Bebes

NIANG

2022-12-07

***Objectif* :** Déterminer si le poids des bébés varie selon plusieurs critères de la mère comme le poids, l'age, la taille et le comportement vis-à-vis du tabac des mères.

```
## [1] "C:/Users/abdou/Documents/cours analyse de données"
```

1. Import fichier bébé dans R

```
##    bwt gestation parity age height weight smoke tension
## 1 3.43       284  FALSE  27  155.0  45.30 FALSE    16.1
## 2 3.23       282  FALSE  33  160.0  61.16 FALSE    12.7
## 3 3.66       279  FALSE  28  160.0  52.09  TRUE    14.8
## 4 3.51        NA  FALSE  36  172.5  86.07 FALSE    12.8
## 5 3.09       282  FALSE  23  167.5  56.62  TRUE    13.3
## 6 3.89       286  FALSE  25  155.0  42.13 FALSE    13.3
```

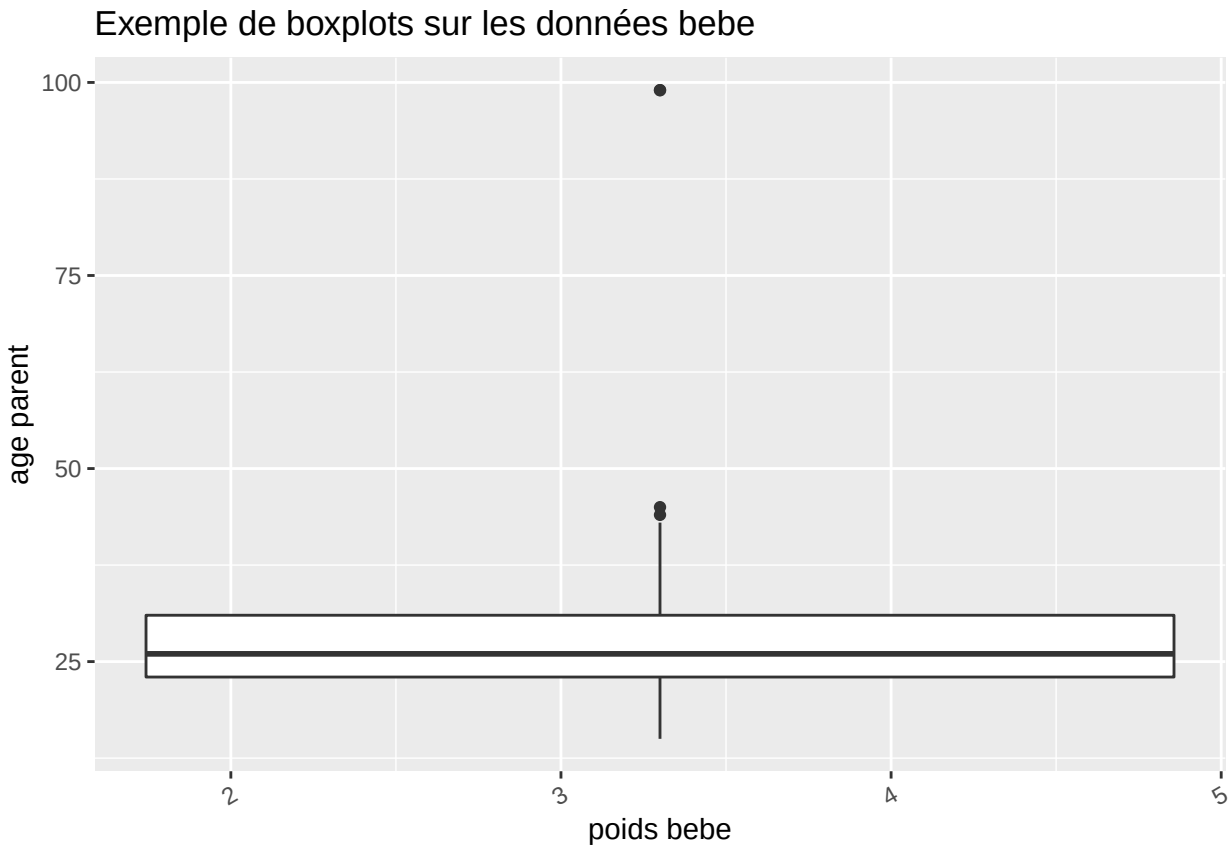
2. Histogramme age des mères Commencons d'abord par une recherche du maximum et du minimum des ages :

```
## [1] 99
```

```
## [1] 15
```

99 ans nous parait comme un chiffre aberrant. Une grossesse à 15 ans peut etre envisagée sous forme de grossesse précoce donc nous nous séparerons de 99 ans et nous conserverons l'age 15 ans.

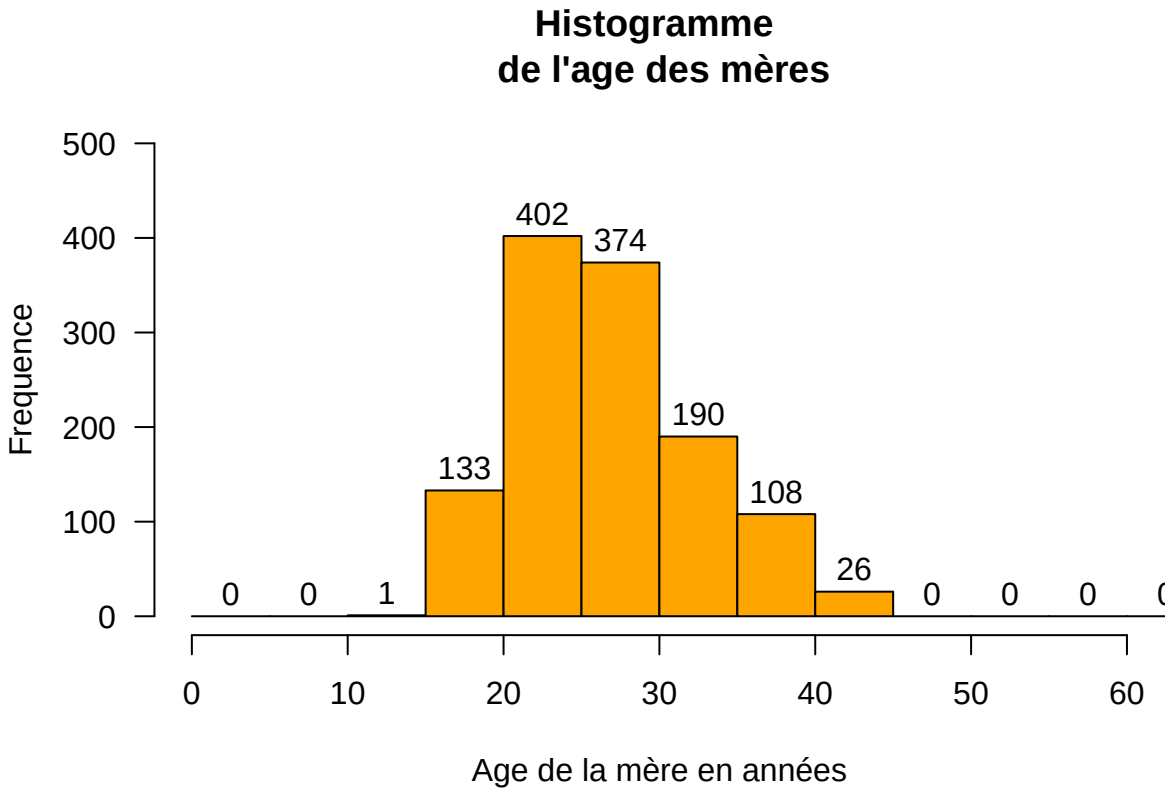
Un boxplot permet de mieux isoler la valeur aberrante 99 ans.



On voit bien que l'age 99 ans est un outlier. Les autres valeurs au delà du 3ème quartile de la moustache sont envisageables comme ages de grossesse (grossesses pour age entre [40,50[)

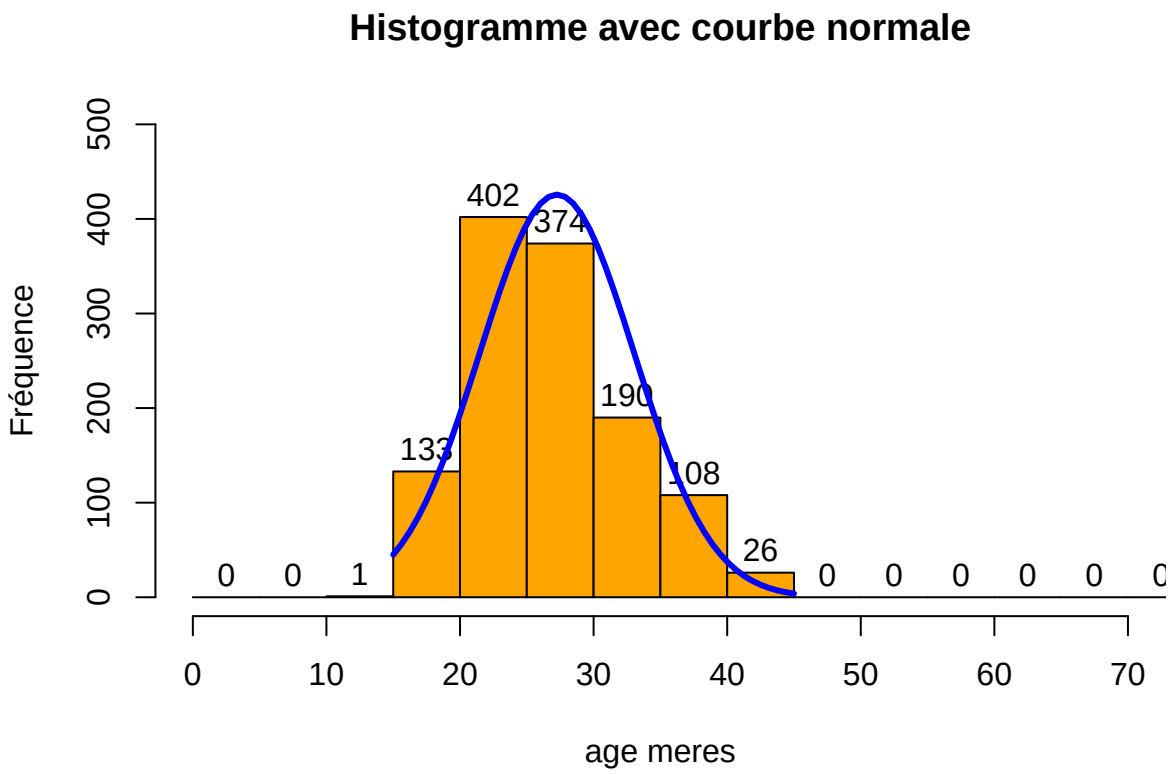
Nous faisons le choix de conserver des ages entre 15 et 50 max.

Nous obtenons l'histogramme suivant:

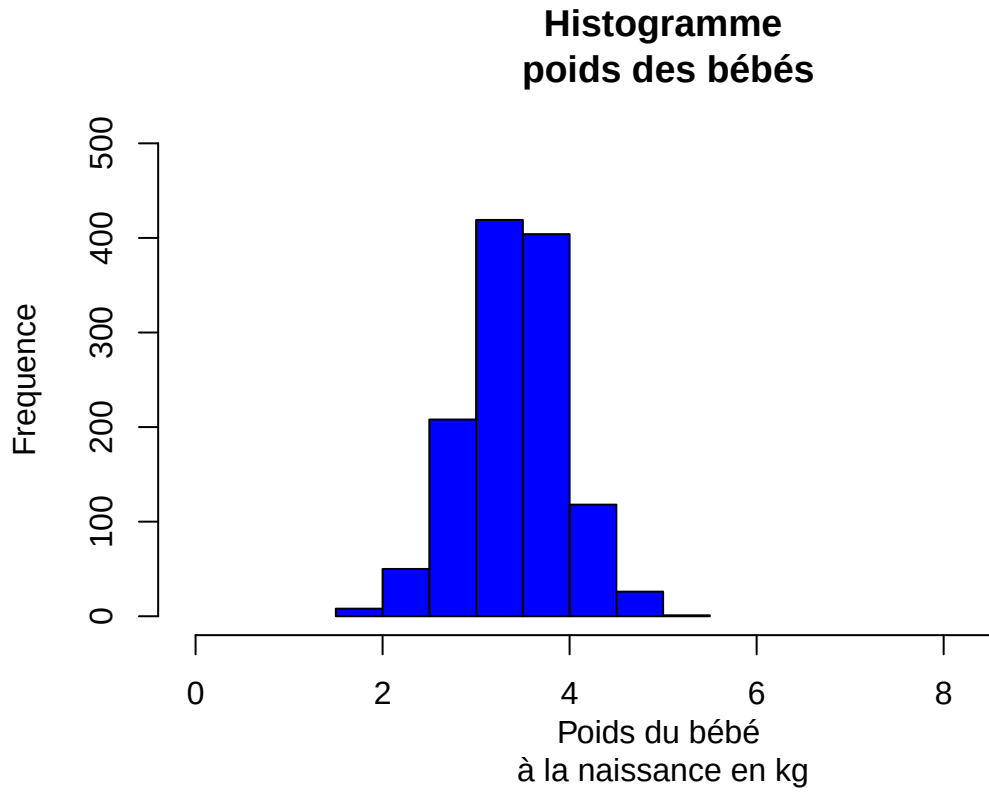


Nous remarquons les naissances sont généralement concentrées sur la période 20 à 30 ans. Quelques rares cas de naissances au delà de 40 ans (26) et une naissance à 15 ans, l'age minimum ici.

Par ailleurs si on superposait la courbe de la loi normale a notre histogramme nous aurions ceci :

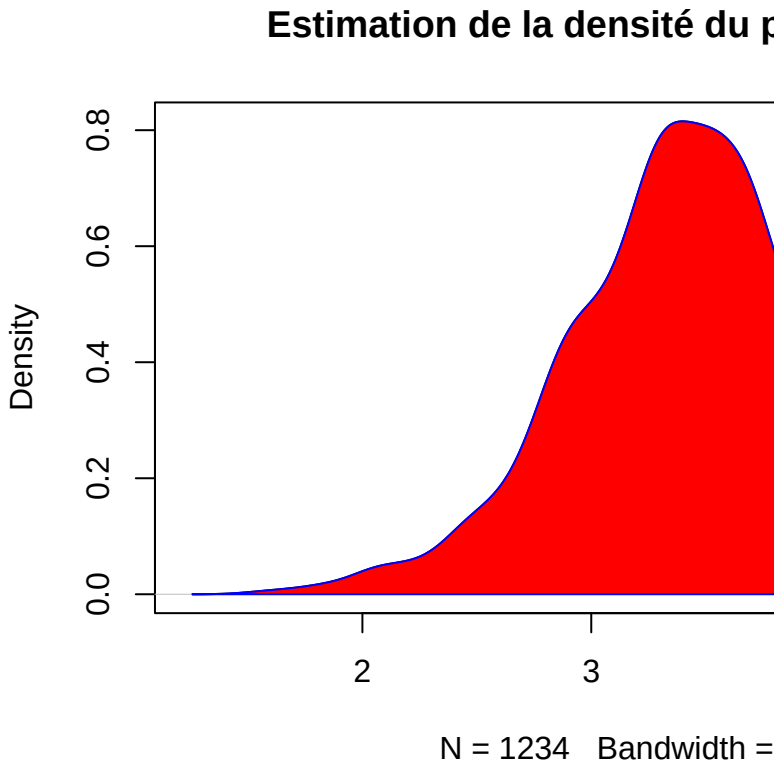


On voit bien que la distribution de la variable age des mères est similaire à celle d'une loi normale.



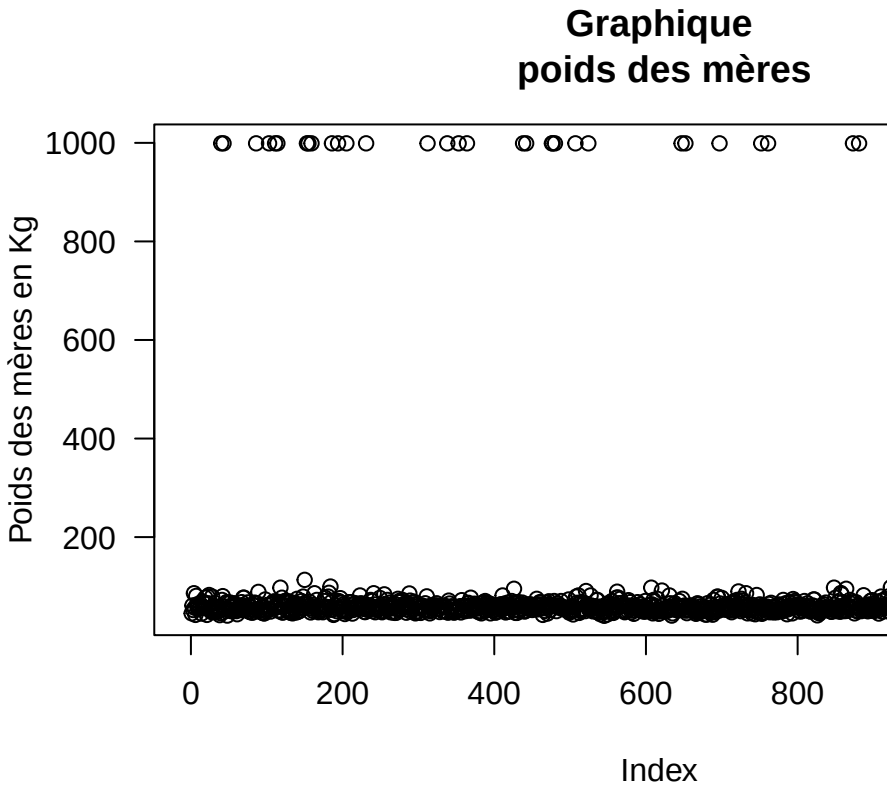
3. Histogramme poids des bébés

Les histogrammes peuvent etre une méthode peu efficace pour déterminer la forme d'une distribution parce qu'ils sont fortement affectés par le nombre de breaks qu'on utlise (si je change la taille de mon break, la forme de la distribution évolue). D'ou l'interet de faire un tracé de l'estimation de la densité.



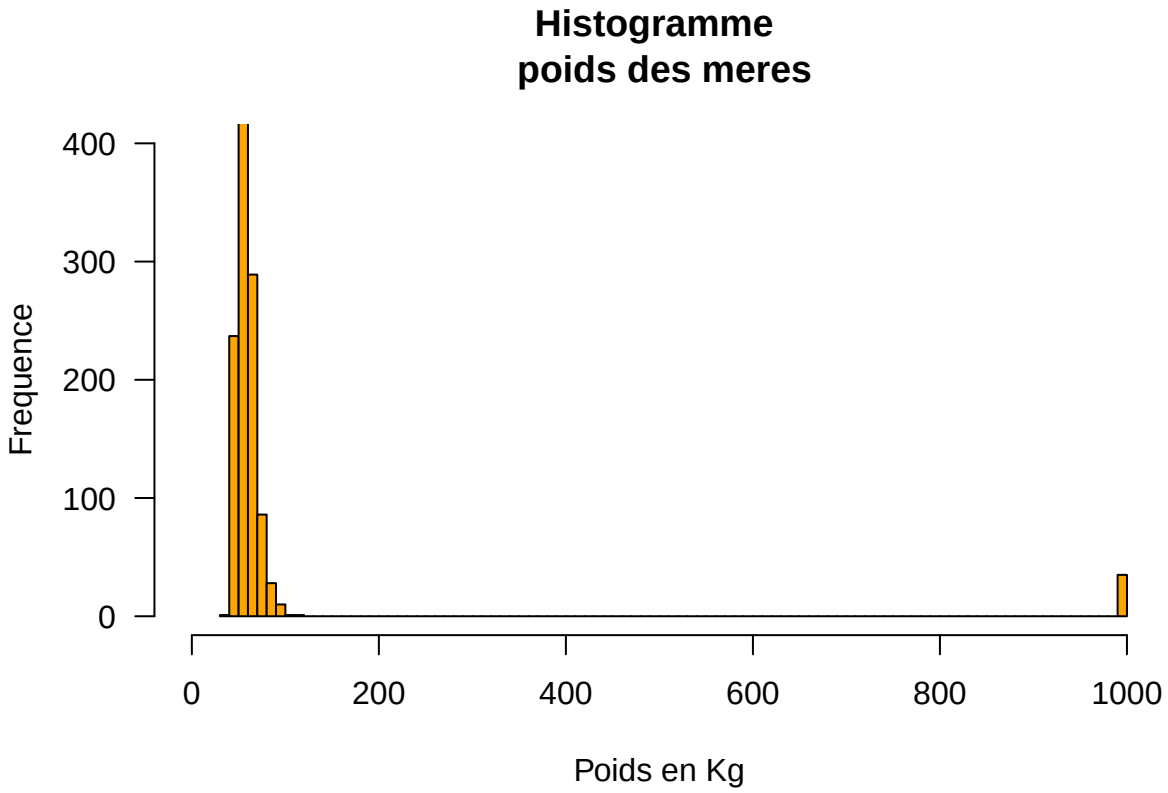
4. Tracé de l'estimation densité poids des bébés :

Elle est similaire à la densité de la distribution d'une loi normale.



5. Etude graphique du poids des mères :

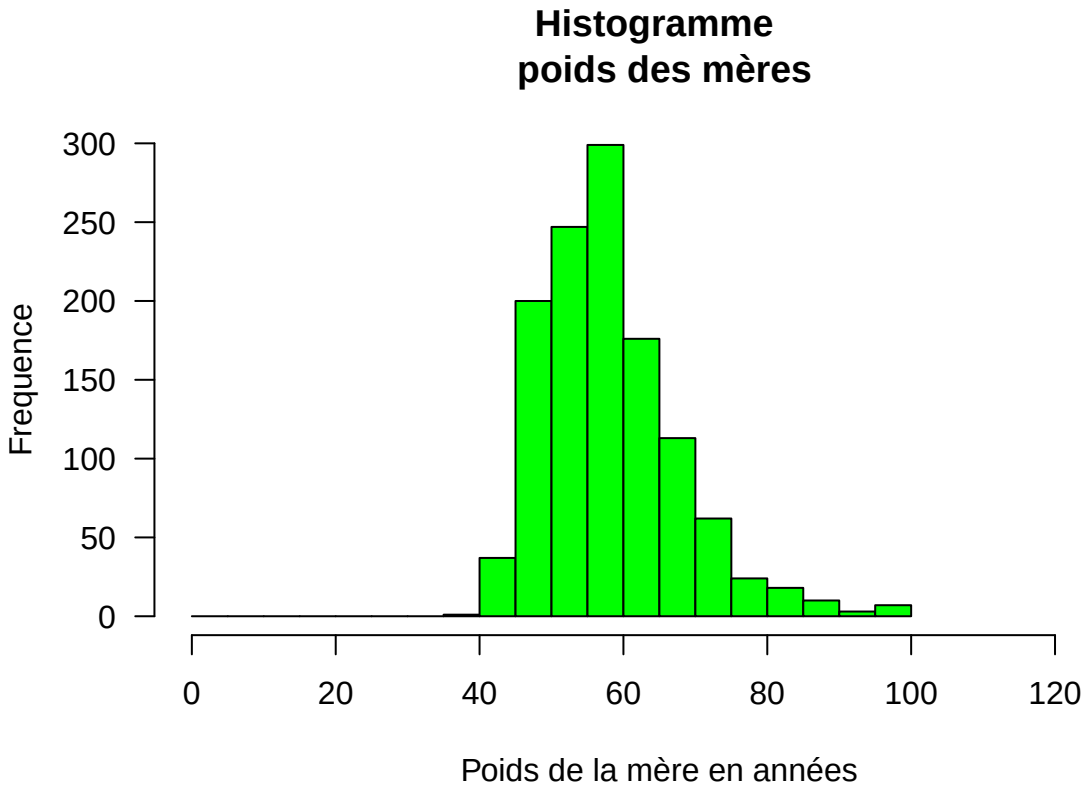
Les poids sont globalement dans un meme intervalle sauf une petite partie qui est proche des 1000kg.

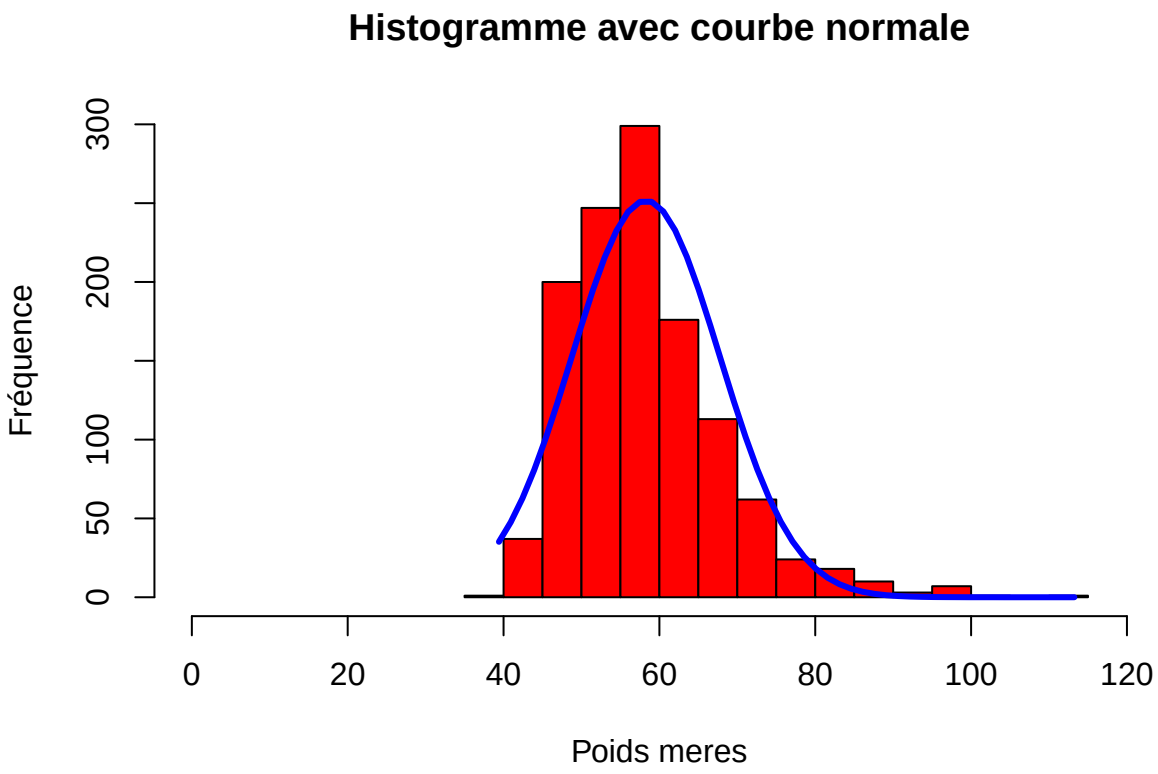


Nous pouvons voir ainsi que les poids atteignant des valeurs proches de 1000kg sont aberrantes. Nous allons choisir un intervalles de poids raisonnables et éliminer celles qui dépassent. Nous choisirons tous les poids ne dépassant pas 120kg :

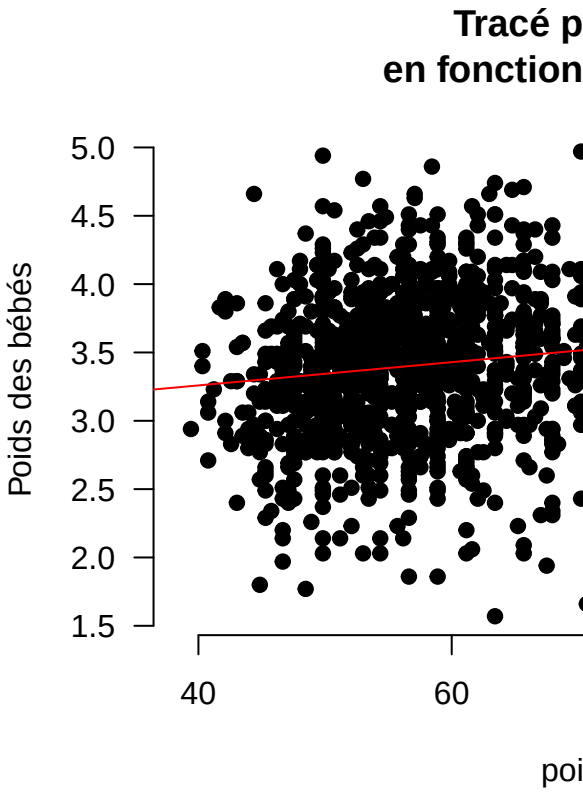
```
## [1] 39.41
```

39.41 correspond au minimum des poids mères


Histogramme
poids des mères
Frequence
300
250
200
150
100
50
0
0
20
40
60
80
100
120
Poids de la mère en années



On peut remarquer que si le poids des bébés semble proche d'une loi normale, celle des mères s'en éloigne.



6. Etude graphique de la relation entre les poids mères et bébés

On devine qu'il y'a une relation entre poids mères et poids bébés. Un intervalle *poids des mères* entre [50kg,70kg] dans lequel le poids des bébés varie entre[2.5kg,4kg] et des valeurs en dehors de l'intervalle ou l'évolution est très dispersée. Nous ne pouvons pas conclure de liaison car elle n'est pas très claire.

```
## [1] 0.1540447
```

La valeur de la corrélation entre poids des bébés et poids des mères est très faible.

7. Création des classes poids de la mère et poids des enfants Nous avons créé un tableau croisé des effectifs poids mères/bébés :

	(0,50]	(50,60]	(60,70]	(70,120]
(0,3]	74	108	59	20
(3,3.5]	92	192	82	40
(3.5,4]	54	189	104	44
(4,6]	18	57	44	22

8. Test d'indépendance du Khi2 entre *bwtClass* et *weightClass* Test d'indépendance du Khi2 au niveau 0.001 :

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  bwt_weight_Cross
## X-squared = 37.257, df = 9, p-value = 2.368e-05
```

La p-value $< 0.001\%$. On peut donc rejeter l'hypothèse d'indépendance. C'est à dire il existe une liaison forte entre *bwtClass* et *weightClass*

NB : On peut regarder les ecarts à l'indépendance : Cet écart à l'indépendance représente l'écart entre l'effectif observé et l'effectif théorique, et ceci pour chacune des cases du tableau de contingence.

```
##          weightClass
## bwtClass      (0,50]    (50,60]    (60,70]   (70,120]
##   (0,3]    3.0831447 -0.9955994 -0.4929568 -1.4182983
##   (3,3.5]  1.2709399  0.5233376 -1.6032367 -0.4080834
##   (3.5,4] -2.6803043  0.8203610  1.0049110  0.4540903
##   (4,6]   -1.8880094 -0.8995992  1.7177756  1.8659483
```

On peut remarquer que l'ecart à l'indépendance entre les modalités (0,50] et (0,3] est positif. Il correspond donc à une attraction entre les deux modalités.

Alors que les modalités (3.5,4] et (0,50] s'opposent.

Les autres attractions oppositions ne sont pas fortes.

9. AFC poids entre les poids de la mère et du bébé *AFC*

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  4  categories; the column variable has 4 categories
## The chi square of independence between the two variables is equal to 37.25709 (p-value =  2.367823e-0
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

Valeurs propres

```
##         eigenvalue variance.percent cumulative.variance.percent
## Dim.1 0.0260614181        83.870324                     83.87032
## Dim.2 0.0040608358        13.068499                     96.93882
## Dim.3 0.0009512139         3.061177                    100.00000
```

Il y'a 3 valeurs propres. La première explique 82.5% de l'information.

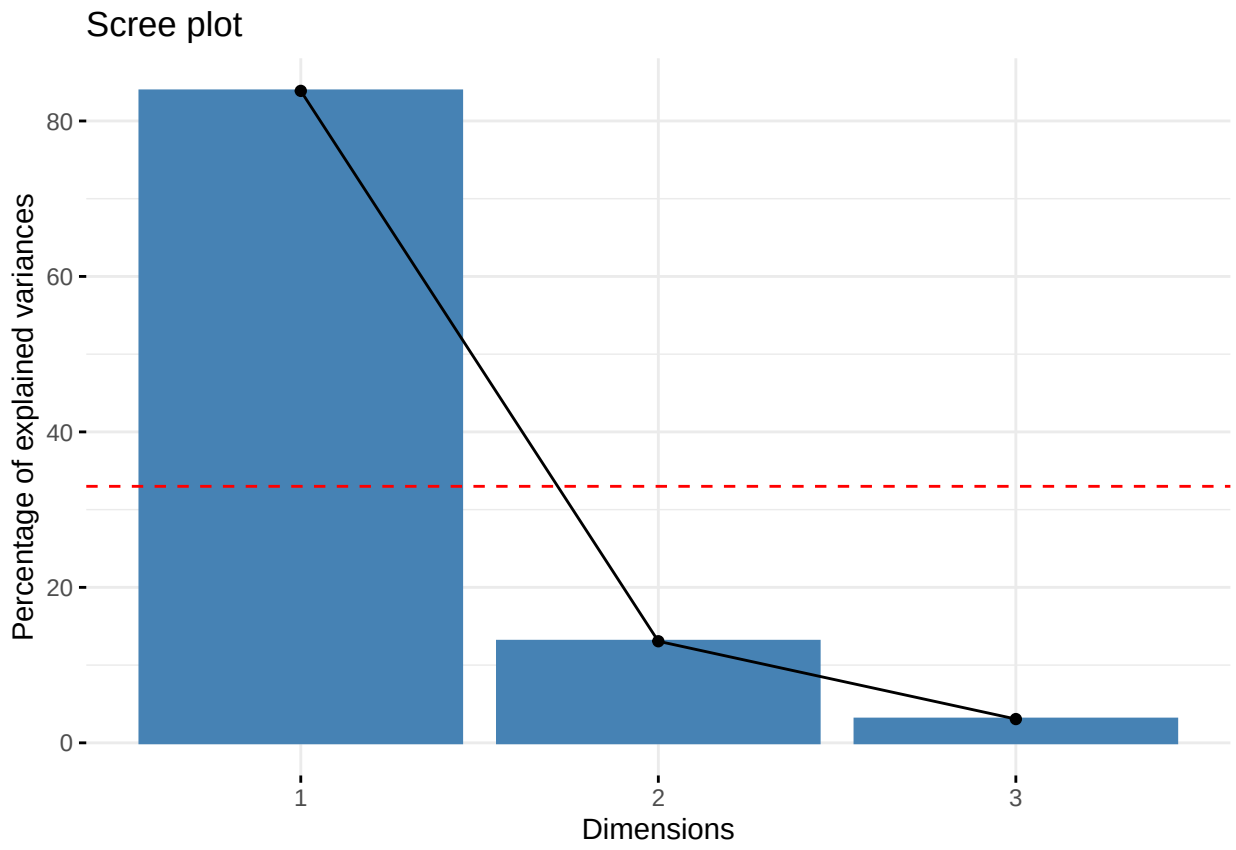
Nous choisissons de conserver un seul axe par la règle du seuil, ***82.5%*** nous semblant un pourcentage acceptable pour expliquer l'information totale.

Il est également possible de calculer une valeur propre moyenne au-dessus de laquelle l'axe doit être conservé dans le résultat:

Selon le graphique ci-dessus, seule la dimensions 1 doit être considérée pour l'interprétation de la solution. Les dimensions 2 et 3 expliquent seulement 17,4% de l'inertie totale, ce qui est inférieur à la valeur moyenne des axes (33,33%) et trop petit pour être conservé pour une analyse plus approfondie.

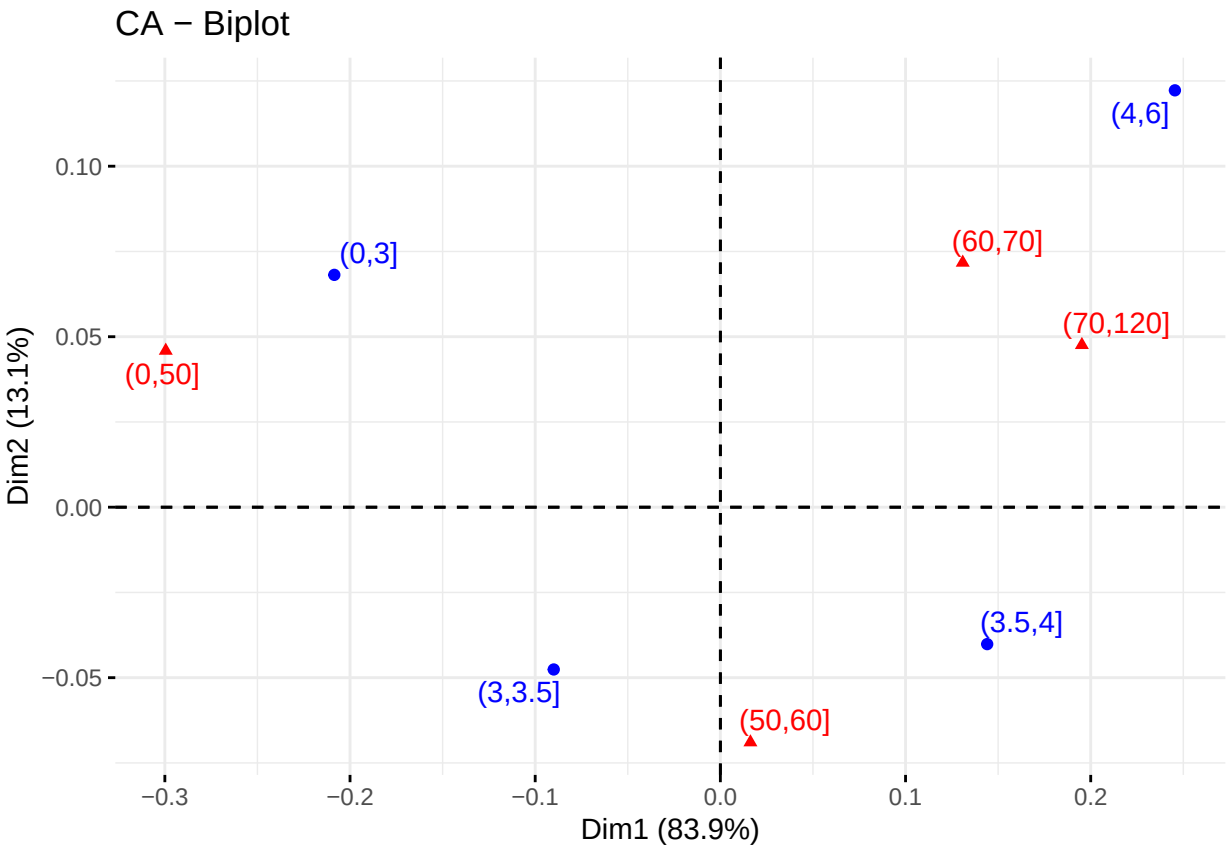
```
## [1] "la somme des valeurs propres est :"
```

```
## eigenvalue
## 0.03107347
```



NB : La valeur propre ou inertie mesurant l'intensité de la liaison, on peut dire qu'elle est très faible ici : en effet $\sum \lambda_i = 0.03100034$ «3 . La liaison ici n'est pas intense.

Biplot

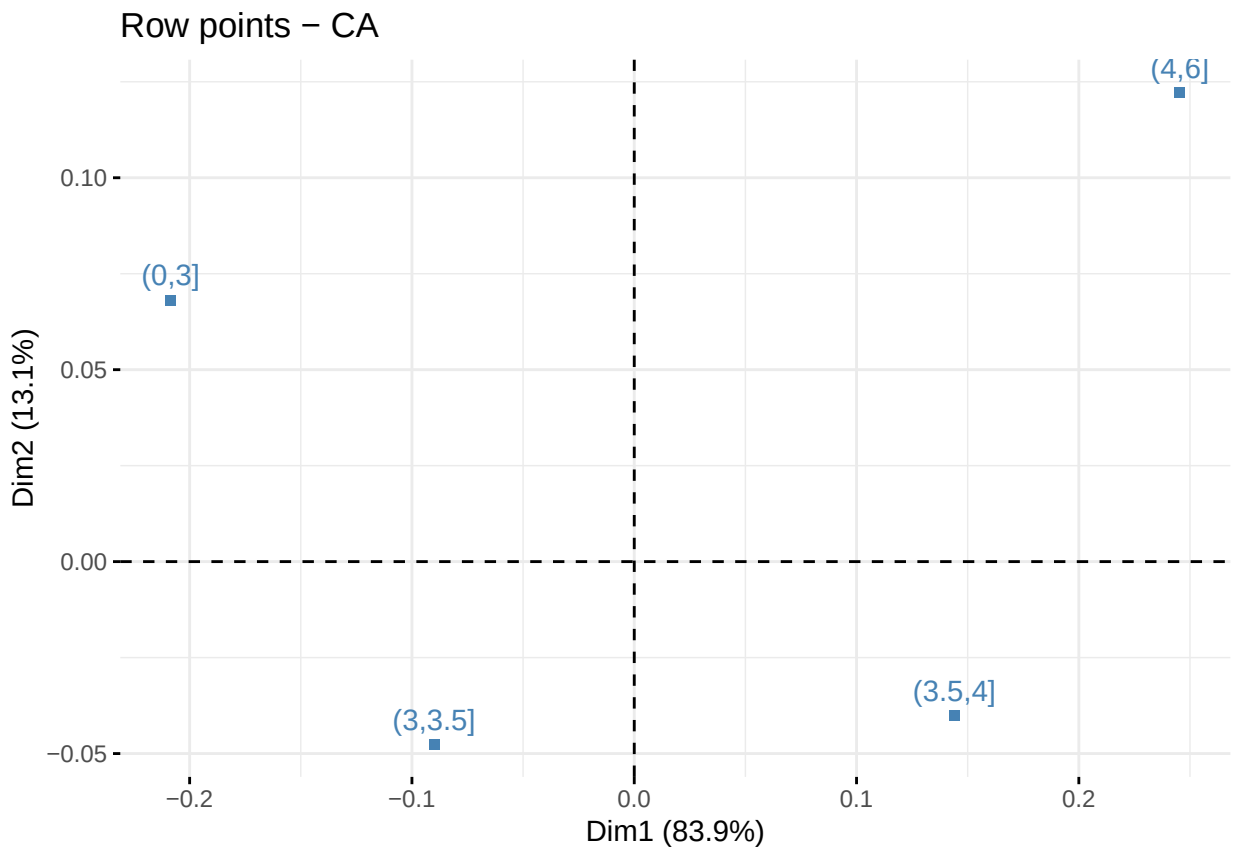


On remarque qu'il y'a une forte attractivié entre l'intervalle (0,3] et l'intervalle [0,50]. Donc ces deux profils colonne et ligne s'associent le plus.

On ne peut interpreter cette proximité si ce n'est dire que les profils colonne sont du coté des profils lignes auxquels ils s'associent le plus.

Graphique des points lignes

```
##               Dim 1       Dim 2       Dim 3
## (0,3]   -0.20849351  0.06813673 -0.02728474
## (3,3.5] -0.08998721 -0.04758878  0.03212277
## (3.5,4]  0.14409105 -0.04015397 -0.02881327
## (4,6]    0.24547526  0.12225220  0.03791108
```



Ici nous n'avons pas de points regroupés donc il n'y a pas de profils-lignes aux comportements similaires.

Les intervalles (0,3] et (4,6] sont correlés negativement car opposés par rapport à l'origine du graphique (quadrants opposés).

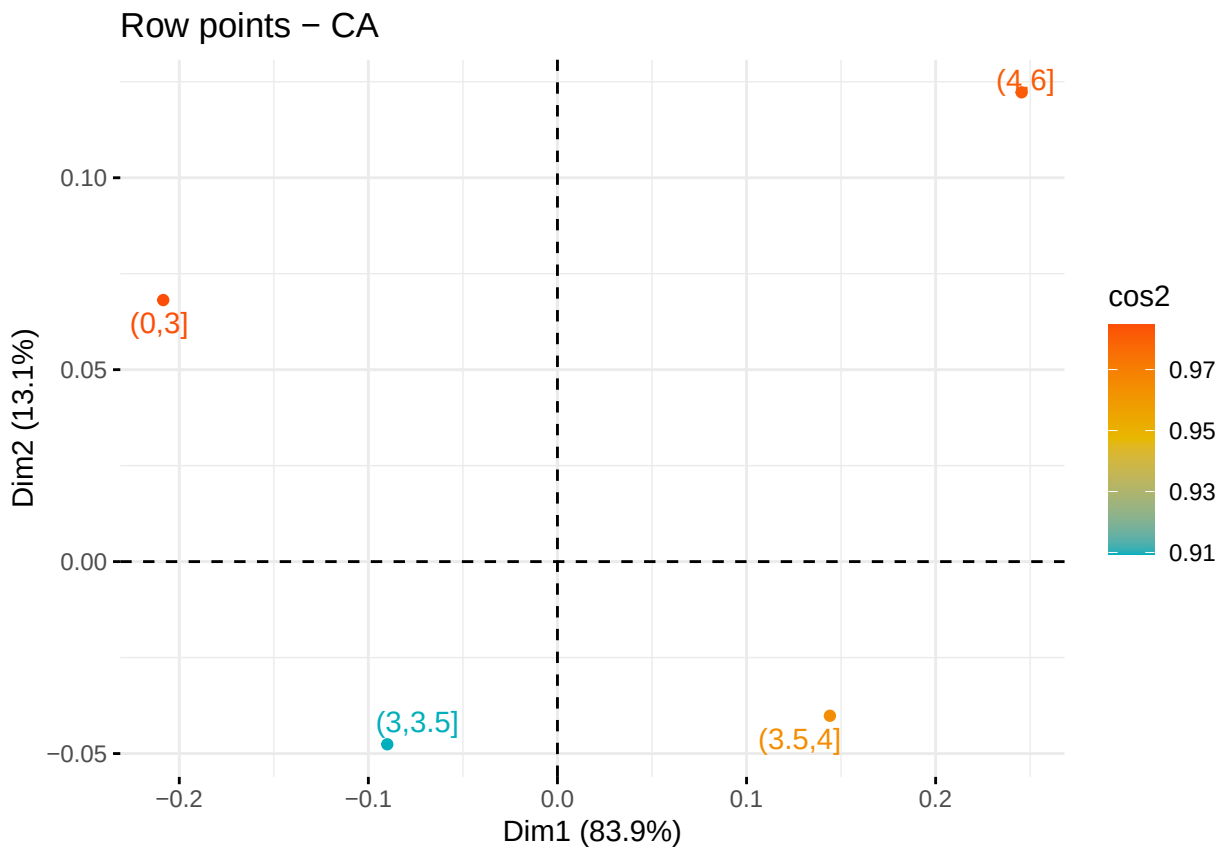
Meme chose pour les intervalles (3.5,4] et(4,6].

La distance (3,3.5] à l'origine étant petite, nous pouvons déjà le considérer pas très bien représenté sur l'axe 1.

Qualité de représentation des lignes

```
##             Dim 1      Dim 2      Dim 3
## (0,3]   0.8897371 0.09502529 0.01523759
## (3,3.5] 0.7106821 0.19875724 0.09056068
## (3.5,4] 0.8947395 0.06948316 0.03577731
## (4,6]   0.7862390 0.19500802 0.01875303
```

Visualisation rapide



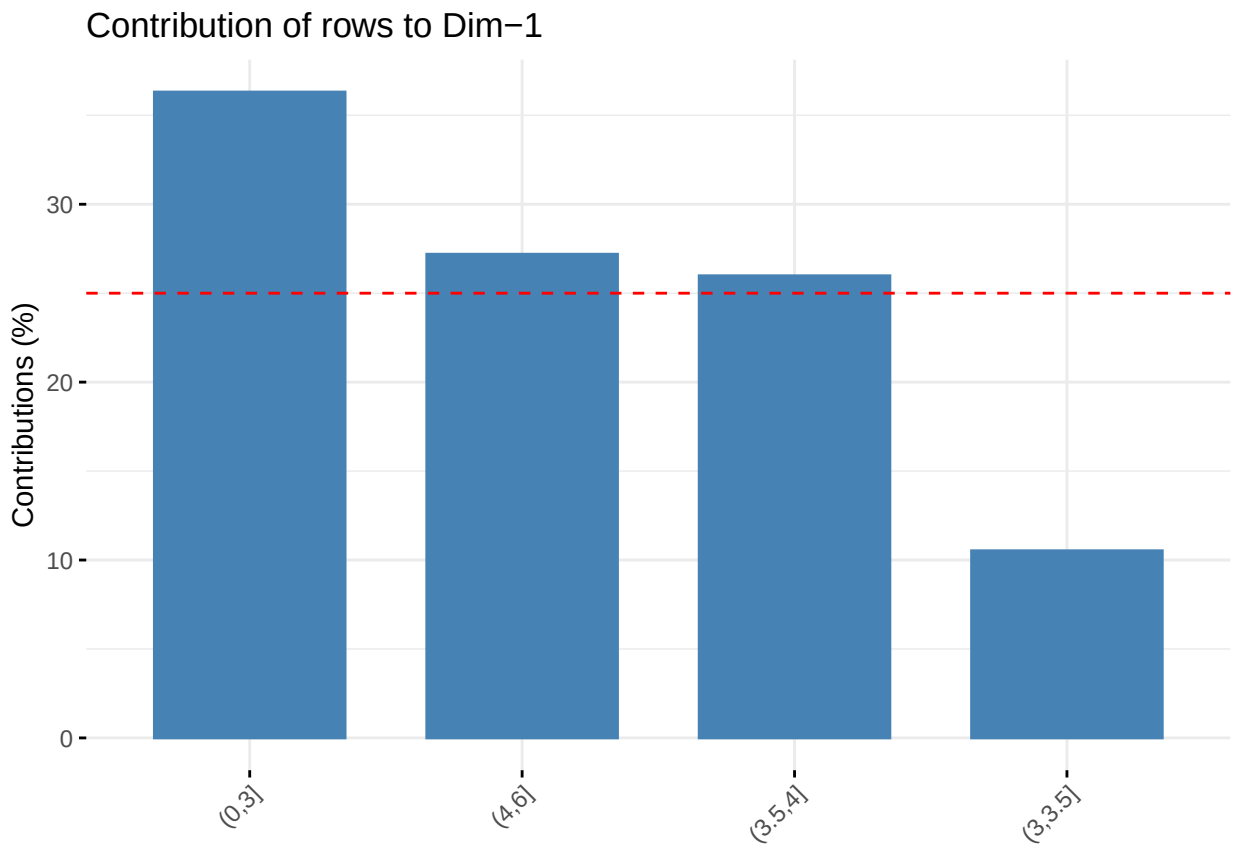
Les intervalles sont tous bien représentés sur l'axe 1 sauf l'intervalle (3,3.5] comme on l'avait anticipé.

Contribution des lignes à la dimension 1

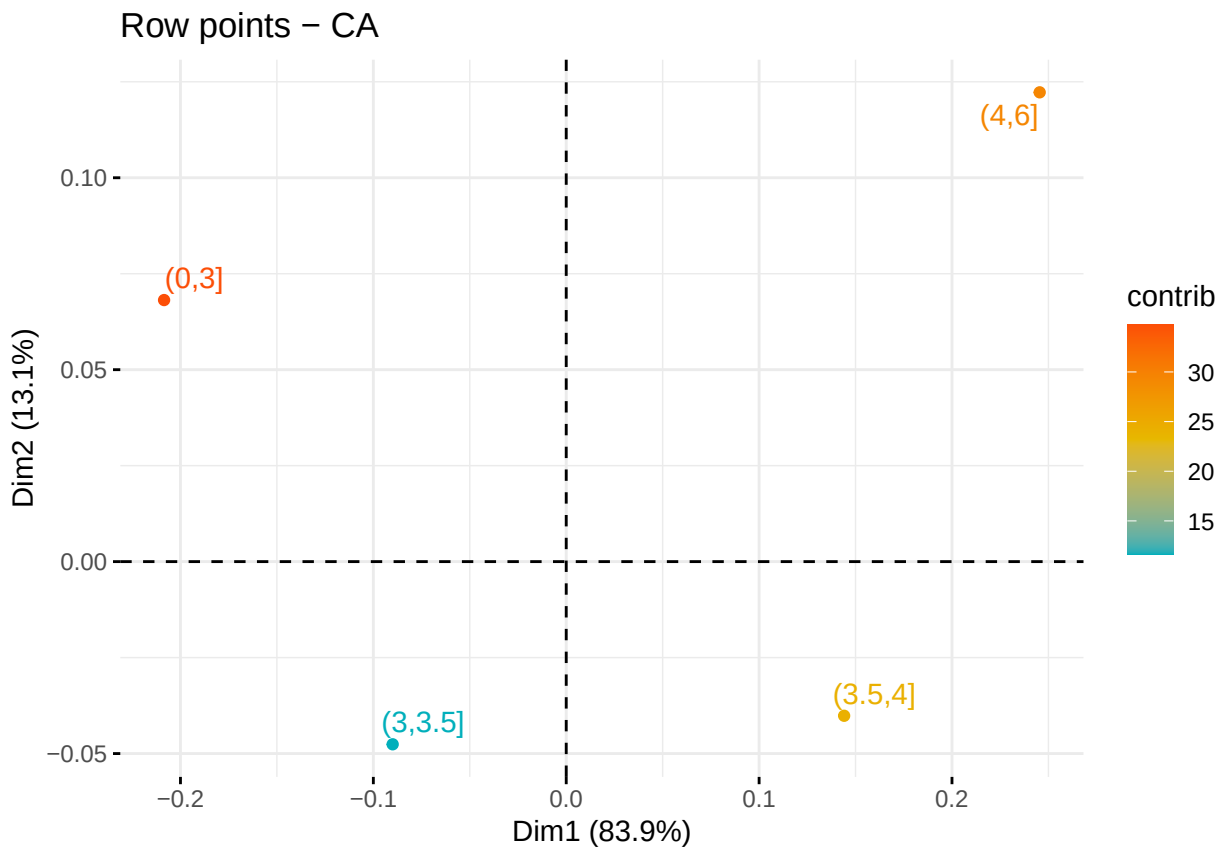
Il y'a 4 catégories de poids de bébé. Donc le seuil envisagé est donc à 25% de contribution à l'inertie du premier axe On peut ensuite obtenir les indivus dont la contribution est supérieure à la contribution moyenne : Les contributions à l'axe 1 sont (en pourcentage) :

	x
(0,3]	36.30850
(3,3.5]	10.52132
(3.5,4]	25.97966
(4,6]	27.19051

On peut mieux le voir ici :



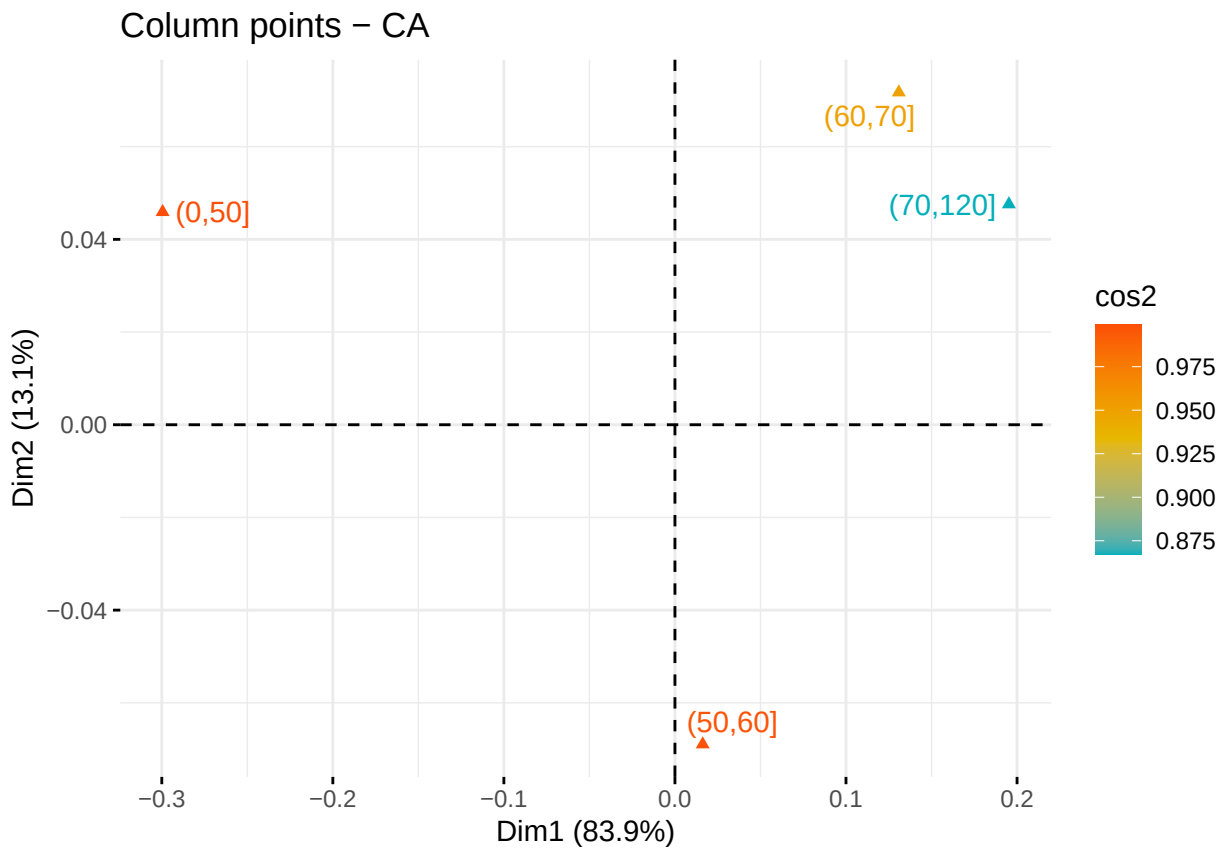
Les intervalles (0,3] , (4,6] et (3.5,4] ont une contribution supérieure à la moyenne (La droite en pointillés rouges, sur le graphique ci-dessus, indique la valeur moyenne attendue)



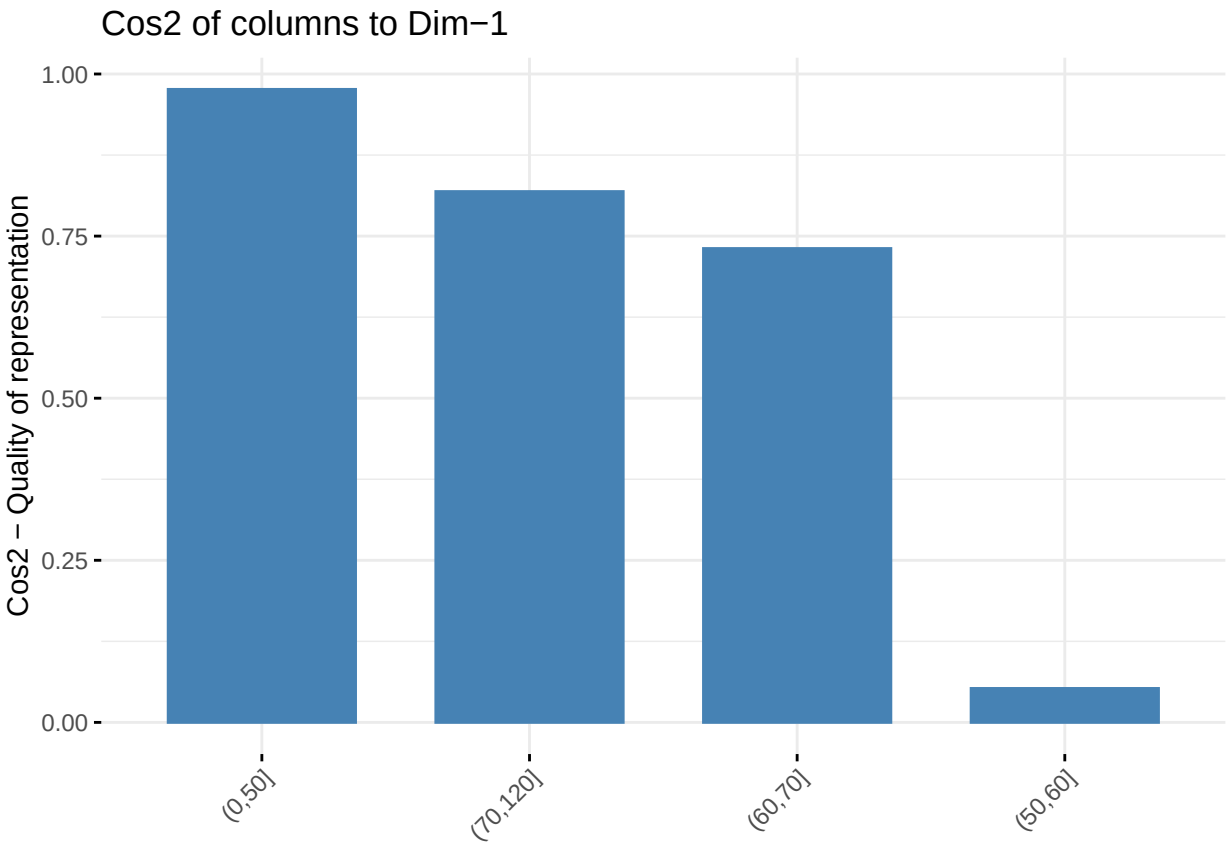
Il est évident que les catégories (4,6] et (3.5,4] ont une contribution importante au pôle positif de la première dimension, tandis que la catégorie (0,3] une contribution majeure au pôle négatif de la première dimension.

En d'autres termes, la dimension 1 est principalement définie par l'opposition (4,6] et (3.5,4] (pôle positif) avec [0,3] (pôle positif).

Graphique des colonnes

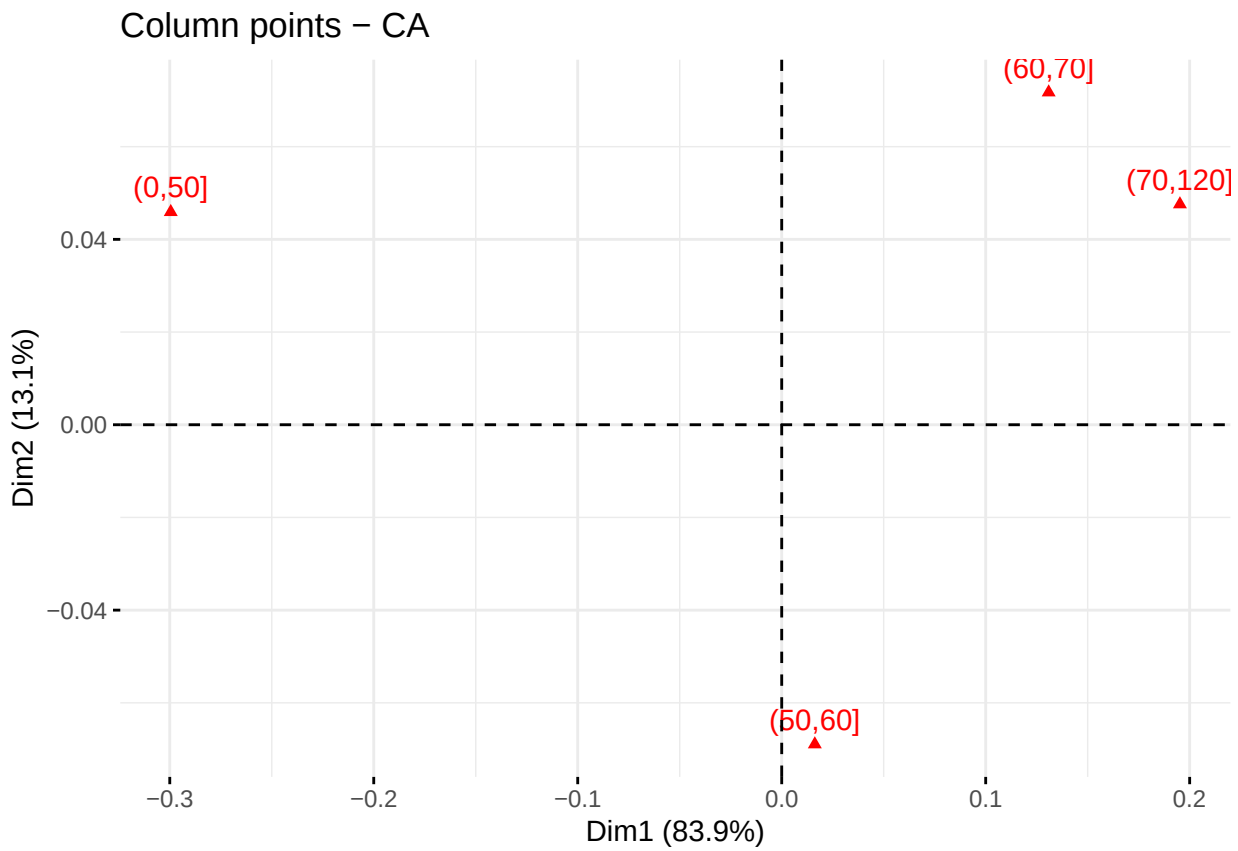


Tous les profils-colonnes sont très bien représentés sur l'axe (1,2). Essayons de voir leur qualité de représentation sur l'axe 1 d'étude.



Le profil (0,50] est très bien représenté sur l'axe 1. Les profils (70,120] et (60,70] aussi dans une moindre mesure.

Seul le profil (50,60] n'est pas bien représenté. Donc son interprétation doit etre faite avec prudence.

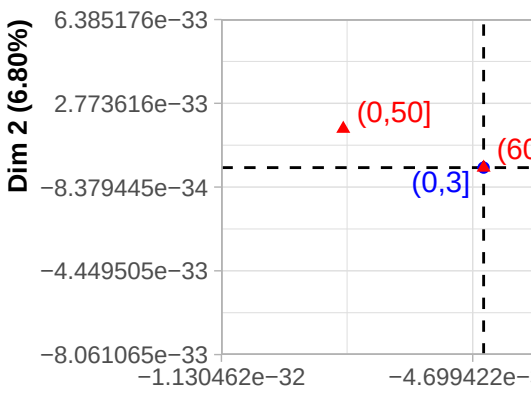


On peut voir que les profils-colonnes (60,70] et (70,120] sont très proches (avec bonne qualité de représentation) et que ce petit groupe est très eloigné du profil-colonne (0,50].

On peut dire que (60,70] et (70,120] ont des comportements similaires. Les femmes de cet intervalles donnent naissance à des enfants aux poids à peu près similaires. On l'a vu grace au biplot.

Alors que les enfants issus de parents dont les poids sont dans l'intervalle (0,50] ont des enfants dont les poids sont dans l'intervalle [0,3].

On ne peut pas s'exprimer clairement sur le comportement des personnes sur l'intervalle (50,60] dans la mesure ou il mal représenté sur l'axe 1.



***Conclusion* : Essayons une afc sous l'hypothèse H0 d'indépendance**

```
## **Results of the Correspondence Analysis (CA)**
## The row variable has  4  categories; the column variable has 4 categories
## The chi square of independence between the two variables is equal to 2.404243e-29 (p-value =  1 ).
## *The results are available in the following objects:
##
##    name              description
## 1  "$eig"            "eigenvalues"
## 2  "$col"            "results for the columns"
## 3  "$col$coord"      "coord. for the columns"
## 4  "$col$cos2"       "cos2 for the columns"
## 5  "$col$contrib"    "contributions of the columns"
## 6  "$row"            "results for the rows"
## 7  "$row$coord"      "coord. for the rows"
## 8  "$row$cos2"       "cos2 for the rows"
## 9  "$row$contrib"    "contributions of the rows"
## 10 "$call"           "summary called parameters"
## 11 "$call$marge.col" "weights of the columns"
## 12 "$call$marge.row" "weights of the rows"
```

La représentation peut porter à confusion mais on peut voir que les points sont tous quasiments confondus avec le centre de gravité (les ordres de grandeur sont en 0.0000 quand on regarde l'echelle).

Ceci n'apporte rien à notre analyse si ce n'est que nous avons bien fait un test du Khi2 avant sinon l'AFC n'est pas utile.