STATISTIQUE THÉORIQUE ET APPLIQUÉE

Tome 2

Inférence statistique
à une et à deux dimensions

Pierre Dagnelie

INTRODUCTIONS DES DIFFÉRENTS CHAPITRES

Bruxelles, De Boeck, 2011, 736 p.

ISBN 978-2-8041-6336-5

De Boeck Services, Fond Jean-Pâques 4, B-1348 Louvain-la-Neuve (Belgique)
Tél. : 32 (0)10 48 25 00 – Fax : 32 (0)10 48 25 19
E-mail : commande@deboeckservices.com – Site web : superieur.deboeck.com

Chapitre 1

Le choix d'une méthode d'analyse statistique

Sommaire

⊕ **1.1 Introduction**

⊕ **1.2 Les facteurs de choix d'une méthode d'analyse statistique**

⊕ **1.3 Un canevas général de choix d'une méthode d'analyse statistique**

1.1 Introduction

Le choix d'une méthode d'analyse statistique bien adaptée à une situation donnée est un problème d'autant plus délicat et d'autant plus important que les logiciels statistiques actuels offrent à leurs utilisateurs des solutions et des options toujours plus nombreuses et plus diversifiées. C'est aussi un sujet difficile à traiter d'une manière générale et d'ailleurs très peu développé dans la plupart des ouvrages de statistique appliquée.

Nous nous efforcerons cependant de fournir à ce propos un certain nombre d'indications utiles, en passant en revue les *principaux facteurs de choix* (§ 1.2) et en présentant un *canevas général de choix* des méthodes (§ 1.3).

On trouvera des informations complémentaires à ce sujet dans les livres de Chatfield [1995] et Dytham [2003], ainsi que dans les articles d'Ehrenberg [1996] et Hand [1994].

Chapitre 2

Les conditions d'application des méthodes statistiques et l'examen initial des données

Sommaire

⊕ **2.1 Introduction**

⊕ **2.2 Les conditions d'application des méthodes statistiques**

⊕ **2.3 L'examen initial des données**

⊕ **2.4 Quelques tests du caractère aléatoire et simple d'une série d'observations**

Exercices

2.1 Introduction

1° Les méthodes d'inférence statistique ne sont applicables que dans des *conditions plus ou moins restrictives*, qui concernent notamment les modalités de collecte des données et la forme de la ou des distributions des populations-parents. Telle est la première question que nous aborderons au cours de ce chapitre (§ 2.2).

D'autre part, avant toute analyse statistique quelque peu élaborée, il est en général souhaitable de procéder à un *premier examen des données* disponibles, en tenant compte à la fois du ou des objectifs poursuivis et des exigences des méthodes d'analyse dont l'utilisation est envisagée. Nous consacrerons également un paragraphe à ce sujet (§ 2.3).

Des informations complémentaires générales sont données notamment par CHATFIELD [1995], COX et SNELL [1981], HAHN et MEEKER [1993], et MADANSKY [1988].

2° En outre, nous présenterons de façon plus particulière quelques *tests du caractère aléatoire et simple* d'une série d'observations, qui peuvent servir à compléter l'étude initiale des données (§ 2.4).

3° Les *exemples* 2.3.1 et 2.3.2 illustrent les questions que soulève l'examen des données, à une et à deux dimensions, tandis que les exemples 2.4.1 et 2.4.2 sont relatifs aux tests du caractère aléatoire et simple.

Chapitre 3

Les tests d'ajustement et de normalité et les observations aberrantes

Sommaire

⊕ **3.1 Introduction**
⊕ **3.2 Le test χ^2 d'ajustement de PEARSON**
⊕ **3.3 Les diagrammes de probabilité et quelques tests associés**
⊕ **3.4 Les tests de conformité de quelques paramètres particuliers**
⊕ **3.5 L'identification des observations aberrantes**
⊕ **3.6 Le cas des données à deux dimensions**
Exercices

⊕ 3.1 Introduction

1° La question de savoir si un ensemble d'observations peut être considéré comme provenant d'une *population d'un type donné* (population normale, population possédant une distribution de Poisson, etc.) est relativement fréquente. Elle peut se poser soit parce qu'on s'intéresse spécifiquement à la distribution envisagée, soit parce que l'existence d'un type donné de distributions est une condition préalable à l'utilisation de l'une ou l'autre méthode d'inférence statistique (§ 2.2.3).

Les *exemples* 3.2.1, 3.2.2 et 3.3.1 sont des illustrations de ces différentes situations.

2° Les *tests d'ajustement* ou *d'adéquation* [1] permettent de répondre d'une façon générale à ce type de questions. Les *tests de normalité* [2] ont pour but de traiter le même problème, souvent de façon plus efficace, dans le cas particulier des distributions normales. Et d'autres tests spécifiques existent également pour d'autres types de distributions (distributions binomiales et distributions de Poisson, par exemple).

Nous envisagerons successivement le *test* χ^2 *de* Pearson (§ 3.2), différentes méthodes basées sur la notion de *diagramme de probabilité*, dont le test de Shapiro et Wilk (§ 3.3), et quelques *tests spécifiques*, basés sur le calcul de paramètres particuliers, dont les coefficients de Pearson et de Fisher (§ 3.4).

3° Le contrôle de la conformité des distributions des populations-parents à un modèle donné se double souvent de la question de savoir si les observations considérées ne comportent pas une ou quelques *valeurs anormales* ou *aberrantes*. Nous examinerons aussi ce problème complémentaire, essentiellement dans le cas des distributions normales (§ 3.5).

Enfin, nous considérerons brièvement l'application des tests d'ajustement et de normalité, ainsi que l'identification d'éventuelles observations aberrantes, dans le cas des données *à deux dimensions* (§ 3.6).

Il faut noter que les différentes méthodes envisagées ne sont applicables de façon rigoureuse qu'à des observations résultant d'un *échantillonnage aléatoire et simple*.

4° Comme nous l'avons signalé en parlant des conditions d'application des méthodes classiques d'inférence statistique (§ 2.2.3.4°), le contrôle de la normalité de la distribution est un problème qui ne se présente pas seulement pour les données initiales elles-mêmes, mais aussi, parfois, pour les *écarts* ou les *résidus* par rapport à l'un ou l'autre modèle théorique, tel qu'une équation de régression, linéaire ou non linéaire.

Diverses études ont montré que les tests de normalité pouvaient être appliqués sans inconvénient, et sans modification, aux résidus de la régression linéaire simple,

[1] En anglais : *goodness-of-fit test*.
[2] En anglais : *test of normality*.

dès que le nombre d'observations atteint ou dépasse la vingtaine. Des effectifs plus importants sont par contre nécessaires en principe dans les cas plus complexes que sont, par exemple, l'analyse de la variance et la régression multiple [Pfaffenberger et Dielman, 1991 ; Pierce et Gray, 1982 ; White et MacDonald, 1980].

Les méthodes proposées restent cependant applicables à titre indicatif dans tous les cas.

5° Le problème du contrôle de la normalité se pose fréquemment aussi, non pas pour un seul échantillon suffisamment important, mais pour un *ensemble d'échantillons d'effectifs relativement limités*. La question peut alors être résolue notamment en calculant les écarts réduits par rapport aux moyennes (§ 2.3.3.4°) et en établissant des diagrammes de probabilité, d'une part séparément pour chacun des échantillons, et d'autre part globalement pour l'ensemble des échantillons.

D'autres solutions, dont l'utilisation de la méthode de regroupement des résultats de plusieurs tests de signification, que nous avons présentée antérieurement [STAT1, § 10.3.5.4°], peuvent également être envisagées [Quesenberry *et al.*, 1983 ; Wilk et Shapiro, 1968].

6° De nombreux *autres tests* d'ajustement et de normalité ont été proposés. Nous en mentionnerons occasionnellement certains.

Parmi les multiples publications consacrées à ce sujet, on peut recommander la consultation des travaux de Seier [2002], Thadewald et Büning [2007], Thode [2002], Yazici et Yolacan [2007], et Zhang et Wu [2005].

Chapitre 4

Les transformations de variables

Sommaire

⊕ **4.1 Introduction**
⊕ **4.2 Les principes de base et la transformation logarithmique**
⊕ **4.3 Les principales transformations**
⊕ **4.4 Le choix d'une transformation**
Exercices

4.1 Introduction

1° Nous avons mis l'accent, au cours du chapitre 2, sur l'importance qu'il faut accorder aux conditions d'application des méthodes d'inférence statistique, et sur la nécessité d'utiliser dans certains cas des transformations de variables en vue de mieux répondre à ces conditions (§ 2.2.3 et 2.2.5). Nous avons d'ailleurs déjà effectué à plusieurs reprises des transformations logarithmiques (exemples 2.3.2 et 3.6.1 notamment).

Nous revenons ici sur ce sujet, en considérant plus particulièrement les conditions de *normalité des populations-parents* et d'*égalité de leurs variances*, dans l'optique des comparaisons de moyennes, essentiellement par l'analyse de la variance (chapitres 9, 10 et 11). Nous envisagerons ultérieurement d'autres aspects des transformations de variables, et notamment leur application au cas de la régression non linéaire (§ 15.2.3).

2° Nous présenterons successivement quelques *principes de base* et la *transformation logarithmique* (§ 4.2), les principales *autres transformations* (§ 4.3), et quelques règles de *choix d'une transformation* (§ 4.4).

Les *exemples* 4.2.1 et 4.3.1 sont des illustrations des problèmes rencontrés dans ce domaine.

3° Les publications de synthèse relatives aux transformations de variables sont relativement peu nombreuses. Nous citerons seulement les articles de Box et Cox [1964], Hinkley et Runger [1984], et Hoyle [1973].

Chapitre 5

Les méthodes relatives à une ou deux proportions ou à un ou deux pourcentages

Sommaire

⊕ **5.1 Introduction**
⊕ **5.2 L'estimation et l'intervalle de confiance d'une proportion**
5.3 Les tests de conformité d'une proportion
5.4 La comparaison de deux proportions
Exercices

⊕ 5.1 Introduction

1° Après avoir envisagé les notions générales relatives au choix d'une méthode d'analyse statistique, à l'examen initial des données, au contrôle des conditions d'application des méthodes choisies, et aux transformations de variables (chapitres 1 à 4), nous abordons la présentation systématique des principales méthodes d'inférence statistique à une et à deux dimensions.

Nous commencerons par les méthodes relatives aux données qualitatives, c'est-à-dire aux données qui concernent des caractères ou des attributs, que chacun des individus observés peut posséder ou ne pas posséder [STAT1, § 2.4.1.3°]. Dans cette optique, nous envisagerons successivement les problèmes les plus simples, relatifs à *une ou deux proportions* ou à *un ou deux pourcentages* (chapitre 5), puis les problèmes relatifs à *plus de deux proportions* ou *plus de deux pourcentages*, ces problèmes étant considérés essentiellement sous l'angle des *tableaux de contingence* (chapitre 6).

Nous examinerons aussi, ultérieurement, d'autres aspects de l'étude des données qualitatives, dont la méthode des *probits* et la *régression logistique* (§ 15.5) [1].

2° Les principaux problèmes relatifs à une ou deux proportions sont l'*estimation* et la détermination des *limites de confiance d'une proportion* (§ 5.2), les *tests de conformité d'une proportion* (§ 5.3), et sous différentes formes, la *comparaison de deux proportions* (§ 5.4).

Ces problèmes sont aussi ceux de l'estimation, de la détermination des limites de confiance et des tests de conformité du paramètre p d'une distribution binomiale, et de la comparaison des paramètres p_1 et p_2 de deux distributions binomiales [STAT1, § 6.2.1].

De plus, bien que toutes les méthodes et les formules soient présentées en termes de proportions, comprises entre 0 et 1 , elles peuvent évidemment être adaptées facilement au cas des pourcentages, allant de 0 à 100 .

Les *exemples* 5.2.1, 5.3.1 et 5.4.1 illustrent ces différents problèmes.

3° Sauf mentions particulières, nous supposerons toujours que les *échantillons* considérés sont *aléatoires et simples*, et qu'ils proviennent de *populations infinies ou pratiquement infinies* (populations dont les effectifs sont au moins dix fois plus importants que les effectifs des échantillons).

[1] Au cours des chapitres précédents, nous avons considéré de façon détaillée, pour tous les exemples, tout ce qui concernait l'examen initial des données et le contrôle des conditions d'application des méthodes d'inférence statistique. Dans la suite, nous passerons en général beaucoup plus rapidement sur ces questions, en concentrant chaque fois l'attention sur l'objet principal de chacun des chapitres. Cette façon de faire ne signifie nullement que nous n'avons pas pris en considération au préalable la qualité des données que nous analysons dans les exemples, ni que les problèmes d'examen initial et de contrôle des conditions d'application peuvent être négligés en pratique. Nous consacrerons d'ailleurs encore deux exemples exclusivement à ces questions (exemples 9.3.2 et 10.3.2).

En outre, en ce qui concerne la comparaison de deux proportions, nous ferons la distinction entre le cas des *échantillons prélevés indépendamment l'un de l'autre* et le cas des *échantillons non indépendants*.

4° Des informations complémentaires peuvent être trouvées notamment dans les ouvrages spécialisés d'Everitt [1992], Fleiss *et al.* [2003], et Lloyd [1999].

On notera également l'existence de logiciels statistiques particuliers, tels que *StatXact* (<www.cytel.com>), qui sont très largement consacrés à l'étude des données qualitatives [Oster, 2002, 2003].

Chapitre 6

Les tableaux de contingence

Sommaire

⊕ **6.1 Introduction**
⊕ **6.2 Les tableaux de contingence à deux dimensions**
6.3 Les tableaux de contingence à trois dimensions
Exercices

6.1 Introduction

1° D'une manière générale, les *tableaux de contingence* [1], auxquels nous avons déjà fait allusion en ce qui concerne le cas particulier 2×2 (§ 5.4.1.2°), sont des distributions de fréquences qui ont trait à deux ou plusieurs caractères qualitatifs considérés simultanément. Les caractères envisagés peuvent être binaires, nominaux ou ordinaux [STAT1, § 2.4.1.3°].

2° Quand deux caractères seulement sont pris en considération, les tableaux de contingence se présentent comme des distributions de fréquences *à deux dimensions* tout à fait classiques [STAT1, § 4.2.2.1°], les différentes lignes correspondant aux différentes modalités d'un des deux caractères et les différentes colonnes aux différentes modalités de l'autre caractère. Nous envisagerons cette situation au cours du paragraphe 6.2.

Nous considérerons ensuite brièvement le cas des tableaux de contingence *à trois dimensions*, qui peuvent intervenir notamment dans l'étude simultanée de plusieurs tableaux à deux dimensions (§ 6.3).

Des illustrations de ces questions sont données par les *exemples* 6.2.1, 6.2.2 et 6.3.1.

3° Comme au chapitre 5, sauf mentions particulières, nous supposerons toujours que les *échantillons* considérés sont *aléatoires et simples*, et qu'ils proviennent de *populations infinies ou pratiquement infinies* (populations dont les effectifs sont au moins dix fois plus importants que les effectifs des échantillons).

4° La bibliographie relative aux tableaux de contingence et, d'une manière plus générale, à l'analyse des données qualitatives est particulièrement abondante. Les livres d'Agresti [2002, 2007], Everitt [1992], Fleiss *et al.* [2003], et Simonoff [2003], parmi d'autres, en témoignent.

[1] En anglais : *contingency table*.

Chapitre 7

Les méthodes relatives à la dispersion

Sommaire

⊕ **7.1 Introduction**

⊕ **7.2 Les estimations et les intervalles de confiance des paramètres de dispersion**

7.3 Les tests de conformité des paramètres de dispersion

⊕ **7.4 La comparaison de deux populations**

⊕ **7.5 La comparaison de plus de deux populations**

Exercices

⊕ 7.1 Introduction

1° Au cours de cette troisième partie, nous présenterons les principales méthodes relatives à l'étude des moyennes et de la dispersion. Il s'agit là d'un des domaines les plus importants de l'inférence statistique.

Nous envisagerons tout d'abord les méthodes relatives à la dispersion ou, de façon plus précise, aux *variances*, aux *écarts-types* et subsidiairement aux *coefficients de variation* (chapitre 7). En effet, l'égalité des variances est souvent une condition préalable à l'étude des moyennes, et les problèmes de variances sont en conséquence fréquemment pris en considération avant les problèmes de moyennes.

Nous présenterons ensuite les méthodes relatives à l'étude d'*une ou deux moyennes* (chapitre 8), puis les méthodes qui concernent l'étude de plus de deux moyennes, c'est-à-dire essentiellement l'*analyse de la variance* (chapitres 9 à 11), ainsi que les méthodes de *comparaisons particulières et multiples de moyennes* (chapitre 12).

2° Le plan que nous suivrons au cours de ce chapitre 7 est fort semblable à celui que nous avons adopté dans le cas des méthodes relatives à une ou deux proportions (chapitre 5), et aussi à celui que nous adopterons ultérieurement, notamment pour l'étude d'une ou deux moyennes (chapitre 8). Nous aborderons en effet successivement les questions d'*estimation* et de détermination de *limites de confiance* (§ 7.2), de *tests de conformité* (§ 7.3), de *comparaison de deux populations* (§ 7.4), et de *comparaison de plus de deux populations* (§ 7.5).

Les *exemples* 7.2.1, 7.3.1, 7.4.1 et 7.5.1 illustrent ces différentes situations.

3° Sauf indications contraires, toutes les méthodes présentées au cours de ce chapitre ne sont applicables que pour des *populations normales* et des *échantillons aléatoires et simples*. En outre, en ce qui concerne les comparaisons de deux ou plusieurs populations, la distinction doit être faite entre le cas des *échantillons prélevés indépendamment les uns des autres* et le cas des *échantillons non indépendants*.

Il faut souligner le fait que la condition de normalité est relativement restrictive pour les méthodes relatives à la dispersion, même dans le cas d'échantillons d'effectifs assez importants, contrairement notamment à ce qui se passe pour l'étude des moyennes [Box, 1953 ; Geary, 1956 ; Pearson et Please, 1975].

Chapitre 8

Les méthodes relatives à une ou deux moyennes

Sommaire

⊕ **8.1 Introduction**

⊕ **8.2 L'estimation et l'intervalle de confiance d'une moyenne**

⊕ **8.3 Les tests de conformité d'une moyenne**

⊕ **8.4 La comparaison de deux moyennes dans le cas des échantillons indépendants**

⊕ **8.5 La comparaison de deux moyennes dans le cas des échantillons non indépendants**

Exercices

⊕ 8.1 Introduction

1° Nous abordons ici les méthodes d'inférence statistique relatives aux moyennes, pour une ou deux populations. Ces méthodes figurent parmi celles qui sont les plus couramment utilisées.

Comme pour l'étude des proportions et de la dispersion (chapitres 5 et 7), nous envisagerons successivement les questions d'*estimation* et de détermination des *limites de confiance* d'une moyenne (§ 8.2), de *tests de conformité* d'une moyenne (§ 8.3), et de *comparaison de deux moyennes*. En ce qui concerne ce dernier point, nous consacrerons deux paragraphes distincts, l'un au cas des *échantillons indépendants* (§ 8.4), et l'autre au cas des *échantillons non indépendants* (§ 8.5).

Les *exemples* 8.2.1, 8.3.1, 8.4.1 et 8.5.1 illustrent les différentes questions qui sont envisagées.

2° Sauf indications contraires, les méthodes classiques qui sont présentées au cours de ce chapitre, et qui sont essentiellement basées sur les distributions t de STUDENT, ne sont applicables que pour des *populations normales* et des *échantillons aléatoires et simples*. En outre, en ce qui concerne la comparaison de deux moyennes dans le cas d'échantillons indépendants, il y a lieu d'être attentif également à la question de l'*égalité des variances*.

En raison de la rapide convergence des distributions d'échantillonnage de la moyenne vers les distributions normales [STAT1, § 8.3.1.5°], la condition de normalité est toutefois très peu restrictive. Ce n'est que pour des effectifs très limités (distributions t à moins de 10 degrés de liberté) que cette condition a une réelle importance.

D'une manière générale, les données étudiées peuvent être non seulement de nature continue, même fortement arrondies, mais aussi éventuellement de nature discontinue [CRESSIE, 1980 ; PEARSON et PLEASE, 1975 ; POSTEN, 1978, 1979 ; TRICKER, 1990a, 1990b, 1990c]. On évitera cependant de traiter, sans transformation, des ensembles de données caractérisés par de fortes dissymétries.

3° Nous présenterons aussi quelques *tests non paramétriques*, qui concernent parfois les médianes plus que les moyennes.

Chapitre 9

L'analyse de la variance à un critère de classification

Sommaire

⊕ **9.1 Introduction**
⊕ **9.2 Les aspects descriptifs**
⊕ **9.3 Les aspects inférentiels**
⊕ **9.4 La puissance et la détermination des nombres d'observations**
Exercices

⊕ 9.1 Introduction

1° D'une manière tout à fait générale, l'*analyse de la variance* [1] a comme objectif de comparer des ensembles de plus de deux moyennes, en identifiant les sources de variation qui peuvent expliquer les différences existant entre elles. À ce titre, l'analyse de la variance est un des principaux outils de l'inférence statistique.

Dans le cas le plus simple, l'analyse de la variance *à un critère de classification* ou *à un facteur* ou *à une voie* [2] concerne des ensembles de moyennes qui ne présentent aucune structure particulière, liée par exemple à l'existence de deux ou plusieurs facteurs sous-jacents [3] (§ 1.2.2.2°).

2° Bien que l'analyse de la variance ait été conçue essentiellement dans l'optique de la réalisation d'estimations et de tests d'hypothèses, elle peut également être considérée dans une certaine mesure comme une méthode descriptive. En vue de clarifier au maximum l'exposé, nous distinguerons les deux approches, en présentant dans un premier temps les *aspects descriptifs* (§ 9.2), puis les *aspects inférentiels* (§ 9.3). Nous envisagerons en outre les questions de détermination de la *puissance* de l'analyse et des *nombres d'observations* à effectuer (§ 9.4).

Les *exemples* 9.2.1 et 9.3.3 sont des illustrations des problèmes envisagés ici.

3° En ce qui concerne l'approche inférentielle, l'analyse de la variance s'applique dans les mêmes conditions que le test t de STUDENT, à savoir des *populations normales et de même variance*, et des *échantillons aléatoires, simples et indépendants* (§ 8.1.2°).

Les mêmes remarques qu'au paragraphe 8.1.2° peuvent être formulées à ce sujet. Comme le test t de STUDENT, l'analyse de la variance est en effet peu sensible à la non-normalité des populations-parents et, pour des échantillons de même effectif, à l'inégalité des variances [DONALDSON, 1968 ; KANJI et LIU, 1983 ; KRUTCHKOFF, 1988 ; TIKU, 1971].

Une réserve doit cependant être formulée en ce qui concerne ce dernier point. En effet, si l'analyse de la variance est peu sensible à une éventuelle inégalité des variances dans le cas des échantillons de même effectif, il n'en est pas de même pour les méthodes de comparaisons particulières et multiples de moyennes, qui sont très fréquemment utilisées en complément à l'analyse de la variance (chapitre 12). Il y a donc lieu, le plus souvent, d'être malgré tout attentif à cette condition, notamment par la réalisation de transformations de variables (chapitre 4).

On notera aussi qu'en particulier, l'analyse de la variance peut être appliquée sans inconvénients majeurs à des données discontinues, telles que des notations

[1] En anglais : *analysis of variance, ANOVA.*
[2] En anglais : *one-way analysis of variance.*
[3] L'analyse de la variance à un critère de classification est parfois appelée aussi analyse de la variance *à deux composantes*, en raison du fait que la variation totale y est divisée en deux parties (variation factorielle et variation résiduelle).

effectuées selon des échelles comportant au moins cinq degrés (appréciations sensorielles pouvant aller de 1 à 5 ou de 1 à 7, par exemple) [Rayner *et al.*, 1986 ; Tricker, 1992].

L'exemple 9.3.2 sera exclusivement consacré, à titre d'illustration, à la question du contrôle des conditions d'application de l'analyse de la variance.

4° La bibliographie relative à l'analyse de la variance est extrêmement abondante. De nombreux livres y sont notamment consacrés, souvent en relation avec les questions de régression ou d'expérimentation. On peut citer, entre autres, les livres de Christensen [1998], Lindman [1992], Mickey *et al.* [2004], Miller [1997], et Sahai et Ageel [2000].

Certains de ces ouvrages présentent l'analyse de la variance comme un cas particulier du modèle linéaire ou modèle linéaire général, qui englobe également la régression linéaire. Nous introduirons ce type de présentation au paragraphe 16.4, en utilisant alors des notations matricielles.

Chapitre 10

L'analyse de la variance à deux critères de classification

Sommaire

10.1 Introduction
10.2 Les modèles croisés à effectifs égaux : aspects descriptifs
10.3 Les modèles croisés à effectifs égaux : aspects inférentiels
10.4 Les modèles croisés à effectifs inégaux
10.5 Les modèles hiérarchisés
10.6 La puissance et la détermination des nombres d'observations
Exercices

10.1 Introduction

1° L'*analyse de la variance à deux critères de classification* [1] peut être considérée comme une généralisation de l'analyse à un critère, qui permet de tenir compte simultanément de deux facteurs sous-jacents, et non plus d'un seul facteur.

Les deux facteurs envisagés peuvent être soit placés sur pied d'égalité, soit au contraire subordonnés l'un à l'autre. Dans le premier cas, les modèles d'analyse de la variance sont dits *croisés* [2], alors que dans le deuxième cas, ils sont dits *hiérarchisés* [3]. Le cas hiérarchique est parfois qualifié aussi de *multi-niveaux* [4].

Dans les différents cas, on doit également faire la distinction entre les modèles *fixes*, les modèles *aléatoires* et les modèles *mixtes* [5]. Enfin, une distinction importante intervient entre le cas des effectifs égaux, parfois qualifié d'équilibré ou orthogonal, et le cas des effectifs inégaux, parfois qualifié de non équilibré ou non orthogonal.

Les *exemples* 10.2.1, 10.3.4 et 10.5.1 sont des illustrations de quelques-unes de ces situations.

2° Comme pour l'analyse de la variance à un critère de classification, nous considérerons tout d'abord les *aspects descriptifs* (§ 10.2), puis les *aspects inférentiels* (§ 10.3) de l'analyse à deux critères, en nous limitant dans un premier temps aux modèles croisés à effectifs égaux. Nous envisagerons ensuite les *modèles croisés à effectifs inégaux* (§ 10.4) et les *modèles hiérarchisés* (§ 10.5). Nous terminerons par quelques informations relatives à la notion de *puissance* et à la détermination des *nombres d'observations* (§ 10.6).

Nous travaillerons toujours par analogie avec l'analyse de la variance à un critère, ce qui devrait nous permettre de ne pas être trop long. C'est ainsi que nous éviterons au maximum de donner des démonstrations, en matière d'espérances mathématiques et de distributions d'échantillonnage notamment.

3° Globalement, les conditions d'application sont, en analyse de la variance à deux critères de classification, de la même nature qu'à un critère : *populations normales et de même variance*, et *échantillons aléatoires, simples et indépendants*. Les mêmes remarques que précédemment peuvent être formulées ici également à ce sujet (§ 8.1.2° et 9.1.3°).

À ces conditions de base, s'ajoute parfois une condition d'additivité, que nous définirons ultérieurement (§ 10.2.4.2°).

Comme en analyse de la variance à un critère de classification (exemple 9.3.2), nous illustrerons par un exemple le contrôle des conditions d'application (exemple 10.3.2).

[1] En anglais : *two-way analysis of variance*.
[2] En anglais : *cross-classification*.
[3] En anglais : *hierarchical classification*.
[4] En anglais : *multilevel analysis*.
[5] En anglais : *mixed model, mixed effects model*.

4° Les références bibliographiques mentionnées au paragraphe 9.1.4° peuvent être utiles aussi pour compléter l'information relative à l'analyse de la variance à deux critères de classification. Éventuellement, on pourra consulter en outre les publications relatives au modèle linéaire qui sont citées au paragraphe 16.1.5°.

Chapitre 11

L'analyse de la variance à trois et plus de trois critères de classification

Sommaire

11.1 Introduction

11.2 L'analyse de la variance à trois critères de classification : modèles croisés à effectifs égaux

11.3 L'analyse de la variance à trois critères de classification : modèles hiérarchisés à effectifs égaux

11.4 L'analyse de la variance à plus de trois critères de classification

11.1 Introduction

1° L'*analyse de la variance à trois critères de classification* [1] et, d'une manière plus générale, l'analyse de la variance *à un nombre quelconque de critères de classification* [2] présentent la même diversité de modèles que l'analyse à deux critères (§ 10.1.1°) : modèles croisés et hiérarchisés, modèles fixes, aléatoires et mixtes, et modèles à effectifs égaux et inégaux. Cette diversité s'accentue même, en raison de l'existence dans chaque cas de plusieurs modèles mixtes et de différents types de modèles hiérarchisés.

Nous envisagerons successivement l'analyse à *trois critères* de classification, en ce qui concerne les *modèles croisés* (§ 11.2) et les *modèles hiérarchisés* (§ 11.3), puis l'analyse à un *nombre quelconque de critères* de classification (§ 11.4). Nous procéderons toujours par analogie avec ce qui a été vu antérieurement, mais de manière sensiblement plus rapide, et nous nous en tiendrons ici au cas des échantillons de même effectif, en considérant ultérieurement le cas des effectifs inégaux, sous l'angle du modèle linéaire (§ 16.4.5).

Les *exemples* 11.2.1, 11.2.4 et 11.3.2 sont des illustrations des problèmes considérés au cours de ce chapitre.

2° Les mêmes principes que précédemment restent en vigueur en ce qui concerne les conditions d'application de l'analyse de la variance : *populations normales et de mêmes variances*, et *échantillons aléatoires, simples et indépendants* (§ 9.1.3° et 10.1.3°).

3° Les références bibliographiques générales du paragraphe 9.1.4° peuvent toujours être consultées ici également, en plus de celles qui sont citées dans le texte. Elles peuvent être complétées par les références relatives au modèle linéaire (§ 16.1.5°).

Il faut noter en outre que nous nous limitons à la présentation de l'analyse de la variance classique, à l'exclusion d'autres possibilités, telles que les modèles à effets principaux additifs et interactions multiplicatives (modèles AMMI) et les méthodes non paramétriques et robustes. Certaines des références données aux paragraphes 10.3.1.2° et 10.3.8.5° peuvent éventuellement fournir des indications relatives à ces diverses possibilités, dans le cas de trois ou plus de trois critères de classification. On peut y ajouter le travail de VAN EEUWIJK et KROONENBERG [1998].

[1] En anglais : *three-way analysis of variance.*
[2] En anglais : *multi-way analysis of variance.*

Chapitre 12

Les comparaisons particulières et multiples de moyennes

Sommaire

12.1 Introduction
12.2 L'utilisation des contrastes
12.3 Les comparaisons avec un ou plusieurs témoins et la recherche de la ou des variantes les meilleures
12.4 Les comparaisons des moyennes considérées sur pied d'égalité
Exercices

12.1 Introduction

1° Sauf dans le cas particulier des critères de classification qui ne possèdent que deux modalités ($p = 2$, $q = 2$, etc.), les hypothèses nulles relatives aux facteurs fixes des analyses de la variance font toujours intervenir plusieurs signes d'égalité (§ 9.3.2.3°, 10.3.2.3°, 10.3.4.4°, etc.). Le rejet de telles hypothèses soulève alors la question d'interpréter et, éventuellement, de localiser les inégalités de moyennes.

De nombreuses solutions, très diversifiées, ont été proposées pour répondre ou tenter de répondre à cette question. Nous en parlons ici sous l'appellation générale de méthodes de comparaisons particulières et multiples. Le choix entre les différentes approches est très largement fonction de la nature, qualitative ou quantitative, des facteurs considérés (§ 1.2.2.3°) et de l'objectif qui a été fixé, ou qui aurait dû être fixé, au moment où la collecte des données a été décidée.

2° Que le ou les facteurs fixes considérés soient de nature qualitative ou quantitative, si un certain nombre de *questions particulières* ont été définies a priori de façon précise, et si ces questions peuvent être exprimées sous la forme de fonctions linéaires des moyennes, il est généralement possible de traiter le problème par l'utilisation de contrastes.

Pour des facteurs quantitatifs uniquement, cette procédure permet également, dans certains cas, d'ajuster aux moyennes observées des équations représentatives de courbes ou de surfaces de réponse. Ces équations peuvent alors être utilisées en vue notamment de rechercher des maximums ou des minimums, ou d'une manière plus générale, des conditions optimales.

Nous examinerons ces problèmes au cours du paragraphe 12.2. Les *exemples* 12.2.1 et 12.2.2 en sont des illustrations.

3° Pour des facteurs qualitatifs, l'équivalent de la recherche de conditions optimales est la recherche de la ou des modalités, ou des *variantes les meilleures*, c'est-à-dire de la ou des variantes dont les moyennes sont maximales ou minimales.

Un autre problème, étroitement lié à ce dernier, est la comparaison d'une série de variantes avec un ou plusieurs *témoins*.

Ces deux questions seront le thème du paragraphe 12.3. Les *exemples* 12.3.1 et 12.3.2 en donnent aussi des illustrations.

4° Dans le cas des facteurs qualitatifs, on peut également souhaiter comparer entre elles une série de modalités ou de variantes qui ne présentent aucune structure particulière et au sujet desquelles on ne se pose a priori aucune question précise.

Une première solution est alors de procéder à toutes les *comparaisons deux à deux*, les moyennes étant considérées sur pied d'égalité. Cette approche, dite de comparaisons multiples, est l'objet de très nombreuses méthodes.

Une autre solution consiste à tenter de définir des groupes de variantes aussi homogènes que possible, par des méthodes de *classification numérique*.

Nous envisagerons ces approches au paragraphe 12.4, par la présentation d'un nombre limité de méthodes. Ici également, des illustrations peuvent être trouvées en considérant les *exemples* 12.4.1 et 12.4.2.

Il faut savoir cependant que les méthodes de comparaisons multiples sont l'objet de nombreuses utilisations abusives, qui résultent le plus souvent d'un manque de définition précise, a priori, des objectifs poursuivis, et aussi de leur grande généralité et leur grande facilité d'utilisation automatique. Ces méthodes devraient en réalité être considérées plutôt comme des pis-aller, que comme des méthodes d'usage courant [Dawkins, 1983 ; Pearce, 1993].

5° D'autres approches et d'autres situations ont aussi été étudiées. Ainsi, d'une manière générale, la plupart des problèmes que nous envisagerons au cours de ce chapitre, essentiellement sous l'angle des tests d'hypothèses, peuvent également être abordés dans l'optique des intervalles de confiance, alors appelés *intervalles de confiance simultanés* [1].

D'autre part, on peut considérer en outre des problèmes tels que la réalisation de comparaisons multiples dans le cas des facteurs quantitatifs, en présence d'hypothèses alternatives ordonnées (§ 9.3.2.9°), en relation par exemple avec des doses croissantes d'une même substance [Liu et Somerville, 2004 ; Nashimoto et Wright, 2005 ; Peng *et al.*, 2006 ; Strassburger *et al.*, 2007]. L'objectif peut être notamment de déterminer une *dose efficace minimale* [2] ou une *dose tolérée maximale* [3] [Bauer, 1997 ; Nakamura et Douke, 2007 ; Tamhane *et al.*, 1996].

Nous pouvons encore ajouter la méthode dite d'*analyse des moyennes* [4], qui peut remplacer à la fois l'analyse de la variance et les comparaisons multiples de moyennes, en vue de mettre en évidence les moyennes particulières qui diffèrent signicativement de la moyenne générale de l'ensemble des observations [Nelson *et al.*, 2005 ; Rao, 2005 ; Ryan, 2006].

6° D'une manière générale, les conditions d'utilisation des méthodes que nous présentons ici sont celles de l'analyse de la variance : *populations normales et de même variance*, et *échantillons aléatoires, simples et indépendants* (§ 9.1.3°). En particulier, l'hypothèse d'égalité des variances, qui peut être considérée comme relativement secondaire en analyse de la variance, dans le cas d'échantillons d'effectifs égaux, est toujours importante ici, même pour des effectifs constants.

Souvent, les méthodes de comparaisons particulières et multiples de moyennes sont présentées en ne considérant que le cas des échantillons de même effectif. Nous nous efforcerons au contraire d'envisager, dans la mesure du possible, des solutions tout à fait générales.

[1] En anglais : *simultaneous confidence intervals*.
[2] En anglais : *minimum effective dose*.
[3] En anglais : *maximum tolerated dose*.
[4] En anglais : *analysis of means, ANOM*.

Les problèmes de comparaisons particulières et multiples peuvent bien sûr être traités également pour d'autres distributions que les distributions normales (distributions exponentielles par exemple), et aussi pour d'autres paramètres que les moyennes (proportions ou pourcentages, coefficients de corrélation et de régression, etc. [Levin et Leu, 2007 ; Schaarschmidt *et al.*, 2008 ; Wu et Chen, 1998].

7° Comme les paragraphes précédents en témoignent déjà, la bibliographie relative aux méthodes dont il sera question au cours de ce chapitre est extrêmement abondante. Nous ajoutons encore les références de quelques livres [Hochberg et Tamhane, 1987 ; Hsu, 1996 ; Klockars et Sax, 1986 ; Miller, 1981]. On trouvera notamment dans ces ouvrages des tables plus diversifiées que celles que nous donnons, ainsi que des algorithmes, dont l'emploi peut se substituer aux tables.

Chapitre 13

Les méthodes relatives à la corrélation simple

Sommaire

13.1 Introduction
13.2 Les distributions d'échantillonnage
13.3 L'estimation et l'intervalle de confiance d'un coefficient de corrélation
13.4 Les tests de conformité et de signification d'un coefficient de corrélation
13.5 La comparaison de deux ou plusieurs coefficients de corrélation
Exercices

13.1 Introduction

1° La quatrième et dernière partie de cet ouvrage est essentiellement consacrée à l'inférence statistique à deux dimensions, dans le cas des données quantitatives. Nous y envisagerons tout d'abord les méthodes relatives à la *corrélation simple* (chapitre 13) et les méthodes relatives à la *régression simple*, *linéaire* (chapitre 14) et *non linéaire* (chapitre 15).

Nous présenterons ensuite quelques notions de *régression multiple*, ainsi que le concept plus général de *modèle linéaire*, en introduisant également diverses extensions de ce modèle (chapitre 16). Enfin, nous considérerons l'*analyse de la covariance*, qui met en jeu simultanément des principes d'analyse de la variance et de régression (chapitre 17).

2° En ce qui concerne la corrélation simple, les différents problèmes à étudier sont de la même nature que ceux que nous avons envisagés antérieurement au sujet des proportions, des paramètres de dispersion et des moyennes. Nous les passerons en revue dans le même ordre que précédemment.

Au préalable, nous consacrerons un paragraphe aux *distributions d'échantillonnage* des coefficients de corrélation (§ 13.2). Nous examinerons ensuite successivement les questions d'*estimation* et d'*intervalle de confiance* (§ 13.3), les *tests de signification et de conformité* (§ 13.4), et la *comparaison de deux ou plusieurs coefficients de corrélation*, ainsi que certaines notions connexes (§ 13.5).

Nous nous intéresserons principalement au coefficient de corrélation simple classique, au sens de Bravais-Pearson [STAT1, § 4.6.1], mais nous donnerons aussi fréquemment des informations relatives à certains paramètres qui en sont dérivés, dont les coefficients de corrélation de rang et intraclasse [STAT1, § 4.6.3]. Rappelons également, à cet égard, que nous avons déjà évoqué antérieurement divers problèmes relatifs aux relations qui peuvent exister entre des caractéristiques qualitatives, y compris les notions de coefficients de corrélation de point et de contingence (§ 6.2.5).

Des illustrations des questions qui sont considérées au cours de ce chapitre sont données par les *exemples* 13.3.1, 13.4.1 et 13.5.1.

3° Le coefficient de corrélation classique concerne principalement des couples de variables continues interdépendantes. On suppose généralement que ces variables possèdent des *distributions normales à deux dimensions* [STAT1, § 7.4.3]. En outre, comme pour les autres paramètres, les *échantillons* doivent toujours être *aléatoires et simples*, et sauf indication contraire, *indépendants* les uns des autres dans le cas des comparaisons de deux ou plusieurs populations.

Pour des échantillons d'effectifs suffisamment élevés (20 ou 30 observations au moins), la condition de normalité à deux dimensions n'est toutefois pas très contraignante. En pratique, le coefficient de corrélation de Bravais-Pearson est d'ailleurs assez fréquemment utilisé aussi pour des variables discontinues, pour des

données qualitatives ordinales codées sous forme numérique, et pour des couples de caractéristiques de natures différentes (données qualitatives ordinales associées à des données quantitatives, par exemple). Il faut cependant s'assurer en toute circonstance du caractère linéaire ou approximativement linéaire des relations entre les variables ou les caractéristiques étudiées, et de l'absence de valeurs aberrantes (§ 3.6.3).

Pour éviter toute erreur systématique dans l'estimation des coefficients de corrélation, il faut supposer en outre que les valeurs observées des variables considérées sont connues *sans erreurs de mesure* ou, en tout cas, sans erreurs de mesure importantes par rapport à la variabilité propre de ces variables (§ 13.3.4°). En particulier, il y a lieu de s'abstenir autant que possible de tout calcul de coefficients de corrélation à partir de distributions de fréquences groupées en classes [STAT1, § 4.2.2.3°].

Des informations relatives à la robustesse des méthodes qui concernent la corrélation simple sont données notamment par Srivastava et Lee [1984], et Subrahmaniam et Gajjar [1980].

4° Très peu de livres généraux sont spécifiquement consacrés aux problèmes de corrélation, ces problèmes étant en fait considérés le plus souvent en marge des questions de régression, pour lesquels les ouvrages spécialisés sont beaucoup plus nombreux (§ 14.1.5°). On peut toutefois citer ici le livre de Lindeman *et al.* [1980], ainsi que celui de Kendall et Gibbons [1990] en ce qui concerne plus particulièrement les coefficients de corrélation de rang.

Chapitre 14

Les méthodes relatives à la régression linéaire simple

Sommaire

⊕ **14.1 Introduction**

⊕ **14.2 Les distributions d'échantillonnage**

⊕ **14.3 L'ajustement et la validation d'une droite des moindres carrés**

14.4 L'estimation à l'aide d'une droite des moindres carrés

14.5 Les tests de conformité, de signification et de linéarité pour les droites des moindres carrés

14.6 La comparaison de deux ou plusieurs droites des moindres carrés

14.7 La droite des moindres rectangles

Exercices

⊕ 14.1 Introduction

1° Comme pour la corrélation simple (chapitre 13), nous consacrerons tout d'abord un paragraphe aux *distributions d'échantillonnage* des paramètres caractéristiques des droites de régression (§ 14.2).

Nous examinerons ensuite les différents problèmes relatifs à la régression au sens des *moindres carrés*, à savoir : l'*ajustement* et la *validation* d'une droite de régression (§ 14.3), l'*estimation à l'aide d'une droite de régression* (§ 14.4), les *tests de conformité, de signification et de linéarité* (§ 14.5), et la *comparaison de deux ou plusieurs droites de régression* (§ 14.6).

Enfin, nous aborderons, beaucoup plus rapidement, les problèmes relatifs à la régression au sens des *moindres rectangles* (§ 14.7).

Les *exemples* 14.3.1, 14.4.1, 14.5.1, 14.6.1 et 14.7.1 constituent quelques illustrations de ces différents problèmes.

2° En ce qui concerne la régression au sens des *moindres carrés*, c'est-à-dire la relation qui lie une variable dépendante à une variable explicative [STAT1, § 4.7], nous considérerons le modèle théorique suivant :

$$\boxed{Y = \alpha + \beta\, x + D} \quad \text{ou} \quad \boxed{Y_i = \alpha + \beta\, x_i + D_i},$$

α étant l'ordonnée à l'origine, β le coefficient de régression, x la variable explicative, non aléatoire, D les écarts ou les résidus aléatoires par rapport à la droite, et Y la variable dépendante, entachée des fluctuations aléatoires dues à D. Comme en analyse de la variance, on suppose alors que les résidus D_i sont des *variables normales, de moyennes nulles, de même variance et indépendantes* les unes des autres [1].

Le caractère non aléatoire de la variable explicative implique que les valeurs x_i sont connues *sans erreurs*, ou en tout cas sans erreurs importantes. La nullité des moyennes des résidus est liée à la linéarité de la régression. La variance qui est supposée constante est en fait la variance résiduelle $\sigma^2_{Y.x}$ [STAT1, § 7.3.4.2°]. Et l'indépendance des résidus peut être assurée par le caractère aléatoire et simple de l'échantillonnage.

En outre, en ce qui concerne la comparaison de deux ou plusieurs droites de régression, on doit également supposer que les différents *échantillons* considérés sont *indépendants* les uns des autres, et que les différentes régressions sont *de même variance résiduelle*.

3° D'autres situations et d'autres solutions doivent aussi être envisagées dans certains cas (régression par l'origine, régression pondérée, régression avec erreurs sur les deux variables, méthodes non paramétriques et robustes, etc.). Nous donnerons diverses informations à ce sujet aux paragraphes 14.3.5 à 14.3.7.

[1] Le symbole α, qui désigne ici une ordonnée à l'origine, n'a bien sûr rien de commun avec le même symbole désignant un risque d'erreur ou un niveau de signification.

4° En ce qui concerne la régression au sens des *moindres rectangles*, c'est-à-dire la relation entre deux variables interdépendantes [STAT1, § 4.8], on suppose que les deux variables, X et Y, possèdent une *distribution normale à deux dimensions* [STAT1, § 7.4.3], et que l'*échantillonnage* est également *aléatoire et simple*. Ces conditions sont les mêmes que pour les méthodes relatives à la corrélation simple (§ 13.1.3°).

On remarquera que les conditions émises pour les deux types de régression ne sont pas fondamentalement différentes. En effet, la condition de normalité à deux dimensions relative au deuxième cas implique, comme dans le premier cas, la linéarité de la régression, la normalité des écarts par rapport aux droites de régression, la nullité des moyennes de ces écarts et l'égalité de leurs variances [STAT1, § 7.4.3].

La distinction entre ces deux situations est souvent faite par l'emploi des expressions *relation fonctionnelle* [2], dans le cas d'une variable explicative connue sans erreur, et *relation structurelle* [3], dans le cas de deux variables interdépendantes.

5° La littérature relative à la régression est particulièrement abondante. D'une manière générale, on peut recommander notamment les livres de Dodge [2004b], Draper et Smith [1998], Ryan [2009], Tomassone *et al.* [1992], et Weisberg [2005]. Ces livres sont aussi partiellement consacrés à la régression non linéaire et à la régression multiple, et dépassent donc largement le cadre de la régression linéaire simple.

On peut citer en outre le livre de Cook et Weisberg [1999], relatif plus particulièrement aux aspects graphiques de la régression.

[2] En anglais : *functional relationship*.
[3] En anglais : *structural relationship*.

Chapitre 15

La régression non linéaire simple et la modélisation

Sommaire

15.1 Introduction
15.2 Les modèles constitués d'une seule équation
15.3 Les modèles à deux ou plusieurs équations
15.4 Les méthodes non paramétriques et robustes
15.5 Les relations entre données qualitatives et quantitatives
15.6 Les séries chronologiques

15.1 Introduction

1° La diversité des problèmes que nous avons présentés en régression linéaire simple (ajustement, validation, estimation directe et estimation inverse, tests de conformité, etc.) subsiste en matière de *régression non linéaire* ou *curvilinéaire* [1], c'est-à-dire pour des *courbes de régression* [2]. Cette diversité se double en outre d'une grande variété de modèles pouvant être pris en considération. L'élaboration de ces modèles est l'objet de ce qui est parfois appelé la *modélisation* [3].

Nous ne reviendrons pas de manière détaillée sur l'ensemble des problèmes abordés à propos de la régression linéaire, en nous limitant ici à passer en revue, assez rapidement, les principaux modèles de régression non linéaire.

2° Dans un premier temps, nous envisagerons les modèles relatifs aux données quantitatives, en considérant successivement le cas le plus classique des phénomènes représentés par *une seule équation* de régression (§ 15.2), le cas des phénomènes qui peuvent être représentés par *deux ou plusieurs équations*, dont la régression segmentée et les modèles à compartiments (§ 15.3), et les *méthodes non paramétriques et robustes*, dont les méthodes de lissage, qui ne font intervenir a priori aucun modèle particulier (§ 15.4). Nous consacrerons ensuite un paragraphe aux *relations entre données qualitatives et quantitatives*, à savoir les notions de probit et de régression logistique (§ 15.5), et un paragraphe aux *séries chronologiques* § 15.6).

Les *exemples* 15.2.3, 15.3.1, 15.4.1, 15.5.1 et 15.6.1 constituent quelques illustrations de ces différentes possibilités.

3° Comme références générales, on peut citer les livres de Bates et Watts [1988], Huet *et al.* [1992, 2004], et Seber et Wild [2003], ainsi que le livre de Carroll *et al.* [1995], en ce qui concerne le cas où la variable explicative et la variable dépendante sont toutes deux entachées d'erreurs de mesure. Nous y ajouterons progressivement diverses références plus particulières.

[1] En anglais : *non-linear regression, curvilinear regression.*
[2] En anglais : *regression curve.*
[3] En anglais : *modelling.*

Chapitre 16

La régression multiple et le modèle linéaire

Sommaire

16.1 Introduction
16.2 La régression linéaire à deux variables explicatives
16.3 La régression linéaire à p variables explicatives
16.4 Le modèle linéaire et l'analyse de la variance
16.5 Quelques extensions du modèle linéaire

16.1 Introduction

1° La *régression multiple* [1] a pour but d'exprimer une variable dépendante y en fonction, non plus d'une seule variable explicative x, mais bien de deux ou plusieurs variables explicatives $x_1, \ldots, x_p$. Comme dans le cas de la régression simple, la relation utilisée à cette fin peut être linéaire ou non linéaire.

Le modèle de base de la *régression linéaire multiple* [2] est une généralisation relativement élémentaire du cas de la régression linéaire simple (§ 14.1.2°). Ce modèle s'écrit en effet :

$$\boxed{Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + D} \quad \text{ou} \quad \boxed{Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + D_i},$$

β_0 étant le terme indépendant (désigné précédemment par α), $\beta_1, \ldots, \beta_p$ étant les coefficients de régression relatifs aux p variables $x_1, \ldots, x_p$, et $x_{i1}, \ldots, x_{ip}$ étant les valeurs de ces variables pour les différents individus observés ($i = 1, \ldots, n$).

Les conditions d'application de ce modèle sont semblables à celles de la régression linéaire simple. Les résidus D_i sont considérés comme des *variables normales, de moyennes nulles, de même variance et indépendantes* les unes des autres, et les valeurs des variables explicatives sont supposées connues *sans erreurs* ou, au moins, sans erreurs importantes (§ 14.1.2°).

2° Le modèle qui vient d'être présenté peut être appliqué notamment aux différents cas d'analyse de la variance et de la covariance. On le désigne souvent sous le nom de *modèle linéaire* ou *modèle linéaire général* [3], et cela éventuellement dans des conditions moins restrictives, que nous évoquerons ultérieurement (§ 16.5.2).

Sauf dans les cas les plus simples, le recours au modèle linéaire s'impose pratiquement toujours en vue de traiter les problèmes d'analyse de la variance relatifs à des échantillons d'effectifs inégaux. Dans de nombreux ouvrages, cette approche est d'ailleurs introduite en premier lieu, l'analyse de la variance n'étant considérée que comme un cas particulier.

D'autre part, la notion de modèle linéaire a été étendue de différentes manières, notamment sous la forme de modèles qualifiés de linéaire mixte et de linéaire généralisé.

3° Nous envisagerons successivement la régression linéaire multiple dans le cas particulier de *deux variables explicatives* (§ 16.2) et dans le cas général de p *variables explicatives* (§ 16.3), puis le *modèle linéaire* et son utilisation en *analyse de la variance* (§ 16.4), et enfin, assez brièvement, les *extensions* du modèle linéaire (§ 16.5). Le paragraphe 16.2 sera entièrement présenté à l'aide de notations algébriques classiques, mais à partir du paragraphe 16.3, nous serons amené à utiliser des notations matricielles.

[1] En anglais : *multiple regression.*
[2] En anglais : *multiple linear regression.*
[3] En anglais : *linear model, general linear model, GLM.*

Les *exemples* 16.2.1, 16.2.3 et 16.4.1 sont des illustrations des diverses situations considérées.

4° La présentation des différentes notions sera relativement sommaire, en ce qui concerne notamment la régression multiple. En particulier, nous ne reviendrons pas sur des questions telles que la validation des équations de régression par l'étude des résidus et la recherche des valeurs influentes (§ 14.3.3 et 14.3.4).

De même, nous n'aborderons pas les questions, importantes en régression multiple, de *colinéarité* ou *multicolinearité* [4], et de choix des variables explicatives, ni les méthodes alternatives que sont par exemple la *régression par les composantes principales* ou *régression orthogonalisée* [5], la *régression par les moindres carrés partiels* ou *régression PLS* [6], la « *ridge regression* », et les *méthodes à rétrécisseurs* [7] [Palm et Iemma, 1995].

5° La bibliographie relative à la régression multiple et au modèle linéaire, ainsi qu'aux extensions de ce modèle, est extrêmement abondante. On peut se référer tout d'abord à certains des ouvrages que nous avons déjà cités antérieurement à propos de l'analyse de la variance et de la régression linéaire simple, dont ceux de Draper et Smith [1998], Mickey *et al.* [2004], et Tomassone *et al.* [1992]. On peut y ajouter les livres de Hocking [2003], Rencher [2000], et Searle [1997], et nous donnerons aussi, ultérieurement, des références plus spécifiques en ce qui concerne notamment les modèles linéaires mixte et généralisé (§ 16.5.3 et 16.5.4).

En outre, des notions de calcul matriciel appliqué à la statistique peuvent être trouvées dans les ouvrages spécialisés de Graybill [2002], Healy [2000], et Searle [1982]. Et des éléments de calcul matriciel figurent également dans certains livres plus généraux, tels que ceux de Draper et Smith [1998], et Rencher [2000].

[4] En anglais : *collinearity, multicollinearity.*
[5] En anglais : *principal component regression.*
[6] En anglais : *partial least squares regression, PLS regression.*
[7] En anglais : *shrinkage method.*

Chapitre 17

L'analyse de la covariance

Sommaire

17.1 Introduction

17.2 L'analyse de la covariance à un critère de classification

17.3 L'analyse de la covariance à deux et plus de deux critères de classification

17.1 Introduction

1° L'*analyse de la covariance* [1] a pour but d'effectuer des comparaisons de moyennes en tenant compte d'un ou plusieurs critères de classification, comme en analyse de la variance, mais en faisant intervenir en outre, par régression, une ou plusieurs *variables auxiliaires*, aussi appelées *variables concomitantes* ou *covariables* [2]. La raison d'être de cette ou de ces variables auxiliaires est très souvent d'éliminer l'influence de cette ou de ces variables, en vue d'augmenter la puissance des comparaisons de moyennes.

D'autres objectifs, dont nous parlerons au paragraphe 17.2.3, peuvent également être poursuivis.

Les *exemples* 17.2.1 et 17.3.1 illustrent les questions qui sont présentées ici.

2° Nous envisagerons l'analyse de la covariance en en exposant les principes pour un nombre limité de modèles et par quelques exemples, essentiellement avec une seule covariable. Nous traiterons successivement de l'analyse de la covariance à *un critère de classification* (§ 17.2) et de l'analyse de la covariance à *deux et plus de deux critères de classification* (§ 17.3).

Dans un cas comme dans l'autre, la présentation que nous adopterons sera très semblable à celle de l'analyse de la variance (chapitres 9 et 10) et de la régression linéaire simple (chapitre 14). Nous indiquerons cependant aussi comment le problème peut être abordé sous l'angle du modèle linéaire (§ 16.4).

3° Les conditions d'application de l'analyse de la covariance sont tout d'abord celles de l'analyse de la variance, à savoir la *normalité des populations*, l'*égalité de leurs variances*, et le *caractère aléatoire, simple et indépendant des échantillons* (§ 9.1.3°). À ces conditions, s'ajoutent, pour les différentes populations, la *linéarité* et le *parallélisme* des relations entre les variables considérées.

Comme en analyse de la variance (§ 9.1.3°), certaines de ces conditions ne sont pas essentielles pour l'analyse de la covariance proprement dite, en particulier dans le cas d'échantillons de même effectif. Ces conditions sont cependant importantes pour les comparaisons de moyennes qui peuvent suivre l'analyse de la covariance.

Dans de nombreux cas, la validité de l'ensemble des conditions d'application, et notamment de la condition de parallélisme, peut difficilement être vérifiée. Il y a lieu d'utiliser alors l'analyse de la covariance avec prudence, voire même d'éviter dans une certaine mesure son emploi, comme nous l'indiquerons au paragraphe 17.2.3.4°.

La robustesse de l'analyse de la covariance a été envisagée notamment par Atiqullah [1964] et Hamilton [1976].

[1] En anglais : *analysis of covariance, ANCOVA, ANOCOVA.*
[2] En anglais : *concomitant variable, covariable.*

4° On trouvera des informations complémentaires relatives à l'analyse de la covariance dans la plupart des ouvrages que nous avons cités à propos de l'analyse de la variance et du modèle linéaire (§ 9.1.4° et 16.1.5°), ainsi que dans le livre spécialisé de MILLIKEN et JONHSON [2002]. On peut mentionner également deux numéros particuliers de la revue *Biometrics*, déjà fort anciens, mais toujours intéressants à consulter [COCHRAN, 1957 ; etc. ; COX et MCCULLAGH, 1982 ; etc.].