Leçon d'oral du CAPES de mathématiques

Séries statistiques à deux variables numériques

Table des matières

1 Introduction et enjeux de la leçon **2**
- 1.1 Prérequis et objectifs . . . 2

2 Définition d'une série statistique à deux variables **2**
- 2.1 Définition . . . 2
- 2.2 Nuage de points . . . 3

3 Caractéristiques numériques **3**
- 3.1 Moyennes . . . 3
- 3.2 Variances et écarts-types . . . 3
- 3.3 Covariance . . . 4
- 3.4 Coefficient de corrélation linéaire . . . 4

4 Ajustement affine **5**
- 4.1 Principe . . . 5
- 4.2 Ajustement graphique . . . 5
- 4.3 Droite de régression de Y en X . . . 5
- 4.4 Propriétés et interprétation . . . 5
- 4.5 Coefficient de détermination . . . 6
- 4.6 Résidus . . . 6

5 Exemple détaillé **6**
- 5.1 Présentation des données . . . 6
- 5.2 Calcul des moyennes . . . 6
- 5.3 Calcul des variances . . . 6
- 5.4 Calcul de la covariance . . . 7
- 5.5 Coefficient de corrélation . . . 7
- 5.6 Droite de régression . . . 7

6 Outils numériques **8**

7 Applications pédagogiques **8**
- 7.1 Exemple de mise en activité . . . 8
- 7.2 Objectifs d'apprentissage . . . 8
- 7.3 Difficultés et erreurs fréquentes . . . 8

8 Limites de l'approche **9**

9 Conclusion **9**

10 Références bibliographiques **9**

1 Introduction et enjeux de la leçon

L'étude statistique de deux variables numériques permet de mettre en évidence une éventuelle liaison entre deux caractères quantitatifs observés simultanément sur une même population.

Elle intervient notamment en économie, en physique, en biologie, en sciences sociales et dans l'analyse de données.

L'objectif est de décrire la relation entre deux variables, de rechercher une tendance et, lorsque cela est pertinent, de proposer un modèle permettant d'effectuer des estimations.

On distinguera soigneusement :

— la description statistique d'une série ;

— la mise en évidence d'une association entre deux variables ;

— l'ajustement d'un modèle mathématique ;

— l'interprétation et les limites des prévisions.

Problématique

Comment décrire la liaison entre deux variables numériques, mesurer son intensité et proposer un ajustement pertinent pour interpréter les données ou effectuer des prévisions ?

1.1 Prérequis et objectifs

Prérequis. La leçon s'appuie sur les connaissances suivantes :

— les notions de population, d'individu, de caractère statistique et de série statistique à une variable ;

— le calcul et l'interprétation d'une moyenne, d'une variance et d'un écart-type ;

— le repérage de points dans un repère cartésien ;

— les fonctions affines et l'équation d'une droite ;

— les calculs algébriques, les sommes et les pourcentages ;

— l'utilisation d'une calculatrice ou d'un tableur pour traiter des données numériques.

Objectifs. À l'issue de cette leçon, on doit être capable de :

— définir une série statistique à deux variables numériques et représenter les données par un nuage de points ;

— calculer et interpréter les moyennes marginales, les variances, les écarts-types et la covariance ;

— définir et interpréter le coefficient de corrélation linéaire ;

— déterminer une droite de régression par la méthode des moindres carrés et interpréter ses coefficients ;

— utiliser un ajustement affine pour effectuer une estimation, en distinguant interpolation et extrapolation ;

— mobiliser des outils numériques pour analyser les données et confronter les résultats aux observations ;

— discuter les limites d'un ajustement statistique et distinguer corrélation et causalité.

2 Définition d'une série statistique à deux variables

2.1 Définition

On considère une population de n individus sur laquelle on observe simultanément deux variables numériques X et Y.

À chaque individu i, on associe un couple de valeurs

$$(x_i, y_i), \qquad i \in \{1, \ldots, n\}.$$

La liste des couples

$$\big((x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\big)$$

constitue une série statistique à deux variables numériques.

Les valeurs x_i et y_i peuvent être des valeurs brutes ou des valeurs regroupées selon le contexte.

2.2 Nuage de points

À chaque couple (x_i, y_i), on associe le point M_i de coordonnées (x_i, y_i) dans un repère du plan. L'ensemble des points M_i constitue le *nuage de points* associé à la série.

L'observation graphique du nuage permet notamment de repérer :

— une tendance croissante ou décroissante ;

— une éventuelle relation approximativement affine ;

— une relation non linéaire ;

— des groupes de points ;

— des valeurs atypiques.

Un nuage de points ne suffit toutefois pas à établir une relation de causalité entre les deux variables.

3 Caractéristiques numériques

3.1 Moyennes

Les moyennes des deux variables sont définies par

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Le point

$$G(\overline{x}, \overline{y})$$

est appelé *point moyen* du nuage.

Il constitue un point de référence important pour les méthodes d'ajustement affine.

3.2 Variances et écarts-types

Les variances descriptives sont définies par

$$V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$

et

$$V(Y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Les écarts-types correspondants sont

$$\sigma_X = \sqrt{V(X)}, \qquad \sigma_Y = \sqrt{V(Y)}.$$

On utilise également les identités

$$V(X) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^{\,2},$$

$$V(Y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^{\,2}.$$

Ces indicateurs décrivent la dispersion marginale de chacune des variables, indépendamment de leur liaison.

3.3 Covariance

La covariance descriptive de X et Y est définie par

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}).$$

Elle peut aussi s'écrire

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\,\overline{y}.$$

La covariance renseigne sur le sens de la liaison linéaire :

— une covariance positive traduit une tendance globale à l'évolution dans le même sens ;

— une covariance négative traduit une tendance globale à l'évolution en sens opposés ;

— une covariance nulle signifie l'absence de liaison linéaire mesurée par cet indicateur.

Attention : une covariance nulle n'implique pas nécessairement l'indépendance statistique, ni l'absence de toute relation entre les variables.

La covariance dépend des unités utilisées. Il est donc difficile de comparer directement des covariances provenant de séries différentes.

3.4 Coefficient de corrélation linéaire

Lorsque $\sigma_X > 0$ et $\sigma_Y > 0$, le coefficient de corrélation linéaire de Pearson est défini par

$$r = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

Il vérifie

$$-1 \leq r \leq 1.$$

Son interprétation est la suivante :

— si r est proche de 1, la liaison linéaire est fortement positive ;

— si r est proche de -1, la liaison linéaire est fortement négative ;

— si r est proche de 0, la liaison linéaire est faible ou absente.

Les valeurs intermédiaires doivent être interprétées en tenant compte du nuage de points, des valeurs atypiques et du contexte de l'étude.

Les cas $r = 1$ et $r = -1$ correspondent, pour une série finie, à un alignement exact des points sur une droite de pente respectivement positive et négative.

Le coefficient r ne mesure que la liaison linéaire : une relation non linéaire peut exister même si r est proche de zéro.

4 Ajustement affine

4.1 Principe

Lorsque le nuage de points présente une tendance approximativement rectiligne, on peut chercher à approcher les données par une droite d'équation

$$y = ax + b.$$

Cette opération est appelée *ajustement affine*.

Le choix d'un ajustement dépend de la forme du nuage, de l'objectif de l'étude et de la qualité attendue des prévisions.

4.2 Ajustement graphique

On peut tracer une droite qui passe au mieux au voisinage des points du nuage.

Cette méthode est simple mais dépend de l'appréciation graphique de l'observateur. Elle ne fournit pas, à elle seule, un critère numérique unique permettant de choisir la droite.

4.3 Droite de régression de Y en X

Lorsque $V(X) > 0$, la droite de régression de Y en X au sens des moindres carrés est la droite

$$y = ax + b$$

qui minimise la somme des carrés des écarts verticaux

$$S(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2.$$

Ses coefficients sont

$$a = \frac{\mathrm{Cov}(X, Y)}{V(X)}, \qquad b = \overline{y} - a\overline{x}.$$

La droite passe donc par le point moyen

$$G(\overline{x}, \overline{y}).$$

On peut également écrire, lorsque les deux écarts-types sont non nuls,

$$a = r\frac{\sigma_Y}{\sigma_X}.$$

Cette relation montre que la pente dépend du coefficient de corrélation et des dispersions des deux variables.

4.4 Propriétés et interprétation

La droite de régression de Y en X est adaptée à l'estimation de Y à partir d'une valeur connue de X.

La pente a indique la variation estimée de Y pour une augmentation d'une unité de X dans le modèle affine.

L'ordonnée à l'origine b est la valeur prédite pour $X = 0$. Elle n'a une interprétation concrète que si cette valeur est pertinente dans le contexte étudié.

La droite de régression de X en Y est différente en général de celle de Y en X. Elle s'utilise lorsque l'objectif consiste à estimer X à partir de Y.

4.5 Coefficient de détermination

Pour une régression linéaire simple avec constante, le coefficient de détermination est

$$R^2 = r^2.$$

Il représente la proportion de la variabilité observée de Y expliquée par le modèle affine ajusté, dans le sens usuel de la décomposition de la somme des carrés.

Il ne mesure pas la causalité et ne garantit pas la validité des prévisions hors de la zone observée.

Un coefficient R^2 élevé ne suffit pas à valider un modèle : il faut également examiner le nuage de points, les résidus et le contexte.

4.6 Résidus

Pour chaque observation, le résidu est défini par

$$e_i = y_i - \widehat{y}_i, \qquad \widehat{y}_i = ax_i + b.$$

Le résidu mesure l'écart entre la valeur observée et la valeur estimée par le modèle.

L'étude graphique des résidus peut mettre en évidence une structure non expliquée par le modèle affine. Une courbure systématique, par exemple, peut indiquer qu'un ajustement non linéaire serait plus pertinent.

5 Exemple détaillé

5.1 Présentation des données

On étudie le temps de révision x, exprimé en heures, et la note y obtenue à une évaluation, sur un groupe de cinq élèves.

Les données fictives sont les suivantes :

Temps x_i	1	2	3	4	5
Note y_i	8	9	12	13	18

On cherche à décrire la liaison entre le temps de révision et la note obtenue.

5.2 Calcul des moyennes

On a

$$\overline{x} = \frac{1+2+3+4+5}{5} = 3$$

et

$$\overline{y} = \frac{8+9+12+13+18}{5} = 12.$$

Le point moyen est donc

$$G(3, 12).$$

5.3 Calcul des variances

Pour X :

$$V(X) = \frac{1}{5}(1+4+9+16+25) - 3^2 = 11 - 9 = 2.$$

Pour Y :

$$\frac{1}{5}\sum_{i=1}^{5} y_i^2 = \frac{64+81+144+169+324}{5} = 156{,}4.$$

Ainsi,

$$V(Y) = 156{,}4 - 12^2 = 12{,}4.$$

Les écarts-types sont

$$\sigma_X = \sqrt{2}, \qquad \sigma_Y = \sqrt{12{,}4}.$$

5.4 Calcul de la covariance

On calcule

$$\frac{1}{5}\sum_{i=1}^{5} x_i y_i = \frac{8 + 18 + 36 + 52 + 90}{5} = \frac{204}{5} = 40{,}8.$$

Par conséquent,

$$\mathrm{Cov}(X, Y) = 40{,}8 - 3 \times 12 = 4{,}8.$$

La covariance est positive : les données présentent une tendance à l'augmentation simultanée des deux variables.

5.5 Coefficient de corrélation

On obtient

$$r = \frac{4{,}8}{\sqrt{2}\sqrt{12{,}4}} = \frac{4{,}8}{\sqrt{24{,}8}} \approx 0{,}964.$$

La liaison linéaire observée est forte et positive. Cette conclusion doit néanmoins être accompagnée de l'observation du nuage et de la prise en compte de la petite taille de l'échantillon.

5.6 Droite de régression

La pente vaut

$$a = \frac{4{,}8}{2} = 2{,}4.$$

L'ordonnée à l'origine est

$$b = 12 - 2{,}4 \times 3 = 4{,}8.$$

La droite de régression de Y en X est donc

$$\boxed{\widehat{y} = 2{,}4x + 4{,}8}.$$

Le modèle estime une augmentation moyenne de 2,4 points de note par heure supplémentaire de révision, dans le cadre de cet ajustement descriptif.

Pour $x = 6$, il donne

$$\widehat{y} = 2{,}4 \times 6 + 4{,}8 = 19{,}2.$$

Cette valeur constitue une extrapolation, car le temps de six heures dépasse l'intervalle observé $[1, 5]$. Elle doit donc être interprétée avec prudence.

Enfin,

$$R^2 = r^2 \approx 0{,}929.$$

Le modèle affine rend compte d'une part importante de la variabilité observée dans cet exemple, sans établir pour autant que le temps de révision est la seule cause des résultats scolaires.

6 Outils numériques

Une calculatrice scientifique, un tableur ou un logiciel de statistique permet de :

— construire le nuage de points ;

— calculer les moyennes, variances et covariance ;

— déterminer le coefficient de corrélation ;

— obtenir la droite de régression ;

— calculer les valeurs estimées et les résidus ;

— comparer plusieurs modèles d'ajustement.

L'usage d'un logiciel ne dispense pas de savoir interpréter les résultats et de vérifier leur cohérence avec les données.

Dans un tableur, on peut représenter les observations par un graphique de type nuage de points et ajouter une courbe de tendance linéaire en affichant son équation et son coefficient de détermination.

7 Applications pédagogiques

7.1 Exemple de mise en activité

On peut proposer aux élèves d'étudier la relation entre la température extérieure et la consommation électrique d'un bâtiment, ou entre la masse et la taille d'individus d'une population étudiée.

Les élèves peuvent :

1. identifier les deux variables et leurs unités ;
2. représenter le nuage de points ;
3. formuler une conjecture sur la liaison ;
4. calculer ou faire calculer les indicateurs ;
5. déterminer un ajustement affine ;
6. discuter la pertinence des estimations obtenues.

7.2 Objectifs d'apprentissage

Les objectifs sont de savoir :

— lire et construire un nuage de points ;

— calculer et interpréter des indicateurs statistiques ;

— déterminer et interpréter une droite de régression ;

— distinguer interpolation et extrapolation ;

— exercer un regard critique sur un modèle.

7.3 Difficultés et erreurs fréquentes

Les principales difficultés concernent :

— la confusion entre covariance et corrélation ;

— l'interprétation causale abusive d'une corrélation ;

— l'utilisation d'une droite sans examiner le nuage ;

— la confusion entre les régressions de Y en X et de X en Y ;

— l'extrapolation au-delà du domaine observé ;

— la confusion entre une estimation et une valeur réelle.

8 Limites de l'approche

Une série statistique à deux variables permet de décrire une association observée, mais ne suffit pas à établir une relation de cause à effet.

La corrélation peut être influencée par des valeurs atypiques, par un choix particulier de population ou par une troisième variable.

Par ailleurs, une liaison non linéaire peut être mal décrite par un modèle affine. Il convient donc de toujours commencer par examiner graphiquement les données.

Enfin, les prévisions obtenues par régression ne sont fiables que dans la mesure où le modèle choisi est adapté et où les conditions ayant produit les données restent pertinentes.

9 Conclusion

L'étude d'une série statistique à deux variables numériques repose sur une démarche complémentaire : représentation graphique, calcul d'indicateurs, ajustement éventuel et interprétation critique.

La covariance et le coefficient de corrélation permettent de caractériser la liaison linéaire, tandis que la droite de régression fournit un modèle d'estimation lorsque sa pertinence est justifiée.

Dans une perspective d'enseignement, l'essentiel est de faire comprendre que les outils statistiques éclairent les données, mais que leur utilisation nécessite toujours une analyse du contexte et des limites du modèle.

10 Références bibliographiques

Références

[1] Ministère de l'Éducation nationale, *Programmes de mathématiques du lycée général et technologique*, notamment les parties relatives aux statistiques descriptives et à l'analyse de données. Bulletins officiels de l'Éducation nationale. https://www.education.gouv.fr/

[2] Ministère de l'Éducation nationale, *Ressources d'accompagnement des programmes de mathématiques du lycée*. Éduscol. https://eduscol.education.fr/

[3] K. Pearson, "Notes on regression and inheritance in the case of two parents", *Proceedings of the Royal Society of London*, vol. 58, 1895, p. 240–242.

[4] P. Dagnelie, *Statistique théorique et appliquée*, tomes 1 et 2, De Boeck Supérieur.

[5] G. Saporta, *Probabilités, analyse des données et statistique*, Éditions Technip, 2006.

[6] A. Monfort, *Cours de statistique mathématique*, Économica.