Cours de statistiques
pour l'agrégation de sciences économiques et sociales

Simon Coste, Pierre Montagnon
avec quelques pages inspirées des cours de Hugo Harari-Kermadec

2020-2021

Table des matières

1
Programme de l'épreuve et commentaires

La partie statistique du programme de mathématiques porte uniquement sur des notions classiques que l'on peut attendre d'un étudiant en licence de sociologie ou d'économie : statistiques descriptives, estimation, tests d'hypothèses, tests du chi-deux, régression linéaire, ACP. L'accent est mis sur la compréhension des concepts et l'utilisation honnête des outils statistiques. Les rapports des jurys insistent presque toujours sur la nécessité absolue de bien savoir utiliser, calculer et interpréter les outils qui permettent de décrire des données, et particulièrement les paramètres simples (moyennes, écarts-types, médianes, r-deux, statistiques du chi-deux). D'autre part, il est indispensable de savoir calculer ces paramètres à l'aide d'une calculatrice, et de savoir lire les sorties des principaux logiciels de statistiques.

Dans ce cours, nous avons autant que possible donné quelques exemples de sorties de logiciels (langage R). Il n'est pas demandé de savoir programmer soi-même. En revanche, il est de votre responsabilité de vous assurer que vous savez **utiliser votre calculatrice** pour savoir rapidement effectuer quelques taches statistiques élémentaires. Au vu des quelques exercices que l'on voit dans les rapports des jurys, ces taches sont : le calcul d'une moyenne, d'une médiane, d'un écart-type, de quantiles ; l'accès à des tables statistiques élémentaires (disons, les tables d'une loi normale, d'une chi-deux et d'une Fisher) ; le calcul des coefficients d'une régression linéaire simple, et éventuellement le calcul du r-deux associé.

Nous avons rassemblé ci-dessous tous les points du programme officiel avec l'endroit où ils sont traités. Il y a aussi un index des termes à la fin du document.

Statistique inférentielle

Estimateurs, section 3 à partir de la page 12.
— Propriétés : biais, risque quadratique, convergence, sous-section 3.2 page 12.
— Estimation ponctuelle (section 3 page 12) ou par intervalle de confiance (section 4 page 16).

Tests d'hypothèses, section 5 à partir de la page 23.
— risques d'erreur, sous-section 5.1 page 23.
— région critique, sous-section 5.4 page 25.

Application aux tests du Chi-deux, section 6 à partir de la page 27.
— Ajustement à une loi, page 27
— Liaison de 2 variables qualitatives, test d'indépendance page 29.
— Test de la moyenne.
— Test d'égalité des moyennes par analyse de la variance (ANOVA), sous-section 6.6 page 30.

Modèle linéaire (cas de la régression linéaire simple ou multiple) : section 7 page 33.
— estimateur des moindres carrés, théorème 7.1 page 35.
— test de Student de signification des coefficients de régression, sous-section 8.4 page 40.

Lecture de sorties de logiciels dans le cas de traitements informatiques de données, sous-section 8.4 page 40 pour les tests de Student en régression, section 6.6 pour l'analyse de la variance, section 9.1 pour l'analyse en composantes principales.

Interprétation des résultats d'une analyse statistique unidimensionnelle ou multidimensionnelle de données socio-économiques, sous-sections 8.4 et 8.5.

Statistique descriptive univariée

Présentation de données statistiques : tableaux à simple entrée. Diagrammes en bâtons, histogrammes (avec classes de même amplitude ou non). Diagrammes circulaires, en barre, box-plots ou « boîte à moustaches ».

Polygones des effectifs ou des fréquences cumulés, page 9.

Paramètres de position : moyenne, mode, médiane, quartiles, déciles... Section 2.3 page 7.

Paramètres de dispersion :
— étendue
— écart interquartile, page 8
— variance, écart-type, coefficient de variation, page 5.

Paramètres de concentration.
— Courbe de Lorenz, section 2.6 page 10.
— Médiale, page 8.
— indice de Gini (défini uniquement comme rapport de deux aires), définition 2.18 page 11.

Indices simples et synthétiques : définitions et propriétés (Laspeyres, Paasche, Fisher), sous-section 2.5 page 9.

Statistique descriptive multivariée

Présentation de données statistiques : tableaux à double entrée, distributions conjointe, marginales et conditionnelles.

Formules de décomposition de la moyenne et de la variance, variances inter et intra : proposition 6.13 page 31, et toute la section sur ANOVA.

Ajustement affine, section 7 page 33.
— principe de la méthode de Mayer, page 33.
— de la méthode des moindres carrés, page 34.
— Coefficient de corrélation linéaire, page 36.

Cas de p variables quantitatives (régression multiple), sous-section 7.4 page 36.
— matrice de covariance, de corrélation linéaire : définition 7.3 page 37.
— Analyse en composantes principales, section 9.1 à partir de la page 43.
— régression linéaire multiple, théorème 8.6 page 40 et discussions un peu avant.

Analyse des correspondances simples dans le cas de 2 variables qualitatives, page 49.

2
Statistiques descriptives élémentaires

Dans tous les rapports récents, les jurys insistent sur deux choses : 1) les paramètres de description élémentaires des données statistiques doivent être parfaitement maîtrisés, et 2) **il faut savoir les calculer à l'aide d'une calculatrice**. Il est hors de question de vous amuser à calculer la moyenne empirique d'un échantillon de 15 données à la main, et encore moins l'écart-type.

Dans cette section, il n'y a pas de probabilités. Il n'y a que des outils mathématiques qui permettent de *décrire* des données connues et accessibles : on ne dit rien sur le modèle statistique qui aurait pu générer ces données. C'est pour cela qu'on parle de *statistiques descriptives*, par opposition aux *statistiques inférentielles* qui ont pour objectif de rechercher, à partir des données, les processus qui auraient pu les générer.

2.1 — Moyennes, écarts-types et coefficient de variation

Si $x_1, \dots, x_n$ sont des données statistiques numériques, leur moyenne empirique est

$$\mu = \frac{x_1 + \cdots + x_n}{n}$$

et leur variance empirique (naïve [1]) est

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Exercice 2.1. Vérifier que la variance empirique peut aussi se calculer via la formule

$$\sigma^2 = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right) - \mu^2.$$

L'écart-type empirique est σ, la racine carrée de la variance empirique. Elle est plus utile que la variance, car elle a la même dimension que les données : si les données sont en mètres, la moyenne est en mètres mais la variance est en « mètres au carré ».

DÉFINITION 2.2. — *Le **coefficient de variation** d'un échantillon statistique est $\rho = \sigma/\mu$.*

Ce coefficient s'interprète comme un écart-type relatif et en soi, c'est une excellente mesure de dispersion des données. L'intérêt du coefficient de variation est qu'il permet de mieux jauger la dispersion de certaines données. Par exemple, supposons qu'une série statistique ait un écart-type de 10. En soi, cela ne dit absolument rien sur la dispersion de ces données. Si les données ont tendance à être très grandes (disons que leur moyenne est 10000) alors une variation de ± 10 représente un écart de seulement $\rho = 0,1\%$ de la valeur moyenne. En revanche, si les données sont d'amplitude modérée (disons que la moyenne est 100), alors un écart type de 10 représente une variation de $\rho = 10\%$ par rapport à la moyenne.

Attention, le principal inconvénient du coefficient de variation est qu'il n'a plus d'intérêt lorsque la moyenne est proche de zéro ; il n'est même pas défini pour des variables centrées. Ce qu'il faut retenir, c'est qu'un coefficient de variation très petit signifie que l'écart-type est beaucoup plus petit que l'écart-type, donc les données sont très concentrées. Lorsque le coefficient de variation est grand, cela peut avoir deux interprétations : soit les données sont en moyenne très proches de zéro, et dans ce cas ρ ne dit rien, soit ce n'est pas le cas et dans ce cas les données ne sont pas très concentrées.

Exercice 2.3. Les notes de Jean-Patrick en latin sont les suivantes : 11, 11, 11, 11. Celles de Marie-Ursuline sont : 10, 10, 12, 12. Celles de Charles-Jason sont 15, 15, 1, 15. Quant à celles de Timéo-Kylian, elles sont de 17, 17, 19, 19. Calculer les coefficients de variation de chacun.

1. On reviendra plus tard sur un meilleur estimateur de la variance des données lorsqu'elles proviennent de réalisations de variables aléatoires.

2.2 — Histogrammes

Un histogramme permet de représenter et de visualiser d'un coup d'oeil la répartition de données numériques. Pour construire un histogramme, on identifie d'abord des classes (les « boîtes ») ; on compte le nombre d'observations qui appartiennent à chaque classe (l'*effectif*), puis on trace un rectangle dont la hauteur [2] est proportionnelle à l'effectif.

Mathématiquement, on choisit les bords des intervalles qui vont représenter nos classes, disons

$$b_1 < b_2 < \cdots < b_h$$

où h est le nombre de classes. La classe i sera composée de toutes les observations entre b_i et b_{i+1}. S'il y a n_i observations dans cette classe, on dessine le rectangle de base $[b_i, b_{i+1}]$ et de hauteur n_i. L'*amplitude* d'une classe est la longueur de sa base, c'est-à-dire $b_{i+1} - b_i$.

REMARQUE 2.4. La question du choix des classes est cruciale. En particulier, insistons sur un point : il est possible de choisir des classes qui n'ont pas la même amplitude. Ce sera le cas par exemple lorsque les données comportent des valeurs aberrantes : supposons que dans un jeu de données, 1000 valeurs sont entre -1 et 1, mais quelques-unes (disons, 3 ou 4) sont très éloignées de cet intervalle. Il peut être intéressant de diviser l'intervalle $[-1,1]$ en classes de même amplitude pour bien visualiser la répartition des données dans cet intervalle, et d'ajouter deux classes $]-\infty,-1[$ et $[1,+\infty[$ qui regrouperont les données aberrantes.

Un même jeu de données peut avoir une allure bien différente selon le choix des classes, comme en témoigne la figure 1.

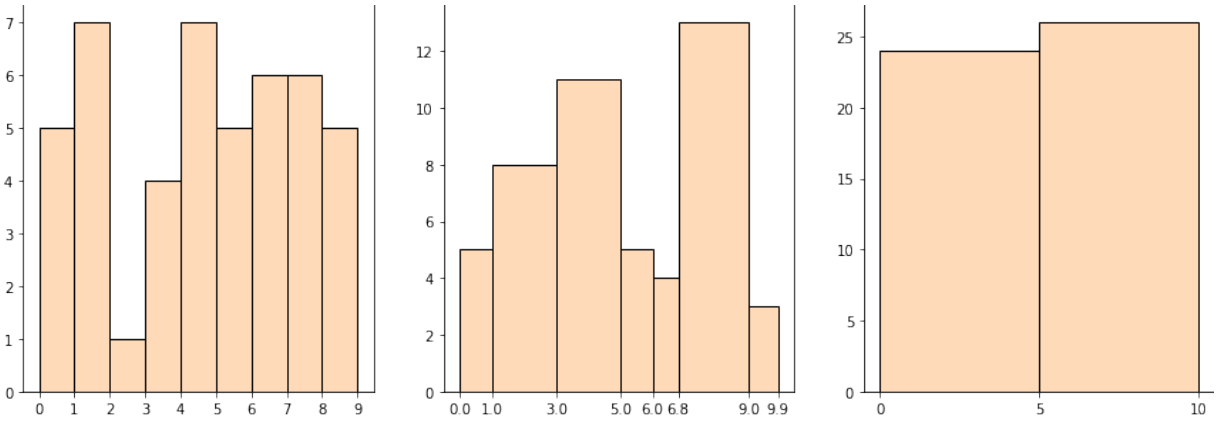


FIGURE 1 – Ces trois histogrammes représentent les mêmes 50 données visibles ci-dessous... mais évidemment, les boîtes ne sont pas les mêmes. es coordonnées des boîtes sont sur le graphique. Pour le premier histogramme on a utilisé des boîtes de même amplitude, c'est-à-dire qu'on a regroupé les données qui étaient entre 0 et 1, entre 1 et 2, etc. Il est par exemple possible de voir qu'il y a 7 observations dans l'intervalle [4,5] alors qu'il n'y en a qu'une dans l'intervalle [2,3], et vous pouvez vérifier cela dans le tableau ci-dessous. Attention, ici les histogrammes ne sont pas normalisés, c'est-à-dire que leur hauteur est égale à leur effectif.

```
7.73758879 8.55486039 5.57317639 7.40163887 8.90398262 0.49244671
1.3229121  1.88289874 5.53974594 5.50729498 1.28480919 7.52289461
6.89911707 7.53623245 3.59944876 0.04543076 8.86375818 5.67736335
4.76600856 2.53338482 4.52923367 4.48551095 3.89973424 1.45357723
1.52385813 9.82116848 4.79383897 4.56044908 9.53507463 6.45979969
3.06989745 6.22678533 4.44466447 7.88809468 0.34304044 6.92446781
8.44946871 1.42276336 3.96894957 6.03034629 9.98164926 1.81290755
5.1177333  6.13538887 7.60008069 0.61181801 4.85440412 0.16406851
8.12691921 9.60018777
```

On dit qu'un histogramme d'un jeu de n données est *normalisé* lorsque les hauteurs de tous les rectangles s'additionnent à 1. Autrement dit, la hauteur du rectangle qui représente la classe i n'est pas n_i, mais n_i/n. Lorsqu'un histogramme est normalisé, il donne une bonne approximation de la densité empirique des données. Si les

2. Ce n'est pas la seule façon de faire. On aurait pu demander à ce que l'aire soit proportionnelle à l'effectif.

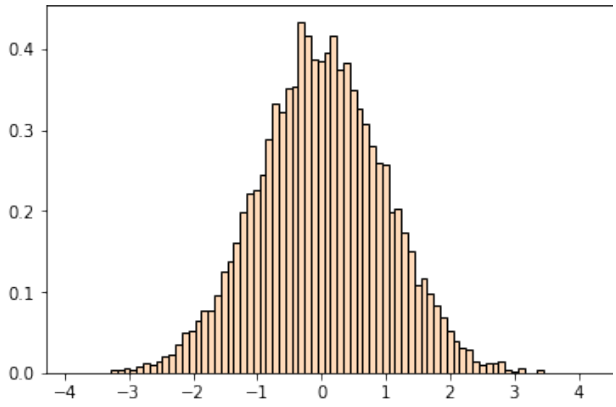


FIGURE 2 – Histogramme de 10000 réalisations d'une loi normale centrée réduite. Cette fois, l'histogramme est normalisé (cf l'axe des ordonnées).

données sont les réalisations de variables iid, cela se voit très bien. Par exemple, dans l'histogramme suivant, on a représenté un histogramme de 10000 réalisations d'une variable gaussienne centrée réduite.

REMARQUE 2.5 (histogramme et convergence)**.** Supposons que les données x_i sont des réalisations de variables aléatoires indépendantes de même loi X_i. La loi des grands nombres, lorsqu'elle s'applique, dit que la proportion d'observations entre les valeurs a et b converge vers la probabilité que la variable aléatoire sous-jacente soit entre a et b. Autrement dit, si l'histogramme est normalisé, la hauteur de la classe correspondant à l'intervalle $[a,b[$ est asymptotiquement très proche de $\mathbf{P}(\mathrm{X} \in [a,b[)$. Ceci est valable pour une classe fixée, et donc aussi pour un nombre fini de classes, mais pas nécessairement pour un quantité « infinie » de classes. En réalité, c'est la convergence en loi qui correspond à ce type de convergence. Il est même possible de voir que la convergence en loi d'une suite de variables aléatoires est en fait équivalence à la « convergence de ses histogrammes ».

2.3 — Quantiles, médiane, médiale

Les quantiles d'un échantillon de données numériques x_i sont des nombres qui divisent l'échantillon en parties de même taille. Par exemple, la médiane est un nombre q tel qu'il y a autant d'observations inférieures à q que d'observations supérieures à q. Mathématiquement, la définition est la suivante.

DÉFINITION 2.6. — *Soient $x_1, \dots, x_n$ des nombres réels. Des quantiles d'ordre m de ces observations sont des nombres $q_{m,1}, \dots, q_{m,m}$, avec $q_{0,m}$ éventuellement égal à $-\infty$ et $q_{m,m}$ éventuellement égal à $+\infty$, tels que le nombre d'observations x_i dans chaque intervalle $\mathrm{I}_k = [q_{k,m}, q_{k+1,m}[$ est toujours le même, c'est-à-dire n/m.*

Les quantiles d'ordre 3 portent le nom de terciles, ceux d'ordre 4 quartiles, ceux d'ordre 5 quintiles, ceux d'ordre 6 sexiles, ceux d'ordre 7 septiles, ceux d'ordre 8 octiles, deux d'ordre 9 noniles, ceux d'ordre 10 déciles, ceux d'ordre 11 ondéciles, ceux d'ordre douze duodéciles, ceux d'ordre 13 tridéciles, deux d'ordre 100 centiles, ceux d'ordre 1000 millésimiles, mais ce sont surtout les mots *quartiles*, *quintiles*, *déciles* et *centiles* qui sont utilisés — et évidemment, les quantiles d'ordre 2 sont très utilisés et s'appellent médiane.

Les quantiles définis ainsi ne sont pas forcément uniques. Par exemple, dans l'échantillon donné par $0,0,0,10,10,10$, le nombre 5 est une médiane, mais le nombre 6 également, ainsi que n'importe quel nombre entre 0 et 10 (strictement). Avec cette définition, les quantiles ne sont même pas toujours définis : par exemple, si le nombre d'observations est impair, il n'y a pas de médiane au sens de la définition ci-dessous. Il faut donc un peu adapter la définition. Dans le cas de la médiane, c'est facile : il suffit de dire que le nombre « du milieu ». En d'autres termes, si l'on ordonne nos observations de façon croissante

$$x_{i_1} \leqslant \cdots \leqslant x_{i_n}$$

alors la médiane est n'importe quel nombre entre $x_{i_{n/2}}$ et $x_{i_{1+n/2}}$ si n est pair, et sinon on convient de dire que c'est le nombre $i_{\lceil n/2 \rceil}$.

Par ailleurs, les quantiles les plus utilisés sont ceux d'ordre 4. On note souvent Q_1 le premier et Q_3 le dernier. Celui du milieu est une médiane (pourquoi ?).

Exercice 2.7. Trouver une médiane pour les données suivantes :

$$84, 58, 7, 12, 63, 99, 588, 6, 65, 23, 74, 25, 5, 56545, 6, \pi, 55.$$

Calculer la moyenne.

Exercice 2.8. La moyenne et la médiane ne sont pas forcément égales. Comment interpréter le fait qu'une moyenne soit, par exemple, inférieure à la médiane ?

Exercice 2.9. Dans la bibliothèque de l'agrégation, on a compté les livres en fonction de leur nombre de pages :

pages	[0, 100[	[100, 250[	[250, 400]	[400, ∞[
nombre de livres	19	29	48	12

Quelles sont les médianes possibles ? Peut-on réellement identifier une médiane ?

Il existe aussi des façons rigoureuses de définir les quantiles d'ordre m même si n n'est pas divisible par m. Nous n'en parlerons pas [3]. En revanche, il peut être utile d'avoir en tête la définition des quantiles pour variables aléatoires. Soit donc Z une variable aléatoire.

DÉFINITION 2.10. — *Si m est un entier naturel, les quantiles d'ordre m pour la loi de Z sont $m+1$ nombres $q_{m,1}, \dots, q_{m,m}$, avec $q_{0,m}$ éventuellement égal à $-\infty$ et $q_{m,m}$ éventuellement égal à $+\infty$, tels que pour chaque $i \in \{1, \dots, m\}$,*

$$\mathbf{P}(q_{i-1,m} < \mathrm{Z} \leqslant q_{i,m}) = \frac{1}{m}. \tag{1}$$

La médiane est n'importe quel nombre q tel que $\mathbf{P}(\mathrm{Z} > q) = \mathbf{P}(\mathrm{Z} < q) = 1/2$.

Exercice 2.11. Les quantiles ne sont pas uniques en général. Montrer toutefois qu'ils le sont lorsque Z possède une densité continue sur $\mathbb{R}$ et strictement positive.

Exercice 2.12. Montrer que $\mathbf{P}(\mathrm{Z} \leqslant q_{k,m}) = k/m$.

Exercice 2.13. Quelle est la médiane d'une variable aléatoire centrée ?

Dans le cas d'un jeu de données numériques $(x_1, \dots, x_n)$, les quantiles et la médiane se définissent de façon analogue : des nombres $q_{i,m}$ sont des quantiles d'ordre m pour ce jeu de données si le nombre des observations entre $q_{m,i-1}$ et $q_{m,i}$ est égal à n/m, autrement dit si la fréquence des observations entre ces deux nombres est $1/m$.

Écart interquartile. Il s'agit de l'écart entre le premier et le troisième quartile d'une distribution statistique, c'est-à-dire de la quantité $\mathrm{Q}_3 - \mathrm{Q}_1$. Il s'agit d'un indicateur de dispersion : un grand écart interquartile indique (dans une certaine mesure) une distribution dispersée. Cet indicateur est souvent présent dans les sorties des logiciels statistiques (par exemple, pour analyser les résidus dans une régression linéaire). Si l'écart interquantile est très proche de zéro, cela signifie que les valeurs sont très concentrées, puisque la moitié des données sont comprises entre les deux valeurs très proches que sont Q_1 et Q_3.

Médiale. Beaucoup moins utilisée que la médiane [4], la médiale d'une distribution est la valeur de la variable qui sépare la masse totale de cette variable en deux parties de même poids [5].

La médiale est donc la valeur qui sépare la *masse totale* de la variable en deux parties, tandis que la médiane est la valeur qui sépare les *effectifs* en deux. Cette distinction est plus digeste avec un exemple :

3. Go wikipédia

4. En fait, la médiale est strictement et rigoureusement inutilisée en dehors du champ restreint des membres du jury de l'oral de mathématiques de l'agrégation de sciences sociales (à confirmer).

5. Ce n'est pas tout à fait exact : il est possible qu'une telle valeur n'existe pas ou ne soit pas unique. Comme dans le cas de la médiane, on définit rigoureusement la médiale m' comme la *plus petite valeur* de x telle que la somme des observations x_i strictement inférieures à x soit inférieure ou égale à la somme des obervations strictement supérieures à x, c'est-à-dire que

$$m' = \inf\left\{ x \in \mathbb{R} : \sum_{x_i \leqslant x} x_i \leqslant \sum_{x_i > x} x_i \right\}.$$

	Salaire	Cumul
M. Acarien	1 200	1 200
M. Moucheron	1 220	2 420
Mlle Araignée	1 250	3 670
Mme Musaraigne	1 300	4 970
Mme Souris	1 350	6 320
M. Rat	1 450	7 770
M. Cobaye	1 450	9 220
M. Lapin	1 560	10 780
M. Chat	1 600	12 380
Mme Gazelle	1 800	14 180
M. Daim	1 900	16 080
M. Cheval	2 150	18 230
M. Élan	2 310	20 540
M. Bison	2 600	23 140
M. Rhinocéros	3 000	26 140
M. Éléphant	3 400	29 540
Mme Baleine	4 800	34 340

Sur le tableau ci-dessus, la médiane de la distribution des salaires est le salaire de M. Lapin (1560 euros). La médiale, quant à elle, est le plus petit salaire correspondant à un cumul supérieur à la moitié de la masse salariale totale, c'est-à-dire, puisque cette masse totale est de 34340 euros, le premier salaire correspondant à un cumul supérieur à 17170 euros, donc celui de M. Cheval (18230 euros).

Exercice 2.14. Montrer que la médiale d'une variable positive est toujours supérieure à sa médiane.

Exercice 2.15. Montrer que la médiale est l'abscisse du premier point d'ordonnée supérieure ou égale à $0,5$ sur la courbe de Lorenz de la distribution.

2.4 — Polygone des effectifs, ou fréquences cumulées

Supposons que l'on dispose de n observations $(x_1,\dots,x_n)$ d'une variable quantitative : salaire, taille, nombre total de cheveux, longueur de l'intestin grêle.

Ce que le programme nomme pompeusement « polygone des effectifs ou fréquences cumulées » n'est rien d'autre que la fonction de répartition des x_i, ou sa courbe représentative. Autrement dit, c'est la fonction f continue par morceaux de la façon suivante :

$$f(t) = \text{ nombre d'observations } x_i \text{ telles que } x_i < t. \tag{2}$$

La courbe représentative d'une telle fonction est une fonction « en escalier », et il y a des gens qui n'aiment pas les fonctions en escalier parce qu'elles ont des trous. En effet, supposons que les x_i sont tous distincts et notons $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ le réordonnement croissant des x_i. À chaque $x_{(i)}$, la fonction f saute de la valeur $i-1$ à la valeur i.

Pour circonvenir ce problème, on introduit le polygone des effectifs, qui est tout simplement le graphe que l'on obtient en reliant les points $(x_{(i)},i)$. Lorsque certains des $x_{(i)}$ sont égaux, c'est la même chose, sauf que les « trous » dans le graphe de f peuvent prendre les valeurs $1,2,3...$ si par exemple plusieurs x_i ont la même valeur.

Qu'il s'agisse du polygone ou du graphe de f, ce genre de graphiques permet de saisir d'un coup d'oeil la répartition d'une variable quantitative. Comme la fonction f prend des valeurs allant de 0 à n, on la divise parfois par n afin d'obtenir de vraies fréquences. Le maximum des x_i est le plus petit x tel que $f(x) = 1$; le minimum est le plus grand x tel que $f(x) = 0$.

Exercice 2.16. Comment calculer les quantiles d'un jeu de données en regardant ses fréquences cumulées ?

2.5 — Indices de Laspeyres-Paasche et Fisher

Les indices des prix de Laspeyres-Paasche décrivent l'évolution entre deux périodes du prix d'un panier de biens de référence.

On note $p_{t,i}$ le prix du bien i au temps t et $q_{t,i}$ la quantité de bien i faisant partie du panier de biens au temps t. Les indices de références du programme sont les suivants.

1. **Indice de Paasche des prix**.
$$\Delta \mathrm{P_P} = 100 \times \frac{\sum_i p_{t,i} q_{t,i}}{\sum_i p_{0,i} q_{t,i}}. \tag{3}$$
C'est la variation de prix (en pourcentage) d'un panier de biens typique de la période courante depuis la période de référence. Cet indice est insensible au changement des habitudes de consommation, c'est-à-dire aux changements de composition du panier.
2. **Indice de Paasche des quantités**.
$$\Delta \mathrm{Q_P} = 100 \times \frac{\sum_i p_{t,i} q_{t,i}}{\sum_i p_{t,i} q_{0,i}}$$
C'est la variation de prix (en pourcentage) entre un panier de biens typique de la période de référence et un panier de biens typique de la période courante, calculé par rapport aux prix actuels. Cet indice rend compte des changements des habitudes de consommation.
3. **Indice de Laspeyres des prix**.
$$\Delta \mathrm{P_L} = 100 \times \frac{\sum_i p_{t,i} q_{0,i}}{\sum_i p_{0,i} q_{0,i}}$$
C'est la variation de prix (en pourcentage) d'un panier de biens typique de la période de référence depuis cette période. Cet indice est celui utilisé dans le calcul de l'inflation en France ; il est insensible au changement des habitudes de consommation.
4. **Indice de Laspeyres des quantités**.
$$\Delta \mathrm{Q_L} = 100 \times \frac{\sum_i p_{0,i} q_{t,i}}{\sum_i p_{0,i} q_{0,i}}$$
C'est la variation de prix (en pourcentage) entre un panier de biens typique de la période de référence et un panier de biens typique de la période courante, calculé par rapport aux prix de la période de référence. Comme l'indice de Paasche des quantités $\Delta \mathrm{Q_P}$, cet indice rend compte des changements des habitudes de consommation, c'est-à-dire des changements de composition du panier.
5. **Indice de Fisher des prix**.
$$\Delta \mathrm{F_P} = \sqrt{\Delta \mathrm{P_P} \Delta \mathrm{P_L}}$$
Cet indice incorpore les indices de Paasche et de Laspeyres dont il est la moyenne géométrique[6]. Il est notamment utilisé par l'institut Statistique Canada.
6. **Indice de Fisher des quantités**.
$$\Delta \mathrm{F_Q} = \sqrt{\Delta \mathrm{Q_P} \Delta \mathrm{Q_L}}$$

Tous ces indices sont *composites*, c'est-à-dire qu'ils intègrent des informations sur plusieurs grandeurs (les quantités et les prix aux deux périodes considérées). À l'inverse, un indice *simple* est un indice qui ne prend en compte qu'une seule grandeur, par exemple l'évolution en pourcentage du prix d'un bien unitaire donné entre deux périodes.

2.6 — Courbe de Lorenz

Soient $x_1, \ldots, x_n$ des observations numériques, par exemple des revenus pécuniaires[7]. Si les revenus étaient équitablement distribués dans la population, la moitié de la population devrait recevoir à peu près la moitié de la masse totale des revenus. Ce n'est évidemment pas le cas : en règle générale, la moitié des revenus totaux est allouée à une part de la population beaucoup plus faible. La courbe de Lorenz permet de visualiser cela. Cette courbe représente la part cumulée des revenus en fonction de la part cumulée de la population.

Mathématiquement, on note $\mathrm{S} = x_1 + \cdots + x_n$ la quantité totale de la variable mesurée (dans le cas des revenus, c'est la richesse totale de la population). On ordonne par ordre croissant toutes les observations, disons

$$x_{i_1} \leqslant \cdots \leqslant x_{i_n}.$$

Les k premières observations sont $x_{i_1}, \cdots, x_{i_k}$. Elles représentent donc $k/n\%$ de la population, mais la part de variable qui lui revient est $x_{i_1} + \cdots + x_{i_k}$, ce qui représente $(x_{i_1} + \cdots + x_{i_k})/\mathrm{S}\%$ de la masse totale. La courbe de Lorenz relie les points d'abscisse k/n et d'ordonnée $(x_{i_1} + \cdots + x_{i_k})/\mathrm{S}$. Elle passe donc nécessairement par les points $(0,0)$ et $(1,1)$, et elle est toujours inférieure à la droite qui relie ces deux points (pourquoi ?). On notera souvent $\ell(t)$ la fonction de Lorenz, celle qui est représentée par la courbe de Lorenz.

6. On appelle *moyenne géométrique* de deux nombres positifs a et b la quantité $\sqrt{ab}$, tandis que la quantité $\frac{a+b}{2}$, habituellement appelée tout simplement « moyenne » de ces deux nombres, en est la *moyenne arithmétique*.

7. C'est l'exemple originellement utilisé par Max Lorenz pour étudier les inégalités de revenus aux États-Unis.

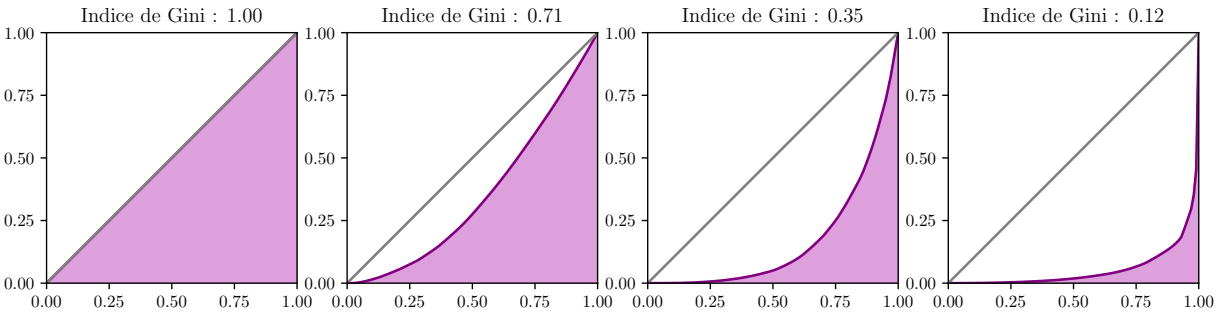


FIGURE 3 – Quelques courbes de Lorenz avec leurs indices correspondants.

Exercice 2.17. Comment trouver la courbe de Lorenz à partir du polygone des fréquences cumulées ?

Si les variables étaient équitablement réparties dans la population, la courbe de Lorenz serait précisément cette droite reliant $(0,0)$ et $(1,1)$. L'aire entre ces deux droites est donc d'autant plus grande que les données sont inégalement réparties dans la population. Il existe plusieurs façons de mesurer la différence entre ces deux aires, ce qui donne naissance aux divers indices de Gini. La définition exigée pour l'agrégation est celle du *rapport* entre ces deux aires.

DÉFINITION 2.18. — *Le coefficient ou indice de Gini est égal défini par*

$$\text{Gini} = \frac{\textit{Aire sous la courbe de Lorenz}}{\textit{Aire si les données étaient équitablement réparties}} = \frac{\int_0^1 \ell(t)\mathrm{d}t}{1/2}. \tag{4}$$

Avec cette définition, **plus l'indice est faible, plus la répartition est inégalitaire, et plus l'indice est proche de 1, plus la répartition est égalitaire**.

2.7 — Boîtes à moustache

Ce qu'on appelle coquettement « boîte à moustache » est une représentation de données numériques qui fait clairement apparaître tous les indicateurs dont nous avons parlé plus tôt : moyenne, médiane, quantiles, écart-type. Pour des données $x_1, \ldots, x_n$, les données sont représentées sous la forme d'une boîte, dont la largeur n'a aucune importance, mais dont le segment supérieur est Q_3, le segment inférieur Q_1, et dans laquelle on a représenté la médiane et/ou la moyenne. Enfin, deux traits (les moustaches) indiquent les valeurs maximales et minimales prises par les données. Parfois, certaines valeurs extrêmes sont représentées.

L'intérêt des box-plots (leur petit nom étranger) est principalement de comparer des données entre différentes populations.

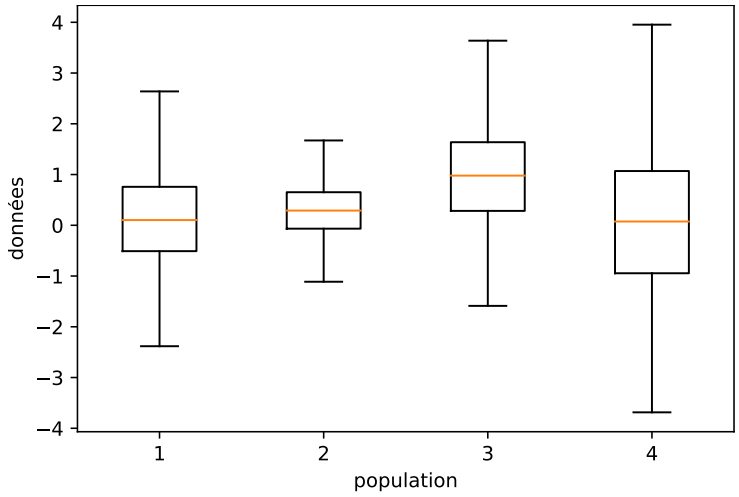


FIGURE 4 – Un exemple de boîtes à moustache. On voit d'un coup d'oeil la différence de dispersion des données entre les quatre populations.

3
Estimation ponctuelle

L'estimation ponctuelle d'un paramètre inconnu θ, comme par exemple une moyenne, consiste à calculer une valeur « probable » de θ compte tenu des observations. Le mot « ponctuel » semble renvoyer au fait que l'on ne cherche à estimer qu'un seul point ou paramètre, à savoir θ. Cela pré-suppose que le statisticien a fait une hypothèse probabiliste sur les données, une hypothèse de type « les données que j'observe sont les réalisations d'une variable aléatoire dont la loi dépend du paramètre θ ».

Un autre paradigme, beaucoup plus complexe et évidemment non-exigible dans ce cours, est l'estimation non-paramétrique qui cherche à estimer la loi de probabilité directement à partir des données, sans hypothèse supplémentaire.

3.1 — Notion d'estimateur

On suppose que l'on observe un phénomène aléatoire que l'on peut rapprocher d'une loi connue (uniforme, binomiale, normale, exponentielle...) mais de paramètre θ inconnu ; pour l'exemple, on pensera à une loi normale $\mathcal{N}(\theta,1)$. Le problème de l'estimation paramétrique revient à approcher la valeur de θ en utilisant des observations indépendantes du phénomène.

EXEMPLE 3.1 (problème du sondage)**.** Un problème d'estimation élémentaire auquel nous ferons souvent référence est le suivant : on considère une population constituée d'un grand nombre d'individus et on cherche à estimer la proportion inconnue $p \in [0,1]$ de personnes au sein de cette population prêtes à voter pour un candidat donné lors des prochaines élections. Ce problème, que nous désignerons sous le nom de *problème du sondage*, est la source de la plupart des problèmes d'estimation proposés à l'oral de l'agrégation ; il correspond à l'estimation du paramètre d'une loi de Bernoulli. D'autres exemples seront présentés en TD.

Si $n \geqslant 1$, on appelle n-échantillon tout n-uplet $(X_1,\dots,X_n)$ de variables aléatoires réelles indépendantes et identiquement distribuées.

DÉFINITION 3.2 (Estimateur)**.** — *Soit $(X_1,\dots,X_n)$ un n-échantillon. On appelle estimateur de θ toute variable aléatoire réelle Z s'écrivant en fonction de $(X_1,\dots,X_n)$, c'est-à-dire telle qu'il existe une fonction f à valeurs dans $\mathbb{R}$ telle que $Z = f(X_1,\dots,X_n)$.*

Par exemple, si $(X_1,\dots,X_n)$ est un n-échantillon, alors $\overline{X_n} = \frac{1}{n}\sum_{i=1}^n X_i$, $\min_i X_i$, $X_1 - X_n$, $\exp(X_1^2 - 1)$... sont des estimateurs de θ.

On note souvent $\hat{\theta}_n$ un estimateur du paramètre θ ; le $\hat{}$ (élégamment prononcé *chapeau*) souligne le caractère empirique de la grandeur $\hat{\theta}_n$, tandis que l'indice $_n$ signifie que $\hat{\theta}_n$ est construit à partir d'un n-échantillon.

Un estimateur de θ est donc simplement une variable aléatoire **qui dépend uniquement des X_i et non de θ** (ou plus généralement de paramètres inconnus). On cherche bien sûr à obtenir de *bons* estimateurs de θ, c'est-à-dire des estimateurs dont on a de bonnes raisons de penser qu'ils prendront en moyenne des valeurs proches de θ, mais aussi, pour des raisons de coût de collecte des données, des estimateurs dont la variabilité est faible.

3.2 — Performance d'un estimateur

DÉFINITION 3.3 (Biais d'un estimateur)**.** — *Soit $\hat{\theta}_n$ un estimateur de θ admettant une espérance.*

1. *Le biais de $\hat{\theta}_n$ (en tant qu'estimateur de θ) est la quantité $\mathbf{E}[\hat{\theta}_n] - \theta$.*
2. *On dit que $\hat{\theta}_n$ est un estimateur sans biais (de θ) lorsque $\mathbf{E}[\hat{\theta}_n] = \theta$.*
3. *On dit que $\hat{\theta}_n$ est un estimateur asymptotiquement sans biais (de θ) lorsque*

$$\lim_{n\to+\infty} \mathbf{E}[\hat{\theta}_n] - \theta = 0$$

DÉFINITION 3.4 (Convergence d'un estimateur)**.** — *Soit $\hat{\theta}_n$ un estimateur. On dit que l'estimateur $\hat{\theta}_n$ de θ est convergent (ou consistant) si et seulement si on a*

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbf{P}} \theta$$

Rappelons que cette dernière propriété signifie que la probabilité pour que $\hat{\theta}_n$ et θ soient éloignés de plus d'une quantité arbitrairement petite tend vers 0 lorsque le nombre d'observations augmente :

$$\hat{\theta}_n \xrightarrow[n\to+\infty]{\mathbf{P}} \theta \quad \Leftrightarrow \quad \forall \varepsilon > 0, \mathbf{P}\left(\left|\hat{\theta}_n - \theta\right| > \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} 0$$

Exercice 3.5. Démontrer à l'aide de l'inégalité de Bienaymé-Tchebychev que si $\hat{\theta}_n$ est un estimateur sans biais de θ admettant une variance σ_n^2 et si $\sigma_n^2 \underset{n\to+\infty}{\longrightarrow} 0$, alors $\hat{\theta}_n$ est convergent.

EXEMPLE 3.6. Dans le cadre du problème du sondage, la moyenne empirique $\overline{\mathrm{X}_n}$ est un estimateur convergent de p. Il suffit pour le voir d'utiliser l'exercice 3.5, puisque pour tout $n \geqslant 1$ on a

$$\mathrm{Var}\left(\overline{\mathrm{X}_n}\right) = \frac{\mathrm{Var}(\mathrm{X}_1)}{n}$$

DÉFINITION 3.7 (Risque quadratique)**.** — *Si l'estimateur $\hat{\theta}_n$ admet un moment d'ordre 2, on définit le risque quadratique de $\hat{\theta}_n$ (en tant qu'estimateur de θ) comme la quantité*

$$\mathrm{RQ}(\hat{\theta}_n, \theta) = \mathbf{E}\left[(\hat{\theta}_n - \theta)^2\right]$$

Dans le cadre de la définition ci-dessus, si $\hat{\theta}_n$ est sans biais alors son risque quadratique est tout simplement sa variance. En général, le risque quadratique de $\hat{\theta}_n$ décrit la dispersion de $\hat{\theta}_n$ autour de θ : il s'agit de la moyenne théorique des carrés des écarts de $\hat{\theta}_n$ par rapport à θ. L'obtention d'une estimation à la fois fine et fiable de θ passe bien sûr par la construction d'un indicateur dont le risque quadratique est faible, et c'est le risque quadratique (et non le biais !) que l'on retiendra comme principal indicateur de la qualité d'un estimateur paramétrique.

PROPOSITION 3.8 (Décomposition biais-variance du risque quadratique)**.** — *Si $\hat{\theta}_n$ admet un moment d'ordre 2, son risque quadratique est donné par*

$$\mathrm{RQ}(\hat{\theta}_n, \theta) = \left(\mathbf{E}[\hat{\theta}_n] - \theta\right)^2 + \mathrm{Var}(\hat{\theta}_n). \tag{5}$$

Pour que le risque quadratique d'un estimateur soit faible, il faut donc qu'à la fois son biais (c'est-à-dire son inadéquation structurelle avec le paramètre) et sa variance (c'est-à-dire sa dispersion statistique) soient faibles.

EXEMPLE 3.9. Dans le cadre du problème du sondage, le risque quadratique de l'estimateur de p donné par la moyenne empirique est $\mathrm{RQ}(\overline{\mathrm{X}_n}, p) = \mathrm{Var}(\overline{\mathrm{X}_n}) = \frac{p(1-p)}{n}$.

PROPOSITION 3.10 (Risque quadratique et convergence)**.** — *Si $\hat{\theta}_n$ admet un risque quadratique (relativement à θ) tendant vers 0 lorsque n tend vers l'infini, alors $\hat{\theta}_n$ est convergent.*

La proposition précédente est intéressante puisqu'elle permet d'étudier la convergence d'un estimateur en regardant simplement son risque quadratique, qui est parfois plus simple à étudier. Lorsque l'on cherche à (ou qu'un énoncé nous demande de) comparer les performances de deux estimateurs, il suffit souvent de calculer leurs risques quadratiques : on choisira systématique l'estimateur dont le risque quadratique tend vers 0 le plus rapidement.

3.3 — Un exemple à connaître : estimation sans biais d'une variance

On dispose d'une n-échantillon de variables aléatoires $(\mathrm{X}_1, \ldots, \mathrm{X}_n)$, qui ont toutes la mêmes distributions qu'une variable de référence X possédant une espérance μ et une variance σ^2 toutes deux inconnues.

L'estimation de l'espérance μ est très facile à obtenir : le lecteur sait bien que la moyenne empirique $\bar{\mathrm{X}}_n$ est un estimateur sans biais de μ. Cependant, l'estimation de la variance est légèrement plus subtile. On sait que la variance est définie par $\sigma^2 = \mathbf{E}[\mathrm{X}^2] - \mu^2$. Pour estimer σ^2, une procédure naturelle est d'estimer chacun de ces deux termes :

1. Le premier terme $\mathbf{E}[\mathrm{X}^2]$ est l'espérance de la variable aléatoire $\mathrm{Y} = \mathrm{X}^2$. On peut donc simplement utiliser la moyenne empirique des carrés des X_i pour l'estimer. Notons

$$\bar{\mathrm{Y}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{X}_i^2.$$

   Il est immédiat de vérifier que $\mathbf{E}[\bar{\mathrm{Y}}_n] = \mathbf{E}[\mathrm{X}^2]$. Jusqu'ici, tout va bien.

2. Il faut maintenant estimer μ^2. Là encore, c'est facile : puisque $\bar{X}_n$ est un estimateur de μ, on pourrait prendre $(\bar{X}_n)^2$ pour estimateur de μ^2. On se hasardera peut-être à supposer que c'est un estimateur sans biais de μ^2... C'est ici que le bât blesse : en règle générale, si Z est une variable aléatoire, il est évidemment faux [8] de dire $\mathbf{E}[Z^2] = (\mathbf{E}[Z])^2$. Le calcul de $\mathbf{E}[X_n^2]$ est assez facile [9]. On a :

$$\begin{aligned}
\mathbf{E}[(\bar{X}_n)^2] &= \mathbf{E}\left[\frac{1}{n^2}\sum_{i,j=1}^{n} X_i X_j\right] \\
&= \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{E}[X_i X_j] \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}[X_i^2] + \frac{1}{n^2}\sum_{i\neq j}\mathbf{E}[X_i X_j] \\
&= \frac{\mathbf{E}[X^2]}{n} + \frac{n(n-1)}{n}\mu^2 \\
&= \frac{\mathbf{E}[X^2]}{n} + \left(1-\frac{1}{n}\right)\mu^2.
\end{aligned}$$

Ceci n'est pas égal à μ^2.

Par conséquent, l'estimateur 'naïf' défini par

$$\bar{\sigma}_n^2 = \bar{Y}_n - (\bar{X}_n)^2$$

n'est pas un estimateur sans biais de σ^2, puisque son espérance est donnée par :

$$\begin{aligned}
\mathbb{E}[\bar{\sigma}_n^2] = \mathbf{E}[\bar{Y}_n] &= \mathbb{E}[(\bar{X}_n)^2] \\
&= \mathbf{E}[X^2] - \frac{\mathbf{E}[X^2]}{n} - \left(1-\frac{1}{n}\right)\mu^2 \\
&= \left(1-\frac{1}{n}\right)\sigma^2.
\end{aligned}$$

Le biais de $\bar{\sigma}_n^2$ est donc égal à $-\sigma^2/n$. Fort heureusement, il est très facile d'en déduire un estimateur sans biais de la variance : il suffit de multiplier des deux côtés par $1-1/n$. L'estimateur sans biais que l'on obtient est donc $\hat{\sigma}_n^2 = (1-n^{-1})\bar{\sigma}_n^2$. Le lecteur *doit* vérifier (par un calcul très facile) que cet estimateur possède l'expression suivante, un peu plus parlante et facile à retenir :

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2. \tag{6}$$

REMARQUE 3.11. Remarquons tout de même que l'estimateur $\bar{\sigma}_n^2$ est *asymptotiquement sans biais*, au sens où l'on voit tout de suite que $\lim_{n\to\infty}\mathbf{E}[\bar{\sigma}_n^2] = \sigma^2$.

REMARQUE 3.12 (Convergence)**.** On pourra également se demander si cet estimateur $\hat{\sigma}_n^2$ est un estimateur *convergent* pour σ^2. C'est le cas, et c'est aussi le cas pour l'estimateur naïf $\bar{\sigma}_n$. En effet, nous savons que la moyenne empirique d'un échantillon est toujours un estimateur convergent de la moyenne correspondant. Ainsi, $\bar{Y}_n$ est un estimateur convergent de $\mathbf{E}[X^2]$, et $\bar{X}_n$ est un estimateur convergent de μ :

$$\bar{Y}_n \xrightarrow{\mathbf{P}} \mathbf{E}[X^2] \qquad \text{et} \qquad \bar{X}_n \xrightarrow{\mathbf{P}} \mu. \tag{7}$$

Or, nous avons vu dans les rappels de cours que la convergence d'un estimateur se comporte bien vis-à-vis des opérations élémentaires [10]. Cela implique notamment que $(\bar{X}_n)^2 \xrightarrow{\mathbf{P}} \mu^2$, et que

$$\bar{\sigma}_n^2 = \bar{Y}_n - (\bar{X}_n)^2 \xrightarrow{\mathbf{P}} \sigma^2.$$

8. Quelles sont les variables pour lesquelles il y a égalité ? De façon générale, il peut être bon de rappeler si f est une fonction, il est en général faux de dire que $\mathbf{E}[f(X)] = f(\mathbf{E}[X])$.

9. Évidemment, on peut écrire astucieusement que $\mathbf{E}[(\bar{X}_n)^2] = \mathrm{Var} + \mathbf{E}[\bar{X}_n]^2$, ce qui donne directement le résultat. Mais il est bon de comprendre les étapes du calcul ci-dessus et de savoir éventuellement les reproduire.

10. Par exemple, si f est une fonction continue et si X_n converge en probabilité vers X, alors $f(X_n)$ converge en probabilité vers $f(X)$. On appelle souvent ce résultat *théorème de continuité pour la convergence en probabilité*.

Le lecteur vérifiera par lui-même la convergence de $\hat{\sigma}_n^2$. Nous aurons par la suite besoin d'estimer la convergence de l'écart-type σ. Or, comme remarqué plus haut, il est faux de dire que $\sqrt{\hat{\sigma}_n^2}$ est un estimateur sans biais de σ. Pourtant, par le théorème de continuité ci-dessus, cet estimateur est bien un estimateur convergent de l'écart-type σ.

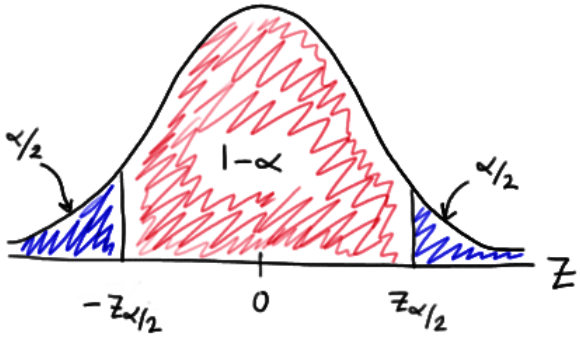


FIGURE 5 – Schéma à reproduire en toute circonstance lors de la manipulation des quantiles d'une loi gaussienne (ou plus généralement d'une loi à densité paire). Ici $z_{\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$.

4
Estimation par intervalles de confiance

Alors que l'estimation ponctuelle d'un paramètre inconnu θ consiste à calculer une valeur « probable » $\hat{\theta}$ à partir des observations, l'estimation par intervalle de confiance consiste à formuler des énoncés de type : « il est probable que le paramètre θ se trouve dans un certain intervalle ».

4.1 — Définition

DÉFINITION 4.1 (Intervalle de confiance)**. —** *Soit $\alpha \in [0,1]$. On appelle intervalle de confiance de niveau $1-\alpha$ pour θ tout intervalle de la forme*

$$I_n = [A_n, B_n]$$

où les bornes A_n, B_n sont des statistiques du modèle[11]*, et tel que*

$$\mathbf{P}(\theta \in I_n) = 1-\alpha. \tag{8}$$

Un intervalle de confiance de niveau $1-\alpha$ pour θ est donc un intervalle aléatoire dont les bornes sont des estimateurs de θ et dans lequel on peut affirmer *a priori* avec un *niveau de confiance* $1-\alpha$ que le paramètre θ se trouvera après avoir réalisé les observations numériques.

REMARQUE 4.2. Soulignons une fois encore que les bornes d'un intervalle de confiance, A_n et B_n dans la définition ci-dessus, sont *aléatoires* et ne sont pas exprimées en fonction du paramètre θ, contrairement à un *intervalle de fluctuation* pour X_1 qui est un intervalle dont les bornes sont uniquement fonction de θ et dans lequel X_1 prend ses valeurs avec une probabilité donnée. La différence entre les deux notions est régulièrement l'objet de questions lors de l'oral d'agrégation !

Pour éviter toute confusion, il suffit de s'astreindre à parler d'*intervalle de confiance pour* θ (c'est-à-dire dans lequel se trouve θ avec une certaine probabilité) et d'*intervalle de fluctuation pour* X_1 (c'est-à-dire dans lequel se trouve X_1 avec une certaine probabilité).

En pratique, on détermine *a priori* un intervalle de confiance de niveau donné, dont les bornes, rappelons-le une dernière fois, sont aléatoires, puis on collecte un échantillon formé de n observations numériques. On sait alors que le paramètre θ à estimer a une probabilité au moins égale à $1-\alpha$ de se trouver dans l'intervalle (cette fois non aléatoire) obtenu en remplaçant les X_i dans l'expression de l'intervalle de confiance par les valeurs observées... pour peu que les observations aient bien été réalisées de façon indépendante et identiquement distribuées, et chacun sait que ce n'est pas une question anodine ! Notons que α est souvent pris égal à 5% ou 10% ; l'estimation par intervalle de confiance est alors fiable respectivement à 95% et 90%.

Passons maintenant à quelques exemples simples.

11. On rappelle qu'une statistique ne doit dépendre que des observations $(X_1, \dots, X_n)$, et pas de paramètres inconnus dont θ.

EXEMPLE 4.3. Si l'on cherche à estimer $\mu \in \mathbb{R}$ à partir d'une unique observation X_1 de loi $\mathcal{N}(\mu,1)$, on peut construire un intervalle de confiance pour μ en utilisant le fait que $Z = X_1 - \mu$ suit une loi normale centrée réduite. En effet, ce dernier point et un regard vers la figure 5 permettent d'écrire pour tout $\alpha \in]0,1[$:

$$\mathbf{P}\left(|Z| \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha$$

or l'événement $\left(|Z| \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) = \left(|X_1-\mu| \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)$ est réalisé si et seulement si X_1 et μ sont à distance au plus $\Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ l'un de l'autre, c'est-à-dire si et seulement si

$$\mu \in \left[X_1 - \Phi^{-1}\left(1-\frac{\alpha}{2}\right); X_1 + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

Nous venons donc d'établir que $\left[X_1 - \Phi^{-1}\left(1-\frac{\alpha}{2}\right); X_1 + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$ est un intervalle de confiance de niveau $1-\alpha$ pour μ.
Supposons que l'on dispose d'une unique observation numérique X_1, par exemple $X_1(\omega) = 7$. Nous ne connaissons pas μ et nous cherchons à l'estimer avec un niveau de confiance de 95% (soit $\alpha = 5\%$). On sait (ou on saura bientôt !) que $\Phi^{-1}(0.975) = 1,96$. En utilisant le résultat précédent, on montre qu'il y a 95% de chances pour que μ se situe dans l'intervalle (numérique cette fois) $[7-1.96, 7+1.96] = [5.04, 8.96]$. On s'attend donc avec un niveau de confiance égal à 95% à ce que μ se situe quelque part entre 5.04 et 8.96.

EXEMPLE 4.4 (Intervalle de confiance pour la moyenne de lois normales)**.** De manière générale, si l'on travaille non plus avec une loi $\mathcal{N}(\mu,1)$ mais avec une loi $\mathcal{N}(\mu,\sigma^2)$ (avec $\sigma > 0$) et avec $n \geqslant 1$ observations, on peut vérifier (exercice : faites-le !) qu'un intervalle de confiance de niveau $1-\alpha$ pour μ est donné par

$$\left[\overline{X_n} - \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right); \overline{X_n} + \frac{\sigma}{\sqrt{n}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$$

On remarquera que l'intervalle est d'autant plus étroit (donc précis) que le nombre d'observations n augmente, ce qui est conforme à l'intuition donnée par la loi des grands nombres.

DÉFINITION 4.5 (Intervalle de confiance asymptotique)**.** — *Si $\alpha \in [0,1]$, on appelle intervalle de confiance asymptotique pour θ de niveau asymptotique $1-\alpha$ tout intervalle aléatoire $I_n = [f_n(X_1,\ldots,X_n), g_n(X_1,\ldots,X_n)]$ (où f_n et g_n sont des fonctions à n variables et à valeurs réelles) dont les bornes dépendent de $(X_1,\ldots,X_n)$ mais pas de θ et tel qu'il existe une suite $(v_n)_{n\in\mathbb{N}*}$ de limite $1-\alpha$ telle que :*

$$\mathbf{P}(\theta \in I_n) \geqslant v_n$$

pour tout $n \in \mathbb{N}^*$.

Notez que l'on ne demande pas à $\mathbf{P}(\theta \in I_n)$ d'admettre une limite quand n tend vers l'infini, mais qu'en particulier I_n est un intervalle de confiance asymptotique de niveau asymptotique $1-\alpha$ dès lors que

$$\lim_{n\to+\infty} \mathbf{P}(\theta \in I_n) = 1-\alpha$$

REMARQUE 4.6 (positivité d'une moyenne)**.** L'exemple 4.4 possède un corollaire intéressant et très largement utilisé en sciences expérimentales : lorsque des observations indépendantes et identiquement distribuées d'une grandeur suivant une loi gaussienne donnent une moyenne empirique supérieure à environ 2 fois l'écart-type empirique [12], on peut affirmer avec 95% de certitude que la moyenne théorique de ces observations est strictement positive. La plupart du temps, cela signifie qu'un effet que l'on cherche à tester existe bel et bien. Nous reviendrons sur ce point dans l'étude de la régression linéaire et des tests d'hypothèses. Dans le cas (fréquent !) d'observations non gaussiennes, on peut utiliser le théorème central limite et les intervalles de confiance asymptotiques associés pour montrer que cette observation est toujours valable lorsque n est assez grand.

Le point de vue inverse [13], celui de la *fluctuation* (voir la remarque 4.2), implique qu'à espérance et variance donnée, la répartition des données d'un grand échantillon obéit à la loi empirique représentée dans la figure 6.

12. C'est-à-dire que l'intégralité de l'intervalle de confiance donné dans l'exemple 4.4 se trouve dans $\mathbb{R}_+^*$.
13. ... ou plutôt *dual*, compte tenu de la correspondance bijective entre intervalles de fluctuation et intervalles de confiance.

4.2 — Un exemple à connaître : intervalles de confiance pour le problème du sondage

On détaille ici les différentes constructions possibles d'un intervalle de confiance dans le cadre du problème du sondage. La lecture de cet exemple est fortement conseillée, mais il n'est pas nécessaire de savoir le reproduire en détail.

On se donne $\alpha \in]0,1[$ et on cherche à construire un intervalle de confiance de niveau $1-\alpha$ pour p.

Remarquons tout d'abord que quelle que soit la valeur de p, on a $p(1-p) \leqslant 1/4$: en effet, le maximum de la fonction polynomiale $x \mapsto x(1-x) = -x^2 + x$ sur $\mathbb{R}$ est atteint en $1/2$ et vaut $1/4$. On utilisera à plusieurs reprises cette inégalité grossière mais très pratique.

Tentons dans un premier temps d'obtenir un intervalle de confiance non asymptotique pour p : il s'agit de trouver un événement impliquant un encadrement de p par des valeurs dépendant des X_i et qui soit de probabilité au plus $1-\alpha$. L'inégalité de Bienaymé-Tchebytchev semble toute indiquée pour construire notre intervalle, mais il faut encore décider à quelle variable l'appliquer. Parmi les différentes variables d'espérance p dont on dispose, on distingue notamment X_1 et $\overline{X_n}$. Or les exemples précédents suggèrent qu'il est préférable d'utiliser toute l'information disponible dans l'échantillon pour réaliser notre estimation si l'on veut que celle-ci soit le plus précise possible ; aussi choisit-on d'appliquer l'inégalité de Bienaymé-Tchebytchev à $\overline{X_n}$, ce qui permet d'écrire pour tout $\varepsilon > 0$:

$$\mathbf{P}\left(\left|\overline{X_n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{V(\overline{X_n})}{\varepsilon^2} = \frac{p(1-p)}{n\varepsilon^2} \leqslant \frac{1}{4n\varepsilon^2}$$

où l'on a utilisé la majoration obtenue au paragraphe précédent pour se débarrasser de l'expression $p(1-p)$ qui contient le paramètre inconnu p et qui empêcherait de choisir par la suite une valeur de ε indépendante de p.

Par passage à l'événement contraire, notre inégalité devient :

$$\mathbf{P}\left(\left|\overline{X_n} - p\right| < \varepsilon\right) \geqslant 1 - \frac{1}{4n\varepsilon^2}$$

c'est-à-dire

$$\mathbf{P}\left(p \in \left]\overline{X_n} - \varepsilon ; \overline{X_n} + \varepsilon\right[\right) \geqslant 1 - \frac{1}{4n\varepsilon^2}$$

Il s'agit ensuite de choisir ε pour que le terme de droite soit égal à $1-\alpha$. Un simple calcul permet de voir que la valeur de ε qui convient est $\frac{1}{2\sqrt{n}\sqrt{\alpha}}$, et donc que l'intervalle aléatoire

$$I_n = \left]\overline{X_n} - \frac{1}{2\sqrt{n}\sqrt{\alpha}} ; \overline{X_n} + \frac{1}{2\sqrt{n}\sqrt{\alpha}}\right[$$

est bien un intervalle de confiance de niveau $1-\alpha$ pour p. On remarque que sa longueur décroît à mesure que n augmente, c'est-à-dire que le fait de disposer d'un grand nombre d'observations permet de réaliser une estimation plus précise. Elle décroît aussi lorsque α, qui représente le degré d'incertitude avec lequel on accepte de formuler la proposition « p est dans l'intervalle de confiance », augmente. Si ces deux mécanismes ne vous semblent pas intuitifs, vous n'avez pas tout à fait compris la définition d'un intervalle de confiance : rendez-vous à la définition 8.

Notez enfin que si nous avions choisi de travailler sur l'indicateur X_1 au lieu de l'indicateur $\overline{X_n}$, notre intervalle de confiance (construisez-le !) aurait été indépendant de n, et donc asymptotiquement infiniment moins précis que celui que nous avons construit.

On veut à présent construire un intervalle de confiance *asymptotique* pour p, dans l'espoir que l'intervalle obtenu soit meilleur (c'est-à-dire plus petit) que l'intervalle non asymptotique que nous venons de présenter. On utilise à nouveau la moyenne empirique des observations $\overline{X_n}$, mais cette fois c'est le théorème central limite qui nous donne notre première relation :

$$\forall \varepsilon > 0, \quad \mathbf{P}\left(-\varepsilon \leqslant \sqrt{n}\frac{\overline{X_n} - p}{\sqrt{p(1-p)}} \leqslant \varepsilon\right) \underset{n\to+\infty}{\longrightarrow} \Phi(\varepsilon) - \Phi(-\varepsilon) = 2\Phi(\varepsilon) - 1$$

Remarquez que l'on choisit d'emblée [14] de construire un intervalle de confiance asymptotique centré sur $\overline{X_n}$, ce qui explique que l'on s'intéresse à la probabilité d'un événement de la forme $(-\varepsilon \leqslant \ldots \leqslant \varepsilon)$, mais qu'il aurait

14. ... parce que l'on sait que la densité de la gaussienne est concentrée autour de sa moyenne et que centrer l'intervalle de confiance est un bon moyen de réduire sa longueur !

tout à fait été possible de considérer un événement de la forme $(a \leqslant \ldots \leqslant b)$, ou même de la forme $(a \leqslant \ldots)$ par exemple, ce dernier ayant alors permis d'obtenir un intervalle de confiance asymptotique unilatère.

La relation ci-dessus se réécrit :

$$\forall \varepsilon > 0, \mathbf{P}\left(p \in \left[\overline{X_n} - \frac{\varepsilon\sqrt{p(1-p)}}{\sqrt{n}}; \overline{X_n} + \frac{\varepsilon\sqrt{p(1-p)}}{\sqrt{n}} \right] \right) \underset{n\to+\infty}{\longrightarrow} 2\Phi(\varepsilon) - 1$$

En remarquant que

$$\left[\overline{X_n} - \frac{\varepsilon\sqrt{p(1-p)}}{\sqrt{n}}; \overline{X_n} + \frac{\varepsilon\sqrt{p(1-p)}}{\sqrt{n}} \right] \subset \left[\overline{X_n} - \frac{\varepsilon}{2\sqrt{n}}; \overline{X_n} + \frac{\varepsilon}{2\sqrt{n}} \right]$$

grâce à l'inégalité $p(1-p) \leqslant \frac{1}{4}$ démontrée plus haut, on voit alors que $\left[\overline{X_n} - \frac{\varepsilon}{2\sqrt{n}}; \overline{X_n} + \frac{\varepsilon}{2\sqrt{n}} \right]$ est un intervalle de confiance asymptotique pour p de niveau asymptotique $2\Phi(\varepsilon) - 1$.

Il suffit alors de choisir ε tel que $2\Phi(\varepsilon) - 1 = 1 - \alpha$ pour obtenir un intervalle de confiance asymptotique pour p de niveau asymptotique $1-\alpha$. Un calcul rapide donne $\varepsilon = \Phi^{-1}(1 - \frac{\alpha}{2})$, d'où l'intervalle de confiance asymptotique pour p de niveau $1-\alpha$ suivant :

$$I'_n = \left[\overline{X_n} - \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{2\sqrt{n}}; \overline{X_n} + \frac{\Phi^{-1}(1-\frac{\alpha}{2})}{2\sqrt{n}} \right]$$

Il est encore une fois possible d'observer que la longueur de I'_n est d'autant plus faible (c'est-à-dire, l'estimation d'autant plus précise) que n est grand et que la tolérance α est élevée. On peut aussi comparer la longueur de I_n et celle de I'_n : pour $\alpha = 5\%$, la longueur de I_n est approximativement égale à $\frac{10}{\sqrt{n}\sqrt{5}} \approx \frac{4{,}47}{\sqrt{n}}$ et celle de I'_n à $\frac{1{,}96}{\sqrt{n}}$. En d'autres termes, pour un même niveau de confiance $1-\alpha$, l'intervalle de confiance asymptotique I'_n est plus de deux fois plus précis que l'intervalle de confiance non asymptotique I_n ! Mais bien entendu, il est n'est valable que lorsque n est grand [15].

4.3 — Sur quelques propriétés de la loi normale

Dans de nombreux tests, on utilise la loi normale d'une façon ou d'une autre, soit que l'échantillon de base soit gaussien, soit qu'il soit iid et qu'on utilise le TCL. Il est donc important de connaître quelques quantiles classiques de cette loi.

La règle 68-95-99. Une règle très utile en pratique est la règle dite « 68-95-99 » : si X suit une loi normale d'écart-type σ et de moyenne μ, la probabilité qu'une réalisation de X se trouve à distance σ de μ est 68%. La probabilité qu'une réalisation de X se trouve à distance σ de μ est 95%. Enfin, la probabilité qu'une réalisation de X se trouve à distance 3σ de μ est 99.7%. Un simple regard sur la figure 6 rendra cette règle assez claire. Dans le langage des intervalles de fluctuation, on dira donc que si $X \sim \mathcal{N}[\mu, \sigma^2)$, alors les intervalles

$$[\mu - \sigma, \mu + \sigma] \qquad [\mu - 2\sigma, \mu + 2\sigma] \qquad [\mu - 3\sigma, \mu + 3\sigma] \tag{9}$$

sont des intervalles de fluctuations de niveaux respectifs $68\%, 95\%, 99.7\%$.

Quantiles. En outre, il est bon de connaître la valeur de quelques quantiles remarquables de la loi normale. Notons Φ la fonction de répartition d'une loi gaussienne centrée réduite, ie $\Phi(t) = \mathbf{P}(\mathcal{N}(0,1) < t)$. Trois valeurs de Φ^{-1} sont à connaître :

$$\Phi^{-1}(0{,}95) \approx 1{,}64 \qquad \Phi^{-1}(0{,}975) \approx 1{,}96 \qquad \Phi^{-1}(0.995) \approx 2.58$$

Ces valeurs apparaissent dans la construction de la plupart des intervalles de confiance classiques.

4.4 — Intervalles de confiance classiques

On trouvera ci-dessous un certain nombre d'intervalles de confiance dont il est possible d'apprendre la forme par cœur... pour peu que l'on ait une idée de leur construction, qui repose toujours sur des techniques comparables à celles que nous venons d'étudier.

15. Notons qu'en économie, n est généralement considéré comme « grand » dès que l'on dispose de plus de 30 observations.

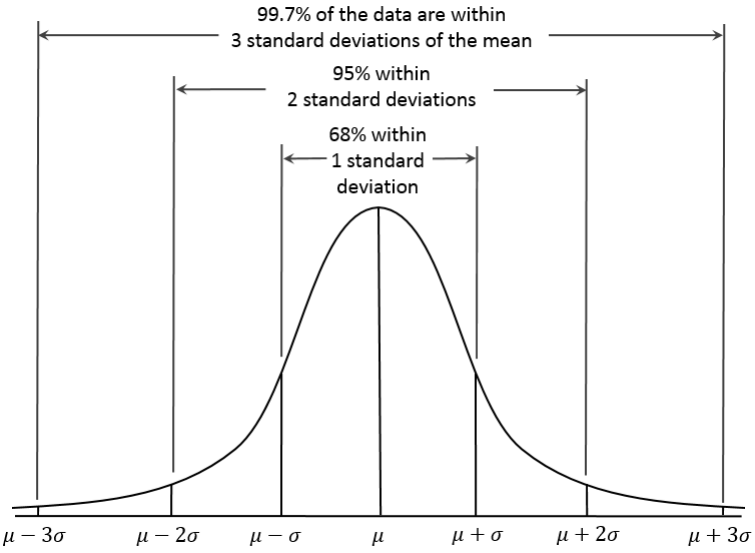


FIGURE 6 – Répartition empirique des données d'un grand échantillon de moyenne μ et de variance σ^2.

Les intervalles bilatères constituent le choix par défaut si l'on n'a pas de raison de circonscrire le paramètre dans une région particulière de $\mathbb{R}$; les intervalles unilatères, quant à eux, permettent d'obtenir des majorations ou des minorations fines. Notons que les bornes $-\infty$ et $+\infty$ dans les intervalles unilatères peuvent souvent être remplacées par des bornes adaptées au problème considéré (on remplace par exemple $-\infty$ par 0 et $+\infty$ par 1 dans le cas du problème du sondage).
On fixe $\alpha \in]0,1[$.

PROPOSITION 4.7 (Intervalle de confiance pour la moyenne, variance connue)**.** — *Si* $(X_1,\dots,X_n)$ *est un* n*-échantillon, si* $\mathbf{E}[X_1] = \mu$ *est inconnue et* $V(X_1) = \sigma^2$ *est connue, alors*

$$I_n = \left[\overline{X_n} - \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma}{\sqrt{n}}, \overline{X_n} + \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma}{\sqrt{n}}\right],$$

$$I'_n = \left]-\infty, \overline{X_n} + \frac{\Phi^{-1}(1-\alpha)\sigma}{\sqrt{n}}\right]$$

et

$$I''_n = \left[\overline{X_n} - \frac{\Phi^{-1}(1-\alpha)\sigma}{\sqrt{n}}, +\infty\right[$$

sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ *pour* μ*. Si les* X_i *sont des variables gaussiennes, ce sont des intervalles de confiance non asymptotiques (c'est-à-dire valables même lorsque* n *est petit).*

On rappelle à présent que $\hat{\sigma}_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \overline{X_n}\right)^2}$ est l'écart-type empirique des X_i. Nous avons déjà croisé cet estimateur dans la remarque 3.12, où nous avions démontré que c'est un estimateur *convergent* de σ.

En l'absence d'informations sur la variance des X_i, les intervalles donnés par la proposition ci-dessus ne sont pas des intervalles de confiance puisqu'ils dépendent du paramètre inconnu σ. Une démarche naturelle consiste donc à remplacer σ par son estimateur $\hat{\sigma}_n$; une telle substitution n'est pas anodine ! Il existe des théorèmes abstraits permettant de vérifier que cette substitution ne change pas fondamentalement les mécanismes à l'oeuvre. On pourra ainsi retenir l'adage suivant [16] :

Dans un problème d'estimation asymptotique, il est permis de remplacer n'importe quel paramètre inconnu par un estimateur convergent de ce paramètre, et cela sans changer les lois asymptotiques des objets étudiés.

16. Cet adage a seulement une valeur indicative. Une formulation rigoureuse nécessite le lemme de Slutsky, qui n'est pas au programme mais qui s'exprime de la façon suivante : si X_n converge en loi vers X et Y_n converge en probabilité vers une constante c, alors pour toute fonction continue f, la variable aléatoire $f(X_n, Y_n)$ converge en loi vers la variable aléatoire $f(X,c)$.

La proposition suivante est donnée à titre d'illustration de cet adage : on a remplacé σ par $\hat{\sigma}_n$ sans rien changer aux résultats.

PROPOSITION 4.8 (Intervalle de confiance pour la moyenne, variance inconnue)**.** — *Si* $(X_1, \dots, X_n)$ *est un n-échantillon et si* $\mathbf{E}[X_1] = \mu$ *et* $V(X_1) = \sigma^2$ *sont inconnues, alors*

$$I_n = \left[\overline{X_n} - \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_n}{\sqrt{n}}, \overline{X_n} + \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_n}{\sqrt{n}} \right],$$

$$I'_n = \left] -\infty, \overline{X_n} + \frac{\Phi^{-1}(1-\alpha)\hat{\sigma}_n}{\sqrt{n}} \right]$$

et

$$I''_n = \left[\overline{X_n} - \frac{\Phi^{-1}(1-\alpha)\hat{\sigma}_n}{\sqrt{n}}, +\infty \right[$$

sont des intervalles de confiance asymptotiques de niveau $1-\alpha$ *pour* μ*.*

On aurait donc pu améliorer notre intervalle de confiance asymptotique dans l'exemple 4.2 en faisant apparaître l'écart-type empirique [17] $\hat{\sigma}_n = \sqrt{\overline{X_n}(1-\overline{X_n})}$ dans les bornes de I'_n au lieu d'avoir recours à la majoration brutale $\sqrt{p(1-p)} \leqslant \frac{1}{2}$. En tant qu'exercice, on pourra vérifier (par un calcul direct ou grâce à l'intervalle bilatère donné ci-dessus) que l'amélioration en termes de largeur de I'_n est significative dès lors que p est éloigné de $\frac{1}{2}$.

Cependant, l'adage ci-dessus et la proposition qui l'illustrent nécessitent une mise en garde : ces résultats sont *asymptotiques*. Même dans le cas gaussien, les intervalles donnés par la proposition 4.8 ne sont pas exacts pour tout n comme dans le cas où la variance est connue, du fait de l'approximation commise σ. Le cas où n est petit et où les X_i sont gaussiennes est donc un peu particulier : il se trouve que dans ce cas précis, en remplaçant σ par $\hat{\sigma}_n$, on a modifié la loi de l'estimateur en question, et que l'on connaît exactement la nouvelle loi obtenue : il s'agit d'une loi de Student. Plus précisément (on ne demande pas aux étudiants de connaître ce raisonnement), si $(X_1, \dots, X_n)$ est un échantillon gaussien, on sait que la variable aléatoire

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$$

suit exactement une loi normale centrée réduite. Or, lorsqu'on remplace σ par $\hat{\sigma}_n$, on obtient la variable aléatoire

$$Z_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\hat{\sigma}_n}$$

et il se trouve que la loi de Z_n est une loi de Student à $n-1$ degrés de liberté. Sa fonction de répartition est souvent notée F_m, où m est le degré de liberté. Cette loi est bien connue des statisticiens, qui utilisent fréquemment ses tables.

17. La formule $\hat{\sigma}_n = \sqrt{\overline{X_n}(1-\overline{X_n})}$ est valable parce que les observations suivent une loi de Bernoulli. On peut la démontrer en établissant d'abord le résultat

$$\hat{\sigma}_n = \sqrt{\overline{X^2}_n - \overline{X}_n^2}$$

valable pour n'importe quelle loi, où l'on a noté $\overline{X^2}_n = \frac{1}{n}\sum_{i=1}^n X_i^2$, et en remarquant que dans le cas particulier d'une loi de Bernoulli on a $X_i^2 = X_i$ pour tout $i \in [\![1, n]\!]$.

PROPOSITION 4.9 (Intervalle de confiance pour la moyenne, variance inconnue (cas gaussien))**.** — *Si* $(X_1,\dots,X_n)$ *est un* n*-échantillon de loi* $\mathcal{N}(\mu,\sigma^2)$ *avec* μ *et* σ^2 *inconnus, alors en notant* F_{n-1} *la fonction de répartition de la loi de Student à* $n-1$ *degrés de liberté,*

$$I_n = \left[\overline{X_n} - \frac{F_{n-1}^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_n}{\sqrt{n}}, \overline{X_n} + \frac{F_{n-1}^{-1}\left(1-\frac{\alpha}{2}\right)\hat{\sigma}_n}{\sqrt{n}}\right],$$

$$I'_n = \left]-\infty, \overline{X_n} + \frac{F_{n-1}^{-1}(1-\alpha)\hat{\sigma}_n}{\sqrt{n}}\right]$$

et

$$I''_n = \left[\overline{X_n} - \frac{F_{n-1}^{-1}(1-\alpha)\hat{\sigma}_n}{\sqrt{n}}, +\infty\right[$$

sont des intervalles de confiance (non asymptotiques) de niveau $1-\alpha$ *pour* μ*.*

Bien sûr, si n est grand, les intervalles donnés par les propositions 4.8 et 4.9 coïncident à peu près. Cela est dû au fait que la loi de Student se rapproche de la loi normale centrée réduite lorsque son nombre de degrés de liberté augmente [18] :

$$\forall x \in \mathbb{R}, \lim_{n\to+\infty} F_{n-1}(x) = \Phi(x) \quad \text{et} \quad \forall q \in]0,1[, \lim_{n\to+\infty} F_{n-1}^{-1}(q) = \Phi^{-1}(q)$$

18. ... ce qui explique que la loi de Student à m degrés de liberté ne soit pas tabulée pour des valeurs de m supérieures à 30 !

5
Tests

L'objectif d'un test statistique est d'énoncer des affirmations sur la plausibilité d'une hypothèse de base, dite hypothèse nulle et notée H_0, contre une autre hypothèse, dite hypothèse alternative et souvent notée H_1. Par exemple, on cherche à tester l'hypothèse H_0 : « le PIB de la France est supérieur au PIB de la Suisse » contre l'hypothèse alternative : « le PIB de la France est inférieur au PIB de la Suisse ».

5.1 — Risques d'erreurs pour des tests simples

Pour commencer, on restreindra le type d'hypothèses que l'on peut faire.

Une *hypothèse simple* sur un modèle statistique est une affirmation de la forme : « la loi de probabilité sous-jacente est égale à P », où P est une certaine loi. Par exemple, si l'on dispose d'observations $(X_0, \dots, X_n)$, une hypothèse simple sur le modèle est

$$\text{Les } X_i \text{ sont des réalisations iid d'une loi normale } \mathcal{N}(0,1). \tag{H_0}$$

Sous l'hypothèse H_0, on connaît la loi de beaucoup de statistiques du modèle. Par exemple, si H_0 est vraie, la moyenne empirique $\bar{X}_n$ est une variable aléatoire de loi normale centrée et de variance $1/n$.

On peut avoir envie de tester l'hypothèse (H_0) ci-dessus contre l'hypothèse alternative suivante :

$$\text{Les } X_i \text{ sont des réalisations iid d'une loi normale } \mathcal{N}(0,4). \tag{H_1}$$

Dans le cadre simple où H_0 et H_1 sont deux hypothèses simples, un test se définit de la façon suivante.

DÉFINITION 5.1. — *Un test de niveau $\alpha \in [0,1]$ d'une hypothèse nulle H_0 contre une hypothèse alternative H_1 est une statistique Φ pouvant prendre uniquement deux valeurs :*

— *$\Phi = 0$, auquel cas on ne rejette pas l'hypothèse nulle ;*

— *$\Phi = 1$, auquel cas on rejette l'hypothèse nulle et on accepte l'hypothèse alternative H_1 ;*

de plus, si l'hypothèse nulle est vraie, la probabilité de rejeter le test doit être égale à α.

On écrit souvent cette dernière condition sous la forme $\mathbf{P}_{H_0}(\Phi = 1) = \alpha$: la notation $\mathbf{P}_H$ signifie « la loi de probabilité du modèle si l'hypothèse H est vraie ».

DÉFINITION 5.2. — *L'erreur de première espèce d'un test est la probabilité de rejeter l'hypothèse nulle, à tort : $\alpha = \mathbf{P}_{H_0}(\Phi = 1)$. Le niveau d'un test est $1 - \alpha = \mathbf{P}_{H_0}(\Phi = 0)$. L'erreur de seconde espèce d'un test est la probabilité de ne pas rejeter l'hypothèse nulle, à tort : $\beta = \mathbf{P}_{H_1}(\Phi = 0)$. La puissance d'un test est $1 - \beta = \mathbf{P}_{H_1}(\Phi = 1)$.*

5.2 — Hypothèses composites

Il arrive que les hypothèses que l'on teste ne soient pas simples : dans ce cas elles sont dites composées, et la définition des erreurs est légèrement plus compliquée, parce que sous l'hypothèse nulle H_0, le modèle peut suivre de nombreuses lois.

On dit qu'une hypothèse est composite lorsqu'elle est de la forme : « la loi de probabilité sous-jacente est égale à P_i pour une certaine loi de probabilité P_h appartenant à la famille $\{P_i : i \in I_0\}$ ». Évidemment, lorsque la famille $\{P_i : i \in I\}$ ne possède qu'un seul élément, l'hypothèse est simple.

EXEMPLE 5.3. Voici quelques exemples d'hypothèses composites :

1. Le modèle suit une loi normale. Ici, la famille associée est l'ensemble des lois $\mathcal{N}(\mu, \sigma^2)$ où $\mu \in \mathbb{R}$ et $\sigma > 0$.
2. L'espérance du modèle est nulle.
3. La loi du modèle est soit $\mathcal{N}(0,1)$, soit χ^2_3, soit $\mathrm{Poi}(5)$.
4. La variance du modèle est plus petite que 1.

Lorsque les hypothèses H_0 et H_1 sont composites, la définition des erreurs se fait « dans le pire des cas ».

DÉFINITION 5.4. — *On suppose que* H_0 *et* H_1 *sont des hypothèses composites associées à des familles de probabilités* $\{P_i : i \in I_0\}$ *et* $\{P_j : j \in I_1\}$.

L'erreur de première espèce d'un test Φ *de* H_0 *contre* H_1 *est la probabilité de rejeter l'hypothèse nulle à tort dans le pire des cas :*

$$\alpha = \max_{i \in I_0} P_i(\Phi = 1) \tag{10}$$

Le niveau est $1-\alpha$.

L'erreur de seconde espèce est la probabilité de ne pas rejeter l'hypothèse nulle, à tort, et dans le pire des cas :

$$\beta = \sup_{j \in I_1} P_j(\Phi = 0). \tag{11}$$

La puissance d'un test est $1-\beta$.

5.3 — Exemples

Dans le cas de tests sur des hypothèses simples pour un paramètre, si l'on dispose d'intervalles de confiance, il est très facile de faire des tests.

EXEMPLE 5.5 (moyenne de lois gaussiennes)**.** Prenons par exemple un n-échantillon de variables aléatoires suivant une loi $\mathcal{N}(\mu, \sigma^2)$. On va effectuer le test des hypothèses suivantes :

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

où μ_0 est une valeur fixée à l'avance, par exemple $\mu_0 = 0$. On sait que l'intervalle

$$I_n = \left[\bar{X}_n - \frac{\sigma F^{-1}(1-\alpha/2)}{\sqrt{n}} ; \bar{X}_n + \frac{\sigma F^{-1}(1-\alpha/2)}{\sqrt{n}} \right]$$

est un intervalle de confiance de niveau $1-\alpha$ pour μ_0. Ainsi, sous l'hypothèse nulle, la probabilité pour que μ_0 soit dans cet intervalle est égale à $1-\alpha$. Cela nous pousse à définir le test suivant :

$$\Phi = \mathbf{1}_{\mu_0 \text{ n'appartient pas à } I_n}.$$

Ce test est bien de niveau $1-\alpha$.

EXEMPLE 5.6 (test de Student)**.** Dans l'exemple précédent, la variance σ^2 était connue. Ce n'est pas forcément le cas. Fort heureusement, la proposition 4.9 page 22 nous donne aussi des intervalles de confiance dans ce cas, grâce à l'utilisation de la loi de Student. On a vu en effet que l'intervalle

$$J_n = \left[\bar{X}_n - \frac{\hat{\sigma}_n F_{n-1}^{-1}(1-\alpha/2)}{\sqrt{n}} ; \bar{X}_n + \frac{\hat{\sigma}_n F_{n-1}^{-1}(1-\alpha/2)}{\sqrt{n}} \right]$$

est un intervalle de confiance de niveau $1-\alpha$. Le test défini par

$$\Phi' = \mathbf{1}_{\mu_0 \text{ n'appartient pas à } J_n}$$

est donc bien un test de niveau $1-\alpha$.

Il est nécessaire de savoir utiliser tous les intervalles de confiance connus, y compris unilatères, pour construire des tests d'hypothèses variés.

REMARQUE 5.7. Nous avons défini les tests précédents en utilisant les intervalles de confiance I_n et J_n, mais il peut être utile de revenir encore plus tôt dans la construction de ces intervalles et d'utiliser une formulation plus simple. Par exemple, les deux tests précédents s'écrivent aussi :

$$\Phi = \mathbf{1}_{\sqrt{n}\left|\frac{\bar{X}_n - \mu_0}{\sigma}\right| > F^{-1}(1-\alpha/2)} \quad \text{et} \quad \Phi' = \mathbf{1}_{\sqrt{n}\left|\frac{\bar{X}_n - \mu_0}{\hat{\sigma}_n}\right| > F_{n-1}^{-1}(1-\alpha/2)}. \tag{12}$$

5.4 — Région critique et lien avec les intervalles de confiance

La région critique W d'un test Φ est l'ensemble des réalisations $(X_1,\dots,X_n)$ qui amènent à rejeter l'hypothèse nulle, c'est-à-dire les réalisations pour lesquelles $\Phi = 1$. Autrement dit,

$$W = \{(X_1,\dots,X_n) \in \mathbb{R}^n : \text{ au vu de } (X_1,\dots,X_n), \text{ on rejette } H_0\}. \tag{13}$$

Dans de nombreux cas, la région critique d'un test est un intervalle de confiance. C'est par exemple de cette façon qu'on a construit les tests de la section précédente. De façon générale, supposons que l'on dispose d'intervalles de confiance pour un paramètre μ : pour un niveau fixé $1-\alpha$, il est possible de construire un intervalle I_n tel que $\mu \in I_n$ avec probabilité $1-\alpha$. On peut alors tester des hypothèses de type H_0 : « $\mu = \mu_0$ » en posant tout simplement

$$\Phi = \begin{cases} 0 \text{ si } \mu_0 \text{ est dans l'intervalle } I_n, \\ 1 \text{ sinon.} \end{cases} \tag{14}$$

5.5 — p-valeur

La p-valeur est un outil essentiel de la méthodologie statistique, mais son interprétation est délicate et sujette à de nombreuses erreurs [19]. Le programme officiel de l'épreuve ne mentionne pas la p-valeur, mais il est à peu près certain que vous devrez commenter une p-valeur dans un exercice d'application, lors de votre oral. Il faut donc impérativement être au point, sans quoi vous risquez fort de vous perdre dans vos explications.

La p-valeur se calcule dans le cadre du test d'une hypothèse nulle H_0, dans lequel on cherche à réfuter l'hypothèse H_0. Mathématiquement, elle se définit comme suit.

DÉFINITION 5.8. — *Soit Φ_α une famille de tests, le test Φ_α étant de niveau $1-\alpha$. La p-valeur d'une réalisation $(X_1,\dots,X_n)$ est la plus petite valeur de α telle que $\Phi_\alpha = 1$.*

Un exemple rendra la chose plus claire.

Supposons que l'on désire tester l'hypothèse nulle H_0 : « $\mu = 0$ », dans un modèle gaussien où l'on dispose d'un n-échantillon $(X_1,\dots,X_n)$. La variance est connue : $\sigma^2 = 1$. Il est maintenant parfaitement assimilé que sous H_0, l'intervalle

$$I_n(1-\alpha) = \left[\bar{X}_n - \frac{t_\alpha}{\sqrt{n}}, \bar{X}_n + \frac{t_\alpha}{\sqrt{n}}\right] \tag{15}$$

contient le paramètre $\mu_0 = 0$ avec probabilité supérieure à $1-\alpha$. Pour l'application numérique, on suppose que $n = 100$ et que la réalisation observée donne une moyenne empirique de $\bar{X}_n = 0.21$.

Si l'on choisit un niveau $1-\alpha = 95\%$, alors $t_\alpha = 1.96$. L'intervalle est égal à

$$I_{100}(0.95) = \left[0.21 - \frac{1.96}{10}, 0.21 + \frac{1.96}{10}\right] = [0.014, 0.406].$$

Comme cet intervalle ne contient pas zéro, on est amené à rejeter l'hypothèse nulle, et à formuler une affirmation de type : « si mon modèle suivait l'hypothèse nulle, les observations que j'ai réalisées auraient moins de 5% de chances de se produire. Par conséquent, je rejette l'hypothèse nulle » .

Pourtant, 0 est vraiment très proche d'une des bornes de l'intervalle : si l'on avait été plus exigeant, on aurait certainement conservé H_0... Supposons donc que le niveau de test soit $1-\alpha = 99\%$. Dans ce cas, on a $t_\alpha = 2.58$. Le nouvel intervalle est

$$I_{100}(0.99) = \left[0.2 - \frac{2.58}{10}, 0.2 + \frac{2.58}{10}\right] = [-0.048, 0.468].$$

Cette fois, l'intervalle contient μ_0, donc notre procédure de test ne permet pas de rejeter H_0.

En diminuant le risque auquel on consent, on est donc incité à conserver l'hypothèse nulle plus souvent. **La p-valeur du problème est le plus petit risque auquel on doit consentir si l'on veut rejeter l'hypothèse nulle compte tenu des observations**.

19. Il s'agit certainement du paramètre statistique le plus mal compris ou mal utilisé de tous les temps. On renvoie à l'excellent article « The earth is round ($p < .05$) » (Jacob Cohen, *The American Psychologist*, 1994).

Ce que la p-valeur n'est pas.

1. *La p-valeur n'est pas la probabilité que l'hypothèse nulle soit fausse*. D'ailleurs, en statistiques, on ne mesure pas la probabilité qu'une hypothèse soit vraie ou fausse. On mesure la compatibilité d'une hypothèse avec des observations.
2. *La p-valeur n'est pas la probabilité de se tromper en rejetant l'hypothèse nulle*. Cela, c'est l'erreur de première espèce.
3. *La p-valeur n'est pas une caractéristique intrinsèque d'un test*. Elle dépend des observations, c'est une variable aléatoire. Pour une même procédure de test avec les mêmes paramètres, deux expériences différentes peuvent donner deux p-valeurs différentes.

MÉTHODE 5.9 (lire une p-valeur)**.** Supposons que sous l'hypothèse nulle, une certaine statistique T suive une certaine distribution connue, disons Z (loi normale, loi de Student, loi du chi-deux), dont la fonction de répartition est F. On construit un test Φ_α en choisissant, par exemple, une région de rejet de la forme $W = \{T > t_\alpha\}$, où $\mathbf{P}(Z > t_\alpha) = \alpha$.

Évidemment, lorsque $1-\alpha$ augmente, t_α augmente : si l'on veut rejeter l'hypothèse nulle avec une faible erreur de première espèce, on a intérêt à rejeter H_0 uniquement si t_α est vraiment très grand.

Supposons que lors d'une expérience, la statistique de test T prend une certaine valeur t. Si t_α est plus petit que t, on rejette H_0 ; si t_α est plus grand que t, on conserve H_0. La p-valeur est la valeur limite : c'est le α^* tel que $t_{\alpha*} = t$, autrement dit

$$p(t) = 1 - F(t) = \text{probabilité sous } H_0 \text{ que la statistique du test soit plus grande que } t. \tag{16}$$

C'est pour cela que dans de nombreux logiciels de statistiques, lorsqu'un test est fondé sur une certaine statistique T, on peut souvent lire une colonne étiquetée t, donnant la valeur de la statistique de test, et une colonne étiquetée "$P > |t|$", ce qui est une façon un peu sibylline de définir la p-valeur correspondant à l'observation t.

6
Tests du chi-deux

Les tests dits « du χ^2 » sont une famille de tests permettant de mesurer à quel point deux distributions se ressemblent ou pas. Ces tests utilisent tous les lois dites « du χ^2 ». Il est vivement recommandé de connaître leur définition, et c'est pour cela que nous faisons une section de rappels.

6.1 — La loi du chi-deux

DÉFINITION 6.1. — *La loi du chi-deux à m degrés de liberté, notée χ^2_m, est la loi de la somme des carrés de m variables aléatoires indépendantes normales centrées réduites.*

Autrement dit, si $(\mathrm{N}_1,\dots,\mathrm{N}_m)$ est un échantillon gaussien standard $\mathscr{N}(0,1)$, alors

$$\mathrm{N}_1^2+\cdots+\mathrm{N}_m^2 \sim \chi^2_m. \tag{17}$$

La loi du chi-deux possède une densité sur $\mathbb{R}_+$, mais son expression [20] n'est pas exigible.

REMARQUE 6.2. Les lois χ^2_k sont tabulées pour les petites valeurs de k. Pour les grandes valeurs de k, on invite le lecteur à appliquer le théorème central limite pour constater que si $\mathrm{T}_k \sim \chi^2_k$, alors $\mathrm{T}_k/\sqrt{k}$ converge en loi vers $\mathscr{N}(0,1)$.

Tests du chi-deux : un exemple simple. L'idée derrière tous les tests du χ^2 résulte d'une extrapolation du théorème central-limite. Supposons en effet que l'on dispose de deux variables aléatoires indépendantes, disons $\mathrm{A}_n,\mathrm{B}_n$, de loi binomiale $\mathrm{Bin}(n,p),\mathrm{Bin}(n,q)$. Le théorème central-limite dit que

$$\tilde{\mathrm{A}}_n = \frac{\mathrm{A}_n - np}{\sqrt{np(1-p)}} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1) \qquad \tilde{\mathrm{B}}_n = \frac{\mathrm{B}_n - nq}{\sqrt{nq(1-q)}} \xrightarrow{\mathscr{L}} \mathscr{N}(0,1). \tag{18}$$

Par conséquent, à supposer que les deux limites sont indépendantes [21], on devrait avoir $\tilde{\mathrm{A}}_n^2+\tilde{\mathrm{B}}_n^2 \xrightarrow{\mathscr{L}} \chi^2_2$. Un statisticien cherchant à tester l'hypothèse H_0 : « $p=q=1/2$ » pourra ainsi utiliser la convergence suivante :

$$\frac{2}{n}\left(\left(\mathrm{A}_n-\frac{n}{2}\right)^2+\left(\mathrm{B}_n-\frac{n}{2}\right)^2\right) \xrightarrow{\mathscr{L}} \chi^2_2 \tag{19}$$

et définir un test en utilisant les quantiles de la loi du χ^2_2. Les sections suivantes ont pour objectif de présenter cette procédure sous une forme manipulable, mais il s'agit bien de la même chose.

6.2 — Tests du chi-deux d'adéquation à une loi uniforme

On dispose de n observations $x_1,\dots,x_n$, chaque x_i pouvant prendre une valeur parmi d. Par exemple, la variable x_i indique 0 si la taille d'une personne est inférieure à 170 centimètres ou 1 si elle est supérieure. On divise les observations en d classes, la classe $\mathscr{C}_k$ regroupant tous les i tels que $x_i=k$. Dans la plupart des tests d'adéquation, on cherche à savoir si les observations sont uniformément réparties parmi les classes. L'hypothèse nulle qu'on veut tester est que les x_i sont indépendants, et uniformément distribués sur chaque classe, autrement dit $\mathbf{P}(x_i=k)=1/d$.

EXEMPLE 6.3 (Scarabées)**.** Jean-Eugène est un éleveur de scarabées. Il en possède $n=132$. Il y a des scarabées qui ont des élytres bleus, d'autres qui ont des élytres noirs, certains ont des élytres violets. Jean-Eugène se demande si ces couleurs sont uniformément réparties et son amie Jeanne-Eugénie lui propose de faire un test du chi-deux. Les trois classes sont naturellement les trois couleurs. Il y a $n_1=48$ scarabées bleus, $n_2=29$ scarabées noirs, et $n_3=55$ scarabées violets.

L'hypothèse nulle est H_0 : « les trois couleurs sont uniformément réparties parmi la population de Scarabée élevée par Jean-Eugène », et l'hypothèse alternative H_1 : « les trois couleurs *ne* sont *pas* uniformément réparties parmi la population de Scarabée élevée par Jean-Eugène »

20. La densité est $g(x)=2^{-k/2}\Gamma(k/2)^{-1}x^{k/2-1}e^{-x/2}$ si $x>0$, zéro sinon.
21. Ce n'est pas absolument trivial. N'en parlons pas.

Quelques définitions sont nécessaires.

DÉFINITION 6.4. —

— *Le nombre d'observations dans la classe $\mathscr{C}_k$ est appelé effectif empirique et on le note $\hat{n}_k$. La fréquence empirique est simplement notée $\hat{p}_k = \hat{n}_k/n$.*
— *Sous l'hypothèse nulle, le nombre d'observations dans la classe k suit une loi binomiale* $\mathrm{Bin}(n, 1/d)$*, et son espérance est égale à n/d : c'est l'effectif théorique, noté n_k. La fréquence théorique correspondante est $p_k = n_k/n = 1/d$.*

EXEMPLE 6.5 (Scarabées)**.** Jeanne-Eugénie calcule les effectifs théoriques et empiriques pour Jean-Eugène. Clairement, sous l'hypothèse nulle, l'effectif théorique de chacune des trois classes est $n/3 = 132/3 = 44$ et la fréquence théorique est $1/3$.

DÉFINITION 6.6. — *Le contraste du χ^2 associé à l'observation $(x_1, \dots, x_n)$ est défini par*

$$\mathrm{D}_n = \sum_{k=1}^{d} \frac{(\hat{n}_k - n_k)^2}{n_k} = n \sum_{k=1}^{d} \frac{(\hat{p}_k - p_k)^2}{p_k}. \tag{20}$$

Il se trouve que sous l'hypothèse nulle H_0, la variable aléatoire D_n converge vers une loi classique très connue, la loi du chi-deux à $d-1$ paramètres de liberté, notée χ^2_{d-1}. Avec un abus de notation notoirement pratique, nous noterons $k_{d-1,t}$ le quantile associé à $1-t$; autrement dit, si $\mathrm{Z} \sim \chi^2_{d-1}$, alors $\mathbf{P}(\mathrm{Z} < k_{d-1,t}) = 1-t$.

THÉORÈME 6.7. — *La fonction Φ définie par*

$$\Phi = \begin{cases} 1 \text{ si } \mathrm{D}_n > k_{d-1,\alpha}, \\ \text{zéro sinon} \end{cases} \tag{21}$$

est un test de niveau asymptotique $1-\alpha$.

EXEMPLE 6.8 (Scarabées)**.** Jeanne-Eugénie calcule maintenant le contraste du chi-deux associé aux scarabées de Jean-Eugène. Elle choisit un niveau de confiance $1-\alpha = 0.95$. Il y a trois classes, donc $d-1$ degrés de liberté. La procédure sera la suivante : d'abord, elle calcule D_n. Puis, si $\mathrm{D}_n > k_{2,0.05}$, l'hypothèse nulle sera rejetée. Ici, on lit sur une table du chi-deux que $k_{d-1,0.05} = 5.99$. Le calcul donne

$$\begin{aligned} \mathrm{D}_n &= \frac{(\hat{n}_1 - n_1)^2}{n_1} + \frac{(\hat{n}_2 - n_2)^2}{n_2} + \frac{(\hat{n}_3 - n_3)^2}{n_3} \\ &= \frac{(48-44)^2}{44} + \frac{(29-44)^2}{44} + \frac{(55-44)^2}{44} \\ &= \frac{16}{44} + \frac{225}{44} + \frac{121}{44} = \frac{362}{44} \\ &\approx 8.227 \end{aligned}$$

Jean-Eugène et Jeanne-Eugénie rejettent l'hypothèse nulle : il est plausible, compte tenu des observations, d'affirmer que les trois couleurs n'apparaissent pas de façon uniforme sur les élytres des scarabées. Cependant, s'ils avaient été plus exigeants et choisi $1-\alpha = 0.99$, ils auraient constaté que $k_{2,0.01} = 9.21$, et ils n'auraient pas pu rejeter l'hypothèse nulle. La p-valeur se situe donc entre 0.01 et 0.05.

6.3 — Tests du chi-deux d'adéquation à une loi

Il s'agit du même test que dans la section précédente, mais avec une loi qui n'est pas nécessairement uniforme. Il est donc nécessaire de décrire la fréquence théorique p_i de la classe i, puis d'effectuer le calcul des effectifs théoriques $n_i = p_i n$ avec ces valeurs. À part cela, aucune différence.

EXEMPLE 6.9 (Scarabées)**.** Jeanne-Eugénie décide de tester une autre répartition de la couleur des élytres des scarabées. Elle remarque qu'il y a peu de scarabées noirs et se demande si la loi de référence ne serait pas plutôt $p_1 = 2/5, p_2 = 1/5$ et $p_3 = 2/5$. Son hypothèse nulle est donc H_0 : « la répartition des couleurs dans la population des scarabées de Jean-Eugène est $2/5, 1/5, 2/5$ comme ci-dessus ».

Les effectifs théoriques sont

$$n_1 = p_1 n = 52.8 \qquad n_2 = p_2 n = 26.4 \qquad n_3 = p_3 n = 52.8.$$

Le contraste du chi-deux est :

$$\begin{aligned} \mathrm{D}_n &= \frac{(\hat{n}_1 - n_1)^2}{n_1} + \frac{(\hat{n}_2 - n_2)^2}{n_2} + \frac{(\hat{n}_3 - n_3)^2}{n_3} \\ &= \frac{(48 - 52.8)^2}{52.8} + \frac{(29 - 26.4)^2}{26.04} + \frac{(55 - 52.8)^2}{52.8} \\ &= \frac{23.04}{52.8} + \frac{8.76}{26.4} + \frac{4.84}{52.8} \\ &= 0.43 + 0.33 + 0.09 \\ &\approx 0.85. \end{aligned}$$

Cette fois, la statistique de test est vraiment plus petite que $k_{2,0.05}$. L'hypothèse nulle ne peut pas être rejetée.

REMARQUE 6.10. Le programme de l'épreuve ne mentionne pas les tests d'adéquation à une famille de lois. Ce sont les tests d'hypothèses de type H_0 : « une certaine variable aléatoire X suit une loi de Poisson », mais on ne précise pas le paramètre : la variable aléatoire peut suivre une loi de Poisson de paramètre 10 ou 100.

Évidemment, pour faire ce genre de tests, on estime d'abord un éventuel paramètre $\hat{\lambda}$, puis on teste l'adéquation à la loi $\mathrm{Poi}(\hat{\lambda})$ comme ci-dessus. Cependant, comme on a estimé un paramètre, on perd un degré de liberté supplémentaire et le contraste du chi-deux suit une loi χ^2_{d-2}. Évidemment, certaines lois ont plus d'un paramètre (penser à la loi normale), et l'estimation de chaque paramètre fait perdre un degré de liberté supplémentaire.

6.4 — Tests du chi-deux d'indépendance

Un cas particulier de la procédure sur les tests d'adéquation ci-dessus est le test d'*indépendance*. On suppose que nos observations sont de la forme (x_i, y_i) : on cherche à déterminer si les x_i et les y_i sont indépendants. Par exemple, supposons que les x_i soient comme ci-dessus (la taille), et que y_i vaut 1 si la personne i possède des chaussettes vertes, et 0 sinon. Le fait de mesurer plus de 170 centimètres est-il vraiment indépendant du fait de posséder des chaussettes vertes ? Maintenant, l'hypothèse nulle est légèrement différente de ci-dessus : on cherche à tester H_0 : « les (x_i) et les (y_j) sont indépendants ».

On suppose que les x_i sont divisiés en c classes, et les y_j en d classes. On s'intéresse aux classes $\mathscr{C}_{i,j}$. Comme ci-dessus, on définit l'effectif empirique et la fréquence empirique par $\hat{n}_{i,j}$ et $\hat{p}_{i,j} = \hat{n}_{i,j}/n$.

1. les effectifs marginaux empiriques $\hat{n}_{i,\star} = \sum_j n_{i,j}$ et les fréquences marginales empiriques $\hat{p}_{i,\star} = \hat{n}_{i,\star}/n$.
2. de même, $\hat{n}_{\star,j} = \sum_i n_{i,j}$ et $\hat{p}_{\star,j} = \hat{n}_{\star,j}/n$.
3. Les fréquences théoriques $p_{i,j} = \hat{p}_{i,\star} p_{\star,j}$, et les effectifs théoriques :

$$n_{i,j} = n p_{i,j}.$$

Le contraste du chi-deux dans ce cas est défini par :

$$\mathrm{D}_n = \sum_{i=1}^{c} \sum_{j=1}^{d} \frac{(\hat{n}_{i,j} - n_{i,j})^2}{n} = n \sum_{i=1}^{c} \sum_{j=1}^{d} \frac{(\hat{p}_{i,j} - p_{i,j})^2}{p_{i,j}}. \tag{22}$$

Le résultat principal que nous utilisons dorénavant dit que sous l'hypothèse nulle H_0, la statistique D_n converge en loi vers une loi du χ^2. Son nombre de degrés de libertés est égal à $\ell = (c-1)(d-1)$.

THÉORÈME 6.11. — *On pose $\ell = (c-1)(d-1)$. La fonction Φ définie par*

$$\Phi = \begin{cases} 1 \text{ si } \mathrm{D}_n > k_{\ell,\alpha}, \\ \text{zéro sinon} \end{cases} \tag{23}$$

est un test de niveau asymptotique $1 - \alpha$.

6.5 — Règles de bienséance relatives aux tests du chi-deux

Les tests du chi-deux sont omniprésents en sciences humaines. Pourtant, leur utilisation raisonnée est soumise à certaines règles qui sont parfois négligées, et que vous devez avoir en tête.

Choix des classes et nombre d'observations. Les classes ne sont pas forcément données par le problème : souvent, c'est au statisticien de les construire.

Or, lorsque les effectifs théoriques sont trop faibles, le test perd de sa puissance. Par exemple, il peut arriver que l'effectif théorique d'une classe soit égal à — mettons — 0.01. En pratique, sous H_0, la probabilité pour que l'effectif de cette classe soit non nul est donc très faible [22] et il ne sera jamais possible de distinguer cette loi d'un effectif théorique nul. Pour contrevenir à ce problème, on adopte la règle suivante :

Dans un test du chi-deux, les effectifs théoriques doivent être tous supérieurs à 5.

Cela veut dire que si vous devez créer les classes par vous-mêmes, vous devez le faire de sorte que cette règle soit satisfaite. Si, dans un exercice, les classes sont déjà données, vous devez alors vérifier que cette règle est satisfaite, puis si elle ne l'est pas, vous devez regrouper les observations jusqu'à ce qu'elle le soit.

Degrés de liberté, moins un. Lorsqu'on divise les n observations en d classes, le choix des $d-1$ premières classes détermine entièrement le choix de la dernière. Par conséquent, les variables $\hat{n}_1, \dots, \hat{n}_d$ ne sont pas indépendantes, puisque leur somme est n. C'est cette dépendance qui fait perdre un degré de liberté dans la loi limite, χ^2_{d-1}.

Calcul de la puissance. La probabilité de rejeter le modèle théorique à tort est connue, c'est α ; par contre on ne peut calculer la probabilité d'accepter le modèle théorique à tort. D'autre part, un test, faute de preuves expérimentales suffisantes, se rabat sur l'hypothèse H_0. On se gardera d'utiliser les facilités des logiciels pour tester à tort et à travers une multitude de modèles théoriques. Il faut au préalable avoir de bonnes raisons pour soupçonner tel ou tel modèle théorique.

6.6 — Analyse de la variance (ANOVA)

L'analyse de la variance est un test du chi-deux.

Elle est utilisée dans le cadre suivant : il y a une variable Y à expliquer (disons, le salaire), et une variable explicative X qui ne peut prendre qu'un nombre fini de valeurs, disons $c_1, \dots, c_p$ (par exemple, la variable qui indique la CSP d'un individu). On dispose donc d'un échantillon $(y_1, x_1), \dots, (y_n, x_n)$; l'objectif des procédures dites ANOVA (*analysis of variance*) est de scinder ces n observations en d classes, la classe k étant constituées des observations (y, x) telles que $x = c_k$. Ensuite, on compare la variance des variables expliquées au sein de chaque classe, avec la variance globale du modèle.

Avant de résumer les principales définitions, on rappelle que la variance empirique (naïve [23]) d'un échantillon $(y_1, \dots, y_n)$ est l'estimateur défini par :

$$\bar{\sigma}_n^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

où $\bar{y}$ est la moyenne empirique.

Dans la suite, on notera n_k le nombre d'observations dans la classe k, de sorte que $n = n_1 + \cdots + n_d$, et on notera $y_1^{(k)}, \dots, y_{n_k}^{(k)}$ ces observations. On calcule d'abord la variance empirique observée à l'intérieur de chaque classe, et on la note $\bar{\sigma}_k^2$:

$$\bar{\sigma}^2_{\text{intra},k} = \frac{1}{n_k} \sum_{i=1}^{n_k} (y_i^{(k)} - \bar{y}^{(k)})^2 = \left(\frac{1}{n_k} \sum_{i=1} (y_i^{(k)})^2 \right) - (\bar{y}^{(k)})^2$$

où $\bar{y}^{(k)}$ est la moyenne empirique des observations dans la classe k.

DÉFINITION 6.12. — *On appelle* variance intra*(-classes) la moyenne des* $\bar{\sigma}^2_{\text{intra},k}$*. Autrement dit, c'est la moyenne pondérée des variances observées dans chacune des classes :*

$$\bar{\sigma}^2_{\text{intra}} = \frac{1}{n} \sum_{k=1}^{d} n_k \bar{\sigma}_k^2. \tag{24}$$

22. $\mathbf{P}(\text{ effectif de cette classe } > 0.1) \leqslant 0.01/0.1 = 0.1$ par Markov.

23. Le moment ne saurait être mieux choisi pour relire le paragraphe 3.3 page 13 sur l'estimation de la variance et la différence entre $\hat{\sigma}_n^2$ et $\bar{\sigma}_n^2$, bien que cette différence ne soit pas franchement cruciale ici.

On appelle variance inter(-classes) la 'variance pondérée' des moyennes $\bar{y}^{(k)}$. Ainsi :

$$\bar{\sigma}^2_{\text{inter}} = \frac{1}{n}\sum_{k=1}^{d} n_k(\bar{y}^{(k)} - \bar{y})^2 \tag{25}$$

où $\bar{y}$ est simplement la moyenne des y_i.

PROPOSITION 6.13 (décomposition de la variance)**.** — *Désignons par $\bar{\sigma}^2_n$ la variance empirique globale des observations. Alors,*

$$\bar{\sigma}^2_n = \bar{\sigma}^2_{\text{intra}} + \bar{\sigma}^2_{\text{inter}}. \tag{26}$$

Bien que théoriquement absente du programme, la démonstration pourra être posée en question d'oral par le jury.

Démonstration. Il suffit de développer, puis de trier en fonction des classes :

$$\begin{aligned}
n\bar{\sigma}^2_n &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\
&= \sum_{k=1}^{d}\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y})^2 \\
&= \sum_{k=1}^{d}\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y}^{(k)} + \bar{y}^{(k)} - \bar{y})^2 \\
&= \sum_{k=1}^{d}\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y}^{(k)})^2 + (\bar{y}^{(k)} - \bar{y})^2 + 2(y_j^{(k)} - \bar{y}^{(k)})(\bar{y}^{(k)} - \bar{y}) \\
&= \sum_{k=1}^{d} n_k\bar{\sigma}^2_{\text{intra},k} + \sum_{k=1}^{d} n_k(\bar{y}^{(k)} - \bar{y})^2 + 2\sum_{k=1}^{d}\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y}^{(k)})(\bar{y}^{(k)} - \bar{y}) \\
&= n\bar{\sigma}^2_{\text{intra}} + n\bar{\sigma}^2_{\text{inter}} + 2\sum_{k=1}^{d}(\bar{y}^{(k)} - \bar{y})\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y}^{(k)}).
\end{aligned}$$

Or, en développant la définition de $\bar{y}^{(k)}$ dans le tout dernier terme, on voit que $\sum_{j=1}^{n_k}(y_j^{(k)} - \bar{y}^{(k)}) = n_k\bar{y}^{(k)} - n_k\bar{y}^{(k)} = 0$. Dans la dernière ligne, le troisième des termes est donc nul et on a l'identité recherchée. □

La proportion $\frac{\bar{\sigma}^2_{\text{intra}}}{\bar{\sigma}^2_n}$ est la *part de la variance de* Y *expliquée par* X. Par exemple, supposons que $\bar{\sigma}^2_{\text{inter}} = 0$: cela signifie que les moyennes empiriques au sein de chaque classe, les $\bar{y}^{(k)}$, sont toutes identiques. La variance du modèle est alors égale à la moyenne de la variance au sein de chaque classe. Dans l'autre sens, si la variance intra-classe est nulle, cela veut dire qu'au sein d'une classe, toutes les observations y_i sont égales. C'est donc la variance entre les différentes classes qui contribuera à la variance totale.

Exercice 6.14. Expliquer le tableau suivant, dans lequel Y est le nombre d'enfants d'une famille et X le type de famille considérée (famille monoparentale ou couple), et calculer la part de la variance de Y expliquée par X.

	Effectif	$\sum y_i$	$\sum y_i^2$	$\overline{y_i}$	σ_i^2	$n_{i.}\,\overline{y_i}^2$	$n_{i.}\,\sigma_i^2$	
Mono	1035	1669	3199	1,6126	0,4903	2691,4955	507,4605	
Couples	6536	11036	22802	1,6885	0,6376	18634,3468	4167,3536	
Total	7571	12705	26001			21325,8423	4674,8141	
Moyenne		1,6781	3,4343			2,8168	0,6175	
		Variance	0,6183		Inter	0,0008	0,6175	Intra

Le théorème suivant permet d'effectuer des *tests d'égalité des moyennes par analyse des variances*.

THÉORÈME 6.15. — *Sous l'hypothèse* H_0 *selon laquelle les espérances conditionnelles* $\mathbf{E}[Y|X = c_i]$ *sont identiques entre les différentes classes, on a*

$$(n-d)\frac{\bar{\sigma}^2_{\text{inter}}}{\bar{\sigma}^2_{\text{intra}}} \xrightarrow[n\to+\infty]{\mathscr{L}} \chi^2_{d-1}. \tag{27}$$

Pour effectuer un test ANOVA d'égalité des moyennes, on procède donc de la façon suivante. On calcule d'abord la statistique $T = (n-d)\frac{\bar{\sigma}^2_{\text{inter}}}{\bar{\sigma}^2_{\text{intra}}}$. Sous l'hypothèse nulle selon laquelle la moyenne des variables expliquées ne dépend pas de la variable explicative, T_{n-d} suit approximativement une loi du χ^2_{d-1}. On rejettera donc l'hypothèse nulle si $T_{n-d} > k_{d-1,\alpha}$, c'est-à-dire si la variance inter-classes est vraiment plus grande que la variance au sein de chaque classe.

Exercice 6.16. Réaliser le test pour les données de l'exercice précédent, avec un niveau de confiance 95%.

7
Ajustement affine

Le principe de l'ajustement affine est de rechercher une relation affine permettant d'expliquer le comportement d'une variable quantitative Y en fonction d'une autre variable quantitative X à partir d'observations empiriques (x_i, y_i). Ces deux variables peuvent être unidimensionnelles (âge, revenu, nombre d'années d'études, nombre d'heures consacrées quotidiennement à une activité, indicatrice d'une variable qualitative...) ou multidimensionnelles (vecteur d'indicatrices codant les réponses à certaines questions, liste des âges des enfants de l'individu, couple de variables unidimensionnelles...).

Dans ces notes, on utilisera de manière interchangeable les termes *ajustement affine* et *régression linéaire*.

7.1 — Cas unidimensionnel

On se place dans cette section dans le cas où X et Y sont toutes deux unidimensionnelles, qui se présente par exemple lorsque l'on cherche à expliquer le niveau de consommation d'un individu par son revenu. Le problème d'ajustement affine revient alors à trouver des réels a et b tels que pour chaque observation i, la variable expliquée y_i soit le plus proche possible de $ax_i + b$; autrement dit, tels que la distance entre le nuage de points (x_i, y_i) observé et la droite d'équation $y = ax + b$ soit aussi petite que possible (voir figure 7).

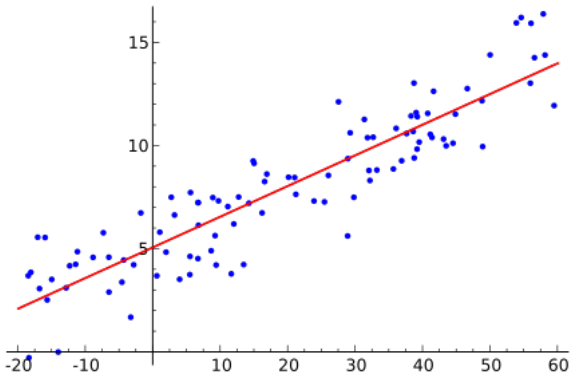


FIGURE 7 – Ajustement (« fit ») d'une droite à un nuage de points

Pourquoi s'amuse-t-on à faire cela ? Au-delà du caractère plaisant de sa grande simplicité, une relation linéaire $y = ax + b$ permet d'effectuer aisément des prédictions et d'énoncer des formules élégantes du type « une année d'études supplémentaires permet de gagner a euros supplémentaires par an », ou « un kilogramme de surcharge pondérale supplémentaire fait passer 8 minutes de plus devant la télévision chaque jour ». Ce dernier exemple, volontairement provocateur, invite bien sûr à prendre le recul nécessaire sur l'interprétation causale de données d'ajustement...

7.2 — La méthode de Mayer

Une méthode naïve pour trouver une droite d'ajustement consiste à partitionner l'ensemble des observations en deux classes d'effectif égal en définissant les groupes de part et d'autre de la médiane des abscisses, puis à placer les deux points correspondant aux moyennes des variables sur ces deux sous-groupes. On trace alors la droite qui passe par ces deux points : c'est la *méthode de Mayer*[24]. La figure 8 représente la mise en pratique d'une telle méthode. Notons qu'elle est sensible au choix des sous-groupes et qu'il est possible de choisir ceux-ci de plusieurs façons lorsque le nombre d'observations est impair ou que plusieurs observations correspondent à la valeur médiane des x_i.

24. Johann Tobias Mayer (1723-1762), astronome allemand. Précisons qu'en dehors du cercle assez restreint des astronomes allemands du milieu du XVIIIeme siècle, personne n'utilise la méthode de Mayer.

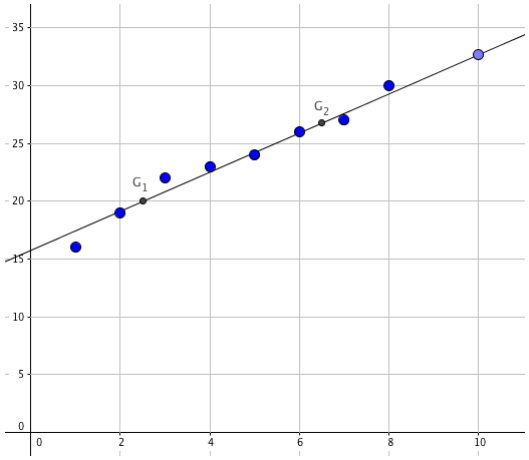


FIGURE 8 – Ajustement d'une droite à un nuage de points par la méthode de Mayer. Le premier groupe contient les quatre premiers points, le second les cinq derniers, et G_1 et G_2 sont les points moyens (barycentres de chacun des groupes).

La méthode de Mayer permet à de grandes erreurs d'ajustement de se compenser au sein d'un sous-groupe. Par exemple, le sous-groupe constitué des observations $(5,5)$, $(7,7)$ et $(9,9)$ correspond au même point moyen que le sous-groupe constitué des observations $(5,3.5)$, $(7,5.5)$ et $(9,12)$. *A priori*, cela ne choque pas l'intuition. Considérons toutefois le problème suivant : lequel des deux points $(1,5)$ et $(3,3)$ est à plus petite distance du point $(0,0)$? En plaçant ces points dans un repère, on se convainc rapidement du fait que $(3,3)$ est le moins éloigné de $(0,0)$ (au sens classique de la norme euclidienne), même si l'on peut passer de $(1,5)$ à $(3,3)$ par une simple « compensation » des coordonnées. Cela résulte du fait que la distance entre deux points (x_1,y_1) et (x_2,y_2) du plan n'est pas donnée par

$$d\big((x_1,y_1),(x_2,y_2)\big) = |x_1-x_2|+|y_1-y_2|$$

mais plutôt par

$$d\big((x_1,y_1),(x_2,y_2)\big) = \sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$$

La règle de compensation qui prévaut dans la méthode de Mayer n'est donc pas vérifiée dans le cas de la distance euclidienne dans le plan, qui pénalise beaucoup les grandes déviations de coordonnées (en effet, $d\big((0,0),(1,5)\big) > d\big((0,0),(3,3)\big)$!).

7.3 — Moindres carrés ordinaires

Pour mieux se rapprocher de la notion de distance euclidienne, on peut choisir de définir la distance d'ajustement $\mathrm{D_{aj}}$ entre les observations $(x_1,y_1),\dots,(x_n,y_n)$ et les prédictions réalisées à partir des x_i par la droite d'ajustement, c'est-à-dire $(x_1,ay_1+b),\dots,(x_n,ay_n+b)$ de la façon suivante :

$$\mathrm{D_{aj}} = \sqrt{\sum_{i=1}^{n}\Big[(x_i-x_i)^2+(y_i-(ax_i+b))^2\Big]} = \sqrt{\sum_{i=1}^{n}(y_i-ax_i-b)^2}$$

$\mathrm{D_{aj}^2}$ est donc la somme des *carrés* des erreurs entre les données et leurs valeurs prédites. Choisir $\mathrm{D_{aj}^2}$ comme distance revient à pénaliser fortement les grandes erreurs d'ajustement [25], de la même façon que la distance euclidienne pénalise fortement les grands écarts de coordonnées entre deux points. La figure 9 représente les erreurs, appelées *résidus* (en vert), associées à la droite d'ajustement censée approcher les différentes données ; $\mathrm{D_{aj}^2}$ est donc la somme des carrés des longueurs des segments verts. On se convaincra facilement que si cette somme est nulle, alors l'ajustement affine est parfait.

La méthode de régression par les moindres carrés consiste donc à trouver a et b minimisant

$$\mathrm{D_{aj}} = \sqrt{\sum_{i=1}^{n}\Big[(x_i-x_i)^2+(y_i-(ay_i+b))^2\Big]} = \sqrt{\sum_{i=1}^{n}(y_i-ax_i-b)^2}.$$

25. Notons que c'est exactement cette démarche de généralisation de la distance euclidienne qui mène à définir la variance d'une variable aléatoire réelle X non pas comme $\mathbf{E}\left(|\mathrm{X}-\mathbf{E}(\mathrm{X})|\right)$ mais plutôt comme $\mathbf{E}\left((\mathrm{X}-\mathbf{E}(\mathrm{X}))^2\right)$, la moyenne des *carrés* des écarts à la moyenne.

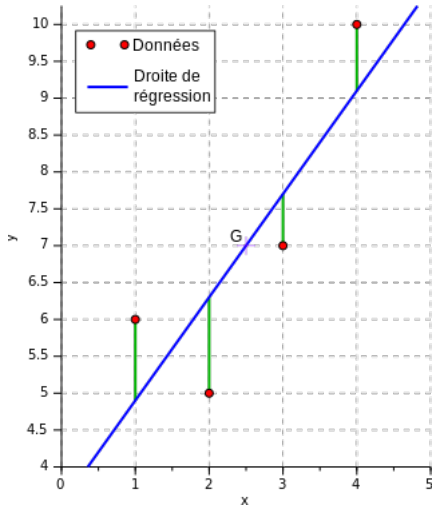


FIGURE 9 – Calcul des résidus associés à une droite de régression

On parle souvent du problème de minimisation des MCO. Il arrive que ce problème soit formulé de la façon suivante : trouver a, b qui minimisent la fonction [26] $f(a,b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2$. Supposons qu'au moins deux des x_i sont différents (sinon le problème est trivial !). On peut alors montrer les formules suivantes, dont la démonstration n'est pas exigée, mais qui doivent être connues. Ces formules s'expriment

THÉORÈME 7.1 (Formule qui donne l'estimateur des MCO)**. —** *Les nombres $\hat{a}, \hat{b}$ qui donnent le meilleur ajustement affine pour les données (x_i, y_i) sont*

$$\hat{a} = \frac{\sum_{i=1}^{n} (x_i - \overline{x_n})(y_i - \overline{y_n})}{\sum_{i=1}^{n} (x_i - \overline{x_n})^2} \tag{28}$$

et

$$\hat{b} = \overline{y_n} - \hat{a}\overline{x_n}, \tag{29}$$

La première formule peut paraître compliquée, mais elle est en réalité assez intuitive et se comprend mieux à l'aide de quelques notations habiles. On introduit ainsi la covariance empirique entre les données explicatives x_i et les données à expliquer y_i par

$$\mathrm{Cov}(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x_n})(y_i - \overline{y_n}).$$

De même, la variance empirique des variables explicatives a déjà été rencontrée dans le cours (attention, ici on adopte la normalisation naïve en $1/n$) :

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x_n})^2.$$

Avec ces notations, le coefficient $\hat{a}$ se reformule assez simplement :

$$\hat{a} = \frac{\mathrm{Cov}(x,y)}{\sigma_x^2}. \tag{30}$$

Pour bien comprendre pourquoi ce résultat est intuitif, on peut faire l'expérience suivante : que se passe-t-il si les données sont effectivement parfaitement affines, autrement dit si on a déjà $y_i = ax_i + b$ pour deux nombres a et b ?

26. Dans l'expression de f, on a enlevé la racine carrée que l'on voyait dans D_{aj}. C'est parce que minimiser une quantité positive ou minimiser sa racine carrée ne change rien au point en lequel est atteint le minimum — même si la valeur de ce minimum n'est pas la même.

Dans ce cas, il est très facile de voir que $\overline{y_n} = a\overline{x_n} + b$. On peut alors faire le calcul de la covariance empirique :

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(y_i - \overline{y_n}) \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(ax_i + b - a\overline{x_n} - b) \\
&= \frac{a}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})(x_i - \overline{x_n}) \\
&= \frac{a}{n}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 \\
&= a\sigma_x^2.
\end{aligned}$$

En faisant passer la variance de l'autre côté, on trouve donc que $\mathrm{Cov}(x,y)/\sigma_x^2 = a$.

Ainsi, dans le cas précis où les données sont déjà affines avec coefficients a et b, on constate que la formule (30) montre que $\hat{a}$ est effectivement égal à a, et il n'est pas difficile de se convaincre que $\hat{b} = b$, ce qui est tout de même bon signe sur la performance de notre méthode.

DÉFINITION 7.2 (Coefficient de corrélation linéaire)**. —** *Comme en probabilités, on définit le coefficient de corrélation linéaire entre les x_i et les y_i, dans le cas où ces observations prennent au moins deux valeurs différentes, par :*

$$\rho_{x,y} = \frac{\mathrm{Cov}(x,y)}{\sigma_x \sigma_y}.$$

Ce coefficient est toujours compris entre -1 et 1. Il vaut 1 (resp. -1) en cas d'adéquation linéaire *croissante* (resp. *décroissante*) parfaite entre les x_i et les y_i. Lorsqu'il est nul, on dit que les x_i et les y_i sont *linéairement décorrélés*. Enfin, ce coefficient a le bon goût d'être insensible aux changements d'unités (si l'on multiplie tous les x_i par 10, on peut vérifier que $\rho_{x,y}$ ne change pas de valeur) et d'être symétrique : il constitue donc un bon indicateur de la force de la corrélation linéaire entre les variables x et y.

7.4 — Régression linéaire multiple et matrice de corrélations

Pour une régression linéaire simple, on veut expliquer une variable réelle Y par une autre variable réelle X. Dans le cas de la régression linéaire multiple, on veut expliquer Y par plusieurs variables réelles, disons $\mathrm{X}^{(1)}, \ldots, \mathrm{X}^{(d)}$. Par exemple, on veut expliquer le revenu Y d'un individu par son âge, son niveau de formation et le revenu de ses parents.

On disposera alors de n observations de la forme (y_i, X_i), où y_i est le revenu de l'individu i, et X_i est un *vecteur* (et non plus un *réel*) possédant autant de coordonnées qu'il y a de variables explicatives, la coordonnée t correspondant à la caractéristique t de l'individu i.

Par exemple, on pourra avoir $\mathrm{X}_i = (x_i^1, x_i^2, x_i^3)$ où x_i^1 est l'âge de i, $\mathrm{X}_i^{(2)}$ son niveau de formation, etc. Attention, dans ce chapitre on écrit $\mathrm{X}_i = (x_i^1, \ldots, x_i^d)$ mais ce ne sont pas des exposants. On écrira d'ailleurs X^t pour le vecteur des n observations de la variable t, c'est-à-dire $\mathrm{X}^t = (x_1^t, x_2^t, x_3^t, \ldots, x_n^t)$.

La régression linéaire consiste alors à trouver des coefficients réels $a_0, a_1, \ldots, a_d$ tels que l'estimation

$$y_i \approx a_0 + \sum_{k=1}^{d} a_k x_i^k$$

soit aussi bonne que possible. Cela revient non plus à faire passer une droite dans un nuage de points du plan mais un *hyperplan* dans le nuage $\left\{(x_i^1, \ldots, x_i^d, y_i)\right\}_{i \in [\![1,n]\!]}$ de points de $\mathbb{R}^{d+1}$.

Une question naturelle que l'on pourrait se poser est la suivante : puisque la régression linéaire simple que nous avons vue à la section précédente permet d'isoler l'effet d'une variable sur la variable expliquée, pourquoi ne suffirait-il pas de réaliser d régressions linéaires pour connaître les effets a_k correspondants ? Les choses ne sont hélas pas si simples en raison de la *corrélation des variables explicatives entre elles*.

Par exemple, lorsque l'on réalise une régression linéaire simple du revenu d'un individu sur le revenu de ses parents, on isole effectivement l'effet du revenu des parents sur le revenu de l'enfant, mais il ne s'agit pas d'un effet

direct : cet effet agit, pour ne citer que des canaux évidents, à travers celui de la PCS des parents et du niveau de formation de l'individu. Si l'on régresse maintenant le revenu de l'individu à la fois sur le revenu de ses parents et sur son niveau de formation, il semble clair que le coefficient associé au revenu des parents sera moindre que celui obtenu précédemment puisque l'incorporation de la variable *niveau de formation* aura capturé une partie de l'effet indirect que nous venons de décrire. En ajoutant des variables représentant la PCS des parents dans la régression, on s'attend à obtenir pour la variable *revenu des parents* un coefficient encore plus faible, qui correspondra cette fois à l'effet du revenu des parents sur celui de l'individu indépendamment des variables relatives à la PCS et au niveau de formation. Cet effet, que l'on imagine bien être toujours positif, peut par exemple tenir à l'existence des transferts monétaires ou d'opportunités professionnelles offertes directement par les parents à leurs enfants.

Il apparaît donc qu'isoler l'effet pur[27] a_k d'une variable X^k sur la variable Y suppose de prendre en compte les corrélations existant entre les différentes variables explicatives.

Pour représenter ces corrélations, on utilise deux outils privilégiés. Ces outils ne sont pas propres aux modèles linéaires, et sont des outils statistiques puissants pour décrire des données numériques.

DÉFINITION 7.3 (Matrices de covariance et de corrélation linéaire)**.** — *Si $x^k = (x_1^k, \dots, x_n^k)$ est le vecteur des observations de la variable X^k, la matrice de covariance empirique des observations est définie par*

$$\mathrm{Cov}(x^1,\dots,x^d) = \begin{pmatrix} \mathrm{Cov}(x^1,x^1) & \mathrm{Cov}(x^1,x^2) & \dots & \mathrm{Cov}(x^1,x^d) \\ \mathrm{Cov}(x^2,x^1) & \mathrm{Cov}(x^2,x^2) & \dots & \mathrm{Cov}(x^2,x^d) \\ \vdots & \vdots & & \vdots \\ \mathrm{Cov}(x^d,x^1) & \mathrm{Cov}(x^d,x^2) & \dots & \mathrm{Cov}(x^d,x^d) \end{pmatrix}$$

où l'on a noté $\mathrm{Cov}(x^k,x^k) = \sigma^2_{x^k}$. On définit de la même façon la matrice de corrélation linéaire empirique par

$$\rho_{x^1,\dots,x^d} = \begin{pmatrix} \rho_{x^1,x^1} & \rho_{x^1,x^2} & \dots & \rho_{x^1,x^d} \\ \rho_{x^2,x^1} & \rho_{x^2,x^2} & \dots & \rho_{x^2,x^d} \\ \vdots & \vdots & & \vdots \\ \rho_{x^d,x^1} & \rho_{x^d,x^2} & \dots & \rho_{x^d,x^d} \end{pmatrix}$$

où $\rho_{x^k,x^k} = 1$.

Notons que les matrices définies ci-dessus sont symétriques (pourquoi ?). On expose sans plus de cérémoniel la méthode suivante, que l'on rapprochera avec profit du cas unidimensionnel traité plus haut :

MÉTHODE 7.4 (Régression linéaire multiple par la méthode des moindres carrés)**.** La régression linéaire multiple consiste à minimiser la somme des carrés des résidus

$$\mathrm{R}^2_{\mathrm{aj}} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i^1 - a_2 x_i^2 - \dots - a_d x_i^d \right)^2 .$$

Si les vecteurs $\mathrm{X}^k = (x_1^k, \dots, x_n^k)$ sont deux à deux non colinéaires[28], il existe un unique choix optimal de $(a_0, a_1, \dots, a_d)$ dont on peut donner une expression explicite[29].

Exercice 7.5. Montrer que le coefficient de corrélation entre deux variables aléatoires, ou entre deux séries statistiques, est toujours compris entre -1 et 1.

Exercice 7.6. Soit X une variable gaussienne centrée réduite. On pose $\mathrm{Y} = \mathrm{X}^2$. Calculer le coefficient de corrélation $\mathrm{Cov}(\mathrm{X},\mathrm{Y})/\sqrt{\mathrm{Var}(\mathrm{X})\mathrm{Var}(\mathrm{Y})}$ entre ces deux variables ; que faut-il en penser ?

Exercice 7.7. On dispose de données statistiques x_i, y_i qui représentent deux variables sociologiques. Pourquoi les termes diagonaux de la matrice des corrélations sont-ils tous égaux à 1 ? Que dire de la relation entre les deux variables lorsque leur matrice de corrélation prend les formes suivantes ?

$$\begin{pmatrix} 1 & -0.000001 \\ -0.000001 & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & 0.998 \\ 0.998 & 1 \end{pmatrix}.$$

27. Sans prétention aucune à accéder à la vérité ontologique des choses par la régression linéaire, on parle d'effet *pur* d'une variable explicative pour désigner son effet sur la variable expliquée indépendamment des autres variables explicatives prises en compte. De façon plus générale, ce que l'on désigne sous le nom d'*effet pur* d'une variable sur une autre est son effet pur dans un modèle idéal qui contiendrait toutes les variables explicatives pertinentes.

28. Ceci mérite une explication. Supposons que l'on dispose de trois variables explicatives, et que les vecteurs $\mathrm{X}^1, \mathrm{X}^2, \mathrm{X}^3$ sont colinéaires dans $\mathbb{R}^n$. Le cours d'algèbre vous explique que l'un d'entre eux peut s'exprimer comme une combinaison linéaire des deux autres : par exemple, $\mathrm{X}^3 = \alpha \mathrm{X}^1 + \beta \mathrm{X}^2$. Cela signifie que la variable X^3 n'apporte strictement aucune information supplémentaire par rapport aux deux premières variables : l'intégrer comme variable explicative est une erreur méthodologique.

29. ... matricielle et trop complexe pour être exigible à l'oral de l'agrégation.

8
Le modèle linéaire gaussien

Dans la section précédente, nous avons expliqué comment trouver de bonnes relations affines entre des variables expliquées et des variables explicatives, sans faire aucune véritable hypothèse sur les données. Or, pour pouvoir mesurer la qualité des estimateurs $\hat{a},\hat{b}$, il est vraiment utile de commencer à faire sur les données des hypothèses statistiques. La plus simple de ces hypothèses est aussi la plus naturelle : elle consiste à supposer que les données sont effectivement linéaires, c'est-à-dire que la variable à expliquer est effectivement *égale* à $ax_i + b$, mais que chaque observation y_i est polluée par un bruit ou une erreur de mesure, disons ε_i. Ainsi, nos observations seraient de la forme

$$y_i = ax_i + b + \varepsilon_i \tag{31}$$

et il faut retrouver les coefficients a, b.

Le modèle gaussien consiste à hasarder une hypothèse raisonnable sur ces erreurs de mesure : on supposera donc que les ε_i sont des variables aléatoires, indépendantes, de moyenne nulle (c'est-à-dire qu'il n'y a pas d'erreur systématique dans un sens ou dans l'autre), et d'ampleur à peu près identique. Une façon statistiquement propre de formaliser ceci est donc de supposer que les ε_i sont des variables aléatoires gaussiennes iid, de moyenne 0 et de variance donnée, égale à un certain σ^2. C'est le *modèle linéaire gaussien* unidimensionnel.

Exercice 8.1. On a vu que $\hat{a} = \mathrm{Cov}(x,y)/\sigma_x^2$. Montrer qu'on a aussi l'expression suivante :

$$\hat{a} = a + \frac{\sum_{i=1}^n (x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

(La solution est dans le cours, un peu plus loin).

Le cas multidimensionnel fonctionne de la même façon. Supposons que les variables explicatives soient $\mathrm{X}^1, \ldots, \mathrm{X}^d$. Dans ce cas, on suppose que le modèle prend la forme

$$y_i = a_0 + a_1 x_i^1 + \cdots + a_d x_i^d + \varepsilon_i, \tag{32}$$

où les ε_i sont des variables aléatoires iid $\mathcal{N}(0,\sigma^2)$.

On insiste sur le fait que dans les modèles linéaires gaussiens, les variables x_i ne sont forcément considérées comme des variables aléatoires. Lorsqu'elles le sont, on suppose systématiquement qu'elles sont indépendantes des erreurs ε_i, ce qui permet de les traiter comme des données extérieures au problème.

8.1 — Pour bien comprendre : le cas unidimensionnel

Dans un modèle linéaire gaussien, il est possible d'expliciter facilement de nombreuses propriétés mathématiques des estimateurs $\hat{a},\hat{b}$. Le résultat général (et à connaître) sera énoncé un peu plus tard dans le théorème 8.6. Dans ce petit paragraphe, on effectue quelques calculs qui devraient permettre de convaincre le lecteur que ce résultat est assez intuitif. Il est vivement conseillé de vérifier que les calculs ci-dessous n'ont aucun mystère pour vous, même s'il n'est pas nécessaire de les savoir.

Plaçons-nous donc dans le cadre du modèle linéaire gaussien unidimensionnel (31), c'est-à-dire lorsque x_i, y_i sont des nombres réels et que $y_i = ax_i + b + \varepsilon_i$ avec les ε_i des gaussiennes iid centrées et de variance σ^2, et tentons de comprendre la loi des estimateurs $\hat{a},\hat{b}$. D'abord, on calcule la covariance empirique :

$$\begin{aligned}
\mathrm{Cov}(x,y) &= \frac{1}{n}\sum (x_i - \overline{x_n})(y_i - \overline{y_n}) \\
&= \frac{1}{n}\sum (x_i - \overline{x_n})(ax_i + b + \varepsilon_i - a\overline{x_n} - b - \overline{\varepsilon_n}) \\
&= \frac{1}{n}\sum (x_i - \overline{x_n})(a(x_i - \overline{x_n}) + \varepsilon_i - \overline{\varepsilon_n}) \\
&= \frac{a}{n}\sum (x_i - \overline{x_n})^2 + \frac{1}{n}\sum (x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}) \\
&= a\sigma_x^2 + \frac{1}{n}\sum (x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}).
\end{aligned}$$

Ainsi, en revenant à la définition de $\hat{a}$, on obtient que

$$\hat{a} = a + \frac{1}{n\sigma_x^2}\sum(x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n}).$$

Prenons l'espérance : on rappelle que les erreurs ε_i sont totalement indépendantes du problème. Ainsi, la linéarité de l'espérance montre que

$$\begin{aligned}\mathbf{E}[\hat{a}] &= \mathbf{E}[a] + \mathbf{E}\left[\frac{1}{n\sigma_x^2}\sum(x_i - \overline{x_n})(\varepsilon_i - \overline{\varepsilon_n})\right] \\ &= a + \frac{1}{n\sigma_x^2}\sum(x_i - \overline{x_n})\mathbf{E}\left[\varepsilon_i - \overline{\varepsilon_n}\right]. \end{aligned} \tag{33}$$

Or, comme les erreurs ε_i sont d'espérance nulle, l'espérance de leur moyenne $\overline{\varepsilon}_n$ est également nulle, et on a $\mathbf{E}\left[\varepsilon_i - \overline{\varepsilon_n}\right] = \mathbf{E}\left[\varepsilon_i\right] - \mathbf{E}\left[\overline{\varepsilon_n}\right] = 0 - 0 = 0$. Par conséquent,

$$\mathbf{E}[\hat{a}] = a, \tag{34}$$

donc l'estimateur $\hat{a}$ est sans biais. On en déduit facilement que $\hat{b}$ est également sans biais. En fait, il est possible de court-circuiter les calculs faits-ci dessus à partir de (33), puisque cette identité montre que $\hat{a} - a$ est simplement une combinaison linéaire de variables gaussiennes centrées ... donc c'est elle-même une variable gaussienne centrée ! Ainsi, $\hat{a}$ est une variable gaussienne de moyenne a, et possédant une certaine variance. Le calcul de cette dernière n'est pas exigible, mais il n'est pas difficile non plus. Pour s'entraîner, il est vivement recommandé d'essayer de faire ce calcul dans le cas unidimensionnel ci-dessus. Le résultat est :

$$\mathrm{Var}(\hat{a}) = \frac{1}{n}\frac{\sigma^2}{\sigma_x^2}. \tag{35}$$

On constate plusieurs choses : d'abord, la variance est d'autant plus grande que la variance empirique des données explicatives, σ_x^2, est petite. En effet, lorsque σ_x^2 est petite, cela signifie que les x_i sont toutes très proches, et donc le nuage de points (x_i, y_i) ne sera pas très étalé. Cela rend plus difficile l'approximation affine. Ensuite, si la variance σ_x^2 est « stable » avec la taille du problème, c'est-à-dire si σ_x^2 converge vers une valeur non nulle, alors la variance de $\hat{a}$ tend vers zéro, et donc l'estimateur $\hat{a}$ est convergent.

Exercice 8.2. On se place dans un modèle linéaire gaussien $y = ax + 1\varepsilon$ où la variance des erreurs est égale à 10. On dispose de $n = 100$ observations (y_i, x_i) et des données suivantes :

$$\bar{x} = 30 \qquad \bar{y} = 2 \qquad \sigma_x = 2 \qquad \sigma_y = 5 \qquad \frac{1}{n}\sum x_i y_i = 20.$$

Estimer a et proposer un test de niveau 0.99 de l'hypothèse nulle « $a = 0$ » contre l'hypothèse alternative « $a = 1$ ». Calculer la puissance de ce test.

8.2 — Résidus

Une fois que l'on a estimé les paramètres a, b par $\hat{a}, \hat{b}$, on peut comparer à quel point la prédiction $\hat{y}_i = \hat{a}x_i + \hat{b}$ est éloignée de sa véritable valeur y_i.

Définition 8.3. — *Le résidu associé à l'observation i et aux estimateurs $\hat{a}, \hat{b}$ est $\hat{\varepsilon}_i = \hat{y}_i - y_i$. La somme des carrés des résidus (SCR) est*

$$\mathrm{SCR} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2. \tag{36}$$

Il est possible de démontrer que les ε_i sont en réalité des variables aléatoires gaussiennes, mais elles sont dépendantes. Dans tous les cas, la somme de leurs carrés suit une loi du chi-deux : plus précisément, SCR/σ^2 suit une loi du chi-deux à $n-2$ degrés de liberté. Ce résultat n'est pas mentionné dans le programme de l'agrégation, mais il peut être salvateur de le retenir.

Exercice 8.4. Démontrer que $\mathrm{SCR}/(n-2)$ est un estimateur sans biais de la variance σ^2.

Exercice 8.5. On note $s^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$, et $\mathrm{SCT} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ (la somme des carrés totale). Montrer que

$$\mathrm{SCT} = s^2 + \mathrm{SCR}.$$

8.3 — Le modèle linéaire gaussien : comportement statistique

Les calculs de la section précédente peuvent être faits dans le cadre général du modèle gaussien multidimensionnel, et les résultats sont résumés dans le théorème suivant, qu'il faut absolument connaître.

THÉORÈME 8.6 (Modèle linéaire gaussien)**. —** *On se place dans le modèle linéaire gaussien multidimensionnel, avec des variables explicatives* X^t*, une variable expliquée* Y*, et des erreurs* σ^2 *:*

$$y_i = a_0 + \sum_{k=1}^{d} a_k x_i^k + \varepsilon_i$$

où $a_0, a_1, \dots, a_d$ *sont des paramètres inconnus et où les* ε_i *sont iid de loi* $\mathcal{N}(0, \sigma^2)$*. On suppose que l'on dispose de plus de* d *observations, et que les variables explicatives ne sont pas colinéaires.*

Alors, les coefficients de régression linéaire multiple $\hat{a}_0, \hat{a}_1, \dots, \hat{a}_d$ *fournis par la méthode des moindres carrés 7.4 sont des estimateurs sans biais des paramètres* $a_0, a_1, \dots, a_d$ *correspondants. Sous des hypothèses supplémentaires disant que les variables explicatives « conservent de la variance » lorsque* $n \to \infty$*, ces estimateurs sont également convergents.*

Par ailleurs, $\hat{a}_0, \dots, \hat{a}_d$ *suivent chacun une loi normale. Pour tout* k*, la variance de* $\hat{a}_k$ *est d'autant plus forte que la corrélation entre la* k*-ème variable explicative et les autres est elle-même forte.*

Dans l'énoncé, on mentionne cette hypothèse mystérieuse sur le fait que les variables explicatives doivent « conserver de la variance » lorsque le nombre d'observations n est grand. C'est en réalité le même phénomène que celui expliqué en dessous de l'équation (35) : quand on calcule la variance de $\hat{a}_k$, on obtient une expression (compliquée), qui dépend notamment de *l'inverse* de la variance des x_i^k et de *l'inverse* la covariance avec les autres variables explicatives. Il est donc nécessaire que ces variances et covariances ne tendent pas vers zéro, et c'est à peu près ce que l'on entend par « conserver de la variance ». Notons que le dernier point du théorème 8.6 est tout à fait intuitif : si X^i est fortement corrélée avec les autres variables, il est d'autant plus difficile d'estimer précisément son effet pur à partir d'un jeu de données fixé.

8.4 — Lecture des résultats logiciels et tests de Student

La principale leçon du théorème 8.6 est que, dans le modèle linéaire gaussien, il est possible d'expliciter la loi des estimateurs $\hat{a}_k$, et que cette loi est une loi gaussienne. À partir de cela, on peut donc construire toute une batterie de tests statistiques, par exemple pour quantifier à quel point on peut dire qu'un coefficient a_k est grand (ce qui signifie que la variable X^k explique une grosse partie de la variable expliquée Y), ou au contraire nul (ce qui signifie que la variable X^k n'a pas d'influence sur la variable expliqée et que l'on peut donc considérer les deux comme peu dépendantes selon ce modèle).

Le test le plus important est le test de Student de significativité d'un coefficient. L'hypothèse qui est testée est la suivante :

Le coefficient a_k est nul. (H0)

La statistique de test, appelée comme toujours « le t de Student », suit une loi de Student. Elle est donnée par les logiciels de statistique, en règle générale avec la p-valeur — on en profite pour rappeler que plus celle-ci est faible, plus il semble légitime de rejeter l'hypothèse nulle. Il n'est pas nécessaire de connaître la démonstration du résultat suivant, mais l'intuition est la même que celle de tous les tests de Student : « loi normale sur variance empirique = loi de Student ». Évidemment, il faut estimer la variance de $\hat{a}$, qui n'est pas connue en général ; cependant, la formule (35) semble indiquer que l'estimation $\sigma_{\hat{a}} = \mathrm{SCR}/(n(n-2)\sigma_x^2)$ est une bonne estimation de la variance. Lorsqu'il y a plusieurs variables, des estimateurs analogues $\sigma_{\hat{a}_k}$ existent.

PROPOSITION 8.7. — *La variable aléatoire* $T_k = \hat{a}_k / \sigma_{\hat{a}_k}$ *suit une loi de Student à* $n-2$ *degrés de libertés, notée* $\mathcal{T}_{n-2}$*. En conséquent, le test*

$$\Phi = \mathbf{1}_{|T_k| > \Phi^{-1}_{\mathcal{T}_{n-2}}(1-\alpha/2)}$$

est un test de niveau $1-\alpha$ *de l'hypothèse* (H_0)*.*

Par exemple, la figure 10 présente un détail d'une sortie de régression linéaire correspondant au modèle de régression linéaire simple

$$\texttt{score} = a \cdot \texttt{age} + b + \varepsilon$$

où la variable `score` représente la note obtenue par un professeur [30] et `age` l'âge dudit professeur.

```
> lm(formula = score ~ age, data = evals)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9185 -0.3531  0.1172  0.4172  0.8825

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.461932   0.126778  35.195   <2e-16 ***
age         -0.005938   0.002569  -2.311   0.0213 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5413 on 461 degrees of freedom
Multiple R-squared:  0.01146,   Adjusted R-squared:  0.009311
F-statistic: 5.342 on 1 and 461 DF,  p-value: 0.02125
```

FIGURE 10 – Extrait de sortie de la régression linéaire décrite ci-dessus (logiciel R).

Voici une explication de ce qu'on lit sur ce tableau.

1. La partie « Résidus » (la moins importante) donne quelques indications sur... les résidus, c'est-à-dire l'erreur $\hat{\varepsilon}_i = y_i - (\hat{a}x_i + \hat{b})$. On y lit la plus grande erreur, la plus petite erreur, les quartiles Q_1, Q_3 et la médiane.
2. Les données importantes sont celles relatives aux coefficients. Le terme (intercept) représente ici la constante.
3. La colonne *estimate* donne l'estima*tion* des coefficients de la régression linéaire. Ici, l'estimation du coefficient a devant la variable `age` est $-0,005938$.
4. La colonne suivante contient **les estimations des écarts-types $\sigma_{\hat{a}}$ et $\sigma_{\hat{b}}$ prédits par le modèle**.
5. La troisième colonne contient la **valeur d'une statistique de test de référence**, notée t dont on a parlé plus haut.
6. Le coefficient en première ligne de la quatrième colonne est d'un intérêt capital : **il s'agit de la p-valeur** [31] **du test de l'hypothèse** $H_0 : a = 0$ **contre l'hypothèse** $H_1 : a \neq 0$, c'est-à-dire de l'hypothèse selon laquelle le coefficient liant la qualité d'un professeur à son âge est nul. La p-valeur n'est pas si faible : elle est proche de 2%. Dans ces conditions, il paraît assez hardi de rejeter l'hypothèse nulle, qui s'interprète peu ou prou comme ceci : « la qualité d'un professeur dépend de son année de naissance ». Les p-valeurs, ici, sont glorieusement ornées d'étoiles qui expriment leur significativité, pas d'étoile si p est trop grande, trois étoiles si p est remarquablement petite.
7. Tous les logiciels de statistiques ont leurs propres sorties, et ici on voit (panel du bas) plusieurs indicateurs supplémentaires ; le plus important est le r^2 (adjuste r-square), que nous commentons dans la section suivante et qui est très important.

8.5 — L'approximation linéaire est-elle efficace ?

Les grandeurs que nous venons de présenter ne permettent pas d'évaluer la capacité du modèle linéaire à rendre compte des relations existant entre les différentes variables. Pourtant, il est clair que si le modèle linéaire est adapté pour expliquer, par exemple, le niveau de consommation d'un individu par son revenu, il l'est beaucoup moins lorsque l'effet d'un accroissement marginal des différentes variables explicatives sur la variable expliquée n'est pas constant (comme dans le cas du niveau d'étude et du revenu), et encore moins lorsqu'il n'est pas de signe constant (comme dans celui du niveau d'imposition et de la recette fiscale) ! Un coup d'oeil à la figure 9 permet de se rendre compte du fait que l'approximation d'un nuage de points (x_i, y_i) du plan par une droite d'équation $y = ax + b$ est bonne si et seulement si les résidus $y_i - ax_i - b$ de la régression sont petits par rapport à l'ordre de

30. Dans certaines universités étrangères, les étudiants peuvent noter leurs professeurs à la fin de chaque trimestre.
31. Prenons une pause bien méritée, et relisons calmement la section 5.5 page 25 sur les p-valeurs.

grandeur des variations des y_i. Un indicateur couramment utilisé pour connaître la qualité d'une régression est le R^2 (« R-deux ») défini, dans le cas d'une régression univariée, par

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - ax_i - b)^2}{\sum_{i=1}^{n}(y_i - \overline{y_n})^2}$$

Cette définition se généralise sans peine au cas de plusieurs variables explicatives. R^2 s'interprète comme la part de la variance des y_i expliquée par la variance des x_i à travers le modèle linéaire $y = ax + b$. On a toujours $R^2 \in [0,1]$; lorsque le R^2 est proche de 1, l'approximation linéaire de la relation entre x et y est excellente et la régression fournit une description synthétique et fiable des données, ce qui rend légitime une démarche de prédiction.

Exercice 8.8. Montrer que l'on a également $R^2 = \rho_{x,y}^2$ où $\rho_{x,y}$ est le coefficient de corrélation.

Cette observation est compatible avec le fait que $\rho_{x,y}$ mesure le degré de dépendance linéaire de y par rapport à x. En pratique, il est de bon ton de regarder le R^2 d'une régression *avant* d'en commenter les résultats : un R^2 faible indique en effet que l'approximation linéaire est mauvaise, soit parce que la relation de dépendance linéaire n'est pas crédible, soit parce qu'il manque des variables explicatives capables de capturer une plus grande quantité de la variance de la variable expliquée.

Le R^2 que l'on lit dans la figure 10 est particulièrement faible. Cela ne signifie pas forcément que la variable age n'est pas corrélée au score obtenu par le professeur... c'est peut-être le cas, mais il se peut tout aussi bien qu'une corrélation très forte existe entre les deux variables, une corrélation qui ne serait absolument pas linéaire [32].

Exercice 8.9. Comment calculer la somme des carrés des résidus en connaissant le R^2 ?

32. Penser à une relation quadratique de type $y = x^2$, par exemple.

9

Réduction de la dimension : l'analyse en composantes principales

9.1 — L'analyse en composantes principales

L'ACP est une technique de **réduction de la dimension**. Le point de départ est un ensemble d'observations, où chaque observation dispose de nombreuses modalités, disons d. Dans l'exemple que nous allons développer dans cette partie, on dispose de 27 observations correspondant à des athlètes : chaque observation est l'ensemble des performances d'un athlète dans une discipline olympique (100 mètres, triple saut, etc.), et $d = 10$ disciplines olympiques ont été mesurées. On notera

$$x_i = (x_i^1, \dots, x_i^d)$$

l'observation des performances de l'athlète i, de sorte que nos données sont un nuage de $n = 27$ points qui vivent dans l'espace $\mathbb{R}^{10}$.

	100m	Long.jump	Shot.put	High.jump	400m	110m.hurdle	Discus	Pole.vault	Javeline	1500m
SEBRLE	11.04	7.58	14.83	2.07	49.81	14.69	43.75	5.02	63.19	291.70
CLAY	10.76	7.40	14.26	1.86	49.37	14.05	50.72	4.92	60.15	301.50
KARPOV	11.02	7.30	14.77	2.04	48.37	14.09	48.95	4.92	50.31	300.20
BERNARD	11.02	7.23	14.25	1.92	48.93	14.99	40.87	5.32	62.77	280.10
YURKOV	11.34	7.09	15.19	2.10	50.42	15.31	46.26	4.72	63.44	276.40
WARNERS	11.11	7.60	14.31	1.98	48.68	14.23	41.10	4.92	51.77	278.10
ZSIVOCZKY	11.13	7.30	13.48	2.01	48.62	14.17	45.67	4.42	55.37	268.00
McMULLEN	10.83	7.31	13.76	2.13	49.91	14.38	44.41	4.42	56.37	285.10
MARTINEAU	11.64	6.81	14.57	1.95	50.14	14.93	47.60	4.92	52.33	262.10
HERNU	11.37	7.56	14.41	1.86	51.10	15.06	44.99	4.82	57.19	285.10
BARRAS	11.33	6.97	14.09	1.95	49.48	14.48	42.10	4.72	55.40	282.00
NOOL	11.33	7.27	12.68	1.98	49.20	15.29	37.92	4.62	57.44	266.60
BOURGUIGNON	11.36	6.80	13.46	1.86	51.16	15.67	40.49	5.02	54.68	291.70
SEBRLE	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.00	70.52	280.01
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.90	69.71	282.00
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.60	55.54	278.11
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.40	58.46	265.42
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.90	55.39	278.05
Zsivoczky	10.91	7.14	15.31	2.12	49.40	14.95	45.62	4.70	63.45	269.54
Hernu	10.97	7.19	14.65	2.03	48.73	14.25	44.72	4.80	57.76	264.35
Nool	10.80	7.53	14.26	1.88	48.81	14.80	42.05	5.40	61.33	276.33
Bernard	10.69	7.48	14.80	2.12	49.13	14.17	44.75	4.40	55.27	276.31
Schwarzl	10.98	7.49	14.01	1.94	49.76	14.25	42.43	5.10	56.32	273.56
Pogorelov	10.95	7.31	15.10	2.06	50.79	14.21	44.60	5.00	53.45	287.63
Schoenbeck	10.90	7.30	14.77	1.88	50.30	14.34	44.41	5.00	60.89	278.82
Barras	11.14	6.99	14.91	1.94	49.41	14.37	44.83	4.60	64.55	267.09
smith	10.85	6.81	15.24	1.91	49.27	14.01	49.02	4.20	61.52	272.74

FIGURE 11 – Les données brutes sur lesquelles on va effectuer notre ACP. On constate par exemple que le dénommé Nool a effectué un très beau lancer de disque à 37,92 mètres. Si les noms apparaissent plusieurs fois, c'est qu'ils correspondent à des performances lors d'événements sportifs différents.

Il n'est évidemment pas possible de visualiser 10 dimensions simultanément. Nous allons donc chercher à représenter les observations en deux dimensions pour avoir un joli dessin [33] ; cela revient à trouver un sous-espace vectoriel de $\mathbb{R}^d$ de dimension 2, qui « représente le mieux » le nuage, et l'ACP trouve un tel espace vectoriel et permet de quantifier intégralement la qualité de cette représentation. Dans toute cette section, on va expliquer les étapes conduisant à trouver cette représentation optimale. Cependant, ces étapes sont faites exclusivement par des méthodes informatiques, et il n'est pas du tout question de savoir les mettre en oeuvre soi-même : ce serait absurde

33. Parfois, on trouve des représentations en 3 dimensions, mais il est assez rare que le gain donné par la troisième dimension soit significatif. On expliquera plus loin des critères disant quand il est plus intéressant de choisir trois dimensions plutôt que deux.

et inefficace. En revanche, comprendre les idées est essentiel pour savoir interpréter une ACP, car les principaux indicateurs de la qualité d'une ACP résultent précisément de la façon dont on la fait.

Dans toute la suite, on décrira le nuage de points par la matrice X dont les lignes sont les observations, autrement dit cette matrice possède n lignes (autant que d'observations) et d colonnes (autant que de variables observées) :

$$\mathrm{X} = \begin{pmatrix} x_1^1 & \dots & x_1^d \\ \vdots & & \vdots \\ x_n^1 & \dots & x_n^d \end{pmatrix} \in \mathscr{M}_{n,d}(\mathbb{R}).$$

Normalisation des données. Cette étape n'est pas très importante. Elle consiste à centrer toutes les colonnes (ce qui revient en fait à centrer tout le nuage de points), puis à les normaliser pour que la norme de chaque colonne soit égale à 1. On supposera dorénavant que c'est déjà le cas pour X.

La notion d'inertie. L'inertie du nuage est la somme des carrés des normes des colonnes, autrement dit

$$\sigma^2 = \sum_{i=1}^{n} |x_i|^2 = \sum_{i=1}^{n} (x_i^1)^2 + \dots + (x_i^d)^2$$

Rappelons que le nuage est centré : si tous les points étaient identiques, ils seraient tous égaux au centre de gravité du nuage, c'est-à-dire zéro, et l'inertie du nuage serait nulle. Ce qu'il faut retenir avec des pincettes, c'est que *plus l'inertie d'un nuage est faible, moins il est riche et complexe*. Tout l'enjeu des méthodes de réduction de la dimension consiste à réduire la dimension des données (donc, à les simplifier un peu), tout en préservant le plus possible leur inertie et leur richesse (donc, simplifier mais pas trop). Il arrive fréquemment que l'on parle de *variance* au lieu d'inertie ; même si σ^2 n'est pas la variance d'une variable aléatoire (il n'y a aucune probabilité dans une ACP !), elle s'interprète de la même façon comme un indicateur de dispersion des données.

Recherche du premier axe. On passe maintenant à la recherche des axes optimaux. Un axe est une droite vectorielle. Elle possède donc un vecteur directeur, disons u, de longueur 1. On note $p_u(x)$ la projection d'un point $x \in \mathbb{R}^d$ sur cet espace vectoriel. Étant donné u, si l'on projetait tout le nuage sur cette droite vectorielle, on obtiendrait un nouveau nuage formé des points

$$p_u(x_1), \dots, p_u(x_n).$$

Il est possible de démontrer que lorsqu'on projette un nuage sur un sous-espace vectoriel [34], l'inertie ne peut que diminuer. On ne peut pas rendre un nuage plus complexe en réduisant sa dimension. Autrement dit, la nouvelle inertie du nuage projeté,

$$\sum_{i=1}^{n} |p_u(x_i)|^2,$$

est plus petite que σ^2. Mais à quel point ? La réponse dépendra évidemment de u, et tout l'art de l'ACP consiste à trouver la direction u telle que l'inertie l'inertie du nouveau nuage est la plus proche possible de la vraie inertie, σ^2 ce qui revient à trouver la direction u d'inertie maximale, que nous appellerons u_1.

Un point de vocabulaire. Ce nouvel axe u_1 est souvent appelé « composante principale » du nuage, et parfois « nouvelle variable » : on verra plus loin qu'il peut s'interpréter comme la construction d'une nouvelle variable, qui remplace les anciennes variables (ici, les variables sont les performances au triple saut, 100 mètres, etc.).

Valeurs propres. L'inertie du nuage projeté sur le premier axe est souvent appelée première valeur propre. Nous la noterons σ_1^2. Elle est plus petite que l'inertie initiale du nuage ; plus elle est proche de l'inertie initiale, meilleure est la projection. Pour cette raison, les logiciels donnent systématiquement le rapport entre les deux,

$$\frac{\sigma_1^2}{\sigma^2}$$

qui est une quantité entre zéro et 1. Cette quantité est souvent dénommée *la part de variance expliquée par le nouvel axe* u_1. Il est possible de démontrer que lorsque cette quantité est égale à 1, cela signifie que toutes les observations sont en fait « essentiellement les mêmes » : dans notre cas, cela signifierait que les performances de tous les athlètes sont proportionnelles les unes avec les autres.

34. Et pas forcément une droite vectorielle, d'ailleurs.

Cosinus carré. Posons-nous la question suivante : à quel point l'observation i est-elle fidèlement représentée lorsqu'on la projette sur le premier axe u_1 ? On dispose d'un bon indicateur pour cela, et il répond au doux nom de *cosinus carré*. Un cosinus carré proche de 1 implique que pour l'observation en question, aucune information n'a été perdue durant la projection. Un cosinus carré proche de zéro indique que cette observation était « orthogonale » à cette nouvelle variable u_1. Les cosinus carrés sont souvent donnés dans les sorties des logiciels ; une représentation habile en sera faite dans la figure 12.

Itérations. Il est rare que la projection du nuage sur la droite vectorielle dirigée par u_1 soit exceptionnellement bonne (disons, que la part de variance expliquée ci-dessus soit supérieure à 95%). Il faut donc continuer à simplifier le nuage... ou plutôt ce qu'il reste à simplifier. La projection $p_{u_1}(x_i)$ est la projection de l'observation x_i sur le premier axe, et il est probable que son cosinus carré ne soit pas très proche de 1. Que reste-t-il donc à expliquer sur cette observation ? Eh bien, tout simplement le *résidu*,

$$x_i - p_{u_1}(x_i).$$

On notera x'_i ce résidu. Il s'interprète comme *la part de l'observation i qui n'a pas du tout été expliquée par le premier axe*. On va donc recommencer exactement comme ci-dessus, mais en cherchant le meilleur axe possible pour réduire la dimension du nuage des x'_i. Évidemment, il faut chercher un axe qui n'ait rien à voir avec u_1, c'est-à-dire que le nouvel axe u_2 devra être « décorrélé » du premier axe [35]. Accessoirement, le nouveau nuage x'_i a pour inertie $\sigma^2 - \sigma_1^2$, ce qui est tout à fait conforme à l'intuition. L'inertie du nuage projeté sur le second axe sera notée σ_2^2, et il est possible de montrer qu'elle est plus petite que σ_1^2. C'est bien normal, car le premier axe était celui qui devait « conserver le plus d'inertie ».

La part cumulative de variance expliquée. En continuant de la sorte jusqu'à épuisement du nuage, ce qui survient au plus après d étapes, on obtient une suite d'axes u_k, qui sont tous décorrélés (orthogonaux). À chacun de ces axes est associée une valeur propre σ_k^2 de plus en plus petite, qui correspond à la part de la variance de σ^2 expliquée par cet axe. Évidemment, la somme de ces variances est égale à σ^2. Ce que l'on appelle la *variance cumulée* des k premières composantes est la proportion

$$v_k = \frac{\sigma_1^2 + \cdots + \sigma_k^2}{\sigma^2}$$

qui correspond donc à la part d'inertie expliquée par les k premiers axes. Cette quantité est d'une importance capitale, et elle est systématiquement donnée par les sorties des logiciels. En effet, dans la pratique, elle permet de déterminer le nombre de dimensions sur lequel on va projeter notre nuage, c'est-à-dire deux ou trois en général. La quantité v_2 donne donc la qualité de la réduction en dimension 2, et v_3 la qualité de la réduction en dimension 3. Il est clair qu'avec trois dimensions, on explique plus de choses qu'en deux dimensions (c'est-à-dire $v_3 \geqslant v_2$). Mais à quel point ? Souvent, l'ajout d'une troisième dimension n'augmente presque pas la part de variance expliquée : cela veut dire que le nuage initial, même s'il possède d dimensions, ne dépend intrinsèquement que de *deux dimensions*. Une v_2 faible est signe d'une trop grande complexité du nuage : il ne se laisse pas facilement réduire à deux dimensions, et toute tentative de le faire se fera au détriment d'une trop grande simplification de sa richesse.

Les valeurs propres et les parts cumulées de la variance expliquée, pour notre ACP sur les données sportives, sont les suivantes :

```
> get_eigenvalue(PCA(decathlon.active, graph = FALSE))

       eigenvalue variance.percent cumulative.variance.percent
Dim.1   3.7550471        37.550471                    37.55047
Dim.2   1.6792071        16.792071                    54.34254
Dim.3   1.2922790        12.922790                    67.26533
Dim.4   1.0470192        10.470192                    77.73552
Dim.5   0.7433940         7.433940                    85.16947
Dim.6   0.5489725         5.489725                    90.65919
Dim.7   0.3268999         3.268999                    93.92819
Dim.8   0.2761597         2.761597                    96.68979
Dim.9   0.1950251         1.950251                    98.64004
Dim.10  0.1359963         1.359963                   100.00000
```

35. En réalité, c'est automatique. Le second axe optimal sera toujours « décorrélé » (en langage algébrique, *orthogonal*) du premier.

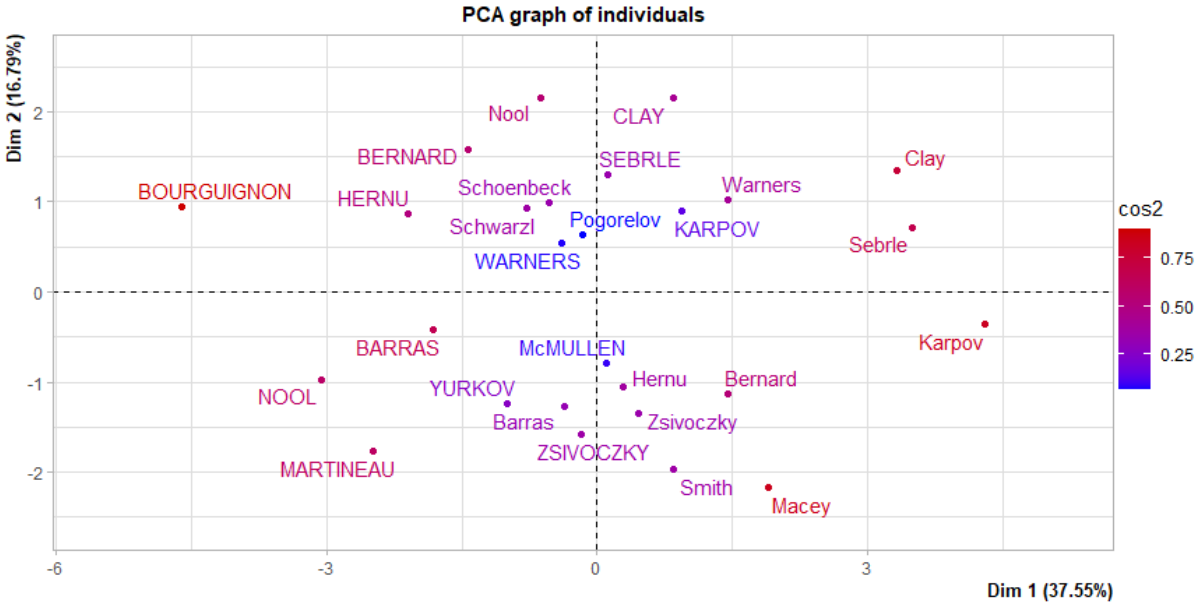


FIGURE 12 – Dans cette représentation de l'ACP pour les données sportives ci-dessus, on a également représenté les cosinus carrés. Plus un athlète est rouge, plus son cosinus carré est élevé, et donc plus sa représentation dans ces deux dimensions est bonne. Ici, les athlètes centraux semblent moins fidèlement représentés que les autres, mais ce n'est pas toujours le cas.

On constate par exemple que la première valeur propre représente $37,5\%$ de l'inertie totale du nuage, et que les deux premières valeurs propres cumulées expliquent $54,3\%$ de la variance. Ce n'est pas un score particulièrement élevé.

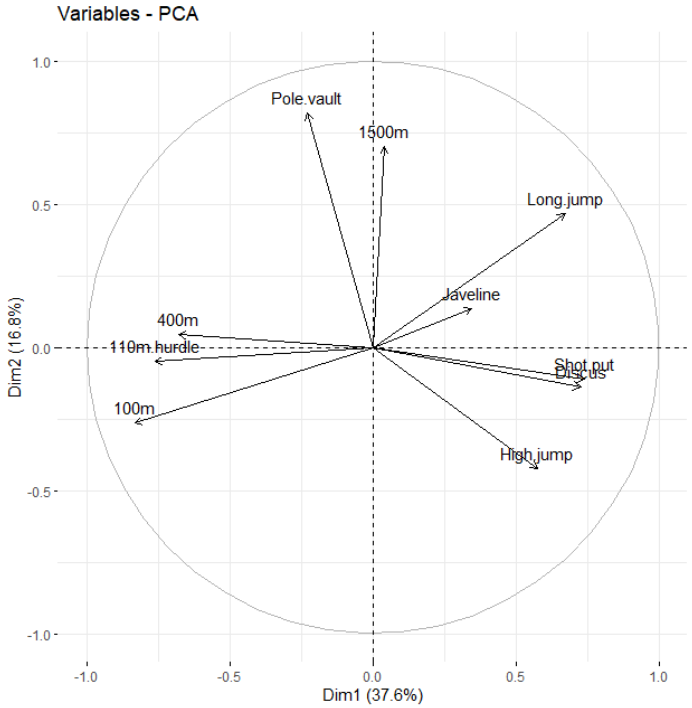
Le choix du nombre d'axes. Le choix du bon nombre d'axes pour représenter nos données se fait en regardant les v_k. Ici, le sociologue doit faire un choix : on dira qu'une représentation est satisfaisante si elle explique au moins 80% de la variance. Ce nombre de 80% est évidemment purement subjectif.

Comme les v_k sont croissants et que le dernier est égal à 1 (pourquoi ?), il existe un rang optimal, disons r, tel que v_r est plus grand que 80%, mais pas v_{r-1}. Dans la plupart des exercices que l'on vous posera, ce r sera égal à 2 ou à 3. Cependant, il faut savoir interpréter des cas exotiques.

— Si $r = 1$, cela signifie que le nuage est intrinsèquement de dimension 1, autrement dit... très pauvre. Ce cas est plutôt rare, car lorsqu'on fait une ACP sur un jeu de données, c'est certainement que le jeu de données présente un peu de complexité et de richesse.

— Si r est plus grand que 4, cela signifie a contrario que le jeu de données est très riche, et qu'une représentation en deux ou trois dimensions se fera au détriment d'une certaine simplification. Ce cas n'est pas forcément très rare. Il ne se prête pas à des représentations graphiques, donc l'interprétation sera moins aisée, mais elle peut être faite. Dans le cas des données sportives, on voit par exemple que si on voulait conserver au moins 80% de l'information, il faudrait projeter les données sur les 5 premiers axes...

Représentation graphique en deux dimensions. Nous arrivons (enfin) à ce qui a fait la renommée de l'ACP dans le multivers des sociologues français, à savoir les représentations graphiques des ACP comme on en croise notamment dans *La Distinction*. Supposons que les deux premiers axes expliquent une grande part de la variance du nuage. On peut alors représenter les données en dimension deux en les projetant sur l'espace vectoriel engendré par u_1, u_2. Si un point x_i est représenté par le couple (a,b), cela veut dire que a est la composante de x_i sur le premier axe, et b celle sur le second axe. Pour nos données sportives, la représentation sur les deux premiers axes est dans la figure 12.



De telles représentations ont surtout pour intérêt de visualiser rapidement le nuage de points, et d'aider à l'interprétation des nouveaux axes.

Interprétation des nouveaux axes et cercle des corrélations. Chaque nouvel axe u_k peut s'interpréter comme une nouvelle variable. Les anciennes variables sont, rappelons-le, les résultats de nos athlètes à certaines épreuves sportives. Chaque nouvelle variable u_k est en quelque sorte un *mélange* de ces anciennes variables, et le mélange est fait de sorte à conserver un maximum d'informations possibles. Tout l'art du sociologue consiste à interpréter ces axes, et cela doit se faire au vu et au su de tous les indicateurs nommés ci-dessus, et en particulier les valeurs propres et les cosinus carrés. En fait, comme le vecteur u_k est lui-même un vecteur dans un espace de dimension 10, il peut s'écrire en fonction des vecteurs de base e_j, disons

$$u_k = (u_k^1, \ldots, u_k^d).$$

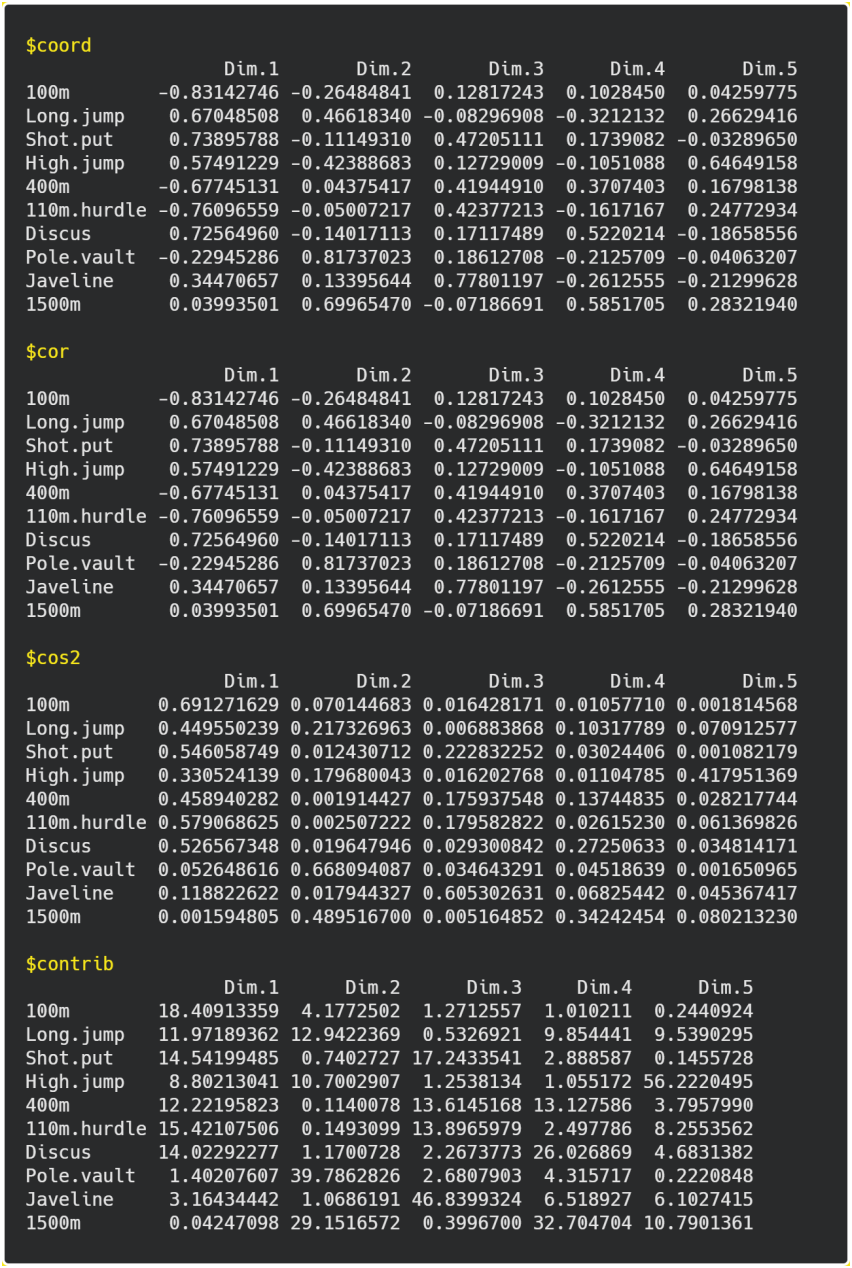
Or, le vecteur de base e_j correspond précisément à la variable j. Par exemple, lorsqu'on écrit l'observation i sous la forme d'un vecteur, disons $x_i = (x_i^1, \cdots, x_i^d)$, on dit que la composante de la variable $j = 2$ (le triple saut ici) est égale à x_i^2. Ainsi, le nombre u_k^j peut s'interpréter comme la corrélation entre le nouvel axe u_k et l'ancienne variable j, ou encore comme la quantité de l'ancienne variable j qui est présente dans la nouvelle variable u_k. Et il est possible d'obtenir les quantités réciproques, c'est-à-dire la quantité de la nouvelle variable u_k qui est présente dans l'ancienne variable j. Cela se visualise bien sur un cercle des corrélations, lorsque le nombre de composantes conservées est égal à 2. En abscisse et en ordonnées, sont représentés les deux axes principaux. Pour chaque ancienne variable j, on trace une flèche qui pointe vers « la quantité de la nouvelle variable qui est présente dans cette ancienne variable ». Deux remarques sur ce point.

(i) Plus la longueur de la flèche est grande (au maximum, elle est égale à 1), mieux cette ancienne variable est exprimée par les nouveaux axes. Réciproquement, si la flèche est très petite, cela veut dire que l'ancienne variable n'a presque pas été intégrée dans la construction des nouveaux axes.

(ii) Plus deux flèches sont proches, plus les deux variables qu'elles représentent sont proches du point de vue des deux premières composantes principales. Si la qualité de l'ACP est bonne, cela veut dire que ces deux variables sont en fait très fortement liées.

Enfin, il est possible de mesurer à quel point une ancienne variable est bien représentée dans le système des nouveaux axes, avec le même outil que pour les observations : il s'agit des cosinus carrés. On ne revient pas sur leur construction, mais l'interprétation est la même, plus le cosinus carré d'une variable est proche de 1, meilleure sera sa représentation.

Toutes les mesures de qualité de l'ACP relatives aux variables sont dans la sortie logiciel de la figure 13 ci-dessous.



```
$coord
                Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
100m        -0.83142746 -0.26484841  0.12817243  0.1028450  0.04259775
Long.jump    0.67048508  0.46618340 -0.08296908 -0.3212132  0.26629416
Shot.put     0.73895788 -0.11149310  0.47205111  0.1739082 -0.03289650
High.jump    0.57491229 -0.42388683  0.12729009 -0.1051088  0.64649158
400m        -0.67745131  0.04375417  0.41944910  0.3707403  0.16798138
110m.hurdle -0.76096559 -0.05007217  0.42377213 -0.1617167  0.24772934
Discus       0.72564960 -0.14017113  0.17117489  0.5220214 -0.18658556
Pole.vault  -0.22945286  0.81737023  0.18612708 -0.2125709 -0.04063207
Javeline     0.34470657  0.13395644  0.77801197 -0.2612555 -0.21299628
1500m        0.03993501  0.69965470 -0.07186691  0.5851705  0.28321940

$cor
                Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
100m        -0.83142746 -0.26484841  0.12817243  0.1028450  0.04259775
Long.jump    0.67048508  0.46618340 -0.08296908 -0.3212132  0.26629416
Shot.put     0.73895788 -0.11149310  0.47205111  0.1739082 -0.03289650
High.jump    0.57491229 -0.42388683  0.12729009 -0.1051088  0.64649158
400m        -0.67745131  0.04375417  0.41944910  0.3707403  0.16798138
110m.hurdle -0.76096559 -0.05007217  0.42377213 -0.1617167  0.24772934
Discus       0.72564960 -0.14017113  0.17117489  0.5220214 -0.18658556
Pole.vault  -0.22945286  0.81737023  0.18612708 -0.2125709 -0.04063207
Javeline     0.34470657  0.13395644  0.77801197 -0.2612555 -0.21299628
1500m        0.03993501  0.69965470 -0.07186691  0.5851705  0.28321940

$cos2
                Dim.1       Dim.2       Dim.3      Dim.4       Dim.5
100m        0.691271629 0.070144683 0.016428171 0.01057710 0.001814568
Long.jump   0.449550239 0.217326963 0.006883868 0.10317789 0.070912577
Shot.put    0.546058749 0.012430712 0.222832252 0.03024406 0.001082179
High.jump   0.330524139 0.179680043 0.016202768 0.01104785 0.417951369
400m        0.458940282 0.001914427 0.175937548 0.13744835 0.028217744
110m.hurdle 0.579068625 0.002507222 0.179582822 0.02615230 0.061369826
Discus      0.526567348 0.019647946 0.029300842 0.27250633 0.034814171
Pole.vault  0.052648616 0.668094087 0.034643291 0.04518639 0.001650965
Javeline    0.118822622 0.017944327 0.605302631 0.06825442 0.045367417
1500m       0.001594805 0.489516700 0.005164852 0.34242454 0.080213230

$contrib
                Dim.1      Dim.2      Dim.3     Dim.4      Dim.5
100m        18.40913359  4.1772502  1.2712557  1.010211  0.2440924
Long.jump   11.97189362 12.9422369  0.5326921  9.854441  9.5390295
Shot.put    14.54199485  0.7402727 17.2433541  2.888587  0.1455728
High.jump    8.80213041 10.7002907  1.2538134  1.055172 56.2220495
400m        12.22195823  0.1140078 13.6145168 13.127586  3.7957990
110m.hurdle 15.42107506  0.1493099 13.8965979  2.497786  8.2553562
Discus      14.02292277  1.1700728  2.2673773 26.026869  4.6831382
Pole.vault   1.40207607 39.7862826  2.6807903  4.315717  0.2220848
Javeline     3.16434442  1.0686191 46.8399324  6.518927  6.1027415
1500m        0.04247098 29.1516572  0.3996700 32.704704 10.7901361
```

FIGURE 13 – Tous les paramètres de l'ACP sur les données sportives qui sont relatifs aux variables. Les deux premiers panels sont identiques (une bizarrerie des sorties du logiciel R) et représentent les anciennes variables dans la base des nouveaux axes. Comme toutes ces variables sont normées, les cosinus carrés des variables sont simplement égaux aux carrés des coordonnées (troisième panel). Le dernier panel décrit la matrice inverse du premier panel, exprimée en pourcentages : pour chaque nouvel axe, on lit la proportion des anciennes variables qui a servi à le fabriquer. Les colonnes somment donc à 100.

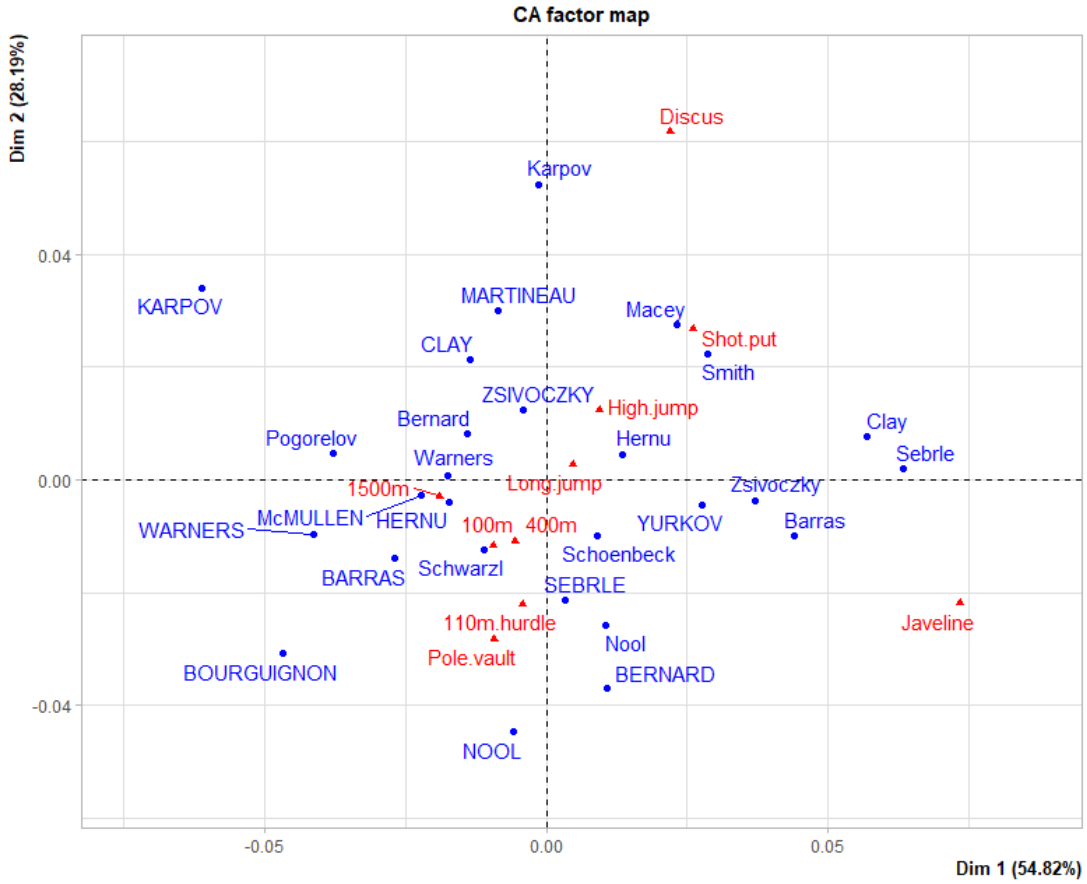


FIGURE 14 – Un exemple de sortie synthétique d'une analyse factorielle des correspondances sur le même jeu de données que pour l'ACP.

9.2 — L'analyse des correspondances en dimension 2

Dans la section précédente, on a effectué une ACP sur les observations x_i ; chaque observation correspondait au relevé des performances d'un athlète pour différentes épreuves. Seulement, on aurait pu au contraire considérer chaque épreuve comme une observation, et considérer que ce sont les athlètes qui forment les variables. Cela revient à transposer la matrice X, c'est-à-dire faire l'ACP non pas sur les lignes de X, mais sur les colonnes de X. **L'analyse factorielle des correspondances consiste à faire les deux simultanément et à comparer les résultats**. Comme pour l'ACP, aucune connaissance des mathématiques sous-jacentes n'est exigible ; il faut juste avoir foi en l'algèbre linéaire, et surtout, savoir interpréter les résultats.

On pratique une AFC lorsqu'on dispose d'un tableau de données avec deux modalités que l'on veut en quelque sorte comparer. **Le cadre de base est donc exactement le même que celui du chi-deux** (et on recommande vivement de relire la partie du cours sur les tests d'indépendance, c'est-à-dire la section 6.4 page 29) : avant de faire une AFC sur des données, il est nécessaire d'effectuer un test d'indépendance entre les deux modalités du tableau. Les sorties de logiciels pour une AFC donnent toujours la statistique du chi-deux : si le test ne permet pas de rejeter l'hypothèse d'indépendance, alors l'intérêt de faire une AFC n'est pas bien grand. Ce n'est que lorsqu'une liaison entre les deux variables est statistiquement acceptable que l'on désirera expliciter la nature de cette liaison, et l'AFC permet d'aider à le faire.

Intérêt et limites L'AFC permet d'interpréter le résultat d'un éventuel test du chi-deux lorsque celui-ci indique que des variables ne sont pas indépendantes. Elle permet de visualiser leur liaison et de trouver des « axes principaux » qui capturent à la fois beaucoup d'informations sur les lignes et sur les colonnes ; en ce sens, elle n'a d'intérêt que si la proportion de la variance expliquée par les axes est grande, mais ce sera toujours le cas si le test du chi-deux a été correctement effectué. En revanche, comme l'ACP, l'AFC n'est pas une procédure de statistique permettant de mesurer la significativité d'un modèle. Ce n'est qu'une méthode descriptive qui évite de choisir des

indicateurs a priori (moyenne, variance), et qui au contraire vous donne les meilleurs indicateurs possibles compte tenu des données.

Index