Analyse de données

ENSIAS

EPREUVE : M4.6.1 : ECHANTILLONNAGE ET ANALYSE DE DONNEES
DATE : LE 21 MAI 2012
DUREE : 2H15

Directives pédagogiques :

- Deux pages synthèses de documents sont autorisées.
- Il sera tenu compte de la clarté des raisonnements.

Toute tentative de fraude sera sanctionnée par la note zéro.

Vous devez numéroter les figures dans les annexes et les commenter si vous les utilisez dans les justifications des réponses. Bien sur les retourner avec les réponses portant votre nom

Nom : KABit
Prénom : MarouaN
N° SALLE : LC2
N° TABLE: .....

On étudie les performances des athlètes ayant participé en 2004 aux épreuves de décathlon des Jeux Olympiques et du Decastar dont le nom de la variable est **(COMPET)**.

Nous analysons dans cette étude les performances des athlètes pour les dix épreuves du décathlon : course sur 100 m **(c100)**, saut en longueur **(long)**, lancer de poids **(poids)**, saut en hauteur **(haut)**, course sur 400 m **(c400)**, course de haies sur 110 m **(c110)**, lancer de disque **(disq)**, saut à la perche **(perch)**, lancer de javelot **(javel)** et course sur 1500 m **(c1500)**.

Les résultats des épreuves de course sont en secondes, alors que pour les autres, on mesure des distances en mètres. Nous nous intéressons également au rang **(RANG)** et au nombre de points obtenus **(POINTS)**.

Notons que les noms des participants sont en majuscules pour le Decastar, afin de permettre de différencier les participations d'un même athlète aux deux épreuves (Cf. aperçu du fichier des données).

NOM	c100	long	poids	haut	c400	c110	disque	perche	javel	c1500	Rang	Points	COMPET
SEBRLE	11,04	7,58	14,83	2,07	49,81	14,69	43,75	5,02	63,19	291,70	1	8217	Decastar
CLAY	10,76	7,40	14,26	1,86	49,37	14,05	50,72	4,92	60,15	301,50	2	8122	Decastar
KARPOV	11,02	7,30	14,77	2,04	48,37	14,09	48,95	4,92	50,31	300,20	3	8099	Decastar
BERNARD	11,02	7,23	14,25	1,92	48,93	14,99	40,87	5,32	62,77	280,10	4	8067	Decastar
YURKOV	11,34	7,09	15,19	2,10	50,42	15,31	46,26	4,72	63,44	276,40	5	8036	Decastar
WARNERS	11,11	7,60	14,31	1,98	48,68	14,23	41,10	4,92	51,77	278,10	6	8030	Decastar
ZSIVOCZKY	11,13	7,30	13,48	2,01	48,62	14,17	45,67	4,42	55,37	268,00	7	8004	Decastar
McMULLEN	10,83	7,31	13,76	2,13	49,91	14,38	44,41	4,42	56,37	285,10	8	7995	Decastar
MARTINEAU	11,64	6,81	14,57	1,95	50,14	14,93	47,60	4,92	52,33	262,10	9	7802	Decastar
HERNU	11,37	7,56	14,41	1,86	51,10	15,06	44,99	4,82	57,19	285,10	10	7733	Decastar
BARRAS	11,33	6,97	14,09	1,95	49,48	14,48	42,10	4,72	55,40	282,00	11	7708	Decastar
NOOL	11,33	7,27	12,68	1,98	49,20	15,29	37,92	4,62	57,44	266,60	12	7651	Decastar
BOURGUIGNON	11,36	6,80	13,46	1,86	51,16	15,67	40,49	5,02	54,68	291,70	13	7313	Decastar
Sebrle	10,85	7,84	16,36	2,12	48,36	14,05	48,72	5,00	70,52	280,01	1	8893	OlympicG
Clay	10,44	7,96	15,23	2,06	49,19	14,13	50,11	4,90	69,71	282,00	2	8820	OlympicG
Karpov	10,50	7,81	15,93	2,09	46,81	13,97	51,65	4,60	55,54	278,11	3	8725	OlympicG
Macey	10,89	7,47	15,73	2,15	48,97	14,56	48,34	4,40	58,46	265,42	4	8414	OlympicG
Warners	10,62	7,74	14,48	1,97	47,97	14,01	43,73	4,90	55,39	278,05	5	8343	OlympicG

L'étude comporte quatre parties indépendantes mais il est recommandé de faire l'analyse selon l'ordre logique :

- **Partie 1 :** Analyse exploratoire des points **(Points)** obtenus par les athlètes aux épreuves de décathlon aussi bien dans des Jeux Olympiques qu'à Decastar (Voir les sorties de SPSS données en Annexe 1).
- **Partie 2 :** Etude de la liaison linéaire entre les rangs **(Rang)** des athlètes et leurs performances **aux dix épreuves** toute compétition confondue à l'aide d'une **régression multiple pas à pas** (Voir les sorties de SPSS données en Annexe 2)
- **Partie 3 : Analyse en composantes principales** pour obtenir une typologie des athlètes basées sur leurs performances aux 10 épreuves toute compétition confondue. (Voir les sorties de SPSS données en Annexe 3 et 4)
- **Partie 4 : Classification des athlètes** ayant participés aux jeux Olympiques G (Voir les sorties de SPSS données en Annexe 5)

Partie 1 : Analyse exploratoire (Voir Annexe 1)

Q1.1 : Faire une analyse exploratoire des points obtenus par les athlètes dans chaque compétition en interprétant les différents indicateurs vus dans le cours (indicateurs de position central, de dispersion et de forme) et ce pour chaque compétition.

a) Analyse exploratoire des points obtenus dans les jeux les jeux Olympics
b) Analyse exploratoire des points obtenus dans les jeux les jeux Decastar
c) Donnez les valeurs de Q1 : ....., Q3 : ..... dans les jeux Olympics
d) Quel est le nombre minimal de points obtenus par 5% des athlètes dans les jeux Olympics : ..... ?

Q1.2 : En se basant sur le test de normalité de Kolmogorof Smirnov, que pouvez vous conclure sur :
a) La normalité de la distribution des rangs dans les jeux Olympics ? justifier
b) La normalité de la distribution des rangs dans les jeux Decastar ? justifier

Q1.3 : En effectuant **le test de comparaison de moyenne** des points obtenus par les athlètes dans les deux compétitions, peut-on affirmer ou infirmer l'égalité des moyennes des points ? Justifiez votre réponse.

Q1.4 : Interprétez la boite à moustache des points obtenus dans les deux compétitions en termes de symétrie, de comparaison de moyenne, d'existence de données extrêmes, etc.

Partie 2 : Régression linéaire multiple pas à pas (Voir Annexe 2)

Q2.1 : Question de cours : Décrire les différentes étapes de construction d'un modèle de régression linéaire.

Q2.2 : Validation de l'étape 1 : Analysez la matrice des diagrammes de dispersion entre le rang obtenu par les athlètes et les résultats des épreuves. Existe-t-il une liaison linéaire entre le Rang et chaque performance ? Justifiez en commentant le graphique des nuages de points dans l'Annexe 2.

Q2.3 : La Régression pas à pas a convergé en deux itérations et a aboutit au modèle 2 donné en Annexe2. Quelle est la qualité d'ajustement de ce modèle ? ..... Justifier

Q2.4 : Effectuez l'estimation des paramètres du modèle linéaire entre le Rang et les différentes performances retenues par le modèle en justifiant la significativité.

a) Quels sont les performances qui sont le plus liées avec le Rang ?
b) En déduite l'équation du modèle après avoir validé leur significativité.

Q2.5 : Analyser les résidus et vérifiez si les hypothèses de validation du modèle de régression sont vérifiées en justifiant par les différentes figures données en Annexe 2. Retenez-vous alors le modèle de la question 2.4 suite à tous vos justificatifs ?

Q2.6 : Si on voulait prédire le Rang d'un Athlète futur dont les performances sont similaires à celle de CLAY et BOURGUIGNON à Descastar, comment faut-il calculer l'équation du modèle ? Proposez une démarche.

Partie 3 : Analyse en composantes principales (voir Annexe 3 et 4)

Q3.1 : Faut-il centrer et réduire les performances aux épreuves pour conduire l'ACP ? Justifiez
Q3.2 : En analysant **la matrice de corrélation**,
a) quelles sont les couples de variables remarquables (les plus corrélées, les moins corrélées, les plus opposées) ?
b) Y a-t-il une corrélation entre les performances de courses et les performances de lancer et de saut ?
Q3.3 : Peut-on conduite une ACP ? Justifiez par tous les indicateurs qui valident la première étape de l'ACP.
Q3.4 : Quels sont les performances aux épreuves qui vont le plus contribuer à la construction des facteurs ? Justifiez en spécifiant la sortie qui vous permet de répondre à cette question.
Q3.5 : En analysant le graphique des valeurs propres et des inerties expliqués, combien faut-il retenir de nombre de facteurs ? ..... Justifiez. Expliquer la démarche qui consiste à retenir un nombre de facteur.

Q3.6 : Nous avons extrait 4 facteurs par l'ACP.
a) Quel est le % d'inertie expliquée par les 4 facteurs ?
b) Quelle est la qualité de représentation de ces quatre facteurs ? Quelles sont les performances mal expliquées ?
c) Combien de résidus restent-ils dans le modèle ? .....

Q3.7 : Pour interpréter les 4 facteurs,

a) Est-il nécessaire de faire une rotation ? Justifiez
b) Interprétez les facteurs suite à votre réponse de a)

INT. Modèle d'interprétation : Le facteur Z est expliqué par : performance X (+) ou (-) & performance Y (+) ou (-), etc. et proposez un nom en terme de performance de courses ou d'endurances, de lancer, de saut etc.

Q3.8 : Interprétez les groupes d'Athlètes 1 et 2 des graphiques de l'Annexe 4. Citez leur plus grandes et/ou plus faibles performances par compétition. Faite une analyse de performance d'un même Athlète qui se trouve dans les deux groupes ou dans le même groupe.

Q3.9 : Interprétez le tableau de performances extrêmes par rapport aux différents facteurs.

Q3.10 : Une fois que vous avez effectué l'analyse sur les 4 facteurs, pensez vous qu'ils sont suffisants pour interpréter les performances des athlètes aux différentes compétions ou fallait-il pousser l'analyse à d'autres dimension ?

Partie 4 : Classification hiérarchique (voir Annexe 5)

Q4.1 : Nous avons effectué une classification hiérarchique pour regrouper les athlètes ayant participé uniquement aux jeux Olympiques. D'après la chaine des agrégations, Combien retenez-vous de classes ?

Q4.2 : D'après l'arbre hiérarchique,

a) quelle méthode d'agrégation a été utilisée ?
b) D'après la coupe proposée sur la figure, combien de classes reste t-il ? Donnez-les ?

Q4.3 : Nous avons décidé de retenir pour la suite de l'analyse trois classes. D'après le tableau d'ANOVA, quels sont les performances qui ont le pouvoir le plus discriminant (i.e. qui séparent le plus les classes) ?

Q4.4 : Interprétation des classes

a) Donnez une interprétation de chacune des classes 1 et 3 en termes d'effectif et de performances en se basant sur le cube OLAP.
b) En quoi performent alors les athlètes de chaque classe ?

Q4.5 : Est-ce que les résultats de la classification sont cohérents avec ceux de l'ACP ?

***B**onne **C**hance*

RECOMMANDATIONS

- Avant de répondre, prenez le temps de lire l'ensemble des questions. Il s'agit là d'une étude complète de bout en bout sur une même étude.

- Il y a un enchaînement logique des questions afin que l'analyse soit complète à la fin de l'épreuve. Ne vous précipitez donc pas à répondre en sautant des questions. Le barème établi ne favorise pas forcément ceux qui répondent au maximum de questions mais plutôt ceux qui analyseront correctement les données de manière logique. J'accepte également toute analyse détaillée complémentaire aux questions posées.

- Pour justifier vos réponses, référer dans le texte aux sorties correspondantes (la figure ou le tableau de SPSS) après les avoir numéroté dans les pages de sortie et commenté en y incluant une légende.