Université Montpellier 2 Année 2016-2017
M2 ICAP

TD 1 : Statistique descriptive

1 Rappels sur les variables aléatoires

Exercice 1. Somme de deux variables aléatoires

1. On lance deux dés non pipés. Quelle est la moyenne de la somme ? la variance de la somme ? la loi de la somme ?
2. Calculer la loi de la somme de 2 v.a. discrètes indépendantes. Interpréter graphiquement.

Exercice 2. Loi Binomial Soit X et Y deux v.a indépendante. On suppose que $X \sim \mathcal{B}(n_1, p_2)$ et $Y \sim \mathcal{B}(n_2, p_2)$ où $n_1, n_2 \in \mathbb{N}$ et $0 < p_1, p_2 < 1$.

1. Si $p_1 = p_2$: Interpréter les quantités X, Y et $Y + X$. Quelle loi suit $X + Y$?
2. Que peut on dire si $p_1 \neq p_2$?

Exercice 3. Loi Normale On considère deux variables aléatoires indépendantes X et Y. On suppose que X suit une loi normale de moyenne μ_X et d'écart-type σ_X et que Y suit une loi normale de moyenne μ_Y et de variance σ_Y^2.

1. Quel est la loi de aX pour $a \in \mathbb{R}$? Interpréter graphiquement.
2. Quel est la loi de $X + Y$? Interpréter graphiquement.
3. On pose $\mu_X = 50$, $\sigma_X = 10$ et $\mu_Y = 20$, $\sigma_Y^2 = 25$. Calculer $\mathbb{P}(2X - Y \leq 90)$?

Exercice 4. Loi binomiale On souhaite comparer deux recettes de sodas : A contient du sucre et B contient un édulcorant. On dispose d'un jury de 24 testeurs et on fait gouter à chaque juge trois échantillons de ces deux sodas (l'un est présenté deux fois l'autre une seule fois). Ils doivent trouver quel est le produit singulier, mais ils sont incapables de le discerner et il répondent "au hasard" de manière indépendante. On note N la v.a donnant le nombre de bonnes réponses :

1. Quelle est la loi de N ? sa moyenne ? sa variance ?
2. Calculer la probabilité des évènements $N = 0$, $N = 1$, $N = 2$, $N = 3$ et $N > 3$?
3. Approximer la loi de N par une loi normale $\tilde{N}$ dont on déterminera les paramètres. Donner une approximation de la probabilité des évènements $N = 3$ et $N \geq 12$?

2 Estimations ponctuelles et intervalles de confiances

Exercice 5. Estimation de moyenne de v.a. gaussienne On a relevé la note d'amertume donnée à une bière brune par un jury de $n = 10$ experts. Cela donne les résultats suivants

$$5.3, 7.8, 6.1, 8.1, 5.0, 3.8, 3.8, 7.4, 6.1, 6.1$$

On suppose que les observations sont issues d'une loi normale $\mathcal{N}(\mu, \sigma^2)$.

1. En se basant sur les études antérieures, on suppose que $\sigma^2 = 3.5$. Donner un intervalle de confiance pour l'espérance μ au niveau de confiance 0.90.
2. Calculer le niveau de l'intervalle de confiance (symétrique) pour μ donné par $[5.36, 6.54]$ (on suppose toujours que $\sigma^2 = 3.5$).
3. Les techniciens pensent que l'hypothèse $\sigma^2 = 3.5$ est fausse dans ce cas... Ils proposent donc de supposer que σ^2 est inconnue. Calculer alors l'intervalle de confiance à 0.90 pour l'espérance μ.

Exercice 6. Estimation de moyenne On relève cette fois ci $n = 32$ notes d'amertume pour la même bière. On a note X_j la note donnée par le j-ième juge. On a $\sum_j x_j = 176.29$ et $\sum_j x_j^2 = 1023.2$. Calculer un intervalle de confiance au niveau 98% de la note moyenne d'amertume.

3 Statistique bivariée pour les variables qualitatives

Exercice 7. On teste une nouvelle composition de nourriture canine. On présente à un panel de chiens 3 gamelles : une gamelle A témoin avec la recette initiale, une gamelle B contenant la nouvelle composition, et une gamelle C contenant un produit d'une marque concurrente. On observe le comportement du chien lors de la présentation des gamelles et on relève la gamelle vers laquelle il se dirige en premier.

	Chihuahua	Cocker	Border Collie
A	5	8	2
B	4	9	6
C	4	9	4

1. Donner les distributions conditionnelles et la distribution marginale de la variable race. Les représenter sur un même graphique.
2. Calculer la statistique du χ^2.

4 Analyse en composantes principales

Exercice 8. Description d'apéritifs On dispose des notes sensorielles (rond à astringent) et hédoniques (chaleur et profondeur) de 5 juges (A à E) pour 4 apéritifs (1 à 4) :

juge,produit	rond	confit	vert	alcool	fruite	terreux	metal	piquant	acide	amer	astringent	chaleur	profondeur
A1	8	9	3	7	4	2	3	0	5	1	3	8	8
A2	9	9	0	3	7	1	1	0	2	4	1	9	5
A3	2	3	6	4	3	0	1	4	6	7	2	5	5
A4	7	1	2	5	6	5	4	0	2	2	5	7	9
B1	6	8	1	8	4	3	3	1	7	2	5	7	7
B2	7	9	0	3	9	0	2	1	3	3	2	8	5
B3	1	4	5	5	4	1	2	6	6	5	3	7	5
B4	5	2	1	5	7	4	5	2	3	1	6	8	8
C1	10	9	3	9	5	2	2	1	6	3	3	10	9
C2	9	8	1	4	7	0	1	0	2	4	2	9	6
C3	1	2	4	5	3	2	2	5	5	7	2	4	5
C4	6	1	1	6	5	4	4	1	2	1	4	6	7
D1	6	9	1	7	4	1	3	1	7	3	4	7	6
D2	7	9	1	3	7	0	1	1	2	3	2	8	5
D3	1	2	5	4	4	2	2	5	6	7	2	5	5
D4	4	1	1	4	7	3	5	1	3	1	5	6	7
E1	8	8	3	8	5	3	2	1	6	2	3	8	8
E2	7	8	0	3	9	0	2	0	3	4	2	9	5
E3	1	4	5	5	3	1	2	6	5	5	2	5	5
E4	6	2	1	5	5	4	4	2	2	1	4	5	7

1. Si l'on considère que chaque ligne de ce tableau code un "individu", que représente un "individu" ici ?
2. On procède à l'ACP des notes sensorielles, en projetant les notes hédoniques en variables supplémentaires. Expliquer brièvement ce choix.

Variable supplémentaire : variable qui ne rentre pas dans le calcul des axes principaux. On la représente a posteriori dans les différentes graphiques ou tableaux pour la situer par rapport aux axes principaux.

1. Réduction dimensionnelle

On obtient les valeurs propres suivantes (exprimées en % de l'inertie totale) :

	Inertie	Cumul
1	40.572956	40.572956
2	31.070484	71.643440
3	18.166437	89.809875
4	4.392309	94.202187
5	1.7374828	95.939667
6	1.3977849	97.337456
7	1.1650453	98.502502
8	.68606353	99.18856
9	.51422048	99.702782
10	.20849274	99.911278
11	.08872653	100

1. Combien faut-il de dimensions pour représenter environ 90% des phénomènes de disparité entre individus ?
2. Quels sont à votre avis les deux meilleurs choix d'axes pour l'étude du tableau ? Nommer les deux critères vous permettant de faire ces choix. Quels sont les pourcentages de l'information que ces choix permettent de visualiser ?

2. Analyse des variables

Correlations variables-facteurs:

_varname	CorF1	CorF2	CorF3	CorF4
rond	-.7193681	-.3940986	.3925047	-.2417834
confit	-.164698	-.6762229	.6063042	.1575604
vert	.8307227	.2871425	.0921744	-.1795613
alcool	-.0712314	.465321	.8019046	-.0974704
fruite	-.6361061	-.5000548	-.347111	.2029558
terreux	-.4052357	.7815533	-.0410103	-.2726775
metal	-.5120475	.736494	-.1864064	.1798701
piquant	.8291297	.2906325	-.2179837	.0923126
acide	.5622748	.2455091	.6335065	.3068211
amer	.8405781	-.2813234	-.152596	-.0269766
astringent	-.4968601	.7387611	.0348275	.2444043
chaleur	-.5089256	-.5133962	.3666813	.0432578
profondeur	-.5504313	.503978	.3779484	-.2693975

CO2 des variables:

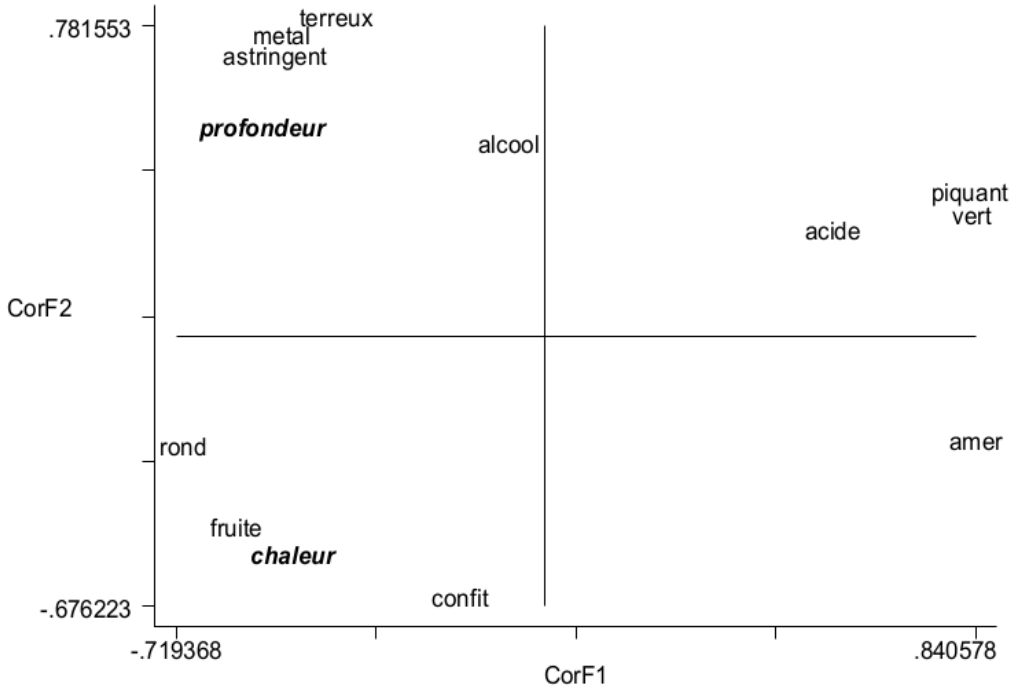
_varname	Cos2F1	Cos2F2	Cos2F3	Cos2F4
rond	.5174904	.1553137	.1540599	.0584592
confit	.0271254	.4572774	.3676047	.0248253
vert	.6901002	.0824508	.0084961	.0322423
alcool	.0050739	.2165236	.643051	.0095005
fruite	.404631	.2500548	.1204861	.0411911
terreux	.164216	.6108256	.0016818	.074353
metal	.2621927	.5424234	.0347474	.0323533
piquant	.6874561	.0844673	.0475169	.0085216
acide	.3161529	.0602747	.4013305	.0941392
amer	.7065716	.0791428	.0232856	.0007277
astringent	.24687	.545768	.001213	.0597334
chaleur	.2590053	.2635757	.1344552	.0018712
profondeur	.3029746	.2539938	.142845	.072575

1. Quelles sont les variables mal représentées dans ce plan ?

2. Qualifier les couples de variables suivants de : positivement corrélés, négativement corrélés, décorrélés, lorsque cela est possible : (rond , vert) ; (alcool , vert) ; (fruité , chaleur) ; (fruité , rond) ; (confit , amer)
3. Combien vous semble-t-il y avoir de faisceaux de variables sensorielles ? Quelles sont les variables qui entrent dans chacun d'eux ?
4. Les notes hédoniques sont-elles corrélées à l'un ou l'autre de ces faisceaux ? Si oui, lequel ?
5. Quelle est la qualité de représentation de rond dans le plan (1,2) ? Quel pourcentage des disparités en termes de rondeur ne seront pas visibles dans ce plan ?
6. Même question pour la sensation d'alcool. Jusqu'à quel axe faudrait-il pousser l'analyse pour visualiser l'essentiel des disparités alcooliques ?
7. L'axe 1 a-t-il une interprétation simple ? Si oui, laquelle ?
8. Même question pour l'axe 2.
9. Les notes hédoniques chaleur et profondeur sont elles bien représentées dans ce plan ? Que peut-on en déduire ?
10. Si la note chaleur devait être exprimée sous la forme :

$$\text{chaleur} = a \times \text{rond} + b \times \text{métal}$$

quels seraient les signes de a et de b ?



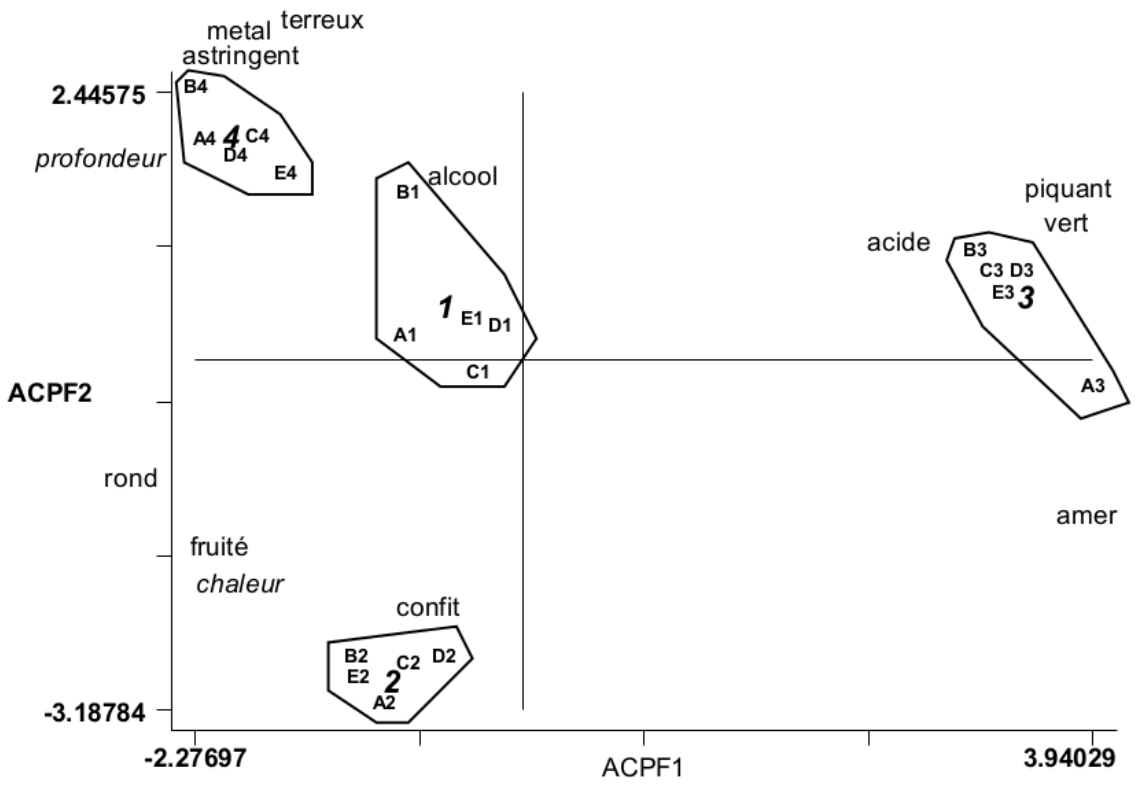
3. Analyse des individus

Chaque ensemble d'"individus" correspondant à un même produit a été entouré. Pour chaque produit et chaque variable, on a calculé la moyenne de la note des 5 juges. On obtient ainsi le jugement moyen de chaque produit sous la forme d'une ligne contenant ses notes moyennes. Ces 4 lignes ont été projetées en supplémentaire sur le plan (1,2). Elles y sont identifiées par les numéros 1 à 4.

1. Où se trouve la projection de la ligne représentant les notes moyennes d'un produit, par rapport aux projections des lignes représentant les jugements des 5 juges pour ce même produit ?
2. Lorsque cela est possible, tracer sur le graphique direct l'axe correspondant à chacune des variables suivantes, et classer les produits dans l'ordre croissant de leur note moyenne pour cette variable :

astringent ; alcool ; amer ; piquant ; profondeur

3. Quels sont les individus les mieux représentés sur l'axe 3 ? En déduire la composante sensorielle qui distingue le mieux le produit 1 des trois autres.



COS2 des individus sur les axes:

	id	CO2F1	CO2F2	CO2F3	CO2F4
1.	A1	.1114631	.0060365	.6562633	.0237801
2.	A2	.0753288	.8230201	.0172711	.0465541
3.	A3	.9273273	.0053001	.0057271	.0019026
4.	A4	.4176932	.3374724	.0996535	.0803263
5.	B1	.0670976	.2314948	.5605586	.0716898
6.	B2	.1283377	.7128298	.0575762	.0753639
7.	B3	.8552936	.0570546	.0067148	.0359859
8.	B4	.3786353	.4368481	.1120448	.0472578
9.	C1	.0106408	.0024776	.8699428	.0339826
10.	C2	.0682151	.8542702	.0039535	.0293765
11.	C3	.8376347	.0466894	.0462553	.0131416
12.	C4	.3965988	.3943879	.0685767	.0873551
13.	D1	.0039335	.01121	.6567702	.2235283
14.	D2	.0351449	.8552856	.0416467	.0001245
15.	D3	.8611803	.0282562	.064343	.0002378
16.	D4	.3524109	.2921459	.2606521	.0655917
17.	E1	.0210436	.017153	.8191473	.048536
18.	E2	.1197285	.7218954	.0734972	.0501869
19.	E3	.8934291	.0269831	.0133444	.0001236
20.	E4	.3628988	.3671432	.1446364	.0557078

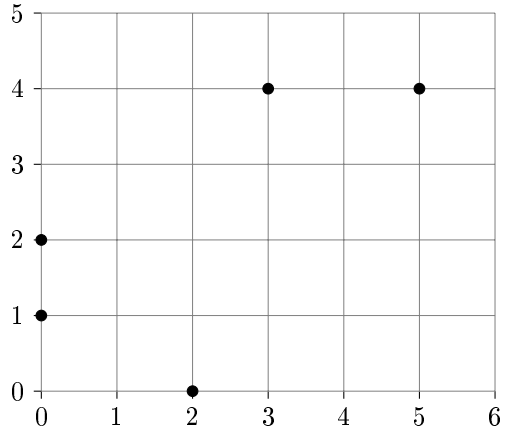
CTR des individus sur les axes:

	id	CTRF1	CTRF2	CTRF3	CTRF4
1.	A1	.7917957	.055996	10.41183	1.560409
2.	A2	1.096882	15.64939	.5616742	6.261808
3.	A3	18.30938	.1366513	.2525478	.3470054
4.	A4	5.710521	6.024835	3.04283	10.14425
5.	B1	.7431421	3.348074	13.86606	7.334427
6.	B2	1.608605	11.66728	1.611776	8.725739
7.	B3	11.44979	.9973828	.2007613	4.449987
8.	B4	6.114079	9.211471	4.04081	7.048998
9.	C1	.1193501	.0362884	21.79246	3.520859
10.	C2	.7471624	12.21851	.0967134	2.972203
11.	C3	12.3081	.8958668	1.517975	1.783728
12.	C4	4.004111	5.199564	1.546315	8.146791
13.	D1	.0300516	.1118362	11.20644	15.77479
14.	D2	.3705609	11.77598	.9807199	.0121272
15.	D3	14.00013	.5998464	2.336179	.0357042
16.	D4	4.727875	5.118056	7.809886	8.128473
17.	E1	.1534768	.1633618	13.34294	3.26987
18.	E2	1.529886	12.04549	2.097487	5.923729
19.	E3	12.98162	.5119753	.4330474	.0165928
20.	E4	3.203482	4.232148	2.851552	4.542517

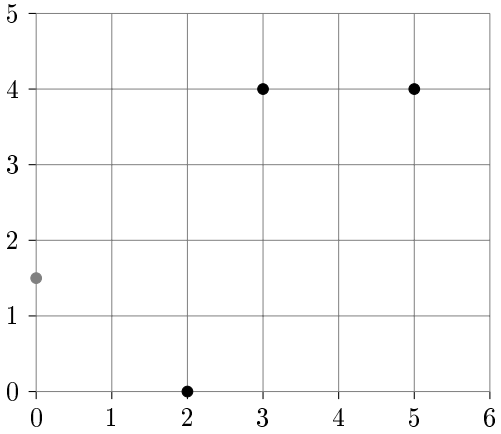
5 Classification ascendante hiérarchique

Exercice 9. Algorithme de la CAH Les quatre dessins suivants correspondent aux étapes successives d'une classification hiérarchique ascendantes des cinq points $M_1 = (2,0), M_2 = (0,1), M_3 = (0,2), M_4 = (3,4)$ et $M_5 = (5,4)$ progressivement regroupées en classes de deux ou trois points dont les centres de gravité sont notés G_6, G_7 et G_8. On suppose que les cinq points initiaux sont tous affectés du poids 1. La distance choisie pour cette classification, qui apparait *au carré* dans les quatre matrices de distances, est l'écart de Ward.

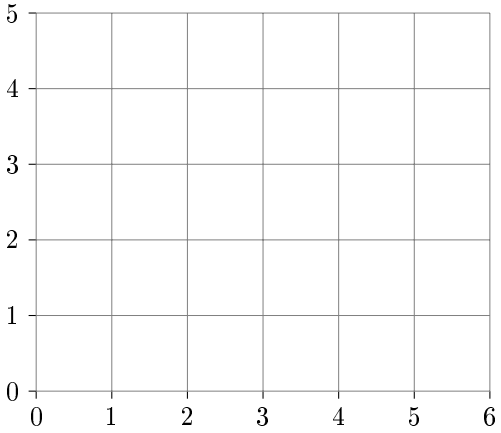
1. Compléter les figures.
2. Compléter les distances manquantes dans les matrices de distances.
3. Préciser les coordonnées des points G_6 , G_7 et G_8 et calculer les coordonnées du centre de gravité G_9 des cinq points.
4. Calculer l'inertie totale du nuage.
5. Tracer un dendrogramme résumant cette classification.



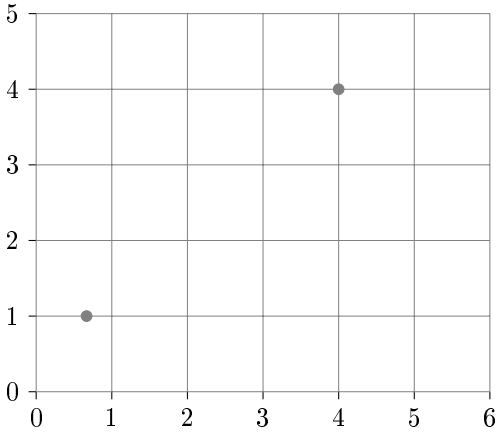
	M_1	M_2	M_3	M_4	M_5
M_1	0	2.50	4.00	8.50	12.50
M_2	2.50	0	0.50	9.00	17.00
M_3	4.00	0.50	0	6.50	14.50
M_4	8.50	9.00	6.50		
M_5	12.50			2.00	0



	M_1	G_6	M_4	M_5
M_1	0	4.17	8.50	12.50
G_6	4.17	0		20.83
M_4	8.50		0	2.00
M_5	12.50	20.83	2.00	0



	M_1	G_6	G_7
M_1	0	4.17	13.33
G_6	4.17	0	22.25
G_7	13.33	22.25	0



	G_8	G_7
G_8	0	24.13
G_7	24.13	0