TD6 : Paramètres d'association (paramètres d'association (suite) et droite de régression)

Questions de réflexion

1/ La **corrélation** a pour objectif de déterminer le sens et l'intensité d'une relation de type linéaire existant entre deux variables. Son coefficient (r) se calcule à partir de la covariance et des écarts-types des variables étudiées : $\frac{Cov_{xy}}{\sigma_{x^*}\,\sigma_y}$.

La **variance expliquée** permet de savoir quelle part de la variance d'une variable peut s'expliquer par les différences sur une autre variable. Pour la calculer, il suffit de mettre le coefficient de corrélation au carré et d'exprimer la valeur obtenue en pourcentage.

Une **relation de causalité** signifie que les variations sur une variable causent des variations sur une seconde variable. Attention, car si une relation de causalité s'accompagne d'une corrélation et d'une variance expliquée importantes, l'inverse n'est pas forcément vrai ! En effet, le fait que l'on puisse prédire avec précision les valeurs d'une variable en fonction des valeurs d'une autre variable (ce qui est le cas lorsque la corrélation et la variance expliquée sont fortes) ne signifie pas forcément que la variable prédictrice est la cause de la variable prédite (ex : corrélation entre la taille des oreilles des enfants et les notes à un exercice de math ➔ l'âge est plus vraisemblablement le facteur causal de ces deux variables). Expliquer ou prédire ne veut pas dire causer !

2/ La régression linéaire et la corrélation ont pour but de déceler l'existence d'une relation linéaire entre deux variables.

En revanche, si la corrélation s'intéresse uniquement à déterminer la force et le sens de la relation entre les variables, la régression vise par contre à estimer l'une des variables (le critère, la VD) en fonction des valeurs de l'autre variable (le prédicteur, la VI). Dans la corrélation, seule la relation entre les variables nous intéresse, alors que dans la régression les variables ont un rôle différent, l'une étant le prédicteur, et l'autre la variable prédite.

3/ La covariance n'est pas suffisante pour estimer la relation existant entre deux variables, car sa valeur est fonction de la variabilité existant au sein des deux variables étudiées. Ainsi, une covariance de 5 pourra refléter une relation faible si les écarts-types des variables sont élevés, et par contre une relation élevée si les écarts-types sont faibles.

Afin de résoudre ce problème, il nous suffit de diviser la valeur de la covariance par le produit des écarts-types des deux variables, ce qui correspond au calcul du coefficient de corrélation de Pearson (r). Contrairement à la covariance qui ne nous permet que d'estimer le sens (positif ou négatif) d'une relation potentiellement existante entre deux variables, le coefficient de corrélation nous permet de déterminer l'existence et l'intensité de cette relation.

4/ Le premier outil à votre disposition pour estimer la relation entre deux variables est le **diagramme de dispersion**. Le diagramme de dispersion permet de représenter les données sur un nuage de n points (n correspondant au nombre de sujets ou de situations évalués sur les deux variables). Le diagramme de dispersion représente donc chaque sujet expérimental inclus dans l'étude par un point dans un espace bidimensionnel. La forme du nuage de points sur le diagramme de dispersion permet de faire une première estimation du sens et de la force de la relation entre les deux variables.

Le second outil est le **coefficient de corrélation de Pearson** (r). Il nous permet de déterminer de manière quantitative l'intensité et le sens de la liaison linéaire entre nos deux variables. Son signe (- ou +) nous indique le sens de la relation, et sa valeur absolue nous indique l'intensité de la relation.

Enfin, le dernier outil est la **variance expliquée**. Cet outil permet de savoir dans quelle mesure les variations observées sur une variable peuvent s'expliquer par les variations sur une autre variable.

Ces trois outils vont nous permettre de conclure à deux niveaux sur la relation existant entre deux variables :

Le premier niveau sera visuel et s'appuiera sur l'analyse du diagramme de dispersion. Cette analyse nous permettra de dire que la relation entre les variables « semble » être de telle intensité et de tel signe (ex : la relation semble être importante et positive).

Le second niveau sera quantitatif et s'appuiera sur les calculs du coefficient de corrélation de Pearson (r) et de la variance expliquée (r^2). Le coefficient de corrélation nous permettra de conclure sur l'intensité et le sens de la relation entre les variables (ex : il existe une corrélation linéaire forte et positive entre les variables). La variance expliquée nous permettra de conclure sur la part des variations d'une variable qui peut être expliquée par les variations sur la seconde variable (ex : pour une variance expliquée d'une valeur de 0.9, 90% des variations sur la première variable peuvent être expliqués par les différences sur la seconde variable).

5/ Lorsque nous souhaitons calculer la covariance entre deux variables, le n correspond au nombre de sujets ou de situations différentes sur lesquels les deux variables ont été mesurées. Dans notre exemple, le n a donc une valeur de 6, puisque nous avons mesuré chacune de nos deux variables simultanément sur 6 participants différents, chaque participant ayant un temps de révision avant de dormir et une note finale à l'examen.

Pour information, notez que si ne pas réviser avant l'examen n'est pas souhaitable (sujet 1 ayant révisé 2 mn avant de se coucher et ayant obtenu 8), il n'est pas non plus souhaitable de réviser massivement et uniquement avant l'examen (sujet 2 ayant révisé 240 mn avant de se coucher et ayant obtenu 9). Il est préférable de réviser régulièrement avant l'examen, et de relire rapidement la synthèse de cours la veille de l'examen pour un dernier rappel.

Quelques conseils : http://www.lemonde.fr/campus/article/2015/06/08/bac-et-examens-comment-optimiser-son-cerveau_4649686_4401467.html.

6/ Pour déterminer si une observation est un outlier, il faut regarder si elle se trouve à plus de deux écarts-types au-dessus ou en-dessous de la moyenne.

Par exemple, si nous avons un jeu de données avec une moyenne de 14 et un écart-type de 3, une observation d'une valeur de 24 est un outlier car elle se trouve à plus de deux écarts-types au-dessus de la moyenne ($14 + 2*3 = 20 < 24$). Une observation de 5 est également un outlier car elle se trouve à plus de deux écarts-types en-dessous de la moyenne ($14 – 2*3 = 8 > 5$).

Pour information, il est en général légitime d'écarter une donnée si elle se trouve au-delà de deux écarts-types de la moyenne car cette donnée aura une grande influence sur les calculs de moyenne, variance et coefficient de corrélation, surtout si l'échantillon est faible.

Exercices

NB : jusqu'ici, les corrections des exercices étaient essentiellement présentées sous forme de tableaux. Une alternative est de reporter les calculs grâce aux formules, comme fait dans les exercices ci-dessous.

A) 1/ Pour plus de facilités, nous allons renommer la variable « Nombre de personnes naïves présentes » X et la variable « % de situations où une aide a été apportée à la victime » Y. Pour calculer le coefficient de corrélation de Pearson, nous avons besoin au préalable de calculer la covariance et les écarts-types des deux variables.

La **moyenne sur la variable X** est $\frac{1+2+3+4+5+6}{6}$ = **3,5**.

La **moyenne sur la variable Y** est $\frac{78+60+40+20+10+5}{6}$ = **35,5**.

La **variance sur la variable X** est $\frac{(1-3,5)^2+(2-3,5)^2+\ldots+(6-3,5)^2}{5}$ = **3,5**[1]. Son écart-type est $\sqrt{3,5}$ = **1,87**.

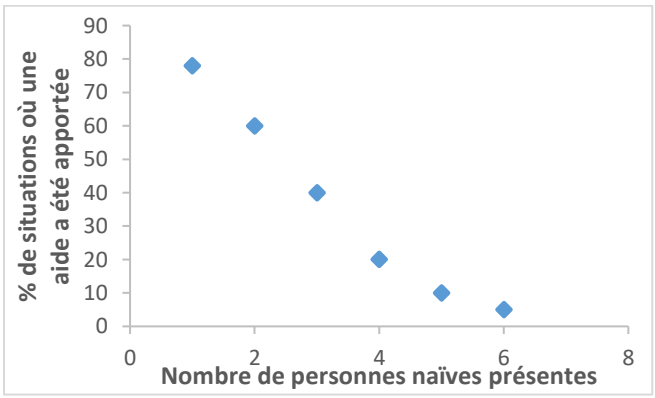
La **variance sur la variable Y** est $\frac{(78-35,5)^2+(60-35,5)^2+\ldots+(5-35,5)^2}{5}$ = **849,5**[1]. Son écart-type est $\sqrt{849,5}$ = **29,15**.

La **covariance** des variables est $\frac{(1-3,5)(78-35,5)+(2-3,5)(60-35,5)+\ldots+(6-3,5)(5-35,5)}{5}$ = **-53,5**[1].

Le **coefficient de corrélation** des variables est $\frac{-53,5}{(1,87*29,15)}$ = **-0,98**.

La **variance expliquée** des variables est $-0,98^2$ = **0,96**.

2/ Voici le diagramme de dispersion correspondant aux données (NB : il est important ici de mettre le taux d'assistance en ordonnée et le nombre de personnes présentes en abscisse car les chercheurs souhaitaient prédire le taux d'assistance en fonction du nombre de personnes présentes) :



[1] Les variances et covariance sont calculées sur n – 1 car le but des chercheurs était de généraliser leurs résultats à la population dont est issu l'échantillon.

3/ Pour déterminer la formule de régression linéaire correspondant à nos données, il nous faut déterminer les valeurs du a et du b de la formule.

a = $\frac{Cov_{xy}}{\sigma_x^2}$ (nb : $\sigma_x^2 = Var_x$) = $\frac{-53,5}{3,5}$ = **-15,29**.

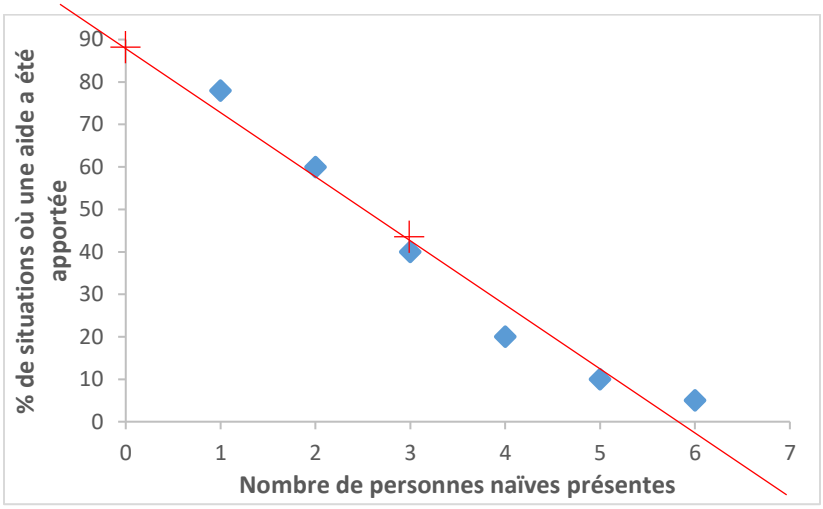
b = $\bar{Y} - a\bar{X}$= 35,5 - (-15,29) * 3,5 = **89**.

La formule de régression linéaire est donc Y = -15,29 X + 89. Pour tracer la droite, prenons deux valeurs aléatoires de X :

Si X = 0, Y = - 15,29 * 0 + 89 = 89.

Si X = 3, Y = - 15,29 * 3 + 89 = 43,13.

Nous pouvons maintenant tracer notre droite de régression sur le diagramme de dispersion.



4/ Nous pouvons conclure à deux niveaux sur la relation existant entre nos deux variables.

Le diagramme de dispersion nous indique qu'il semble visuellement exister une relation linéaire forte et négative entre le Nombre de personnes naïves présentes et le % de situations où il a été porté assistance à la victime, supposition que vient appuyer la pente de la droite de régression.

Le coefficient de corrélation et la variance expliquée nous permettent de conclure sur le plan quantitatif. Le coefficient de corrélation, d'une valeur de -0,98, nous permet de confirmer l'existence de cette relation linéaire forte et négative entre les deux variables. Plus le nombre de personnes naïves présentes est important, moins il va être porté assistance à la victime[2].

La variance expliquée nous permet de dire que 96% de la variance sur la variable Taux d'assistance à la victime peuvent être expliqués par le Nombre de personnes naïves présentes.

[2] Cette corrélation négative entre nos deux variables est un phénomène psycho-social des situations d'urgence dans lesquelles le comportement d'aide d'un sujet est inhibé par la simple présence d'autres personnes sur les lieux. En d'autres mots, plus le nombre de personnes qui assistent à une situation exigeant un secours est important, plus les chances que l'une d'entre elles décide d'apporter son aide sont faibles. On appelle cela l'effet du témoin (https://fr.wikipedia.org/wiki/Effet_du_t%C3%A9moin).

B) 1/ Pour plus de facilités, nous allons renommer la variable « Nombre moyen de critiques négatives sur le comportement » X et la variable « Moyenne générale » Y. Pour calculer le coefficient de corrélation de Pearson, nous avons besoin au préalable de calculer la covariance et les écarts-types des deux variables.

La **moyenne sur la variable X** est $\frac{2+1+5+3+4+0+5+0}{8}$ = **2,5**.

La **moyenne sur la variable Y** est $\frac{8+9+3+5+6+12+7+13}{8}$ = **7,9**.

La **variance sur la variable X** est $\frac{(2-2{,}5)^2+(1-2{,}5)^2+\ldots+(0-2{,}5)^2}{8}$ = **3,75**[3]. Son écart-type est $\sqrt{3{,}75}$ = **1,94**.

La **variance sur la variable Y** est $\frac{(8-7{,}9)^2+(9-7{,}9)^2+\ldots+(13-7{,}9)^2}{8}$ = **10,11**[3]**.** Son écart-type est $\sqrt{10{,}11}$ **= 3,18**.

La **covariance** des variables est $\frac{(2-2{,}5)\,(8-7{,}9)+(1-2{,}5)\,(9-7{,}9)+\ldots+(0-2{,}5)\,(13-7{,}9)}{8}$ = **-5,44**[3].

Le **coefficient de corrélation** des variables est $\frac{-5{,}44}{(1{,}94*3{,}18)}$ = **-0,88**.

La **variance expliquée** des variables est $-0{,}88^2$ = **0,78**.

2/ Pour déterminer la formule de régression linéaire correspondant à nos données, il nous faut déterminer les valeurs du a et du b de la formule.

a = $\frac{Cov_{xy}}{\sigma_x^2}$ (nb : $\sigma_x^2 = Var_x$) = $\frac{-5{,}44}{3{,}75}$ = **-1,45**.

b = $\bar{Y} - a\bar{X}$= 7,9 - (-1,45) * 2,5 = **11,5**.

La formule de régression linéaire est donc Y = -1,45 X + 11,5.

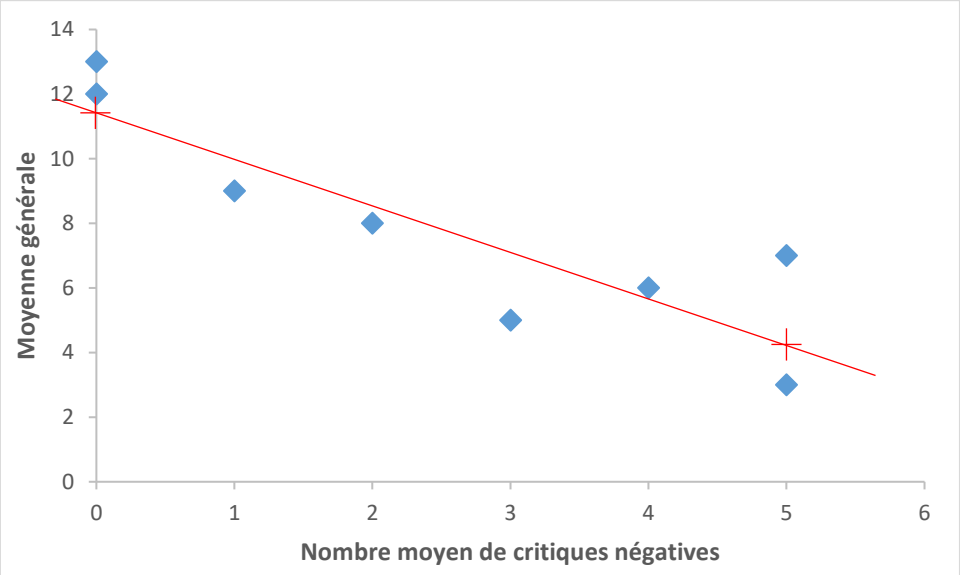
[3] Nous avons calculé les variances et covariance sur n car ici les chercheurs s'intéressaient uniquement aux données recueillies et ne souhaitaient pas généraliser leurs résultats à la population.

3/ Pour tracer la droite, prenons deux valeurs aléatoires de X :

Si X = 0, Y = - 1,45 * 0 + 11,5 = 11,5.

Si X = 5, Y = - 1,45 * 5 + 11,5 = 4,25.

Grâce à ces deux points, nous pouvons maintenant tracer notre droite de régression sur le diagramme de dispersion (NB : il est important ici de mettre la moyenne générale en ordonnée et le nombre moyen de critiques négatives en abscisse car les chercheurs souhaitaient savoir si l'on pouvait prédire la moyenne générale en fonction du nombre de critiques négatives) :



4/ Nous pouvons conclure à deux niveaux sur la relation existant entre nos deux variables.

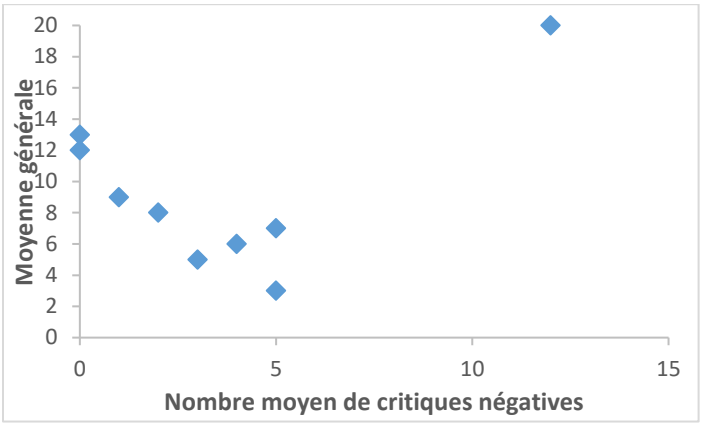
Le diagramme de dispersion nous permet de conclure qu'il semble visuellement exister une relation linéaire négative entre la variable Nombre moyen de critiques négatives sur le comportement et la variable Moyenne générale, supposition que vient appuyer la pente de la droite de régression.

Le coefficient de corrélation et la variance expliquée nous permettent de conclure sur le plan quantitatif. Le coefficient de corrélation, d'une valeur de -0,88, nous permet de confirmer l'existence de cette relation linéaire négative entre les deux variables, et d'ajouter qu'elle est d'intensité élevée. Plus le nombre moyen de critiques négatives sur le comportement est important, plus la moyenne générale de l'enfant sera faible[4].

La variance expliquée nous permet de dire que 78% de la variance sur la variable Moyenne générale peuvent être expliqués par le Nombre moyen de critiques négatives sur le comportement.

[4] Cette corrélation négative entre nos deux variables se rapproche de ce que l'on appelle effet Golem en psychologie sociale (l'effet inverse, plus connu, étant l'effet Pygmalion). Cet effet est une prophétie autoréalisatrice qui se traduit par une performance moindre sous l'effet d'un potentiel jugé limité par une autorité (parent, professeur…). Sources : https://fr.wikipedia.org/wiki/Effet_Golem ; https://fr.wikipedia.org/wiki/Effet_Pygmalion.

5/ Voici le diagramme de dispersion et les calculs nécessaires pour déterminer le coefficient de corrélation sur la base de nos nouvelles données :



La **moyenne sur la variable X** est $\frac{2+1+5+3+4+0+5+0+12}{9}$ = **3,6**.

La **moyenne sur la variable Y** est $\frac{8+9+3+5+6+12+7+13+20}{9}$ = **9,2**.

La **variance sur la variable X** est $\frac{(2-3{,}6)^2+(1-3{,}6)^2+\ldots+(12-3{,}6)^2}{9}$ = **12,25**. Son écart-type est $\sqrt{12{,}25}$ = **3,50**.

La **variance sur la variable Y** est $\frac{(8-9{,}2)^2+(9-9{,}2)^2+\ldots+(20-9{,}2)^2}{9}$ = **23,51.** Son écart-type est $\sqrt{23{,}51}$ = **4,85**.

La **covariance** des variables est $\frac{(2-3{,}6)\,(8-9{,}2)+(1-3{,}6)\,(9-9{,}2)+\ldots+(12-3{,}6)\,(20-9{,}2)}{9}$ = **6,54**.

Le **coefficient de corrélation** des variables est $\frac{6{,}54}{(3{,}50*4{,}85)}$ = **0,39**.

On peut noter sur le diagramme de dispersion que le point correspondant au nouveau sujet est très éloigné du reste de la distribution. D'autre part, on peut noter que le rajout de ce sujet a fait passer la valeur du coefficient de corrélation de -0,88 à 0,39, soit d'une relation linéaire négative forte à une relation linéaire positive modérée.

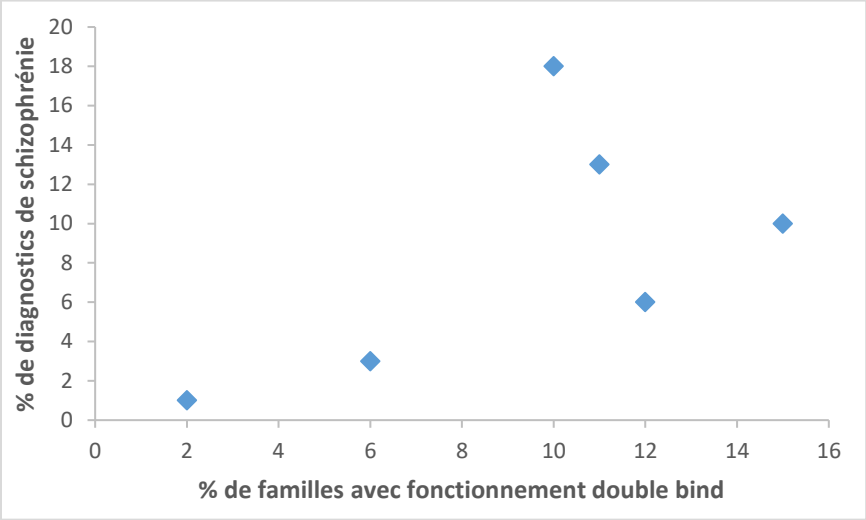
Sur la base de ces observations, nous pouvons supposer que les données d'Isabella sont des données outliers. Pour s'en assurer, nous pouvons vérifier qu'elles se situent à plus de deux écarts-types de la moyenne.

La moyenne sur la variable X est de 3,6 et l'écart-type de 3,50. La valeur d'Isabella sur la variable X (12) est donc outlier car 3,6 + 2 * 3,50 = 10,6 < 12.

La moyenne sur la variable Y est de 9,2 et l'écart-type de 4,85. La valeur d'Isabella sur la variable Y (20) est donc outlier car 9,2 + 2 * 4,85 = 18,9 < 20.

Par conséquent, les données d'Isabella sur les deux variables sont outliers, ce qui explique le diagramme de dispersion et la nouvelle valeur du coefficient de corrélation.

C) 1/ Voici le diagramme de dispersion correspondant aux données (NB : il est important ici de mettre le % de diagnostics de schizophrénie en ordonnée et le % de familles avec fonctionnement double bind en abscisse car les chercheurs souhaitaient savoir si l'on pouvait prédire le % de diagnostics de schizophrénie en fonction du % de familles présentant un fonctionnement double bind) :



2/ Pour plus de facilités, nous allons renommer la variable « % de familles présentant un fonctionnement double bind » X et la variable « % de diagnostics de schizophrénie » Y. Pour calculer le coefficient de corrélation de Pearson, nous avons besoin au préalable de calculer la covariance et les écarts-types des deux variables.

La **moyenne sur la variable X** est $\frac{12+6+2+15+11+10}{6}$ = **9,33**.

La **moyenne sur la variable Y** est $\frac{6+3+1+10+13+18}{6}$ = **8,5**.

La **variance sur la variable X** est $\frac{(12-9{,}33)^2 + (6-9{,}33)^2 + \ldots + (10-9{,}33)^2}{5}$ = **21,47**[5]. Son écart-type est $\sqrt{21{,}47}$ = **4,63**.

La **variance sur la variable Y** est $\frac{(6-8{,}5)^2 + (3-8{,}5)^2 + \ldots + (18-8{,}5)^2}{5}$ = **41,1**[5]**.** Son écart-type est $\sqrt{41{,}1}$ **= 6,41**.

La **covariance** des variables est $\frac{(12-9{,}33)(6-8{,}5) + (6-9{,}33)(3-8{,}5) + \ldots + (10-9{,}33)(18-8{,}5)}{5}$ = **17,8**[5].

Le **coefficient de corrélation** des variables est $\frac{17{,}8}{(4{,}63*6{,}41)}$ = **0,60**.

La **variance expliquée** des variables est $0{,}60^2$ = **0,36**.

[5] Les variances et covariance sont calculées sur n – 1 car le but des chercheurs était de généraliser leurs résultats à la population dont sont issues les données.

3/ Nous pouvons conclure à deux niveaux sur la relation existant entre nos deux variables.

Le diagramme de dispersion nous permet de conclure qu'il semble visuellement exister une relation positive entre la variable % de familles présentant un fonctionnement double bind et la variable % de diagnostics en schizophrénie.

Le coefficient de corrélation et la variance expliquée nous permettent de conclure sur le plan quantitatif. Le coefficient de corrélation, d'une valeur de 0,60, nous permet de confirmer l'existence de cette relation positive entre les deux variables, et nous permet d'ajouter qu'elle est d'intensité modérée, voire relativement élevée. Plus le % de familles présentant un fonctionnement double bind est élevé, plus le % de diagnostics en schizophrénie est élevé.

La variance expliquée nous permet de dire que 36% de la variance sur la variable % de diagnostics en schizophrénie peuvent être expliqués par le % de familles avec fonctionnement double bind[6].

4/ Pour déterminer la formule de régression linéaire correspondant à nos données, il nous faut déterminer les valeurs du a et du b de la formule.

a = $\frac{Cov_{xy}}{\sigma_x^2}$ (nb : $\sigma_x^2 = Var_x$) = $\frac{17,8}{21,47}$ = **0,83**.

b = $\bar{Y} - a\bar{X}$= 8,5 – 0,83 * 9,33 = **0,76**.

La formule de régression linéaire est donc Y = 0,83 X + 0,76.

Pour tracer la droite, prenons deux valeurs aléatoires de X :

Si X = 0, Y = 0,83 * 0 + 0,76 = 0,76.

Si X = 20, Y = 0,83 * 20 + 0,76 = 17,36.

[6] Attention cependant à ne pas tirer trop vite des conclusions sur la relation entre les variables sur la base de ces résultats, car un pré-requis indispensable pour que cette corrélation ait réellement un sens est de vérifier que les familles présentant un fonctionnement double bind et celles ayant un adolescent avec diagnostic de schizophrénie sont les mêmes familles ! D'autre part, s'il existe une théorie sur le lien entre le fonctionnement double bind et la schizophrénie, il n'existe à ce jour pas de confirmation empirique de ce modèle. Pour plus d'informations sur ce modèle théorique : http://laboiteame.unblog.fr/double-contrainte-injonction-paradoxale-ecole-de-palo-alto/.

Grâce à ces deux points, nous pouvons maintenant tracer notre droite de régression sur le diagramme de dispersion :