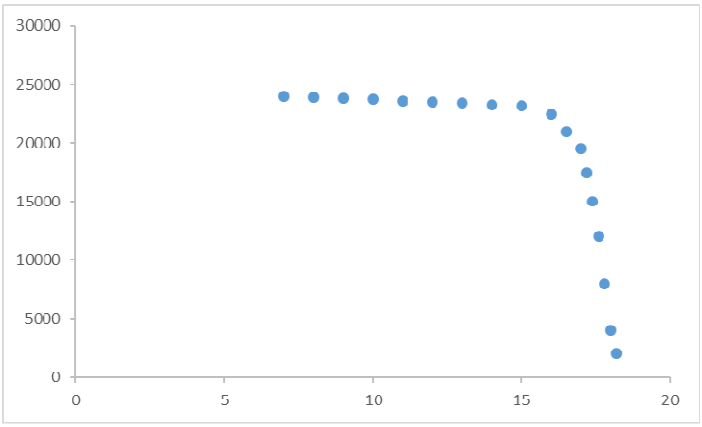
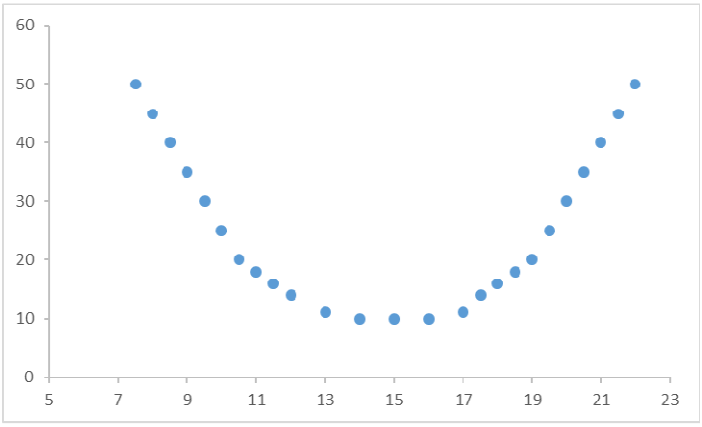
TD4 : Paramètres d'association (covariance)

Questions de réflexion

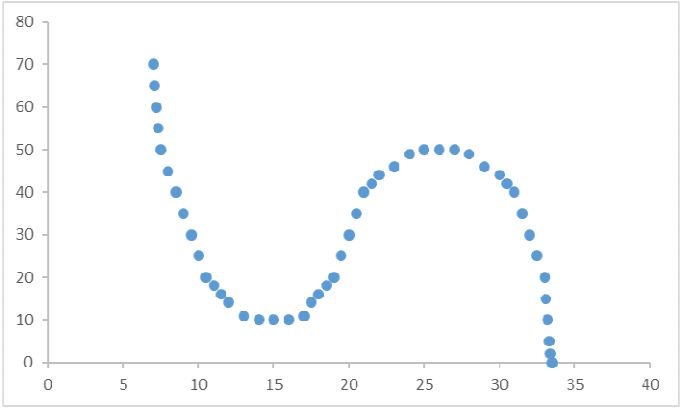
1/ Le type de relation ci-dessous peut être qualifié de **monotone non linéaire**. Monotone car la relation entre les variables est strictement décroissante, et non linéaire car la courbe ne correspond pas à une relation entre les variables de type Y=aX+b (en d'autres termes, ce n'est pas une droite).



Le type de relation ci-dessous est **non monotone non linéaire**. Non monotone car la courbe n'est pas strictement croissante ou décroissante, et non linéaire car la courbe ne correspond pas à une relation entre les variables de type Y=aX+b.



Le dernier type de relation ci-dessous est également **non monotone non linéaire**, pour les mêmes raisons que le deuxième type de relation ci-dessus.



2/ Une covariance d'une valeur de 231 959 183 entre deux variables ne nous dit pas grand-chose en soi. En effet, si cette covariance est issue de deux variables ayant des échelles allant de 0 à 200 000 par exemple et ayant une variabilité importante (c'est-à-dire, des écarts-types élevés), la covariance ne signifiera pas forcément une relation importante entre les variables. Par contre, si cette covariance est calculée sur des variables ayant des échelles plus réduites et une variabilité plus faible, la covariance pourra indiquer une forte relation entre les variables.

Cependant, **pour déterminer précisément l'intensité d'une relation unissant deux variables, il est nécessaire de passer par le calcul du coefficient de corrélation** (voir le TD5).

3/ Les étapes du calcul de la variance sont les suivantes :

- **Calcul de la moyenne sur la variable étudiée**
- **Calcul des écarts à la moyenne pour chacune des données**
- Calcul des carrés des écarts à la moyenne
- Somme des carrés des écarts à la moyenne obtenus à l'étape précédente
- Division de cette somme par n (statistiques descriptives) ou n – 1 (statistiques inférentielles)

Les étapes du calcul de la covariance sont les suivantes :

- **Calcul de la moyenne sur les deux variables étudiées**
- **Calcul des écarts à la moyenne pour chacune des données sur les deux variables**
- Calcul des produits des écarts pour chaque paire de données simultanées (obtenues sur un même sujet ou dans une même situation, par exemple le temps (variable 1) et le nombre d'erreurs (variable 2) sur une tâche pour le sujet 1)
- Somme des produits des écarts obtenus à l'étape précédente
- Division de cette somme par n (statistiques descriptives) ou n – 1 (statistiques inférentielles)

Les étapes communes aux deux calculs sont celles en gras, la seule différence étant que les calculs sont faits sur une seule variable à la fois pour la variance, alors qu'ils sont faits sur deux variables recueillies sur un même échantillon pour la covariance.

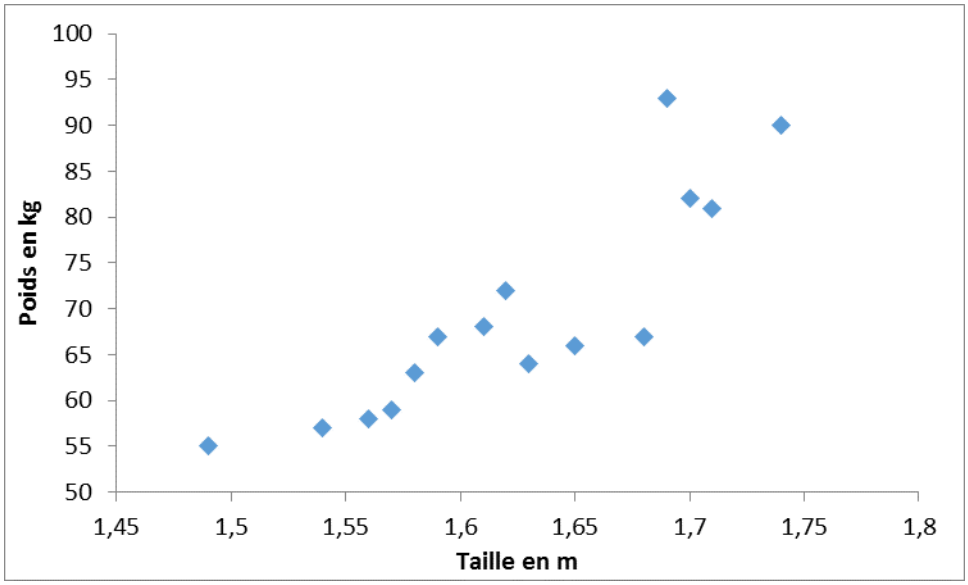
Etant donné ces étapes communes entre la variance et la covariance, il vous est vivement conseillé de construire au moins les colonnes des écarts à la moyenne pour les variables étudiées, afin de vous éviter d'avoir à faire plusieurs fois les mêmes calculs.

Notons également que la variance et la covariance ont en commun de se terminer par un rapport sur n ou sur n-1 en fonction du type de statistiques dans lequel on se situe (descriptif ou inférentiel). Retenez que cette distinction de calcul entre descriptif et inférentiel est propre à la variance et à la covariance. **Pour rappel, en statistiques descriptives on ne s'intéresse qu'aux données qui ont été recueillies sur notre échantillon, alors qu'en statistiques inférentielles on souhaite faire des inférences sur la population à partir des résultats recueillis sur notre échantillon.**

Exercices

A) Pour rappel, nous nous intéressions uniquement à notre groupe de patients, et n'avions pas pour but de généraliser nos résultats à la population dont ce groupe est issu.

1/ Voici le diagramme de dispersion des données (NB : si vous avez inversé l'ordre des variables, cela est également correct).



2/ La variance sur la variable taille est de 0.0047 et l'écart-type de 0.068. La variance sur la variable poids est de 131.72, et l'écart-type de 11.48. Ci-après, vous retrouverez les étapes de calcul de la variance pour les deux variables (pour plus de facilités, nous allons appeler la variable Taille X, et la variable Poids Y).

La moyenne du groupe pour la variable X est de 1.624, et elle est de 69.47 pour la variable Y. Dans les troisièmes colonnes de chaque tableau, 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°1, 1,65 (x1) – 1,624 (moyenne de x), va nous donner un résultat de 0,026, qui est reporté dans la première 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°1 et pour la variable X, $0{,}026^2$ va nous donner un résultat de 0,000676, qui est reporté dans la première case de la quatrième colonne.

Numéro du patient	Taille (X)	xi-mx	$(xi-mx)^2$
1	1,65	0,026	0,000676
2	1,7	0,076	0,005776
3	1,68	0,056	0,003136
4	1,74	0,116	0,013456
5	1,54	-0,084	0,007056
6	1,58	-0,044	0,001936
7	1,49	-0,134	0,017956
8	1,62	-0,004	1,6E-05
9	1,71	0,086	0,007396
10	1,56	-0,064	0,004096
11	1,59	-0,034	0,001156
12	1,63	0,006	3,6E-05
13	1,69	0,066	0,004356
14	1,61	-0,014	0,000196
15	1,57	-0,054	0,002916

Numéro du patient	Poids (Y)	yi-my	$(yi-my)^2$
1	66	-3,46666667	12,0177778
2	82	12,5333333	157,084444
3	67	-2,46666667	6,08444444
4	90	20,5333333	421,617778
5	57	-12,4666667	155,417778
6	63	-6,46666667	41,8177778
7	55	-14,4666667	209,284444
8	72	2,53333333	6,41777778
9	81	11,5333333	133,017778
10	58	-11,4666667	131,484444
11	67	-2,46666667	6,08444444
12	64	-5,46666667	29,8844444
13	93	23,5333333	553,817778
14	68	-1,46666667	2,15111111
15	59	-10,4666667	109,551111

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 0.0047 pour la variable X (taille), et 131.72 pour la variable Y (poids). Les racines carrées de ces valeurs, les écarts-type, sont respectivement 0.068 et 11.48.

3/ La covariance des variables taille et poids est de 0.69. Ci-après, vous retrouverez les étapes de calcul de cette covariance (nous reprenons nos appellations X pour la variable Taille, et Y pour la variable Poids).

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 sujet 1 qui a une taille de 1,65 m et un poids de 66 kg a des écarts correspondants de 0,026 et -3,47. Ainsi, le produit des écarts que nous allons calculer pour ce sujet est (0,026*-3,47), soit -0,09013, valeur reportée dans la première case de la sixième colonne.

Numéro du patient	Taille (X)	Poids (Y)	xi-mx	yi-my	(xi-mx)*(yi-my)
1	1,65	66	0,026	-3,46666667	-0,09013333
2	1,7	82	0,076	12,5333333	0,95253333
3	1,68	67	0,056	-2,46666667	-0,13813333
4	1,74	90	0,116	20,5333333	2,38186667
5	1,54	57	-0,084	-12,4666667	1,0472
6	1,58	63	-0,044	-6,46666667	0,28453333
7	1,49	55	-0,134	-14,4666667	1,93853333
8	1,62	72	-0,004	2,53333333	-0,01013333
9	1,71	81	0,086	11,5333333	0,99186667
10	1,56	58	-0,064	-11,4666667	0,73386667
11	1,59	67	-0,034	-2,46666667	0,08386667
12	1,63	64	0,006	-5,46666667	-0,0328
13	1,69	93	0,066	23,5333333	1,5532
14	1,61	68	-0,014	-1,46666667	0,02053333
15	1,57	59	-0,054	-10,4666667	0,5652

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 0.69.

➔ **La covariance étant dépendante de la variabilité au sein des deux variables dont elle étudie la relation, nous ne pouvons pas conclure sur l'intensité du lien existant entre les deux variables. Néanmoins, le diagramme de dispersion et le signe de la covariance nous indiquent que si une relation existe entre les variables taille et poids chez nos sujets, a priori elle est positive.**

4/ La covariance des variables taille (en cm cette fois) et poids est de 68.55. Pour la déterminer, il nous suffit de refaire la même démarche qu'à la question 3, en actualisant les données de la variable taille (X) dans le tableau, la moyenne sur la variable X étant maintenant de 162.4 :

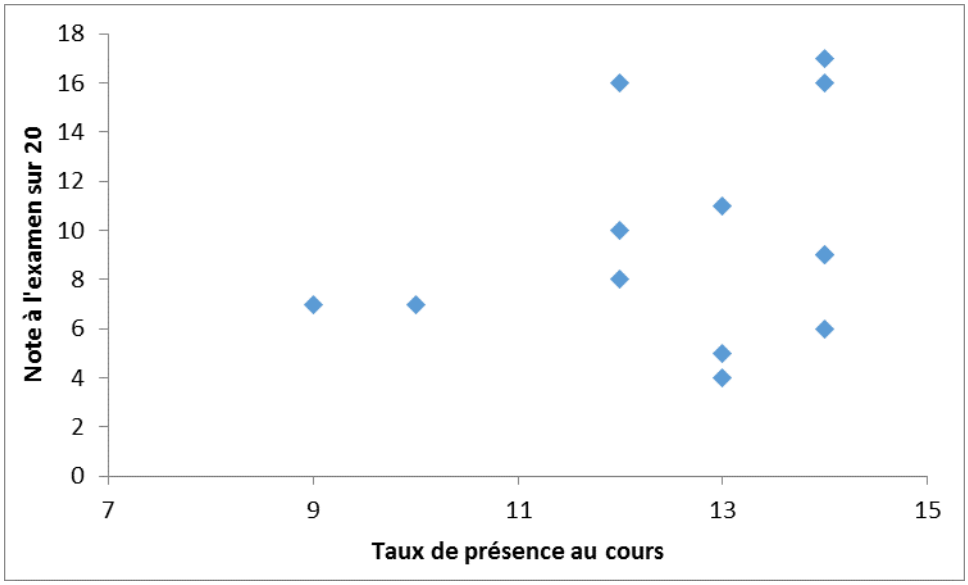
Numéro du patient	**Taille (X)**	Poids (Y)	**xi-mx**	yi-my	(xi-mx)*(yi-my)
1	**165**	66	**2,6**	-3,46666667	**-9,01333333**
2	**170**	82	**7,6**	12,5333333	**95,2533333**
3	**168**	67	**5,6**	-2,46666667	**-13,8133333**
4	**174**	90	**11,6**	20,5333333	**238,186667**
5	**154**	57	**-8,4**	-12,4666667	**104,72**
6	**158**	63	**-4,4**	-6,46666667	**28,4533333**
7	**149**	55	**-13,4**	-14,4666667	**193,853333**
8	**162**	72	**-0,4**	2,53333333	**-1,01333333**
9	**171**	81	**8,6**	11,5333333	**99,1866667**
10	**156**	58	**-6,4**	-11,4666667	**73,3866667**
11	**159**	67	**-3,4**	-2,46666667	**8,38666667**
12	**163**	64	**0,6**	-5,46666667	**-3,28**
13	**169**	93	**6,6**	23,5333333	**155,32**
14	**161**	68	**-1,4**	-1,46666667	**2,05333333**
15	**157**	59	**-5,4**	-10,4666667	**56,52**

En comparant cette valeur de la covariance (68.55) avec celle que nous avons obtenue à la question 3 (0.69), on s'aperçoit que la seconde valeur que nous avons calculée est simplement la première multipliée par 100. Cela est logique car la seule différence entre les deux jeux de données est que nous avons changé l'unité de mesure de la variable Taille (nous sommes passés du mètre au centimètre, ce qui revient à multiplier les valeurs par 100).

➔ **Ce résultat nous conforte dans notre idée qu'il est complexe d'interpréter directement la covariance car sa valeur dépend de la grandeur des échelles sur lesquelles se situent les variables, et de la variabilité existant au sein de ces variables. Afin d'avoir une idée plus précise de l'intensité de la relation existant entre le poids et la taille chez ces patients, nous aurions besoin d'utiliser le coefficient de corrélation (voir TD5 pour le calcul de ce coefficient).**

B) Pour rappel, nous avions pour but de généraliser nos résultats à l'ensemble de la promo dont sont issus ces étudiants.

1/ Voici le diagramme de dispersion des données. Si vous avez inversé l'ordre des variables, cela est également correct. Cependant, il est plus logique de se dire que l'on va prédire la note à l'examen en fonction du taux de présence au cours plutôt que l'inverse, et par convention nous plaçons la variable prédite en ordonnée et le prédicteur en abscisse.



2/ La covariance des variables taux de présence et note à l'examen est de 1.84. Ci-après, vous retrouverez les étapes de calcul de cette covariance (pour plus de facilités, nous renommerons la variable Taux de présence X et la variable Note à l'examen Y).

La moyenne du groupe pour la variable X est de 12.62, et elle est de 9.62 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 premier étudiant, 8 (y1) – 9,62 (moyenne de y) va nous donner un résultat approximatif de -1,62, qui est reporté dans la première case de la cinquième 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, l'étudiant 3 qui a un taux de présence de 13 et une note à l'examen de 5 a des écarts correspondants de 0,38 et -4,62. Ainsi, le produit des écarts que nous allons calculer pour cet étudiant est (0,38*-4,62), soit approximativement -1,78, valeur reportée dans la troisième case de la sixième colonne.

Numéro de l'étudiant	Taux de présence (X)	Note à l'examen (Y)	xi-mx	yi-my	(xi-mx)*(yi-my)
1	12	8	-0,61538462	-1,61538462	0,99408284
2	10	7	-2,61538462	-2,61538462	6,84023669
3	13	5	0,38461538	-4,61538462	-1,77514793
4	14	16	1,38461538	6,38461538	8,84023669
5	9	7	-3,61538462	-2,61538462	9,4556213
6	12	10	-0,61538462	0,38461538	-0,23668639
7	13	11	0,38461538	1,38461538	0,53254438
8	14	9	1,38461538	-0,61538462	-0,85207101
9	14	17	1,38461538	7,38461538	10,2248521
10	12	16	-0,61538462	6,38461538	-3,92899408
11	14	6	1,38461538	-3,61538462	-5,00591716
12	13	4	0,38461538	-5,61538462	-2,15976331
13	14	9	1,38461538	-0,61538462	-0,85207101

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 12 (13 – 1). En effet, notre but ici est de généraliser les résultats obtenus sur ces 13 étudiants à l'ensemble de la promotion 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 12 nous renvoie la valeur 1.84.

3/

- ➔ **Au vu du diagramme de dispersion, il ne semble pas exister de relation linéaire entre les variables taux de présence et note à l'examen, le nuage de points ne constituant pas une forme croissante ou décroissante qui pourrait être correctement représentée par une droite.**

- ➔ **La covariance, d'une valeur de 1.84, nous suggère que si la relation entre les deux variables existe, elle est positive. Cependant, étant donné que la valeur absolue de la covariance**

dépend de la variabilité au sein des variables, nous ne pouvons conclure sur l'existence d'une relation entre les deux variables sur la base de cet indice.

(Question subsidiaire) Le fait que seulement 7% de la variance de la note à l'examen soient expliqués par le taux de présence nous suggère que le taux de présence au cours de statistiques nous permet très faiblement de prédire quelle note les étudiants auront à l'examen. Cela nous suggère que d'autres facteurs que le taux de présence entrent certainement en ligne de compte et pourraient peut-être permettre de davantage expliquer les notes des étudiants. Par exemple, le temps de travail personnel à la maison, le niveau de compréhension du cours par l'étudiant, le temps passé à écouter et à travailler pendant le cours, etc…

NB : Attention à ne pas confondre la capacité à prédire une variable (ex : la note à l'examen) en fonction d'une autre variable (ex : le taux de présence en cours) et l'existence d'une relation de causalité. **La capacité de prédiction et la causalité peuvent coexister, mais ne sont pas synonymes !**