Exercices statistiques pour le cours «Cardiologie préventive: facteurs de risques cardiovasculaire, style de vie ou génétique ? » du 29 janvier 2024

**** Travail à rendre jusqu'au 4 février: un fichier .pdf contenant différents graphiques** : à renvoyer à M Dobretz, MS au kevin.dobretz@unige.ch ***

Toute la démarche est décrite ici, c'est une suite logique qui nécessite une simple exécution.

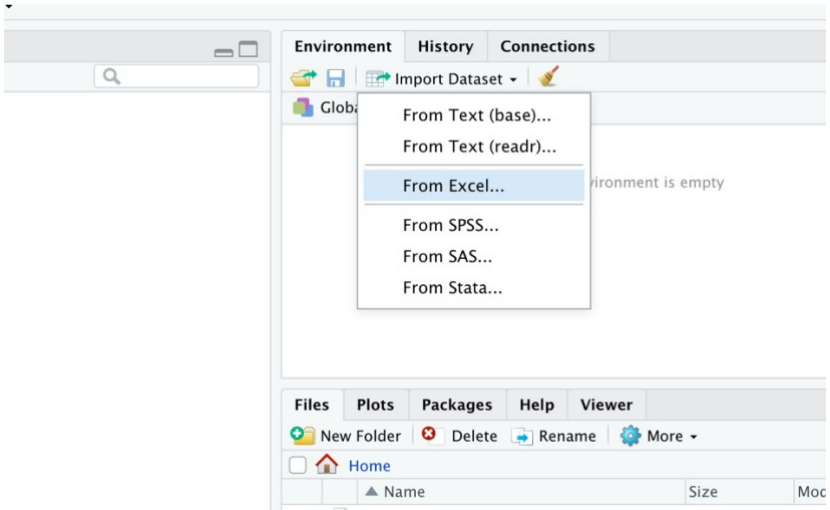
1. Télécharger les autres fichiers du cours sur www.risquecardiovasculaire.ch , à savoir : « Préparation.pdf » et « data.xls ».
2. Installer R et R-studio (instructions dans « Préparation.pdf »)
3. Dans R-studio, créez une nouvelle page de script :

   Dans le menu File en haut, faites New file > R Script.

   Ce script servira de page blanche pour écrire vos bouts de code.
4. Activez les packages readxl et ggplot2 avec la fonction library :

   library(readxl)

   library(ggplot2)
5. Importer les données du fichier dans R-studio en utilisant la commande « import dataset » puis « From Excel » (en haut à droite).



Dans le menu qui s'ouvre, sélectionnez le fichier "data.xls" dans le dossier où vous l'avez sauvé en cliquant sur « Browse ». Vous verrez ensuite une preview de votre fichier, cliquez sur « Import » (en bas à droite).

Les données ont été sauvées sous le nom « data » dans la mémoire de votre environnement R visible en haut à droite sur R-studio.

Retrouvez-le et cliquez dessus pour en observer le contenu, Il y a plusieurs colonnes de données :

- ID (faux numéro d'identification)
- l'âge en années, le LDL sanguin (en mmol/L)
- la tension artérielle systolique (TAS, en mmHg)
- le génotype pour deux variants génlétiques rs12345 G>T et rs6789 A>G. indiqué avec le dosage, 0,1 ou 2, c'est le nombre d'allèles porteurs du variant.

  0 = aucun allèle

  1= hétérozygote

  2= homozygote

5. Refaites un histogramme pour la tension artérielle, comme celui montré en cours avec la fonction hist().

Rappel, écrire : hist(data$LDL)

Dans la console (en bas à gauche) et appuyer « entrée » vous fera un histogramme de base pour le LDL , sans titre ni noms pour les axes.

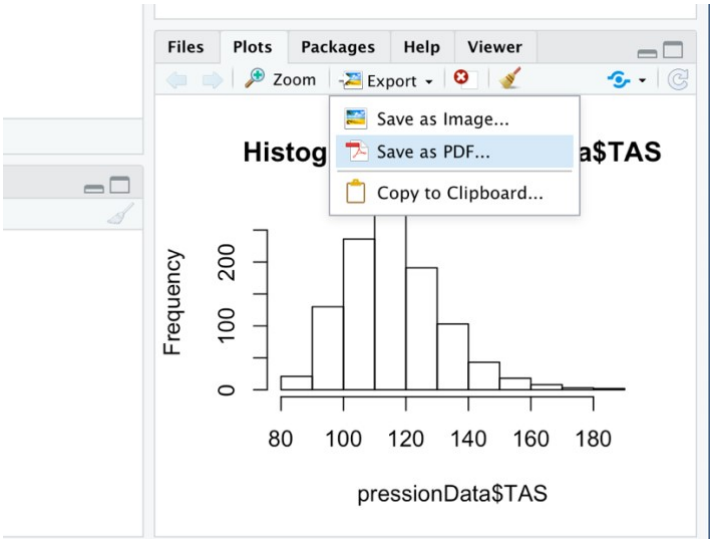
Et écrire :

hist(data$LDL, main= "mon titre", xlab= "nom axe des x", ylab= "nom axe des y")

Vous fera un histogramme avec des titres et noms (de votre choix).

Exercice : Refaites un histogramme pour la tension artérielle.

Vous pouvez soit l'exporter en pdf et l'envoyer par courriel, soit plutôt avancer à la question optionnelle 6.



6. ***Optionnel* : Fonctions graphiques de base, plot et boxplot**

Pour la représentation graphique de deux variables, on peut utiliser la fonction plot() et augmenter les détails du graphique avec des options ou des commandes supplémentaires.

Copiez-collez et validez l'entièreté de la commande suivante:

```
x= data$age y=
data$LDL
model=lm(y~x)
plot(x,y)
abline(model, col="red")
```

Et vous aurez un nuage de points avec la droite de régression linéaire, vous pouvez ajouter des noms et des titres en recopiant les options utilisées pour l'histogramme (main= "mon titre" etc…).
Vous pouvez voir qu'à la 3ème ligne on crée l'objet « model », ce sont les solutions pour la régression linéaire exprimant la tension par rapport à l'âge. Vous pouvez explorer ces coefficients avec la fonction summary() (pente et ordonnée de la droite de régression).

```
summary(model)
```

Pour le génotype, il n'y a que trois valeurs possibles par variant, par conséquent la représentation par nuage de points et droite de régression n'est pas parlante.

Faites plutôt un boxplot avec la fonction suivante :

```
x = data$rs6789_A_G y=
data$TAS
```

```
boxplot(y~x  )
```

Une fois de plus vous pouvez éditer les options main=, xlab= pour les titre et labels :

Option supplémentaire : Réutilisez ces lignes de code en changeant les variables x et y pour exposer celles qui semblent corrélées entre elles. Et si vous voulez vous arrêter là, exportez un ou deux pdf pour le montrer. Si vous voulez continuer avancez vers la question optionnelle 7.

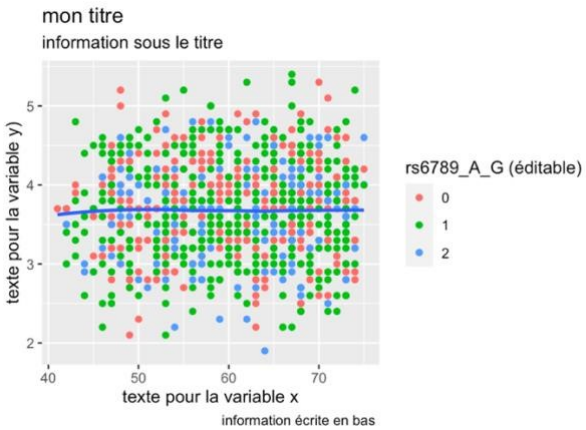
7. Option supplémentaire : Graphique avancé avec ggplot2

Sélectionnez entièrement et copiez-collez d'un coup le code suivant dans la console:

```
fg<-ggplot(data=data, aes(x=age, y=LDL)) +
    geom_point(aes(col=paste(rs6789_A_G) ), na.rm=TRUE) +
    geom_smooth(method="loess", se=F, na.rm=TRUE) +
    labs(subtitle="information sous le titre",
            y="texte pour la variable y)",
            x="texte pour la variable x",
            title="mon titre",
            caption="information écrite en bas ",
            col="rs6789_A_G (éditable)")
    suppressMessages(print(fg))
```

(il ne s'agit pas d'expliquer cette formule, juste de s'en servir pour modèle, si ça bug c'est peut être que vous n'avez pas fait library(ggplot2) avant d'entrer le code.)

Vous devriez voir un graphique plus coloré apparaître :



Chaque point représente un individu, sa couleur son génotype pour le variant rs12345_G_T, en y le taux de LDL, en x son âge. Il s'agit donc du graphe exprimant le LDL en fonction de l'âge, montrant le nuage de points (colorés pour le génotype) et la courbe de régression.

1. Maintenant changez les titres, légendes et autre texte apparents (tout c'est qui est **vert et entre guillemets** dans la fonction) pour qu'on voie de quelles variables il s'agit (en **bleu** dans la fonction).

2. Ensuite, j'aimerais que vous changiez le **LDL** pour **TAS** dans la formule et recopiez-la dans la console, vous devriez voir un autre graphique exposant la tension en fonction de l'âge.

3. Et enfin j'aimerais que vous changiez **rs6789_A_G** pour **rs12345_G_T** pour changer le variant marqué en couleur.

Exportez-ce graphique là en pdf pour valider cette démonstration.