Frédéric Bertrand
Université Louis Pasteur

Examen de Statistique Appliquée II

****

• *Le cours, les exercices de travaux dirigés, leurs corrigés ainsi que les notes de cours sont autorisés. Tout autre document est interdit.*
• *Les téléphones portables sont formellement interdits.*

• *Les calculatrices sont autorisées.*

• *Tous les tests seront effectués au seuil de signification $\alpha = 5$ %.*

• *Les deux exercices sont indépendants.*

• *Afin de pouvoir traiter les questions, plusieurs résultats numériques et graphiques ont été intégrés au document.*
• *On prendra un soin particulier à préciser quelles sont les hypothèses testées.*

Durée de l'épreuve 2 heures

****

Exercice 1. Détection de patterns

Une expérience est conçue pour étudier la détection de certains patterns théoriquement importants sur un écran. Les patterns varient par leur forme et leur solidité. La variable dépendante est un Score de Reconnaissance, arrondi à l'entier le plus proche et que l'on considérera comme une variable quantitative continue, et les facteurs sont la Forme (Cercle, Carré ou Triangle) et la Solidité (Plein ou Contours). Le même échantillon de participants est utilisé pour toutes les combinaisons possibles de traitements. Les valeurs ont été reportées dans le tableau ci-après.

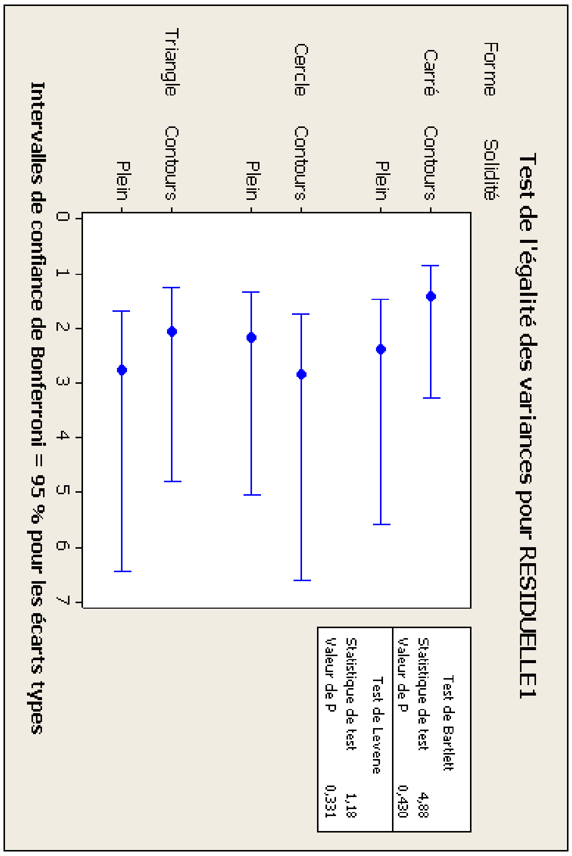
Participant	Forme	Solidité	Score	Participant	Forme	Solidité	Score
1	Cercle	Plein	4	6	Cercle	Plein	3
1	Cercle	Contours	2	6	Cercle	Contours	6
1	Carré	Plein	2	6	Carré	Plein	4
1	Carré	Contours	8	6	Carré	Contours	5
1	Triangle	Plein	7	6	Triangle	Plein	5
1	Triangle	Contours	5	6	Triangle	Contours	10
2	Cercle	Plein	3	7	Cercle	Plein	7
2	Cercle	Contours	6	7	Cercle	Contours	12
2	Carré	Plein	2	7	Carré	Plein	2
2	Carré	Contours	6	7	Carré	Contours	6
2	Triangle	Plein	8	7	Triangle	Plein	4
2	Triangle	Contours	9	7	Triangle	Contours	8
3	Cercle	Plein	2	8	Cercle	Plein	6
3	Cercle	Contours	10	8	Cercle	Contours	10
3	Carré	Plein	2	8	Carré	Plein	9
3	Carré	Contours	5	8	Carré	Contours	5
3	Triangle	Plein	5	8	Triangle	Plein	0
3	Triangle	Contours	3	8	Triangle	Contours	10
4	Cercle	Plein	1	9	Cercle	Plein	4
4	Cercle	Contours	8	9	Cercle	Contours	5
4	Carré	Plein	5	9	Carré	Plein	7
4	Carré	Contours	5	9	Carré	Contours	6
4	Triangle	Plein	2	9	Triangle	Plein	8
4	Triangle	Contours	9	9	Triangle	Contours	12
5	Cercle	Plein	4	10	Cercle	Plein	2
5	Cercle	Contours	6	10	Cercle	Contours	12
5	Carré	Plein	4	10	Carré	Plein	12
5	Carré	Contours	5	10	Carré	Contours	8
5	Triangle	Plein	5	10	Triangle	Plein	10
5	Triangle	Contours	10	10	Triangle	Contours	12

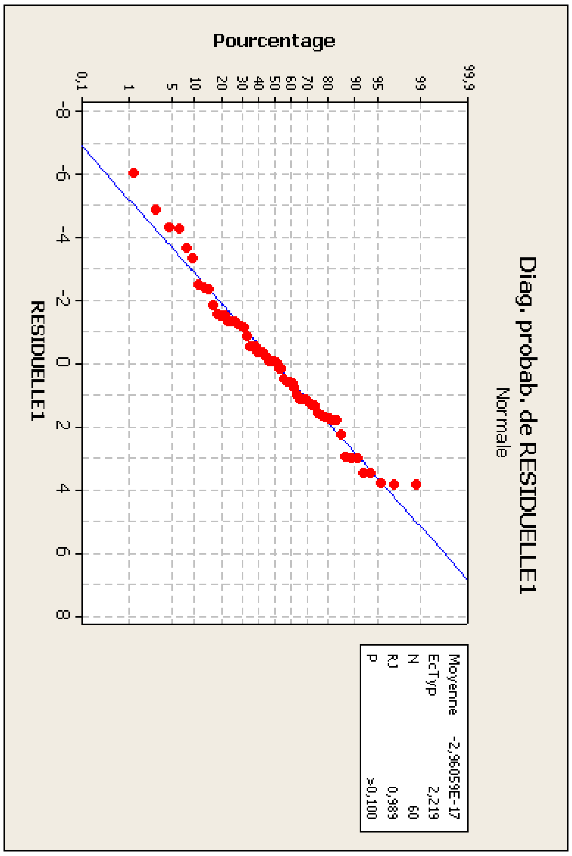
1. Écrire le modèle d'analyse de la variance relatif à cette étude. On précisera la nature des facteurs explicatifs ainsi que les hypothèses faites.

2. Les hypothèses du modèle sont-elles vérifiées ? On se contentera de vérifier les hypothèses pour lesquelles des graphiques sont fournis. On mentionnera explicitement celles qui restent et que l'on suppose désormais vérifiées.

3. Procéder à l'étude du tableau de l'analyse de la variance.

4. Compte tenu de la nature des facteurs peut-on procéder à des comparaisons multiples ? Il y a-t-il un facteur pour lequel cette procédure se justifie, si oui procéder aux tests correspondants et interpréter les résultats.





Modèle linéaire général : Score en fonction de Participant; Forme; ...

Facteur	Type	Niveaux	Valeurs
Participant	aléatoire	10	1; 2; 3; 4; 5; 6; 7; 8; 9; 10
Forme	fixe	3	Carré; Cercle; Triangle
Solidité	fixe	2	Contours; Plein

Analyse de la variance pour Score, avec utilisation de la somme des carrés ajustée pour les tests

Source	DL	SomCar séq	SomCar ajust	CM ajust	F	P
Participant	9	109,683	109,683	12,187	1,89	0,079
Forme	2	33,700	33,700	16,850	2,61	0,085
Solidité	1	120,417	120,417	120,417	18,65	0,000
Forme*Solidité	2	26,433	26,433	13,217	2,05	0,141
Erreur	45	290,617	290,617	6,458		
Total	59	580,850				

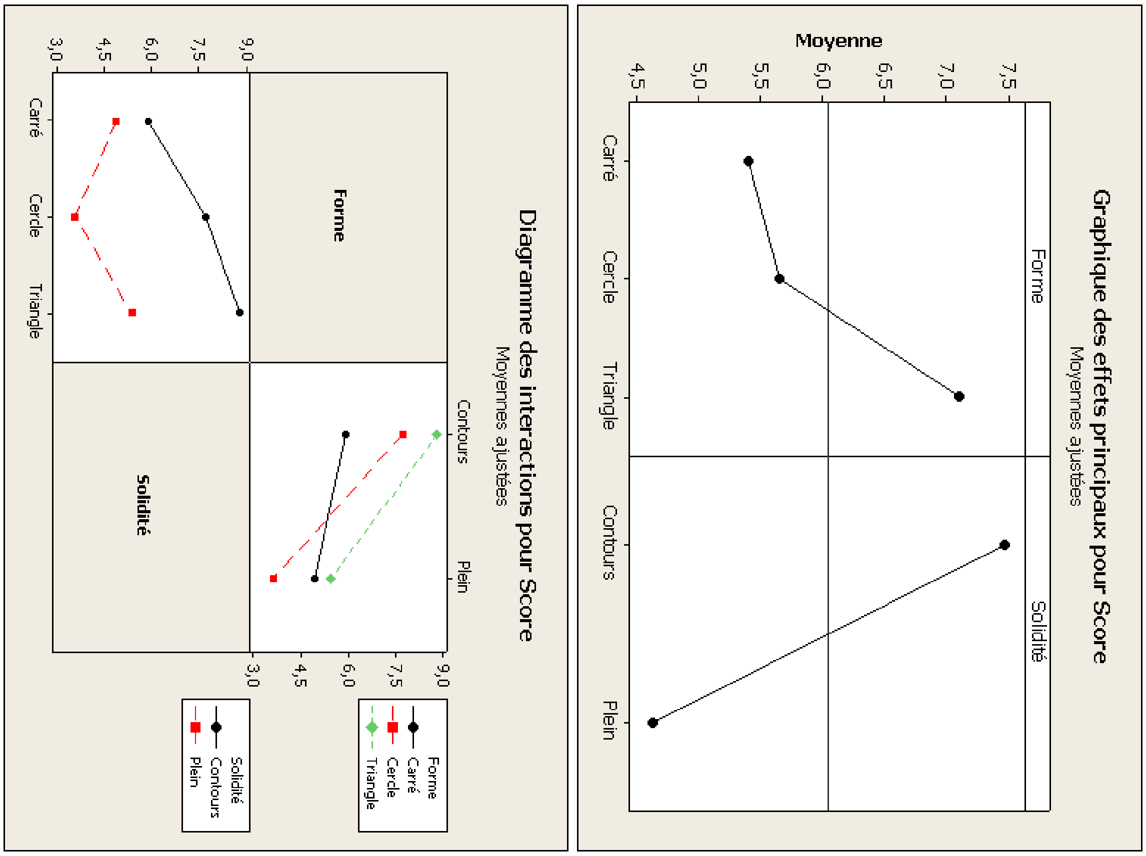
S = 2,54129 R carré = 49,97 % R carré (ajust) = 34,40 %

Terme		Coeff	Coef ErT	T	P
Constante		6,0500	0,3281	18,44	0,000
Participant					
1		-1,3833	0,9842	-1,41	0,167
2		-0,3833	0,9842	-0,39	0,699
3		-1,5500	0,9842	-1,57	0,122
4		-1,0500	0,9842	-1,07	0,292
5		-0,3833	0,9842	-0,39	0,699
6		-0,5500	0,9842	-0,56	0,579
7		0,4500	0,9842	0,46	0,650
8		0,6167	0,9842	0,63	0,534
9		0,9500	0,9842	0,97	0,340
Forme					
Carré		-0,6500	0,4640	-1,40	0,168
Cercle		-0,4000	0,4640	-0,86	0,393
Solidité					
Contours		1,4167	0,3281	4,32	0,000
Forme*Solidité					
Carré	Contours	-0,9167	0,4640	-1,98	0,054
Cercle	Contours	0,6333	0,4640	1,37	0,179

Observations aberrantes pour Score

Observation	Score	Valeur ajustée	Ajust ErT	Valeur résiduelle	Valeur résiduelle normalisée	
47	0,0000	6,0167	1,2706	-6,0167	-2,73	R
55	2,0000	6,8833	1,2706	-4,8833	-2,22	R

R indique une observation ayant une valeur résiduelle normalisée importante




```
Intervalles de confiance simultanés de Tukey = 95,0 %
Variable de réponse Score
Toutes les comparaisons deux à deux sur les niveaux de Forme
Forme = Carré  soustrait de :

Forme      Inférieur  Centré  Supérieur    -+---------+---------+---------+-----
Cercle        -1,699  0,2500      2,199    (-----------*-----------)
Triangle      -0,249  1,7000      3,649             (-----------*-----------)
                                           -+---------+---------+---------+-----
                                          -1,5       0,0       1,5       3,0


Forme = Cercle  soustrait de :

Forme      Inférieur  Centré  Supérieur    -+---------+---------+---------+-----
Triangle     -0,4991   1,450      3,399            (-----------*-----------)
                                           -+---------+---------+---------+-----
                                          -1,5       0,0       1,5       3,0


Tests de simultanéité de Tukey
Variable de réponse Score
```

```
Toutes les comparaisons deux à deux sur les niveaux de Forme
Forme = Carré  soustrait de :

                        Erreur type
           Différence        de la  Valeur  Valeur de
Forme     des moyennes   différence    de T  p ajustée
Cercle          0,2500       0,8036  0,3111     0,9481
Triangle        1,7000       0,8036  2,1154     0,0981


Forme = Cercle  soustrait de :

                        Erreur type
           Différence        de la  Valeur  Valeur de
Forme     des moyennes   différence    de T  p ajustée
Triangle         1,450       0,8036   1,804     0,1797


Intervalles de confiance simultanés de Tukey = 95,0 %
Variable de réponse Score
Toutes les comparaisons deux à deux sur les niveaux de Solidité
Solidité = Contours  soustrait de :

Solidité  Inférieur  Centré  Supérieur  -----+---------+---------+---------+-
Plein        -4,155  -2,833     -1,512  (----------*----------)
                                        -----+---------+---------+---------+-
                                          -3,6      -2,4      -1,2       0,0


Tests de simultanéité de Tukey
Variable de réponse Score
Toutes les comparaisons deux à deux sur les niveaux de Solidité
Solidité = Contours  soustrait de :

                        Erreur type
           Différence        de la  Valeur  Valeur de
Solidité  des moyennes   différence    de T  p ajustée
Plein           -2,833       0,6562  -4,318     0,0001
```

Exercice 2. Condition de sang et consommation quotidienne moyenne de cigarettes et d'alcool

Le tableau suivant montre les huit premières observations sur l'existence d'une condition de sang pré-morbide chez 100 personnes différentes avec leur moyenne de consommation de cigarettes et d'alcool par jour. Les unités ont été sélectionnées pour couvrir l'intervalle entier de consommation pour chaque variable : une unité de Fumer est dix cigarettes et une unité de Boire est l'équivalent d'un verre de vin.

Individu	Sang	Fumer	Boire	Individu	Sang	Fumer	Boire
1	Oui	7	18	5	Oui	5	11
2	Oui	6	15	6	Oui	2	18
3	Oui	1	10	7	Non	0	0
4	Oui	7	16	8	Oui	6	12

1. Écrire le modèle de régression permettant d'expliquer la condition du sang à l'aide de toutes les variables explicatives proposées mais ne comportant pas de terme d'interaction. On précisera la nature des facteurs explicatifs ainsi que les hypothèses faites.
2. Les hypothèses du modèle sont-elles vérifiées ? Calculer les estimations de tous les paramètres du modèle.
3. Ce modèle est-il adéquat ?
4. Ce modèle est-il intéressant ?
5. Quelles sont les variables explicatives qui influent significativement sur la condition du sang ?
6. Est-il pertinent de simplifier le modèle introduit en **1.** ? Les résultats obtenus avec ce nouveau modèle sont-ils en accord avec les précédents ?

```
Régression logistique binaire : Sang en fonction de Fumer; Alcool

Fonction de liaison : Logit


Informations de réponse

Variable  Valeur  Dénombrement
Sang      1                 44  (Evénement)
          0                 56
          Total            100


Tableau de régression logistique

                                                         Rapport des
```

Prédicteur	Coeff	Coef ErT	Z	P	probabilités de succès	IC à 95 % Inférieur
Constante	-1,39440	0,372948	-3,74	0,000		
Fumer	2,26416	0,512861	4,41	0,000	9,62	3,52
Alcool	-0,0781604	0,0849944	-0,92	0,358	0,92	0,78

Prédicteur	Supérieur
Constante	
Fumer	26,29
Alcool	1,09

Log de vraisemblance = -38,999
Vérifier que toutes les pentes sont à zéro : G = 59,187, DL = 2, Valeur de p = 0,000

Tests d'adéquation de l'ajustement

Méthode	Khi deux	DL	P
Pearson	17,6285	23	0,777
Somme des carrés d'écart	14,0732	23	0,925
Hosmer-Lemeshow	3,8283	4	0,430

Mesures d'association :
(entre la variable de réponse et les prévisions de probabilité)

Paires	Nombre	Pourcentage	Mesures récapitulatives	
Concordant	2151	87,3	D de Somers	0,82
Discordant	131	5,3	Gamma de Goodman-Kruskal	0,89
Ex aequo	182	7,4	Tau a de Kendall	0,41
Total	2464	100,0		

Régression logistique binaire : Sang en fonction de Fumer

Fonction de liaison : Logit

Informations de réponse

Variable	Valeur	Dénombrement	
Sang	1	44	(Evénement)
	0	56	
	Total	100	

Tableau de régression logistique

Prédicteur	Coeff	Coef ErT	Z	P	Rapport des probabilités de succès	IC à 95 % Inférieur
Constante	-1,56091	0,332617	-4,69	0,000		
Fumer	2,19947	0,527099	4,17	0,000	9,02	3,21

Prédicteur	Supérieur
Constante	
Fumer	25,35

Log de vraisemblance = -39,398
Vérifier que toutes les pentes sont à zéro : G = 58,389, DL = 1, Valeur de p = 0,000

Tests d'adéquation de l'ajustement

Méthode	Khi deux	DL	P
Pearson	1,49289	7	0,983
Somme des carrés d'écart	1,22893	7	0,990

Mesures d'association :
(entre la variable de réponse et les prévisions de probabilité)

Paires	Nombre	Pourcentage	Mesures récapitulatives	
Concordant	1843	74,8	D de Somers	0,72
Discordant	61	2,5	Gamma de Goodman-Kruskal	0,94
Ex aequo	560	22,7	Tau a de Kendall	0,36
Total	2464	100,0		