CAPES EXTERNE DE MATHEMATIQUES

Epreuve sur dossier

Thème : Statistiques, statistiques à deux variables

1. L'exercice proposé au candidat

(d'après Bac ES France 1999)

Le tableau suivant donne l'indice mensuel des dépenses d'assurance maladie d'août 94 à juin 95 (tendances observées à fin juillet 1995 - base 100 janvier 1990).

Mois	Août 94	Octobre 94	Déc. 94	Février 95	Avril 95	juin 95
Rang du mois xi	1	3	5	7	9	11
Indice yi	123,4	125,9	127,5	127,9	129	131,4

(Source : Département statistique de la Caisse Nationale de l'Assurance Maladie des Travailleurs Salariés).

1. Représenter le nuage de point dans un repère orthogonal en plaçant le point G, point moyen du nuage.

2. *Le modèle étudié dans cette question sera appelé « droite de Mayer ».*

a. G1 désigne le point moyen des trois premiers points du nuage et G2 celui des trois derniers points. Déterminer les coordonnées de G1 et de G2 puis l'équation réduite de la droite (G1G2) sous la forme $y = Ax + B$. Tracer la droite (G1G2).

b. En utilisant la calculatrice, déterminer la somme des résidus pour cet ajustement affine : $S_1 = \sum_{i=1}^{6} (yi - Axi - B)^2$

3. *Le deuxième modèle proposé est celui des moindres carrés.*

La calculatrice donne :

- L'équation de la droite (D) d'ajustement de y en x : $y = 0{,}71x + 123{,}26$.
- La somme des résidus pour cet ajustement $S_2 = 1{,}7$ (arrondie avec un chiffre après la virgule).

a. Des droites (D) et (G1G2) quelle est celle qui réalise le meilleur ajustement affine ? Justifier.

b. Tracer (D) sur le graphique précédent.

4. a. Quels sont les indices mensuels que l'on pouvait prévoir en utilisant l'ajustement affine par la méthode des moindres carrés (question 3) pour les mois cités dans le tableau ci-dessous ?

	16	17	18
Mois	nov. 95	déc. 95	jan. 96
Indices prévisionnels calculés par l'ajustement affine des moindres carrés	134,62	135,33	136,04
Tendances réellement observées	134,3	133,4	133,5

b. Quel commentaire peut-on faire ?

2. Travail demandé au candidat.

En aucun cas, le candidat ne doit rédiger sur sa fiche sa solution de l'exercice. Celle-ci pourra néanmoins lui être demandée partiellement ou en totalité lors de l'entretien avec le Jury.

Pendant sa préparation, le candidat traitera les questions suivantes :

Q1) Indiquer le (ou les) niveau(x) auquel(s) peut-être posé cet exercice, les connaissances et les méthodes requises pour sa résolution.

Q2) Quelle réponse donner à la question 3 **a.** « Des droites (D) et (G1G2) quelle est celle qui réalise le meilleur ajustement affine ? »

Q3) Dans le cas général d'un nuage de points $(x_i ; y_i)$, calculer une équation de la droite de régression de Y en X par la méthode des moindres carrés.

Sur sa fiche, le candidat rédigera et présentera :

- sa réponse à la question Q3)
- un ou deux exercices sur le thème des statistiques à deux variables.

Thème : Statistiques à deux variables

1) L'exercice proposé au candidat

1) Cf Calculatrice. $G(5 ; 127,5166) \sim 127,5166 = \frac{7651}{60}$.

2) a) $G_1(3 ; 125,6)$

$G_2(9 ; 129,4333..) \sim 129,4333 = \frac{3883}{30}$.

$(G_1G_2) : y = \frac{23}{36}x + \frac{7421}{60} \sim 0,63888x + 123,683333$

b) $S_1 = \sum_{i=1}^{6} (y_i - Ax_i - B)^2 \simeq 2,06$

3) a) Cf question du jury.

b) Cf calculatrice

4) a) Cf tableau

b) L'erreur entre les indices prévisionnels et les tendances observées s'accroît du fait que la période de relevés statistiques est nettement trop restreinte.

2) Le travail demandé au candidat

Q1) Les statistiques apparaissent

- en 5ème : Lire et interpréter un tableau, un diagramme à barres, un diagramme circulaire ou semi-circulaire
  Regrouper des données statistiques en classes, calculer des effectifs.
  Calcul de fréquences

- en 4ème : Calculer des effectifs cumulés, des fréquences cumulées
  Calculer la moyenne d'une série statistique
  Calculer une valeur approchée de la moyenne d'une série statistique regroupée en classes d'intervalles.

- en 3ème : Caractéristiques de position d'une série statistique (médiane)
  Approche de caractéristiques de dispersion d'une série statistique. (l'étendue)

- en seconde : • Résumé numérique par une ou plusieurs mesures de tendance centrale (moyenne, médiane, classe modale, moyenne élaguée) et une mesure de dispersion (étendue)
  • Définition de la distribution des fréquences d'une série prenant un petit nombre de valeurs et de la fréquence d'un événement
  Simulation et fluctuation d'échantillonnage.

- en première : S : dispersion : écart interquartile et écart-type.
  ES : tableau de contingence, fréquences marginales
  Fréquences conditionnelles. taux d'évolution, moyenne mobile

- Tem S → lien avec les probabilités
- Tem ES → stats à deux variables

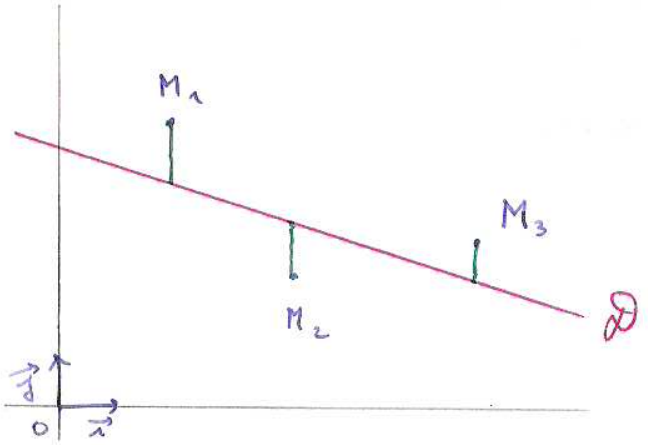
on a un nuage de points, et on cherche à prévoir des tendances pour le futur (extrapolation) ou estimer des valeurs manquantes dans l'intervalle connu (interpolation). → Cf calculette et présenter l'exercice.

Connaissances et méthodes :

- Représenter une série statistique à deux variables graphiquement (nuage de points).
- Calculer les coordonnées du point moyen G $(\overline{x}\,;\overline{y})$
- Connaissant les coordonnées de deux points, trouver l'équation de la droite passant par ces deux points (résolution d'un système)
- Calculer la somme des résidus (calculatrice)
- Tracer une droite en connaissant son équation
- Interpréter l'erreur constatée entre une tendance prévisionnelle et une tendance observée.

Q2) Parfois, le nuage de points associé à une série statistique à deux variables a une forme allongée.
il semble qu'on peut tracer une droite (et même plusieurs) autour de laquelle sont situés les points du nuage.
On dit alors que chacune de ces droites réalise un ajustement affine du nuage.
Il convient alors de se demander si une droite est "meilleure" qu'une autre et si oui selon quel critère.

On considère un nuage de points $(M_i(x_i,y_i))_{1\leq i\leq n}$ et une droite $\mathcal{D}$ d'équation $y=ax+b$.



On appelle somme des résidus associée à la droite $\mathcal{D}$ le nombre réel S défini par $S=\sum_{i=1}^{n}(y_i-(ax_i+b))^2$

Si P_i désigne le point d'abscisse x_i sur $\mathcal{D}$: $S=\sum_{i=1}^{n}M_iP_i^2$

(Faire la somme des résidus ne serait pas satisfaisante car les erreurs positives et négatives peuvent se compenser même si la droite passe loin de tous les points).

On appelle méthodes moindres carrés la méthode qui consiste à rechercher les coefficients a et b tels que la somme S soit minimale.
Dans notre exemple, il est donc logique que $S_2\leq S_1$ et que ce soit (D) qui réalise le meilleur ajustement.

En Tem ES: Seule la méthode des moindres carrés est au programme

Q3 + exo) Cf feuille du jury.

Fiche du Jury.

Réponse à la question Q_3 :

> Dans le cas général d'un nuage de points (x_i, y_i), la droite d'équation $y = ax + b$ qui rend minimale la somme des résidus est la droite :
> - qui passe par le point moyen $G(\bar{x}; \bar{y})$
> - qui a pour coefficient directeur $a = \dfrac{\mathrm{Cov}(X,Y)}{V(X)}$
>
> où $\mathrm{Cov}(X,Y)$ désigne la covariance de X et Y.
> et $V(X)$ désigne la variance de X.

preuve : $S = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \sum_{i=1}^{n} ((y_i - ax_i) - b)^2$

$= n b^2 - 2b \sum_{i=1}^{n} (y_i - ax_i) + \sum_{i=1}^{n} (y_i - ax_i)$

- On suppose a fixé et on considère S comme un polynôme du second degré en b.
  S est donc minimum lorsque $b = \frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i) = \bar{y} - a\bar{x}$

La droite recherchée doit donc vérifier $y - \bar{y} = a(x - \bar{x})$.
Ainsi, parmi toutes les droites de coefficient directeur donné a, celle qui rend S minimale est celle qui passe par le point moyen $G(\bar{x}; \bar{y})$
C'est donc une condition nécessaire.

- Cherchons à présent le coefficient directeur.
On ne considère désormais que les droites qui passent par G.
Pour simplifier les écritures, on se place dans le repère $(G, \vec{i}, \vec{j})$.
Les droites $\mathcal{D}$ ont alors pour équation $Y = aX$ avec les formules de changement de repère définies par : $\forall i \in [\![1, n]\!]$ $X_i = x_i - \bar{x}$ et $Y_i = y_i - \bar{y}$

Relativement à $(G, \vec{i}, \vec{j})$: $S = \sum_{i=1}^{n} (y_i - ax_i - b)^2$

$= \sum_{i=1}^{n} ((y_i - \bar{y}) - a(x_i - \bar{x}))^2$

$= \sum_{i=1}^{n} (Y_i - aX_i)^2$

$= \sum_{i=1}^{n} Y_i^2 - 2a \sum_{i=1}^{n} X_i Y_i + a^2 \sum_{i=1}^{n} X_i^2$

Ce polynôme du second degré en a est minimum lorsque $a = \dfrac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$

ie $a = \dfrac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} = \dfrac{\mathrm{Cov}(X,Y)}{V(X)}$

Exercice supplémentaire :

Le tableau suivant donne l'évolution du nombre de passagers sur une ligne aérienne entre 1994 et 1998 :

Année x_i	1994	1995	1996	1997	1998
Nombre de passagers p_i	7523	9101	12589	18065	47546

1) Tracer le nuage de points $(x_i ; p_i)$
quelle allure a-t-il ?

2) on définit une nouvelle série statistique en posant $(q_i) = (\ln(p_i))$

a) Tracer le nuage de points $(x_i ; q_i)$

b) Calculer l'équation de la droite de régression de q en x.

c) En déduire alors un ajustement de p en x.

3) Combien de passagers peut-on espérer en 2007 ?