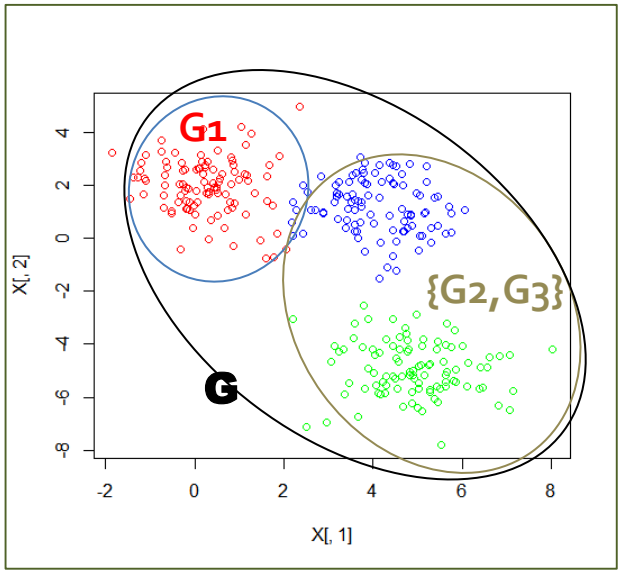
Comprendre la taille d'effet (effect size)

Dans le contexte de la caractérisation des groupes en classification automatique (clustering)

Ricco Rakotomalala

Classification automatique – Interprétation des classes

L'algorithme de classification automatique se charge de mettre en évidence les groupes « naturels » c.-à-d. qui se démarquent significativement les uns des autres.



A l'issue de la constitution des groupes, il faut comprendre leur nature : Qu'est-ce qui caractérise tel ou tel groupe ? Qu'est-ce qui le distingue (des autres) ?

- G1 vs. G (population globale)
- G1 vs. {G2, G3} (les autres)

Ces deux lectures sont proches mais ne sont pas strictement identiques (G1 participe à la population G)

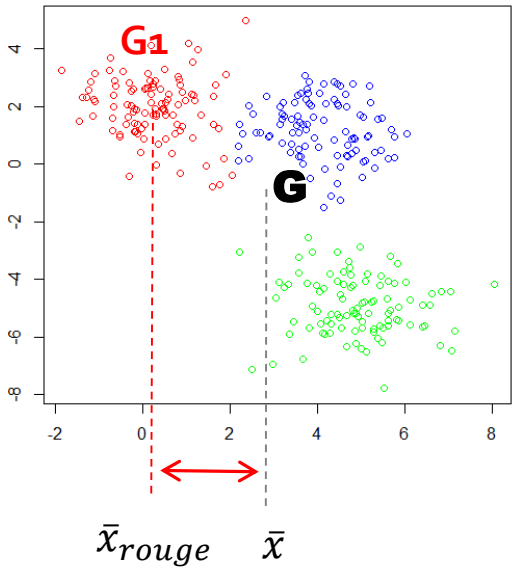
Plan

1. Principe de la valeur test
2. Taille d'effet pour les variables quantitatives
3. Taille d'effet pour les variables qualitatives
4. Etude de cas n°1 (petits effectifs)
5. Etude de cas n°2 (grands effectifs)
6. Conclusion

Principe de la valeur test

Evaluer l'impact des variables prises individuellement.

Pour les variables quantitatives, mesurer l'importance de l'écart de la moyenne du groupe avec la moyenne globale (*comparer les proportions pour les variables qualitatives*).



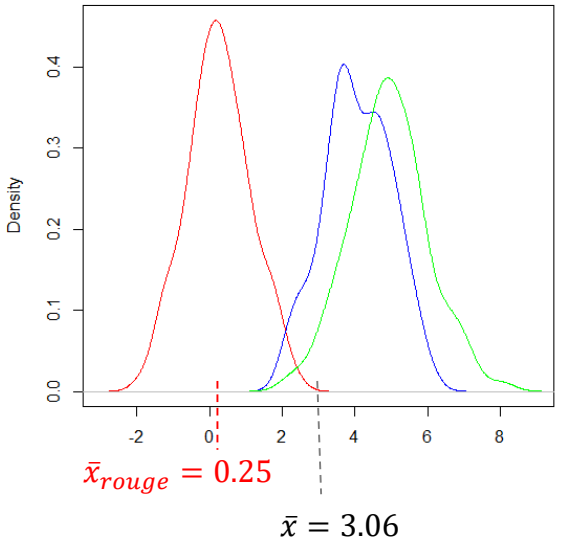
Valeur test

- σ^2 est la variance empirique calculée sur l'ensemble de l'échantillon
- n, n_g sont respectivement la taille de l'échantillon global et celle du groupe « g »

$$vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \frac{\sigma^2}{n_g}}}$$

La statistique suit *très approximativement* une loi normale (|vt| > 2, écart significatif à 5%). Valable surtout pour les variables illustratives.

Si on s'intéresse aux fonctions de densité….



Valeur test – Problème de la volumétrie

La valeur test est très sensible à la taille de l'échantillon. Sur les grosses volumétries, les vt prennent des valeurs très élevées, indiquant systématiquement des écarts « significatifs ».

$$vt = \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \frac{\sigma^2}{n_g}}} = \sqrt{n_g} \times \frac{\bar{x}_g - \bar{x}}{\sqrt{\frac{n - n_g}{n-1} \times \sigma^2}}$$

➔ On le comprend en réécrivant la formule. La vt dépend de la taille absolue du groupe (n_g) et non de sa taille relative ($\frac{n_g}{n}$).

Ex. Toutes choses égales par ailleurs, entre des échantillons de taille n = 30 et n = 3000, les vt sont multipliées mécaniquement par 10. La région critique (|vt| > 2) n'est absolument pas discriminante.

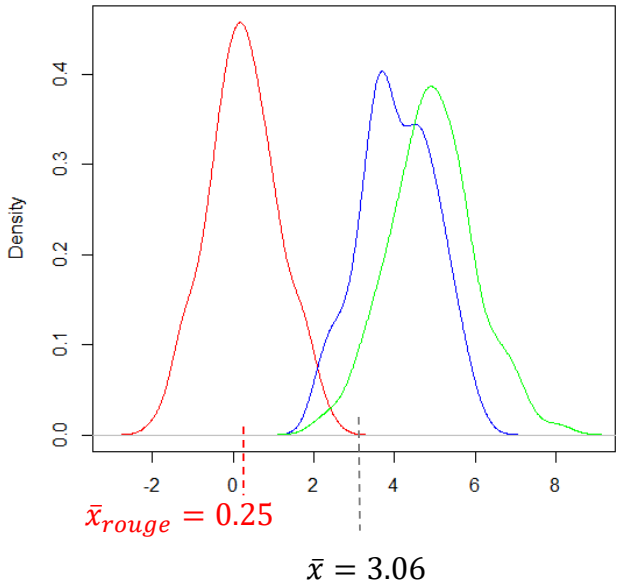
Une solution consisterait à travailler sur les pourcentages c.-à-d. fixer artificiellement n' = 100 et donc utiliser $n'_g = \frac{n_g}{n} \times 100$: c'est l'indicateur **VT-100** (*développé dans le cadre des règles d'association mais transposable ici*).

Mais la valeur (n' = 100) reste empirique, peut-être discutable et/ou paramétrable. Un indicateur plus intéressant devrait être totalement indépendant des effectifs.

La valeur test – Groupe vs. population globale ou Groupe vs. les autres

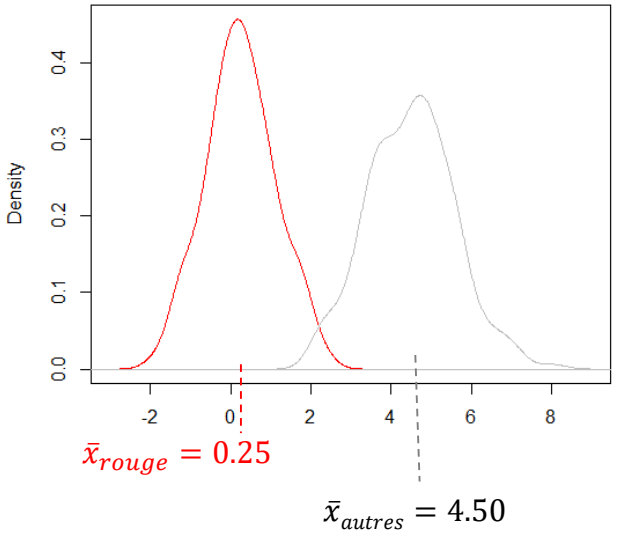
Pour caractériser un groupe, on peut comparer sa moyenne avec la moyenne globale comme réalisé pour la vt....

(peut se comprendre via l'équation d'analyse de variance)



... mais on pourrait aussi la comparer avec la moyenne des autres groupes.

(peut être vu comme un test post hoc – Test de Dunnett)



Ce n'est pas mieux ou moins bien, c'est un autre point de vue. !

Spécifications pour un indicateur de caractérisation des groupes

1. Il doit mesurer l'amplitude des écarts (l'intensité des écarts) entres les moyennes conditionnelles (ou proportions)

2. Facile à calculer, obtenu à partir des indicateurs statistiques usuels (moyenne, variance, proportion…)

3. Compréhensible et interprétable. Ou tout du moins correspondre à des notions statistique connues et reconnues. Proposer un nouvel indicateur illisible ne sert à rien.

4. Confrontable à des valeurs seuils permettant de situer la significativité des écarts

5. Non exposé au problème des grands effectifs.

Comparaison de moyennes

TAILLE D'EFFET POUR LES VARIABLES QUANTITATIVES

Données utilisées pour l'exposé

$\mathcal{N}(m_a=2, \sigma_a=1)$

$n_a = 1000$

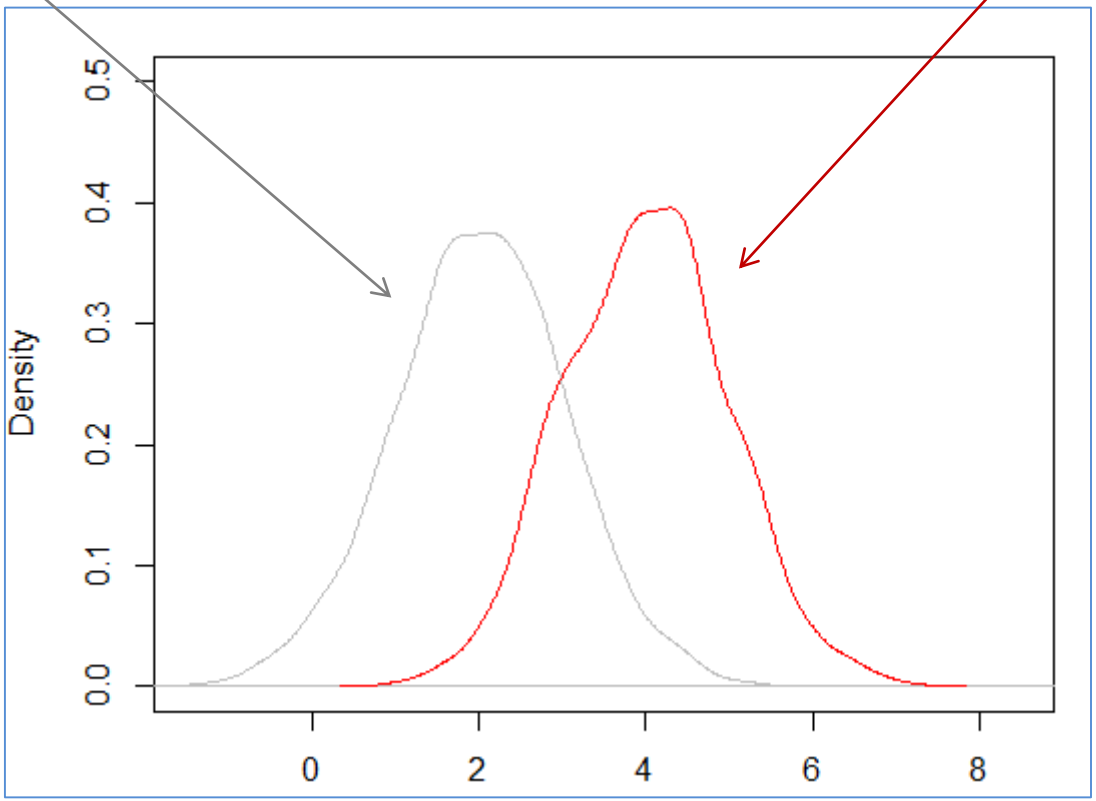
« Autres »

$n = n_g + n_a$

$\mathcal{N}(m_g=4, \sigma_g=1)$

$n_g = 1000$

Groupe (à caractériser)



$\bar{x}_a = 2.01$

$s_a = 1.03$

$\bar{x}_g = 4.00$

$s_g = 0.98$

d de Cohen – La taille d'effet

Le d de Cohen (population) normalise la différence entre les moyennes par l'écart-type

$$d = \frac{\bar{x}_g - \bar{x}_a}{\sigma_{pooled}} \quad \text{Où} \quad \sigma_{pooled} = \sqrt{\frac{(n_g - 1)s_g^2 + (n_a - 1)s_a^2}{n_g + n_a}}$$

Cohen définit une mesure « échantillon » (d_s) que l'on retrouve également sous le nom de g de Hedges

$$g = \frac{\bar{x}_g - \bar{x}_a}{s_{pooled}} \quad \text{Où} \quad s_{pooled} = \sqrt{\frac{(n_g - 1)s_g^2 + (n_a - 1)s_a^2}{n_g + n_a - 2}}$$

Une échelle permet d'apprécier l'intensité de l'écart (Cohen, 1988 ; Sawilowsky, 2009)

(Attention, ce sont des repères, les « vrais » seuils dépendent des domaines dans lesquels nous travaillons)

Evaluation	Traduction	Seuil (en valeur absolue)
Very small	Très faible	0.01
Small	Faible	0.20
Medium	Moyenne	0.50
Large	Elevée	0.80
Very large	Très élevée	1.20
Huge	Immense	2.00

Dans notre exemple, $g = 1.975$ c.-à-d. l'écart équivaut à ≈2 fois l'écart-type.

Intérêt de la taille d'effet

a. Nous disposons d'une grille de référence qui permet d'apprécier l'intensité des écarts. g = 1.97, nous savons que l'écart est représentatif d'un phénomène important.

b. La mesure est insensible aux effectifs. Nous multiplions par 1000 la taille des échantillons, la mesure n'est pas modifiée (ex. avec des données simulées sous R où n_g = n_a = 1 000 000, nous obtenons $g = 1.99$).

c. C'est une notion qui ne vient pas de nulle part. Elle est liée au t de Student du test de comparaison de moyennes (**sous** hypothèse d'égalité des variances)

$$g = t\sqrt{\frac{n_g + n_a}{n_g \times n_a}}$$

J'aurais fait clignoter si j'avais pu !

d. Sous hypothèse de normalité des distributions, elle se prête à des interprétations qui se révèlent particulièrement instructives. Voir pages suivantes…

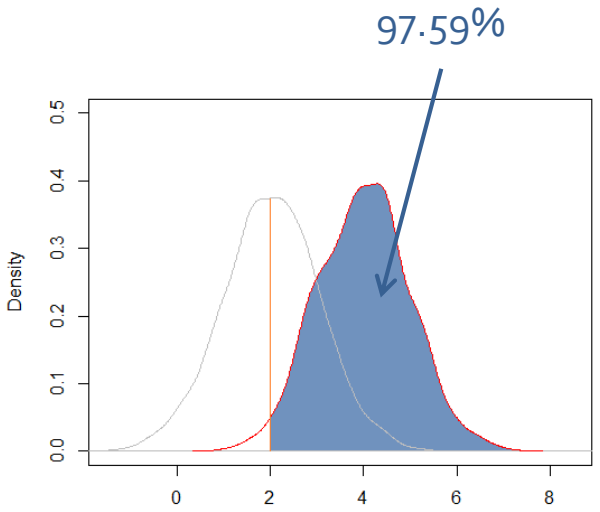
Interprétations probabilistes de la taille d'effet

Φ() est la fonction de répartition de la loi normale centrée et réduite.

$$U_3 = \Phi(g) = 0.9759$$

Il y a 97,59% de chances que les valeurs du groupe soient au dessus de la médiane des autres.

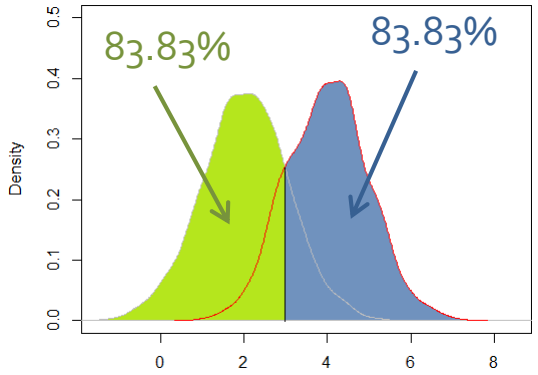
Puisque les données sont simulées, nous connaissons la « vraie » valeur, elle est égale à 97.972 %



$$U_2 = \Phi\left(\frac{|g|}{2}\right) = 0.8383$$

83,83% des valeurs du groupe excèdent 83,83% des valeurs des autres

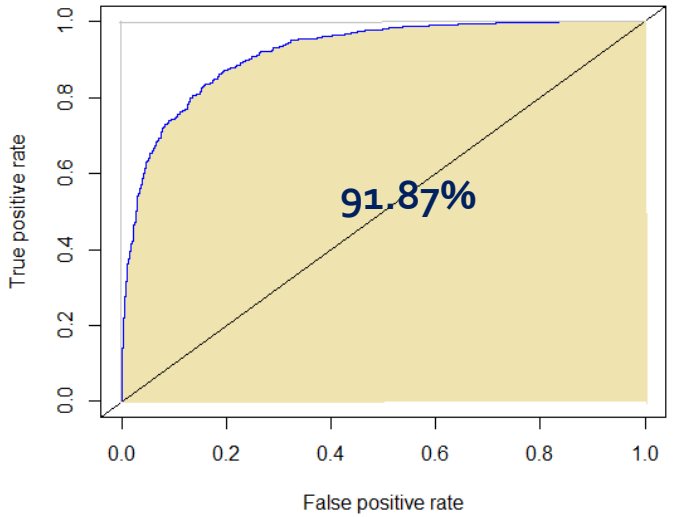
La vraie valeur est 84.13%, et le point permettant de séparer au mieux les deux distributions est 3 (en abscisse)



Common language effect size

On peut trier les individus selon leurs valeurs et construire la courbe ROC en considérant que la classe cible est le groupe « g »

Courbe ROC : en abscisse, taux de faux positifs ; en ordonnée, taux de vrais positifs.



Common language effect size

$$CLES = \Phi\left(\frac{|\bar{x}_g - \bar{x}_a|}{\sqrt{s_g^2 + s_a^2}}\right) = 0.9187$$

$$(0 \leq \text{CLES} \leq 1)$$

91,87% correspond à la probabilité pour qu'un individu du groupe « g » (pris au hasard) présente une valeur plus élevée qu'un autre individu pris parmi les « autres »

Binomial effect size display (BESD)

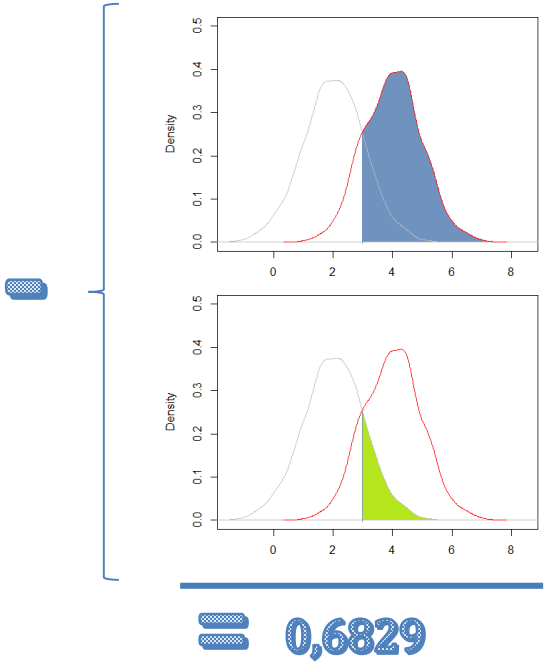
Comme l'une des variables est binaire, on peut obtenir cette corrélation à partir des moyennes conditionnelles : corrélation bisériale ponctuelle

$$r_{pb} = \frac{\bar{x}_g - \bar{x}_a}{s}\sqrt{\frac{n_g \times n_a}{n(n-1)}} = 0.7029$$

Où $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$

Le coefficient de corrélation r_{pb} entre une indicatrice de classe (0/1) et les valeurs permet de qualifier l'amplitude de l'écart (par convention, le signe r = sens de l'écart des moyennes)

Il s'agit d'une estimation (pas très précise) de l'écart entre les probabilités des distributions conditionnelles à être supérieure à la médiane globale.



= 0,6829

Φ() est la fonction de répartition de la loi normale centrée et réduite.

$$\tau = 2 \times \Phi\left(\frac{r_{pb}}{\sqrt{1 - r_{pb}^2}}\right) - 1 = 0.677$$

Remarque : une mesure corrigée permet d'obtenir une estimation plus précise (Rosenthal & Rubin, 1982)

Binomial effect size display (BESD) - Suite

Le BESD est lié à la taille d'effet (g de Hedge) par la relation

$$r_{pb} = \sqrt{\frac{g^2}{g^2 + \frac{n-2}{n\left(\frac{n_g}{n}\right)\left(\frac{n_a}{n}\right)}}} = 0.7029$$

La corrélation est une notion connue et reconnue, elle est bornée ($0 \leq |r_{pb}| \leq 1$) et des seuils sont proposés dans la littérature (Cohen, 1988 ; pages 79 et 80)

Evaluation	Traduction	Seuil (en valeur absolue)
Small	Faible	0.10
Medium	Moyenne	0.30
Large	Elevée	0.50

En la montant au carré, elle peut également se lire comme la proportion de variance expliquée par l'appartenance au groupe (le groupe cible vs. les autres)

$$(r_{pb})^2 = 0.494032$$!

49.40% de la variance de la variable d'intérêt s'explique par la dichotomie « groupe cible vs. les autres ».

Analyse en deux étapes - Rapport de corrélation et BESD

Partition en K groupes

La variabilité totale d'une variable peut se décomposer en variabilités expliquées par l'appartenance aux K groupes (K ≥ 2) et résiduelles (intra-groupes). On peut dégager le rapport de corrélation dont le carré correspond à la proportion de variance expliquée ($0 \leq \eta^2 \leq 1$).

$$\eta^2 = \frac{\sum_{k=1}^{K} n_k (\bar{x}_k - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Dans notre cas, si on se cantonne à K = 2, nous avons

$$\eta^2 = \frac{1974.763}{3997.238} = 0.497032$$

Un groupe vs. les autres

Le BESD (ou corrélation bisériale ponctuelle) caractérise un groupe cible « g » contre les « autres que 'g' » ($-1 \leq r_{pb} \leq 1$).

$$r_{pb} = \frac{\bar{x}_g - \bar{x}_a}{s} \sqrt{\frac{n_g \times n_a}{n(n-1)}} = 0.7029$$

Et, $(r_{pb})^2 = 0.494032$

Le η^2 ne se décompose pas en $(r_{pb})^2$, mais il n'en reste pas moins qu'il y a une forme de cohérence entre ces deux mesures : l'une caractérise la partition globale en K classes ; l'autre caractérise une classe cible « g » par rapport aux autres. Quand nous n'avons que 2 classes (K = 2), elles se rejoignent.

Identifier les modalités caractérisant les groupes – Comparaison de proportions

TAILLE D'EFFET POUR LES VARIABLES QUALITATIVES

Données utilisées pour l'exposé

Modalités de la variable X

Groupes issus du clustering	Modalité d'intérêt « ℓ »	Autres modalités	**Total**
Groupe cible « g »	**72**	**14**	$n_g = 86$
Autres groupes « a »	**24**	**40**	$n_a = 64$
Total	96	54	$n = 150$

L'objectif est de mesurer la sur-représentativité ou la sous-représentativité de la modalité d'intérêt « ℓ » dans le groupe

$$p_{l/g} = \frac{72}{86} = 0.837$$

En la comparant à :

- Soit la prévalence de la modalité dans la population $p_l = \frac{96}{170} = 0.640$
- Soit sa proportion dans les autres groupes $p_{l/a} = \frac{24}{64} = 0.375$

⇨ On souhaite que l'indicateur prenne une valeur positive (> 0) s'il y a surreprésentation, négatif sinon.

Valeur test pour les variables qualitatives

	Modalité d'intérêt « ℓ »	Autres modalités	Total
Groupe cible « g »	72	14	86
Autres groupes « a »	24	40	64
Total	96	54	150

Statistique de test de comparaisons de proportions. La prévalence dans la population est la référence

$$vt = \sqrt{n_g} \times \frac{p_{l/g} - p_l}{\sqrt{\frac{n - n_g}{n - 1} p_l (1 - p_l)}}$$

$$vt = \sqrt{86} \times \frac{0.837 - 0.640}{\sqrt{\frac{150 - 86}{150 - 1} 0.640(1 - 0.640)}} = \mathbf{4.36}$$

vt suit une loi normale de manière très approximative, surtout valable pour les variables illustratives. Valeur critique ±2 pour un test bilatéral à 5%

vt est aussi très sensible à la taille de l'échantillon, les proportions étant exactement les mêmes, tout devient significatif sur les gros effectifs.

Se ramener à des pourcentages (VT-100) est une piste. Etre indépendant de la taille d'échantillon est préférable. La notion de taille d'effet peut être aussi utilisée pour les comparaisons de proportions (Cohen, 1988 ; chapitre 6).

Taille d'effet pour la comparaison de proportions

La variance de la proportion p dépend de sa valeur [$Var(p) = \frac{p(1-p)}{n}$], les proportions $p_{l/g}$ et $p_{l/a}$ ne sont pas comparables directement.

Une piste de travail consiste à réaliser une transformation de variables de manière à supprimer cette relation : transformation arcsinus

$$\varphi = 2\arcsin(\sqrt{p})$$

$$Var(\varphi) = \frac{1}{n}$$

Cohen (1988 ; page 181) propose la mesure de taille d'effet suivante

$$h = \varphi_{l/g} - \varphi_{l/a}$$

	Modalité d'intérêt « ℓ »	Autres modalités	Total
Groupe cible « g »	72	14	86
Autres groupes « a »	24	40	64
Total	96	54	150

$$\varphi_{l/g} = 2\arcsin\left(\sqrt{\frac{72}{86}}\right) = 2.311 \qquad \varphi_{l/a} = 2\arcsin\left(\sqrt{\frac{24}{64}}\right) = 1.318$$

$$h = 2.311 - 1.318 = 0.993$$

Avec des seuils permettant de situer l'amplitude des écarts

Evaluation	Traduction	Seuil (en valeur absolue)
Small	Faible	0.20
Medium	Moyenne	0.50
Large	Elevée	0.80

Taille d'effet exprimée par la corrélation

Le croisement entre les indicatrices peut être représenté par un tableau de contingence.

Modalités

Classes	1	0	Total
1	72	14	$n_g = 86$
0	24	40	$n_a = 64$
Total	96	54	$n = 150$

$\phi = 0.462$

Le coefficient de corrélation r ($0 \leq r \leq 1$) entre les indicatrices de classe (0/1) et de modalité (0/1) permet de mesurer leur attraction ou leur répulsion.

Pour mesurer l'association entre les indicatrices via un tableau de contingence (2 x 2), nous passons par l'indicateur ϕ

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Le signe de ϕ correspond au signe de la différence ($p_{\ell/g} - p_{\ell/a}$)

Les valeurs repères sont les mêmes que pour la corrélation bisériale ponctuelle (Cohen, 1988 ; pages 224 et 225).

Evaluation	Traduction	Seuil (en valeur absolue)
Small	Faible	0.10
Medium	Moyenne	0.30
Large	Elevée	0.50

L'indicateur ϕ^2 s'interprète comme la proportion de variance partagée par les deux variables indicatrices (par le groupe cible et la modalité d'intérêt).

Analyse en deux étapes – V de Cramer et ϕ

Partition en K groupes

La proportion de variation de variation partagée entre deux variables matérialisant l'appartenance aux K groupes (K ≥ 2) et les modalités (L ≥ 2) peut être quantifiée avec le carré du v de Cramer ($0 \leq v^2 \leq 1$).

$$v^2 = \frac{\chi^2}{n \times \min(K-1, L-1)}$$

⇨ Dans le tableau de contingence global croisant les K groupes (en ligne) et les L modalités (en colonne)

Un groupe vs. les autres

L'indicateur ϕ mesure l'attraction (ou la répulsion) entre un un groupe cible « g » et une modalité d'intérêt « ℓ » ($-1 \leq \phi \leq 1$).

$$\phi = \text{sgn}(p_{l/g} - p_{l/a})\sqrt{\frac{\chi^2}{n}}$$

⇨ Dans le tableau de contingence (2 x 2) opposant le groupe « g » avec la modalité « ℓ ».

Le v^2 ne se décompose pas en $(\phi)^2$, mais il n'en reste pas moins qu'il y a une forme de cohérence entre ces deux mesures : l'une caractérise la partition globale en K classes par rapport aux L modalités ; l'autre caractérise l'association entre la classe cible « g » et la modalité d'intérêt « ℓ ».

Quand nous n'avons que 2 classes (K = 2) et 2 modalités (L = 2), les mesures se rejoignent.

Travailler sur un petit effectif

ETUDE DE CAS N°1

Classification des automobiles

(voir http://tutoriels-data-mining.blogspot.fr/2016/09/clustering-caracterisation-des-classes.html)

Variables « actives »

Variables « illustratives »

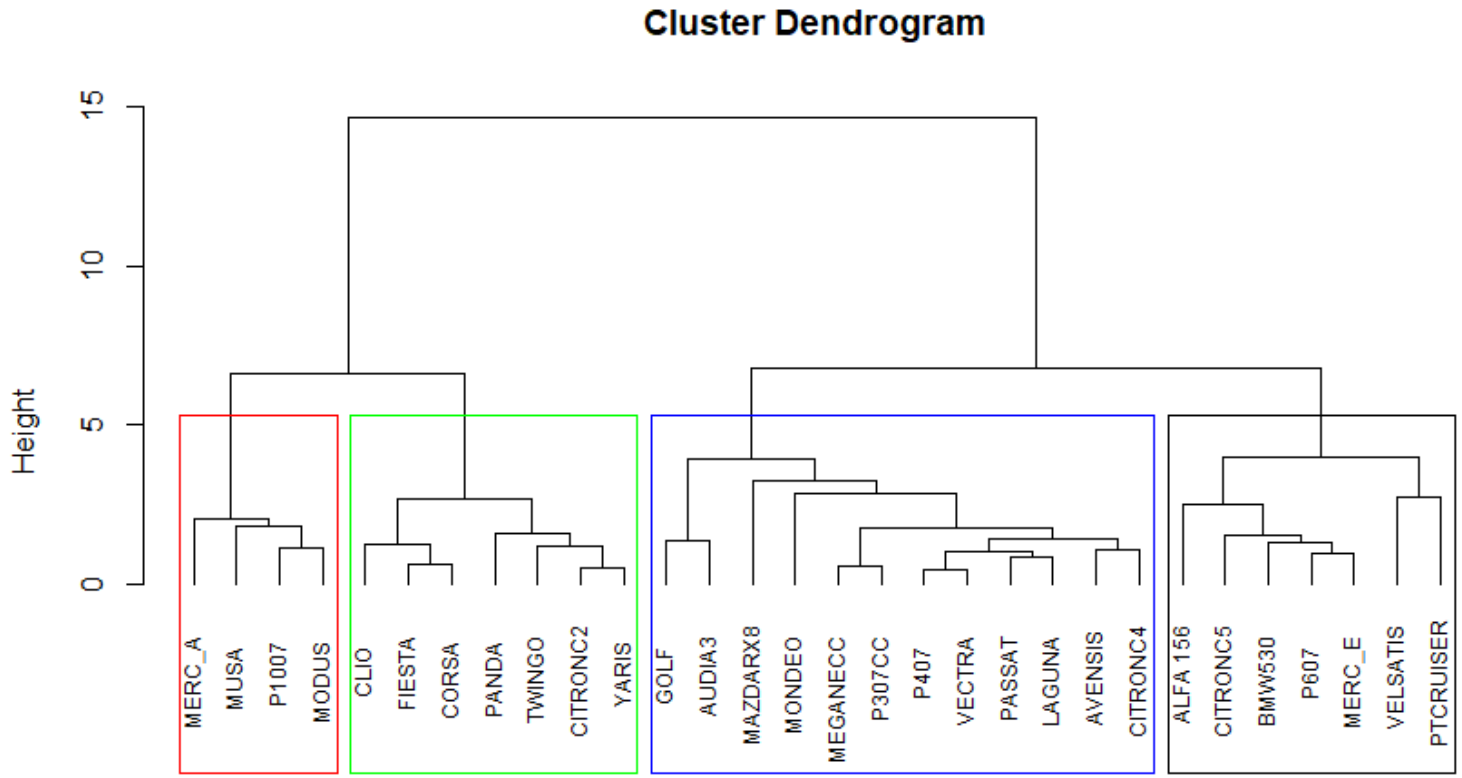
Modele	puissance	cylindree	vitesse	longueur	largeur	hauteur	poids	CO2	prix	origine	carburant
PANDA	54	1108	150	354	159	154	860	135	8070	Europe	Essence
TWINGO	60	1149	151	344	163	143	840	143	8950	France	Essence
CITRONC2	61	1124	158	367	166	147	932	141	10700	France	Essence
YARIS	65	998	155	364	166	150	880	134	10450	Autres	Essence
FIESTA	68	1399	164	392	168	144	1138	117	14150	Europe	Diesel
CORSA	70	1248	165	384	165	144	1035	127	13590	Europe	Diesel
GOLF	75	1968	163	421	176	149	1217	143	19140	Europe	Diesel
P1007	75	1360	165	374	169	161	1181	153	13600	France	Essence
MUSA	100	1910	179	399	170	169	1275	146	17900	Europe	Diesel
CLIO	100	1461	185	382	164	142	980	113	17600	France	Diesel
AUDIA3	102	1595	185	421	177	143	1205	168	21630	Europe	Essence
MODUS	113	1598	188	380	170	159	1170	163	16950	France	Essence
AVENSIS	115	1995	195	463	176	148	1400	155	26400	Autres	Diesel
P407	136	1997	212	468	182	145	1415	194	23400	France	Essence
CITRONC4	138	1997	207	426	178	146	1381	142	23400	France	Diesel
MERC_A	140	1991	201	384	177	160	1340	141	24550	Europe	Diesel
MONDEO	145	1999	215	474	194	143	1378	189	23100	Europe	Essence
VECTRA	150	1910	217	460	180	146	1428	159	26550	Europe	Diesel
PASSAT	150	1781	221	471	175	147	1360	197	27740	Europe	Essence
VELSATIS	150	2188	200	486	186	158	1735	188	38250	France	Diesel
LAGUNA	165	1998	218	458	178	143	1320	196	25350	France	Essence
MEGANECC	165	1998	225	436	178	141	1415	191	27800	France	Essence
P307CC	180	1997	225	435	176	143	1490	210	28850	France	Essence
P607	204	2721	230	491	184	145	1723	223	40550	France	Diesel
MERC_E	204	3222	243	482	183	146	1735	183	46450	Europe	Diesel
CITRONC5	210	2496	230	475	178	148	1589	238	33000	France	Essence
PTCRUISER	223	2429	200	429	171	154	1595	235	27400	Autres	Essence
MAZDARX8	231	1308	235	443	177	134	1390	284	34000	Autres	Essence
BMW530	231	2979	250	485	185	147	1495	231	46400	Europe	Essence
ALFA 156	250	3179	250	443	175	141	1410	287	40800	Europe	Essence

n = 30 obs.

Objectif de l'étude : Identifier les catégories (groupes) de voitures « similaires » (c.-à-d. qui se ressemblent au regard de leurs propriétés)

CAH (Classification Ascendante hiérarchique)

Critère de Ward (ward.D2 sous R)



G3 — Petits monospaces

G1 — Petites voitures

G2 — Familiales

G4 — Berlines

Caractérisation de la partition – Var. quantitatives

Carré du rapport de corrélation (η^2)

Moyennes conditionnelles

	G 1	G 3	G 2	G 4	% epl.
poids	952.14	1241.50	1366.58	1611.71	85.8
longueur	369.57	384.25	448.00	470.14	83.0
cylindree	1212.43	1714.75	1878.58	2744.86	81.7
puissance	68.29	107.00	146.00	210.29	73.8
vitesse	161.14	183.25	209.83	229.00	68.2
largeur	164.43	171.50	178.92	180.29	67.8
hauteur	146.29	162.25	144.00	148.43	65.3
prix	11930.00	18250.00	25613.33	38978.57	82.48
CO2	130.00	150.75	185.67	226.43	59.51

La constitution des groupes s'est appuyée avant tout sur le poids, la longueur et la cylindrée (les autres variables contribuent quand même pas mal).

La segmentation se traduit par une différenciation des véhicules par les prix.

Caractérisation des groupes – Var. quantitatives

Corrélation bisériale ponctuelle (r_{pb})

G4

	Moyenne G4	Moyenne (!G4)	r_pb
cylindree	2744.86	1647.35	**0.791**
puissance	210.29	115.57	**0.688**
poids	1611.71	1218.70	**0.669**
longueur	470.14	413.04	**0.546**
vitesse	229.00	190.39	**0.540**
largeur	180.29	173.22	**0.388**
hauteur	148.43	147.87	**0.033**
prix	38978.57	20168.26	**0.755**
CO2	226.43	162.65	**0.599**

Le groupe 4 des « berlines » se distingue par une cylindrée, puissance, poids, longueur et vitesse fortement élevés par rapport aux autres ($r_{pb} > 0.5$), qui se traduit par des prix et Co2 élevés (toujours par rapport aux autres).

G3

	Moyenne (G3)	Moyenne (!G3)	r_pb
hauteur	162.25	145.81	**0.772**
poids	1241.50	1321.00	**-0.109**
cylindree	1714.75	1932.46	**-0.126**
largeur	171.50	175.38	**-0.171**
puissance	107.00	142.38	**-0.206**
vitesse	183.25	201.88	**-0.209**
longueur	384.25	432.85	**-0.373**
CO2	150.75	181.65	**-0.233**
prix	18250.00	25527.69	**-0.235**

Le groupe 3 des « petits monospaces » se distingue par une hauteur fortement élevée ($r_{pb} > 0.5$) et une longueur modérément faible ($r_{pb} < -0.3$).

Caractérisation de la partition et des groupes – Var. qualitatives

Carré du v de Cramer (v^2) et coefficient ϕ

Caractérisation de la partition à l'aide des variables. Les deux variables sont très faiblement liées avec les partitions !

	Cramer v^2
carburant	1.46%
origine	1.24%

Croisement des groupes avec les modalités de « Origine »

	Origine			
Groupe	Autres	Europe	France	Total
G1	1	3	3	7
G2	2	5	5	12
G3	0	2	2	4
G4	1	3	3	7
Total	4	13	13	30

La somme de la colonne fait 100%

G4

$\left(\frac{1}{7}\right)$ $\left(\frac{1+2+0}{7+12+4}\right)$

	p_l/g	p_l/a	phi
Autres	14.3%	13.0%	0.0155
Europe	42.9%	43.5%	-0.0053
France	42.9%	43.5%	-0.0053

$\left(\frac{3}{7}\right)$ $\left(\frac{3+5+2}{7+12+4}\right)$

Il n'en ressort rien (ϕ). On pouvait s'y attendre au regard du v^2 d'Origine

G3

	p_l/g	p_l/a	phi
Autres	0.0%	15.4%	-0.154
Europe	50.0%	42.3%	0.053
France	50.0%	42.3%	0.053

Malgré le v^2, on constate (ϕ) une *légère* sous-représentation de « Autres » parmi le groupe G3

Travailler sur un effectif (plus) élevé

ETUDE DE CAS N°2

Classification des clients d'une banque (quelconque)

Avec un tel effectif (n = 39 919 obs.), les vt auront tendance à gonfler exagérément. D'où l'intérêt des tailles d'effet.

Dataset description

15 attribute(s)

39919 example(s)

Attribute	Category	Modalités	Description
bank_seniority	Discrete	3 values	ancienneté
account	Discrete	2 values	compte bancaire o/n
credit_card	Discrete	2 values	carte bancaire o/n
privilege_card	Discrete	2 values	carte privilège o/n
savings_ldd	Discrete	2 values	détenteur LDD o/n
savings_cel	Discrete	2 values	détenteur CEL o/n
savings_pel	Discrete	2 values	détenteur PEL o/n
savings_pep	Discrete	2 values	détenteur PEP o/n
revolving_credit	Discrete	2 values	détenteur crédit revolving o/n
revolving_card	Discrete	2 values	détenteur carte revolving o/n
personnal_credit	Discrete	2 values	détenteur crédit personnel o/n
housing_credit	Discrete	2 values	détenteur crédit immobilier o/n
financial_titles	Discrete	2 values	détenteur titres financiers o/n
financial_savings	Discrete	2 values	détenteur épargne o/n

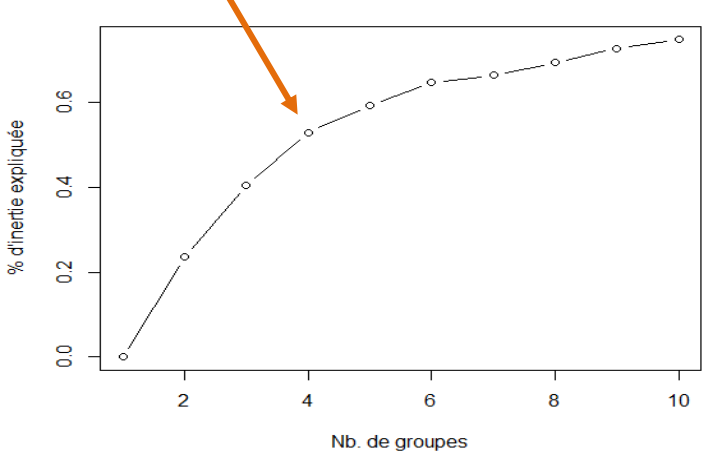
Constitution des groupes

AFCM (analyse factorielle des correspondances multiples) + K-Means

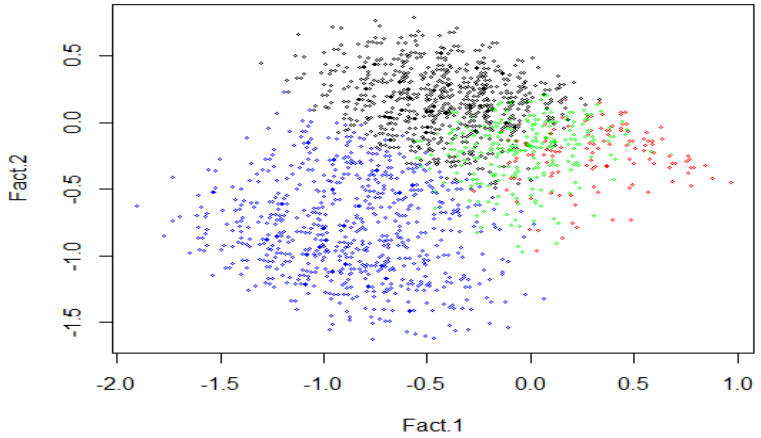
2 étapes (sous R) :

1. Réalisation d'une ACM (package « ca »)
2. K-Means sur les 5 premiers facteurs de l'ACM (package « stats »)

« coude » ?



Courbe de l'évolution de l'inertie expliquée en fonction du nombre de groupes. Un choix de K = 4 est décidé.



Position des groupes dans le premier plan factoriel

	Effectifs
G1	4653
G2	9052
G3	3085
G4	23129

Caractérisation des partitions et des groupes

Les clients récents n'ont pas de compte (encore) apparemment. Ce qui les distingue des autres (il s'agit du groupe G1 si on creuse un peu)

Variable	v^2
revolving_credit	0.777
revolving_card	0.746
financial_titles	0.521
account	0.491
bank_seniority	0.423
savings_ldd	0.147
financial_savings	0.134
credit_card	0.107
personnal_credit	0.104
privilege_card	0.045
housing_credit	0.038
savings_pep	0.037
savings_cel	0.031
savings_pel	0.026

(revolving credit, revolving card) sont les plus déterminants dans la constitution des classes. Ensuite viennent (financial titles, account et bank seniority).

On s'intéresse à deux classes en particulier

G3

		p_l/g	p_l/a	phi
revolving_credit	n	0.013	0.980	-0.881
	y	0.987	0.020	0.881
revolving_card	n	0.237	1.000	-0.864
	y	0.763	0.000	0.864
financial_titles	n	0.130	0.285	-0.093
	y	0.870	0.715	0.093
account	n	0.000	0.067	-0.074
	y	1.000	0.933	0.074
bank_seniority	<=1	0.029	0.087	-0.057
	>=5	0.824	0.731	0.056
	1<senior<5	0.148	0.182	-0.024

Les clients qui abusent du crédit revolving.

G4

		p_l/g	p_l/a	phi
revolving_credit	n	0.978	0.805	0.290
	y	0.022	0.195	-0.290
revolving_card	n	1.000	0.860	0.293
	y	0.000	0.140	-0.293
financial_titles	n	0.034	0.603	-0.631
	y	0.966	0.397	0.631
account	n	0.000	0.146	-0.300
	y	1.000	0.854	0.300
bank_seniority	<=1	0.011	0.180	-0.304
	>=5	0.960	0.432	0.593
	1<senior<5	0.029	0.388	-0.462

Les clients anciens, qui ont des titres financiers, et qui ne sont pas appétents au crédit revolving.

- Les seuils ne sont pas à prendre pour argent comptant, la gradation des valeurs importe aussi.
- Malgré n = 39 919 obs., les indicateurs conservent des valeurs intelligibles ! (*à la différence de la vt*)

Conclusion

Contexte

- La caractérisation des groupes est essentielle en classification automatique
- Les techniques descriptives étudiant le rôle individuel des variables a le mérite de la simplicité
- Mais nous devons disposer d'un indicateur numérique permettant de les hiérarchiser

Solution

- Les mesures basées sur le concept de « taille d'effet » correspond à ce cahier des charges
- Les concepts sous-jacents sont reconnus (proportion de variance expliquée, corrélation)
- Elles sont normalisées, varient entre [0 ; 1] pour les uns, entre [-1; 1] pour les autres
- Elles ne subissent pas d'inflation lorsque nous traitons de grands effectifs

Utilisation

Une utilisation en deux temps est préconisée

- Proportion de variance pour caractériser la partition (η^2 et v^2)
- Corrélation pour interpréter les groupes (r_{pb} et ϕ)

Plus loin

D'autres mesures de taille d'effet existent (Z-factor, SSMD, etc.), mais notre propos était avant tout d'identifier des indicateurs reposant sur des notions statistiques simples permettant de caractériser les classes issues d'un processus de classification automatique.

Références

Références

- Cohen J., « Statistical Power Analysis for the behavioral sciences », Psychology Press, 1988.
- Michael Furr R., « Summary of effect size and their links to inferential statistics », Psychology Department, Wake Forest University, 2008.

- Hsu L. M., « Biases of success rate differences shown in binomial effect display », in Psychological Methods, 9(2), pp. 183-197, 2004.
- McGraw K.O., Wong S.P., « A common language effect size statistic », in Psychological bulletin, 111(2), pp. 361-365, 1992.
- Morineau A., « Note sur la caractérisation statistique d'une classe et les valeurs-tests », in Bulletin Technique du Centre de Statistique et Informatique Appliquées, 2(1-2), pp. 20-27, 1984.
- Morineau A., Rakotomalala R., « Critère VT100 de sélection des règles d'association », in EGC'2006, RNTI-E-6, pp. 581-592, 2006.
- Rakotomalala R. « Interpréter la ''valeur test'' », Tutoriel Tanagra, avril 2008.
- Rosenthal R., Rubin D. B., « A simple, general purpose display of magnitude of experimental effect », in Journal of Educationnal Psychology, 74(2), pp. 166-169, 1982.
- Rosnow R.L., Rosenthal R., Rubin D. B., « Contrasts and correlations in effect-size estimation », in Psychological Science, 11(6), pp. 446-453, 2000.