير.

Ex1: Using the term compare the dispersion of the three groups.

G1: 6, 6, 6, 6, 6.

G2: 4, 5, 6, 7, 8.

G3: 0, 1, 6, 10, 13.

Sol:

R(G1) = 6 – 6 = 0

R(G2) = 8 – 4 = 4

R(G3) = 13 – 0 = 13

بما أن مدى المجموعة الأولى يساوي صفراً ، فيعني ذلك أنه لا يوجد اختلاف أو تشتت بين قيم هذه المجموعة ، أي أن كل القيم داخل هذه المجموعة متساوية.

ونلاحظ أن مدى المجموعة الثانية أصغر من مدى المجموعة الثالثة، وهذا يعني أن تشتت القيم داخل المجموعة الثانية أقل من تشتت القيم داخل المجموعة الثالثة.

H.W1: Compare the dispersion of the statistics subject's scores for three groups of students:

G1: 50, 62, 30, 70, 45, 80.

G2: 0, 20, 70, 55, 65, 82, 35, 77.

G3: 49, 54, 85, 50, 60, 63, 75, 65.

Ex2: Calculate the range for the following data showing weights in kilograms of 100 students:

Weight (kg)	Number of students
60 – 63	5
63 – 66	15
66 – 69	40
69 – 72	28
72 – 75	12

Sol:

R = The upper limit for the last class - The minimum limit for the first class

= 75 – 60 = 15.

Range properties:

1. Range is a measure that is easy to calculate and simple in concept and significance.
2. It is concerned with only two values in the data and neglects the rest.
3. It depends in its calculation on the smallest and largest value only, and therefore it is considered a misleading measure, because it is when the largest or minor values, or both, are outliers, then the range is large while the group values are not divergent, for example if we have the following data:

   55, 14, 11, 9, 8, 12, 10.

   R = 55 – 8 = 47

فالمدى كبير ، ويشير إلى وجود تشتت كبير في المجموعة في حين أن القيم متقاربة ، ولذلك فهو مقياس مضلل ، وسبب ذلك هو اعتماده على القيم المتطرفة ، فلو حذفنا القيمة المتطرفة وهي القيمة 55 فنجد أن قيمة المدى تساوي :

R = 14 – 8 = 6

وهي قيمة صغيرة وواقعية .

2. Variance(التباين)

The variance is defined as the arithmetic mean of the squares of the deviations of values from their arithmetic mean (S^2).

A. Non- classified data:

- نحسب الوسط الحسابي للبيانات.
- نحسب انحراف كل قيمة عن الوسط الحسابي ،

  حيث : انحراف القيمة عن الوسط الحسابي = القيمة - الوسط الحسابي
- نوجد مربع انحراف كل قيمة عن الوسط الحسابي .
- نوجد مجموع مربعات انحرفات القيم عن الوسط الحسابي .
- نحسب قيمة التباين باستخدام القانون التالي :

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}$$

Ex3: Calculate the variance for data representing 8 students' grades:

7, 6, 10, 9, 8, 5, 5, 6.

Sol:

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} : \frac{56}{8} = 7$$

Value (X_i)	($\overline{X}$ - X_i)	($\overline{X}$ - X_i)²
7	0	0
6	-1	1
10	3	9
9	2	4
8	1	1
5	-2	4
5	-2	4
6	-1	1
	Σ	24

$$S^2 = \frac{24}{7} = 3.43$$

B. Classified data:

في حالة البيانات المعروضة في جداول تكرارية نتبع الخطوات التالية :

- نحسب مراكز الفئات.
- نحسب الوسط الحسابي للتوزيع .
- نحسب انحراف كل مركز عن الوسط الحسابي .
- نوجد مربع انحراف كل مركز عن الوسط الحسابي .
- يضرب مربع انحراف كل مركز عن الوسط الحسابي لفئة في تكرار هذه الفئة .
- نجمع حواصل الضرب التي تحصلنا عليها في الخطوة السابقة .
- نحسب قيمة التباين باستخدام القانون التالي :

$$S^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \overline{x})^2}{\sum_{i=1}^{k} f_i - 1}$$

حيث : x_i : مركز الفئة ، f_i : تكرار الفئة ، $\overline{x}$: الوسط الحسابي

Ex4: Calculate the variance for the following data, which represent the distribution of 200 workers by overtime spend on the job at the factory every month:

Overtime (h)	Number of Workers
0 – 10	20
10 – 20	80
20 – 30	50
30 – 40	40
40 – 50	10
Σ	200

Sol:

$$S^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}$$

class	f_i	X_i	$X_i f_i$	$(X_i - \bar{x})$	$(X_i - \bar{x})^2$	$f_i(X_i - \bar{x})^2$
0 – 10	20	5	100	-17	289	5780
10 – 20	80	15	1200	-7	49	3920
20 – 30	50	25	1250	3	9	450
30 – 40	40	35	1400	13	169	6760
40 – 50	10	45	450	23	529	5290
Σ	200		4400			22200

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{4400}{200} = 22$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}$$

$$= \frac{22200}{200-1} = \frac{22200}{199} = 111.56\ h^2$$

ونلاحظ أن التباين وحداته هي مربع وحدات القياس الأصلية، وكثيراً ما تكون غير ذات معنى فمثلاً في هذا المثال التباين = $111.56\ h^2$ فهنا ساعة تربيع ليس لها أي معنى، وهذا هو عيب التباين ؛ ولذلك نُرجع الوحدات إلى أصلها بأخذ الجذر التربيعي للتباين ، ويسمى المقياس الجديد بالانحراف المعياري .

3. Standard deviation (الانحراف المعياري)

Standard deviation is the measure of dispersion of a set of data from its mean. It measures the absolute variability of a distribution; the higher the dispersion or variability, the greater is the standard deviation and greater will be the magnitude of the deviation of the value from their mean.

It is the positive square root of the mean of the squares of deviations of values from their mean arithmetic, that is, it is the positive square root of the variance, and it is denoted by a symbol (S).

هو الجذر التربيعي الموجب للوسط الحسابي لمربعات انحرافات القيم عن وسطها الحسابي ، أي: هو الجذر التربيعي الموجب للتباين.

A. Non- classified data:

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

B. Classified data:

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \bar{x})^2}{\sum_{i=1}^{k} f_i - 1}}$$

For Ex :

$S = \sqrt{3.43} = 1.85$

For Ex :

$S = \sqrt{111.56} = 10.56$

إذا لم يكن الوسط الحسابي عدداً صحيحاً فإن حساب التباين ومن ثم الانحراف المعياري باستخدام الصيغ السابقة الذكر ، يصبح أمراً غير سهلٍ ، ولذلك اشتقت من الصيغة الأساسية للتباين والتي تعتمد على الانحرافات عن الوسط الحسابي ، صيغة أخرى تعتمد على القيم مباشرة ، وذلك لتسهيل العمليات الحسابية ، وبالطبع الصيغتان تعطيان نفس النتيجة تماماً ، وصيغ التباين التي تعتمد على القيم مباشرة هي :

A. Non- classified data:

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

B. classified data:

$$S^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1} \left[\sum_{i=1}^{k} x_i^2 f_i - \frac{\left(\sum_{i=1}^{k} x_i f_i \right)^2}{\sum_{i=1}^{k} f_i} \right]$$

حيث : X_i : مركز الفئة ، f_i : تكرار الفئة

يجب الانتباه بأن هناك فرقاً بين المقدار $\sum_{i=1}^{k} x_i^2 f_i$ والمقدار $\left(\sum_{i=1}^{k} x_i f_i \right)^2$.

Ex5: Calculate the variance and standard deviation of the data mentioned in an example 3, so using the second variance formula (direct values formula).

Sol:

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

X_i	7	6	10	9	8	5	5	6	$\Sigma X_i = 56$
X_i^2	49	36	100	81	64	25	25	36	$\Sigma X_i^2 = 416$

$$S^2 = \frac{1}{8-1}\left[416 - \frac{(56)^2}{8}\right]$$

$$= \frac{1}{7}\left[416 - 392\right] = \frac{24}{7} = 3.43$$

$$S = \sqrt{3.43} = 1.85$$

نلاحظ انها نفس النتيجة في المثال السابق.

Ex6: Calculate the variance and standard deviation of the data mentioned in an example 4, so using the second variance formula (direct values formula).

Sol:

$$S^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1}\left[\sum_{i=1}^{k} x_i^2 f_i - \frac{\left(\sum_{i=1}^{k} x_i f_i\right)^2}{\sum_{i=1}^{k} f_i}\right]$$

class	f_i	X_i	$X_i f_i$	X_i^2	$f_i X_i^2$
0 – 10	20	5	100	25	500
10 – 20	80	15	1200	225	18000
20 – 30	50	25	1250	625	31250
30 – 40	40	35	1400	1225	49000
40 – 50	10	45	450	2025	20250
Σ	200		4400		119000

$$S^2 = \frac{1}{200-1}\left[119000 - \frac{(4400)^2}{200}\right]$$

$$= \frac{1}{199}\left[119000 - 96800\right] = 111.56$$

$$S = \sqrt{111.56} = 10.56$$

نلاحظ انها نفس النتيجة التي حصلنا عليها في المثال الرابع.

Properties of Standard Deviation:

- Standard deviation is the most important and widely used measure of dispersion.
- Enter into his calculation all the observed values without neglecting any value.
- It is characterized by its ability to treat algebra.
- It cannot be calculated for open frequency distributions.

Coefficient of variation (معامل الاختلاف):

كل مقاييس التشتت السابقة تعتمد على وحدات القياس، وبالتالي لا يمكن استعمالها لمقارنة تشتت توزيعين مختلفين في وحدات القياس كمقارنة تشتت الأطوال بتشتت الأوزان مثلاً ، ولذلك يجب التعامل مع مقياس نسبي لا يعتمد على الوحدات المستعملة ويسمى هذا المقياس معامل الاختلاف ويحسب كما يلي :

$$\text{Coefficient of variation} = \frac{s}{\bar{x}} * 100$$

$$\text{معامل الاختلاف} = \frac{\text{الانحراف المعياري}}{\text{الوسط الحسابي}} \times 100$$

وبالطبع كلما زاد التشتت أخذ معامل الاختلاف نسبة أكبر .

Ex7: Compare the dispersion of the students' lengths and their weights, using the following data, which represent the lengths and weights 100 students.

Lengths (cm)	Number of students
110 – 120	12
120 – 130	15
130 – 140	25
140 – 150	30
150 – 160	10
160 - 170	8

Weights (Kg)	Number of students
60 – 63	6
63 – 66	16
66 – 69	40
69 – 72	28
72 – 75	10

Sol:

نحسب معامل الاختلاف لكل من التوزيعين ثم نقوم بالمقارنة ، ولحساب معامل الاختلاف لكل توزيع يلزمنا حساب الوسط الحسابي والانحراف المعياري لكل منهما ، وذلك كما يلي :

A. Distribution of lengths:

نوضح العمليات الحسابية التي تلزمنا لحساب الوسط الحسابي والانحراف المعياري في الجدول التالي:

class	f_i	X_i	$X_i f_i$	$(X_i - \bar{x})$	$f_i(X_i - \bar{x})^2$
110 – 120	12	115	1380	-23.5	6627
120 – 130	15	125	1875	-13.5	2733.75
130 – 140	25	135	3375	-3.5	306.25
140 – 150	30	145	4350	6.5	1267.5
150 – 160	10	155	1550	16.5	2722.5
160 - 170	8	165	1320	26.5	5618
Σ	100		13850		19275

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = \frac{13850}{100} = 138.50$$

$$S^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1}\left[\sum_{i=1}^{k} f_i\,(x_i - \bar{x})^2\right]$$

$$= \frac{1}{100\text{-}1}\ 19275 = 194.7$$

$$S = \sqrt{194.7} = 13.95$$

Coefficient of variation for lengths $= \frac{13.95}{138.5} * 100 = 10.07\%$

B. Distribution of weights:

الجدول التالي يوضح لنا العمليات الحسابية اللازمة للحصول على الوسط الحسابي والانحراف المعياري للاوزان:

class	f_i	X_i	$X_i f_i$	$(X_i - \bar{x})$	$f_i(X_i - \bar{x})^2$
60 – 63	6	61.5	369	-6.6	261.36
63 – 66	16	64.5	1032	-3.6	207.36
66 – 69	40	67.5	2700	-0.6	14.4
69 – 72	28	70.5	1974	2.4	161.28
72 – 75	10	73.5	735	5.4	291.6
Σ	100		6810		936

$$\overline{X} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i} = 68.1$$

$$S^2 = \frac{1}{\sum_{i=1}^{k} f_i - 1}\left[\sum_{i=1}^{k} f_i \, (x_i - \bar{x})^2 \right]$$

$$= \frac{936}{99} = 9.45$$

$$S = \sqrt{9.45} = 3.07$$

Coefficient of variation for weights $= \frac{3.07}{68.1} * 100 = 4.51\%$

بما أن معامل الاختلاف للأطوال أكبر من معامل الاختلاف للأوزان ، إذن تشتت الأطوال أكبر من تشتت الاوزان.

H.W 2: If we had two sets of data:

set 1: 21, 15, 23, 16, 14, 12, 25.

set 2: 30, 42, 64, 60, 28, 70, 82, 56.

A- For each of these two groups, calculate the following:

Range, Variance, Standard deviation.

B- Compare the dispersion of the two groups using the coefficient of variation.

H.W 3: If you know the coefficient of variation = 25% and that the Variance = 25, that for the following data:

10, 18, 25, 15, X. Calculate the missing value X.