Introduction au traitement statistique des données sensorielles

Benjamin Charlier (2013-2015)
Xavier Bry (2006-2012)

Institut de Montpelliérain Alexander Groethendieck
Université Montpellier
Case courrier 051
34095 Montpellier, Cedex 5
Benjamin.Charlier@umontpellier.fr

janvier 2016

Plan du cours

Références

Source : ce cours est en (très grande) partie issu du cours de X. Bry.

Livres : spécifiques à l'analyse sensorielle ou plus généralistes :

- I. Urdapilleta et al. *Traiter d'évaluation sensorielle - Aspects cognitifs et métrologiques des perceptions*. Dunod (2001)
- F. Depledt et al. *Évaluation sensorielle - manuel méthodologique* - Lavoisier (2009)
- J.H. Zar *Biostatistical analysis* - Prentice Hall (2010)

Internet : vous pourrez y trouver :

- Le site de F. Husson, `http://math.agrocampus-ouest.fr/infoglueDeliverLive/membres/Francois.Husson`
- Un package R spécialisé, `http://sensominer.free.fr/`
- Nombreux articles spécialisés (*e.g* recherche avec google scholar)

Sommaire

Objectifs

Objectifs : analyser, étudier, comparer un ensemble de produits à l'aide d'un ou plusieurs sens (goût, odorat, toucher, ouïe, vue)

Domaines d'applications : produits alimentaires, cosmétiques, industrie automobile (description de tableaux de bord, bruit de portières...), ergonomie (ressenti du confort thermique de la pièce), sport (grip de raquette de tennis, confort vêtements. . .)

Particularité des mesures : les instruments de mesure sont des êtres humains (problème de qualité de la mesure, fiabilité et répétabilité $\Longrightarrow$ traitement statistique particulier)

Quelques problématiques : nous allons aborder les questions suivantes

- estimation (créer le profil sensoriel d'un produit)
- comparaison (peut on caractériser une gamme de produits ?)
- discrimination (ces produits sont-ils différents ?),
- classement (ordre de préférence),

. . . mais d'autres questions peuvent être posées :

- lien entre les descriptions sensorielles des produits et leur description physico-chimique ?
- lien entre les préférences des produits et leur description sensorielle ?

Outils

Disciplines concernées :

Origine des *stimuli* : chimie, physique, biologie

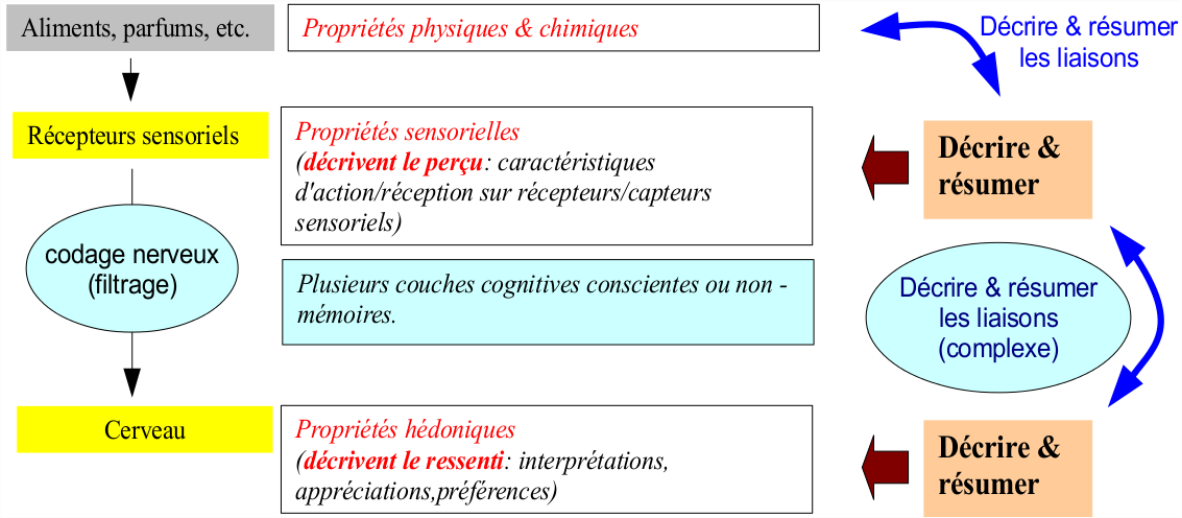
Perception des *stimuli* : neurobiologie, sciences cognitives, physiologie

Recueillir les données : métrologie sensorielle (quelles données ? Comment recueillir les données ?)

Traiter les données : statistique. C'est une branche des mathématiques qui utilise le formalisme des probabilités et qui a pour objet d'étude l'analyse d'informations à des fins de décisions. Démarche statistique générale :

- modélisation
- traitement des données (statistique descriptive, data mining)
- aide à la décision (statistique inférentielle)

Différents niveaux de descriptions



Les données

On en distingue deux grands types :

Évaluations qualitatives et quantitatives : description d'un produit par des données "objectives" (saveurs : salé, amer ; arômes : boisés, fruits rouges ; textures : lisse, pâteux...) avec existence de références (*e.g* concentration en caféine pour évaluer l'amertume...). Mise en place d'un profil sensoriel du produit.
Jury : panel restreint "d'experts" (10 à 12 testeurs entrainés à partir de références, par exemple avec des solutions de différentes concentrations)

Évaluations hédoniques : données subjectives (on teste la notion de plaisir, l'appréciation du produit). Elles sont liées au sujet : par exemple, certains apprécient le très sucré, d'autres non (cible marketing).
Jury : panel plus large de sujets "naïfs" (minimum 50-100 consommateurs)

Dans ce cours on ne traitera que le cas des évaluations qualitatives et quantitatives.

Recueil des données

Instrument de mesure : Sélection, entrainement, étalonnage du jury à l'aide de séances spécifiques (améliorer la reproductibilité et la cohérence des réponses, . . .)

Protocole : Le déroulement d'une évaluation est codifié (normes sur les conditions atmosphériques, la salle avec box séparés, pause entre les évaluations,. . .) pour éviter les biais et tenter de garantir l'*indépendance* des jugements (très utile pour le traitement des données !).

Plan d'expérience : La présentation des produits et le questionnaire jouent un rôle important car ils orientent le jugement :

- monadique séquentielle, "un par un" (biais par effet de rang, effet de report)
- tous les produits sur un plateau (le juge peut revenir sur son jugement)
- par paire (test la préférence ; lequel préférez vous ?)
- . . .

En pratique : Fortes contraintes de coûts qui influent sur le recueil des données (effectif du jury, nombre de répétitions, . . .)

Type de données

Suivant les besoins, on recueille les données suivant :

Échelle continue : on observe des nombres réels.

Exemple : On cherche à évaluer l'intensité du caractère "sucré" de deux produits A et B. On demande de faire une marque sur une ligne continue

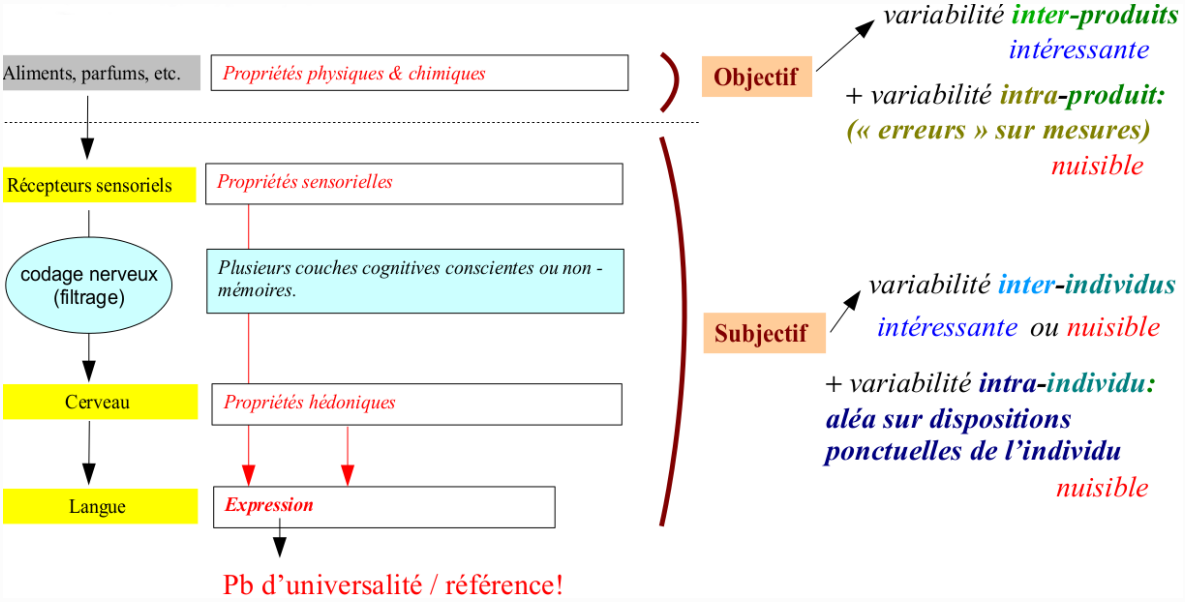
Échelle discrète : on a des variables ordinales.

Exemple : On demande de cocher la case correspondante sur une échelle discrète :

0	1	2	3	4	5	6	7

Classement : classer A et B du plus au moins sucré. Les données sont les rangs.

Différentes sources de variabilité



Un premier modèle additif

Variabilité d'une mesure :
- Part de se que l'on veut saisir : *structurelle*
- Part de se que l'on ne veut pas saisir : *aléatoire*

Estimer une intensité par un juge : On suppose que l'on a un produit dont on veut évaluer la note sur une échelle d'intensité. Le juge va noter ce produit L fois. Pour chaque note $\ell = 1, \cdots, L$ on a suppose que l'on a :

où
- S_ℓ : note donnée par le juge au ℓ-ième essai.
- c^0 : note "théorique" du produit
- β_ℓ : perturbation aléatoire

Idée : estimer c^0 par la moyenne des notes . De plus, plus le nombre L de mesures est grand meilleur devrait être l'estimation.

Échelle d'intensité

Échelle de référence : elle concerne une dimension sensorielle précise pour permettre l'évaluation d'un produit. On étalonne les degrés de l'échelle à l'aide de solutions de différentes concentrations.

Loi de Fechner : L'intensité i de la sensation perçue est proportionnelle au logarithme de l'intensité de la stimulation. Ici, on considère des concentrations c_i :

Si on prend une suite géométrique de concentrations :

on devrait définir une échelle d'intensités linéaire. La raison r doit être voisine (supérieurement) du seuil de détectabilité pour que l'échelle soit assez précise.

Exemple : Évaluer une solution d'aspartame à $200mg/L$ sur une échelle de concentrations en saccharose (en g/L) :

intensité i	1	2	3	4	5	6	7	8	9
concentration c_i	7.5	10.6	15					84.9	120

Ici $r =$ et $c_1 =$

Méthode Up & down

Objectif : Estimer l'intensité d'un *stimulus* S sur une échelle d'intensité.

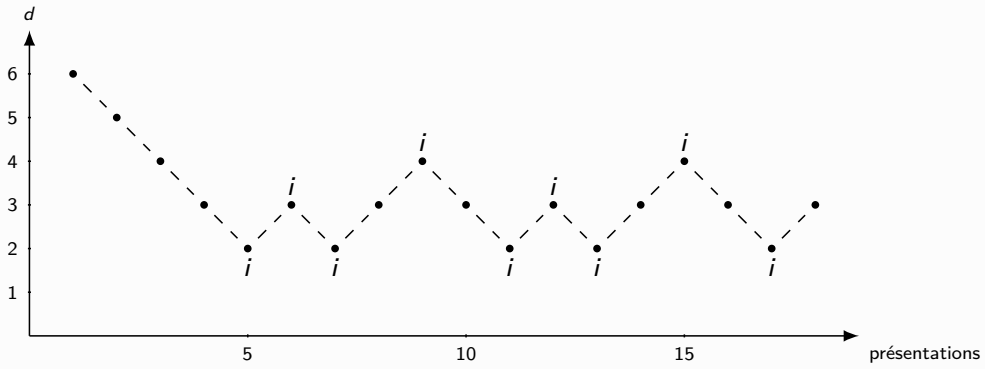
Épreuve : On présente successivement à un juge entraîné des paires (*stimulus*, référence) $= (S, c_i)$. Le *stimulus* reste constant, et la référence varie comme suit :

- Initialisation : la 1ère référence est un c_i bien choisi (proche de S).
- Règle de progression :
  - Si on juge $c_i < S$ alors on incrémente $i \leftarrow i + 1$ (*i.e* on augmente l'intensité de la référence)
  - Si on juge $c_i > S$ alors on décrémente $i \leftarrow i - 1$ (*i.e* on baisse l'intensité de la référence)
- Règle d'arrêt : On fixe le nombre n d'inversions du sens de variation des degrés successifs, (*e.g* $n = 9$). On arrête les présentations lorsque ce nombre est atteint.

Résulat : On ne tient pas compte de ce qu'il se passe avant les deux premières inversions (phase de rodage). Le niveau du *stimulus* est estimé par la moyenne des niveaux d'inversion situés après les deux premières inversions.

Méthode Up & down

Exemple (aspartame) :



L'intensité perçue i est estimé à : La concentration en saccharose équivalente est alors

Limitations : Le nombre N d'essais par juge pour un seul produit peut être grand : On peut abaisser N pour être plus expéditif, ce qui augmente la variabilité de l'estimation pour le juge. Mais en utilisant plusieurs juges et en faisant la moyenne de leurs estimations, on peut compenser cette augmentation de la variabilité.

Définition

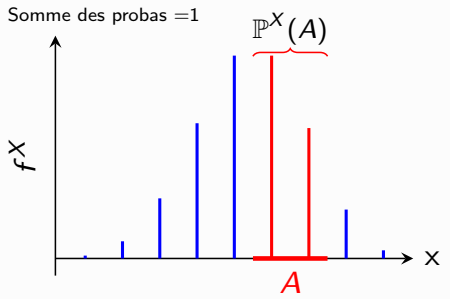
Variable aléatoire (v.a) : c'est une fonction X définie sur l'ensemble des éventualités (noté Ω) d'une expérience aléatoire. Étant donné un évènement $\omega \in \Omega$, la valeur $X(\omega)$ prise par X est notée x (en minuscule). On note $X(\omega) = x$ ou $X = x$.

Support d'une v.a : c'est l'ensemble $X(\Omega)$ des valeurs qui peuvent être prises par X. On dit que la v.a est quantitative si elle renvoie des nombres ou qualitative si elle renvoie des noms.

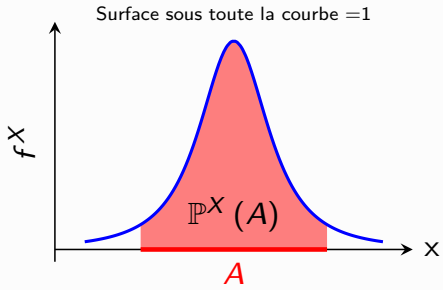
Évenement : c'est un ensemble d'éventualités. Autrement dit, c'est une partie de $X(\Omega)$. On note l'ensemble des évènements possibles $\mathcal{P}(X(\Omega))$.

Loi d'une v.a : c'est la distribution théorique des valeurs prises par X. C'est une probabilité $\mathbb{P}^X : \mathcal{P}(X(\Omega)) \to [0,1]$. On note pour tout évènement $A \in \mathcal{P}(X(\Omega))$

$$\mathbb{P}^X(A) = \mathbb{P}(X \in A)$$



(a) v.a discrète : $X(\Omega) = \{x_1, \ldots, x_n, \ldots\}$



(b) v.a continue : $X(\Omega) \subset \mathbb{R}$, admet une densité f^X

Indépendance et échantillon

Indépendance : deux v.a X et Y sont indépendantes si la distribution des valeurs prises par X n'influe pas sur la distribution des valeurs prises par Y. Autrement dit, si on a

Échantillon : L'expérience aléatoire fournissant la v.a X est répétée indépendamment n fois. Cela fournit une nouvelle v.a qui est un n-uplet $(X_1, \cdots, X_n)$: c'est l'échantillon. On dit que les v.a X_i sont indépendantes et identiquement distribuées (i.i.d). Les valeurs effectivement observées au cours de ces réalisations sont notées $(x_1, \cdots, x_n)$ en minuscules.

Moyenne empirique : c'est la moyenne d'un échantillon. C'est donc la valeur prise par la v.a

où $(X_1, \cdots, X_n)$ est un n-échantillon i.i.d. Si les valeurs observées sont $(x_1, \cdots, x_n)$ alors on a $\bar{X}_n = \bar{x}_n$ avec

Lois Gaussiennes

Lois normales/gaussiennes : $\mathcal{N}(\mu, \sigma^2)$ pour tout $\mu \in \mathbb{R}$ et tout $\sigma^2 > 0$. Elle a pour densité

pour tout $x \in \mathbb{R}$.

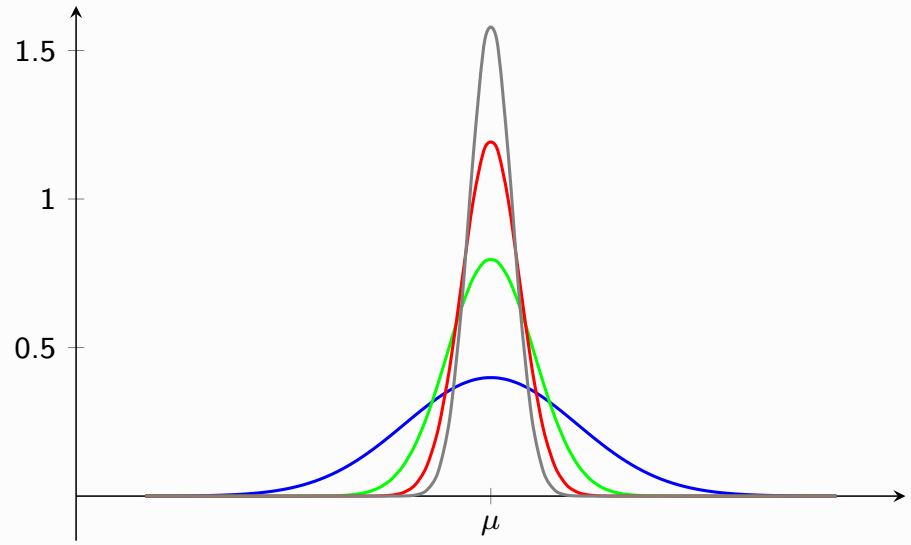


FIGURE – Bleu : $\mathcal{N}(\mu, 1)$. Vert : $\mathcal{N}(\mu, \frac{1}{2})$. Rouge : $\mathcal{N}(\mu, \frac{1}{3})$. Gris : $\mathcal{N}(\mu, \frac{1}{4})$

Lois continues issues de la Gaussienne

Lois du Khi-deux : $\chi^2(d)$ pour tout $d \in \mathbb{N}^*$. Le paramètre d est appelé nombre de degrés de liberté. Elle est construite à partir de la loi normale de la façon suivante : soient $X_1, \cdots, X_d \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, et soit

Alors $K_d \sim \chi^2(d)$.

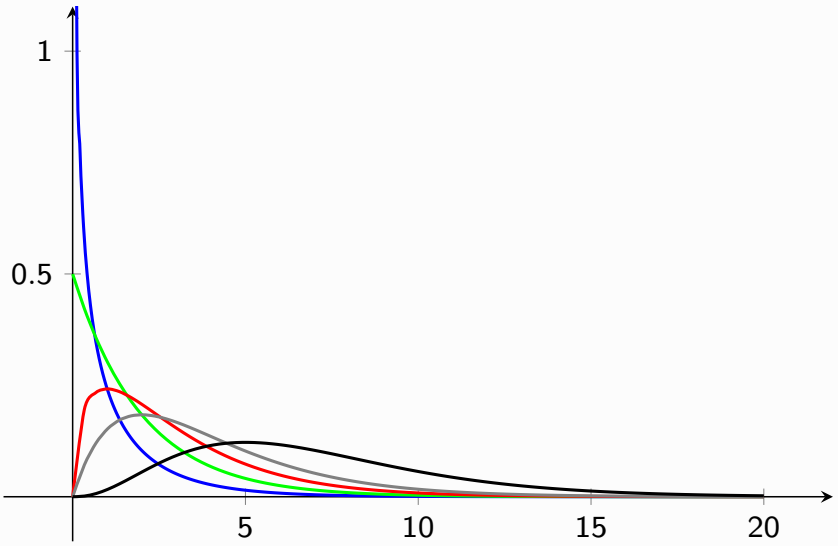


FIGURE – Bleu : $\chi^2(1)$. Vert : $\chi^2(2)$. Rouge : $\chi^2(3)$. Gris : $\chi^2(4)$. Noir : $\chi^2(7)$.

Lois continues issues de la Gaussienne

Lois de Student : $\mathcal{T}(d)$ pour tout $d \in \mathbb{N}^*$. Le paramètre d est appelé nombre de degré de liberté. Elle est construite à partir de la loi normale et de la loi du χ^2. Soient deux v.a indépendantes $X \sim \mathcal{N}(0,1)$ et $K_d \sim \chi^2(d)$. Alors

Approximation de loi : Lorsque d est grand, on peut approximer les quantiles de la loi de Student par ceux d'une loi normale de paramètre $\mathcal{N}(0,1)$ (disons, $d \geq 30$ pour le quantile 0.95). Graphiquement cela donne :

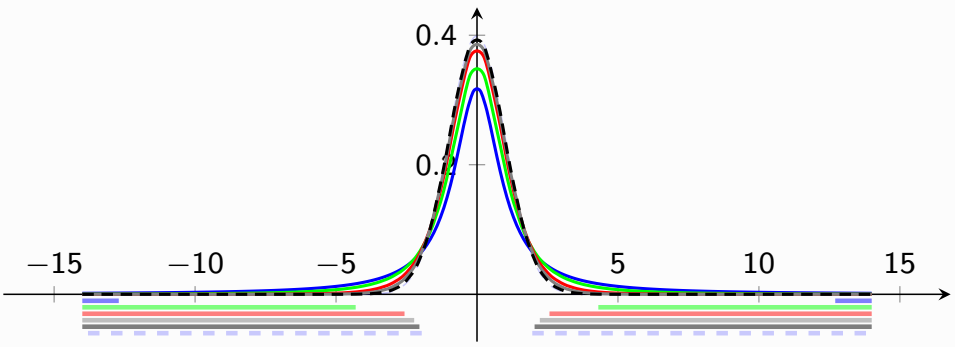


FIGURE – Bleu : $\mathcal{T}(1)$. Vert :$\mathcal{T}(2)$. Rouge : $\mathcal{T}(5)$. Gris : $\mathcal{T}(10)$. Noir : $\mathcal{T}(30)$

Lois discrètes issues du schéma de Bernoulli

Lois de Bernoulli : $Bern(p)$, pour tout $0 < p < 1$. Les plus simples : $X(\Omega) = \{0, 1\}$ et

Loi binomiale : $Bin(n, p)$ pour tout $n \in \mathbb{N} \setminus \{0\}$ et $0 < p < 1$. C'est la loi du nombre de succès lorsque l'on répète n fois de manière indépendante un schéma de Bernoulli. On a $X(\Omega) = \{0, 1, \cdots, n\}$

Approximation de loi : Lorsque n, np et $n(1-p)$ sont grands (disons, $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 10$ ou $np \geq 10$ et $n(1-p) \geq 10$ pour une meilleur précision) on peut approximer la loi dicrète binomiale par une loi normale (qui est continue) de moyenne np et variance $np(1-p)$). Graphiquement cela donne :

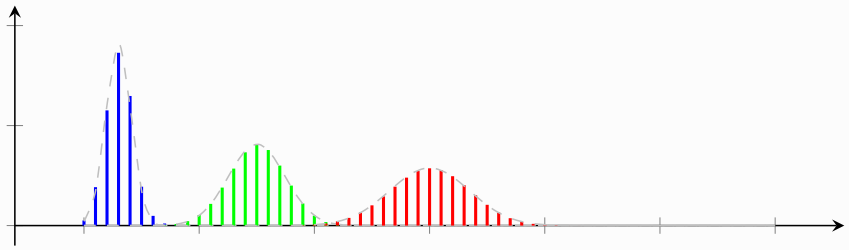


FIGURE – Bleu : $Bin(5, .6)$. Vert : $Bin(25, .6)$. Rouge : $Bin(50, .6)$

Espérance

Espérance d'une v.a X : c'est la moyenne des valeurs (théoriques) prises par X. C'est un paramètre de localisation noté $\mathbb{E}(X)$. On a

v.a discrète :

v.a continue :

Propriétés

- *Pour tout X, Y v.a et tout $a, b \in \mathbb{R}$:* $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$
- *Pour tout X, Y v.a* **indépendantes** *:* $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$

Conséquence : Pour un n échantillon $(X_1, \cdots, X_n)$ i.i.d satisfaisant $\mathbb{E}(X_i) = \mu$ pour tout $i = 1, \cdots, n$ on a

Cela signifie que la moyenne empirique est centrée sur la moyenne (théorique).

Variance

Variance de X : C'est une mesure de dispersion pour les valeurs (théoriques) prises par X. C'est un nombre réel positif ou nul défini par

Remarque : $Var(X) = 0$ ssi $X = cte$ ssi X n'est pas aléatoire.

Propriétés

Soit X une v.a et $a, b \in \mathbb{R}$

$$Var(aX + b) = a^2 Var(X).$$

Si X et Y sont deux v.a **indépendantes** *alors*

$$Var(X + Y) = Var(X) + Var(Y)$$

Conséquence : Pour un n échantillon $(X_1, \cdots, X_n)$ i.i.d on a $\mathbb{E}(X_i) = \mu$ et $Var(X_i) = \sigma^2$ pour tout $i = 1, \cdots, n$. Alors

Cela signifie que les fluctuations de la moyenne empirique diminuent lorsque n augmente...

Covariance

Covariance de X et Y : c'est le nombre

Propriétés

Pour tout X, Y v.a et $a, b \in \mathbb{R}$

$$Cov(X, X) = Var(X)$$
$$Cov(X, Y) = Cov(Y, X)$$
$$Cov(aX + b, Y) = aCov(Y, X)$$

Si X et Y sont indépendantes alors

$$Cov(X, Y) = 0.$$

Attention : Deux v.a avec une covariance nulle ne sont pas nécessairement indépendantes !

Généralités

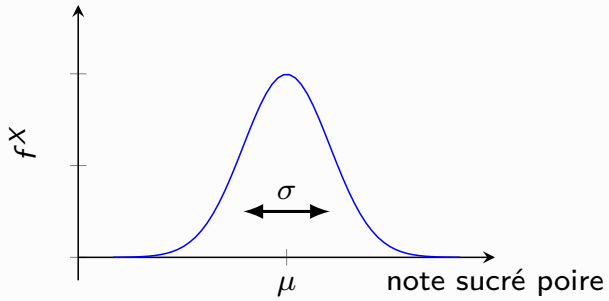
Objectif : On cherche à déterminer la loi d'une v.a à partir des valeurs d'un échantillon. Pour simplifier un peu le problème : on suppose que la loi de X appartient à une famille décrite par un nombre *fini* de paramètres. On cherche alors a estimer ces paramètres au vu des observations.

Modélisation : la note sucrée d'une poire william issue d'une production donnée est notée X. On suppose que

où

- $\mu \in \mathbb{R}$: note sucrée moyenne théorique
- $\varepsilon \sim \mathcal{N}(0, \sigma^2)$: v.a centrée modélisant la variabilité suivant les observations

La loi de X est une loi Normale à deux paramètres inconnus : la moyenne μ et la variance σ^2. On note $X \sim \mathcal{N}(\mu, \sigma^2)$ ou $P^X = \mathcal{N}(\mu, \sigma^2)$.



Limitations et garanties théoriques

Non Adéquation du modèle : Attention ! La variable étudiée ne suit, le plus souvent, une loi paramétrique que sous des hypothèses simplicatrices, parfois peu réalistes... Beaucoup de phénomènes courants sont trop complexes pour avoir une loi simple et connue...

Principe d'invariance : Heureusement, il existe une série de résultats connus sous le nom de théorèmes de la limite centrale (TCL) qui fait que même si on ne connaît pas la loi de X, on connaît à peu près celle de $\bar{X}_n$: c'est (asymptotiquement) une loi Normale.

Théorème Central Limite

X de loi *quelconque* P^X d'espérance μ et de variance σ^2. Soit $(X_1, \cdots, X_n)$ un n-échantillon i.i.d de loi P^X. Alors, si "n est assez grand" :

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \approx \mathcal{N}(0,1) \Leftrightarrow \bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Un corollaire important : si n est grand, on a $\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \approx \mathcal{N}(0,1)$. (On peut remplacer σ par un estimateur)
- si X est Gaussien alors on peut remplacer les $\approx$ ("est approximativement de loi") par des $\sim$ ("est de loi").

Modéliser l'incertitude

Variabilité d'une mesure :
- Part de se que l'on veut saisir : *structurelle*
- Part de se que l'on ne veut pas saisir : *aléatoire*

Modèle statistique : On suppose que l'on a K biscuits que l'on fait goûter par J juges. Tous les juges goûtent tous les biscuits et donnent une note de sucré :

où
- S_j^k : note de sucré du produit k par le juge j (observation)
- b_j : différence de niveau dans les notes de chaque juge (effet juge)
- c^k : différence de niveau de sucre dans les biscuits (effet produit)
- ε_j^k : perturbation aléatoire, résidu (erreur)

Estimer l'effet produit : Dans la pratique on s'intéresse souvent à un produit $A \in \{1, \cdots, K\}$ en particulier. On a donc J observations (chaque juge ne fait qu'une observation) :

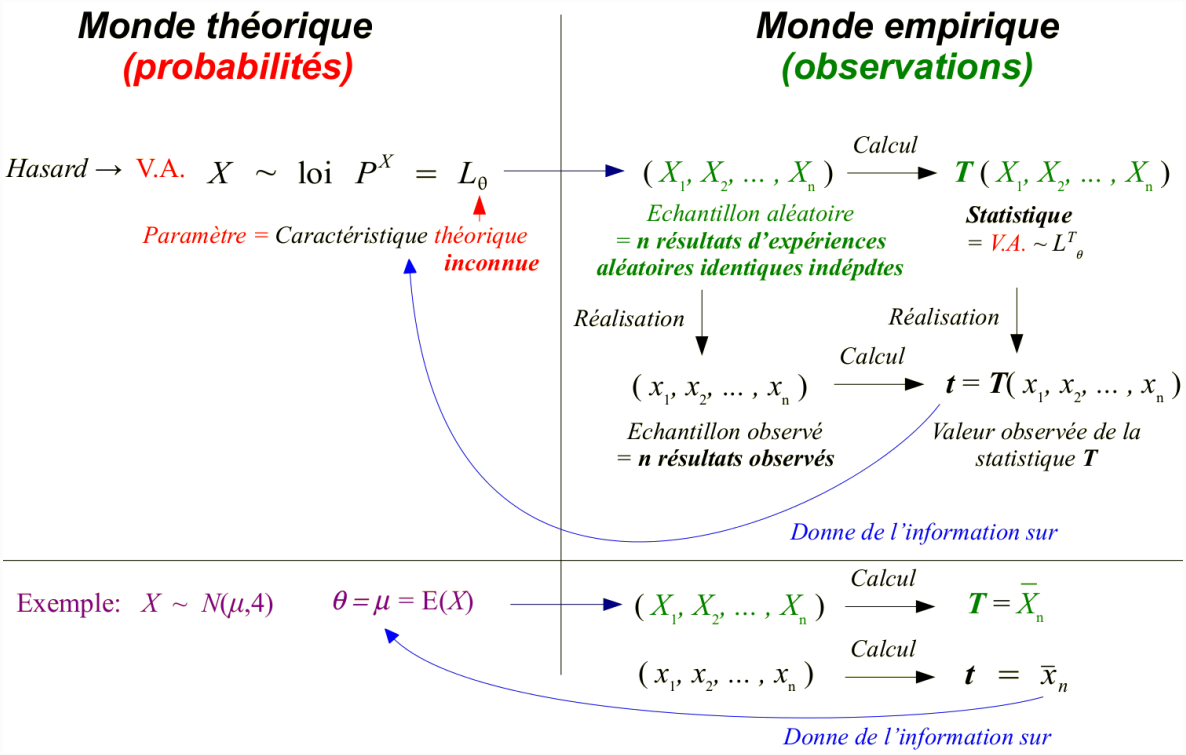
L'effet juge devient une erreur résiduelle qui sera petite en moyenne si elle est centrée...

Pour être identifiable la part structurelle doit se retrouver dans *plusieurs* mesures.

Sommaire

2.1. Estimation :

Principes



Moyenne et variance empirique

Soit $(X_1, \cdots, X_n)$ un n échantillon i.i.d de loi P^X de moyenne μ et de variance σ^2.

Estimateur de la moyenne μ : c'est simplement la moyenne empirique $\bar{X}_n$. On a

- $\mathbb{E}(\bar{X}_n) = \mu$: centrée sur μ (sans biais)
- $Var(\bar{X}_n) \simeq \frac{\sigma^2}{n}$ dispersion $\to 0$ quand $n \to \infty$ (convergent).
- dans le cas Gaussien : $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

Estimateur de la variance : C'est la variance empirique définie par

$$S_n^2 =$$

On peut montrer que l'on a

- $\mathbb{E}(S_n^2) = \sigma^2$: centrée sur μ (sans biais)
- S_n^2 est bien convergent vers σ^2.
- Dans le cas Gaussien : $\frac{n-1}{\sigma^2} S_n^2$ suit une loi du $\chi^2(n-1)$.

Remarques :

- Dans le cas Gaussien : $\bar{X}_n$ et S_n^2 sont indépendants (*cf* Lois de Student et TCL).
- Ne pas confondre estimateur de la variance (*e.g* S_n^2) et variance d'un estimateur (*e.g* $Var(\bar{X}_n)$)

Principe et exemple

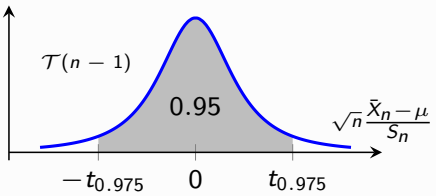
Objectifs : A partir de l'échantillon, on veut construire une "fourchette" de valeurs ayant une probabilité suffisante $1-\alpha$ (*e.g* $\alpha = 0.05$) de contenir la vraie valeur (inconnue) du paramètre.

Construction : On part d'un intervalle de valeurs auquel l'estimateur du paramètre appartient avec la probabilité $1-\alpha$ (celui-ci ne peut être calculé que si on connaît la loi de l'estimateur). Puis, on "retourne" l'intervalle...

Exemple : Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ où σ^2 et μ sont inconnus. On cherche un intervalle de confiance sur μ : on a $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \underbrace{\sim}_{\text{car } X \text{ est Gaussien}}$. Reste à "injecter" l'estimateur S_n^2 de σ^2 dans la formule précédente :

$$\sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \underbrace{\sim}_{\text{loi de Student}} \underbrace{\approx}_{\text{si } n \text{ assez grand } \geq 30}$$

En utilisant une table de quantiles on a la valeur $t_{1-\frac{\alpha}{2}} = t_{1-\frac{\alpha}{2}}(n-1) > 0$ telle que



$$\mathbb{P}\left(-t_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}_n - \mu}{S_n} \leq t_{1-\frac{\alpha}{2}}\right) = 1-\alpha$$

$$\Leftrightarrow \mathbb{P}\left(\bar{X}_n - t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}} \leq \mu \leq \bar{X}_n + t_{1-\frac{\alpha}{2}}\frac{S_n}{\sqrt{n}}\right) = 1-\alpha$$

l'IC pour μ de niveau α est $\left[\bar{x}_n - t_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}, \bar{x}_n + t_{1-\frac{\alpha}{2}}\frac{s_n}{\sqrt{n}}\right]$, ou $\bar{x}_n$ et s_n sont les valeurs observées.

En pratique

Variable quantitative : On considère le modèle

$$X = \mu + \varepsilon \text{ où } \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

On estime μ à l'aide de la moyenne empirique $\bar{X} \sim$ et σ^2 à l'aide de la variance empirique $S^2 \sim$. On a

$$\sqrt{n}\left(\frac{\bar{X}-\mu}{S}\right) \sim T(n-1) \underbrace{\approx}_{\text{si } n \text{ grand}} \quad \text{et } IC_{1-\alpha} = \left[\bar{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right] \underbrace{\simeq}_{\text{si } n \text{ grand}} .$$

Variable qualitative : (X prend des valeurs discrètes) On veut estimer $p = \mathbb{P}(X = m)$ pour chaque modalité m. On considère l'indicatrice $\mathbb{1}_{X=m} \sim Bern(p)$. On estime $p = \mathbb{E}(\mathbb{1}_{X=m})$ par la moyenne empirique $\hat{P} = \overline{\mathbb{1}_{X=m}}$. Si n est grand :

$$\sqrt{n}\frac{\hat{P}-p}{\sqrt{\hat{P}(1-\hat{P})}} \approx \mathcal{N}(0,1) \quad \text{et} \quad IC_{1-\alpha} \simeq \left[\hat{P} \pm u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}\right]$$

Remarque : on a utilisé l'estimateur $\hat{P}(1-\hat{P})$ pour la variance de $\mathbb{1}_{X=m}$. Mais comme on a $t(1-t) \leq \frac{1}{4}$ pour tout $t \in [0,1]$, on peut prendre $IC_{1-\alpha} \simeq \left[\hat{P} \pm u_{1-\frac{\alpha}{2}} \frac{1}{2\sqrt{n}}\right]$ pour simplifier.

Généralités

Contexte : un unique produit est noté par n juges à l'aide de descripteurs quantitatifs (*e.g* sucré) et/ou qualitatifs (*e.g* saveur).

Outils : Statistique univariée (description de la distribution). Différentes techniques suivant le type de variables :

- variables quantitatives : moyenne, quantile, variance, boxplot,...
- variables qualitatives : fréquences des modalités, diagrammes,...

Convergence du jury : Les disparités inter-juges ne sont pas intéressantes ici. Un jury est dit *non-convergent* si (pour des variables quantitatives) : il y a beaucoup d'observations atypiques, une distribution des observations non symétrique, une variance trop élevée, ...

Profil sensoriel

Contexte : un unique produit est noté par n juges sur J descripteurs quantitatifs.
Profil sensorielle : c'est l'ensemble des descripteurs et de la mesure associée.
Exemple : profils d'un ketshup A

descripteurs	confit	sucré	métal	astringent	acide	vert
juge 1	10	8	5	2	4	6
juge 2	8	9	3	3	4	4
⋮			⋮			

Représentation : par exemple en graphique "radar" :

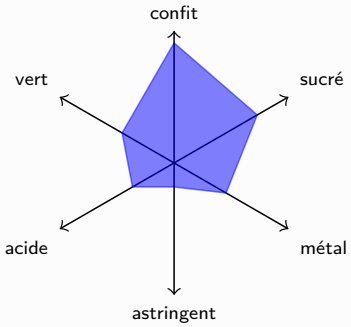


FIGURE – un seul juge

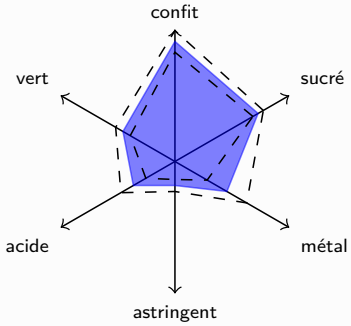


FIGURE – plusieurs juges (intervalles de confiance)

Généralisation : pour les variables qualitatives on représente l'indicatrice.

Relation entre les variables

Problématique : Deux variables X et Y ont été mesurées sur un même échantillon. On dit qu'il existe une *relation* entre X et Y si l'attribution des modalités de X et de Y ne se fait pas au hasard, c'est à dire si les valeurs de X *dépendent* des valeurs de Y ou si les valeurs de Y dépendent des valeurs de X. En bref, la connaissance des valeurs de X permet "de prédire" celles de Y et/ou inversement.

Questions : comment mesurer la relation entre deux variables, l'influence de l'une sur l'autre, tester l'indépendance entre les deux ? Les techniques utilisées seront différentes suivant la nature des variables.

Tableau de contingence

La distribution d'une variable est la donnée des modalités accompagnées de leur effectif. On note n la taille de l'échantillon.

Distribution jointe : modalités : $m_1, \cdots, m_K$ pour X et $\ell_1, \cdots, \ell_L$ pour Y. Effectif : $n_{i,j}$ nombre d'individus pour lesquels $X = m_i$ et $Y = \ell_j$. Résumés dans le tableau de contingence :

X \ Y	ℓ_1	$\cdots$	ℓ_L
m_1	$n_{1,1}$	$\cdots$	$n_{1,L}$
$\vdots$	$\cdots$	$\cdots$	$\cdots$
m_K	$n_{K,1}$	$\cdots$	$n_{K,L}$

On a $n = \sum_{i=1}^{K} \sum_{j=1}^{L} n_{i,j}$. Les fréquences associées à cette distribution jointe sont données par : $f_{i,j} = \frac{n_{i,j}}{n}$

Exemple : On étudie la présence ou non de diverses tonalités aromatiques de fruits rouges/fruits noirs d'un vin en fonction des stades de maturations du raisin.

F \ Mat	s_1	s_2	s_3	s_4
oui	113	54	27	56
non	33	14	15	24

F \ Mat	s_1	s_2	s_3	s_4
oui		0.1607	0.0804	
non	0.0982	0.0417		

on a $n =$

Distributions marginales

Distributions marginales de X et de Y : obtenues à partir de la distribution jointe. Effectifs marginaux de X : $n_{i,.} = \sum_{j=1}^{L} n_{i,j}$ pour tout $i = 1, \cdots, K$ (additionne les effectifs par ligne). Effectifs marginaux de Y : $n_{.,j} = \sum_{i=1}^{K} n_{i,j}$ pour tout $j = 1, \cdots, L$ (somme sur les colonnes). Ils sont écrits dans les marges du tableau de contingence.

X \ Y	ℓ_1	$\cdots$	ℓ_L	Total
m_1	$n_{1,1}$	$\cdots$	$n_{1,L}$	$n_{1,.}$
$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\vdots$
m_K	$n_{K,1}$	$\cdots$	$n_{K,L}$	$n_{K,.}$
Total	$n_{.,1}$	$\cdots$	$n_{.,L}$	n

On a bien $\sum_{i=1}^{K} n_{i,.} = \sum_{j=1}^{L} n_{.,j} = n$. Les fréquences marginales sont données par $f_{i,.} = \frac{n_{i,.}}{n}$ et $f_{.,j} = \frac{n_{.,j}}{n}$

Exemple : effectifs marginaux

F \ Mat	s_1	s_2	s_3	s_4	Total
oui	113	54	27	56	250
non	33	14	15	24	
Total		68	42		

Distributions conditionnelles

On compare les distributions de Y conditionnellement aux valeurs de X (On peut intervertir le rôle des deux variables). On partitionne l'échantillon en K sous-échantillons : on note E_{m_i} le sous-échantillon ne contenant que les $n_{i,.}$ individus pour lesquels $X = m_i$.

Distributions conditionnelles de Y : ce sont les K distributions de Y sur ces K sous-échantillons. Par exemple, si $i_0 \in \{1, \cdots, K\}$ est fixé alors la distribution conditionnelle de Y sachant que $X = m_{i_0}$ est notée $Y_{X=m_{i_0}}$ ou $Y_{m_{i_0}}$. On a le tableau :

$Y_{m_{i_0}}$	ℓ_1	$\cdots$	ℓ_L	Total
Effectifs	$n_{i_0,1}$	$\cdots$	$n_{i_0,L}$	$n_{i_0,.}$
Fréquences	$\frac{n_{i_0,1}}{n_{i_0,.}}$	$\cdots$	$\frac{n_{i_0,L}}{n_{i_0,.}}$	1

Les fréquences conditionnelles sont les proportions des modalités de Y parmi les individus du sous-échantillon E_{m_i} (*i.e* le rapport entre effectifs conditionnels et effectif du sous échantillon).

Remarque : De même on peut calculer les lois conditionnelles de X suivant les valeurs de Y. On peut partitionner l'échantillon en L sous-échantillons F_{ℓ_j} (de taille $n_{.,j}$) suivant la valeur de Y. Les distributions conditionnelles de X sont les L distributions de X (notée $X_{Y=\ell_j}$ ou X_{ℓ_j}) sur ces L sous-échantillons.

Exemple

On représente les distributions conditionnelles de *Mat* (distributions de *Mat* conditionnellement aux valeur de F). Il y en a autant que de modalités pour F :

$Mat_{F=\text{oui}}$	s_1	s_2	s_3	s_4	Total
Effectifs	113	54	27	56	250
Fréquence		0.2160			1

$Mat_{F=\text{non}}$	s_1	s_2	s_3	s_4	Total
Effectifs	33	14	15	24	86
Fréquence		0.1628	0.1744		

On représente de plus la marginale de *Mat* déjà calculée plus haut :

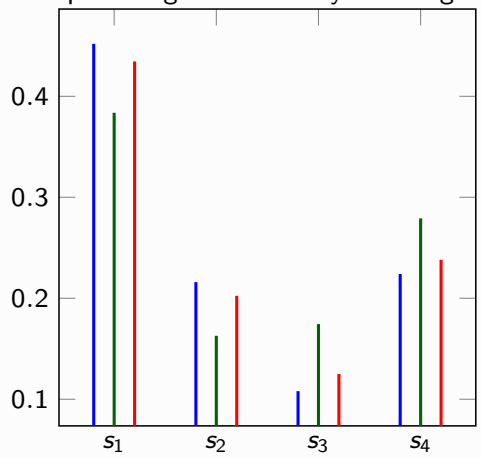
Mat	s_1	s_2	s_3	s_4	Total
Effectifs	146	68	42	80	
Fréquence	0.4345			0.2381	1

Indépendance et représentations graphiques

On parle d'indépendance entre les variables X et Y lorsque la valeur de X n'a aucune influence sur la valeur de Y, et inversement.

Vérification graphique : Comparer les distributions conditionnelles et la distribution marginale (en fréquences). Par exemple, si X n'a que peu d'influence sur Y, les distributions conditionnelles de Y en fréquences seront toutes sensiblement identiques à la distribution marginale de Y.

Exemple : diagramme en tuyaux d'orgues



En bleu : $Mat_{F=\text{oui}}$.
En vert : $Mat_{F=\text{non}}$.
En rouge : Mat.

Effectifs théoriques d'indépendance

Supposons que X et Y soient indépendants. Dans ce cas, les distributions conditionnelles de X en fréquence sont toutes égales à la distribution marginale de X et on a donc pour tout $i = 1, \cdots, K$ et $j = 1, \cdots, L$:

$$\frac{n_{i,j}}{n_{.,j}} = \frac{n_{i,.}}{n} \Longrightarrow n_{i,j} = \frac{n_{i,.}\, n_{.,j}}{n}.$$

Effectifs théoriques d'indépendance : Ce sont les quantités,

$$\bar{n}_{i,j} \doteq \frac{n_{i,.}\, n_{.,j}}{n}$$

et il y aura donc indépendance dans l'échantillon entre X et Y uniquement si $n_{i,j} = \bar{n}_{i,j}$ pour tout couple (i, j).

Remarque : la comparaison entre les effectifs observés $n_{i,j}$ et les effectifs théoriques d'indépendance $\bar{n}_{i,j}$ nous permet quelques interprétations :

- si $n_{i,j} > \bar{n}_{i,j}$, il existe un phénomène d'attraction entre les modalités m_i et ℓ_j. Chez les individus pour lesquels $X = m_i$, on observe plus souvent $Y = \ell_j$ que dans le reste de la population. De même, chez les individus pour lesquels $Y = \ell_j$, on observe plus souvent $X = m_i$ que dans le reste de la population.
- si $n_{i,j} < \bar{n}_{i,j}$, il existe un phénomène de répulsion entre les modalités m_i et ℓ_j. Chez les individus pour lesquels $X = m_i$, on observe moins souvent $Y = \ell_j$ que dans le reste de la population. De même, chez les individus pour lesquels $Y = \ell_j$, on observe moins souvent $X = m_i$ que dans le reste de la population.

Statistique du χ^2 et indice de Cramer

Statistique du χ^2 : Pour mesurer l'écart à l'indépendance entre X et Y, on utilise généralement la statistique :

$$\chi^2 \doteq$$

Elle est toujours positive, vaut 0 uniquement en cas d'indépendance et est d'autant plus grande que les variables X et Y sont liées dans l'échantillon. Nous (re)verrons plus tard comment utiliser cette statistique pour tester l'indépendance de X et Y dans la population.

Indice de Cramer : On peut normaliser la statistique χ^2 pour récupérer une grandeur invariante par changement de taille de l'échantillon. C'est l'indice de Cramer

$$V \doteq \sqrt{\frac{\chi^2}{(\min\{K, L\} - 1)n}}$$

qui vaut également 0 en cas d'indépendance et est inférieur à 1.

Exemple

Le tableau des effectifs théoriques :

F \ Mat	s_1	s_2	s_3	s_4
oui	108.6310		31.2500	
non		17.4048	10.7500	20.4762

La statistique du χ^2 est $\chi^2 =$ et l'indice de Cramer est $V =$.

Comme vous l'avez peut être déjà vu, le test d'indépendance du χ^2 (H_0 :les deux variables sont indépendantes, contre H_1 : elles ne sont pas indépendantes) compare cette valeur au quantile de la distribution $\chi^2(3)$ (où $3 = (2-1) \times (4-1)$). Ici la p-valeur est l'ordre de 0.2...

Le problème

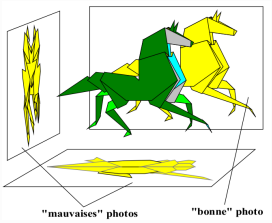
Contexte : On observe un grand nombre de données que l'on stocke dans un tableau. Les lignes représentent les différentes observations tandis que les colonnes représentent les variables.

Exemple :

	Astringence	Boisé	Souplesse	Piquant	Persistance	
Vin 1	11	17	6	8	22	
Vin 2	16	12	15	9	20	Juge 1
Vin 3	5	16	12	5	17	
Vin 1	10	14	9	6	24	
Vin 2	18	13	13	7	17	Juge 2
Vin 3	10	13	12	9	15	
...			...			...

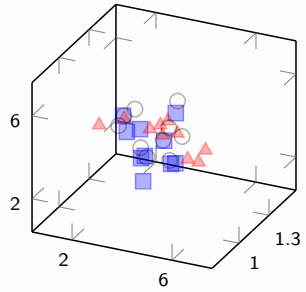
Objectif : trouver une représentation, dans un espace de dimension réduite, permettant de mettre en évidence d'éventuelles structures au sein des données.

Illustration : trouver l'angle de vue et faire une "photo" (dimension réduite) fidèle (détails discernables, étalement maximale) d'un objet 3d



Les méthodes linéaires

Idée : on cherche un sous-espace vectoriels (sev) de faible dimension (droite, plan,...) qui "résume" au mieux les données (simplifie mais conserve les structures).



Plusieurs choix : le choix de ce sous-espace dépend des besoins de l'étude :

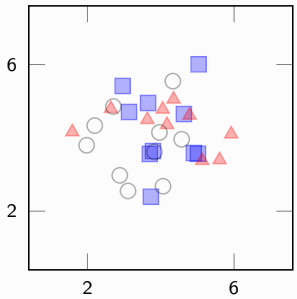


FIGURE – ACP : s.e.v avec la plus grande variabilité

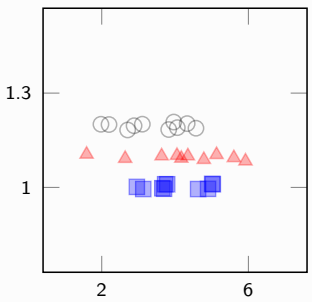


FIGURE – Analyse discriminante : s.e.v où les groupes sont bien séparés.

Présentation

Analyse en composantes principales : Une méthode de statistique exploratoire permettant de décrire un grand tableau de données quantitatives de type individus / variables.

Question : Lorsqu'on étudie simultanément un nombre important de variables (ne serait-ce que 4 !), comment en faire une représentation globale ? On recherche les sous-espaces dans lesquels la projection du nuage déforme le moins possible le nuage initial.

Idée : L'ACP est un simple changement de base : passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des *facteurs*. Les facteurs sont des combinaisons linéaires des variables : ce sont les vecteurs propres de la matrice des corrélations. On ne sélectionne qu'une petite partie des facteurs et cela permet de réaliser des graphiques dans cet espace de faible dimension (dimension égale au nombre de facteurs retenus).

Interprétation : les graphiques peuvent éventuellement permettre de comprendre la structure des données analysées. Cette interprétation est guidée par des indicateurs numériques et graphiques.

Les données

Les données : Un tableau à n lignes et p colonnes représenté par une matrice X de taille $n \times p$:

- la i-ème ligne représente les différentes valeurs des p variables prise par l'individu i.
- la j-ème colonne représente les valeurs de la variable j pour les n individus.

Exemple : Lors d'un concours agricole, un jury a donné des notes à 10 marques de cidres relativement à 10 critères de dégustation. Les marques de cidres sont les individus et les critères gustatifs sont les variables.

cidre	odeur	sucre	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
1	2,14	1,86	3,29	2,29	2	0,14	2,29	1,86	1,29	1,29
2	2,43	0,79	2,71	2,57	2	0,43	2,57	2,86	0,43	0,14
3	2,71	3,14	2,57	2,57	1,43	0,14	2,14	0,86	2,29	1,71
4	3	3,71	2,14	2,07	1,57	0	1,29	1	3,14	3,14
5	3,43	1,29	2,86	3,14	2,17	1	1,86	2,86	1,14	0,29
6	3,14	0,86	2,86	3,79	2,57	0,14	1,71	3,29	0,14	0
7	3,14	1,14	2,86	2,86	2	0,43	1,71	1,86	0,14	0
8	2,43	3,71	3,21	1,57	1,71	0	1	0,57	2,57	2,86
9	5,1	2,86	2,86	3,07	1,79	1,71	0,43	1,43	0,57	2,71
10	3,07	3,14	2,57	3	2	0	0,43	1,29	2,57	3,07

Dualité

Espace E des individus (direct) : Une matrice $n \times p$ code un nuage de n points dans $\mathbb{R}^p$. À chaque individu i on associe le vecteur colonne $x_i = (x_i^1, \cdots, x_i^p)^t \in \mathbb{R}^p$ ("les valeurs d'une ligne mis en colonne"). Ce sont des éléments d'un espace vectoriel E de dimension p : prendre $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice M lui conférant une structure d'espace euclidien.

Espace F des variables (dual) : de la même manière, cette même matrice $n \times p$ code un nuage de p points dans $\mathbb{R}^n$... À la j-ème variable, on associe le vecteur colonne $x^j \in \mathbb{R}^n$. C'est un élément d'un espace vectoriel noté F de dimension n : prendre $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice diagonale $D = diag(w_1, \cdots, w_n)$ lui conférant une structure d'espace euclidien.

En pratique : le plus souvent on a $M = Id_p$ et $D = Id_n$ (*i.e* $w_1 = \cdots = w_n = 1$). Cette formulation plus générale permet d'inclure nombre de cas particuliers dans le formalisme de l'ACP.

ACP normée

Le plus souvent (par défaut dans de nombreuses implémentations de l'ACP), le nuage de points des individus est centré et réduit (*i.e* "on centre et on réduit colonne par colonne")

Centrage : On note $\bar{x} = (\bar{x}^j)_{j=1}^p = \left(\frac{1}{n}\sum_{i=1}^n x_i^j\right)_{j=1}^p \in \mathbb{R}^p$ le vecteur ligne des moyennes pour les p variables. La matrice des données centrées est

$$Y = X - \mathbb{1}_n \bar{x}. \qquad \text{où } \mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = (1, \cdots, 1)^t \in \mathbb{R}^n$$

Réduction : pour éviter les problèmes d'échelles, on réduit chaque colonne de X à variance 1 (cela ne détruit pas la structure de corrélation du nuage de point). On note $s_j^2 = \frac{1}{n}\sum_{i=1}^n (x^j - \bar{x}^j)^2$ la variance de la colonne j et on pose

$$M_{1/s} = Diag\left(\frac{1}{s_1}, \cdots, \frac{1}{s_p}\right) \in \mathbb{R}^{p \times p}.$$

Travailler avec la métrique $M = M_{1/s^2}$ sur le tableau centré Y revient donc à travailler avec la métrique identité $M = Id_p$ sur le tableau centré réduit

$$Z = YM_{1/s}.$$

L'ACP "usuelle" revient donc à centrer et réduire les observations puis à utiliser la métrique identité : c'est l'**ACP normée**.

Matrice de variance

Les n valeurs relevées pour chaque variable peuvent être vues comme n réalisations indépendantes d'une v.a. :

Lien statistique et géométrie : Pour des variables centrées et $D = Id_n$, on a $\langle y^j, y^{j'} \rangle_F = (y^j)^y y^{j'} = \sum_{i=1}^n y_i^j y_i^{j'}$. Ainsi :

$$\|y^j\|_F \rightsquigarrow \qquad \text{et} \qquad \langle y^j, y^{j'} \rangle_F \rightsquigarrow$$

Matrice de variance : c'est la matrice $p \times p$ notée V dont l'entrée j, j' contient $cov(x^j, x^{j'})$. On a

$$V =$$

Exemple : **Matrice de corrélation** (matrice de variance des données centrée réduites Z et pour $D = Id_n$) pour l'exemple des cidres :

	odeur	sucre	acide	amer	astringence	suffocante	piquante	alcool	parfum	fruitée
odeur	1,00	0,08	-0,16	0,49	0,04	**0,84**	-0,61	0,03	-0,29	0,18
sucre	0,08	1,00	-0,29	-0,60	**-0,77**	-0,19	-0,61	**-0,92**	**0,87**	**0,95**
acide	-0,16	-0,29	1,00	-0,08	0,34	0,14	0,14	0,15	-0,40	-0,27
amer	0,49	-0,60	-0,08	1,00	**0,71**	0,38	-0,03	**0,70**	**-0,63**	-0,50
astringence	0,04	**-0,77**	0,34	**0,71**	1,00	0,07	0,14	**0,86**	**-0,66**	**-0,64**
suffocante	**0,84**	-0,19	0,14	0,38	0,07	1,00	-0,23	0,22	-0,50	-0,10
piquante	-0,61	-0,61	0,14	-0,03	0,14	-0,23	1,00	0,48	-0,33	**-0,73**
alcool	0,03	**-0,92**	0,15	**0,70**	**0,86**	0,22	0,48	1,00	**-0,76**	**-0,83**
parfum	-0,29	**0,87**	-0,40	**-0,63**	**-0,66**	-0,50	-0,33	**-0,76**	1,00	**0,80**
fruitée	0,18	**0,95**	-0,27	-0,50	**-0,64**	-0,10	**-0,73**	**-0,83**	**0,80**	1,00

Inertie

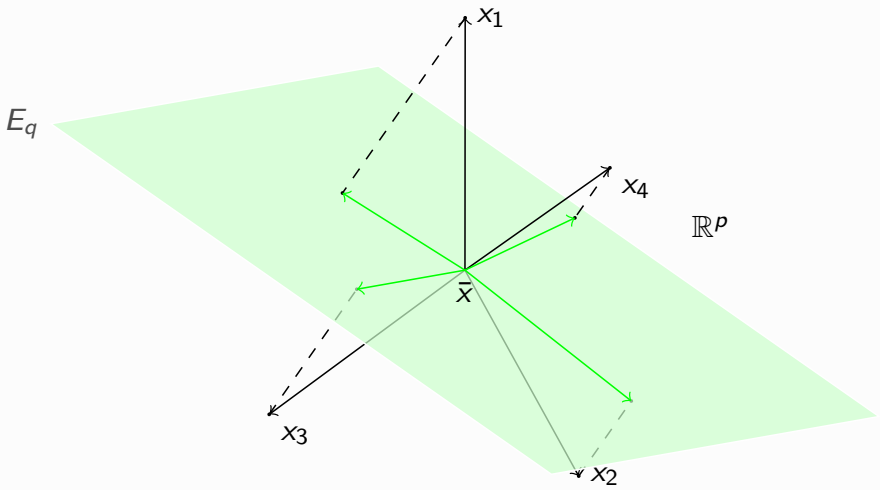
L'inertie généralise la notion de variance des variables réelles aux variables dans $\mathbb{R}^p$.

Inertie totale : du nuage de points décrit par X est

$$\mathcal{I}(X) = \frac{1}{n}\sum_{i=1}^n \|x_i - \bar{x}\|_M^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^t M (x_i - \bar{x})$$

On a $\mathcal{I}(X) = \mathcal{I}(Y)$ où Y est le nuage centrée.

Inertie dans un sous espace E_q : inertie du nuage projeté orthogonalement sur E_q :



Espaces d'inertie maximale

Les sev d'inertie maximale sont les sous espaces propres de la matrice $VM = Y\,{}^tDYM$.

Inertie des axes propre : On calcule les valeurs/vecteurs propres de la matrice $VM \in \mathbb{R}^{p\times p}$ et on ordonne les vecteurs propres par ordre décroissant de valeur propre :

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$
$$v_1 \perp v_2 \perp \cdots \perp v_p$$

Pour tout $k = 1, \cdots, p$, l'inertie de l'axe engendré par $v_k \in \mathbb{R}^p$ est $\lambda_k \geq 0$ la part d'inertie expliqué par cet axe est $\frac{\lambda_k}{\lambda_1+\cdots+\lambda_p}$.

Sous espaces d'inertie maximale : On note E_q "le" sev d'inertie maximum et de dimension $q = 1, \cdots, p$. Les sev E_q sont emboîtées :

$$E_1 = Vect\{v_1\} \subset E_2 = Vect\{v_1, v_2\} \subset \cdots \subset E_p = Vect\{v_1, \cdots, v_p\} = E$$

Qualité de la représentation : La part de dispersion expliquée par le sous-espace E_q est

$$\frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_p}.$$

Exemple des cidres :

	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8	λ_9	λ_{10}
valeur propre	5,154	2,502	1,097	0,834	0,194	0,14	0,049	0,024	0,006	0
inertie expliquée (%)	51,54	25,02	10,97	8,34	1,94	1,40	0,49	0,24	0,06	0
inertie expl. cumulée (%)	51,54	76,56	87,53	95,87	97,81	99,21	99,70	99,94	100	100

Vocabulaire

Espaces factoriels directs : ce sont les p sev $Vect\{v_k\}$ engendrés par les $v_k \in \mathbb{R}^p$ (directions de plus grande inertie).

Espaces factoriels duals : ce sont les sev engendrés par les composantes principales. La k-ème composante principale est le vecteur $c^k \in \mathbb{R}^n$ des coordonnées des n individus sur l'axe factoriel $Vect\{v_k\}$. En ACP normé on a :

La composante principale c^k est donc une combinaison linéaire des variables x^j et peut être vue comme une nouvelle variable dans le sous-espace factoriel dual.

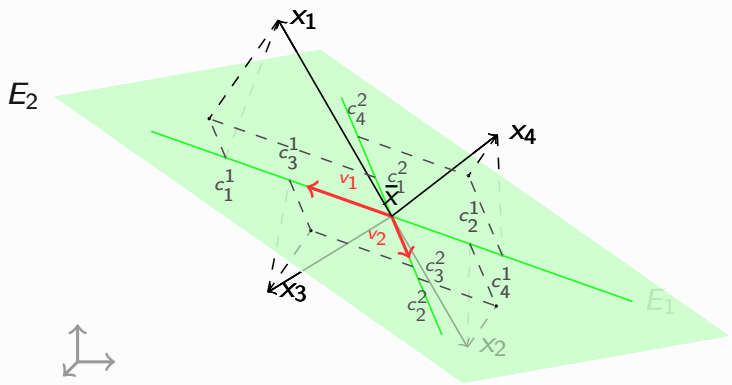


FIGURE – Espaces factoriels et composantes principales ($n = 4$ et $p = 3$)

Qualité de la représentation

Individus : La qualité de représentation d'un individu i sur un axe factoriel direct est mesurée par le cosinus carré de l'angle entre l'axe factoriel et le vecteur z_i . Plus le cosinus est grand, plus z_i sera proche de l'axe factoriel et donc sera bien représenté sur cet axe. Pour connaître la qualité de la représentation sur un espace E_q on ajoute simplement les q cosinus carrés.

Exemple : tableau des $\cos^2$ des angles entre les cidres x_i et les trois premiers axes factoriels engendrés par v_1, v_2, v_3 :

	Dim.1	Dim.2	Dim.3
1	0.0470	0.5799	0.3250
2	0.5664	0.2450	0.0075
3	0.4919	0.1207	0.0358
4	0.7755	0.0029	0.1801
5	0.7808	0.0619	0.0001
6	0.8235	0.0001	0.0567
7	0.6969	0.0035	0.0162
8	0.6392	0.0748	0.2211
9	0.0046	0.9185	0.0565
10	0.4235	0.0775	0.1599

Tous les individus sont correctement représentés dans le premier plan factoriel sauf peut être le 10 (il est "représenté" à $QLT_{1,2}(x_{10}) =$)

Qualité de la représentation

Variable : La qualité de représentation d'une variable j sur le k-ème axe factoriel dual est exprimée par le coefficient de corrélation linéaire $r(c^k, z^j)$ entre la variable initiale z^j et la nouvelle variable c^k. La valeur de cette corrélation sera également très importante pour interpréter ces nouveaux axes factoriels en fonction des variables initiales.

Exemple : tableau des corrélations entre les variables x^j et les composantes principales c^1, c^2, c^3 :

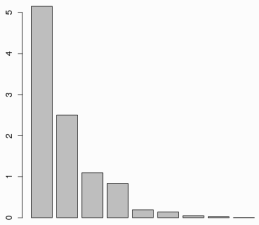
	Dim.1	Dim.2	Dim.3
odeur	0.082	0.984	0.005
sucre	-0.973	0.162	0.022
acide	0.326	-0.153	0.874
amer	0.716	0.465	-0.381
astringence	0.833	0.033	-0.089
suffocante	0.307	0.790	0.309
piquante	0.489	-0.717	-0.009
alcool	0.943	-0.039	-0.192
parfum	-0.906	-0.196	-0.211
fruitée	-0.914	0.293	0.017

Toutes les variables (sauf) sont bien représentées dans le premier plan factoriel

Remarque : bien qu'ayant retenu un sous-espace factoriel expliquant une part importante de l'inertie totale, il est possible que certaines variables ou individus d'intérêt soient mal représentés dans ce sous-espace. Il sera alors intéressant de compléter le sous-espace factoriel en ajoutant des axes factoriels supplémentaires de sorte que ces variables ou individus d'intérêt soient bien représentés

Choisir le nombre de dimension :

Exemple des cidres : éboulis des valeurs propres.

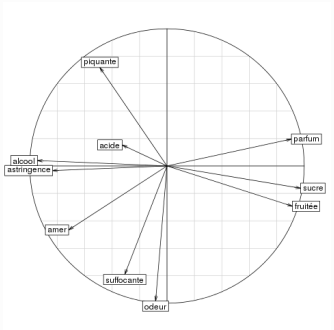


Méthodes empiriques : le choix du nombre de dimension se fait sur des critères empiriques. Cela dépend aussi de la finalité de l'ACP : pour de la description de données on a souvent $q \leq 4$ (au delà difficultés d'interprétation), pour de la compression q peut être beaucoup plus grand.

Exemple de méthodes : pour l'analyse descriptive :

- Méthode de Kaiser : Ne retenir que les directions principales qui expliquent une proportion de l'inertie supérieure à $\frac{1}{p}$. Dans l'exemple : cela correspond à $q = 3$. Mais mieux vaut couper avant/apres un décrochement...
- Méthode du coude : sur "l'éboulis" des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière : sélectionner les axes avant le décrochement.
- Méthode pratique : ne retenir que les dimensions que l'on peut interpréter...

Interprétation des composantes principales



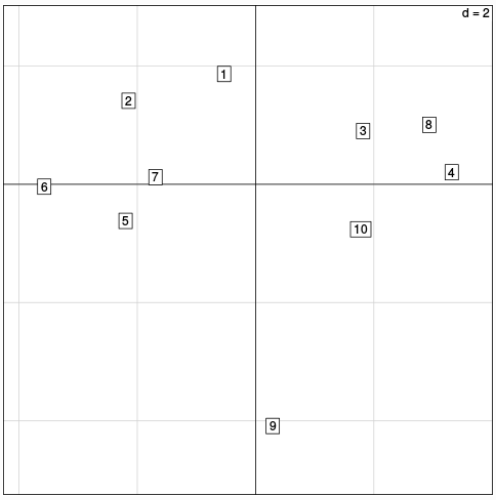
Interprétation des composantes principales : On mesure la relation entre c^k et la variable z^j l'aide de la corrélation $r(c^k, z^j)$. Étant donnés c^1 et c^2, on trace le "cercle de corrélations" : chaque variable z^j est représentée par un point de coordonnées $(r(c^1, z^j), r(c^2, z^j))$ dans le plan. Les variables les mieux expliquées correspondent aux points les plus proches du cercle unité.

Toutes les variables (sauf acidité) sont bien représentées dans ce plan factoriel puisque leurs corrélations avec les axes sont relativement importantes (les projections sont proches du cercle de corrélation). Interprétation possible des deux premiers axes factoriels :

- le premier axe factoriel semble opposer le cidre doux (fruité, sucré, parfumé) au cidre brut (plus alcoolisé et astringent).
- le second axe factoriel semble opposer les cidres ayant une particularité olfactive (forte odeur) aux cidres ayant une certaine particularité gustative (piquante).

Exemple : représentation des données

Les coordonnées des individus dans le premier plan factoriel (composantes principales) :

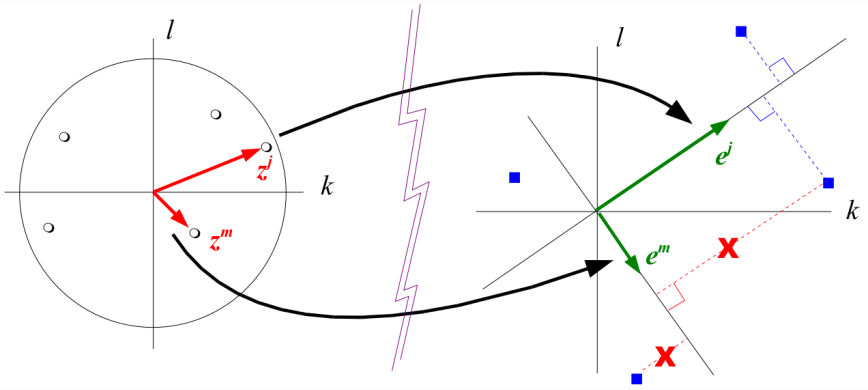


Les valeurs :

i	1	2	3	4	5	6	7	8	9	10
c_i^1	-0.53	-2.15	1.82	3.32	-2.20	-3.57	-1.69	2.94	0.29	1.78
c_i^2	1.87	1.41	0.90	0.20	-0.62	-0.04	0.12	1.01	-4.09	-0.76

Biplot

Objectif : On superpose les deux représentations duales (les variables et les individus) sur un même graphique pour aider à l'interprétation.



Attention : n'est pertinent que pour les variables et les individus bien représentés !