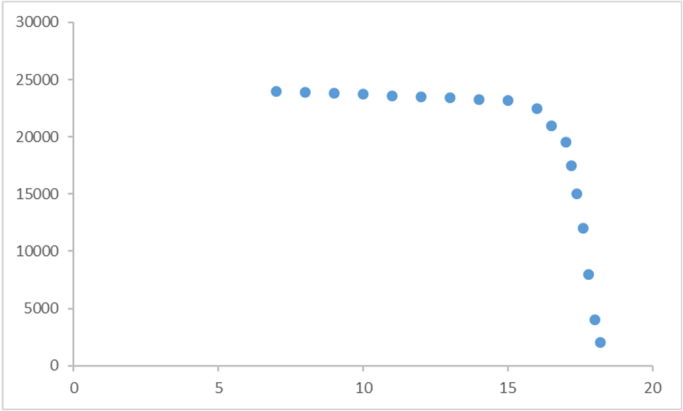
TD5 : Paramètres d'association (covariance suite, coefficient de corrélation et variance expliquée)

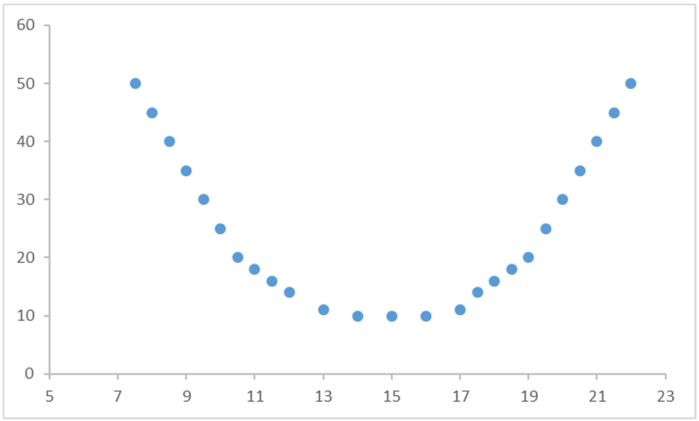
Questions de réflexion

NB : dans les corrections de ces questions de réflexion, à la différence des notations simplifiées mx et my ou X et Y que nous avons utilisées dans les corrections ou dans les énoncés de TD pour symboliser la moyenne de la variable X ou de la variable Y, nous utiliserons maintenant les notations scientifiques de la moyenne : $\overline{x}$ et $\overline{y}$.

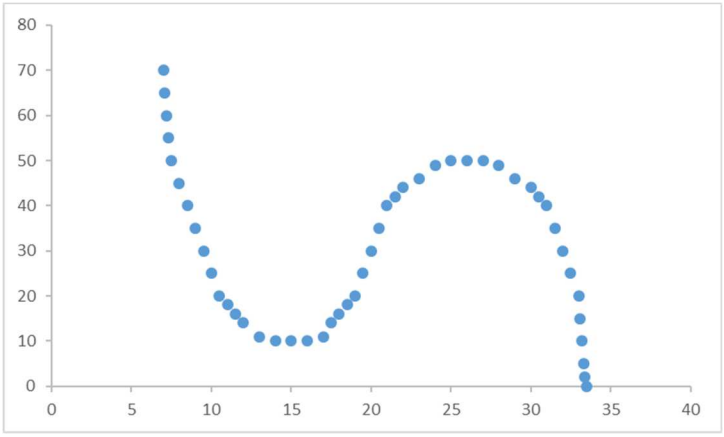
1/ Pour le premier diagramme de dispersion représenté ci-dessous, nous pouvons nous attendre à obtenir un coefficient de corrélation négatif. En effet, même si la moitié de la courbe est plutôt plate, l'autre moitié est en revanche très décroissante, ce qui suggère une relation inverse entre les deux variables sur cette partie de la courbe. Le cumul de la partie plate (qui conduira plutôt à des produits des écarts soit positifs soit négatifs selon la position des valeurs par rapport aux moyennes des variables X et Y) et de la partie décroissante de la courbe (qui conduira plutôt à des produits des écarts très négatifs) donnera au final un **coefficient de corrélation plutôt négatif**. D'autre part, nous pouvons nous attendre à une valeur absolue élevée pour ce coefficient de corrélation, étant donné que la pente décroissante de la courbe est très forte.



Pour le deuxième diagramme de dispersion ci-dessous, nous pouvons nous attendre à un coefficient de corrélation proche de 0. En effet, on remarque que la courbe (que l'on peut qualifier de quadratique ou parabolique) comprend une partie décroissante et une partie croissante qui sont symétriques. La partie décroissante va donc correspondre à des produits des écarts négatifs (relation inverse entre les variables) alors que la partie croissante va correspondre à des produits des écarts positifs (relation positive entre les variables). Le fait d'additionner des produits des écarts de signe opposé va nous conduire à un **coefficient de corrélation proche de 0**.



Le dernier diagramme de dispersion ci-dessous va également conduire à l'obtention d'un coefficient de corrélation proche de 0. Comme pour le deuxième diagramme, on remarque que la courbe comprend des parties décroissantes et des parties croissantes symétriques. Les parties décroissantes correspondant à des produits des écarts négatifs, et les parties croissantes correspondant à des produits des écarts positifs, la somme de ces produits des écarts de signe opposé va également nous conduire à un **coefficient de corrélation proche de 0**.



Au regard de ces différents diagrammes de dispersion, nous pouvons dire que **le coefficient de corrélation n'est pas particulièrement adapté pour rendre compte des relations de type monotone non linéaire ou non monotone non linéaire pouvant exister entre deux variables.**

Le premier diagramme de dispersion, correspondant à une relation de type monotone non linéaire, conduit à un coefficient de corrélation négatif. Ce coefficient permet de rendre compte de la décroissance de la courbe, mais la partie plate de cette même courbe n'est en revanche pas représentée par ce coefficient négatif.

Les second et troisième diagrammes de dispersion, correspondant à des relations de type non monotone non linéaire, semblent indiquer l'existence d'une relation entre les deux variables. Cependant, cette relation ne peut être mise en évidence par le coefficient de corrélation, étant donné que les parties décroissantes et croissantes de ces deux courbes vont se compenser et conduire à l'obtention d'un coefficient proche de 0.

Le coefficient de corrélation de Bravais-Pearson est donc adapté pour déterminer l'existence d'une relation linéaire entre des variables, mais pas pour les autres types de relation.

2/ Dans le cadre d'une relation linéaire négative entre deux variables X et Y, si (xi – $\bar{x}$) est négatif, nous pouvons nous attendre à une valeur de (yi – $\bar{y}$) positive. En effet, le fait qu'une relation linéaire négative existe entre deux variables implique que des valeurs faibles sur la variable X (en-dessous de la moyenne) soient associées à des valeurs élevées sur la variable Y (au-dessus de la moyenne). En retour, nous pouvons nous attendre à ce qu'un (xi – $\bar{x}$) positif soit associé à une valeur de (yi – $\bar{y}$) négative. Pour résumer, **dans les relations linéaires négatives, (xi – $\bar{x}$) et (yi – $\bar{y}$) sont généralement de signe opposé, ce qui conduit à des valeurs de la covariance et du coefficient de corrélation négatives**.

Dans le cadre d'une relation linéaire positive entre deux variables X et Y, si (xi – $\bar{x}$) est négatif, cette fois nous pouvons nous attendre à une valeur de (yi – $\bar{y}$) également négative. En effet, le fait qu'une relation linéaire positive existe entre deux variables implique que des valeurs faibles sur la variable X (en dessous de la moyenne) soient associées à des valeurs également faibles sur la variable Y (en dessous de la moyenne). Nous pouvons nous attendre en retour à ce qu'un (xi – $\bar{x}$) positif soit associé à une valeur

également positive de (yi – $\bar{y}$). Pour résumer, **dans les relations linéaires positives, (xi – $\bar{x}$) et (yi – $\bar{y}$) sont généralement de même signe, ce qui conduit à des valeurs de la covariance et du coefficient de corrélation positives**.

Dans le cadre d'une relation linéaire nulle entre deux variables X et Y, si (xi – $\bar{x}$) est négatif, nous pouvons nous attendre à ce que (yi – $\bar{y}$) soit tantôt négatif, tantôt positif. En effet, lorsqu'il y a une absence de relation linéaire entre deux variables, cela signifie que nous ne pouvons pas prédire les valeurs de la variable Y en fonction des valeurs de la variable X. Un (xi – $\bar{x}$) positif sera donc également tantôt associé à un (yi – $\bar{y}$) positif, et tantôt à un (yi – $\bar{y}$) négatif. Pour résumer, **en cas d'absence de relation linéaire, dans certains cas (xi – $\bar{x}$) et (yi – $\bar{y}$) seront de même signe, et dans d'autre cas de signe opposé, ce qui conduira à une somme des produits des écarts proche de 0**.

3/ Les bornes du coefficient de corrélation sont -1 et +1. Il n'est pas possible d'obtenir un coefficient de corrélation en dehors de ces limites. Par conséquent**, un résultat de -2.6 pour le coefficient provient d'une erreur de calcul**.

4/ Seule la valeur absolue du coefficient de corrélation indique l'intensité de la relation, et non le signe du coefficient. La valeur absolue 0.76 étant plus élevée que la valeur absolue 0.48, **-0.76 est le coefficient reflétant la relation la plus intense**.

5/ La variance expliquée correspond au carré du coefficient de corrélation : $0.99^2 = 0.98$. En d'autres termes, 98% de la variance du nombre de personnes décédées en tombant dans les escaliers entre 2007 et 2010 peuvent être expliqués par le nombre d'iphones vendus durant cette période. Ce résultat signifie que **le nombre d'iphones vendus entre 2007 et 2010 permet très efficacement de <u>prédire</u> le nombre de personnes décédées en tombant dans les escaliers durant cette période**. Cela ne veut par contre nullement dire que les ventes d'iphones causent les décès par chute dans les escaliers. La capacité de prédire n'est pas synonyme de causalité.

6/ Un coefficient de corrélation de 0 signifie une absence de relation linéaire entre les deux variables. Nous ne pouvons par contre aucunement conclure que les deux variables sont indépendantes, puisqu'il peut exister entre elles un autre type de relation que nous ne pouvons pas identifier grâce au coefficient de corrélation.

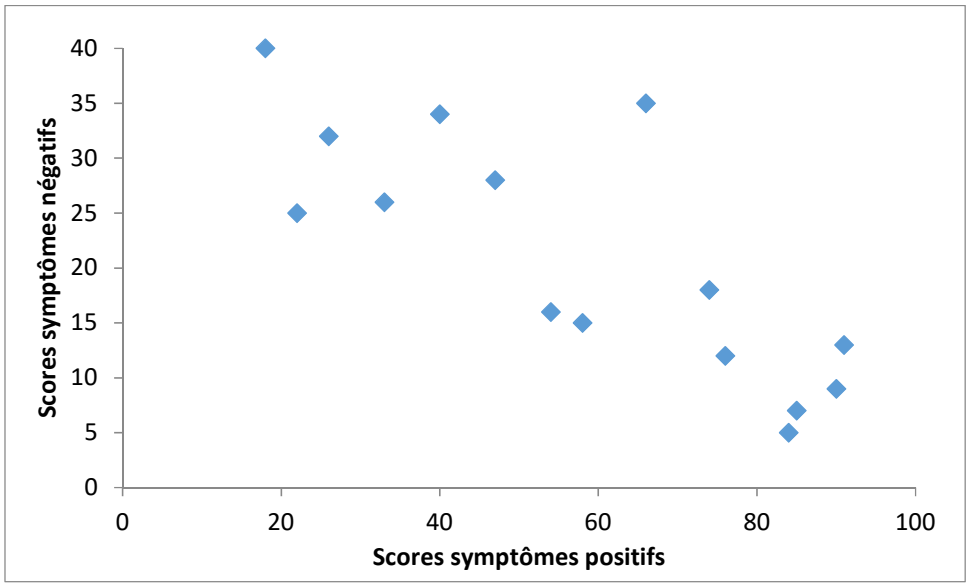
Exercices

A) Pour rappel, nous souhaitions généraliser nos résultats à la population à partir des données que nous avons recueillies.

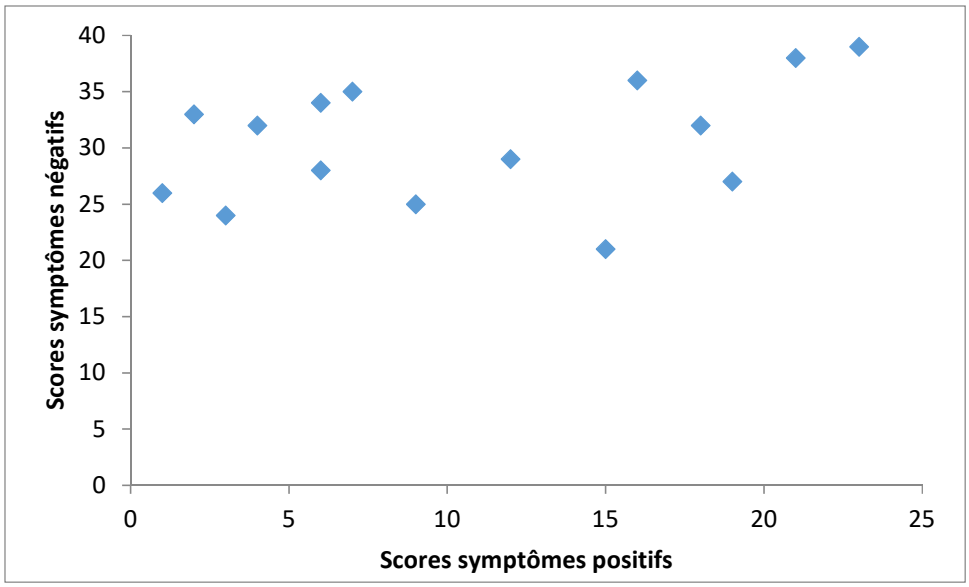
NB : dans cet exercice, à la différence des notations simplifiées mx et my ou X et Y que nous avons parfois utilisées dans les énoncés corrections pour symboliser la moyenne de la variable X ou de la variable Y, nous utiliserons maintenant les notations scientifiques de la moyenne $\overline{x}$ et $\overline{y}$.

Pour rappel, les moyennes des variables X et Y pour le groupe psychotique étaient respectivement de 57.6 et 21, et de 10.8 et 30.6 pour le groupe dépressif.

1/ Voici le diagramme de dispersion pour les patients psychotiques :



Et le diagramme de dispersion pour les patients dépressifs :



(NB : si vous avez inversé l'ordre des variables, cela est également correct)

2/ La covariance entre les variables symptômes positifs et négatifs est de -229.43 pour les patients psychotiques, et de 14.27 pour les patients dépressifs. Ci-après, vous retrouverez les étapes de calcul de cette covariance pour les patients psychotiques, la procédure à suivre étant la même pour les patients dépressifs (pour plus de facilités, nous avons renommé la variable Symptômes positifs X et la variable Symptômes négatifs Y).

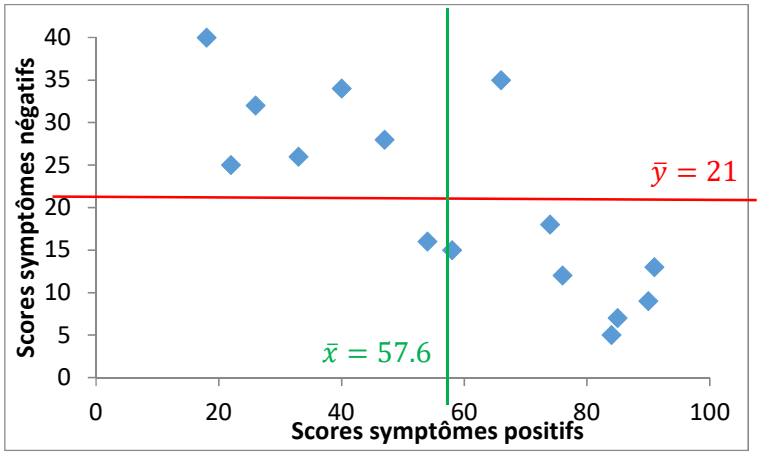
La moyenne du groupe pour la variable X est de 57.6, et elle est de 21 pour la variable Y. Dans les quatrième et cinquième colonnes du tableau, on calcule les écarts à la moyenne pour chaque variable. Par exemple, pour le second patient, 47 (x1) – 57,6 (moyenne de x, ou $\bar{x}$) va nous donner un résultat de -10.6, qui est reporté dans la première case de la quatrième colonne.

L'étape suivante est le calcul des produits des écarts. Pour chaque sujet, deux valeurs ont été observées (une sur chaque variable), et nous allons calculer le produit des écarts à la moyenne de ces valeurs. Par exemple, le cinquième patient qui a un score de symptômes positifs de 74 et un score de symptômes négatifs de 18 a des écarts correspondants de 16.4 et -3. Ainsi, le produit des écarts que nous allons calculer pour ce patient est (16.4*-3), soit approximativement -49.2, valeur reportée dans la cinquième case de la sixième colonne.

Numéro du patient	Score symptômes positifs (X)	Score symptômes négatifs (Y)	xi-$\bar{x}$	yi-$\bar{y}$	(xi-$\bar{x}$)*(yi-$\bar{y}$)
1	76	12	18,4	-9	-165,6
2	47	28	-10,6	7	-74,2
3	85	7	27,4	-14	-383,6
4	22	25	-35,6	4	-142,4
5	74	18	16,4	-3	-49,2
6	66	35	8,4	14	117,6
7	26	32	-31,6	11	-347,6
8	90	9	32,4	-12	-388,8
9	33	26	-24,6	5	-123
10	58	15	0,4	-6	-2,4
11	40	34	-17,6	13	-228,8
12	91	13	33,4	-8	-267,2
13	18	40	-39,6	19	-752,4
14	54	16	-3,6	-5	18
15	84	5	26,4	-16	-422,4

Pour l'étape finale du calcul de la covariance, il nous suffit de diviser la somme de la sixième colonne par (n – 1), soit par 14 (15 – 1). En effet, notre but ici est de généraliser les résultats obtenus sur ces 15 patients à l'ensemble de la population des patients psychotiques dont ils sont issus. Par conséquent, **nous nous situons en statistiques inférentielles et cherchons à estimer les paramètres de la population dont provient l'échantillon. Dans ces conditions, le calcul de la covariance est effectué sur (n – 1)**, et non sur n. Cette division de la somme des produits des écarts à la moyenne (soit la division de la somme de la sixième colonne) par 14 nous renvoie la valeur -229.43.

3/ Voici le diagramme de dispersion des patients psychotiques avec les droites des moyennes représentées.



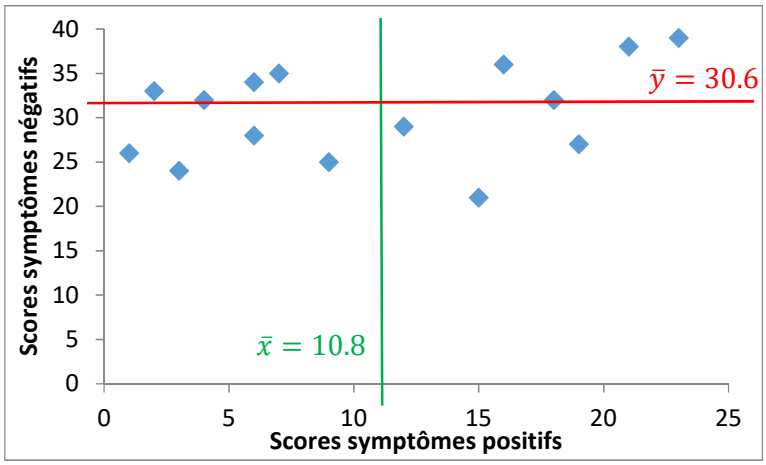
Si nous observons ce diagramme, nous pouvons constater que les points se répartissent majoritairement dans deux des quatre cadrans délimités par le croisement des droites des moyennes.

Le premier cadran, en haut à gauche, contient les résultats de participants dont les produits des écarts correspondants sont négatifs. En effet, les participants situés dans ce cadran ont des scores de symptômes positifs (X) inférieurs à la moyenne, soit un écart (xi - $\bar{x}$) négatif, et des scores de symptômes négatifs (Y) supérieurs à la moyenne, soit un écart (yi - $\bar{y}$) positif. Le produit de ces deux écarts (négatif * positif) nous donnera donc un résultat négatif.

Le second cadran où se trouvent majoritairement les points, en bas à droite, contient les résultats de participants dont les produits des écarts correspondants seront également négatifs. En effet, les participants situés dans ce cadran ont des scores de symptômes positifs (X) supérieurs à la moyenne, soit un écart (xi - $\bar{x}$) positif, et des scores de symptômes négatifs (Y) inférieurs à la moyenne, soit un écart (yi - $\bar{y}$) négatif. Le produit de ces deux écarts (positif * négatif) nous donnera également un résultat négatif.

➔ **Le diagramme de dispersion et la représentation des moyennes nous indiquent donc que le calcul de la covariance du groupe psychotique nous conduit à additionner des produits des écarts essentiellement négatifs, ce qui explique la valeur négative de la covariance que nous avons obtenue (-229.43).**

Voici le diagramme de dispersion des patients dépressifs avec les droites des moyennes représentées :



Si nous observons ce diagramme, nous pouvons constater que les points se répartissent équitablement dans les quatre cadrans délimités par le croisement des droites des moyennes.

Nous avons dit que les cadrans en haut à gauche et en bas à droite correspondaient à des produits des écarts négatifs. A l'inverse, les cadrans en haut à droite ((xi - $\bar{x}$) et (yi - $\bar{y}$) positifs) et en bas à gauche ((xi - $\bar{x}$) et (yi - $\bar{y}$) négatifs) correspondent tous deux à des produits des écarts positifs.

➔ **Le diagramme de dispersion et la représentation des moyennes nous indiquent que le calcul de la covariance pour le groupe dépressif nous conduit à additionner des produits des écarts négatifs et positifs, ce qui explique la valeur de la covariance a priori peu élevée (en comparaison de celle des patients psychotiques) que nous avons obtenue (14.27).**

4/ Pour calculer le coefficient de corrélation, il nous suffit de diviser la covariance par le produit des écarts-types des deux variables.

Nous avons déjà calculé les écarts-types des deux variables pour les deux groupes dans l'exercice supplémentaire B du TD3 (je vous invite à vous reporter au corrigé de cet exercice pour la méthode).

Les valeurs des écarts-types des variables X et Y pour le groupe des patients psychotiques sont respectivement de 25.59 et 11.17. Par conséquent, le coefficient de corrélation de ce groupe est :

-229.43 / (25.59 * 11.17) = -0.80.

➔ **Ce coefficient de corrélation à la fois négatif et élevé nous indique qu'il existe une relation importante et négative entre les variables X et Y chez les patients psychotiques. Cela signifie que plus le nombre de symptômes positifs augmente chez ces patients, plus le nombre de symptômes négatifs diminue, et inversement.**

Les valeurs des écarts-types des variables X et Y pour le groupe des patients dépressifs sont respectivement de 7.38 et 5.38. Par conséquent, le coefficient de corrélation de ce groupe est :

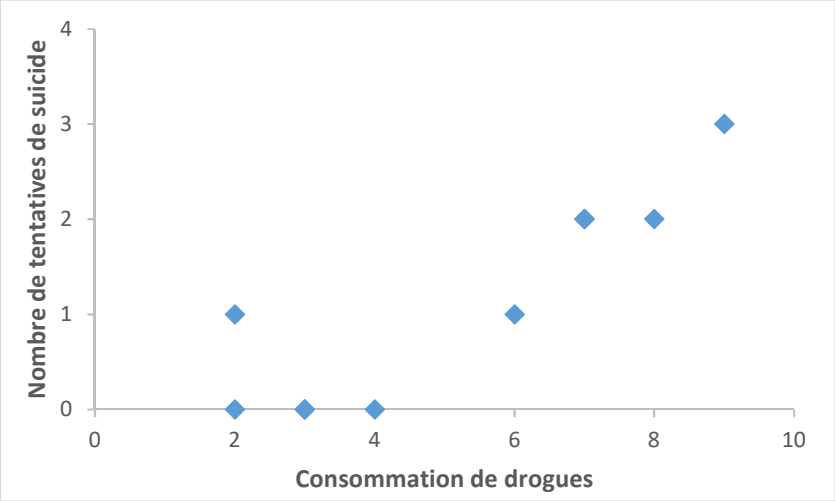
14.27 / (7.38 * 5.38) = 0.36.

➔ **Ce coefficient de corrélation à la fois positif et peu élevé nous indique qu'il existe une relation faible et positive entre les variables X et Y chez les patients dépressifs. Cela signifie que, modérément, plus le nombre de symptômes positifs augmente chez ces patients, plus le nombre de symptômes négatifs augmente également.**

B) Pour rappel, dans cette étude, nous ne souhaitions pas généraliser nos résultats à la population des adolescents souffrant de toxicomanie en général.

NB : dans cet exercice, à la différence des notations simplifiées mx et my ou X et Y que nous avons parfois utilisées dans les énoncés et corrections pour symboliser la moyenne de la variable X ou de la variable Y, nous utiliserons maintenant les notations scientifiques de la moyenne $\overline{x}$ et $\overline{y}$.

1/ Voici le diagramme de dispersion représentant les données des adolescents souffrant de toxicomanie sur les variables Consommation de drogues et Nombre de tentatives de suicide (NB : si vous avez inversé l'ordre des variables, cela est également correct. Cependant, il est plus logique de placer la Consommation de drogues en abscisse et le Nombre de tentatives de suicide en ordonnée, car nous nous attendons davantage à ce que la consommation de drogues permette de prédire le nombre de tentatives de suicide plutôt que l'inverse.)



2/ La variance sur la variable Consommation de drogues est de 6.09 et l'écart-type de 2.47. Sur la variable Tentatives de suicide, la variance est de 1.09, et l'écart-type de 1.04. Ci-après, vous retrouverez les étapes de calcul de la variance pour les deux variables (pour plus de facilités, nous avons renommé la variable Consommation de drogues X et la variable Nombre de tentatives de suicide Y).

La moyenne sur la variable Consommation de drogues est de 5.1, et de 1.1 sur la variable Tentatives de suicide. Dans les troisièmes colonnes de chaque tableau ci-dessous, on calcule les écarts à la moyenne pour chaque donnée de la deuxième colonne. Par exemple, l'écart à la moyenne de X pour le patient n°6, 9 (x1) – 5.1 (moyenne de x, ou $\bar{x}$), va nous donner un résultat de 3.9, qui est reporté dans la sixième case de la troisième colonne.

Dans la quatrième colonne on met les écarts à la moyenne de la troisième colonne au carré. Par exemple, toujours pour le patient n°6 et pour la variable X, 3.9^2 va nous donner un résultat de 15.21, qui est reporté dans la sixième case de la quatrième colonne.

Numéro du patient	Conso de drogues (X)	$(xi - \bar{x})$	$(xi - \bar{x})^2$
1	3	-2,1	4,41
2	7	1,9	3,61
3	6	0,9	0,81
4	8	2,9	8,41
5	2	-3,1	9,61
6	9	3,9	15,21
7	4	-1,1	1,21
8	3	-2,1	4,41
9	7	1,9	3,61
10	2	-3,1	9,61

Numéro du patient	Tentatives de suicide (Y)	$(yi - \bar{y})$	$(yi - \bar{y})^2$
1	0	-1,1	1,21
2	2	0,9	0,81
3	1	-0,1	0,01
4	2	0,9	0,81
5	1	-0,1	0,01
6	3	1,9	3,61
7	0	-1,1	1,21
8	0	-1,1	1,21
9	2	0,9	0,81
10	0	-1,1	1,21

Pour le calcul de la variance, il nous suffit de faire la moyenne de la quatrième colonne pour chacune des deux variables. En effet, nous avons dit dans l'énoncé que nous nous intéressions uniquement aux sujets que nous avons étudiés. En d'autres termes, **nous sommes en statistiques descriptives et n'avons pas pour but de généraliser nos résultats à la population, ce qui signifie que nous calculons la variance sur n.** Ainsi, pour calculer la variance il nous suffit de faire la moyenne des carrés des écarts à la moyenne (soit la moyenne de la quatrième colonne), ce qui nous renvoie la valeur 6.09 pour la variable X (consommation de drogues), et 1.09 pour la variable Y (tentatives de suicide). Les racines carrées de ces valeurs, les écarts-types, sont respectivement 2.47 et 1.04.

3/ La covariance des variables Consommation de drogues et Tentatives de suicide est de 2.29. Ci-après, vous retrouverez les étapes de calcul de cette covariance (nous reprenons nos appellations X pour la variable Consommation de drogues, et Y pour la variable Tentatives de suicide).

Etant donné que nous avons déjà calculé les variances des deux variables, et que nous savons qu'il est nécessaire de calculer les écarts à la moyenne sur les deux variables pour obtenir la covariance, nous pouvons réutiliser ces colonnes des écarts à la moyenne (colonnes 4 et 5 du tableau ci-dessous).

L'étape suivante est le calcul des produits des écarts. Pour chaque sujet, deux valeurs ont été observées (une sur chaque variable), et nous allons calculer le produit des écarts à la moyenne de ces valeurs. Par exemple, le patient n°4 qui a une consommation de drogues de 8 et un nombre de tentatives de suicide de 2 a des écarts correspondants de 2.9 et 0.9. Ainsi, le produit des écarts que nous allons calculer pour ce sujet est (2.9 * 0.9), soit 2.61, valeur reportée dans la quatrième case de la sixième colonne.

Numéro du patient	Conso de drogues (X)	Tentatives de suicide (Y)	(xi - $\bar{x}$)	(yi - $\bar{y}$)	(xi - $\bar{x}$)* (yi - $\bar{y}$)
1	3	0	-2,1	-1,1	2,31
2	7	2	1,9	0,9	1,71
3	6	1	0,9	-0,1	-0,09
4	8	2	2,9	0,9	2,61
5	2	1	-3,1	-0,1	0,31
6	9	3	3,9	1,9	7,41
7	4	0	-1,1	-1,1	1,21
8	3	0	-2,1	-1,1	2,31
9	7	2	1,9	0,9	1,71
10	2	0	-3,1	-1,1	3,41

Pour l'étape finale du calcul de la covariance, il nous suffit de faire la moyenne de la sixième colonne. Comme pour la variance, étant donné que nous nous intéressons uniquement aux données que nous avons recueillies, nous calculons également la covariance sur n. Ainsi, pour calculer la covariance il nous suffit de faire la moyenne des produits des écarts (soit la moyenne de la sixième colonne), ce qui nous renvoie la valeur 2.29.

4/ Pour calculer le coefficient de corrélation, il nous suffit de diviser la covariance par le produit des écarts-types des deux variables.

Comme calculées dans la question 2, 1es valeurs des écarts-types des variables X et Y sont respectivement de 2.47 et 1.04. Par conséquent, le coefficient de corrélation est :

2.29 / (2.47 * 1.04) = 0.89.

- ➔ **Ce coefficient de corrélation à la fois positif et élevé nous indique qu'il existe une relation importante et positive entre les variables X et Y. Cela signifie que plus la consommation de drogues augmente chez ces adolescents souffrant de toxicomanie, plus le nombre de tentatives de suicide augmente également.**

La variance expliquée correspond au carré du coefficient de corrélation : $0.89^2 = 0.79$. En d'autres termes, 79% de la variance du nombre de tentatives de suicide peuvent être expliqués par la consommation de drogues chez les adolescents souffrant de toxicomanie. Cela signifie que la consommation de drogues permet très efficacement de prédire le nombre de tentatives de suicide.