

LA SEGMENTATION

Spécialisé dans le traitement statistique des études cliniques,
le Département Biométrie de FOVEA vous propose
au travers de ce fascicule quelques notions simples
concernant la segmentation.



Traitement statistique

DES ETUDES CLINIQUES

LA SEGMENTATION EN PRATIQUE

L'un des problèmes posés en recherche clinique est de définir précisément **le profil du patient qui a le plus de chance de répondre à un traitement donné**, et inversement le profil de celui pour lequel le traitement sera de peu d'utilité.

En d'autres termes, à facteur traitement égal, quels sont le ou les facteurs qui vont le plus contribuer au succès ou à l'échec du traitement ?

Cela revient donc à étudier le "comportement", à l'égard du traitement, de sous-groupes de patients définis par un certain nombre de variables ainsi que les classes de valeurs qui y sont rattachées.

La segmentation, **technique d'analyse multivariée utilisée à l'origine en marketing**, permet d'aider à répondre à cette question.

En effet, la segmentation a été utilisée pour expliquer le comportement d'achat de consommateurs en fonction de certains paramètres les définissant : âge, sexe, habitat, catégorie socio-professionnelle, niveau d'études, niveau de revenus, ...

En recherche clinique, la segmentation permet donc d'identifier **les caractéristiques les plus discriminantes** afin de découper la population en sous-groupes de patients allant de ceux qui ont le plus fort taux de succès à ceux qui ont le plus fort taux d'échec.

1

Principe de la segmentation

Le principe de la segmentation est d'étudier **les relations** existant entre **une variable Y** dite **variable à expliquer** et **des variables X_i** dites **variables explicatives**.

La variable Y peut être qualitative et bimodale (*guérison/échec, existence ou non de tel effet secondaire, ...*), mais aussi quantitative. **Les variables explicatives X_i** sont obligatoirement **qualitatives**.

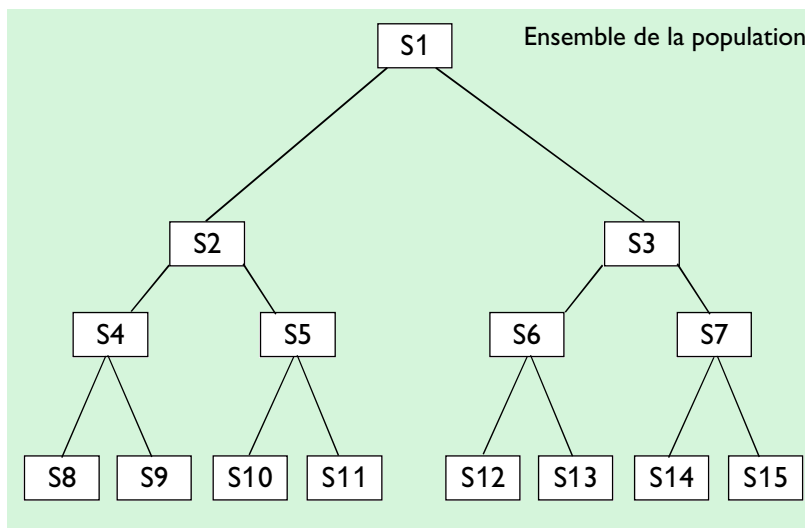
Il s'agit donc de constituer à l'aide des X_i des segments homogènes : **les caractéristiques des patients doivent être aussi homogènes que possible à l'intérieur d'un même segment et aussi**

différentes que possible entre les différents segments.

La partition se fait par rapport à la variable Y. On détermine ainsi **une structure arborescente** que l'on peut représenter par un graphe "en arbre" où chaque segment, constitué lui-même par partition d'un segment père en fonction des modalités de la variable Y, donne naissance à 2 segments fils.

Le processus se poursuit sur ce modèle avec un fractionnement de plus en plus poussé de la population.

A son terme, il fournit un certain nombre de segments définis à partir des différentes variables explicatives X_i. A l'étape n, 2ⁿ segments ont été créés. Chaque chemin déterminé en parcourant les différentes branches de l'arbre donne un





profil de patient fixé par rapport à la variable Y. Les branches les plus externes de l'arbre, à gauche et à droite, correspondent aux profils des patients les plus opposés vis-à-vis de la variable à expliquer.

2

Réalisation pratique d'une segmentation

Pratiquement une segmentation se réalise selon le processus suivant :

- on considère une variable explicative qualitative, sur laquelle on effectue un certain nombre de subdivisions en classes,
- on identifie la subdivision pour laquelle il existe la plus forte relation avec la variable à expliquer,
- on procède de la même façon pour l'ensemble des variables explicatives,
- on retient la variable explicative (et la subdivision optimale qui lui est attachée) pour laquelle il existe la plus forte relation avec la variable à expliquer.

La population est alors subdivisée par rapport à cette variable en plusieurs sous-populations et le processus est réitéré sur chacune de ces sous-populations indépendamment.

Pour ce faire, les informations suivantes sont nécessaires :

- choix de la variable à expliquer Y : qualitative et bimodale ou quantitative,

- choix des variables explicatives X_i qualitatives.
 - **les critères d'arrêt**, à savoir :
 - **L'effectif minimum** de chacun des segments : un segment ayant un nombre de patients inférieur à cet effectif minimum ne sera plus subdivisé,
 - **La distance minimum** entre 2 segments : elle est comprise entre 0 et 1.
- => Si $c = 0$, il n'y a aucune liaison entre la variable explicative X_i et la variable à expliquer Y : aucune subdivision n'est possible.
- => Si $c = 1$, la liaison entre X_i et Y est maximum, tous les patients sont sur la diagonale du tableau de contingence : la subdivision est totale.
- => Les valeurs intermédiaires de c donnent les subdivisions plus ou moins nettes. c s'interprète donc globalement comme un **coefficient de corrélation**, bien qu'il soit toujours positif.

3

Méthodes de segmentation

Toutes les méthodes de segmentation utilisent le processus de fractionnement de plus en plus poussé de la population selon une structure arborescente.

Une règle d'arrêt, définie à l'avance, détermine l'interruption du processus.

Le critère permettant de mesurer l'intensité de la liaison entre

la variable explicative et la variable à expliquer est le même à toutes les étapes de la segmentation.

n Critère de Belson

Soit Y, la variable à expliquer, qualitative et bimodale, présentant les modalités Y_1 et Y_2 . Les variables explicatives X_i , se découpent de façon dichotomique ($i = 1, \dots, k$). La population étant subdivisée en 2 segments selon la variable X_i , on obtient le tableau ci-dessous :

Y	X_i		
	Segment 1	Segment 2	
Y_1	n_{11}	n_{12}	b_1
Y_2	n_{21}	n_{22}	b_2
	n_1	n_2	n

On utilise les notations suivantes :

$$b_i = n_{i1} + n_{i2} \quad (i = 1, 2)$$

$$n_i = n_{i1} + n_{i2} \quad (i = 1, 2)$$

Les b_i sont des constantes indépendantes de la subdivision effectuée.

Pour une variable X_i donnée, on choisit parmi toutes les subdivisions en 2 segments que l'on s'autorise à effectuer, celle qui rend maximale l'expression :

$$c = \left| n_{11} - \frac{b_1}{n} n_1 \right|$$

n_{11} étant la valeur observée pour une case donnée, $\frac{b_1}{n} n_1$ étant la valeur théorique ou calculée de la case considérée.

Le critère de Belson "c" est donc égal à la valeur absolue de la différence entre l'effectif observé et l'effectif théorique de chacune des cases.

LA SEGMENTATION EN PRATIQUE

La méthode de Belson "emprunte" donc une des étapes du calcul du χ^2 .

On réalise cette opération sur l'ensemble des variables explicatives X_i ($i = 1, \dots, k$) et on retient celle qui donne le **meilleur résultat en termes de valeur de c**.

La population ayant été ainsi subdivisée en 2 segments par rapport à la première variable retenue, on réitère le processus sur chaque nouveau segment ainsi formé avec les autres variables explicatives non encore retenues.

Le processus est interrompu par un critère d'arrêt pré-déterminé.

n La méthode du Chi χ^2

De même que la méthode précédente, pour une variable explicative X_i donnée, on choisit parmi toutes les subdivisions en 2 segments que l'on s'autorise à effectuer, celle qui rend maximale la valeur du χ^2 .

Cette opération étant effectuée sur l'ensemble des variables explicatives X_i ($i = 1, \dots, k$), on retient celle qui donne le **meilleur résultat en termes de valeur du χ^2** .

On réitère ensuite le processus jusqu'à atteindre le critère d'arrêt.

n Méthode AID (Automatic Interaction Detector)

La variable à expliquer Y est quantitative et les variables X_i sont qualitatives. La méthode consiste à **décomposer la variance de Y en fonction de chaque X_i** . On utilise l'équation suivante :

$$\text{Variance totale} = \text{variance inter-groupe} + \text{variance intra-groupe}$$

Le **meilleur découpage** est obtenu avec la variable X_i qui partitionnée, donne la plus grande variance inter-groupe.

4

Conclusion

La segmentation a pour objectif essentiel de mettre en évidence de manière aussi précise que possible les relations existant entre variables explicatives X_i et variable à expliquer Y .

Elle permet ainsi de définir, dans une population, des sous-groupes (ou profils) de patients qui, en fonction de caractéristiques X_i qualitatives, vont se déterminer par rapport à la variable à expliquer Y .

En recherche clinique, la segmentation permet d'émettre des pronostics en fonction de certaines caractéristiques présentées par le patient. Elle permet également de définir le critère le plus déterminant pour un résultat clinique donné.

