DISPOSITIF FORMATION STATISTIQUE
MODULE “MODÈLE LINÉAIRE GÉNÉRALISÉ”

QUESTIONNAIRE DE POSITIONNEMENT DES CONNAISSANCES - CORRIGÉ

Ce questionnaire est destiné à faire connaître au formateur le niveau de connaissance des stagiaires qu'il doit former. Il l'utilisera pour composer des groupes homogènes, adapter la formation au groupe et éventuellement proposer à ceux qui n'ont pas le minimum nécessaire de connaissances de suivre au préalable un autre module du dispositif, le module 1 “Introduction à la décision statistique” ou le module 2/3 “Le modèle linéaire”.

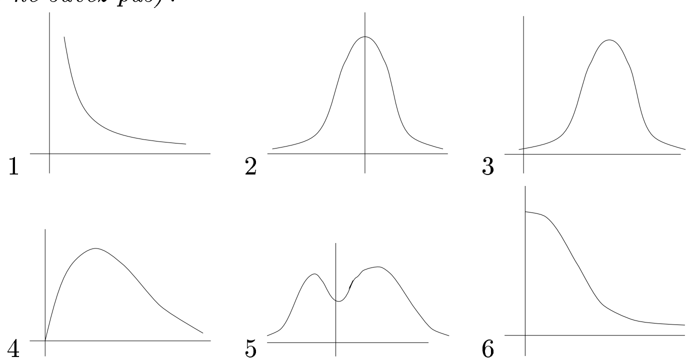
1 Probabilités

1.1

Si la variance de la variable X est σ^2, complétez :
$Var X = \sigma^2$
$Var(X+1) = \sigma^2$
$Var(2X) = 4\sigma^2$
$Var(-2X+1) = 4\sigma^2$
écart-type$(X) = \sigma$
écart-type$(-X) = \sigma$
écart-type$(2X) = 2\sigma$
écart-type$(2X+1) = 2\sigma$

1.2

Les graphiques suivants représentent la fonction de densité d'une variable gaussienne
(entourer les bonnes réponses, barrer les mauvaises réponses, ne faites rien si vous ne savez pas) :



entourer 2 et 3, barrer le reste.

1.3

Si deux variables aléatoires sont indépendantes alors : *(entourer les bonnes réponses, barrer les mauvaises réponses, ne faites rien si vous ne savez pas)*
(a) $Var(X+Y) = VarX + VarY$
(b) la covariance de X et Y est nulle
(c) la corrélation de X et Y est nulle
(d) quand X augmente, Y augmente
(e) quand X augmente, Y diminue
(f) la loi de Y est la même que la valeur de X soit connue ou non.
entourer a, b, c, f, barrer le reste.

2 Statistiques

2.1

Le test de Student pour un échantillon sert à : *(entourer les bonnes réponses, barrer les mauvaises réponses, ne faites rien si vous ne savez pas)*
(a) calculer la moyenne théorique (ou moyenne sur toute la population) d'une variable
(b) estimer la moyenne théorique d'une variable
(c) comparer la moyenne d'un échantillon à une valeur donnée
(d) comparer la moyenne théorique à une valeur donnée en utilisant la moyenne et la variance d'un échantillon
entourer d, barrer a et c, entourer aussi b.

2.2

On se demande si les rendements en blé sont meilleurs quand le cultivateur a apporté 100 unités d'azote à l'ha plutôt que 50 unités à l'ha, au début du tallage. On dispose de :

- 8 mesures de rendement avec 100 unités, de moyenne $\overline{x}$
- 11 mesures de rendement avec 50 unités, de moyenne $\overline{y}$.

L'estimation de la variance commune s'obtient en divisant la somme des 2 sommes de carrés d'écarts à la moyenne, par la somme des degrés de liberté : 7+10=17.
La statistique de Student

$$T = \frac{\overline{x} - \overline{y}}{\text{écart-type estimé de } (\overline{x} - \overline{y})}$$

vaut

$$T = \frac{27.4}{18.038} = 1.52$$

A quoi comparez-vous cette valeur pour faire un test de niveau 5 % (ou 0.05)? Quelle est votre réponse à la question? (utilisez la table donnée en annexe).
Je la compare à $1.74 = t(17)_{5\%}$. *Les rendements ne sont pas meilleurs avec 100 unités d'azote à l'ha, autant n'apporter que 50 unités d'azote à l'ha.*

2.3

Le niveau d'un test de comparaison de deux populations est : *(entourer les bonnes réponses, barrer les mauvaises réponses, ne faites rien si vous ne savez pas)*
(a) 5 %
(b) 10 %
(c) ce qu'on choisit comme risque d'erreur de 1ère espèce
(d) la probabilité de se tromper quand les deux populations sont différentes
(e) la probabilité de se tromper quand les deux populations sont identiques
(f) la probabilité que les deux populations soient identiques.
entourer c et e, barrer le reste.

2.4

Compléter le texte suivant :
A partir d'un échantillon $X_1, X_2, \ldots, X_9$ on calcule la moyenne de l'échantillon notée $\overline{X}$. La moyenne théorique ou moyenne de la population est notée μ. La moyenne de l'échantillon est une variable aléatoire. C'est une estimation de $\boxed{mu}$. Si σ^2 est la variance théorique de chaque X_i, la variable $S^2 = \frac{1}{8}\sum_i (X_i - \overline{X})^2$ est $\boxed{\textit{une estimation}}$ de σ^2. La variance de $\overline{X}$ vaut $\boxed{\frac{\sigma^2}{n}}$ et peut s'estimer avec $\boxed{\frac{\widehat{\sigma^2}}{n} = \frac{S^2}{n}}$. Si les nombres t et γ sont tels que

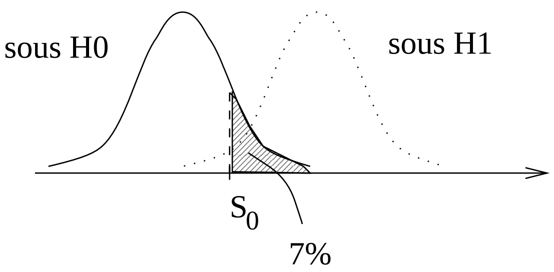
$$Proba\left\{-t \le \frac{\overline{X} - \mu}{\left(\frac{S}{3}\right)} \le t\right\} = \gamma$$

l'intervalle aléatoire $[\overline{X} - tS/3, \overline{X} + tS/3]$ est un intervalle de confiance autour de $\boxed{\overline{X}}$. Il contient $\boxed{\mu}$ avec une probabilité égale à $\boxed{\gamma}$. Si la distribution de chaque X_i est gaussienne, les valeurs de t et $gamma$ se lisent dans une table de $\boxed{\textit{student}}$ à $\boxed{8}$ degrés de liberté.

2.5

Pour décider entre H_0 ou H_1 vous voulez utiliser une statistique S calculée à partir des données. Cette statistique a une distribution tabulée sous H_0 et une distribution décalée vers la droite sous H_1. Sur votre expérience cette statistique vaut S_0. Dans la table vous lisez que sous H_0 : $Pr(S > S_0) = 0.07$. Faire un schéma pour résumer ce problème et compléter le texte suivant :
Au niveau 0.05 vous $\boxed{\textit{acceptez}}$ l'hypothèse H_0.

Par contre au niveau $\boxed{10\%}$ vous rejetez l'hypothèse H_0.



3

On ajuste un modèle de régression multiple, et on obtient les résultats suivants :

$$\widehat{Y} = 1.2 + 0.5X_1 - 0.3X_2$$

La corrélation entre le coefficient de X_1 et le coefficient de X_2 est estimé à 0.88. L'interprétation est la suivant :

- une augmentation de X_1 provoque une augmentation de Y,
- une augmentation de X_2 provoque une diminution de Y.

Vrai ou faux? argumenter.
Ça n'est pas forcément vrai : la corrélation entre les paramètres étant élevée, les paramètres ne peuvent pas être interprétés indépendamment l'un de l'autre.

4

Supposons que M_0 et M_1 sont deux modèles linéaires, tels que M_1 est un sous-modèle de M_0. On note SC_1 la somme de carrés résiduels du modèle 1, dl_1 le nombre de degrés de liberté résiduels associés, SC_0 la somme de carrés résiduels du modèle 0, dl_0 le nombre de degrés de liberté résiduels associé. Quelle statistique de test utilisez-vous pour comparer le modèle M_1 au modèle M_0 ? donnez son expression.

$$\frac{\frac{SC_1 - SC_0}{dl_1 - dl_0}}{\frac{SC_0}{dl_0}}$$

où SC_0 et SC_1 sont les sommes de carrés résiduels des modèles M_0 et M_1 respectivement, et où dl_0 et dl_1 sont les degrés de libertés résiduels associés.

5

Parmi les modèles qui suivent, trouver les séquences de modèles emboîtés. Un même modèle peut apparaître dans plusieurs séquences. Quels sont les modèles non linéaires?

1. $Y = \beta X$
2. $Y = \alpha + \beta\, X + \gamma\, X^2$
3. $Y = \alpha + \beta\, X$
4. $Y = \alpha$
5. $Y = \alpha + \beta\, X^\gamma$
6. $Y = \alpha + \beta\, X + \gamma\, X^2, \quad \gamma = 2\,\beta$

On peut trouver 3 séquences de modèles emboîtés :

- $Y = \alpha + \beta\, X + \gamma\, X^2$
  $Y = \alpha + \beta\, X + \gamma\, X^2, \quad \gamma = 2\,\beta$
  $Y = \alpha + \beta\, X$
  $Y = \alpha$

- $Y = \alpha + \beta\, X^\gamma$
  $Y = \alpha + \beta\, X$
  $Y = \alpha + \beta\, X, \quad \beta = 1$

- $Y = \alpha + \beta\, X^\gamma$
  $Y = \alpha + \beta\, X$
  $Y = \beta\, X$

Seul le modèle $Y = \alpha + \beta\, X^\gamma$ est non linéaire (en les paramètres).

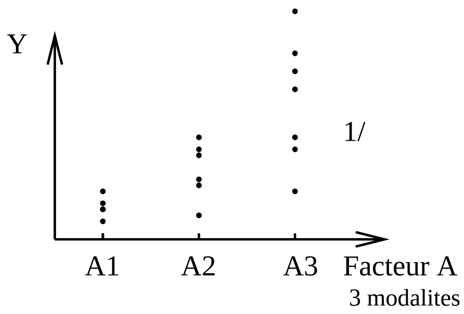
6 Question subsidiaire :

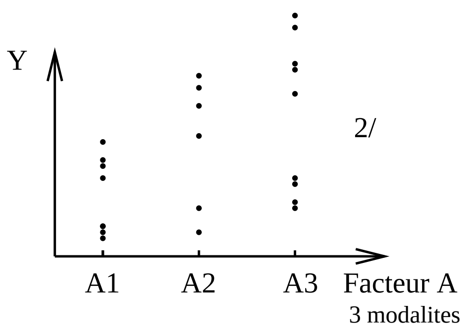
Que definit une famille de lois de probabilité sur R^n, lorsqu'elle peut être plongée de manière lisse dans une structure canonique saturée et régulière?
C'était un brin d'humour ...

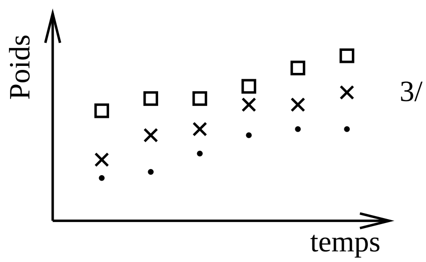
7

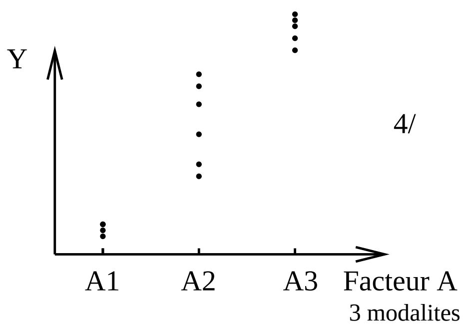
Pour chaque graphe, on ajuste un modèle à un facteur (A), ou un régresseur (le temps). Commentez les situations représentées sur les graphiques ci-après en cochant et en remplissant les cases de cette grille. (dans le schéma 3, 3 individus sont suivis dans le temps).

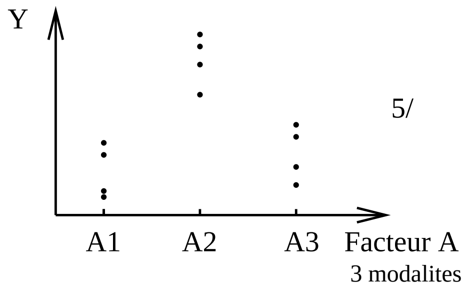
	pas de problème	Problèmes				
		postulat(s) posant problème			autres problèmes	solution(s) possible(s)
		$E(\varepsilon) = 0$	$Var(\varepsilon_i) = \sigma^2$	indépendance		
1			X			transformation monotone
2		X			facteur caché non pris en compte	le chercher
3				X		
4			X			aucune transformation monotone possible
5	X					
6			X			transformation monotone
7			X			aucune transformation monotone possible
8	X					

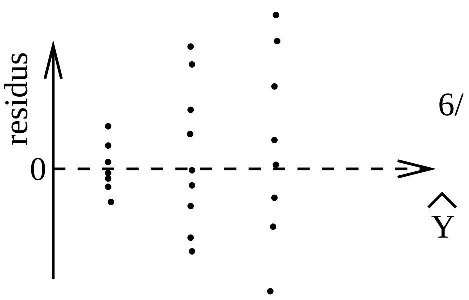

Y
1/
A1
A2
A3
Facteur A
3 modalites

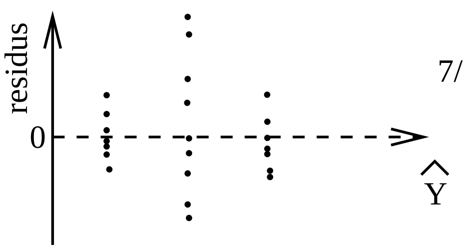

Y
2/
A1
A2
A3
Facteur A
3 modalites

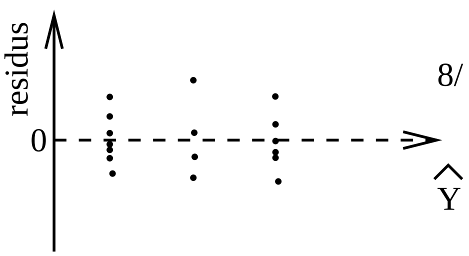

Poids
3/
temps


Y
4/
A1
A2
A3
Facteur A
3 modalites


Y
5/
A1
A2
A3
Facteur A
3 modalites


residus
6/
0
$\hat{Y}$


residus
7/
0
$\hat{Y}$


residus
8/
0
$\hat{Y}$