TP2 : Analyse univariée

— Les comptes rendus TP et commentaires statistiques : fichiers **.rmd et .html** sont à envoyer à **l'enseignant.**
— Ne pas oublier de commenter les résultats numériques et graphiques (comme en TD).

....A partir de la section 3.2 : Variables quantitatives

— Attention ! certains codes ne marche pas en l'état, c'est fait exprès afin de faire réagir, il faudra prendre le temps de diagnostiquer le pourquoi et corriger. C'est la cas par exemple des blocs R
  * Section 3.1.1 : Autres fonctions de descriptions de bases. Expliquer pourquoi ça ne marche pas.
  * Exercice 4 : blocs 1 (Caractéristiques de position) et 2(Caractéristiques de dispersion)
  * En cas de problème, comparer vos codes R avec ceux de vos voisins directs.

L'analyse peut se faire sur la base des 2 approches de programmation..
— Classique ou
— Dans l'univers Tydiverse : https ://www.tidyverse.org/
  * plus sofistiqué qui conduit à des présentations plus "pro"

1 Les données

1.1 Faire une présentation de votre problématiques

Exercice 1. Importation des données sous R

1. Importer le jeu de données
   — Voir TP1 ou suivre les instructions en séance ou....
   — https ://support.rstudio.com/hc/en-us/articles/218611977-Importing-Data-with-RStudio
   — Dans tydiverse –>
   https ://rstudio-education.github.io/tidyverse-cookbook/import.html
   https ://uomresearchit.github.io/r-day-workshop/03-loading-data-into-R/
2. Pour faciliter les échanges avec l'enseignant, on donnera le même nom au tableau de données sous R/Rstudio
   — Nom tableau de données –> "df" ...

1.2 Pré-traitement : nettoyage, gestion des données manquantes, codage des variables,...

En fonction du jeu de données, il faudra procéder à un prétraitement,...

2 Analyses univariées de variables quantitatives

2.1 Premier résumé statistique : informations générales sur tout le tableau de données.

Dans ce qui suit, le tableau dat est générique, il faudra appliquer le code à votre tableau de données.

Exercice 2. Résumé statistique : la fonction summary()

— Résumé statistique de tout le tableau. La fonction **summary()** donne un résumé statistique des variables.

```
summary(df)
```

Dans Tidyverse

— https ://dplyr.tidyverse.org/reference/summarise.html

— https ://cran.r-project.org/web/packages/gtsummary/vignettes/tbl_summary.html

2.1.1 Autres fonctions de descriptions de bases.

— La fonction générique pour les résumés statistiques est ***summary()*** en voici d'autres

— Quantitative : Mean ; mean() / Standard ; deviation sd() / Variance ; var() / Minimum ; min() / Maximum ; maximum() / Median ; median() / Range of values (minimum and maximum) ; range() / Sample quantiles ; quantile() / InterQuartile Range IQR(), table(), et bien d'autres

— Comme les matrices, il est possible d'appliquer ces fonctions à chaque colonne d'un data.frame ici **d**f. (générique)

```
###En utilisant la fonction apply(df,2,...) #l'option 2 renvoie à une opération sur les
colonnes et l'option 1, sur les lignes. Si dat est le tableau de données
apply(df,2,mean)# applique la fonction mean à chaque colonne de df .
apply(df,1,mean)## applique la fonction mean à chaque ligne de df .
```

Voir plus loin pour les applications.

2.2 Variables quantitatives ou catégorielles

Exercice 3. Le tableau de données des variables quantitatives

— Si votre tableau de données comportent des variables quantitatives et qualitatives, faire une séparation en 2 tableaux :

— "quanti" (regroupe les variables quantitatives), "quali" (les quali)..

Appliquer les codes ci-dessous à vos données.

1. `Caractéristiques de position`

```
###En utilisant la fonction apply()
apply(dat,2,mean) # sur un tableau générique nommé dat.
```

Faire de même pour la médiane, les quartiles, les premier et dernier déciles.

```
apply(dat,2,function(x)quantile(x,0.1))#pour le 1er décile
apply(dat,2,function(x)quantile(x,probs=c(0.1,0.9))# le 1er et dernier déciles
apply(dat,2,function(x)quantile(x,probs=seq(0,1,by=0.3))#....
```

2. `Caractéristiques de dispersion`

```
apply(dat,2,sd)#écart-type

apply(dat,2,sd)/apply(dat,2,mean)
```

Donner :

(a) Intervalle entre le 1er et dernier quartile ($[q_1, q_3]$) et commenter le résultat ;

(b) Intervalle entre le 1er et le dernier décile ($[d_1, d_9]$) et commenter le résultat,

Dans Tidyverse
https ://dplyr.tidyverse.org/reference/across.html

3. Graphiques

```
##Les graphiques##
boxplot(df$nom de la variable ,horizontal=T)
title(main="poids de l\'enfant a la naissance")
hist(df$nom de la variable)
hist(df$nom de la variable,probability=TRUE) #version courte hist(penf_n,prob=T)
lines(density(df$nom de la variable),col= "red")
```

— **Dans Tidyverse**...

Il possible de faire des graphiques plus sophistiqués et élégants en utilisant le package ggplot2.
https ://www.r-graph-gallery.com/boxplot.html

— Vivement conseiller !!!!

Exercice 4. Refaire avec au moins 3 autres variables quantitatives et commenter.

3 Analyses univariées de variables qualitatives

3.1 Transformation en objet Factor

Les variables qualitatives sont des objets “factor” pour R. En fonction de l’importation, dans certains cas, les variables quali ne sont pas transformées automatiquement en “factor” (en particulier dans le cas de variable quantative discrète ou codée avec du numérique) dans ce cas... il faut

— Transformer vos variables qualitatives en “factor”

Exemples :
x=c(rep(1,4),rep(3,5),rep(2,3))
x
as.factor(x)
x=c(rep(1,4),rep("a",5),rep(2,3))
x
y=as.factor(x)
y

— **Dans Tidyverse**...
https ://forcats.tidyverse.org/

3.2 Caractéristiques de variables qualitatives

1. Les caractéristiques numériques

   ##Les tableaux des effectifs, fréquences ##
   — summary(df$nom de la variable) ou
   #summary(quali,maxsum=9)
   — table(df$nom de la variable)# donne les effectifs
   — table(df$nom de la variable)/nrow(df) # donne les proportions

2. Des graphiques : diagramme circulaire, en barre

   pie(table(df$nom de la variable))
   title(main="Genre de l\'enfant")
   #Regarder help(title) pour connaître les options.

   barplot(table(df$sexenf))
   title(main="Genre de l\'enfant")
   Modifier les couleurs du diagramme circulaire.
   #voir help(barplot) utiliser l'option col=c("red", "blue",... autant que de modalités)
   barplot(table(df$sexenf),col=c("red", "blue"))

Exercice 5. Refaire l'exercice précédent avec les variables suivantes et commenter les résultats :
— Pour au moins 2 variables qualitatives.

— **Dans Tidyverse**...
https ://r-graph-gallery.com/218-basic-barplots-with-ggplot2.html
https ://r-graph-gallery.com/piechart-ggplot2.html

4 D'autres références pour le traitement univarié et fusion de données

1. http ://www.sthda.com/english/wiki/descriptive-statistics-and-graphics#frequency-tables
2. https ://mgimond.github.io/ES218/Week05a.html
3. https ://cengel.github.io/R-data-viz/data-visualization-with-ggplot2.html
4. https ://www.r-graph-gallery.com/
5. Fusion https ://larmarange.github.io/analyse-R/fusion-de-tables.html

5 Pour exporter un tableau de données,...

Exporter le tableau de données

```
write.table(df,file="don.txt")
#ou
write.csv(df,file="don.csv")
```