Etude ANOVA

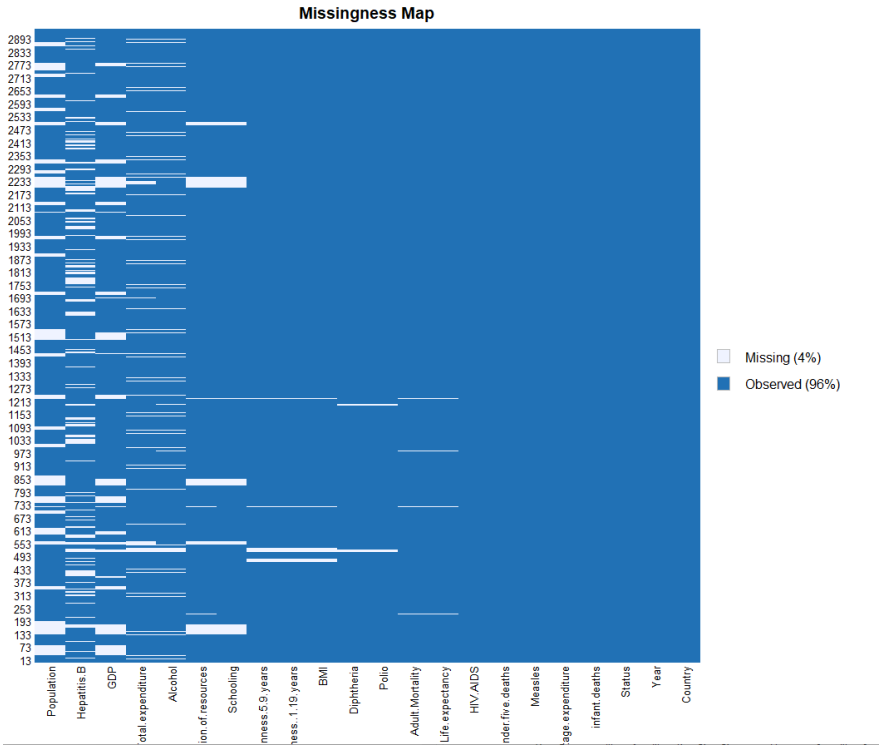
Réalisé par : *BENABOU Mohamed El Ghali*

Introduction :

Plusieurs études statistiques ont été faites sur l'espérance de vie et de comment celle-ci a évolué pendant les années. Le but de cette étude est d'une part de voir s'il y a une évolution (notamment une augmentation) de l'espérance de vie entre 2000 et 2015 dans le monde, et d'une autre part de voir si le fait qu'un pays soit développé ou en voie de développement a aussi un impact sur l'espérance de vie de ses habitants. On essayera aussi de voir s'il existe un effet d'interaction.
Ceci se fera à travers une étude d'ANOVA sur une base de données fournie par l'Organisation Mondiale de la Santé qui assure le suivi de l'état de santé ainsi que de nombreux autres facteurs connexes pour 193 pays. On se limitera à une ANOVA à deux facteurs ici : le facteur année -year et le facteur statut -Status qui décrit si un pays est développé ou en voie de développement.

Description du plan d'expérimentation :

Une inspection visuelle initiale des données a montré qu'il y avait quelques valeurs manquantes. Les données manquantes ont été traitées dans le logiciel R à l'aide de la commande Missmap. Le résultat a indiqué que la plupart des données manquantes concernaient la population, l'hépatite B et le PIB. Les données manquantes provenaient principalement de pays moins connus comme le Vanuatu, les Tonga, le Togo, le Cap-Vert, etc. Il a été difficile de trouver toutes les données pour ces pays. Mais compte tenu du fait que 191 sur 193 des pays étudiés ont au moins une valeur manquante et que les variables qu'on compte utiliser pour notre ANOVA n'en contiennent aucune, il a été décidé d'inclure quand même toutes les pays dans le modèle final.



On a trouvé que 83 % des pays (161 pays) sont considérés comme en voie de développement contre 17 % développés (32 pays). Puisqu'on compte faire une ANOVA à 2-facteur avec répétitions (puisqu'on a plusieurs pays/observations par année et par statut), on va devoir choisir un échantillon aléatoire de taille 32 depuis la modalité «developing » dans le facteur statut pour avoir un plan équilibré. On choisira ces derniers aléatoirement pour pouvoir faire de l'inférence à la fin sur la totalité des pays en voie de développement. Mais, il manquait des années dans la base de données pour certains pays et que la MissMap n'arrivait donc pas à détecter. On a choisit donc d'enlever ces pays car sinon cela aurait rendu la tâche de générer un plan aléatoire presque impossible. On a donc choisit de générer plusieurs plans aléatoires pour ne prendre que ceux où on a les 32 pays avec les 4 années voulues (ce qui est équivalent à enlever les pays avec un manque d'informations de l'étude) ce qui nous permettra d'inférer quand même sur tous les pays dont on a assez de données.
On se limitera aux années 2000, 2005, 2010 et 2015 dans le facteur année pour représenter l'évolution de l'espérance de vie chaque 5 ans au lieu d'une évolution annuelle.

Le fichier final (ensemble de données final) se compose de 3 colonnes (Year, Status et Life.expectancy) et de 256 lignes, ce qui signifie 2 variables prédictives avec 4 modalités pour « Year » et 2 pour Status avec 32 observations dans chaque cas.

Pour avoir les mêmes résultats aléatoires que dans cette étude, on prendra la random seed =935 avec la commande set.seed(935). Celle-ci a été choisi aléatoirement parmi celles qui nous donnaient des pays n'ayant aucune année qui manque.

Validation des hypothèses d'application :

Préparation des données :

On commence par préparer les données pour les différents tests qui viennent et on estime leurs statistiques de bases

```
data = read.csv("C:/Users/ben-g/Downloads/Life Expectancy Data.csv")

data = data[, 1:4]

data = data[data$Year %in% c(2000, 2005, 2010, 2015),]


data_developed = data[data$Status=="Developed",]
view(data_developed)

set.seed(935)
developing = unique(data[data$Status=="Developing",]$Country)
sample = sample(developing, 32)
data_developing = data[data$Country %in% sample,]

data = rbind(data_developed,data_developing)

data = data[, 2:4]
data_developed = data_developed[, 2:4]
data_developing = data_developing[, 2:4]
data_2000 = data[data$Year=="2000",]
data_2005 = data[data$Year=="2005",]
data_2010 = data[data$Year=="2010",]
data_2015 = data[data$Year=="2015",]
```

```
summary(data)
summary(data_2000$Life.expectancy)
summary(data_2005$Life.expectancy)
summary(data_2010$Life.expectancy)
summary(data_2015$Life.expectancy)

summary(data_developed$Life.expectancy)
summary(data_developing$Life.expectancy)
```

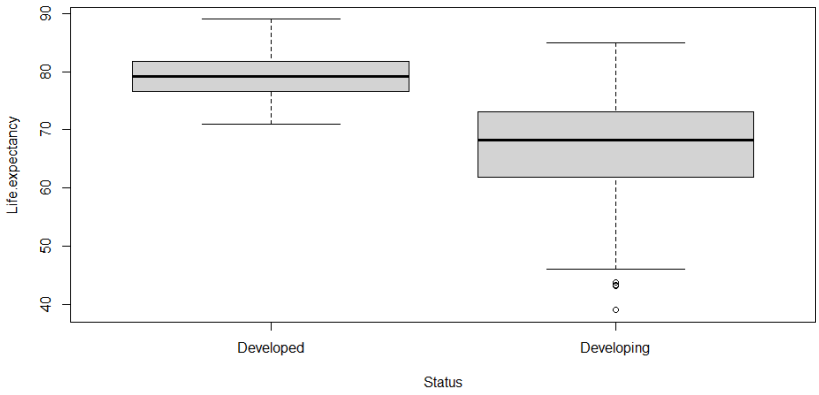
```
> summary(data_2000$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   65.70   73.30   70.11   77.80   81.10
> summary(data_2005$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  43.30   68.20   74.05   71.92   78.83   88.00
> summary(data_2010$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  48.10   69.00   74.35   73.85   81.00   89.00
> summary(data_2015$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  51.00   69.78   75.90   74.92   81.62   88.00
>
> summary(data_developed$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  71.00   76.67   79.15   79.06   81.72   89.00
> summary(data_developing$Life.expectancy)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  39.00   61.80   68.30   66.33   73.03   85.00
```

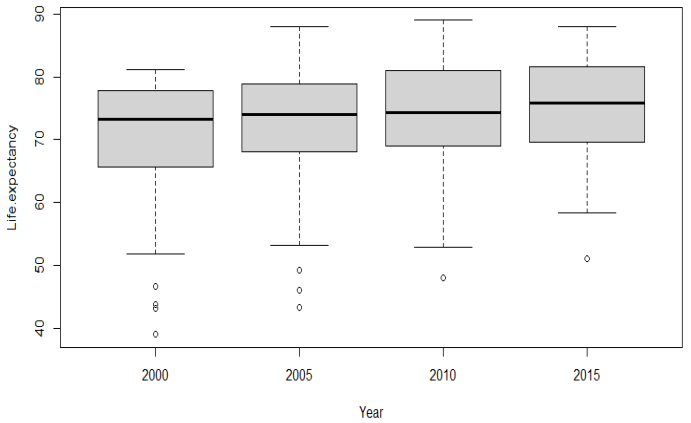
Normalité de la distribution :

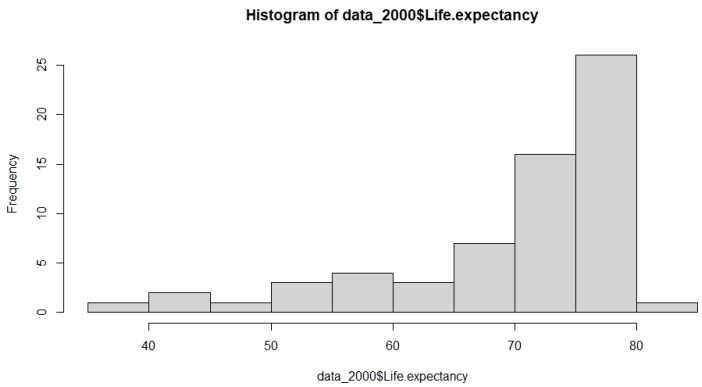
On trace après les histogrammes et les boîtes à moustaches des espérances de vie pour les deux modalités de Status et les quatre modalités de Year :

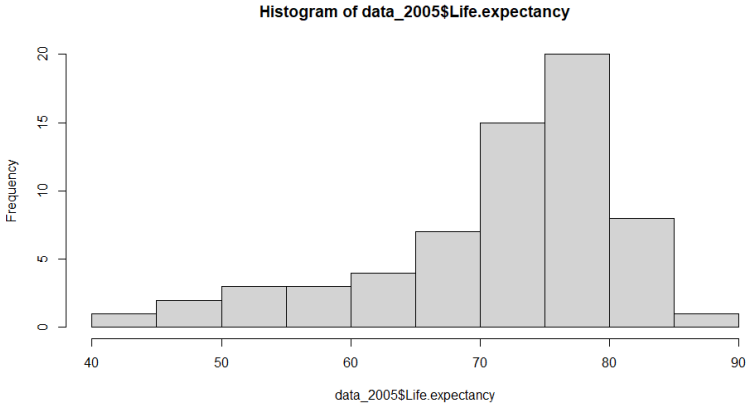
```
hist(data_developing$Life.expectancy)

hist(data_developed$Life.expectancy)

boxplot(Life.expectancy~Status,data)

boxplot(Life.expectancy~Year,data)
```

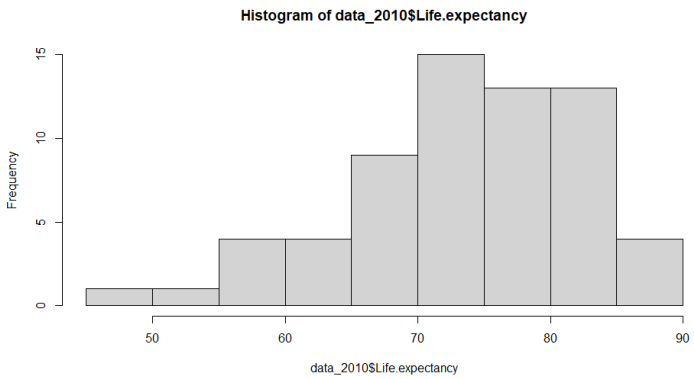
```
hist(data_2000$Life.expectancy)
hist(data_2005$Life.expectancy)
hist(data_2010$Life.expectancy)
hist(data_2015$Life.expectancy)
```

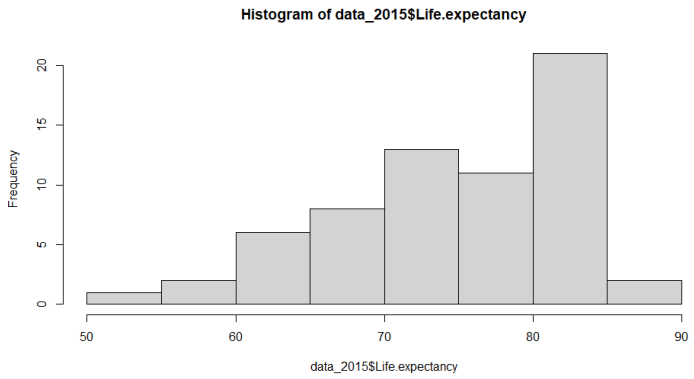












Les histogrammes suivent tous les deux une distribution plus ou moins similaire pour ce qui du facteur Status. On remarque des similitudes aussi pour ce qui est de la distribution de l'espérance de vie par rapport aux différentes modalités de Year. Les boîtes à moustaches nous montre la symétrie de la distribution pour la modalité « developed » de Status ainsi qu'une asymétrie et des valeurs aberrantes pour la modalité « developing » de Status et pour toutes les modalités de Year. On fait alors le test de Shapiro-Wilk pour trancher quand à la normalité de ces ditributions.

On installe et lance, alors, la librairie « car » qui contient le test de Shapiro-Wilk et on teste pour les différentes modalités :

```
library(car)
```

```
shapiro.test(data_2000$Life.expectancy)
shapiro.test(data_2005$Life.expectancy)
shapiro.test(data_2010$Life.expectancy)
shapiro.test(data_2015$Life.expectancy)


shapiro.test(data_developed$Life.expectancy)
shapiro.test(data_developing$Life.expectancy)
```

```
> shapiro.test(data_2000$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_2000$Life.expectancy
W = 0.83148, p-value = 4.693e-07

> shapiro.test(data_2005$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_2005$Life.expectancy
W = 0.89463, p-value = 5.049e-05

> shapiro.test(data_2010$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_2010$Life.expectancy
W = 0.95535, p-value = 0.02116

> shapiro.test(data_2015$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_2015$Life.expectancy
W = 0.93757, p-value = 0.002928
```

```
> shapiro.test(data_developed$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_developed$Life.expectancy
W = 0.98486, p-value = 0.1662

> shapiro.test(data_developing$Life.expectancy)

        Shapiro-Wilk normality test

data:  data_developing$Life.expectancy
W = 0.94412, p-value = 4.72e-05
```

D'après les résultats des tests de Shapiro-Wilk, on peut conclure, au seuil 5 %, de la normalité de la distribution de pour ce qui est de la modalité « developed » du facteur « Statuts » et de la non normalité des autres distributions ce qui confirme les résultats retrouvés graphiquement.

On doit alors utiliser une ANOVA non paramétrique pour pouvoir conclure quand à l'égalité ou à la différence de l'espérance de vie suivant ces différentes modalités mais d'abord on commence par tester l'égalité des variances.

Égalité des variances :

On préfère utiliser le test de Levene qui est moins sensible à la normalité pour tester l'égalité des variances pour chaque modalité.

On recode la variable Year en variable factorielle et on fait le test :

```
data$Year <- factor(data$Year, levels = c("2000","2005","2010","2015"))
```

```
> leveneTest(Life.expectancy~Year,data=data)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   3  0.3031 0.8231
      252
 > leveneTest(Life.expectancy~Status,data=data)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   1  48.849 2.427e-11 ***
      254
```

On a la p-value du test de la variance du facteur Year est 0,82 qui est supérieure à 0.05 donc on accepte H0 du test de Levene qui nous donne l'égalité des variances pour ce facteur.
Tandis que la p-value du test de l'égalité de la variance du facteur Status nous montre une inégalité des variances pour le facteur Status.

Test d'égalité des moyennes :

Lors des tests d'hypothèses, on a trouvé que les distributions suivant chaque modalité des deux facteurs, à l'exception de la modalité « developed » du facteur « Statuts », ne suivaient pas la loi normale. Et on n'a pu confirmer l'égalité des variances que pour le facteur Year. Donc, on va utiliser l'ANOVA de Kruskal-Wallis pour le facteur Year, et tester l'égalité des moyennes pour le facteur Status grâce au t-test de Welch vu qu'on n'a que 2 modalités dans ce facteur et une inégalité des variances.

```
> kruskal.test(Life.expectancy~Year,data = data)

        Kruskal-Wallis rank sum test

data:  Life.expectancy by Year
Kruskal-Wallis chi-squared = 10.087, df = 3, p-value = 0.01784
```

Le test de Kruskal-Wallis pour le facteur Year nous donne une p-value de 0.01<0.05. Donc, au seuil de 5 %, on peut confirmer l'existence d'une inégalité des moyennes de la variable Life.expectancy par rapport au facteur Year.

```
> t.test(data_developed$Life.expectancy,data_developing$Life.expectancy)

        Welch Two Sample t-test

data:  data_developed$Life.expectancy and data_developing$Life.expectancy
t = 14.833, df = 171.52, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.03635 14.42459
sample estimates:
mean of x mean of y
 79.06250  66.33203
```

Le test de Welch nous donne une p-value très petite donc on peut conclure de l'existence d'une différence statistiquement significative entre les moyennes des deux modalités du facteur Status. On peut aussi comparer directement les moyennes pour en conclure que celle de la modalité « developed » est plus grande que la moyenne de « developing ».

Interprétation :

Notre étude nous montre deux points essentiels :

- L'existence d'une évolution de l'espérance de vie dans le monde de l'année 2000 à l'année 2015. On peut remarqué notamment une augmentation dans les boîtes à moustaches qui représente ce cas.

- L'existence d'une différence entre l'espérance de vie des habitants des pays développés et les pays en voie de développement. Notamment, les habitants des pays développés vivent en moyenne plus que ceux dans les pays en voie de développement.