**Module 7 : Le nuage statistique et la corrélation**

1. Indique si chacune des affirmations suivantes est vraie ou fausse (explique si fausse).
	1. **Les associations statistiques existent uniquement pour des variables quantitatives.** *Faux, une association statistique peut aussi exister pour des variables qualitatives.*
	2. **S’il existe une association entre deux variables, alors la connaissance de la valeur de l’une permet de prédire plus précisément la valeur de l’autre.** *Vrai, grâce à l’équation de la droite de régression.*
	3. **Un fort coefficient de corrélation entre deux variables signifie qu’il existe une association statistique linéaire positive entre celles-ci.** *Faux, l’association peut aussi être négative.*
	4. **Si le coefficient de corrélation entre deux variables est près de 0, il n’y a pas d’association entre les deux variables.** *Faux, il peut y avoir une association, mais pas de corrélation, on ne pourrait pas prédire la valeur d’une variable en connaissant l’autre.*
	5. **Une relation entre deux variables est dite causale, si la valeur que prend une variable dépend de la valeur que prend l’autre variable.** *Vrai, on peut aisément prédire la valeur d’une variable en connaissant l’autre.*
	6. **La relation entre l’âge d’un homme et l’âge de sa conjointe est causale.** *Faux, le temps fait en sorte qu’il y a une corrélation. L’âge de l’homme ne cause pas l’âge de la femme.*
	7. **Le coefficient de corrélation est suffisant pour affirmer l’existence d’un modèle linéaire (droite) lorsqu’il vaut exactement 1 ou -1.** *Vrai, cela donne l’équation y = mx + b.*
2. **Si les élèves qui réussissent bien en musique réussissent aussi généralement bien en mathématiques, on peut dire que l’association est :** *positive (c’est difficile de prédire à quelle force).*
3. **Décris l’association pour chacun des nuages de points suivants :**



1. Douze personnes se sont inscrites à un programme d’entraînement physique d’une durée de 10 semaines. On fait passer un test de conditionnement physique à ces personnes au début et à la fin des 10 semaines. Le diagramme de dispersion ci-contre présente pour chacune des 12 personnes, les résultats exprimés en cotes z aux deux tests passés.
2. **Par rapport à l’ensemble du groupe, trouve la personne qui a le plus progressé.** *La personne J.*
3. **Par rapport à l’ensemble du groupe, trouve la personne qui a le plus régressé.** *La personne H.*
4. Le coefficient de corrélation, r, entre la taille des joueurs de basket-ball d’une équipe fictive et leur nombre de paniers réussis dans une saison est de 0,60. Détermine si chacune des affirmations suivantes est vraie ou fausse.
	1. **Les joueurs réussissent en moyenne 60% de leurs lancers au panier.** *FAUX*
	2. **Aucun joueur de petite taille ne réussit beaucoup de paniers.** *FAUX*
	3. **60% des joueurs de grande taille réussissent plus de paniers que la moyenne.** *FAUX*
	4. **Les plus petits joueurs réussissent moins de paniers.** *VRAI*
	5. **Puisqu’on ne connait pas la taille moyenne des joueurs, le coefficient de corrélation n’a pas d’intérêt.** *FAUX*
5. Une technicienne indique la masse et la taille de 20 personnes sur des fiches individuelles. Ces données servent à calculer le coefficient de corrélation entre la masse et la taille. Lesquelles, parmi les affirmations suivantes, modifient le coefficient de corrélation ? :
	1. **Les fiches ont été échappées sur le sol et les données ont été entrées dans l’ordinateur dans un autre ordre.**
	2. ***La masse et la taille ont été interchangés sur une des fiches.***
	3. **La masse et la taille ont été interchangés sur toutes les fiches.**
	4. ***Une fiche a été perdue.***

|  |  |
| --- | --- |
| **Nombre de rondes** | **Handicap** |
| 15 | 12 |
| 28 | 4 |
| 45 | 2 |
| 25 | 13 |
| 31 | 9 |
| 47 | 5 |
| 19 | 17 |
| 26 | 10 |

1. Le tableau ci-contre indique, pour chacun des personnes de 30 à 35 ans membres du Club de golf le Crystal, le nombre de rondes de golf jouées pendant la saison et le handicap. **Calcule et interprète le coefficient de corrélation.**

r = $\frac{\sum\_{}^{}xy-n\overbar{x}\overbar{y}}{(n-1)s\_{x}s\_{y}}$ = $\frac{1804-\left(8\right)(29,5)(9)}{(8-1)(11,364)(5,071)}$ ≈ -0,79

*Avec un coefficient de corrélation de -0,79, l’Association entre les deux variables est forte et négative. Plus un membre joue un grand nombre de ronde, plus son handicap diminue.*

1. Pour chaque relation suivante, indique :
	1. **Si la relation la force et le signe de la corrélation (lorsque c’est possible)**
	2. **Si la relation est causale ou non (sinon indique la cause)**
	3. **Si la relation est causale, indique la variable explicative (indépendant) et la variable réponse (dépendante)**
* La masse d’une personne et sa taille. *Cette corrélation devrait être forte positivement. Elle est causale, la variable explicative est la taille et la variable réponse est la masse.*
* Le nombre de dépanneurs et le nombre salle de cinéma dans une ville. *Cette relation devrait être forte positivement. Elle est non-causale, le 3e facteur, ou la cause, devrait être la taille de la population des villes.*
* La valeur d’une maison et le revenu annuel familial. *Cette corrélation devrait être forte positivement. Elle est causale, la variable explicative est le revenu familial et la variable réponse est la valeur de la maison.*
* La température moyenne et le nombre de policier d’une municipalité. *Il ne devrait pas y avoir une forte corrélation entre ces deux variables. Elle est non-causale, il n’y a pas vraiment de facteur externe qu’il affecte en même temps ces deux variables.*
* L’âge et la valeur des automobiles. *Cette corrélation devrait être forte négativement. Elle est causale, la variable explicative est l’âge de la voiture et la variable réponse est la valeur de la voiture.*
* L’âge du directeur et le rendement académique des élèves. *Il ne devrait pas y avoir une forte corrélation entre ces deux variables. Elle est non-causale, il n’y a pas vraiment de facteur externe qu’il affecte en même temps ces deux variables.*
* Le nombre de parties gagnées et le nombre de parties perdues par une équipe sportive. *Cette corrélation devrait être forte négativement. Elle est causale, la variable explicative est le nombre de victoires et la variable réponse est le nombre de défaites.*

**Module 8 : La droite de régression**

1. Le tableau ci-dessous représente la relation entre l’accès à l’eau potable et le taux de mortalité des enfants de 0-5 ans.
* Variable explicative (x) : Pourcentage de la population ayant accès à une source d’eau potable
* Variable réponse (y) : Mortalité des enfants de moins de 5 ans / 1000 naissances vivantes.

On a calculé les mesures suivantes pour cette relation :

1. **Calcule et interprète la pente de la droite de régression.**

m= $\frac{\sum\_{}^{}xy-n\overline{x}\overline{y}}{(n-1)(s\_{x})^{2}}$ = $\frac{\left(376 236\right)-(42)(63,2)(151,3)}{(42-1)(18,9^{2})}$ =$ \frac{-25 374,72}{14 645,61}$ $≈$ -1,73. *Cela signifie qu’à chaque augmentation d’un pourcent d’accès à l’eau potable, le taux de mortalité diminue de 1,73 pour 1000 naissances.*

1. **Prédis le taux de mortalité (nombre de décès / 1000 naissances vivantes) des enfants de moins de 5 ans pour un pays où 40% de la population ont accès à de l’eau potable.**

*b =* $\overline{y}$ – m$\overline{x}$ = 151,3 – (-1,73)(63,2) = 260,636 🡪 y = -1,73x + 260,636 🡪 -1,73(40) + 260,636 = 191,436. *Il y a environ 191 décès par 1000 naissances vivantes lorsque 40% de la population ont accès à de l’eau potable.*

1. **L’équation de la droite de régression d’une relation est y = 1,82x + 8,54. Détermine quel point, parmi les suivants, n’appartient pas à cette droite.**



1. Le tableau suivant présente la relation entre le nombre de tirs au but et le nombre de buts produits pendant les séries éliminatoire 2003/2004 par les 15 meilleurs pointeurs de l’équipe des Flames de Calgary.
2. **Trace un nuage de dispersion.**
3. **Calcule et interprète le coefficient de corrélation.**
4. **Calcul l’équation de la droite de régression.**



*Coefficient de corrélation : r =* $\frac{\sum\_{}^{}xy-n\overbar{x}\overbar{y}}{(n-1)s\_{x}s\_{y}}$ *=* $\frac{\left(3052\right)-(15)(41,06667)(3,86667)}{(15-1)(17,5559)(3,3778)}$ *≈* $\frac{670,13}{830,2}$ *≈ 0,81*

*Le coefficient de corrélation est de 0,81, ce qui est positif et fort. Cela signifie qu’il y a une bonne relation entre les deux variables, plus un joueur lance au but, plus il devrait faire des buts.*

m= $\frac{\sum\_{}^{}xy-n\overline{x}\overline{y}}{(n-1)(s\_{x})^{2}}$ = $\frac{\left(3052\right)-(15)(41,06667)(3,86667)}{(15-1)(17,5559^{2})}$ *≈*$\frac{670,13}{4 314,935}$ *≈ 0,1553*

*b =* $\overline{y}$ – m$\overline{x}$ =3,86667 – 0,1553(41,06667) *≈ -2,5112*

*Équation de la droite de regression : y = 0,1553x – 2,5112*

1. **Utilise ton équation pour prédire le nombre de but que devrait produire un joueur qui tire 60 fois au filet.**

*y = 0,1553x – 2,5112 🡪 y = 0,1553(60) – 2,5112 ≈6,8. Une personne qui tire 60 fois au filet devrait scorer environ 6,8 (7) fois.*

1. Un chauffeur de taxi a noté le nombre de kilomètres parcourus et le coût des sept dernières courses qu’il a faites. Voici les résultats :

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Nombre de km** | 14,4 | 12,9 | 8,4 | 9,2 | 10,4 | 7,2 | 13,5 |
| **Coût ($)** | 24,10 | 22,90 | 17,20 | 16,50 | 18,80 | 12,80 | 23,70 |

1. **Quelles sont les variables étudiées ? Laquelle est la variable explicative (indépendante) et la variable réponse (dépendante) ? Quels types de variables sont chacune d’elles ?**

*Variable explicative (indépendante, x) est le nombre de kilomètre. Il s’agit d’une variable quantitative continue.*

*Variable réponse (dépendante, y) est le coût. Il s’agit d’une variable quantitative continue.*

1. **Calculer et interpréter le coefficient de corrélation.**



*Coefficient de corrélation : r =* $\frac{\sum\_{}^{}xy-n\overbar{x}\overbar{y}}{(n-1)s\_{x}s\_{y}}$ *=* $\frac{\left(1 546,36\right)-(7)(10,857)(19,429)}{(7-1)(2,771)(4,282)}$ *≈* $\frac{69,775429}{71,192532}$ *≈ 0,98*

*Il y a une très forte corrélation positive entre les deux variables. Plus le nombre de kilomètre augmente, plus le coût augmente. On pourrait vraiment estimer une interpolation ou une extrapolation avec un coefficient de corrélation de 0,98.*

1. **Donner l’équation de la droite de régression.**

m= $\frac{\sum\_{}^{}xy-n\overline{x}\overline{y}}{(n-1)(s\_{x})^{2}}$ = $\frac{\left(1546,36\right)-(7)(10,857)(19,429)}{(7-1)(2,771^{2})}$ *≈*$\frac{69,775429}{46,070646}$ *≈ 1,5146*

*b =* $\overline{y}$ – m$\overline{x}$ = 19,429 – 1,5146(10,857) *≈ 2,9843*

*Équation de la droite de regression : y = 1,5146x +2,9843*

1. **Dans le contexte de ce problème, donner une signification aux valeurs m et b de la droite de régression.**

*Le m représente le taux de variation. Cela veut dire qu’à chaque kilomètre, le coût augmente de 1,5146 $.*

*Le b représente la valeur initiale. Cela veut dire qu’il y a un coût initiale (de base) sans avoir parcouru de kilométrage de 2,9843 (environ 3 $).*

1. **Interpoler le prix si une personne parcourrait 12 kilomètres dans ce taxi.**

*Y = 1,5146(12) + 2,9843 = 21,16 $ Cela lui coûterait 21,16 pour parcourir 12 kilomètres.*

1. **Extrapoler le nombre de kilomètres d’une personne dont le coût serait de 35 $.**

*35 = 1,5146x + 2,9843 🡪 1,5146x = 32,0157 🡪 x = 21,138 km. Il pourrait parcourir 21,138 km avec 35 $.*

1. **Tracer le nuage de point et tracer la droite de régression.**

