

RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA
RECHERCHE SCIENTIFIQUE



UNIVERSITÉ FRÈRES MENTOURI CONSTANTINE
FACULTÉ DES SCIENCES EXACTES
DÉPARTEMENT DE MATHÉMATIQUES

Cours Analyse de données2
Master 2

OPTION
STATISTIQUES APPLIQUÉES

Mme LABDAOUI Ahlam

2021/2022

Table des matières

Introduction générale	v
I Notions préliminaires	1
1 Le recueil de données	2
1.1 Introduction	2
1.2 L'enquête	3
1.2.1 C'est quoi une enquête ? Et pourquoi mener une enquête ?	3
1.2.2 Les types d'enquête	3
1.2.3 Les étapes de faire une enquête	4
1.3 Les principales méthodes du recueil de données	4
1.3.1 L'interview	5
1.3.2 L'observation	7
1.3.3 L'étude de documents	8
1.3.4 Le questionnaire	9
2 L'échantillonnage	15
2.1 Introduction	15
2.2 Quelques définitions	15
2.3 L'échantillonnage	17
2.3.1 Rôle de l'échantillonnage	18
2.3.2 Les types d'échantillonnage	18
2.3.3 Les règles de base de l'échantillonnage	19
2.3.4 L'échantillon par rapport au recensement	21
2.3.5 Le plan d'échantillonnage	21
2.4 Avantages de l'échantillonnage	22
2.5 Inconvénient de l'échantillonnage	23

3	La théorie d'échantillonnage	24
3.1	Introduction	24
3.2	Les principes fondamentaux de la théorie de l'échantillonnage	24
3.2.1	Le principe de validité	24
3.2.2	Le principe de constante statistique	25
3.2.3	Le principe d'optimisation	25
3.3	Statistiques descriptives	25
3.4	La statistique inférentielle	26
3.4.1	Objectif de l'inférence statistique	26
3.4.2	Les hypothèses de la statistique inférentielle	27
3.4.3	La taille d'échantillonnage	27
3.4.4	La distribution d'échantillonnage	28
3.4.5	L'estimation	32
3.4.6	Test d'hypothèse	33
3.4.7	Les erreurs d'échantillonnage dans une enquête	33
3.5	Application dans R	34
II	Les méthodes d'échantillonnages auxiliaires	38
4	Méthodes d'échantillonnages probabilistes avec des exemples	40
4.1	Introduction	40
4.2	Les différentes méthodes d'échantillonnages probabilistes	41
4.2.1	Échantillonnage aléatoire simple	41
4.2.2	Échantillonnage systématique	45
4.2.3	Échantillonnage stratifié	47
4.2.4	Échantillonnage par grappes	51
4.2.5	Échantillonnage à plusieurs degrés	53
4.3	Les avantages des méthodes d'échantillonnage probabiliste	55
4.4	Les inconvénients des méthodes d'échantillonnage probabiliste	55
5	Méthodes d'échantillonnages non probabilistes avec des exemples	56
5.1	Introduction	56
5.2	Les différentes méthodes d'échantillonnages non probabilistes	57
5.2.1	Échantillonnage de commodité « convenance »	57
5.2.2	Échantillonnage de l'aveuglette « accidentelle »	58
5.2.3	Échantillonnage de volontaire	59
5.2.4	Échantillonnage par quotas	61

5.2.5	Échantillonnage par boule de neige	65
5.2.6	Échantillonnage par jugement	67
5.3	Les avantages de méthodes non probabilistes	68
5.4	Les inconvénients de méthodes non probabilistes	69
6	La différence entre les méthodes probabilistes et non proba-	
	bilistes	70
6.1	Introduction	70
6.2	Principales différences entre les échantillons probabilistes et	
	non probabilistes	71
6.3	Résumé	72
III	Analyse des correspondances multiples et applications	73
7	Analyse des correspondances multiples (ACM)	74
7.1	Introduction	74
7.2	Principe de l'ACM	74
7.3	Le problème	75
7.4	Les objectifs	75
7.5	Aide à l'interprétation	76
	7.5.1 Le choix du nombre des axes retenir	76
	7.5.2 L'interprétation des axes	77
7.6	L'interprétation d'un ACM	77
7.7	Conclusion	78
8	Application 1 : Une des Méthodes probabiliste avec un ACM	
	sur R	79
8.1	La démarche de notre travail	79
	8.1.1 Introduction	79
	8.1.2 Démarches d'enquête	80
	8.1.3 La méthode de recueille les données	81
	8.1.4 Echantillonnage	81
	8.1.5 Le but d'étude	82
	8.1.6 Découvrez nos variables et leurs encodage	83
8.2	Partie analyse	85
	8.2.1 Les données	85

8.2.2	Analyse descriptive	85
8.2.3	Analyse des correspondances multiples	96
9	Application 2 : Une des Méthodes non probabiliste avec un ACM sur R	126
9.1	La démarche de notre travail	126
9.1.1	Introduction	126
9.1.2	Démarches d'enquête	127
9.1.3	Echantillonnage	132
9.1.4	Le but d'étude	135
9.2	Partie analyse	135
9.2.1	Les données	135
9.2.2	Analyse descriptive	136
9.2.3	Analyse des correspondances multiples	146
	Conclusion générale	176
	Bibliographie	177

Introduction générale

L'étude de propriétés caractéristiques d'un ensemble, quand on ne dispose pas encore de données, nécessite d'examiner, d'observer des éléments de cet ensemble. La manière de recueillir ces données fait l'objet d'une théorie mathématique appelée théorie des sondages ou encore théorie de l'échantillonnage, en anglais *sampling theory*. Cette théorie concerne l'optimisation de la collecte des données selon divers critères et répond à certaines interrogations sur la façon de procéder à cette collecte en rapport avec l'information disponible et l'effort d'échantillonnage consenti.

Donc, Il est important sinon primordial de savoir c'est quoi l'échantillonnage.

Selon Colin (1970), un échantillon est un fragment d'un ensemble prélevé pour juger de cet ensemble. De nombreuses méthodes d'observations et de mesures appliquées à de tels fragments peuvent être proposées, adaptés à chaque cas particulier en vue d'obtenir une représentation satisfaisante de l'objet étudié. Il faut exprimer explicitement de quelle propriété on veut juger avant de pouvoir concevoir un plan d'échantillonnage. L'échantillonnage doit être adapté à tester l'hypothèse que l'on a fait, à une échelle spatiale et temporelle donnée. Il est impératif de prendre le temps de planifier son échantillonnage.

Il existe deux types de méthodes d'échantillonnage : L'échantillonnage probabiliste et l'échantillonnage non probabiliste. La différence entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une « chance » d'être sélectionnée et que cette chance peut être quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste ; dans ce cas, chaque unité incluse à l'intérieur d'une population n'a pas une chance égale d'être sélectionnée. La section suivante décrit les caractéristiques des deux types d'échantillonnages et fournit des détails sur certaines des méthodes reliées à chaque type.

Dans ce cours , qui s’articule autour de huit chapitres divisés en trois parties, on essaie d’étudier :

Première partie : Notions Préliminaires

Dans cette partie, nous avons mis les notions préliminaires d’échantillonnage. Elle est concernée les méthodes de prélever l’échantillon à travers de faire une enquête par l’un de leur méthodes.

Chapitre 1 : Recueil des données.

Dans ce chapitre, nous avons introduit le recueil des données et leurs méthodes. Comme il s’agit de faire Une enquête pour collecte d’informations portant sur l’ensemble de la population étudiée. Nous avons ainsi défini leurs méthodes : l’interview, l’observation, l’étude d’un document et le questionnaire.

Chapitre 2 : L’échantillonnage.

Ce chapitre est concernant l’échantillonnage qui constitue la phase qui consiste à sélectionner les individus que l’on souhaite interroger au sein de la population de base. Nous avons mis comment préparer un échantillon. Aussi nous avons parlés dans se chapitre au type d’échantillonnage, le rôle d’échantillonnage, et le principe de le faire sur une population cible.

Chapitre 3 : La théorie d’échantillonnage.

Ce dernier chapitre de la première partie a introduit la théorie d’échantillonnage qui consiste d’étudier des liaisons existant entre une population et les échantillons de cette population, prélevés par sondage. Nous avons mentionné comment calculer différents paramètres statistiques de position (moyenne, ...) ou de dispersion (écart type, ...) issus de la statistique descriptive et déterminer des paramètres statistiques d’une population par son recensement exhaustif ou déduire des propriétés de la population à partir de celles de l’échantillon par inférence statistique.

Deuxième partie : Les méthodes d’échantillonnages auxiliaires

Dans cette partie, nous avons mis les différentes méthodes d’échantillonnages. Elle est consterné la définition, le principe et les étapes de chaque type de chaque méthode (probabiliste ou non probabiliste).

Chapitre 4 : Méthodes d’échantillonnages probabilistes avec des exemples.

Dans ce chapitre on introduit les méthodes d’échantillonnages probabilistes. Nous avons défini leurs types : l’échantillonnage aléatoire simple, l’échantillonnage systématique, l’échantillonnage stratifié, l’échantillonnage par grappes et l’échantillonnage a plusieurs degrés avec un ou deux exemples sur chaque type. Aussi nous l’avons mentionné les avantages et les inconvénients de chaque type d’échantillonnage probabiliste et de l’échantillonnage

probabiliste en générale.

Chapitre 5 : Méthodes d'échantillonnages non probabilistes avec des exemples.

Ce chapitre concernée les méthodes d'échantillonnages non probabiliste et leur déférente types : l'échantillonnage de commodité, l'échantillonnage de l'aveuglette, l'échantillonnage de volontaire, l'échantillonnage par quotas, l'échantillonnage par boule de neige et l'échantillonnage par jugement avec un exemple sur chaque type. Aussi nous l'avons mentionné les avantages et les inconvénients de chaque type d'échantillonnage non probabiliste et de l'échantillonnage non probabiliste en générale.

Chapitres 6 : La différence entre les méthodes probabilistes et non probabilistes

Dans ce chapitre nous avons mentionné la différence entre les méthodes d'échantillonnages probabilistes et non probabilistes. Nous fiassions les Principales différences entre les échantillons probabilistes et non probabilistes, et à la fin nous avons donné un résumé.

Troisième partie : Applications sur les méthodes d'échantillonnages auxiliaires.

Dans cette partie, nous avons appliqué deux méthodes d'échantillonnage probabiliste et non probabiliste, et nous avons étudié les échantillons obtenus dans logiciel R par analyse correspondantes multiples.

Chapitre 07 : Analyse des correspondances multiples (ACM).

Dans ce chapitre, nous définissons le principe ; le problème et l'objectif de l'analyse correspondants multiples(ACM), et comment interprété les résultats de l'ACM.

Chapitre 08 : Application 1 : Une des Méthodes probabiliste avec un ACM sur R.

Ce chapitre contient une petite définition de l'échantillon étudié, qui est un échantillon des patients du cancer de VADS. Aussi, Ce chapitre explique la démarche de notre travail qui contient : la méthode de recueille les données, la méthode d'échantillonnage, le but d'étude, la découverte des variables et l'encodage des données. Enfin, Nous avons analysé les données par l'analyse descriptive et l'analyse correspondantes multiples et nous avons conclus quelque résultat sur les patients de VADS.

Chapitre 09 : Application 2 : Une des Méthodes non probabiliste avec un ACM sur R.

Ce chapitre contient une introduction expliquant l'épidémie de COVID-19 et l'étendue de l'importance de la sensibilisation du peuple algérien. Aussi, Ce chapitre explique la démarche de notre travail à savoir la méthode de recueille

CHAPITRE 0. INTRODUCTION GÉNÉRALE

les données ; la méthode d'échantillonnage ; le but d'étude ; la découverte des variables et l'encodage des données. Nous avons analysé les données par l'analyse descriptive et l'analyse correspondantes multiples et nous avons conclu quelque résultat sur la conscience du peuple algérien pendant la pandémie du COVID-19.

Première partie

Notions préliminaires

Chapitre 1

Le recueil de données

1.1 Introduction

Le recueil de données correspond à toute opération qui a pour objet la recherche, la collecte et l'exploitation d'informations. Il permet par exemple de collecter des informations précises et pertinentes sur un patient afin de mieux appréhender sa situation.

C'est une recherche d'informations primaires, de données, qui se fait par le biais d'une enquête. Il est du domaine de l'investigation.

De façon générale, on est amené à rechercher de l'information ou recueillir des données lorsqu'on désire cerner une situation soit :

- Pour évaluer.
- Pour analyser des besoins.
- Pour prendre une décision.
- Pour améliorer un fonctionnement, des performances.
- Pour former.
- Pour résoudre un problème.
- Pour cerner un phénomène.
- Pour tester des hypothèses scientifiques.

1.2 L'enquête

1.2.1 C'est quoi une enquête ? Et pourquoi mener une enquête ?

Une enquête est une recherche méthodique d'informations auprès d'une population ciblée. Elle s'articule autour de plusieurs techniques afin de combiner quantité et qualité des données.

Lorsqu'on ne dispose pas d'information, il est indispensable de mener une enquête. Elle constitue une «garantie» précieuse en permettant de recueillir des renseignements. Une enquête peut être réalisée pour :

- Prendre une décision.
- Repérer des dysfonctionnements.

1.2.2 Les types d'enquête

Enquêtes périodiques

- Il fondées sur un plan d'échantillonnage où la taille d'échantillon augmente en même temps que la population.
- L'enquêteur doit S'assurer que le plan des enquêtes périodiques soit aussi souple que possible, pour pouvoir faire face aux changements futurs, comme l'augmentation ou la réduction de la taille de l'échantillon, la restratification, le ré-échantillonnage et l'actualisation des probabilités de sélection.
- L'enquêteur peut utiliser un plan d'échantillonnage avec rotation, qui remplace une partie de l'échantillon à chaque période.
- Ce type a aussi l'avantage de réduire les coûts, lorsque le premier contact est beaucoup plus coûteux que les contacts subséquents.
- L'enquêteur mit en place une stratégie d'actualisation pour élaborer des procédures visant à surveiller la qualité du plan d'échantillonnage au fil du temps.

Enquêtes longitudinales

- Pour les enquêtes longitudinales par panel, l'enquêteur détermine la durée du panel (sa durée dans l'échantillon) en tentant de maintenir un juste équilibre entre la satisfaction des besoins de l'enquête (données relatives à la durée), d'une part, et les effets d'attrition et de conditionnement de l'échantillon, d'autre part.

- Il importe grandement de choisir un plan d'échantillonnage dont les caractéristiques sont simples (c.-à-d. base de sondage unique, réduction du nombre de degrés et de phases au minimum).
- Il est recommandé de réserver – principalement – les enquêtes longitudinales à la production d'estimations longitudinales.

1.2.3 Les étapes de faire une enquête

Chaque enquête est spécifique. Mais un certain nombre d'étapes sont incontournables.

1. En premier lieu, lorsque l'enquête est réalisée sur un échantillon (i.e. une partie de la population seulement), il faut choisir un plan de sondage, c'est-à-dire une méthode de tirage de l'échantillon adaptée aux contraintes financières et logistiques, tout en étant le plus efficace possible d'un point de vue statistique.

2. En second lieu, une fois l'enquête réalisée, se pose la question de l'estimation, c'est-à-dire la manière de produire, à partir de l'échantillon obtenu, des grandeurs valables pour l'ensemble de la population (une extrapolation). Cette extrapolation passe souvent par le calcul d'une pondération, qui tient compte du plan de sondage mais aussi des non-réponses et des sources de données externes permettant d'améliorer l'estimation. On parle alors de redressement.

3. Enfin, on peut chercher à évaluer la qualité de l'échantillon obtenu à travers des indicateurs comme le biais ou la précision.

1.3 Les principales méthodes du recueil de données

Pour recueillir des données, il est nécessaire d'élaborer une stratégie qui va faire appel à des méthodes. Le choix d'une de ces méthodes doit être basé sur un ensemble de critères fixé au préalable à savoir : le rapport entre le degré de détail et de précision souhaité et les coûts que cela entraîne en termes d'énergie, ainsi qu'en termes de temps d'exécution, de traitement et de disponibilité des résultats finaux.

Les principales méthodes sont :

- Les méthodes face à face "La pratique d'interviews".
- L'observation.

1.3. LES PRINCIPALES MÉTHODES DU RECUEIL DE DONNÉES

- Les enquêtes par questionnaires.
- L'étude de documents.

1.3.1 L'interview

Définition 1.3.1 *L'interview est une méthode de recueil d'informations qui consiste en des entretiens oraux, individuels ou de groupes, avec plusieurs personnes sélectionnées, afin d'obtenir des informations sur des faits ou des représentations, dont on analyse le degré de pertinence, de validité et de fiabilité en regard des objectifs du recueil d'informations.*

Une interview est un entretien individuel en face à face avec des personnes issues du groupe cible. Elle peut être basée sur un questionnaire ou consister en une conversation ouverte, ou combiner les deux.

Les questions peuvent être :

Ouvertes.

Semi-ouvertes.

Fermées.

Les types de l'interview

Entretien dirigé est caractérisé par :

- Discours qui suit l'ordre des questions posées.
- Questions préparées à l'avance et posées dans un ordre bien précis Information partielle et raccourcie.
- Information recueillie rapidement ou très rapidement.
- Inférence assez faible.

Entretien semi-dirigé est caractérisé par :

- Discours «par paquets», dont l'ordre peut être plus ou moins déterminé.
- Quelques points de repère pour l'interviewer Information de bonne qualité, orientée vers le but poursuivi.
- Information recueillie dans un laps de temps raisonnable.
- Inférence modérée.

Entretien libre est caractérisé par :

- Discours continu.
- Aucune question préparée à l'avance Information de très bonne qualité, mais pas nécessairement pertinente

- Durée de recueil d'informations non prévisible.
- Inférence exclusivement fonction du mode de recueil.

La préparation de l'interview

1. Préciser les questions générales et spécifiques de recherche.
2. Faire une première rédaction des questions.
3. Organiser la séquence des questions.
4. Adapter le processus de l'interview aux objectifs visés (transitions entre thèmes).
5. Préparer l'introduction et les conclusions.
6. Préparer le système de notation des réponses.
7. Tester le protocole de l'interview La réalisation de l'interview.

Quelques compétences de base

- Écouter attentivement.
- Adopter une attitude ouverte et empathique.
- Paraphraser.
- Résumer les contenus lorsque vous changez de thème ou à la fin d'une phase importante.

Avantages de l'interview

- Un plus haut pourcentage de réponses peut être obtenu.
- Les personnes réellement sélectionnées peuvent être atteintes.
- Tout type de population peut être atteint.
- On est mieux informé sur les caractéristiques de ceux qui ne répondent pas.
- Les réponses sont spontanées.
- Les réponses sont définitives.
- Les réponses sont surtout plus riches, plus nuancées.
- Permet d'explorer le ressenti, les sentiments, les impressions, les opinions, les attitudes.
- L'enquête peut porter sur des sujets plus délicats et personnels.
- L'enquête peut être rendue plus divertissante et donc mieux acceptée.

Inconvénients de l'interview

- L'apprentissage est indispensable pour bien manipuler la technique de l'entretien semi-directif.
- C'est un outil qui demande du temps (temps d'entretien, temps de dépouillement, temps d'analyse).
- Méthode très coûteuse.
- Réactivité maximale.
- L'enregistrement papier des réponses peut être difficile.
- L'attitude de l'interviewer peut affecter la validité des réponses.

1.3.2 L'observation

L'observation est un processus dont l'action première immédiate est de recueillir de l'information sur l'objet pris en considération en fonction de l'objectif de l'étude. Ce recueil suppose une activité de codage : l'information brute sélectionnée est traduite grâce à un code pour ensuite être transmise.

Il existe différentes procédures d'observation :

1. L'observation systématique

- Grille d'observation standardisée utilisée de manière systématique.
- Pas d'implication personnelle.
- Degré de réactivité des sujets.

2. L'observation participante

- Implication active.
- Degré de participation variable.
- Grille d'observation.

3. L'observation libre

- Pas de grille.

Avantages de l'observation

- Les observations permettent d'appréhender une réalité vécue, plutôt que d'en obtenir un écho éventuellement déformé au travers des représentations que les gens s'en forgent.

Inconvénients de l'observation

- Les événements sont non répétitifs.
- Les limites sont aussi d'ordre géographique.

- Dans toute procédure d'observation, les personnes observées ont tendance à modifier leur comportement.

1.3.3 L'étude de documents

Définition 1.3.2 *Document = toute trace, déjà existante, de l'activité humaine, qu'elle soit sonore, visuelle ou informatique.*

L'étude de documents est une des méthodologies qualitatives utilisées dans les sciences sociales et humaines. Elle consiste à faire de la recherche en analysant des documents pour mieux comprendre le phénomène à étudier. Elle dépend de :

- *La nature des documents à analyser.*
- *De la quantité des documents à analyser.*
- *De l'objet et du but de l'investigation.*

La nature des documents à analyser

Différents types de documents peuvent être concernés :

- Documents scientifiques.
- Documents officiels (textes de lois, programmes d'enseignement, normes...).
- Documents à usage limité dans le temps.

La quantité des documents à analyser

La quantité des documents à analyser déterminera le type d'analyse à effectuer : une analyse exhaustive, ou une analyse par échantillonnage ou par sélection.

L'objet et le but d'investigation par document

Deux types d'analyse des documents sont distingués :

- Le type d'analyse de documents présentant un caractère exploratoire comme par exemple la recherche documentaire, dont l'objectif est l'exploration en vue d'élaborer une problématique (cadre théorique et hypothèse).
- Le type d'analyse présentant un caractère confirmatoire où le document sélectionne selon une stratégie est traité comme une donnée, au même titre que le discours recueilli par l'entretien ou un comportement recueilli par l'observation, et dont l'objectif est de vérifier une hypothèse.

1.3.4 Le questionnaire

Définition 1.3.3 *Selon qu'il s'agisse d'une évaluation des performances de personnes, ou au contraire d'une évaluation d'un fonctionnement, de l'évaluation d'un système, d'une recherche descriptive ou expérimentale, le questionnaire prendra deux sens différents.*

Sens 1 : Le questionnaire de contrôle de connaissances qui permettra de recueillir des données sur soit une activité demandée à un individu (donner une définition, construire. . .) soit un contenu sur lequel s'exerce cette activité (une formule, une démarche à suivre etc.)

Sens 2 : Le questionnaire d'enquête concerne l'enquête au sens d'une étude d'un thème précis auprès d'une population dont on détermine un échantillon afin de préciser certains paramètres.

Les étapes d'une enquête par questionnaire

Les étapes essentielles de la réalisation d'un questionnaire sont :

1. Définir l'objet de l'enquête.
2. Définir les hypothèses de l'enquête.
3. Inventaire des moyens matériels mis à la disposition des réalisateurs de l'enquête (problème de la durée de l'enquête, le budget alloué à l'enquête.).
4. Recherches préalables.
5. Détermination des objectifs et des hypothèses de travail (que veut-on montrer ?).
6. Déterminer la population ciblée ou l'univers de l'enquête.
7. Construction de l'échantillon (taille de l'échantillon, procédé de l'échantillonnage, etc.).
8. Rédaction du projet du questionnaire (choix de la méthode d'administration, choix des questions, mise en forme etc.).
9. Mise à l'épreuve du questionnaire (pré-test ou enquête pilote).
10. Rédaction du questionnaire définitif (mise en page, pré codage, etc.).
11. Formation des enquêteurs.
12. Réalisation matérielle de l'enquête (collecte des questionnaires remplis, contrôle des questionnaires etc.).
13. Le dépouillement et la transcription des réponses.
14. Codage des questionnaires.
15. Dépouillement des questionnaires.
16. Analyse des résultats.

Rédaction du questionnaire

De la qualité de votre questionnaire vont dépendre la qualité des réponses et l'utilité des données recueillies pour votre étude. Il faut donc soigner les questions, structurer correctement le questionnaire et éviter certains écueils :

L'introduction du questionnaire Votre questionnaire doit impérativement s'accompagner en introduction d'un mot explicatif concis et efficace précisant :

- Ta présenter.
- Vos références.
- Dans quel cadre vous effectuez cette enquête.
- Indiquer les objectifs de l'enquête (dans quel but vous réalisez cette enquête).
- Pourquoi vous interrogez cette personne.
- Préciser le délai de réponse.
- L'anonymat de l'enquête.
- Des remerciements pour le temps consacré.

Exemple 1.3.1 *Madame, Monsieur, étudiant en Sociologie politique à L'université X, je réalise un mémoire de fin d'études sur les pratiques électorales des seniors.*

Dans ce cadre, je vous remercie de bien vouloir consacrer quelques minutes pour répondre au questionnaire ci-joint.

Vos réponses sont anonymes.

Les types des questions

Questions fermées Dans ces questions, les réponses sont fixées à l'avance et le répondant doit choisir parmi l'éventail qui lui est présente.

Ce type de question est celui qui se prête le mieux au dépouillement et à l'analyse statistique.

En effet, on peut répartir les différents répondants selon la réponse qu'ils ont fournie sans passer par des étapes d'analyse intermédiaires.

- Ces questions se comprennent facilement et on peut répondre facilement.
- Elles garantissent un certain anonymat.
- Elles peuvent servir de questions filtres, c'est-à-dire de discriminants servant de répartir les répondants entre plusieurs séries de réponses ultérieures.

1.3. LES PRINCIPALES MÉTHODES DU RECUEIL DE DONNÉES

Cependant, l'inconvénient de ces questions est de dicter la réponse au répondant n'autorisant aucune expression de nuance. Leur champ d'application se limite au recueil de caractéristiques objectives.

Exemple 1.3.2 *Lisez-vous des périodiques rédigés dans une autre langue que votre langue maternelle ?*

- *Régulièrement.*
- *Occasionnellement.*
- *Rarement.*
- *Jamais.*

Exemple 1.3.3 *(questions filtres)*

Jugez-vous la mise en place d'un centre de documentation nécessaire :

- *Oui.*
- *Non.*

Si oui, pourquoi ?

- *Accès rapide à l'information.*
- *Disponibilité d'information actualisée.*

Questions ouvertes Dans les questions ouvertes, la réponse n'est pas prévue et le répondant est libre de s'exprimer comme il veut.

On réserve, dans le questionnaire, un emplacement suffisant pour que la réponse de la personne puisse être enregistrée en son entier.

Caractéristiques :

- Si ces questions sont bien formulées, on pourra obtenir de bonnes informations sur n'importe quel sujet.
- Elles sont indispensables pour recueillir des renseignements sur des problèmes délicats.
- Leur utilisation s'impose lorsqu'on ne peut prévoir les réponses possibles, dans ce cas on essayera d'abord d'énoncer une question de type semi ouverte.
- Leur formulation est très délicate : elles doivent être aisément comprises et ne comporter aucune ambiguïté ou contresens.
- Leur dépouillement est plutôt difficile, il faut souvent ressortir les occurrences pour pouvoir faire l'analyse statistique.

Exemple 1.3.4 *Quelle est, à votre avis, la meilleure formule d'examen possible dans l'enseignement universitaire ?*

Questions semi ouvertes ou semi fermées Les principales réponses possibles sont prévues, comme dans une question fermée, mais on laisse la possibilité d'ajouter des réponses libres, en dehors de l'éventail proposé, comme dans une question ouverte.

Caractéristiques :

Elles contribuent surtout à faciliter le dépouillement étant donné que le nombre de réponses est déjà prévu. Elles risquent d'influencer la réaction du répondant, par la suggestion de réponses qui peuvent paraître plus habituelles ou moins habituelles et la partie questions fermées doit être bien établie.

Exemple 1.3.5 *Avez-vous déjà utilisé des produits documentaires ?*

- Bulletin signalétique.*
- Bulletin des sommaires.*
- Liste des acquisitions.*
- Catalogue imprimé de la bibliothèque.*
- Autres :*

Questions mixtes Il y a un item de réponses style « autre » et on a en même temps la question fermée avec les modalités de réponses présentées à la personne interviewée. Dans ce cas on flexibilise l'instrument ; on a la fois la standardisation et on évite l'oubli d'offrir certaines modalités de réponses qui est une erreur fatale, car on ne peut la réparer. D'autre part en cas de doute cela permet de laisser une réponse ouverte.

Question « cafeteria » Renvoie aux questions fermées avec à une liste plus longue de réponses, d'autre part les modalités de réponses peuvent être unique ou multiples.

La structuration du questionnaire L'enchaînement des questions à toute son importance. Quelques conseils :

- Regrouper les questions de même type.
- Classer les questions par thème.
- Aller des questions générales aux questions spécifiques.
- Commencer le questionnaire par des questions simples qui mettent en confiance la personne interrogée.
- Terminer le questionnaire par les questions d'identification (âge, sexe, CSP...).

1.3. LES PRINCIPALES MÉTHODES DU RECUEIL DE DONNÉES

- Intercaler entre les groupes de questions difficiles ou délicates des questions plus faciles ;
- Limiter le nombre de questions afin de ne pas décourager la personne enquêtée.
- Veillez à la longueur du questionnaire. Plus il sera long, plus il découragera les personnes d'y répondre. Vérifiez que toutes les questions posées sont utiles pour traiter votre objet d'étude. Certaines sont peut être intéressantes mais n'ont aucun lien avec votre sujet. Vérifiez que deux questions ne se répètent pas.
- Que le questionnaire ne dépasse pas une longueur optimale : 45 à 60 minutes, 20 à 30 questions au maximum.
- Donner au questionnaire un aspect cohérent et logique, tout en tenant compte de la possibilité de perturbations qui peuvent être provoquées chez le répondant.

Les biais à éviter Pour faire un bon questionnaire, il doit :

- Evitez les mots trop techniques, ambigus ou trop vagues. Les questions doivent être simples et compréhensibles, sans double sens.
- Evitez les questions à double négation. Par exemple : « N'estimez-vous pas que les campagnes publicitaires n'ont pas d'impact sur les personnes de plus de 80 ans ? »
- Pour les choix de réponses, préférez les fréquences (plus de 3 fois par semaine, plus de 10 fois par semaine) plutôt que les expressions d'appréciation (régulièrement, rarement...) car ces dernières sont subjectives.
- Détournez les questions qui mettent en jeu la désirabilité sociale, le prestige... Optez pour des questions indirectes. Par exemple : si vous posez la question : « Lisez-vous une histoire à votre enfant le soir ? » Beaucoup vous répondront oui car en répondant « oui » les personnes interrogées auront le sentiment d'être de bons parents, mais la réalité est peut être différente. Pour éviter ce biais, passez par une question indirecte. Par exemple : « Quelles sont vos activités de 19h à 22h ? »
- Evitez les questions dont la formulation sous-entend une réponse évidente. Par exemple : « La confiance est-elle selon vous un facteur de durabilité du couple ? » Tout le monde vous répondra « oui ». Préférez alors la question suivante : « Quels sont, selon vous, les facteurs de durabilité du couple ? »
- Veillez toujours à la neutralité du ton des questions pour ne pas mettre

mal à l'aise la personne interrogée. Aucune des questions ne doit refléter un jugement sur la personne enquêtée.

Avantages du questionnaire

- C'est un outil facile à manipuler.
- Un public nombreux peut être touché.
- La technique est rapide.
- Elle permet d'atteindre des personnes fortes éloignées.
- Le coût apparent est inférieur à celui de l'interview.
- L'enquête reste plus anonyme.
- Il n'y a pas de biais introduit par la présence de l'enquêteur.
- Les réponses peuvent être réfléchies.
- Les répondants peuvent choisir de répondre quand cela leur convient le mieux, ce qui peut faciliter leur participation.

Inconvénients du questionnaire

- Il peut apparaître comme un outil d'élaboration facile mais il requiert une certaine compétence basée sur une technique sûre
- Il prive le chercheur de beaucoup d'observations accessibles lors d'un contact direct.
- Certains thèmes sont difficilement abordables dans un questionnaire.
- Difficulté liée aux et au contrôle des non-réponses.

Chapitre 2

L'échantillonnage

2.1 Introduction

Pour effectuer une étude statistique (enquête, observation ou expérimentation), on se sert généralement d'un échantillon. Celui-ci doit refléter le plus exactement possible l'image de la population. En fait, choisir un échantillon, c'est mettre ensemble un certain nombre d'individus qui composeront une sorte de modèle réduit de la population à laquelle ils appartiennent. Mais comment choisir ces individus ? C'est à cette question que nous répondrons, en présentant les méthodes d'échantillonnage les plus courantes.

Echantillonner, c'est choisir une partie d'une population pour représenter l'ensemble de la population. Si vous travaillez sur un échantillon, c'est pour chercher une information sur l'ensemble de sa population mère. Les résultats observés sur cet échantillon n'auront de sens que s'ils sont rapportés à toute la population. On a donc recours à des techniques bien précises pour construire un échantillon et s'assurer qu'il est représentatif de sa population mère.

2.2 Quelques définitions

Définition 2.2.1 *Population est un ensemble que l'on observe et qui sera soumis à une analyse statistique, chaque élément de cet ensemble est un individu ou unité statistique.*

Définition 2.2.2 *Population cible est une partie de la population totale pour laquelle nous voulons généraliser les résultats de l'échantillon.*

Définition 2.2.3 *Population finie est une population qui consiste en un nombre fini d'éléments.*

Définition 2.2.4 *Population infinie est une population est infinie s'il n'y a pas de limite au nombre d'éléments qu'il contient.*

Définition 2.2.5 *Population homogène est une population avec des éléments qui possèdent les mêmes caractéristiques.*

Définition 2.2.6 *Population non homogène est une population avec des éléments qui ne possèdent pas les mêmes caractéristiques.*

Définition 2.2.7 *Population échantillonnée est une population à partir de laquelle l'échantillon est constitué.*

Définition 2.2.8 *Echantillon est un sous ensemble de la population considérée, le nombre d'individus dans l'échantillon est la taille de l'échantillon.*

Définition 2.2.9 *Unité d'échantillonnage est un élément choisi de la population pour faire partie de l'échantillon.*

Définition 2.2.10 *Distribution d'échantillonnage est une distribution de probabilité composée de toutes les valeurs possibles d'une statistique d'échantillon.*

Définition 2.2.11 *Schème ou plan d'échantillonnage est une procédure servant à définir la population cible, la méthode d'enquête, le choix de l'échantillon, et les caractéristiques des unités.*

Définition 2.2.12 *Théorie d'échantillonnage est une étude des liaisons existant entre une population et les échantillons de cette population.*

Définition 2.2.13 *Méthodes d'échantillonnage est un ensemble des méthodes permettant de réaliser un sondage (de prélever un échantillon de données) au sein d'une population, de manière à reproduire un échantillon aussi représentatif que possible de cette population.*

Définition 2.2.14 *Echantillonnage sans remise c-à-d chaque élément ne peut être sélectionné qu'une fois :*

$$\text{Nombre d'échantillons possibles : } C = \frac{N!}{n!(N-n)!}.$$

$$\text{Probabilité qu'un échantillon particulier soit tiré} = \frac{1}{C}.$$

Définition 2.2.15 *L'inférence statistique est définie comme le processus d'utilisation des données d'un échantillon pour estimer ou tester des hypothèses sur les caractéristiques numériques (« paramètres ») d'une population.*

Définition 2.2.16 *Tirage exhaustif est une tirage sans remise.*

Définition 2.2.17 *Paramètre est une caractéristique numérique d'une population telle que la moyenne de la population μ , l'écart type de la population σ et la proportion de la population p .*

Définition 2.2.18 *Les individus statistiques ou unités statistiques sont les éléments de la population statistique.*

Définition 2.2.19 *Unité de base est une unité d'échantillonnage ou unité de sondage, c'est l'élément pris en considération dans l'enquête.*

Définition 2.2.20 *Sondage c-à-d enquête incomplète, enquête partielle ou enquête par échantillonnage, c'est une enquête au cours de laquelle seulement une partie des unités de base de la population sont observée.*

Définition 2.2.21 *Base de sondage est une énumération ou présentation ordonnée de toutes les unités de base constituant la population.*

Définition 2.2.22 *Erreur d'échantillonnage est un écart entre les résultats obtenus auprès d'un échantillon et ce que nous apprendrait un recensement comparable de la population. Plus la taille de l'échantillon est grande plus l'erreur d'échantillonnage diminue.*

Définition 2.2.23 *Fraction ou taux de sondage est une proportion des unités de la population qui font partie de l'échantillon. C'est le rapport entre la taille de l'échantillon n , et la taille de la population N .*

2.3 L'échantillonnage

La théorie mathématique des probabilités suppose que, pour connaître les événements qui peuvent survenir dans une population donnée, il n'est possible d'étudier ou d'interroger qu'une petite partie de celle-ci, à condition de respecter des règles rigoureuses de sélection de cette fraction de population. Seules garanties de sa représentativité.

L'échantillonnage est un procédé qui permet de définir un échantillon dans un travail d'enquête. Il s'agit d'étudier une partie sélectionnée pour établir des conclusions applicables à un tout.

En d'autres termes, l'échantillonnage est une sélection précise de personnes ciblées pour réaliser un entretien, un focus group, un sondage ou un questionnaire.

2.3.1 Rôle de l'échantillonnage

Lorsqu'on souhaite collecter les informations sur une population, deux possibilités s'offrent : La première solution consiste à observer ou interroger tous les éléments de la population, c'est ce qu'on appelle une enquête complète ou enquête exhaustive ou recensement. La seconde solution consiste à observer ou interroger une partie de la population, c'est ce qu'on appelle enquête partielle ou sondage. Les éléments de la population qui sont réellement observés constituent l'échantillon et l'opération qui consiste à choisir ces éléments est appelée échantillonnage. L'alternative décrite ci-dessus se présente dans beaucoup de situations et le recours à la deuxième solution c'est à dire l'enquête partielle et la pratique la plus courante. Par rapport à l'enquête complète, l'enquête partielle offre une série d'avantages. Le coût global de l'enquête partielle est en général plus réduit que le coût global d'une enquête complète. L'enquête par sondage est plus rapide que l'enquête complète, surtout lorsque la caractéristique étudiée présente des modifications assez importantes au cours du temps. Les erreurs d'observations sont plus réduites que dans l'enquête exhaustive. En fin dans certaines situations particulières, l'enquête partielle est la seule solution possible, c'est le cas lorsque l'observation présente un caractère destructif.

2.3.2 Les types d'échantillonnage

Pour effectuer un échantillonnage, vous avez deux options :

1. **L'échantillonnage représentatif.**
2. **L'échantillonnage aléatoire (aussi appelée « méthode probabiliste »).**

L'échantillonnage représentatif

Un échantillon représentatif, souvent utilisé dans une étude quantitative (questionnaire ou sondage), est défini comme représentatif lorsqu'il a les mêmes caractéristiques que la population étudiée (population mère).

Un échantillon représentatif peut se faire à travers l'utilisation de la technique des quotas.

Exemple 2.3.1 *Dans le cadre d'un mémoire sur la consommation d'alcool en France, l'étudiant chercheur doit réaliser un échantillon représentatif de la population française en termes d'âge, de genre, de CSP... L'étudiant étudiera donc un échantillon de 1000 personnes ayant les caractéristiques suivantes :*

- 51% de femmes et 49% d'hommes.
- 20% de population rurale et 80% de population urbaine.
- 20,5% d'ouvriers, 27,5% d'employés, 25,5% de professions intermédiaires, 18,5% de cadres, 6,5% d'artisans et 1,5% d'agriculteurs.

L'échantillonnage aléatoire

L'échantillonnage aléatoire (ou “méthode d'échantillonnage probabiliste”) est déterminé à partir d'une procédure de tirage aléatoire statistique.

Malgré le hasard, la représentativité de l'échantillon aléatoire est assurée par les lois statistiques de la probabilité.

Pour connaître les habitudes radiophoniques de la population française, un étudiant pourrait réaliser un échantillonnage aléatoire et contacter des personnes au hasard. C'est la technique utilisée par Médiamétrie, une entreprise qui étudie les audiences radiophoniques en France.

2.3.3 Les règles de base de l'échantillonnage

Comme l'indique la définition des chercheurs ci-dessus, l'échantillonnage est efficace “à condition de respecter des règles rigoureuses”. Nous allons par conséquent vous expliquer comment réaliser un échantillonnage correctement.

Bien choisir les personnes à interroger

Les caractéristiques de l'échantillonnage sont strictes : pour être efficace, il ne vous faut interroger que les individus susceptibles de vous apporter des informations pertinentes pour votre recherche !

Pour respecter la règle de la représentativité de l'échantillonnage, vous devez cibler les personnes interrogées.

La taille de l'échantillon en fonction de l'étude

Une étude empirique peut être menée par le biais d'une étude qualitative ou une étude quantitative. À travers le type d'étude et la technique d'enquête utilisée, la taille de l'échantillon varie.

Etude qualitative Lors d'une étude qualitative, la taille de l'échantillon peut être extrêmement restreinte : 1 à 2 personnes.

En fonction des informations que vous voulez obtenir, il est important de sélectionner la bonne personne à interroger.

Etude quantitative Dans le cadre d'une étude quantitative, l'échantillon à étudier doit être le plus représentatif possible de la population ciblée.

Cet échantillon doit également être assez important, afin d'obtenir des données statistiques pertinentes et une conclusion efficace.

Les techniques d'interrogation

Selon le type de l'étude et de l'outil employé pour récolter des informations, l'échantillon à interroger le sera de différentes façons.

Pour l'entretien semi-directif ou libre, l'échantillon à questionner est de petite taille. L'étudiant a donc le temps d'interroger la personne en face à face, pendant plus d'une heure.

Pour l'entretien directif, qui se trouve à mi-chemin entre l'étude qualitative et l'étude quantitative, l'étudiant a le choix entre réaliser ces entretiens en face à face s'il a le temps, ou bien par e-mail, téléphone ou via les réseaux sociaux, si le nombre d'entretiens devient trop conséquent.

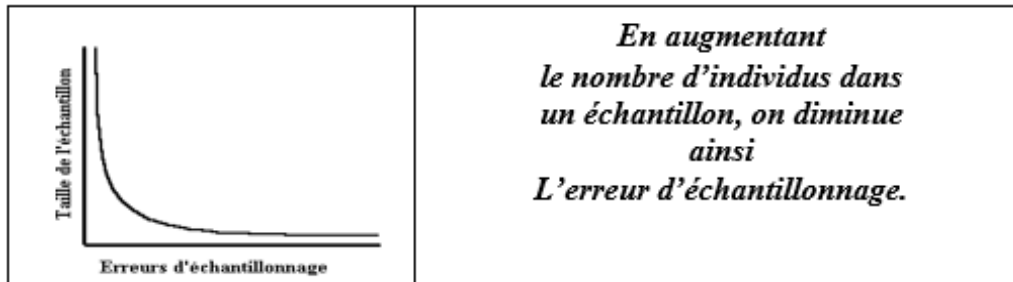
Pour un focus group, l'échantillon d'une dizaine de personnes doit obligatoirement être interrogé lors d'une réunion organisée avec tous les participants.

Pour un sondage ou un questionnaire, l'échantillon doit souvent représenter plus de 500 *personnes*.

2.3.4 L'échantillon par rapport au recensement

Pour diminuer le risque d'erreurs d'échantillonnage, vous n'avez qu'à inclure toute la population dans votre enquête. Les erreurs de toutes sortes seront réduites à zéro, ou presque.

L'erreur d'échantillonnage, c'est l'écart entre les paramètres de la population cible et les valeurs estimées de l'échantillon. Les erreurs d'échantillonnage diminuent à mesure que l'échantillon augmente. Elles atteignent la valeur zéro lorsque l'échantillon se substitue à la population : l'échantillon comprend la population cible au complet. Comme le montre le graphique suivant, la diminution des erreurs est inversement proportionnelle à la racine carrée de la taille de l'échantillon.



Vous constatez en même temps qu'il est utopique de penser que pour la moindre expérience ou enquête il faut faire appel à toute la population. Il n'est donc pas nécessaire de tenir un recensement pour toute enquête. Un échantillon suffit amplement, car il présente des avantages intéressants et souvent insoupçonnés.

2.3.5 Le plan d'échantillonnage

Une fois les objectifs, les lignes directrices et les définitions élaborées, le statisticien peut travailler au plan de sondage, qui comporte trois parties :

- Le plan d'échantillonnage : la façon dont on prélèvera l'échantillon.
- Les techniques d'estimation : la façon dont on appliquera les résultats établis à partir de l'échantillon à l'ensemble de la population.
- Les mesures de la précision : la façon dont on mesurera l'erreur d'échantillonnage.

Nous étudierons le plan d'échantillonnage. Les étapes suivantes amènent à définir complètement le plan d'échantillonnage :

1. Définir une problématique et des hypothèses.

2. Choisir le type d'étude (qualitative ou quantitative).
3. Choisir l'outil le plus adéquat (entretien, focus group, questionnaire, sondage) car la taille de l'échantillon n'est pas la même selon l'outil employé.
4. Déterminer ce que sera la population observée (p. ex., des élèves, des hommes de 20 à 35 *ans*, des nouveau-nés, etc.).
5. Définir un échantillon efficace : un questionnaire ou un sondage demanderont un échantillon large alors qu'un focus group ou entretien beaucoup moins .
6. Sélectionner une méthode d'échantillonnage.
7. Cadrer votre échantillonnage : définissez un début et une fin à votre étude (saturation théorique).

2.4 Avantages de l'échantillonnage

- Par rapport au recensement, l'échantillon permet de réduire les couts à bien des égards. Un recensement mobilise tellement de ressources humaines et matérielles que seul un pays peut se le permettre à intervalles réguliers.

- L'échantillon permet de mieux suivre les unités échantillonnées. Si votre enquête consiste à rejoindre 250 personnes par téléphone, vous pourrez aisément exécuter des rappels téléphoniques pour les personnes absentes lors de votre premier appel.

- Puisque votre échantillon est restreint, vous pourrez poser un plus grand nombre de questions aux personnes interrogées. Ainsi, votre recherche gagnera en qualité.

- La qualité de données que vous obtiendrez auprès de votre échantillon pourra être manipulée aisément à l'aide de la technologie information courante. Aussi, les analyses statistiques seront exécutées avec plus de rapidité.

- Si nous échantillon a été tiré selon les règles de l'art, il comportera moins d'erreurs d'échantillonnage que le recensement. N'y a-t-il pas là contradiction avec le texte précédent ? L'échantillon ne peut être plus exact que le recensement, en théorie. Mais, en pratique, les experts soutiennent qu'il est impossible de rejoindre toute la population dans un court laps de temps pour l'interroger à fond. Par contre, si l'échantillon est restreint, bien peu d'individus choisis échapperont à l'enquête.

2.5 Inconvénient de l'échantillonnage

- L'échantillonnage a pour but de fournir suffisamment d'informations pour pouvoir faire des d'édutions sur les caractéristiques de la population. Mais bien entendu, les résultats obtenus d'un échantillon à l'autre vont être en général différents et différents également de la valeur de la caractéristique correspondante dans la population. On dit qu'il y a des fluctuations d'échantillonnage. Comment, dans ce cas, peut-on tirer des conclusions valables ?

Heureusement, l'erreur d'échantillonnage est quantifiable et les résultats obtenus à partir des échantillons sont utilisables pour la prise de décision.

En effet, associé à un certain degré de confiance, il sera possible de faire des approximations de la population.

Chapitre 3

La théorie d'échantillonnage

3.1 Introduction

Sur un échantillon, on peut calculer différents paramètres statistiques de position (moyenne, ...) ou de dispersion (écart type, ...) issus de la statistique descriptive. De la même manière que l'on peut déterminer des paramètres statistiques d'une population par son recensement exhaustif. Mais on peut aussi déduire des propriétés de la population à partir de celles de l'échantillon par inférence statistique. D'après la loi des grands nombres, Plus la taille de l'échantillon augmente, plus ses propriétés seront proches de celle de la population. En particulier, on peut estimer une probabilité sur les individus d'une population par la fréquence observée sur un échantillon si sa taille est suffisamment grande. Par ailleurs, une moyenne ou une proportion dans une population peuvent être estimés par un intervalle de confiance. Cette démarche est employée dans le cadre des sondages d'opinions ou le contrôle statistique de la qualité.

3.2 Les principes fondamentaux de la théorie de l'échantillonnage

3.2.1 Le principe de validité

Selon ce principe, l'échantillonnage doit permettre d'obtenir une estimation valide des paramètres de la population. L'estimation des paramètres sera valide pour autant que soit connue au préalable la probabilité de chaque in-

dividu de faire partie de l'échantillon. Un paramètre est une caractéristique mesurée de la population, p.ex. La langue parlée au foyer, l'état civil, la scolarité, et ainsi de suite.

3.2.2 Le principe de constante statistique

La nécessité et l'importance de choisir un individu comme membre de l'échantillon doivent découler de la théorie des probabilités, et non d'autres considérations. Ainsi, on prétend qu'un groupe assez nombreux d'individus choisis au hasard comportera les mêmes caractéristiques que celles de la population d'où il origine.

3.2.3 Le principe d'optimisation

L'optimisation des résultats consiste à obtenir le maximum d'efficacité avec le minimum de ressources et de dépenses. Sans jamais perdre de vue la réduction des erreurs d'échantillonnage, il faut réduire les coûts et éviter les dépenses inutiles d'énergie pour former un échantillon représentatif.

Ces principes n'ont de valeur que si tout est mis en œuvre pour qu'ils soient respectés. Ils sont bien plus que des vœux pieux. La technique d'échantillonnage que vous utilisez doit respecter la loi des probabilités coûte que coûte si vous voulez être capables de généraliser les résultats obtenus auprès de votre échantillon à la population générale. Sinon, votre recherche ou votre enquête aura servi à décrire un groupe quelconque de personnes et non un échantillon.

3.3 Statistiques descriptives

Une fois les données collectées, il convient de synthétiser et de résumer l'information contenue dans ces données. On utilise pour cela des représentations des données sous forme de tableaux, de graphiques ou d'indicateurs numériques (tels que la moyenne, la variance, la corrélation linéaire, . . . pour des variables quantitatives).

Cette phase est connue sous le nom de statistique descriptive. On parle de statistique descriptive univariée lorsque l'on regarde une seule variable, de statistique descriptive bivariée lorsque l'on regarde simultanément deux variables, et de statistique descriptive multidimensionnelle lorsque l'on regarde

simultanément p variables.

Dans le but est la description de données exhaustives à l'aide de moyens appropriés, qu'ils correspondent à des valeurs calculées (moyenne, médiane, écart-type, quartile, ...) ou à des représentations graphiques (tableaux, histogramme, camembert, résumés numériques, ...). Il s'agit de méthodes exploratoires et descriptives.

Donc on résume que la statistique descriptive consiste de trois principaux étape :

- Collecter les données (synthétiser et résumer l'information contenue dans les données).
- Organiser les données (mise en forme les données sous forme de tableau, de graphique ou de carte).
- Traiter les données (la description de données exhaustives à l'aide de moyens appropriés pour déceler les liens qui existent entre eux).

3.4 La statistique inférentielle

L'inférence statistique est définie comme le processus d'utilisation des données d'un échantillon pour estimer ou tester des hypothèses sur les caractéristiques numériques (« paramètres ») d'une population. Lorsqu'une moyenne, une fréquence ou un écart-type est mesuré sur un échantillon, on définit un intervalle de confiance probabilisé dans lequel se situent la moyenne, la proportion ou l'écart-type de la population. Son but est d'étendre (d'inférer) les propriétés constatées sur l'échantillon (grâce l'analyse exploratoire par exemple) à la population toute entière, et de valider ou d'infirmer des hypothèses.

3.4.1 Objectif de l'inférence statistique

L'objectif de l'inférence statistique est de tester la généralisable des conclusions de l'analyse statistique descriptive pour trois objectifs statistiques :

1. Comparaison d'un groupe d'observation à une distribution connue.
2. Comparaison de deux groupes d'observations.
3. Evaluation de la liaison entre deux variables.

3.4.2 Les hypothèses de la statistique inférentielle

- La population est considérée comme infinie (très grande).
- Les variables statistiques qui la décrivent peuvent être considérées comme des v.a. (La valeur prise par la variable statistique X pour un individu donné de la population ne peut pas être déterminée a priori et dépend d'un grand nombre de paramètres : On peut considérer sa valeur comme fonction du résultat d'une expérience aléatoire.)
- La répartition des valeurs de ces variables sont caractérisées par des lois de probabilités. (Caractérisée par une densité de probabilité (X continue) ou une séquence de fréquences relatives à chacune de ses valeurs (X discrète), Possédant des caractéristiques ($E(X)$, $V(X)$, autres paramètres résumant la distribution.)
- Les variations simultanées de deux ou plusieurs variables statistiques sont décrites par une loi jointe (caractérisée par une densité jointe (variables continues) ou une séquence des fréquences jointes (variables discrètes). Possédant différentes caractéristiques, par exemple un vecteur espérance, une matrice de variance covariance, un coefficient de corrélation linéaire.)

3.4.3 La taille d'échantillonnage

Cela dépend de plusieurs facteurs dont celui de :

- Le plan de sondage.
- La précision souhaitée et l'intervalle de confiance.
- Le coût. . .

Détermination de la taille d'échantillonnage

La formule donnant la taille minimale de l'échantillon est :

$$n = \frac{t^2 p(1-p)}{e^2}$$

Rappel des notations :

- n est la taille minimale de l'échantillon.
- t coefficient de marge traduit du tableau de confiance ($t = 1,96$ à 95 % de niveau de confiance $t = 2,58$ à 99 % de niveau de confiance).
- p Proportion de la population qui représente une caractéristique donnée. (Si p est inconnue, on utilise $p = 0.5$).

- e est la marge d'erreur choisie pour l'estimation (par exemple, si $e = 0.01$, alors cela signifie que l'on veut connaître la proportion réelle à 1% près).

Exemple 3.4.1 *Présumons que $p = 0,50$ et que $q = 0,50$ et que l'on considère une très grande population ; calculer la taille de l'échantillon à un niveau de confiance de 95% et un intervalle de confiance à $e = 0,05$:*

$$n = \frac{t^2 p(1-p)}{e^2} = \frac{1.96^2 \times 0.5 \times 0.5}{0.05^2} = 384.16$$

Si le niveau de confiance est à 99 %, alors « n » devient :

$$n = \frac{t^2 p(1-p)}{e^2} = \frac{2.58^2 \times 0.5 \times 0.5}{0.05^2} = 665.64$$

À un niveau de confiance de 99 %, la taille de l'échantillon requise est beaucoup plus élevée qu'à 95 %.

3.4.4 La distribution d'échantillonnage

Les principales distributions d'échantillonnage sont la distribution d'échantillonnage de la moyenne, de la variance, et de la proportion.

Une distribution d'échantillonnage d'une statistique (moyenne d'échantillon, variance d'échantillon) est l'ensemble des valeurs prises par cette statistique sur chaque échantillon issu d'une même population. Dans ce chapitre, on étudie comment se comportent le pourcentage de l'échantillon et la moyenne de l'échantillon lorsqu'on répète un grand nombre de fois l'expérience (l'échantillon).

La distribution d'échantillonnage est l'étude de la de probabilité de l'échantillon en fonction de la distribution de la variable parente lorsque la taille de l'échantillon augmente.

Pour résoudre les problèmes d'estimation de paramètres inconnus, il faut tout d'abord étudier les distributions d'échantillonnage, c'est à dire la loi de probabilité suivie par l'estimateur.

Considérons tous les échantillons possibles de taille n extraits d'une population de taille N , de moyenne μ , de variance σ^2 , ...etc. Pour chaque échantillon, il est possible de calculer les paramètres statistiques $\bar{x}$, s , s^2 ...ect qui varient d'un échantillon à l'autre. Chaque paramètre possédera ainsi une distribution d'échantillonnage au même titre que la variable aléatoire X .

3.4. LA STATISTIQUE INFÉRENTIELLE

On utilise souvent :

	Population	Echantillon	Distribution des moyennes	Distribution des variances
Taille	N	n	/	/
Moyenne	μ	$\bar{x}$	$\mu_{\bar{x}}$	μ_{s^2}
Ecart-type	σ	s	$\sigma_{\bar{x}}$	σ_{s^2}
variance	σ^2	s^2	$\sigma_{\bar{x}}^2$	$\sigma_{s^2}^2$
Proportion	π	p	/	/

Distribution d'échantillonnage des moyennes

Soit une population de taille N , on désigne par μ et σ la moyenne et l'écart-type de cette population respectivement. On extrait de la population une série d'échantillons de taille n , chacun de ces échantillons a une moyenne $\bar{x}$, les différentes moyennes obtenues constituent une distribution d'échantillonnage de moyenne $\bar{X}$ on désigne par $\mu_{\bar{x}}$ et $\sigma_{\bar{x}}$ la moyenne et l'écart-type de la distribution d'échantillonnage de la moyenne.

On a :

$$\mu_{\bar{x}} = \mu.$$

Si le tirage est exhaustif (sans remise) :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}.$$

Dans le cas où la population est infinie ou le tirage est non exhaustif (avec remise) :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Remarque :

- Si n est petit devant N , la distinction entre exhaustivité et non exhaustivité est sans objet car $\frac{N-n}{N-1} \approx 1$.
- Si la taille des échantillons est assez grande (en pratique $n \geq 30$), la distribution d'échantillonnage de la moyenne approche la distribution normale quelle que soit la distribution de la population.
- Si la population est normalement distribuée, la distribution d'échantillonnage de la moyenne est une loi normale quelle que soit la valeur n de la taille des échantillons.

Distribution d'échantillonnage des variances

Chaque échantillon de taille n de la population a une variance $S^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, ces variances sont des valeurs observées d'une même variable aléatoire.

$$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

$$E(S^2) = \frac{n-1}{n} \sigma^2.$$

$$V(S^2) = \frac{n-1}{n^2} [(n-1)\mu^3 - (n-3)\sigma^2].$$

Distribution d'échantillonnage des fréquences

La probabilité de la réalisation d'un événement est supposée être égale à p . on considère les échantillons de taille n extraits, avec remise, d'une population de taille N . à chaque échantillon extrait correspond une fréquence f_n de réalisation de l'évènement considéré.

On a :

- Si le tirage est avec remise :

$$E(f_n) = p.$$

$$V(f_n) = \frac{p(1-p)}{n}.$$

- Si le tirage est sans remise :

$$E(f_n) = p.$$

$$V(f_n) = \frac{p(1-p)}{n} \frac{N-n}{N-1}.$$

Exemple 3.4.2 (la distribution de la moyenne et la variance dans le cas d'un tirage sans remise)

On a une population finie composée de 3 éléments, $P = \{1, 2, 3\}$.

1/- Calculer la moyenne et l'écart-type de cette population.

2/- Quelle est le nombre des échantillons qui peuvent être prélevés à partir de cette population ?

3/- Effectuez le prélèvement de ces échantillons.

3.4. LA STATISTIQUE INFÉRENTIELLE

4/- Etablir une distribution d'échantillonnage des moyennes.

5/- Calculez la moyenne de la distribution d'échantillonnage des moyenne $\mu_{\bar{X}}$.

6/- Calculez l'écart-type de la distribution d'échantillonnage des moyenne $\sigma_{\bar{X}}$.

7/- Peut-on nous confirmer les deux formules suivantes $\mu_{\bar{X}} = \mu$ et $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Réponse

1/- On sait que :

$$\mu = \frac{\sum_{i=1}^n x_i}{n} = \frac{(1 + 2 + 3)}{3} = 2$$

Et on sait aussi que : $\sigma = \sqrt{(\sigma^2)}$

et que :

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 = \frac{1 + 4 + 9}{3} = \frac{2}{3}$$

Donc :

$$\sigma = \sqrt{(\sigma^2)} = \sqrt{\frac{2}{3}}$$

Résultat fondamentaux : $\mu = 2$ et $\sigma = \sqrt{\frac{2}{3}}$.

2/- Parce que le tirage est sans remise donc le nombre des échantillons possible à être prélevés est une combinaison de n éléments pris parmi N éléments c-à-d :

$$C_3^2 = \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3$$

3/- Les échantillons sont les suivants :

$$E = \{(1, 2)(2, 3)(1, 3)\}$$

4/- La distribution d'échantillonnage des moyennes est la suivante :

La distribution d'échantillonnage des moyennes = $\{(1, 5)(2, 5)\}$

5/- La moyenne de la distribution d'échantillonnage des moyennes :

$$\mu_{\bar{X}} = \frac{\sum_{i=1}^n x_i}{n} = \frac{6}{3} = 2 = \mu$$

6/- L'écart-type de la distribution d'échantillonnage des moyennes :

$$\sigma_{\bar{X}}^2 = \frac{\sum_{i=1}^n x_i^2}{n} - \bar{x}^2 = \frac{12.5}{3} - 4 = \frac{12.5 - 12}{3} = \frac{0.5}{3} = \frac{1}{6}$$

$$\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \sqrt{\frac{1}{6}}$$

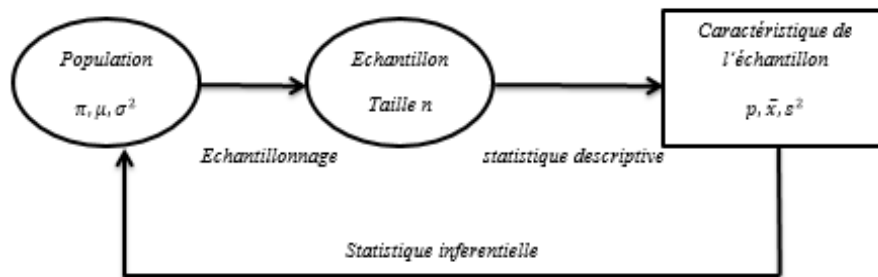
7/- Confirmation :

$$\mu_{\bar{X}} = 2 = \mu$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} = \frac{\sqrt{\frac{2}{3}}}{\sqrt{2}} \sqrt{\frac{1}{2}} = \sqrt{\frac{1}{3}} \sqrt{\frac{1}{2}} = \sqrt{\frac{1}{6}}$$

3.4.5 L'estimation

L'estimation consiste à donner des valeurs approchées aux paramètres d'une population (π, μ, σ^2) où (proportion, moyenne, variance) à partir des données de l'échantillon $(p, \bar{x}, s^2)$. On supposera vérifiée l'hypothèse d'échantillonnage aléatoire simple. La statistique inférentielle peut se résumer par le schéma suivant :



Estimations par intervalle de confiance

Dans le cadre de l'estimation, on associe un nombre, une estimation à un paramètre dont la valeur est inconnue. La précision de cette estimation peut être déterminée en calculant un intervalle de confiance pour ce paramètre,

c'est-à-dire un intervalle contenant la valeur inconnue du paramètre avec une grande probabilité donnée.

3.4.6 Test d'hypothèse

Un test statistique est une procédure de décision qui permet de choisir entre deux hypothèses antagonistes (contraires) faites sur une population ou sur un ou plusieurs paramètres au vu d'un échantillon. Un test statistique est basé sur les données d'une expérience et l'essence d'un test est une règle qui dit au statisticien si les données qu'il a recueillies le conduisent à accepter ou rejeter l'hypothèse.

On distingue deux catégories de tests :

a) **Tests paramétriques** : tester certaine hypothèse relative à un ou plusieurs paramètres d'une variable aléatoire de densité de probabilité connue.

b) **Tests non paramétriques** : utilisés lorsqu'on a à faire à un échantillon dont la loi est complètement inconnue.

Un test statistique permet de trancher entre deux hypothèses :

-Hypothèse nulle notée H_0 : c'est l'hypothèse à tester.

-Hypothèse alternative notée H_1 : hypothèse contraire.

3.4.7 Les erreurs d'échantillonnage dans une enquête

En principe, les erreurs d'échantillonnage diminuent au fur et à mesure que s'accroît la taille de l'échantillon i.e. que plus la taille d'échantillon se rapproche de celle de la population plus les erreurs diminuent. Au-delà d'un certain seuil, la précision d'un estimé ou d'un paramètre s'accroît peu. Généralement, les erreurs d'échantillonnage diminuent de façon inversement proportionnelle à la taille de l'échantillon :

$$E \approx \frac{1}{n}$$

Où :

E = erreurs d'échantillonnage.

n = taille de l'échantillon.

Les types d'erreurs

Pour n'importe quel échantillonnage, des erreurs surviennent. Elles sont inévitables. Par contre, elles peuvent être réduites de plusieurs manières, surtout en augmentant la taille de l'échantillon. Plus on se rapproche de la population, plus les erreurs d'échantillonnage diminuent.

Deux principaux types d'erreurs peuvent être mentionnés :

Les erreurs systématiques : est attribuable aux fluctuations dans l'estimation des paramètres. Ces erreurs sont mesurables par l'erreur-type i.e la différence entre les valeurs ou statistiques d'un échantillon et les paramètres de la population.

Les erreurs dues au hasard : constitue des biais, cette fois-ci attribuables à plusieurs facteurs. Par exemple, ces biais peuvent survenir au moment de la collecte des données et/ou à cause de l'instrument de mesure et/ou à l'étape de la compilation des données et/ou dans le codage et/ou le traitement des données.

3.5 Application dans R

A l'aide de la fonction "rnorm()" commencez par créer la population "THpop2012" de 24,4 millions d'homme adulte observée en France en 2012 et dont la taille (cm) suit une loi Normale de moyenne 175 cm et d'écart type 6 cm :

```
> THpop2012 <- rnorm(24400000, m=175, sd=6)
> str(THpop2012)
 num [1:24400000] 170 181 175 178 184 ...
```

Créez une matrice nulle de taille 1000×10 à l'aide de la fonction "matrix()" :

```
> MILLE <- matrix(data=0, ncol=10, nrow=1000)
> str(MILLE)
 num [1:1000, 1:10] 0 0 0 0 0 0 0 0 0 0 ...
```

Remplissez ensuite votre matrice des 1000 échantillons de 10 individus issus de la population "THpop2012" à l'aide d'une boucle for et de la fonction "sample()" :

```
> for (i in 1:1000)
+ {
+ MILLE[i, ] <-sample(THpop2012, size=10, replace=TRUE)
+ }
> str(MILLE)
  num [1:1000, 1:10] 170 172 185 168 179 ...
```

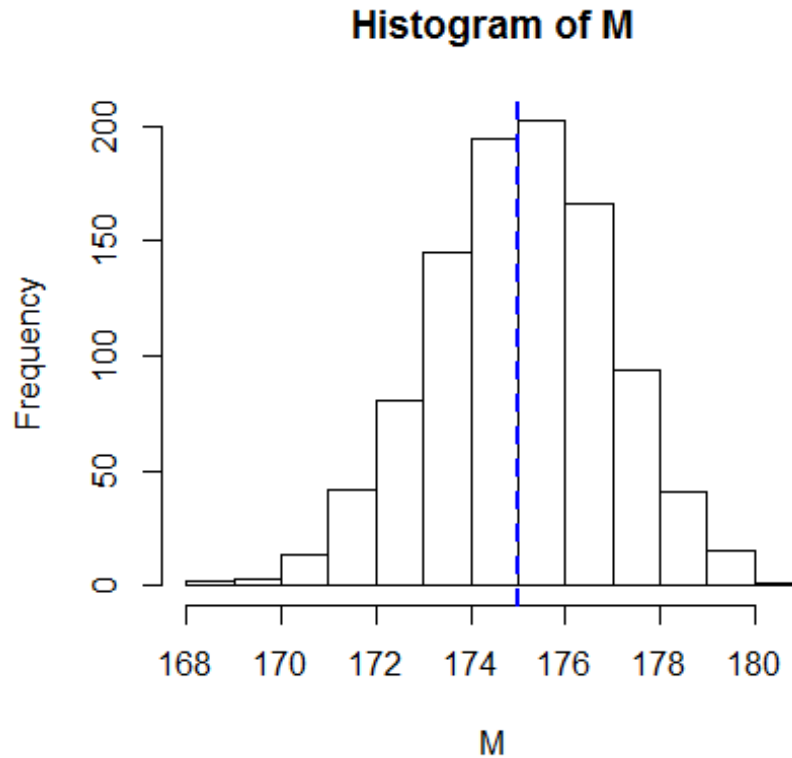
A l'aide de la fonction "apply()", calculez les 1000 moyennes des 1000 échantillons et stockez le résultat dans un objet de classe "numeric" :

```
> M <-apply(MILLE, 1, mean)
> str(M)
  num [1:1000] 173 173 177 176 175 ...
> mean(M)
[1] 175.0457
```

A l'aide de la fonction "hist()", tracez l'histogramme de la distribution des 1000 moyennes :

```
> hist(M)
> abline(v=175, col="blue", lwd=2, lty=2)
```

A l'aide de la fonction "abline()", ajouter une droite verticale de couleur bleue représentant la position de la moyenne de la population :



La moyenne de l'échantillon est un estimateur non biaisé de la moyenne de la population.

A partir de la formule de la variance utilisée en statistique descriptive, calculez les 1000 variances des 1000 échantillons et stockez le résultat dans un objet de classe "numeric" :

```
> V1 <-apply(MILLE, 1, function(x) sum((x-mean(x))^2)/10)
> str(V1)
 num [1:1000] 36.4 26.5 21.5 39.5 21.4 ...
> mean(V1)
[1] 31.72029
```

Utilisez maintenant la fonction "var()" de R (cf. statistique inférentielle) pour calculer les 1000 variances des 1000 échantillons et stockez le résultat dans un nouvel objet de classe "numeric" :

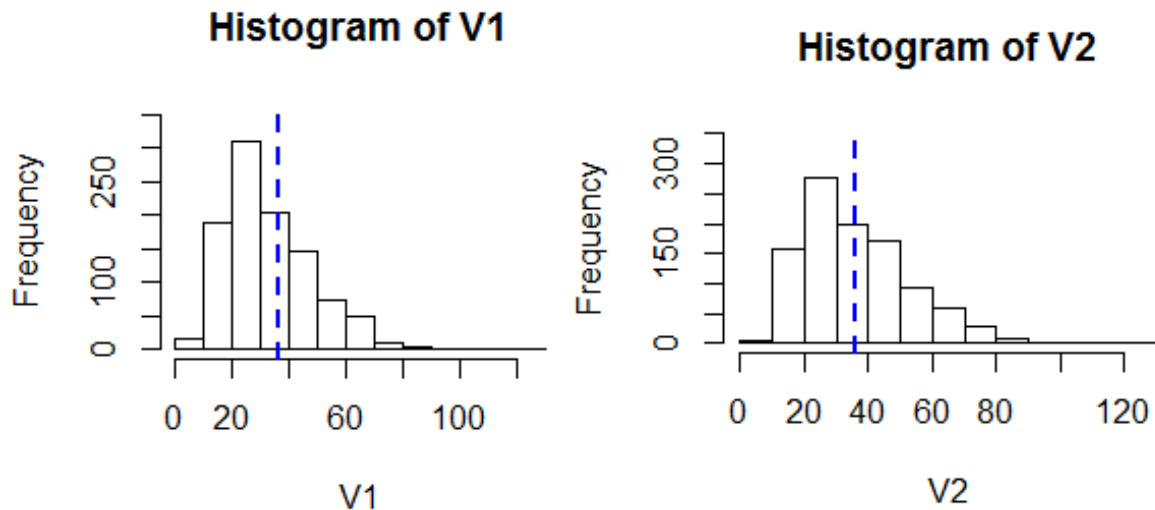
```
> V2 <-apply(MILLE, 1, var)
> str(V2)
 num [1:1000] 40.4 29.5 23.8 43.9 23.7 ...
> mean(V2)
[1] 35.24477
```

Tracez l'histogramme de la distribution des 1000 variances calculées à partir de la formule utilisée en statistique descriptive et positionnez la variance de la population dans la distribution :

```
> hist(V1, breaks=c(seq(0, 130, 10)), ylim=c(0, 350))
> abline(v=36, col="blue", lwd=2, lty=2)
```

Tracez l'histogramme de la distribution des 1000 variances issues de la fonction "var()" de R (cf. statistique inférentielle) et positionner la variance de la population dans la distribution :

```
> hist(V2, breaks=c(seq(0, 130, 10)), ylim=c(0, 350))
> abline(v=36, col="blue", lwd=2, lty=2)
```



A l'aide de la fonction "sd()", calculez l'écart type des 1000 moyennes stockées dans l'objet "M" :

```
> sd(M)
[1] 1.898256
```

Sachant que l'écart type de la population est de 6 cm et que la taille de l'échantillon est de 10, calculez l'erreur standard à partir de la formule :

```
> 6/sqrt(10)
[1] 1.897367
> sd(THpop2012)/sqrt(10)
[1] 1.897261
```

Calculez l'erreur standard de la moyenne à partir d'un échantillon :

```
> sd(THech2012)/sqrt(10)
[1] 1.792072
```

Deuxième partie

**Les méthodes
d'échantillonnages auxiliaires**

Pour que les résultats d'une enquête par sondage puissent être extrapolés à l'ensemble de la Population faisant l'objet de l'étude, il est indispensable que cette enquête soit conduite selon Des règles bien définies et que les calculs conduisant à ces extrapolations soient conformes à la procédure d'échantillonnage utilisée. L'échantillon choisi doit être le plus représentatif possible de la population étudiée, c'est à dire le degré de correspondance entre l'information recueillie et ce que nous apprendrait un recensement comparable de la population dépend en grande partie de la façon dont l'échantillon a été choisi. La théorie moderne de l'échantillonnage nous propose une distinction fondamentale entre échantillons basés sur la probabilité : échantillons probabilistes ; et échantillons non basés sur la probabilité : échantillons non probabilistes ou empiriques. La différence entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une « chance » d'être sélectionnée et que cette chance peut être quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste ; dans ce cas, chaque unité incluse à l'intérieur d'une population n'a pas une chance égale d'être sélectionnée. La section suivante décrit les caractéristiques des deux types d'échantillonnages et fournit des détails sur certaines des méthodes reliées à chaque type.

Chapitre 4

Méthodes d'échantillonnages probabilistes avec des exemples

4.1 Introduction

L'échantillonnage probabiliste est la méthode définie comme étant la plus complexe et celle qui demande le plus de temps. De plus, elle coûte habituellement plus cher que l'échantillonnage non probabiliste (non aléatoire). Pour ce type d'échantillonnage, la sélection des unités de la population se réalise de manière aléatoire (au hasard). Ainsi, toutes les unités de la population ont une chance égale d'être sélectionnées pour constituer l'échantillon. Ce dernier a donc de meilleures chances d'être représentatif de la population, car la probabilité de sélection de chaque unité peut être calculée, des estimations fiables sont possibles ainsi que des estimations d'erreur d'échantillonnage. Il est reconnu que les déductions réalisées à partir de données d'un échantillon qui est représentatif de la population sont les plus fiables. Voici les méthodes d'échantillonnage probabiliste les plus courantes :

- L'échantillonnage aléatoire simple.
- L'échantillonnage systématique.
- L'échantillonnage stratifié.
- L'échantillonnage par grappes.
- L'échantillonnage à plusieurs degrés.

4.2 Les différentes méthodes d'échantillonnages probabilistes

4.2.1 Échantillonnage aléatoire simple

Définition 4.2.1 *Un échantillonnage est aléatoire si tous les individus de la population ont la même chance de faire partie de l'échantillon ; il est simple si les prélèvements des individus sont réalisés indépendamment les uns des autres.*

Principe

Le principe de l'ESA est que chaque membre d'une population a une chance égale d'être inclus à l'intérieur de l'échantillon. Chaque combinaison de membres de la population a aussi une chance égale de composer l'échantillon.

Étapes

1. On constitue la base de sondage qui correspond à la liste complète et sans répétition des éléments de la population.

2. On se procure une liste de toutes les unités statistiques de la population et on les numérote de 1 à N .

3. On choisit au hasard « n » nombres différents correspondant aux « n » unités statistiques qui devront faire partie de l'échantillon. Ce choix peut se faire avec ou sans remise :

- Avec remise : Un individu peut être choisi plusieurs fois.
- Sans remise : Un individu déjà choisi ne peut l'être de nouveau.

C'est le cas habituel.

4. Procéder, à la sélection des éléments pour constituer l'échantillon à l'aide de : **la table des nombres aléatoires**.

(Il existe trois techniques : le tirage de jetons numérotés ou de papiers avec le nom de chaque individu dans une urne, la désignation de façon aléatoire de l'échantillon à l'aide d'un ordinateur et la sélection de nombre identifiant les individus à l'aide d'une table de nombres aléatoires)

5. On utilise la table des nombres aléatoires en suivant 3 étapes :

1. On choisit aléatoirement un nombre (dans la table) compris entre 1 et N .

CHAPITRE 4. MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES AVEC DES EXEMPLES

2. On choisit notre chemin de lecture.
3. On suit notre chemin en choisissant, seulement, les nombres compris entre 1 et N .

Exemple 4.2.1 *Il faudrait numérotter dans un ordre séquentiel chaque entrée ou inscription pour prélever un échantillon aléatoire simple d'un annuaire téléphonique. S'il y avait 10000 entrées dans l'annuaire téléphonique et si la taille de l'échantillon était 2000 numéros, un ordinateur devrait alors générer au hasard 2000 numéros entre 1 et 10000. Chaque numéro aurait la même chance qu'un autre d'être généré par l'ordinateur (ce qui respecterait l'exigence de l'échantillonnage aléatoire simple : une chance égale pour chaque unité). Les 7 entrées dans l'annuaire téléphonique correspondant aux 2000 numéros aléatoires générés par l'ordinateur composeraient l'échantillon.*

Un échantillonnage aléatoire simple peut s'effectuer avec ou sans remplacement. Un échantillon avec remplacement signifie qu'il est possible que l'entrée dans l'annuaire téléphonique échantillonnée soit sélectionnée deux fois ou plus. Habituellement, l'échantillonnage aléatoire simple est effectué sans remplacement, parce qu'il est plus pratique et donne des résultats plus précis. Nous ferons référence à l'échantillonnage sans remplacement lorsque nous traiterons de l'échantillonnage aléatoire simple aux fins des présentes descriptions.

Exemple 4.2.2 *Nous prendrons un extrait des tables de Kendall et Babington Smith.*

TABLE DE KENDALL ET BABINGTON SMITH

	5	10	15	20	25	30	35	40	45	50
1	13407	62899	78937	90525	25033	56358	78902	47008	72488	72488
2	50230	63237	94083	93634	71652	02656	57532	60307	91619	48916
3	84980	62458	09703	78397	66179	46982	67619	39254	90763	74056
4	22116	33646	17545	31321	65772	86506	09811	82848	92211	51178
5	68645	15068	56898	84021	40115	27524	42221	88293	67592	06430
6	26518	39122	96561	56004	50260	68648	85596	83979	09041	62350
7	36493	41666	27871	71329	69212	57932	65281	57233	07732	58439
8	77402	12994	59892	85581	70823	53338	34405	67080	16568	00854
9	83679	97154	40341	84741	08967	73287	94952	59008	95774	44927
10	71802	39356	02981	89107	79788	51330	37129	31898	34011	43304

4.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES

On veut prélever, de façon aléatoire, 8 individus dans une population de 850 individus.

Il s'agit donc ici de 8 tirages successifs au hasard sans remise.

La première opération consiste à numéroté les 850 individus de 1 à 850.

Ensuite, on prend une table de nombres au hasard et on décide de la façon de l'utiliser.

Première méthode

- *On part de la ligne 10 et de la colonne 6 (point d'entrée dans la table).*
- *On va de la gauche vers la droite.*
- *On prend les chiffres par "paquets" de 3.*

Dans la table, on lit les nombres suivants : 393, 560, 298, 189, 107, 797, 885, 133, 037, 129, 318, 983, 401...

On ne retient dans cette liste que les 8 premiers nombres inférieurs ou égaux à 850.

Les 8 individus constituant l'échantillon sont donc les individus respectivement numérotés : 393, 560, 298, 189, 107, 797, 133 et 037.

En effet, le nombre 885 n'est pas retenu, car il ne correspond à aucun individu.

Deuxième méthode

- *On part de la ligne 1 et de la colonne 31.*
- *On parcourt la table de haut en bas et de la gauche vers la droite.*
- *On prélève les nombres par "paquets" de 5 (la table s'y prête).*
- *On ne conserve dans chaque "paquet" que les trois premiers chiffres.*

Dans la table, on lit :

78902	57532	67619	09811	42221	85596	65281	34405	94952	37129
-------	-------	-------	-------	-------	-------	-------	-------	-------	-------

Dans ce cas, les individus prélevés sont les individus respectivement numérotés : 789, 575, 676, 98, 422, 652, 344, 371 (855 et 949 étant supérieurs à 850 ne sont pas retenus).

Remarque :

Si le numéro 789, par exemple, était sorti une deuxième fois, comme il s'agissait d'un tirage "sans remise", nous n'aurions pas tenu compte du deuxième 789.

Si on avait du prélever 10 individus, les individus prélevés seraient les individus respectivement numéros : 789, 575, 676, 98, 422, 652, 344, 371, 470, 603. Le nombre 470 provient de 47008 et 603 de 60307.

CHAPITRE 4. MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES AVEC DES EXEMPLES

Troisième méthode

On peut envisager des déplacements particuliers, par exemple celui indiqué sur le tableau ci-dessous ; les nombres choisis sont soulignés.

	5	10	15	20	25	30	35	40	45	50
1	13407	62899	78937	90525	25033	56358	78902	47008	72488	72488
2	<u>50230</u>	63237	94083	93634	71652	02656	57532	60307	91619	48916
3	84980	<u>62458</u>	09703	78397	66179	46982	67619	39254	90763	74056
4	22116	33646	17545	31321	65772	86506	09811	82848	92211	51178
5	68645	15068	<u>56898</u>	<u>84021</u>	40115	27524	42221	88293	67592	06430
6	26518	39122	96561	56004	<u>50260</u>	68648	85596	83979	09041	62350
7	36493	41666	27871	71329	69212	<u>57932</u>	65281	57233	07732	58439
8	77402	12994	59892	85581	70823	53338	34405	67080	16568	00854
9	83679	97154	40341	84741	08967	73287	94952	<u>59008</u>	95774	44927
10	71802	39356	02981	89107	79788	51330	37129	31898	<u>34011</u>	43304

Les nombres ainsi prélevés sont : 502, 624, 175, 840, 502, 579, 344, 590, 340.

Le nombre 502 n'est pas retenu une deuxième fois car ...

Dans ce cas, les individus retenus sont les individus respectivement numérotés : 502, 624, 175, 840, 579, 344, 590, 340.

Avantages

- C'est simple.
- C'est une analyse statistique classique.
- L'échantillon représente bien la population.

Inconvénients

- Il faut posséder une bonne base de sondage, c'est-à-dire une liste complète (à jour) de toutes les unités statistique de la population, sans répétition.
- Cette technique suppose de disposer de la liste exhaustive de tous les individus de la population source, sinon, il y aura un biais de sélection. Tous les sujets ont la même probabilité d'être sélectionné, qui correspond à la fraction de sondage : $f = \frac{1}{N}$.

4.2.2 Échantillonnage systématique

Définition 4.2.2 *Sous-ensemble fini de données choisies selon un procédé systématique appliqué à tous les individus d'une population et considéré comme étant représentatif d'une population ou d'un phénomène à étudier.*

Principe

La procédure de tirage systématique consiste à choisir le premier élément de manière aléatoire, les éléments suivants étant ensuite sélectionnés à intervalles réguliers.

Étapes

1. On se procure une liste de toutes les unités statistiques de la population et on les numérote de 1 à N .
2. On détermine la taille « n » de l'échantillon voulu.
3. On calcule r et appelée le pas de sondage (ou la raison de sondage)
 $r = \frac{N}{n}$.
4. On choisit au hasard un nombre entier « d » entre 1 et N . ce sera le numéro de départ.
5. Pour sélectionner les autres, il suffit d'ajouter à d la raison de sondage : les individus choisis seront alors les individus dont les numéros correspondent à : d ; $d + r$; $d + 2r$; $d + 3r$; ; $d + (n - 1)r$.

Selon la raison de sondage et le point de départ choisi, il peut arriver que le nombre obtenu se situe à l'extérieur de la liste. En pareil cas, il faut revenir au début de la liste.

Exemple 4.2.3 *Imaginez que vous devez mener une enquête pour votre collège ou votre université sur les logements pour les étudiants. Dix mille (10000) étudiants sont inscrits dans votre établissement d'enseignement et vous voulez en prélever un échantillon systématique de 500. Pour ce faire, vous devez premièrement déterminer ce que serait votre intervalle d'échantillonnage (r) :*

Population totale $\div$ taille de l'échantillon = intervalle d'échantillonnage.

$$N \div n = r = 10000 \div 500 = 20.$$

CHAPITRE 4. MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES AVEC DES EXEMPLES

Il faudrait attribuer un numéro séquentiel à chaque étudiant pour entreprendre cet échantillonnage systématique. On choisirait le point de départ en sélectionnant un numéro au hasard entre 1 et 20. Si ce numéro était 9, on sélectionnerait alors le 9 étudiant inscrit sur la liste et chaque 20 étudiant par la suite. L'échantillon d'étudiants serait constitué de ceux qui correspondraient aux numéros d'étudiant 9, 29, 49, 69... 9929, 9949, 9969 et 9989.

Exemple 4.2.4 *On veut choisir par échantillonnage systématique 6 étudiants parmi un groupe de 60. La raison de sondage (r) sera 10, car*

$$\frac{N}{n} = 10.$$

(Si l'on devait trouver un échantillon de taille $n = 7$, la raison serait plutôt 9 puisque $60/7 = 8,57$, arrondi à 9).

Si l'origine choisie au hasard est disons, 3, les étudiants inclus dans l'échantillon correspondront aux numéros suivants :

3, c'est-à-dire le point de départ ; 13, c'est-à-dire $3 + 10$; 23, c'est-à-dire $3 + (2 \times 10)$; 33, c'est-à-dire $3 + (3 \times 10)$; 43, c'est-à-dire $3 + (4 \times 10)$; 53, c'est-à-dire $3 + (5 \times 10)$.

Notre échantillon est maintenant complet ; il sera composé des étudiants portant les numéros suivants : 3, 13, 23, 33, 43, 53.

Il faudra reprendre au début lorsque la liste sera épuisée. Si l'on reprenait l'exemple précédent et si l'on devait continuer à prélever des étudiants, on obtiendrait le numéro 63, c'est-à-dire $3 + (6 \times 10)$. Comme ce numéro se situe à l'extérieur de la liste, on reviendrait au début de la liste et on choisirait le numéro 3 (puisque $63 - 60 = 3$).

Avantages

- Bonne répartition de l'échantillon dans la base de sondage.
- Sélectionne d'un seul nombre aléatoire.
- Rapide.
- L'échantillon est facile à sélectionner car un seul nombre est choisi au hasard.
- Cette méthode tend à répartir l'échantillon sur l'ensemble de la liste de la population, ce qui permet d'obtenir une assez bonne précision.

Inconvénients

- Il faut posséder une bonne base de sondage, c'est-à-dire une liste complète (à jour) de toutes les unités statistiques de la population, sans répétition.
- Il faut méfier des effets de périodicité qui peuvent fournir un échantillon biaisé.
- On ne peut obtenir qu'un nombre limité d'échantillons différents à partir de la même liste.

4.2.3 Échantillonnage stratifié

Définition 4.2.3 *Est une méthode plus avancée de sélection des échantillons. Elle est particulièrement utile si les phénomènes étudiés sont répartis d'une manière irrégulière ou si la zone concernée est très hétérogène.*

Principe

La base de sondage hétérogène d'effectif N est divisée en P groupes homogènes (strates), d'effectif N_i de telle sorte que

$$N = N_1 + N_2 + \dots + N_p.$$

Un échantillon, d'effectif n_i , est par la suite, prélevé indépendamment au sein de chacune des strates en appliquant un plan d'échantillonnage au choix de l'utilisateur. Le plus souvent, on procédera par un échantillonnage aléatoire et simple à l'intérieur de chaque strate. Selon un critère lié à la variable à estimer (exemple tranche d'âge, sexe, habitat urbain-rural etc..). Puis un tirage au sort est réalisé dans chaque strate.

Utilisation de strates

Pourquoi doit-on créer des strates ?

Pour bien des raisons, la principale étant que leur création peut rendre la stratégie d'échantillonnage plus efficace. Nous avons mentionné précédemment que vous aviez besoin d'un échantillon plus grand pour obtenir une estimation plus exacte d'une caractéristique qui varie beaucoup d'une unité à l'autre. Si chaque personne incluse dans une population, par exemple, avait le même salaire, il suffirait alors d'un échantillon d'une seule unité pour obtenir une estimation précise du salaire moyen des membres de cette population.

Pour la répartition de l'effectif total, n , de l'échantillon dans les différentes strates, La première solution, dite proportionnelle, consiste à conserver la même fraction d'échantillonnage dans chaque strate. Une seconde solution, dite optimale, tient compte du budget de l'enquête.

Étapes

1. On se procure une liste de toutes les unités statistiques de la population.
2. On sépare la population en différentes strates de manière à ce que chaque strate regroupe les individus de la population possédant une caractéristique commune qui pourront avoir une influence sur les résultats de l'étude, on numérote ensuite les individus dans chaque strate. attention « chaque individu doit faire partie d'une seule strate ».
3. On détermine la proportion de chaque strate dans la population.

$$Proportion = \frac{N_1}{N}, \frac{N_2}{N}, \dots, \frac{N_P}{N}.$$

4. On détermine la taille « n » de l'échantillon voulu.
5. On détermine le nombre d'individus qu'il faudra dans chaque strate de l'échantillon.

La première solution, **dite proportionnelle**, consiste à conserver la même fraction d'échantillonnage dans chaque strate.

$$f = \frac{n}{N} = \frac{n_i}{N_i}.$$

telle que

$$w_i = \frac{N_i}{N}$$

Le nombre d'unités à choisir dans chacune des strates est donc :

$$n_i = w_i \times n = f \times N_i.$$

Une seconde solution, **dite optimale**, tient compte du budget de l'enquête.

G : Budget total de l'enquête.

w_i : Le Poids de la strate.

c_i : Le Coût de la collecte de l'information dans la strate.

σ_i : La dispersion à l'intérieur de la strate mesurée par l'écart type.

Le nombre d'unités à choisir dans chacune des strates est donné par :

4.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES

$$n_i = \frac{k(w_i \times \sigma_i)}{\sqrt{c_i}}.$$

avec :

$$k = \frac{G}{(\sum_{i=1}^P (w_i \times \sigma_i \sqrt{(c_i)}))}.$$

Nombre d'individus = $n_1 + n_2 + \dots + n_p$.

6. On sélectionne le nombre d'individus voulu dans chaque strate de la population par échantillon aléatoire simple.

Exemple 4.2.5 (*Répartition proportionnelle*)

Dans une population de 10000 entreprises, réparties en 5000 petites entreprises, 3000 moyennes entreprises et 2000 grandes entreprises, on souhaite avoir un échantillon de 500 entreprises.

-comment on peut répartir cet échantillon ?

Fraction de sondage constante :

$$f = \frac{n}{N} = \frac{500}{10000} = 0.05$$

Les poids de strates :

$$\begin{aligned} w_1 &= \frac{N_1}{N} = \frac{5000}{10000} = 0.5 \\ w_2 &= \frac{N_2}{N} = \frac{3000}{10000} = 0.3 \\ w_3 &= \frac{N_3}{N} = \frac{2000}{10000} = 0.2 \end{aligned}$$

Strates	Effectif de la strate	Effectif de l'échantillon
Petites entreprises	$N_1 = 5000$	$n_1 = 500 \times 0.5 = 250$
moyennes entreprises	$N_2 = 3000$	$n_2 = 500 \times 0.3 = 150$
Grandes entreprises	$N_3 = 2000$	$n_3 = 500 \times 0.2 = 100$
Total	$N = N_1 + N_2 + N_3 = 10000$	$n = n_1 + n_2 + n_3 = 500$

Exemple 4.2.6 (*Répartition optimale*)

Dans la population des 10000 entreprises, on a pu avoir les informations suivantes :

$$G = 5000dh$$

Strates	Le Poids	Le Coût	l'écart type	L'effectif de l'échantillon
Petites	0.5	50	0.8	26
Moyennes	0.3	75	1.5	24
Grandes	0.2	100	2.2	20

$$\begin{aligned} K &= \frac{5000}{(0,5 \times 0,8 \times \sqrt{50} + 0,3 \times 1,5 \times \sqrt{75} + 0,2 \times 2,2 \times \sqrt{100})} = 449,42 \\ n_1 &= \frac{449,42 \times (0,5 \times 0,8)}{\sqrt{50}} = 26 \text{ petites entreprises.} \\ n_2 &= \frac{449,42 \times (0,3 \times 1,5)}{\sqrt{75}} = 24 \text{ moyennes entreprises.} \\ n_3 &= \frac{449,42 \times (0,2 \times 2,2)}{\sqrt{100}} = 20 \text{ grandes entreprises.} \end{aligned}$$

Avantages

- Le sondage stratifié permet d'être certain d'avoir des individus dans chaque strate.
- Le gain de précision dû à la stratification est d'autant plus grand qu'il existe une liaison plus forte entre le critère de stratification et la variable étudiée.
- L'échantillon obtenu représente bien chacune des caractéristiques de la population.

Inconvénients

- Il faut connaître chacune des caractéristiques de la population.
- Il peut être difficile de rejoindre des individus faisant partie d'une strate peu nombreuse de la population.
- Ce type d'échantillonnage souvent très coûteux.

4.2.4 Échantillonnage par grappes

Définition 4.2.4 *Cette méthode est utile lorsque la population à l'étude comporte une hiérarchie comme les habitants d'une ville qui occupent des maisons; ces derniers forment des îlots, et l'ensemble des îlots compose la ville, Ainsi, un échantillon d'îlots où seront prélevés les sujets ou les objets constitue les groupes d'un échantillonnage par grappes. Les grappes sont des sous-groupes de la population-cible.*

Principe

Pour l'échantillonnage par grappes, il faut diviser la population en grappes, c'est-à-dire en sous-ensembles. Chacun de ces sous-ensembles devant être représentatif de la population mère. Ce type de sondage est utilisé lorsqu'on n'a pas de liste des individus, mais qu'on dispose d'une liste de grappes réunissant plusieurs individus.

Cependant, il faut que l'ensemble des grappes soient relativement semblables, tant par leur taille que par leur composition.

L'échantillonnage par grappes constitue donc à tirer aléatoirement des individus au sein des grappes choisies et mener l'étude sur ces individus.

L'exemple le plus commun de ce type d'échantillonnage est celui d'études menées à l'échelle d'une ville, que l'on divise en quartier. Un échantillonnage aléatoire simple est alors effectué.

Pour obtenir un échantillonnage par grappes ayant les propriétés statistiques aussi précises que possible, il faut :

- Un nombre de grappes non conséquent.
- La taille des grappes uniformes.
- Une homogénéité des individus composant les grappes.

Étapes

1. On sépare la population en grappes hétérogènes de tailles semblables qu'on numérote.
2. On détermine la taille « n » de l'échantillon voulu.
3. On calcule le nombre de grappes qu'il faudra choisir pour constituer l'échantillon.

$$\text{Nombre de grappes} = \frac{n}{\text{moyenne d'individus par grappe}}$$

CHAPITRE 4. MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES AVEC DES EXEMPLES

4. On choisit le nombre de grappes nécessaires par la méthode d'échantillonnage simple.
5. On sélectionne tous les individus des grappes choisies.

Exemple 4.2.7 *Les élèves appartiennent à une classe. Les classes sont regroupées par niveau. Plusieurs niveaux constituent une école. Plusieurs écoles font partie du district scolaire. Une grappe, ce peut être une classe, un niveau, une école.*

L'échantillonnage par grappes offre l'avantage d'avoir accès à peu de frais à un grand nombre d'unités. Le choix des unités d'échantillonnage ne se fait pas à partir des unités, mais en fonction d'une catégorie d'unités. Au lieu d'une liste exhaustive d'élèves du primaire de votre région, vous choisissez quelques classes au hasard. À cause de l'homogénéité des élèves d'une région, l'échantillonnage par grappes est tout aussi efficace que l'échantillonnage aléatoire simple.

<i>Echantillonnage par grappes</i>				
<i>Ecole</i>	<i>XYZ</i>			
<i>Niveau</i>	<i>1</i>		<i>2</i>	
<i>Classes</i>	<i>A</i>	<i>B</i>	<i>D</i>	<i>E</i>
<i>Enlevés</i>				

Dans le tableau précédent, les icônes représentent les élèves. Les classes sont assez semblables entre elles, si bien que vous pourriez choisir une classe au hasard. Un niveau complet pourrait tout aussi bien servir de grappes. En choisissant un niveau, toutes les classes appartenant à ce niveau feront alors partie de l'échantillon.

Exemple 4.2.8 *Au moyen de L'échantillonnage par grappes, il s'agit de choisir 12 étudiants; dans un groupe de 60. On demande aux étudiants de se regrouper par 6. On choisit ensuite au hasard deux regroupements, par exemple les grappes numéro 4 et 7 en retenant tous les individus de ces deux grappes, on constitue un; échantillon de 12 étudiants.*

Avantages

- Réduit les déplacements et les couts lorsque la population est répartie sur un grand territoire.
- Les grappes sont plus hétérogènes en leur sein pour la variable étudiée (**variance intra-grappe élevée**).

4.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES PROBABILISTES

- Les grappes sont plus semblables entre elles (**variance inter-grappe faible**).
- La méthode ne nécessite pas une liste globale de la population, puisque seuls les individus inclus dans les grappes comptent.
- Limiter l'échantillon à des groupes proches géographiquement, ce qui permet de réduire les frais de déplacement et de supervision.

Inconvénients

- Si les grappes sont constituées d'individus statistiques homogènes, l'échantillon produit par cette méthode ne représentera pas bien la population.
- L'échantillonnage par grappes peut entraîner des résultats imprécis.

4.2.5 Échantillonnage à plusieurs degrés

Définition 4.2.5 *Méthode qui ressemble à la méthode d'échantillonnage en grappes, sauf qu'il faut dans ce cas prélever un échantillon à l'intérieur de chaque grappe sélectionnée, plutôt que d'inclure toutes les unités dans la grappe. Ce type d'échantillonnage exige au moins deux degrés.*

Principe

En effet, la méthode part de la présomption qu'il y a une hiérarchie d'espaces emboîtés où les unités spatiales du premier degré peuvent être subdivisées en unités du deuxième degré et ainsi de suite. Par exemple, une ville est composée d'arrondissements, les arrondissements de quartiers, les quartiers d'îlots, etc. Pour chaque degré d'échantillonnage, il existe une base de sondage correspondante. En d'autres mots, un découpage géographique correspondant à chaque échelle géographique et chaque unité spatiale à l'échelle du premier degré se subdivise en unités du deuxième degré. À chaque niveau, un échantillonnage du 1^{er} du 2^{ème} ou 3^{ème} degré (2^{ème} ou 3^{ème} degré sont des sous-échantillonnages).

Les unités d'échantillonnage du premier degré sont appelées unités primaires d'échantillonnage, et les unités d'échantillonnage du deuxième degré, unités secondaires d'échantillonnage. Ils ne sont pas obligatoirement de la même taille.

Étapes

1. À partir de la base de sondage des unités du premier degré (des grappes), on détermine un nombre déterminé d'unités du premier degré (des grappes). Ces unités renferment plus d'unités de la population qu'il n'en faut pour l'échantillon final.

2. Puis, on sélectionne un nombre déterminé d'unités au deuxième degré à partir des grappes sélectionnées (à l'aide de l'une des méthodes d'échantillonnage probabiliste possibles).

3. Si l'on utilise plus de deux degrés, le processus de sélection d'unités de la population à l'intérieur des grappes continue et s'arrête lorsque les sujets ou les objets de la population-cible ont été sélectionnés aléatoirement.

Exemple 4.2.9 *Pour étudier les ravages d'animaux sauvages dans un parc d'une superficie environ 13000 km² a été divisé en 500 parcelle-échantillon de 26 km² et 20 d'entre elle ont été sélectionnées au hasard. La superficie des différents peuplements végétaux a été mesurée pour caractériser le couvert végétal. Chaque parcelle échantillon a été subdivisée en 59 placettes de 0.44 km² et 6 d'entre elle ont été sélectionnées de façon aléatoire au sein de chacune des 20 parcelles-échantillons.*

Deux plans d'échantillonnage, l'un du 1er degré et l'autre du second degré, ont été appliqués.

*$n = 20$ parcelles-échantillon (**unités primaires**) parmi $N = 500$ pour caractériser le parc.*

*$m = 6$ placettes (**unités secondaires**) sur $M = 59$ ont été mesurées tertiaires.*

Avantages

- L'échantillonnage par degrés est relativement peu contraignant, car il n'exige pas la liste complète des éléments de la population.

- Ce plan est économique, car il diminue généralement les frais directs et indirects de déplacements.

- Ce plan est flexible, car il permet un ajustement du protocole de sélection des unités d'échantillonnages aux conditions particulières de l'étude.

Inconvénients

- L'échantillonnage par degrés est moins efficace que l'échantillonnage aléatoire simple et systématique lorsque le coefficient de corrélation intra-grappe s'avère très élevé, c'est-à-dire, lorsque les grappes se composent d'éléments qui se ressemblent beaucoup.
- Cette perte d'efficacité qui apparaît notamment dans les populations fortement auto corrélées, doit être compensée par l'économie de temps ou d'argent générée par le rapprochement spatiotemporel des éléments examinés.

4.3 Les avantages des méthodes d'échantillonnage probabiliste

- L'échantillon a de meilleures chances d'être représentatif de la population et les déductions réalisées à partir de données d'un échantillon qui est représentatif de la population sont les plus fiables.
- Calcul possible de la marge d'erreur.
- Un échantillon relativement petit peut être utilisé pour formuler des déductions sur une grande population.

4.4 Les inconvénients des méthodes d'échantillonnage probabiliste

- La collecte des données est plus complexe.
- Méthode plus coûteuse.
- Nécessite une base de sondage.

Chapitre 5

Méthodes d'échantillonnages non probabilistes avec des exemples

5.1 Introduction

Dans celui de l'échantillonnage non probabiliste, on suppose que la distribution des caractéristiques à l'intérieur de la population est égale. C'est ce qui fait que le chercheur croit que n'importe quel échantillon serait représentatif et que les résultats, par conséquent, seront exacts.

L'échantillonnage non probabiliste repose sur un choix arbitraire des unités, Elles regroupent les techniques où les éléments de l'échantillon sont choisis sur le terrain en fonction de jugements sur les caractéristiques de la population ou bien en reconstituant le hasard. Elles ne nécessitent pas de base de sondage.

Afin de faire des déductions sur la population à partir d'un échantillon non probabiliste, l'analyse des données doit supposer que cet échantillon est représentatif de la population. Or, cette supposition est souvent risquée, car la probabilité qu'une personne soit incluse dans l'échantillon ne peut être calculée, à la différence d'un échantillon probabiliste où il est possible d'estimer la marge d'erreur.

Il existe plusieurs méthodes d'échantillonnage non probabiliste, les plus connues sont les suivantes :

- L'échantillonnage de commodité ou de convenance.

- L'échantillonnage de l'aveuglette ou accidentelle.
- L'échantillonnage volontaire.
- L'échantillonnage par quotas.
- L'échantillonnage boule de neige.
- L'échantillonnage au jugé.

5.2 Les différentes méthodes d'échantillonnages non probabilistes

5.2.1 Échantillonnage de commodité « convenance »

Définition 5.2.1 *L'échantillonnage de commodité est une technique d'échantillonnage non probabiliste où les sujets sont choisis en raison de leur accessibilité et de la proximité du chercheur.*

Le principe

Le principe de l'échantillonnage de commodité est prélevé à partir d'un groupe de personnes faciles à contacter ou à atteindre de sujets de l'étude proviennent d'un groupe qui est facilement accessible à un chercheur. Est un échantillon choisi lors d'une étude pour des raisons pratiques d'accessibilité et de coût. Ce type d'échantillonnage est également connu comme l'échantillonnage ou l'échantillonnage grab de disponibilité. Il n'y a pas d'autres critères à la méthode d'échantillonnage, sauf que les gens soient disponibles et prêts à participer. De plus, ce type de méthode d'échantillonnage ne nécessite pas qu'un échantillon aléatoire simple est généré, car le seul critère est de savoir si les participants acceptent de participer.

Étapes

1. L'interviewer qui fait l'échantillonnage présume que la population est homogène.
2. Choisir un échantillon d'individus facilement interrogeables.
3. On ne sélectionne que les individus les plus faciles d'accès. Un enquêteur qui interroge par exemple, ses amis, les membres de sa famille ou encore ses collègues de travail utilise une procédure de commodité.

Exemple 5.2.1 *Une société de jeu qui veut savoir comment l'un de leurs jeux est en train de faire sur le marché un jour après sa sortie. Son analyste peut choisir de créer un sondage en ligne sur Facebook pour noter ce jeu. Le défi majeur de cette approche sera d'atteindre les gens qui jouent à des jeux. Alors que les médias sociaux est un vaste lieu, il est toujours difficile de recueillir des échantillons de la population d'intérêt. La plupart des gens peuvent ne pas être intéressés ou de prendre au sérieux l'enquête tout en complétant, ce qui se traduit par erreur d'échantillonnage. L'enquête peut être grandement améliorée si les postes d'analystes il a des pages de fans dédiés au jeu amoureux. Il peut trouver beaucoup plus de personnes dans ce groupe qui serait enclin à juger et évaluer de façon critique le jeu.*

Avantages

- Elle est facile d'accès, exigeant peu d'effort de la part du chercheur.
- Rapide, peu coûteuse et simple et les sujets sont facilement disponibles.
- Une technique simple et peu coûteuse.
- L'absence d'influence de l'investigateur.
- Donner des résultats exacts lorsque la population est homogène.
- Cette méthode est demande peu de planification.

Inconvénients

- N'est pas une représentation exacte de la population.
- Les résultats d'une étude menée avec un tel échantillon ne peuvent être généralisés à la population dans son ensemble.
- Le non représentativité.
- L'impossibilité d'évaluer le biais associé.

5.2.2 Échantillonnage de l'aveuglette « accidentelle »

Définition 5.2.2 *C'est une Méthode par laquelle on sélectionne les individus qui feront partie de l'échantillon de façon tout à fait arbitraire, Il s'agit d'un échantillon constitué d'individus qui se trouvaient accidentellement à l'endroit et au moment où l'information a été collectée. Cette méthode consiste à sélectionner des unités arbitraires sans utiliser de véritable processus de randomisation.*

Principe

Cette méthode non-probabiliste est tout à fait subjective : l'enquêteur choisit ses répondants sur les lieux de l'enquête et cette sélection se poursuit jusqu'à la taille de l'échantillon prévue. Ainsi, les individus sont-ils choisis sur la base de leur présence en un lieu et un temps donné. Tous les autres sont exclus de l'échantillon. Cette méthode comporte des biais importants car d'une part une partie de la population-cible est exclue si elle ne se trouve pas sur le lieu de l'enquête et, d'autre part, la probabilité de sélection des cas est impossible à calculer. Cette méthode est efficace si la population étudiée est homogène.

Étapes

1. L'interviewer qui fait l'échantillonnage présume que la population est homogène.
2. L'enquêteur choisit ses répondants sur les lieux de l'enquête et cette sélection se poursuit jusqu'à la taille de l'échantillon prévue. Ainsi, les individus sont-ils choisis sur la base de leur présence en un lieu et un temps donné.

Exemple 5.2.2 *Lorsqu'on interroge des individus passant dans la rue, à la porte d'un commerce ou dans manifestation par exemple. On fait un échantillon à l'aveuglette.*

Avantages

- La facilité d'application.
- L'absence d'influence de l'investigateur.

Inconvénients

- Le non représentativité.
- L'impossibilité d'évaluer le biais associé.

5.2.3 Échantillonnage de volontaire

Définition 5.2.3 *L'échantillonnage par volontaire est une des méthodes les plus utilisées. C'est une méthode par laquelle on sélectionne les individus*

CHAPITRE 5. MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES AVEC DES EXEMPLES

qui feront partie de l'échantillon en faisant appel à des volontaires, unités d'échantillonnage offrent volontairement leurs participations à l'étude en question.

Dans lequel l'échantillon est dit auto sélectionné (car les individus se sélectionnent eux-mêmes).

Principe

On prélève l'échantillon à partir d'un groupe de volontaires. Il arrive parfois qu'un chercheur offre de l'argent à des gens pour les inciter à participer à son étude. En échange, les volontaires acceptent la possibilité d'avoir à se prêter à des processus longs, exigeants ou quelques fois désagréables.

Étapes

1. Le chercheur fait Appel à participation par : annonces dans les journaux locaux ou nationaux et/ou d'un dépliant distribué.
2. Les individus se sélectionnent eux-mêmes.
3. Le chercheur recrute au hasard pour y participer des gens du grand public.
4. On prélève l'échantillon à partir d'un groupe de volontaires (Exemples : expériences médicales ou psychologiques). Dans ce cas, l'échantillon est composé de toutes les personnes que l'analyse considère concernées (population cible) et volontaires (acceptent d'y participer) par le sujet de l'étude.

Remarque

Le fait d'échantillonner des participants volontaires plutôt que la population en général peut introduire des biais marqués. Souvent, à l'occasion des sondages d'opinion, seuls les gens qui se soucient assez fortement d'une façon ou d'une autre de la question étudiée ont tendance à y répondre. La majorité silencieuse n'y répond généralement pas, ce qui entraîne un important biais sur le plan de la sélection.

Exemple 5.2.3 *Expéri-pharm*

Nous recherchons présentement des hommes et des femmes âgés de 18 à 55 ans, en bonne santé, pour participer à une étude clinique sur un médicament déjà offert en vente libre.

Une récompense de 20000 DZ sera offerte aux participants.

5.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES

Tenant de telles études les patients sont amenés à consommer certains médicaments puis on observe chez eux les effets secondaires les effets thérapeutiques et cetera on peut même avoir à prendre plusieurs prélèvements sur eux et à tester donc les effets de toutes sorte que le médicament peut avoir sur l'individu bien sûr.

Avantages

- Attractif du point de vue de l'éthique.
- Utile pour les phases exploratoires.
- Cette méthode est très facile à appliquer car l'enquêteur ne se préoccupe pas de convaincre les enquêtés. Il ne s'intéresse qu'à ceux qui ont accepté de répondre.
- Son principal avantage est de ne pas avoir à rechercher au hasard des personnes à interroger, mais de sélectionner sur une liste celles qui répondent aux caractéristiques recherchées et de les contacter.
- Peut être utile lorsque l'étude comporte des inconvénients importants pour les participants.

Inconvénients

- Le non représentativité.
- L'impossibilité d'évaluer le biais associé.
- On ne connaît pas la population de référence.
- Échantillon biaisé car aucune stratégie de sélection n'est suivie pour s'assurer que l'échantillon ainsi constitué est bien représentatif de la population cible.
- Si la base n'est pas actualisée, on risque de ne pas retrouver les enquêtés au cas où ils avaient changé d'adresse ou de numéro de téléphone.

5.2.4 Échantillonnage par quotas

Définition 5.2.4 *La méthode de quotas est une technique d'échantillonnage qui s'impose à chaque fois qu'il faut utiliser un échantillon représentatif alors qu'il n'est pas facile de disposer d'une liste de sondage et que la structure de la population à étudier est connue et reproduire le plus fidèlement la population. Elle repose non plus sur le hasard, au sens statistique du terme, mais sur le raisonnement de l'analyste qui choisit des éléments composant l'échantillon*

CHAPITRE 5. MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES AVEC DES EXEMPLES

en fonction de critères de sélection jugés pertinents en regard du problème à résoudre.

Cette méthode consiste à construire un échantillon représentatif à partir de certains critères clés appelés « variable de quota ou de contrôle ».

Principe

Cette méthode repose sur un principe simple : reproduire le plus fidèlement possible les caractéristiques de la population étudiée grâce à l'application de règles très précises qu'on appelle variables de contrôle.

Étapes

1. On choisit les variables de contrôle.
2. On choisit les individus en fonction de certaines caractéristiques de la population qu'on appelle variables de contrôle.
3. Ces groupes sont constitués en fonction du lieu géographique, de l'âge, du sexe, du revenu, etc.
4. Préparer un plan d'échantillonnage au sein duquel le nombre de cas est prévu, conforme ou correspondant aux variables de contrôle.
5. Puis, chaque enquêteur est libre d'interviewer les personnes à la condition de répondre aux quotas fixés. Il faut noter qu'on ne doit retenir qu'un nombre restreint de quotas (au maximum *trois variables* de quotas, Au-delà de 3 *critères*, la démarche devient complexe.).
6. Une fois les quotas sont fixés, les individus sont sélectionnés à la convenance de l'enquêteur.

Les critères servant de base à la définition des quotas ne doivent pas être nombreux. Les quotas doivent être construits sur une base de données fiable (statistiques disponibles) indiquant la répartition de la population sur les critères choisis.

-La variable de contrôle : c'est une variable qui permet d'assurer la conformité d'un échantillon.

-Le taux de sondage = $\frac{n}{N}$.

Exemple 5.2.4 *Sur une population de 15000 entreprises étudiées selon le chiffre d'affaires (variable X) et le nombre d'employés (variable Y). on obtient les résultats suivants :*

5.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES

X (en milliers de Dh)	Effectifs populations N_s
0 – 50	5500
50 – 100	4500
100 – 200	1750
200 – 500	2000
500 – 1000	500
> 1000	750
<i>Total</i>	15000

Y (en milliers de Dh)	Effectifs populations N_s
0 – 10	7050
10 – 50	3700
50 – 100	2500
100 – 500	1500
> 500	250
<i>Total</i>	15000

Soit un taux de sondage de 0.10. Quelle méthode peut on utilise pour déterminer un échantillon ?

On supposera que l'institut responsable du sondage de 30 enquêteurs.

Expliquer votre démarche et les concepts utilisés.

solution

$$\text{Quotas population} = \frac{N_s}{N} \times 100.$$

$$n_i = n \times \text{Quotas population}.$$

$$n_i = N_s \times 10\%.$$

X	N_s	$\frac{N_s}{N}$	n_i	Répartition des enquêteurs
0 – 50	5500	36.67%	550	11
50 – 100	4500	30%	450	9
100 – 200	1750	11.66%	175	3
200 – 500	2000	13.34%	200	4
500 – 1000	500	3.33%	50	1
> 1000	750	5%	75	2
<i>Total</i>	$N = 15000$	100%	$n = 1500$	30

CHAPITRE 5. MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES AVEC DES EXEMPLES

Y	N_s	$\frac{N_s}{N}$	n_i	Répartition des enquêteurs
0 – 50	7050	47%	705	14
50 – 100	3700	24.66%	370	7
100 – 200	2500	16.66%	250	5
200 – 500	1500	10%	150	3
500 – 1000	250	1.68%	25	1
<i>Total</i>	$N = 15000$	100%	$n = 1500$	30

Exemple 5.2.5 *S'il y a dans la population 48% d'hommes et 52% de femmes, 25% de jeunes de 25 à 30 ans, etc. L'enquêteur chargé d'enquêter auprès de cent personnes devra interroger 48 hommes, 52 femmes et 25 personnes de 25 à 30 ans, etc.*

Les quotas retenus peuvent être « simples » c'est à dire avec plusieurs séries de distribution qui seront respectées de façon indépendante ou encore croisée en combinant plusieurs caractéristiques.

Revenons à notre exemple si l'on sait que 19% des femmes de la population mère a entre 25 à 30 ans et 31% des hommes ont de 25 à 30 ans. L'enquêteur devra respecter ces quotas et interroger 10 femmes et 15 hommes ayant de 25 à 30 ans.

*Pour assurer une certaine objectivité aux résultats il faut fixer quelques règles aux enquêteurs comme la gestion rigoureuse des feuilles de quotas, le contrôle strict de la démarche « **aléatoire** » de l'enquêteur (interdiction d'interroger des personnes qui se connaissent entre elles ou qui connaissent l'enquêteur ; l'obligation d'une certaine dispersion géographique avec imposition d'un itinéraire précis à respecter ; ou encore la réalisation des enquêtes à des moments différents.*

Avantages

- Un coût de réalisation plus faible que ceux d'une enquête aléatoire.
 - Des délais de réalisation plus faibles que ceux d'une enquête aléatoire.
- Ce qui justifie son emploi lorsqu'erreurs de mesure et erreurs de non-réponse sont difficilement compressibles comme dans les sondages d'opinion.
- Ne nécessite pas de base de sondage mais uniquement la connaissance de la répartition dans la population selon certaines caractéristiques
 - Il est également facile à administrer.

Inconvénients

- Le non représentativité.
- L'impossibilité d'évaluer le biais associé.
- Non connaissance des taux de réponse.
- Certaines unités peuvent n'avoir aucune chance d'être sélectionnées.
- On ne connaît pas la probabilité d'inclusion des individus de la population, et donc, l'inférence des résultats obtenus à la population doit être faite avec plus de précaution que dans les méthodes aléatoires.
- La représentativité de l'échantillon ainsi constitué peut être sujette à caution.
- Existence de biais que ne peuvent éviter les consignes de recherche complétant généralement les quotas. En particulier, si le nombre de refus qu'a dû subir l'enquêteur est élevé, la représentativité peut être sujette à caution.
- Impossibilité de calculer des marges d'erreur. (On donne généralement comme approximation grossière, faute de mieux, les marges d'erreur du sondage aléatoire simple.)

5.2.5 Échantillonnage par boule de neige

Définition 5.2.5 *Dans ce type d'échantillon, appelé aussi échantillon par réseau, les individus sont sélectionnés en fonction de leurs liens avec un « noyau » d'individus. La méthode consiste à diffuser votre questionnaire d'enquête à des personnes ayant les caractéristiques que vous recherchez puis de leur demander de le diffuser à d'autres personnes de profil similaire.*

C'est ce qu'on appelle l'effet boule de neige c'est à dire un phénomène où des réactions en chaîne produisent une accumulation, à la manière d'une boule de neige roulant le long d'une pente enneigée.

Principe

Cette méthode s'avère utile lorsque le chercheur éprouve des difficultés à trouver les sujets de l'enquête parce qu'ils constituent une population minoritaire ou une catégorie sociale bien précise. Elle est basée sur le principe des contacts personnels ou professionnels permettant de retracer les sujets recherchés. En d'autres mots, il s'agit de constituer un réseau permettant de trouver les répondants au fur et à mesure de la progression de l'enquête.

Étapes

1. On choisissant d'abord arbitrairement un petit groupe d'individus ayant les caractéristiques recherchées pour l'étude.
2. On demande à ces individu sélectionner d'autres personnes de leur entourage qui présentant les mêmes caractéristiques pour qu'ils prennent part à l'échantillon à leur tour.
3. Ces nouvelles personnes devront elles aussi sélectionner d'autres de la même manière et ainsi de suite jusqu'à ce que l'échantillon compte le nombre d'individu voulu.

Exemple 5.2.6 *Vous créez un nouveau produit pour les sportifs et vous voudriez savoir quelles sont les attentes des consommateurs dans ce domaine et quel prix ils seraient prêts à payer pour votre nouveau produit : commencez à envoyer le questionnaire à vos amis sportifs et demandez-leur de le diffuser dans leur propre entourage.*

Petit à petit, le nombre de répondants grossit par effet cumulatif, comme la boule de neige qui descend la pente !

Toutefois, si les caractéristiques recherchées sont très spécifiques et donc rares, ou si les personnes ne jouent pas le jeu de diffuser à leur tour le questionnaire, le risque est d'obtenir trop peu de répondants.

Pour booster le taux de réponse, rien ne vous empêche de mettre en place un invetive pour les répondants, c'est-à-dire une prime ou une récompense.

Exemple 5.2.7 *Imaginons une société quelconque qui décide de s'intéresser aux conditions de vie des immigrants illégaux qui ont tendance à se cacher malheureusement, si on a la chance d'en trouver quelqu'un la meilleur façon d'en réunir plus c'est de leur demander de prendre contact avec des personnes dans la même situation qu'eux et qui accepterait de participer à l'étude, si par exemple on demande à deux personne et chacun en trouve 3 de leur côté on se retrouvera avec un échantillon de 8 personnes, et si ces six nouveaux individus décident aussi de vous aider et en trouve à leur tour quelques-uns on finira avec un échantillon de taille considérable, ce procédé nous permettra même d'atteindre la taille d'échantillon voulu, si comme on l'avait prédit ces 6 personnes trouvent à leur tour 3 autres personnes on se retrouvera avec 18 nouveaux individu ce qui revient à un total de 26.*

Avantages

- Ce type d'échantillon est largement utilisé en milieu industriel.

5.2. LES DIFFÉRENTES MÉTHODES D'ÉCHANTILLONNAGES NON PROBABILISTES

- Peu couteuse.
- Est très facile à mettre.
- Est utile pour atteindre des populations difficiles à rejoindre, mais ne permet pas de déterminer la taille ou pour pénétrer des milieux fermés.

Inconvénients

- Difficulté d'obtention d'un échantillon représentatif si vous n'avez pas pu avoir des recommandations.
- Cette méthode produit presque toujours un échantillon biaisé.

5.2.6 Échantillonnage par jugement

Définition 5.2.6 *L'échantillonnage au jugé est une technique basée uniquement sur les connaissances du chercheur sur la population tel que Le chercheur choisit les unités statistiques qu'il juge possédant les caractéristiques typiques de la population à l'étude.*

On appelle l'échantillon ainsi prélevé échantillon au jugé ou échantillon par choix raisonné.

Principe

Cette technique consiste finalement à se fier au jugement du chercheur pour déterminer la composition d'un échantillon, ce dernier jugeant quelles sont les unités statistiques les plus susceptibles de fournir l'information adéquate (C'est le chercheur qui décidera ce qui est typique).

Pour cette méthode, l'échantillonnage est réalisé sur la base de certains jugements et d'idées préalables sur l'ensemble de la population. Par ses connaissances de cette dernière, le responsable de l'enquête juge que les unités qui constitueront l'échantillon sont représentatives de la population. Des biais importants peuvent découler du fait que les idées préconçues du chercheur, si elles sont inexactes, soient reflétées dans l'échantillon.

Étapes

1. On sélectionne les "unités considérées caractéristiques" du phénomène à l'étude.
2. Choix des éléments basé sur le jugement du chercheur par rapport à leur caractère typique ou atypique

3. L'échantillonnage à l'intérieur de chacune des grandes unités se fait ensuite au hasard.
4. On ne peut alors généraliser qu'aux unités choisies.
5. L'enquêteur sélectionnera des unités qui seront caractéristiques de la population.

Exemple 5.2.8 *Un conseil municipal choisit de tenir son enquête socioéconomique dans un seul quartier de la ville. En considérant le rapport cout/bénéfices, l'enquête n'a pas touché d'autres quartiers.*

Une telle technique risque de produire un échantillon très biaisé, mais elle peut garantir, néanmoins d'éviter que des éléments indésirables se glissent dans un échantillon .

Avantages

- La réduction du coût et du temps qu'exige l'acquisition de l'échantillon est l'un des avantages de l'échantillonnage au jugé.
- Permet d'étudier des phénomènes rares ou inusités ; peu de représentativité de l'ensemble de la population.

Inconvénients

- Le risque de ce genre d'échantillonnage est de considérer des individus apparemment représentatifs de la population étudiée qui ne le sont pas.
- Étant donné que l'échantillonnage au jugé reflète toutes les idées préconçues que risque d'avoir le chercheur, il peut y avoir introduction de biais importants si ces idées sont inexactes.

5.3 Les avantages de méthodes non probabilistes

- Méthode rapide.
- Méthode relativement peu coûteuse.
- Une base de sondage n'est pas nécessaire.
- Les méthodes d'échantillonnage non probabiliste peuvent être utiles lorsqu'on désire des commentaires descriptifs au sujet des échantillons eux-mêmes.

5.4. LES INCONVÉNIENTS DE MÉTHODES NON PROBABILISTES

- Leur utilisation prend peu de temps tout en étant plus économique et plus pratique.
- Contrairement au sondage aléatoire, la sélection des individus à interroger dans un sondage empirique n'exige pas l'application d'un algorithme informatique sur une base de sondage.
- Echantillonnage moins complexe, plus rapide et moins coûteux que l'échantillonnage probabiliste.

5.4 Les inconvénients de méthodes non probabilistes

- On ne peut mesurer la fiabilité d'un échantillonnage non probabiliste.
- Puisqu'on choisit arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon.
- On ne peut estimer la variabilité de l'échantillonnage ni identifier le biais possible.
- L'échantillon a de moins bonnes chances d'être représentatif de la population. Il est peu fiable de généraliser et d'appliquer les résultats de l'échantillon à toute la population (Il faut avoir des hypothèses solides sur la représentativité de l'échantillon pour formuler des déductions sur la population).
- La Mesure de la précision de la procédure d'estimation et calcul rigoureux des marges d'erreurs «impossibles».
- Biais de l'échantillon apporté par la liberté laissée aux enquêteurs dans le choix des personnes à interroger.

Chapitre 6

La différence entre les méthodes probabilistes et non probabilistes

6.1 Introduction

L'échantillonnage consiste à sélectionner un groupe ou un échantillon particulier pour représenter l'ensemble de la population. Les méthodes d'échantillonnage sont principalement divisées en deux catégories : échantillonnage probabiliste et échantillonnage non probabiliste. Dans le premier cas, chaque membre dispose d'une possibilité connue d'appartenir à l'échantillon, alors que dans le second cas, il n'y a pas de probabilité spécifique qu'un individu fasse partie de l'échantillon.

Pour un profane, ces deux concepts sont identiques, mais en réalité, ils sont différents en ce sens que, dans l'échantillonnage probabiliste, chaque membre de la population a une chance de sélection non identique à celle utilisée dans l'échantillonnage non probabiliste. D'autres différences importantes entre l'échantillonnage probabiliste et l'échantillonnage non probabiliste sont rassemblées dans l'article ci-dessous.

6.2 Principales différences entre les échantillons probabilistes et non probabilistes

Les différences significatives entre échantillonnage probabiliste et non probabiliste :

- La technique d'échantillonnage, dans laquelle les sujets de la population ont la même chance de choisir un échantillon représentatif, est connue sous le nom d'échantillonnage probabiliste. Une méthode d'échantillonnage dans laquelle on ne sait pas quel individu de la population sera choisi comme échantillon est appelée échantillonnage non probable.

- La base de l'échantillonnage probabiliste est la randomisation ou le hasard ; on l'appelle aussi échantillonnage aléatoire. Au contraire, la technique de randomisation par échantillonnage non probabiliste n'est pas appliquée pour la sélection d'un échantillon. Par conséquent, il est considéré comme un échantillonnage non aléatoire.

- Dans l'échantillonnage probabiliste, l'échantillonneur choisit le représentant pour faire partie de l'échantillon de manière aléatoire, tandis que dans l'échantillonnage non probabiliste, le sujet est choisi de manière arbitraire pour appartenir à l'échantillon par le chercheur.

- Les chances de sélection dans l'échantillonnage probabiliste sont fixes et connues. Par opposition à l'échantillonnage non probabiliste, la probabilité de sélection est égale à zéro, c'est-à-dire qu'elle n'est ni spécifiée ni connue.

- L'échantillonnage probabiliste est utilisé lorsque la recherche est de nature concluante. D'autre part, lorsque la recherche est exploratoire, il convient d'utiliser un échantillonnage non probable.

- Les résultats générés par échantillonnage probabiliste sont exempts de biais alors que les résultats d'échantillonnage non probabiliste sont plus ou moins biaisés.

- Comme les sujets sont sélectionnés de manière aléatoire par le chercheur dans le cadre d'un échantillonnage probabiliste, la mesure dans laquelle il représente l'ensemble de la population est plus élevée par rapport à l'échantillonnage non probabiliste. C'est pourquoi l'extrapolation des résultats à l'ensemble de la population est possible dans l'échantillonnage probabiliste mais pas dans l'échantillonnage non probabiliste.

- Hypothèse d'un test d'échantillonnage probabiliste mais un échantillonnage non probabiliste le génère.

Tableau de comparaison

Base de comparaison	Echantillonnage probabiliste	Echantillonnage non probabiliste
Alternativement appelé	Echantillonnage aléatoire	Echantillonnage non aléatoire
Base de sélection	Au hasard	Arbitrairement
Possibilité de sélection	Fixe et connu	Non spécifié et inconnu
Recherche	Concluante	Exploratoire
Résultat	Impartial	Biaisé
Méthode	Objectif	Subjectif
Inférences	Statistique	Analytique
Hypothèse	Testé	Généré

6.3 Résumé

L'échantillonnage non probabiliste est privilégié par les Instituts de sondage et de marketing tandis que l'échantillonnage probabiliste est utilisé par les statisticiens professionnels.

Chacune de ces méthodes a ses avantages et ses inconvénients. Bien que l'échantillonnage probabiliste soit basé sur le principe de la randomisation selon lequel chaque entité a la chance de faire partie de l'échantillon, l'échantillonnage non probabiliste repose sur l'hypothèse que les caractéristiques sont réparties de manière égale dans la population, ce qui laisse croire à l'échantillonneur que L'échantillon ainsi sélectionné représenterait l'ensemble de la population et les résultats seraient précis.

Un autre aspect de la complémentarité de ces deux méthodes c'est la combinaison de ces deux méthodes pour la construction d'échantillons complexes.

En conclusion, sur ces techniques d'échantillonnage, il faut retenir que l'utilisation de ces techniques va vous permettre de reproduire fidèlement la population totale que vous étudiez et de ce fait vous garantir la représentativité de votre échantillon. Mais il faut savoir qu'en matière de sondage il n'existe pas de certitude même si l'échantillon a été obtenu par utilisation d'une méthode rigoureuse. Les pourcentages obtenus, résultats du sondage, ont donc toujours une marge d'erreur liée au fait qu'on interroge une fraction de la population.

Troisième partie

Analyse de correspondance multiples et applications

Chapitre 7

Analyse des correspondances multiples (ACM)

7.1 Introduction

L'analyse des correspondances multiple (ACM) permet d'étudier une population de I individus décrits par J variables qualitatives.

Une variable qualitative ou (nominale) est une application de l'ensemble I des individus dans un ensemble fini sur lequel on ne considère aucune structure : par exemple un ensemble de trois couleurs (bleu, blanc, rouge). Les éléments de cet ensemble sont appelés modalités de la variable et l'on dit par exemple qu'un individu bleu possède la modalité bleu.

L'application la plus courante de l'ACM est le traitement de l'ensemble des réponses à une enquête. Chaque question constitue une variable dont les modalités sont les réponses proposées (parmi lesquelles chaque enquêté doit faire un choix unique).

Nous commençons par passer en revue différentes façons de transcrire

7.2 Principe de l'ACM

Le principe consiste à conduire une AFC à partir d'un tableau disjonctif complet (individus ou autres unités statistiques en lignes, variables et modalités en colonnes) ou d'un tableau de Burt. Rappelons que ce dernier est une présentation de tous les tableaux de contingence des variables prises deux à deux et réunis en une seule matrice.

Ainsi, il est possible de déterminer des proximités entre des modalités de variables différentes ou entre individus pour en tirer des enseignements. Par exemple, un profil particulier de clientèle peut émerger à partir de divers critères observés.

Cette technique est une généralisation de l'AFC mais elle peut aussi être considérée comme une extension aux variables nominales de l'ACP sur individus et de l'ACP sur variables. . .

Ajoutons que l'ACM permet aussi d'analyser simultanément des variables qualitatives et numériques. Il suffit pour cela de décomposer ces dernières en classes. De simples graphiques descriptifs peuvent aider à choisir le découpage le plus judicieux. Par ailleurs, il est habituel d'intégrer des variables illustratives, c'est-à-dire non prises en compte dans les calculs mais situées par rapport aux autres dans les résultats.

Pour la procédure, peu de différences par rapport à une AFC dont le principe consiste à comparer le tableau des observations avec un tableau théorique de totale indépendance. Les écarts entre les deux sont mesurés avec la distance du χ^2 . Une distinction entre ces deux techniques est que l'AFC se fonde sur des fréquences marginales alors que l'ACM est soit réalisée à partir d'un tableau disjonctif, soit à partir d'un tableau de Burt.

Rappelons que la distance du χ^2 est la somme des carrés des écarts entre des valeurs observées et théoriques, rapportés aux théoriques. Cette métrique a beaucoup inspiré les concepteurs de l'ACM. . .

7.3 Le problème

Il s'agit d'étudier les liaisons pouvant exister entre les p variables considérées. En fait, dans la mesure où les données se présentent sous forme d'un tableau de Burt, juxtaposition de tables de contingence, seules les liaisons entre variables prises deux à deux sont considérées (il s'agit de ce que l'on appelle en statistique les interactions d'ordre deux). Pour étudier ces liaisons, la démarche sera de même nature qu'en A.F.C.

7.4 Les objectifs

Les objectifs que cette méthode spécifique, l'ACM, doit remplir sont les mêmes que ceux de l'ACP ou de l'AFC. Il s'agit d'obtenir une typologie des

lignes et des colonnes et relier ces deux typologies. Nous aurons ici trois familles d'éléments à étudier, les individus, les variables et les modalités des variables. Cette étude se fait par la définition de ressemblances et liaisons pour ces trois familles que nous détaillons dans la section suivante. An d'établir un bilan des ressemblances entre individus, comme en ACP nous cherchons à répondre à des questions du type :

- Quels sont les individus qui se ressemblent ?
 - Quelles sont ceux qui sont différents ?
 - Existe-t-il des groupes homogènes d'individus ?
 - Est-il possible de mettre en évidence une typologie des individus ?
- Les mêmes types de questions se posent pour les variables et les modalités.

7.5 Aide à l'interprétation

Nous allons nous intéresser principalement aux :

7.5.1 Le choix du nombre des axes retenir

Sélection des axes retenus principalement par rapport aux valeurs propres :

Commençons par les valeurs propres ; pour comprendre comment sont calculés ces axes, il faut se faire une idée, même approximative, de la logique de l'algorithme de l'ACM,

- Le premier axe calculé par l'algorithme de l'ACM est toujours celui qui permet de conserver la plus grande part possible des variations de l'information contenue dans le tableau des variables actives.

- Le deuxième axe est ensuite calculé de la même façon sur ce reste, donc de telle façon qu'il conserve la plus grande part possible des variations de l'information contenue dans ce « reste ».

On traduit en général cette information sous la forme d'un « pourcentage d'inertie » et correspondant donc à la proportion des variations totales des variables originales expliquées par chaque axe.

Le but de la démarche est de rendre compte de la plus grande partie possible de l'information originale, avec le nombre le plus petit possible de « variables ».

On peut utiliser l'histogramme des valeurs propres pour effectuer visuellement cet arbitrage à l'aide des méthodes vues pour l'ACP et l'AFC, critère du « coude » et critère de « Kaiser ».

- Critère de « Kaiser » : suggère de garder tous les axes dont la valeur propre est supérieure à la valeur propre moyenne.
- Critère du « coude » : on observe un décrochement (coude), suivi d'une décroissance régulière, entre le deuxième axe et le troisième axe, ce qui pourrait inciter à n'examiner que les deux premiers axes.

7.5.2 L'interprétation des axes

L'interprétation des axes factoriels se fait séquentiellement, pour chaque axe et chaque nuage de points, en regardant les contributions à la formation des axes.

Axe1

- Profile ligne (individus) :
 - Contribution des individus.
 - Qualités de représentation des individus.
- Profile colonne (modalités) :
 - Contribution des modalités.
 - Qualités de représentation des modalités.
- Conclusion.

Axe2

- Profile ligne (individus) :
 - Contribution des individus.
 - Qualités de représentation des individus.
- Profile colonne (modalités) :
 - Contribution des modalités.
 - Qualités de représentation des modalités.
- Conclusion.

Et ainsi de suite jusqu'à le dernier axe choisit.

7.6 L'interprétation d'un ACM

- L'analyse se fait axe par axe, en parallèle sur les profils ligne (individus) et colonne (variables/modalités).
- Plus la contribution est grande, plus l'influence du profil est grande.

- La contribution est considéré comme positive si le profil est dans la partie positive de l'axe.
- La contribution est considérée comme négatif si le profil est dans la partie négative de l'axe
- Le bilan des contributions peut être présenté pour un axe donné sous forme d'un tableau avec les principales contributions positifs (+) et contributions négatives (−) des profils (lignes/colonnes).
- Un profile est bien représenté si la qualité de représentation ($\cos^2$) de ce profile est proche de 1.

7.7 Conclusion

L'ACM est une méthode d'étude de plusieurs variables qualitatives définies sur un ensemble d'individus. Sa problématique est très riche et va bien au-delà d'une simple mise en œuvre de l'ACF sur un tableau particulier.

C'est là un des aspects de l'équivalence entre l'AFC sur le TDC et sur le tableau de Burt. L'existe d'ailleurs d'autres équivalences que celles déjà citées ; des points de vue très différents sur l'étude de variables qualitatives ont induit la conception de méthodes qui conduisent, au moins partiellement, aux mêmes résultats que l'AFC sur le TDC.

Outre qu'elles permettent de considérer l'ACM comme une méthode à part entière, ces convergences la renforcent. Les mécanismes de l'ACM, supportant plusieurs interprétations, sont d'une part adaptés à une vaste palette de problèmes concrets et d'autre part fournissent des résultats en accord avec plusieurs points de vue.

Chapitre 8

Application 1 : Une des Méthodes probabiliste avec un ACM sur R

Les patients du cancer de VADS.

8.1 La démarche de notre travail

8.1.1 Introduction

Les cancers des voies aérodigestives supérieures (VADS) sont un ensemble hétérogène de tumeurs. Les principaux facteurs de risque de cancer du larynx, de la cavité buccale, de l'oropharynx et de l'hypopharynx sont, par ordre d'importance, le tabac, l'alcool, une consommation de fruits de moins de 300 g/j et de légumes non féculents de moins de 300 g/j, une infection par un papillomavirus humain (HPV) et, enfin, des expositions professionnelles (amiante, vapeurs d'acides forts, dérivés du nickel et du chrome). Les deux premiers facteurs, tabac et alcool, augmentent chacun les risques indépendamment et, quand ils sont associés, ce qui est fréquent, les risques sont à peu près multiplicatifs.

Le cancer des voies aérodigestives supérieures (VADS) correspond à la présence d'une ou plusieurs tumeurs malignes localisées dans la partie haute des oranges respiratoires et /ou digestifs. Ces tumeurs ont pour origine une cellule normale qui a subi une transformation la rendant cancéreuse.

Prolifération anarchique de cette cellule cancéreuse mène à la formation d'une masse plus ou moins organisée appelée tumeur maligne. Dans un premier temps, celle-ci se développe de manière locale mais peut affecter d'autres endroits de l'organisme si elle n'est pas traitée, formant ainsi des métastases. Dans 90% des cas, les cancers des voies aérodigestives supérieures se développent à partir d'une muqueuse (tissu en contact avec l'environnement extérieur ; ex : la muqueuse nasale) et sont qualifiés de carcinomes épidermoïdes.

Remarque

- L'anatomie pathologique est la science qui étudie les lésions des tissus de l'organisme.

Elle permet d'établir le diagnostic, l'évolution et le pronostic des différentes maladies.

- La cytologie est la science qui étudie les différentes cellules normales ou pathologique ainsi leurs morphologies.

8.1.2 Démarches d'enquête

- Déterminer les informations nécessaires pour traiter l'analyse : ici des réponses liées aux effets et les caractéristiques du cancer des voies aérodigestives supérieures (VADS).

- Identifier les « sources » qui pourront nous donner ces informations : les patients des tumeurs de VADS qui étaient traités dans l'hôpital de BETNA.

- Identifier les méthodes de recueil de données pertinentes : étude d'un document.

- Elaborer des outils : un document pris de l'hôpital de BETNA.

- Recueil de données : c.-à-d. réaliser l'enquête à l'aide des outils cités précédemment.

Plusieurs éléments doivent être pris en compte pour déterminer le choix d'un ou plusieurs modes de recueil et le ciblage d'une ou plusieurs sources d'information :

- Les objectifs du travail : étudier le cancer des voies aérodigestives supérieures (Etudier leur effet, caractéristiques, médicaments...).

- La nature de ce qui doit être recueilli : ici des informations personnelles, existe-t-il des risques d'informations biaisées ? Oui.

- Les possibilités de traitement des données : dans notre cas, l'analyse des correspondances multiples.

- Les compétences liées à l'exploitation : temps d'encodage des réponses, rapide puisque tout se passe sur le l'Excel, techniques statistiques descriptives voire inférentielles.

8.1.3 La méthode de recueil les données

Le type d'enquête : c'est une enquête par l'étude d'un document. Cette étude dépend de :

La nature des documents

C'est un document scientifique.

La quantité des documents à analyser

On fait une analyse par échantillonnage. On prend un échantillon de 100 patients de VADS qui sont déjà traité dans l'hôpital de BETNA.

L'objet et le but d'investigation

Dans notre exemple, le type de l'analyse du document sélectionné est le type confirmatoire selon une stratégie est traité comme des données.

8.1.4 Echantillonnage

Méthode d'échantillonnage

On utilisant la méthode probabiliste d'échantillonnage « échantillonnage par grappe » pour faire une analyse de correspondance multiple (ACM) sur notre jeu de donnée, ces données nous le choisissent comme suit :

Nous avons pris un document contient 211 des patients de VADS qui étaient traité dans l'hôpital de BETNA en 2016.

On sépare 211 patients en grappes hétérogènes d'une taille semblable qu'on numérote de 1 à 4 :

Grappe 1 : 60 individus.

Grappe 2 : 50 individus.

Grappe 3 : 61 individus.

Grappe 4 : 40 individus.

On détermine la taille « 100 » de l'échantillon voulu.

On calcule la moyenne d'individus par grappes et le nombre de grappes qu'il faudra choisir pour constituer notre échantillon.

Moyenne d'individus par grappes

$$\frac{60 + 50 + 40 + 61}{4} = 52.27$$

Nombre de grappes

$$\frac{n}{\text{moyenned'individuspargrappes}} = \frac{100}{52.27} = 1.89 \approx 2.$$

Enfin, on choisit le nombre de grappes nécessaires par la méthode d'échantillonnage simple 2 grappes 1 et 4, et on sélectionne tous les individus de ces grappes choisies.

Détermination de la taille de l'échantillon

La formule donnant la taille minimale de l'échantillon est :

Rappel des notations :

- n est la taille de l'échantillon
 - t coefficient de marge traduit du tableau de confiance
 - p Proportion de la population qui représente une caractéristique donnée.
- (Si p est inconnue, on utilise p=0.5)
- e est la marge d'erreur choisie pour l'estimation.

On a :

Pour e=0.1 , et une proportion estimé à 50% soit p=0.50, si on se fixe un seuil de confiance de (1- α)=95%, on se reporte à la table de la loi normale, et on cherche la valeur correspondante à une probabilité (1- α)= 0,95, ce qui donne t = 1.96

La taille de l'échantillon sera donc d'au moins :

$$n = \frac{t^2 p(1-p)}{e^2} = \frac{1.96^2 \times 0.5 \times 0.5}{0.1^2} = 96,04$$

Notre échantillon doit contenir, donc, au moins 96 individus.

8.1.5 Le but d'étude

Comme nous l'avons expliqué précédemment, le VADS est l'un des types de cancers les plus dangereux qui affecte l'humanité. Cette maladie est classé

3eme en termes de taux de prévalence dans le monde.

Notre but est d'étudier le comportement des tumeurs VADS vis-à-vis des médicaments et si ces médicaments sont-ils bénéficient les patients ou non ? Est-ce qu'ils sont suffisants et efficace d'éliminer cette maladie ? Quel est le groupe d'âge le plus touché par cette maladie ? Est ce qu'il y a une relation entre la métastase du VADS et les types des médicaments ? Quel sont les effets de cette maladie ? Est ce qu'il y a une relation entre les différents types de traitements et le retour des tumeurs du VADS ?

8.1.6 Découvrez nos variables et leurs encodage

Notre jeu de données consiste 13 variables et 40 modalités. Ces variables se composent de 12 variables qualitatives et un seul variable quantitatif. Afin de faciliter le travail nous encodons ces variables.

Nous rappelons que chaque question est une variable et chaque réponse est une modalité. Voila l'encodage :

- 1/- L'âge : décompose de 3 intervalle, il encodé (Age)
 - [10,30[: encodé (A)
 - [30,60[: encodé (A+)
 - [60,90[: encodé (A++)
- 2/- Le sexe : masculin ou féminin, il encodé (Sexe)
 - Masculin : encodé (M)
 - Féminin : encoé (F)
- 3/- Le diagnostic : déterminer le type de maladie , il encodé (DGT)
 - Processus caviar : encodé (PC)
 - KC du cavum : encodé (KCC)
 - Carcinome caviar : encodé (CC)
 - UCNT : encodé (UCNT)
 - Récidive larynx : encodé (RL)
- 4/- La base de diagnostic : les tests médicaux utilisés pour diagnostiquer le type de maladie, il encodé (BD-DGT)
 - Anapath :encodé (ANA)
 - Cytologie : encodé (CT)
- 5/- La localisation de la maladie : déterminé où se situe la maladie, il encodé (LOC)
 - Nasopharynx : encodé (NASO)
 - Larynx : encodé (LAR)
 - Bouche : encodé (BC)

- Glandes salivaires : encodé (GC)
- Hypopharynx : encodé (HYPO)
- Sinus : encodé (SIN)
- 6/- La morphologie : déterminer si la tumeur bénigne ou maligne, il encodé (MOR)
 - Bénigne : encodé (CAR)
 - Maligne : encodé (NCAR)
- 7/- Le grade histologique : la phase où la tumeur a été découverte, il encodé (GD-HIS)
 - 1er : encodé (G)
 - 2ème : encodé (G+)
 - 3ème : encodé (G++)
 - 4ème : encodé (G+++)
- 8/- La méthode de traitement : quelle est la méthode de traitement que le patient a subi, il encodé (M-TRT)
 - la chirurgie : encodé (R)
 - la chimiothérapie : encodé (CH)
 - la radiothérapie : encodé (CG)
- 9/- La métastase : la tumeur s'est-elle propagée ou non, il encodé (MTS)
 - Propagé : encodé (M+)
 - Pas propagée : encodé (M-)
- 10/- La récurrence : après le traitement, les tumeurs sont-elles réapparues , il encodé (RCD)
 - Revenue : encodé (RC+)
 - Pas revenue : encodé (RC-)
- 11/- L'effet indésirables : Effets secondaires des tumeurs, il encodé (E-INDRB)
 - Perdu de vue : encodé (PV)
 - Osseuse : encodé (OS)
 - Pulmonaire : encodé (PL)
 - Hépatique : encodé (HP)
 - Aucune effet : encodé (NO)
- 12/- Vivant : Le patient est-t-il vive ou est-il mort , il encodé (VIV)
 - Vivant : encodé (V+)
 - Mort : encodé (V-)
- 13/- L'encore de traitement : Le patient est-t-il traite à nouveau , il encodé (ED- TRT)
 - Oui : encodé (TRT+)

- Non : encodé (TRT-)

8.2 Partie analyse

8.2.1 Les données

Nous avons rassemblées notre données dans un tableau de type individus variables sous forme « CSV » appelé ‘MALADIE 1’.

AGE	SEXE	DGT	BD-DGT	LOC	MOR	GD-HIS	M-TRT	MTS	RCD	ED-TRT	E-INDRB	VIV
A+	M	PC	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	OS	V-
A+	M	KCC	ANA	NASO	CAR	G++	R	M+	RC-	TRT-	HP	V-
A+	M	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A	F	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	R	M+	RC-	TRT-	NO	V-
A++	F	CC	ANA	NASO	CAR	G++	R	M+	RC-	TRT-	HP	V-
A++	M	UCNT	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	NO	V-
A	M	CC	ANA	NASO	NCAR	G++	R	M-	RC-	TRT-	NO	V+
A	M	UCNT	ANA	NASO	CAR	G+++	R	M+	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT+	NO	V+
A+	M	KCC	ANA	NASO	CAR	G+	CH	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	CAR	G++	R	M-	RC+	TRT+	NO	V+
A	F	KCC	ANA	NASO	CAR	G	CH	M-	RC-	TRT-	NO	V+
A+	F	CC	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	NO	V-
A++	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC+	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	R	M+	RC-	TRT-	NO	V+
A++	M	KCC	ANA	NASO	CAR	G++	R	M-	RC-	TRT-	NO	V+
A++	M	RL	CT	LAR	NCAR	G++	CG	M-	RC+	TRT-	NO	V+
A+	F	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	F	KCC	ANA	NASO	NCAR	G	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	CH	M+	RC-	TRT+	NO	V+
A++	M	KCC	ANA	NASO	NCAR	G++	R	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G+++	R	M-	RC-	TRT+	NO	V+
A+	M	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	NCAR	G++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	CAR	G	R	M-	RC-	TRT+	NO	V+
A	F	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G+++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	CT	LAR	CAR	G+	CH	M-	RC-	TRT-	NO	V-
A	F	KCC	CT	LAR	CAR	G	CH	M-	RC-	TRT-	NO	V-

8.2.2 Analyse descriptive

Description de la population selon la variable âge

La population globale est répartie en 3 classes :

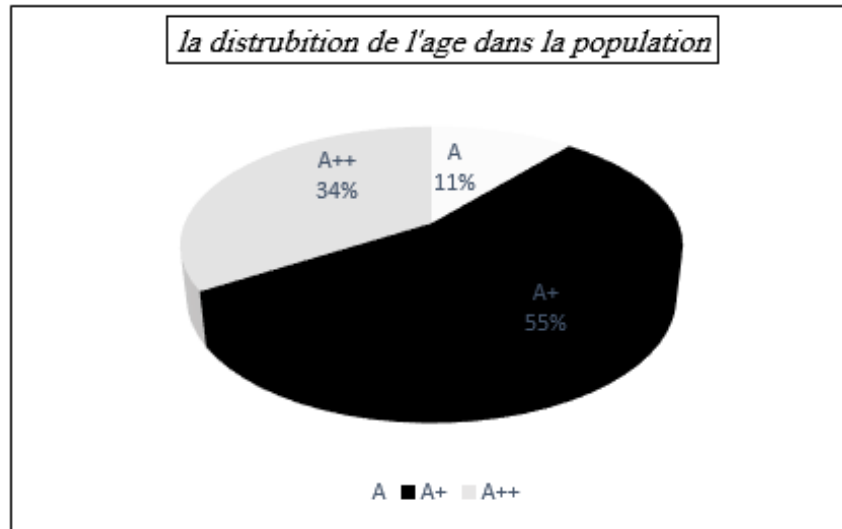
A : [10 ; 30[	A+ : [30 ; 60[	A++ : [60 ; 90[	Totale
11	55	34	100

- Moyenne de l'âge : 33.3333.

- Maximum : 85.

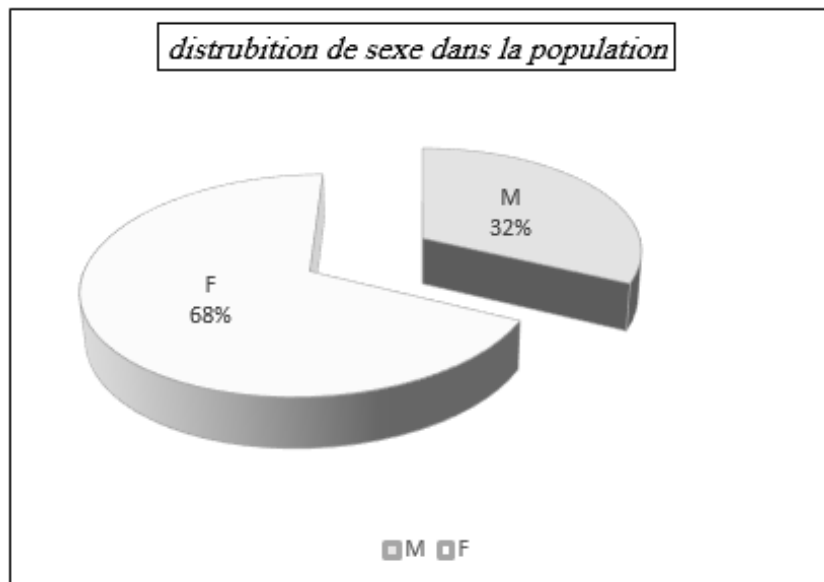
- Minimum : 10.

- Ecart type : 22.007.



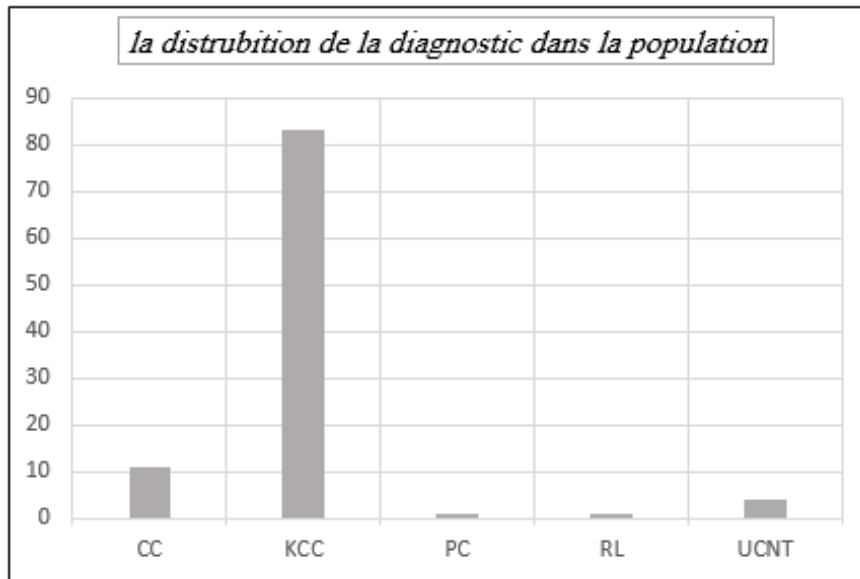
Effet du sexe

Femme	Homme	Totale
68	32	100



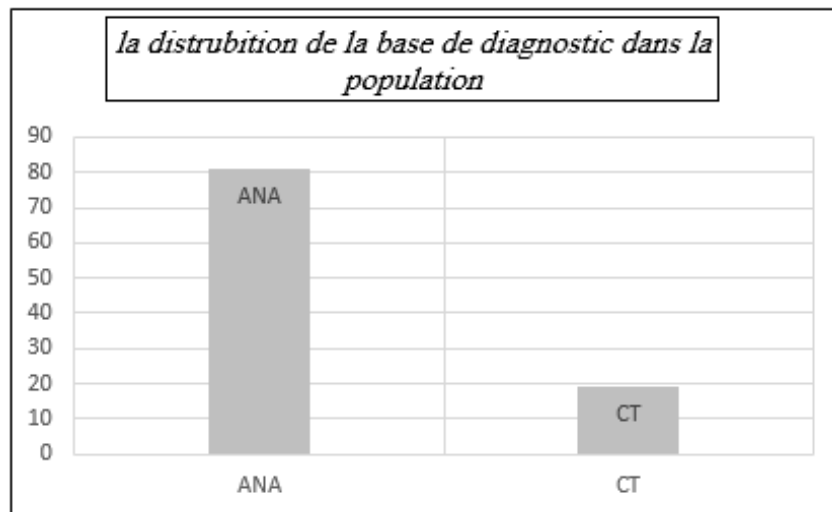
Effet du Diagnostic

CC	KCC	PC	RL	UCNT
11	83	1	1	4



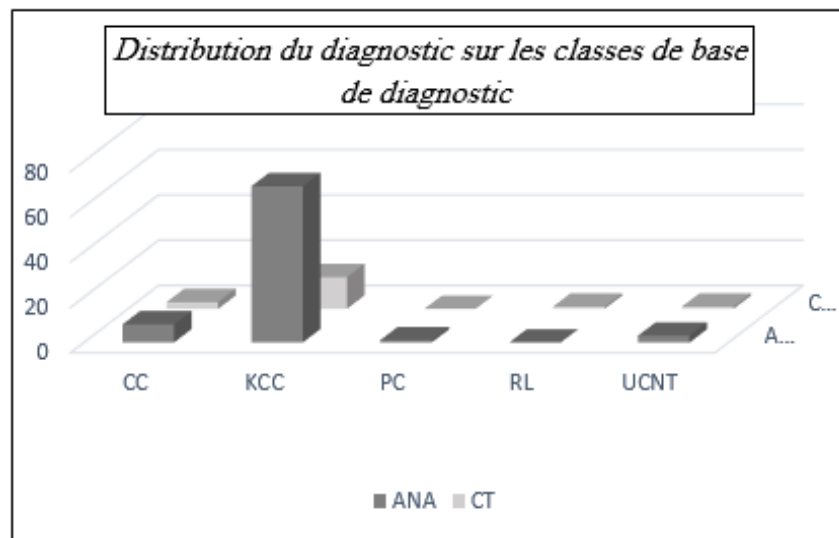
La base de diagnostic

ANA	CT
81	19



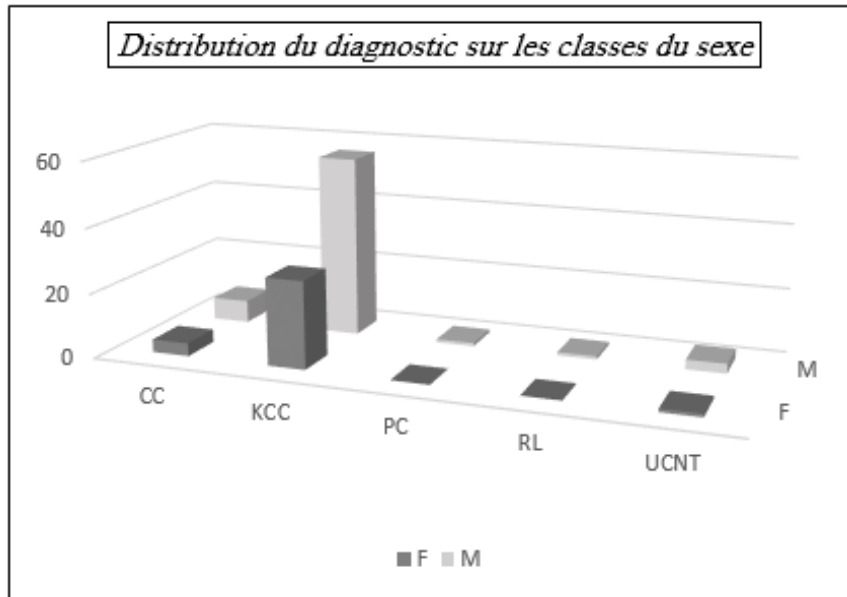
Distribution du diagnostic selon les classes de base de diagnostic

	CC	KCC	PC	RL	UCNT
ANA	8	69	1	0	3
CT	3	14	0	1	1



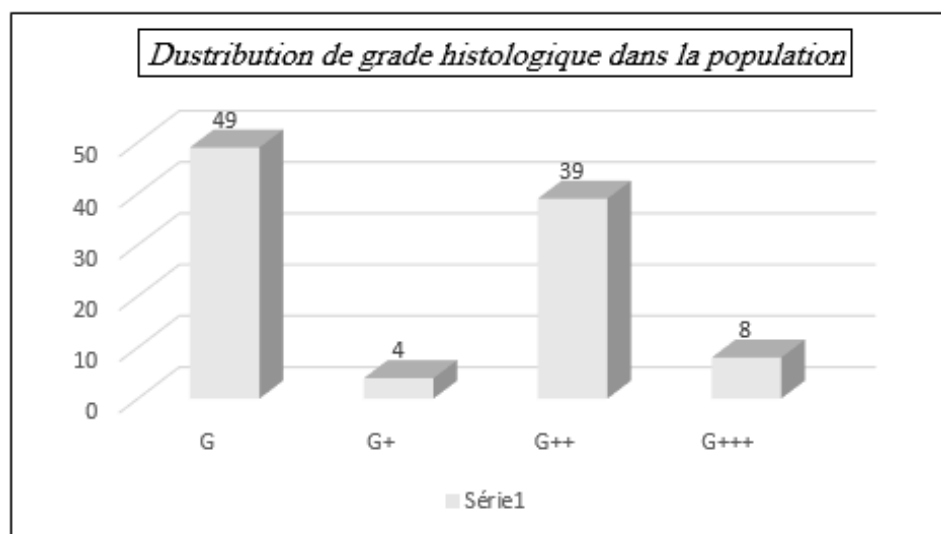
Distribution du diagnostic selon les classes du sexe

	CC	KCC	PC	RL	UCNT
F	4	27	0	0	1
M	7	56	1	1	3



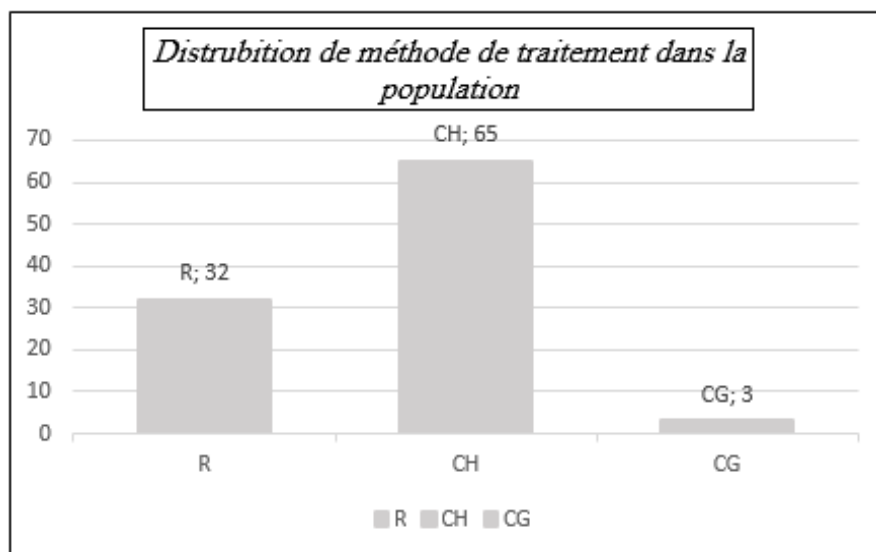
Grade histologique

G	G+	G++	G+++
49	4	39	8



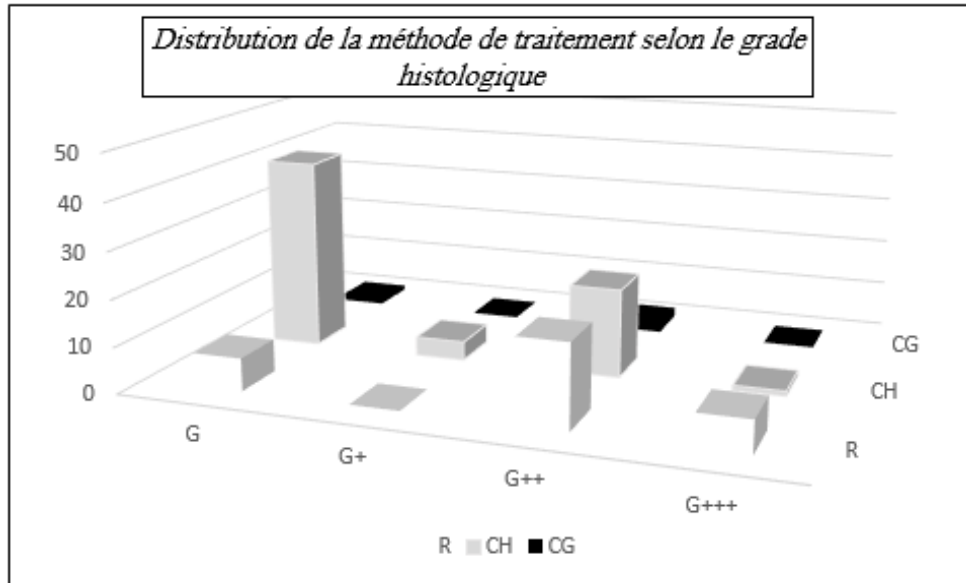
Méthode de traitement

R	CH	CG
32	65	3



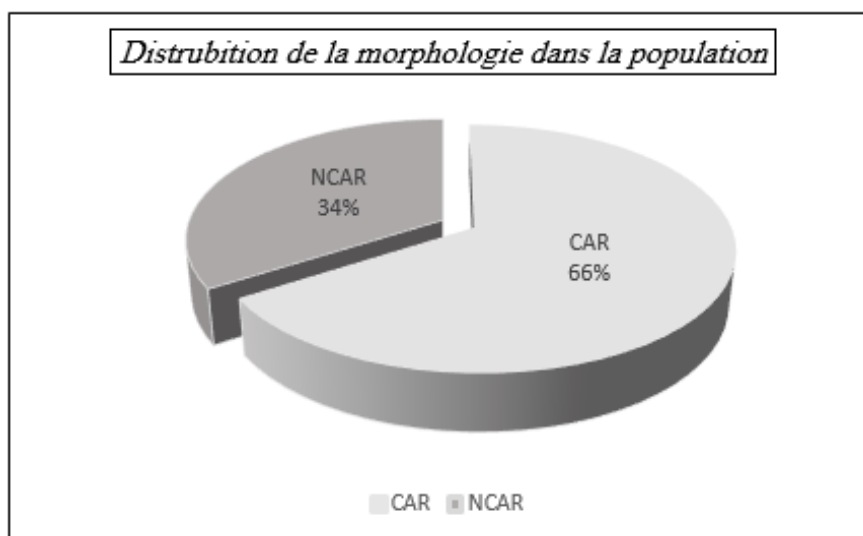
Distribution de la méthode de traitement selon le grade histologique

	G	G+	G++	G+++
R	7	0	18	7
CH	41	4	19	1
CG	1	0	2	0



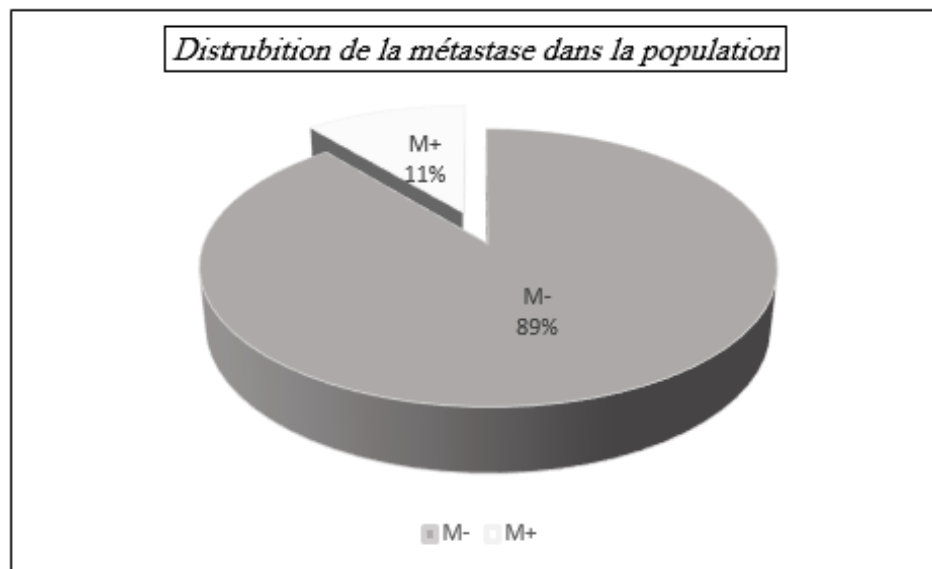
Morphologie

CAR	NCAR
66	34



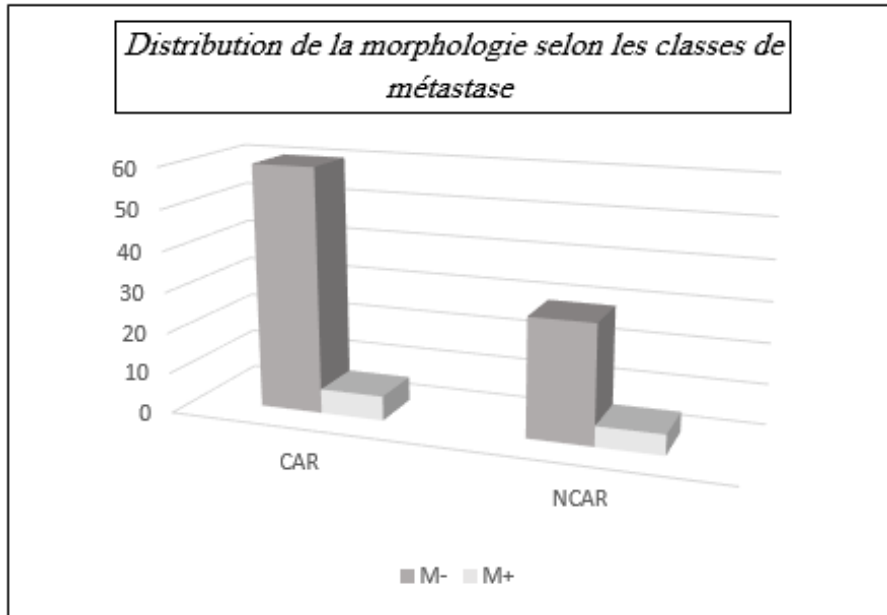
Métastase

M+	M-
11	89



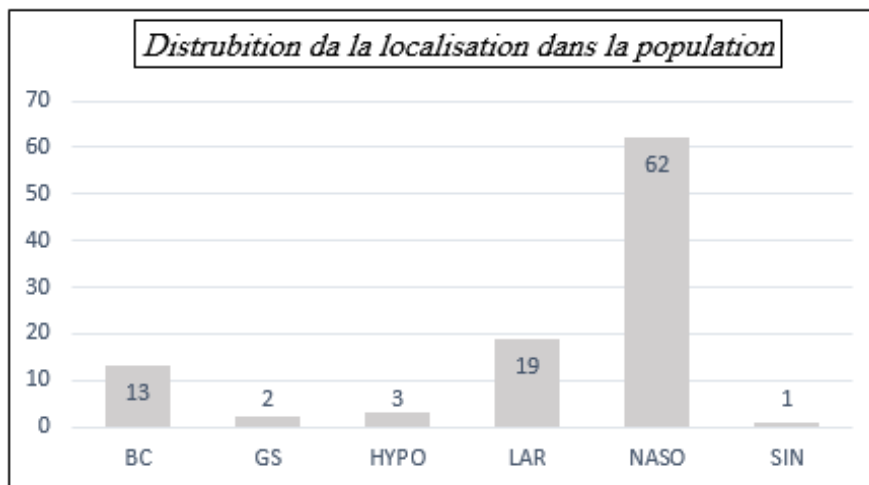
Distribution de la morphologie selon les classes de métastase

	CAR	NCAR
M-	60	29
M+	6	5



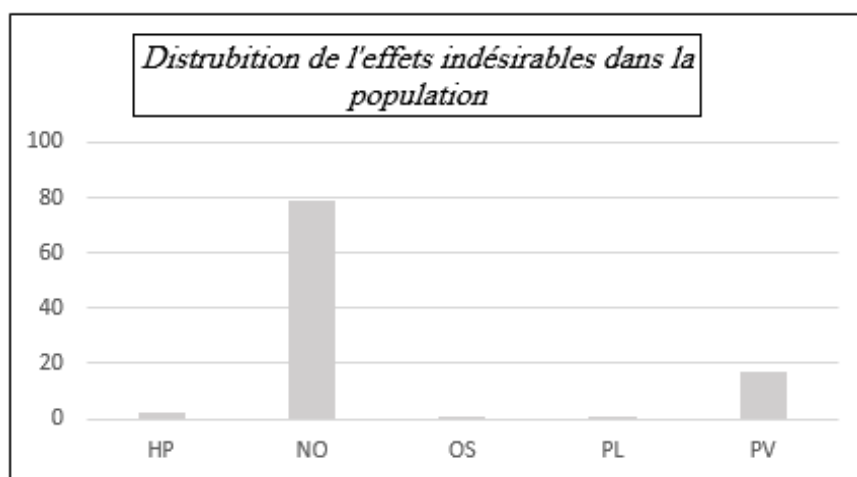
Localisation

BC	GS	HYPO	LAR	NASO	SIN
13	2	3	19	62	1



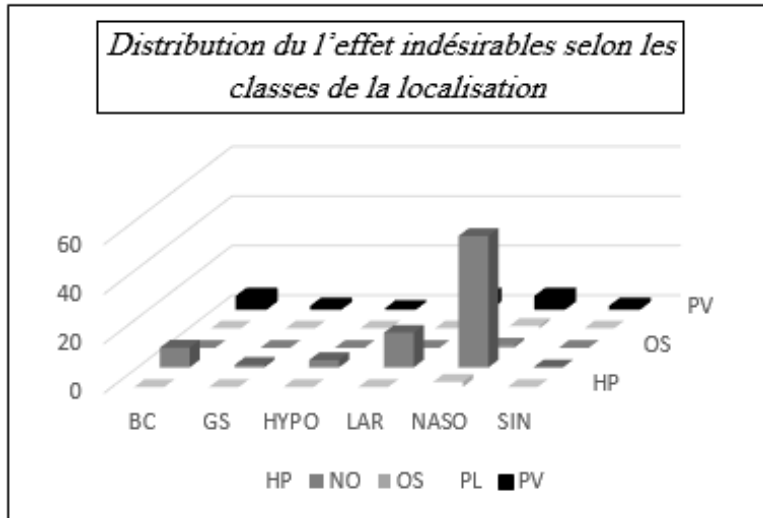
Effets indésirables

HP	NO	OS	PL	PV
2	79	1	1	17



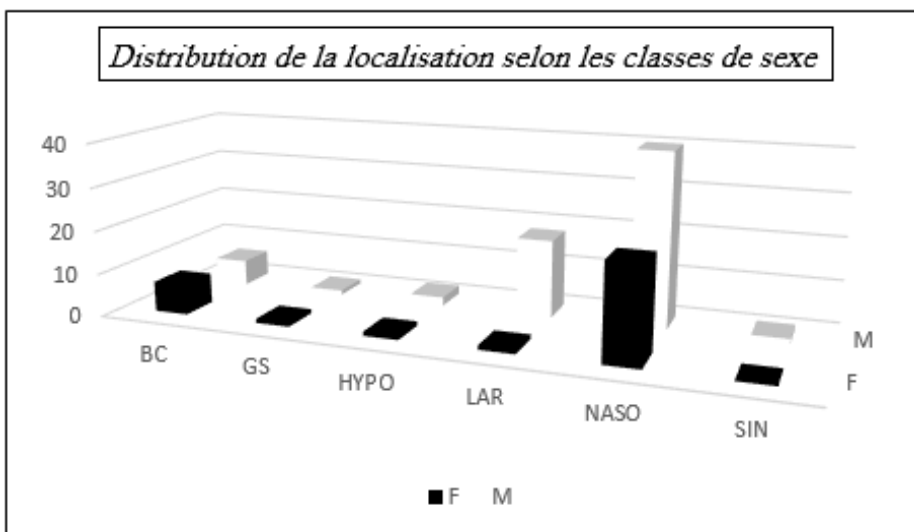
Distribution du l'effet indésirables selon les classes de la localisation

	BC	GS	HYPO	LAR	NASO	SIN
HP	0	0	0	0	2	0
NO	8	1	3	14	53	0
OS	0	0	0	0	1	0
PL	0	0	0	0	1	0
PV	5	1	0	5	5	1



Distribution de la localisation selon les classes de sexe

	BC	GS	HYPO	LAR	NASO	SIN
F	7	1	1	1	22	0
M	6	1	2	18	40	1



8.2.3 Analyse des correspondances multiples

Importation du jeu de données

Nous installons d'abord les packages suivantes :

-Le package « FactoMiner » pour exécuter ACM l'instruction qui construire ce tableau sur « R » est la suivante :

```
library (FactoMineR)
```

-Le package « ggplot2 » pour visualiser les graphes sous « R » l'instruction qui construire ce visualisation :

```
library (ggplot2)
```

-Le package « factoextra » pour facilitant l'extraction et la visualisation de la sortie d'analyses exploratoires de données multivariées.

```
library (factoextra)
```

Importation des données sur « R » par l'instruction suivante :

```
donnees<-read.csv("MALADIE 1.CSV ",header=TRUE,sep=";",dec  
=".",row.names =1)
```

```
donnees
```

Nous verrons les données comme suit :

8.2. PARTIE ANALYSE

```
donnees <-read.csv("MALADIE 1.CSV ",header=TRUE,sep=";",dec=".",row.names =1)
donnees
```

AGE	SEXE	DGT	BD.DGT	LOC	MOR	GD.HIS	M.TRT	MTS	RCD	ED.TRT	E.INDRB	VIV
A+	M	PC	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	OS	V-
A+	M	KCC	ANA	NASO	CAR	G++	R	M+	RC-	TRT-	HP	V-
A+	M	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A	F	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	R	M+	RC-	TRT-	NO	V-
A++	F	CC	ANA	NASO	CAR	G++	R	M+	RC-	TRT-	HP	V-
A++	M	UCNT	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	NO	V-
A	M	CC	ANA	NASO	NCAR	G++	R	M-	RC-	TRT-	NO	V+
A	M	UCNT	ANA	NASO	CAR	G+++	R	M+	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G+++	R	M-	RC-	TRT+	NO	V+
A+	M	KCC	ANA	NASO	CAR	G+	CH	M-	RC-	TRT-	NO	V-
A+	M	KCC	ANA	NASO	CAR	G++	R	M-	RC+	TRT+	NO	V+
A	F	KCC	ANA	NASO	CAR	G	CH	M-	RC-	TRT-	NO	V+
A+	F	CC	ANA	NASO	CAR	G++	CH	M+	RC-	TRT-	NO	V-
A++	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC+	TRT-	NO	V-
A+	M	KCC	ANA	NASO	NCAR	G++	R	M+	RC-	TRT-	NO	V+
A++	M	KCC	ANA	NASO	CAR	G++	R	M-	RC-	TRT-	NO	V+
A++	M	RL	CT	LAR	NCAR	G++	CG	M-	RC+	TRT-	NO	V+
A+	F	KCC	ANA	NASO	NCAR	G++	CH	M-	RC-	TRT-	NO	V+
A+	M	KCC	ANA	NASO	CAR	G++	CH	M-	RC-	TRT-	NO	V+

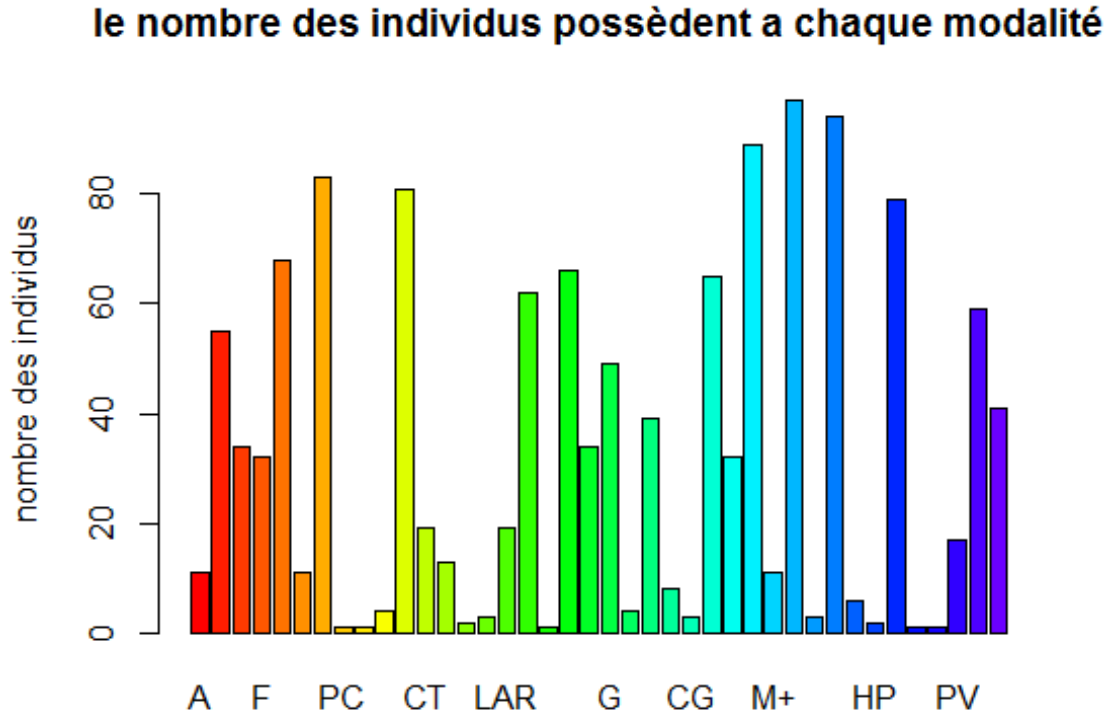
Traitement du jeu de données

Après l'importation des données nous commençons le traitement, et La première chose que nous commençons est le tableau disjonctif complet.

```
disj<- tab.disjonctif(donnees)
```

```
disj
```

Nous avons vu dans R :



Remarque : La somme de Chaque ensemble de modalité du même variable est égale à 100 qui est le nombre de participants da cette enquête.

Pour généraliser l'analyse des correspondances à l'étude des croisement entre plus de deux variables on peut construire un tableau contenant l'ensemble des tableaux de contingence entre les variables prises 2 à 2. C'est le tableau de Burt.

```
Burt <- t(disj)%*%disj
```

Burt

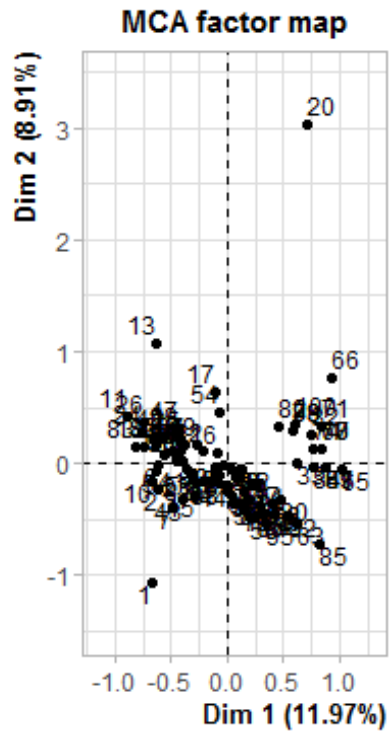
CHAPITRE 8. APPLICATION 1 : UNE DES MÉTHODES PROBABILISTE AVEC UN ACM SUR R

```
> Burt <- t(cbind(j1**w1,j2
> Burt
      A A+ A++ F H CC ECC FC FL UCRT AMA CT SC GS HYPO LAR NASO JIN CAR NCAR G G+ G++ G+++ CG CH K-M M+ SC- SC+ INT- INT+ HP HO CO FL FY V- V+
A      11 0 0 6 5 1 9 0 0 1 10 1 2 0 0 1 8 0 6 5 6 0 3 2 0 6 5 10 1 11 0 11 0 0 8 0 0 2 8 5
A+     0 55 0 20 35 4 49 1 0 1 47 8 7 1 1 8 38 0 34 21 22 3 24 8 1 35 19 49 6 53 2 49 4 1 45 1 0 8 25 50
A++    0 0 34 6 28 6 25 0 1 2 24 10 4 1 2 10 16 1 26 8 21 1 12 0 2 24 8 30 4 33 1 34 0 1 28 0 1 7 28 4
F      6 20 6 32 0 4 27 0 0 1 51 1 7 1 1 1 22 0 18 14 16 1 12 3 1 21 10 50 2 32 0 32 0 1 27 0 0 4 18 14
H      5 35 28 0 68 7 84 1 1 3 50 18 6 1 2 18 40 1 40 20 33 3 27 5 2 44 22 59 8 65 3 62 6 1 82 1 1 13 41 27
CC     1 4 6 6 7 11 0 0 0 0 8 2 2 1 1 2 3 2 1 9 2 4 1 6 0 0 7 4 9 2 11 0 11 0 1 6 0 0 2 7 4
ECC    9 49 25 27 54 0 83 0 0 0 69 14 11 1 2 14 55 0 53 30 44 3 30 4 2 55 26 77 6 81 2 77 6 1 66 0 1 15 48 35
FC     0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 0 1 0 1 0 0 1 0 0 1 0 1 0
FL     0 0 1 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 1 0 0 0 1 0 0 0 1 0 1
UCRT   1 1 1 1 2 0 0 0 0 0 4 3 1 0 0 0 1 3 0 3 1 1 0 1 2 0 2 2 2 2 8 0 4 0 0 4 0 0 3 1
AMA    10 47 24 31 50 8 69 1 0 3 81 0 13 2 3 0 42 1 48 33 35 2 34 8 1 50 30 70 11 79 2 78 4 2 45 1 1 12 46 35
CT     1 8 10 1 18 3 14 0 1 1 0 19 0 0 0 19 0 0 18 1 14 2 3 0 2 15 2 19 0 18 1 19 0 0 14 0 0 5 13 4
SC     2 7 4 7 4 2 11 0 0 8 13 0 13 0 0 0 0 0 9 4 9 0 4 0 1 10 2 13 0 13 0 13 0 0 8 0 0 3 11 2
GS     0 1 1 1 1 1 1 0 0 0 2 0 0 2 0 0 0 0 2 0 2 0 0 0 0 2 0 2 0 2 0 2 0 0 0 1 0 0 1 1 1
HYPO   0 1 2 1 2 1 2 0 0 0 3 0 0 0 3 0 0 0 2 1 2 1 0 0 0 3 3 3 0 3 0 3 0 0 3 0 0 3 0 2 1
LAR     1 8 10 1 18 3 14 0 1 1 0 19 0 0 0 19 0 0 18 1 14 2 3 0 2 15 2 19 0 18 1 19 0 0 14 0 0 5 13 4
NASO   8 38 16 32 40 3 53 1 0 3 62 0 0 0 0 0 42 0 34 28 21 1 32 8 0 34 28 51 11 40 2 54 4 2 53 1 1 5 31 31
JIN     0 0 0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 1 1 0 0 1 0 0 0 0 1 0 1 0 1 0 0 0 0 0 1 1 0
CAR     4 34 26 18 48 9 33 1 0 3 48 18 8 2 2 18 34 1 46 0 37 4 22 3 1 49 16 60 4 44 2 42 4 2 30 1 0 13 42 24
NCAR    3 21 0 14 20 2 30 0 1 1 33 1 4 0 1 1 28 0 0 34 12 0 17 5 2 16 16 29 5 33 1 32 2 0 29 0 1 4 17 17
G       4 22 21 16 35 4 44 0 0 1 35 14 9 2 2 14 21 1 37 12 49 0 0 0 1 41 7 47 3 49 0 48 1 0 31 0 1 17 34 13
G+      0 5 1 1 3 1 3 0 0 0 2 2 0 0 1 2 1 0 4 0 0 4 0 0 0 4 0 4 0 4 0 4 0 0 4 0 0 0 2 2
G++     3 24 12 12 27 6 30 1 1 1 36 3 4 0 0 3 32 0 22 17 0 0 39 0 2 19 18 31 8 34 3 36 3 2 34 1 0 0 17 22
G+++    2 6 0 3 3 0 6 0 0 2 8 0 0 0 0 3 8 0 3 5 0 0 0 8 0 1 7 7 1 8 0 6 2 0 6 0 0 0 4 4
CG      0 1 2 1 2 0 2 0 1 0 1 2 1 0 0 2 0 0 1 2 1 0 2 0 3 0 0 3 0 2 1 3 0 0 3 0 0 0 2 1
CH      8 35 24 21 44 7 55 1 0 2 50 15 10 2 3 15 34 1 49 16 41 4 19 1 0 65 0 61 4 64 1 62 2 0 42 1 0 16 41 24
K-M     3 19 9 10 22 4 26 0 0 2 30 2 2 0 0 2 28 0 16 16 7 0 18 7 0 0 32 25 7 31 1 28 4 2 28 0 1 1 16 14
M+     10 49 50 30 59 9 77 0 1 2 70 19 13 2 3 19 51 1 60 29 47 4 31 7 3 41 25 89 0 86 3 84 0 0 72 0 0 17 50 29
SC-     1 6 4 2 9 2 4 1 0 2 11 0 0 0 0 0 11 0 6 5 2 0 8 1 0 4 7 0 11 11 0 10 1 2 7 1 1 0 9 2
SC+    11 53 35 32 65 11 81 1 0 4 79 18 13 2 3 18 60 1 44 35 49 4 34 8 2 44 31 86 11 37 0 92 3 2 74 1 1 17 58 39
INT-    0 2 1 0 3 0 2 0 1 0 2 1 0 0 0 1 2 0 2 1 0 0 3 0 1 1 1 3 0 0 3 1 0 3 0 0 0 1 2
INT+   11 49 34 32 62 11 77 1 1 4 75 19 13 2 3 19 86 1 42 32 48 4 34 4 3 43 29 84 20 92 2 94 0 2 73 1 1 17 53 35
HP      0 6 0 0 4 0 4 0 0 2 6 0 0 0 0 0 6 0 4 2 1 0 3 2 0 2 4 5 1 3 1 0 4 0 4 0 0 0 4
HO      0 1 1 1 1 1 1 0 0 8 2 0 0 0 0 0 2 0 2 0 0 0 2 0 0 0 2 0 2 2 0 2 0 2 0 0 0 0 2 0
CO      9 45 25 27 52 8 66 0 1 4 65 14 8 1 3 14 53 0 50 29 31 4 34 8 3 48 28 72 7 74 3 73 4 0 79 0 0 0 38 41
FL      0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 1 0 0 1 1 0 1 0 0 0 1 0 0 1 0
FY      0 0 1 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 1 0 1 0
V-      2 8 7 4 13 2 19 0 0 0 12 3 8 1 0 5 5 1 13 4 17 0 0 0 16 1 17 0 17 0 0 0 0 0 17 17 0
V+     6 25 28 18 41 7 48 1 0 3 46 13 11 1 2 13 31 1 42 17 36 2 17 4 2 41 16 50 9 58 1 59 0 2 38 1 1 17 59 0
```

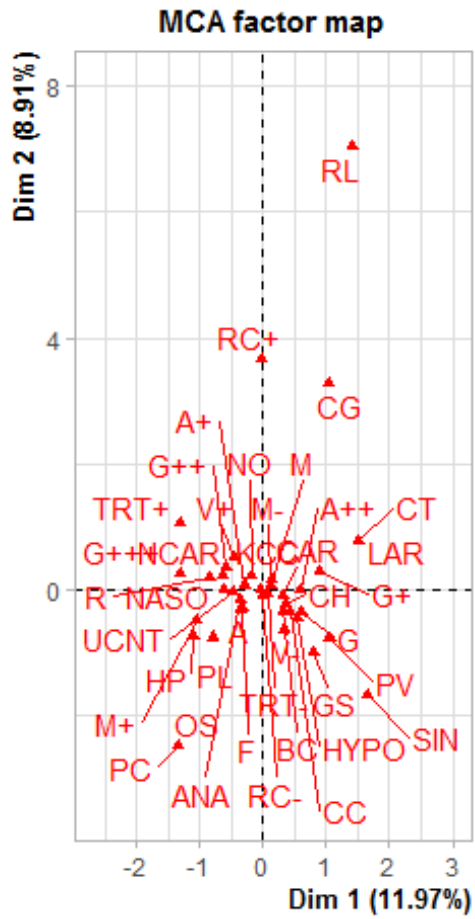
Nous avons vu la représentation graphique préliminaire des individus, des modalités et des variables par l'instruction suivante :

```
M <- MCA(donnees,ncp=2,graph=TRUE)
```

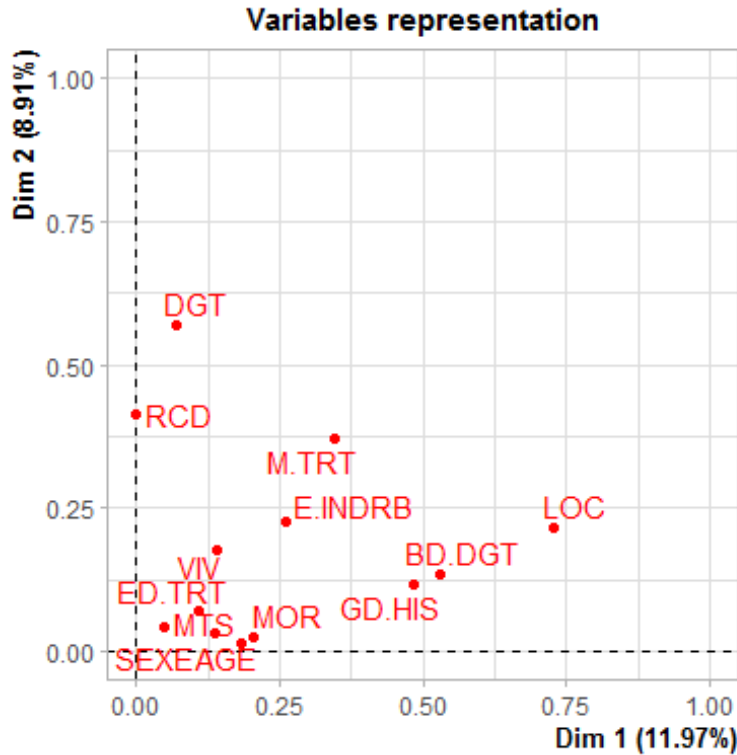
- Représentation graphique des individus



- Représentation graphique des modalités



- Représentation graphique des variables



Tous ces représentations graphiques sont avant le centrage des coordonnés des individus ou des variables. C'est pour ça on observe des figures mal déformées avec une grande variance entre les éléments soit individu ou variable.

Les coordonnées factorielles standardisées des individus et des modalités :

Et comme nous dit précédemment, les anciennes représentations sont mal déformées c'est pour ça on construit une base standardisée qui donne des nouveaux coordonnés. Se dernière aide nous à représenter les individus et les variables avec la meilleure forme qui conserve le maximum d'information incluse dans l'ensemble des données à étudié.

Et ce sont les deux structures qu'ils nous habilitent de faire ça :

Xstand <- X %*% diag(1/d) #coordonnées factorielles standardisées des individus.

Ystand <- Y %*% diag(1/d) #coordonnées factorielles standardisées des modalités.

```
Xstand <- X %*% diag(1/d)
Xstand
      [,1]      [,2]
-1.345787782 -2.47515407
-1.236682336 -0.55487507
-1.105138434  0.17182129
-0.744309599  0.39811877
-1.269074098 -0.17200338
-1.245291048 -0.02009519
-0.961469272 -0.92279641
-0.433850076 -0.37198576
-0.993033516  0.26285102
-1.326341464 -0.36499046
-1.768918805  0.99074198
  0.248008168 -0.12349388
-1.255363374  2.47069657
-0.244520327 -0.36965192
-0.577135207 -0.65288317
-0.173420770  0.22785381
-0.218942670  1.48035669
-1.481690849  0.33528834
-0.546061103  0.39927300
  1.411962211  7.03152731

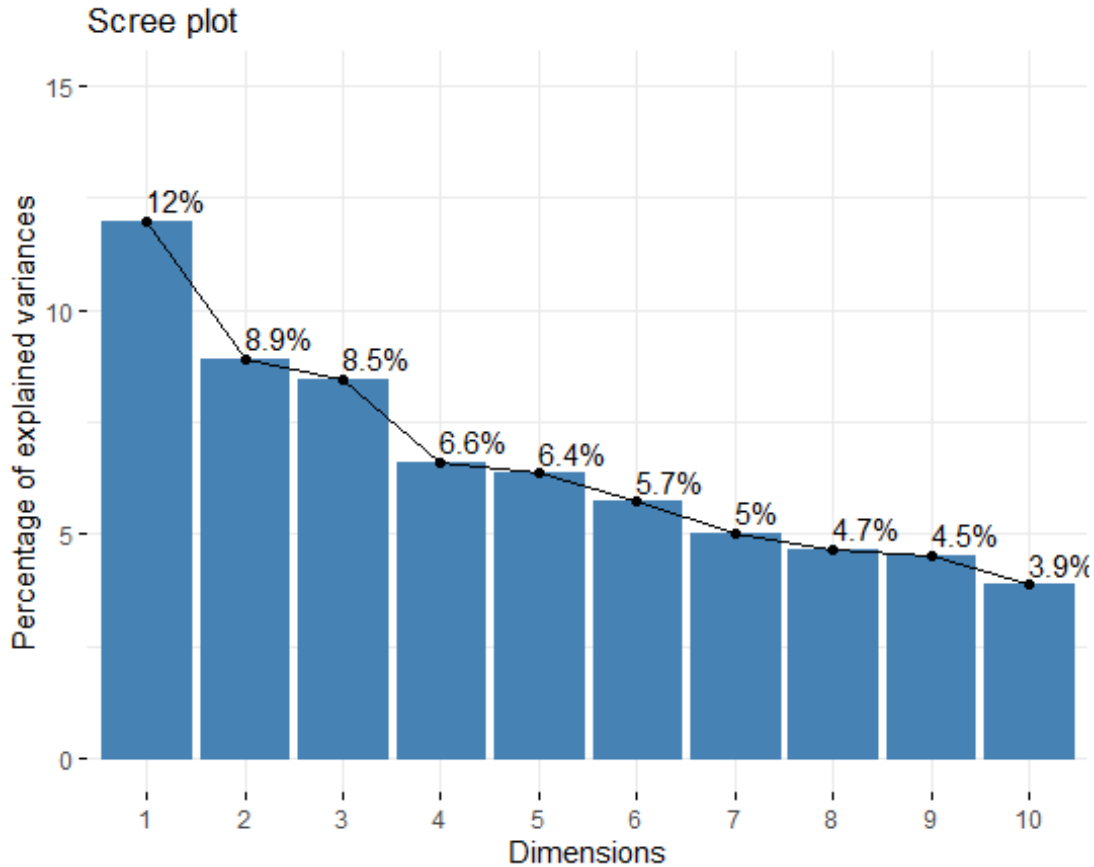
Ystand <- Y %*% diag(1/d)
Ystand
      [,1]      [,2]
-0.721925411 -0.72564843
-0.592481243  0.17172561
  1.191989643 -0.04302282
-0.634575295 -0.69699684
  0.298623668  0.32799851
  0.861583969 -0.83138315
-0.071095595 -0.01402981
-2.699018974 -5.75500595
  2.831733835 16.34907580
-0.927301035 -0.07109532
-0.707662842 -0.41295198
  3.016878433  1.76047949
  0.683094329 -1.47650891
  1.608235994 -2.35918555
  1.109289552 -1.03932895
  3.016878433  1.76047949
-1.226841847 -0.04037943
```

Aide à l'interprétation

Le choix du nombre des axes retenir Combien en retenir ? C'est l'examen de l'histogramme des valeurs propres qui va permettre de choisir. On trace cet histogramme de la façon suivante, avec la fonction suivant :

```
fviz_screplot (M, addlabels = TRUE, xlim=c(1,10), ylim = c
(1, 15))

title("Eboulis des valeurs propres en % ")
```



On peut utiliser l'histogramme des valeurs propres pour effectuer visuellement cet arbitrage à l'aide des méthodes vues pour l'ACP et l'AFC, critère du « coude » et critère de « Kaiser ».

- Critère de « Kaiser » : suggère de garder tous les axes dont la valeur propre est supérieure à la valeur propre moyenne, Ici, le critère de Kaiser incite à retenir les axes dont le pourcentage d'inertie est supérieur à $(100/13) = 7.69\%$, on garde alors les 2 premiers axes.

- Critère du « coude » : on observe un décrochement (coude), suivi d'une décroissance régulière, entre le deuxième axe et le troisième axe, ce qui pourrait inciter à n'examiner que les deux premiers axes.

Les deux premiers axes sont bien interprétables, ils cumulent presque de 20.88% l'inertie totale (11.97% pour le premier axe 8.91% pour le deuxième axe expliquant par les valeurs propres 0.25 et 0.18 respectivement).

La fonction suivante nous donne la distribution du le pourcentage d'inertie

aux 27 dimensions expliquant par les valeurs propres.

`round(M$eig,2)`

```
> round(M$eig,2)
      eigenvalue percentage of variance cumulative percentage of variance
dim 1      0.25          11.97          11.97
dim 2      0.18           8.91          20.88
dim 3      0.18           8.46          29.34
dim 4      0.14           6.61          35.95
dim 5      0.13           6.36          42.31
dim 6      0.12           5.72          48.03
dim 7      0.10           5.03          53.06
dim 8      0.10           4.68          57.74
dim 9      0.09           4.51          62.25
dim 10     0.08           3.87          66.13
dim 11     0.08           3.81          69.93
dim 12     0.07           3.57          73.50
dim 13     0.07           3.23          76.74
dim 14     0.06           3.11          79.85
dim 15     0.06           2.96          82.81
dim 16     0.05           2.63          85.43
dim 17     0.05           2.55          87.98
dim 18     0.05           2.25          90.23
dim 19     0.04           1.94          92.17
dim 20     0.04           1.75          93.92
dim 21     0.03           1.59          95.51
dim 22     0.03           1.37          96.88
dim 23     0.03           1.25          98.13
dim 24     0.02           1.10          99.23
dim 25     0.02           0.77         100.00
dim 26     0.00           0.00         100.00
dim 27     0.00           0.00         100.00
```

Noter dans la deuxième colonne le Pourcentage d'inertie explique par chaque valeur propre et le Pourcentage d'inertie cumulé, Quel total jusqu'à 100 dans la troisième colonne.

L'interprétation des axes L'interprétation des axes factoriels se fait séquentiellement, pour chaque axe et chaque nuage de points, en regardant les contributions à la formation des axes.

L'interprétation d'un ACM

Les graphiques peut devenir illisible, il ne permet pas de distinguer entre les modalités qui contribuent le plus fortement et celles qui ne contribuent pas aux axes représentés.

Dans ce cas-là on peut se servir de leur contribution aux axes représentés pour sélectionner celles qui ont justement les contributions les plus importantes.

Un critère possible est de ne retenir que les modalités dont la contribution est supérieure à deux fois la contribution moyenne.

(PS : Ce seuil est évidemment arbitraire, et rien n'empêche de retenir un seuil plus élevé ou plus faible).

Suivant le nombre de modalités qu'on veut représenter, dans notre cas, le nombre de modalités actives étant faible.

On peut garder celles dont la contribution à l'un ou l'autre des deux premiers axes est simplement supérieure à la contribution moyenne.

La contribution moyenne est l'inverse du nombre de modalités actives.

Dans notre cas, les 13 variables actives se décomposent en un total de 40 modalités.

La contribution moyenne est donc égale à :

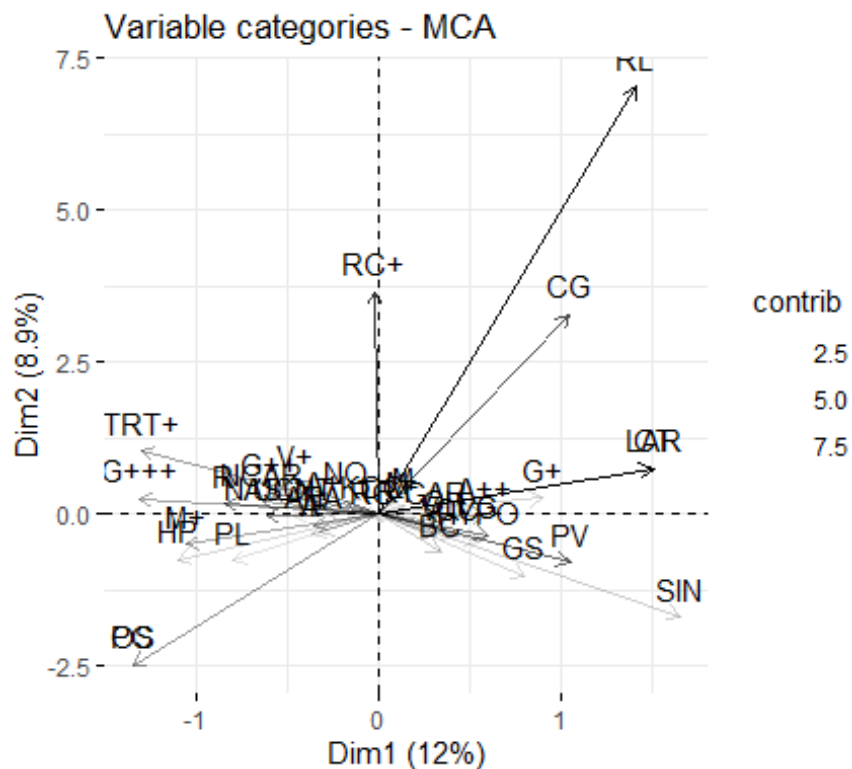
$$\text{Contribmoyenne} = 100/40 = 2.5\%.$$

On ne retiendra donc dans l'interprétation des deux premiers axes de notre ACM que les modalités dont la contribution est supérieure à 2.5%.

Seules les modalités dont la contribution est élevée sont à considérer pour l'interprétation d'un axe.

L'axe 1 Profile colonne (modalités)

```
fviz_pca_var(M, pointsize = "contrib", alpha.var = "contrib", select.var = list(contrib = 40))
```



Pour repérer les contributions significatives, on utilise le tableau de contribution suivant :

```
axe1<-matrix(nrow=40, ncol=3)

axe1[,1]<-round(M$var$coord[,1],1)

axe1[,2]<-round(M$var$contrib[,1],1)

axe1[,3]<-round(M$var$cos2[,1],1)

rownames(axe1)<-rownames(M$var$coord)

colnames(axe1)<-c('coord1','contrib1', 'cos²1')

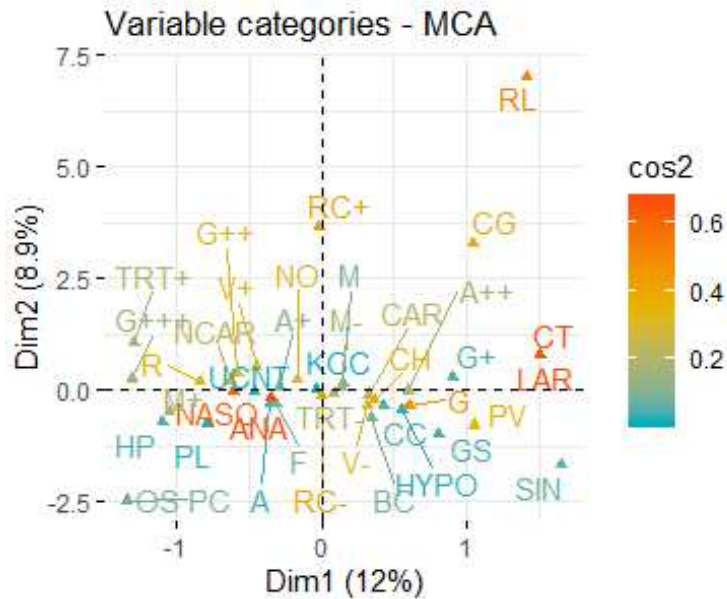
axe1
```

8.2. PARTIE ANALYSE

```
> axis1
      coord1 contrib1 cos*1
A          -0.4         0.4    0.0
A+         -0.3         1.5    0.1
A++        -0.6         3.7    0.2
F          -0.3         1.0    0.0
M           0.1         0.5    0.0
CC          0.4         0.6    0.0
KCC         0.0         0.0    0.0
PC         -1.3         0.6    0.0
RL          1.4         0.6    0.0
UCNT        -0.5         0.3    0.0
ANA        -0.4         3.1    0.5
CT          1.5        13.3    0.5
BC          0.3         0.5    0.0
GS          0.0         0.4    0.0
HYPO        1.6         0.3    0.0
LAR         1.5        13.3    0.5
NASO       -1.6         7.2    0.6
SIN         1.7         0.8    0.0
CAR         0.3         2.2    0.2
NCAR        0.6         4.2    0.2
G           0.0         5.6    0.4
G+          0.9         1.0    0.0
G++         0.6         4.1    0.2
G+++        0.3         4.2    0.2
CG          1.0         1.0    0.0
CH          0.4         2.7    0.2
R           0.8         7.0    0.3
M-          0.1         0.5    0.1
M+         -1.1         3.7    0.1
RC-          0.0         0.0    0.0
RC+          0.0         0.0    0.0
TRT-        0.1         0.2    0.1
TRT+       -1.3         3.1    0.1
HP          1.1         0.7    0.0
NO          0.2         0.7    0.1
OS          0.3         0.6    0.0
PL          0.8         0.2    0.0
PV          1.1         5.8    0.2
V-          0.3         1.8    0.1
V+         -0.6         2.6    0.1
```

On constate sur l'axe 1 que les contributions varient de 0,00% à 13,30%.

```
fviz_mca_var(mca, repel = TRUE, col.var = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



CHAPITRE 8. APPLICATION 1 : UNE DES MÉTHODES PROBABILISTE AVEC UN ACM SUR R

On sait que les modalités contribuant le plus à la formation de l'axe 1 sont celles dont les cosinus sur cet axe sont proches de 1 en valeur absolue.

les modalités les plus excentrés ne sont pas nécessairement celles qui contribuent le plus sur l'axe 1, on remarque que :

Les modalités actives qui Contribuent (+)	Les modalités actives qui Contribuent (-)
CT ; LAR	NASO
PV	R
G	G+++
A++	NCAR
	G++
	M+
	TRT+
	ANA

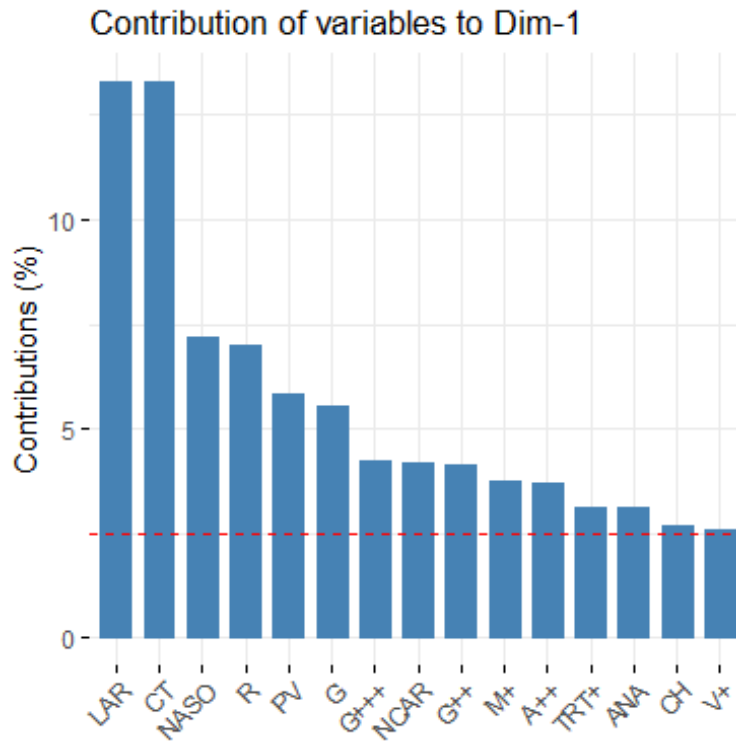
Donc on a :

Les modalités LAR (larynx) ; et CT (cytologie) ont la plus forte contribution positivement avec un taux de 13.30%.

De plus, la modalité NASO (nasopharynx) a la plus forte contribution négative avec un taux de 7.18%.

L'histogramme suivant confirme nos résultats :

```
fviz_contrib(mca, choice = "var", axes = 1, top = 15)
```



La qualité de représentation des modalités sur l'axe 1 D'après le tableau de « la qualité de représentation des modalités » obtenu par la commande :

```
round(M$var$cos2[,1],1)
> round(M$var$cos2[,1],1)
```

A	A+	A++	F	M	CC	KCC	PC	RL	UCNT	ANA
0.0	0.1	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.5
CT	BC	GS	HYPO	LAR	NASO	SIN	CAR	NCAR	G	G+
0.5	0.0	0.0	0.0	0.5	0.6	0.0	0.2	0.2	0.4	0.0
G++	G+++	CG	CH	R	M-	M+	RC-	RC+	TRT-	TRT+
0.2	0.1	0.0	0.2	0.3	0.1	0.1	0.0	0.0	0.1	0.1
HP	NO	OS	PL	PV	V-	V+				
0.0	0.1	0.0	0.0	0.2	0.1	0.1				

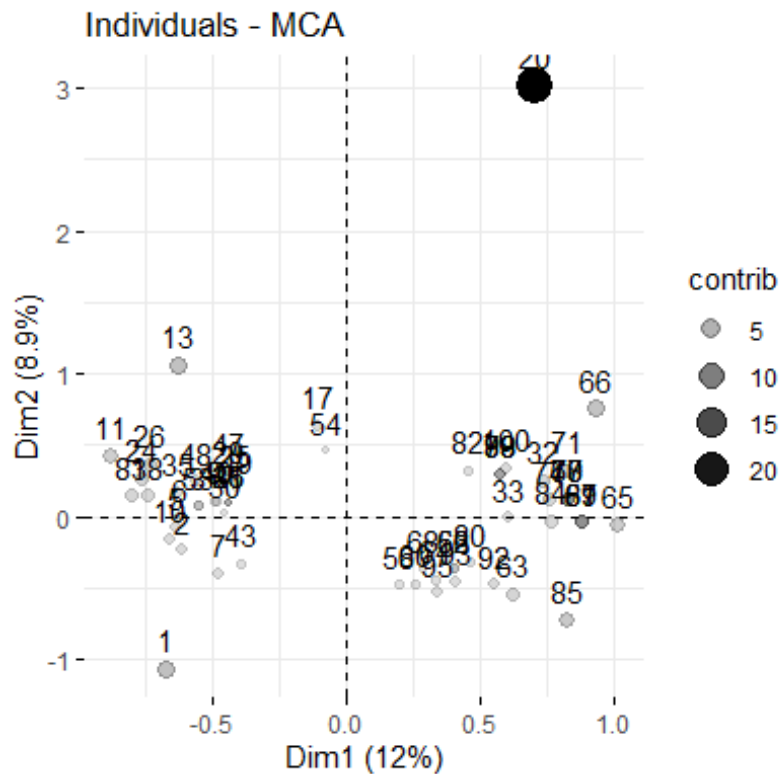
On remarque que :

- Les modalités NASO , LAR, ANA, CT sont moyennement bien représentées sur l'Axe1.
- Toutes les autres modalités actives sont plutôt mal représentés sur

l'Axe1.

L'axe 1 Profile ligne (individus)

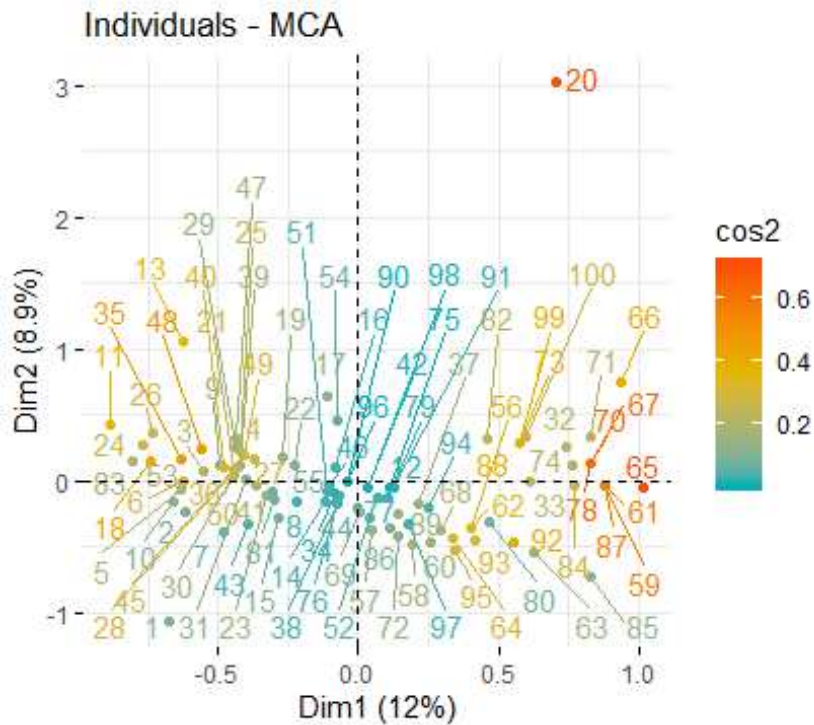
```
fviz_pca_ind(M, pointsize = "contrib", alpha.ind = "contrib", select.ind
= list(contrib = 40))
```



On remarque que les individus excentrés sont les mieux représentés alors que les individus proches du centre ont une mauvaise qualité de représentation.

A chaque fois qu'on s'éloigne du centre on voit une augmentation dans la qualité de représentation.

```
fviz_mca_ind(mca, repel = TRUE, col.ind = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



On remarquons que :

Les individus qui contribuent (-)	Les individus qui contribuent (+)
86,26 ,24 ,11	85 ,65,74 ,87,70,71,74,78,66,61,59

La qualité de représentation des individus sur l'axe 1 D'après le tableau de « la qualité de représentation des individus » obtenu par la commande :

```
round(M$ind$cos2[,1],1)
```

```
> round(M$ind$cos2[,1],1)
  1   2   3   4   5   6   7   8   9  10  11  12  13
0.0 0.1 0.2 0.2 0.2 0.3 0.0 0.0 0.1 0.1 0.3 0.0 0.1
 14  15  16  17  18  19  20  21  22  23  24  25  26
0.0 0.0 0.0 0.0 0.4 0.1 0.0 0.3 0.1 0.1 0.2 0.2 0.2
 27  28  29  30  31  32  33  34  35  36  37  38  39
0.2 0.3 0.1 0.2 0.1 0.2 0.2 0.0 0.4 0.3 0.1 0.0 0.1
 40  41  42  43  44  45  46  47  48  49  50  51  52
0.3 0.2 0.0 0.0 0.0 0.3 0.0 0.1 0.4 0.3 0.2 0.0 0.0
 53  54  55  56  57  58  59  60  61  62  63  64  65
0.2 0.0 0.1 0.2 0.0 0.0 0.6 0.0 0.6 0.2 0.1 0.1 0.7
 66  67  68  69  70  71  72  73  74  75  76  77  78
0.2 0.6 0.1 0.0 0.6 0.2 0.0 0.3 0.2 0.0 0.0 0.0 0.6
 79  80  81  82  83  84  85  86  87  88  89  90  91
0.0 0.1 0.1 0.2 0.2 0.3 0.1 0.0 0.6 0.3 0.0 0.0 0.0
 92  93  94  95  96  97  98  99 100
0.2 0.1 0.0 0.1 0.0 0.0 0.0 0.3 0.2
```

On remarque que :

- Les individus 59,61,65,67,70,78 sont moyennement bien représenté sur l'Axe1.
- Toutes les autres modalités actives sont plutôt mal représentées sur l'Axe1.

Conclusion On conclure que :

- Les individus possédant les modalités (actives) biens représentées du côté positif de l'axe 1, ont plus de chance de posséder des modalités comme «CT, LAR, PV, G » et les individus bien représentés du côté négatif, ont plus de chance de posséder des modalités comme «NASO, G+++, ANA»
- Le premier axe distingue par les pients qui sont détecté le type de maladie par le citologie CT ,ils ont été subute par la même maladie LAR , ce qui les a fait perdre de vue PV et a été découvert au premier grad d'histologie G et leurs tumeurs ne récidive pas et aussi les patients qui ont été subite par la meme maladie NASO été découvert au dèrnier grade histologuie G+++,ils sont détecté la maladie pa l'analyse anapathe ANA.

L'axe 2 Profile colonne (modalités)

```
fviz_pca_var(M, pointsize = "contrib", alpha.var = "contrib", select.var = list(contrib = 40))
```

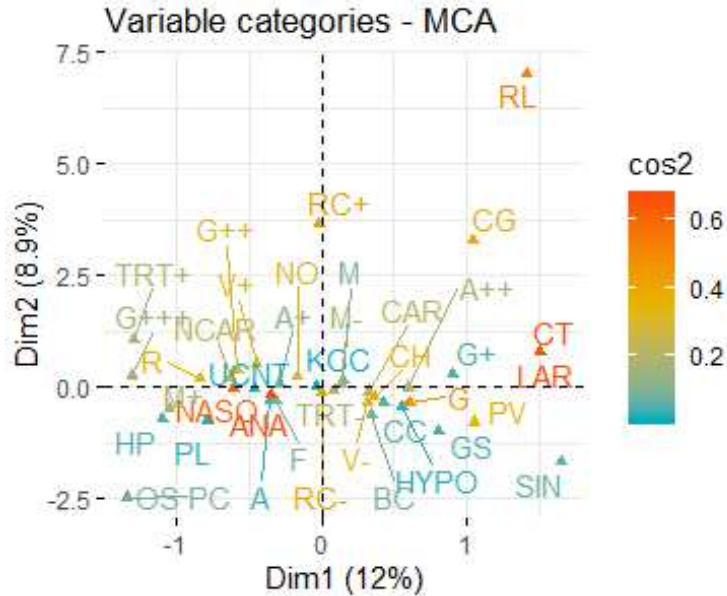

CHAPITRE 8. APPLICATION 1 : UNE DES MÉTHODES PROBABILISTE AVEC UN ACM SUR R

```
> axe2
      coord2 contrib2 cos*2
A      -0.3      0.4      0.0
A+     -0.1      0.1      0.0
A++    -0.0      0.0      0.0
F      -0.3      1.2      0.0
M      -0.1      0.6      0.0
CC     -0.4      0.6      0.0
KCC    -0.0      0.0      0.0
PC     -2.5      2.5      0.1
RL     -0.0      2.0      0.5
UCNT   -0.0      0.0      0.0
ANA    -0.2      1.1      0.2
CT     -0.8      4.4      0.2
BC     -0.6      2.2      0.2
GS     -1.0      0.0      0.0
HYPO   -0.4      0.5      0.0
LAR     0.5      4.5      0.1
NASO   -0.0      0.0      0.0
SIN    -1.7      1.2      0.0
CAR    -0.1      0.3      0.0
NCAR   -0.2      0.6      0.0
G      -0.3      2.5      0.1
G+     -0.3      0.1      0.0
G++    -0.4      2.1      0.1
G+++   -0.3      0.2      0.0
CG     -3.3      13.6      0.3
CH     -0.2      1.6      0.1
R      -0.2      0.4      0.0
M-     -0.1      0.1      0.0
M+     -0.5      1.1      0.0
RC-    -0.1      0.5      0.4
RC+    -3.7      16.7      0.4
TRT-   -0.1      0.2      0.1
TRT+   -1.0      2.7      0.1
HP     -0.7      0.5      0.0
NO     -0.2      1.7      0.2
OS     -2.5      2.5      0.1
PL     -0.8      0.2      0.0
PV     -0.8      4.4      0.1
V-     -0.4      3.0      0.2
V+      0.5      4.3      0.2
```

On constate sur l'axe 2 que les contributions varient de 0,00% à 20%.

Voire le graphe suivant :

```
fviz_mca_var(mca, repel = TRUE, col.var = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



D'après la table de contribution et les graphes précédents, on obtient le tableau suivant :

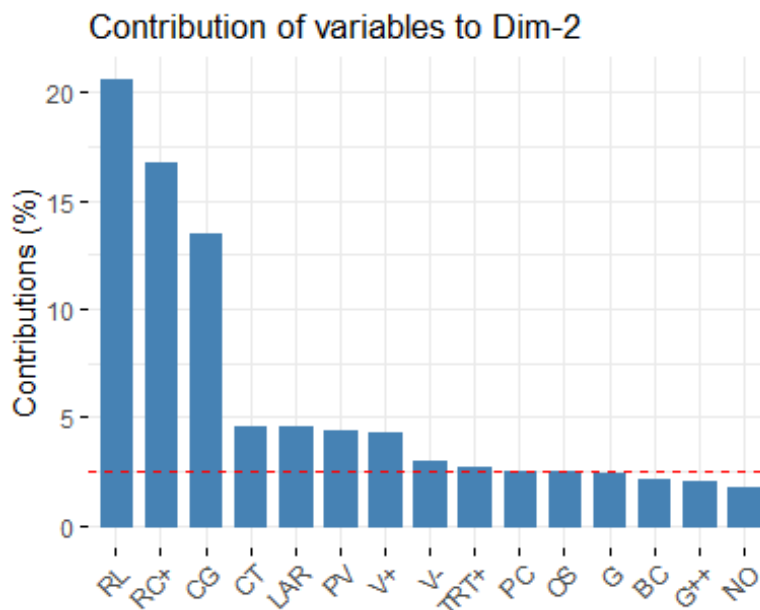
Les modalités actives qui Contribuent (+)	Les modalités actives qui Contribuent (-)
RL	PV
RC+	V-
CG	OS
CT ; LAR	
V+	
TRT+	

Remarquant que :

- La modalité RL (récidive larynx) a la plus forte contribution positive avec un taux de 20.56%.
- La modalité PV (perde de vu) a la plus forte contribution négative avec un taux de 4.43%.

L'histogramme suivant confirme nos résultats :

fviz_contrib(mca, choice = "var", axes = 2, top = 15)



La qualité de représentation des modalités sur l'axe 2 D'après le tableau de « la qualité de représentation des modalités » obtenu par la commande :

```
round(M$var$cos2[,2],1)
```

```
> round(M$var$cos2[,2],1)
```

A	A+	A++	F	M	CC	KCC	PC	RL	UCNT
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.0
ANA	CT	BC	GS	HYPO	LAR	NASO	SIN	CAR	NCAR
0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0
G	G+	G++	G+++	CG	CH	R	M-	M+	RC-
0.1	0.0	0.1	0.0	0.3	0.1	0.0	0.0	0.0	0.4
RC+	TRT-	TRT+	HP	NO	OS	PL	PV	V-	V+
0.4	0.1	0.1	0.0	0.2	0.1	0.0	0.1	0.2	0.2

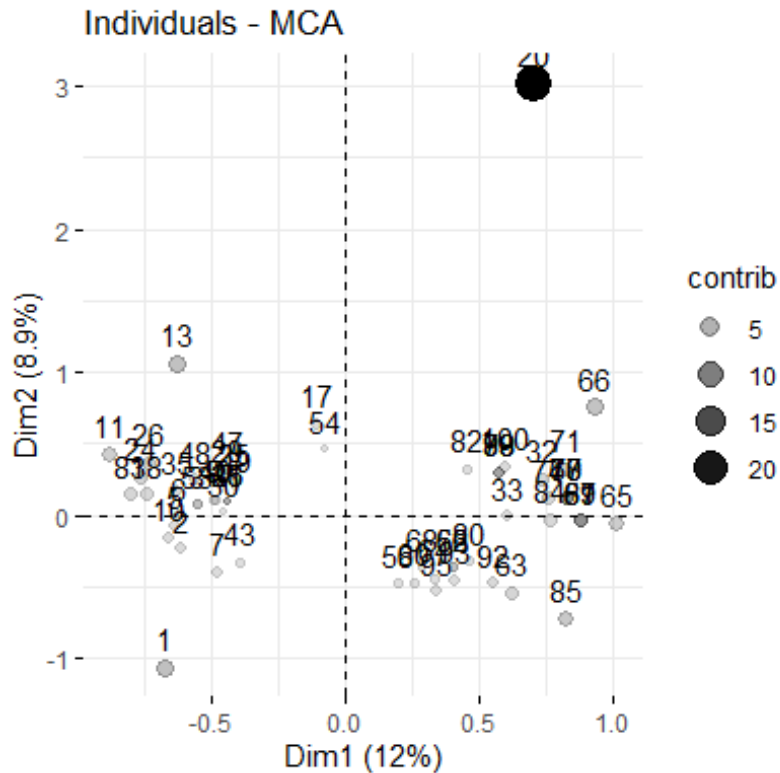
On remarque que :

- La modalité 20 est moyennement bien représenté sur l'Axe2.
- Toutes les autres modalités actives sont plutôt mal représentées sur l'Axe2.

(remarque : Notons que les modalités qui ont un $\cos^2$ proche de 1 ont des grandes contributinon dans les axes).

L'axe 2 Profile colonne (individus)

```
fviz_pca_ind(M, pointsize = "contrib", alpha.ind = "contrib", select.ind
= list(contrib = 40))
```



A chaque fois qu'on s'éloigne du centre on voit une augmentation dans la qualité de représentation.

On remarquons que :

Les individus qui contribuent (-)	Les individus qui contribuent (+)
1,85	13,20,17,66

La qualité de représentation des individus sur l'axe 2 D'après le tableau de « la qualité de représentation des individus » obtenu par la commande :

```
round(M$ind$cos2[,2],1)
```

On remarque que :

- L'individu 20 est moyennement bien représenté sur l'Axe2.

CHAPITRE 8. APPLICATION 1 : UNE DES MÉTHODES PROBABILISTE AVEC UN ACM SUR R

- Toutes les autres modalités actives sont plutôt mal représentées sur l'Axe2.

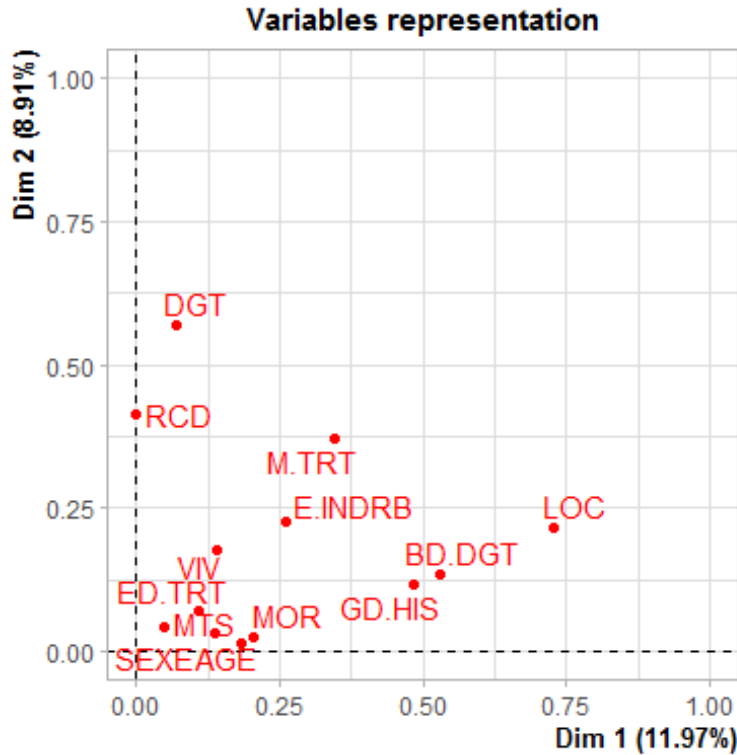
```
> round(M$ind$cos2[,1],1)
  1   2   3   4   5   6   7   8   9  10  11  12  13
0.0 0.1 0.2 0.2 0.2 0.3 0.0 0.0 0.1 0.1 0.3 0.0 0.1
 14  15  16  17  18  19  20  21  22  23  24  25  26
0.0 0.0 0.0 0.0 0.4 0.1 0.0 0.3 0.1 0.1 0.2 0.2 0.2
 27  28  29  30  31  32  33  34  35  36  37  38  39
0.2 0.3 0.1 0.2 0.1 0.2 0.2 0.0 0.4 0.3 0.1 0.0 0.1
 40  41  42  43  44  45  46  47  48  49  50  51  52
0.3 0.2 0.0 0.0 0.0 0.3 0.0 0.1 0.4 0.3 0.2 0.0 0.0
 53  54  55  56  57  58  59  60  61  62  63  64  65
0.2 0.0 0.1 0.2 0.0 0.0 0.6 0.0 0.6 0.2 0.1 0.1 0.7
 66  67  68  69  70  71  72  73  74  75  76  77  78
0.2 0.6 0.1 0.0 0.6 0.2 0.0 0.3 0.2 0.0 0.0 0.0 0.6
 79  80  81  82  83  84  85  86  87  88  89  90  91
0.0 0.1 0.1 0.2 0.2 0.3 0.1 0.0 0.6 0.3 0.0 0.0 0.0
 92  93  94  95  96  97  98  99 100
0.2 0.1 0.0 0.1 0.0 0.0 0.0 0.3 0.2
```

Conclusion On conclure que :

- Le deusieme axe distingue par le patient qui détecte lamaladie par le citologie CT, il a été traité par chirurgie CG et blésser une autre fois à la maladie RC+.

- Les pients qui sont dignostiquée des tumeus localisé sur larynx LAR . Il sont détecté le type de maladie par le citologie CT ,ils ont été blessés par la même maladie Récidive larynx RL , ce qui les a fait perdre de vue PV. Maleureusement, ces patients sont morts V-.

Interprétation des variables pour interprété les variables on voir ce graphe :



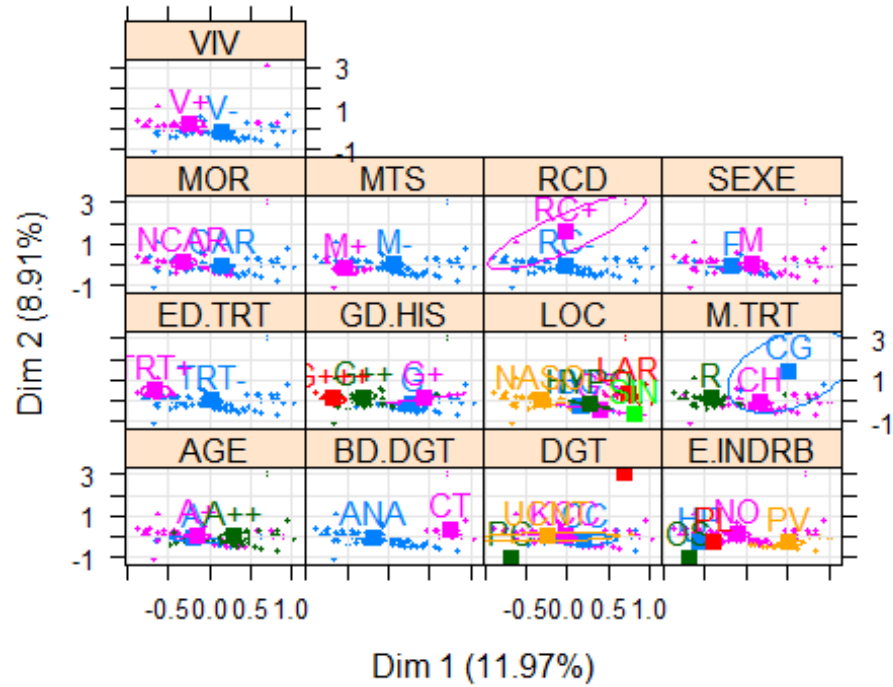
- Les variables LOC, BD-DGT , GD-HID ont une bonne corrélation avec le premier axe.
- Les variables DGT, RCD, M.TRT sont bien corrélées avec le second axe.
- Les autres variables sont quant à elle pas corrélé avec les deux axes puisqu'elles ne participent pas du tout à la construction des axes, ce qui expliquera ce faible rapport de corrélation.

La réparation des individus en fonction des modalités

La fonction `plotellipses` colore les individus et trace des ellipses de confiance autour des modalités de variables qualitatives. L'objectif est de voir si les modalités d'une variable qualitative sont significativement différentes les unes des autres.

Pour cela nous téléchargeons les deux package (`ggplot2`) et (`factoextra`). Le résultat est le suivant :

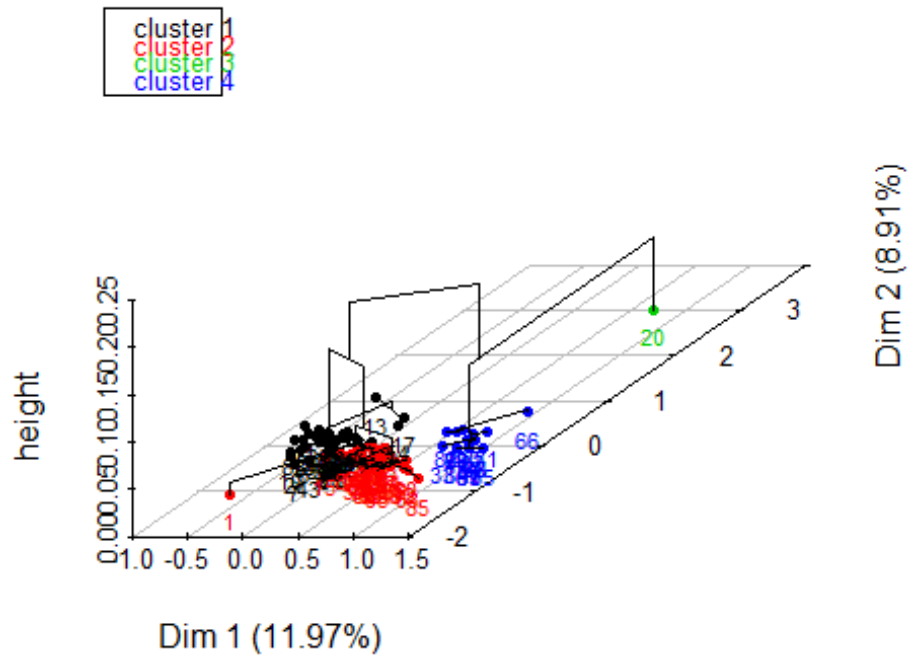
```
library(ggplot2)
library(factoextra)
plotellipses(M,keepvar=donnees)
```

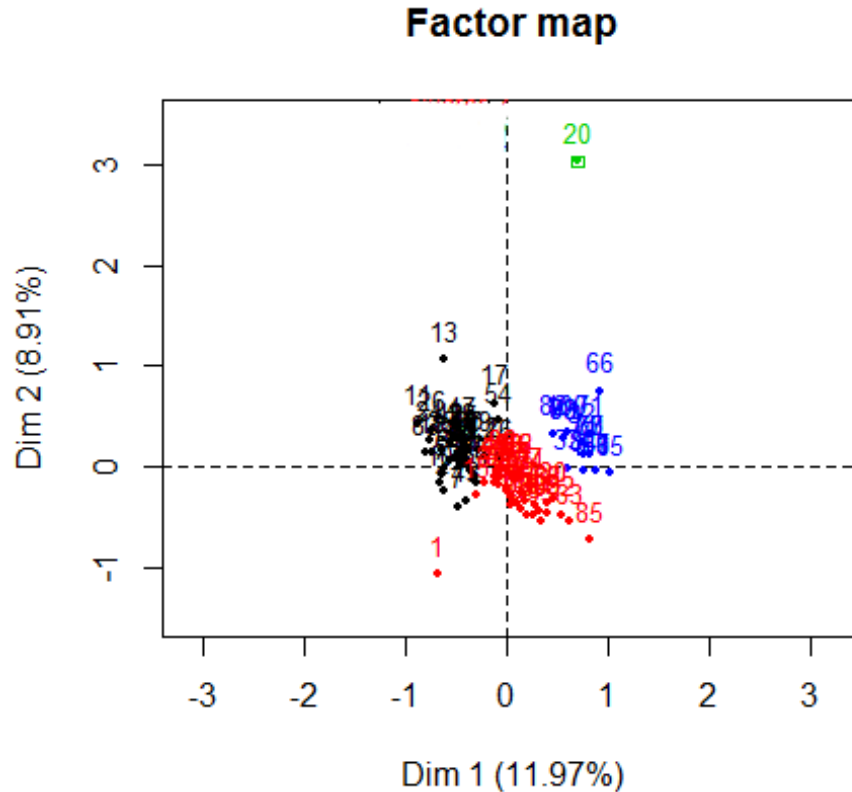


Classification

```
res.hcpc<-HCPC(M)
```

Hierarchical clustering on the factor map





Notre échantillon distingue entre 4 classes bien distinctes, chaque catégorie inclut l'ensemble des individus dont le profil est presque le même :

- La première catégorie (en vert) possède les patients qui détecte la maladie par le citologie CT, il a été traité par chirurgie CG et blessé une autre fois à la maladie RC+.

- La deuxième catégorie (en bleu) s'oppose avec la première ; les patients de cette classe âgé de 60 à 90 ans. Ils ont découvert la maladie à ses débuts G par le citologie CT, ils ont été souffrir de la maladie récursive larynx RL. Ces patients étaient blessés sur larynx LAR, ils sont traités par la chirurgie mais ils ont été blessés par la même maladie une autre fois, ce qui les a fait perdre de vue PV. Heureusement, ces patients vivent et V+ et continue de traiter TRT+.

- La troisième catégorie (en noir) représente un groupe de patients qui découvre la maladie dans ses deuxième et troisième phases par l'anapathie ANA. Ces patients étaient blessés par la maladie NASO, leurs tumeurs sont malignes NCAR. Malgré ils sont traités par la radiothérapie, leurs tumeurs

étaient propagée M+. Ils sont continué de traiter pour gérer TRT+.

- La quatrième catégorie (en rouge) représente les patients qui ont été blessé par la maladie NASO ou BC, ce qui les a fait perdre de vue PV ou OS. Maleureusement, ils sont morts V-.

Conclusion

dire que :

- Les catégories d'âge les plus touchées par le VADS sont les deux catégories de 10 ans à 30 ans et de 60 ans à 90 ans. La plupart d'entre eux sont des hommes, cela est dû au principal facteur de risque qui cause cette maladie qui est le tabagisme.

- Il n'y a pas de remède complet pour ce type de cancer, car les traitements existants aident les patients à vivre le plus longtemps possible seulement.

- Le stade du diagnostic est lié à la méthode de traitement utilisée. Par exemple, le traitement chirurgical est utilisé pour les patients qui sont diagnostiqué la maladie à ses débuts. Aussi le traitement par la radiothérapie est utilisé pour les patients qui sont diagnostiqué la maladie à ses dernières stades.

- La méthode de traitement dépend également de l'emplacement de la tumeur.

- Les symptômes du cancer de Vads n'apparaissent pas aux premiers stades, ils sont donc souvent détectés aux stades ultérieurs.

- Les patients atteints de tumeurs bénignes vivent longtemps avec la maladie et ne se traitent pas d'autre fois.

- Les patients atteints de tumeurs malignes ne vivent pas longtemps avec la maladie parce que la tumeur propagée ; à chaque fois que cette tumeur se propage, ils sont traités à nouveau.

- Bien que les cancers de vads soient liés au système respiratoire, il affecte beaucoup plus la vue.

Chapitre 9

Application 2 : Une des Méthodes non probabiliste avec un ACM sur R

L'étendue de la sensibilisation du peuple algérien pour affronter la pandémie du COVID-19.

9.1 La démarche de notre travail

9.1.1 Introduction

Les virus Corona sont un large éventail de virus qui peuvent provoquer des maladies chez les animaux et les humains. On sait qu'un certain nombre de virus corona provoquent des maladies respiratoires humaines d'une gravité allant du rhume à des maladies plus graves telles que le syndrome respiratoire du Moyen-Orient (MERS) et le syndrome respiratoire aigu sévère (SRAS). Le virus Corona nouvellement découvert provoque la maladie Covid-19.

La maladie Covid-19 est une maladie infectieuse causée par le dernier virus détecté du Coronavirus. Il n'y avait aucune connaissance de l'existence et de la maladie de ce nouveau virus avant son épidémie à Wuhan, en Chine, en décembre 2019. Covid-19 est maintenant devenu une pandémie affectant de nombreux pays du monde.

Alors que la nouvelle épidémie de Corona (Covid 19) continue sa propagation mondiale avec plus d'un demi-million de blessés et plus de vingt mille morts, et alors que les gouvernements du monde entier exploitent tous leurs efforts et capacités pour arrêter la propagation dévastatrice de cette épidémie, les yeux sont concentrés et les paris sont principalement tournés autour de la prise de conscience des sociétés et des peuples.

La gravité de cette épidémie et la participation active aux efforts pour y faire face en tant que facteur décisif et le plus influent pour combattre et vaincre cette épidémie, à travers l'engagement strict des peuples et des groupes de la société aux mesures préventives prises par les gouvernements, Surtout en ce qui concerne la restriction des mouvements, l'isolement sanitaire, le maintien à domicile et la prévention des rassemblements et des mélanges entre les gens, car c'est la mesure défensive la plus importante à ce jour pour empêcher la propagation de l'épidémie jusqu'à ce que les traitements nécessaires soient développés pour l'éliminer.

La sensibilisation de la communauté n'est pas seulement un engagement envers toutes les procédures, instructions et instructions, mais aussi dans un sens profond de responsabilité envers le pays et de soutien aux efforts du gouvernement pour traverser cette crise, par un engagement à la distance physique, et en ne quittant la maison que dans les limites les plus étroites et par nécessité absolue, et en ne prêtant aucune attention à Nouvelles ou informations peu fiables, en particulier celles diffusées sur les médias sociaux et les plateformes, et faisant la promotion de fausses informations sur Corona, que ce soit en relation avec sa récession mondiale ou sa gravité ; Parce que cela nuirait aux grands efforts du gouvernement face à cette épidémie.

9.1.2 Démarches d'enquête

Le type d'enquête

C'est une enquête périodique.

La méthode de recueil les données

La méthode de recueil de données pertinentes : un sondage sur internet (C'est un questionnaire fermé car les réponses sont fixées à l'avance et le répondant doit choisir parmi l'éventail qui lui est présente.)

Les étapes de la réalisation du questionnaire

- L'objet de l'enquête :
 - L'étendue La sensibilisation du peuple algérien pour affronter la pandémie du COVID-19.
 - Le peuple algérien contribue-t-il à la propagation de ce virus ? Ou est-elle due à d'autres facteurs ?
 - Quel est L'effet de la pandémie du COVID-19 sur le peuple algérien ?
 - Les hypothèses de l'enquête :
 - Le peuple algérien est conscient lorsqu'il est rencontré la pandémie du COVID-19.
 - Le peuple algérien n'est pas conscient lorsqu'il est rencontré la pandémie du COVID-19.
 - Les moyens matériels mis à la disposition des réalisateurs de l'enquête :
 - La durée de l'enquête : 7jours.
 - Le budget alloué à l'enquête : la facture d'internet / gratuit.
 - Nous sommes collectés les questions remplis du questionnaire, puis nous les partageons sur les médias sociaux et surveillons les réponses.
 - Recherches préalables : nous faisons une petite recherche sur notre problème (voir introduction).
 - Que veut-on montrer ?
 - Nous voulons voir si le peuple est consciente ou non pendant la période de la pandémie du COVID-19.
 - La population ciblée ou l'univers de l'enquête : le peuple algérien.
 - La construction de l'échantillon :
 - Taille de notre échantillon : 266.
 - L'échantillonnage : échantillonnage par jugement.
 - La rédaction du projet du questionnaire
 - Le choix de la méthode d'administration : par internet.
- Nous préparons le questionnaire. Puis, nous le publions sur les réseaux sociaux. Et lorsque l'enquêté répond seul au questionnaire, nous avons continué à surveiller et à suivre les réponses dans Google Drive.
- Choix des questions : nous respectons les instructions, Les étapes et la Rédaction du questionnaire (L'introduction, les types, La structuration, et Les biais à éviter) qui mentionné dans la partie théorique.
 - Mise en forme : Nous avons essayé de mettre en forme les questions en fonction du sujet à l'étude Aussi, nous avons essayé de déterminer toutes les

réponses possibles pour chaque question, afin que les répondants du notre questionnaire puisse choisir la réponse qui lui convient.

- La mise à l'épreuve du questionnaire (pré-test ou enquête pilote) :
 - Il s'agit d'une phase fondamentale, souvent négligée, qui consiste à mettre à l'épreuve le questionnaire par rapport à quelques individus, autrement dit à le tester. Elle est donc centrée sur l'évaluation du questionnaire lui-même. L'approche est ici plus qualitative que quantitative. Il s'agit d'évaluer la clarté et la précision des termes utilisés et des questions posées, la forme des questions, l'ordre des questions, l'efficacité de la mise en page, éliminer toutes les questions ambiguës ou refusées, repérer les omissions, voir si le questionnaire est jugé trop long, ennuyeux, indiscret, etc. . .
- La Rédaction du questionnaire définitif :
 - La mise en page : Selon toutes les règles mentionnées dans la partie théorique.
 - La pré codage : Nous avons encodé toutes les questions et réponses que nous avons définies précédemment (Comme indiqué dans ce qui suit).
- La formation des enquêteurs :
 - Nous avons cherché des moyens d'administrer le questionnaire, et examiné des nombreux questionnaires et nous avons essayé aussi de respecter au maximum leurs règles qui nous l'avons trouvé.

Le questionnaire

Madame, Monsieur, nous sommes des étudiantes en statistique appliqué à L'université Frères Mentouri Constantine -1 Faculté des Sciences exactes, nous réalisons un mémoire de fin d'études sur les méthodes d'échantillonnages auxiliaires et applications.

Dans ce cadre, nous vous remercions de bien vouloir consacrer quelques minutes pour répondre au questionnaire ci-joint.

Vos réponses sont anonymes.

1/- Croyez-vous que virus corona (covid-19) existe ? : encodé (EXISTE)

- Oui : encodé (E+).

- Non : encodé (E-).

2/- Dans quelle mesure êtes-vous d'accord avec les informations de la santé qui provenant des différentes sources (télévision, médias sociaux) ? : encodé (INF-SM)

- Vous l'Acceptez : encodé (INF+).

- Vous ne l'acceptez pas : encodé (INF-).

- Vous vérifiez d'abord sa vérité : encodé (INF).
- 3/- Où obtenez-vous les dernières Développements de virus ? : encodé (DR-DIV).
 - Sites des réseaux sociaux : encodé (DV-).
 - Chaînes de télévision : encodé (DV-).
 - Communauté (amis, voisins) : encodé (DV+).
 - Vous n'êtes pas intéressés par les développements quotidiens : encodé (DV++).
- 4/- Les opinions de vus des gens (pensées, croyances) sur le virus affectent-elles le reste des gens ? : encodé (OP-VU)
 - Oui : encodé (O+).
 - Non : encodé (O-).
- 5/- Pensez-vous des mesures préventives pour éviter l'infection par le corona virus ? : encodé (EV-CRN).
 - Oui : encodé (EV+).
 - Non : encodé (EV-).
 - Parfois : encodé (EV).
- 6/- Avez-vous respecté toute la période de quarantaine et êtes-vous resté chez vous ? : encodé (RSP).
 - Oui : esncodé (R+).
 - Non : encodé (R-).
 - Sortez uniquement lorsque cela est nécessaire : encodé (R).
- 7/- Que faites-vous lorsque vous toussez ou surchauffé ? : encodé (CH).
 - Vous vous asseyez à la maison vous et vous isolez : encodé (C+).
 - Vous allez à l'hôpital : encodé (C-).
 - Vous n'inquiétez pas et complétez les activités de la vie quotidienne : encodé (C).
- 8/- Lorsque vous soupçonnez que quelqu'un a le virus, que faites-vous ? : encodé (SUP).
 - Vous appelez la protection civile : encodé (S+).
 - Ne prenez que des précautions : encodé (S-).
 - Vous s'en moquez : encodé (S).
- 9/- Dans quelle mesure pensez-vous que le peuple algérien s'engage à prendre les précautions nécessaires ? : encodé (PRE-NS).
 - Un petit peu : encodé (NS-).
 - Un peu : encodé (NS).
 - Beaucoup : encodé (NS+).

10/- À votre avis, quelle est la meilleure façon de contrôler le virus ? : encodé (MF+).

- Le Quarantaine : encodé (MF).

- Politique de dispersion des troupeaux (distance sociale) : encodé (MF-).

11/- Souffrez-vous d'une maladie chronique ou d'une immunité affaiblie ? : encodé (M-FC).

- Ou i : encodé (MC+).

- Non : encodé (MC-).

12/- Avez-vous peur d'être infecté par le virus ? : encodé (IFC-CRN).

- Oui toujours : encodé (IF+).

- Non : encodé (IF-).

- Parfois : encodé (IF).

13/- Avez-vous ressenti des troubles ou de la peur lorsque le virus a commencé à se propager ? : encodé (TRB).

- Oui : encodé (T+).

- Non : encodé (T-).

- Parfois : encodé (T).

14/- À votre avis, quels sont les effets négatifs causés par la longue survie du Corona virus et sa propagation rapide ? : encodé (NG).

- Dépression : encodé (NG- -).

- Le besoin Surtout pour les personnes à faible revenu : encodé (NG+).

- Augmentation du prix : encodé (NG++).

15/- À votre avis, quelles sont les raisons de l'augmentation de la propagation du virus ? : encodé (RS-AUG).

- L'indifférence : encodé (RS).

- L'ignorance : encodé (RS-).

16/- Quel est votre lieu de résidence ? : encodé (LR).

- Village : encodé (LR+).

- Ville : encodé (LR-).

17/- Quel est votre travail ? : encodé (TRV).

- Employé du gouvernement : encodé (TR+).

- Au chômage : encodé (TR-).

- Affaires Libre : encodé (TR).

18/- Quel est votre niveau d'études ? : encodé (NV-ED).

- Les personnes avec un niveau moyen au plus : encodé (ED).

- Les personnes avec un niveau universitaires au plus : encodé (ED+).

19/- Quel est votre statut matrimonial ? : encodé (SM).

- Marié : encodé (Ma).

- Célibataire : encodé (Cé).
- 20/- Quel est votre âge ? : encodé (AGE).
- [18 :38[: encodé (A).
- [38 :58[: encodé (A+).
- 58 et plus : encodé (A++).
- 21/- Quel est votre sexe ? : encodé (SEXE).
- Femme : encodé (F).
- Homme : encodé (M).

9.1.3 Echantillonnage

Le plan d'échantillonnage

Le statisticien peut travailler au plan de sondage, qui comporte trois parties :

- Le plan d'échantillonnage : la façon dont on prélèvera l'échantillon.
- Les techniques d'estimation : analyse correspondantes multiple (ACM).
- Les mesures de la précision : la façon dont on mesurera l'erreur d'échantillonnage.

Nous étudierons le plan d'échantillonnage. Les étapes suivantes amènent à définir complètement le plan d'échantillonnage :

1. La problématique et des hypothèses :
 - Le peuple algérien est conscient lorsqu'il est rencontré la pandémie du COVID-19.
 - Le peuple algérien n'est pas conscient lorsqu'il est rencontré la pandémie du COVID-19.
2. le type d'étude : qualitative.
3. l'outil le plus adéquat : questionnaire.
4. Déterminer ce que sera la population observée : peuple algérien.
5. Définir un échantillon efficace : un questionnaire ou un sondage demanderons un échantillon large alors qu'un focus group ou entretien beaucoup moins.
6. une méthode d'échantillonnage : méthode par jugements.
7. Cadrer votre échantillonnage : définissez un début et une fin à votre étude (saturation théorique).

Méthode d'échantillonnage

Echantillonnage par jugement

Nous utilisons la méthode non probabiliste d'échantillonnage « échantillonnage par jugement » pour faire une analyse de correspondance multiple (ACM) sur notre jeu de donnée.

Pour cette méthode, l'échantillonnage est réalisé sur la base de certains jugements et d'idées préalables sur l'ensemble de la population. D'après nos connaissances, nos précédents jugements sont :

- Le peuple algérien est conscience pour affronter la pandémie du COVID-19.

- Nous pouvons accès au peuple algérien par les médias réseaux.

D'abord, nous présumons que la plus par du peuple algérien utilise les médias sociaux, alors nous jugeons que les unités qui constitueront l'échantillon sont représentatives de la population (le peuple algérien conscient).

Par nos connaissances sur le peuple algérien nous sélectionnons les "unités considérées caractéristiques" du phénomène à l'étude qui sont les gens du peuple algérienne qui répondent sur notre questionnaire dans les médias sociaux. Puis, nous Choisissons les individus basés sur nos jugements précédents. De plus, L'échantillonnage à l'intérieur de chacune des grandes unités se fait ensuite au hasard. Nous ne pouvons alors généraliser qu'aux unités choisies. Enfin, nous sélectionnerons les unités qui seront caractéristiques de la population.

Des biais importants peuvent découler du fait que nos idées préconçues, si elles sont inexactes, soient reflétées dans l'échantillon.

Détermination de la taille de l'échantillon

La formule donnant la taille minimale de l'échantillon est :

Rappel des notations :

- n est la taille de l'échantillon

- t coefficient de marge traduit du tableau de confiance

- p Proportion de la population qui représente une caractéristique donnée.

(Si p est inconnue, on utilise $p = 0.5$)

- e est la marge d'erreur choisie pour l'estimation.

On a :

Pour $e = 0.1$, et une proportion estimé à 50% soit $p = 0.50$, si on se fixe un seuil de confiance de $(1 - \alpha) = 95\%$, on se reporte à la table de la loi normale,

et on cherche la valeur correspondante à une probabilité $(1 - \alpha) = 0,95$, ce qui donne $t = 1.96$

La taille de l'échantillon sera donc d'au moins :

$$n = \frac{t^2 p(1-p)}{e^2} = \frac{1.96^2 \times 0.5 \times 0.5}{0.1^2} = 96,04$$

Notre échantillon doit contenir, donc, au moins 96 individus.

La distribution d'échantillonnage

Dans notre étude, nous disposant que de variables qualitatives nous permettant ainsi d'étudier les effectifs et les fréquences... Nous nous intéresseront donc exclusivement à la distribution d'échantillonnage de la proportion.

Soit $f_n = \frac{X_n}{n}$ la fréquence ou proportion des individus possédant, dans l'échantillon, le caractère étudié, ainsi chacun des échantillons possède une fréquence. Ces fréquences peuvent être considérées comme des valeurs observées d'une même variable aléatoire : $F_n = \frac{X_n}{n}$.

On sait que : $E(f_n) = p$ et $V(f_n) = \frac{pq}{n}$, tel que : p est La probabilité de la réalisation d'un événement, et $q = (1 - p)$ est La probabilité de la non réalisation d'un événement

Dans le cas d'une population d'effectif N (ici 40 million d'algérien), au sein de laquelle est prélevé, sans remise, un échantillon par jugement d'effectif n (ici 266), la variance de la fréquence est :

$$V(F_n) = V\left(\frac{X_n}{n}\right) = \frac{1}{n^2 V(X_n)} = \frac{1}{\frac{n^2(N-n)}{(N-1)npq}} = \frac{(N-n)}{\frac{(N-1)pq}{n}} \approx \frac{pq}{n}$$

Car

$$N - n = 40.000.000 - 250 \approx 40.000.000 - 1 \approx 40.000.000$$

Ainsi

$$\frac{(N-n)}{(N-1)} \approx 1$$

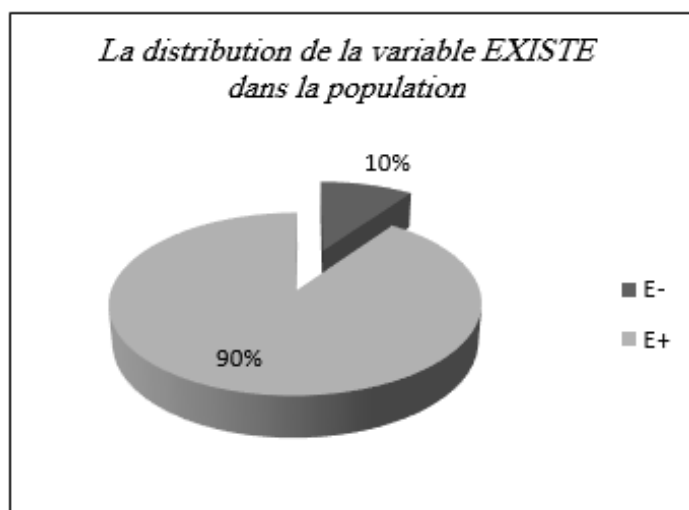
On peut donc affirmer que la distribution de la proportion suit une loi normale de moyenne p et d'écart type σ à condition que la taille de l'échantillon soit supérieure ou égale à 30 ($n \geq 30$) et le produit $np \geq 5$.

La population algérienne dépassant les 40 millions, on peut la considérer comme infini.

9.2.2 Analyse descriptive

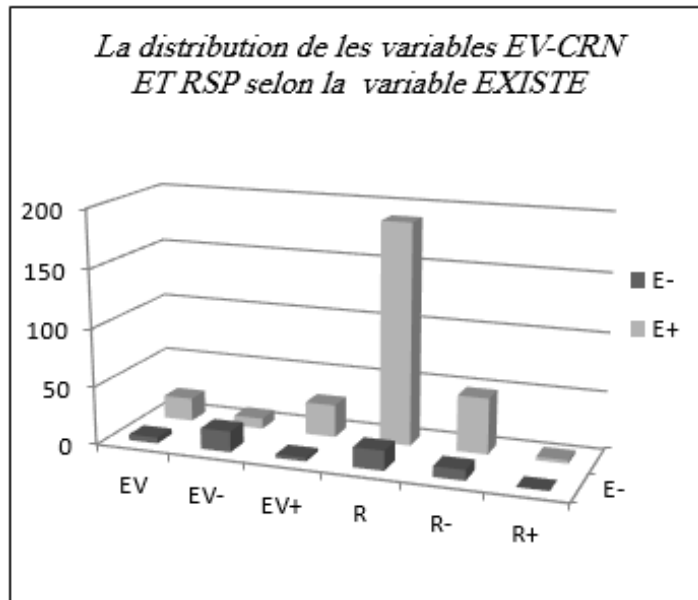
Description de la variable EXISTE

E-	E+
26	240



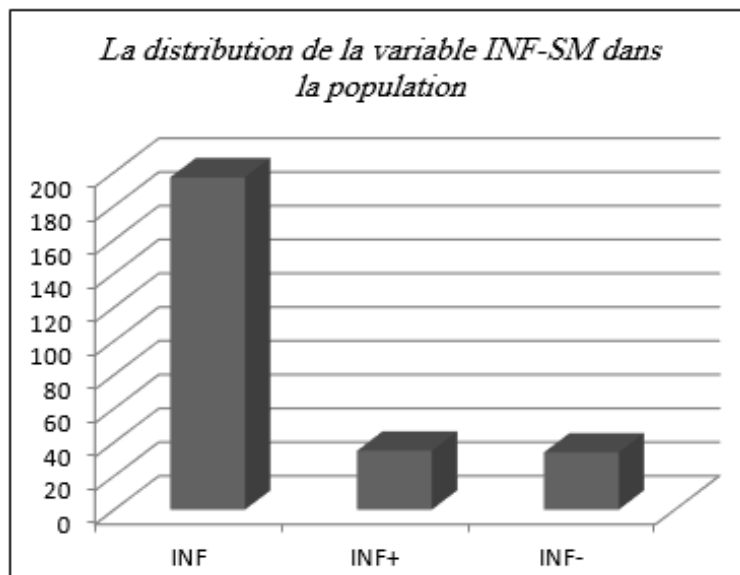
La distribution de les variables EV-CRN et RSP selon la variable EXISTE

	EV	EV-	EV+	R	R-	R+
E-	5	18	3	17	9	0
E+	20	9	28	189	48	3



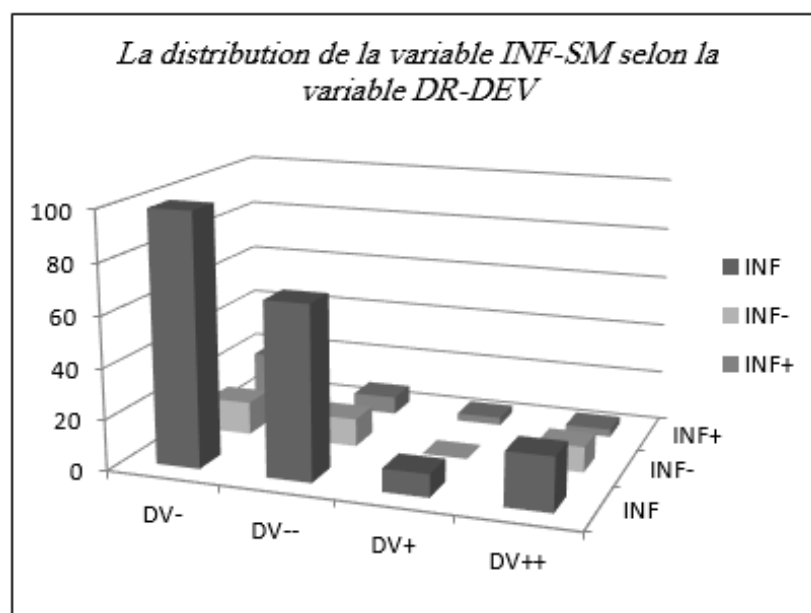
La distribution de la Variable INF-SM

INF	INF+	INF-
197	35	34



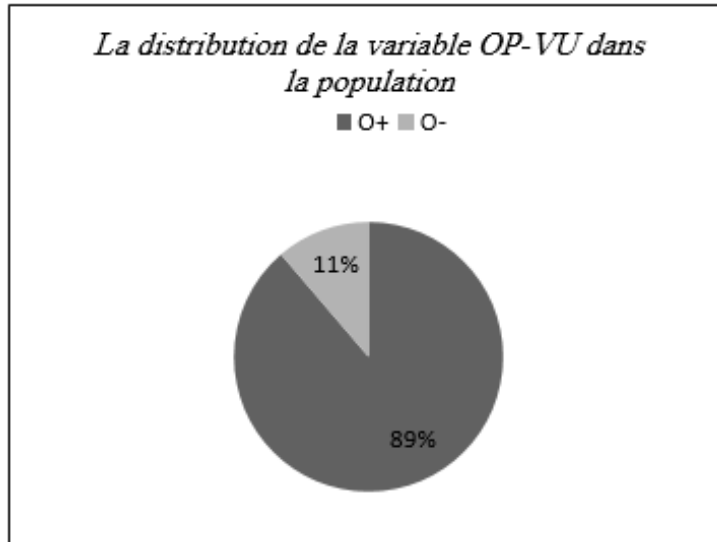
La distribution de la variable INF-SM selon la variable DR-DEV

	INF	INF-	INF+
DV-	99	13	20
DV-	68	11	7
DV+	9	0	3
DV++	21	10	3



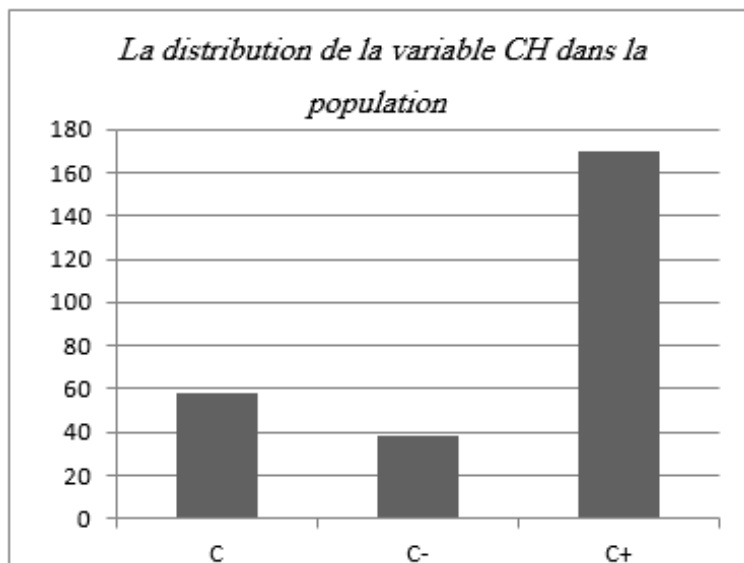
La distribution de la variable OP-VU dans la population

O+	O-
236	30



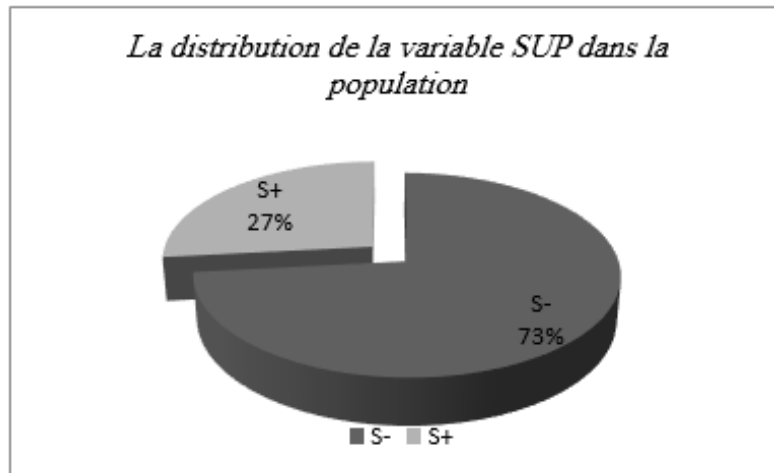
La distribution de la variable CH dans la population

C	C-	C+
58	38	170



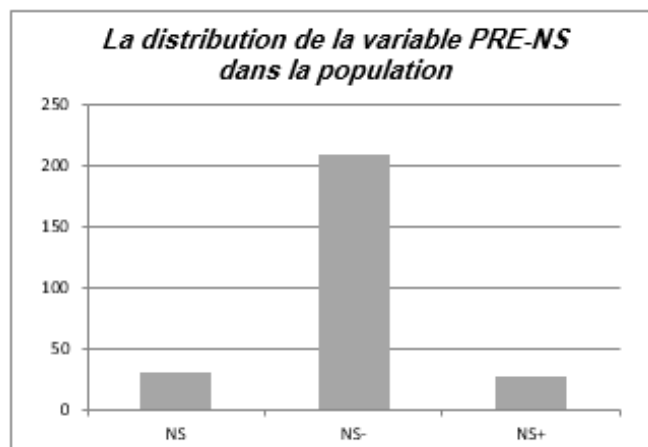
La distribution de la variable SUP dans la population

S-	S+
195	71



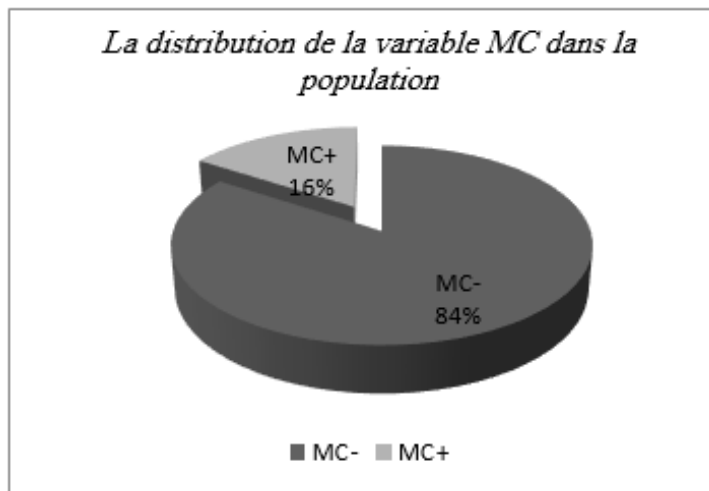
La distribution de la variable PRE-NS dans la population

NS	NS-	NS+
30	209	27



La distribution de la variable MC dans la population

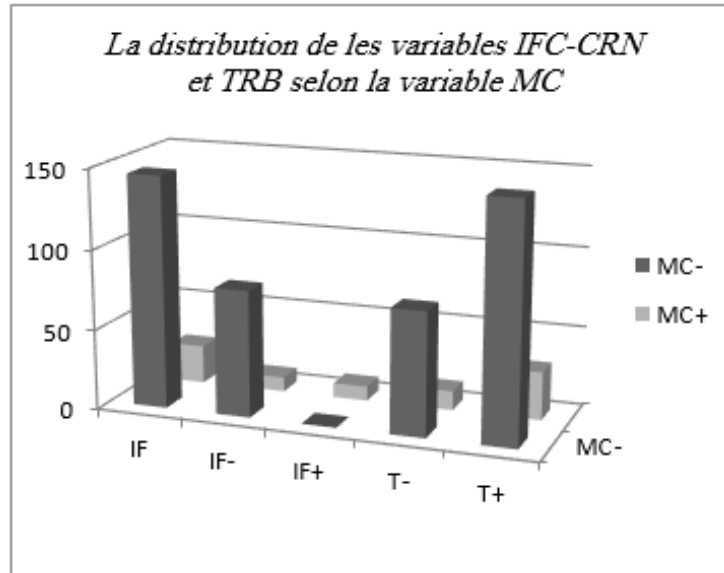
MC-	MC+
224	42



La distribution de les variables IFC-CRN et TRB selon le variable MC

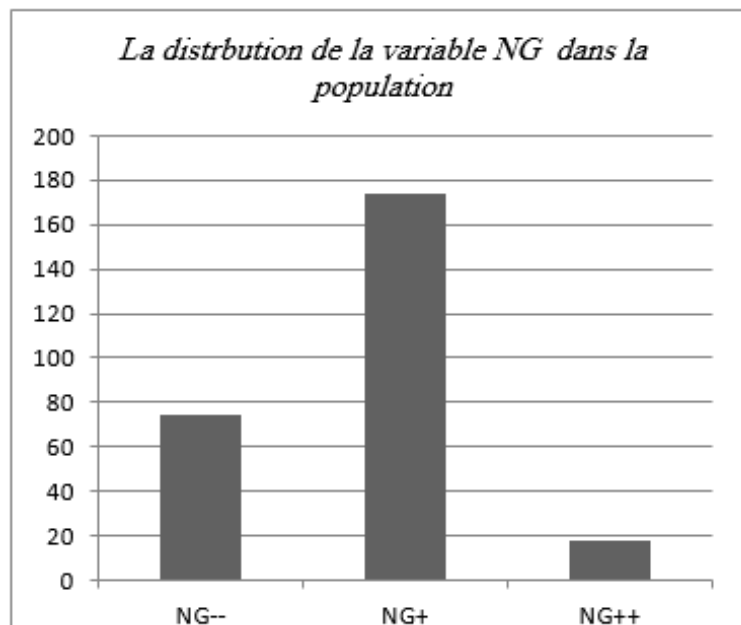
	IF	IF-	IF+	T-	T+
MC-	145	79	0	77	147
MC+	24	9	9	12	30

CHAPITRE 9. APPLICATION 2 : UNE DES MÉTHODES NON
PROBABILISTE AVEC UN ACM SUR R



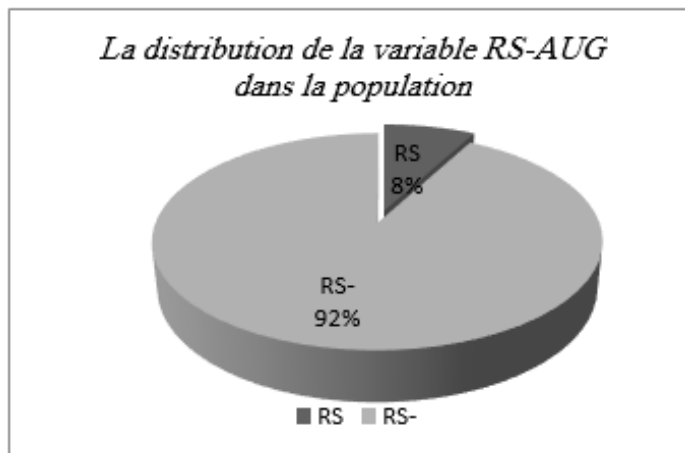
La distribution de la variable NG dans la population

NG-	NG+	NG++
74	174	18



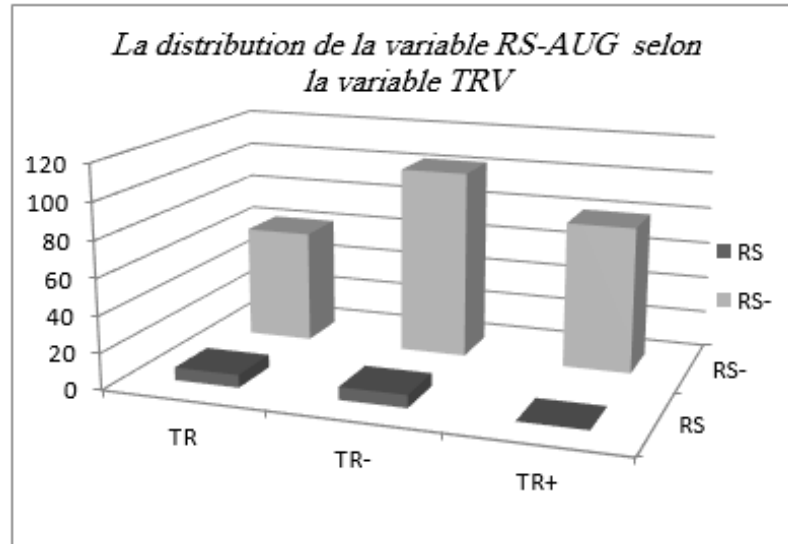
La distribution de la variable RS-AUG dans la population

RS	RS-
21	245



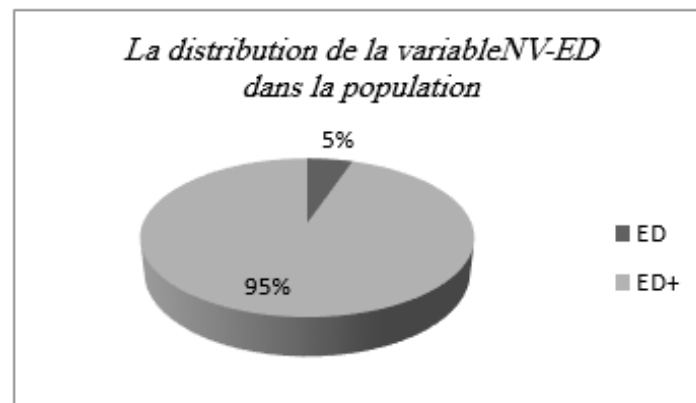
La distribution de la variable RS-AUG selon la variable TRV

	RS	RS-
TR	7	62
TR-	7	103
TR+	0	80



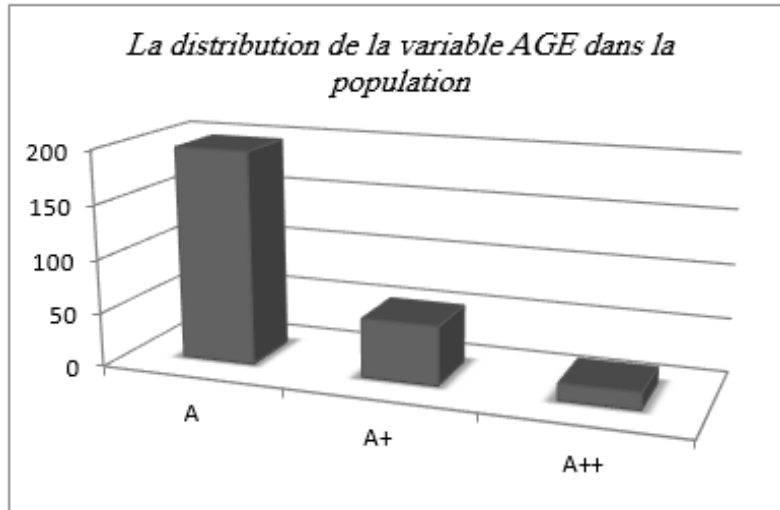
La distribution de la variable NV-ED dans la population

ED	ED+
14	252



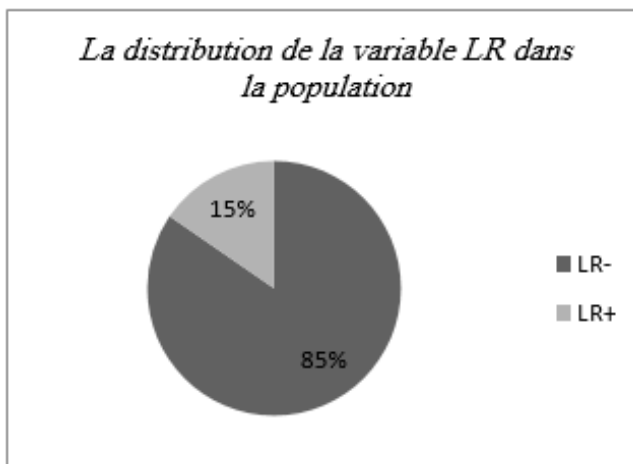
La distribution de la variable AGE dans la population

A	A+	A++
197	54	15



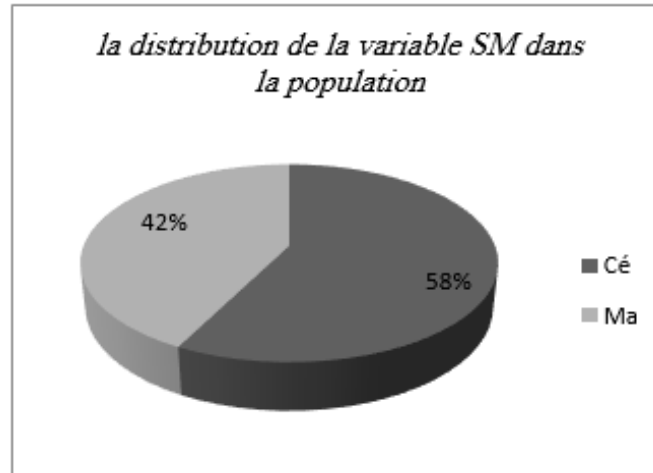
La distribution de la variable LR dans la population

LR-	LR+
225	41



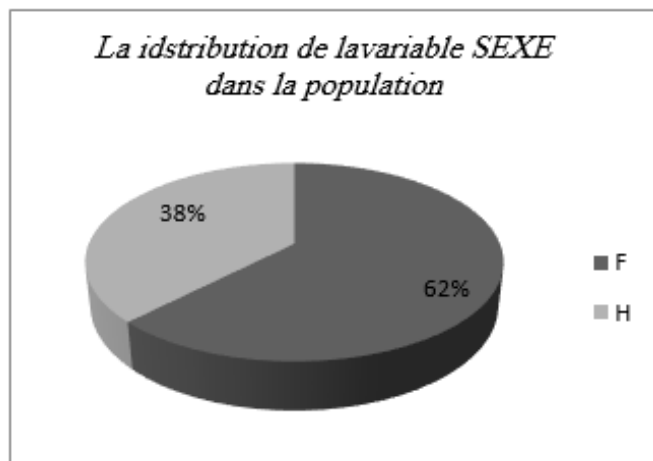
La distribution de la variable SM dans la population

Cé	Ma
153	113



La distribution de la variable SEXE dans la population

F	H
166	100



9.2.3 Analyse des correspondances multiples

Importation du jeu de données

Nous installons d'abord les packages suivantes :

-Le package « FactoMiner » pour exécuter ACM l'instruction qui construire ce tableau sur « R » est la suivante :

library (FactoMineR)

-Le package « ggplot2 » pour visualiser les graphes sous « R » l'instruction qui construire ce visualisation :

library (ggplot2)

-Le package « factoextra » pour facilitant l'extraction et la visualisation de la sortie d'analyses exploratoires de données multivariées.

library (factoextra)

Importation des données sur « R » par l'instruction suivante :

```
DONNEES<-read.csv("COVID-19.CSV ",header=TRUE,sep=";",dec
=".",row.names =1)
```

DONNEES

Nous verrons les données comme suit :

```
DONNEES <-read.csv("COVID-19.CSV",header=TRUE,sep=";",dec =".",row.names =1)
DONNEES
EXISTE INF.SM DR.DEV OP.VU EV.CRN RSP CH SUP PRS.NS M.FC MC IFC.CRN TRB NG RS.AUG LR TRV NV.ED SM AGE SEXE
E+ INF+ DV+ O+ EV+ R+ C+ S- NS+ MF- MC+ IF T+ NG++ RS- LR- TR- ED Ma A++ M
E+ INF+ DV- O+ EV R C+ S- NS MF- MC- IF T+ NG-- RS- LR- TR+ ED+ Ma A+ F
E+ INF DV- O- EV- R C- S+ NS MF- MC- IF- T- NG-- RS LR+ TR- ED+ Cé A H
E+ INF DV- O+ EV+ R- C+ S- NS- MF- MC- IF T- NG+ RS- LR- TR ED+ Ma A+ M
E+ INF DV+ O+ EV+ R+ C- S- NS+ MF+ MC+ IF+ T- NG-- RS- LR- TR- ED+ Ma A+ F
E+ INF+ DV- O- EV R- C S- NS MF+ MC- IF T- NG-- RS- LR- TR+ ED+ Ma A+ H
E+ INF DV- O+ EV R C+ S- NS MF- MC- IF T+ NG+ RS- LR- TR+ ED+ Ma A+ F
E+ INF DV- O+ EV R C+ S+ NS+ MF+ MC- IF T- NG+ RS- LR- TR ED+ Ma A+ H
E+ INF DV-- O+ EV R- C S+ NS MF+ MC- IF T+ NG+ RS- LR+ TR+ ED+ Ma A+ H
E- INF DV- O+ EV+ R C S- NS- MF+ MC+ IF- T- NG+ RS- LR- TR ED+ Ma A++ H
E+ INF DV- O- EV+ R+ C S- NS- MF- MC+ IF+ T- NG-- RS- LR- TR- ED+ Ma A+ F
E+ INF+ DV- O+ EV R C S- NS- MF+ MC+ IF- T+ NG-- RS- LR- TR- ED+ Ma A+ F
E+ INF- DV-- O+ EV R C+ S+ NS- MF+ MC- IF T+ NG+ RS- LR- TR ED+ Cé A+ F
E+ INF DV- O+ EV R C+ S- NS+ MF+ MC- IF T+ NG+ RS- LR+ TR+ ED+ Cé A+ H
E+ INF+ DV- O+ EV R- C S+ NS- MF+ MC- IF T- NG-- RS- LR+ TR- ED+ Cé A H
E+ INF DV+ O+ EV+ R- C S+ NS- MF- MC- IF T- NG++ RS- LR- TR ED+ Ma A H
E+ INF- DV-- O+ EV+ R- C+ S+ NS+ MF+ MC- IF T- NG+ RS- LR- TR ED+ Cé A H
E+ INF DV- O+ EV R C+ S- NS- MF+ MC- IF T+ NG+ RS- LR+ TR- ED+ Cé A+ F
E+ INF- DV-- O+ EV R C+ S- NS MF+ MC+ IF T+ NG+ RS- LR- TR- ED+ Ma A+ F
E+ INF DV- O- EV R C- S- NS- MF+ MC- IF- T- NG+ RS- LR- TR- ED+ Cé A H
```

Traitement du jeu de données

Après l'importation des données nous Nous commençons le traitement, et La première chose que nous commençons est le tableau disjonctif complet.

```
disj<- tab.disjonctif(DONNEES)
```

```
disj
```

CHAPITRE 9. APPLICATION 2 : UNE DES MÉTHODES NON PROBABILISTE AVEC UN ACM SUR R

```

#K<- tab.dimension(DONNEES)
#K
#  E-  INF  INF-  INF+  DV-  DV--  DV+  DV++  O-  O+  EV  EV-  EV+
# 26  240  197   34   35  132   86   14   34   30  236  208   27   31
#  R  R-   R+   C   C-   C+   S-   S+   NS  NS-  NS+  MF-  MF+  MC-
#206   57    3   58   38  170  195   71   30  209   27  117  149  224
#MC+  IF  IF-  IF+  T-   T+  NG--  NG+  NG++  RS  RS-  LR-  LR+  TR
# 42  169   88    9   89  177   74  174   18   21  245  225   41   69
#TR-  TR+  ED  ED+  Cé  Ma   A   A+  A++   F   H
#110   87   14  252  153  113  197   54   15  166  100

```

L'instruction suivante nous donne chaque modalité contient combien d'individus.

```
K<-apply(disj,2,sum)
```

K

```

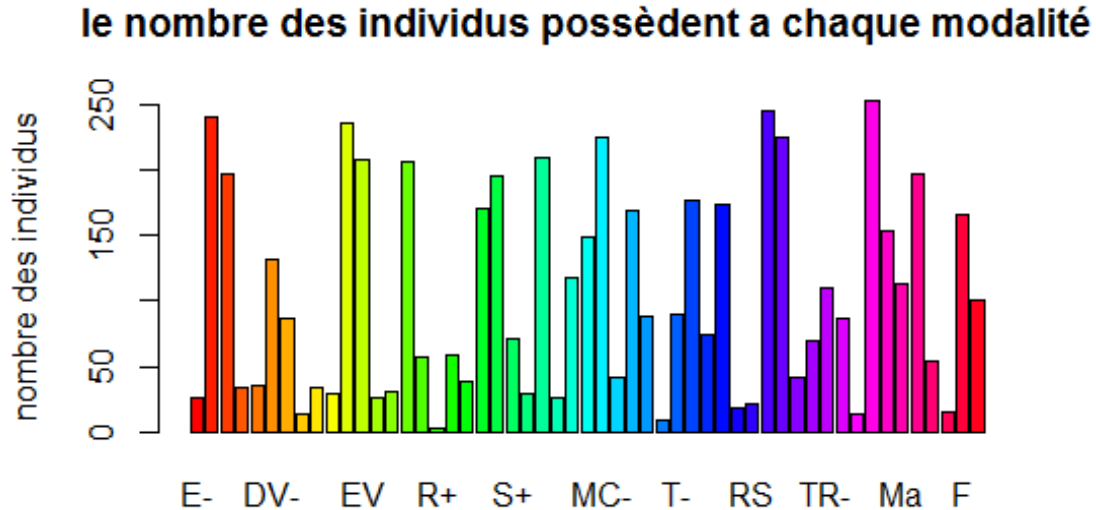
> K<-apply(disj,2,sum)
> K
  E-   E+  INF INF-  INF+  DV-  DV--  DV+  DV++  O-   O+  EV  EV-  EV+
26  240  197   34   35  132   86   14   34   30  236  208   27   31
  R  R-   R+   C   C-   C+   S-   S+   NS  NS-  NS+  MF-  MF+  MC-
206   57    3   58   38  170  195   71   30  209   27  117  149  224
MC+  IF  IF-  IF+  T-   T+  NG--  NG+  NG++  RS  RS-  LR-  LR+  TR
 42  169   88    9   89  177   74  174   18   21  245  225   41   69
TR-  TR+  ED  ED+  Cé  Ma   A   A+  A++   F   H
110   87   14  252  153  113  197   54   15  166  100

```

Nous représentons graphiquement les résultats du tableau précédent par cette instruction.

```
barplot(K,ylab="nombre des individus",col=rainbow(dim(Burt)))
```

```
title("le nombre des individus possèdent à chaque modalité")
```



Remarque : La somme de Chaque ensemble de modalité du même variable est égale à 266 qui est le nombre de participants da cette enquête.

Pour généraliser l'analyse des correspondances à l'étude des croisement entre plus de deux variables on peut construire un tableau contenant l'ensemble des tableaux de contingence entre les variables prises 2 à 2. C'est le tableau de Burt.

```
Burt <- t(disj)%*%disj
```

```
Burt
```

```
> Burt <- t(disj)%*%disj
```

```
> Burt
```

	E-	E+	INF	INF-	INF+	DV-	DV--	DV+	DV++	O-	O+	EV	EV-	EV+	R	R-	R+
E-	26	0	16	4	6	15	3	4	4	2	24	5	18	3	17	9	0
E+	0	240	181	30	29	117	83	10	30	28	212	203	9	28	189	48	3
INF	16	181	197	0	0	99	68	9	21	21	176	158	16	23	155	40	2
INF-	4	30	0	34	0	13	11	0	10	6	28	24	5	5	24	10	0
INF+	6	29	0	0	35	20	7	5	3	3	32	26	6	3	27	7	1
DV-	15	117	99	13	20	132	0	0	0	14	118	102	13	17	103	28	1
DV--	3	83	68	11	7	0	86	0	0	11	75	76	3	7	71	15	0
DV+	4	10	9	0	5	0	0	14	0	0	14	6	5	3	9	3	2
DV++	4	30	21	10	3	0	0	0	34	5	29	24	6	4	23	11	0
O-	2	28	21	6	3	14	11	0	5	30	0	23	4	3	23	6	1
O+	24	212	176	28	32	118	75	14	29	0	236	185	23	28	183	51	2
EV	5	203	158	24	26	102	76	6	24	23	185	208	0	0	184	24	0
EV-	18	9	16	5	6	13	3	5	6	4	23	0	27	0	16	11	0
EV+	3	28	23	5	3	17	7	3	4	3	28	0	0	31	6	22	3

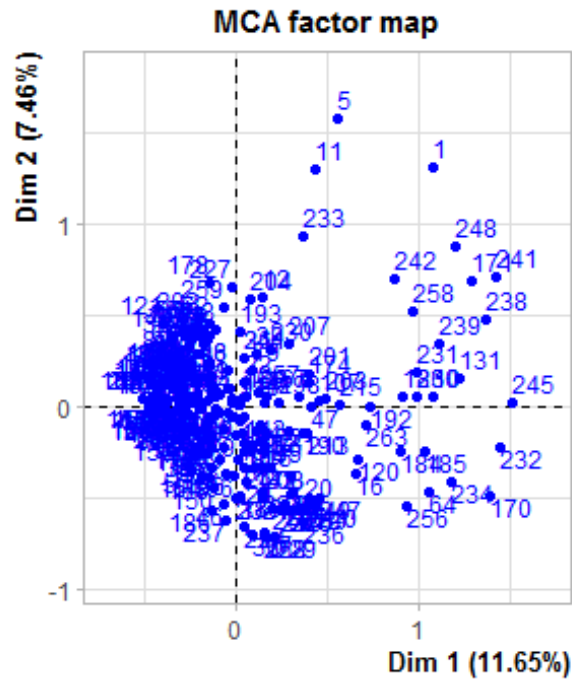
Nous avons vu la représentation graphique préliminaire des individus, des

*CHAPITRE 9. APPLICATION 2 : UNE DES MÉTHODES NON
PROBABILISTE AVEC UN ACM SUR R*

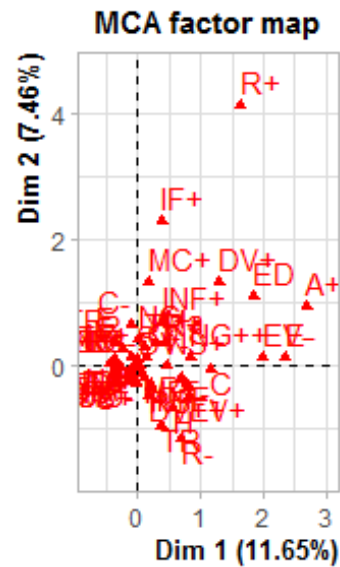
modalités et des variables par l'instruction suivante :

```
M <- MCA(DONNEES,ncp=2,graph=TRUE)
```

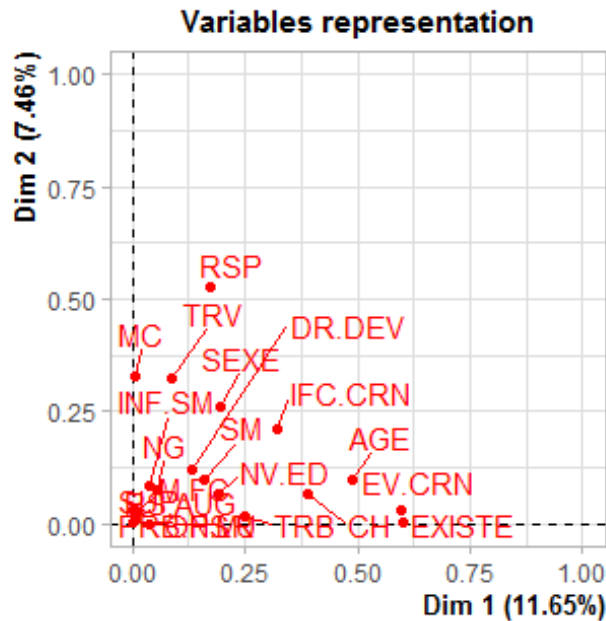
- Représentation graphique des individus.



- Représentation graphique des modalités.



- Représentation graphique des variables.



Tous ces représentations graphiques sont avant le centrage des coordonne des individus ou des variables. C'est pour sa on observe des figures mal déformées avec une grande variance entre les éléments soit individu ou variable.

- Les coordonnés factorielles standardisé des individus et des modalités :

Et comme nous dit précédemment, les anciennes représentations sont mal déformées c'est pour ça on construire une base standardisée qui donne des nouveaux coordonnés. Se dernière aidons nous à représenter les individus et les variables avec la meilleure forme qui conserve le maximum d'information inclue dans l'ensemble des données a étudié.

Et ce sont les deux structures qu'ils nous habilitent de faire ça :

Xstand <- X %%% diag(1/d) #coordonnées factorielles standardisées des individus.

Ystand <- Y %%% diag(1/d) #coordonnées factorielles standardisées des modalité.

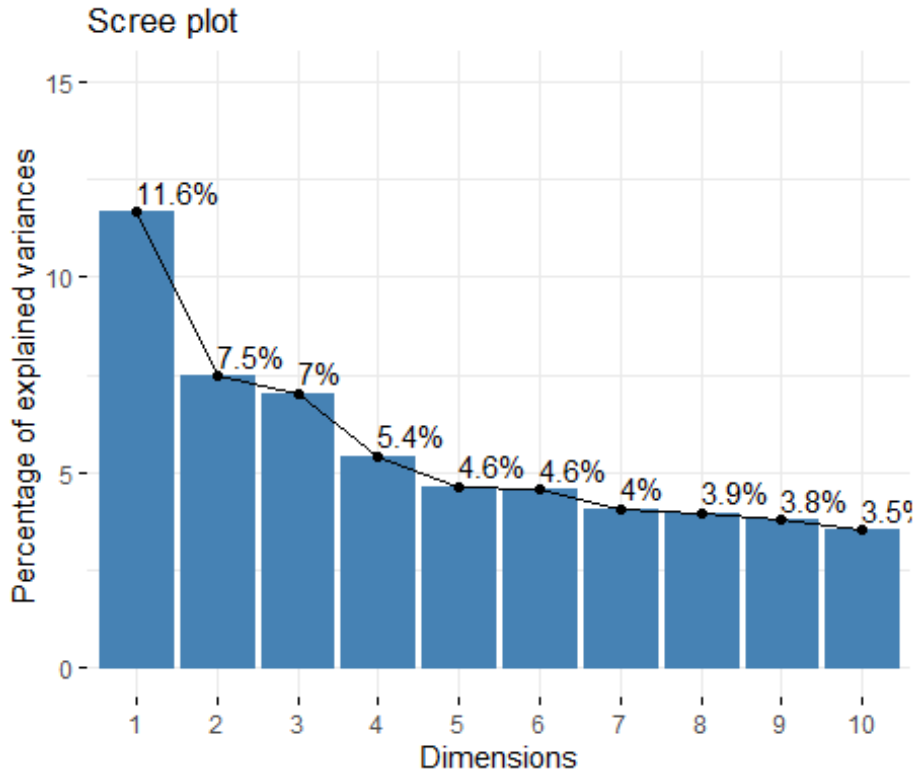
Xstand			Ystand		
	[,1]	[,2]		[,1]	[,2]
2.5665855014	3.867670072		5.58733695	0.404074292	
-0.3715637125	1.073505305		-0.60529484	-0.043774715	
0.6782269361	-0.407578901		-0.26128671	-0.230092016	
0.7105603019	-1.431688499		0.49656932	-0.897650221	
1.3258109664	4.666867170		0.98828930	2.167092418	
0.8068723624	0.149465505		0.09576058	0.001132543	
-0.5469851380	0.476445319		-0.81564372	-0.098748099	
0.2023154489	-0.652549701		3.06398147	3.952758388	
0.2135088736	-0.640342356		0.42968305	-1.382228722	
2.5439927326	0.147838816		0.37447235	0.353719579	
1.0182409475	3.850690116		-0.04760242	-0.044964353	
0.3354556994	1.770973461		-0.90747352	0.158132775	
-0.6875196507	-0.471100562		4.72969701	0.365765912	
-0.3343680582	-0.175017798		1.96944105	-1.379590220	
0.3677271143	-0.621316867		-0.51522989	0.766242267	
1.5435784435	-1.094558041		1.65758559	-3.413783293	
0.3800216167	-2.057135269		3.88499260	12.246580201	
-0.8247670003	0.288954618		2.76688238	-0.149432396	
-0.6805201549	1.186693024		-0.21209472	1.895941399	
-0.0358201170	-0.185816433		-0.89658576	-0.372815848	

Aide à l'interprétation

Le choix du nombre des axes retenir Combien en retenir ? C'est l'examen de l'histogramme des valeurs propres qui va permettre de choisir. On trace cet histogramme de la façon suivante, avec la fonction suivant :

```
fviz_screplot (M, addlabels = TRUE, xlim=c(1,10), ylim = c(1, 15))
```

```
title("Eboulis des valeurs propres en % ")
```



On peut utiliser l'histogramme des valeurs propres pour effectuer visuellement cet arbitrage à l'aide des méthodes vues pour l'ACP et l'AFC, critère du « coude » et critère de « Kaiser ».

- Critère de « Kaiser » : suggère de garder tous les axes dont la valeur propre est supérieure à la valeur propre moyenne, Ici, le critère de Kaiser incite à retenir les axes dont le pourcentage d'inertie est supérieur à $(100/21) = 4.76\%$, on garde alors les 2 premiers axes.

- Critère du « coude » : on observe un décrochement (coude), suivi d'une décroissance régulière, entre le deuxième axe et le troisième axe, ce qui pourrait inciter à n'examiner que les deux premiers axes.

Les deux premiers axes sont bien interprétables, ils cumulent presque de 19.11% l'inertie totale (11.65% pour le premier axe 7.46% pour le deuxième axe expliquant par les valeurs propres 0.18 et 0.11 respectivement).

La fonction suivante nous donne la distribution du le pourcentage d'inertie aux 32 dimensions expliquant par les valeurs propres.

```
round(M$eig,2)
```

CHAPITRE 9. APPLICATION 2 : UNE DES MÉTHODES NON
PROBABILISTE AVEC UN ACM SUR R

```
> round(M$eig,2)
      eigenvalue percentage of variance cumulative percentage of variance
dim 1          0.18                11.65                11.65
dim 2          0.11                 7.46                19.11
dim 3          0.11                 7.02                26.13
dim 4          0.08                 5.40                31.52
dim 5          0.07                 4.63                36.15
dim 6          0.07                 4.55                40.70
dim 7          0.06                 4.05                44.75
dim 8          0.06                 3.95                48.70
dim 9          0.06                 3.78                52.48
dim 10         0.05                 3.52                56.00
dim 11         0.05                 3.39                59.39
dim 12         0.05                 3.30                62.69
dim 13         0.05                 2.99                65.68
dim 14         0.04                 2.93                68.61
dim 15         0.04                 2.89                71.50
dim 16         0.04                 2.65                74.16
dim 17         0.04                 2.48                76.64
dim 18         0.03                 2.26                78.90
dim 19         0.03                 2.20                81.10
dim 20         0.03                 2.09                83.18
dim 21         0.03                 2.02                85.21
dim 22         0.03                 1.84                87.05
dim 23         0.03                 1.76                88.81
dim 24         0.03                 1.71                90.53
dim 25         0.02                 1.58                92.11
dim 26         0.02                 1.53                93.63
dim 27         0.02                 1.41                95.04
dim 28         0.02                 1.35                96.39
dim 29         0.02                 1.20                97.59
dim 30         0.02                 0.99                98.58
dim 31         0.01                 0.78                99.36
dim 32         0.01                 0.64               100.00
```

Noter dans la deuxième colonne le Pourcentage d'inertie explique par chaque valeur propre et le Pourcentage d'inertie cumulé, Quel total jusqu'à 100 dans la troisième colonne.

L'interprétation des axes L'interprétation des axes factoriels se fait séquentiellement, pour chaque axe et chaque nuage de points, en regardant les contributions à la formation des axes.

L'interprétation d'un ACM

Les graphiques peut devenir illisible, il ne permet pas de distinguer entre les modalités qui contribuent le plus fortement et celles qui ne contribuent pas aux axes représentés.

Dans ce cas-là on peut se servir de leur contribution aux axes représentés pour sélectionner celles qui ont justement les contributions les plus importantes.

Un critère possible est de ne retenir que les modalités dont la contribution est supérieure à deux fois la contribution moyenne.

(PS : Ce seuil est évidemment arbitraire, et rien n'empêche de retenir un seuil plus élevé ou plus faible).

Suivant le nombre de modalités qu'on veut représenter, dans notre cas, le nombre de modalités actives étant faible.

On peut garder celles dont la contribution à l'un ou l'autre des deux premiers axes est simplement supérieure à la contribution moyenne.

La contribution moyenne est l'inverse du nombre de modalités actives.

Dans notre cas, les 21 variables actives se décomposent en un total de 53 modalités.

La contribution moyenne est donc égale à :

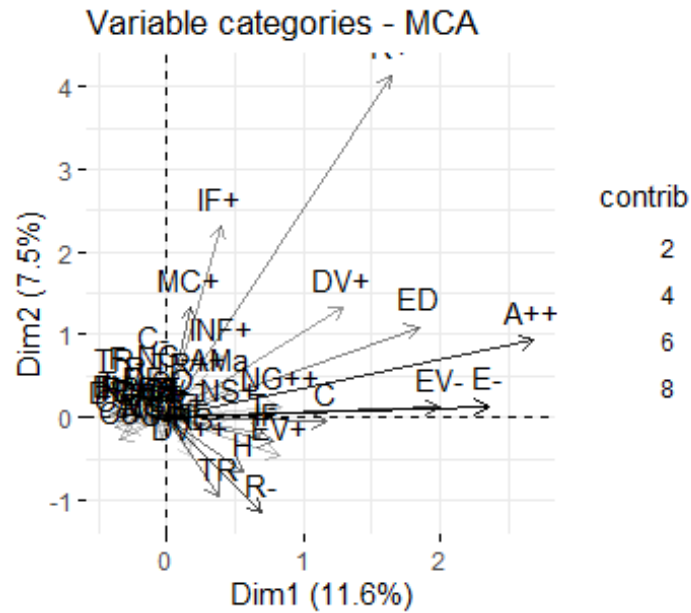
$$\text{Contrib moyenne} = 100/53 = 1.88\%.$$

On ne retiendra donc dans l'interprétation des deux premiers axes de notre ACM que les modalités dont la contribution est supérieure à 1.88%.

Seules les modalités dont la contribution est élevée sont à considérer pour l'interprétation d'un axe.

L'axe 1 Profile colonne (modalités)

```
fviz_pca_var(M, pointsize = "contrib", alpha.var = "contrib", select.var = list(contrib = 53))
```



```

> axe1
      coord1 contrib1 cos*1 MF+   -0.1    0.1    0.0
E-      2.4     14.5    0.6 MC-   0.0    0.0    0.0
E+     -0.3      1.6    0.6 MC+   0.2    0.1    0.0
INF     -0.1      0.2    0.0 IF    -0.4    3.1    0.3
INF-     0.2      0.2    0.0 IF-   0.8    5.4    0.3
INF+     0.4      0.6    0.0 IF+   0.4    0.1    0.0
DV-      0.0      0.0    0.0 T-    0.7    4.4    0.2
DV--    -0.3      1.0    0.1 T+   -0.4    2.2    0.2
DV+      1.3      2.4    0.1 NG--   0.0    0.0    0.0
DV++     0.2      0.1    0.0 NG+   -0.1    0.2    0.0
O-       0.2      0.1    0.0 NG++   0.8    1.3    0.1
O+       0.0      0.0    0.0 RS     0.2    0.1    0.0
EV      -0.4      3.1    0.5 RS-    0.0    0.0    0.0
EV-      2.0     10.8    0.4 LR-    0.0    0.0    0.0
EV+      0.8      2.2    0.1 LR+    0.1    0.0    0.0
R       -0.2      1.0    0.2 TR     0.4    1.0    0.1
R-       0.7      2.8    0.1 TR-   -0.3    1.2    0.1
R+       1.6      0.8    0.0 TR+    0.1    0.1    0.0
C        1.2      7.9    0.4 ED     1.8    4.8    0.2
C-      -0.1      0.0    0.0 ED+   -0.1    0.3    0.2
C+      -0.4      2.4    0.3 Cé    -0.3    1.8    0.2
S-       0.0      0.0    0.0 Ma     0.5    2.5    0.2
S+       0.0      0.0    0.0 A     -0.3    1.6    0.2
NS      -0.3      0.3    0.0 A+     0.3    0.5    0.0
NS-      0.0      0.0    0.0 A++    2.7   10.9    0.4
NS+      0.5      0.6    0.0 F     -0.3    1.9    0.2
MF-      0.1      0.2    0.0 H      0.6    3.2    0.2

```

On constate sur l'axe 1 que les contributions varient de 0,00% à 14,50%.

On sait que les modalités contribuant le plus à la formation de l'axe 1 sont celles dont les cosinus sur cet axe sont proches de 1 en valeur absolue.

les modalités les plus excentrés ne sont pas nécessairement celles qui contribuent le plus sur l'axe 1, on remarque que :

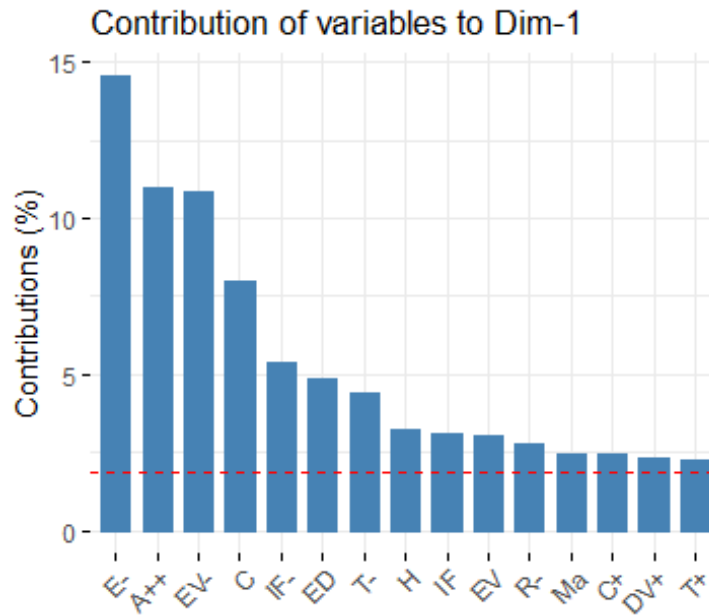
Les modalités actives qui Contribuent (+)	Les modalités actives qui Contribuent (-)
E-	IF
A++	EV
EV-	C+
C	T+
IF-	
ED	
T-	

Donc on a :

- La modalité E- (le COVID_19 n'existe pas) a la plus forte contribution positivement avec un taux de 14.5%.
- De plus, les modalités IF (parfois) ; EV (parfois) ont la plus forte contribution négative avec un taux de 3.1%.

L'histogramme suivant confirme nos résultats :

`fviz_contrib(mca, choice = "var", axes = 1, top = 15)`



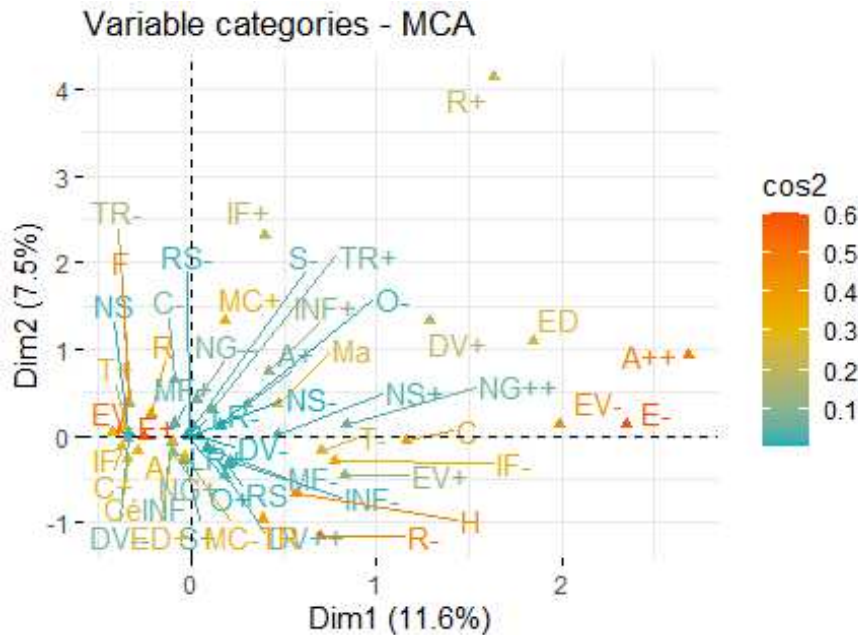
La qualité de représentation des modalités sur l'axe 1 D'après le tableau de « la qualité de représentation des modalités » obtenu par la commande :

`round(M$var$cos2[,1],1)`

`> round(M$var$cos2[,1],1)`

E-	E+	INF	INF-	INF+	DV-	DV--	DV+	DV++	O-	O+	EV	EV-
0.6	0.6	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.5	0.4
EV+	R	R-	R+	C	C-	C+	S-	S+	NS	NS-	NS+	MF-
0.1	0.2	0.1	0.0	0.4	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
MF+	MC-	MC+	IF	IF-	IF+	T-	T+	NG--	NG+	NG++	RS	RS-
0.0	0.0	0.0	0.3	0.3	0.0	0.2	0.2	0.0	0.0	0.1	0.0	0.0
LR-	LR+	TR	TR-	TR+	ED	ED+	Cé	Ma	A	A+	A++	F
0.0	0.0	0.1	0.1	0.0	0.2	0.2	0.2	0.2	0.2	0.0	0.4	0.2
H												
0.2												

```
fviz_mca_var(mca, repel = TRUE, col.var = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```

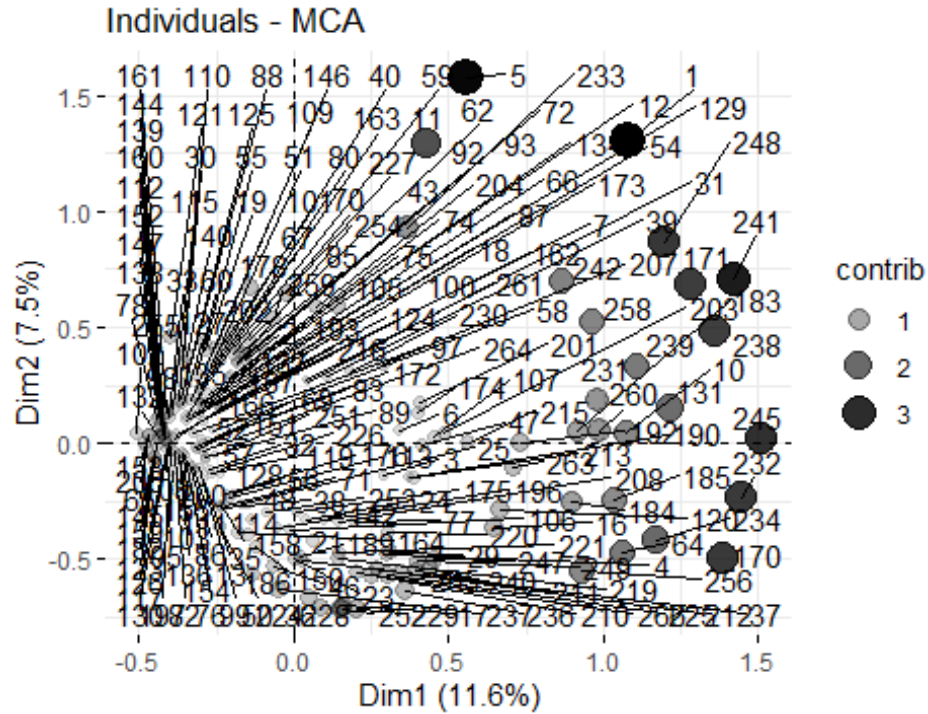


On remarque que :

- Les modalités EV , E^- , E^+ sont moyennement bien représentées sur l'Axe1.
- Toutes les autres modalités actives sont plutôt mal représentés sur l'Axe1.

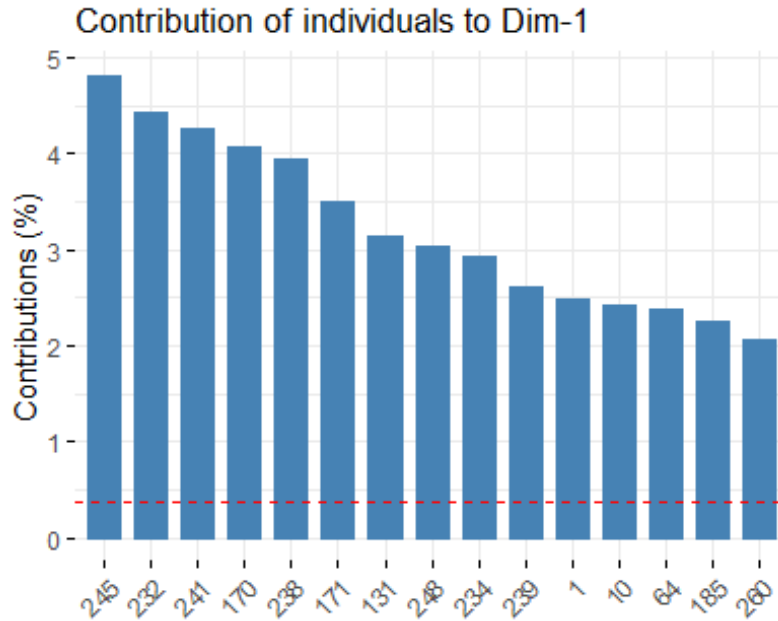
L'axe 1 Profile ligne (individus)

```
fviz_pca_ind(M, pointsize = "contrib", alpha.ind = "contrib", select.ind
= list(contrib = 200), jitter = list(what = "label", width = 0.1,
height = 0.15))
```



On remarque que les individus excentrés sont les mieux représentés alors que les individus proches du centre ont une mauvaise qualité de représentation.

```
fviz_contrib(M, choice = "ind", axes = 1, top = 15)
```

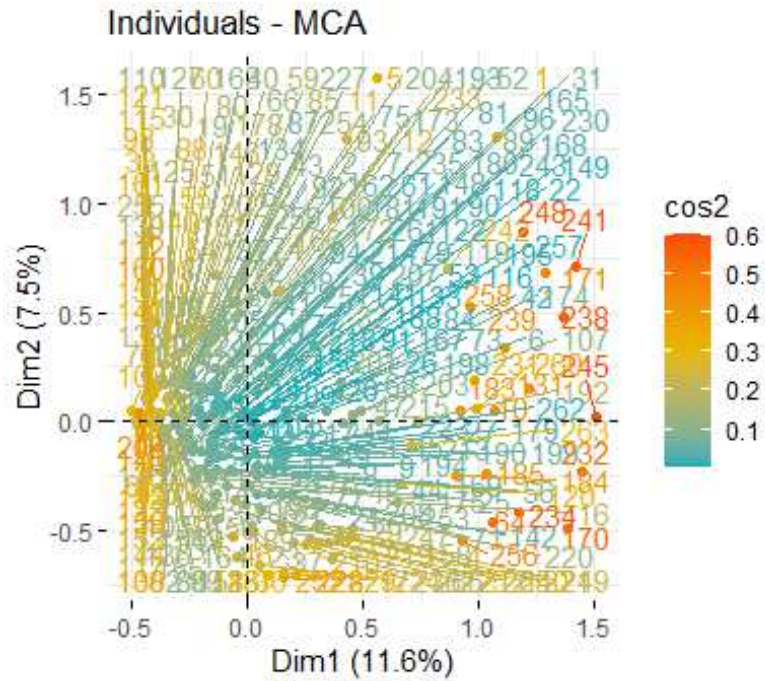


On remarquons que :

Les individus qui contribuent (-)	Les individus qui contribuent (+)
206,145,112, 98, 65.	245, 232, 241, 170, 238, 171, 131

La qualité de représentation des individus sur l'axe 1 A chaque fois qu'on s'éloigne du centre on voit une augmentation dans la qualité de représentation.

```
fviz_mca_ind(mca, repel = TRUE,col.ind = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



D'après le tableau de « la qualité de représentation des individus » obtenu par la commande :

```
round(M$ind$cos2[,1],1)
```

```

> round(M$ind$cos2[,1],1)
  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
0.1 0.0 0.0 0.1 0.0 0.1 0.1 0.0 0.0 0.4 0.0 0.0 0.1 0.0 0.0 0.1 0.0 0.1 0.1 0.0
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.1 0.0 0.1 0.3 0.0 0.1 0.0 0.0 0.1 0.0 0.1
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.1 0.1 0.1 0.0 0.2 0.1 0.0 0.1 0.1 0.0 0.1
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
0.0 0.0 0.0 0.4 0.4 0.1 0.1 0.0 0.1 0.0 0.1 0.0 0.0 0.1 0.0 0.2 0.0 0.2 0.0 0.1
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
0.0 0.1 0.0 0.0 0.0 0.1 0.0 0.2 0.1 0.0 0.0 0.0 0.0 0.0 0.2 0.0 0.1 0.3 0.1 0.1
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
0.0 0.3 0.1 0.0 0.0 0.0 0.1 0.3 0.1 0.1 0.2 0.4 0.0 0.0 0.2 0.0 0.2 0.0 0.1 0.3
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
0.1 0.0 0.3 0.1 0.1 0.2 0.1 0.1 0.1 0.3 0.4 0.3 0.0 0.0 0.2 0.1 0.1 0.2 0.2 0.1
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
0.0 0.0 0.0 0.2 0.4 0.1 0.3 0.0 0.0 0.0 0.1 0.3 0.3 0.1 0.2 0.0 0.0 0.0 0.4 0.4
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
0.3 0.1 0.1 0.0 0.0 0.2 0.0 0.0 0.0 0.5 0.4 0.2 0.0 0.1 0.0 0.1 0.0 0.0 0.0 0.1
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
0.1 0.3 0.4 0.4 0.4 0.0 0.1 0.0 0.0 0.1 0.0 0.2 0.0 0.0 0.0 0.0 0.3 0.0 0.0 0.2
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
0.1 0.0 0.1 0.0 0.0 0.4 0.0 0.0 0.0 0.1 0.1 0.1 0.1 0.0 0.1 0.3 0.0 0.0 0.0 0.1
221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.0 0.0 0.0 0.3 0.5 0.0 0.5 0.0 0.0 0.0 0.5 0.4 0.1
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
0.5 0.2 0.0 0.0 0.6 0.1 0.1 0.4 0.1 0.0 0.1 0.0 0.0 0.0 0.2 0.3 0.0 0.3 0.0 0.3
261 262 263 264 265 266
0.0 0.0 0.3 0.1 0.0 0.0

```

On remarque que :

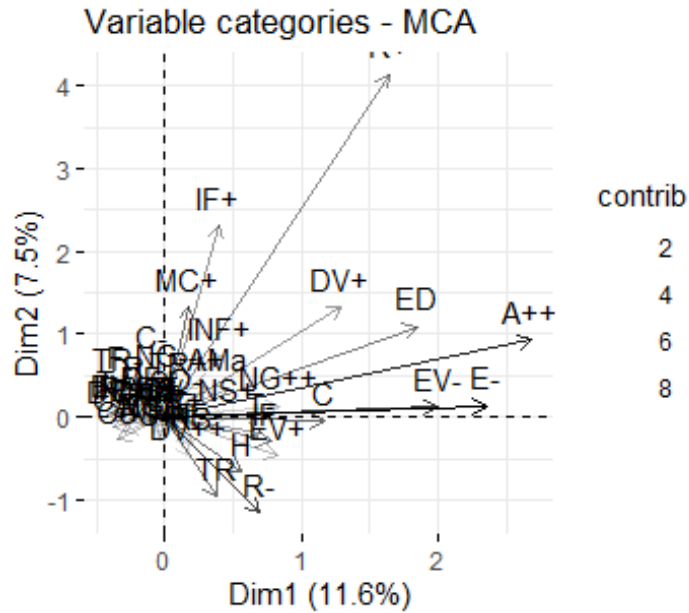
- Les individus 245, 241, 238, 232, 234, 131, 112, 65, 10 sont moyennement bien représenté sur l'Axe1.
- Toutes les autres individus actives sont plutôt mal représentées sur l'Axe1.

Conclusion On conclue que :

- Les individus possédant les modalités (actives) biens représentées du côté positif de l'axe 1, ont plus de chance de posséder des modalités comme « EV-, E-, A++ , C, IF- » et les individus bien représentés du côté négatif, ont plus de chance de posséder des modalités comme « ».
- Le premier axe distingue par :
 - les gens âgées (A++) qui croient que le COVID-19 n'existe pas (E-). Ces gens arrêtent leurs étude au niveau moyenne (ED). ils ne respecte pas le quarentaire (R-) ,et ne prend pas des mesures préventives pour éviter l'infection par le virus (EV-). Aussi ils ne sont pas peur d'être infecté par le virus (IF-) et ne ont pas des troubles ou de la peur lorsque le virus a commencé à se propage (T-). de plus, lorsque ils toussent ou surchauffé , ils n'inquiètent pas et complètent les activités de la vie quotidienne (C).

L'axe 2 Profile colonne (modalités)

```
fviz_pca_var(M, pointsize = "contrib", alpha.var = "contrib", select.var = list(contrib = 53))
```



D'après le tableau des contributions des modalités on voit bien que les modalités qui ont une contribution supérieure à la contribution moyenne (1.88%) sont ceux qui contribuent le plus à la formation de l'Axe2.

```
axe2<-matrix(nrow=53, ncol=3)

axe2[,1]<-round(M$var$coord[,2],1)

axe2[,2]<-round(M$var$contrib[,2],1)

axe2[,3]<-round(M$var$cos2[,2],1)

rownames(axe2)<-rownames(M$var$coord)

colnames(axe2)<-c('coord2','contrib2', 'cos²2')

axe2
```

```

> axe2
      coord2 contrib2 cos^2 MF+      0.1      0.5      0.0
E-      0.1      0.1      0.0 MC-     -0.2      2.2      0.3
E+      0.0      0.0      0.0 MC+      1.3     11.6      0.3
INF     -0.1      0.2      0.0 IF      0.0      0.0      0.0
INF-    -0.3      0.5      0.0 IF-     -0.3      1.2      0.0
INF+     0.7      2.9      0.1 IF+      2.3      7.6      0.2
DV-      0.0      0.0      0.0 T-      -0.2      0.5      0.0
DV--     0.0      0.0      0.0 T+       0.1      0.2      0.0
DV+      1.3      3.9      0.1 NG--     0.4      2.1      0.1
DV++    -0.5      1.2      0.0 NG+     -0.2      1.0      0.1
O-       0.1      0.1      0.0 NG++     0.1      0.1      0.0
O+       0.0      0.0      0.0 RS      -0.3      0.4      0.0
EV       0.1      0.1      0.0 RS-      0.0      0.0      0.0
EV-      0.1      0.1      0.0 LR-      0.0      0.0      0.0
EV+     -0.5      1.1      0.0 LR+     -0.1      0.1      0.0
R        0.3      2.2      0.2 TR      -1.0     10.0      0.3
R-      -1.2     11.9      0.4 TR-      0.4      2.3      0.1
R+       4.1      8.1      0.2 TR+      0.3      1.3      0.0
C       -0.1      0.0      0.0 ED       1.1      2.6      0.1
C-       0.6      2.4      0.1 ED+     -0.1      0.1      0.1
C+      -0.1      0.4      0.0 C       -0.3      1.7      0.1
S-       0.1      0.3      0.0 Ma       0.4      2.3      0.1
S+      -0.3      1.0      0.0 A       -0.2      0.9      0.1
NS       0.0      0.0      0.0 A+       0.4      1.1      0.0
NS-      0.0      0.0      0.0 A++      0.9      2.1      0.1
NS+      0.0      0.0      0.0 F        0.4      4.1      0.3
MF-     -0.2      0.6      0.0 H       -0.7      6.9      0.3

```

On constate sur l'axe 2 que les contributions varient de 0,00% à 11.9%.

Voire le graphe suivant :

Dapr  s la table de contrubution et les graphes pr  c  dent, on obtient le tableau suivant :

*CHAPITRE 9. APPLICATION 2 : UNE DES MÉTHODES NON
PROBABILISTE AVEC UN ACM SUR R*

Les modalités actives qui Contribuent (+)	Les modalités actives qui Contribuent (-)
MC+	R-
R+	TR
IF+	H
INF+	F
TR-	MC-
C-	
ED	
R	
NG-	

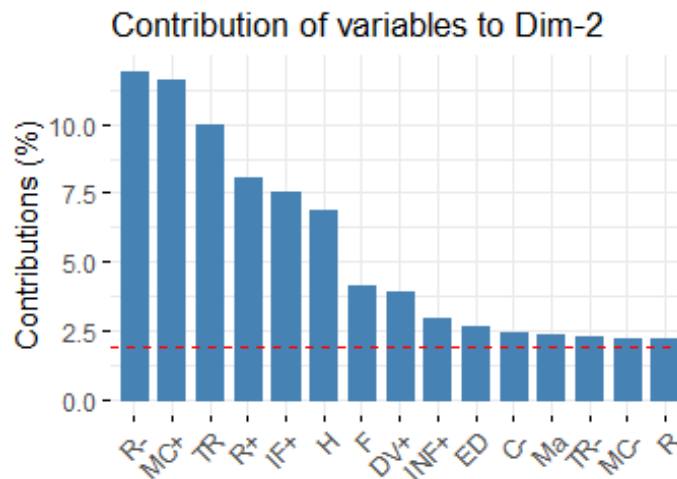
Remarquant que :

- La modalité MC+ (oui) a la plus forte contribution positive avec un taux de 11.6%.

- La modalité R- (non) a la plus forte contribution négative avec un taux de 11.9%.

L'histogramme suivant confirme nos resultats :

fviz_contrib(mca, choice = "var", axes = 2, top = 15)



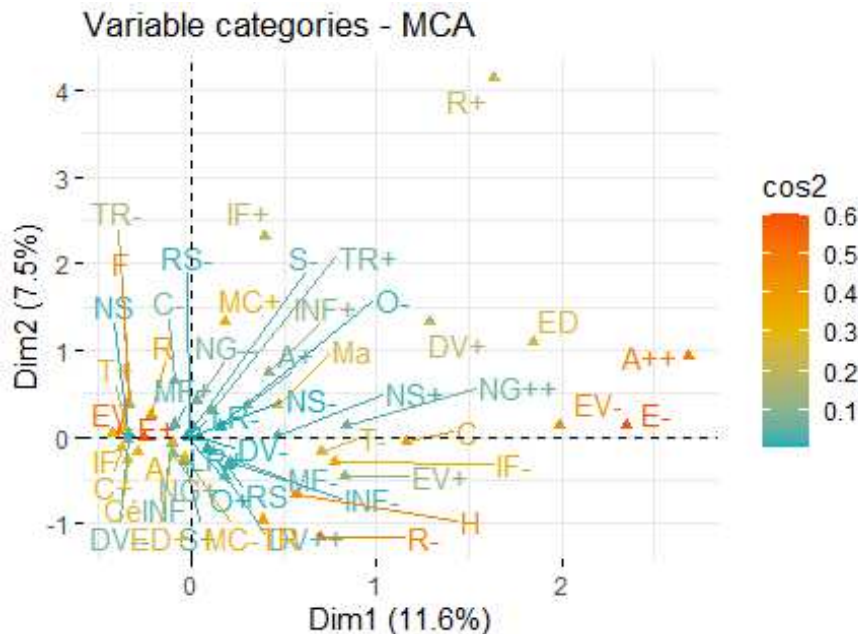
La qualité de représentation des modalités sur l'axe 2 D'après le tableau de « la qualité de représentation des modalités » obtenu par la commande :

```
round(M$var$cos2[,2],1)
```

```
> round(M$var$cos2[,2],1)
```

E-	E+	INF	INF-	INF+	DV-	DV--	DV+	DV++	O-	O+	EV	EV-
0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0
EV+	R	R-	R+	C	C-	C+	S-	S+	NS	NS-	NS+	MF-
0.0	0.2	0.4	0.2	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
MF+	MC-	MC+	IF	IF-	IF+	T-	T+	NG--	NG+	NG++	RS	RS-
0.0	0.3	0.3	0.0	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0
LR-	LR+	TR	TR-	TR+	ED	ED+	Cé	Ma	A	A+	A++	F
0.0	0.0	0.3	0.1	0.0	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.3
H												
0.3												

```
fviz_mca_var(mca, repel = TRUE, col.var = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



On remarque que :

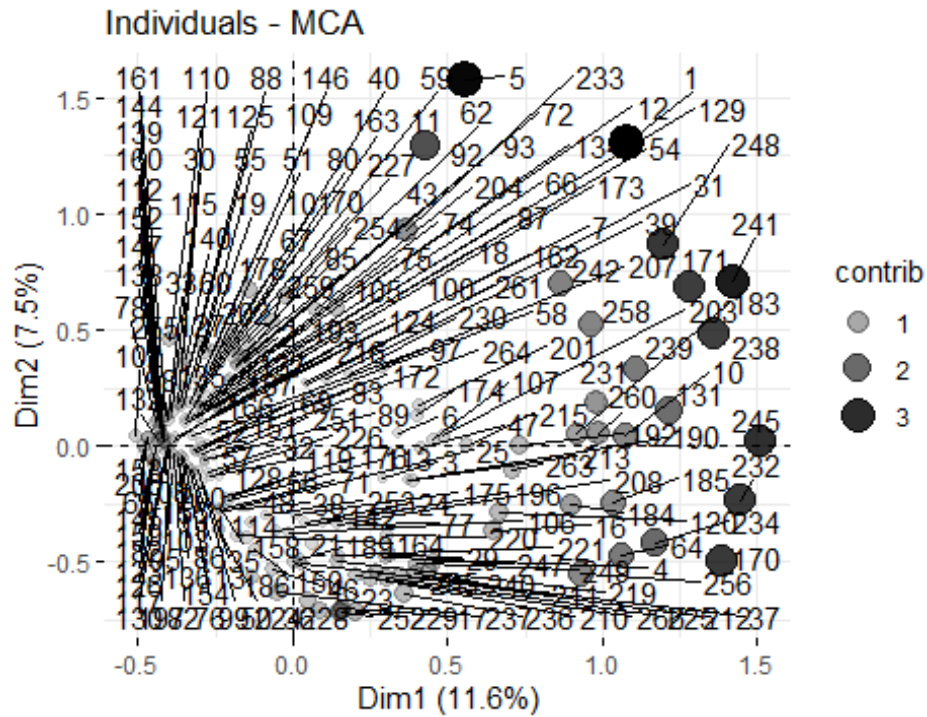
- Toutes les autres modalités actives sont plutôt mal représentées sur l'Axe2.

(remarque : Notons que les modalités qui ont un $\cos^2$ proche de 1 ont des grandes contributinon dans les axes .)

L'axe 2 Profile colonne (individus)

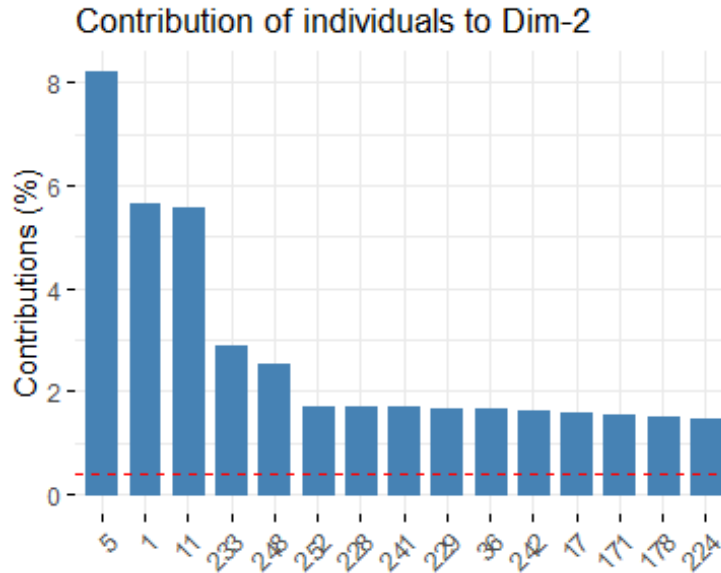
```
fviz_pca_ind(M, pointsize = "contrib", alpha.ind = "contrib", select.ind
```

```
= list(contrib = 200), jitter = list(what = "label", width = 0.1,
height = 0.15))
```



On remarque que les individus excentrés sont les mieux représentés alors que les individus proches du centre ont une mauvaise qualité de représentation.

```
fviz_contrib(M, choice = "ind", axes = 2, top = 15)
```

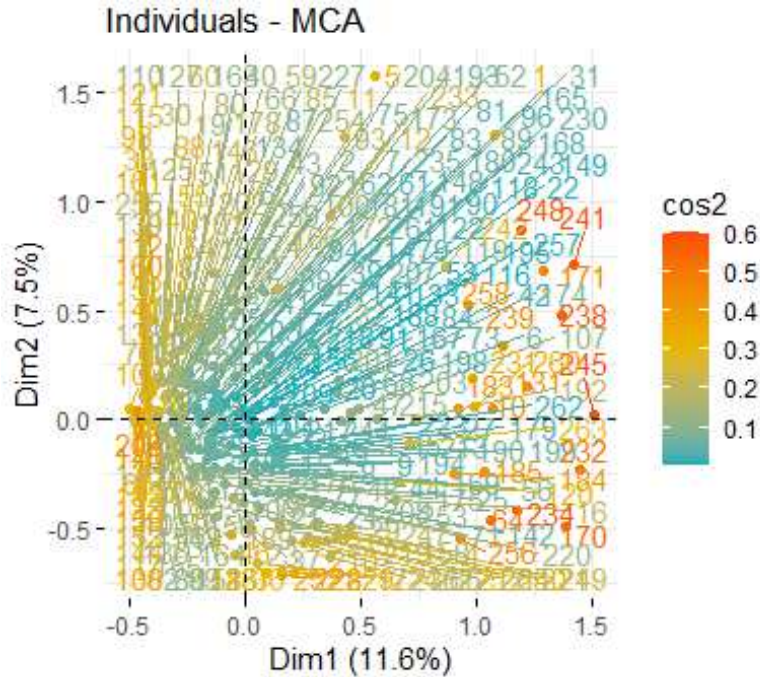


On remarquons que :

Les individus qui contribuent (-)	Les individus qui contribuent (+)
17, 36, 186, 212, 210	5, 1, 11, 171, 241

La qualité de représentation des individus sur l'axe 2 A chaque fois qu'on s'éloigne du centre on voit une augmentation dans la qualité de représentation.

```
fviz_mca_ind(mca, repel = TRUE, col.ind = "cos2", gradient.cols
= c("#00AFBB", "#E7B800", "#FC4E07"), ggtheme = theme_minimal())
```



D'après le tableau de « la qualité de représentation des individus » obtenu par la commande :

```
round(M$ind$cos2[,2],1)
```

```
> round(M$ind$cos2[,2],1)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
0.2 0.1 0.0 0.2 0.3 0.0 0.0 0.0 0.0 0.0 0.2 0.2 0.0 0.0 0.0 0.0 0.2 0.0 0.1 0.0
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
0.1 0.0 0.2 0.0 0.1 0.0 0.0 0.0 0.1 0.1 0.1 0.0 0.0 0.0 0.1 0.3 0.2 0.1 0.0 0.1
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
0.0 0.0 0.1 0.0 0.0 0.3 0.0 0.1 0.0 0.2 0.1 0.0 0.0 0.0 0.2 0.0 0.0 0.0 0.1 0.1
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
0.0 0.1 0.0 0.1 0.0 0.0 0.1 0.0 0.0 0.1 0.1 0.1 0.1 0.0 0.1 0.0 0.0 0.0 0.0 0.1
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
0.0 0.0 0.0 0.0 0.2 0.0 0.0 0.1 0.0 0.0 0.0 0.1 0.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0
101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120
0.1 0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.1 0.0 0.0 0.0 0.0 0.0
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140
0.2 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1
141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160
0.0 0.0 0.0 0.0 0.0 0.2 0.0 0.0 0.0 0.1 0.0 0.0 0.0 0.0 0.0 0.0 0.0 0.1 0.0 0.0
161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
0.0 0.0 0.0 0.1 0.0 0.0 0.0 0.0 0.1 0.1 0.1 0.0 0.1 0.0 0.1 0.1 0.0 0.2 0.0 0.0
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200
0.0 0.0 0.0 0.0 0.0 0.3 0.0 0.0 0.2 0.0 0.0 0.0 0.1 0.1 0.0 0.1 0.0 0.0 0.0 0.0
201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220
0.0 0.2 0.0 0.1 0.0 0.0 0.0 0.1 0.0 0.3 0.1 0.3 0.0 0.0 0.0 0.0 0.0 0.0 0.2 0.1
221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
0.2 0.0 0.0 0.3 0.2 0.0 0.1 0.4 0.3 0.0 0.0 0.0 0.2 0.1 0.0 0.1 0.3 0.1 0.0 0.2
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260
0.1 0.1 0.0 0.0 0.0 0.0 0.1 0.2 0.1 0.0 0.0 0.4 0.1 0.1 0.0 0.1 0.0 0.1 0.1 0.0
261 262 263 264 265 266
0.0 0.0 0.0 0.0 0.2 0.1
```

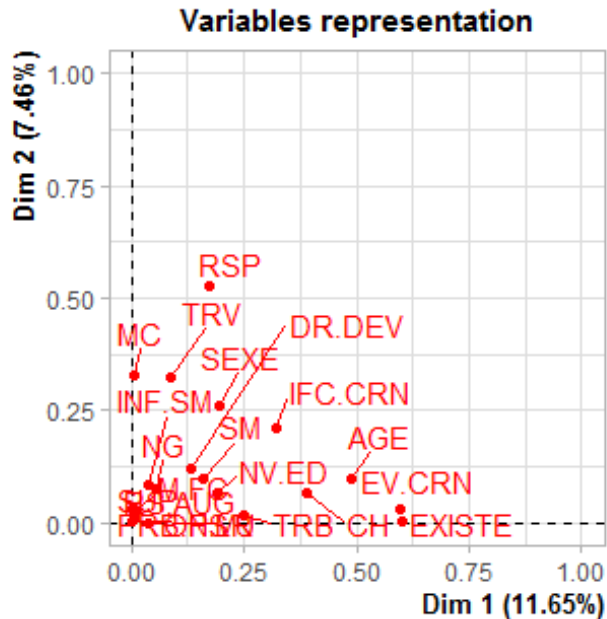
On remarque que :

- L'individu 5, 36, 186, 228, 237 est moyennement bien représenté sur l'Axe2.
- Toutes les autres individus actives sont plutôt mal représentées sur l'Axe2.

Conclusion On conclue que :

- Les hommes (H) qui sont travaillés libre (TR) ne respecte pas le quarantenaire (R) ,mais ils prennent des mesures préventives (EV+) pour éviter l'infection par le covid19.
- Les femmes (F) qui sont au chômage (TR-) , sortent uniquement lorsque cela est nécessaire .si elles toussent ou surchauffé (C-), elles vont à l'hôpital. Aussi elles sont croyer que l'effet négatif causés par la longue survie du Corona virus et sa propagation rapide est la Suicide (NG-).

Interprétation des variables pour interprété les variables on voir ce graphe :

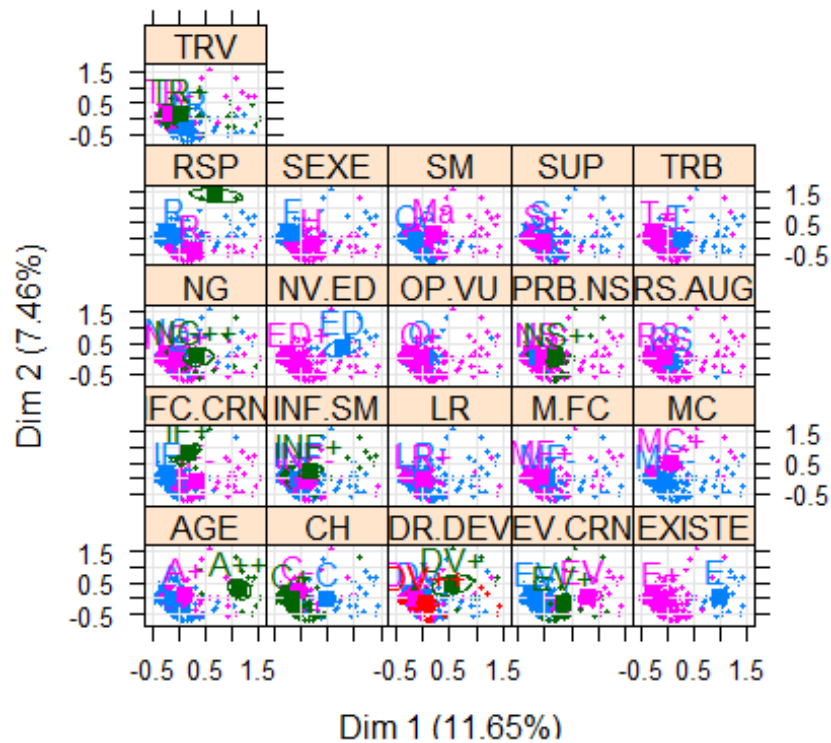


- Les variables EXISTE;EV.CRN;AGE ont une bonne corrélation avec le premier axe.
- Les variables RSP;MC;TRV sont bien corrélées avec le second axe.

- Les autres variables sont quant à elle pas corrélé avec les deux axes puisqu'elles ne participent pas du tout à la construction des axes, ce qui expliquera ce faible rapport de corrélation.

La réparation des individus en fonction des modalités : La fonction `plotellipses` colore les individus et trace des ellipses de confiance autour des modalités de variables qualitatives. L'objectif est de voir si les modalités d'une variable qualitative sont significativement différentes les unes des autres.

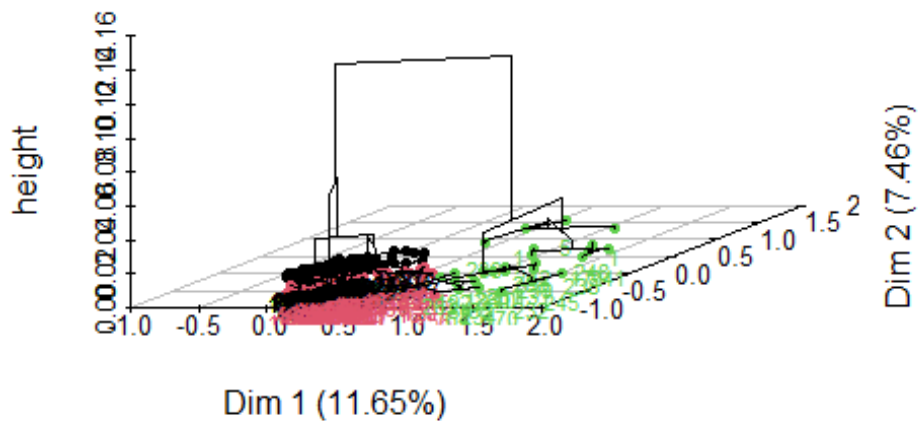
`plotellipses(M,keepvar=DONNEES)`



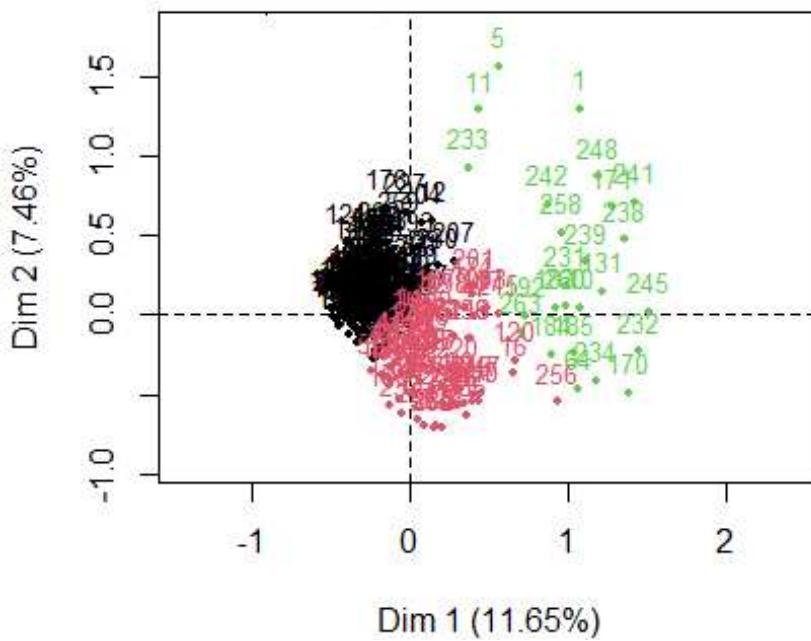
Classification :

`res.hcpc<-HCPC(M)`

Hierarchical clustering on the factor map



Factor map



Notre échantillon distingue entre 3 classes bien distinctes, chaque catégo-

rie inclut l'ensemble des individus dont le profil est presque le même :

- La première catégorie (en vert) possède Les gens âgées mariés qui sont arrêts leurs études au niveau primaire ou moyen. Ils croient que le covid19 n'existe pas, et ne sont pas d'accord avec les informations de la santé qui provenant des différentes sources (télévision, médias sociaux) sur ce virus, c'est pour ça ils ne respectent pas la quarantaine, et ne prend aucuns mesures préventives pour éviter l'infection par le virus peur d'infecter .de plus, lorsque ils toussent ou surchauffé, ils n'inquiètent pas et complètent les activités de la vie quotidienne.

- La deuxième catégorie (en noir) possède Les gens célibataires qui Souffrent d'une maladie chronique ou d'une immunité affaiblie. ils sont croyés que le virus corona (covid-19) existe, et ont ressenti des troubles ou de la peur lorsque le virus a commencé à se propager.Ces gens sont d'accord avec les informations de la santé qui provenant des différentes sources (télévision, médias sociaux) sur le développement le covid19. Ils représentent le groupe quit croient que l'effet nécessaire négatif du verus est le besoin. de plus, Ils ont respecté toute la période de quarantaine, mais à cause de la nature de leurs travaux(les travailles libres), ils doivent sortir lorsque cela est nécessaire . mais lorsque ils doivent sortir, ils prennent des mesures préventives pour éviter l'infection par le corona virus parcequ'ils sont peur de l'infecter.

- La troisième catégorie (en rouge) possède les gens en chômage (des hommes et des femmes) qui ne souffrent pas d'une maladie chronique ou d'une immunité affaiblie . Ils ne respectent pas toute la période de quarantaine car ils ne sont pas peur d'infecter par le virus et ils ne prennent aucun des mesures préventives pour éviter l'infection par covid19. De plus, ces gens n'acceptent pas les informations de la santé qui provenant des différentes sources (télévision, médias sociaux).

Conclusion

- Le niveau d'éducation des gens du peuple algérien correspond au degré de prise de conscience dans la société algérienne, tel que les gens qui ont un niveau d'éducation élevé sont plus consciences que les gens qui ont un niveau d'éducation moyen.

- La prise de conscience des gens ne suffit pas de les adhérer à la quarantaine parce qu'il y a des gens doivent sortir pour travailler et apporter ses besoins quotidiens. Donc la besoin peut reçoit une des causes de la propagation du covide19.

- La jeunesse du peuple algérien est plus instruit et consciente que Les personnes âgées.
- Le peuple algérien croit que la quarantaine est la meilleure façon d'affronter le covid-19.

Conclusion générale

Dans ce cours, nous avons étudié les différents types des méthodes d'échantillonnage auxiliaires. Nous avons montré leurs intérêts d'obtenir des conclusions et des informations sur la population cible. De plus nous avons étudié les notions préliminaires et la théorie d'échantillonnage et les techniques utilisées dans chaque type des méthodes probabilistes ou non probabilistes. Ces derniers sont intéressants puisqu'ils permettent que l'échantillon ainsi sélectionné représente l'ensemble de la population.

Nous avons montré le rôle d'échantillonnage.

Nous avons pu voir des exemples précis fournissent un résumé utile pour aider au choix de la taille de l'échantillon représentatif pour prédire des informations sur les populations qui nous avons des difficultés de l'accéder à tous leur individus. Mais toujours ce type d'étude contient une marge d'erreur liée au fait qu'on interroge une fraction de la population.

Bibliographie

- [1] AgroParisTech,- Eléments de la théorie de l'échantillonnage-, dépt. MMIP- 28 septembre 2007- Echantillonnage- Robinrobin@agroparistech.fr-.
- [2] Adil EL MARHOUM- COURS D'ECHANTILLONNAGE ET ESTIMATIONS- ANNEE UNIVERSITAIRE 2014/2015- UNIVERSITE MOHAMMED RABAT FACULTE DES SCIENCES JURIDIQUES ECONOMIQUES
- [3] ET SOCIALES AGDAL.
- [4] BardinBahouayila- Cours de pratiques des enquêtes- Submitted on 2 Aug 2016, HAL Id : cel-01317629- [https ://hal.archives-ouvertes.fr/cel-01317629](https://hal.archives-ouvertes.fr/cel-01317629).
- [5] BOURAINE née BERDJOU DJ Louiza- Statistique Inférentielle-mi-lessons06-stat-inferentielle.
- [6] Claire Durand- Méthodes de sondage - SOL3017 et SOL 6448- 5 mars 2009- Département de sociologie, Université de Montréal.
- [7] David Haziza- Echantillonnage STT-2000- le 10 Août 2008- Département de mathématiques et de statistique Université de Montréal.
- [8] Dr ALLAOUA- Technique d'Echantillonnage et Analyse des Données // - Cours destiné aux étudiants 3ème année licence / - Ecologie en sciences environnementales - / polycopier-Allaoua-Noua-Technique-d'échantillonnage-pdf.
- [9] Gumuchian, Claude Marois- Chapitre 6. Les méthodes d'échantillonnage et la détermination de la taille de l'échantillon : p. 265-294/Hervé-INITIATION A LA RECHERCHE EN GEOGRAPHIE.
- [10] Joëlle Berrewaerts- EDUS 1101 - Méthodologie de l'observation-berrewaerts-methodologie-observation-0.
- [11] JY Baudot/ [http ://www.jybaudot.fr/Analdonnees/acm.html](http://www.jybaudot.fr/Analdonnees/acm.html).
- [12] Jonathan Lenoir (MCU)- Introduction à la statistique inférentielle-jonathan.lenoir@u-picardie.fr, Unité "Ecologie et Dynamique des Systèmes Anthropisés" [http ://www.u-picardie.fr/edysan/](http://www.u-picardie.fr/edysan/).
- [13] Jean-Marc Meunier- Inférence Statistique : Résumés et exercices-R2442T.

- [14] Jonathan Lenoir- Introduction à la statistique inférentielle-(MCU), jonathan.lenoir@u-picardie.fr /-Unité "Ecologie et Dynamique des Systèmes Anthropisés" /- <http://www.u-picardie.fr/edysan>.
- [15] Jean VAILLANT –Initiation à la théorie de l'échantillonnage- Octobre 2005.
- [16] Jean-Jacques Daudin- Quelques éléments de la théorie de l'échantillonnage- AgroParisTech, 2009.
- [17] <http://www.statcan.gc.ca/edu/power-pouvoir/ch13/5214895-fra.htm>.
- [18] Hababou- Département SEG- Les méthodes d'échantillonnage- Cours statistique inférentielle-53340ca9cc969.
- [19] KHERRI Abdenacer- ECOLE DES HAUTES ETUDES COMMERCIALES - Support pédagogique de cours N° 02 :Echantillonnage-6-2 -(www.sgehec.jimdo.com)-.
- [20] Laurent SMOCH- PROBABILITES ET STATISTIQUE INFERENTIELLE- Septembre 2016-/(smoch@lmpa.univ-littoral.fr)-probas.
- [21] M.Mrabet- L'ECHANTILLONNAGE- L'ECHANTILLONNAGE.
- [22] NEKRI- :Recueil et Traitement Statistique des Données : Introduction Générale à la Statistique-cours-RTSD.
- [23] Nathalie Cheze- Statistique inférentielle I - Estimation/- July 9, 2007- courseestimaion-2007-1183977405151.
- [24] Philippe Letourmy, Cirad- décembre 2017- Recueil planifié des données : compléments sur l'échantillonnage- ComplémentEchantillonnage-RPD5.
- [25] P. Ribereau- Cours de Statistiques Inferentielles-6 janvier 2016.
- [26] Ricco.Rakotomalala.<http://eric.univ-lyon2.fr/~ricco/cours>.
- [27] Sara Elder-Module 3 : Sampling methodology- ILO school-to-worktransition survey : A methodological guide.
- [28] S. Robin- Eléments de la théorie de l'échantillonnage-28 septembre 2007- du livre "Statistique Inférentielle", Daudin, R., Vuillet (2001) ,robin@agroparistech.fr.AgroParisTech, d'épt. MMIP.
- [29] Tenenhaus, " Statistique -Méthodes pour décrire, expliquer et prévoir ",Dunod, 2006 ; pages 253 à 264, partie théorique ; pages 266 à 276, pour l'exemple que nous traiterons.
- [30] cours4(Cours 4 :Statistique inférentielle ?chantillonnage).
- [31] ANNEXE-1-Mthodologie-du-sondage (ANNEXE 1 :Méthodologie sondage).
- [32] c3a9chantillonnage(Echantillonnage non probabiliste Source/ <http://www.statcan.gc.ca/pouvoir/ch13/nonprob/5214898-fra.htm>2).

-
- [33] Methodes d'échantillonnage-Stat-Can (<http://www.statcan.gc.ca/edu/power-pouvoir/ch13/5214895-fra.htm>).
- [34] module14(L'échantillonnage).
- [35] STA108Sondages/ MagalieSenaux- cours n°10 :01DEC2017- sta108-37-cours-07-enquetes-par-quotas-cle8e4d22.
- [36] Chapitre 3 :Distributions d'échantillonnage- échantillonnage.
- [37] Chapitre 4 : DISTRIBUTIONS D'ECHANTILLONNAGE Et INTERVALLES DE VARIATION /- Bases de la statistique inférentielle- chapitre4-sec1-1134399360300.
- [38] inference statistique (L'inference statistique).
- [39] Statistique-ch1 (Chapitre 1 : L'ECHANTILLONNAGE).
- [40] <https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch13/prob/5214899-fra.htm>.
- [41] <http://www.statcan.gc.ca/edu/power-pouvoir/ch13/5214895-fra.htm>.