



# Projet Logiciel Statistique : Fertility

Présenté par :

Alexandra Marques
Lucie Guillaumin
Ninon Hersant
Sarra Chahdoura

# Sommaire

# Introduction

Pour ce projet, 100 volontaires ont fourni un échantillon de sperme. Celui-ci a été analysé selon les critères de l'OMS 2010.
La concentration de sperme est liée aux données sociodémographiques, aux facteurs environnementaux, à l'état de santé et aux habitudes de vie. Notre but ici est d'étudier grâce à plusieurs variables les causes de l'infertilité chez les hommes.
Nous avons décidé de faire ce projet sous SAS plutôt que sous R.
Pour simplifier les codes nous avons décidé d'utiliser des abréviations pour chaque variable (mise entre parenthèses après chaque variable énoncée).

Voici les différentes variables mises à notre disposition :

- Saison dans laquelle l'analyse a été effectuée ("saison") : 1) hiver 2) printemps 3) été 4) automne. (-1, -0,33, 0,33, 1). C'est une variable qualitative.

- Âge au moment de l'analyse ("age") : 18-36 (0, 1). C'est une variable quantitative discrète.

- Maladies infantiles (par exemple : varicelle, rougeole, oreillons, poliomyélite) ("maladies_infantiles") : 1) oui 2) non. (0, 1). C'est une variable qualitative.

- Accident ou traumatisme grave ("accident") : 1) oui 2) non. (0, 1). C'est une variable qualitative.

- Intervention chirurgicale ("chirurgie") : 1) oui 2) non. (0, 1). C'est une variable qualitative.

- Forte fièvre au cours de la dernière année ("fièvre") : 1) il y a moins de trois mois 2) il y a plus de trois mois 3) non. (-1, 0, 1). C'est une variable qualitative.

- Fréquence de la consommation d'alcool ("alcool") : 1) plusieurs fois par jour 2) tous les jours 3) plusieurs fois par semaine 4) une fois par semaine 5) presque jamais ou jamais (0, 1). C'est une variable continue, et ici nous allons la traiter comme une variable quantitative.

- Habitude à fumer ("fumer") : 1) jamais 2) à l'occasion 3) chaque jour. (-1, 0, 1). C'est une variable qualitative.

- Nombre d'heures passées assis par jour (“assis”) : (0, 1). C’est une variable quantitative continue.

- Sortie : Diagnostic normal (“diagnostic”) : (N) : diagnostic négatif, (O) : diagnostic positif. C’est une variable qualitative.

En vue de ces 10 variables, on remarque qu’il aurait pu manquer certaines variables afin d’avoir un échantillon plus large de facteurs pour lesquels l’homme soit infertile, comme par exemple :

- Consommation de drogue : 1) jamais  2) à l’occasion  3) régulièrement
- Continent d’habitat de la personne concernée : 1) Afrique  2) Europe 3) Océanie 4) Asie  5) Amérique
- Antécédent Obésité : 1) Oui  2) Non

De plus, parmis ces 10 variables nous trouvons surprenantes les variables suivantes :

- Nombre d'heure assis par jour.
- Saison dans laquelle l'analyse a été effectuée.

Nous verrons par la suite si ces dernières sont utiles et ont un réel impact sur l'infertilité des hommes. D’un premier abord ces variables nous semblent avoir peu de rapport avec le diagnostic final posé pour chaque homme.

# Importation des données dans SAS

La première étape de notre projet consiste à mettre notre jeu de données sous forme d'un tableau pour que cela soit plus lisible.

On commence par créer une bibliothèque que l'on appelle projet.
**libname projet "/users/licence/ii04686/Bureau/Fertility/";**

On va ensuite créer une macro pour ne pas se perdre dans l'ensemble de nos codes.
**%let var** = age maladies_infantiles accident chirurgie fievre alcool fumer assis winter spring summer fall;
*/*variables que l'on utilise le plus souvent dans les codes */*

Les codes ci-dessous nous permettront par la suite de construire notre table avec notre jeu de données.

Code : **data projet.fertility;**
infile "/users/licence/ii04686/Bureau/Fertility/fertility_Diagnosis.txt" firstobs=1 dlm=",";
input  saison age maladies_infantiles accident chirurgie fievre alcool fumer assis diagnostic $;
**run;**

Ici, nous avons utilisé un autre fichier nommé "fertility2" avec les nouvelles variables modifiées. Pour le reste du projet nous allons continuer avec ce fichier.

Code **: data projet.fertility2;**
set projet.fertility;
age=age*18+18;
age=floor(age);
**run;**

Code **: data projet.fertility2;**
set projet.fertility2;
if saison = - 1
then winter = 1;
if winter ="."
then winter = 0;
**run;**

Code : **data projet.fertility2;**
set projet.fertility2;
if diagnostic= "N" then diagnostic2 = 1;
if diagnostic = "O" then diagnostic2 = 0;
**run;**

**Commentaire :**

Pour que cela soit plus simple et plus compréhensible dans nos interprétations, nous avons décidé de changer plusieurs variables,

La variable “**âge**” codée entre 0 et 1, est transformée en âge concret entre 18 et 36 ans.

La variable “**saison**” qui est une variable circulaire, a elle aussi été modifiée afin de transformer cette unique variable en quatre variables indépendantes binaires (on remarquera que tout le codage n’a pas été retranscrit dans ce rapport, seul le codage pour “winter” l’est).

La variable “**alcool**” quant à elle, va être traitée quantitativement.

De plus, nous avons préféré mettre la variable “**diagnostic**” avec des valeurs binaires, 1 pour N(non) et 0 pour O(oui) afin de faciliter les calculs,de pouvoir calculer des moyennes mais aussi réaliser des comparaisons avec les autres variables.

Voici une partie du tableau du jeu de données que nous obtenons :

| Obs. | saison | age | maladies_infantiles | accident | chirurgie | fievre | alcool | fumer | assis | diagnostic | winter | spring | summer | fall | diagnostic2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | -0.33 | 30 | 0 | 1 | 1 | 0 | 0.8 | 0 | 0.88 | N | 0 | 1 | 0 | 0 | 1 |
| 2 | -0.33 | 34 | 1 | 0 | 1 | 0 | 0.8 | 1 | 0.31 | O | 0 | 1 | 0 | 0 | 0 |
| 3 | -0.33 | 27 | 1 | 0 | 0 | 0 | 1.0 | -1 | 0.50 | N | 0 | 1 | 0 | 0 | 1 |
| 4 | -0.33 | 31 | 0 | 1 | 1 | 0 | 1.0 | -1 | 0.38 | N | 0 | 1 | 0 | 0 | 1 |
| 5 | -0.33 | 30 | 1 | 1 | 0 | 0 | 0.8 | -1 | 0.50 | O | 0 | 1 | 0 | 0 | 0 |
| 6 | -0.33 | 30 | 1 | 0 | 1 | 0 | 0.8 | 0 | 0.50 | N | 0 | 1 | 0 | 0 | 1 |
| 7 | -0.33 | 30 | 0 | 0 | 0 | -1 | 0.8 | -1 | 0.44 | N | 0 | 1 | 0 | 0 | 1 |
| 8 | -0.33 | 36 | 1 | 1 | 1 | 0 | 0.6 | -1 | 0.38 | N | 0 | 1 | 0 | 0 | 1 |
| 9 | 1.00 | 29 | 0 | 0 | 1 | 0 | 0.8 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 10 | 1.00 | 28 | 1 | 0 | 0 | 0 | 1.0 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 11 | 1.00 | 30 | 1 | 1 | 0 | -1 | 0.8 | 0 | 0.31 | N | 0 | 0 | 0 | 1 | 1 |
| 12 | 1.00 | 32 | 1 | 1 | 1 | 0 | 0.6 | 0 | 0.13 | N | 0 | 0 | 0 | 1 | 1 |
| 13 | 1.00 | 31 | 1 | 1 | 1 | 0 | 0.8 | 1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 14 | 1.00 | 32 | 1 | 0 | 0 | 0 | 1.0 | -1 | 0.38 | N | 0 | 0 | 0 | 1 | 1 |
| 15 | 1.00 | 34 | 1 | 1 | 1 | 0 | 0.2 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 16 | 1.00 | 32 | 1 | 1 | 0 | 0 | 1.0 | 1 | 0.50 | N | 0 | 0 | 0 | 1 | 1 |
| 17 | 1.00 | 29 | 1 | 0 | 1 | 0 | 1.0 | -1 | 0.38 | N | 0 | 0 | 0 | 1 | 1 |
| 18 | 1.00 | 30 | 1 | 0 | 1 | 0 | 0.8 | -1 | 0.25 | O | 0 | 0 | 0 | 1 | 0 |
| 19 | 1.00 | 31 | 1 | 1 | 1 | 0 | 1.0 | 1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 20 | 1.00 | 30 | 1 | 0 | 0 | 0 | 0.8 | 1 | 0.38 | O | 0 | 0 | 0 | 1 | 0 |
| 21 | 1.00 | 30 | 0 | 0 | 1 | 0 | 0.8 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 22 | 1.00 | 31 | 1 | 0 | 0 | 0 | 0.6 | 0 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 23 | 1.00 | 30 | 1 | 1 | 0 | 0 | 0.8 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 24 | 1.00 | 30 | 1 | 0 | 1 | -1 | 1.0 | -1 | 0.44 | O | 0 | 0 | 0 | 1 | 0 |
| 25 | 1.00 | 28 | 1 | 0 | 1 | 0 | 1.0 | -1 | 0.63 | N | 0 | 0 | 0 | 1 | 1 |
| 26 | 1.00 | 30 | 1 | 0 | 0 | 0 | 1.0 | -1 | 0.25 | N | 0 | 0 | 0 | 1 | 1 |
| 27 | 1.00 | 30 | 1 | 0 | 1 | 0 | 0.6 | -1 | 0.38 | O | 0 | 0 | 0 | 1 | 0 |
| 28 | 1.00 | 32 | 1 | 1 | 0 | 1 | 0.6 | -1 | 0.38 | O | 0 | 0 | 0 | 1 | 0 |
| 29 | 1.00 | 28 | 0 | 0 | 1 | 0 | 1.0 | -1 | 0.19 | N | 0 | 0 | 0 | 1 | 1 |
| 30 | 1.00 | 30 | 0 | 0 | 1 | 0 | 0.6 | 0 | 0.50 | O | 0 | 0 | 0 | 1 | 0 |
| 31 | 1.00 | 28 | 1 | 0 | 1 | 0 | 1.0 | -1 | 0.63 | N | 0 | 0 | 0 | 1 | 1 |
| 32 | 1.00 | 28 | 1 | 0 | 0 | 0 | 1.0 | -1 | 0.44 | N | 0 | 0 | 0 | 1 | 1 |

# Analyse descriptive du jeu de données

Deuxièmement, nous souhaitons étudier et analyser variable après variable, les moyennes, écarts-types, médianes et fréquences de celles-ci.

Tout d'abord, grâce à la procédure **means**, nous obtenons :

La procédure MEANS

| Variable | N | Moyenne | Ec-type | Minimum | Maximum |
|---|---|---|---|---|---|
| saison | 100 | -0.0789000 | 0.7967255 | -1.0000000 | 1.0000000 |
| age | 100 | 29.6900000 | 2.1541363 | 27.0000000 | 36.0000000 |
| maladies_infantiles | 100 | 0.8700000 | 0.3379977 | 0 | 1.0000000 |
| accident | 100 | 0.4400000 | 0.4988877 | 0 | 1.0000000 |
| chirurgie | 100 | 0.5100000 | 0.5024184 | 0 | 1.0000000 |
| fievre | 100 | 0.1900000 | 0.5807519 | -1.0000000 | 1.0000000 |
| alcool | 100 | 0.8320000 | 0.1675009 | 0.2000000 | 1.0000000 |
| fumer | 100 | -0.3500000 | 0.8087276 | -1.0000000 | 1.0000000 |
| assis | 100 | 0.4068000 | 0.1863953 | 0.0600000 | 1.0000000 |
| winter | 100 | 0.2800000 | 0.4512609 | 0 | 1.0000000 |
| spring | 100 | 0.3700000 | 0.4852366 | 0 | 1.0000000 |
| summer | 100 | 0.0400000 | 0.1969464 | 0 | 1.0000000 |
| fall | 100 | 0.3100000 | 0.4648232 | 0 | 1.0000000 |
| diagnostic2 | 100 | 0.8800000 | 0.3265986 | 0 | 1.0000000 |

Code : **proc means** data = projet.fertility2;
**run;**

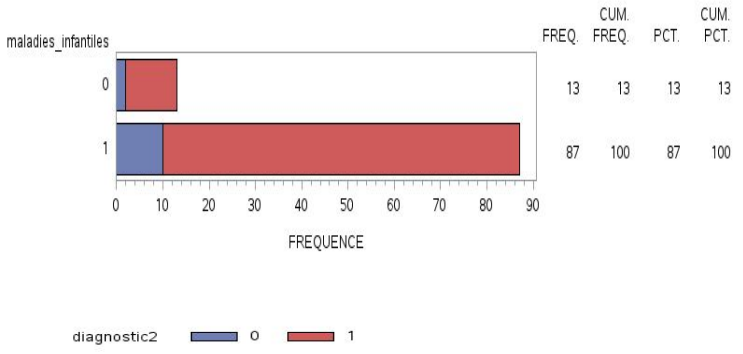
**Commentaire** :
Ici, nous allons faire une analyse des variables dont les histogrammes (on verra par la suite les variables dont il est intéressant d'avoir un histogramme) ne nous donnent pas assez d'informations car ce sont des variables binaires.

De plus, à l'aide de la procédure **gchart**, qui produit des diagrammes en bâtons, nous pourrons étudier la répartition des individus pour chaque variable, notamment pour savoir si il y a plus de diagnostic normal ou modifié.

Code : **proc gchart** data = projet.fertility2;
hbar age maladies_infantiles accident chirurgie fievre alcool fumer assis saison /discrete
subgroup = diagnostic2;
pie age maladies_infantiles accident chirurgie fievre alcool fumer assis saison;**run;**

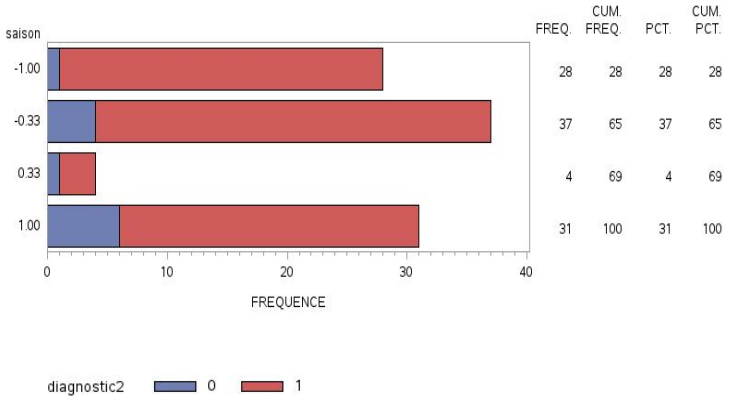
**quit;**

**Variable "maladies_infantiles"**




**Commentaire :**
On rappelle ici que 0 correspond à "Oui la personne sondée a eu une maladie lorsqu'elle été enfant" et 1 correspond à "Non elle n'en a pas eu".
La moyenne de la variable **"maladies_infantiles"** est de 87%, ce qui est relativement proche de 100%, en moyenne les hommes n'ont donc pas souffert de maladies lorsqu'ils étaient jeunes. On peut le confirmer grâce à la fréquence : sur nos 100 personnes sondées seulement 13 ont eu des maladies infantiles dont une diagnostiquée stérile.

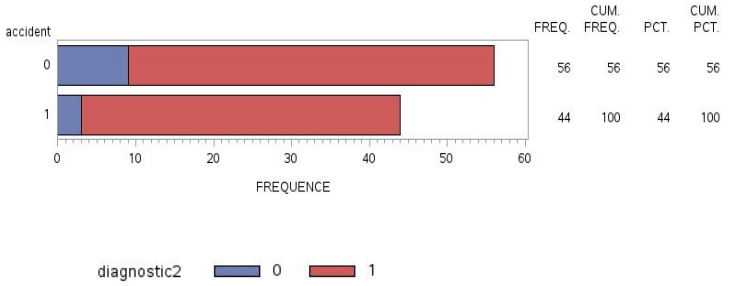
## Variable “saison”




## Commentaire :

On rappelle que -1 correspond à l’hiver, -0.33 au printemps, 0.33 à l’été et 1 à l’automne.
On remarque que pour la variable **“saison”,** le printemps, l’automne et l’hiver sont les saisons où les tests ont été effectués en majorité.
Seulement 4 personnes ont fait leur test en été, on peut aussi le voir grâce à la moyenne qui est de 4% pour la variable **“summer”.**

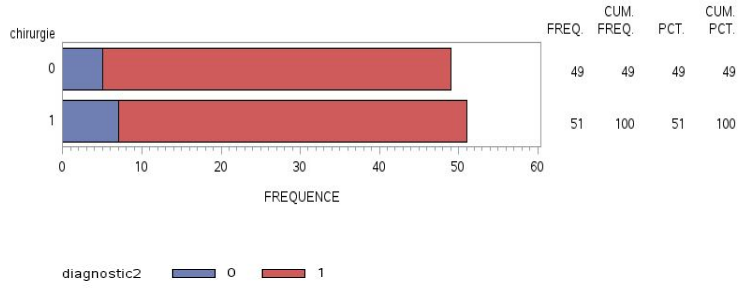
## Variable “accident”




## Commentaire :

On rappelle ici que 0 correspond à “Oui la personne sondée a eu un accident ou un traumatisme grave” et 1 correspond à “Non elle n’en a pas eu”.
La moyenne de la variable **“accident”** est de 44%, proche de la moitié.
Ce chiffre est plus proche de 0 que de 100%, on peut en déduire qu’en moyenne ils ont eu un accident ou un traumatisme grave.
A l’aide de la fréquence, on remarque qu’il y a presque autant de personnes qui ont eu un accident ou traumatisme grave que ceux qui n’en ont pas eu l’expérience.
On peut supposer que si la personne a eu un accident ou un traumatisme grave, celle-ci a plus de chance d’être infertile.

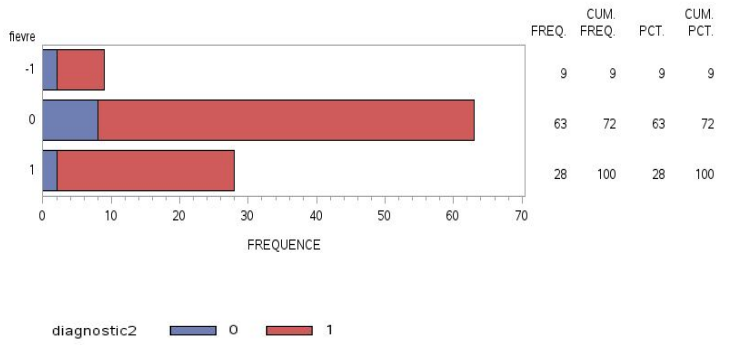
## Variable “chirurgie”




## Commentaire :

On rappelle ici que 0 correspond à “Oui la personne sondée a fait de la chirurgie”, donc 1 correspond à “Non elle n’en a pas fait”.
La moyenne de la variable **“chirurgie”** est de 51%, celle-ci est plus proche de 100% (de très peu) que de 0, il y a donc plus d’hommes qui n’ont pas eu recours à la chirurgie. Sur les 49 personnes ayant fait de la chirurgie, 5 sont stériles.

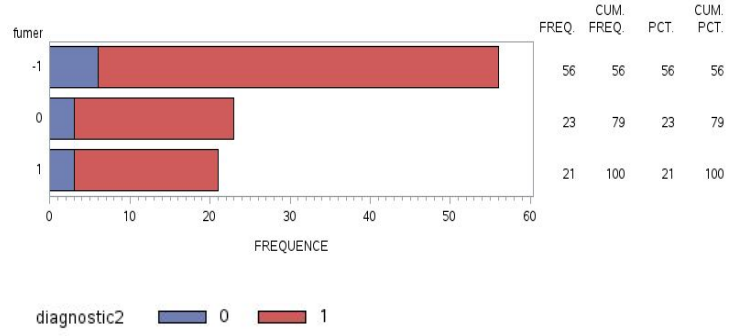
## Variable “fièvre”




## Commentaire :

On rappelle ici que -1 correspond à “La personne sondée a eu une forte fièvre il y a moins de 3 mois”, 0 correspond à “La personne sondée a eu une forte fièvre il a plus de 3 mois” et 1 correspond à “Non elle n’en a pas eu”.
La moyenne de la variable **“fièvre”** est de 19%, relativement proche de 0 qui dans notre cas correspond à une forte fièvre il y a plus de trois mois.
Avec la fréquence, on remarque que la plupart des personnes sondées ont eu une forte fièvre il y a plus de 3 mois.
On peut poser l’hypothèse qu’une personne qui a eu une forte fièvre il y a plus de 3 mois a plus de risque d’être infertile.

## Variable “fumer”



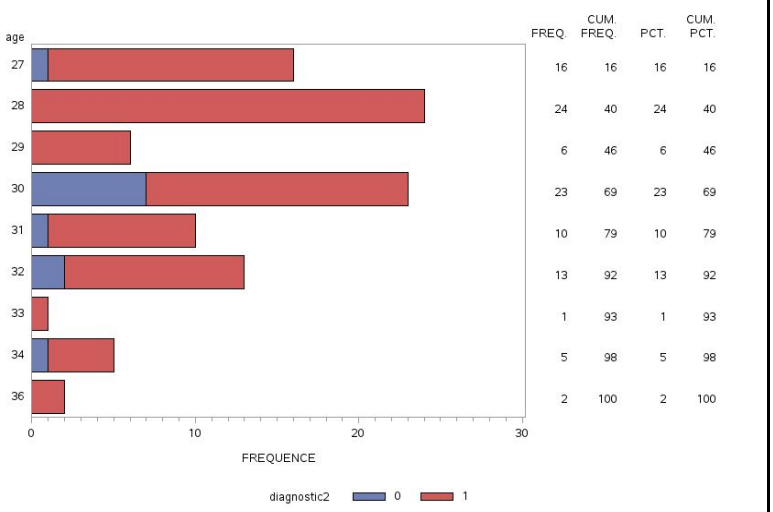

## Commentaire :

On rappelle ici que -1 correspond à “La personne sondée ne fume jamais”, 0 correspond à “La personne sondée fume de temps en temps” et 1 correspond à “La personne sondée fume chaque jour”.
La moyenne de notre variable **“fumer”** est de 35% ainsi en moyenne les personnes fument occasionnellement. On remarque que la plupart des personnes sondées ne fument pas ou juste occasionnellement. Sur les 21 personnes qui fument de façon journalière, 3 sont infertiles.

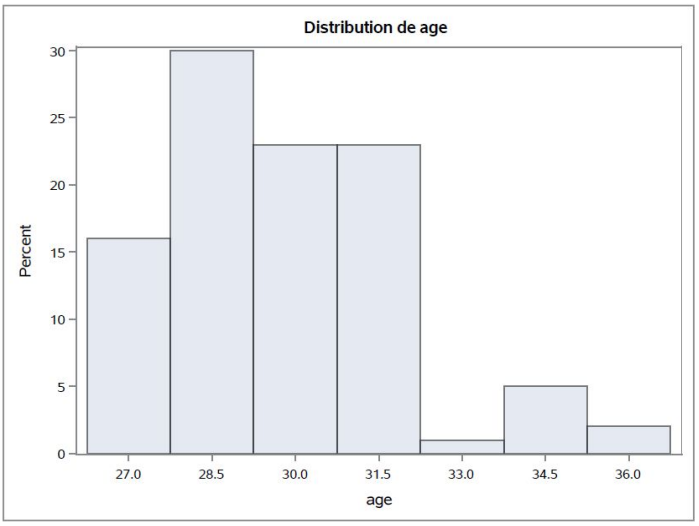
La procédure **univariate** produit les indicateurs statistiques traditionnels et analyse de manière approfondie la distribution d’une série de variables numériques, nous pourrons alors faire une étude plus précise de chaque variable. Pour certaines variables il sera avantageux d’afficher l'histogramme correspondant, pour d’autres, comme dit précédemment, cela n’a pas forcément d'intérêt notamment pour les variables binaires.

Code : **proc univariate** data = projet.fertility2;
var age;
histogram age;
title “Distribution de age”;
**run;** (on utilise aussi ce code pour les variables alcool et assis)

## Variable “age”



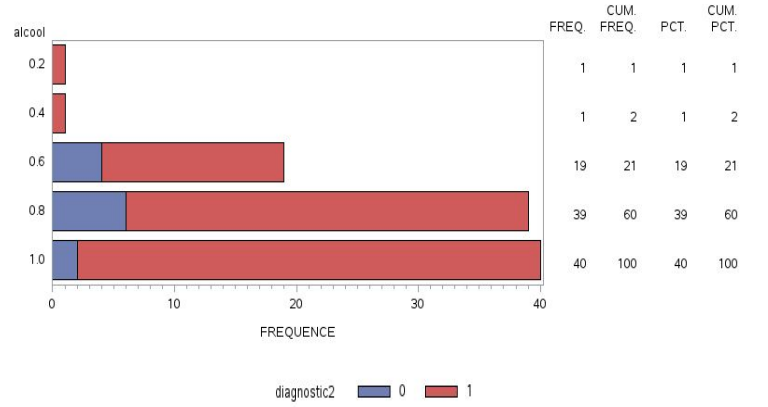

| age | FREQ. | CUM. FREQ. | PCT. | CUM. PCT. |
|---|---|---|---|---|
| 27 | 16 | 16 | 16 | 16 |
| 28 | 24 | 40 | 24 | 40 |
| 29 | 6 | 46 | 6 | 46 |
| 30 | 23 | 69 | 23 | 69 |
| 31 | 10 | 79 | 10 | 79 |
| 32 | 13 | 92 | 13 | 92 |
| 33 | 1 | 93 | 1 | 93 |
| 34 | 5 | 98 | 5 | 98 |
| 36 | 2 | 100 | 2 | 100 |



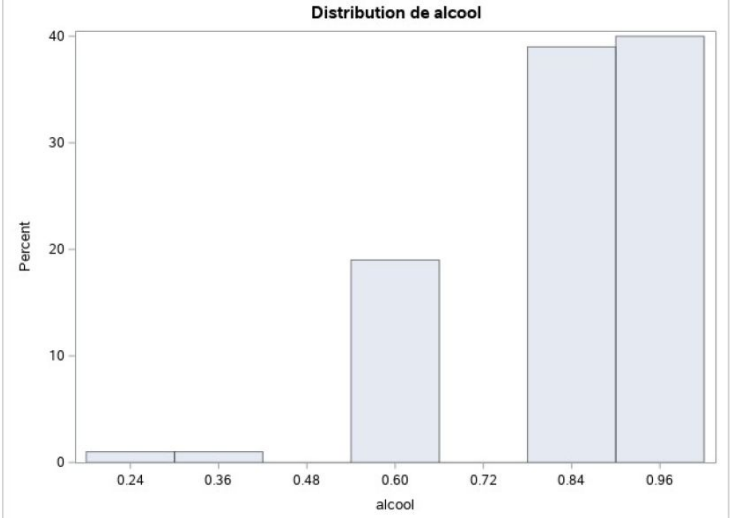

**Commentaire:**

Sur les 100 personnes que l’on sonde, la moyenne d'âge est de 29 ans et demi. L’âge minimum est de 27 ans et le maximum de 36 ans. On remarque ici que la plupart des personnes interrogées ont entre 27 et 31.5 ans.

## Variable “alcool”



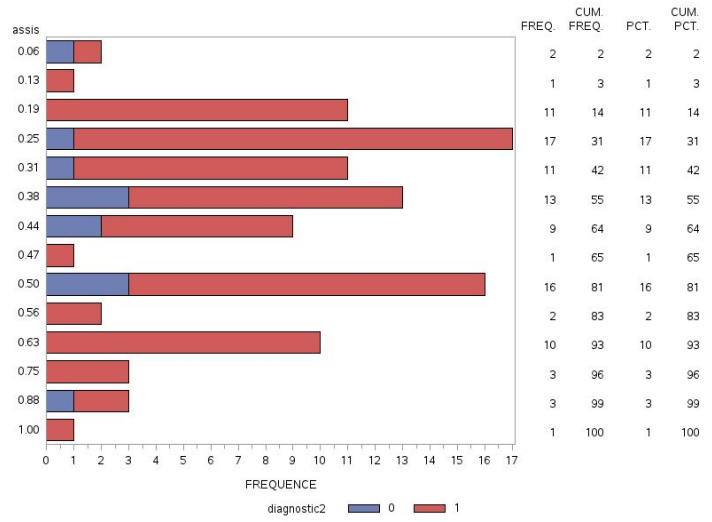

| alcool | FREQ. | CUM. FREQ. | PCT. | CUM. PCT. |
|---|---|---|---|---|
| 0.2 | 1 | 1 | 1 | 1 |
| 0.4 | 1 | 2 | 1 | 2 |
| 0.6 | 19 | 21 | 19 | 21 |
| 0.8 | 39 | 60 | 39 | 60 |
| 1.0 | 40 | 100 | 40 | 100 |



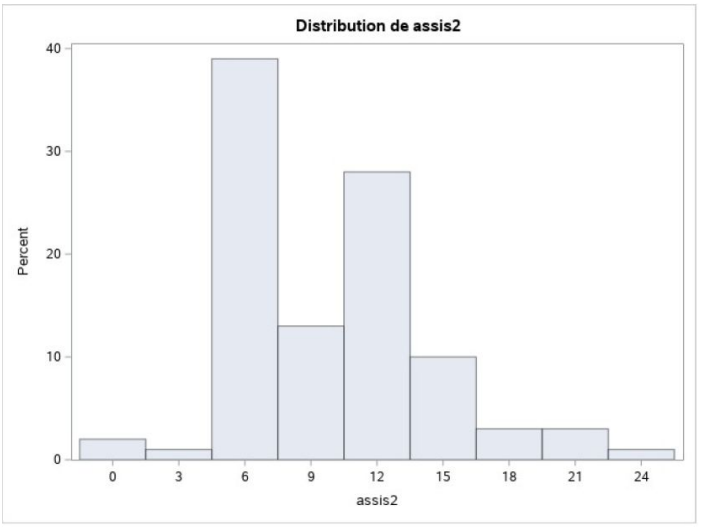

**Commentaire :**

On rappelle que la variable alcool est traitée comme variable quantitative.
La moyenne est de 0.83, globalement les personnes interrogées ne boivent quasiment jamais. Il y a un fort pourcentage (entre 39 et 40%) de personnes sondées qui boivent soit une fois par semaine, soit presque jamais d’alcool. Sur les 39 personnes buvant de l’alcool une fois par semaine, 6 sont stériles.
On peut donc penser qu’à partir du moment où la personne bois plusieurs fois par semaine, elle a plus de risque d’être infertile.

## Variable “assis”




| assis | FREQ. | CUM. FREQ. | PCT. | CUM. PCT. |
|---|---|---|---|---|
| 0.06 | 2 | 2 | 2 | 2 |
| 0.13 | 1 | 3 | 1 | 3 |
| 0.19 | 11 | 14 | 11 | 14 |
| 0.25 | 17 | 31 | 17 | 31 |
| 0.31 | 11 | 42 | 11 | 42 |
| 0.38 | 13 | 55 | 13 | 55 |
| 0.44 | 9 | 64 | 9 | 64 |
| 0.47 | 1 | 65 | 1 | 65 |
| 0.50 | 16 | 81 | 16 | 81 |
| 0.56 | 2 | 83 | 2 | 83 |
| 0.63 | 10 | 93 | 10 | 93 |
| 0.75 | 3 | 96 | 3 | 96 |
| 0.88 | 3 | 99 | 3 | 99 |
| 1.00 | 1 | 100 | 1 | 100 |




## Commentaire :

Concernant la variable “**assis”**, nous allons la modifier. Nous allons multiplier nos variables par 24 pour avoir un résultat réaliste en terme d’heures afin de faire une analyse plus approfondie de cette variable.
Nous avons donc créé une autre variable nommée “**assis2**” (où assis2 = assis * 24)
On remarque que les hommes interrogés restent en moyenne 9,6 heures assis par jour, ce qui paraît raisonnable car cela équivaut à une journée de travail en bureau.
De plus, on voit que seulement 35 personnes sur les 100 restent assises plus d’une demi journée, et notamment une personne qui reste immobile pendant 24h.
Sur les 16 personnes restant une demi journée assises, 3 sont stériles.

Ensuite, étudions la normalité de nos variables continues “**assis**” et “**alcool**” :

Code : **proc univariate** data = projet.fertility2 **NORMALTEST**;
var assis ;
histogram assis / NORMAL ;
probplot assis / NORMAL(MU=est SIGMA=est COLOR=red L=1) ;
qqplot assis / NORMAL(MU=est SIGMA=est COLOR=red L=1);
**run;**

## Test de normalité pour la variable “assis”

| Tests de normalité | | | | |
|---|---|---|---|---|
| Test | Statistique | | p-value | |
| Shapiro-Wilk | W | 0.943341 | Pr < W | 0.0003 |
| Kolmogorov-Smirnov | D | 0.118533 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 0.259782 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 1.714959 | Pr > A-Sq | <0.0050 |

**Commentaire** :

On pose les hypothèses h0 la variable “**assis**” est gaussienne et h1 elle ne l’est pas. Grâce au test de Shapiro, on remarque que la p-value est très faible, et on peut donc affirmer que la variable n’est pas gaussienne avec un risque d’erreur de 0.03%.

## Test de normalité pour la variable “alcool”

| Tests de normalité | | | | |
|---|---|---|---|---|
| Test | Statistique | | p-value | |
| Shapiro-Wilk | W | 0.811623 | Pr < W | <0.0001 |
| Kolmogorov-Smirnov | D | 0.242065 | Pr > D | <0.0100 |
| Cramer-von Mises | W-Sq | 1.189618 | Pr > W-Sq | <0.0050 |
| Anderson-Darling | A-Sq | 7.29269 | Pr > A-Sq | <0.0050 |

**Commentaire** :

On pose les hypothèses h0 la variable “**alcool**” est gaussienne et h1 elle ne l’est pas. Grâce au test de Shapiro, on remarque que la p-value est très faible, et on peut donc affirmer que la variable n’est pas gaussienne avec un risque d’erreur inférieur à 0.01%

# Analyse Bivariée

L'analyse descriptive bivariée permet d'analyser le lien entre deux variables, que ce soit grâce à des coefficients de corrélation ou à des nuages de points.

Nous allons commencer par étudier la corrélation entre la variable **"diagnostic"** et les autres variables afin de connaître celle qui influe le plus sur le diagnostic.
La procédure **corr** permet d'obtenir des coefficients de corrélations entre les variables ainsi que plusieurs statistiques descriptives et différents tests que l'on va supprimer de la sortie de la procédure avec l'option **NOSIMPLE**.

La procédure CORR

| | |
|---|---|
| 1 Avec les variables : | diagnostic2 |
| 12 Variables : | age maladies_infantiles accident chirurgie fievre alcool fumer assis winter spring summer fall |

Coefficients de corrélation de Pearson, N = 100
Proba > |r| sous H0: Rho=0

| | age | maladies_infantiles | accident | chirurgie | fievre | alcool | fumer |
|---|---|---|---|---|---|---|---|
| diagnostic2 | -0.13955<br>0.1661 | 0.04026<br>0.6908 | 0.14135<br>0.1607 | -0.05417<br>0.5924 | 0.12142<br>0.2288 | 0.14476<br>0.1507 | -0.04589<br>0.6503 |

| assis | winter | spring | summer | fall |
|---|---|---|---|---|
| -0.02296<br>0.8206 | 0.16175<br>0.1079 | 0.02804<br>0.7818 | -0.08166<br>0.4193 | -0.15170<br>0.1319 |

Code :
```
proc corr data = projet.fertility2 NOSIMPLE;
    var &var;
    with diagnostic2;
    run;
```

## Commentaire :

La méthode de corrélation de Pearson calcule un coefficient de corrélation appelé paramétrique. Le coefficient de corrélation peut avoir une valeur comprise entre -1 et +1. Plus la valeur absolue du coefficient est importante, plus la relation linéaire entre les variables est forte.
Le signe du coefficient indique la direction de la relation. Si les deux variables ont tendance à augmenter ou à diminuer ensemble, le coefficient est positif, et la ligne qui représente la corrélation s'incline vers le haut. Si une variable a tendance à augmenter lorsque l'autre diminue, le coefficient est négatif, et la ligne représentant la corrélation s'incline vers le bas.
Ici, nous avons autant de coefficients de corrélation positifs que négatifs.
Le coefficient de corrélation le plus important est 0.16175, donc plus les tests sont faits en hiver, plus il y a de chance qu'une grande majorité de personnes soit diagnostiquée stériles. Également, plus la personne boit, plus elle a de chance d'avoir un diagnostic modifié (d'être stérile).
Au contraire, nous voyons que la corrélation entre la variable diagnostic et la variable chirurgie est négative, ce n'est donc pas parce que la personne a fait de la chirurgie qu'elle aura forcément des chances d'être infertile.

# Régressions logistiques

Dans notre cas nous ne pouvons faire de régression simple et multiple car notre variable de sortie est binaire, nous allons donc faire une régression logistique simple et multiple.
Par ailleurs, la régression logistique est un modèle de régression binomiale. Elle vise à construire un modèle permettant de prédire et d'expliquer les valeurs prises par une variable qualitative, ici notre variable est binaire.
Nous commencerons par la régression logistique multiple, en effet celle-ci va nous permettre de choisir la ou les variables pour notre régression logistique simple.

## a) Régression Logistique Multiple

Grâce à la **proc logistic** nous obtenons le tableau suivant :

| Estimation du rapport de cotes | | | |
|---|---|---|---|
| Effet | Estimation du point | Intervalle de confiance de Wald à95% | |
| age | 0.698 | 0.469 | 1.041 |
| maladies_infantiles | 0.716 | 0.104 | 4.937 |
| accident | 7.048 | 1.185 | 41.915 |
| chirurgie | 0.796 | 0.184 | 3.433 |
| fievre | 2.546 | 0.636 | 10.198 |
| alcool | 16.439 | 0.288 | 937.856 |
| fumer | 0.734 | 0.310 | 1.740 |
| assis | 0.046 | <0.001 | 3.703 |
| winter | 6.833 | 0.593 | 78.742 |
| spring | 1.782 | 0.364 | 8.715 |
| summer | 0.626 | 0.038 | 10.308 |

Code : **proc logistic** data = projet.fertility2;
model diagnostic = &var;
**run;**

**Commentaire :**

Grâce au rapport de cotes, on remarque que :

- "Être stérile" est indépendant de la variable spring car l'odds ratio n'est pas loin de 1 (1.782).
- "Être stérile" est plus fréquent chez les individus qui ont eu un accident ou traumatisme grave, de la fièvre, boivent de l'alcool et ont fait leur test en hiver. L'odds ratio est supérieur à 1 donc le rapport de probabilités chez les sujets exposés est plus grand que le rapport chez les sujets non exposés, on en déduit que le risque d'infertilité est plus élevé chez les sujets exposés.
- "Être stérile" n'est pas influencé ou très peu par les maladies infantiles qu'ils ont pu avoir, le recours à la chirurgie, le fait de fumer, de rester longtemps assis mais également par le fait d'avoir fait leurs test en été ou en automne. De plus l'âge n'influe pas sur la stérilité.

L'odds est inférieur à 1 donc le rapport de probabilité chez les sujets exposés est plus petit que le rapport chez les sujets non exposés, on en déduit que le risque de maladie est moindre chez les sujets exposés que chez les sujets non exposés.

Afin de vérifier ces analyses, nous allons enlever au fur et à mesure de notre régression logistique les variables non significatives et comparer l'AIC (Critère d'information d'Akaike), qui mesure la qualité d'un modèle statistique, pour savoir si elles influent sur le diagnostic ou non. Et d'obtenir le modèle le plus efficace.

Code : **proc genmod** data = projet.fertility2 descending;
model diagnostic2 = age maladies_infantiles accident chirurgie fievre alcool fumer assis winter spring summer fall / DIST=BINOMIAL LINK=LOGIT;
**run;**

Nous obtenons un grand nombre de tableaux et décidons de présenter celui ou l'AIC est présent. Ici nous avons utiliser la procédure **genmod** avec toutes nos variables :

| Critères d'évaluation de l'adéquation | | | |
|---|---|---|---|
| Critère | DDL | Valeur | Valeur/DDL |
| Log-vraisemblance | | -29.0866 | |
| Log-vraisemblance complète | | -29.0866 | |
| AIC (préférer les petites valeurs) | | 82.1732 | |
| AICC (préférer les petites valeurs) | | 85.7594 | |
| BIC (préférer les petites valeurs) | | 113.4352 | |

L'algorithme a convergé.

Nous commençons par enlever la variable "**spring**" qui est pour nous une des variables les moins importantes pour le diagnostic. On obtient le tableau suivant :

| Critères d'évaluation de l'adéquation | | | |
|---|---|---|---|
| Critère | DDL | Valeur | Valeur/DDL |
| Log-vraisemblance | | -29.0866 | |
| Log-vraisemblance complète | | -29.0866 | |
| AIC (préférer les petites valeurs) | | 82.1732 | |
| AICC (préférer les petites valeurs) | | 85.7594 | |
| BIC (préférer les petites valeurs) | | 113.4352 | |

L'algorithme a convergé.

Nous remarquons que l'AIC ne change pas, ce qui appuie sur le fait que la variable "spring" n'influe pas sur la stérilité et qu'elle est indépendante.
Comme précédemment, nous avons enlever dans cet ordre les variables :

1) fall
2) summer
3) assis
4) âge
5) fumer
6) chirurgie
7) maladie_infantile

Nous avons, au vu des résultats précédents, retiré les variables une à une, en commençant par celles qui étaient le moins impactante pour le diagnostic.
En effet, toutes les variables ci-dessus font baisser la valeur de l'AIC, nous observons alors qu'elles n'ont aucune influence sur la stérilité des hommes. Si une de celles-ci avait fait augmenter l'AIC la variable aurait donc été importante de garder dans le modèle.
C'est ce qu'il c'est passé pour les variables suivantes :

| | |
|---|---|
| En effet, lorsque nous enlevons la variable "**winter**" nous obtenons une augmentation de l'AIC par rapport à celle enregistrée une fois avoir enlevé les 8 variables précédentes : |  |
| De même pour la variable "**fièvre**" :<br>(AIC a augmenté de 0.3548) |  |
| Ainsi que pour la variable "**accident**" :<br>(AIC a augmenté de 1.9437) |  |

Critères d'évaluation de l'adéquation

| Critère | DDL | Valeur | Valeur/DDL |
|---|---|---|---|
| Log-vraisemblance | | -32.7452 | |
| Log-vraisemblance complète | | -32.7452 | |
| AIC (préférer les petites valeurs) | | 73.4903 | |
| AICC (préférer les petites valeurs) | | 73.9114 | |
| BIC (préférer les petites valeurs) | | 83.9110 | |

L'algorithme a convergé.

Critères d'évaluation de l'adéquation

| Critère | DDL | Valeur | Valeur/DDL |
|---|---|---|---|
| Log-vraisemblance | | -33.9225 | |
| Log-vraisemblance complète | | -33.9225 | |
| AIC (préférer les petites valeurs) | | 73.8451 | |
| AICC (préférer les petites valeurs) | | 74.0951 | |
| BIC (préférer les petites valeurs) | | 81.6606 | |

L'algorithme a convergé.

Critères d'évaluation de l'adéquation

| Critère | DDL | Valeur | Valeur/DDL |
|---|---|---|---|
| Log-vraisemblance | | -35.7170 | |
| Log-vraisemblance complète | | -35.7170 | |
| AIC (préférer les petites valeurs) | | 75.4340 | |
| AICC (préférer les petites valeurs) | | 75.5577 | |
| BIC (préférer les petites valeurs) | | 80.6444 | |

L'algorithme a convergé.

Nous constatons grâce à ces différents tableaux, que les seules variables influant sur la stérilité sont les suivantes : **"winter"**, **"fièvre"**, **"accident"** et **"alcool"**. Le modèle le plus intéressant est le modèle qui a la plus faible valeur de AIC. On s'arrête ainsi après avoir enlevé **"maladie_infantile"**.

## b) Régression Logistique Simple

Dans la régression logistique simple contrairement à la multiple, on évalue à chaque étape si certaines variables devraient être retirées en se basant sur :

Le **rapport de vraisemblance** (*likelihood-ratio,* LR) : on conserve la variable si le changement du LR est significatif quand la variable est retirée, ce qui indique que cette variable contribue à la qualité de l'ajustement.

La **statistique conditionnelle** : il s'agit d'un critère moins exigeant que le LR, donc il est préférable de prioriser le 1er

La **statistique Wald** : cette fois, on retire toutes les variables pour lesquelles la statistique Wald est inférieure à 0,1. Cette méthode peut être utilisée avec un petit échantillon. Sinon, il est préférable de privilégier le LR.

Dans un premier temps nous avons décidé de faire une régression logistique simple sur les quatre variables les plus significatives que nous avons pu identifier grâce à la régression logistique multiple. Nous allons ensuite examiner leurs valeurs pour le test de Wald ainsi que le rapport de vraisemblance du modèle. Pour finir, nous analyserons le coefficient B.

Code : **proc logistic** data = projet.fertility2;
model diagnostic = alcool;
**run;**

| | |
|---|---|
| Pour la variable "**alcool"**, nous constatons que la valeur du test de Wald ainsi que le rapport de vraisemblance est élevée. | Test de l'hypothèse nulle globale : BETA=0<br>Test / khi-2 / DDL / Pr > khi-2<br>Rapport de vrais / 1.9510 / 1 / 0.1625<br>Score / 2.0956 / 1 / 0.1477<br>Wald / 2.0066 / 1 / 0.1566 |
| Pour la variable "**accident"**, nous observons également une valeur du test de Wald et une vraisemblance élevée par rapport aux autres variables essayées ( "summer", "fall", "assis"... Nous n'avons pas trouvé utile de les afficher dans ce rapport). | Test de l'hypothèse nulle globale : BETA=0<br>Test / khi-2 / DDL / Pr > khi-2<br>Rapport de vrais / 2.1054 / 1 / 0.1468<br>Score / 1.9979 / 1 / 0.1575<br>Wald / 1.8883 / 1 / 0.1694 |

Pour la variable “**winter”**, on voit comme les 2 variables précédentes un rapport de vraisemblance et un rapport de Wald élevé.

| Test de l'hypothèse nulle globale : BETA=0 | | | |
|---|---|---|---|
| Test | khi-2 | DDL | Pr > khi-2 |
| Rapport de vrais | 3.1971 | 1 | 0.0738 |
| Score | 2.6162 | 1 | 0.1058 |
| Wald | 2.1894 | 1 | 0.1390 |

Il en est de même pour la variable “**fièvre”** .

| Test de l'hypothèse nulle globale : BETA=0 | | | |
|---|---|---|---|
| Test | khi-2 | DDL | Pr > khi-2 |
| Rapport de vrais | 1.4824 | 1 | 0.2234 |
| Score | 1.4743 | 1 | 0.2247 |
| Wald | 1.4472 | 1 | 0.2290 |

Ces analyses nous autorise à dire que ces variables sont également importantes pour notre étude. Ceci nous permet encore une fois de voir que ces variables ont un impact sur le diagnostic final. Nous n’apprenons donc rien de plus avec ces coefficients. Intéressons nous maintenant à l’analyse de B, ce coefficient nous donne un peu plus d’informations sur nos variables.

## Variable alcool

Nous constatons que le coefficient B vaut 2.4268.
On en déduit que exp(2.4268) est l’odds ratio qui associe la variable explicative (alcool) à la variable à expliquer (diagnostic) et vaut 11.32, ce qui signifierait que boire de l’alcool multiplierait par 11 la probabilité d’être stérile.

| Analyse des valeurs estimées du maximum de vraisemblance | | | | | |
|---|---|---|---|---|---|
| Paramètre | DDL | Estimation | Erreur type | Khi-2 de Wald | Pr > khi-2 |
| Intercept | 1 | 0.0371 | 1.3716 | 0.0007 | 0.9784 |
| alcool | 1 | 2.4268 | 1.7132 | 2.0066 | 0.1566 |

## Variable accident

Nous constatons que le coefficient B vaut 0.9620.
On en déduit que exp(0.9620) est l’odds ratio qui associe la variable explicative (accident) à la variable à expliquer (diagnostic) et vaut 2.61, ce qui signifierait qu’avoir eu un accident ou un traumatisme grave multiplierait par plus de 2.5 la probabilité d’être stérile.

| Analyse des valeurs estimées du maximum de vraisemblance | | | | | |
|---|---|---|---|---|---|
| Paramètre | DDL | Estimation | Erreur type | Khi-2 de Wald | Pr > khi-2 |
| Intercept | 1 | 1.6529 | 0.3639 | 20.6375 | <.0001 |
| accident | 1 | 0.9620 | 0.7001 | 1.8883 | 0.1694 |

## Variable fièvre

Nous constatons que le coefficient B vaut 0.6550.
On en déduit que exp(0.6550) est l'odds ratio qui associe la variable explicative (fièvre) à la variable à expliquer (diagnostic) et vaut 1.92, ce qui signifierait qu'avoir eu de la fièvre peu ou fortement multiplierait par presque 2 la probabilité d'être stérile.

| Analyse des valeurs estimées du maximum de vraisemblance | | | | | |
|---|---|---|---|---|---|
| Paramètre | DDL | Estimation | Erreur type | Khi-2 de Wald | Pr > khi-2 |
| Intercept | 1 | 1.9217 | 0.3105 | 38.3058 | <.0001 |
| fievre | 1 | 0.6550 | 0.5444 | 1.4472 | 0.2290 |

## Variable winter

Nous constatons que le coefficient B vaut 1.5828.
On en déduit que exp(1.5828) est l'odds ratio qui associe la variable explicative (winter) à la variable à expliquer (diagnostic) et vaut 4.86, ce qui signifierait qu'avoir fait ces tests en période hivernale multiplierait par presque 5 la probabilité d'être stérile.

| Analyse des valeurs estimées du maximum de vraisemblance | | | | | |
|---|---|---|---|---|---|
| Paramètre | DDL | Estimation | Erreur type | Khi-2 de Wald | Pr > khi-2 |
| Intercept | 1 | 1.7130 | 0.3276 | 27.3460 | <.0001 |
| winter | 1 | 1.5828 | 1.0697 | 2.1894 | 0.1390 |

# Analyse en composantes principales

L'analyse en composantes principales permet de réduire la dimension de l'échantillon pour pouvoir mieux faire une analyse statistique descriptive.
L'ACP sert à décrire un jeu de données comportant de nombreux individus et variables quantitatives. L'analyse permet d'extraire l'information pertinente et la synthétise sous forme de composantes principales, nouveaux axes pour décrire le jeu de données. Elle permet aussi de quantifier les corrélations entre les variables du jeu de données. Des groupes de variables ayant des tendances identiques sont identifiés sur le cercle des corrélations.

| Matrice de corrélation | | | |
|---|---|---|---|
| | age | alcool | assis |
| age | 1.0000 | -.2522 | -.4307 |
| alcool | -.2522 | 1.0000 | 0.1114 |
| assis | -.4307 | 0.1114 | 1.0000 |

| Valeurs propres de la matrice de corrélation | | | | |
|---|---|---|---|---|
| | Valeur propre | Différence | Proportion | Cumulé |
| 1 | 1.55249716 | 0.64838742 | 0.5175 | 0.5175 |
| 2 | 0.90410974 | 0.36071664 | 0.3014 | 0.8189 |
| 3 | 0.54339310 | | 0.1811 | 1.0000 |

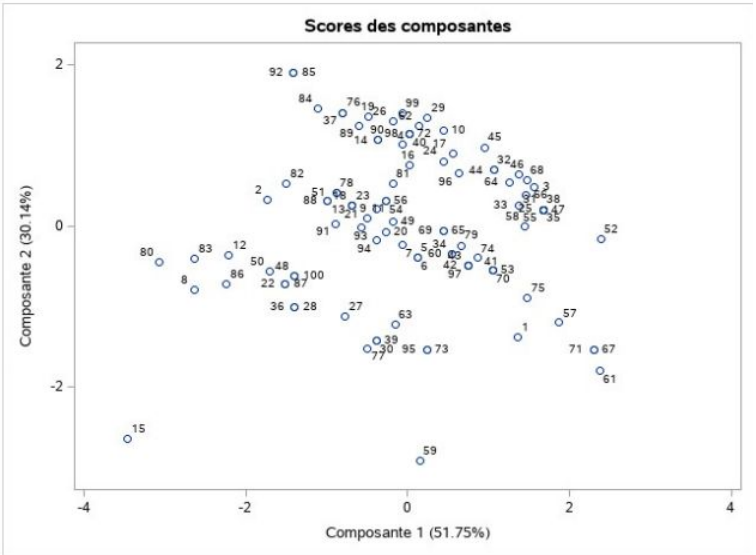
| Vecteurs propres | | | |
|---|---|---|---|
| | Prin1 | Prin2 | Prin3 |
| age | -.669017 | 0.117128 | 0.733960 |
| alcool | 0.427865 | 0.868158 | 0.251462 |
| assis | 0.607740 | -.482269 | 0.630927 |

Code : **proc princomp** data= projet.fertility2 plots=ellipse;
var age alcool assis;
**run;**

**Commentaire :**

On va choisir de retenir les axes dont les valeurs propres sont supérieures à 1. La première valeur propre vp1 = 1.55 > 1. On choisit donc de retenir le premier axe. De plus, la seconde valeur propre vp2 = 0.90 est proche de 1, on choisit donc de retenir le deuxième axe.
Pour conclure, on choisit de retenir les 2 premiers axes qui conservent environ 82% de l'inertie totale, en effet le premier axe conserve environ 52% de l'inertie du nuage, le second axe conserve une part importante de l'inertie totale qui est de 30%.
La chute est importante dès le troisième axe qui ne conserve plus que 18% de l'inertie totale.
On décide donc de ne retenir que les deux premiers axes.

Voici le nuage de point obtenus :




**Commentaire :**

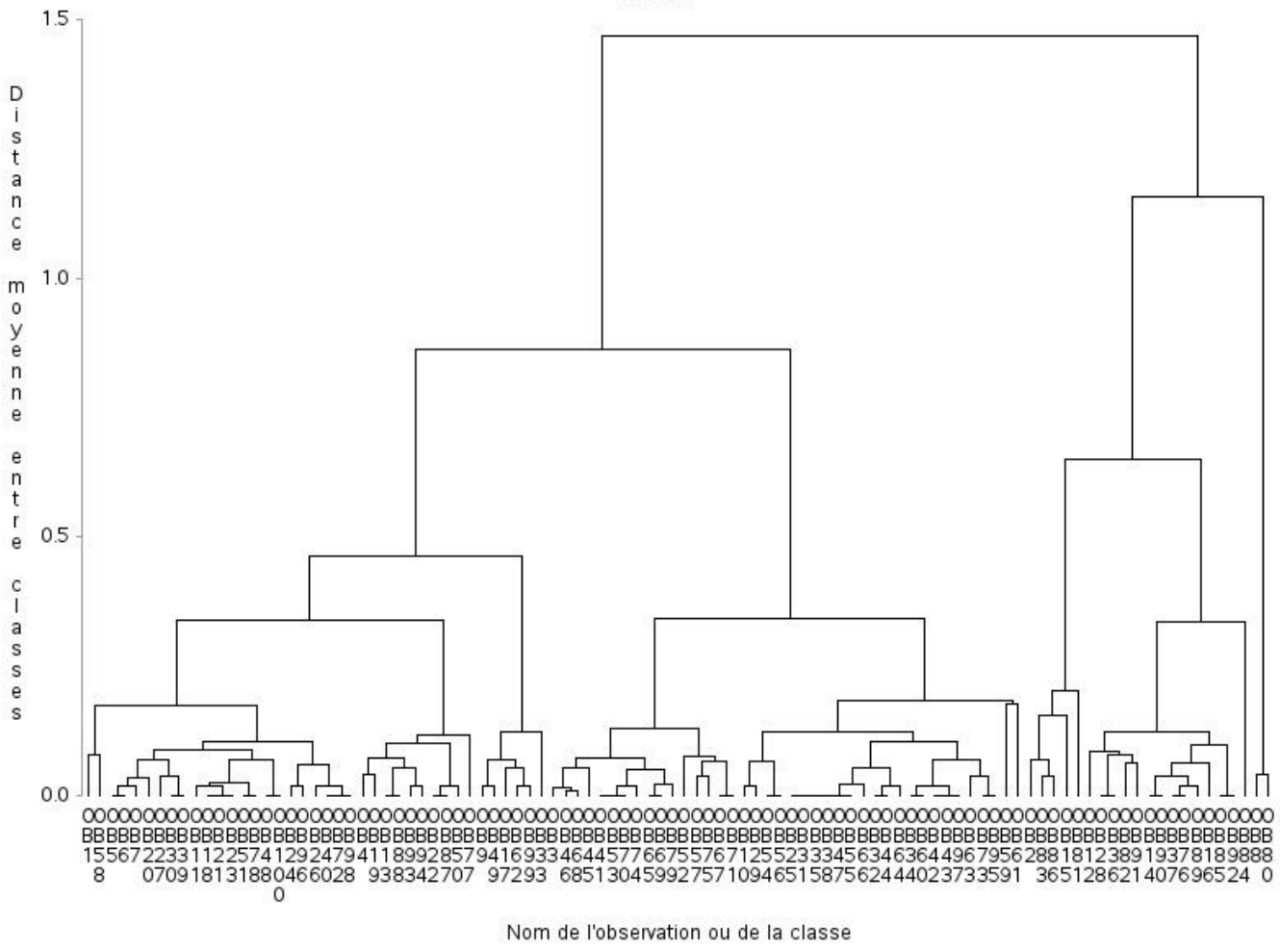
Les valeurs aberrantes, qui sont des valeurs de données très éloignées des autres valeurs de données, peuvent avoir une incidence importante sur nos résultats.
Nous notons ici que l'individu 15 est assez loin des autres, tout comme l'individu 59.
Seulement deux individus sur les 100 sont mal représentés par le premier plan principal et sont très éloignés du point moyen, cela ne justifie pas de rajouter un axe supplémentaire.
On peut également remarquer qu'il semble y avoir trois groupes sur le nuage de points, en effet on voit 3 diagonales assez visibles. On supposera qu'il peut y avoir 3 types de comportement différents chez les individus.

# Classification ascendante hiérarchique

La Classification Ascendante Hiérarchique est une technique statistique visant à mettre une population en différentes classes ou sous-groupes. Elle vise à ce que les individus au sein d'une même classe soient les plus semblables possible tandis que les classes soient le plus dissemblables.
Il sera utile ici de faire une classification ascendante hiérarchique afin de regrouper nos 100 individus en fonction des 10 variables.

Voici l'arbre hiérarchique fait uniquement avec les variables quantitatives :

Code : **proc cluster** data = projet.fertility2 method = average;
var age alcool assis;
proc tree;
**run;**



**Commentaire :**
La classification automatique ou analyse typologique vise à regrouper les individus en paquets homogènes. Les individus qui ont des caractéristiques similaires sont réunis dans un même groupe (cluster, classe); les individus présentant des caractéristiques dissemblables (éloignées) sont associés à des groupes différents.

Le dendrogramme suggère ici un découpage en 4 groupes (en 0.75).
On note qu'une classe se démarque fortement des autres, c'est la classe composé des individus 80 et 8, on peut supposer que ces deux individus ont des comportements extrêmements différents des autres, de plus grâce au nuage de points, on remarque qu'il y a une forte corrélation entre ces deux individus.
On observe également que la classe de droite comporte nettement moins d'individus et se distingue des autres, donc on aurait pu envisager aussi un découpage en 2 groupes seulement.

# Conclusion

A l'aide de l'analyse descriptive du jeu de données, nous avons étudié et analysé nos variables afin de bien les maîtriser. On constate qu'il n'y a pas énormément de variables qui influent sur l'infertilité (4 sur les 10 proposées), on aurait pu ajouter les ondes électromagnétiques, la génétique,  l'obésité, la chaleur (être plus exposé à celle-ci) ou encore savoir si la personne interrogée a eu des antécédents de toxicologie. De plus, la tranche d'âge est assez petite, elle va de 27 ans à 36 ans. Il nous aurait peut être fallu des hommes plus jeunes ou plus âgés pour faire une étude plus approfondie et pouvoir vraiment conclure sur l'influence de l'âge sur l'infertilité. (a t-on plus de chance d'être stérile lorsque l'on est jeune ou plus âgé ?)
De même savoir depuis combien de temps la personne boit de l'alcool par exemple.

Nos suppositions du début n'étaient pas totalement fausses par rapport aux variables importantes ou non.
La variable **"winter"** nous surprend quant à ces résultats car on remarque que cette variable joue un rôle important quant au diagnostic final. Nous pensons qu'avec un plus large panelle d'individus (au moins 10 fois plus) cette variable ne serait pas forcément ressortis de manière aussi importante pour le diagnostic. Au contraire pour la variable **"assis"** nous avions fait un bon jugement (cf. p.4). En effet la variable "winter" est assez compliqué, car cela nous paraît étrange d'affirmer que si les tests ont été fait en hiver, la personne a plus de chance d'être stérile. On pourrait alors aussi dire que lorsque les tests sont fait en période hivernale, une majorité de personnes est diagnostiquée infertile, comparées aux autres saisons, le climat aurait donc un rôle à jouer sur la stérilité.

Au vu de nos variables, nous n'avons pas pu faire de régression linéaire. Nous avons donc décidé de nous orienter vers une régression logistique. Nous avons rencontré quelques difficultés sur la régression logistique simple tandis que la régression logistique multiple nous a permis de confirmer nos suppositions.
En utilisant celle-ci, nous avons remarqué que les individus ayant eu un accident ou un traumatisme grave, de la fièvre et qui boivent de l'alcool sont plus enclin à être infertile.

La régression logistique simple nous a finalement permis de vérifier ces informations et de connaître la probabilité d'augmentation de l'infertilité pour ces quatre variables.
Si nous devions écrire un classement concernant les causes de la stérilité de la plus importante à la moins importante, nous constatons que l'abus d'alcool est le facteur le plus dangereux causant la stérilité, suivi du facteur saison "winter" pour la période de test, le fait d'avoir eu un accident ou un traumatisme grave et pour terminer avoir eu de la fièvre, contrairement au fait de rester assis, avoir fait de la chirurgie ou encore avoir eu des maladies dans son enfance.

Ensuite,  nous avons réalisé une ACP puis une CAH, celle-ci nous a permis de regrouper nos individus dans différentes classes. On distingue des classes plutôt distinctes (cf. nuage de points p.20).
Notre panelle d'individus est donc assez varié ce qui est plutôt pratique afin de réaliser une étude statistique.
On voit cependant quelques individus éloignés des autres, ce qui peut perturber cette étude.
Pour finir un élément important à prendre en compte pour nuancer ces résultats est le faible échantillon d'individus qui est seulement de 100. Il est recommandé de prendre 10 à 20 fois plus d'observations que de variables.