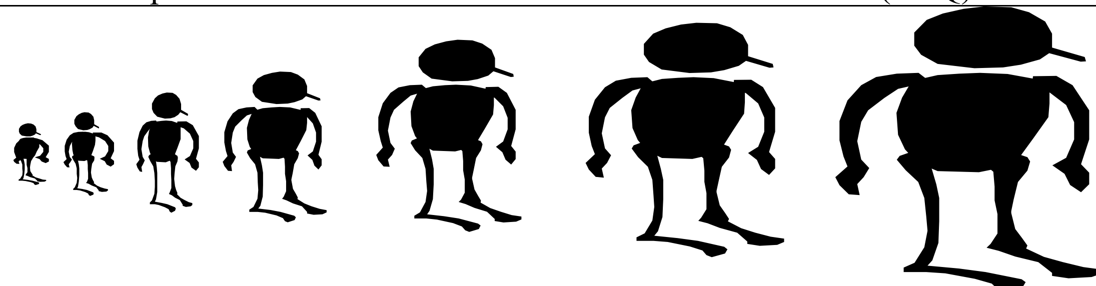
Approche quantitative

La variance

Les objectifs pédagogiques

- Approfondir la notion de variabilité dans une distribution de scores
- Mieux comprendre la relation entre la moyenne et l'écart-type
- Comprendre comment la relation entre deux variables est calculée
- Connaître les facteurs qui font varier le coefficient de corrélation
- En savoir plus concernant l'effet de diverses échelles de mesure
- sur le coefficient de corrélation
- Savoir calculer le pourcentage de variance d'une variable expliquée par une autre variable
- Connaître les réponses à des questions souvent posées quant à la variabilité dans une distribution de scores (FAQ)



Le sommaire

Université de Moncton, Moncton, N.-B. Canada E1A 3E9 longd@umoncton.ca 858-4886

Pour comprendre le rôle véritable de l'écart-type dans les analyses statistiques, il convient avant tout d'en connaître davantage sur la nature de la moyenne.

1. La moyenne, cette vedette statistique

La moyenne est la valeur par excellence pour représenter un ensemble de scores appartenant à un groupe. La moyenne de groupe définit fort bien le groupe, pourvu que la distribution des scores s'approche de la normalité. Pourquoi la moyenne est-elle si efficace pour représenter le groupe ?

Prenons, en guise d'exemple, une distribution hypothétique de scores obtenus par 23 élèves de 7^{e} année en mathématiques. Les scores varient de 39 à 94. Vous désirez trouver une **valeur centrale** qui va le mieux représenter la performance de ce groupe d'élèves.

Pour ce faire, vous pourriez proposer une étendue de scores et dire que, par exemple, 75 % des élèves ont obtenu une note située entre 63 et 78. Cependant, cette étendue de scores ne vous révèle rien sur la nature des résultats obtenus par les autres élèves à l'extérieur de ces limites. De plus, à l'intérieur de ces limites, vous ne savez pas non plus comment vos élèves sont répartis les uns par rapport aux autres. Il est préférable, par conséquent, de se fier à un indicateur unique, un point de repère spécifique sur l'échelle de mesure, plutôt qu'à une étendue de scores.

2. Ça finit par donner zéro !

Allons droit au but. L'indicateur de tendance centrale par excellence d'une distribution normale de scores, c'est la moyenne. La raison en est que la moyenne représente le point le plus stratégique pour calculer la distance de chacun des scores à un point unique de la distribution de manière à ce qu'il y ait le moins de valeurs résiduelles possibles.

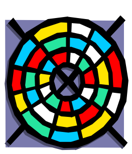
La moyenne est avant tout un concept mathématique, le résultat d'un calcul exécuté sur des symboles, des nombres. Cependant, pour mieux saisir sa véritable nature, il est pratique de s'en faire **une représentation spatiale**. La moyenne n'est pas le juste milieu entre deux extrêmes, mais plutôt **le point milieu le moins éloigné des deux extrêmes**. Ouf, voilà qui mérite une explication !

En additionnant les nombres suivants, 4, 5, 12 et 27, on arrive à 48. En divisant cette somme par 4, on obtient une moyenne de 12. Quel savant calcul !!!
Ainsi, nous pouvons soutenir que la distribution des scores est bien représentée par la moyenne de 12. Ce qui n'est vrai, cependant, que pour un seul score.
Le point médian entre les deux scores extrêmes 4 et 27 se situe à 15,5. Le point médian 15,5 reflète-t-il mieux la distribution que la moyenne 12 ?

Dans le tableau suivant, on constate que **la moyenne représente un point où la différence de chacun de ces scores avec elle est minimale. En fait, elle est zéro (0).** Par ailleurs, le juste milieu entre les deux scores extrêmes, 4 et 27, est 15,5. La différence de chaque score avec ce point médian aboutit à une erreur ou à une valeur résiduelle de 3,5 pour chacun des scores. **Chaque score est, pour ainsi dire, mal placé par une valeur résiduelle de 3,5.** Le point médian d'une distribution n'est pas nécessairement le *juste milieu* de cette distribution.

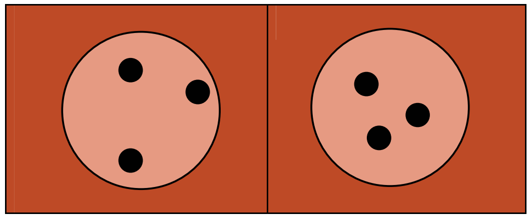
Tableau montrant le résultat de la soustraction de chacun des quatre scores avec la moyenne (12) et avec le point médian (15,5)			
Paramètre	**Score**	**Différence avec la moyenne (12)**	**Différence avec le point médian (15,5)**
	4	-8	-11,5
	5	-7	-10,5
	12	0	-3,5
	27	15	11,5
Somme	48	0	-14,0
Moyenne	12	0	-3,5

Vous comprenez sûrement mieux maintenant pourquoi la moyenne constitue l'indicateur qui représente le mieux une distribution normale de scores. C'est ainsi qu'en connaissant la moyenne et l'écart de chaque score à la moyenne on peut deviner chacun de ces scores en refaisant, pour ainsi dire, le processus inverse. Cependant, si la distribution des scores comporte, par exemple, des scores extrêmes, la moyenne ne constitue pas une mesure centrale qui représente fidèlement une distribution.



3. La distance minimale

Pour mieux comprendre le concept de variance et d'écart-type, faisons appel à une représentation graphique. Chacun des deux cercles dans le tableau suivant contient trois points. Les trois points du cercle de gauche sont éloignés du centre, tandis que ceux du cercle de droite en sont rapprochés. Il est donc évident que la distance moyenne des points du cercle de gauche est plus grande que pour celle des points du cercle de droite.



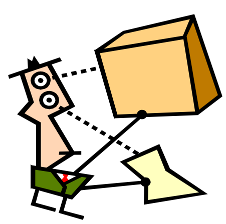
La moyenne représente le juste milieu entre des scores. C'est le point où sa distance avec chaque score est la plus petite. Le centre d'un cercle est aussi le point où sa distance avec chacun des autres points à l'intérieur est la plus petite.

La distance moyenne des points du cercle de droite est inférieure à la distance moyenne des points du cercle de gauche. Dans la discipline de la statistique, cette distance moyenne est comparable à l'écart-type.

Vous constatez déjà que l'écart-type est un concept dynamique qui implique une comparaison avec un critère fixe, en l'occurrence la moyenne.

L'écart-type,
c'est la moyenne
de tous les écarts
de chaque score avec la moyenne.

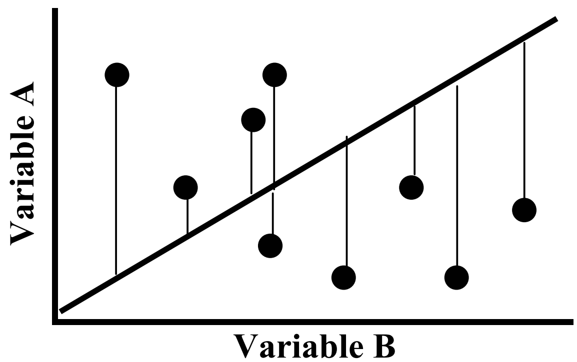
L'écart-type est donc une mesure de variabilité qui nous renseigne sur la distance de chaque score par rapport à la moyenne. C'est, tout à la fois, un indicateur de variabilité et une sorte de moyenne. N'est-ce pas contradictoire ?



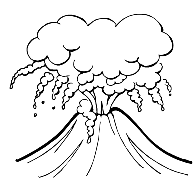
4. La solution des moindres carrés

Bien des analyses statistiques font appel à l'écart de chaque score avec la moyenne de groupe. Lorsqu'il s'agit de mesurer la relation entre deux ou plusieurs variables, certaines analyses font même appel à l'écart de chaque score avec une *ligne imaginaire* située entre ces deux variables. Pour la régression multiple, par exemple, il s'agit de déterminer sur un tableau cartésien à quel endroit passe une ligne droite de façon à ce que la somme des distances au carré de chaque score avec la ligne soit la moindre pour l'ensemble des scores. En fait, on cherche à obtenir les plus petites valeurs résiduelles possibles.

Il existe de nombreux avantages à se représenter la moyenne sur un plan spatial. Certaines analyses statistiques s'appuient même sur des représentations graphiques pour calculer les écarts à une ligne imaginaire dite pente de régression. Bien des analyses multivariées ne sont ni plus ni moins que des extensions de la solution des moindres carrés (en anglais, *least squares solution*). Pour illustrer le concept d'écart optimum à une ligne de régression, voici un graphique dans lequel chaque point représente la jonction entre un score à la variable A et un score à la variable B. La ligne entre le point et la ligne de régression représente l'écart entre le point de jonction des deux scores et la ligne de régression produisant la plus petite valeur résiduelle pour l'ensemble du groupe.



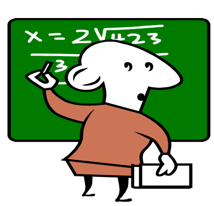
5. De la tendance centrale (moyenne) à la dispersion (écart-type)



Si la moyenne représente bien une distribution normale, elle ne nous en révèle pas toutes les caractéristiques. D'autres valeurs statistiques, entre autres l'écart-type, nous apportent de précieuses informations. Si la moyenne représente le point *le plus approprié* entre des scores extrêmes, **l'écart-type, lui, représente à quel point les scores sont éloignés (ou rapprochés) de la moyenne**.

L'écart-type est, comme la moyenne, une caractéristique déterminante d'une distribution de scores. Il nous informe sur l'étendue des scores. C'est comme si l'écart-type nous disait qu'en général les scores se trouvent à telle distance de la moyenne. Donc, plus l'écart-type est grand, plus grand est l'écart entre la moyenne et chacun des scores. Oui, mais…

L'écart-type est une mesure de la variance, de la variabilité des scores autour de la moyenne. Il est



6. Est-ce que l'écart-type serait, lui aussi, une sorte de moyenne ?

donc une moyenne des écarts de chaque score à la moyenne de la distribution. Ouf ! J'en perds mon latin.

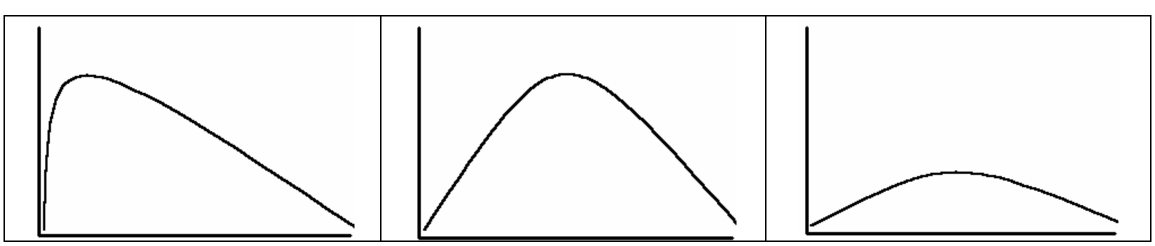
Cessons de jongler avec les mots et disons simplement que, pour obtenir la variance, on doit forcément calculer la différence de chaque score avec la moyenne, élever chacune des différences au carré, en faire la somme qui sera, finalement, divisée par le nombre d'entrées ou de scores. **L'écart-type s'obtient en retirant la racine carrée de la variance.** Il n'est donc pas surprenant que nous parlions tantôt de variance, tantôt d'écart-type.

?	Pourquoi doit-on élever au carré la différence de chaque score à la moyenne pour calculer la variance totale d'une distribution ?
R	En soustrayant chaque score à la moyenne, la somme des scores négatifs serait égale à la somme des scores positifs. Comme le résultat serait toujours zéro (0), on élève chaque score au carré. Ensuite, on calcule la somme de ces scores. Enfin, on élimine le carré de chaque score en tirant la racine carrée de la somme.

Puisque nous calculons la variance à partir de plusieurs scores, on se rend vite compte **qu'elle est sensible à diverses caractéristiques** d'une distribution. Ces scores sont-ils distribués selon une courbe normale ? Trouve-t-on une forte concentration de scores dans une partie de la distribution ? Des scores éloignés de la moyenne rendent-ils la forme de la distribution fortement irrégulière ?

Quant à la variance, ce qui nous intéresse par-dessus tout, c'est son influence sur les résultats d'une multitude d'analyses statistiques. À vrai dire, **l'effet produit par la dispersion des scores** d'une distribution sur les résultats des analyses statistiques est incommensurable.

Le graphique de gauche dans le tableau suivant représente une distribution anormale de scores à cause de la présence de scores extrêmes, éloignés de la moyenne. La plupart des scores, cependant, sont concentrés du côté gauche du graphique. Le graphique de droite est caractérisé par des scores éloignés de chaque côté de la moyenne. Le graphique du centre représente une distribution dite normale.



La variabilité des individus par rapport à une caractéristique quelconque : une notion cruciale en sciences humaines

Que nous révèle donc l'écart-type sur la nature des scores comme tels ? Un petit écart-type indique que les scores varient très peu sur l'échelle de mesure utilisée. Et plus grand est l'écart-type, plus les limites extrêmes s'éloignent. Nul besoin d'insister sur le fait que la moyenne peut rester inchangée, même lorsque la variance augmente.

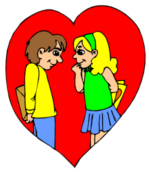
Selon Nunnally (1978), « les questions d'ordre scientifique ne se posent que dans la mesure où les objets ou les personnes en question varient par rapport à des attributs particuliers ». Les caractéristiques constantes chez les êtres humains ou ailleurs dans l'univers sont intéressantes en soi. Mais, il faut admettre que **la plupart des phénomènes qui influent sur nous sont des variables**. Nous cherchons donc à les comprendre en vue de les modifier ou de les prédire. C'est ainsi que nous formulons des hypothèses de recherche qui prétendent expliquer la relation entre les composantes d'une problématique ou d'un phénomène. Au fil du temps et de l'accumulation de données empiriques sur le phénomène, des théories explicatives émergent.

Comme nous cherchons à préciser la relation entre les composantes (variables) d'un phénomène, il ne faut pas se surprendre si une forte proportion des analyses statistiques se fonde sur la variance qui rend compte de la variabilité entre les individus. Des analyses statistiques qui dépendent beaucoup de la variance dans une distribution sont la corrélation et la régression multiple, entre autres. La corrélation, en apparence simple et banale, constitue **la pierre angulaire de nombreuses analyses statistiques beaucoup plus complexes**.

Le coefficient de corrélation nous informe sur le degré d'association entre deux variables. Ce degré d'association est calculé au moyen de la variabilité des scores de chacune des variables. Cependant, **le coefficient de corrélation ne nous indique rien sur la dispersion des scores**. Un coefficient de corrélation élevé ne signifie pas que les scores de chaque variable sont éloignés de la moyenne.

Le coefficient nous informe sur
le degré d'association ou de rapport
entre des variables,
et non sur la dispersion respective des scores
des variables en cause.

Avant d'aborder le calcul du coefficient de corrélation, il est impérieux de comprendre la nature véritable des rapports ou de l'absence de rapports entre des variables.



7. Le rapport entre des variables

Question : Quand deux variables sont-elles associées ou reliées l'une à l'autre ?
Réponse : **Lorsqu'une variation dans une variable s'accompagne d'une variation dans une autre variable.**

Le tableau qui suit montre qu'il existe trois types de relation entre deux variables selon que les deux varient dans la même direction, dans des directions opposées ou selon qu'elles sont indépendantes l'une de l'autre.

Type de relation	Direction	Description
Positive	Unidirectionnelle	**Une augmentation dans une variable s'accompagne d'une augmentation dans l'autre variable.**
Négative	Bidirectionnelle	**Une augmentation dans une variable s'accompagne d'une diminution dans une autre variable.**
Nulle	Imprévisible	**La variation dans une variable est indépendante de la variation dans une autre variable.**

La représentation graphique qui suit pourrait vous aider à mieux comprendre la nature de ces trois types de relation qui peuvent exister entre deux variables.

Dans la partie située à gauche du tableau qui suit, on constate que le sujet n° 1 obtient la note la plus faible à la variable A et la note la plus faible à la variable B. Le sujet n° 4 obtient la deuxième note la plus élevée aux deux variables.

	Relation positive			Relation négative	
Sujet	A	B		A	B
N° 1	2	12		2	44
N° 2	3	16		3	36
N° 3	6	32		6	32
N° 4	12	36		12	16
N° 5	17	44		17	12
Moyenne	8	28		8	28

On voit clairement que la position de chaque individu à la variable A est la même que pour la variable B. Autrement dit, si le sujet obtient une note supérieure à la moyenne à la variable A, il obtient aussi une note supérieure à la moyenne à la variable B. En faisant, pour ainsi dire, la somme des relations des cinq sujets, on conclut que, plus un score augmente à la variable A, plus il augmente aussi à la variable B. L'inverse est tout aussi vrai : plus un score diminue à la variable A, plus il diminue à la variable B.

Le scénario est tout à fait différent pour la partie droite du tableau. Le sujet n° 1 obtient la note la plus faible à la variable A en même temps que la note la plus élevée à la variable B. Pour chacun des autres sujets, une note au-dessus de la moyenne à une variable s'accompagne d'une note au-dessous de la moyenne à l'autre variable. C'est une relation inverse.

La relation de chaque score à la moyenne d'une distribution est cruciale pour bien comprendre le calcul de l'écart-type et de la variance.

Le tableau suivant présente une corrélation nulle. En fait, ce n'est pas une relation à proprement parler, car elle représente plutôt une absence de relation entre deux variables. Pour chaque sujet on ne peut donc pas prédire le score sur la variable B en connaissant son score sur la variable A. **Les deux variables sont, par conséquent, indépendantes l'une de l'autre**.

	Relation nulle	
Sujet	A	B
N° 1	2	12
N° 2	3	16
N° 3	6	32
N° 4	12	36
N° 5	17	44
Moyenne	8	28

8. La corrélation

Question : Pourquoi doit-on s'intéresser à partir de maintenant à la corrélation ?
Réponse : Le calcul du coefficient de corrélation se fait à partir de l'écart de chaque score avec la moyenne de sa distribution.

Quelle stratégie pouvons-nous utiliser pour mieux comprendre l'analyse de corrélation ? On peut tenter de s'en faire une *image mentale*, même si cette image comporte des imperfections. Voici une façon de s'imaginer la nature de la corrélation, même si nous venons à peine d'en parler.

La corrélation est en quelque sorte *une mesure de position* : pour chaque individu elle révèle, pour ainsi dire, la position qu'il occupe sur chacune des variables. Le score sur une variable est comparé au score sur l'autre variable.

La corrélation sera positive si, pour l'ensemble du groupe, un score se trouve à une position similaire dans les deux distributions. (Cette façon de s'imaginer la corrélation s'applique mieux, il est vrai, à la corrélation des rangs, mais qu'importe.)

Une corrélation sera négative si les scores d'une distribution se trouvent à une position inverse sur l'autre variable. L'individu typique de ce groupe obtient un score plutôt élevé sur une variable en même temps qu'un score plutôt bas sur l'autre variable.

Les trois types de corrélation					
Négative		Positive		Nulle	
A	**B**	**A**	**B**	**A**	**B**
1	**8**	**1**	**1**	**1**	**7**
2	**7**	**2**	**2**	**2**	**2**
3	**6**	**3**	**3**	**3**	**5**
4	**5**	**4**	**4**	**4**	**3**
5	**4**	**5**	**5**	**5**	**8**
6	**3**	**6**	**6**	**6**	**6**
7	**2**	**7**	**7**	**7**	**1**
8	**1**	**8**	**8**	**8**	**4**

Par ailleurs, une corrélation sera nulle si le coefficient s'approche de zéro. Elle indique qu'un score sur une variable ne nous aide pas à savoir où se situe son score sur l'autre variable. Et cette constatation est vraie pour l'ensemble des sujets. Dans ce cas, il n'y a donc pas d'association entre les deux variables.

Le fait que la corrélation soit élevée entre deux variables ne signifie pas pour autant que la variabilité soit grande pour les deux variables. **Une corrélation élevée signifie plutôt que les deux varient simultanément (dans la même direction ou dans des directions opposées).**

C'est la relation entre les deux variables qui fait varier le coefficient de corrélation, et non la variabilité à elle seule.
Une augmentation dans la variabilité des scores ne fera augmenter le coefficient de corrélation **que si** elle améliore la relation entre les deux variables.

Vous voulez établir le lien entre le nombre de services reçus et la satisfaction de la clientèle d'un organisme gouvernemental.
Il y a donc deux variables à mettre en relation.
La **première**, c'est le nombre de services reçus qui se mesurent aisément en indiquant un nombre approximatif de services.
La **deuxième**, c'est le degré de satisfaction que vous pourriez définir de maintes façons. Vous pourriez, entre autres, offrir aux personnes interrogées une échelle de mesure de satisfaction plutôt courte.
Chacun cocherait l'une des réponses suivantes :
(1) Insatisfait ; (2) Satisfait.
Par contre, vous pourriez leur présenter une échelle de mesure plus élaborée.

Voici l'exemple d'une échelle de mesure plus élaborée :

1 Totalement insatisfait
2 Modérément insatisfait
3 Un peu insatisfait
4 Incertain
5 Un peu satisfait
6 Modérément satisfait
7 Totalement satisfait

En quoi cette échelle de mesure plus longue est-elle plus avantageuse que la première échelle plus courte, l'échelle dichotomique ? L'individu qui s'est dit insatisfait sur l'échelle courte pourrait cocher 1, 2 ou 3 sur la deuxième échelle. Il est possible qu'en augmentant la variabilité dans les scores le coefficient augmente parce qu'une échelle plus longue ajoute plus de précision au comportement mesuré.

Il n'est pas surprenant que les chercheurs et les statisticiens recommandent fortement d'utiliser des échelles de mesure ordinales et de proportion de préférence aux échelles dichotomiques et ordinales.

Question : *Que se passe-t-il vraiment lorsqu'on remplace une échelle nominale par une échelle ordinale, ou une échelle à intervalles par une échelle de proportion ?* **Réponse** : *La variabilité augmente.*

De plus, les scores représentent probablement mieux les variations réelles dans le comportement mesuré, apportant ainsi une plus grande précision qui pourrait, par ricochet, se refléter par un coefficient de corrélation plus élevé. Une plus grande variabilité dans les scores de chacune des distributions ne doit pas faire gonfler le coefficient de corrélation de façon artificielle. Une plus grande variabilité doit plutôt permettre plus de précision. Ainsi, la relation partagée entre les variables en question sera mieux servie.

Voici un exemple convaincant de l'importance qu'il y a à utiliser une échelle de mesure la plus longue qui soit si nous voulons mesurer la relation entre deux variables avec le plus de précision possible. Ainsi, le coefficient de corrélation devrait traduire ce raffinement accru.

D'abord, nous avons une variable nommée A. Elle comprend des chiffres allant de 4 à 12. Ensuite, nous avons une autre variable, qui porte le nom de B et qui est de nature dichotomique pour la circonstance. Imaginons que les valeurs 1 et 2 de la variable B soient le résultat d'un lancement d'un dé. On le lance à 10 reprises. À chaque lancement, si le résultat est un chiffre de 1 à 3, nous inscrivons le score 1. Par ailleurs, si le score se situe entre 4 et 6, nous lui attribuons le score 2. En fait, à partir des scores réels obtenus, nous réduisons, pour ainsi dire, la variabilité de l'échelle de mesure : elle passe de 6 possibilités (A) à 2 possibilités seulement (B).

Voyez ci-dessous un tableau de calcul de la corrélation. Il est évident que l'écart-type de la variable B est restreint (0,527). Maintenant, comment le coefficient de corrélation peut-il être affecté à la baisse ?

Sujet	A	B	a	b	a*a	b*b	a*b
a	5	1	-3,0	-0,5	9,0	0,25	1,5
b	10	1	2,0	-0,5	4,0	0,25	-1,0
c	7	2	-1,0	0,5	1,0	0,25	-0,5
d	10	2	2,0	0,5	4,0	0,25	1,0
e	4	1	-4,0	-0,5	16,0	0,25	2,0
f	9	2	1,0	0,5	1,0	0,25	0,5
g	10	2	2,0	0,5	4,0	0,25	1,0
h	5	1	-3,0	-0,5	9,0	0,25	1,5
i	8	1	0,0	-0,5	0,0	0,25	0,0
j	12	2	4,0	0,5	16,0	0,25	2,0
Somme	**80**	**15**	**0,0**	**0,0**	**64,0**	**2,5**	**8,0**
Moyenne	**8**	**1,5**					
Écart-type	**2,67**	**0,527**					

$$\mathbf{r_{ab}} = \frac{\mathbf{S_{ab}}}{\sqrt{\mathbf{a^2b^2}}} = \frac{\mathbf{8}}{\mathbf{12{,}65}} = \mathbf{0{,}63}$$

La dernière colonne (a*b) du tableau sert de numérateur au coefficient. Plus le numérateur est petit, plus le coefficient est petit lui aussi. Le résultat de cette colonne est le produit des colonnes a et b, lesquelles comprennent l'écart de chaque score à sa moyenne respective. On constate que la variabilité est petite dans la colonne b. On ne peut donc s'attendre à un numérateur imposant ! De fait, le coefficient de corrélation est de 0,63, résultat tout de même significatif à 0,05 de probabilité.

Le tableau suivant comporte une différence substantielle avec le précédent. Dans ce tableau, il y a non plus la variable B, mais la variable C. Cette dernière variable comprend des scores allant de 1 à 6, soit les scores obtenus en lançant le dé 10 fois. On voit, par exemple, pour le *sujet j* que son score n'est plus 2, mais bien 6. Quel sera donc l'effet de cette augmentation dans la variabilité des scores ?

Sujet	A	C	a	c	a*a	c*c	a*c
a	5	2	-3,0	-1,6	9,0	2,56	4,8
b	10	3	2,0	-0,6	4,0	0,36	-1,2
c	7	4	-1,0	0,4	1,0	0,16	-0,4
d	10	6	2,0	2,4	4,0	5,76	4,8
e	4	1	-4,0	-2,6	16,0	6,76	10,4
f	9	4	1,0	0,4	1,0	0,16	0,4
g	10	5	2,0	1,4	4,0	1,96	2,8
h	5	2	-3,0	-1,6	9,0	2,56	4,8
i	8	3	0,0	-0,6	0,0	0,36	0,0
j	12	6	4,0	2,4	16,0	5,76	9,6
Somme	**80**	**36**	**0,0**	**0,0**	**64,0**	**26,4**	**36,0**
Moyenne	**8**	**3,6**					
Écart-type	**2,67**	**1,713**					

$$\mathbf{r_{ac} = \frac{S_{ac}}{\sqrt{a^2c^2}} = \frac{36}{41,10} = 0,88}$$

L'écart-type de la variable C est de 1,713. On constate surtout que la somme dans la dernière colonne est passée de 8 pour la variable A à 36 pour la variable C. Ce nombre servira de numérateur au coefficient de corrélation. On peut déjà prévoir que le coefficient de corrélation va augmenter. De fait, il est maintenant de 0,88 et il atteint le niveau de signification de 0,001.

La variabilité des scores influence la grandeur d'un coefficient de corrélation. Ce n'est cependant pas vrai dans tous les cas, parce que le coefficient de corrélation est avant tout influencé par **le rapport des deux scores pour chacun des sujets**.

Lorsque les deux scores d'un sujet se trouvent du même côté de leur moyenne respective, ils contribuent à un **coefficient de corrélation positif**. Si un des deux scores est moins élevé que la moyenne de sa distribution, ils contribuent à un **coefficient de corrélation négatif**.

?	Quand un coefficient de corrélation devient-il **positif** pour un groupe d'individus ou pour un ensemble de paires de scores ?
R	Lorsque la variabilité totale accumulée pour les paires de scores *positifs* (les deux scores se trouvant du même côté de leur moyenne respective) est supérieure à la variabilité totale accumulée par l'ensemble des paires de scores *négatifs* (les deux scores se trouvant à des côtés opposés de leur moyenne respective).
?	Quand un coefficient de corrélation devient-il **négatif** pour un groupe d'individus ou pour un ensemble de paires de scores ?
R	Lorsque la variabilité totale accumulée par les paires de scores *négatifs* est supérieure à la variabilité totale accumulée par les paires de scores *positifs*.

?	Quand un coefficient de corrélation s'approche-t-il de **zéro** ?
R	Quand la variabilité totale accumulée par les paires de scores *positifs* est équivalente à la variabilité totale accumulée par les paires de scores *négatifs*.

La grandeur du coefficient de corrélation dépend du **montant de variabilité** comme tel et non pas du nombre de paires de scores en soi. Il faut bien reconnaître, cependant, que plus le nombre de paires de scores dans un ensemble va dans la même direction, plus grand sera le coefficient de corrélation positive. C'est ainsi que quelques paires de scores ajoutant beaucoup de variabilité peuvent parfois faire grimper de façon appréciable le coefficient de corrélation. Les scores extrêmes peuvent affecter sérieusement le coefficient de corrélation. De là l'importance de s'assurer que les distributions de scores se rapprochent le plus possible de la normalité.

9. L'effet de scores extrêmes sur le coefficient de corrélation

Qu'arrive-t-il à un coefficient de corrélation lorsque des scores extrêmes sont présents dans une distribution ? Pour en constater l'effet, nous allons reprendre l'exemple précédent en ne changeant qu'un seul score pour la variable C. Le *sujet j* aura un score de 26 au lieu de 6 à la variable C.

Dans le tableau qui suit, on constate à quel point un seul score extrême peut affecter le coefficient de corrélation. Que s'est-il passé pour que le coefficient de corrélation passe de 0,88 à 0,66 après que nous avons changé un seul score ?

Pourtant, la variabilité a augmenté pour la variable C. Et, comme la variabilité a augmenté sous l'incidence d'un score extrême, le coefficient de corrélation ne devrait-il pas lui aussi augmenter au lieu de diminuer ?

Sujet	A	C	a	c	a*a	c*c	a*c
a	5	2	-3,0	-3,6	9,0	12,96	10,8
b	10	3	2,0	-2,6	4,0	6,76	-5,2
c	7	4	-1,0	-1,6	1,0	2,56	1,6
d	10	6	2,0	0,4	4,0	0,16	0,8
e	4	1	-4,0	-4,6	16,0	21,16	18,4
f	9	4	1,0	-1,6	1,0	2,56	-1,6
g	10	5	2,0	-0.6	4,0	0,36	-1,2
h	5	2	-3,0	-3,6	9,0	12,96	10,8
i	8	3	0,0	-2,6	0,0	6,76	0,0
j	12	26	4,0	20,4	16,0	416,16	81,6
Somme	**80**	**56**	**0,0**	**0,0**	**64,0**	**482,4**	**116,0**
Moyenne	**8**	**5,6**					
Écart-type	**2,67**	**7,32**					

$$\mathbf{r_{ac}} = \frac{\mathbf{S_{ac}}}{\sqrt{\mathbf{a^2c^2}}} = \frac{\mathbf{116}}{\mathbf{175{,}71}} = \mathbf{0{,}66}$$

Voici comment expliquer cette diminution. Avant de changer le score du *sujet j* à la variable C, il y avait cinq scores au-dessus de la moyenne. Après avoir changé le score de 6 à 26, il n'y a plus que deux scores au-dessus de la moyenne, laquelle est passée de 3,6 à 7,32. Comme le calcul du coefficient de corrélation est fondé sur la différence de variabilité entre les paires de scores positifs et les paires de scores négatifs, il n'est donc pas surprenant que le coefficient ait diminué sensiblement.

Pour saisir l'effet d'un score extrême sur le coefficient de corrélation, il faut faire **la comparaison entre les paires de scores quant à leur position relative par rapport à la moyenne**. Un score peut être égal (=), supérieur (+) ou inférieur (-) à la moyenne. Pour le tableau où la corrélation est de 0,88, voici comment se répartissent les paires de scores.

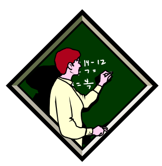
A	B	C	D	E
+ -	- -	- +	+ +	= -
	- -		+ +	
	- -		+ +	
			+ +	

Les paires de scores dans les colonnes B et D contribuent à une corrélation positive (7 paires). Les paires de scores dans les colonnes A et C contribuent à une corrélation négative. Il y a donc énormément plus de paires de scores qui contribuent à une corrélation positive qu'à une corrélation négative (350 % de plus). Voyons maintenant ce qui se produit lorsque nous changeons un score pour le *sujet j*.

A	B	C	D	E
+ -	- -		+ +	= -
+ -	- -		+ +	
+ -	- -			
	- -			

Il y a maintenant 6 paires de scores, au lieu de 7 paires, qui contribuent à la corrélation positive. Par ricochet, 3 paires de scores, au lieu de 2, contribuent à la corrélation négative. Avant le changement, il y avait 3 fois et demi plus de paires contribuant à la corrélation positive ; après le changement, ces paires ne sont plus que deux fois plus nombreuses. Voilà pourquoi le coefficient de corrélation a subi un tel contrecoup lorsqu'un score a été changé pour le *sujet j*.

10. Le calcul du coefficient de corrélation



Pour bien montrer comment divers aspects d'une distribution de scores influent sur le coefficient de corrélation, voici un exemple fictif de corrélation entre deux variables. L'exemple est repris en apportant à chaque fois un changement quelconque de manière à mieux montrer comment un tel changement agit sur le coefficient de corrélation.

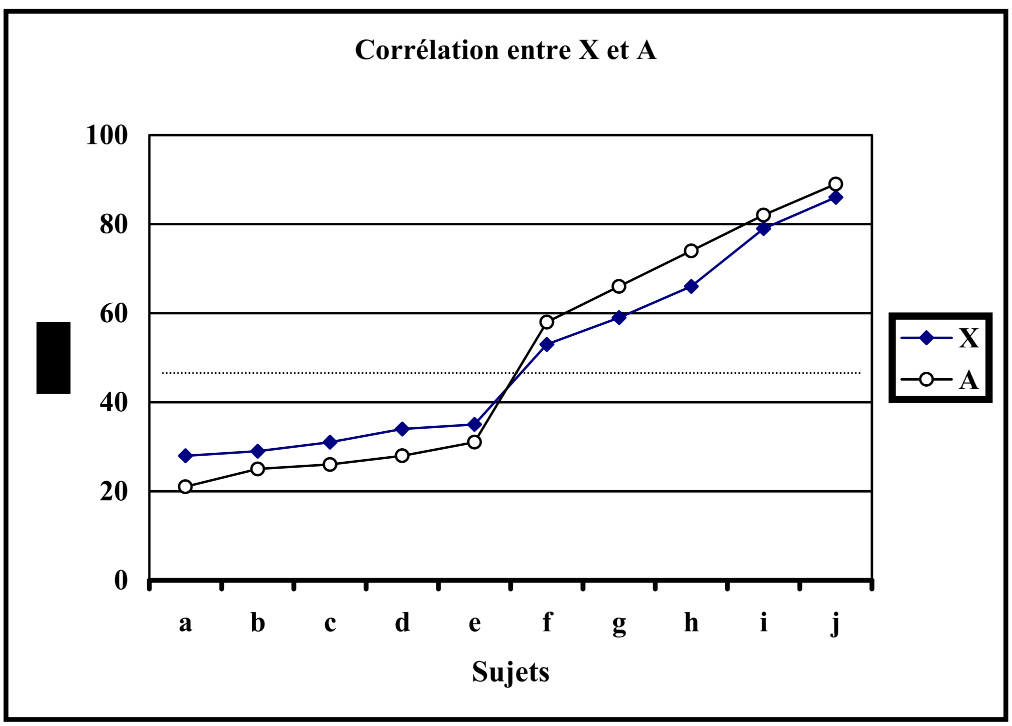
Quel nom porte cette formule magique ?

$$\sqrt{\frac{Sx^2}{N - 1}}$$

Exemple 1

Pour les cinq premiers sujets, le score X est plus élevé que le score A. Pour les cinq derniers sujets, le score X est moins élevé que le score A. Cependant, aucun score plus bas que la moyenne pour une variable n'est associé à un score plus haut que la moyenne pour l'autre variable. Pour chaque paire, les deux scores se situent du même côté de leur moyenne respective.

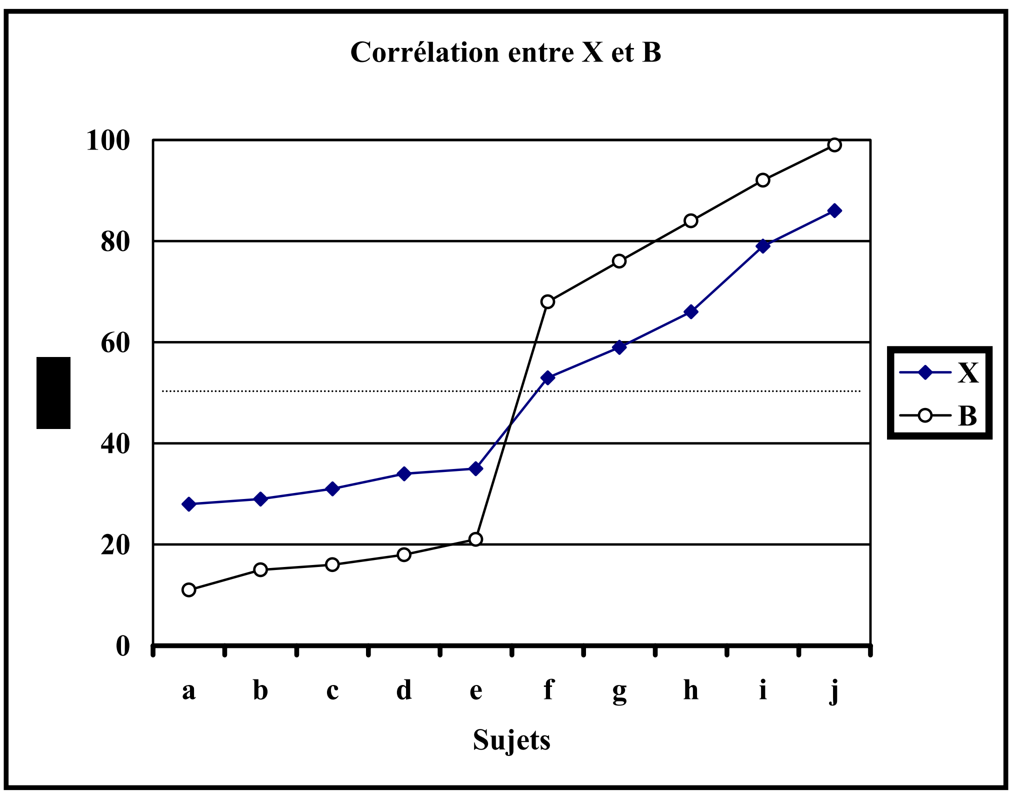
Sujet	X	A	B	C	D
1	28	**21**	11	21	21
2	29	**25**	15	18	84
3	31	**26**	16	11	11
4	34	**28**	18	16	68
5	35	**31**	21	15	99
6	53	**58**	68	92	92
7	59	**66**	76	68	16
8	66	**74**	84	99	15
9	79	**82**	92	84	18
10	86	**89**	99	76	76
Somme	500	**500**	500	500	500
Moyenne	50	**50**	50	50	50
Écart-type	21,7	**26,5**	36,6	36,6	36,6
Variance	472,2	**703,1**	1343,1	1343,1	1343,1
Corrélation		**0,99**	0,98	0,85	-0,15



Exemple 2

Par rapport à la variable A, chacun des scores de la variable B a été diminué de 10 ou augmenté de 10. Cependant, la même relation entre les deux scores d'une même paire a été conservée (par rapport à la variable A). Même si la variance a augmenté,
la relation entre les scores d'une même paire n'a pas été modifiée.

Sujet	X	A	B	C	D
1	28	21	**11**	21	21
2	29	25	**15**	18	84
3	31	26	**16**	11	11
4	34	28	**18**	16	68
5	35	31	**21**	15	99
6	53	58	**68**	92	92
7	59	66	**76**	68	16
8	66	74	**84**	99	15
9	79	82	**92**	84	18
10	86	89	**99**	76	76
Somme	500	500	**500**	500	500
Moyenne	50	50	**50**	50	50
Écart-type	21,7	26,5	**36,6**	36,6	36,6
Variance	472,2	703,1	**1343,1**	1343,1	1343,1
Corrélation		0,99	**0,98**	0,85	-0,15

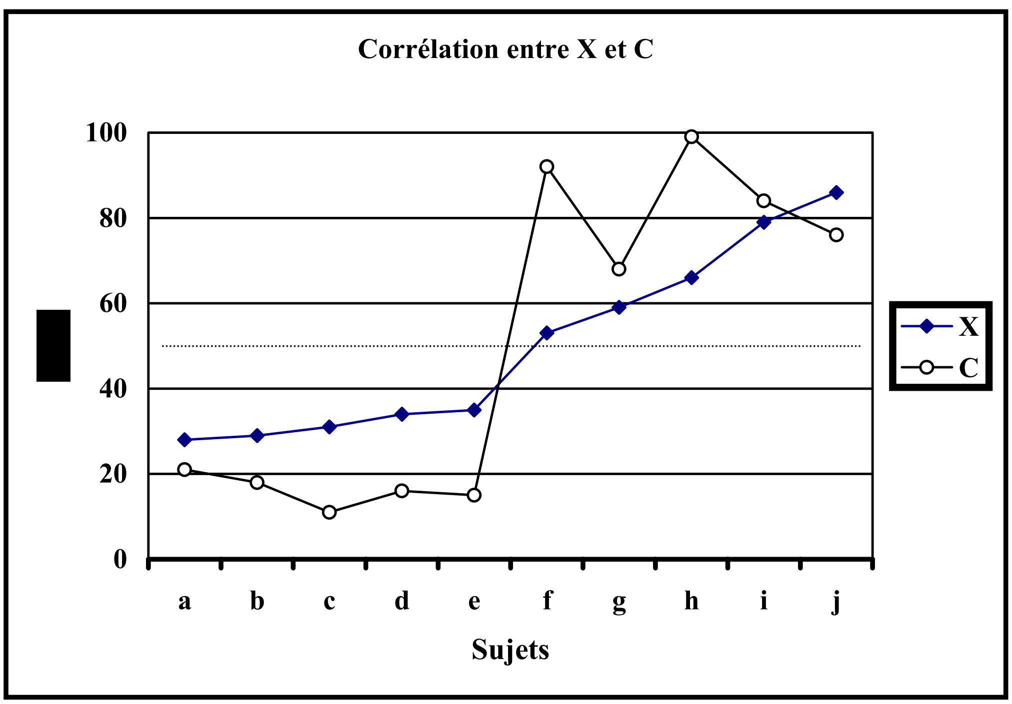


Université de Moncton, Moncton, N.-B. Canada E1A 3E9 longd@umoncton.ca 858-4886

Exemple 3

Par rapport à la variable B, les scores de la variable C ont été mélangés à l'intérieur du groupe des cinq premiers sujets. La même opération a été faite pour les scores des cinq derniers sujets. Pour une seule paire de scores, la relation a changé quelque peu, ce qui explique la diminution dans le coefficient de corrélation. La relation entre les deux scores de neuf paires n'a pas changé.

Sujet	X	A	B	C	D
1	28	21	11	**21**	21
2	29	25	15	**18**	84
3	31	26	16	**11**	11
4	34	28	18	**16**	68
5	35	31	21	**15**	99
6	53	58	68	**92**	92
7	59	66	76	**68**	16
8	66	74	84	**99**	15
9	79	82	92	**84**	18
10	86	89	99	**76**	76
Somme	500	500	500	**500**	500
Moyenne	50	50	50	**50**	50
Écart-type	21,7	26,5	36,6	**36,6**	36,6
Variance	472,2	703,1	1343,1	**1343,1**	1343,1
Corrélation		0,99	0,98	**0,85**	-0,15

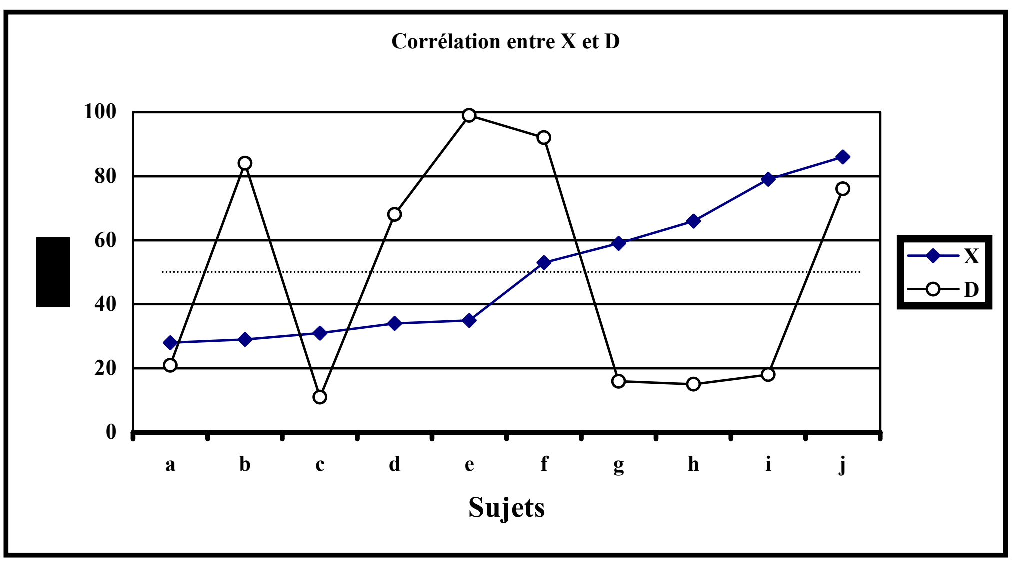


On constate ici que le coefficient de corrélation a diminué parce que, pour le *sujet j*, la relation entre les scores a changé.

Exemple 4

Par rapport à la variable C, six scores de la variable D ont été déplacés dans la distribution de sorte que, pour ces trois paires, un score plus bas que la moyenne sur une variable (X ou D) est maintenant associé à un score plus élevé que la moyenne sur l'autre variable. Pour six paires la relation entre les variables va dans une direction, tandis que pour les quatre autres paires elle va dans l'autre direction. La corrélation apparaît maintenant avec un signe négatif, traduisant la suprématie des paires de scores inverses sur les autres paires qui vont dans la même direction.

Sujet	X	A	B	C	D
1	28	21	11	21	**21**
2	29	25	15	18	**84**
3	31	26	16	11	**11**
4	34	28	18	16	**68**
5	35	31	21	15	**99**
6	53	58	68	92	**92**
7	59	66	76	68	**16**
8	66	74	84	99	**15**
9	79	82	92	84	**18**
10	86	89	99	76	**76**
Somme	500	500	500	500	**500**
Moyenne	50	50	50	50	**50**
Écart-type	21,7	26,5	36,6	36,6	**36,6**
Variance	472,2	703,1	1343,1	1343,1	**1343,1**
Corrélation		0,99	0,98	0,85	**-0,15**



Le coefficient de corrélation est tombé à zéro parce que la relation a changé pour six paires de scores par rapport aux exemples précédents.

11. Comment calculer le pourcentage de variance expliquée

Le coefficient de corrélation mesure le degré d'association entre deux variables, X et Y, à partir de la variance commune que les deux partagent. Si les deux variables sont fortement associées, cela signifie que le sujet qui s'éloigne à une telle distance de la moyenne sur la variable X s'éloigne de la moyenne sur la variable Y à peu près à la même distance. Et cela est vrai pour l'ensemble du groupe. La corrélation nous informe, pour ainsi dire, sur la ***distance commune*** partagée par les deux variables.

Avant d'explorer ce concept de **variance commune** entre deux variables, examinons le concept de **variance individuelle** d'une seule variable. La variance individuelle se définit comme la somme des distances de chaque score avec la moyenne de sa variable. Chaque score contribue à cette variance individuelle. On parle alors de la proportion ou du pourcentage de variance. Calculons la variance individuelle des cinq scores suivants :

Score	Moyenne	X-M	$(X-M)^2$	% de variance
12	14,6	-2,6	6,76	16,41
16	14,6	1,4	1,96	4,76
19	14,6	-4,4	19,36	46,99
11	14,6	3,6	12,96	31,46
15	14,6	-0,4	0,16	0,39
			41,2	100,00

Le score 12, par exemple, s'éloigne de la moyenne par une distance de -2,6. Cette variance représente 16,41 % de la variance totale de la distribution (6,76/41,2 x 100 = 16,41 %). La valeur de 6,76, soit $(X-M)^2$, a été obtenue en soustrayant le score 12 de la moyenne (14,6) pour élever ensuite au carré cette différence de -2,6.

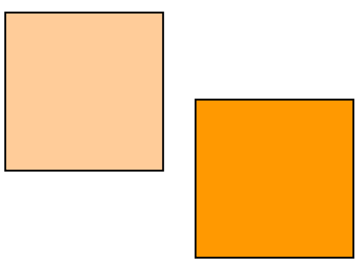
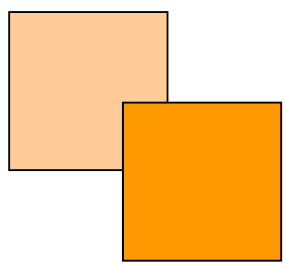
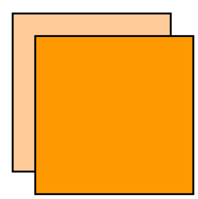
Le score 19, le plus éloigné de la moyenne, représente 46,99 % de la variance totale. Vous comprenez déjà que, dans le cas d'une corrélation entre deux variables, plus la variabilité augmente pour chacune d'elles, plus il y a de chances que le coefficient de corrélation augmente. Cependant, et au risque de se répéter,

……une grande variabilité
dans les scores de chaque distribution
n'entraîne pas, ipso facto,
une augmentation
dans le coefficient de corrélation.

Pour comprendre ce que signifie le pourcentage de variance de la variable X expliquée par la variable Y, il faut obtenir auparavant un indicateur du degré d'association entre ces deux variables. Cet indicateur est justement le coefficient de corrélation. Aussi bien dire que le coefficient de corrélation sert à mesurer aussi le pourcentage de variance d'une variable expliquée par une autre variable. Dans ce cas, il porte le nom de **coefficient de détermination** et est représenté par la formule suivante :

$$\textbf{coefficient de détermination} = r^2_{xy} \times 100$$

La variance expliquée est en quelque sorte une variance partagée, une variance commune entre deux variables. La variance partagée se représente facilement en joignant les carrés dans le tableau qui suit. Les variables peuvent ne partager aucune variance commune entre elles (ensemble 1). Elles peuvent partager un faible pourcentage de variance commune (ensemble 2) ou un fort pourcentage de variance commune (ensemble 3). Dans l'exemple suivant, la variance commune, c'est la surface partagée par les deux carrés. La surface non partagée représente la variance spécifique ou la *variance erreur* attribuable à divers facteurs connus ou inconnus.

Ensemble 1	**Ensemble 2**	**Ensemble 3**
		

Le concept de variance commune est difficile à saisir en examinant chaque paire de scores, une à la fois. **La variance commune est, avant tout, un résultat de groupe**. Néanmoins, s'il existe une variance commune entre deux variables, c'est qu'il existe une association entre elles. Et pour qu'il y ait une association, chaque score de la paire doit s'écarter de sa moyenne respective d'un montant similaire. Les deux scores peuvent être supérieurs ou inférieurs à la moyenne comme dans le cas de la corrélation positive. Pour la corrélation négative, l'un des scores est inférieur à la moyenne, tandis que l'autre est supérieur à la moyenne de sa distribution.

Le coefficient de corrélation varie de -1,0 à +1,0	-1,0 ⇔ 0 ⇔ +1,0

Comme le coefficient de corrélation mesure le degré d'association entre deux variables par le biais de la variance qu'elles partagent, faisons le calcul du coefficient de corrélation entre les scores obtenus par cinq élèves en français et en anglais.

Élève	**Français (F)**	**Anglais (A)**	f	a	f*a
a	54	59	-16	-9,8	156,8
b	63	68	-7	-0,8	5,6
c	72	61	2	-7,8	-15,6
d	78	69	8	0,2	1,6
e	83	87	13	18,2	236,6
Somme	350	344	0	0	385,0
Moyenne	70,0	68,8			
Écart-type	11,64	11,05			
Variance	135,5	122,2			

En examinant la dernière colonne, on constate que certaines paires de scores sont éloignées de leur moyenne respective (sujets a et e), tandis que d'autres sont tout près de leur moyenne de groupe (sujets b et d). Il y a donc beaucoup de variabilité dans les scores en français ainsi qu'en anglais. Mais, les scores sont-ils reliés ? Partagent-ils de la variance commune ? Dans quelle direction les scores varient-ils ?

La dernière colonne représente la covariation ou la covariance entre les deux distributions :

$$Cov_{fa} = \frac{\Sigma fa}{N - 1} = \frac{385}{4} = 96,25$$

Afin que les deux distributions soient comparées sans subir l'influence d'une différence entre les échelles de mesure, la covariance est divisée par le produit des deux écarts-types.

$$\frac{Cov_{fa}}{S_f \times S_a} = \frac{96,25}{128,62} = 0,75$$

La corrélation entre les scores en français et les scores en anglais est donc de 0,75. On élève au carré ce coefficient, ensuite on le multiplie par 100 pour obtenir la variance partagée par les deux distributions.

$$r^2_{fa} \times 100 = 0,75 \times 100 = 56,25 \text{ \% de variance expliquée}$$

Il y a donc 43,75 % de variance qui n'est pas partagée par les deux variables. Même si cette relation semble importante, elle n'est pas significative statistiquement (p = 0,15). À l'examen du tableau précédent, on voit que trois élèves ont obtenu un score moins élevé en français qu'en anglais, tandis que les deux autres ont obtenu une note plus forte en français qu'en anglais. Ainsi, on comprend pourquoi le coefficient de corrélation n'est pas significatif, même s'il est de 0,75.

Les étapes à suivre pour connaître le pourcentage de variance d'une variable expliquée par une autre variable se résument ainsi :

1. Construire un tableau similaire au tableau qui précède.
2. Calculer la moyenne et l'écart-type de chaque distribution.
3. Soustraire chaque score de sa moyenne.
4. Pour chaque individu, multiplier les deux différences.
5. À partir des deux formules qui précèdent, calculer la covariance, puis le coefficient de corrélation.

? 12. Foire aux questions (FAQ) quant à la variance et à l'écart-type ?

Q1	**La moyenne d'une distribution est-elle plus importante que l'écart-type ?**
R1	**Fausse question!** La moyenne est une mesure de tendance centrale, alors que l'écart-type reflète la dispersion des scores autour de la moyenne. Il est aussi facile de répondre à cette question que de répondre à la question suivante : les roues avant de votre auto sont-elles plus nécessaires que les roues arrière ?

Q2	**En augmentant la variabilité dans une distribution de scores, la corrélation de cette distribution avec une autre distribution va-t-elle augmenter par le fait même ?**
R2	Non, pas nécessairement. Pourquoi ? Le coefficient de corrélation varie selon la relation que les scores d'une distribution entretiennent avec celle d'une autre distribution, et non selon leur grandeur comme telle. En d'autres termes, pour chaque individu, c'est la relation entre son score sur la variable X et son score sur la variable Y qui fait varier le coefficient de corrélation. Il ne suffit donc pas d'accroître la variabilité pour produire une augmentation significative dans le coefficient de corrélation. D'ailleurs, nous pouvons obtenir un coefficient de corrélation élevé ou parfait entre deux variables dichotomiques (coefficient de contingence).

Q3	**Pourquoi le concept de variance est-il important en recherche ?**
R3	Si les phénomènes ne variaient pas, il n'y aurait que des constantes. L'univers entier se retrouverait dans un état statique, immuable. En fait, existerait-il vraiment ? Car, pour exister, il faut passer d'un état à un autre, ce qui implique de la variabilité. Pour évoluer, c'est pareil. Revenons sur terre. La variabilité d'un phénomène est intéressante à connaître surtout pour savoir à quel point il est relié à un ou plusieurs autres phénomènes. Cette variabilité, dans les études corrélationnelles par exemple, nous permet surtout de savoir à quel point une variable est associée à une autre variable. Par la suite, on peut établir des relations de toutes sortes entre des variables, et même les organiser en un système conceptuel qui nous aide à comprendre un phénomène dans toute sa complexité. La plupart des analyses statistiques font appel au concept de variance d'une manière plus ou moins transparente. Un grand nombre d'analyses statistiques ont, comme point de départ, la corrélation.

Q4	**Une échelle de mesure plus longue est-elle meilleure qu'une échelle plus courte ?**
R4	Oui, si elle est plus précise. Comme elle est plus longue, elle risque fort d'être plus précise. Que signifie vraiment *plus longue*, ou même *plus courte* ? *Plus longue* par rapport à quoi... ? Contentons-nous de dire qu'une échelle de mesure ordinale, par exemple, est plus précise qu'une échelle nominale. La meilleure échelle serait, en définitive, l'échelle qui pourrait refléter toutes les variations du phénomène à l'étude.

Q5	**Quelle caractéristique d'une distribution influence le plus les résultats d'analyses statistiques faisant appel à cette distribution ?**
R5	Plusieurs caractéristiques d'une distribution peuvent influencer les résultats d'analyses statistiques. C'est surtout le degré de sévérité de cette caractéristique qui détermine à quel point les résultats statistiques seront influencés par elle. Cependant, en analysant des résultats, quelques caractéristiques sont à surveiller parce qu'elles arrivent souvent à l'improviste. L'une d'elles c'est la présence de scores extrêmes (en anglais, *outliers*). Leurs effets sur la plupart des analyses statistiques sont dits aberrants. Une autre caractéristique lourde de conséquences, c'est la présence de deux modes dans une distribution : on parle alors d'une distribution bimodale.

Q6	**Pour calculer la variance d'une distribution, pourquoi porte-t-on au carré la différence entre chaque score et la moyenne ?**
R6	Pour parvenir à la variance, on calcule la différence de chaque score avec la moyenne de la distribution. La somme de ces scores est toujours égale à zéro (la somme des + égale la somme des -). Afin de contourner cet obstacle, le résultat de chaque soustraction est multiplié par lui-même, ce qui maintient, par ailleurs, la relation de grandeur observée entre les scores. Une fois que tous les scores sont devenus positifs, on extraie la racine carrée de cette somme pour obtenir l'écart-type.

Q7	**Quelle différence y a-t-il entre l'écart-type et la variance d'une distribution ?**
R7	La variance se définit comme la somme de tous les écarts à la moyenne sans égard à la valeur du signe. La variance constitue, en quelque sorte, la *moyenne* des écarts. L'écart-type est la racine carrée de la variance.

Q8	**Dans un petit échantillon, l'écart-type est-il plus petit que celui d'un grand échantillon ?**
R8	L'écart-type dépend non pas du nombre d'observations ou de scores, mais de l'étendue des scores par rapport à la moyenne.

Q9	**À la lumière de l'écart-type, comment peut-on dire que les scores d'une distribution sont plus fortement ou faiblement dispersés que ceux d'une autre distribution ?**
R9	Pour bien faire, il faut transformer les scores bruts en scores normalisés, ce qui permet l'application d'une mesure universelle, d'un commun dénominateur. Ainsi, il est possible de comparer les écarts-types de plusieurs distributions. Faute de mieux, on peut tout de même juger de l'étendue des scores pour chaque distribution isolément : après tout, l'écart-type reflète justement la dispersion des scores. Chaque écart-type est lié à l'échelle de mesure qui l'a produit. Un écart-type de 2,5 est plutôt grand si l'échelle varie de 1 à 4. Le même écart-type de 2,5 indiquerait une forte concentration des scores autour de la moyenne si l'échelle allait de 1 à 10. Si l'écart-type est aussi grand que la moyenne, alors... Il est utile, voire nécessaire, pour bien saisir une distribution de scores, d'en obtenir un histogramme. Les logiciels statistiques fournissent, évidemment, la description d'une distribution sous diverses formes, en même temps que bien des valeurs statistiques importantes : mode, médiane, variance, écart-type, kurtose, et d'autres. Deux valeurs nous aident à juger à quel point une distribution déroge de la normalité : la kurtose et l'assymétrie.

Q10	**Comment une transformation de scores influence-t-elle une distribution ?**
R10	La position relative de chacun des scores de la distribution reste la même. Cependant, la moyenne change, ainsi que l'écart-type. Ce qui n'est pas le cas lorsqu' on multiplie, par exemple, chaque score par une constante. En ajoutant une constante à une distribution de scores, on ne modifie pas la relation entre les scores, seulement la grandeur des scores. Ferguson (1966, p.71) fait la démonstration mathématique de l'effet de l'ajout d'une constante aux scores d'une distribution et affirme que “si toutes les mesures contenues dans un échantillon sont multipliées par une constante, l'écart-type est lui aussi multiplié par la valeur de cette constante”.

Q11	**Quel effet une distribution qui déroge fortement à la normalité peut-elle avoir sur un coefficient de corrélation ?**
R11	Terrible ! Terrible ! Terrible !

Q12	**Est-ce préférable de mesurer, par exemple, le revenu des particuliers en leur demandant d'indiquer leur revenu exact ou la catégorie de salaire à laquelle ils appartiennent ?**
R12	Plus l'échelle de mesure est longue, plus il y a de variabilité dans la distribution. Tout découpage d'une distribution continue de scores ne fait que réduire sa variabilité. Une variable continue est préférable à une variable discrète, et une variable discrète est préférable à une variable dichotomique. De plus, si vous tenez à présenter dans votre rapport de recherche des catégories de salaires, vous pouvez toujours le faire à partir d'une distribution continue, une fois vos données recueillies. Par contre, l'inverse n'est pas possible. Le motif principal pour lequel il faut allonger autant que possible une échelle de mesure, c'est pour gagner le maximum de précision, c'est-à-dire se rapprocher le plus près possible du *comportement* que nous voulons définir ou mesurer. Dans les sciences sociales ou en éducation, lorsqu'on agrandit l'échelle de mesure, on augmente la plupart du temps la variabilité. Cette augmentation n'est cependant pas automatique.

Q13	**Les distributions avec de *grosses* moyennes engendrent-elles de “gros” coefficients de corrélation ?**
R13	Le coefficient de corrélation est sensible à la dispersion des scores autour de la moyenne, et non à la taille de la moyenne.

Q14	**En éliminant les individus d'un groupe ayant des scores faibles, est-ce que le coefficient de corrélation va augmenter ou diminuer par rapport à celui où tous les individus sont considérés dans le calcul ?**
R14	Il va diminuer parce que la variabilité dans les scores a été diminuée. En éliminant des individus d'un groupe, on met en branle un processus de *régression vers la moyenne*. La régression vers la moyenne se produit lorsque les scores extrêmes tendent à revenir vers la moyenne après des évaluations subséquentes, diminuant ainsi la variabilité. En éliminant systématiquement des scores extrêmes dans une distribution, on provoque une sorte de régression vers la moyenne.

13. Je fais mes exercices

1. L'écart-type est une mesure de dispersion des scores.

O Vrai O Faux

2. La variance est un concept indépendant de l'écart-type.

O Vrai O Faux

3. Un score extrême dans une distribution n'influence pas la moyenne de cette distribution.

O Vrai O Faux

4. La moyenne représente bien une distribution normale de scores.

O Vrai O Faux

5. L'écart-type est, en quelque sorte, une moyenne des écarts de chaque score à la moyenne de la distribution.

O Vrai O Faux

6. Est-ce possible de calculer la variance des scores pour une seule variable ?

O Oui O Non

7. Laquelle des deux distributions suivantes possède le plus grand écart-type ?

A	3 9 8 2 1 5 7 8
B	132 129 130 131 132 130 131

8. La corrélation est un indicateur d'association entre deux variables.

O Vrai O Faux

9. Le coefficient de corrélation est plus sensible au rapport entre les deux variables qu'à la variabilité de chacune prise individuellement.

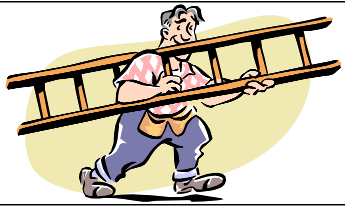
O Vrai O Faux

10. Un coefficient de corrélation négatif indique qu'il n'y a pas de relation entre les variables.

O Vrai O Faux

11. Une corrélation positive exprime une association toujours plus forte entre deux variables qu'une corrélation négative.

O Vrai O Faux

	12. Pour obtenir un coefficient de corrélation plus élevé, il est parfois avantageux d'utiliser une échelle de mesure longue plutôt que courte.	O Vrai O Faux

13. Le coefficient de détermination exprime la variance partagée par deux variables.

O Vrai O Faux

14. Pour calculer le coefficient de détermination, il s'agit de mettre au carré le coefficient de corrélation.

O Vrai O Faux

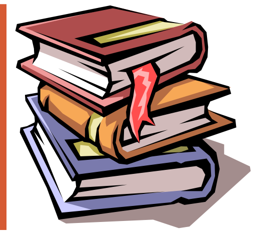
15. En tirant la racine carrée du coefficient de détermination, on obtient le coefficient de corrélation.

O Vrai O Faux

16. Plus le coefficient de corrélation s'approche de zéro (0), moins le pourcentage de variance expliquée est élevé.

O Vrai O Faux

14. Sources



Angers, Maurice (1996). *Initiation pratique à la méthodologie des sciences humaines*. Les Éditions CEC Inc., 2^{e} édition.

Dayhaw, Laurence T. (1963). *Manuel de statistique*. Éditions de l'Université d'Ottawa. 2^{e} édition.

Ferguson, George A. (1966). *Statistical Analysis in Psychology and Education*. McGraw-Hill Book Co., 2nd Edition.

Fortin, Marie-Fabienne (1996). *Le processus de la recherche*. Décarie Éditeur Inc.

Hayduk, Leslie A. (1987). *Structural Equation Modeling with Lisrel*. The John Hopkins University Press, Baltimore.

Hair, Joseph F. Jr., Rolph E. Anderson et Ronald L. Tatham (1987). *Multivariate Data Analysis*. 2nd Edition. Macmillan Publishing Company, New York.

Hays, William L. (1963). *Statistics for Psychologists*. Holt, Rinehart and Winston.

McCall, Chester H. Jr. (1982). *Sampling and Statistics Handbook for Research*. 1^{st} Edition. The Iowa State University Press, Ames, Iowa 50010.

Nunnally, Jim C. (1978). *Psychometric Theory*. McGraw-Hill Book Co., 2nd Edition.

Schumacker, Randall E., et Richard G. Lomax (1996). *A Beginner's Guide to Structural Equation Modeling*. Lawrence Erlbaum Associates. Mahwah, New Jersey.

Shavelson, Richard J. *Statistical Reasoning for the Behavioural Sciences.* Allyn & Bacon, Needham Heights, Massachusetts 02194.

Tabachnik, Barbara. G. et Linda S. Fidell (1983). *Using Multivariate Statistics*. New York, Harper & Row Publishers.