PROJET DE STATISTIQUE

Tests d’hypothèses

(2021 – 2022)

I. Introduction

Description de la population : L'étude que nous réalisons porte sur une population d'hommes venus consulter pour infécondité masculine au Centre de Stérilité Masculine (CSM) à Toulouse entre 2000 et 2004. Notre fichier de données comporte 1000 individus. Tous ces patients ont réalisés un spermogramme au cours de leur première consultation. Cela consiste en l'analyse du sperme en laboratoire. En particulier, le spermogramme mesure avec prévision un certain nombre de paramètres comme le volume de l'éjaculat, la concentration en spermatozoïdes, ou la vitalité des spermatozoïdes.

Description des variables : Notre étude porte, elle, sur l'analyse d'un certain nombre de variables, 11 dans notre base de données. Tout d'abord, nous avons la variable « ***alcool*** » qui prend la valeur 1 si le patient boit plus d'un litre d'alcool par jour et 0 sinon. Ensuite, nous avons la variable « **tabagisme** » qui prend la valeur 1 si le patient fume, 0 sinon. La variable suivante est la variable « ***vol*** » qui est le volume de l'éjaculat (en ml). La variable « ***conc*** » est la concentration en spermatozoïdes de l'éjaculat (en millions par ml). La variable « ***vit*** » représente la vitalité, c'est-à-dire le pourcentage de spermatozoïdes vivants dans l'éjaculat. La variable « ***exam*** » compte le nombre d'examens anormaux à la consultation. La variable « ***age*** » est l'âge du patient lors de la consultation (en années). La variable « ***dureeinf*** » représente la durée d'infertilité, c'est-à-dire le nombre de mois pendant lesquels le couple a essayé d'avoir un enfant sans succès. La variable « ***imc*** » est l'indice de masse corporelle du patient, indice définit par l'Organisation Mondiale de la Santé comme le standard pour évaluer les risques liés au surpoids. Il se calcule en faisant le rapport du poids (en kg) sur la taille (en mètres) au carré. La variable « ***mob*** » indique la mobilité, c'est-à-dire le pourcentage de spermatozoïdes mobiles dans l'éjaculat. Enfin, la variable « ***fievre*** » prend la valeur 1 si le patient a souffert d'une forte fièvre moins de trois mois avant la consultation, 0 sinon.

But de l'étude : Ainsi, nous nous intéressons à deux principaux types de variables : certaines portent sur les caractéristiques de l'éjaculat (*vol, conc, vit, mob*), d'autres sur les caractéristiques du patient (*alcool, tabagisme, exam, age, dureeinf, imc, fievre*). Le but de notre étude est en effet d'étudier les potentielles liaisons qui peuvent exister entre ces différentes variables, et notamment d'étudier l'impact des conditions de vie (représentées par les variables portant sur les caractéristiques du patient) sur les éléments du spermogramme (représentés par les variables portant sur les caractéristiques de l'éjaculat). En effet, les éléments du spermogramme caractérisent la fécondité des hommes et donc leur capacité à avoir des enfants. Si des études ont déjà pu mettre en évidence l'impact de la corpulence sur la fertilité des hommes, nous souhaitons ici voir si nous pouvons retrouver cet impact au sein de notre population et nous cherchons à étudier d'autres potentielles liaisons entre nos variables.

II. Création de variables

2) a) Nous créons la nouvelle variable « *alcool2* » avec la fonction « factor ». Nous reprenons la variable « *alcool* » en remplaçant le 0 par « non » et le 1 par « oui ». De la même manière, nous créons également la variable « *tabagisme2* » avec la fonction « factor », en reprenant la variable « *tabagisme* » et en remplaçant le 0 par « non » et le 1 par « oui ».

2) b) Nous créons ensuite la variable « *num* » en multipliant la variable « *vol* » pour le volume de l'éjaculat (en ml) par la variable « *conc* » pour la concentration en spermatozoïdes de l'éjaculat (en millions par ml). Nous obtenons ainsi le nombre de spermatozoïdes présents dans l'éjaculat (en millions).

2) c) Nous créons ensuite la variable « *imcc* » qui correspond à l'IMC en classes. Pour cela, nous créons trois classes : la classe « maigre ou normal » (puisque les individus maigres sont trop peu nombreux dans notre échantillon) qui correspond aux IMCs allant jusqu'à 25 exclu (d'après les intervalles définis par l'OMC) ; la classe « surpoids », correspondant aux IMCs de 25 à 30 exclu ; puis la classe « obésité » pour les IMCs supérieurs ou égaux à 30. Nous utilisons pour cela la fonction « cut » pour laquelle nous indiquons les délimitations des classes afin de créer ces trois classes : entre 11 et 24, entre 25 et 29 et entre 30 et 54. Les choix des bornes minimales et maximales ont été guidés par le minimum et le maximum des valeurs de la variable *imc*. Les bornes inférieures des deux dernières « coupures » ne sont pas comprises, ainsi les classes sorties par le logiciel correspondent bien aux classes énoncées ci-dessus.

2) d) Enfin, nous créons la variable « *examc* ». Pour cela, nous créons tout d'abord deux classes afin de séparer les données entre les patients qui n'ont aucun examen anormal et ceux qui ont un ou plusieurs examens anormaux. Nous utilisons comme dans la question précédente la fonction « cut ». Ensuite, nous utilisons la fonction « factor » afin d'associer les classes aux intitulés « aucun examen anormal » et « au moins un examen anormal ».

III. Étude descriptive

• Etude descriptive des variables :

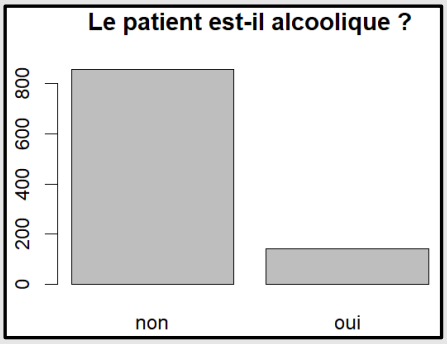
• ***alcool2*** : C'est une variable **qualitative** à 2 modalités (oui/non).

Par la fonction *summary*, nous trouvons :

```
> summary(alcool2)
non oui
859 141
```

Parmi les 1000 individus observés, 141 sont alcooliques et 859 ne le sont pas.

Graphiquement, nous trouvons :



Cela nous donnes les mêmes informations.

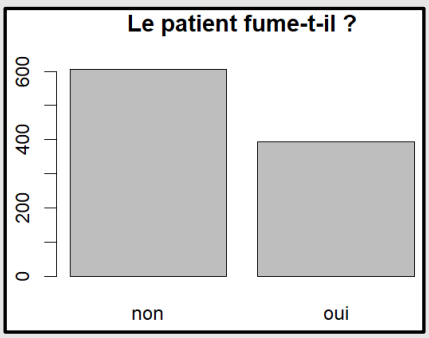
• ***tabagisme2*** : C'est une variable **qualitative** à 2 modalités (oui/non).

Par la fonction *summary*, nous trouvons :

```
> summary(tabagisme2)
non oui
607 393
```

Parmi les 1000 individus observés, 393 sont fumeurs et 607 ne le sont pas.

Graphiquement, nous trouvons :



Cela nous donne les mêmes informations.

• ***fievre*** : C'est une variable **qualitative** à 2 modalités (1 si le patient a eu une forte fièvre moins de trois mois avant la consultation / 0 sinon).

Par la fonction *table*, nous trouvons :

```
> table(fievre)
fievre
  0   1
981  19
```

Parmi les 1000 individus observés, 19 ont eu une fièvre récente et 981 n'en ont pas eu.

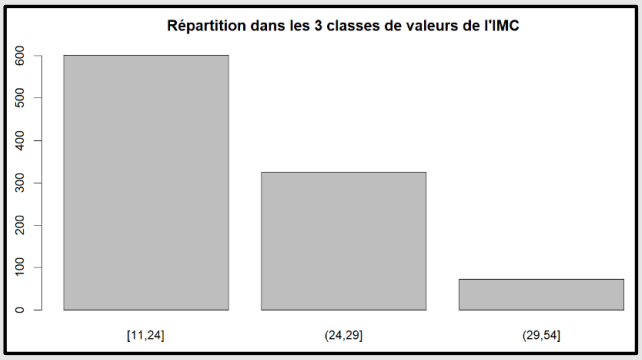
• ***imcc*** : C'est une variable **qualitative** à 3 modalités (correspondant aux 3 classes « maigre ou normal », « surpoids », « obésité »).

Par la fonction *summary*, nous trouvons :

```
> summary(imcc)
[11,24] (24,29] (29,54]
    602     325      73
```

Parmi les 1000 individus observés, 602 ont une corpulence maigre ou normale, 325 sont en surpoids et 73 sont obèses.

Graphiquement, nous trouvons :



Cela nous donne les mêmes informations.

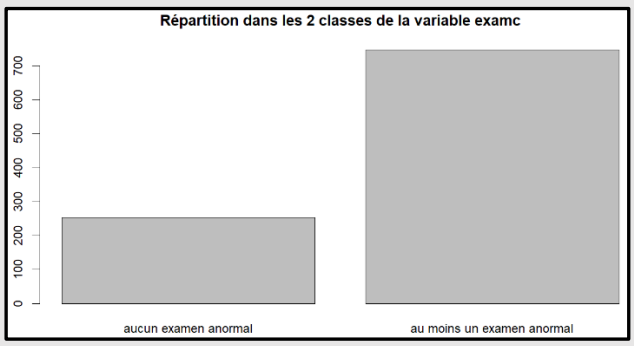
• ***examc*** : C'est une variable **qualitative** à 2 modalités (correspondant aux 2 classes « aucun examen anormal », « au moins un examen anormal »).

Par la fonction *summary*, nous trouvons :

```
> summary(examc)
     aucun examen anormal au moins un examen anormal
                      253                       747
```

Parmi les 1000 individus observés, 253 n'ont eu aucun examen anormal et 747 en ont eu au moins un.

Graphiquement, nous trouvons :



Cela nous donne les mêmes informations.

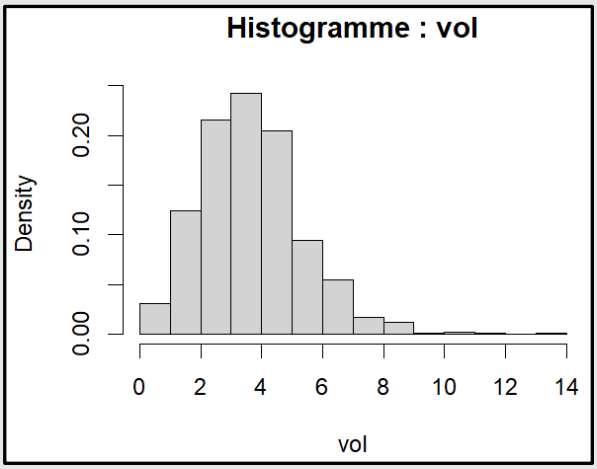
• ***vol*** : C'est une variable **quantitative** continue (nombres à virgules).

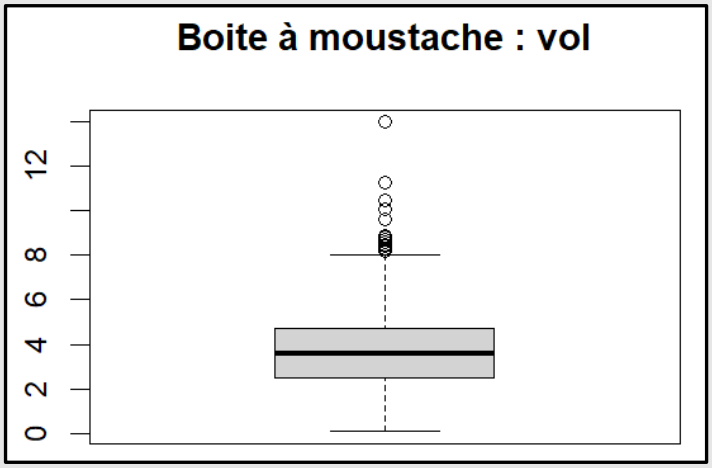
Par la fonction *summary*, nous trouvons :

```
> summary(vol)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.100   2.500   3.600   3.723   4.700  14.000
```

Pour les 1000 individus observés, le volume de l'éjaculat moyen est de 3,723 ml. La valeur minimale est de 0,1 ml et la valeur maximale de 14 ml. La médiane est à 3,6 ml, ce qui signifie que, par exemple, 50 % des individus observés ont un volume inférieur à 3,6 ml. De même pour les quartiles : par exemple, 25 % des individus observés ont un volume inférieur à 2,5 ml (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *vol*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut. Nous pouvons visualiser facilement le 1er quartile, la médiane, et le 3ème quartile pour chaque variable. Dans le carré gris, il y a 50 % des individus et on peut aussi voir le minimum et le maximum : max = Q3 + 1,5(Q3-Q1) et min = Q1 –1,5(Q3-Q1). Les points isolés au-dessus ou en-dessous des extremums sont des valeurs extrêmes.

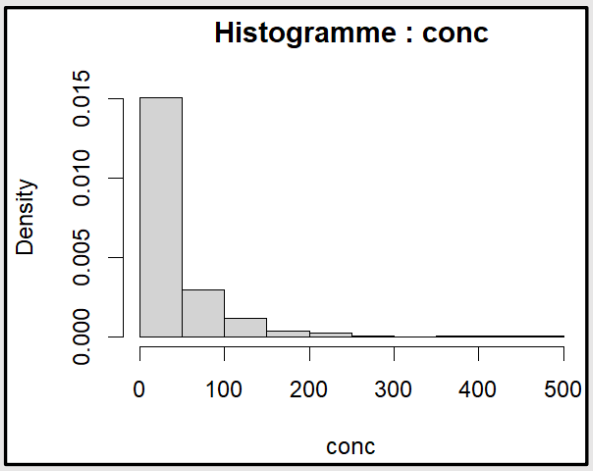
• ***conc*** : C'est une variable **quantitative** continue. (nombre à virgules).

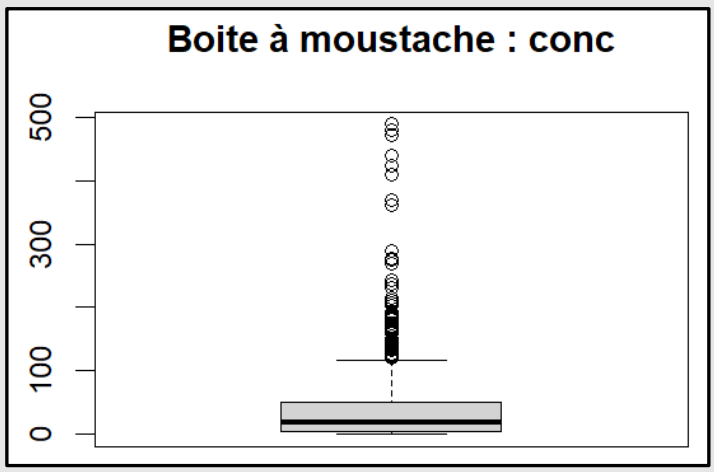
Par la fonction *summary*, nous trouvons :

```
> summary(conc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.010   4.575  17.000  38.381  50.000 490.000
```

Pour les 1000 individus observés, la concentration en spermatozoïdes de l'éjaculat moyenne est de 38,381 millions de spermatozoïdes par ml. La valeur minimale est de 0,01 millions de spermatozoïdes par ml et la valeur maximale de 490 millions de spermatozoïdes par ml. La médiane est à 17 millions de spermatozoïdes par ml, ce qui signifie que, par exemple, 50 % des individus observés ont une concentration inférieure à 17 millions de spermatozoïdes par ml. De même pour les quartiles : par exemple, 25 % des individus observés ont une concentration inférieure à 4,575 millions de spermatozoïdes par ml (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *conc*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut. Notons que de très nombreuses « valeurs aberrantes » sont visibles sur la boite à moustache.

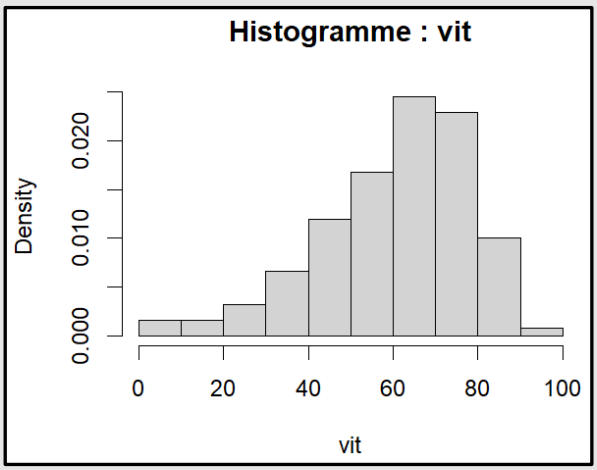
• ***vit*** : C'est une variable **quantitative** continue.

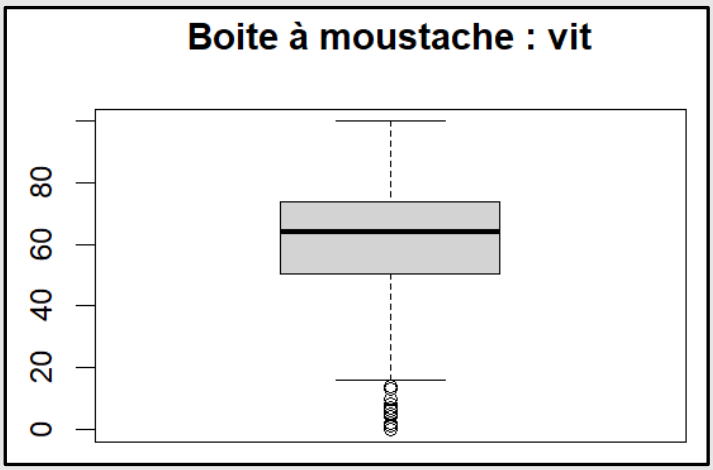
Par la fonction *summary*, nous trouvons :

```
> summary(vit)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   50.75   64.00   61.11   74.00  100.00
```

Pour les 1000 individus observés, la vitalité moyenne est de 61,11% (ce qui signifie que 61,11% des spermatozoïdes sont vivants dans l'éjaculat). La valeur minimale est de 0% et la valeur maximale de 100%. La médiane est à 64%, ce qui signifie que, par exemple, 50 % des individus observés ont une vitalité inférieure à 64%. De même pour les quartiles : par exemple, 25 % des individus observés ont une vitalité inférieure à 50,75% (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *vit*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut.

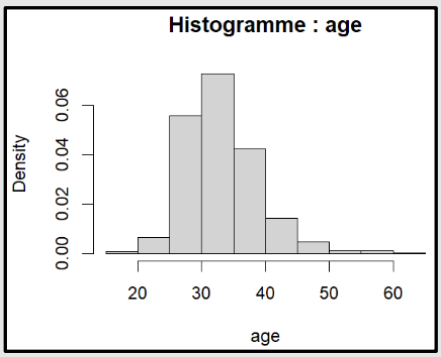
• ***age*** : C'est une variable **quantitative** discrète.

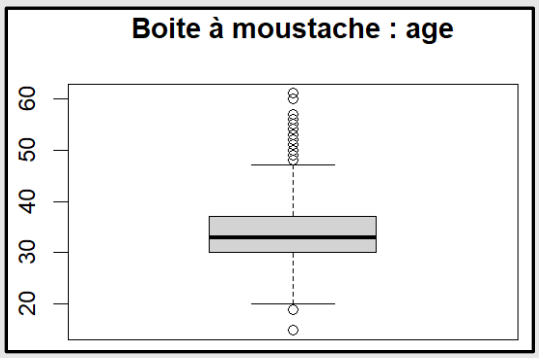
Par la fonction *summary*, nous trouvons :

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.00   30.00   33.00   33.63   37.00   61.00
```

Pour les 1000 individus observés, l'âge moyen est de 33,63 ans. La valeur minimale est de 15 ans et la valeur maximale de 61 ans. La médiane est à 33 ans, ce qui signifie que, par exemple, 50 % des individus observés ont un âge inférieur à 33 ans. De même pour les quartiles : par exemple, 25 % des individus observés ont un âge inférieur à 30 ans (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *age*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut.

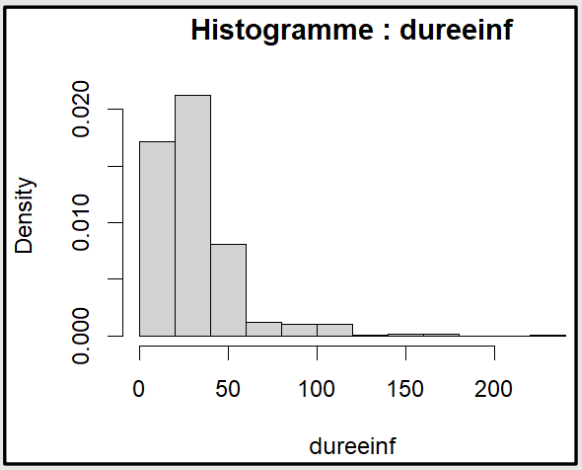
• ***dureeinf*** : C'est une variable **quantitative** discrète.

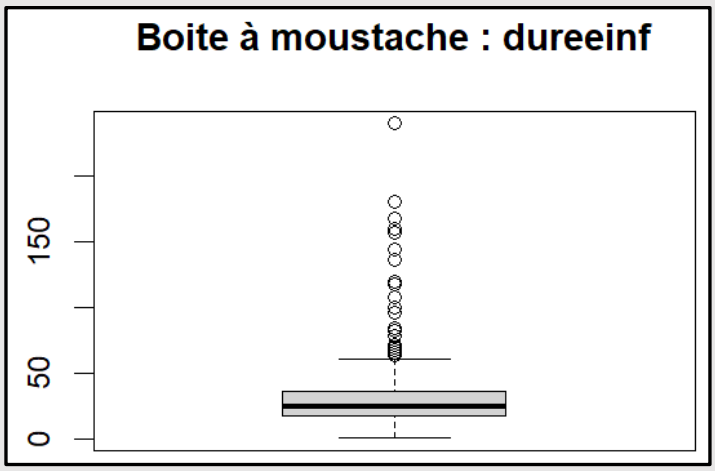
Par la fonction *summary*, nous trouvons :

```
summary(dureeinf)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   18.00   24.00   32.73   36.00  240.00
```

Pour les 1000 individus observés, la durée d'infertilité moyenne est de 32,73 mois. La valeur minimale est de 1 mois et la valeur maximale de 240 mois. La médiane est à 24 mois, ce qui signifie que, par exemple, 50 % des individus observés ont une durée d'infertilité inférieure à 24 mois. De même pour les quartiles : par exemple, 25 % des individus observés ont une durée d'infertilité inférieure à 18 mois (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *dureeinf*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut. Notons que de nombreuses « valeurs aberrantes » sont visibles sur la boite à moustache.

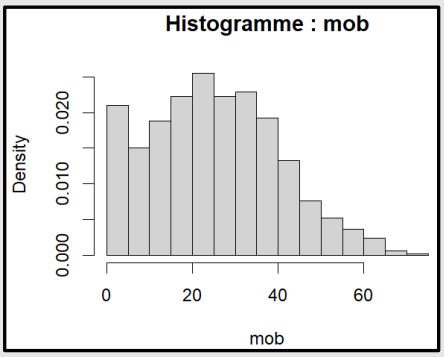
• ***mob*** : C'est une variable **quantitative** continue.

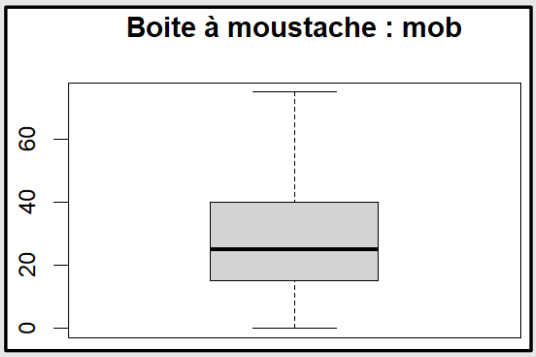
Par la fonction *summary*, nous trouvons :

```
> summary(mob)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   15.00   25.00   27.54   40.00   75.00
```

Pour les 1000 individus observés, la mobilité moyenne est de 27,54% (ce qui signifie que 27,54% des spermatozoïdes étaient mobiles dans l'éjaculat). La valeur minimale est de 0% et la valeur maximale de 75%. La médiane est à 25%, ce qui signifie que, par exemple, 50 % des individus observés ont une mobilité inférieure à 25%. De même pour les quartiles : par exemple, 25 % des individus observés ont une mobilité inférieure à 15% (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *mob*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut. Notons qu'il n'y a pas de « valeurs aberrantes » visibles.

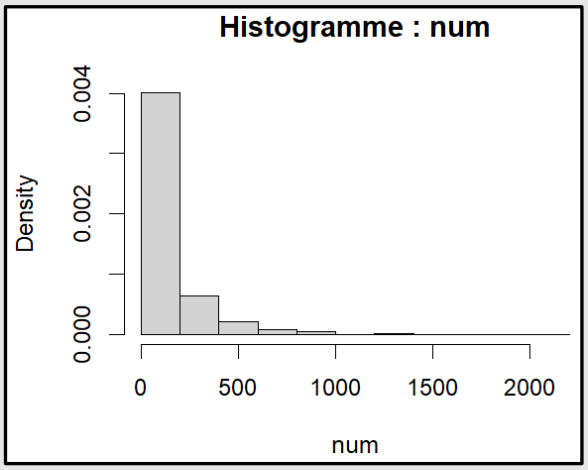
• ***num*** : C'est une variable **quantitative** continue (car exprimée en millions donc c'est un nombre à virgule).

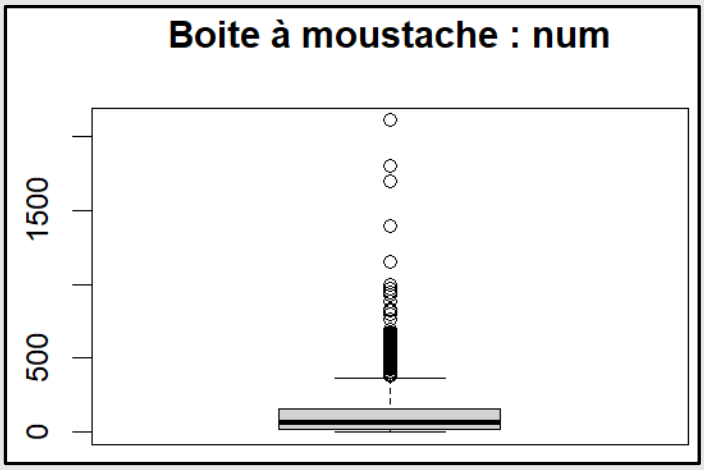
Par la fonction *summary*, nous trouvons :

```
> summary(num)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.016  14.963  59.155 127.052 159.135 2112.000
```

Pour les 1000 individus observés, la numération moyenne est de 127,052 millions. La valeur minimale est de 0,016 millions et la valeur maximale de 2112 millions. La médiane est à 59,155 millions, ce qui signifie que, par exemple, 50 % des individus observés ont une numération inférieure à 59,155 millions. De même pour les quartiles : par exemple, 25 % des individus observés ont une numération inférieure à 14,963 millions (= 1er quartile).

Par les fonctions *hist* et *boxplot*, nous trouvons :





L'histogramme à gauche représente la distribution de la variable *num*. La boite à moustache à droite permet de synthétiser l'analyse du summary effectuée en haut. Notons que de très nombreuses « valeurs aberrantes » sont visibles sur la boite à moustache.

• Test de la normalité de la variable *vit* :

Nous testons ensuite la normalité de la variable « *vit* » qui correspond à la vitalité, c'est-à-dire le pourcentage de spermatozoïdes vivants dans l'éjaculat. Pour cela, nous réalisons un **test d'adéquation de Kolmogorov-Smirnov** avec la fonction « ks.test » : il permet de tester l'adéquation à une loi donnée à partir d'un échantillon de données individuelles continues.

Nous posons :

H_0 : "***La loi dont proviennent les données a comme fonction de répartition celle d'une loi normale***".

H_1 : "***La loi dont proviennent les données n'a pas comme fonction de répartition celle d'une loi normale***".

En effectuant le test, nous obtenons **p-value = 1.308e-07**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Nous concluons que pour la variable « *vit* », les données ne proviennent pas d'une loi gaussienne.**

IV. Comparaison de la population d'où est extrait l'échantillon avec la population française en 2002

4) a) Afin de pouvoir observer si les couples de notre échantillon ont suivi le conseil d'attendre deux ans d'essais infructueux avant de consulter pour infertilité, nous réalisons tout d'abord un résumé de la variable *dureeinf* qui correspond à la durée d'infertilité du couple (en mois) :

```
summary(dureeinf)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   18.00   24.00   32.73   36.00  240.00
```

Grâce à cela, nous observons que la **moyenne** de durée d'infertilité est de 32,73 mois, ce qui est supérieur aux 24 mois qui constituent deux ans. **On peut donc dire qu'en moyenne les couples de notre échantillon ont respecté le conseil**. Néanmoins, si l'on observe la **médiane** de la durée d'infertilité, on peut observer qu'elle est de 24 mois, autrement dit de 2 ans. **Cela signifie donc que 50% des couples de notre échantillon n'ont pas respecté le conseil.**

4) b) Nous devons ici tester **l'adéquation des observations** (répartition empirique de l'IMC en classes) avec une **distribution théorique donnée** (répartition française de l'IMC en classes).

En étudiant *imcc*, nous obtenons la **répartition empirique** de la population observée au sein des 3 classes de la variables :

```
imcc
[11,24] (24,29] (29,54]
    602     325      73
```

Par exemple, on peut voir qu'il y a 602 individus de notre échantillons qui ont un ICM entre 11 et 24, autrement dit de corpulence maigre ou normale.

La **répartition théorique** nous est donnée par l'énoncé.

Comme nous disposons d'observations groupées **en classes, nous allons effectuer un test d'adéquation du X^2**. Vérifions d'abord les conditions de validité. En calculant les effectifs attendus nous obtenons :

```
584 292 124
```

Tous les **effectifs attendus** sont supérieurs à 5, donc on peut faire le **test qui est valide**.

Nous posons :

$$H_0 : \textit{La population étudiée présente la même répartition de l'IMC en classes.}$$

$$H_1 : \textit{Les répartitions de l'IMC en classe observé et de celui français sont différentes.}$$

En effectuant le test, nous obtenons **p-value = 3.272e-06**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Nous concluons donc que la population que nous étudions ne présente pas la même répartition de l'IMC en classe que ce que l'on peut voir dans la population française en 2002. En effet, dans notre population observée, nous avons notamment beaucoup moins d'obèses que ce que l'on peut voir dans la population française en 2002 (7,3 % dans notre échantillon VS. 12,4 %).**

4) c) Nous devons ici tester **l'adéquation des observations** (répartition empirique des fumeurs ou non-fumeurs dans notre échantillon) avec une **distribution théorique donnée** (nous connaissons la proportion

de fumeurs en France en 2002). Nous devons en réalité tester l'hypothèse que la variable *tabagisme2* suit une loi de Bernoulli de paramètre 0,36 (il y a deux issues : =1 (i.e. fumeur) en proportion 0,36 et =0 (i.e. non-fumeur) en proportion 0,64). Plus simplement, nous allons tester si p (la proportion empirique de fumeurs) est égale à 0,36 (la proportion théorique de fumeurs).

En étudiant *tabagisme2*, nous obtenons la **répartition empirique** de la population observée au sein des 2 modalités de la variables :

On peut voir qu'il y a 393 individus de notre échantillons qui fument.

Comme nous sommes dans le cas particulier où k=2**, nous allons effectuer un test d'adéquation du X^2**.

Nous posons :

$$\boldsymbol{H_0 : p = 0,36.}$$

$$\boldsymbol{H_1 : p \neq 0,36.}$$

Vérifions d'abord les conditions de validité. En calculant les effectifs attendus nous obtenons :

Les 2 **effectifs attendus** sont supérieurs à 5, donc on peut faire le **test qui est valide**.

En effectuant le test, nous obtenons **p-value = 0.0297**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Nous concluons donc que la population que nous étudions ne fume pas autant que la population française en 2002. Notre population observée fume plus (39,3% VS. 36%).**

4) d) Ici, nous voulons tester **l'égalité d'une moyenne** (de la variable *num* ; nous la notons μ) **à une valeur donnée** (250 millions de spermatozoïdes en moyenne dans un éjaculat).

Nous posons :

$$\boldsymbol{H_0 : \mu = 250.}$$

$$\boldsymbol{H_1 : \mu \neq 250\ (test\ bilatéral).}$$

(Nous aurions pu tester avec $H_1 : \mu < 250$ car la question nous pousse à faire un test unilatéral ; néanmoins, le sujet nous demande de faire des tests bilatéraux).

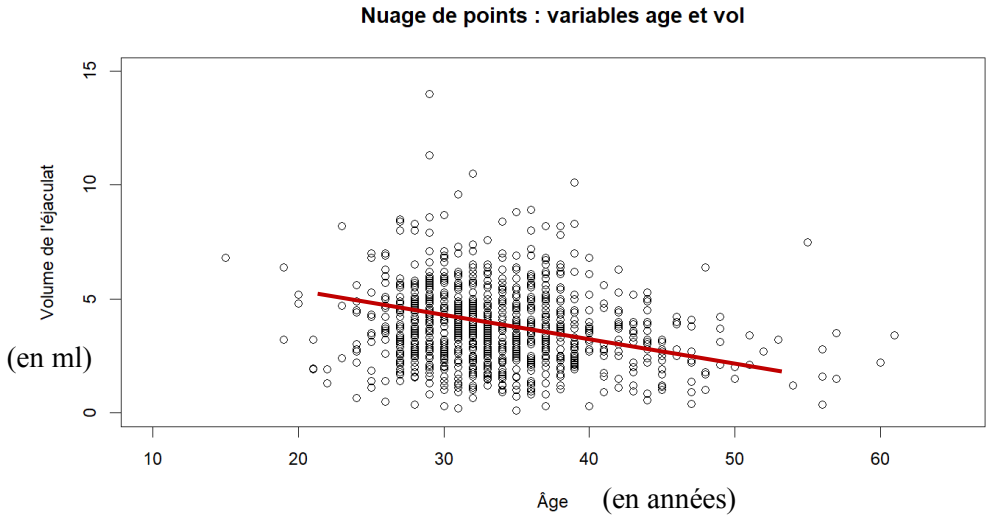
Comme n est grand (1000 > 30), on peut se passer de l'hypothèse gaussienne car la moyenne empirique de la variable peut être considérée comme gaussienne quand n est grand d'après le théorème central limite (TCL). On peut donc appliquer le **test de Student de comparaison d'une espérance à une valeur donnée** dans ce cas.

En effectuant le test, nous obtenons **p-value < 2.2e-16**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est très significatif. **Nous concluons que les individus qui consultent pour infertilité au CSM ont une numération différente (test bilatéral) de la moyenne de 250 millions** (et même très largement inférieure dans le cas unilatéral : la moyenne empirique est de 127.052 millions).

V. Influence de l'âge sur les autres variables

5) a) L'âge (en années) est une variable quantitative, tout comme le volume de l'éjaculat (en ml). Nous voulons tester la liaison entre ces **deux variables quantitatives**. Nous allons alors utiliser le **test du coefficient de corrélation linéaire de Pearson**.

En préalable au test de la liaison, on trace le **nuage de points** des couples des réalisations de *age* et *vol*. Si le nuage de points montre une tendance linéaire, le coefficient de corrélation linéaire est adapté pour mesurer la liaison entre les deux variables.



On peut observer une (légère) **liaison linéaire négative** entre l'âge et le volume de l'éjaculat. **Donc le test que nous allons effectuer est adapté.**

Par ailleurs, nous pouvons calculer que le coefficient de corrélation linéaire est **- 0.1671656**, qui est négatif, et nous savons que n = 1000 > 30.

Nous posons :

$$\boldsymbol{H_0 : \rho(age, vol) = 0 \ (absence\ de\ liaison\ linéaire\ entre\ les\ 2\ variables)}.$$

$$\boldsymbol{H_1 : \rho(age, vol) \neq 0 \ (test\ bilatéral)}.$$

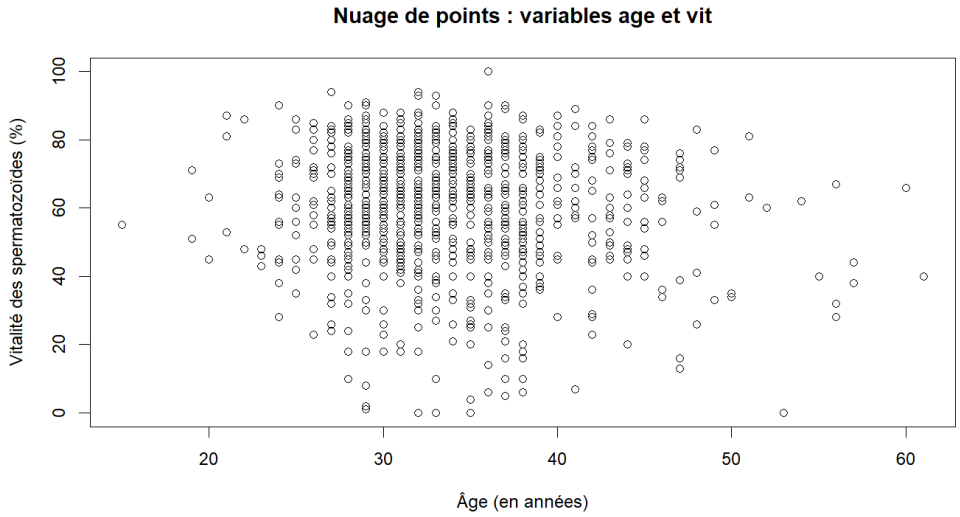
(Nous aurions pu faire un test unilatéral avec $H_1 : \rho(age, vol) < 0$ *car le sujet nous laissait penser à une liaison négative, mais le sujet nous demande de faire des tests bilatéraux).*

En effectuant le test par la commande *cor.test(age,vol)*, nous obtenons **p-value = 1.054e-07**. Cette p-value est **< 0.05**. Donc **nous rejetons H₀**. Le test est très significatif (p-valeur très inférieure à 5%). Nous concluons qu'il y a une **liaison linéaire (négative)** entre l'âge et le volume de l'éjaculat. **On peut observer que plus l'âge est élevé, plus le volume de l'éjaculat diminue.**

5) b) La vitalité des spermatozoïdes et la mobilité des spermatozoïdes sont des **variables quantitatives** (en pourcentages). Nous voulons tester la liaison entre **deux variables quantitatives** (âge et vitalité ; âge et mobilité). Nous allons alors utiliser le **test du coefficient de corrélation linéaire de Pearson**.

- **<u>Pour la vitalité des spermatozoïdes :</u>**

En préalable au test de la liaison, on trace le **nuage de points** des couples des réalisations de *age* et *vit*. Si le nuage de points montre une tendance linéaire, le coefficient de corrélation linéaire est adapté pour mesurer la liaison entre les deux variables.



Nous n'observons pas vraiment de forme linéaire distincte.

Le calcul du coefficient de corrélation linéaire donne **- 0.09917721**, qui est négatif, et nous savons que n = 1000 > 30.

Nous posons :

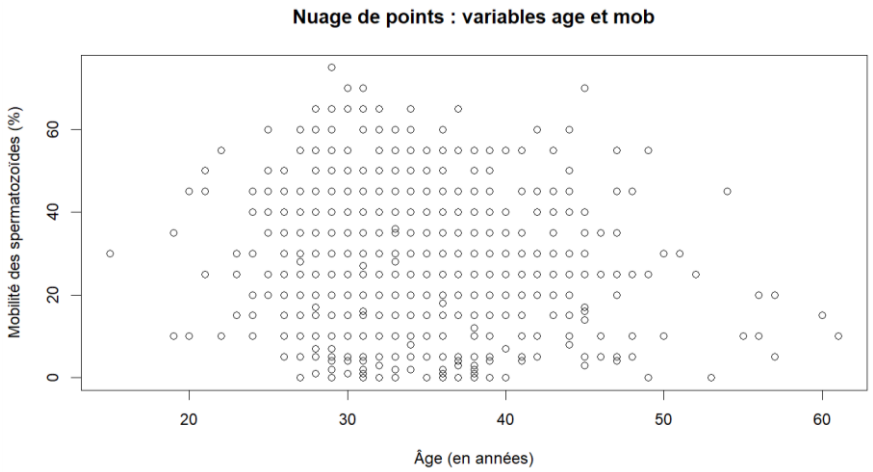
$$\boldsymbol{H_0 : \rho(age, vit) = 0 \ (absence\ de\ liaison\ linéaire\ entre\ les\ 2\ variables)}.$$

$$\boldsymbol{H_1 : \rho(age, vit) \neq 0 \ (test\ bilatéral)}.$$

En effectuant le test par la commande *cor.test(age,vit)*, nous obtenons **p-value = 0.001689**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. Nous concluons qu'il y a une **liaison linéaire** entre l'âge et la vitalité des spermatozoïdes. **La liaison est négative (légèrement) : plus l'âge est élevé, plus la vitalité des spermatozoïdes diminue.**

- **<u>Pour la mobilité des spermatozoïdes :</u>**

En préalable au test de la liaison, on trace le **nuage de points** des couples des réalisations de *age* et *mob*. Si le nuage de points montre une tendance linéaire, le coefficient de corrélation linéaire est adapté pour mesurer la liaison entre les deux variables.



Nous n'observons pas vraiment de forme linéaire distincte.

Le calcul du coefficient de corrélation linéaire donne **- 0.05669931**, qui est négatif ; et nous savons que n = 1000 > 30.

Nous posons :

$$\boldsymbol{H_0 : \rho(age, mob) = 0 \ (absence\ de\ liaison\ linéaire\ entre\ les\ 2\ variables)}.$$

$$\boldsymbol{H_1 : \rho(age, mob) \neq 0 \ (test\ bilatéral)}.$$

En effectuant le test par la commande *cor.test(age,mob)*, nous obtenons **p-value = 0.07354**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence l'existence d'une liaison linéaire entre l'âge et la mobilité des spermatozoïdes.**

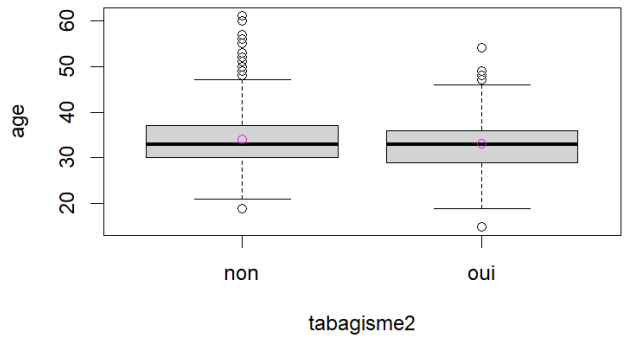
5) c) L'âge est une variable **quantitative**. Le comportement tabagique est une **variable qualitative à 2 modalités**.

Calculons d'abord les moyennes conditionnelles des âges selon le statut tabagique. Nous trouvons :

- les non-fumeurs ont en moyenne 34 ans.

- les fumeurs ont en moyenne 33,1 ans.

Effectuons les boites à moustache :



On voit sur ce graphique que les dispersions des âges sont relativement similaires, tous comme les médianes.

Nous n'avons pas l'hypothèse gaussienne pour l'âge. Néanmoins, comme on a bien des effectifs assez grands (> 30 ; 607 et 393 respectivement), on ne peut pas utiliser le test de comparaison des variances de Fisher, mais on va donc faire le **test de Welch**.

Posons :

$$\boldsymbol{H_0 : \mu_1 = \mu_2}.$$

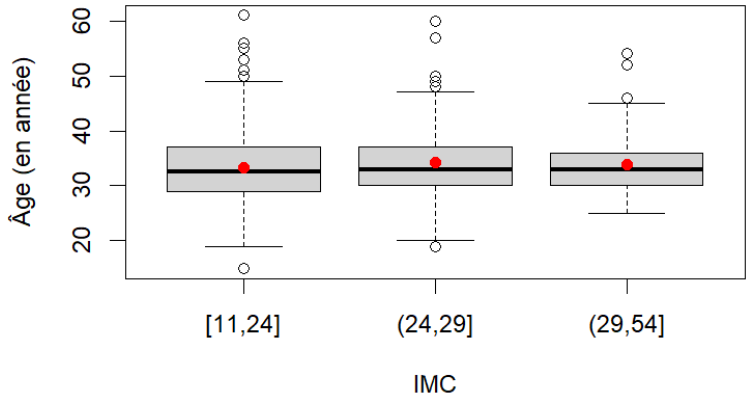
$$\boldsymbol{H_1 : \mu_1 \neq \mu_2}.$$

(Où μ_1 est l'espérance de l'âge des non-fumeurs ; μ_2 est l'espérance de l'âge des fumeurs).

En effectuant le test de Welch, nous obtenons **p-value = 0.01822**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Ainsi, l'âge moyen des non-fumeurs est différent de l'âge moyen des fumeurs. Il y a donc une liaison entre âge et comportement tabagique : les moyennes conditionnelles plus haut laissent à penser que les jeunes fument plus que les individus plus âgés.**

5) d) L'âge est une **variable quantitative**. La corpulence (avec l'IMC en classes) est une **variable qualitative avec 3 modalités**. Commençons par faire des boites à moustaches pour se donner une idée de

la liaison potentielle entre les 2 variables. Si l'on calcule les **moyennes conditionnelles** des âges sachant les classes d'IMC, on obtient tout d'abord : 33,3 34,3 33,9. Les **boites à moustaches** donnent :



Les médianes sont équivalentes entre les 3 groupes. Nous ne pouvons pas mettre en évidence de liaison claire entre l'âge et l'IMC en classes avec ces boites à moustaches.

Nous allons alors effectuer un **test paramétrique de comparaison de k moyennes (analyse de la variance à un facteur – ANOVA)**. Tout d'abord, nous devons vérifier les conditions de validité :

- **<u>Normalité des échantillons</u>** : Calculons les effectifs par groupes :

```
imcc
[11,24] (24,29] (29,54]
    602     325      73
```

Comme tous les effectifs par groupes sont suffisamment grands (> 30), alors on peut se passer de l'hypothèse de normalité.

- **<u>Homogénéité des variances</u>** : Posons **H_0 : « Les variances sont égales pour chaque groupe »**. Nous effectuons le **test de Brown et Forsythe**. Nous obtenons une **p-value = 0.2003**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas l'hypothèse nulle**. **Nous concluons donc que nos données satisfont l'hypothèse d'homogénéité des variances.**

Dès lors, comme nous n'avons pas rejeté l'homoscédasticité, nous pouvons effectuer le test ANOVA.

Nous posons :

$$H_0 : \mu_1 = \mu_2 = \mu_3.$$

$$H_1 : \exists i, j \; tels \; que \; \mu_i \neq \mu_j.$$

En effectuant le test, nous obtenons **p-value = 0.04287**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Au moins deux des classes d'IMC ont des âges moyens différents.**

(On peut aussi regrouper les résultats dans le tableau d'analyse de la variance suivant :)

```
                  Df Sum Sq Mean Sq F value Pr(>F)
as.factor(imcc)    2    212  105.88   3.159 0.0429
Residuals        997  33410   33.51
```

Il faut alors faire des tests de comparaisons multiples après le test ANOVA significatif. Ce sont des tests de comparaison des moyennes deux à deux pour pouvoir préciser entre quels groupes se situent les différences.

Tout d'abord, nous effectuons un **test LSD**, qui compare deux à deux les moyennes. Nous posons :

$$H_0^{ij} : \mu_i = \mu_j.$$

$$H_1^{ij} : \mu_i \neq \mu_j.$$

En effectuant le test, on obtient :

	[11,24]	(24,29]
(24,29]	0.013	-
(29,54]	0.364	0.656

Ainsi, on peut voir qu'un seul test est significatif, dans le cas où p-value = 0,013 < 0,05. On a donc :

- on rejette $H_0^{[11,24];(24,29]}$ **: l'âge moyen est différent entre les individus de la classe d'IMC de corpulence maigre/normale et de la classe de surpoids. Grâce aux moyennes conditionnelles, DANS CE CAS-LA UNIQUEMENT on peut voir que l'âge est plus élevé dans la classe de corpulence la plus élevée (33,3 ans VS 34,3 ans).**

- on ne rejette pas $H_0^{[11,24];(29,54]}$ **: nous n'avons pas réussi à mettre en évidence une différence d'âge moyen entre les individus de la classe d'IMC de corpulence maigre/normale et de la classe d'obésité.**

- on ne rejette pas $H_0^{(24,29];(29,54]}$ **: nous n'avons pas réussi à mettre en évidence une différence d'âge moyen entre les individus de la classe d'IMC de corpulence surpoids et de la classe d'obésité.**

Nous pouvons également utiliser le test t de Bonferroni : pour comparer deux à deux les moyennes de 3 groupes, on doit faire 3 test de comparaisons. Si on prend α = 5%, chacun des tests a 5% de risque de rejeter son H_0 à tort, donc en tout, la probabilité de commettre au moins une erreur avec les 3 tests est supérieure à 5%. On veut contrôler ce risque d'erreur global, qui est d'autant plus élevé que k est grand. Une solution possible est de prendre pour chaque test $\varepsilon = \alpha/3$ = 0,0166. Effectuons le ***test t de Bonferroni*** *:*

	[11,24]	(24,29]
(24,29]	0.04	-
(29,54]	1.00	1.00

Ici, on peut voir que tous les tests sont non-significatifs.

VI. Influence des addictions et conditions de vie

6) a) L'addiction au tabac (*tabagisme2* en oui/non) et l'addiction à l'alcool (*alcool2* en oui/non) sont **deux variables qualitatives** à 2 modalités chacune. Nous allons donc utiliser un **test du X^2 d'indépendance**.

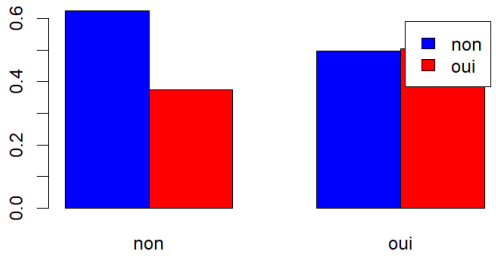
La catégorisation de l'échantillon selon les deux variables donne le **tableau de contingence** suivant :

		tabagisme2		
		Non	Oui	Total
alcool2	Non	537	322	859
	Oui	70	71	141
	Total	607	393	1000

En préalable au test du X^2 d'indépendance, effectuons les graphiques des **profils lignes / colonnes**.

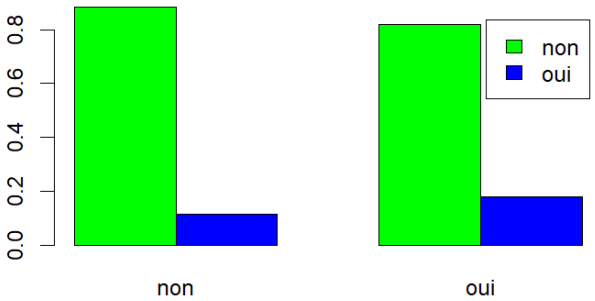
Profils lignes :

Distributions conditionnelles du tabagisme sachant l'alcoolisme



Profils colonnes :

Distributions conditionnelles de l'alcoolisme sachant le tabagisme



Dans les 2 cas, les distributions sont relativement similaires, donc nous ne pouvons pas prévoir l'existence potentielle d'une liaison entre les 2 variables. Faisons alors le **test du X^2 d'indépendance**.

Nous posons :

$$\boldsymbol{H_0 : alcool2 \text{ et } tabagisme2 \text{ sont indépendantes.}}$$

$$\boldsymbol{H_1 : alcool2 \text{ et } tabagisme2 \text{ sont liées.}}$$

Calculons les **effectifs attendus** sous H_0 :

```
          tabagisme2
alcool2      non      oui
    non  521.413  337.587
    oui   85.587   55.413
```

Tous les effectifs attendus sont supérieurs ou égaux à 5, donc le test est valide.

En effectuant le test, nous obtenons **p-value = 0.003734**. Cette p-value est **< 0.05**. Donc **nous rejetons H_0**. Le test est significatif. **Nous concluons qu'il y a une liaison entre l'addiction au tabac et l'addiction à l'alcool.**

Comme nous sommes dans le cas particulier du test où chaque variable à deux modalités, alors, pour conclure, nous comparons les deux proportions observées de « oui » dans deux populations (= deux groupes) $\widehat{p_1}$ *et* $\widehat{p_2}$.

Pour les non-fumeurs ($n_1 = 607$) : $\widehat{p_1} = \frac{o_1}{n_1} = \frac{70}{607} = 0{,}11$

Pour les fumeurs ($n_2 = 393$) : $\widehat{p_2} = \frac{o_2}{n_2} = \frac{71}{393} = 0{,}18$

Pour les non-alcooliques ($n_1 = 859$) : $\widehat{p_1} = \frac{o_1}{n_1} = \frac{322}{859} = 0{,}37$

Pour les alcooliques ($n_2 = 141$) : $\widehat{p_2} = \frac{o_2}{n_2} = \frac{71}{141} = 0{,}50$

Ainsi, un fumeur a plus de chances d'être alcoolique qu'un non-fumeur ; un alcoolique a plus de chances d'être fumeur qu'un non-alcoolique.

6) b) La corpulence (en classes) est une **variable qualitative à 3 modalités**. Le statut tabagique est une **variable qualitative à 2 modalités**. Nous allons donc utiliser un **test du X^2 d'indépendance**.

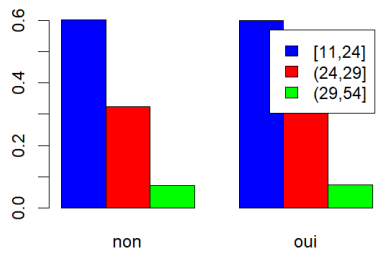
La catégorisation de l'échantillon selon les deux variables donne le **tableau de contingence** suivant :

	imcc			
tabagisme2	**[11,24]**	**(24,29]**	**(29,54]**	**Total**
non	366	197	44	607
oui	236	128	29	393
Total	602	325	73	1000

En préalable au test du X^2 d'indépendance, effectuons les graphiques des **profils lignes / colonnes**.

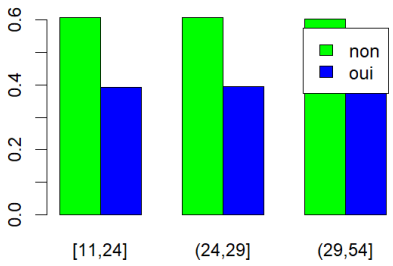
Profils lignes :

Distributions conditionnelles de la corpulence en classe sachant le statut tabagique



Profils colonnes :

Distributions conditionnelles du statut tabagique sachant la corpulence en classes



Dans les 2 cas, les distributions sont très similaires, donc nous ne pouvons pas prévoir l'existence potentielle d'une liaison entre les 2 variables. Faisons alors le **test du X^2 d'indépendance**.

Nous posons :

$$H_0 : tabagisme2 \text{ et } imcc \text{ sont indépendantes}.$$

$$H_1 : tabagisme2 \text{ et } imcc \text{ sont liées}.$$

Calculons les **effectifs attendus** sous H_0 :

tabagisme2	imcc [11,24]	(24,29]	(29,54]
non	365.414	197.275	44.311
oui	236.586	127.725	28.689

Tous les effectifs attendus sont supérieurs ou égaux à 5, donc le test est valide.

En effectuant le test, nous obtenons **p-value = 0,9955**. Cette p-value est **> 0,05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence l'existence d'une liaison entre la corpulence en classes et le statut tabagique.**

6) c) • Cas 1 : Abus d'alcool :

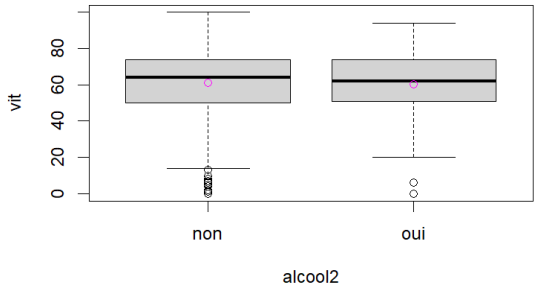
La vitalité des spermatozoïdes est une **variable quantitative**. L'alcoolisme est une **variable qualitative à 2 modalités**.

Calculons d'abord les moyennes conditionnelles des vitalités des spermatozoïdes selon le type de consommation d'alcool. Nous trouvons :

- pour les non-alcooliques, 61,2 % des spermatozoïdes sont vivants dans l'éjaculat.

- pour les alcooliques, 60,5 % des spermatozoïdes sont vivants dans l'éjaculat.

Effectuons les boites à moustache :



On voit sur ce graphique que les dispersions des vitalités sont relativement similaires, tous comme les moyennes et les médianes.

Nous savons que la vitalité des spermatozoïdes ne suit pas une loi normale (d'après la question 3). Comme on n'a pas l'hypothèse gaussienne mais qu'on a néanmoins bien des effectifs assez grands (> 30), on ne peut pas utiliser le test de comparaison des variances de Fisher, on va donc faire le **test de Welch**.

Posons :

$$H_0 : \mu_1 = \mu_2.$$

$$H_1 : \mu_1 \neq \mu_2.$$

(Où μ_1 est l'espérance de la vitalité des spermatozoïdes pour les non-alcooliques ; μ_2 est l'espérance de la vitalité des spermatozoïdes pour les alcooliques).

En effectuant le test de Welch, nous obtenons **p-value = 0.6296**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence une différence entre les moyennes. Nos données ne nous permettent pas de mettre en évidence un impact de l'abus d'alcool sur la vitalité des spermatozoïdes.**

• **Cas 2 : Fumer :**

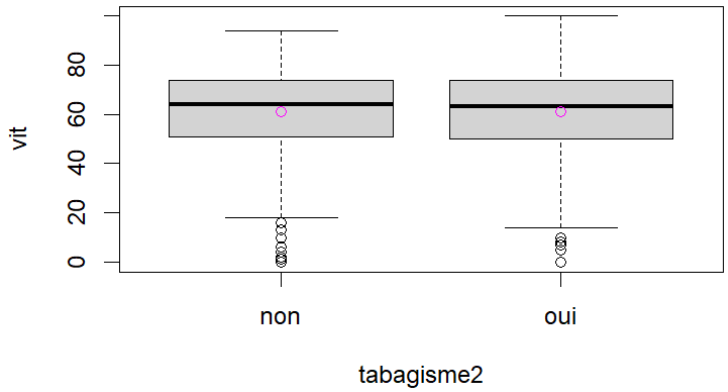
La vitalité des spermatozoïdes est une **variable quantitative**. Le statut de tabagisme est une **variable qualitative à 2 modalités**.

Calculons d'abord les moyennes conditionnelles des vitalités des spermatozoïdes selon le statut de tabagisme. Nous trouvons :

- pour les non-fumeur, 61,2 % des spermatozoïdes sont vivants dans l'éjaculat.

- pour les fumeurs, 61 % des spermatozoïdes sont vivants dans l'éjaculat.

Effectuons les boites à moustache :



On voit sur ce graphique que les dispersions des vitalités sont relativement similaires, tous comme les moyennes et les médianes.

Nous savons que la vitalité des spermatozoïdes ne suit pas une loi normale (d'après la question 3). Comme on n'a pas l'hypothèse gaussienne mais qu'on a néanmoins bien des effectifs assez grands (> 30), on ne peut pas utiliser le test de comparaison des variances de Fisher, on va donc faire le **test de Welch**.

Posons :

$$\boldsymbol{H_0 : \mu_1 = \mu_2.}$$

$$\boldsymbol{H_1 : \mu_1 \neq \mu_2.}$$

(Où μ_1 est l'espérance de la vitalité des spermatozoïdes pour les non-fumeurs ; μ_2 est l'espérance de la vitalité des spermatozoïdes pour les fumeurs).

En effectuant le test de Welch, nous obtenons **p-value = 0.8845**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence une différence entre les moyennes. Nos données ne nous permettent pas de mettre en évidence un impact du tabagisme actif sur la vitalité des spermatozoïdes.**

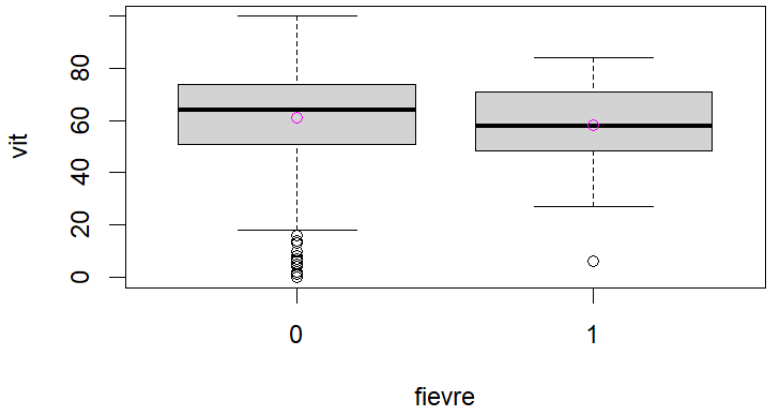
6) d) La vitalité des spermatozoïdes est une **variable quantitative**. Le statut de fièvre est une **variable qualitative à 2 modalités**.

Calculons d'abord les moyennes conditionnelles des vitalités des spermatozoïdes selon le statut de fièvre. Nous trouvons :

- pour les personnes n'ayant pas eu de fièvre récente, 61,2 % des spermatozoïdes sont vivants dans l'éjaculat.

- pour les personnes en ayant eu, 58,1 % des spermatozoïdes sont vivants dans l'éjaculat.

Effectuons les boites à moustache :



On voit sur ce graphique que la dispersion de la vitalité est plus large pour les individus n'ayant pas eu de fièvre récente, et la médiane est un petit peu supérieure.

Nous savons que la vitalité des spermatozoïdes ne suit pas une loi normale (d'après la question 3). De plus, **l'un des effectif de groupe est petit** (19 individus ont eu une fièvre < 30). Ainsi, on ne peut plus utiliser de test paramétrique. On va utiliser le **test non-paramétrique de Wilcoxon-Mann-Whitney** qui compare les médianes.

Posons :

$$\boldsymbol{H_0 : Me_1 = Me_2.}$$

$$\boldsymbol{H_1 : Me_1 \neq Me_2.}$$

(Où Me_1 est la médiane de la vitalité des spermatozoïdes pour les individus sans fièvre récente ; Me_2 est la médiane de la vitalité des spermatozoïdes pour les individus ayant eu une fièvre récente).

En effectuant le test, nous obtenons **p-value = 0.4875**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence une différence entre les médianes. Nos données ne nous permettent pas de mettre en évidence un impact du statut de fièvre sur la vitalité de l'éjaculat.**

6) e) Le statut de fièvre est une **variable qualitative à 2 modalités**. La présence (ou non) d'examens anormaux est une **variable qualitative à 2 modalités**. Nous allons donc utiliser un **test du X^2 d'indépendance**.

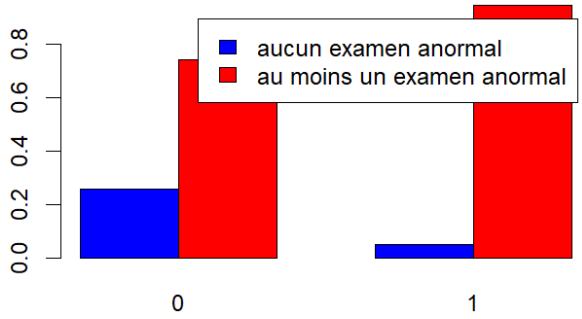
La catégorisation de l'échantillon selon les deux variables donne le **tableau de contingence** suivant :

		examc	
fievre	Aucun exam anormal	Au moins 1 exam anormal	Total
Non	252	729	981
Oui	1	18	19
Total	253	747	1000

En préalable au test du X^2 d'indépendance, effectuons les graphiques des **profils lignes / colonnes**.

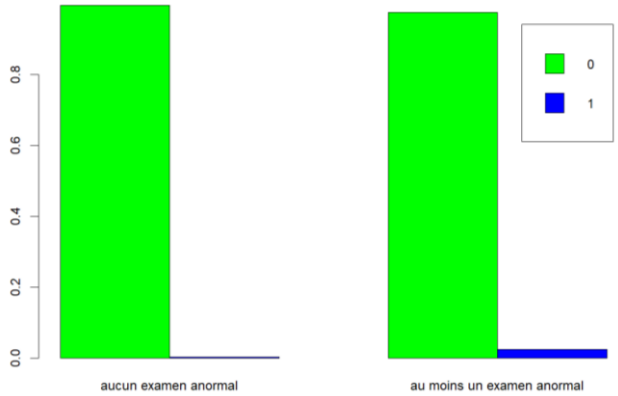
Profils lignes :

Distributions conditionnelles du statut d'examenss anormaux sachant le statut de fièvre



Profils colonnes :

Distributions conditionnelles du statut de fièvre sachant le statut d'examens anormaux



Dans les 2 cas, les distributions sont relativement similaires, donc nous ne pouvons pas prévoir l'existence potentielle d'une liaison entre les 2 variables. Faisons alors le **test du X^2 d'indépendance**.

Nous posons :

$$\boldsymbol{H_0 : fievre\ et\ examc\ sont\ indépendantes.}$$

$$\boldsymbol{H_1 : fievre\ et\ examc\ sont\ liées.}$$

Calculons les **effectifs attendus** sous H_0 :

fievre	examc aucun examen anormal	examc au moins un examen anormal
0	248.193	732.807
1	4.807	14.193

Les effectifs attendus sont supérieurs ou égaux à 5 sauf un. Néanmoins, on en est très proche car 4,807 est très proche de 5. Ainsi, on peut considérer le test comme valide.

En effectuant le test, nous obtenons **p-value = 0.04252**. Cette p-value est **< 0.05** (de peu). Donc **nous rejetons H_0**. Le test est significatif. **Nous concluons qu'il y a une liaison entre un accès de fièvre récent et la présence d'au moins un examen anormal.**

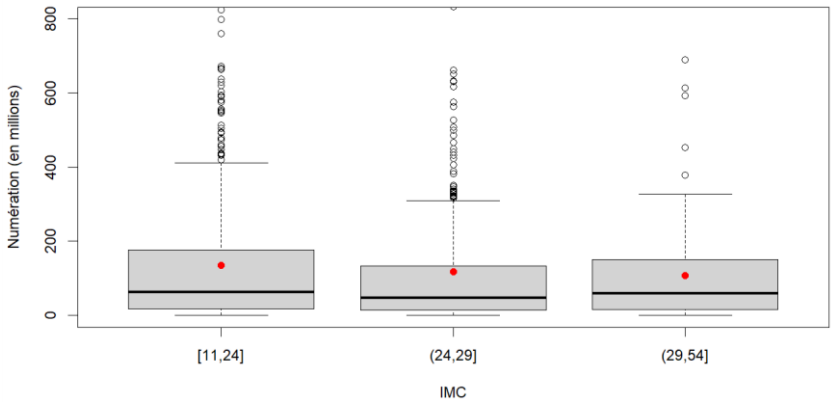
Comme nous sommes dans le cas particulier du test où chaque variable à deux modalités, alors, pour conclure, nous comparons les deux proportions observées de « au moins un examen anormal » dans deux populations (= deux groupes) $\widehat{p_1}$ *et* $\widehat{p_2}$.

- Pour les individus sans fièvre récente ($n_1 = 981$) : $\widehat{p_1} = \frac{o_1}{n_1} = \frac{729}{981} = 0{,}74$

- Pour les individus avec fièvre récente ($n_2 = 19$) : $\widehat{p_2} = \frac{o_2}{n_2} = \frac{18}{19} = 0{,}94$

Ainsi, un individu avec fièvre récente a plus de chances d'avoir au moins un examen anormal qu'un individu sans fièvre récente.

6) f) La numération de l'éjaculat est une **variable quantitative**. La corpulence (avec l'IMC en classes) est une **variable qualitative avec 3 modalités**. Commençons par faire des boites à moustaches pour se donner une idée de la liaison potentielle entre les 2 variables. Si l'on calcule les **moyennes conditionnelles** des numérations sachant les classes d'IMC, on obtient tout d'abord : 134 118 107. Ainsi, on peut déjà observer que, empiriquement, la numération moyenne diminue avec l'augmentation de la corpulence. Les **boites à moustaches** donnent :



Nous ne pouvons pas mettre en évidence de liaison claire avec ces boites à moustaches.

Nous allons alors effectuer un **test paramétrique de comparaison de k moyennes (analyse de la variance à un facteur – ANOVA)**. Tout d'abord, nous devons vérifier les conditions de validité :

- **<u>Normalité des échantillons</u>** : Calculons les effectifs par groupes :

```
imcc
[11,24] (24,29] (29,54]
    602     325      73
```

Comme tous les effectifs par groupes sont suffisamment grands (> 30), alors on peut se passer de l'hypothèse de normalité.

- **<u>Homogénéité des variances</u>** : Posons **H_0 : « Les variances sont égales pour chaque groupe »**. Nous effectuons le test de Brown et Forsythe. Nous obtenons une **p-value = 0.4276**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas l'hypothèse nulle**. **Nous concluons donc que nos données satisfont l'hypothèse d'homogénéité des variances.**

Dès lors, comme nous n'avons pas rejeté l'homoscédasticité, nous pouvons effectuer le test ANOVA.

Nous posons :

$$\boldsymbol{H_0 : \mu_1 = \mu_2 = \mu_3.}$$

$$\boldsymbol{H_1 : \exists i, j \; tels \; que \; \mu_i \neq \mu_j.}$$

En effectuant le test, nous obtenons **p-value = 0.327**. Cette p-value est **> 0.05**. Donc **nous ne rejetons pas H_0**. Le test est non-significatif. **Nous n'avons pas réussi à mettre en évidence qu'au moins deux des classes d'IMC ont des numérations moyennes différentes.**

(On peut aussi regrouper les résultats dans le tableau d'analyse de la variance suivant :)

```
> summary(aov(num~as.factor(imcc)))
                Df   Sum Sq Mean Sq F value Pr(>F)
as.factor(imcc)  2    86167   43083   1.119  0.327
Residuals      997 38384104   38500
```

VII. Question bonus

• Effectuons une régression linéaire multiple pour évaluer l'impact des variables qualitatives sur la vitalité.
Notre modèle est :

$$vit_i = \beta_0 + \beta_1 alcool2_i + \beta_2 tabagisme2_i + \beta_3 fievre_i + \beta_4 imcc_i + \beta_5 examc_i + \varepsilon_i$$

$$i \in \{1, 2, \dots, 1000\}$$

Effectuons la régression linéaire de cette façon :

```
modele = lm(vit ~ alcool2 + tabagisme2 + fievre + imcc + examc, data=fertiproj)
summary(modele)
```

Nous obtenons :

```
Residuals:
   Min     1Q Median     3Q    Max
-62.27 -10.49   2.93  13.31  39.46

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    62.4278     1.2938  48.253   <2e-16 ***
alcool2oui                     -0.8496     1.6154  -0.526    0.599
tabagisme2oui                  -0.1537     1.1489  -0.134    0.894
fievre                         -2.8280     4.1112  -0.688    0.492
imcc(24,29]                    -0.3791     1.2169  -0.312    0.755
imcc(29,54]                     0.7072     2.1925   0.323    0.747
examcau moins un examen anormal  -1.3578     1.2898  -1.053    0.293
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.67 on 993 degrees of freedom
Multiple R-squared:  0.002205,  Adjusted R-squared:  -0.003824
F-statistic: 0.3657 on 6 and 993 DF,  p-value: 0.9008
```

Ainsi, on peut voir que nous avons un **impact négatif** de l'alcoolisme, du tabagisme, de la présence de fièvre récente, d'une corpulence en surpoids, et de la présence d'au moins un examen anormal **sur la vitalité**. Néanmoins, ici, être obèse ne semble pas avoir d'impact négatif sur la vitalité par rapport à si l'individu était de corpulence maigre ou normale.

• Effectuons deux autres régressions linéaires pour étudier l'impact des variables qualitatives sur :

- le volume :

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8332 -1.1732 -0.1380  0.9741 10.0167

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.93323    0.12249  32.110  < 2e-16 ***
alcool2oui                    -0.25251    0.15294  -1.651   0.0991 .
tabagisme2oui                 -0.11206    0.10878  -1.030   0.3032
fievre                        -0.22488    0.38924  -0.578   0.5636
imcc(24,29]                   -0.49216    0.11522  -4.272 2.13e-05 ***
imcc(29,54]                   -0.04530    0.20758  -0.218   0.8273
examcau moins un examen anormal  0.05006    0.12212   0.410   0.6820
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.673 on 993 degrees of freedom
Multiple R-squared:  0.02307,   Adjusted R-squared:  0.01717
F-statistic: 3.909 on 6 and 993 DF,  p-value: 0.0007243
```

Ainsi, on peut voir que nous avons un **impact négatif** de l'alcoolisme, du tabagisme, de la présence de fièvre récente, d'une corpulence en surpoids, d'une corpulence obèse sur le **volume**. Néanmoins, la présence d'au moins un examen anormal ne semble pas avoir d'impact négatif sur le volume.

- la concentration :

```
Residuals:
   Min     1Q Median     3Q    Max
-47.86 -32.43 -21.07  10.77 454.17

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     45.489      4.287  10.610   <2e-16 ***
alcool2oui                      -3.230      5.353  -0.603    0.546
tabagisme2oui                    1.006      3.807   0.264    0.792
fievre                           4.101     13.624   0.301    0.763
imcc(24,29]                      1.484      4.033   0.368    0.713
imcc(29,54]                     -5.341      7.265  -0.735    0.462
examcau moins un examen anormal   -9.664      4.274  -2.261    0.024 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 58.54 on 993 degrees of freedom
Multiple R-squared:  0.006393,  Adjusted R-squared:  0.0003889
F-statistic: 1.065 on 6 and 993 DF,  p-value: 0.3819
```

Ainsi, on peut voir que nous avons un **impact négatif** de l'alcoolisme, d'une corpulence obèse et de la présence d'au moins un examen anormal sur la **concentration**. Néanmoins, le tabagisme, la présence d'une fièvre récente et une corpulence en surpoids ne semble pas avoir d'impact négatif sur la concentration.

VIII. Conclusion

Nos analyses statistiques nous permettent de tirer les conclusions suivantes :
(Nous étudions l'impact des variables lignes sur les variables colonnes)

Variables portant sur la qualité de l'éjaculat → / Variables portant sur les caractéristiques du patient ↓	**Vitalité**	**Volume**	**Mobilité**	**Numération**
Alcoolisme	Nos données ne nous permettent pas de mettre en évidence un impact de l'abus d'alcool sur la vitalité des spermatozoïdes.			
Tabagisme	Nos données ne nous permettent pas de mettre en évidence un impact du tabagisme actif sur la vitalité des spermatozoïdes.			
Âge	**Liaison linéaire (négative, légèrement) : plus l'âge est élevé, plus la vitalité des spermatozoïdes diminue.**	**Liaison linéaire négative : plus l'âge est élevé, plus le volume de l'éjaculat diminue.**	Nous n'avons pas réussi à mettre en évidence l'existence d'une liaison linéaire entre l'âge et la mobilité des spermatozoïdes.	
IMC en classes				Nous n'avons pas réussi à mettre en évidence qu'au moins deux des classes d'IMC ont des numérations moyennes différentes.
Présence ou non d'une fièvre récente	Nos données ne nous permettent pas de mettre en évidence un impact du statut de fièvre sur la vitalité de l'éjaculat.			

Autres influences	**Tabagisme**	**IMC en classes**	**Présence ou non d'un examen anormal**
Âge	**Il y a une liaison entre âge et comportement tabagique : les moyennes conditionnelles laissent à penser que les jeunes ont plus tendance à fumer que les individus plus âgés.**	**Il y a une liaison entre l'âge et la corpulence en classes.**	
Alcoolisme	**Il y a une liaison entre l'addiction au tabac et l'addiction à l'alcool. Un fumeur a plus de chances d'être alcoolique qu'un non-fumeur ; un alcoolique a plus de chances d'être fumeur qu'un non-alcoolique.**		
Tabagisme		Nous n'avons pas réussi à mettre en évidence l'existence d'une liaison entre la corpulence en classes et le statut tabagique.	
Fievre			**Il y a une liaison entre un accès de fièvre récent et la présence d'au moins un examen anormal. Un individu avec fièvre récente a plus de chances d'avoir au moins un examen anormal qu'un individu sans fièvre récente.**

Annexe – Code R

```
# CODE R - Projet statistique - Tests d'hypothèses - BESSEDE DESIRE HADI

fertiproj=read.table("fertiproj11.txt",header=TRUE)
attach(fertiproj) # donc plus besoin de préciser fertiproj$variable

    # QUESTION 2

# Recodage de alcool et tabagisme :
alcool2=factor(alcool,labels=c("non","oui")) # oui pour 1 et non pour 0
tabagisme2=factor(tabagisme,labels=c("non","oui"))

# Création de la variable num :
num=vol*conc

# Création de la variable imcc :
summary(imc) # pour obtenir les bornes min et max
imcc=cut(imc,breaks=c(11,24,29,54),include.lowest=TRUE)

# Création de la variable examc :
summary(exam)
exam1=cut(exam,breaks=c(-1,0,6),include.lowest=TRUE)
examc=factor(exam1,labels=c("aucun examen anormal","au moins un examen anormal"))

    # QUESTION 3

# Etude descriptive des variables :
summary(alcool2)
summary(tabagisme2)
table(fievre)
summary(imcc)
summary(examc)
summary(vol)
summary(conc)
summary(vit)
summary(age)
summary(dureeinf)
summary(mob)
summary(num)

# Graphes :
plot(alcool2,main="Le patient est-il alcoolique ?")
plot(tabagisme2,main="Le patient fume-t-il ?")
plot(imcc,main="Répartition dans les 3 classes de valeurs de l'IMC")
plot(examc,main="Répartition dans les 2 classes de la variable examc")
hist(vol, freq = FALSE, main="Histogramme : vol")
boxplot(vol,main="Boite à moustache : vol")
hist(conc, freq = FALSE, main="Histogramme : conc")
boxplot(conc,main="Boite à moustache : conc")
hist(vit, freq = FALSE, main="Histogramme : vit")
boxplot(vit,main="Boite à moustache : vit")
hist(age, freq = FALSE, main="Histogramme : age")
boxplot(age,main="Boite à moustache : age")
hist(dureeinf, freq = FALSE, main="Histogramme : dureeinf")
boxplot(dureeinf,main="Boite à moustache : dureeinf")
hist(mob, freq = FALSE, main="Histogramme : mob")
boxplot(mob,main="Boite à moustache : mob")
hist(num, freq = FALSE, main="Histogramme : num")
boxplot(num,main="Boite à moustache : num")

# Test de la normalité de vit :
# Test de Kolmogorov-Smirnov d'adéquation à une loi normale quelconque :
# On met mean(vit) et sd(vit) car on teste que la loi est normale
# mais sans préciser avec quels paramètres comme on ne sait pas
# ce que valent mu et sigma, donc on en prend des estimations sur les données
ks.test(vit,"pnorm",mean(vit),sd(vit))

    # QUESTION 4 - A
```

```
summary(dureeinf)

    # QUESTION 4 - B

table(imcc) # Pour obtenir la répartition empirique (dans notre échantillon)
effempiriques = c(602,325,73)
p0 = c(0.584,0.292,0.124)

# Pour vérifier les conditions de validité du test :
chisq.test(effempiriques,p=p0)$expected # Effectifs attendus

# Test du X² d'adéquation :
chisq.test(effempiriques,p=p0)

    # QUESTION 4 - C

table(tabagisme2)
effempiriques = c(607,393)
p0 = c(0.64,0.36)

# Pour vérifier les conditions de validité du test :
chisq.test(effempiriques,p=p0)$expected

# Test du X² d'adéquation :
chisq.test(effempiriques,p=p0)

    # QUESTION 4 - D

# Test de Student de comparaison d'une espérance à une valeur donnée :
t.test(num,mu=250)

    # QUESTION 5 - A

# Nuage de points :
plot(x=age,y=vol,xlab="Âge",ylab="Volume de l'éjaculat",
   xlim=c(10, 65),ylim=c(0, 15),
   main="Nuage de points : variables age et vol")

# Coefficient de correlation :
cor(age,vol,use="pairwise.complete.obs")

# Test du coefficient de corrélation linéaire de Pearson :
cor.test(age,vol)

    # QUESTION 5 - B

# VITALITE :

# Nuage de points :
plot(x=age,y=vit,xlab="Âge (en années)",ylab="Vitalité des spermatozoïdes (%)",
   main="Nuage de points : variables age et vit")

# Coefficient de correlation :
cor(age,vit,use="pairwise.complete.obs")

# Test du coefficient de corrélation linéaire de Pearson :
cor.test(age,vit)

# MOBILITE :

# Nuage de points :
plot(x=age,y=mob,xlab="Âge (en années)",ylab="Mobilité des spermatozoïdes (%)",
   main="Nuage de points : variables age et mob")

# Coefficient de correlation :
cor(age,mob,use="pairwise.complete.obs")

# Test du coefficient de corrélation linéaire de Pearson :
cor.test(age,mob)
```

```
    # QUESTION 5 - C.

# Moyennes conditionnelles des âges selon le comportement tabagique :
moycondtab=tapply(age,tabagisme2,mean)

# Boites à moustaches :
boxplot(age~tabagisme2)
points(moycondtab,col="magenta")

# Effectifs pour chaque modalités de tabagisme2 :
table(tabagisme2)

# Test de Welch :
t.test(age~tabagisme2,var.equal=FALSE)

    # QUESTION 5 - D.

# Moyennes conditionnelles :
moycondage=tapply(age,imcc,mean)

# Boites à moustaches juxtaposées :
boxplot(age~imcc,xlab="IMC",ylab="Âge (en année)")
points(moycondage,col="red",pch=19)

# Vérification des conditions de validité :
# Normalité : Calculons les effectifs par groupe :
table(imcc)
# Homogénéité des variances : Effectuons un test de Brown et Forsythe :
library(car)
leveneTest(age,as.factor(imcc),center=median)

# Anova classique comme on n'a pas rejeté l'homoscédasticité :
oneway.test(age~as.factor(imcc),var.equal=TRUE)
# Si on veut aussi le tableau d'analyse de la variance :
summary(aov(age~as.factor(imcc)))

# Tests de comparaisons multiples :
# Test LSD :
pairwise.t.test(age,imcc,p.adjust.method = "none")
# Test de Bonferroni :
pairwise.t.test(age,imcc,p.adjust.method = "bon")

    # QUESTION 6 - A

# Tableau de contingence des 2 variables qualitatives :
tableaucont=table(alcool2,tabagisme2)
print(tableaucont)

# Profils lignes :
round(prop.table(tableaucont,1),digits=2)
barplot(t(prop.table(tableaucont,1)),beside=TRUE,col=c("blue","red"),
      legend.text=TRUE)

# Profils colonnes :
round(prop.table(tableaucont,2),digits=2)
barplot(prop.table(tableaucont,2),beside=TRUE,col=c("green","blue"),
      legend.text=TRUE)

# Effectifs attendus sous H0 :
chisq.test(tableaucont,correct=FALSE)$expected

# Test du X² d'indépendance :
chisq.test(tableaucont,correct=FALSE)

    # QUESTION 6 - B

# Tableau de contingence des 2 variables qualitatives :
tableaucont6b=table(tabagisme2,imcc)
print(tableaucont6b)

# Profils lignes :
```

```
round(prop.table(tableaucont6b,1),digits=2)
barplot(t(prop.table(tableaucont6b,1)),beside=TRUE,col=c("blue","red","green"),
      legend.text=TRUE)

# Profils colonnes :
round(prop.table(tableaucont6b,2),digits=2)
barplot(prop.table(tableaucont6b,2),beside=TRUE,col=c("green","blue"),
      legend.text=TRUE)

# Effectifs attendus sous H0 :
chisq.test(tableaucont6b,correct=FALSE)$expected

# Test du X² d'indépendance :
chisq.test(tableaucont6b,correct=FALSE)

    # QUESTION 6 - C

# Moyennes conditionnelles des vitalités selon la consommation d'alcool :
moycondvit1=tapply(vit,alcool2,mean)

# Boites à moustaches :
boxplot(vit~alcool2)
points(moycondvit1,col="magenta")

# Test de Welch :
t.test(vit~alcool2,var.equal=FALSE)

# Moyennes conditionnelles des vitalités selon le statut tabagique :
moycondvit2=tapply(vit,tabagisme2,mean)

# Boites à moustaches :
boxplot(vit~tabagisme2)
points(moycondvit2,col="magenta")

# Test de Welch :
t.test(vit~tabagisme2,var.equal=FALSE)

    # QUESTION 6 - D.

# Moyennes condionnelles des vitalités selon le statut de fièvre :
moycondvit3=tapply(vit,fievre,mean)

# Boites à moustaches :
boxplot(vit~fievre)
points(moycondvit3,col="magenta")

table(fievre)

# Test non paramétrique de Wilcoxon-Mann-Whitney :
wilcox.test(vit~fievre)

    # QUESTION 6 - E

# Tableau de contingence des 2 variables :
tableaucont2=table(fievre,examc)
print(tableaucont2)

# Profils lignes :
round(prop.table(tableaucont2,1),digits=2)
barplot(t(prop.table(tableaucont2,1)),beside=TRUE,col=c("blue","red"),
      legend.text=TRUE)

# Profils colonnes :
round(prop.table(tableaucont2,2),digits=2)
barplot(prop.table(tableaucont2,2),beside=TRUE,col=c("green","blue"),
      legend.text=TRUE)

# Effectifs attendus sous H0 :
chisq.test(tableaucont2,correct=FALSE)$expected

# Test :
```

```
chisq.test(tableaucont2,correct=FALSE)

    # QUESTION 6 - F

# Moyennes conditionnelles :
moycond=tapply(num,imcc,mean)

# Boites à moustaches juxtaposées :
boxplot(num~imcc,xlab="IMC",ylab="Numération (en millions)",ylim=c(0,800))
points(moycond,col="red",pch=19)

# Vérification des conditions de validité :
# Normalité : Calculons les effectifs par groupe :
table(imcc)
# Homogénéité des variances : Effectuons un test de Brown et Forsythe :
library(car)
leveneTest(num,as.factor(imcc),center=median)

# Anova classique comme on n'a pas rejeté l'homoscédasticité :
oneway.test(num~as.factor(imcc),var.equal=TRUE)
# Si on veut aussi le tableau d'analyse de la variance :
summary(aov(num~as.factor(imcc)))

    # QUESTION 7

# Régression linéaire multiple :

modele = lm(vit ~ alcool2 + tabagisme2 + fievre + imcc + examc, data=fertiproj)
summary(modele)

modele = lm(vol ~ alcool2 + tabagisme2 + fievre + imcc + examc, data=fertiproj)
summary(modele)

modele = lm(conc ~ alcool2 + tabagisme2 + fievre + imcc + examc, data=fertiproj)
summary(modele)
```