Correction Exercices supplémentaires TD3

Questions de réflexion

1/ La variance correspondant à une moyenne de valeurs élevées au carré, elle ne peut JAMAIS être négative. L'étudiant a donc fait une erreur lors du calcul de sa variance.

Si vous essayez d'en calculer l'écart-type via une calculatrice (c'est-à-dire de calculer la racine carrée de la variance), celle-ci vous renverra un message d'erreur. Un nombre élevé au carré ne pouvant être négatif, il est également impossible de calculer la racine carrée d'un nombre négatif.

2/ Le fait que les deux étudiants obtiennent le même écart-type alors que leurs échelles correspondent à deux ordres de grandeur différents (20 et 100), ce qui se traduit également dans le calcul de leurs moyennes (11.4 et 64.7), est étonnant.

En effet, on s'attend à ce qu'une moyenne plus élevée (64.7), recueillie sur une échelle correspondant à un ordre de grandeur plus élevé (100), soit également associée à un écart-type plus élevé que pour une moyenne plus faible (11.4) recueillie sur une échelle correspondant à un ordre de grandeur inférieur (20).

Un calcul supplémentaire que nous pouvons faire est celui du coefficient de variation. Cet indice, se définissant comme le rapport de l'écart-type normalisé par la moyenne et exprimé en pourcentage, va nous permettre de comparer la dispersion de variables n'ayant pas le même ordre de grandeur.

Le coefficient de variation de l'échelle à 20 est (5.8/11.4) * 100 = 50.9% alors que le coefficient de variation de l'échelle à 100 est (5.8/64.7) * 100 = 8.9%.

Ces deux coefficients de variation comparés nous indiquent que la dispersion est beaucoup plus importante chez les individus ayant été évalués sur l'échelle à 20 points que chez ceux ayant été évalués sur l'échelle à 100 points.

3/ La plus petite valeur étant 5 et l'étendue étant 60, cela signifie que la valeur la plus importante est de 65. Au regard de la valeur du troisième quartile (24), qui est bien éloignée de la valeur 65 et nous indique que 75% des données ont une valeur de 24 ou moins, nous pouvons émettre l'hypothèse que 65 est une valeur aberrante qui nous donne une idée déformée de la variabilité réelle de notre distribution.

En effet, si nous avions seulement pris en compte l'étendue, nous aurions pu imaginer que les individus avaient des temps répartis de manière équitable entre 5 et 65, ce qui a priori, étant donné la valeur du troisième quartile, n'est pas le cas.

Votre collègue aurait pu calculer la variance et l'écart-type de ces données et, étant donné la présence d'une valeur aberrante, aurait obtenu une variance et un écart-type relativement élevés. Ce résultat, qui signifie en général une importante dispersion des données autour de la moyenne, aurait pu lui suggérer la présence de valeurs aberrantes dans son jeu de données, valeurs auxquelles la variance est très sensible. En effet, l'écart d'une valeur aberrante à la moyenne, déjà très important et d'autant plus avec l'élévation au carré, va fortement augmenter la valeur de la variance.

4/ La première explication possible au résultat de votre collègue est qu'il a simplement fait la somme des écarts à la moyenne avant d'élever cette somme au carré pour calculer la variance, ce qui est bien entendu une erreur et conduira toujours au résultat 0.

Prenons un exemple de jeu de données quelconque, dont la moyenne est 10,75, et pour lequel nous avons calculé les écarts à la moyenne (xi – mx) :

Jeu de données quelconque	**xi - mx**
7	-3.75
8	-2.75
12	1.25
15	4.25
9	-1.75
8	-2.75
13	2.25
14	3.25

Vous constaterez en faisant la somme des écarts à la moyenne (donc des valeurs de la seconde colonne) que le résultat est 0.

Ce résultat de 0 lorsque l'on calcule la moyenne des écarts sans les élever au carré est toujours valable, et résulte de la définition de la moyenne. En effet, la moyenne étant le « point d'équilibre » ou « point central » d'une distribution, la somme des écarts positifs est toujours égale à la somme des écarts négatifs. Par conséquent, la somme des écarts positifs et négatifs est obligatoirement égale à zéro. Il est donc indispensable de calculer les carrés de chacun des écarts à la moyenne avant d'en faire la somme et la moyenne pour calculer la variance.

La seconde explication possible au résultat de votre collègue est que votre prof a prononcé le même nombre de « du coup » dans les 4 TDs. En d'autres termes, les 4 valeurs sur lesquelles votre collègue a calculé la variance sont égales. Dans ce cas-là, la moyenne sera égale aux 4 valeurs et la variance sera nulle, car les écarts à la moyenne seront tous égaux à 0. Par conséquent, le seul cas pouvant vous générer une variance nulle (sans que cela soit une erreur) est celui pour lequel l'ensemble de vos données ont la même valeur.

Exercices

A) 1/ L'étendue des deux groupes est de 8, les valeurs extrêmes pour les deux groupes étant 1 et 9.

2/ Pour déterminer l'espace interquartile pour les deux groupes, il nous faut déjà déterminer les valeurs des premier et troisième quartiles.

Pour le groupe des mannequins, le premier quartile correspond à la 35/4 = 8,75^{ème} valeur dans l'ordre croissant (reportez-vous au corrigé de l'exercice supplémentaire B du TD2 pour avoir les données classées). 8,75 n'étant pas un nombre entier, on arrondit à l'entier supérieur -> 9. Le premier quartile correspond donc à la 9ème valeur, soit 2.

Le troisième quartile pour ce même groupe correspond à la (35/4) * 3 = 26,25^{ème} valeur. 26,25 n'étant pas non plus un entier, on arrondit à l'entier supérieur -> 27. Le troisième quartile correspond donc à la 27ème valeur, soit 8.

➔ **Les premier et troisième quartiles du groupe des mannequins sont 2 et 8. L'espace interquartile pour ce groupe est donc de 6 (8 – 2).**

Pour le groupe des postières, nous savons déjà que les premier et troisième quartiles correspondent aux 9ème et 27ème valeurs dans l'ordre croissant, étant donné que l'effectif total est le même que pour le groupe des mannequins. La 9ème valeur correspond à 4, et la 27ème à 7.

➔ **Les premier et troisième quartiles du groupe des postières sont 4 et 7. L'espace interquartile pour ce groupe est donc de 3 (7 – 4).**

➔ **On sait que l'espace interquartile correspond à l'espace dans lequel sont compris 50% des données. On peut donc déjà constater que les 50% d'individus se trouvant au centre de la distribution sont plus dispersés dans le groupe des mannequins que dans celui des postières, l'espace interquartile du premier groupe étant plus grand que celui du second groupe. Cependant, ce seul paramètre ne nous permet pas de savoir ce qu'il en est des 50% restants des deux groupes. Afin que ces données soient également prises en compte, le calcul de la variance est nécessaire.**

3/ La variance du groupe des mannequins est de 7.62 et l'écart-type de 2.76. La variance du groupe des postières est de 3.42, et l'écart-type de 1.85. Ci-après, vous retrouverez les étapes de calcul de la variance pour le groupe des postières (même procédure pour les mannequins) :

La moyenne du groupe est de 5.2 (**Erratum Exercices supplémentaires du TD2 où j'avais indiqué que la moyenne était de 5.4**). Dans la deuxième colonne on calcule les écarts à la moyenne pour chaque donnée de la première colonne. Dans la troisième colonne on met les écarts à la moyenne de la deuxième colonne au carré.

Données	xi-mx	$(xi-mx)^2$
2	-3,2	10,24
7	1,8	3,24
1	-4,2	17,64
8	2,8	7,84
5	-0,2	0,04

5	-0,2	0,04
6	0,8	0,64
4	-1,2	1,44
8	2,8	7,84
6	0,8	0,64
5	-0,2	0,04
5	-0,2	0,04
7	1,8	3,24
6	0,8	0,64
4	-1,2	1,44
7	1,8	3,24
3	-2,2	4,84
5	-0,2	0,04
3	-2,2	4,84
4	-1,2	1,44
4	-1,2	1,44
3	-2,2	4,84
5	-0,2	0,04
2	-3,2	10,24
6	0,8	0,64
7	1,8	3,24
9	3,8	14,44
4	-1,2	1,44
8	2,8	7,84
5	-0,2	0,04
5	-0,2	0,04
7	1,8	3,24
6	0,8	0,64
6	0,8	0,64
4	-1,2	1,44

Et pour le calcul de la variance, il nous suffit de faire la moyenne de la troisième colonne. 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, ce qui nous renvoie la valeur 3.42. La racine carrée de cette valeur, l'écart-type, est de 1.85.

➔ **La comparaison des variances et écarts-types des deux groupes nous permet de voir que les mannequins et les postières se différencient sur ces paramètres de dispersion. Contrairement à la médiane, à la moyenne et à l'étendue qui ne nous permettent pas de mettre en évidence les différences existant entre ces groupes, la variance et l'écart-type nous permettent de constater que le groupe des mannequins est plus dispersé que le groupe des postières.**

B) 1/Pour calculer les variances des deux variables, on reprend la méthode vue à la question 3 de l'exercice précédent. Pour plus de simplicité, nous allons baptiser la variable « Score symptômes positifs » X et la variable « Scores symptômes négatifs » Y :

Participants	X	Y	xi-mx	$(xi-mx)^2$	yi-my	$(yi-my)^2$
1	76	12	18,4	338,56	-9	81
2	47	28	-10,6	112,36	7	49
3	85	7	27,4	750,76	-14	196
4	22	25	-35,6	1267,36	4	16
5	74	18	16,4	268,96	-3	9
6	66	35	8,4	70,56	14	196
7	26	32	-31,6	998,56	11	121
8	90	9	32,4	1049,76	-12	144
9	33	26	-24,6	605,16	5	25
10	58	15	0,4	0,16	-6	36
11	40	34	-17,6	309,76	13	169
12	91	13	33,4	1115,56	-8	64
13	18	40	-39,6	1568,16	19	361
14	54	16	-3,6	12,96	-5	25
15	84	5	26,4	696,96	-16	256

La moyenne sur la variable X est de 57.6, et de 21 sur la variable Y. Etant donné que nous souhaitons généraliser nos résultats à la population, nous allons calculer les variances sur (n – 1).

Pour la variance de la variable X, nous allons diviser la somme de la colonne $(xi-mx)^2$ par 14, et pour la variable Y, nous allons diviser la somme de la colonne $(yi-my)^2$ par 14 également.

➔ **La variance de la variable « Score symptômes positifs » (X) est de 654.69, et celle de la variable « Score symptômes négatifs » (Y) est de 124.86.**

Les écarts-types sont les racines carrées des variances.

➔ **L'écart-type de la variable « Score symptômes positifs » (X) est de 25.59, et celui de la variable « Score symptômes négatifs » (Y) est de 11.17.**

2/ Pour comparer directement la dispersion des données sur les deux échelles, il nous faut utiliser le coefficient de variation, qui correspond au rapport de l'écart-type normalisé par la moyenne et exprimé en pourcentage.

➔ **Le coefficient de variation des scores de symptômes positifs est de (25.59/57.6) * 100 = 44.43%, et celui des scores de symptômes négatifs est de (11.17/21) * 100 = 53.19%. Sur la base de ces résultats, nous pouvons voir que la dispersion des données est un peu plus élevée pour les scores de symptômes négatifs que pour les scores de symptômes positifs.**

3/ Nous reprenons la procédure appliquée lors de la question 1 pour le calcul des variances.

Participants	X	Y	xi-mx	$(xi-mx)^2$	yi-my	$(yi-my)^2$
1	2	33	-8,8	77,44	2,4	5,76
2	7	35	-3,8	14,44	4,4	19,36
3	6	28	-4,8	23,04	-2,6	6,76
4	19	27	8,2	67,24	-3,6	12,96
5	4	32	-6,8	46,24	1,4	1,96
6	15	21	4,2	17,64	-9,6	92,16
7	3	24	-7,8	60,84	-6,6	43,56
8	18	32	7,2	51,84	1,4	1,96
9	1	26	-9,8	96,04	-4,6	21,16
10	6	34	-4,8	23,04	3,4	11,56
11	12	29	1,2	1,44	-1,6	2,56
12	16	36	5,2	27,04	5,4	29,16
13	9	25	-1,8	3,24	-5,6	31,36
14	21	38	10,2	104,04	7,4	54,76
15	23	39	12,2	148,84	8,4	70,56

La moyenne sur la variable X est de 10.8, et de 30.6 sur la variable Y. Etant donné que nous souhaitons généraliser nos résultats à la population, nous allons calculer les variances sur (n – 1).

Pour la variance de la variable X, nous allons diviser la somme de la colonne $(xi-mx)^2$ par 14, et pour la variable Y, nous allons diviser la somme de la colonne $(yi-my)^2$ par 14 également.

➔ **La variance de la variable « Score symptômes positifs » (X) est de 54.46, et celle de la variable « Score symptômes négatifs » (Y) est de 28.97.**

Les écarts-types sont les racines carrées des variances.

➔ **L'écart-type de la variable « Score symptômes positifs » (X) est de 7.38, et celui de la variable « Score symptômes négatifs » (Y) est de 5.38.**

On calcule ensuite les coefficients de variation.

➔ **Le coefficient de variation des scores de symptômes positifs est de (7.38/10.8) * 100 = 68.3%, et celui des scores de symptômes négatifs est de (5.38/30.6) * 100 = 17.58%. Sur la base de ces résultats, nous pouvons voir que la dispersion des données est beaucoup plus élevée pour les scores de symptômes positifs que pour les scores de symptômes négatifs.**

4/ L'écart de dispersion entre les scores de symptômes positifs et négatifs semble plus important chez les patients dépressifs que chez les patients psychotiques. Sur les scores de symptômes positifs, les patients dépressifs présentent une plus grande variabilité que les patients psychotiques. A l'inverse, sur les scores de symptômes négatifs, les patients dépressifs présentent une plus faible variabilité que les patients psychotiques.

➔ La présence de symptômes positifs semble assez variable dans les deux groupes, les patients dépressifs comme les patients psychotiques pouvant présenter soit peu, soit beaucoup de symptômes de ce type. En revanche, si la présence de symptômes négatifs est également variable chez les patients psychotiques, elle est relativement homogène chez les patients dépressifs et est plutôt élevée (score moyen de 30.6/40).