

Statistique

L. Rouvière

JANVIER 2012

Contents

I	Statistiques descriptives	2
1	Introduction	3
2	La statistique exploratoire	4
2.1	Etude d'une variable	5
2.2	Etude de deux variables	6
2.3	Etude de plus de deux variables	7
3	L'analyse en composantes principales	8
3.1	Quelques rappels d'algèbre linéaire	8
3.2	Introduction à l'ACP - Réduction de la dimension	10
3.3	Analyse du nuage des individus	12
3.3.1	Recherche des axes factoriels	14
3.3.2	Contributions et qualités de représentation	17
3.4	Analyse du nuage des variables	19
II	Théorie de l'estimation	22
1	Quelques exemples	23
2	Modèle statistique	25
3	Qualités d'un estimateur	26
3.1	Biais, variance et risque quadratique	26
3.2	Critère de performance asymptotique	29
3.3	2 méthodes d'estimation	30
3.3.1	La méthode des moments	30
3.3.2	La méthode du maximum de vraisemblance	31

4	Information de Fisher et Borne de Cramer Rao	32
4.1	Dimension 1	32
4.2	Dimension p	34
5	Estimation par intervalle de confiance	36
III	Tests d'hypothèses	38
1	Introduction	38
2	Tests paramétriques	40
2.1	Vocabulaire	40
2.2	Le principe de Neyman-Pearson	42
2.3	Puissance de test - Test UPP	44
2.4	Exemples	45
2.5	Comparaison de deux échantillons gaussiens	46
2.6	Cas non gaussien	49
3	Une introduction aux tests non paramétriques	51
3.1	Le test du χ^2 d'adéquation	52
3.2	Le test du χ^2 d'indépendance	54
IV	Le modèle de régression linéaire	56
1	Introduction	56
2	La régression linéaire simple	58
2.1	Ajustement par moindres carrés	58
2.2	Propriétés des estimateurs	60
2.3	Inférence statistique	62
3	La régression multiple	64
3.1	Notations et modélisation	64
3.2	Estimateur des moindres carrés	65
3.3	Propriétés statistiques	66
4	Validation et choix de modèles	68
4.1	Résidus et coefficient de détermination	68
4.2	Tests entre modèles emboîtés	70
5	Analyse de la variance	72
5.1	Modèle à un facteur	72
5.2	ANOVA à deux facteurs	74
5.3	Sélection de modèles	75

Part I

Statistiques descriptives

1 Introduction

Vocabulaire (voir Saporta : Probabilités, analyse des données et statistique)

- **Population** : ensemble d'objets de même nature.
- **Individu** : élément de cette population.
- **Variable** : caractéristique étudiée sur la population.
- **Echantillon** : sous ensemble de la population dont les individus feront l'objet de l'étude

Un exemple

- On s'intéresse aux performances de décathlons de haut niveau dans les 10 disciplines qui composent ce sport
- On dispose des performances de $n = 41$ athlètes réalisées au cours des JO et au décastar :

	X100m	Longueur	Poids	Hauteur	X400m	X110mH	Disque	Perche
Sebrle	10.85	7.84	16.36	2.12	48.36	14.05	48.72	5.0
Clay	10.44	7.96	15.23	2.06	49.19	14.13	50.11	4.9
Karpov	10.50	7.81	15.93	2.09	46.81	13.97	51.65	4.6
Macey	10.89	7.47	15.73	2.15	48.97	14.56	48.34	4.4
Warners	10.62	7.74	14.48	1.97	47.97	14.01	43.73	4.9

	Javelot	X1500m	Classement	Points	Competition
Sebrle	70.52	280.01	1	8893	OlympicG
Clay	69.71	282.00	2	8820	OlympicG
Karpov	55.54	278.11	3	8725	OlympicG
Macey	58.46	265.42	4	8414	OlympicG
Warners	55.39	278.05	5	8343	OlympicG

- **Population** : ensemble des décathlons de haut niveau.
- **Individu** : un décathlonien de haut niveau.
- **Variation** : performances dans chacune des 10 disciplines.
- **Echantillon** : les décathloniens ayant participé au JO ou au décastar.

La statistique exploratoire

- *But* : synthétiser, résumer, structurer l'information contenue dans un tableau de données.
- *Comment ?* : représentations sous forme de tableaux, de graphiques, d'indicateurs numériques.
- *Exemple* : calcul de la longueur moyenne sautée dans l'épreuve de saut en longueur...

La statistique inférentielle

- *But* : étendre les propriétés constatées sur l'échantillon à la population toute entière.
- *Comment ?* : les méthodes font généralement appel à la théorie des probabilités (construction d'intervalles de confiance, de tests d'hypothèses).
- *Exemple* : peut-on dire que les performances des athlètes sont meilleurs aux JO qu'au décastar ?

2 La statistique exploratoire

Typologie des variables

Les méthodes diffèrent selon le type de variables :

- *quantitative* : additionner les modalités à un sens.
 - **continue** : la variable prend ses valeurs dans un intervalle de \mathbb{R} (taille, poids, saut en longueur...);
 - **discrète** : nombre fini ou dénombrable de valeurs (nombre de personnes dans une file d'attente à un moment donnée, classement du décathlon...).
- *qualitative* : additionner les modalités n'a pas de sens.
 - **ordinaire** : relation d'ordre entre les modalités (type de mention à un examen);
 - **nominale** : sinon (type de traitement subi, CSP).

2.1 Étude d'une variable

Mesures de tendance centrale

On note x_1, \dots, x_n n observations d'une variable quantitative X .

- *moyenne* : $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
- *médiane* : "valeur qui coupe l'échantillon en 2". On la définit à partir de la fonction de répartition empirique

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{x_i \leq x\}}.$$

- *médiane* : plus petite valeur M telle que $F_n(x) \geq 0.5$:

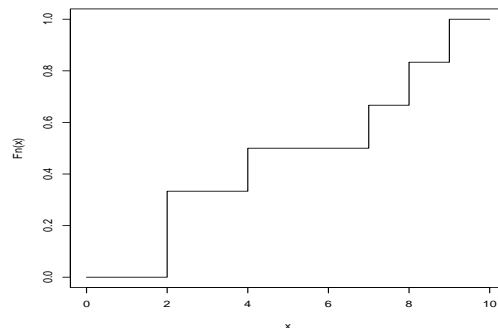
$$M = \inf\{x : F_n(x) \geq 0.5\}.$$

- *quantile d'ordre α* : plus petite valeur q_α telle que $F_n(x) \geq \alpha$:

$$q_\alpha = \inf\{x : F_n(x) \geq \alpha\}.$$

Exemple

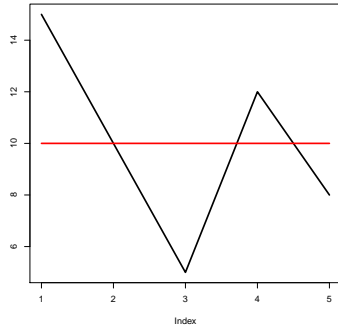
x_1	x_2	x_3	x_4	x_5	x_6
4	2	2	8	7	9



$$M = 4, \quad q_{0.25} = 2, \quad q_{5/6} = 8.$$

- Mesurer la tendance centrale n'est pas suffisant :

x_1	x_2	x_3	x_4	x_5	\bar{x}	M
10	10	10	10	10	10	10
15	10	5	12	8	10	10



• *Variance* :

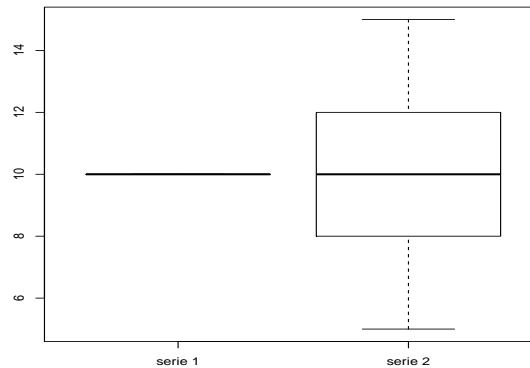
$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2.$$

• série 1 : $S_n^2 = 0$, série 2 $S_n^2 = \frac{1}{5}(5^2 + 0^2 + \dots) = 11.6$.

• *Conclusion* : les observations de la série 2 sont plus dispersées autour de leur moyenne.

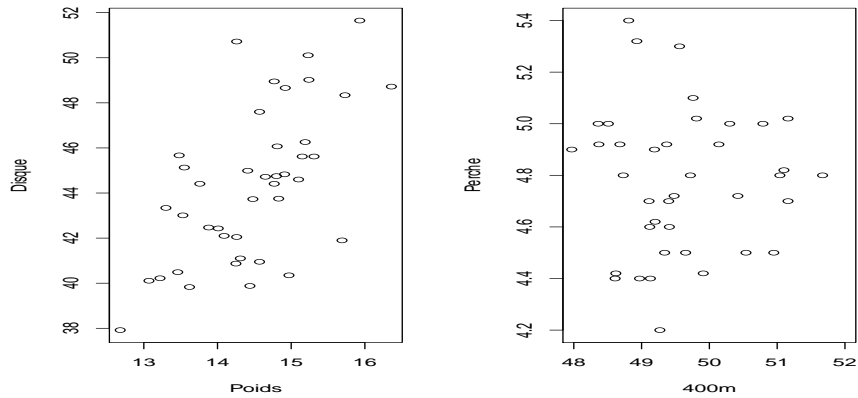
Boxplot

x_1	x_2	x_3	x_4	x_5	\bar{x}	M
10	10	10	10	10	10	10
15	10	5	12	8	10	10



2.2 Étude de deux variables

- On observe 2 variables quantitatives X et Y sur un échantillon de n individus. Les observations sont notées $(x_i, y_i), i = 1, \dots, n$.
- *Problème* : mesurer la relation entre X et Y .
- *Exemple* :



Mesure d'une relation linéaire

- *Définition*

- **covariance** entre X et Y :

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^n x_i y_i - \bar{x} \bar{y}.$$

- **corrélation** entre X et Y :

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(X)\sigma(Y)}.$$

- *Propriété*

- $-1 \leq \rho(X, Y) \leq 1$ et $|\rho(X, Y)| = 1$ si et seulement si il existe a et b tels que $y_i = ax_i + b$.

- Si $|\rho(X, Y)| \approx 1$ on dit que X et Y sont corrélées et si $|\rho(X, Y)| \approx 0$ on dit qu'elles sont non corrélées.

- *Exemple* : $\rho(\text{Poids}, \text{Disque}) = 0.62$ et $\rho(400\text{m}, \text{perche}) = -0.08$.

2.3 Étude de plus de deux variables

- Lorsque l'on cherche à étudier plus de deux variables simultanément, les choses se compliquent...

- Sur l'exemple du décathlon, on a $n = 41$ individus et $p = 10$ variables.

- *Questions* :

- peut-on regrouper certains individus selon leur performance ? on pourrait calculer les $n(n - 1)/2 = 820$ distances entre individus... difficile à analyser.
- peut-on identifier des groupes de variables (des disciplines pour lesquelles certains individus pourraient être très performants ou non) ? Une idée : utiliser la matrice des corrélations.

Un exemple

	X100m	Longueur	Poids	Hauteur	X400m	X110mH	Disque	Perche	Javelot	X1500m
X100m	1,00	-0,60	-0,36	-0,25	0,52	0,58	-0,22	-0,08	-0,16	-0,06
Longueur	-0,60	1,00	0,18	0,29	-0,60	-0,51	0,19	0,20	0,12	-0,03
Poids	-0,36	0,18	1,00	0,49	-0,14	-0,25	0,62	0,06	0,37	0,12
Hauteur	-0,25	0,29	0,49	1,00	-0,19	-0,28	0,37	-0,16	0,17	-0,04
X400m	0,52	-0,60	-0,14	-0,19	1,00	0,55	-0,12	-0,08	0,00	0,41
X110mH	0,58	-0,51	-0,25	-0,28	0,55	1,00	-0,33	0,00	0,01	0,04
Disque	-0,22	0,19	0,62	0,37	-0,12	-0,33	1,00	-0,15	0,16	0,26
Perche	-0,08	0,20	0,06	-0,16	-0,08	0,00	-0,15	1,00	-0,03	0,25
Javelot	-0,16	0,12	0,37	0,17	0,00	0,01	0,16	-0,03	1,00	-0,18
X1500m	-0,06	-0,03	0,12	-0,04	0,41	0,04	0,26	0,25	-0,18	1,00

On mesure les corrélations deux à deux mais difficile d'obtenir une information plus globale...

3 L'analyse en composantes principales

3.1 Quelques rappels d'algèbre linéaire

Projecteurs

- Soit E un espace vectoriel de dimension finie n muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ et F un sous-espace vectoriel de E de dimension p .

Définition

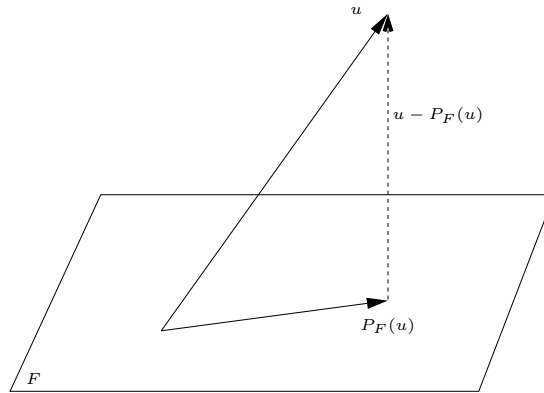
Un projecteur $p : E \rightarrow E$ est une application linéaire qui vérifie $p \circ p = p$.

Projection orthogonale

Projection orthogonale

P_F est la projection orthogonale sur F si

- $\forall u \in E, P_F(u) \in F$;
- $\forall u \in E, u - P_F(u) \in F^\perp$.



Propriétés

1. Soient $(u, v) \in E^2$. Le projeté orthogonal de u sur $F = \text{vect}(v)$ est donné par

$$P_F(u) = \frac{\langle u, v \rangle}{\|v\|^2} v.$$

2. Soit F un sev de E de dimension p et $\mathcal{B} = (v_1, \dots, v_p)$ une base orthogonale de F , alors

$$P_F(u) = \frac{\langle u, v_1 \rangle}{\|v_1\|^2} v_1 + \dots + \frac{\langle u, v_p \rangle}{\|v_p\|^2} v_p.$$

Soient F et G 2 sev orthogonaux de E . Alors

$$P_{F \oplus G} = P_F + P_G.$$

Valeurs propres - Vecteurs propres

Définition

Soit A une matrice $n \times n$.

- $v \in E$ est un vecteur propre de A si et seulement si il existe $\lambda \in \mathbb{R}$ tel que $Av = \lambda v$ (λ est appelé valeur propre de A).
- L'ensemble des vecteurs propres de A associé à la valeur propre λ est appelé **espace propre** E_λ :

$$E_\lambda = \ker(A - \lambda I).$$

Propriété

λ est valeur propre de A si et seulement si $\det(A - \lambda I) = 0$.

Diagonalisation

Définition

A est diagonalisable si il existe une matrice P inversible et une matrice D diagonale telles que $A = P^{-1}DP$.

Propriété

Soit A une matrice admettant pour valeurs propres $\lambda_1, \dots, \lambda_k$. La matrice A est diagonalisable si et seulement si la somme des dimensions des sous-espaces propres est égale à n , c'est-à-dire

$$\sum_{j=1}^k \dim(E_{\lambda_j}) = n.$$

Diagonalisation de matrices semi-définie positive

Propriété

Soit A une matrice symétrique semi-définie positive. Alors

1. A est diagonalisable ;
2. Les valeurs propres de A sont ≥ 0 ;
3. Les espaces propres de A sont deux à deux orthogonaux.

Propriété

Soit X une matrice $n \times p$. Alors

1. la matrice $X'X = \Sigma$ de dimension $p \times p$ est semi-définie positive.
2. la matrice $XX' = A$ de dimension $n \times n$ est semi-définie positive.
3. Σ et A ont les mêmes valeurs propres non nulles.

3.2 Introduction à l'ACP - Réduction de la dimension

Notations

- Tableau des données

$$X = \begin{matrix} & X_1 & \dots & X_p \\ e_1 & \left(\begin{matrix} x_{1,1} & \dots & x_{1,p} \\ \vdots & & \vdots \\ x_{n,1} & \dots & x_{n,p} \end{matrix} \right) \\ \vdots & & & \\ e_n & & & \end{matrix}$$

- $e_i = (x_{i,1}, \dots, x_{i,p})'$ l'individu i et $X_j = (x_{1,j}, \dots, x_{n,j})'$ la variable j .

- $e_i \in \mathbb{R}^p$, la représentation de l'ensemble des individus est un nuage de points dans \mathbb{R}^p , appelé *nuage des individus*, \mathcal{N} .
- $X_j \in \mathbb{R}^n$, la représentation de l'ensemble des variables est un nuage de points dans \mathbb{R}^n , appelé *nuage des variables*, \mathcal{M} .

Si l'œil était capable de visualiser dans \mathbb{R}^n et \mathbb{R}^p , il n'y aurait pas de problème...

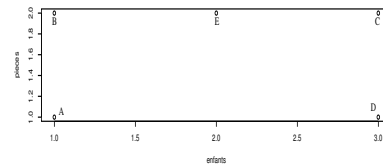
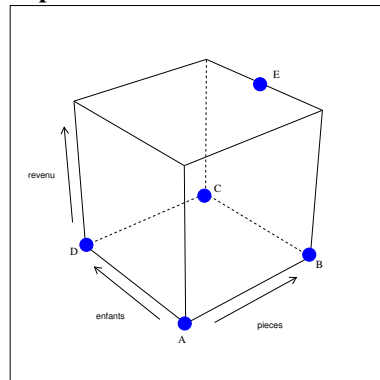
Objectifs

Déterminer un sous-espace de dimension réduite qui soit "compréhensible" par l'œil sur lequel projeter le nuage.

Un exemple "jouet"

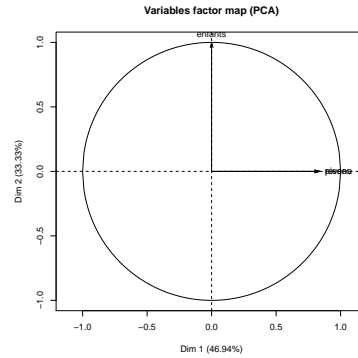
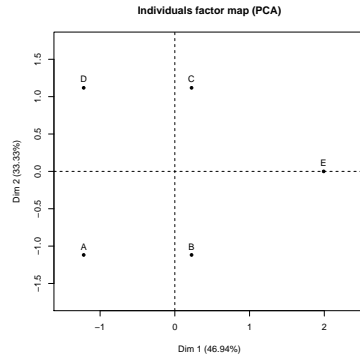
Ménage	Revenu	nb pièces	nb enfants
A	10 000	1	1
B	10 000	2	1
C	10 000	2	3
D	10 000	1	3
E	70 000	2	2

Diverses représentations



Fonction PCA

On obtient sur R avec la fonction `PCA` : `res <- PCA(D)`



Le plan de projection est ici défini par $\mathcal{P} = \text{vect}(u_1, u_2)$ avec $u_1 = X_1 + X_2$ et $u_2 = X_3$.

3.3 Analyse du nuage des individus

Notations

On se place dans l'espace \mathbb{R}^p muni de la distance euclidienne :

- $\langle e_i, e_j \rangle = \sum_{k=1}^p x_{i,k} x_{j,k}$
- $\|e_i\|^2 = \sum_{k=1}^p e_{i,k}^2$
- $d(e_i, e_j)^2 = \sum_{k=1}^p (x_{i,k} - x_{j,k})^2 = \|e_i - e_j\|^2$

Centrage des données :

- Soit $G = \frac{1}{n} \sum_{i=1}^n e_i = (\bar{X}_1, \dots, \bar{X}_p)'$ le centre de gravité du nuage des individus.
- Par simplicité d'écriture de la méthode, on centre le nuage :

$$e_i^c = \begin{pmatrix} x_{i,1} - \bar{X}_1 \\ \vdots \\ x_{i,p} - \bar{X}_p \end{pmatrix} \quad \text{et} \quad \mathcal{N}^c = \{e_1^c, \dots, e_n^c\}.$$

Idée

Chercher à projeter les observations dans un sous-espace \mathcal{F} visible à l'œil qui "restitue au mieux" l'information contenue dans le tableau.

L'inertie

- On appelle **inertie totale** du nuage de points \mathcal{N}

$$I(\mathcal{N}) = \frac{1}{n} \sum_{i=1}^n d(e_i, G)^2 = \frac{1}{n} \sum_{i=1}^n \|e_i - G\|^2 = \frac{1}{n} \sum_{i=1}^n \|e_i^c\|^2 = I(\mathcal{N}^c).$$

- On appelle **inertie portée par un sous espace** \mathcal{F} du nuage de points \mathcal{N}

$$I_{\mathcal{F}}(\mathcal{N}) = \frac{1}{n} \sum_{i=1}^n \|P_{\mathcal{F}}(e_i^c)\|^2,$$

où $P_{\mathcal{F}}(\cdot)$ est la projection orthogonale sur \mathcal{F} .

Une remarque importante

- L'ACP permet de prendre en compte une pondération différente des individus : p_i poids de l'individu i tel que $\sum_{i=1}^n p_i = 1$.
- L'inertie est alors définie par $I(\mathcal{N}) = \sum_{i=1}^n p_i d(e_i, G)^2$.
- Dans ce cours, on supposera que tous les individus ont le même poids : $p_i = 1/n$, $i = 1, \dots, n$.

Il est facile de voir que $I_{\mathcal{F}}(\mathcal{N}) \leq I(\mathcal{N})$: projeter fait perdre de l'inertie.

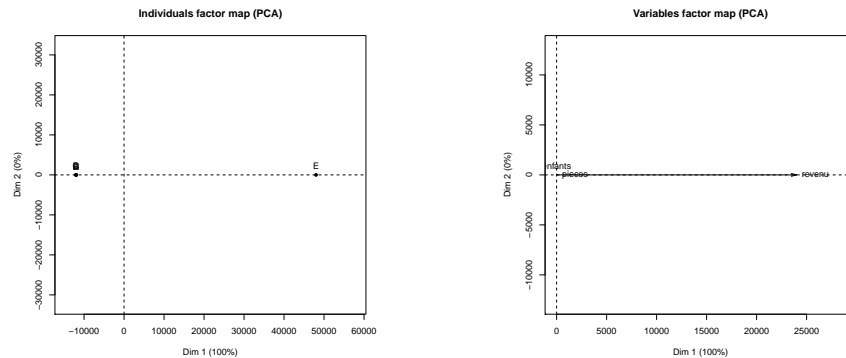
Objectif

Trouver le sous espace \mathcal{F} qui minimise cette perte d'inertie, ou encore trouver le sous espace \mathcal{F} tel que

$$I_{\mathcal{F}}(\mathcal{N}) \text{ soit maximale.}$$

Un "léger" problème

1. Les variables ne sont généralement pas à la même échelle
2. L'inertie est donc généralement "portée" par un sous groupe de variables
3. Sur l'exemple, la variable `revenu` porte à elle seule la quasi totalité de l'inertie...



Pour pallier à cette difficulté, on réduit les données initiales :

$$X = \begin{matrix} & X_1 & \dots & X_p \\ e_1 & \tilde{x}_{11} & \dots & \tilde{x}_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ e_n & \tilde{x}_{n1} & \dots & \tilde{x}_{np} \end{matrix} \quad \text{avec} \quad \tilde{x}_{ij} = \frac{x_{ij} - \bar{X}_j}{\sigma_j} \quad \text{et} \quad \sigma_j = \sigma(X_j).$$

Avec un léger abus, on note $x_{ij} = \tilde{x}_{ij}$.

Centrage et réduction

- On rappelle que

	revenu	pieces	enfants
A	10000	1	1
B	10000	2	1
C	10000	2	3
D	10000	1	3
E	70000	2	2
μ	22000.0	1.6	2.0
σ	24000	0.4898979	0.8944272

- On obtient après centrage et réduction

	revenu	pieces	enfants
A	-0.5	-1.2247449	-1.118034
B	-0.5	0.8164966	-1.118034
C	-0.5	0.8164966	1.118034
D	-0.5	-1.2247449	1.118034
E	2.0	0.8164966	0.000000
μ	0	0	0
σ	1	1	1

3.3.1 Recherche des axes factoriels

"Meilleur" sous-espace de dimension 1

Il s'agit de chercher une droite vectorielle Δ_1 dirigée par un vecteur unitaire $u_1 \in \mathbb{R}^p$ telle que $I_{\Delta_1}(\mathcal{N})$ soit maximale.

Propriété

- $I_{\Delta_1}(\mathcal{N}) = \frac{1}{n} \sum_{i=1}^n \langle e_i, u_1 \rangle^2 = \frac{1}{n} C_1' C_1$ où

$$C_1 = (\langle e_1, u_1 \rangle, \dots, \langle e_n, u_1 \rangle)' = X u_1.$$

Le problème mathématique

Chercher u_1 unitaire qui maximise $I_{\Delta_1}(\mathcal{N})$ revient à résoudre le problème d'optimisation suivant :

$$\text{maximiser } \frac{1}{n} u_1' X' X u_1 \text{ sous la contrainte } \|u_1\| = 1.$$

Propriétés

Un vecteur propre unitaire u_1 rendant l'inertie $I_{\Delta_1}(\mathcal{N})$ maximale est un vecteur propre normé associé à la plus grande valeur propre λ_1 de la matrice $\Sigma = \frac{1}{n}X'X$.

Remarques

- La matrice d'inertie $\Sigma = \frac{1}{n}X'X$ étant symétrique et définie positive, elle est diagonalisable et toutes ses valeurs propres sont positives ou nulles.
- u_1 est appelé premier axe factoriel.

Exemple

Sur l'exemple "jouet", on a

$$\frac{1}{n}X'X = \begin{pmatrix} 1.0000000 & 0.4082483 & 0.000000e + 00 \\ 0.4082483 & 1.0000000 & 3.144186e - 18 \\ 0.0000000 & 0.0000000 & 1.000000e + 00 \end{pmatrix}$$

D'où

```
$values
[1] 1.4082483 1.0000000 0.5917517

$vector
      [,1] [,2] [,3]
[1,] 0.7071068 0 0.7071068
[2,] 0.7071068 0 -0.7071068
[3,] 0.0000000 1 0.0000000
```

On obtient les coordonnées des individus sur le premier axe

```
> X%*%u1 #coordonnées des individus sur les axes
      [,1]
A -1.2195788
B 0.2237969
C 0.2237969
D -1.2195788
E 1.9915638
```

Second axe : première approche

Problème

Trouver une droite vectorielle Δ_2 dirigée par un vecteur normé u_2 telle que

$$\begin{cases} I_{\Delta_2}(\mathcal{N}) = u_2' \Sigma u_2 \text{ maximale} \\ \|u_2\|^2 = u_2' u_2 = 1 \\ \langle u_2, u_1 \rangle = u_2' u_1 = 0 \end{cases}$$

Solution

Un vecteur unitaire u_2 solution du problème précédent est un vecteur propre normé associé à la deuxième plus grande valeur propre λ_2 de la matrice $\Sigma = \frac{1}{n}X'X$.

Question

Le plan $\text{vect}(u_1, u_2)$ est-il le meilleur sous-espace de \mathbb{R}^2 en terme de maximisation d'inertie projeté ?

Réponse

La réponse est oui ! On déduit ainsi qu'un sous-espace de dimension $q < p$ qui maximise l'inertie projeté est donné par $\text{vect}(u_1, \dots, u_q)$ où u_j est un vecteur normé associé à la $j^{\text{ème}}$ plus grande valeur propre λ_j de $\Sigma = \frac{1}{n}X'X$.

Conclusion : chercher les axes factoriels revient à diagonaliser $\Sigma = \frac{1}{n}X'X$.

ACP \approx changement de base

<i>Base canonique</i>	<i>Base $\{u_1, \dots, u_p\}$</i>
$X = \begin{pmatrix} X_1 & \dots & X_p \\ x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$	$X = \begin{pmatrix} C_1 & \dots & C_p \\ c_{11} & \dots & c_{1p} \\ \vdots & \vdots & \vdots \\ c_{n1} & \dots & c_{np} \end{pmatrix}$

Propriété

1. $C_j = Xu_j = \sum_{k=1}^p u_{kj}X_k$
2. C_j centrée et $\mathbf{V}(C_j) = \frac{1}{n}\|C_j\|^2 = \lambda_j = I_{\Delta_j}(\mathcal{N})$.
3. $\rho(C_j, C_k) = 0$ pour $k \neq j$.

Conclusion

L'ACP normée remplace les variables d'origines X_j par de nouvelles variables C_j appelées composantes principales, de variance maximale, non corrélées deux à deux et qui s'expriment comme combinaison linéaire des variables d'origine.

Premier bilan

Calculer les axes factoriels n'est pas difficile. Il reste néanmoins plusieurs problèmes à régler pour mener l'analyse :

1. Comment choisir le sous-espace ? (Ou encore, combien d'axes factoriels doit-on retenir ?)
2. Comment mesurer la qualité de représentation d'un individu sur le sous-espace choisi ?
3. Comment interpréter les axes ?

3.3.2 Contributions et qualités de représentation

Propriétés

1. $I(\mathcal{N}) = \text{tr}(\Sigma) = \sum_{j=1}^p \lambda_j$.
2. $I_{\Delta_j}(\mathcal{N}) = \lambda_j$.
3. L'inertie est additive : si on note \mathcal{F}_k le sous espace de \mathbb{R}^p engendré par les k premiers vecteurs propres associés aux k plus grandes valeurs propres de Σ , alors

$$I_{\mathcal{F}_k}(\mathcal{N}) = \sum_{j=1}^k \lambda_j.$$

Définitions

- La **contribution à l'inertie** de l'axe Δ_k est λ_k .
- La **contribution relative à l'inertie** de l'axe Δ_k est $\lambda_k / \sum_{j=1}^p \lambda_j$.
- La **contribution relative à l'inertie** du plan (Δ_j, Δ_k) est $(\lambda_j + \lambda_k) / \sum_{j=1}^p \lambda_j$.

Choix du nombre d'axes

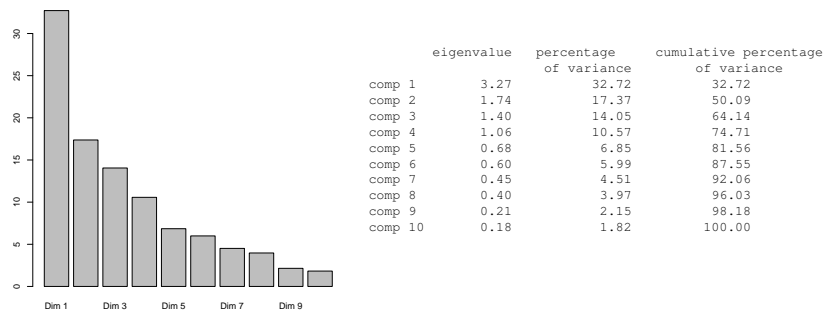
Hélas

Il n'existe pas de méthodes "universelles" permettant de choisir le nombre d'axes. Les critères sont le plus souvent empiriques.

Les critères sont le plus souvent empiriques :

- Pourcentage d'inertie reconstitué par le sous-espace sélectionné
- Etudier la décroissance des valeurs propres (critère dit "du coude")

Exemple du décathlon



Il semble que retenir 4 axes puisse être un choix intéressant.

Questions

1. Quels individus ont le plus contribué à la formation des axes factoriels ?
2. Quels individus sont bien représentés par les axes factoriels ?

Solutions

1. Comme $I_{\Delta_j}(\mathcal{N}) = \frac{1}{n} C_j' C_j = \frac{1}{n} \sum_{i=1}^n c_{ij}^2 = \lambda_j$, on mesure la contribution de l'individu i à l'axe j par

$$CR_j(i) = \frac{c_{ij}^2}{n\lambda_j}.$$

2. Un individu e_i sera bien représenté par un axe Δ_j si il est "proche" de son projeté sur Δ_j , ou encore si le cosinus de l'angle $\theta_{ij} = \widehat{(e_i, u_j)}$ est proche de 1 ou -1 .

Solutions (suite)

La qualité de représentation de l'individu e_i sur l'axe u_j est ainsi mesurée par

$$qlt_j(i) = \cos^2 \theta_{ij} = \frac{\|P_{\Delta_j}(e_i)\|^2}{\|e_i\|^2}.$$

De même la qualité de représentation de e_i sur le plan $\mathcal{F} = (\Delta_j, \Delta_k)$ est mesurée par

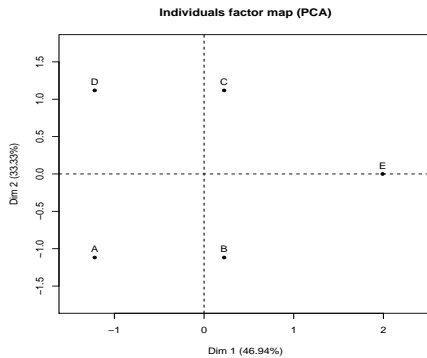
$$qlt_{\mathcal{F}}(i) = \frac{\|P_{\mathcal{F}}(e_i)\|^2}{\|e_i\|^2}.$$

Propriétés

Le cosinus carré étant additif sur des sous-espaces orthogonaux, on déduit

$$qlt_{\mathcal{F}}(i) = \frac{\|P_{\Delta_j}(e_i)\|^2 + \|P_{\Delta_k}(e_i)\|^2}{\|e_i\|^2}.$$

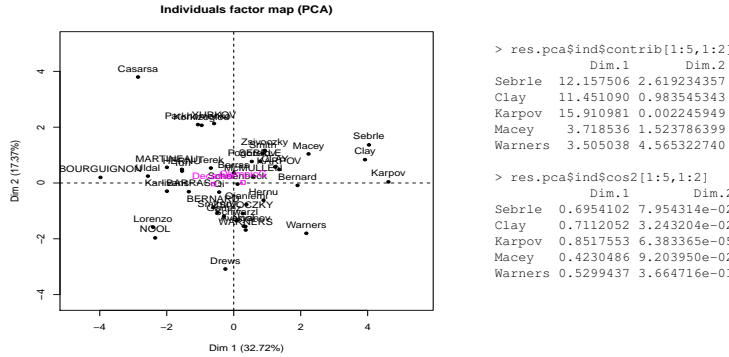
Exemple



```
> res$ind$contrib
      Dim.1      Dim.2      Dim.3
A 21.1237244 2.500000e+01 8.876276
B  0.7113098 2.500000e+01 29.288690
C  0.7113098 2.500000e+01 29.288690
D 21.1237244 2.500000e+01 8.876276
E 56.3299316 3.081488e-31 23.670068

> res$ind$cos2
      Dim.1      Dim.2      Dim.3
A 0.49579081 4.166667e-01 0.08754252
B 0.02311617 5.769231e-01 0.39996075
C 0.02311617 5.769231e-01 0.39996075
D 0.49579081 4.166667e-01 0.08754252
E 0.84992711 3.301594e-33 0.15007289
```

Exemple



Comment interpréter les positions des individus ?

3.4 Analyse du nuage des variables

L'analyse duale

- On s'intéresse maintenant au nuage des variables $\{X_1, \dots, X_p\}$, $X_j \in \mathbb{R}^n$.
- Pour prendre en compte les poids des individus, on munit \mathbb{R}^n de la métrique $P = \text{diag}(p_1, \dots, p_n)$ (on mène l'analyse avec $p_i = 1/n$)

Conséquence

1. $\|X_j\|_P = 1$ et $\cos(X_j, X_k) = \rho(X_j, X_k)$.
2. Les variables X_j se trouvent sur la sphère unité de \mathbb{R}^n .
3. La projection des X_j sur des plans passant par l'origine se trouveront à l'intérieur du cercle unité de \mathbb{R}^2 (que nous appellerons cercle des corrélations).

On souhaite transposer l'analyse du nuage des individus à celui des variables (on parle d'**analyse duale**).

Bonheur...

- Les axes factoriels de \mathbb{R}^n (ceux du nuage des variables) se déduisent de ceux de \mathbb{R}^p (ceux du nuage des individus).
- Les taux d'inerties sont identiques pour des axes du même rang dans les deux analyses.

Propriétés

1. $I(\mathcal{N}) = I(\mathcal{M}) = \text{trace}\left(\frac{1}{n}X'X\right) = p$, les deux dernières égalités viennent du fait que la matrice des données est centrée-réduite.

2. Chercher un vecteur v_1 de \mathbb{R}^n unitaire qui maximise $I_{\text{vect } v_1}(\mathcal{M})$ revient à résoudre le problème

$$\text{maximiser } \frac{1}{n} v_1' X X' v_1 \text{ sous la contrainte } \|v_1\|_P = 1$$

3. La solution est donnée par un vecteur propre normé de $\frac{1}{n} X X'$ associé à la plus grande valeur propre de $\frac{1}{n} X X'$.

Propriétés

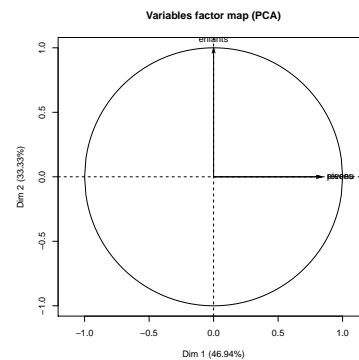
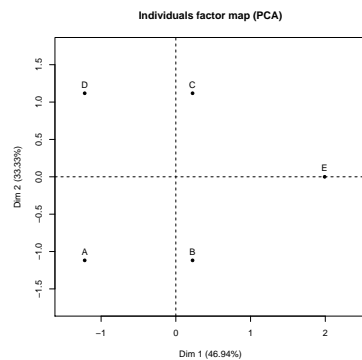
- $v_j = \frac{1}{\sqrt{\lambda_j}} X u_j$ et $u_j = \frac{1}{\sqrt{\lambda_j}} X' P v_j$;
- Coordonnées du projeté de X_j projeté $i^{\text{ème}}$ axe :
 1. $d_{ij} = \langle X_j, v_i \rangle_P \implies D_j = X' P v_j$
 2. $D_j = \frac{1}{\sqrt{\lambda_j}} X' P C_j$ et $C_j = \frac{1}{\sqrt{\lambda_j}} X D_j$.

PROPRIETE

$$d_{ij} = \rho(X_j, C_i)$$

Conséquence

- Si d_{ij} est grand (proche de 1 ce qui signifie que la projection de la variable est proche du cercle des corrélations), cela signifie que :
 - la $j^{\text{ème}}$ est fortement corrélée à la $i^{\text{ème}}$ composante principale.
 - les individus qui possèdent une coordonnée élevée sur l'axe i seront parmi ceux possédant une forte valeur de la variable j .
- Cette propriété permet de faire le lien entre la représentation du nuage des variables et celle du nuage des individus sur un plan factoriel.



Interprétation

- L'axe 1 oppose les individus possédant des revenus élevés et vivant dans de grands appartements à des individus plus pauvres et vivant dans des appartements plus petits.
- L'axe 2 les grandes familles aux petites.
- Contribution de la variable j à l'axe i : d_{ij}^2/λ_j car

$$I_{v_i}(\mathcal{M}) = \sum_{j=1}^p (\langle X_j, v_i \rangle_P)^2 = \sum_{j=1}^p d_{ij}^2 = \lambda_j.$$

- Qualité de représentation de la variable X_j sur l'axe i :

$$\cos^2(X_j, v_i) = \langle X_j, v_i \rangle^2 = d_{ij}^2.$$

- Corrélations entre variables : $\rho(X_j, X_k) = \cos(X_j, X_k)$. Donc, sur le cercle des corrélations, deux variables bien représentées
 - proches sont fortement corrélées ;
 - qui s'opposent sont négativement corrélées ;
 - orthogonales sont non corrélées.

Variables et individus supplémentaires

Certaines analyses peuvent être menées en retirant des variables ou des individus pour construire les axes de l'ACP :

- individus aberrants pouvant "trop" contribuer à l'inertie ;
- variables ayant été construites à partir d'autres variables déjà utilisées dans l'analyse. Les variables `classement` et `points` sur l'exemple du décathlon.

Une fois l'ACP réalisé, il peut néanmoins être intéressant de visualiser comment ces individus ou variables se situent par rapport aux autres.

Variables et individus supplémentaires

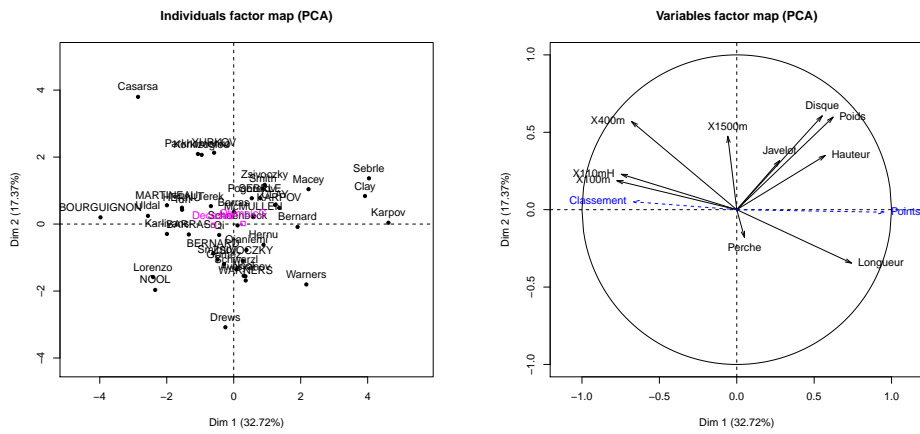
La méthode est simple. Il suffit de

1. faire subir à ces nouveaux éléments les mêmes transformations qu'aux autres (centrage et réduction) ;
2. projeter ces nouveaux éléments sur les axes factoriels du nuage des individus ou des variables.

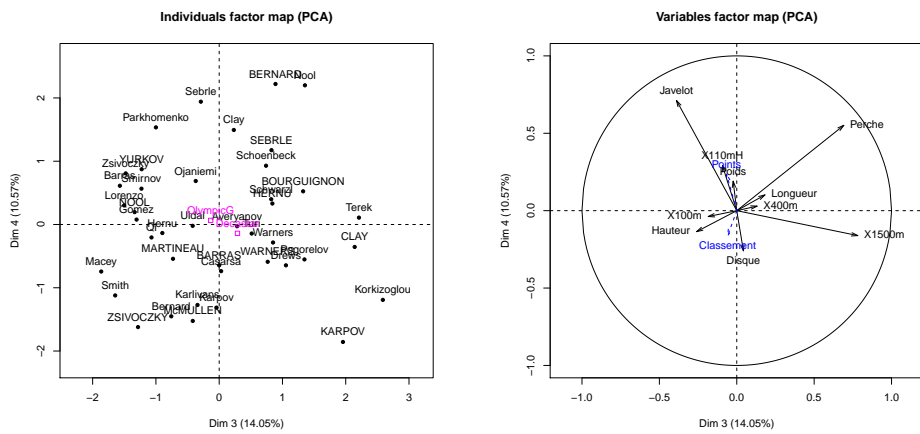
Retour à l'exemple du décathlon

- On fait l'ACP du jeu de données en utilisant les 41 individus et les 10 variables correspondant aux disciplines du decathlon. les variables classement, points et compétition sont mises en supplémentaire
- On fait l'analyse des 4 premiers axes.

Premier plan factoriel



Second plan factoriel



Part II

Théorie de l'estimation

1 Quelques exemples

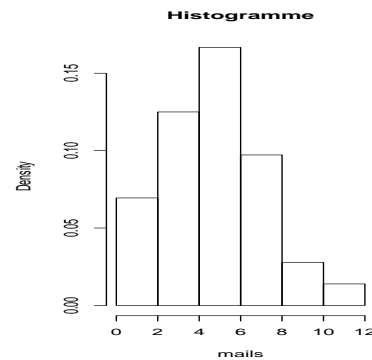
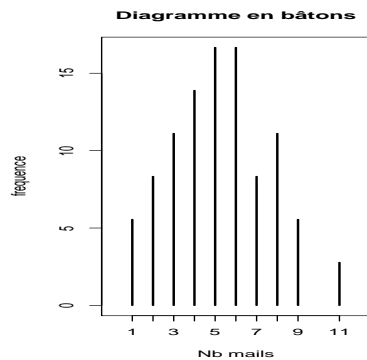
Exemple 1

- On souhaite tester l'efficacité d'un nouveau traitement à l'aide d'un essai clinique.
- On traite $n = 100$ patients atteints de la pathologie.
- A l'issue de l'étude, 72 patients sont guéris.
- Soit p_0 la probabilité de guérison suite au traitement en question.
- On est tenté de conclure $p_0 \approx 72$.

Un tel résultat n'a cependant guère d'intérêt si on n'est pas capable de préciser l'erreur susceptible d'être commise par cette estimation.

Exemple 2

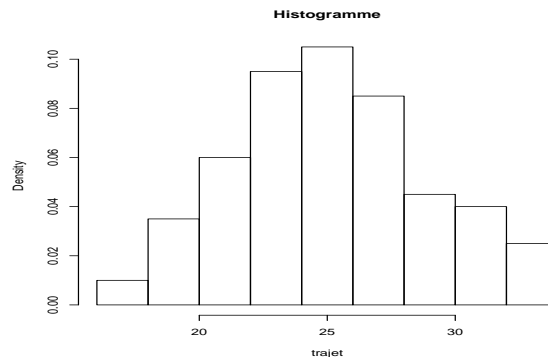
- On s'intéresse au nombre de mails reçus par jour par un utilisateur pendant 36 journées.
- $\bar{x} = 5.22$, $S_n^2 = 5.72$.



Quelle est la probabilité de recevoir plus de 5 mails dans une journée ?

Exemple 3

- Temps mis pour venir de son domicile à Supelec.
- On dispose de $n = 100$ mesures : $\bar{x} = 25.1$, $S_n^2 = 14.46$.



J'ai un devoir à 8h30, quelle est la probabilité que j'arrive en retard si je pars de chez moi à 7h55 ?

- Nécessité de se dégager des observations x_1, \dots, x_n pour répondre à de telles questions.
- Si on mesure la durée du trajet pendant 100 nouveaux jours, on peut en effet penser que les nouvelles observations ne seront pas exactement les mêmes que les anciennes.

Idée

Considérer que les n valeurs observées x_1, \dots, x_n sont des réalisations de variables aléatoires X_1, \dots, X_n .

Attention

X_i est une variable aléatoire et x_i est une réalisation de cette variable, c'est-à-dire un nombre !

Un modèle pour l'exemple 1

- On note $x_i = 1$ si le $i^{\text{ème}}$ patient a guéri, 0 sinon.
- On peut supposer que x_i est la réalisation d'une variable aléatoire X_i de loi de bernoulli de paramètre p_0 .
- Si les individus sont choisis de manière indépendante et ont tous la même probabilité de guérir (ce qui peut revenir à dire qu'ils en sont au même stade de la pathologie), il est alors raisonnable de supposer que les variables aléatoires X_1, \dots, X_n sont indépendantes.

On dit que X_1, \dots, X_n est un n -échantillon de variables aléatoires indépendantes de même loi $B(p_0)$.

2 Modèle statistique

Définitions

Modèle

On appelle **modèle statistique** tout triplet $(\mathcal{H}, \mathcal{A}, \mathcal{P})$ où

- \mathcal{H} est l'espace des observations (l'ensemble de tous les résultats possibles de l'expérience) ;
- \mathcal{A} est une tribu sur \mathcal{H} ;
- \mathcal{P} est une famille de probabilités définie sur $(\mathcal{H}, \mathcal{A})$.

Modèle paramétrique

Si $\mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ alors on parle de modèle paramétrique et Θ est l'espace des paramètres.

Modèle identifiable

Si $\theta \mapsto \mathbf{P}_\theta$ est injective, le modèle est dit **identifiable**.

Exemples

	\mathcal{H}	\mathcal{A}	\mathcal{P}
Exemple 1	$\{0, 1\}$	$\mathcal{P}(\{0, 1\})$	$\{B(p), p \in \{0, 1\}\}$
Exemple 2	\mathbb{N}	$\mathcal{P}(\{0, 1\})$	$\{\mathcal{P}(\lambda), \lambda > 0\}$
Exemple 3	\mathbb{R}	$\mathcal{B}(\mathbb{R})$	$\{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+\}$

- Les modèles que nous allons considérer auront pour espace d'observations un ensemble dénombrable Ω ou \mathbb{R}^d et seront munis des tribus $\mathcal{P}(\Omega)$ ou $\mathcal{B}(\mathbb{R}^d)$.
- Dans la suite, on se donne un modèle $\mathcal{M} = (\mathcal{H}, \mathcal{P} = \{\mathbf{P}_\theta, \theta \in \Theta\})$.

Echantillon

Un échantillon de taille n est une suite X_1, \dots, X_n de n variables aléatoires indépendantes et de même loi \mathbf{P}_{θ_0} , pour $\theta_0 \in \Theta$.

Notations

- L'échantillon définit un vecteur aléatoire (X_1, \dots, X_n) ayant comme loi $\mathbf{P}_{\theta_0}^{\otimes n}$.
- On note $\mathcal{M}_n = (\mathcal{H}^n, \{\mathbf{P}^{\otimes n}, \theta \in \Theta\})$ le modèle produit.
- Le modèle \mathcal{M}_n est un ensemble de loi sur \mathcal{H}^n contenant $\mathbf{P}_{\theta_0}^{\otimes n}$.

La démarche statistique

1. On récolte n observations (n valeurs) x_1, \dots, x_n qui sont le résultats de n expériences aléatoires indépendantes.
2. **Modélisation** : on suppose que les n valeurs sont des réalisations de n variables aléatoires indépendantes X_1, \dots, X_n et de même loi \mathbf{P}_{θ_0} . Ce qui nous amène à définir le modèle $\mathcal{M}_n = (\mathcal{H}^n, \{\mathbf{P}^{\otimes n}, \theta \in \Theta\})$.
3. **Estimation** : Chercher dans le modèle une loi $\mathbf{P}_{\hat{\theta}}$ qui soit le plus proche possible de \mathbf{P}_{θ_0} : chercher un **estimateur** $\hat{\theta}$ de θ_0 .

Estimateurs

Définitions

- Une statistique est une application mesurable définie sur \mathcal{H}^n .
- Un estimateur (de θ_0) est une fonction mesurable de (X_1, \dots, X_n) indépendante de θ à valeurs dans un sur-ensemble de Θ .

Exemple 1

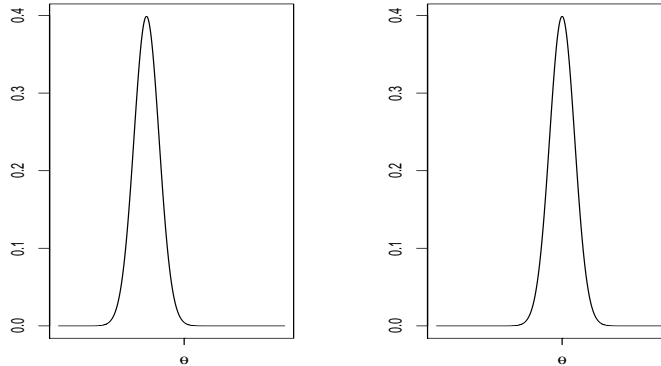
Les variables aléatoires $\hat{\theta}_1 = X_1$ et $\hat{\theta}_2 = \bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ sont des estimateurs de θ_0 .

Remarque : dans la suite, nous supposerons que le paramètre à estimer est θ . Les définitions et les résultats généraux que nous allons présenter s'étendent au cas où le paramètre d'intérêt est une fonction $g(\theta)$ de θ .

3 Qualités d'un estimateur

3.1 Biais, variance et risque quadratique

- Un estimateur $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ est une variable aléatoire. Il va donc posséder une loi.



- Un moyen de mesure la qualité de $\hat{\theta}$ est de regarder sa "valeur moyenne" et de la comparer à θ .

Biais d'un estimateur

- On désigne par \mathbf{E}_θ l'espérance sous la loi \mathbf{P}_θ :

$$\mathbf{E}_\theta(\hat{\theta}) = \mathbf{E}_\theta(\hat{\theta}(X_1, \dots, X_n)) = \int_{\mathcal{H}_n} \hat{\theta}(x) \mathbf{P}_\theta dx$$

où $x = (x_1, \dots, x_n)$.

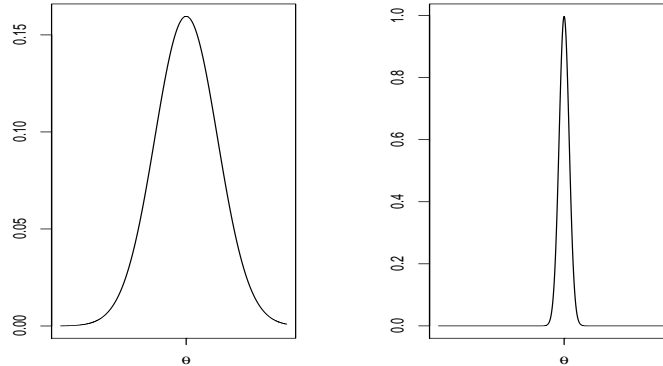
Soit $\hat{\theta}$ un estimateur d'ordre 1.

1. Le **biais** de $\hat{\theta}$ en θ est $\mathbf{E}_\theta(\hat{\theta}) - \theta$.
2. $\hat{\theta}$ est **sans biais** lorsque son biais est nul en chaque $\theta \in \Theta$.
3. $\hat{\theta}$ est asymptotiquement sans biais si pour chaque $\theta \in \Theta$, $\lim_{n \rightarrow \infty} \mathbf{E}_\theta(\hat{\theta}) = \theta$.

Exemple 1

Les estimateurs $\hat{\theta}_1$ et $\hat{\theta}_2$ sont sans biais.

- Mesurer le biais n'est pas suffisant.
- Il faut également mesurer la dispersion des estimateurs.



Risque quadratique

Définitions

Soit $\hat{\theta}$ un estimateur d'ordre 2.

1. Le risque quadratique de $\hat{\theta}$ sous \mathbf{P}_θ est

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{E}_\theta \|\hat{\theta} - \theta\|^2$$

2. Soit $\hat{\theta}'$ un autre estimateur d'ordre 2. On dit que $\hat{\theta}$ est préférable à $\hat{\theta}'$ si

$$\mathcal{R}(\theta, \hat{\theta}) \leq \mathcal{R}(\theta, \hat{\theta}') \quad \forall \theta \in \Theta.$$

Exemple 1

$\hat{\theta}_2$ est préférable à $\hat{\theta}_1$.

Estimateur VUMSB

Propriété décomposition biais variance

1. Si $\hat{\theta}$ est d'ordre 2, on a la décomposition

$$\mathcal{R}(\theta, \hat{\theta}) = \|\mathbf{E}_\theta(\hat{\theta}) - \theta\|^2 + \mathbf{E}_\theta \|\hat{\theta} - \mathbf{E}_\theta \hat{\theta}\|^2.$$

2. Si $\theta \in \mathbb{R}$, on obtient

$$\mathcal{R}(\theta, \hat{\theta}) = b(\hat{\theta})^2 + \mathbf{V}(\hat{\theta}).$$

Définition

Si $\hat{\theta}$ est sans biais, on dit qu'il est de **variance uniformément minimum parmi les estimateurs sans biais (VUMSB)** si il est préférable à tout autre estimateur sans biais d'ordre 2.

Exemple

$\hat{\theta}_2$ est VUMSB.

3.2 Critère de performance asymptotique

Consistance

On dit que l'estimateur $\hat{\theta}$ est **consistant** (ou **convergent**) si $\hat{\theta} \xrightarrow{P} \theta \forall \theta \in \Theta$, c'est-à-dire

$$\forall \theta \in \Theta, \forall \varepsilon > 0 \lim_{n \rightarrow \infty} \mathbf{P}_\theta(\|\hat{\theta} - \theta\| > \varepsilon) = 0.$$

Soit $(v_n)_n$ une suite de réels positifs telle que $v_n \rightarrow \infty$. On dit que $\hat{\theta}$ est asymptotiquement normal, de vitesse v_n si $\forall \theta \in \Theta$

$$v_n(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma_\theta)$$

où Σ_θ est une matrice symétrique définie positive.

Outils

La loi des grands nombres et le théorème central limite sont souvent utilisés pour montrer la consistance et la normalité asymptotique.

Loi des grands nombres

Soit $(X_n)_n$ une suite de vecteurs aléatoires i.i.d. d'espérance $\mu \in \mathbb{R}^d$. Alors $\bar{X}_n \xrightarrow{p.s.} \mu$.

Si de plus X_i est d'ordre 2, on a $\bar{X}_n \xrightarrow{L_2} \mu$.

TCL

Soit $(X_n)_n$ une suite de vecteurs aléatoires i.i.d. d'espérance $\mu \in \mathbb{R}^d$ et de matrice de variance covariance Σ , alors

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma).$$

Exemple 1

$\hat{\theta}_1$ est consistant et asymptotiquement normal (avec la vitesse \sqrt{n}).

Delta méthode

- X_1, \dots, X_n n échantillon i.i.d. de loi exponentielle de paramètre $\lambda > 0$. $\hat{\theta} = 1/\bar{X}_n$ estimateur de λ asymptotiquement normal ?

Delta méthode

Si $v_n(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} X \sim \mathcal{N}(0, \Sigma)$ et si $h : \mathbb{R}^d \rightarrow \mathbb{R}^m$ admet des dérivées partielles au point θ , alors

$$v_n(h(\hat{\theta}) - h(\theta)) \xrightarrow{\mathcal{L}} Dh_\theta X$$

où Dh_θ est la matrice $m \times d$ de terme $(Dh_\theta)_{ij} = \frac{\partial h_i}{\partial \theta_j}(\theta)$.

- On obtient grâce à la delta-méthode :

$$\frac{\sqrt{n}}{\lambda} \left[\frac{1}{\bar{X}_n} - \lambda \right] \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

3.3 2 méthodes d'estimation

3.3.1 La méthode des moments

- modèle $(\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \Theta\})$.
- n échantillon X_1, \dots, X_n i.i.d. de loi \mathbf{P}_θ .

Idée

Trouver le paramètre $\theta \in \Theta$ tel que les moments empiriques coïncident avec les moments théoriques :

$$\hat{m}_j = \frac{1}{n} \sum_{i=1}^n X_i^j \approx m_j(\theta_0) = \mathbf{E}_{\theta_0}(X_i^j), \quad j = 1, \dots, p.$$

Si $p = 1$ la méthode revient à résoudre l'équation en θ :

$$\frac{1}{n} \sum_{i=1}^n X_i = \mathbf{E}_\theta(X_1).$$

L'estimateur des moments

L'estimateur des moments est défini comme la solution du système à p équations :

$$\begin{cases} m_1(\theta) = \hat{m}_1 \\ \vdots \\ m_p(\theta) = \hat{m}_p \end{cases}$$

$$M : \Theta \rightarrow \mathcal{L}$$

$$\theta \mapsto (m_1(\theta), \dots, m_p(\theta))$$

est une bijection. Alors l'estimateur des moments existe et est unique.

Pour le modèle gaussien $\mathbf{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ et $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \hat{\mu}^2$

Propriété

Rappel

Si \mathbf{P}_θ est d'ordre p , la LFGN et le TCL assure la consistance et la normalité asymptotique des moments empiriques.

Théorème

Soit $\hat{\theta}$ l'estimateur des moments.

1. Si \mathbf{P}_θ admet un moment d'ordre p fini et si M est un homéomorphisme alors $\hat{\theta}$ est consistant.
2. Si \mathbf{P}_θ admet un moment d'ordre $2p$ fini et si M est un difféomorphisme alors

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \mathbb{V}_\theta)$$

où $\mathbb{V}_\theta = DM_\theta^{-1} \Sigma_\theta (DM_\theta^{-1})'$ et $\Sigma_\theta = \mathbf{V}(X_1, X_1^2, \dots, X_1^p)$.

3.3.2 La méthode du maximum de vraisemblance

Retour à l'exemple 1

- X_1, \dots, X_n i.i.d. $X_1 \sim B(p)$.
- x_1, \dots, x_n réalisations de X_1, \dots, X_n .

Idée

1. La quantité $L(x_1, \dots, x_n; p) = \mathbf{P}_p(X_1 = x_1, \dots, X_n = x_n)$ peut être vue comme une mesure de la probabilité d'observer les données dont on dispose.
 2. Choisir le paramètre p qui maximise cette probabilité.
- $L(x_1, \dots, x_n; p)$ est appelée **vraisemblance** (elle mesure la vraisemblance des réalisations x_1, \dots, x_n sous la loi \mathbf{P}_p).
 - L'approche consiste à choisir p qui "rend ces réalisations les plus vraisemblables possible".

Vraisemblance

Cas discret

La **vraisemblance** du paramètre θ pour la réalisation (x_1, \dots, x_n) est l'application $L : \mathcal{H}^n \times \Theta$ définie par

$$L(x_1, \dots, x_n; \theta) = \mathbf{P}_\theta^{\otimes n}(\{x_1, \dots, x_n\}) = \prod_{i=1}^n \mathbf{P}_\theta(\{x_i\}).$$

Cas absolument continu

Soit $f(\cdot, \theta)$ la densité associée à \mathbf{P}_θ . La **vraisemblance** du paramètre θ pour la réalisation (x_1, \dots, x_n) est l'application $L : \mathcal{H}^n \times \Theta$ définie par

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta).$$

L'estimateur du maximum de vraisemblance

Définition

Un **estimateur du maximum de vraisemblance (EMV)** est une statistique g qui maximise la vraisemblance, c'est-à-dire $\forall (x_1, \dots, x_n) \in \mathcal{H}^n$

$$L(x_1, \dots, x_n; g(x_1, \dots, x_n)) = \sup_{\theta \in \Theta} L(x_1, \dots, x_n; \theta).$$

L'EMV s'écrit donc $\hat{\theta} = g(X_1, \dots, X_n)$.

- Pour le modèle gaussien $\mathbf{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, L'EMV est

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \text{ et } \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \hat{\mu}^2.$$

- Il coïncide avec l'estimateur des moments.

Propriétés de l'EMV

Invariance

Soit $\psi : \Theta \rightarrow \mathbb{R}^k$ et $\hat{\theta}$ l'EMV de θ . Alors l'emv de $\psi(\theta)$ est $\psi(\hat{\theta})$.

Consistance

On suppose que \mathbf{P}_θ admet une densité $f(x, \theta)$, que Θ est un ouvert et que $\theta \mapsto f(x, \theta)$ est différentiable. Alors l'EMV $\hat{\theta}$ est consistant.

4 Information de Fisher et Borne de Cramer Rao

4.1 Dimension 1

- On se place dans le cas où θ est réel.

Objectif : montrer que sous certaines hypothèses de régularité l'EMV est asymptotiquement VUMSB :

1. $\hat{\theta}$ est asymptotiquement sans biais.
2. il existe une fonction $r(n, \theta)$ telle que pour tout estimateur T sans biais de θ , on a $\mathbf{V}(T) \geq r(n, \theta)$.
3. la variance asymptotique de l'EMV vaut $r(n, \theta)$.

L'information de Fisher

Soit $\mathcal{M} = (\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \Theta\})$ un modèle. On suppose dans cette partie que :

- Θ est un ouvert.
- \mathbf{P}_θ admet une densité $f(x, \theta)$ (par rapport à la mesure de Lebesgue ou à la mesure de comptage) et que f est deux fois dérivable par rapport à θ .
- $\forall h \in L^1(\mathbf{P}_\theta)$ on a

$$\frac{\partial}{\partial \theta} \int h(x) f(x, \theta) dx = \int h(x) \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

On appelle **information de Fisher** du modèle \mathcal{M} au point θ :

$$I(\theta) = \mathbf{E}_\theta \left[\left(\frac{\partial}{\partial \theta} \log(f(X, \theta)) \right)^2 \right].$$

- **Fonction de score :**

$$S(x, \theta) = \frac{\partial}{\partial \theta} \log(f(x, \theta)).$$

Propriété

$$I(\theta) = -\mathbf{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log(f(X, \theta)) \right].$$

$$I(\theta) \geq 0 \text{ et } I(\theta) = 0 \Leftrightarrow f(x, \theta) = f(\theta).$$

$I(\theta)$ mesure en quelque sorte le pouvoir de discrimination du modèle entre deux valeurs proches du paramètre θ :

- $I(\theta)$ grand : il sera "facile" d'identifier quel modèle est le meilleur.
- $I(\theta)$ petit : l'identification sera plus difficile.

Modèle produit

Propriété d'additivité

Si X_1 et X_2 sont deux variables i.i.d. de loi \mathbf{P}_θ , alors

$$I_{(X_1, X_2)}(\theta) = I_{X_1}(\theta) + I_{X_2}(\theta) = 2I_{X_1}(\theta).$$

Corollaire

L'information de Fisher du modèle produit \mathcal{M}_n au point θ vaut $I_n(\theta) = nI(\theta)$.

Sur l'exemple 1, on a $I_n(p) = \frac{n}{p(1-p)}$.

Borne de Cramer-Rao

Théorème

Soit $T = T(X_1, \dots, X_n)$ un estimateur sans biais de θ . Alors

$$\mathbf{V}(T) \geq \frac{1}{I_n(\theta)}.$$

1. La quantité $\frac{1}{I_n(\theta)}$ est appelée borne de Cramer-Rao.
2. Si un estimateur sans biais $\hat{\theta}$ atteint la borne de Cramer-Rao, il est VUMSB.
3. Si T est un estimateur sans biais de $g(\theta)$ avec g dérivable, alors $\mathbf{V}(T) \geq \frac{(g'(\theta))^2}{I_n(\theta)}$.

Exemple 1

$\hat{p} = \bar{X}$ est VUMSB.

Efficacité asymptotique de l'EMV

Théorème

Sous certaines conditions de régularité sur la densité $f(x, \theta)$, l'EMV $\hat{\theta}_n$ est asymptotiquement gaussien et

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} N\left(0, \frac{1}{I(\theta)}\right).$$

- $\hat{\theta}_n$ est asymptotiquement sans biais.
- $\hat{\theta}_n$ est asymptotiquement efficace.
- $\hat{\theta}_n$ converge vers θ en moyenne quadratique.

4.2 Dimension p

BCR en dimension p

- On se place dans un modèle de densité $\mathcal{M} = (\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \mathbb{R}^p\})$, où $\mathbf{P}_\theta \sim f(\cdot, \theta)$;
- On note $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_p)'$ un estimateur du paramètre $\theta = (\theta_1, \dots, \theta_p)'$ de matrice de variance covariance $\Sigma_{\hat{\theta}}$.

Définition

La matrice d'information de Fisher au point θ du modèle ci-dessus est la matrice de dimension $p \times p$ définie par

$$\begin{aligned} I(\theta)_{i,j} &= \mathbf{E}_\theta \left[\frac{\partial}{\partial \theta_i} \log(f(X, \theta)) \frac{\partial}{\partial \theta_j} \log(f(X, \theta)) \right] \\ &= - \mathbf{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log(f(X, \theta)) \right]. \end{aligned}$$

Théorème

La borne de Cramer-Rao du modèle précédent est $\frac{1}{n}I(\theta)^{-1}$. C'est-à-dire que pour tout estimateur sans biais $\hat{\theta}$ de θ , on a

$$\Sigma_{\hat{\theta}} \geq \frac{1}{n}I(\theta)^{-1}$$

(l'inégalité est à prendre au sens des matrices sdp).

Un exemple : le modèle gaussien

Pour le modèle gaussien $\mathbf{P}_\theta = \mathcal{N}(\mu, \sigma^2)$, la matrice d'information de Fisher est donnée par :

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}.$$

La borne de Cramer-Rao vaut

$$BCR = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}.$$

- **Question :** l'estimateur $\tilde{\theta} = (\bar{X}, S^2)$ est-il efficace ?

Rappel : Corollaire de Cochran

Soit X_1, \dots, X_n un échantillon i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$. Alors :

1. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$;
2. \bar{X} et $\hat{\sigma}^2$ sont indépendantes ;
3. $(n\hat{\sigma}^2)/\sigma^2 \sim \chi_{(n-1)}^2$.

- On déduit

$$\Sigma_{\tilde{\theta}} = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n-1} \end{pmatrix}$$

$\tilde{\theta}$ est (presque) efficace.

Effacité asymptotique de l'emv

Théorème

On se place dans un modèle de densité $(\mathcal{H}, \{f(\cdot, \theta), \theta \in \Theta\})$. Sous certaines hypothèses de régularité sur la densité f , l'emv $\hat{\theta}$ de θ est

- consistant ;
- asymptotiquement normal :

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)^{-1}).$$

Retour au modèle gaussien

L'emv est donné par $\hat{\theta} = (\bar{X}, \hat{\sigma}^2)$. On obtient par Cochran

$$\begin{aligned} \sqrt{n}(\hat{\theta} - \theta) &\sim \mathcal{N}\left(0, \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \frac{n-1}{n} \end{pmatrix}\right) \\ &\xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)^{-1}). \end{aligned}$$

5 Estimation par intervalle de confiance

Motivations

- Donner une seule valeur pour estimer un paramètre peut se révéler trop ambitieux.
- *Exemple 1* : le taux de guérison du traitement est de 72% (alors qu'on ne l'a testé que sur 100 patients).
- Il peut parfois être plus raisonnable de donner une réponse dans le genre, le taux de guérison se trouve dans l'intervalle [70%, 74%] avec une confiance de 90%.

Intervalle de confiance

- X_1, \dots, X_n n échantillon i.i.d. de loi \mathbf{P}_{θ_0} .

Définition

Soit $\alpha \in]0, 1[$. On appelle intervalle de confiance pour θ_0 tout intervalle de la forme $[A_n, B_n]$, où A_n et B_n sont des fonctions mesurables telles que $\forall \theta \in \Theta$:

$$\mathbf{P}_{\theta}(\theta \in [A_n, B_n]) = 1 - \alpha.$$

Si $\lim_{n \rightarrow \infty} \mathbf{P}_{\theta}(\theta \in [A_n, B_n]) = 1 - \alpha$, on dit que $[A_n, B_n]$ est un intervalle de confiance asymptotique pour θ_0 au niveau $1 - \alpha$.

Construction d'un IC

- Inégalité de Bienaymé Tchebychev.
- Utilisation d'une **fonction pivotable pour le paramètre θ** : fonction mesurable des observations et du paramètre inconnu mais dont la loi ne dépend pas de θ .

Méthode

1. se donner un niveau $1 - \alpha$.
2. trouver un estimateur $\hat{\theta}_n$ de θ dont on connaît la loi afin de construire une fonction pivotable.

Exemples

1. $P_\theta = \mathcal{N}(\mu_0, \sigma^2)$ avec σ^2 connu :

$$IC_{1-\alpha}(\mu_0) = \left[\bar{X} - q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + q_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

2. $P_\theta = \mathcal{N}(\mu_0, \sigma^2)$ avec σ^2 inconnu :

$$IC_{1-\alpha}(\mu_0) = \left[\bar{X} - t_{1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2} \frac{S_n}{\sqrt{n}} \right].$$

3. P_θ d'espérance μ_0 et de variance σ^2 inconnue :

$$IC_{1-\alpha}^{asympt}(\mu_0) = \left[\bar{X} - t_{1-\alpha/2} \frac{S_n}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2} \frac{S_n}{\sqrt{n}} \right].$$

Part III

Tests d'hypothèses

1 Introduction

Exemple 1 : test de conformité

- On s'intéresse à la longueur de pièces fabriquées par une machine.
- "En théorie" la longueur moyenne de ces pièces doit être de 150cm.
- On décide de mesurer 49 pièces choisies au hasard. La valeur moyenne des mesures est de 149.9.

Peut-on dire que la machine est toujours bien réglée ?

Exemples 2-3

- Un médicament couramment utilisé est connu pour guérir 60% des patients.
- Un nouveau traitement est expérimenté sur 80 patients.
- On observe 60 guérisons.

Doit-on remplacer l'ancien traitement par le nouveau ?

- Une entreprise emploie 40 hommes et 60 femmes.

Peut-on affirmer que les recruteurs sont sexistes ?

Exemple 4

- On s'intéresse au nombre de garçons dans des familles de 4 enfants :

Nb de garçons	0	1	2	3	4	Total
Effectifs	3	30	39	23	5	100

Est-ce que la variable aléatoire nombre de garçons suit une loi Binomiale $\mathcal{B}(4, p)$?

Exemple 5

Le titanic a emporté à son bord :

- 325 passagers en première classe
- 285 passagers en deuxième classe
- 706 passagers en troisième classe
- 885 membres d'équipage.

Parmi les survivants on compte :

- 203 passagers en première classe
- 118 passagers en deuxième classe
- 178 passagers en troisième classe
- 212 membres d'équipage.

Existe-t-il un lien entre le fait d'avoir survécu et la classe ?

Problématique de test

- Ici, il ne s'agit plus d'estimer un paramètre à partir d'un échantillon mais de prendre une décision à l'aide de cet échantillon.
- Répondre aux questions posées revient à choisir une hypothèse parmi deux (on les notera H_0 et H_1).
- Un *test statistique* permet de réaliser un tel choix.

	H_0	H_1
Exemple 1	$\mu = 150$	$\mu \neq 150$
Exemple 2	$p = 0.6$	$p \geq 0.6$
Exemple 3	$p_F = p_H$	$p_F \neq p_H$
Exemple 4	$X \sim \mathcal{B}(4, p)$	$X \approx \mathcal{B}(4, p)$
Exemple 5	$S \amalg C$	$S \vee C$

2 Tests paramétriques

2.1 Vocabulaire

Hypothèse nulle et hypothèse alternative

- **Modèle statistique** $(\mathcal{H}, \{\mathbf{P}_\theta, \theta \in \Theta\})$.
- **Hypothèse nulle** : $H_0 : \theta \in \Theta_0$.
- **Hypothèse alternative** : $H_1 : \theta \in \Theta_1$.

Si $\Theta = \{\theta\}$ l'hypothèse est dite **simple**, sinon elle est dite **multiple**.

A partir de n observations $x = (x_1, \dots, x_n)$, prendre une décision :

- accepter H_0 .
- rejeter H_0 au profit de H_1 .

Fonction de test

Définition

- On appelle **fonction de test** toute statistique $\varphi : \mathcal{H}^n \rightarrow \{0, 1\}$
- L'ensemble $\varphi^{-1}(\{1\})$ noté \mathcal{R}_{H_0} est **la région de rejet ou région critique** du test.
- L'ensemble $\varphi^{-1}(\{0\})$ noté \mathcal{A}_{H_0} est **la région d'acceptation** du test.

Exemple

- $\mathbf{P}_\theta = \mathcal{N}(\mu, 1)$, $n = 10$.
- $H_0 : \mu = 3$ contre $H_1 : \mu = 3.5$.
- $\mathcal{R}_{H_0} = \{x \in \mathbb{R}^n : \bar{x} > s_0\}$.
- $\mathcal{A}_{H_0} = \{x \in \mathbb{R}^n : \bar{x} \leq s_0\}$.

Erreurs de décision

		Réalité	
		H_0	H_1
Décision	H_0	OK	erreur de deuxième espèce
	H_1	erreur de première espèce	OK

- Le **risque de première espèce d'un test** φ est la fonction

$$\begin{aligned} \alpha &: \Theta_0 \rightarrow [0, 1] \\ \theta_0 &\mapsto \mathbf{P}_{\theta_0}(\mathcal{R}_{H_0}) \end{aligned}$$

- Le **risque de deuxième espèce d'un test** φ est la fonction

$$\begin{aligned} \beta &: \Theta_1 \rightarrow [0, 1] \\ \theta_1 &\mapsto \mathbf{P}_{\theta_1}(\mathcal{A}_{H_0}) \end{aligned}$$

Exemple

- $\mathbf{P}_{\theta} = \mathcal{N}(\mu, 1)$, $n = 10$.
- $H_0 : \mu = 3$ contre $H_1 : \mu = 3.5$.
- $\mathcal{R}_{H_0} = \{x \in \mathbb{R}^n : \bar{x} > s_0\}$.
- $\mathcal{A}_{H_0} = \{x \in \mathbb{R}^n : \bar{x} \leq s_0\}$.

- **Risque de première espèce :**

$$\alpha = \mathbf{P}_{H_0}(\{x \in \mathbb{R}^n : \bar{x} > s_0\}) = \mathbf{P}_{H_0}(\bar{X}_n > s_0) = 1 - F_{3,0.1}(s_0)$$

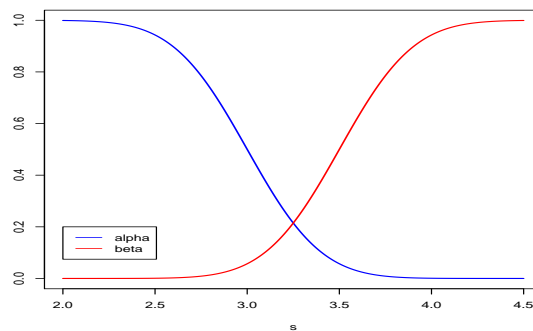
- **Risque de deuxième espèce :**

$$\beta = \mathbf{P}_{H_1}(\{x \in \mathbb{R}^n : \bar{x} \leq s_0\}) = F_{3.5,0.1}(s_0).$$

2.2 Le principe de Neyman-Pearson

Une idée

- Choisir les ensembles \mathcal{R}_{H_0} et \mathcal{A}_{H_0} qui minimisent les risques α et β .
- Si $\mathcal{A}_{H_0} = \mathcal{H}^n$ alors $\mathcal{R}_{H_0} = \emptyset$, $\alpha = 0$ et $\beta = 1$.
- Si $\mathcal{A}_{H_0} = \emptyset$ alors $\mathcal{R}_{H_0} = \mathcal{H}^n$, $\alpha = 1$ et $\beta = 0$.
- Les risques α et β varient généralement en sens inverse.



Le principe de Neyman et Pearson

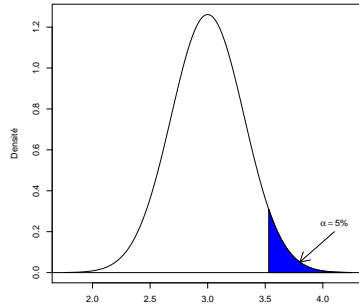
- Neyman et Pearson (1933) proposent de traiter les risques de façon non symétrique.
- On fixe tout d'abord le risque maximal de première espèce $\alpha = \sup_{\theta_0 \in \Theta_0} \alpha(\theta_0)$. Ce risque maximal est appelé **niveau du test**.

La procédure de Neyman et Pearson consiste à chercher dans l'ensemble des tests de niveau α un test optimal (dans le sens où son risque de deuxième espèce sera minimum).

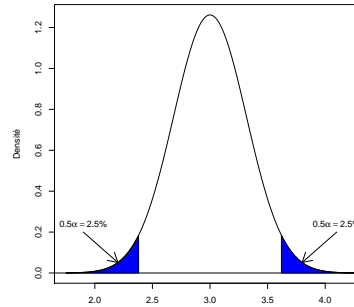
- $\mathbf{P}_\theta = N(\mu, 1)$, $n = 10$, $H_0 = 3$ contre $H_1 = 3.5$
- Sous H_0 : $\bar{X}_n \sim \mathcal{N}(3, \frac{1}{n})$.

$$\varphi(x_1, \dots, x_n) = \mathbf{1}_{\bar{x}_n > s}$$

$$\varphi(x_1, \dots, x_n) = \mathbf{1}_{s_0 \leq \bar{x}_n \leq s_1}$$



$$\beta = \mathbf{P}_{H_1}(\bar{X}_n \leq 3.52) \simeq 0.525.$$



$$\beta = \mathbf{P}_{H_1}(2.38 \leq \bar{X}_n \leq 3.62) \simeq 0.648.$$

En pratique

La construction d'un test de niveau α se compose des étapes suivantes :

1. Détermination des **hypothèses** H_0 et H_1 à partir du problème posé.
2. Détermination d'une **statistique de test** et de la **forme de la fonction de test**.
3. Détermination précise des **constantes** intervenant dans la fonction de test de sorte à ce que le niveau du test soit α .
4. **Conclusion** au vu de l'observation.

Dissymétrie des hypothèses...

- H_0 et H_1 ne jouent pas des rôles symétriques.
- En pratique le seul risque contrôlé est le risque de première espèce (fixé à α), H_0 est ainsi l'hypothèse à privilégier, il faut en tenir compte dans le choix des hypothèses.

Exemple

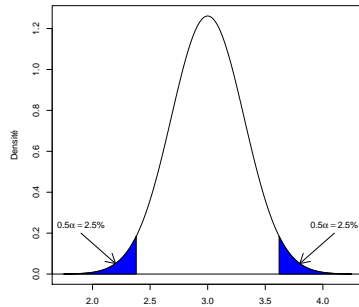
1. Mise en circulation d'un nouveau médicament :

$$H_0 : p_N = p_A \quad \text{contre} \quad H_1 : p_N > p_A.$$

2. Procès d'assise : H_0 : "innocent" contre H_1 : "coupable".

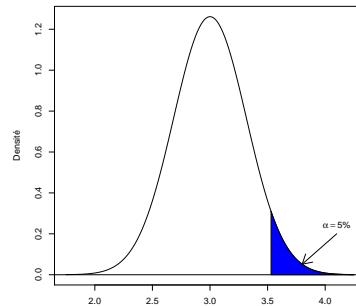
- H_0 doit être l'hypothèse à privilégier.
- Le choix de H_1 intervient dans le choix de la fonction de test (ou encore sur la forme de la zone de rejet du test)

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0$$



$$\mathcal{R}_{H_0} = \{] - \infty, x_1[\cup] x_2, \infty [\}$$

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu > \mu_0$$



$$\mathcal{R}_{H_0} = \{] x, \infty [\}$$

Accepter-rejeter les hypothèses ?

- Sur l'exemple précédent, on voit que la décision est prise en étudiant la loi d'une **statistique de test sous H_0** .
- Pour décider, on regarde si la valeur observée t_{obs} de la statistique de test T tombe dans une zone "raisonnable" sous l'hypothèse nulle H_0 .

Conclusion

- Si $t_{obs} \in \mathcal{A}_{H_0}$, on dit qu'on accepte l'hypothèse H_0 au profit de H_1 au niveau α .
- Si $t_{obs} \in \mathcal{R}_{H_0}$, on dit qu'on rejette l'hypothèse H_0 au profit de H_1 au niveau α .

2.3 Puissance de test - Test UPP

Définitions

- On appelle **puissance d'un test** la probabilité de rejeter H_0 alors qu'elle est effectivement fautive, c'est-à-dire

$$\eta : \Theta_1 \rightarrow [0, 1]$$

$$\theta_1 \mapsto \mathbf{P}_{\theta_1}(\mathcal{R}_{H_0}) = 1 - \beta(\theta_1).$$

- Soit φ_1 et φ_2 deux tests de niveau α . φ_1 est dit **uniformément plus puissant (UPP)** que φ_2 si

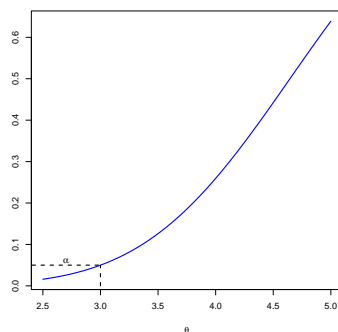
$$\forall \theta_1 \in \Theta_1 \quad \eta_{\varphi_1}(\theta_1) \geq \eta_{\varphi_2}(\theta_1).$$

- Un test φ est dit **UPP** parmi les tests de niveau α si il est de niveau α et si il est UPP que tout test de niveau α .

Un test φ de niveau α est dit **sans biais** si pour tout $\theta_1 \in \Theta_1$ on a $\eta_{\varphi}(\theta_1) \geq \alpha$.

Exemple

- $X \sim \mathcal{N}(\mu, 1)$, $H_0 : \mu = 3$, $H_1 : \mu > 3$, $\alpha = 0.05$.
- $\mathcal{R}_{H_0} = \{x : x > q_{0.95,3,1}\}$.
- $\eta(\mu) = 1 - F_{\mu,1}(q_{0.95,3,1})$ pour $\mu > 3$.



Le test est sans biais.

2.4 Exemples

Tests sur une moyenne - échantillons gaussiens

- X_1, \dots, X_n i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 connu.

H_0	H_1	Stat de test	\mathcal{R}_{H_0}	Propriétés
$\mu = \mu_0$	$\mu > \mu_0$	$\sqrt{n} \frac{X_n - \mu_0}{\sigma}$	$\{x : x > q_{1-\alpha}\}$	UPP
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n} \frac{X_n - \mu_0}{\sigma}$	$\{x : x < q_{\alpha/2} \text{ ou } x > q_{1-\alpha/2}\}$	UPPSB

- X_1, \dots, X_n i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$ avec σ^2 inconnu.

H_0	H_1	Stat de test	\mathcal{R}_{H_0}	Propriétés
$\mu = \mu_0$	$\mu > \mu_0$	$\sqrt{n} \frac{X_n - \mu_0}{S_n}$	$\{x : x > t_{1-\alpha}\}$	UPP
$\mu = \mu_0$	$\mu \neq \mu_0$	$\sqrt{n} \frac{X_n - \mu_0}{S_n}$	$\{x : x < t_{\alpha/2} \text{ ou } x > t_{1-\alpha/2}\}$	UPPSB

Tests sur une variance - échantillons gaussiens

- X_1, \dots, X_n i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$ avec μ connu.

H_0	H_1	Stat de test	\mathcal{R}_{H_0}	Propriétés
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma_0^2}$	$\{x : x > \chi_{1-\alpha}^2(n)\}$	"UPP"
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\sum_{i=1}^n \frac{(X_i - \mu)^2}{\sigma_0^2}$	$\{x : x < \chi_{\alpha/2}^2(n) \text{ ou } x > \chi_{1-\alpha/2}^2(n)\}$	"UPPSB"

- X_1, \dots, X_n i.i.d de loi $\mathcal{N}(\mu, \sigma^2)$ avec μ inconnu.

H_0	H_1	Stat de test	\mathcal{R}_{H_0}	Propriétés
$\sigma^2 = \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma_0^2}$	$\{x : x > \chi_{1-\alpha}^2(n-1)\}$	"UPP"
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\sum_{i=1}^n \frac{(X_i - \bar{X})^2}{\sigma_0^2}$	$\{x : x < \chi_{\alpha/2}^2(n-1) \text{ ou } x > \chi_{1-\alpha/2}^2(n-1)\}$	"UPPSB"

2.5 Comparaison de deux échantillons gaussiens

Exemple

- On dispose de $n_1 = 13$ observations du poids de poulpes femelles et $n_2 = 15$ observations du poids de poulpes males.
- On souhaite vérifier si le sexe a une influence sur le poids.

Modélisation

- X_i v.a. correspondant au poids du $i^{\text{ème}}$ poulpe femelle, $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$.
- Y_i v.a. correspondant au poids du $i^{\text{ème}}$ poulpe male, $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$.
- On souhaite tester les hypothèses $H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$.

Les statistiques de test

	σ_i connus		σ_i inconnus	
	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$	$\sigma_1 = \sigma_2$	$\sigma_1 \neq \sigma_2$
Stat de test	$\frac{\bar{X} - \bar{Y}}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
Lois sous H_0	$\mathcal{N}(0, 1)$	$\mathcal{N}(0, 1)$	$\mathcal{T}(n_1 + n_2 - 2)$	$\simeq \mathcal{N}(0, 1)$

Test de comparaison de variances

- $H_0 : \sigma_1 = \sigma_2$ contre $H_1 : \sigma_1 \neq \sigma_2$. On note

$$\begin{aligned} - \widehat{\sigma}_1^2 &= \frac{1}{n_1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \text{ et } \widehat{\sigma}_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \\ - S_1^2 &= \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \text{ et } S_2^2 = \frac{1}{n_2} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2 \end{aligned}$$

μ_1 et μ_2 connus

Sous H_0 la statistique $\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2}$ suit une loi de Fisher $\mathcal{F}(n_1, n_2)$.

μ_1 et μ_2 inconnus

Sous H_0 la statistique $\frac{S_1^2}{S_2^2}$ suit une loi de Fisher $\mathcal{F}(n_1 - 1, n_2 - 1)$.

Exemple des poulpes avec R

- Test d'égalité des variances.

```
> var.test(Poids~Sexe, conf.level=0.95, data=poulpes)

F test to compare two variances

data: Poids by Sexe
F = 0.2883, num df = 12, denom df = 14, p-value = 0.03713
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.0945296 0.9244467
sample estimates:
ratio of variances
 0.2883299
```

La fonction ne renvoie pas la décision mais une **p-value** (valeur-p en français) également appelée probabilité critique.

Probabilité critique

Définition

La **probabilité critique** est la probabilité que sous H_0 la statistique de test prenne une valeur au moins aussi extrême que celle observée.

Calcul de la pc

On note T la statistique de test et t_{obs} la valeur observée.

- Si la région de rejet est unilatérale, par exemple $\{x : x > c\}$ alors $pc = \mathbf{P}_{H_0}(T > t_{obs})$.
- Si la région de rejet est bilatérale, par exemple $\{x : x > c_1 \text{ ou } x < c_2\}$ alors

$$pc = \begin{cases} 2\mathbf{P}_{H_0}(T > t_{obs}) & \text{si } t_{obs} > M \\ 2\mathbf{P}_{H_0}(T < t_{obs}) & \text{si } t_{obs} < M \end{cases}$$

Interprétation de la probabilité critique

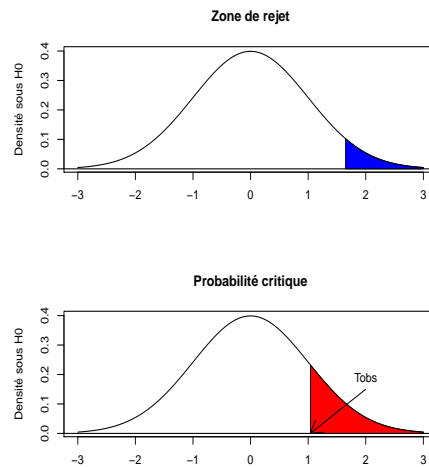
- La probabilité critique correspond au niveau de test minimum pour lequel on rejette H_0 . Ainsi,
 - si $pc \geq \alpha$ l'hypothèse nulle est acceptée au niveau α .
 - si $pc < \alpha$ l'hypothèse nulle est rejetée au niveau α .

En pratique...

- Les logiciels ne renvoient généralement pas la conclusion du test mais la valeur de la probabilité critique.
- La décision est prise en comparant cette valeur au niveau fixé par l'utilisateur.

Exemple pour un test unilatéral

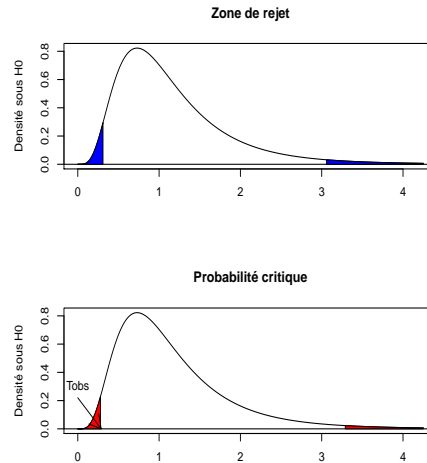
- $H_0 : \mu = \mu_0$ vs $H_1 : \mu > \mu_0$.
- $\alpha = 0.05$.
- $T \sim \mathcal{N}(0, 1)$ sous H_0 .



Conclusion : $pc > \alpha$ donc H_0 est acceptée au profit de H_1 .

Exemple pour un test unilatéral

- Test d'égalité des variances pour les poulpes.
- $H_0 : \sigma_1 = \sigma_2$ vs $H_1 : \sigma_1 \neq \sigma_2$.
- $\alpha = 0.05$.
- $T \sim \mathcal{F}(12, 14)$ sous H_0 .



$pc \leq \alpha$ donc H_0 est rejetée au profit de H_1 . On conclut que la variance de la variable poids diffère selon le sexe.

Comparaison du poids moyen des poulpes

- Pour comparer le poids moyen des poulpes, on fait ainsi un test d'égalité de moyenne avec variances inégales.
- Sur R, on obtient

```
> t.test(Poids~Sexe, alternative="two.sided", conf.level=0.95,
         var.equal=FALSE, data=poulpes)
```

Welch Two Sample t-test

```
data: Poids by Sexe
t = -3.7496, df = 22.021, p-value = 0.001107
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2010.624 -578.607
sample estimates:
mean in group Femelle    mean in group Male
      1405.385             2700.000
```

Au seuil $\alpha = 5\%$, on conclut que le poids des poulpes est différent selon le sexe.

2.6 Cas non gaussien

Question

Que se passe-t-il si les paramètres que l'on souhaite tester ne sont pas gaussiens ?

Une réponse

Dans le cas où le paramètre à tester correspond à l'espérance d'une variable aléatoire, on utilise souvent l'approximation gaussienne via le TCL.

Nous illustrons l'approche via le test de proportions.

Test sur le paramètre d'une loi de Bernoulli

- X_1, \dots, X_n i.i.d de loi de Bernoulli p_0 .
- $H_0 : p = p_0$ contre $H_1 : p \neq p_0$.
- TCL :

$$\sqrt{n} \frac{\hat{p} - p}{\sqrt{p(1-p)}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

- Sous H_0 , on fait l'approximation :

$$T_n = \sqrt{n} \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}} \sim \mathcal{N}(0, 1).$$

- On déduit $\mathcal{R}_{H_0} =] - \infty, -q_{1-\alpha/2}[\cup] q_{1-\alpha/2}, +\infty[$.

Exemple

- Une entreprise emploie 40 hommes et 60 femmes.

Peut-on affirmer que les recruteurs sont sexistes ?

- **Modélisation :** on note p la probabilité de recruter une femme et X_i la va qui prend pour valeur 1 si la $i^{\text{ème}}$ personne recrutée est une femme, 0 sinon.
- $H_0 : p = 0.5$ contre $H_1 : p \neq 0.5$, $\alpha = 0.05$.
- Sous H_0 , la statistique

$$T = \sqrt{100} \frac{\hat{p} - 0.5}{\sqrt{0.5(1-0.5)}}$$

suit (approximativement) une loi $\mathcal{N}(0, 1)$.

- $\mathcal{R}_{H_0} =] - \infty, -1.96[\cup] 1.96, +\infty[$.
- $T_{obs} = 2 \in \mathcal{R}_{H_0}$, on rejette H_0 au niveau 0.05 ($p_c = 0.0455$).

Test de comparaison de deux proportions

- X_1, \dots, X_{n_1} i.i.d de loi $B(p_1)$.
- Y_1, \dots, Y_{n_2} i.i.d de loi $B(p_2)$.
- $H_0 : p_1 = p_2$ contre $H_1 : p_1 \neq p_2$.

- On note $\hat{p} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$ et sous H_0 on approche la loi de

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

par la loi $\mathcal{N}(0, 1)$.

- On rejettera donc H_0 si

$$t_{obs} \in \mathcal{R}_{H_0} =]-\infty, -q_{1-\alpha/2}[\cup]q_{1-\alpha/2}, +\infty[.$$

3 Une introduction aux tests non paramétriques

Motivations

- A un âge donné, on a pu observer que parmi les bébés non prématurés : 50% marchent, 12% ont une ébauche de marche et 38% ne marchent pas.
- Pour le même âge, sur 80 prématurés, on a observé que 35 marchent, 4 ont une ébauche de marche et 41 ne marchent pas.

Les bébés prématurés développent-ils la marche de la même manière que les bébés non-prématurés ?

- Exemple du nombre de garçons qui suit une loi Binomiale ?
- Dépendance entre le fait d'avoir survécu et la classe d'appartenance pour les passagers du Titanic.

Tests non paramétriques

- On peut répondre à ces questions à l'aide de tests statistiques.
- Ici, le problème est de confronter la loi d'une variable à une autre loi, ou encore de tester l'indépendance entre deux variables.
- Les hypothèses ne vont plus porter sur les paramètres de lois de probabilités, c'est pourquoi on parle de **tests non paramétriques**.

La distance du χ^2

- Soit X_1, \dots, X_n n va réelles de loi P . On note $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$.
- Soit $(\mathcal{O}_1, \dots, \mathcal{O}_m)$ une partition de \mathbb{R} , on note $p_k = \mathbf{P}(X \in \mathcal{O}_k)$.
- Soit $N_k = \sum_{i=1}^n \mathbf{1}_{X_i \in \mathcal{O}_k}$.

Définition

La distance du χ^2 entre P et P_n est définie par

$$D(P_n, P) = \sum_{k=1}^m \frac{(N_k - np_k)^2}{np_k}.$$

Théorème

Lorsque $n \rightarrow \infty$, on a

$$D(P_n, P) \xrightarrow{\mathcal{L}} \chi^2(m-1).$$

Remarques

- $D(P_n, P)$ est une sorte de distance entre la loi empirique P_n et la loi théorique P .
- Elle est construite en comparant les effectifs observés N_k aux effectifs théoriques np_k .

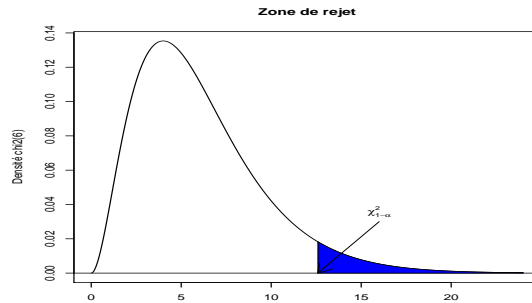
Utile pour tester des hypothèses du genre $H_0 : P = P_0$ contre $H_1 : P \neq P_0$. En effet,

- Sous H_0 , D_n aura tendance à ne pas prendre de trop grande valeurs puisque $N_k/n \xrightarrow{p.s.} p_k$ (LFGN).
- Sous H_1 , D_n prendra de fortes valeurs puisque $D_n \xrightarrow{p.s.} \infty$.

3.1 Le test du χ^2 d'adéquation

Le test d'adéquation du χ^2

- $H_0 : P = P_0$ contre $H_1 : P \neq P_0$.
- Sous H_0 la statistique $D(P_n, P_0) \stackrel{\text{approx}}{\sim} \chi^2(m-1)$.
- $\mathcal{R}_{H_0} =]\chi_{1-\alpha}^2(m-1), +\infty[$.



Remarque

Le test étant asymptotique, il est recommandé de l'appliquer pour $n > 30$ et $np_k^0 > 5$.

Exemple

- X : "Marche à l'âge donné" pour les prématurés. 3 modalités (oui, ébauche, non).
- $H_0 : X \sim P_0$ contre $H_1 : X \not\sim P_0$ avec $P_0(\text{oui}) = 0.5$, $P_0(\text{ébauche}) = 0.12$, $P_0(\text{non}) = 0.38$. Risque $\alpha = 0.05$.
- Sous H_0 la statistique

$$D_n = \sum_{k=1}^3 \frac{(N_k - np_k)^2}{np_k} \underset{\text{approx}}{\sim} \chi^2(2).$$

- $\mathcal{R}_{H_0} =]5.991, +\infty[$.

	oui	ébauche	non	Total
obs	35	4	41	80
théo	40	9.6	30.4	80
écart	0.625	3.267	3.696	7.588

- $D_{obs} \in \mathcal{R}_{H_0}$, on rejette H_0 au risque 5% ($pc = 1 - F_{\chi^2_2}(7.588) = 0.0225$).
- Sur **R**, on peut réaliser le test avec les commandes

```
> x <- c(35, 4, 41)
> p0 <- c(0.5, 0.12, 0.38)
> chisq.test(x, p=p0)
```

Chi-squared test for given probabilities

```
data: x
X-squared = 7.5877, df = 2, p-value = 0.02251
```

Compléments

Propriété

Si la loi P_0 dépend de r paramètres, alors ces paramètres sont estimés (MV par exemple) et dans ce cas, sous H_0 $D(P_n, P_0) \xrightarrow{\mathcal{L}} \chi_{m-r-1}^2$.

Exemple du nombre de garçons

- $H_0 : X \sim B(4, p)$ contre $H_1 : X \approx B(4, p)$.
- Estimateur de p : $\hat{p} = 0.4925$.

	0	1	2	3	4	Total
obs	3	30	39	23	5	100
théo	6.63	25.7	37.48	24.2	5.8	100

$D_{obs} = 2.95 \notin]7.81, +\infty[$. On accepte H_0 au risque 5%.

3.2 Le test du χ^2 d'indépendance

Notations

- Soient X et Y deux variables aléatoires à valeurs dans E et F . On souhaite tester au niveau α les hypothèses $H_0 : X$ et Y sont indépendantes contre $H_1 : X$ et Y ne sont pas indépendantes.
- On se donne (E_1, \dots, E_I) et (F_1, \dots, F_J) deux partitions de E et F .
- On dispose de n mesures du couple (X, Y) et on désigne par N_{ij} l'effectif observé dans la classe $E_i \times F_j$.

	F_1	...	F_j	...	F_J	Total
E_1	N_{11}	...	N_{1j}	...	N_{1J}	$N_{1\bullet}$
\vdots						\vdots
E_i	N_{i1}	...	N_{ij}	...	N_{iJ}	$N_{i\bullet}$
\vdots						\vdots
E_I	N_{I1}	...	N_{Ij}	...	N_{IJ}	$N_{I\bullet}$
Total	$N_{\bullet 1}$...	$N_{\bullet j}$...	$N_{\bullet J}$	n

Le test

Propriété

Sous H_0 la statistique

$$X_n = \sum_{i=1}^I \sum_{j=1}^J \frac{(N_{i\cdot} N_{\cdot j} - N_{ij})^2}{\frac{N_{i\cdot} N_{\cdot j}}{n}}$$

converge en loi vers la loi $\chi_{(I-1)(J-1)}^2$.

- Au niveau α , on rejettera l'hypothèse H_0 si X_{obs} est supérieure au quantile d'ordre $1 - \alpha$ de la loi du $\chi_{(I-1)(J-1)}^2$.
- Le test étant asymptotique, il faudra s'assurer en pratique que $n > 30$ et $\frac{N_{i\cdot} N_{\cdot j}}{n} > 5$.

L'exemple du Titanic

- X : survécu ou pas, Y : classe.
- H_0 : X et Y sont indépendantes contre H_1 : X et Y ne sont pas indépendantes.

		Effectifs observés				
		1	2	3	E	Total
oui	203	118	178	212	711	
non	122	167	528	673	1490	
Total	325	285	706	885	2201	

		Effectifs Théoriques				
		1	2	3	E	Total
oui	105	92	228	286	711	
non	220	192	478	599	1490	
Total	325	285	706	885	2201	

$X_{obs} = 190.4011 > 0.352 = \chi_{0.95}^2(3)$, l'hypothèse nulle est donc (clairement !) rejetée au niveau α .

Part IV

Le modèle de régression linéaire

1 Introduction

Exemple

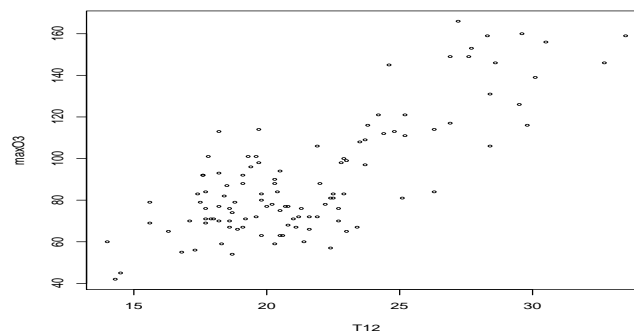
- On cherche à **expliquer** ou à **prédire** la concentration en ozone.
- On dispose de $n = 112$ observations de la concentration en ozone ainsi que de 12 autres variables susceptibles d'expliquer cette concentration :
 - Température relevée à différents moments de la journée.
 - Indice de nébulosité relevé à différents moments de la journée.
 - Direction du vent.
 - Puie.

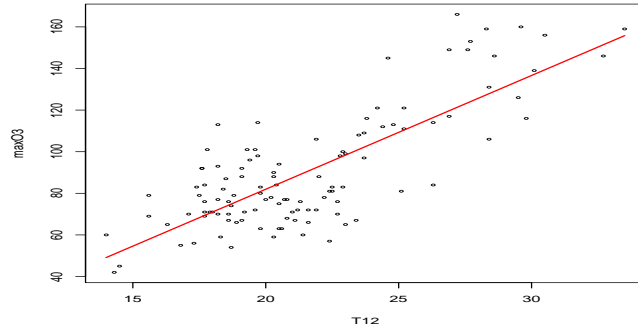
Question

Comment expliquer (modéliser) la concentration en ozone à l'aide de toutes ces variables ?

Ozone en fonction de la température à 12h ?

MaxO3	87	82	92	114	94	80	...
T12	18.5	18.4	17.6	19.7	20.5	19.8	...

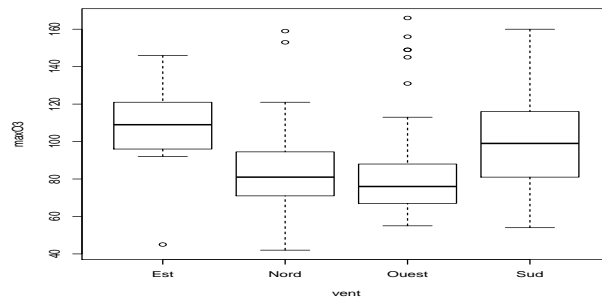




Comment ajuster le nuage de points ?

Ozone en fonction du vent ?

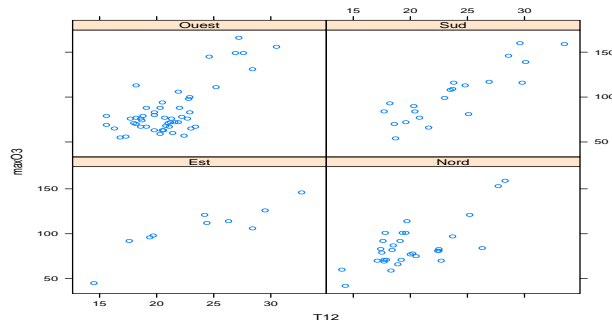
MaxO3	87	82	92	114	94	80	...
Vent	Nord	Nord	Est	Nord	Ouest	Ouest	...



$$MaxO3 \approx \alpha_1 \mathbf{1}_{vent=est} + \dots + \alpha_4 \mathbf{1}_{vent=sud}.$$

$$\alpha_j = ???$$

Ozone en fonction de la température à 12h et du vent



$$\max O3 \approx \begin{cases} \beta_{01} + \beta_{11}T12 & \text{si vent=est} \\ \vdots & \vdots \\ \beta_{04} + \beta_{14}T12 & \text{si vent=ouest} \end{cases}$$

Autre modélisation

- Généralisation

$$\max O3 \approx \beta_0 + \beta_1 V_1 + \dots + \beta_{12} V_{12}$$

Questions

- Comment calculer (ou plutôt **estimer**) les paramètres β_j ?
- Le modèle avec les 12 variables est-il "meilleur" que des modèles avec moins de variables ?
- Comment trouver le "meilleur" sous-groupe de variables ?

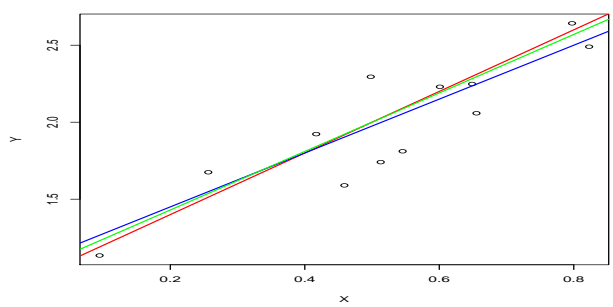
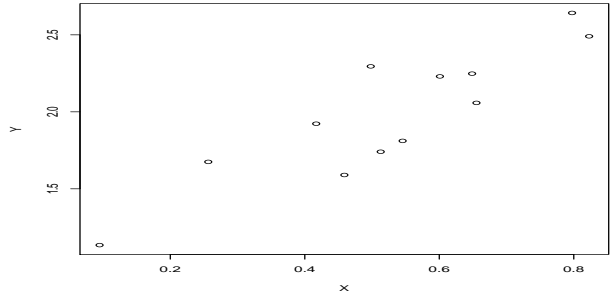
2 La régression linéaire simple

2.1 Ajustement par moindres carrés

Ajustement linéaire d'un nuage de points

Notations

- n observations y_1, \dots, y_n de la **variable à expliquer** (maxO3).
- n observations x_1, \dots, x_n de la **variable explicative** (T12).

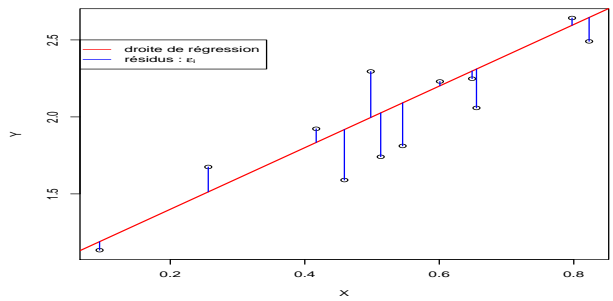


Problème

Trouver la droite qui ajuste "au mieux" le nuage de points $(x_i, y_i)_{i=1, \dots, n}$.

- On cherche $y = \beta_0 + \beta_1 x$ qui ajuste au mieux le nuage des points.
- Toutes les observations mesurées ne se trouvent pas sur une droite :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$



Idée

Chercher à minimiser les **erreurs** ou les **bruits** ε_i .

Le critère des moindres carrés

Critère des MC

On cherche (β_0, β_1) qui minimisent

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Solution

La solution est donnée par :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{et} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

à condition que tous les x_i ne soient pas égaux.

2.2 Propriétés des estimateurs

Début de modélisation statistique

- L'idée sous-jacente est que la variable Y est liée à X par une relation linéaire ou quasi-linéaire.
- Sur les observations, la linéarité n'est généralement pas "parfaite".
- **Hypothèse** : cet "écart" à la linéarité est la conséquence de phénomène que l'on ne peut contrôler de manière déterministe (**phénomènes aléatoires**).

Les erreurs $\varepsilon_i, i = 1, \dots, n$ sont des variables aléatoires.

Le modèle de régression linéaire (VI)

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

avec

- $\varepsilon_1, \dots, \varepsilon_n$ n variables aléatoires indépendantes.

Conséquence

- Y_1, \dots, Y_n sont n variables aléatoires indépendantes.
- Qu'en est-il pour les x_i ?
 - quantités déterministes ?
 - quantités aléatoires ?
- L'étude des propriétés statistiques du modèle linéaire est quasiment identique selon la nature des x_i , nous les supposons déterministes dans la suite.

Premières propriétés

Rappels

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \quad \text{et} \quad \hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Propriétés

Sous les hypothèses :

1. $\mathcal{H}_1 : \mathbf{E}(\varepsilon_i) = 0$ pour $i = 1, \dots, n$.
2. $\mathcal{H}_2 : \mathbf{V}(\varepsilon_i) = \sigma^2$ pour $i = 1, \dots, n$.

on a

1. $\hat{\beta}_0$ et $\hat{\beta}_1$ sont des estimateurs sans biais.
2. Les variances sont données par

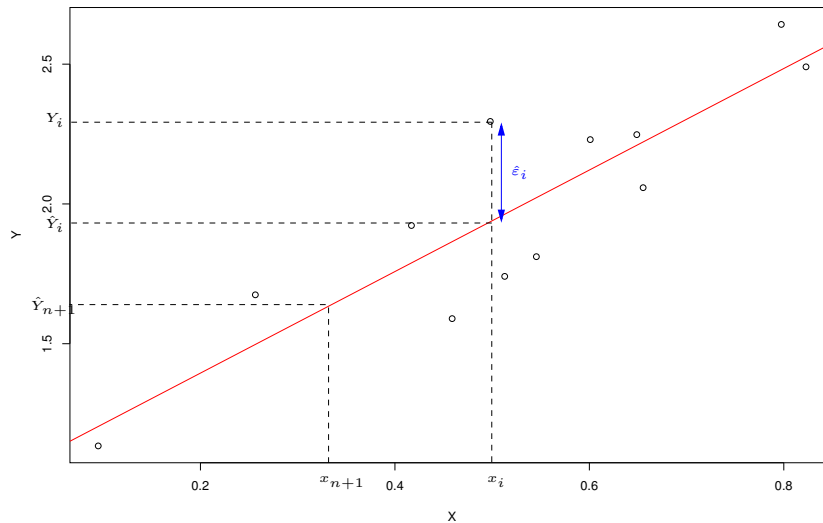
$$\mathbf{V}(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \mathbf{V}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Résidus

Vocabulaire

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: valeur ajustée de Y_i par le modèle.
- $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$ résidu.
- Etant donné x_{n+1} une nouvelle valeur de la variable X , cherche à estimer $y_{n+1} = \beta_0 + \beta_1 x_{n+1}$.
- Un estimateur naturel est la prévision associée à cette nouvelle observation \hat{Y}_{n+1}
:

$$\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}.$$



Propriété

On a

$$\mathbf{V}(\hat{Y}_{n+1}) = \sigma^2 \left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right).$$

La variance de la prévision est d'autant plus faible que :

- σ^2 est petit (on pouvait s'y attendre...).
- x_{n+1} est proche du centre de gravité des x_i (plus difficile de bien prédire vers les extrêmes).

Questions

1. Comment mesurer la qualité du modèle ?
2. Comment tester la valeur des coefficients du modèle ?
3. Peut-on obtenir des intervalles de confiance pour les paramètres β_j ou pour la prévision \hat{Y}_{n+1} ?

2.3 Inférence statistique

Hypothèse de normalité

- Pour pouvoir faire de l'inférence, il nous faut mettre une loi sur la variable à expliquer.

\mathcal{H}_2 : les variables aléatoires ε_i suivent une loi $\mathcal{N}(0, \sigma^2)$.

Remarque

Cette hypothèse revient à supposer que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Elle nous amène donc à considérer le modèle paramétrique $(\mathbb{R}, \{\mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), (\beta_0, \beta_1) \in \mathbb{R}^2\})$.

Lois des estimateurs

σ^2 connu

1. $\hat{\beta}_0 \sim \mathcal{N}(\beta_0, \sigma_{\hat{\beta}_0}^2)$ et $\hat{\beta}_1 \sim \mathcal{N}(\beta_1, \sigma_{\hat{\beta}_1}^2)$

σ^2 inconnu

On pose $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{\|\hat{\varepsilon}\|^2}{n-2}$. On a

1. $(n-2) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{(n-2)}^2$.
2. $(\hat{\beta}_0, \hat{\beta}_1)$ et $\hat{\sigma}^2$ sont indépendants.
3. $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim \mathcal{T}_{n-2}$.
4. $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{T}_{n-2}$.

Il est alors "facile" de construire des intervalles de confiance sur les paramètres ainsi que des tests d'hypothèses du genre $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Exemple de l'ozone

```
> reg.simple <- lm(maxO3~T12,data=donnees)
> summary(reg.simple)

Call:
lm(formula = maxO3 ~ T12, data = donnees)

Residuals:
    Min       1Q   Median       3Q      Max
-38.0789 -12.7352  0.2567  11.0029  44.6714

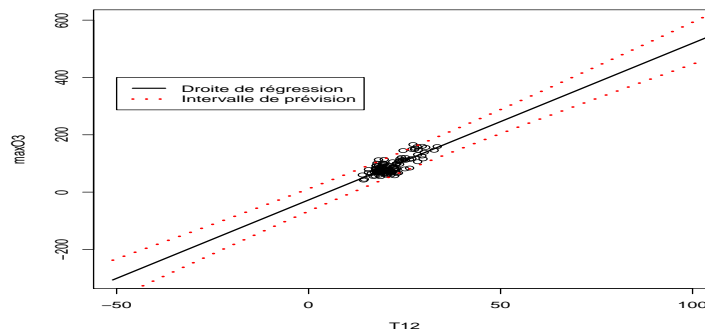
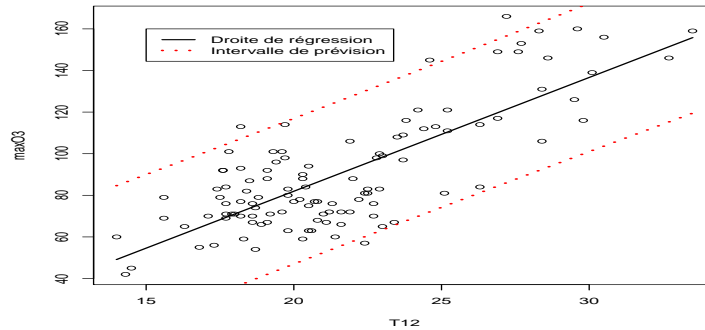
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.4196     9.0335  -3.035  0.003 **
T12           5.4687     0.4125  13.258 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.57 on 110 degrees of freedom
Multiple R-squared:  0.6151, Adjusted R-squared:  0.6116
F-statistic: 175.8 on 1 and 110 DF,  p-value: < 2.2e-16
```

Intervalle de confiance de prévision

IC pour Y_{n+1}

$$\left[\hat{Y}_{n+1} \pm t_{n-2}(1 - \alpha/2) \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right].$$



3 La régression multiple

3.1 Notations et modélisation

Introduction

- La température à 12h n'est pas la seule variable permettant d'**expliquer** ou de **prédire** la concentration en ozone.
- D'autres variables doivent être prise en compte (nébulosité, force et direction du vent...)
- Nécessité d'étendre le modèle linéaire à plus d'une variable explicative.

Notations

- Y : variable (aléatoire) à expliquer à valeurs dans \mathbb{R} .
- X_1, \dots, X_p : p variables explicatives à valeurs dans \mathbb{R} .

- n observations $(x_1, Y_1), \dots, (x_n, Y_n)$ avec $x_i = (x_{i1}, \dots, x_{ip})'$.

Le modèle de régression linéaire multiple

Le modèle s'écrit :

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$$

où les erreurs aléatoires ε_i sont i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

Ecriture matricielle

- On note

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Ecriture matricielle

Le modèle se réécrit

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon$$

où $\varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n)$.

3.2 Estimateur des moindres carrés

Estimateurs des moindres carrés

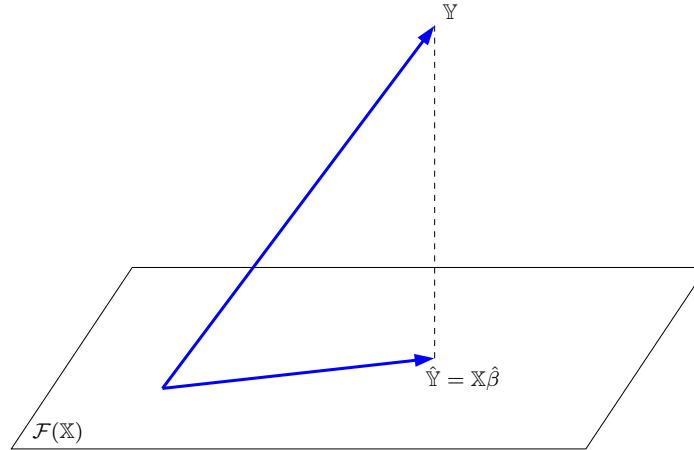
Définition

On appelle **estimateur des moindres carrés** $\hat{\beta}$ de β la statistique suivante :

$$\hat{\beta} = \underset{\beta_0, \dots, \beta_p}{\operatorname{argmin}} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2 = \underset{\beta \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

- On note $\mathcal{F}(\mathbb{X})$ le s.e.v. de \mathbb{R}^n de dimension $p+1$ engendré par les $p+1$ colonnes de \mathbb{X} .
- Chercher l'estimateur des moindres carrés revient à minimiser la distance entre $\mathbb{Y} \in \mathbb{R}^n$ et $\mathcal{F}(\mathbb{X})$.

Représentation géométrique



Expression de l'estimateur des moindres carrés

- On déduit que $\mathbb{X}\hat{\beta}$ est le projeté orthogonal de \mathbb{Y} sur $\mathcal{F}(\mathbb{X})$:

$$\mathbb{X}\hat{\beta} = \mathbf{P}_{\mathcal{F}(\mathbb{X})}(\mathbb{Y}) = \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

Théorème

Si la matrice \mathbb{X} est de plein rang, l'estimateur des MC est donné par :

$$\hat{\beta} = (\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'\mathbb{Y}.$$

3.3 Propriétés statistiques

Propriété

1. $\hat{\beta}$ est un estimateur sans biais de β .
2. La matrice de variance-covariance de $\hat{\beta}$ est donnée par

$$\mathbf{V}(\hat{\beta}) = \sigma^2(\mathbb{X}'\mathbb{X})^{-1}.$$

3. $\hat{\beta}$ est VUMSB.

Remarque

La log-vraisemblance du modèle $(\mathbb{R}^n, \{\mathcal{N}(x_i'\beta, \sigma^2)\}^{\otimes n}, \beta \in \mathbb{R}^{p+1})$ est donnée par

$$\mathcal{L}(y_1, \dots, y_n; \beta) = \frac{n}{2} \log(\sigma^2) - \frac{n}{2} \log(2\pi) - \frac{1}{2\sigma^2} \|\mathbb{Y} - \mathbb{X}\beta\|^2.$$

Conclusion : l'estimateur du maximum de vraisemblance $\hat{\beta}_{MV}$ coïncide avec l'estimateur des moindres carrés $\hat{\beta}$.

Loi des estimateurs

- Soit $\hat{\varepsilon} = \mathbb{Y} - \hat{\mathbb{Y}}$ le vecteur des résidus et $\hat{\sigma}^2$ l'estimateur de σ^2 défini par

$$\hat{\sigma}^2 = \frac{\|\hat{\varepsilon}\|^2}{n - (p + 1)}.$$

Proposition

1. $\hat{\beta}$ est un vecteur gaussien d'espérance β et de matrice de variance-covariance $\sigma^2(\mathbb{X}'\mathbb{X})^{-1}$.
2. $(n - (p + 1))\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-(p+1)}^2$.
3. $\hat{\beta}$ et $\hat{\sigma}^2$ sont indépendantes.

Intervalle de confiance et tests

Corollaire

On note $\hat{\sigma}_j^2 = \hat{\sigma}^2[\mathbb{X}'\mathbb{X}]_{jj}^{-1}$ pour $j = 0, \dots, p$. On a

$$\forall j = 0, \dots, p, \quad \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_j} \sim \mathcal{T}(n - (p + 1)).$$

On déduit de ce corollaire :

- des intervalles de confiance de niveau $1 - \alpha$ pour β_j .
- des procédures de test pour des hypothèses du genre $H_0 : \beta_j = 0$ contre $H_1 : \beta_j \neq 0$.

Prévision

- On dispose d'une nouvelle observation $x_{n+1} = (x_{n+1,1}, \dots, x_{n+1,p})$ et on souhaite prédire la valeur $y_{n+1} = x'_{n+1}\beta$ associée à cette nouvelle observation.

- Un estimateur (naturel) de y_{n+1} est $\hat{y}_{n+1} = x'_{n+1}\hat{\beta}$.
- Un intervalle de confiance de niveau $1 - \alpha$ pour y_{n+1} est donné par

$$\left[\hat{y}_{n+1} \pm t_{n-(p+1)}(\alpha/2)\hat{\sigma} \sqrt{x'_{n+1}(\mathbb{X}'\mathbb{X})^{-1}x_{n+1} + 1} \right].$$

Exemple de l'ozone

- On considère le modèle de régression multiple :

$$MaxO3 = \beta_0 + \beta_1 T_{12} + \beta_2 T_{15} + \beta_3 N_{12} + \beta_4 V_{12} + \beta_5 MaxO3v + \varepsilon.$$

```
> reg_multi <- lm(maxO3~T12+T15+Ne12+Vx12+maxO3v, data=donnees)
> summary(reg_multi)

Call:
lm(formula = maxO3 ~ T12 + T15 + Ne12 + Vx12 + maxO3v, data = donnees)

Residuals:
    Min       1Q   Median       3Q      Max
-54.216  -9.446  -0.896   8.007  41.186

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.04498    13.01591   0.234  0.8155
T12          2.47747     1.09257   2.268  0.0254 *
T15          0.63177     0.96382   0.655  0.5136
Ne12        -1.83560     0.89439  -2.052  0.0426 *
Vx12         1.33295     0.58168   2.292  0.0239 *
maxO3v       0.34215     0.05989   5.713 1.03e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.58 on 106 degrees of freedom
Multiple R-squared:  0.7444, Adjusted R-squared:  0.7324
F-statistic: 61.75 on 5 and 106 DF,  p-value: < 2.2e-16
```

4 Validation et choix de modèles

Questions

- Le modèle linéaire repose sur certaines hypothèses (normalité des erreurs par exemple), comment les vérifier ?
- A partir de p variables explicatives, il est possible de construire (au moins) 2^p modèles linéaires, comment choisir le "meilleur" sous-ensemble de variables à inclure dans le modèle ?

4.1 Résidus et coefficient de détermination

Analyse des résidus

- Le modèle linéaire

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_n).$$

- Les erreurs ε sont inconnues. On les estime par $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$.

Propriété

$$\mathbf{E}(\hat{\varepsilon}_i) = 0 \quad \text{et} \quad \mathbf{V}(\hat{\varepsilon}_i) = \sigma^2(\mathbf{I} - \mathbf{P}_{\mathcal{F}(\mathbb{X})}).$$

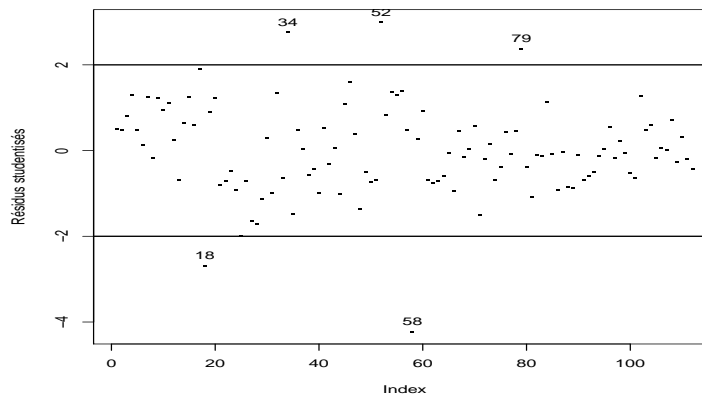
Conséquence

Un moyen de vérifier l'hypothèse de normalité des erreurs est de comparer la distribution des **résidus studentisés**

$$\frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

à la distribution gaussienne centrée réduite.

Tracé d'un index plot



L'analyse de ce type de graphique est généralement accompagnée d'un test d'adéquation des résidus à la loi normale (test de Shapiro-Wilks par exemple).

Le coefficient de détermination R^2

Equation d'analyse de la variance

On a d'après Pythagore :

$$\begin{aligned} \|Y - \bar{y}\mathbf{1}\|^2 &= \|\hat{Y} - \bar{y}\mathbf{1}\|^2 + \|\hat{\varepsilon}\|^2 \\ SCT &= SCE + SCR \end{aligned}$$

Coefficient de détermination R^2

$$R^2 = \frac{\text{V. expliquée par le modèle}}{\text{V. totale}} = \frac{\|\hat{Y} - \bar{y}\mathbf{1}\|^2}{\|Y - \bar{y}\mathbf{1}\|^2} = \frac{SCE}{SCT}.$$

- $0 \leq R^2 \leq 1$.
- Si $R^2 = 1$, la variabilité est entièrement expliquée par le modèle.
- Si $R^2 = 0$, la variabilité se trouve dans la résiduelle (ce qui n'est pas très bon...).

4.2 Tests entre modèles emboîtés

Le test de Fisher global

- Modèle

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- Hypothèses

$$H_0 : \beta_1 = \dots = \beta_p = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \dots, p\} \beta_j \neq 0.$$

- Sous H_0 ,

$$F = \frac{R^2}{1 - R^2} \frac{n - (p + 1)}{p} \sim \mathcal{F}_{p, n - (p + 1)}.$$

- On rejette H_0 si $F_{obs} > F_{p, n - (p + 1)}(1 - \alpha)$.

Généralisation

- On souhaite tester le modèle \mathcal{M}_1

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

contre le modèle \mathcal{M}_0

$$Y_i = \beta_q x_{iq} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2).$$

- Cela revient à tester si les q premiers coefficients de \mathcal{M}_1 sont nuls

$$H_0 : \beta_0 = \dots = \beta_{q-1} = 0 \quad \text{contre} \quad H_1 : \exists j \in \{0, \dots, q-1\} \beta_j \neq 0.$$

On parle de test entre modèles emboîtés car \mathcal{M}_0 est un cas particulier de \mathcal{M}_1 .

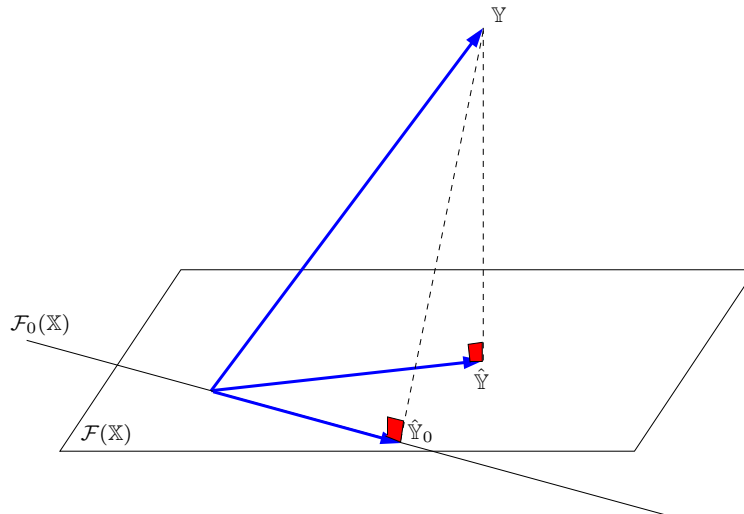
Statistique de test

Idée

On note :

- $\mathcal{F}(\mathbb{X})$ le s.e.v de dimension $p + 1$ engendré par les colonnes de \mathbb{X} et $\hat{\mathbb{Y}}$ la projection de \mathbb{Y} sur \mathcal{F} .
- $\mathcal{F}_0(\mathbb{X})$ le s.e.v de dimension $p - q + 1$ engendré par les $p - q + 1$ colonnes de \mathbb{X} et $\hat{\mathbb{Y}}_0$ la projection de \mathbb{Y} sur \mathcal{F}_0 .

- $\mathcal{F}_0(\mathbb{X}) \subset \mathcal{F}(\mathbb{X})$, l'idée consiste à regarder si \hat{Y}_0 est "proche" de \hat{Y} :
 - si $\hat{Y}_0 \approx \hat{Y}$, on choisira le modèle \mathcal{M}_0 (on acceptera H_0).
 - Sinon, on choisira le modèle \mathcal{M}_1 (on rejettera H_0).



Le test

Cochran...

Sous H_0 ,

$$F = \frac{\|\hat{Y}_0 - \hat{Y}\|^2/q}{\|\mathbb{Y} - \hat{Y}\|^2/(n - (p + 1))} \sim \mathcal{F}_{q, n - (p + 1)}.$$

- On rejette H_0 si $F_{obs} > F_{q, n - (p + 1)}(1 - \alpha)$.

Exemple

- On considère le modèle

$$\text{MaxO3} = \beta_0 + \beta_1 T_{12} + \beta_2 T_{15} + \beta_3 N_{12} + \beta_4 V_{12} + \beta_5 \text{MaxO3v} + \varepsilon$$

et on teste $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$ contre $H_1 : \exists j \in \{2, 3, 4\} : \beta_j \neq 0$.

```
> reg0 <- lm(maxO3~T12+maxO3v, data=donnees)
> reg1 <- lm(maxO3~T12+T15+Ne12+Vx12+maxO3v, data=donnees)
> anova(reg0, reg1)
Analysis of Variance Table

Model 1: maxO3 ~ T12 + maxO3v
Model 2: maxO3 ~ T12 + T15 + Ne12 + Vx12 + maxO3v
  Res.Df  RSS Df Sum of Sq  F    Pr(>F)
1     109 26348
2     106 22540  3    3808.4  5.97 0.000844 ***
```

$pc = 0.000844 > 0.05$, on rejette H_0 et on conserve le modèle à 5 variables par rapport au modèle à 2 variables.

5 Analyse de la variance

5.1 Modèle à un facteur

Motivations

- Jusqu'à présent, les variables explicatives étaient quantitatives.
- Comment généraliser le modèle linéaire à des variables explicatives qualitatives.

Exemple

Comment expliquer la variable `maxO3` par la variable `vent` (ou pluie) ?

Le modèle

- Y variable à expliquer (quantitative) et X variable explicative (qualitative) à J modalités, niveaux ou facteurs.
- On dispose de n observations. Soit n_j le nombre d'individus pour lesquels on a observé la $j^{\text{ème}}$ modalité de X .
- On ordonne les individus selon les modalités de X :

Y	Y_{11}	\dots	Y_{1n_1}	Y_{21}	\dots	\dots	Y_{J,n_J}
X	M_1	\dots	M_1	M_2	\dots	\dots	M_J

Écriture 1

Le modèle d'ANOVA à un facteur s'écrit

$$Y_{ij} = \alpha_j + \varepsilon_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, J,$$

avec ε_{ij} i.i.d. et de loi $\mathcal{N}(0, \sigma^2)$.

Écriture matricielle

- n observations $(x_1, Y_1), \dots, (x_n, Y_n)$ et on écrit le modèle

$$Y_i = \beta_1 \mathbf{1}_{x_i=M_1} + \beta_2 \mathbf{1}_{x_i=M_2} + \dots + \beta_J \mathbf{1}_{x_i=M_J} + \varepsilon_i$$

où ε_i i.i.d. de loi $\mathcal{N}(0, \sigma^2)$.

- Si on note

$$\mathbb{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \mathbb{X} = \begin{pmatrix} \mathbf{1}_{x_1=M_1} & \cdots & \mathbf{1}_{x_1=M_J} \\ \vdots & & \vdots \\ \mathbf{1}_{x_n=M_1} & \cdots & \mathbf{1}_{x_n=M_J} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_J \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

On a alors l'écriture matricielle

$$\mathbb{Y} = \mathbb{X}\beta + \varepsilon.$$

- Il s'agit d'un modèle de régression multiple pour une matrice \mathbb{X} particulière.
- Tout ce qui a été vu dans les sections précédentes (estimation, tests, résidus...) peut s'appliquer à ce nouveau modèle.

Un exemple

```
> reg <- lm(maxO3~vent, data=donnees)
> reg
Call:
lm(formula = maxO3 ~ vent, data = donnees)
Coefficients:
(Intercept)   ventNord   ventOuest   ventSud
  105.600      -19.471      -20.900      -3.076
```

Remarque importante

- La fonction renvoie une estimation pour une constante et pas d'estimation pour la modalité Est.
- Le modèle ajusté est le suivant :

$$Y_i = \beta_0 + \beta_1 \mathbf{1}_{x_i=Est} + \beta_2 \mathbf{1}_{x_i=Nord} + \beta_3 \mathbf{1}_{x_i=Ouest} + \beta_4 \mathbf{1}_{x_i=Sud} + \varepsilon_i$$

muni de la contrainte $\beta_1 = 0$.

- Une telle écriture est moins intuitive mais donne lieu à d'importantes généralisations.

Test de "significativité" du facteur

- On souhaite savoir si la variable X a une influence sur Y : la direction du vent a-t-elle une influence sur la concentration en ozone ?
- On teste donc $H_0 : \beta_1 = \dots = \beta_4 = 0$ contre $H_1 : \exists j \in \{1, \dots, 4\} : \beta_j \neq 0$ pour le modèle

$$Y_i = \beta_0 + \beta_1 \mathbf{1}_{x_i=Est} + \beta_2 \mathbf{1}_{x_i=Nord} + \beta_3 \mathbf{1}_{x_i=Ouest} + \beta_4 \mathbf{1}_{x_i=Sud} + \varepsilon_i.$$

- Il suffit de reprendre le test de Fisher vu dans la partie précédente.

```
> anova(reg)
Analysis of Variance Table

Response: maxO3
      Df Sum Sq Mean Sq F value Pr(>F)
vent    3  7586 2528.69   3.3881 0.02074 *
Residuals 108  80606  746.35
```

$pc = 0.02074 < 0.05$, on rejette H_0 . On peut conclure au risque $\alpha = 5\%$ que la direction du vent a une influence sur la concentration en ozone.

5.2 ANOVA à deux facteurs

Un exemple

- On souhaite expliquer maxO3 par la direction du vent et la pluie.
- On dispose de n observations (x_i, Y_i) et on écrit le modèle

$$Y_i = \beta_0 + \beta_1 \mathbf{1}_{x_{i1}=\text{Est}} + \beta_2 \mathbf{1}_{x_{i1}=\text{Nord}} + \beta_3 \mathbf{1}_{x_{i1}=\text{Ouest}} + \beta_4 \mathbf{1}_{x_{i1}=\text{Sud}} \\ + \beta_5 \mathbf{1}_{x_{i2}=\text{pluie}} + \beta_6 \mathbf{1}_{x_{i2}=\text{Sec}} + \varepsilon$$

muni des contraintes $\beta_1 = 0$ et $\beta_5 = 0$.

```
> reg <- lm(maxO3~vent+pluie,data=donnees)
> reg

Call:
lm(formula = maxO3 ~ vent + pluie, data = donnees)

Coefficients:
(Intercept)   ventNord   ventOuest   ventSud   pluieSec
  85.123      -16.333      -12.709      -2.101      25.597
```

Test de signicativité des facteurs

- Une fois de plus, c'est le test de Fisher qui nous permet de tester la signicativité des variables explicatives dans le modèle.
- On réalise les deux tests de Fisher pour les hypothèses :

$$H_0 : \beta_1 = \dots = \beta_4 = 0 \quad \text{contre} \quad H_1 : \exists j \in \{1, \dots, 4\} : \beta_j \neq 0.$$

$$H_0 : \beta_5 = \beta_6 = 0 \quad \text{contre} \quad H_1 : \exists j \in \{5, 6\} : \beta_j \neq 0.$$

```
> anova(reg)
Analysis of Variance Table

Response: maxO3
      Df Sum Sq Mean Sq F value    Pr(>F)
vent    3  7586 2528.7   4.1984 0.007514 **
pluie    1 16159 16159.4 26.8295 1.052e-06 ***
Residuals 107  64446  602.3
```

Conclusion

- Les variables explicatives quantitatives et qualitatives ont été traitées séparément dans ce chapitre.
- Bien évidemment, en pratique il convient de les traiter ensemble (il suffit d'écrire correctement la partie quantitative et la partie qualitative du modèle).

Exemple

$$\text{maxO3} = \beta_0 + \beta_1 T_{12} + \beta_2 N e_{15} + \beta_3 \mathbf{1}_{\text{pluie}} + \beta_4 \mathbf{1}_{\text{sec}} + \varepsilon$$

muni de la contrainte $\beta_3 = 0$.

```
> reg <- lm(maxO3~T12+Ne15+pluie, data=donnees)
> reg

Call:
lm(formula = maxO3 ~ T12 + Ne15 + pluie, data = donnees)

Coefficients:
(Intercept)      T12      Ne15  pluieSec
   -5.978      4.594     -1.613      8.413
```

5.3 Sélection de modèles

Un exemple de sélection de variables

```
> reg <- lm(maxO3~., data=donnees)
> anova(reg)
Analysis of Variance Table

Response: maxO3
      Df Sum Sq Mean Sq F value    Pr(>F)
T9      1  43138   43138  205.0160 < 2.2e-16 ***
T12     1  11125   11125   52.8706 9.165e-11 ***
T15     1    876    876    4.1619 0.0440614 *
Ne9     1   3244   3244   15.4170 0.0001613 ***
Ne12    1    232    232    1.1035 0.2961089
Ne15    1     5     5    0.0248 0.8752847
Vx9     1   2217   2217   10.5355 0.0016079 **
Vx12    1     1     1    0.0049 0.9443039
Vx15    1    67    67    0.3186 0.5737491
maxO3v  1   6460   6460   30.6993 2.584e-07 ***
vent    3    234    78    0.3709 0.7741473
pluie   1   183   183    0.8694 0.3534399
Residuals 97  20410   210
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Méthode pas à pas

```
> reg1 <- step(reg, direction="backward")
> anova(reg1)
Analysis of Variance Table

Response: maxO3
      Df Sum Sq Mean Sq F value    Pr(>F)
T12     1  54244   54244  276.777 < 2.2e-16 ***
Ne9     1   3579   3579   18.260 4.193e-05 ***
Vx9     1   2035   2035   10.385 0.001684 **
maxO3v  1   7364   7364   37.572 1.499e-08 ***
Residuals 107  20970   196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```