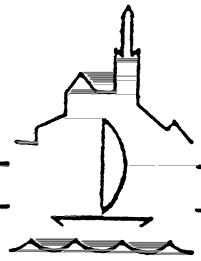
Documents pour la classe

ECART-TYPE

par

Robert Bayard
Alain Brunel
Aimé Gandolfi
Albert Rolland



N° 3

Publications de l'I.R.E.M. d'Aix-Marseille

1990

Ecart - type

Robert BAYARD
Alain BRUNEL
Aimé GANDOLFI
Albert ROLLAND

Documents pour la classe
Publications de l'IREM d'AIX-MARSEILLE

Les 2 paramètres les plus importants en statistique et en calcul des probabilités sont la **moyenne** et l'**écart-type**.

La moyenne est un paramètre de position.
L'écart-type est un paramètre de dispersion.

L'écart-type est une moyenne quadratique.

OBJECTIF

Présentation de l'écart-type

PLAN

1 - Comment introduire la notion d'écart-type

1.1 - Simulation d'un jeu de "Pile" ou "Face" avec l'ordinateur

La méthode, la plus courante, qui consiste à donner de l'écart-type d'abord la définition puis des procédés de calcul, n'est pas motivante.

Avant même de donner la définition de l'écart-type nous pouvons suggérer l'importance de ce paramètre à l'aide de l'expérience suivante.

Faisons jouer un ordinateur à "Pile" ou "Face".
Simulons, par exemple, 100 jets.
La probabilité d'obtenir "Pile" étant de 0,5, il n'y a aucune difficulté à faire admettre que nous devons nous attendre à ce que le nombre de "Pile" obtenu approche 50.

Cependant la probabilité d'obtenir exactement 50 est faible
[0,08 (en utilisant une table de Laplace - Gauss)]
Nous faisons constater que le résultat obtenu appartient à l'intervalle [40 ; 60] (le contraire est très rare : la probabilité est inférieure à 0,05).
Nous devons insister et dire que nous sommes presque *"presque certains"* que le nombre de "Pile" obtenu appartient à l'intervalle [40 ; 60].
Les probabilistes donnent à ce type d'intervalle le nom d'intervalle de confiance.

Nous en déduisons que le hasard est en partie maîtrisable.

Cf: Dans le domaine professionnel, un assureur qui, pour une catégorie de sinistres, connaît la probabilité qu'un assuré soit accidenté au cours d'une année peut prévoir le nombre maximal (et aussi le nombre minimal) d'accidentés dans l'année parmi ses assurés.

Reprenons le nombre 50 qui devait être approché par le nombre de "Pile" obtenu au cours des 100 jets.

Faisons remarquer que 40, la première borne de l'intervalle, est égale à 50 diminuée de 10 (ou 5 x 2) et que 60, la deuxième borne, est égale à 50 augmentée de 10 (ou 5 x 2).

Mais pourquoi ce nombre 5 ?

Eh bien ! Parcequ'il correspond à un paramètre qui est d'une très grande importance : ce paramètre porte le nom d'écart-type.
Mais alors ce paramètre si important :
- qui est-il ?
- comment l'a-t-on créé ?
- comment le calcule-t-on ?

1.2 - Etude de la dispersion des valeurs d'une série par rapport à la moyenne

Le directeur d'un établissement "sport études" examine les notes trimestrielles de mathématiques obtenues par 2 groupes d'élèves qui suivent dans un lycée les disciplines d'enseignement général.

1.2.1 - Chacun des 2 groupes comprend 2 élèves

Groupe A : 04 ; 18

Groupe B : 10 ; 12

Les 2 groupes ont la même moyenne trimestrielle : 11
Les 2 groupes sont néanmoins dissemblables.
Dans le groupe A un élève est faible, l'autre est excellent.
Dans le groupe B les 2 élèves ont chacun une note proche de 11.
Les notes du groupe A sont plus dispersées que celles du groupe B.

Nous désirons définir, en étudiant chacun des groupes, un paramètre qui mesure la dispersion des notes par rapport à 11 (moyenne des notes de chaque groupe).

Comment pouvons-nous procéder ?

Suggérons de représenter, pour le groupe A, le couple (04 ; 18) et le couple (11 ; 11) dans un repère orthonormé de $\mathbb{R}^2$ de base $(i_1 ; i_2)$ [espace vectoriel euclidien de dimension 2]
Cf: Le couple (11 ; 11) se trouve sur la "bissectrice" ou droite d'équation $X_2 = X_1$.

Demandons de donner la distance (euclidienne ; avec la formule de Pythagore) du couple (11 ; 11) au couple (04 ; 18).
Cette distance est :

$\sqrt{(04-11)^2 + (18-11)^2} = \sqrt{98} = 9{,}89...$
Cf: C'est la norme du vecteur ((04 - 11) ; (18 - 11)) (racine carrée du carré scalaire).

Quant au groupe B, la distance du couple (11 ; 11) au couple (10 ; 12) est :

$\sqrt{(10-11)^2 + (11-12)^2} = \sqrt{2} = 1{,}41...$

La distance considérée *pourrait* être choisie comme paramètre de dispersion par rapport à la moyenne.

1.2.2 - Chacun des groupes comprend 5 élèves

Groupe C : 09 ; 10 ; 12 ; 13 ; 16

Groupe D : 06 ; 08 ; 09 ; 18 ; 19

Les 2 groupes sont dissemblables tout en ayant la même moyenne trimestrielle : 12.
Dans le groupe C un seul élève a moins de 10 et encore il s'en approche.
Dans le groupe D 3 élèves ont une note inférieure à 10 ; en revanche 2 élèves sont excellents.

Le 5-uplet (09 ; 10 ; 12 ; 13 ; 16) et le 5-uplet (12 ; 12 ; 12 ; 12 ; 12) sont représentés dans un repère orthonormé de $\mathbb{R}^5$ de base $(i_1 ; i_2 ; i_3 ; i_4 ; i_5)$
Cf. Le 5-uplet (12 ; 12 ; 12 ; 12 ; 12) se trouve sur la droite d'équation :
$X_5 = \ldots = X_1$.

La distance entre les 2 5-uplets
est :

$$\sqrt{(09-12)^2+(10-12)^2+(12-12)^2+(13-12)^2+(16-12)^2}$$
ou $\sqrt{30} = 5,47...$

La distance entre les 2 5-uplets correspondant au groupe D est :
$\sqrt{146} = 12,08...$

Par suite les notes du groupe D sont plus dispersées par rapport à la moyenne que celles du groupe C.

1.2.3 - Le nombre d'élèves de chaque groupe est différent

Groupe E : 06 ; 08 ; 17 ; 19

Groupe F : 07 ; 09 ; 13 ; 16 ; 18 ; 18

La moyenne des notes du groupe E est : 12,5.
La moyenne des notes du groupe F est : 13,5

Les 2 groupes étant d'effectifs différents, pour pouvoir comparer, il paraît normal de faire intervenir l'effectif

Nous calculerons la moyenne des carrés des "différences des valeurs par rapport à la moyenne".

Puis, nous calculerons la racine carrée positive de cette moyenne.

Pour le groupe E le résultat est :

$\sqrt{1/4 \times ((06-12,5)^2+(08-12,5)^2+(17-12,5)^2+(19-12,5)^2)}$

ou $\sqrt{31,25} = 5,59...$

Pour le groupe F le résultat est : 4,27...

Par suite les notes du groupe E sont plus dispersées par rapport à la moyenne que celles du groupe F.

Cf: Le nombre que nous obtenons est une distance euclidienne avec formule de Pythagore si nous nous plaçons dans un repère orthonormé de IR^4 de base $(i_1\sqrt{4}, i_2\sqrt{4}, ..., i_4\sqrt{4})$

Pour le sujet qui nous préoccupe, nous retiendrons ce procédé de calcul, que les effectifs soient différents ou non.

Une justification sur le choix de cette procédure :
Si nous considérons les 2 groupes :

Goupe G : 03 ; 07

Goupe H : 03 ; 03 ; 07 ; 07

notre intuition nous conduit à estimer que ces 2 goupes de même moyenne 5 sont semblables quant à la dispersion par rapport à la moyenne.
Comme nous le souhaitons, la distance que nous retenons est bien la même : $\sqrt{2}$.

1.3 - Définitions

1.3.1 - Variance

La variance est la moyenne des carrés des "différences des valeurs par rapport à la moyenne".
Notation : σ^2
Cf: "L'écart" étant la valeur absolue de la "différence d'une valeur par rapport à la moyenne", nous disons aussi : la variance est la moyenne des carrés des écarts par rapport à la moyenne.

$$\sigma^2 = 1/n \sum_{i=1}^{n} (x_i - x)^2$$

En utilisant les effectifs :

$$\sigma^2 = \sum_{i=1}^{n} a_i (x_i - x)^2 \quad / \quad \sum_{i=1}^{n} a_i$$

En utilisant les fréquences :

$$\sigma^2 = \sum_{i=1}^{n} f_i (x_i - x)^2$$

Nota : Pour toute la suite de notre document, nous écrirons le signe sigma sans préciser i=1 à n ; cela sera sous-entendu.

1.3.2 - Ecart-type

L'écart-type est la racine carrée positive de la variance.
Notation : σ

Cf: Si x_i a la dimension L alors la variance a la dimension L^2 et l'écart-type a la dimension L.

1.3.3 - Application

Reprenons les notes des goupes C et D de l'établissement "Sport études".

Groupe C

Notes	Différences	Carrés
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
9	-3	9
10	-2	4
12	-0	0
13	1	1
16	4	16
	0	30

Variance : 30/5 = 6

Ecart-type : $\sqrt{6}$ = 2,44...

Groupe D

Notes	Différences	Carrés
x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
6	-6	36
8	-4	16
9	-3	9
18	6	36
19	7	49
	0	146

Variance : 146/5 = 29,2

Ecart-type : $\sqrt{29,2}$ = 5,40...

1.3.4. - Propositions

1.3.4.1 - La somme des "différences des valeurs par rapport à la moyenne" est égale à 0.
*En effet, il s'agit de la **somme** de nombres **diminuée** de la **moyenne** de ces nombres **multipliée** par n.*

1.3.4.2 - La moyenne des carrés des "différences des valeurs par rapport à une valeur z" est minimale et égale à la variance si $z = \bar{x}$.

$$f(z) = 1/n \sum (x_i - z)^2 = 1/n \sum (x_i^2 - 2x_i z + z^2)$$

$$= 1/n \sum x_i^2 - 2(1/n \sum x_i) z + z^2$$

f est une fonction du second degré, le coefficient de z^2 étant positif,
f passe par un minimum égal à :

$$2(1/n \sum x_i) / 2 = \bar{x}$$

1.3.5. - Remarques
Nous avons essayé de montrer que la statistique et le calcul des probabilités font appel à la structure d'espace vectoriel.

Dans $\mathbb{R}^n$ l'écart-type est lié à la notion :

- de distance euclidienne
- de norme euclidienne (racine carrée positive du carré scalaire)

par ailleurs, on pourra voir que dans $\mathbb{R}^n$:
- la covariance est liée au produit scalaire
- le coefficient de corrélation est un cosinus

1.4 - Une propriété de la variance

La variance est égale à la moyenne des carrés des valeurs diminuée du carré de la moyenne.

Cf. le résultat de la variance est en général plus rapidement obtenu en utilisant cette propriété qu'en faisant appel à la définition.

Actuellement, les instruments électroniques permettent des résultats immédiats.

Reprenons, par exemple, les notes des élèves du groupe C de l'établissement "Sport études".

09 ; 10 ; 12 ; 13 ; 16

Les carrés respectifs des notes sont :

81 ; 100 ; 144 ; 169 ; 256

La moyenne des carrés est : 750/5 = 150

Le carré de la moyenne des valeurs étant $12^2 = 144$,
la variance est : 150 - 144 = 6

Démonstration

$$\sigma^2 = 1/n \sum (x_i - \bar{x})^2 = 1/n \sum (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= 1/n \sum x_i^2 - 2(1/n \sum x_i)\bar{x} + \bar{x}^2$$

$$= 1/n \sum x_i^2 - 2\bar{x}^2 + \bar{x}^2$$

$$= 1/n \sum x_i^2 - \bar{x}^2$$

1.5 - Un calcul particulier de la variance

Dans le cas d'une loi binomiale (chaque épreuve donne lieu à 2 éventualités et à 2 seulement) la variance est donnée par la formule suivante :

n x p x (p - 1)

n est le nombre d'épreuves

p est la probabilité d'une éventualité

Reprenons l'exemple du jeu de "Pile" ou "Face" avec 100 jets.
Il s'agit d'une loi binomiale.

La variance est :

100 x 0,5 x 0,5 = 25

Par suite l'écart-type est 5

2 - Importance de l'écart-type

2.1 - L'écart-type, avec la moyenne, se rencontre constamment en statistique et en calcul des probabilités.

Grâce à la notion de distance à laquelle il est lié l'écart-type permet la comparaison de 2 échantillons.

L'écart-type joue un rôle de très grande importance :

- dans la loi normale ou loi de Laplace-Gauss
- dans la loi binomiale
- dans la loi de Poisson
- dans les sondages et les intervalles de confiance

2.2 - Un des rôles importants de l'écart-type dans la loi normale

2.2.1 - Etude d'un exemple

Un agriculteur veut étudier le prix de vente de sa récolte de pommes.
Ce prix dépend de la grosseur des fruits cueillis.
Constituer des tas séparés de toute la récolte d'après les diamètres maximaux représentant un travail trop important, l'agriculteur décide de faire appel à la statistique et au calcul des probabilités et considère un échantillon de 40 pommes.

Les mesures (en mm) effectuées sur cet échantillon sont données dans le tableau suivant :

105	**90**	**100**	**90**	**90**	**90**	**85**	**75**
75	**65**	**95**	**70**	**85**	**80**	**100**	**80**
95	**120**	**95**	**90**	**95**	**110**	**100**	**105**
75	**90**	**90**	**95**	**95**	**105**	**85**	**80**
70	**85**	**100**	**100**	**85**	**85**	**110**	**115**

La moyenne de l'échantillon est : 91,25

L'écart-type est : 12,38

Les mesures des diamètres maximaux des pommes peuvent être acceptées comme réparties suivant une loi normale.
Alors, il est justifié que 95/100 des pommes ont un diamètre appartenant à l'intervalle :

$$[91{,}25 - 1{,}96 \times 12{,}38 / \sqrt{39} ; \; 91{,}25 + 1{,}96 \times 12{,}38 / \sqrt{39}]$$

ou

[87,...;.95,...]

L'agriculteur peut considérer que 95/100 des pommes ont un diamètre compris entre 87 mm et 95 mm.

L'intervalle de confiance auquel il a été fait appel est de la forme :

$$[x - 1{,}96 \times \sigma / \sqrt{(n-1)} ; \; x + 1{,}96 \times \sigma / \sqrt{(n-1)}]$$

2.2.2 - Généralisation

Nous voulons étudier la mesure de la valeur d'un caractère quantitatif des éléments d'un ensemble E.
Cette mesure est acceptée comme répartie suivant une loi normale.

Considérons un sous-ensemble de E (ou échantillon) de n éléments.
x et σ sont respectivement la moyenne et l'écart-type de l'échantillon.

Il est justifié que nous pouvons estimer que 95/100 des éléments de E ont un caractère dont la mesure appartient à l'intervalle :

$$[x - 1{,}96 \times \sigma / \sqrt{(n-1)} ; \; x + 1{,}96 \times \sigma / \sqrt{(n-1)}]$$

Cf. Pour simplifier, le coefficient 1,96 est souvent remplacé par 2.

BIBLIOGRAPHIE

Théorie des probabilités

H.Vensel
Editions Mir

La Statistique

A.Vessereau
Collection Que sais-je ? n° 281

Les Probabilités

A.Jacquard
Collection Que sais-je ? n° 1 571

Encyclopaedia universalis (1985)

(Articles sur la statistique et
sur le calcul des probabilités)