كلية العلوم و التقنيات فاس
ⵜⴰⵖⵚⵓⵍⵜ ⵏ ⵜⵇⴰⵙⵙⵉⵢⵏ ⴷ ⵜⵙⵙⵏⵜⵉⵏ
Faculté des Sciences et Techniques de Fès

جامعة سيدي محمد بن عبد الله
ⵜⴰⵙⴷⴰⵡⵉⵜ ⵙⵏⵙ ⵇⴰⵇⵇⴰ ⵙⵉ ⵎⵓⵃⴰⵎⴻⴷ
Université Sidi Mohamed Ben Abdellah

Master Mathématique et Application au Calcul Scientifique

(MACS)

MEMOIRE DE FIN D'ETUDES

Pour l'obtention du Diplôme de Master Sciences et Techniques (MST)

Les statistiques dans la recherche médicale

Réalisé par : MARZOUQ Aïcha

Encadré par: Pr. AMMOR Ouafae

Soutenu le 12/06/2018

Devant le jury composé de:

- **Pr. Ammor Ouafae**
- **Pr. Ezzaki Fatima**
- **Pr. Oudghiri Anisse**
- **Pr. Rahmouni Hassani Aziza**

Année Universitaire 2017 / 2018

FACULTE DES SCIENCES ET TECHNIQUES FES – SAISS

B.P. 2202 – Route d'Imouzzer – FES

Remerciements

Après avoir rendu grâce à Dieu le tout Puissant et le Miséricordieux, je tiens à exprimer ma profonde gratitude à notre chère professeur et encadrant Pr. AMMOR Ouafae pour son suivi et pour son énorme soutien, qu'il n'a cessé de me prodiguer tout au long de la période du projet.

J'adresse mes vifs remerciements aux membres des jurys pour avoir bien voulu examiner et juger ce travail.

Mes remerciements vont aussi à tout le personnel que j'ai contacté durant mon travail de recherche

Finalement, je remercie chaleureusement ma famille pour la confiance, l'amour et les encouragements.

Dédicaces

Je dédie mon travail à tous les malades et les patients dans la terre, et surtout à ceux qui souffrent d'un cancer, et je souhaite à eux une vie saine et heureuse.

A mes chers parents

A ma famille

A mes amies

Sommaire :

Remerciements 2

Dédicaces 3

Introduction : 7

Partie 1: Statistique descriptive et recherche médicale 9

I. Notions de base 10

II. Exemple de la statistique descriptive dans la recherche médicale 11

III. Conclusion 14

Partie 2 : Statistique inférentielle et recherche médicale 15

I. Echantillonnage 16

1) Estimation ponctuelle 16

1.1) Principe 16

1.2) Exemples 16

2) Estimation par intervalle de confiance 17

2.1) Principe 17

2.2) Exemple : Application de la SI sur les données médicales 17

II. Tests statistiques 18

1) Tests paramétriques 19

1.1) Test t de Student 19

1.1.1) Test de Student pour échantillon unique et exemple 19

1.1.2) Test t de Student pour échantillons indépendants et exemple 21

1.1.3) Test-t de Student pour échantillons appariés 23

1.2) Test de khi deux $\chi 2$ 24

1.2.1) principe 24

1.2.2) Exemple dans la recherche médicale 24

1.3) L'analyse de variance l'ANOVA 25

1.3.1) principe 25

1.3.2) Exemple 26

2) Tests non paramétriques 28

2.1) Test de normalité (Shapiro-Wilk) 29

2.1.1) Principe 29

2.1.2) Exemple 29

2.2) Test de Levene (test d'homogénéité des variances) 30

2.2.1) Principe 30

2.2.2) Application 31
2.3) Test de Wilcoxon (le test des rangs signés de Wilcoxon) 32
2.3.1) Principe 32
2.3.2) Exemple 33
2.4) Test de Kruskal-Wallis 34
2.4.1) Principe 34
2.4.2) Exemple 35
3) Conclusion 35
Partie 3 : Analyse statistique multidimensionnelle et recherche médicale 36
I. Analyse des données 37
1) ACP (Analyse en Composante Principales) 37
1.1) Principe 37
1.2) Application et interprétation de l'ACP en recherche médicale 39
2) La classification 42
2.1) La classification ascendante hiérarchique(CAH) 43
2.2) La classification non hiérarchique 43
2.3) Application de Classification 44
2.4) Conclusion 45
II. Segmentation d'images médicales 46
1) Approche par classification 46
1.1) Principe 46
1.2) Application en recherche médicale 47
2) Approche contours 48
III. Le big data (Analyse des données en masse) 49
1) Définition 49
2) Traitement big-data 50
3) Le big data dans la recherche médicale 51
3.1) Les différentes applications du Big Data dans la recherche médicale 52
3.2) Le big data pour une médecine préventive 53
4) Conclusion 54
Conclusion générale 55
Bibliographie et web-graphie 56

Liste des figures

Figure 1 : le diagramme en boite de l'IMC 13
Figure 2 :l'histogramme de l'IMC 14
Figure 3 : graphe de centile pour les trois groupes de scolarisation :A,B et C 28
Figure 4 : les deux histogrammes des poids des 14 patientes avant et après traitement 33
Figure 5 : Le nuage de points (femmes), cas (oui) et témoin (non) 40
Figure 6 : le graphe des variables de l'ACP(cercle de corrélation) 41
Figure 7 : la classification ascendante hiérarchique des patientes (dendrogramme) 44
Figure 8 : les clusters obtenus par une classification non hiérarchique (k-means) 45
Figure 9 : Le coût pour séquencer un génome humain a été divisé par 1 00000 en 15 ans, et est proche d'atteindre la barre symbolique des 1 000 dollars. 52

Liste des tableaux

Tableau 1 : interprétation de la valeur de l'IMC chez les femmes 12
Tableau 2 : le IMC (kg/m²) des 53 patientes 12
Tableau 3 : le tableau de différents cas possibles pour estimer un paramètre en intervalle de confiance 17
Tableau 4 : l'MC des patientes selon le milieu rural(r) ou urbain(u) 22
Tableau 5 : récapitulation de nombre de patientes selon les deux modalité : prendre des contraceptifs oraux et subir un faux accouchement 24
Tableau 6 : le tableau Anova 26
Tableau 7 : l'IMC des patientes selon le niveau de scolarisation 27
Tableau 8 : le poids des patientes avant et après traitement 30
Tableau 9 : le poids, la taille et l'âge des 14 patientes qui ont un cancer du sein 31
Tableau 10 : les données des patientes pour faire une ACP 39

Introduction :

Les statistiques sont une discipline mathématique, qui est en partie liée dans sa partie théorique, avec les probabilités, les deux disciplines forment ce qu'il est convenu d'appeler les sciences de l'aléatoire. Quand on utilise les statistiques dans le domaine médical, elles deviennent des statistiques biomédicales.

En fait les statistiques sont l'ensemble des méthodes scientifiques à partir desquelles on recueille, organise, résume, présente et analyse des données, à fin d'en tirer des conclusions et de prendre des décisions judicieuses, la statistique est utilisée dans tous les domaines et notamment dans la médecine et c'est ce qui constitue notre sujet de fin d'études. Elle peut être utilisée dans la comptabilité des équipements (lits, hôpitaux…) ou des agents (médecins, infirmiers…) ou des consommations (hospitalisations, consultations, médicaments…) ou encore des maladies et des morts**.** Elles sont à la base de :

- La recherche clinique comme la mortalité : nombre de décès survenus pendant une période donnée, au sein d'une population étudiée, en relation avec une maladie déterminée. On peut la calculer selon différents paramètres (tranches d'âge, sexe…).
- L'épidémiologie comme la morbidité (tout ce qui est relatif à la maladie étudiée)…

La statistique doit prendre en considération *la variabilité individuelle* (la variation de la valeur mesurée d'une population ou d'un échantillon) que se soit dans le cas de :

- Variables quantitatives : cette variabilité peut être due aux erreurs de mesures, ou à des variabilités *entre sujets* (les patients ont des poids, des tensions artérielles,…différents.) ou à des variabilités *intra-sujets* (si nous mesurons la tension artérielle d'un même patient à différents moments de la journée ou au même moment mais plusieurs jours de suite, nous obtiendrons des valeurs différentes).
- Ou variables qualitatives: les patients n'ont pas les mêmes statuts (diabètes ou non diabètes).

Donc la variabilité individuelle crée une variabilité au niveau des échantillons (la composition de deux échantillons n'est en général pas la même) ce phénomène qualifié de fluctuations d'échantillonnages est très important dans le domaine biomédical car il est extrêmement fréquent de travailler sur des échantillons. [1]

Mon rapport traite l'étude théorique des deux grands titres de recherche statistique qui sont :

- ✓ Premièrement la statistique *"déductive" ou descriptive* : c'est la statistique de base pour toute étude statistique, elle a pour but de résumer et de présenter les données

observées pour que l'on puisse en prendre connaissance facilement : tableaux, graphiques…

- ✓ Deuxièmement la statistique *"inductive"* ou *inférentielle* : elle permet d'étendre ou de généraliser, dans certaines conditions, les conclusions obtenues. Cette phase comporte certains risques d'erreur qui peuvent être mesurés en faisant appel à la théorie des probabilités. Elle concerne l'échantillonnage et les tests statistiques paramétriques et non paramétriques. Dans l'échantillonnage, l'estimation des paramètres peut être ponctuelle ou par intervalle (appel à des intervalles de confiance).Les tests statistiques représentent un outil indispensable dans la recherche médicale, ceci est pré détaillé et illustré par des exemples dans mon rapport.

  L'étude théorique est suivie d'applications sur des exemples de base de données collecté à partir *d'une étude cas-témoins :* étude rétrospective utilisée pour mettre en évidence des facteurs qui peuvent être incriminés dans l'apparition d'une maladie en comparant des individus qui ont cette maladie (les cas) à d'autres qui en sont indemnes, mais similaires par ailleurs (les témoins ou contrôles)(Par exemple : ce type d'étude a permis de démontrer le lien entre le tabagisme et le cancer du poumon).
- ✓ Un troisième volet de mon rapport traite de façon plus vaste et plus concrète l'utilisation et l'importance des statistiques en général dans le domaine médical composé de :
  - ❖ L'analyse des données (ACP)
  - ❖ La classification hiérarchique et non hiérarchique
  - ❖ Domaine de l'imagerie médicale (segmentation d'image)
  - ❖ Big-data

  Cette partie est concrétisée par la base de donnée de 53 patientes présentes dans un séjour hospitalier à CHU Fès qui sont soit des cas de cancer du sein soit des cas témoin, et je vais traiter les résultats obtenus.
- ✓ Et je finis mon rapport par une conclusion générale résumant l'importance d'utilisation de la statistique dans la recherche médicale, et ouvrant des perspectives pour profiter de cette vaste discipline.

Partie 1: Statistique descriptive et recherche médicale

La statistique descriptive constitue la base de la recherche biomédicale pour un clinicien ou pour un biologiste (laboratoire) ou pour n'importe quel chercheur dans le domaine médicale.

I. Notions de base

La statistique descriptive contient l'ensemble des termes, des définitions et des formules que nous résumons de la manière suivante:

- Une variable statistique est un caractère qui fait le sujet d'une étude statistique dans une *population* ou dans *un échantillon* de **n** individus observé, soumis à une analyse statistique (par exemple le taux des nouveau-nés à CHU Fès), Une variable statistique est dite :
  - *Qualitative* (codage) si elle n'est pas mesurable, une classe contient tous les individus ayant la même modalité. Nous distinguons deux groupes de variables qualitatives :
    - ✓ une variable *qualitative nominale* dont les classes ne sont pas ordonnées (ex: groupe sanguin).
    - ✓ une variable *qualitative ordinale* dont les classes sont ordonnées ex: les classes de l'IMC (voir le chapitre suivant).
  - *Quantitative* si elle est mesurable, Nous distinguons deux groupes de variables quantitatives discrètes ou continues :
    - ✓ Une variable statistique est discrète (discontinue) lorsqu'elle ne peut prendre que certaines valeurs xi, exemple : (le nombre de patients qui ont une hypertension, le poids des enfants juste après la naissance…) une classe contient tous les individus ayant la même valeur du caractère (les patientes qui ont pris des contraceptifs oraux, et les patientes qui ne l'ont pas pris).
    - ✓ Une variable statistique est continue lorsqu'elle peut prendre toutes les valeurs d'un intervalle fini ou infini. une classe est un intervalle représentée par son centre, qui est le milieu de l'intervalle.
- *L'effectif* : le nombre d'individus de la classe, il est noté $\mathbf{n_i}$ (i est l'indice de la classe).
- *La fréquence* : la proportion d'individus de la population ou de l'échantillon appartenant à la classe, elle est notée $f_i=\frac{n_i}{n}$
- *La fréquence cumulée* : c'est la fréquence de la classe augmentée de celles des classes précédentes.
- *L'étendue :* représente la différence entre les valeurs extrêmes de la distribution.

- *La variance* : notée V(x) pour les variables quantitatives uniquement. C'est un indicateur de la dispersion des valeurs des individus autour de la moyenne. La variance est la moyenne des carrés des écarts à la moyenne ; sa racine carrée est *l'écart-type* noté σ.

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{(si la série est non classée)}$$

$$V(x) = \frac{1}{n}\sum_{i=1}^{k} n_i(x_i - \bar{x})^2 \quad \text{(si la série est classée)}$$

- Le mode : la valeur de la variable statistique la plus fréquente (Dans le cas d'une variable statistique continue, on parle plutôt de classe modale)
- Médiane(Me) : pour une variable ordinale ou quantitative, c'est la valeur qui correspond à 50% de l'effectif rangé par valeurs croissantes. Diffère de la moyenne.
- Moyenne (μ) : pour une variable quantitative, c'est la valeur uniforme que devrait présenter chaque individu d'une population ou d'un échantillon pour que le total de l'ensemble reste inchangé.
- La moyenne arithmétique : s'obtient en divisant la somme des valeurs par l'effectif. La moyenne est une statistique dite de tendance centrale.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{(si les observations ne sont pas groupées)}$$

$$\bar{x} = \frac{\sum_{i=1}^{k} n_i x_i}{n} \quad \text{(sinon)}$$

- Quartile (q): même définition que pour la médiane, les valeurs de la variable étant classées à 25%, 50% et 75%
- Diagrammes : Ils jouent le rôle de visualiser la répartition des individus graphiquement, par exemple l'histogramme joue un rôle très important pour estimer la distribution de la variable étudiée.[2]

II. *Exemple de la statistique descriptive dans la recherche médicale*

Définition :

L'**Indice de Masse Corporelle** (IMC) est un outil communément utilisé comme point de référence afin de déterminer si une personne a un surplus de poids. Le principe de base est qu'une personne ayant un surplus de poids soit plus à risque de développer divers problèmes de santé (cholestérol, diabète, hypertension et maladies cardiovasculaires). L'IMC est également utilisé pour déterminer si une personne a un poids anormalement bas, comme dans le cas de l'anorexie.

L'équipe médicale de l'hôpital CHU de Fès a fait un questionnaire sur différentes variétés chez n=53 femmes patientes dans un séjour hospitalier, et nous avons collecté les données nécessaires, alors nous voulons faire une étude statistique sur le IMC des patientes dont l'interprétation se fait comme nous indique le tableau suivant [3] :

IMC ($kg.m^{-2}$)	**Interprétation**
≤16.5	Dénutrition
16.5 à 18.5	Maigreur
18.5 à 25	Corpulence normale
25 à 30	Surpoids
30 à 35	Obésité modérée
35 à 40	Obésité sévère
≥40	Obésité morbide ou massive

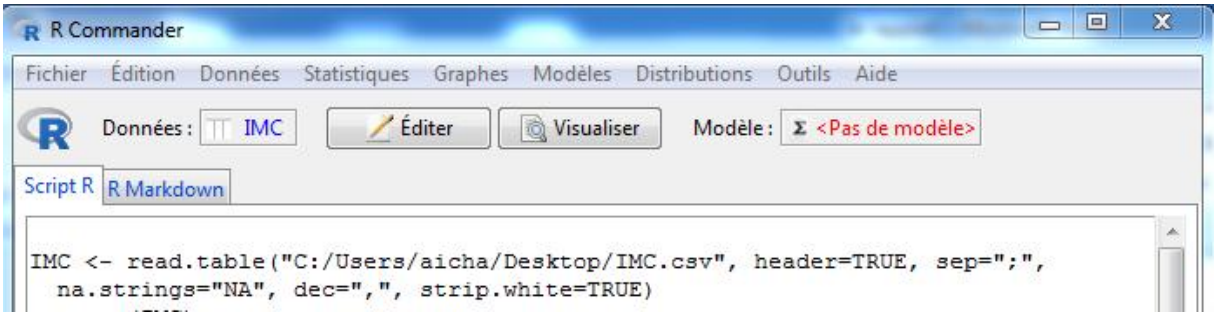
Tableau 1 : interprétation de la valeur de l'IMC chez les femmes

Nous calculons le IMC=$\frac{\text{poids(kg)}}{(\text{la taille(m)})^2}$ sur Excel ou sur le logiciel R nous obtenons les résultats suivants :

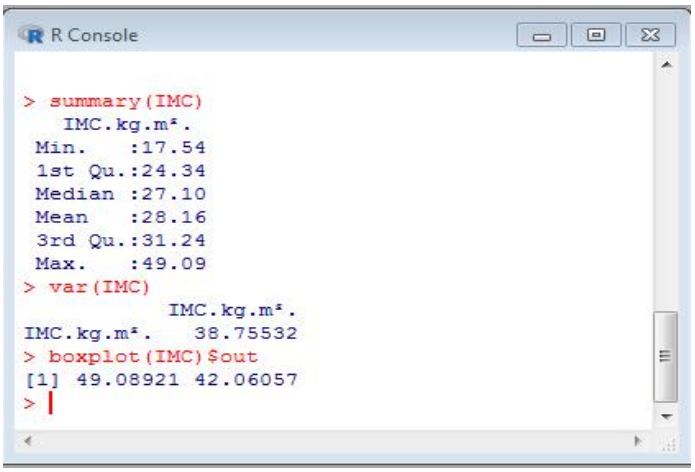
	IMC.kg.m².		
1	24.08822	27	32.89329
2	27.94214	28	40.79016
3	25.10388	29	31.25000
4	25.63692	30	27.76710
5	27.43484	31	23.14050
6	49.08921	32	32.03896
7	18.77834	33	34.54735
8	25.91068	34	27.54821
9	28.71972	35	32.87071
10	25.55933	36	25.71101
11	25.21736	37	17.54309
12	30.09143	38	21.00399
13	24.34176	39	27.09925
14	42.06057	40	33.20312
15	36.14744	41	30.96266
16	21.23057	42	19.53125
17	26.47211	43	31.61579
18	21.09375	44	37.16563
19	26.03749	45	31.23859
20	41.52249	46	28.60476
21	31.11111	47	24.30462
22	22.09317	48	26.56250
23	26.83865	49	27.68878
24	27.09925	50	30.11028
25	23.22543	51	19.81405
26	24.84098	52	23.62445
		53	26.10656

Tableau 2 : le IMC (kg/m^2) des 53 patientes

Les instructions utilisées dans le logiciel R pour importer des données appelées IMC à partir d'un fichier CSV (séparateur point-virgule) sont:



Nous obtenons les résultats suivants :



les résultats désignent que la moyenne de IMC chez les patientes est environ 28.16 $kg.m^{-2}$, le IMC minimal est 17.54 $kg.m^{-2}$, le maximal est 49.09 $kg.m^{-2}$, de plus de cela la médiane divisant le jeu de données en deux groupes de patientes de taille égale est 27.10 $kg..m^{-2}$, La dernière commande dans le 'R-console' boxplot permet de résumer graphiquement la distribution d'un échantillon, elle nous permet de détecter les valeurs aberrantes qui sont : 49.08921 $kg.m^{-2}$ (patiente n°6) et 42.06057 $kg.m^{-2}$(patiente n°14), ces paramètres sont résumés dans la figure suivante:

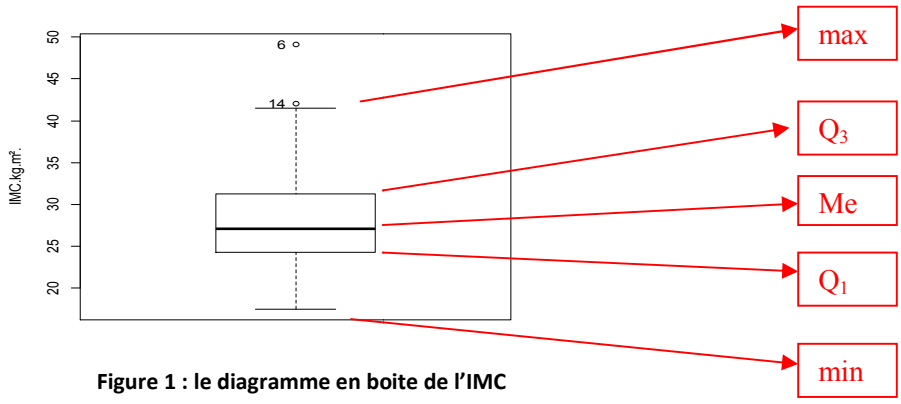


Figure 1 : le diagramme en boite de l'IMC

Alors les deux patientes correspondant ont une obésité morbide ou massive, elles nécessitent un régime urgent et spécial.

Pour visualiser les données dans un histogramme, on utilise la commande suivante

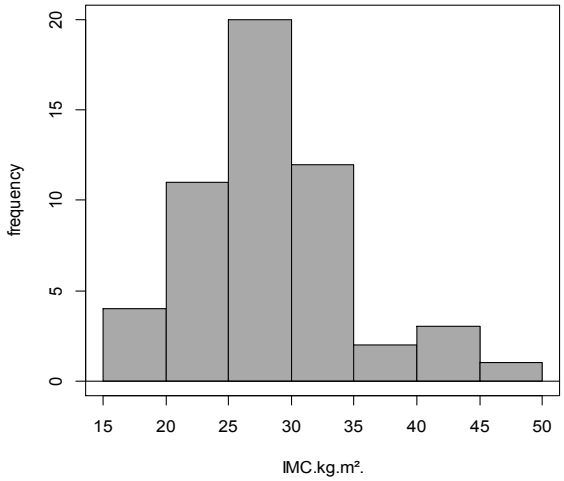


Figure 2 :l'histogramme de l'IMC

C'est un histogramme qui donne les fréquences d'IMC selon les classes en abscisses alors la classe modale est [25 ; 30], l'allure de l'histogramme nous permet de dire que la distribution tend à suivre la loi normale.

III. Conclusion

L'utilisation de la statistique descriptive dans la recherche médicale permet de déterminer les différents paramètres de base de toute étude statistique, qui sont soit des paramètres de position: la moyenne, la médiane, le mode, les quartiles…, ou des paramètres de dispersion: l'étendu, la variance, l'écart type….

La statistique descriptive permet de diviser les données (effectifs) en classes et calculer les fréquences associées pour avoir une représentation graphique (histogrammes,…) et un résumé plus pratique de l'information (diagramme en boites…) pour éviter les listes ennuyeuses.

Partie 2 : Statistique inférentielle et recherche médicale

Dans l'interprétation d'un essai thérapeutique, la signification statistique est un élément important qui assure que le résultat obtenu a de forte chance d'être réel et non pas d'être le fruit du hasard. C'est le but de la statistique inférentielle, cette dernière constitue un volet très utilisé en milieu médicale grâce aux différents résultats qu'on peut obtenir à travers l'échantillonnage et les tests statistiques appropriés et leurs interprétations.

I. Echantillonnage

Dans le domaine médical la situation la plus fréquente pour une étude statistique consiste à trouver une généralisation pour la population à partir d'un échantillon de taille n. c'est aussi le cas pour l'estimation qui vise à donner à partir des observations sur un échantillon la vraie valeur du paramètre dans la population. Et il y'a deux types d'estimation [1] :

1) Estimation ponctuelle

1.1) Principe

Elle fournit une valeur plus proche possible de la vraie valeur du paramètre notée θ.

On sait que des échantillons de composition différente peuvent être observées dans une même population et alors ils donnent deux estimations différentes, donc il y'a des conditions que doit vérifier un estimateur ponctuel (la statistique à calculer notée $\hat{\theta}$) qui sont :

- L'absence de bais : signifie que les estimations obtenues sur des échantillons successifs ne s'écartent pas de la vraie valeur de façon systématique ie $E(\hat{\theta}) \rightarrow \theta$
- Une variance faible : indique que les estimations sont peu dispersées ie $Var(\hat{\theta}) \rightarrow 0$

Par exemple :

$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ est un estimateur sans biais et convergent d'une moyenne μ

$S^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}$ est un estimateur sans biais et convergent d'une variance σ^2

$f = \frac{n_{obs}}{n}$ est un estimateur sans biais et convergent d'une proportion p (avec n_{obs} le nombre d'individus ayant le caractère).

1.2) Exemples

- Dans l'exemple précédent de l'IMC nous avons $\bar{x}$=28.16 est une estimation de la moyenne de l'indice de masse corporelle des patientes qui souffrent des mêmes symptômes, et l'écart de l'IMC autour de la moyenne est estimée par S^2=39.500614.
- Sur un échantillon de 53 patientes à CHU Fès nous avons 14 patientes ont un cancer du sein alors un estimateur de la proportion du cancer du sein est donnée par : f=14/53.

2) Estimation par intervalle de confiance

2.1) Principe

Plutôt qu'une seule valeur, la prise en compte de l'incertain permet de déterminer un intervalle à l'intérieur duquel on a une certaine probabilité de se situer et donc un risque α (fixé à priori) de ne pas y être, d'où la notion d'intervalle de confiance : l'intervalle de confiance à 1- α est un intervalle qui a une probabilité de 1- α de contenir la vraie valeur (inconnue) du paramètre. [4]

Le tableau suivant résume les différentes statistiques utilisées pour en extraire les estimateurs des paramètres en intervalle de confiances dans le domaine biomédical :

Paramètre à estimer	Loi de la population		Statistique	Loi
Moyenne μ	Normale	σ^2connu	$(\bar{x}-\mu)/(\sigma/\sqrt{n})$	N(0,1)
		σ^2inconnu	$(\bar{x}-\mu)/(S/\sqrt{n})$	Student à n-1 d.d.l
	Qlq n>30	σ^2connu	$(\bar{x}-\mu)/(\sigma/\sqrt{n})$	N(0,1)
		σ^2inconnu	$(\bar{x}-\mu)/(S/\sqrt{n})$	N(0,1)
Variance σ^2	Normale	μconnue	$\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{\sigma^2}$	χ^2 à n d.d.l
		μ inconnue	$\frac{(n-1)s^2}{\sigma^2}$	χ^2 à n-1 d.d.l
Proportion p	n>50		$(f-p)/(\sqrt{f(1-f)}/\sqrt{n})$	N(0,1)

Tableau 3 : le tableau de différents cas possibles pour estimer un paramètre en intervalle de confiance

2.2) Exemple : Application de la SI sur les données médicales

- Revenons à l'exemple de l'IMC

On a : $\bar{x} - Z_{1-\alpha/2} \times S/\sqrt{n} \leq \mu \leq \bar{x} + Z_{1-\alpha/2} \times S/\sqrt{n}$

Pour une confiance de 95% la valeur correspondante est $Z_{1-0.05/2} = Z_{0.975} = 1.96$, et sur R nous obtenons :

```
> n=53
> z=1.96
> m=28.16
> v=38.755532
> BI=m-z*sqrt(v/n)
> BS=m+z*sqrt(v/n)
> BI
[1] 26.48396
> BS
[1] 29.83604
```

Les valeurs de l'IMC ont 95 chances sur 100 d'être comprises entre 26.48396 et 29.83604.

- Par exemple sur notre échantillon de 53 patientes, nous avons observé 14 qui ont un cancer du sein, donc la proportion de l'échantillon est f=14/53.

En développant la dernière ligne du tableau nous obtenons l'intervalle de confiance de la proportion suivant : $f-Z_{1-\alpha/2}\times\sqrt{f(1-f)}/\sqrt{n} \leq p \leq f+Z_{1-\alpha/2}\times\sqrt{f(1-f)}/\sqrt{n}$

Pour une confiance de 95% la valeur correspondante est $Z_{1-0.05/2}= Z_{0.975}=1.96$, et sur R nous obtenons :

```
> n=53
> f=14/53
> z=1.96
> BI=f-1.96*sqrt(f*(1-f)/n)
> BS=f+1.96*sqrt(f*(1-f)/n)
> BI
[1] 0.1454543
> BS
[1] 0.3828476
> 
```

On conclut donc que le risque d'avoir un cancer du sein chez ces catégories des patientes a 95 chances sur 100 d'être compris entre 0,14 et 0,38.

II. Tests statistiques

Les tests statistiques fonctionnent tous sur le même principe qui consiste à énoncer une hypothèse nulle (H_0) hypothèse postulant une égalité entre deux données d'un modèle, elle est toujours testée contre une hypothèse alternative (H_1), qui postule soit la différence entre les données (test bilatéral), soit une inégalité entre elles (unilatéral à gauche ou à droite). [5]

- Le risque d'erreur : α = probabilité d'accepter H1 alors que H0 est vraie
- La p-value : à partir de la statistique de test, la p-value calcule le niveau de risque α réel du test. Une p-value de 0,03 se lie « On a 3% de chance de rejeter l'hypothèse nulle à tort ».
- La décision : on compare la p-value au risque α que nous avons choisi (généralement 1 ou 5%), et on l'interprète de la manière suivante :

  Si p-value ≤ α : on peut conclure qu'au niveau de risque p-Value%, de rejeter l'hypothèse nulle H_0 en faveur de l'hypothèse alternative H_1.

  Si p-value > α : on retient l'hypothèse nulle H0.

Cette étape consiste à réaliser le calcul du test le plus approprié à la situation, vérifier ses conditions d'application et à déduire de la table (habituellement celle de la loi normale) la probabilité recherchée.

1) Tests paramétriques

Les tests paramétriques se basent sur des distributions statistiques supposées dans les données. Par conséquent, certaines conditions de validité doivent être vérifiées pour que le résultat d'un test paramétrique soit fiable.

Les tests paramétriques sont adaptés à de multiples situations et doivent êtres privilégiés chaque fois que cela est possible. Leur utilisation facilite les calculs et les interprétations. Ils sont d'usage courant ce qui aide à la lecture des résultats d'une analyse. Ils sont les plus puissants si leurs conditions d'application sont remplies. Ils sont robustes aux faibles écarts à ces conditions.

Les tests paramétriques les plus utilisés dans la recherche médicale sont :

1.1) Test t de Student

Le test t de **Student** est un test paramétrique très utilisé dans la recherche médicale, mais il n'est fiable que si les données associées à chaque échantillon suivent une distribution normale et si les variances des échantillons sont homogènes. [6]

1.1.1) Test de Student pour échantillon unique et exemple

Il s'agit de comparer une moyenne observée (m) à une moyenne théorique (μ).

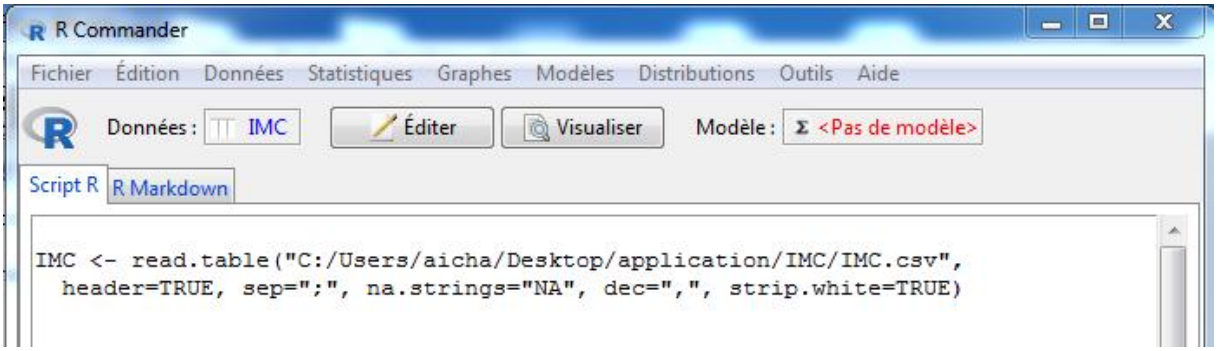
- Données : Soit X une série de valeurs de taille n, de moyenne m et d'écart-type S.
- Déroulement du test :

  La statistique du test est donnée par la formule $t=\frac{m-\mu}{\frac{S}{\sqrt{n}}}$

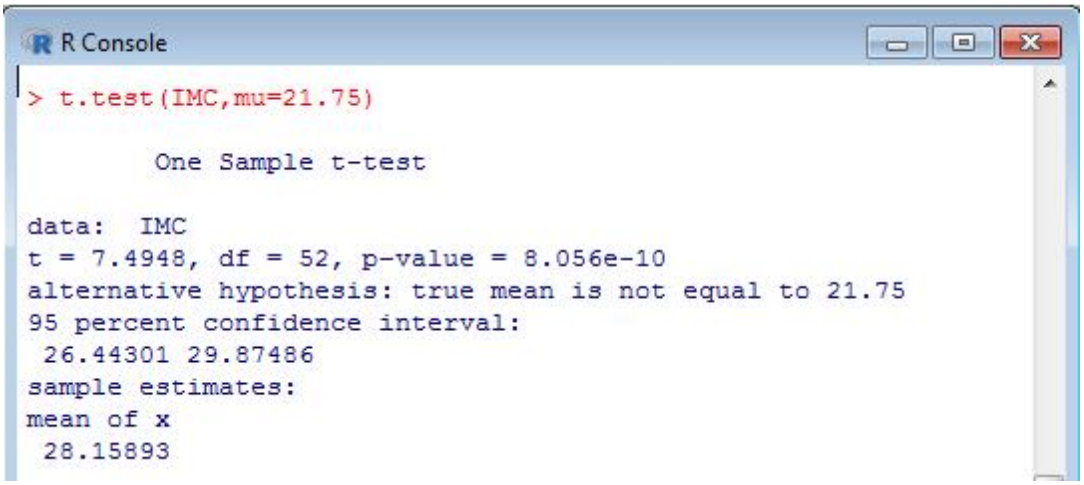
  Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au risque $\alpha = 5\%$ pour un degré de liberté :ddl=n-1

Exemple

Reprenons le même exemple et appliquons le test de Student pour comparer la moyenne observée de l'IMC à la valeur théorique μ=21.75=(18.5+25)/2 c'est la valeur normale de l'IMC.



Procédons par le test bilatéral, on a :



Interprétation : C'est un simple test pour échantillon unique, dont la p-value est très petite, donc H1 est significative contre H0 (on garde H1 et rejette H0).De plus on est confiant à 95% que les valeurs de l'IMC sont dans l'intervalle [26.87486 ; 29.87486], et un simple estimateur de la moyenne est donné par la moyenne arithmétique (elles sont les mêmes valeurs que nous avons trouvé précédemment).

Donc d'après le test de Student nous avons une différence significative entre la moyenne observée et la moyenne théorique, c-à-d le IMC est différent de la valeur normale mondiale. Pour préciser la relation exacte entre elles, faisons le test bilatéral à gauche :

```
> t.test(IMC,mu=21.75,alternative="less")

        One Sample t-test

data:  IMC
t = 7.4948, df = 52, p-value = 1
alternative hypothesis: true mean is less than 21.75
95 percent confidence interval:
   -Inf 29.591
sample estimates:
mean of x
 28.15893
```

Le p-value est grand donc nous n'avons pas la preuve significative contre l'hypothèse nulle, et nous sommes confiant à 95% que les valeurs sont inférieures à 29.591. Donc essayons avec une alternative à droite :

```
> t.test(IMC,mu=21.75,alternative="greater")

        One Sample t-test

data:  IMC
t = 7.4948, df = 52, p-value = 4.028e-10
alternative hypothesis: true mean is greater than 21.75
95 percent confidence interval:
 26.72687      Inf
sample estimates:
mean of x
 28.15893
```

Finalement c'est le test le plus significatif, et nous sommes confiant à 95% que les valeurs de l'IMC sont supérieures à 26.72687, donc l'IMC de nos patients dépasse la valeur normale mondiale.

1.1.2) Test t de Student pour échantillons indépendants et exemple

Le test t de Student est un test paramétrique utilisé pour comparer deux moyennes théoriques μ_1 et μ_2 de deux échantillons indépendants.

$H_0 : \mu_1 = \mu_2$ contre $H_1 : \mu_1 \neq \mu_2$ (ou $H_1 : \mu_1 < \mu_2$) (ou $H_1 : \mu_1 > \mu_2$)

Soient $n_1$et n_2 les tailles respectives des deux échantillons. La valeur t de Student est donnée par la formule suivante : $t = \frac{m_1 - m_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}}$

S^2 est la variance commune aux deux groupes, elle est calculée par la formule suivante :

$$S^2 = \frac{\sum (x_i - m_1)^2 - \sum (x_i - m_2)^2}{n_1 + n_2 - 2}$$

Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t la valeur critique correspondant au risque α = 5% pour un degré de liberté : ddl=n1+n2−2

Exemple :

Pour savoir s'il y'a une différence significative entre l'IMC des patientes selon leur milieu urbain/rural, nous procédons à un test de Student indépendant dont les groupes sont u et r, d'abord on importe les données par la commande suivante :

Script R | R Markdown

```
données <- read.table("C:/Users/aicha/Desktop/T-independant.csv",
  header=TRUE, sep=";", na.strings="NA", dec=",", strip.white=TRUE)
```

Et on visualise les données qui sont regroupé dans le tableau suivant :

donné...

	groupe	IMC
1	r	24.08822
2	r	27.94214
3	r	25.10388
4	r	25.63692
5	r	49.08921
6	r	18.77834
7	r	25.91068
8	r	28.71972
9	r	25.55933
10	r	25.21736
11	r	30.09143
12	r	42.06057
13	r	36.14744
14	r	21.23057
15	r	26.47211
16	r	21.09375
17	r	26.03749
18	r	41.52249
19	r	31.11111
20	r	22.09317
21	r	26.83865
22	r	27.09925
23	r	23.22543
24	r	24.84098
25	r	40.79016
26	r	23.14050
27	r	32.03896
28	r	32.87071
29	r	25.71101
30	r	27.09925
31	r	33.20312
32	r	30.96266
33	r	19.53125
34	r	31.61579
35	r	37.16563
36	r	31.23859
37	r	28.60476
38	r	24.30462
39	r	26.56250
40	r	27.68878
41	r	19.81405
42	r	23.62445
43	r	26.10656
44	u	27.43484
45	u	24.34176
46	u	32.89329
47	u	31.25000
48	u	27.76710
49	u	34.54735
50	u	27.54821
51	u	17.54309
52	u	21.00399
53	u	30.11028

Tableau 4 : l'MC des patientes selon le milieu rural(r) ou urbain(u)

Le test de Student indépendant se fait de la manière suivante :

```
> t.test(IMC~groupe, alternative='two.sided', conf.level=.95, var.equal=FALSE,
+     data=données)

        Welch Two Sample t-test

data:  IMC by groupe
t = 0.45432, df = 15.963, p-value = 0.6557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.231407  4.993827
sample estimates:
mean in group r mean in group u
       28.32520        27.44399
```

Donc la différence n'est pas significative, on ne peut pas juger l'IMC selon le milieu (rural/urbain) d'où viennent les patientes.

1.1.3) Test-t de Student pour échantillons appariés

L'hypothèse nulle se traduit par la différence moyenne des deux mesures est égale 0. Et comme son équivalent non apparié, il est valide si la différence des deux mesures suit une distribution normale, ou dans tous les cas quand le nombre de paires est suffisamment grand. [5]

Pour **comparer les moyennes de deux séries appariées**, on calcule tout d'abord la différence des deux mesures pour chaque paire, soit d la série des valeurs correspondant.

La statistique du test est donnée par la formule : $t=\frac{m}{\frac{S}{\sqrt{n}}}$ où m et S représentent la moyenne et l'écart-type de la différence d, n est la taille de la série d.

Pour savoir si la différence est significative, il faut tout d'abord lire dans la table t, la valeur critique correspondant au risque α= 5% pour un degré de liberté : ddl=n−1

1.2) Test de khi deux χ^2

1.2.1) principe

[7]Le test est utilisé pour étudier la liaison entre deux variables qualitatives X et Y.

L'hypothèse testée : H0 : les deux critères sont indépendants.

- Déroulement du test :

  Calculer la statistique est $\chi^2_{th} = \sum_{\substack{i=1 \\ j=1}}^{\substack{c \\ l}} \frac{(O_{ij}-C_{ij})^2}{C_{ij}}$ suit une loi de khi deux à (l-1) (c-1) d.d.l (avec c est le nombre de modalité sur X et l le nombre de modalité sur Y)

  O_{ij} correspond à la valeur attendue de la cellule (i,j) et C_{ij} à la valeur calculée de la cellule(i,j).

  L'hypothèse nulle est généralement rejetée lorsque p-value <0.05 ou si $\chi^2_{th} \geq \chi^2_{value}$

- Conditions d'application:

Pour appliquer le test, il est nécessaire que chaque observation soit indépendante des autres (par exemples, elles ne peuvent pas avoir été récoltées chez un même sujet) et le critère de Cochran de 1954 soit réalisé, selon lequel toutes les classes i, j doivent avoir une valeur théorique non nulle ($C_{i,j} \geq 1$), et que 80 % des classes doivent avoir une valeur théorique supérieure ou égale à 5 : $C_{i,j} \geq 5$. Lorsque le nombre de classes est petit, cela revient à dire que toutes les classes doivent contenir un effectif théorique supérieur ou égal à 5.

1.2.2) Exemple dans la recherche médicale

Sur ces 53 femmes ,48 ont répondu à deux questions :

-avez-vous utilisé des contraceptifs oraux ? Combien de temps(en mois) ?

-est ce que vous avez subir un faux accouchement? Combien de fois ?

But : savoir si le fait d'apprendre des contraceptifs oraux a une influence significative sur la santé de l'utérus, chose qui peut provoquer ou augmenter le nombre des faux accouchements.

Alors de notre base de données, nous avons le tableau récapitulatif suivant:

	Contraceptif oraux	Non Contraceptif oraux
Nombre de femmes qui ont subit au moins un faux accouchement	6	10
Nombre de femmes qui n'ont ne subit aucun faux accouchement	14	18

Tableau 5 : récapitulation de nombre de patientes selon les deux modalité : prendre des contraceptifs oraux et subir un faux accouchement

Nous effectuons le test d'indépendance de khi deux sur le logiciel R, et nous obtenons les résultats suivants :

```
> library(abind, pos=16)

> .Table <- matrix(c(6,10,14,18), 2, 2, byrow=TRUE)

> dimnames(.Table) <- list("rows"=c("Fcouche", "nonFcouche"),
+   "columns"=c("contraceptif", "noncontraceptif"))

> .Table  # Counts
            columns
rows         contraceptif noncontraceptif
  Fcouche               6               10
  nonFcouche           14               18

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test


        Pearson's Chi-squared test

data:  .Table
X-squared = 0.17143, df = 1, p-value = 0.6788


> .Test$expected # Expected Counts
            columns
rows         contraceptif noncontraceptif
  Fcouche        6.666667        9.333333
  nonFcouche    13.333333       18.666667
```

Les valeurs théoriques données dans le deuxième tableau de sorties sont tous supérieur ou égal à 5, donc le test de khi deux est applicable.

La p-value >α ce qui veut dire que $\chi^2_{th} = 0.17143 \geq \chi^2_{value}$ (ddl=1)

Donc on rejette l'hypothèse nulle, autrement dit le fait d'apprendre des contraceptifs oraux n'influe pas sur la santé de l'utérus, donc il n'a aucun effet sur l'augmentation du taux de faux accouchements.

1.3) L'analyse de variance l'ANOVA

1.3.1) principe

- Données : des mesures des valeurs de X dans k échantillons indépendants notés (X_1,X_2, …,X_k) de tailles respectives n_1 n_2… n_k .On souhaite comparer les k moyennes expérimentales, où $n = \sum_{i=1}^{k} n_i$ est l'effectif global.
- Conditions d'applications : les données sont distribuées suivant la loi normale et les variances des groupes sont homogènes.

- Les tests statistiques : $H_0 : m_1 = m_2 = \ldots = m_k$

  $H_1 : \exists\, i \neq j \ \epsilon \{1,\ldots, k\}$ tel que $m_i \neq m_j$

- Déroulement du test : Le tableau de l'analyse de la variance (Anova) qui résume les données est le suivant :

Source de variation	d.d.l	Somme des carrés	Variance	F_{exp}
Entre les groupes	k-1	$SC_{ent} = \sum_{i=1}^{k} n_i (\bar{X}_i - \bar{x})$	$S^2_{ent} = \frac{SC\ ent}{k-1}$	$\frac{S^2ent}{S^2int}$
A l'intérieur des groupes	n-k	$SC_{int} = \sum_{\substack{1 \le i \le k \\ 1 \le j \le n_i}} (x_{ij} - \bar{x_i})^2$	$S^2_{int} = \frac{SC\ int}{n-k}$	
Totale	n-1	$SC_{tot} = \sum_{\substack{1 \le i \le k \\ 1 \le j \le n}} (x_{ij} - \bar{x})^2$		

Tableau 6 : le tableau Anova

F_{exp} suit une loi de Fisher à k-1 et n-k degrés de liberté

- Décision :

Au seuil de signification α nous comparons $F_{exp} = \frac{S^2ent}{S^2int}$ à la valeur $F_{\alpha\ ;k-1\ ;n-k}$ donnée par le tableau statistique correspondant.

- ♦ Si $F_{exp} = \frac{S^2ent}{S^2int} \leq F_{\alpha\ ;\ k-1;\ n-k}$ on garde H_0, les moyennes sont égales c'est-à-dire qu'il n y'a pas de différence significative entre les sous-populations.
- ♦ Si : $F_{exp} = \frac{S^2ent}{S^2int} > F_{\alpha\ ;\ k-1;\ n-k}$ il y'a une différence significative entre les moyennes des k sous-populations (c'est-à-dire on garde H_1)[8]

1.3.2) Exemple

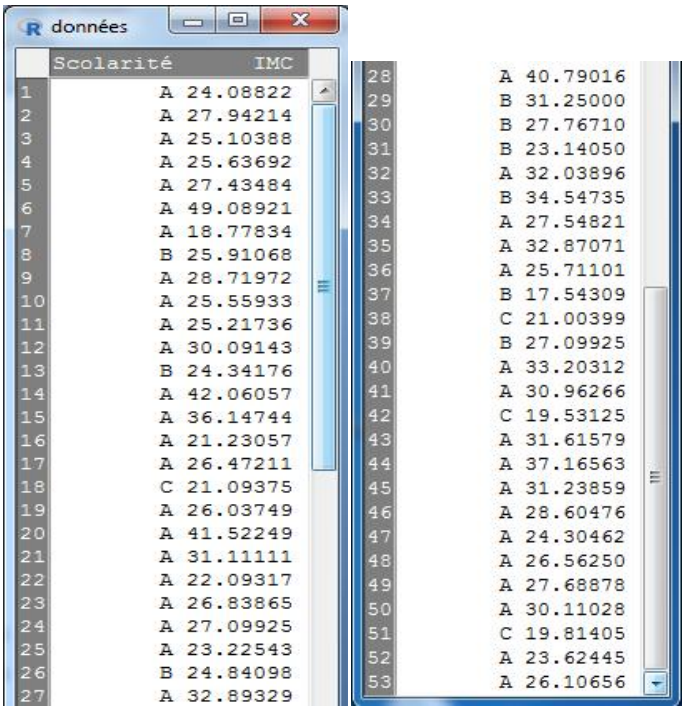
Reprenons toujours le même avec nos 53 patientes, il y'a trois catégories des femmes selon leur niveau d'étude: le niveau A: analphabète, B: passable (niveau primaire ou secondaire), et le niveau C: niveau supérieur, nous voulons utiliser un test Anova pour savoir est ce qu'il y'a une différence significative de l'IMC due au niveau d'études et au niveau intellectuel des femmes.

Nous importons toujours les données de la même façon :

Script R | R Markdown

```
données <- read.table("C:/Users/aicha/Desktop/anova.csv", header=TRUE,
  sep=";", na.strings="NA", dec=",", strip.white=TRUE)
```

Et nous avons la visualisation des données :



	Scolarité	IMC
1	A	24.08822
2	A	27.94214
3	A	25.10388
4	A	25.63692
5	A	27.43484
6	A	49.08921
7	A	18.77834
8	B	25.91068
9	A	28.71972
10	A	25.55933
11	A	25.21736
12	A	30.09143
13	B	24.34176
14	A	42.06057
15	A	36.14744
16	A	21.23057
17	A	26.47211
18	C	21.09375
19	A	26.03749
20	A	41.52249
21	A	31.11111
22	A	22.09317
23	A	26.83865
24	A	27.09925
25	A	23.22543
26	B	24.84098
27	A	32.89329
28	A	40.79016
29	B	31.25000
30	B	27.76710
31	B	23.14050
32	A	32.03896
33	B	34.54735
34	A	27.54821
35	A	32.87071
36	A	25.71101
37	B	17.54309
38	C	21.00399
39	B	27.09925
40	A	33.20312
41	A	30.96266
42	C	19.53125
43	A	31.61579
44	A	37.16563
45	A	31.23859
46	A	28.60476
47	A	24.30462
48	A	26.56250
49	A	27.68878
50	A	30.11028
51	C	19.81405
52	A	23.62445
53	A	26.10656

Tableau 7 : l'IMC des patientes selon le niveau de scolarisation

Nous appliquons l'Anova et nous obtenons les résultats suivant :

```
> summary(AnovaModel.1)
             Df Sum Sq Mean Sq F value Pr(>F)
Scolarité     2  333.4  166.68   4.955 0.0109 *
Residuals    50 1681.9   33.64
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> with(données, numSummary(IMC, groups=Scolarité, statistics=c("mean", "sd")))
      mean        sd data:n
A 29.36349 6.1849611     40
B 26.27119 4.8487870      9
C 20.36076 0.8037404      4
```

Alors nous avons F_{exp}=4.955, et la probabilité globale associé à ce test est 0.0109 < 0.05, ce qui veut dire que nos données(IMC) sont significativement différentes selon les groupes de scolarité des patientes.

Pour révéler où se trouve les différences, nous effectuons la comparaison deux à deux et nous obtenons:

```
         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts
```

```
Fit: aov(formula = IMC ~ Scolarité, data = données)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
B - A == 0   -3.092      2.140  -1.445   0.3163
C - A == 0   -9.003      3.041  -2.960   0.0121 *
C - B == 0   -5.910      3.485  -1.696   0.2091
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

Alors nous trouvons qu'il y'a une différence significative entre les deux groupes A et C. cette différence est visualisée dans la figure suivante :

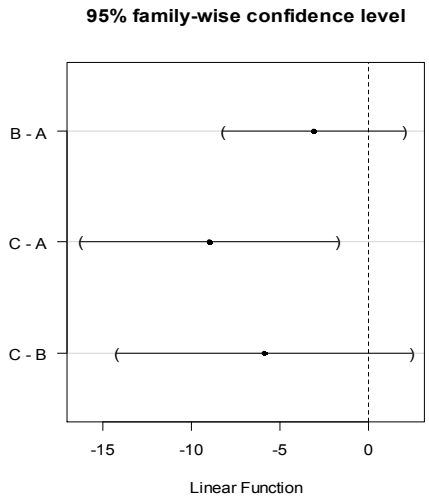


Figure 3 : graphe de centile pour les trois groupes de scolarisation :A,B et C

Cette figure nous permet de dire qu'il y'a une différence significative entre les deux groupes A et C, car la valeur 0 ne se trouve pas dans l'intervalle de confiance de la différence C-A.

Conclusion : les femmes plus instructives s'occupent bien de leur hygiène mieux que les femmes analphabètes.

2) Tests non paramétriques

Les tests non-paramétriques ne se basent pas sur des distributions statistiques. Ils peuvent donc être utilisés même si les conditions de validité des tests paramétriques ne sont pas vérifiées. Ils ne nécessitent pas d'estimation de la moyenne et de la variance. En fait, ils n'utilisent même pas les valeurs x_i recueillies dans les échantillons, mais seulement leur rang dans la liste ordonnée de toutes les valeurs.

Les tests non paramétriques les plus connus dans la recherche médicale sont :

2.1) Test de normalité (Shapiro-Wilk)

2.1.1) Principe

Dans le domaine médical, si nous voulons chercher par exemple une régression linéaire liant la variable indépendante : consommation de l'alcool avec la variable dépendante : l'apparition du cancer du poumon, nous devons s'assurer de la normalité des résidus notés x_i alors le test le plus fiable c'est le test de shapiro-wilk.

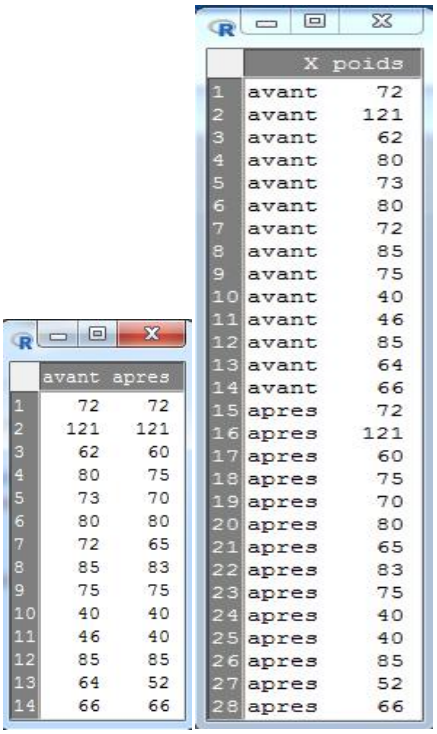
- Hypothèse testée : H0 : l'échantillon suit la loi normale.
- Déroulement du test :

1. Classer les n observations par ordre de grandeur croissante
2. Calculer les différences $d_i=x_{(n-i+1)}-x_i$ (entre le premier et le dernier, le deuxième et l'avant-dernier et ainsi de suite, l'observation médiane est ignorée si *n* est impair).
3. Lire dans la table spécifique de Shapiro-Wilk les coefficients a_i relatifs à chaque valeur.
4. Enfin, calculer la valeur pratique $W_{th}= b^2 /Z^2 = \frac{\Sigma\,(ai \times di)}{\Sigma\,(xi - \bar{x})\,2}$, elle est comprise entre 0 et 1.

- Décision : Avec un intervalle de confiance de 95% l'interprétation se fait selon la valeur $W\boldsymbol{\alpha}$ lue sur la table de la loi ou bien selon la p-value :
  Si $W_{th} > W\alpha$ (ou bien p-value > α) On retient H0 alors nos résidus suivent la loi normale avec un risque de se tromper de p-value%. Sinon on rejette H0. [9]

2.1.2) Exemple

Les patientes qui sont des cas cancer du sein sont soumises à un traitement pendant une semaine, et nous avons les données suivantes concernant leurs poids avant et après traitement. Les données sont regroupées dans le tableau suivant :



	avant	apres
1	72	72
2	121	121
3	62	60
4	80	75
5	73	70
6	80	80
7	72	65
8	85	83
9	75	75
10	40	40
11	46	40
12	85	85
13	64	52
14	66	66

	X	poids
1	avant	72
2	avant	121
3	avant	62
4	avant	80
5	avant	73
6	avant	80
7	avant	72
8	avant	85
9	avant	75
10	avant	40
11	avant	46
12	avant	85
13	avant	64
14	avant	66
15	apres	72
16	apres	121
17	apres	60
18	apres	75
19	apres	70
20	apres	80
21	apres	65
22	apres	83
23	apres	75
24	apres	40
25	apres	40
26	apres	85
27	apres	52
28	apres	66

Tableau 8 : le poids des patientes avant et après traitement

En effectuant le test de normalité de Shapiro-Wilk en logiciel R, nous obtenons les résultats suivants :

```
#importer les données à partir d'un fichier bloc note appelé normalite
> normalite <- read.table("C:/Users/aicha/Desktop/normalite.txt", header=TRUE,
+    sep="\t", na.strings="NA", dec=",", strip.white=TRUE)
#effectuer le test de normalité de scapiro-wilk
> normalityTest(~poids, test="shapiro.test", data=normalite)

        Shapiro-Wilk normality test

data:  poids
W = 0.90362, p-value = 0.01392
```

On a la p-value=0.01392 $\leq$ 0.05, donc la distribution ne suit pas la loi normale.

2.2) Test de Levene (test d'homogénéité des variances)

2.2.1) Principe

Le test de Levene (1960) est basé sur une analyse de la variance effectuée sur les écarts absolus par rapport à la moyenne de chaque échantillon, pour vérifier que les k traitements (échantillons) ont des variances qui sont comparables. [10]

- Hypothèses testées : H0: les variances des traitements sont homogènes.

  H1: au moins une des variances est fortement différente des autres.

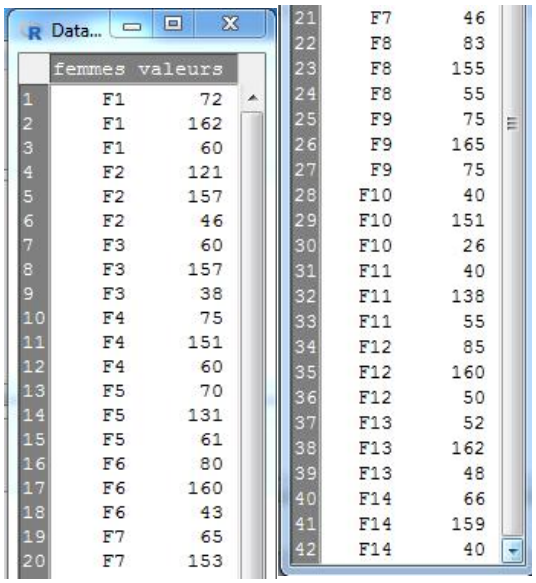
- Déroulement du test : La statistique du test est donnée par: $L=\frac{(n-k)\sum_{i=1}^{k} n_i(\bar{V}_i-\bar{V})^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n}((V_{ij}-\bar{V}_i)^2}$

  Où $V_{ij}=|X_{ij}-\bar{X}|$ et $\bar{V}$ est la moyenne des $\bar{V}_i$ et $\bar{V}_i=\frac{1}{n_i}\sum_{j=1}^{n_i} V_{ij}$

- Décision : La variable L suit une distribution F de Fisher-Snedecor de k-1 et n-k degrés de liberté. On rejette l'hypothèse nulle d'égalité des variances des populations lorsque : $L>F_{1-}\alpha_{/2,k-1,n-k}$

2.2.2) Application

Pour les 14 patientes parmi les 53 qui ont un cancer du sein, nous trouvons que quelques unes nécessitent une chimiothérapie, on veut savoir est ce que cela est fiable pour les autres en prenant en considération trois variables qui sont : le poids, la taille et l'âge, pour cela on pense à faire une Anova, alors nos données sont:



	femmes	valeurs
1	F1	72
2	F1	162
3	F1	60
4	F2	121
5	F2	157
6	F2	46
7	F3	60
8	F3	157
9	F3	38
10	F4	75
11	F4	151
12	F4	60
13	F5	70
14	F5	131
15	F5	61
16	F6	80
17	F6	160
18	F6	43
19	F7	65
20	F7	153
21	F7	46
22	F8	83
23	F8	155
24	F8	55
25	F9	75
26	F9	165
27	F9	75
28	F10	40
29	F10	151
30	F10	26
31	F11	40
32	F11	138
33	F11	55
34	F12	85
35	F12	160
36	F12	50
37	F13	52
38	F13	162
39	F13	48
40	F14	66
41	F14	159
42	F14	40

Tableau 9 : le poids, la taille et l'âge des 14 patientes qui ont un cancer du sein

D'abord on doit vérifier la normalité et l'homogénéité par le test de Levene (k=14 et $n_1=n_2=\ldots=n_{14}=3$) et on a :

```
> leveneTest(valeurs ~ femmes, data=Dataset, center="mean")
Levene's Test for Homogeneity of Variance (center = "mean")
      Df F value Pr(>F)
group 13  0.2194  0.997
      28
```

On obtient une p-value>α, on conclut à une homogénéité des variances des trois groupes, donc si le test de normalité est aussi assuré, une analyse de variance est faisable.

2.3) Test de Wilcoxon (le test des rangs signés de Wilcoxon)

2.3.1) Principe

Le test des rangs signés de Wilcoxon est l'équivalent non paramétrique du test t de Student apparié. Il repose sur les rangs des différences entre les deux mesures. Il est adapté aux variables quantitatives discrètes ou aux variables continues lorsque la condition de normalité d'utilisation du test t de Student apparié n'est pas remplie. [11]

- Données : deux échantillons appariés c'est-à-dire que chaque valeur du premier échantillon est associée à une valeur du deuxième.
- Hypothèse testée : On souhaite comparer les deux moyennes expérimentales, c'est-à-dire tester l'hypothèse nulle : H0 : les 2 distributions sont identiques, contre H1 : les 2 distributions sont différentes(ou < ou >)
- Déroulement du test :
  - On calcule les différences entre les valeurs appariées, puis on les classe par ordre croissant des valeurs absolues, en omettant les différences nulles.
  - On affecte à chaque différence non nulle son rang dans le classement (ou la moyenne de ses rangs en cas d'ex-æquo : par exemple, une valeur apparaissant aux 8ème et 9ème positions sera associée les deux fois à 8,5).
  - On note w_+ la somme des rangs des différences strictement positives, w_- la somme des rangs des différences strictement négatives
  - On vérifie que $w_+ + w_- = \frac{N(N+1)}{2}$, où N désigne le nombre de différences non nulles.
  - Enfin, on note w le plus petit des deux nombres : $w = \min \{w_+, w_-\}$.
- Décision:
  - ❖ Si $N \leq 25$, on lit dans la table du test de Wilcoxon le nombre w_α tel que, On rejette (H0) au risque d'erreur α si $w \geq w_\alpha$. Autrement on accepte (H0).
  - ❖ Si $N > 25$, sous (H0), w suit approximativement la loi normale $N(\mu, \sigma)$ avec $\mu = \frac{N(N+1)}{2}$ et $\sigma = \sqrt{\frac{N(N+1)(2N+1)}{24}}$ on calcule donc la statistique du test $u = \frac{w-\mu}{\sigma}$ et on la compare à u_α lue dans la table de la loi normale centrée réduite. On conclut comme à l'habitude.
  - ❖ On compare la valeur pratique à la valeur du risque α que nous avons choisi en fonction du sens du test : Si p-value > α (on retient H_0) nos 2

séries de données sont identiques ou proches avec un risque de se tromper de p-value %, sinon on rejet H_0

2.3.2) Exemple

On a le poids des 14 femmes qui ont un cancer du sein (cas) et nous voulons savoir si le traitement pris pendant une semaine réagit sur le poids d'une façon significative. C'est-à-dire tester si le poids diminue après une semaine de traitement alors nous pensons au test de Student apparié, mais les conditions du test ne sont pas vérifiées, car le test de normalité de Shapiro-Wilk qui est fait précédemment montre que les données ne suivent pas la loi normale. Alors le test de Student n'a pas de sens, donc nous procédons par son équivalent non paramétrique qui est le test de Wilcoxon apparié, nous voyons d'abord d'après les deux histogrammes suivants que les distributions des deux échantillons appariés sont presque identiques.

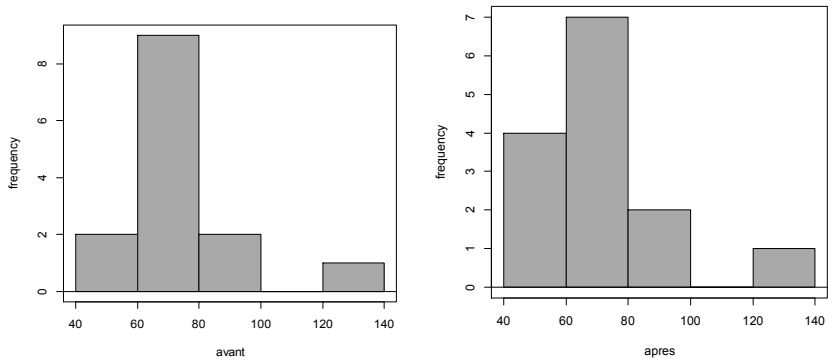


Figure 4 : les deux histogrammes des poids des 14 patientes avant et après traitement

Alors le test des hypothèses sera mieux sur les médianes que sur les moyennes des deux échantillons (avant et après traitement) H0 : $Me_{avant} = Me_{après}$ et H1 : $Me_{avant} > M_{après}$, comme nous montre la deuxième construction en logiciel R :

```
#importer les données à partir d'un fichier csv
> wilcoxon <- read.table("C:/Users/aicha/Desktop/wilcoxon.csv", header=TRUE,
+   sep=";", na.strings="NA", dec=",", strip.white=TRUE)
#tester les hypothèse nulle et alternative unilatéral> selon les médianes
> with(wilcoxon, median(avant - apres, na.rm=TRUE)) # median difference
[1] 1
#calculer la p-value et déduire le résultat
> with(wilcoxon, wilcox.test(avant, apres, alternative='two.sided', paired=TRUE))

        Wilcoxon signed rank test with continuity correction

data:  avant and apres
V = 28, p-value = 0.02225
alternative hypothesis: true location shift is not equal to 0
```

Nous obtenons une p-value = 0.0225≤0.05, et d'après la dernière ligne fournie par le logiciel nous concluons à la fiabilité de l'hypothèse alternative c'est-à-dire le traitement agit sur la diminution des poids des patientes d'une façon significative.

2.4) Test de Kruskal-Wallis

2.4.1) Principe

Le test de Kruskal-Wallis est l'équivalent non paramétrique de l'ANOVA, adapté à la comparaison entre plusieurs échantillons d'une variable quantitative discrète, ou d'une variable continue lorsque les conditions d'utilisations l'ANOVA ne sont pas remplies.

- Déroulement du test :

  On trie les valeurs obtenues dans la réunion des k échantillons par ordre croissant, puis on associe à chaque valeur son rang dans la liste si elle n'apparaît qu'une fois, la moyenne des rangs de ses apparitions si elle apparaît plusieurs fois. Pour chaque échantillon on calcule la somme r_i des rangs des valeurs qui en sont issues.

  On pose $h=\frac{12}{n(n+1)}\left(\sum_{i=1}^{k}\frac{r_i^2}{n_i}\right)-3(n+1)$

  - ❖ Si tous les n_i sont assez grands (> 5 en général), on rejette (H0) au risque d'erreur α si $h \geq \chi_{\alpha}^{2}$ à $k-1$ degrés de liberté. Autrement on accepte (H0).
  - ❖ Dans le cas où on dispose de 3 échantillons de petite taille (≤ 5), on lit dans les tables du test de Kruskal et Wallis ci-dessous le nombre h_α tel que, On rejette (H0) au risque d'erreur α si $h \geq h_\alpha$. Autrement on accepte (H0). [11]

2.4.2) Exemple

Pour faire une Anova avec les données : poids, taille et l'âge des 14 patientes qui ont un cancer on doit vérifier la normalité et l'homogénéité des variances, alors que ce dernier est déjà vérifié avec le test de Levene. Vérifions alors la normalité par le test de Shapiro-Wilk :

```
> Dataset <- read.table("C:/Users/aicha/Desktop/kruskal.csv", header=TRUE,
+   sep=";", na.strings="NA", dec=".", strip.white=TRUE)

> normalityTest(~valeurs, test="shapiro.test", data=Dataset)

        Shapiro-Wilk normality test

data:  valeurs
W = 0.83649, p-value = 0.00002932
```

Pour le test de Shapiro-Wilk nous avons la p-value <0.05 donc nous concluons au rejet de H0, donc nous n'avons pas une normalité.

Alors même si que nous avons conclure à une homogénéité des variances, mais à cause de l'absence de normalité, nous ne pouvons pas appliquer l'Anova d'où l'utilité de son équivalent paramétrique qui est le test de Kruskal-Wallis, alors nous avons :

```
> kruskal.test(valeurs ~ femmes, data=Dataset)

        Kruskal-Wallis rank sum test

data:  valeurs by femmes
Kruskal-Wallis chi-squared = 5.1016, df = 13, p-value = 0.9729
```

Donc nous trouvons une p-value>0.005, ce qui nous amène à retenir H0.

Par suite, toutes les 14 patientes sont des candidates à une chimiothérapie.

3) Conclusion

L'échantillonnage et l'estimation nous aident beaucoup à estimer une variable médicale à partir d'un échantillon, et ils permettent également la prise en compte de l'incertain par la détermination de l'intervalle de confiance qui a une probabilité de 1- α de contenir la vraie valeur (inconnue) du paramètre.

Les tests paramétriques et non paramétriques nous fournissent la signification statistique, c'est-à-dire grâce à eux on peut conclure par exemple à une fiabilité ou à une non-fiabilité d'une thérapie.

Partie 3 : Analyse statistique multidimensionnelle et recherche médicale

Dans la recherche médicale, nous avons toujours beaucoup de données et de variables, chose qui nécessite une analyse statistique multidimensionnelle, c'est la raison pour laquelle nous choisissons de traiter dans ce rapport là les méthodes les plus utilisées à savoir l'analyse des données et la segmentation d'images.

I. Analyse des données

L'analyse des données est très importante dans la recherche médicale pour en tirer des résultats pertinents, c'est pour cela nous avons choisi de traiter les méthodes les plus utiles à savoir l'ACP et la classification.

1) ACP (Analyse en Composante Principales)

1.1) Principe

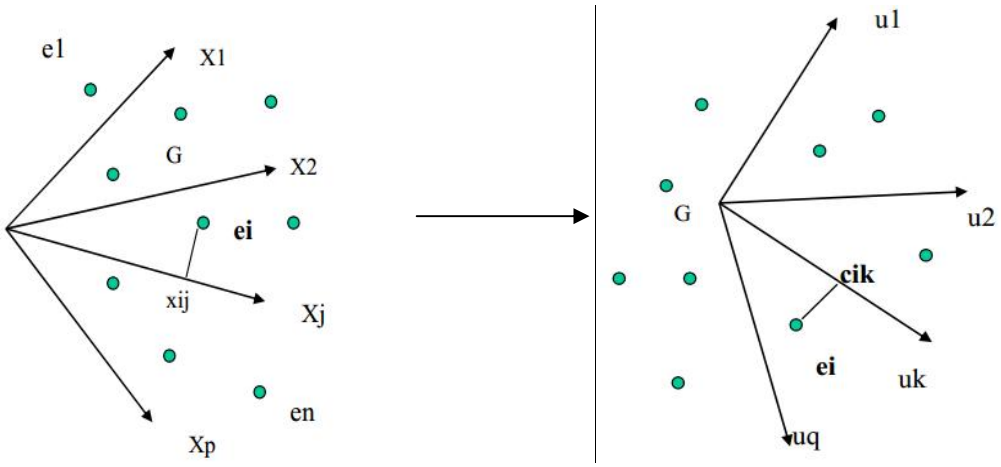
L'ACP est une méthode exploratoire (descriptive) multi-variée de réduction dimensionnelle, son objectif est de condenser l'information d'un tableau (n individus décrits par p>3 variables numériques), de manière à retirer les relations vraiment caractéristiques : proximités (corrélations) entre variables et ressemblance entre individus, ceci en limitant la perte d'information.

C'est équivalent à : déterminer un sous- espace de dimension **q<p** (q nouveaux axes) (ou **q<n**), sur lequel projeter les nuages de points relatifs au tableau de données qui soit :

- « compréhensible » par l'œil: **q** faible, de préférence **q=1,2** ou **3**
- le moins déformant possible (projection la plus fidèle possible).

Ce sous-espace est appelé espace factoriel du nuage, son principe de construction (ex : individus) est définit de la manière suivante :

- On effectue un changement de repère, passant du repère défini par les **p** variables à un repère de dimension **p** le moins déformant possible pour le nuage. Il sera défini par **p** nouveaux axes, appelés axes factoriels.
- On retient ensuite les q premiers axes du nouveau repère, ce qui nous donnera l'espace factoriel de dimension q. Il permet de récupérer les liens les plus significatifs contenus dans le tableau. [12]



$e_i = (x_{i1}, \ldots x_{ij}, \ldots x_{ip})$ $\quad u_k = \alpha_{k.1} X_1 + \ldots + \alpha_{k.p} X_p$

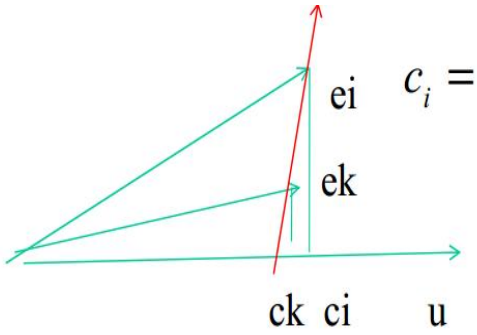
Alors la valeur de la composante principale k prise par l'individu i est donnée par :

$C_{ik} = x_{i1}\, u_{1k} + \ldots + x_{ip}\, u_{pk}$

Les p axes factoriels (composantes principales) sont définis séquentiellement:

- On détermine l'axe (premier axe factoriel) sur lequel le nuage se déforme le moins possible en projection,
- On cherche un second axe, sur lequel le nuage se déforme le moins en projection, après le premier axe, tout en étant orthogonal au premier,
- On réitère jusqu'à l'obtention de p axes.

Et la projection sur un axe s'obtient:



$$c_i = \langle e_i, u \rangle = x_{i1}u_1 + \ldots + x_{ip}u_p$$

C'est-à-dire pour obtenir une déformation minimale, il faut que l'axe sur lequel on projette permette la dispersion maximale : d (ci,ck)≈ d(ei,ek).

Souvent, et surtout en biologie et en médecine, les p variables sont fortement inter-corrélées : il y a une grande redondance dans l'information qu'elles délivrent. Il n'y a donc probablement pas besoin d'un espace à p dimensions pour restituer toute l'information, sachant que la plupart des dimensions apportent peu ou prou la même information.

1.2) Application et interprétation de l'ACP en recherche médicale

Nous appliquons la méthode ACP sur les données des différentes patientes (53) présentes à CHU de Fès, dont les variables quantitatives sont : l'âge, le poids actuel, la taille, le tour de la taille, et le nombre d'heure du sommeil (Cinque variable quantitatif), en plus de la variable qualitatif : cas (oui)/témoin (non).

Acp

	femmes	Poids_actuel	Taille.cm.	Tour_taille.cm.	Sommeil.heure.	Age	Cas.Témoin
1	a	64	163	86	7	66	non
2	b	68	156	109	6	63	non
3	c	58	152	80	6	64	non
4	d	64	158	83	5	57	non
5	e	72	162	127	6	60	oui
6	f	121	157	132	8	46	oui
7	g	53	168	86	5	49	non
8	h	68	162	83	9	62	non
9	i	83	170	114	8	49	non
10	j	73	169	98	7	48	non
11	k	67	163	87	5	61	non
12	l	78	161	100	6	47	non
13	m	60	157	92	8	38	oui
14	n	105	158	108	9	61	non
15	o	76	145	98	5	49	non
16	p	53	158	86	6	52	non
17	q	82	176	99	7	53	non
18	r	54	160	92	7	53	non
19	s	65	158	104	9	55	non
20	t	120	170	118	4	53	non
21	u	70	150	93	4	49	non
22	v	70	178	97	5	60	non
23	w	67	158	98	7	59	non
24	x	72	163	96	6	63	non
25	y	64	166	89	6	47	non
26	z	66	163	96	6	48	non
27	aa	75	151	105	4	60	oui
28	bb	70	131	96	6	61	oui
29	cc	80	160	98	5	43	oui
30	dd	65	153	97	6	46	oui
31	ee	63	165	94	4	50	non
32	ff	75	153	98	4	64	non
33	gg	83	155	93	6	55	oui
34	hh	75	165	130	9	75	oui
35	ii	72	148	112	7	52	non
36	jj	65	159	98	7	46	non
37	kk	40	151	65	10	26	oui
38	ll	40	138	84	8	55	oui
39	mm	72	163	101	8	67	non
40	nn	85	160	140	8	50	oui
41	oo	66	146	107	10	46	non
42	pp	50	160	74	7	54	non
43	qq	84	163	165	7	58	non
44	rr	76	143	122	8	45	non
45	ss	77	157	135	6	55	non
46	tt	76	163	106	7	64	non
47	uu	63	161	98	8	47	non
48	vv	68	160	86	9	52	non
49	ww	70	159	98	7	48	non
50	xx	80	163	95	5	49	non
51	yy	52	162	98	6	48	oui
52	zz	62	162	96	8	48	non
53	µ	66	159	87	9	40	oui

Tableau 10 : les données des patientes pour faire une ACP

Regardons l'aspect de nuage des individus, et des groupes semblent se former selon la modalité : cas/témoin.

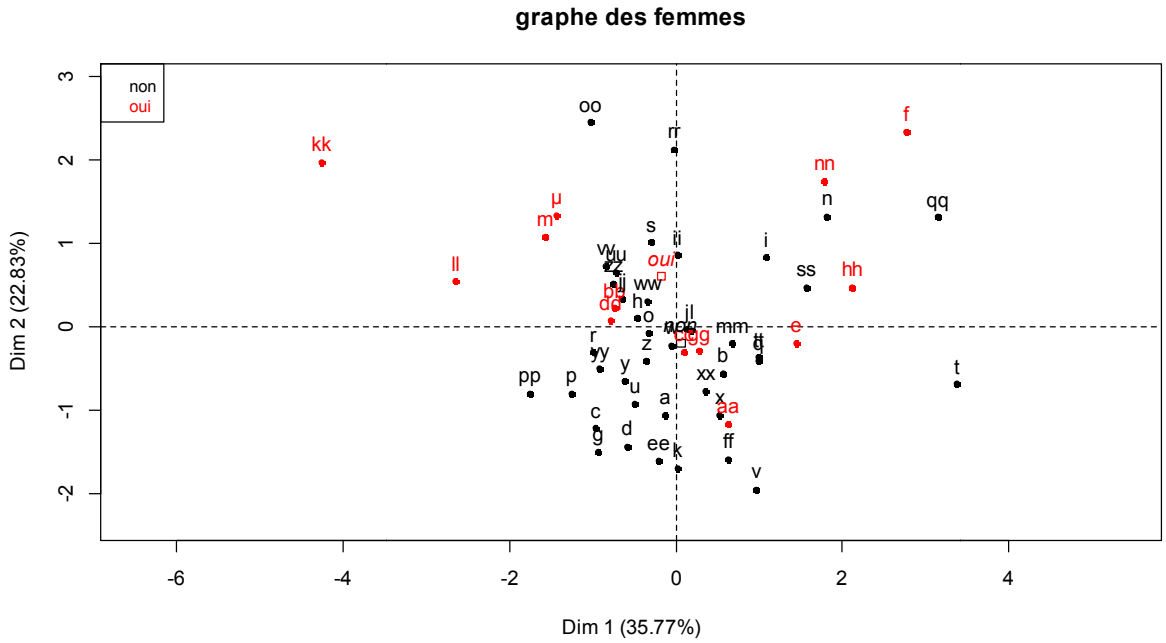


Figure 5 : Le nuage de points (femmes), cas (oui) et témoin (non)

Nous avons la liste des résultats concernant les individus s'affichant sur la console suivante :

```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5
Variance               1.789   1.142   0.925   0.798   0.346
% of var.             35.774  22.832  18.510  15.965   6.919
Cumulative % of var.  35.774  58.607  77.116  93.081 100.000

Individuals (the 10 first)
                     Dist    Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
a                |  1.883 | -0.135  0.019  0.005 | -1.055  1.841  0.314 | -0.189
b                |  1.399 |  0.571  0.344  0.167 | -0.561  0.521  0.161 | -0.990
c                |  2.156 | -0.965  0.982  0.200 | -1.221  2.462  0.321 | -1.257
d                |  1.619 | -0.589  0.366  0.132 | -1.437  3.414  0.788 | -0.342
e                |  1.836 |  1.452  2.225  0.626 | -0.198  0.065  0.012 | -0.337
f                |  4.020 |  2.769  8.088  0.474 |  2.343  9.075  0.340 |  0.286
g                |  2.182 | -0.943  0.937  0.187 | -1.503  3.733  0.474 |  1.073
h                |  2.100 | -0.467  0.230  0.050 |  0.103  0.017  0.002 |  0.170
i                |  2.005 |  1.090  1.253  0.295 |  0.828  1.134  0.171 |  1.444
j                |  1.404 |  0.127  0.017  0.008 | -0.034  0.002  0.001 |  1.397
```

Eigenvalues : les valeurs propres et les pourcentages d'inerties associées à chaque dimension.

La colonne **dist** désigne la distance de l'individu (femme) au centre de gravité de nuage.

Dim.1,…dim.5 désigne la coordonnée sur la première,…la cinquième dimension.

La qualité des individus de représentation est mesurée par l'entrée cos2. Un individu mal représenté se situe généralement (à tort) près du centre du repère(les patientes : jl, w, gg et cc sont mal représentées) et sa spécificité est mal prise en compte par l'ACP pour les composantes principales considérées.

Et nous avons la même chose pour les variables:

```
Variables
                   Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3    ctr
Poids_actuel     |  0.851 40.519  0.725 |  0.203  3.595  0.041 |  0.049  0.262
Taille.cm.       |  0.386  8.311  0.149 | -0.338 10.013  0.114 |  0.816 71.871
Tour_taille.cm.  |  0.808 36.515  0.653 |  0.405 14.370  0.164 | -0.133  1.897
Sommeil.heure.   | -0.209  2.441  0.044 |  0.796 55.475  0.633 |  0.146  2.295
Age              |  0.467 12.214  0.218 | -0.435 16.548  0.189 | -0.468 23.676
                   cos2
Poids_actuel     0.002 |
Taille.cm.       0.665 |
Tour_taille.cm.  0.018 |
Sommeil.heure.   0.021 |
Age              0.219 |
```

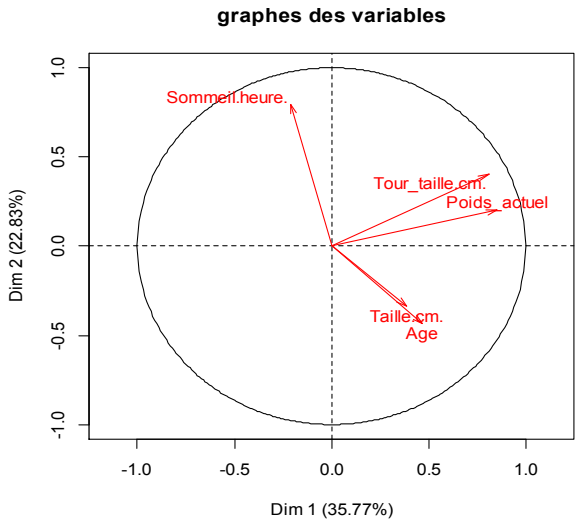


Figure 6 : le graphe des variables de l'ACP(cercle de corrélation)

Plus une variable possède une qualité de représentation élevée dans l'ACP, plus sa flèche est longue, donc nous pouvons dire que toutes les variables ont une bonne qualité de représentation de l'ACP (elle est mesurée par le cos2) à l'exception de la variable taille, alors les premières variables sont bien représentées et la taille est moins bien représentée.

Plus deux variables sont corrélées, plus leurs flèches pointent dans la même direction (dans le cercle de corrélation, le coefficient de corrélation est symbolisé par les angles géométriques entre les flèches), alors les deux variables taille et âge sont fortement corrélées entre eux.

Plus une variable est proche d'un axe principal de l'ACP, plus elle est liée à lui. Cette dernière règle permet généralement de donner un sens concret aux axes de l'ACP. Donc la variable poids est la variable la plus importante pour la construction du premier axe, de même la variable sommeil est la variable la plus importante pour la construction du deuxième axe.

Nous concluons qu'au cours d'un traitement de la maladie concernée il faut chaque fois mesurer le poids et le nombre d'heures qu'effectuent les patientes car ils sont des variables significatives de l'état et de processus de la maladie.

2) La classification

Les méthodes de classification ou de typologie ont une utilité assez importante dans le domaine médicale, elles ont pour but de regrouper les individus en un nombre restreint de classes homogènes. Par exemple, imaginez une étude dans laquelle un chercheur en médecine a collecté des données sur diverses mesures d'aptitude physique (variables) dans un échantillon de patients cardiaques (observations). Le chercheur peut créer des classes d'observations (patients) afin de détecter les groupes de patients présentant des symptômes similaires. Dans le même temps, le chercheur peut réaliser des classes de variables (mesures de l'aptitude physique) afin de détecter les classes de mesures qui semblent révéler les mêmes capacités physiques. [13]

Il s'agit de décrire les données en procédant à une réduction du nombre des individus.

Les méthodes de classification reposent sur la notion de dissimilarité (distance) entre les objets que l'on souhaite regrouper en classes homogènes.
Soit *E* un ensemble de *n* objets, une application d de E × E dans $\mathbb{R}$ + est appelée dissimilarité (distance) si elle vérifie :

(1) d(i,j) = d(j,i)

(2) d(i,j) >= 0

(3) d(i,j) = 0 ssi i = j

(4) d(i,j) <= d(i,k) + d(k,j)

2.1) La classification ascendante hiérarchique(CAH)

Le principal problème des méthodes de classification consiste à définir le bon critère de regroupement de 2 classes (clusters), ce qui revient à définir une distance adéquate entre 2 classes.

- Au départ chaque objet représente un groupe.
- A chaque étape on regroupe les deux groupes d'objets dont la ressemblance est la plus forte, selon les critères d'agrégation :

  Saut minimum : la distance entre deux classes est ici déterminée par la distance entre les deux objets les plus proches (les plus proches voisins) dans les différentes classes. Cette règle provoque des chaînes d'objets assemblés en classes, et les résultats obtenus ressemblent à de longues chaînes.

  Méthode de Ward : Il tente de minimiser la Somme des Carrés de tous les couples de classes pouvant être formés à chaque étape.

  On continue jusqu'à ce qu'il n'y ait plus qu'une seule classe.

Cette méthode aboutit à un emboîtement de partitions (homogènes et séparés) visualisé graphiquement par un arbre hiérarchique indicé (dendrogramme). [13]

2.2) La classification non hiérarchique

Contrairement à d'autres méthodes dites hiérarchiques, qui créent une structure en « arbre de clusters » pour décrire les groupements, les méthodes non hiérarchiques ne créent qu'un seul niveau de k classes aussi différentes entre elles que possible. [14]

La méthode la plus connue et utilisée dans la recherche clinique et médicale c'est la méthode de k-means, du fait de sa simplicité de mise en œuvre. Il partitionne les données en K clusters tirés au hasard de l'ensemble d'individus. L'algorithme associé renvoie une partition des données, dans laquelle les objets à l'intérieur de chaque cluster sont aussi proches que possible les uns des autres et aussi loin que possible des objets des autres clusters. Chaque partition est définie par ses objets et son centroïde.

Les principales étapes de l'algorithme k-means sont :

1. Choix aléatoire de la position initiale des **k** clusters.)
2. Affecter les points (données) à un cluster suivant un critère de minimisation des distances (généralement selon une mesure de distance euclidienne).

3. Une fois tous les points placés, recalculer les **k** centroïdes.
4. Réitérer les étapes 2 et 3 jusqu'à ce que plus aucune ré- affectation ne soit faite.

2.3) Application de Classification

Nous utilisons les mêmes données des femmes (patientes) à CHU, en négligeant le fait qu'elles sont des cas ou des témoins (négliger la variable qualitatif pour éviter d'avoir des groupes selon cette modalité) pour faire des groupes de patientes selon leurs caractéristiques (poids, âge…) afin que nous puissions trouver les traitements adéquats à chaque patiente, également chercher un moyen de prévention pour les femmes (témoins) chaque groupe.

- Donc, en utilisant la classification hiérarchique, nous avons le schéma suivant :

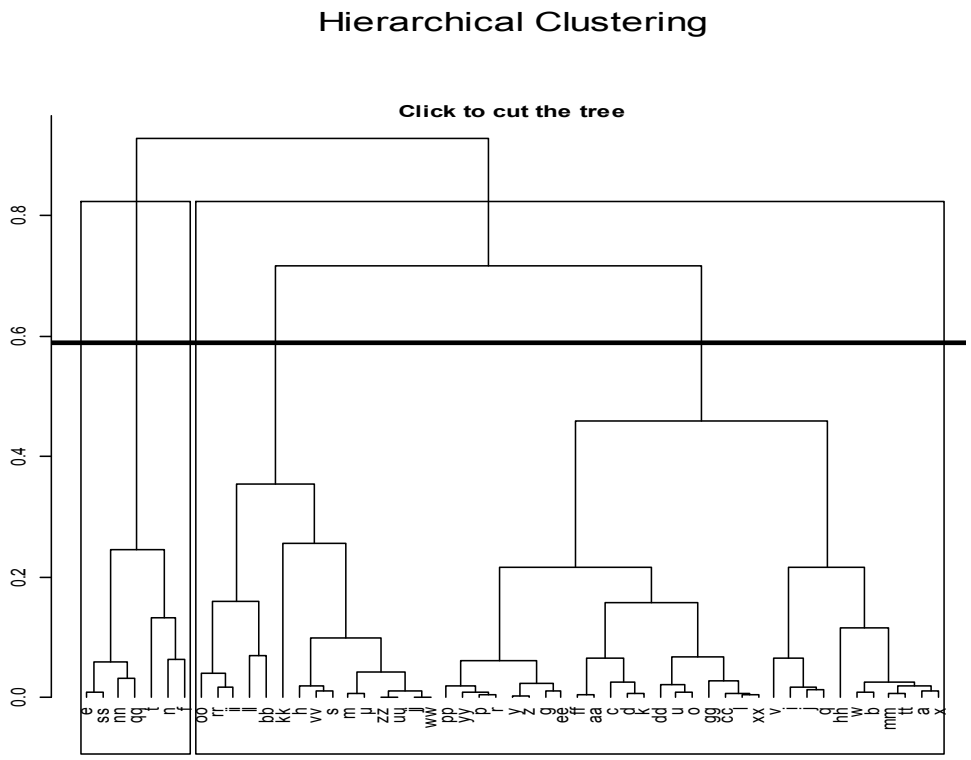


Figure 7 : la classification ascendante hiérarchique des patientes (dendrogramme)

Cette figure nous montre qu'on part des clusters de petites taille, et chaque fois on diminue le nombre de clusters en réaffectant les patientes au cluster le plus homogène et le plus adéquat, jusqu'à avoir uniquement deux groupes dont les membres sont liées et ils ont le maximum de similarité entre leurs caractéristiques. Alors les deux clusters finaux sont : C1={e,f,n,t,qq,nn,ss} et C2 contenant le reste. Cela permet de préciser les traitements adéquats à ces clusters de patientes selon la disponibilité.

- Nous utilisons les mêmes données de la classification hiérarchique, mais en procédant maintenant à une classification non hiérarchique (K-means). Nous obtenons la figure suivante :

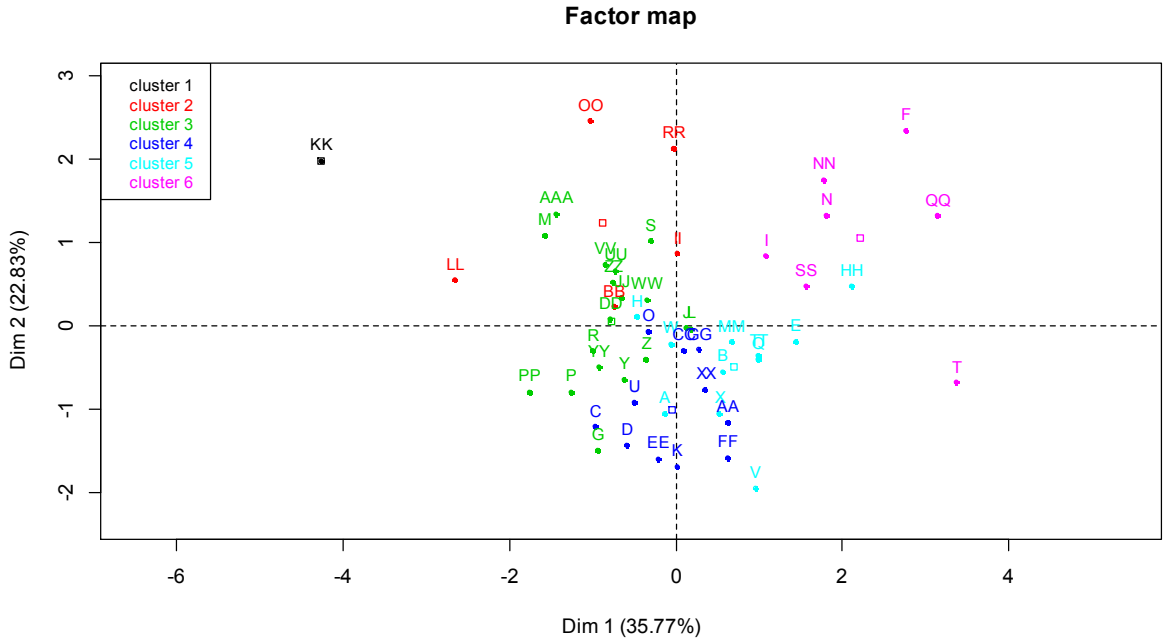


Figure 8 : les clusters obtenus par une classification non hiérarchique (k-means)

Nous voyons qu'il y'a possibilités à 6 catégories de patientes qu'il faut prendre en considération dans la recherche du traitement convenable, mais nous pouvons isoler la femme indexée par KK et la traiter à part, c'est une patiente qui nécessite une étude spéciale, d'une part c'est un cas cancer et d'autre part nous avons déjà vu qu'elle a un IMC très faible dans la partie de la statistique descriptive. Nous concluons qu'elle est dans un état avancé de la maladie.

2.4) Conclusion

Le fait d'imposer de hiérarchiser les données aboutit à une perte des cas spéciaux car nous voyons par exemple la patiente indexée par KK est dans un groupe de 46 patientes similaires par la classification hiérarchique, par contre dans la classification non hiérarchique cette patiente est restée à part, donc cette dernière classification donne des clusters plus homogènes que la première.

II. Segmentation d'images médicales

Le but du traitement des images médicales est d'extraire à partir des images acquises, les informations utiles au diagnostique, de révéler des détails difficiles à percevoir à l'œil nu, tout en évitant la création d'artefacts, faussement informatifs. Pour cela le traitement fait appel à des outils, des algorithmes, qui permettent d'agir sur l'image numérisée. La reconstruction de forme, les segmentations, les quantifications, l'analyse fonctionnelle, jusqu'aux simulations (organes virtuels, malades virtuels), tous ces outils de traitement ont contribué à l'amélioration de la qualité des images acquises, à leur interprétation et surtout à une meilleure approche au diagnostic. Elle joue un rôle important dans la détection des micro-calcifications à un stade précoce, ce qui va nous aider à éviter d'avoir recours au traitement radical comme l'ablation du sein.

La segmentation d'images est une opération fondamentale et importante, c'est l'étape préliminaire a à tout traitement d'images.

La segmentation consiste à rassembler des pixels d'une image entre eux en formant des régions connexes, homogènes et bien séparées. Ces régions possèdent une certaine uniformité pour une ou plusieurs caractéristiques (intensité, couleur, texture, …) et sont différentes pour au moins une de ses caractéristiques des régions voisines(R).

Formellement la segmentation d'une image numérique **I** consiste à chercher une partition de **I** en un sous-ensemble R = {R1, R2, .., Rn} tels que:[15]

$$\forall\, i \;\; R_i \neq \emptyset$$

$$\forall\, i \neq j \;\; R_i \cap R_i = \emptyset$$

$$I = \cup\, R_i$$

Il existe deux grandes catégories de segmentation : segmentation par classification et segmentation par contours

1) Approche par classification

1.1) Principe

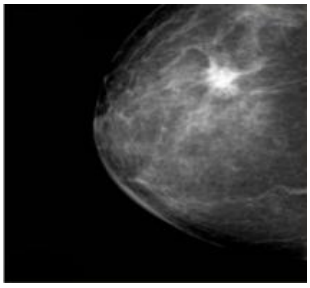
Nous allons utiliser le même algorithme défini précédemment sur l'image à segmenter qui est considérée comme une répartition de pixels, ces derniers sont les points de l'algorithme que

nous allons calculer les distances entre eux, ce qui veut dire que la segmentation est la donnée d'un ensemble de clusters compacts et clairement séparés.

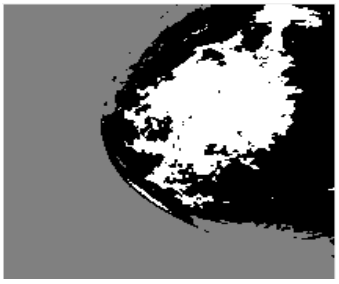
L'affectation des pixels aux différentes classes définies au début se fait grâce au calcul de la distance minimale entre les pixels et les centroïdes des classes, donc à chaque pixel on lui donne le numéro de la classe qu'il est inclus dedans.

1.2) Application en recherche médicale

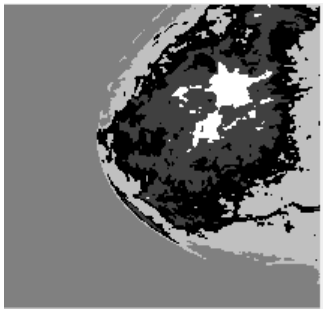
Nous avons une mammographie (une échographie du sein), à partir d'une segmentation d'image et en utilisant la classification k-means, détectons la position exacte de la tumeur si elle existe [15]



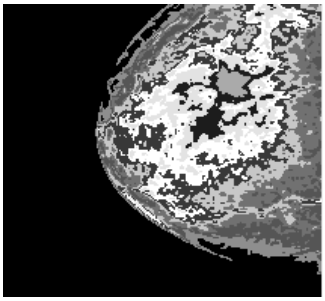
En utilisant MATLAB, on fixe le nombre de clusters en k=3, on obtient l'image segmentée suivante :



On augmente maintenant le nombre de clusters e à k=5, on obtient l'image segmentée suivante :



On remarque qu'on approche de temps en temps à une zone qui prend la couleur la plus claire. On augmente de plus le nombre de cluster en k=10, pour une visualisation plus claire de la position de la tumeur, et on obtient :



On voit maintenant la position de la tumeur pour choisir le traitement qui convient avec son position et sa dimension.

2) Approche contours

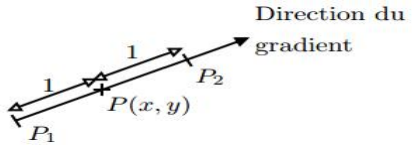
Une image I n'est pas simplement un ensemble discret de pixels mais l'observation d'une fonction continue I : $\Omega \rightarrow \mathbb{R}$ (où $\Omega \subsetneq \mathbb{R}^2$ ou bien $\mathbb{R}^3$).

La méthode permettant d'obtenir une image des contours est donnée par :

1. Estimation du gradient en chaque point de l'image.

   $\nabla I(x, y) = (G_x, G_y) = \left(\frac{\partial I}{\partial x}(x,y) , \frac{\partial I}{\partial y}(x,y) \right)$

2. Extraction des maxima locaux de la norme du gradient $G = \sqrt{G_x^2 + G_y^2}$ dans la direction du gradient $\theta = \arctan(G_y/G_x)$.

   Soient Gr, Gr1 et Gr2 respectivement les gradients en P,P1 et P2.

   Si Gr>Gr1 et Gr>Gr2 ⇒Présence d'un maximum local en P

   Existence d'un maximum local de gradient ⇒ présence d'un contour

3. Sélection des maxima locaux significatifs par seuillage (pour limiter la fragmentation des contours obtenus).
4. Fermeture des contours en traçant les chemins suivant une ligne dans l'image de la norme du gradient.
5. Résultat : une image binaire (image des contours) [16]

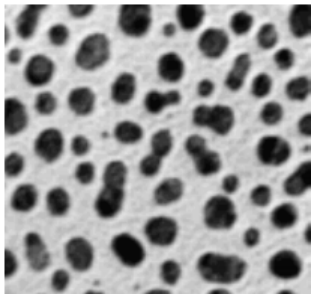
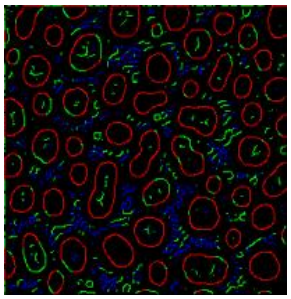


Image à segmenter



Les maxima locaux du gradient

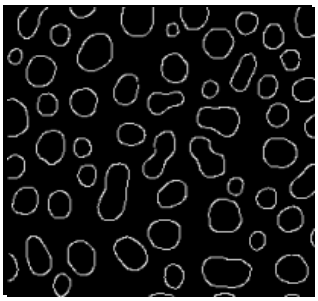


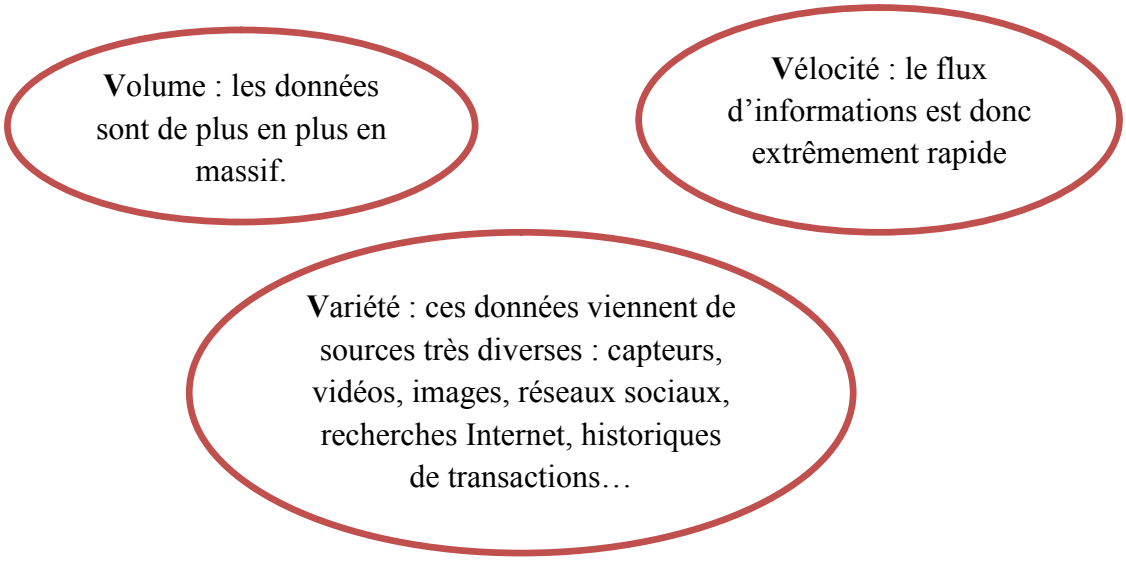
Image des contours

III. Le big data (Analyse des données en masse)

J'ai voulu terminer mon rapport avec un volet d'actualité qui consiste au traitement statistique de grande masse de données à savoir le big-data.

1) Définition

- Le Big Data, ou en Français, "méga-données" désigne à la fois l'explosion des données produites grâce aux nouvelles technologies, mais aussi les outils et pratiques destinées à les collecter, les stocker («"On garde tout! »), les traiter, les analyser pour mieux comprendre le monde qui nous entoure (les visualiser). il est caractérisé par ce qu'on appelle les « 3V »: [17]

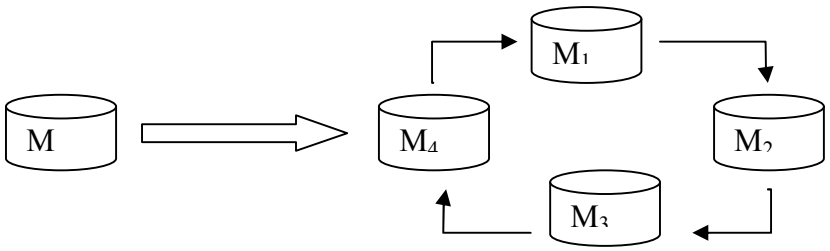


L'essor de big data a suivi l'évolution des systèmes de stockage et de traitement des données avec notamment l'avènement du cloud computing et des supercalculateurs. Désormais, on parle de pétaoctets et de zettaoctets pour désigner les volumes que représentent les big data. Le volume de données produites dans le monde atteindra les 40 zettaoctet (40×10^{18}) en 2020. [18]

De notre Smartphone, véhicules et de nos ordinateurs, des capteurs et des caméras enregistrent notre localisation et toutes nos activités, alors : voix, sens, image…tout est convertit en chiffres. Tout cela pour calculer et prédire l'avenir en se basant sur le passé, c'est ce qu'on appelle l'analyse prédictive du big data.

2) Traitement big-data

La distribution est la seule solution pour réussir un traitement big data, lorsqu'on dit une distribution c-à-d une découpe, un partitionnement de nos données sur une série de machines, c'est ce qu'on appelle cluster.



Alors on distribue des données à travers :

- Un Système de Fichiers Distribué (google file système)
- HDFS (Hadoop File Système) c'est un système de fichiers distribué open source (libre et gratuit)

On distribue donc les données puis, on les découpe, on les met sur un certain nombre de machines et il faut donc distribuer le traitement sur ces données à travers :

- Hadoop, MapReduce (les algorithmes intégrés)
- DAG : graphe orienté acyclique (il désigne qu'on va faire une étape après l'autre dans le traitement des données).

Cette distribution peut se faire de deux façons :

- Traitement centralisé : une machine maitre va régulariser le traitement qui va s'effectuer sur des machine secondaires (recours à la classification hiérarchique)
- Traitement non centralisé : chaque machine a les mêmes responsabilités que les autres (recours à la classification non hiérarchique).

On se base sur un stockage « schared nothing » (rien n'est partagé entre les instances) c-à-d chaque machine va avoir localement ses données, on va découper par exemple un volume de données de 10 pétaocets en 100 partitions, donc on va avoir 100 machines, 100 nœuds, et 1/100 de données sur chaque machine. Alors on conclut que le clusering est très important dans le big data et on a deux niveaux :

- Partitionnement pour la répartition (scharding)
- Réplication : chaque partition de données sera répliquée sur d'autres nœuds de façon à ne pas perdre les données si un nœud tombe en panne. [19]

Finalement, lorsqu'on dit partitionner les données, on a recourt à des échantillons de taille **n**, donc tous les méthodes statistiques peuvent être utilisé comme étude préliminaire avant de déterminer les algorithmes du big data notamment : l'analyse statistique multidimensionnelle, la régression linéaire pour déterminer les corrélations, et les tests sont aussi inclus parce qu'on teste des hypothèses pour une prédiction.

L'exploitation des big data a ouvert de nouvelles perspectives dans de nombreux domaines : la recherche scientifique, la politique, la communication, la médecine, la météorologie, l'écologie, la finance, le commerce, etc. Grâce à des outils analytiques et à la modélisation de données, des chercheurs, des entreprises, des administrations peuvent faire de l'analyse tendancielle ou prédictive, dresser des profils, anticiper des risques et suivre des phénomènes en temps réel...

3) Le big data dans la recherche médicale

Le Big Data, ce sont des montagnes de données stockées sur des serveurs, dans d'immenses entrepôts. Et parmi toutes ces données, certaines concerneront l'histoire médicale de chaque personne, accessible au médecin via un futur *dossier médical informatisé*. Il indiquera tous les éléments utiles du patient depuis la naissance, comme le carnet de vaccination, les maladies chroniques… Il suivra le patient dans le temps et l'espace et deviendra un vrai outil de coordination des soins. [20]

Alors les données de santé peuvent être collectées:

1- Des informations patientes récoltés en ville et à l'hôpital et des données de recherche.
2- Des objets connectés et des conversations sur internet.
3- De l'analyse du génome humain…

   Depuis le séquençage du premier génome humain en 2001, le coût du séquençage a été divisé par plus de 100000, et il est aujourd'hui possible de séquencer un individu en quelques jours pour environ 1000 dollars (Figure 9); le séquençage est ainsi entrain de devenir un examen de routine, générant environ 100 Go de données par échantillon. [21]

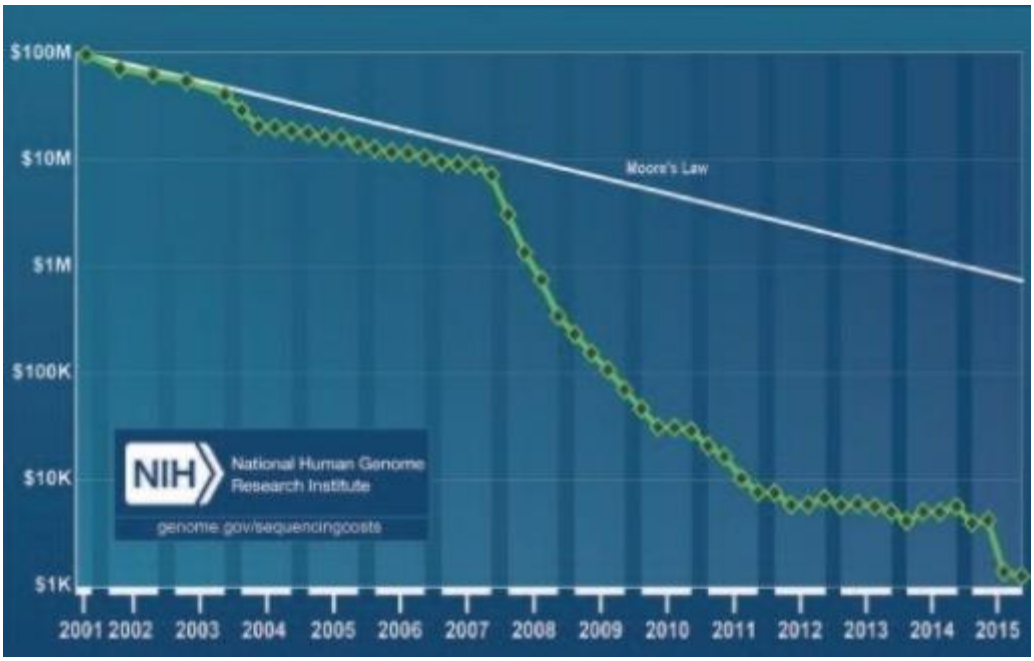


Figure 9 : Le coût pour séquencer un génome humain a été divisé par 1 00000 en 15 ans, et est proche d'atteindre la barre symbolique des 1 000 dollars.

3.1) Les différentes applications du Big Data dans la recherche médicale

Bien exploitées et traitées, ces données sont une mine d'or et d'informations pour la santé publique et la recherche médicale. Elles permettent notamment de :

- ✓ Mieux prévenir et prendre en charge les maladies :

En récoltant diverses données sur la population (habitudes de vie, localisation, données de santé), on peut plus aisément identifier des facteurs de risque pour certaines maladies comme le cancer, le diabète ou l'asthme. On peut alors développer des programmes de prévention, notamment à destination des populations identifiées comme à risque.

- ✓ Vérifier l'efficacité d'un traitement :

Notamment dans le cas des vaccins, on peut à présent relever et traiter des centaines de paramètres lors des essais cliniques.

- ✓ Prédire des épidémies :

Sur une région donnée, on dispose de nombreuses informations sur l'état de santé des habitants et l'on peut ainsi repérer l'expansion d'une malade ou identifier des comportements à risque. Depuis 1984, la France a développé un réseau qui suit plusieurs maladies infectieuses et alerte sur les épidémies.

- ✓ Accompagner la recherche pharmaceutique et améliorer la pharmacovigilance :

En traitant des bases avec des données sur le long terme, on peut plus facilement rapprocher les traitements des problèmes de santé. C'est comme cela qu'on a pu étudier le risque

d'infarctus du myocarde chez les femmes utilisant une pilule contraceptive de 3ème génération. [22]

3.2) Le big data pour une médecine préventive

L'analyse prédictive et les anticipations du big data dans la recherche médicale est diverse mais on peut avoir une idée sur son utilité et sur son principe algorithmique dans les exemples suivants :

Exemple 1 : Les psychiatres veulent identifier les maladies psychiques à un stade précoce bien avant qu'il se déclare, ils ont participé au développement d'une application capable de prédire la survenue d'une maladie de dépression.
L'application transforme le téléphone en motif d'espionnage, elle enregistre les moindres détails :-où vous vous trouvez ?en voyage ?en travail ?..., A quelle fréquence vous communiquez ? en Watsapp ? en Skype ?en Facebook ? Quelle est la mesure de l'intonation de voix ?…
Cette application ne peut pas améliorer le diagnostique mais permet de savoir comment le patient réagit avec le traitement avec beaucoup plus de précisions, elle nous permet à étudier et à analyser le changement et l'évolution des activités quotidienne, hebdomadaire, mensuelle et annuelle (presque 100.000.000de données par jours) [23]

Exemple 2: Dans un laboratoire pharmaceutique, les biologistes et les bio-informaticiens veulent déterminer quelle franche de la population est menacée par le diabète en utilisant les données brutes et anonymes des patients déjà atteint de diabète. La problématique c'est de déterminer : qu'elle est le critère à extraire de ces données ?
Donc l'algorithme utilisé doit d'abord savoir à quoi ressemble le profil d'un diabétique en schéma, chercher ensuite le même schéma chez d'autres personnes, puis calculer les probabilités que ces personnes deviennent des diabétiques.[23]

Exemple 3: Jusqu'à récemment la seule manière de détecter une épidémie de grippes c'est faire la consultation et le test médical et attendre le résultat. Les chercheurs donc ont inversé le procédé : pouvait-on prédire une épidémie de grippes à partir des recherches faites sur internet. Ils se sont attaqué ce qui semblait impossible, ils ont fait le tri des recherches (des milliers de recherches) et en examinant toutes les données ils ont trouvé une corrélation entre les recherches sur la grippe et le nombre de personnes atteintes, ils ont également identifié les

termes de recherches qui permettent de prédire avec justesse les épidémies de grippes. Or en se basant sur notre modèle google on pouvait les avoir instantanément juste en appliquant notre algorithme en recherche effectué en ce moment.

Dans la première fois on a de la rétroaction qui permet de voir en temps réel ce qui se passe, mais cet année l'algorithme a été faussé à cause des médiats qui ont consacré beaucoup de temps à la sensibilisation du danger de grippes, cela a attisé la curiosité de certains, ce qui amène à l'augmentation des recherches sur le web. [24]

4) Conclusion

On peut dire que les algorithmes utilisée dans le big data ne Nous fournissent que des probabilités, Nous ne sommes pas sûr de la Big Data, mais de nombreuses données sont crées, stockées et traitées dans le but de faciliter les échanges patient / médecin, Donc on ne peut pas totalement écarter le hasard, mais on peut tjrs être un peu plus sûr.

Les chercheures ou bien ce qu'on appelle les data-scientist sont en cours de les développer en mieux, alors peut être à un certain moment d'avenir le big data va remplacer le médecin.

Conclusion générale

Dans ce rapport, j'ai essayé de une vue globale sur l'importance de l'utilisation de la statistique dans le domaine médicale. Partant de la statistique descriptive comme étant la base de toute étude statistique médicale : paramètres de position et de dispersion, représentation graphiques, tableaux…,

Puis la statistique inférentielle : L'estimation des paramètres et les tests statistiques paramétriques et non paramétriques sont nécessaire pour savoir la signification statistique des essais thérapeutiques, par exemple pour connaitre l'efficacité de certains remèdes dans la guérison des maladies.

Ensuite, l'analyse des données multidimensionnelle est obligatoire en recherche médicale à cause de la diversité des variables liée à un sujet d'étude médical, dont les méthodes les plus utiles sont l'ACP et la classification.

La troisième partie traite des sujets plus vastes avec le traitement d'images médicales et son étape préliminaire qui est la segmentation d'images médicales est nécessaire pour une meilleure approche au diagnostic. Elle joue un rôle important dans la détection des micro-calcifications à un stade précoce, ce qui va nous aider à éviter d'avoir recours au traitement radical comme l'ablation du sein.

Finalement, le big-data comme nous avons vu est la science de l'avenir, il a pour but de rendre la médecine préventive plus personnalisée au plus prés des besoins des patients, car pour mieux comprendre les maladies et inventer de nouveaux traitements nous devons comprendre comment ils apparaissent et évoluent en temps réel pour chaque patient.

Bibliographie et web-graphie

[1]Méthodes statistiques Médecine-Biologie (Jean Bouyer)

[2]math.univ-lille1.fr/~ayache/cours_SD.pdf

[3]sante.journaldesfemmes.com/poids-calcul-imc/

[4]Statistiques décisionnelle 2015 (Ammor Ouafae)

[5]help.xlstat.com/customer/fr/portal/articles/2062457-what-statistical-test-should-i-use ?b_id=9283

[6]http://www.sthda.com/french/wiki/test-de-student-formules

[7]www.info.univ-angers.fr/~gh/wstat/Perfectionnement_R/mazerolle-khi-carre.pd f

[8]Polycopié du cours de statistique et probabilité de la licence CSA 2011(Ezzaki Fatima)

[9]wikilean.com/Articles/Analyse/5-Les-tests-d-hypotheses-25-articles/Test-de-Shapiro-Wilk

[10]http://jebrane.perso.math.cnrs.fr/m1psy/Tests_%20egalite_%20variances.pdf

[11]http://www.ecofog.gf/IMG/pdf/testsnonparametriques.pdf

[12]https://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf

[13]http://statsoft.fr/concepts-statistiques/classifications/classifications.htm

[14]eric.univ-lyon2.fr/~ricco/cours/slides/classif_centres_mobiles.pdf

[15]http://www.univ-usto.dz/theses_en_ligne/doc_num.php?explnum_id=829

[16]http://dept-info.labri.fr/~vialard/Image3D/cours/cours-segmentation.pdf

[17]https://www.saadrachid.net/bi-big-data/quest-big-data-definition-principe/

[18]https://www.futura-sciences.com/tech/definitions/informatique-big-data-15028/

[19]https://www.youtube.com/watch?v=V1a0RukTSt8

[20]newstoprotect.axa/sante/big-data-revolution-medecine-jean-pierre-thierry

[21]members.cbio.mines-paristech.fr/~jvert/publi/2016revuemines.pdf

[22]blog.mondocteur.fr/big-data-et-sante

[23]https://www.youtube.com/watch?v=Rt-gS4enlNE

[24]https://www.youtube.com/watch?v=UydeL192vL4