Statistiques ⋈ Forensiques

Exploration de données morphométriques pour portrait robot ADN : quelques exemples de graphiques en 3D

Jean R. Lobry

22 mars 2017

Table des matières

1 Origine des données

Les données sont extraites d'un article [3] du consortium académique *Visigen* dirigé par le Professeur Tim Spector au Royaume-Uni et le Professeur Manfred Kayser aux Pays-Bas. Les données brutes sont librement disponibles dans le fichier `Table_S6.xlsx` de la version en ligne de l'article. On ne s'intéresse ici qu'aux données morphométriques et à quelques variables auxiliaires disponibles dans la première feuille de ce fichier de type tableur. Les importer sous

```
rmn <- read.table(url("http://pbil.univ-lyon1.fr/R/donnees/rmn3D.csv"),
                  header = TRUE, sep = '\t', dec = ",")
class(rmn)
```

```
[1] "data.frame"
```

```
dim(rmn)
```

```
[1] 5388   31
```

```
names(rmn)
```

```
 [1] "id"          "sex"         "age"         "cohort"      "ZygionRx"    "ZygionRy"
 [7] "ZygionRz"    "ZygionLx"    "ZygionLy"    "ZygionLz"    "EyeballRx"   "EyeballRy"
[13] "EyeballRz"   "EyeballLx"   "EyeballLy"   "EyeballLz"   "AlareRx"     "AlareRy"
[19] "AlareRz"     "AlareLx"     "AlareLy"     "AlareLz"     "Nasionx"     "Nasiony"
[25] "Nasionz"     "Pronasalex"  "Pronasaley"  "Pronasalez"  "Subnasalex"  "Subnasaley"
[31] "Subnasalez"
```

2 Nature de variables

2.1 Les variables auxiliaires

2.1.1 La variable `id`

La première colonne contient classiquement une clef d'identification des individus dont on peut vérifier facilement qu'elle correspond à leur rang dans les données :

```
identical(rmn$id, 1:nrow(rmn))
```
```
[1] TRUE
```

Elle ne présente pas d'intérêt direct pour nous ici mais est très importante pour assurer la traçabilité des données.

2.1.2 La variable `sex`

La deuxième colonne donne le sexe des individus :

```
summary(rmn$sex)
```
```
Female   Male   NA's
  2076   1684   1628
```

On note ici la présence de données manquantes. Ceci est expliqué dans la feuille `Notes` du fichier d'origine `Table_S6.xlsx` : *For sex and age information in the SHIP cohort, please contact Reiner Biffar*. Ces données manquantes ne sont pas la conséquence d'un défaut de mesure ou d'un problème de perte de l'information mais du choix des propriétaires des données de contrôler la confidentialité de l'information. C'est le premier indice que nous rencontrons de l'hétérogénéité des données.

2.1.3 La variable `age`

La troisième colonne donne l'age des individus.

```
summary(rmn$age)
```
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  20.00   50.57   57.45   54.35   61.90   92.86    1628
```

On note ici la présence de valeurs manquantes pour la même raison que pour les données sur le sexe pour l'échantillon `SHIP`. La valeur maximum 92.86 interpelle avec ses 4 chiffres significatifs. Se pourrait-il que la précision de sur l'âge diffère selon les échantillons ?

Table 1. Characteristics of the study subjects (N = 9,823).

Cohort	Country	Individual	For	Image	N	Male%	Age	±	sd
RS1	Netherlands	Unrelated	discovery	3D head MRI	2,470	46.4	59.7	±	8.0
RS2	Netherlands	Unrelated	discovery	3D head MRI	745	43.1	59.0	±	7.9
QTIMS	Australia	Twins	discovery	3D head MRI	545	39.6	23.7	±	2.3
SHIP	Germany	Unrelated	discovery	3D head MRI	797	47.3	46.0	±	12.8
SHIP-TREND	Germany	Unrelated	discovery	3D head MRI	831	44.8	50.4	±	13.6
SYS	Canada	Siblings	replication	3D head MRI	568	48.1	15.1	±	1.9
TwinsUK	UK	Twins	replication	2D portrait photo	1,530	9.5	58.4	±	12.9
BLTS	Australia	Twins	replication	2D portrait photo	2,337	47.8	23.6	±	4.6

doi:10.1371/journal.pgen.1002932.t001

FIGURE 1 – Copie d'écran de la table 1 de l'article [3] à l'origine des données utilisées ici donnant quelques statistiques descriptives utiles pour vérifier la cohérence des données des échantillons utilisés pour cette compilation. La colonne **Cohort** donne le le nom de code de l'échantillon dont l'individu est issu. Les trois dernières (*viz.* `SYS`, `TwinsUK`, `BLTS`) ne sont pas disponibles ici, ce qui explique la différence entre le $n = 9,823$ de la table et le $n = 5,388$ du jeu de données. Les échantillons `SHIP` et `SHIP-TREND` sont agrégés dans le jeu de données ce qui explique qu'il n'y ait au final que 4 échantillons documentés (*viz.* `RS1`, `RS2`, `QTIMS`, `SHIP`).

2.1.4 La variable `cohort`

La quatrième colonne est très importante puisqu'elle nous donne l'origine de l'échantillon pour les individus.

```
summary(rmn$cohort)
```

```
QTIMS   RS1   RS2  SHIP
  545  2470   745  1628
```

Exercice. En une ligne, retrouvez les statistiques sur la sexe-ratio par groupe données dans la figure 1 page 3.

```
   QTIMS      RS1      RS2     SHIP
39.63303 46.43725 43.08725       NA
```

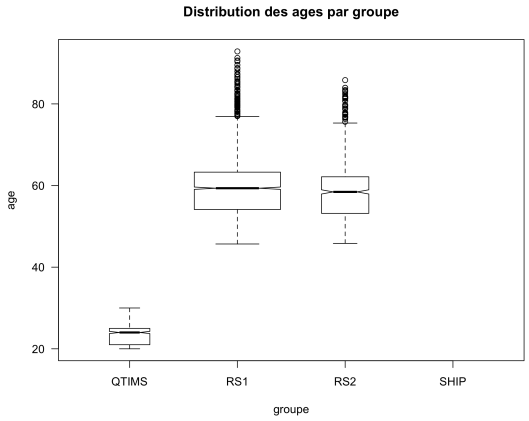
Exercice. En une ligne, retrouvez les statistiques sur l'age par groupe données dans la figure 1 page 3.

```
$QTIMS
[1] 23.691743  2.319282
$RS1
[1] 59.703425  7.968676

$RS2
[1] 59.046658  7.891293

$SHIP
[1] NA NA
```

Exercice. Représenter la distribution des ages en fonction des groupes.

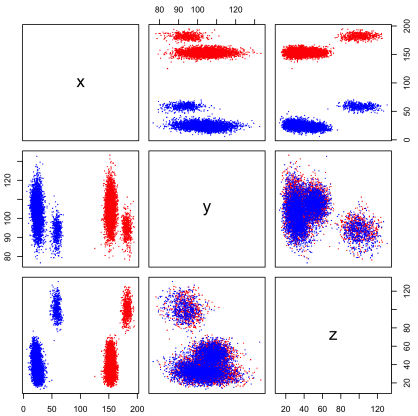


2.2 Les variables morphométriques

2.2.1 Les variables `ZygionR[x|y|z]` et `ZygionL[x|y|z]`

Ces variables contiennent les coordonnées spatiales des extrémités de l'os zygomatique, comme représenté sur la figure 2 page 5. Représenter graphiquement les données en utilisant par convention du rouge pour les points du zygion gauche et du bleu pour les points du zygion droit.

```
ZygionR <- rmn[ , 5:7]
nxyz <- c("x", "y", "z")
names(ZygionR) <- nxyz
ZygionL <- rmn[ , 8:10]
names(ZygionL) <- nxyz
Zygion2 <- rbind(ZygionR, ZygionL)
colRL <- rep(c("red", "blue"), each = nrow(ZygionR))
plot(Zygion2, col = colRL, pch = '.')
```



Le paquet `scatterplot3d` [2] permet de faire une représentation statique en perspective des données :

```
library(scatterplot3d)
scatterplot3d(Zygion2, pch = '.', color = colRL, main = "Position des zygions")
```

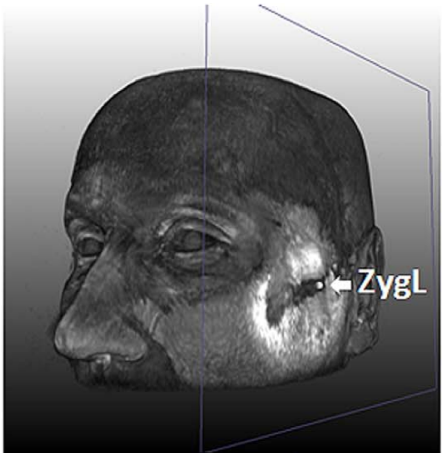
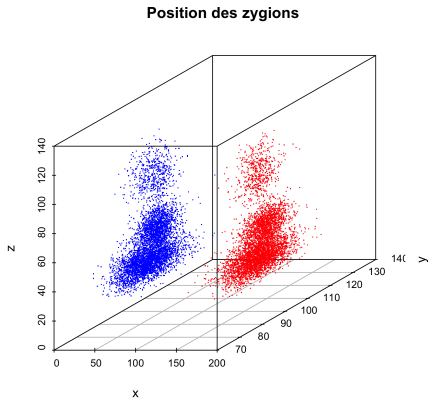
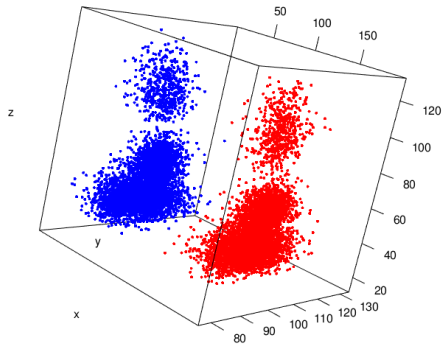


FIGURE 2 – Copie d'écran de la figure 1A de [3]. Il s'agit d'une image obtenue par résonance magnétique nucléaire de la tête du Professeur Manfred Kayser. Un plan de coupe a été utilisé pour enlever une partie de la tête et mettre en évidence le point d'intérêt **ZygL**, pour *left zygion*, le point le plus distal de l'os zygomatique. Notez que la droite et la gauche sont celles du point de vue de l'individu. Le point symétrique **ZygR** n'est pas visible sur cette figure.



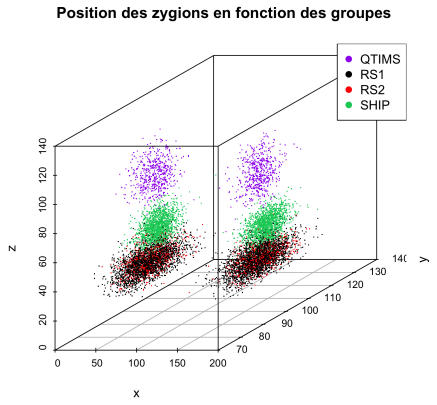
Le paquet `rgl` [1] permet de faire une représentation dynamique en perspective des données :

```
library(rgl)
plot3d(Zygion2, col = colRL)
```



Représenter les groupes en variable supplémentaire pour montrer qu'ils sont déterminants pour la variabilité des données :

```
paletteG <- c("purple", "black", "red", "springgreen3")
colG <- paletteG[rmn$cohort]
scatterplot3d(Zygion2, pch = '.', color = c(colG, colG),
              main = "Position des zygions en fonction des groupes")
legend("topright", legend = levels(rmn$cohort), col = paletteG, pch = 19)
```

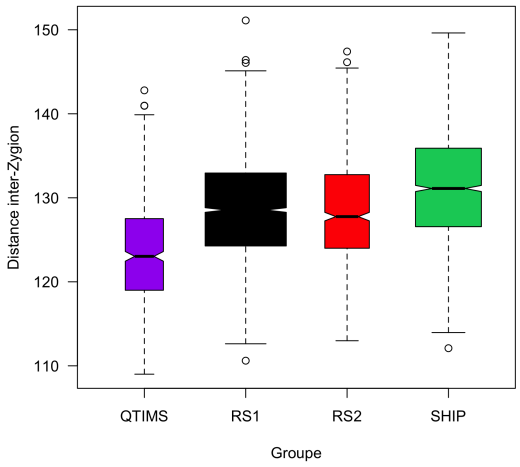


Pour pallier cette variabilité on décide d'utiliser la distance inter-zygion comme variable dérivée. On utilise la distance euclidienne qui pour deux points A et B de $\mathbb{R}^3$ de coordonnées (A_x, A_y, A_z) et (B_x, B_y, B_z), respectivement, est définie par :

$$d(A,B) = \sqrt{(A_x - B_x)^2 + (A_y - B_y)^2 + (A_z - B_z)^2}$$

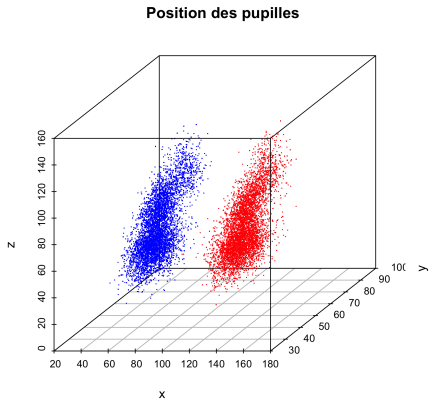
Calculer la distance inter-zygion et la représenter en fonction des groupes :

```
dZyg <- sqrt(rowSums((ZygionL - ZygionR)^2))
boxplot(dZyg~rmn$cohort, notch = T, col = paletteG, las = 1,
        xlab = "Groupe", ylab = "Distance inter-Zygion",  varwidth = T)
```



2.2.2 Les variables `EyeballR[x|y|z]` et `EyeballL[x|y|z]`

Ces variables contiennent les coordonnées spatiales des pupilles des yeux, comme représenté sur la figure 3 page 8. Représenter graphiquement les données en utilisant par convention du rouge pour les pupilles gauches et du bleu pour les pupilles droites.



Mettre en évidence l'effet des groupes :

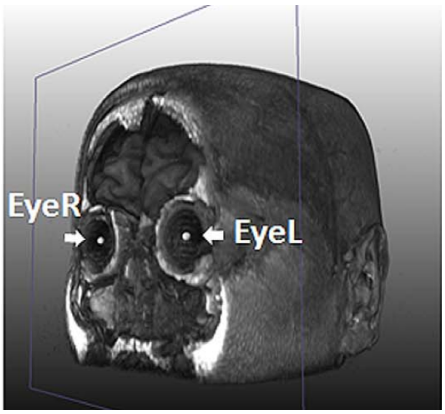
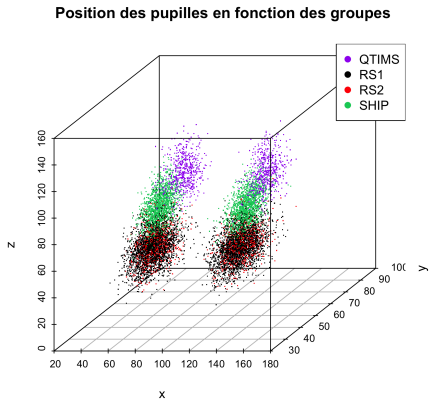
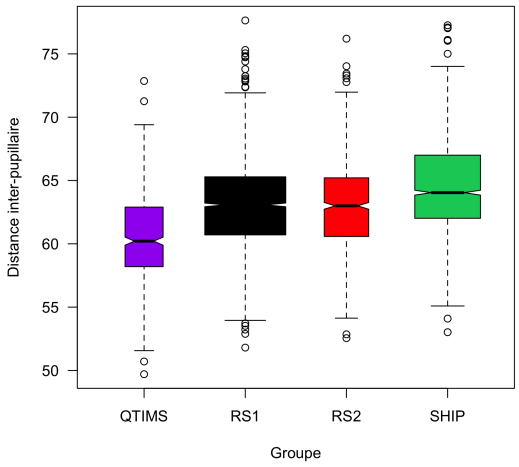


FIGURE 3 – Copie d'écran de la figure 1B de [3]. Un plan de coupe a été utilisé pour enlever une partie de la face et mettre en évidence la position des pupilles **EyeR** et **EyeL**.

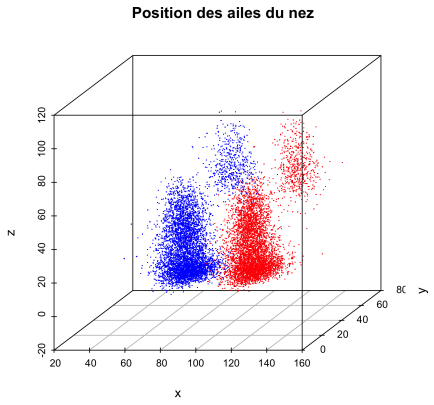


Calculer la distance inter-pupillaires et la représenter en fonction des groupes :



2.2.3 Les variables `AlareR[x|y|z]` et `AlareL[x|y|z]`

Ces variables contiennent les coordonnées spatiales des ailes du nez, comme représenté sur la figure 4 page 10. Représenter graphiquement les données.



Mettre en évidence l'effet des groupes :

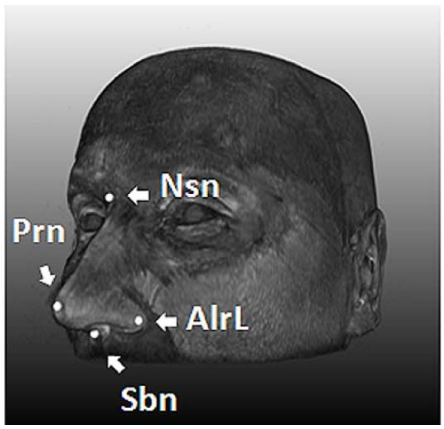
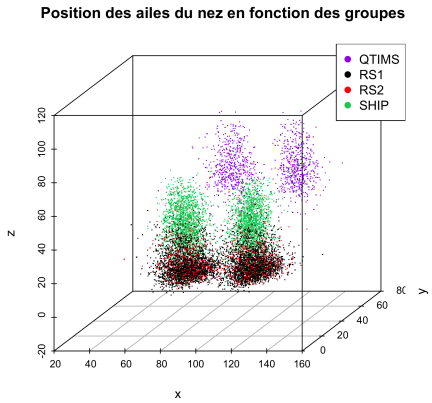
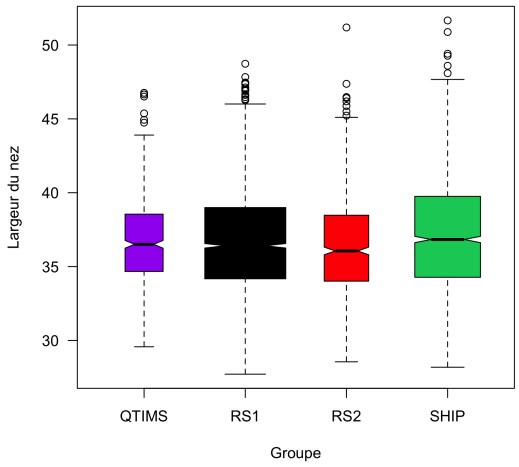


FIGURE 4 – Copie d'écran de la figure 1B de [3]. Les points caractéristiques du nez. Les ailes droites **AlareR** et gauches du nez **AlareL**. La partie supérieure de l'os nasal, *nasion*, **Nsn**. Le bout du nez, *pronasale*, **Prn**. La base du nez, *subnasale*, **Sbn**.

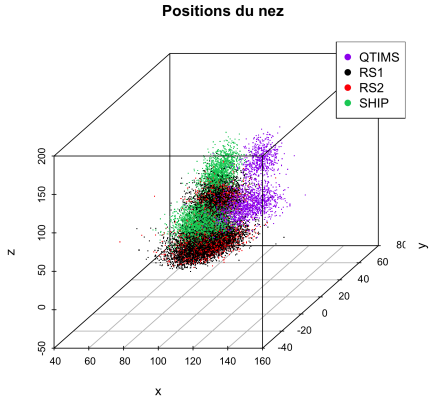


Calculer `dAlare` la largeur du nez et la représenter en fonction des groupes :

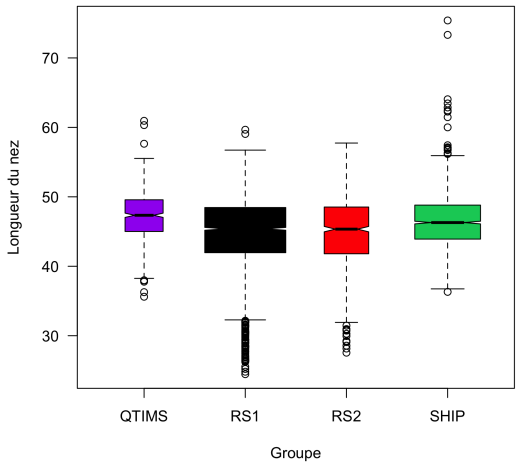


2.2.4 Les variables `Nasion[x|y|z]`, `Pronasale[x|y|z]` et `Subnasale[x|y|z]`

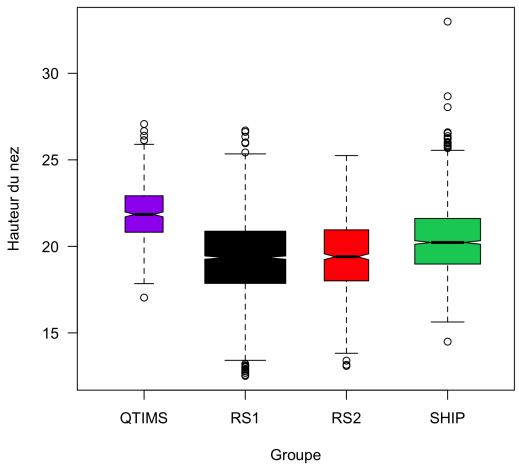
Ces variables contiennent les coordonnées spatiales du nez, comme représenté sur la figure 4 page 10. Représenter graphiquement les données en fonction des groupes.



Calculer `lNez` la longueur du nez et la représenter en fonction des groupes :



Calculer `hNez` la hauteur du nez et la représenter en fonction des groupes :



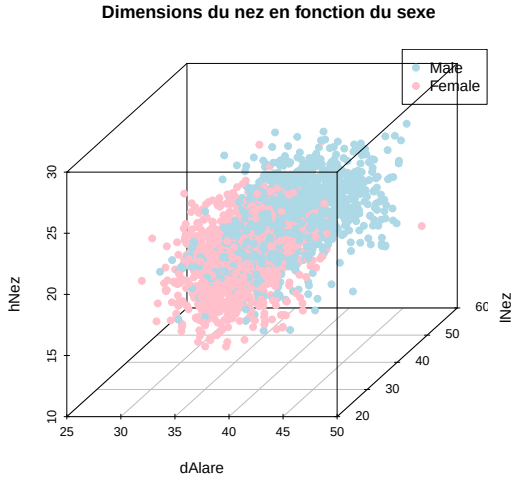
3 Sexe et morphologie du nez

On aimerait savoir s'il y a un effet du sexe sur la morphologie du nez. Pour éliminer l'effet du groupe d'origine on décide de ne conserver que celui qui est le mieux documenté, `RS1`. Construire le jeu de données.

```
rmn$dAlare <- dAlare
rmn$lNez <- lNez
rmn$hNez <- hNez
sexetnez <- rmn[rmn$cohort == "RS1", c("sex", "dAlare", "lNez", "hNez")]
dim(sexetnez)
```

```
[1] 2470    4
```

Représenter les données en mettant en évidence le sexe des individus :

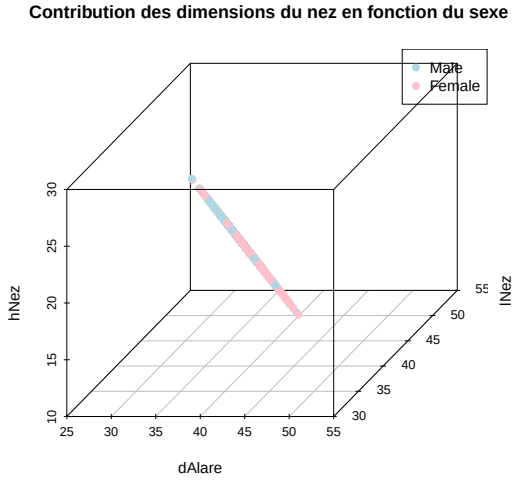


On voit ici essentiellement un effet taille : les dimensions des nez des mâles sont supérieures à celles des femelles. Pour neutraliser l'effet taille on exprime les dimensions du nez en contribution à la somme des dimensions :

```
psexetnez <- sexetnez
psexetnez[, 2:4] <- 100*psexetnez[, 2:4]/rowSums(psexetnez[, 2:4])
head(psexetnez)
```

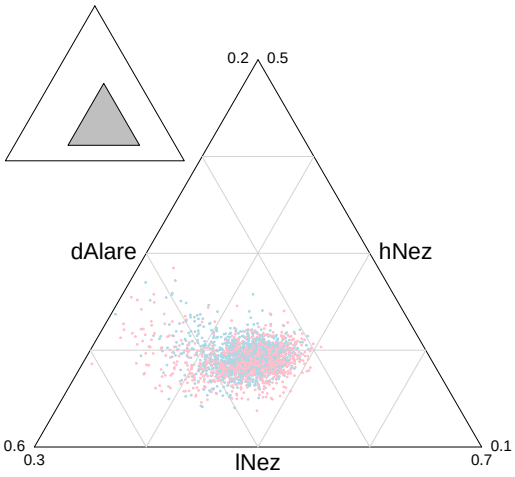
```
     sex   dAlare     lNez     hNez
1   Male 37.52297 45.05680 17.42023
2   Male 34.37477 44.78403 20.84120
3 Female 37.48699 45.16436 17.34865
4   Male 35.41548 44.61255 19.97197
5   Male 36.48861 44.28825 19.22314
6 Female 32.86903 48.54243 18.58854
```

Représenter les données en mettant en évidence le sexe des individus :



Utiliser la fonction `plot3d()` du paquet `rgl` pour comprendre la structure des données. Dans ce cas de figure on peut utiliser une représentation triangulaire :

```
library(ade4)
coord <- triangle.plot(psexetnez[, 2:4], cpoint = 0)
points(coord, col = colSex, pch = 19, cex = 0.25)
```



Exercice. Testez s'il y a un effet significatif du sexe des individus sur les trois variables de `psexetnez`.

Références

[1] Daniel Adler and Duncan Murdoch. *rgl : 3D visualization device system (OpenGL)*, 2014. R package version 0.93.996.

[2] Uwe Ligges and Martin Mächler. Scatterplot3d - an r package for visualizing multivariate data. *Journal of Statistical Software*, 8(11) :1–20, 2003.

[3] F. Liu, F. van der Lijn, C. Schurmann, G. Zhu, M.M. Chakravarty, P.G. Hysi, A. Wollstein, O. Lao, M. de Bruijne, M.A. Ikram, van der Lugt A., F. Rivadeneira, A.G. Uitterlinden, A. Hofman, W.J. Niessen, G. Homuth, G. de Zubicaray, K.L. McMahon, P.M. Thompson, A. Daboul, R. Puls, K. Hegenscheid, L. Bevan, Z. Pausova, S.E. Medland, G.W. Montgomery, M.J. Wright, C. Wicking, S. Boehringer, T.D. Spector, T. Paus, N.G. Martin, R. Biffar, and M. Kayser. A genome-wide association study identifies five loci influencing facial morphology in europeans. *PLOS Genetics*, 8 :e1002932, 2012.