Scénario 5

LIAISON ENTRE 2 VARIABLES

1 Présentation générale

1.1 Objectifs

- Réaliser une représentation graphique caractérisant la liaison entre deux séries de mesures.
- Dire si le calcul d'une corrélation est justifié.
- Calculer une corrélation, citer un exemple où l'interprétation cause-effet est absurde.

1.2 Les étapes

- Caractérisation de la liaison par une représentation graphique :
  1. notion de couple de mesures
  2. réalisation d'un graphique, discussion
  3. présentation de différents types de relation visuelle
- Calcul d'un coefficient de corrélation, interprétation :
  1. calcul du τ (tau) de Kendall
  2. présentation du coefficient de corrélation linéaire r
  3. calcul de r et notion de covariance (*en option*)
  4. présentation du coefficient de corrélation de Spearman
  5. interprétation des valeurs des coefficients
- Présentation rapide d'autres types de graphiques (*en option*)

1.3 Matériel

tableau - transparents

1.4 La forme

Travaux pratiques - discussion

1.5 Durée

1/2 journée

1.6 Prérequis

- scénario 4
- représentation graphique selon deux axes de coordonnées
- en option : notations indice (X_i), puissance (X^2) et somme ($\sum X_i$)

1.7 Evaluation

En groupe, au cours du scénario, sous forme orale et écrite

2 Les étapes

2.1 Etape 1

- distribuer à chaque participant le fichier des 10 mesures sur les sépales d'iris (feuille 1).

  ⟶ discuter ensemble de la nature des mesures et des méthodes de mesure, insister sur la notion de *couple de données.*

  ⟶ demander de résumer chacune des deux séries de nombres (cette vérification des prérequis risque de "redonder" avec l'évaluation du scénario précédent).

- Poser les questions : existe-t-il une relation entre la longueur et la largeur des sépales? et comment peut-on représenter cette relation?

  ⟶ attendre les réponses. On espère que ce sera "faire un graphique" (si c'est "faire un calcul", insister sur le fait qu'on veut voir d'abord). Demander de préciser quel graphique.

  ⟶ faire le graphique ensemble à partir du transparent 1 : choix des abscisses et ordonnées, choix des échelles, légendes et titres, identification des couples (x, y), placement.

- faire ensemble l'interprétation globale du graphique. Le but est de récupérer le vocabulaire exprimé (susciter certains mots qui manquent au besoin), qu'on classera au fur et à mesure en 3 colonnes sur le tableau

liaison	effet	corrélation
dépendance	influence	covariance
correspondance		(régression)
covariation		
association		
concordance		

  ⟶ montrer que la 1ère colonne relève de l'*interprétation*, que la 2ème introduit la notion de *cause-effet*, tandis que la 3ème relève du *calcul.* (Si le mot "régression" sort, dire qu'on y reviendra plus tard et que c'est autre chose, sinon le laisser tranquille).

- demander quels autres types de relations visuelles on peut avoir entre deux variables. Les schématiser au tableau au fur et à mesure. Il serait bon d'avoir en tête des exemples adaptés aux participants pour les cas suivants (au cas où certains ne seraient pas cités) : pas de relation, relation inverse, n groupes, relation non linéaire.

  Présenter une représentation graphique possible dans le cas de nombreux points à partir des transparents 2 et 3.

- montrer sur le transparent 4, dont une copie sera distribuée simultanément aux stagiaires (feuille 2), les 2 graphiques représentant la relation entre longueur et largeur des sépales chez les 2 espèces *iris setosa* et *iris versicolor.* Poser la question : La relation entre longueur et largeur des sépales est-elle la même pour les 2 espèces d'iris?

  ⟶ la réponse attendue est : La relation est différente pour les 2 espèces. Longueur et largeur des sépales apparaissent dépendantes chez *iris setosa* et indépendantes chez *iris versicolor.*

⟶ dire que l'on recherche une *mesure de la dépendance* entre deux variables, un résumé numérique de la relation observée sur le graphique.

2.2 Etape 2

- définir les notions de *concordance* et *discordance* entre 2 points à partir du transparent 5.

  ⟶ compter ensemble le nombre de paires total, le nombre de paires concordantes et discordantes sur la graphique *iris setosa*. A cette occasion, on précise que les paires pour lesquelles les valeurs de X (ou de Y) présentent des ex-aequos sont considérées comme "nulles" : on leur affecte la valeur 0.

  ⟶ pour évaluation immédiate, leur faire traiter de façon identique mais seuls le graphique *iris versicolor*.

- A l'aide du transparent 6, définir le *coefficient de corrélation* τ (tau) de Kendall à partir de la statistique S définie à partir des situations de concordance (+1) et de discordance (-1). Montrer que par construction, il est totalement indépendant de l'échelle de chacune des deux variables et il est compris entre -1 et +1.

  ⟶ calculer ensemble la valeur de ce coefficient de corrélation sur l'exemple *Iris setosa*.

  ⟶ pour évaluation immédiate, leur faire calculer la valeur de ce coefficient sur l'exemple *Iris versicolor*.

  ....
  ..... PAUSE .....
  ....

- Dire qu'il existe plusieurs mesures différentes de la dépendance entre 2 séries de mesures ou coefficients de corrélation :

  1. coefficient de corrélation τ de Kendall
  2. coefficient de corrélation de Pearson ou coefficient de corrélation linéaire, noté r. Il s'agit du coefficient de corrélation que l'on trouve sur les calculettes.

     En option : présenter le calcul de ce coefficient à partir de la notion de covariance : voir annexe 1 et transparents 12 à 18 (pour cela, *prérequis supplémentaires* : connaître les notations indice (X_i), puissance (X^2), et somme ($\sum X$)).
  3. montrer, à partir du transparent 7, que ce coefficient ne mesure que la liaison linéaire entre les 2 variables, alors que le τ de Kendall mesure une liaison plus générale (monotonie). La notion de linéarité est illustrée par le transparent 8.
  4. coefficient de corrélation de Sperman, noté r_s, où on remplace chaque valeur par son rang dans la série de mesures étudiées.
  5. montrer sur le transparent 9 que pour un même jeu de données, ces coefficients peuvent donner des valeurs numériques différentes.

- donner les *conseils de prudence* pour une bonne interprétation de la liaison entre 2 séries de nombres :

  1. rappeler oralement qu'il est indispensable d'avoir des *couples de données*.
  2. montrer qu'un calcul de coefficient de corrélation peut être *non justifié* (transparent 9) : présence de valeurs particulières et/ou aberrantes, mélange de plusieurs populations, relation non linéaire.
  3. souligner le danger d'une *interprétation* de type *cause à effet*.

Raconter deux exemples marquant :

- si, dans une enceinte où l'on fait croître la température, on place côte à côte un fil métallique et un bouillon de culture, et si on mesure à intervalles réguliers l'allongement du fil et le taux de développement des organismes, on trouvera une corrélation étroite entre ces deux variables.
- des observations annuelles montreraient sans doute qu'en région Alsace la baisse de natalité est liée à la diminution du taux de migration des cigognes.

  ⟶ pour évaluation, demander aux stagiaires d'autres exemples (en prévoir quelques-uns).

2.3 Etape 3

(en option) (durée 1/4 heure)

Montrer qu'il existe une autre représentation graphique possible de la relation entre 2 séries de nombres : le *boxplot* ou *boite à pattes par tranche.*

- **Cas de deux variables quantitatives :** exemple montré sur le transparent 10 de la liaison entre la latitude d'origine et le potentiel de croissance de plusieurs provenances (ou origines naturelles) de Douglas (pin d'Orégon).
  Les 4 tranches latitudinales choisies correspondent aux 4 états suivants à l'Ouest du continent Nord-Américain : la Californie, l'Orégon, l'Etat de Washington et la Colombie britannique.

  ⟶ insister sur la nécessité d'avoir un grand nombre de points par tranche pour que ce graphique soit réalisable.
- **Cas d'une variable quantitative et d'une variable qualitative :** exemple montré sur le transparent 11 de la relation entre la consommation des automobiles en 1/100 km et le pays d'origine du constructeur.

2.4 Mots-clé

Couple de données ⟶ évaluation orale lors de la réalisation du graphique
Dépendance
Concordance-discordance ⟶ évaluation écrite par exercices
Coefficient de corrélation ⟶ évaluation orale pour la justification de son calcul et écrite pour le calcul lui-même
Liaison cause-effet ⟶ évaluation orale
Linéarité
Rang d'une observation

2.5 Liste du vocabulaire

Coefficient de corrélation de Pearson (Pearson correlation coefficient)
Coefficient de corrélation t de Kendall (Kendall's tau, Kendall's rank correlation coefficient)
Coefficient de corrélation de Spearman (Spearman's rank correlation)
Concordance
Rang

3 Annexes

3.1 Documents stagiaires

Présentation générale du scénario
Feuilles 1 et 2
Vocabulaire

3.2 Documents formateurs

Transparents 1 à 11 : étapes 1 et 2
Transparents 12 à 18 : avec commentaires : corrélation linéaire

3.3 Transparents 1 à 18