

Aya EL YAOUTI
Mariam N'DIAYE

COMPTE RENDU : TP2 TEST PARAMÉTRIQUES

Pour le 17/12/2023 À Marie Perrot-Dockès

I) Exercice 1

Nous allons ici importer notre fichier CSV nommé "Agri_conv_TP.csv" et le stocker dans une data-frame appelé Agri_conv_TP. Par la suite nous allons extraire les noms de colonnes de notre jeu de donnée Agri_conv_TP, plus précisément de la colonne 2 à la colonne 14 et nous les stockerons dans un vecteur appelé nom_indicateurs.

C'est une base de données publique française qui fournit des données sur les impacts environnementaux des produits agricoles et alimentaires. Nous avons ici 13 indicateurs mais nous allons seulement nous intéresser à l'indicateur co2_eq.

```
```{r}
install.packages("readr")
library(readr)

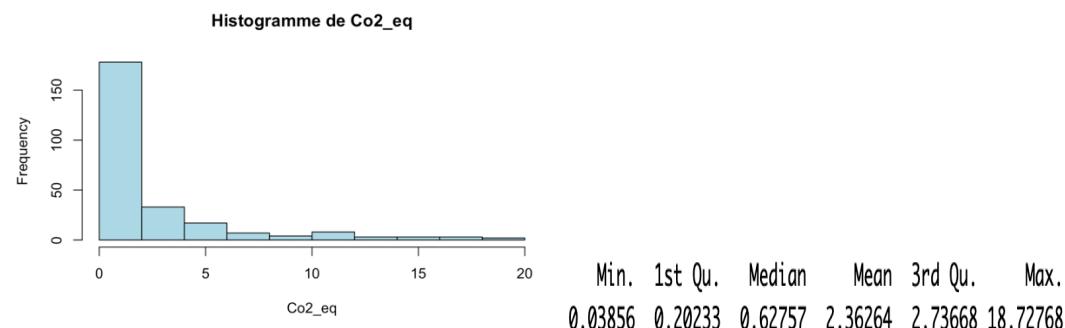
Lisez le fichier CSV
Agri_conv_TP <- read_csv("Agri_conv_TP.csv")
nom_indicateurs <- names(Agri_conv_TP) [2:14]
```
> nom_indicateurs
[1] "Co2_eq"           "Appauvr_Ozone"      "Rayon"
[4] "Ozone_chimie"    "Particules_fines"   "Acidification"
[7] "Eutrophisation_eaux_douces" "Eutrophisation_marine" "Eutrophisation_terrestre"
[10] "Utilisation_du_sol" "Epuisement_eau"       "Epuisement_energie"
[13] "Epuisement_mineraux"
```

Question :

1)

```
for(i in nom_indicateurs){
  s<-summary(Agri_conv_TP[,i])
  hist(unlist(Agri_conv_TP[,i]),main = paste("Histogramme de", variable), xlab = variable, col = "lightblue",
border = "black")
  print(s)
}
```

Sortie R :



On conclut une distribution étalemente, avec la plupart des observations autour de la médiane, mais des valeurs extrêmes élevées qui influencent la moyenne. L'indicateur varie de faibles à très élevées, indiquant une diversité importante dans les observations.

2.a)

```
for( i in nom_indicateurs){
  boxplot( unlist(Agri_conv_TP[,i])~unlist(Agri_conv_TP$Categorie) )}
```


Welch Two Sample t-test

```
data: vectoranimal and vectorplante
t = 9.191, df = 91.315, p-value = 1.223e-14
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.863846 5.994323
sample estimates:
mean of x mean of y
5.6104899 0.6814053
```

Étant donné la p-valeur très faible, on peut conclure que les moyennes des indicateurs pour les aliments animaux et végétaux sont significativement différentes. La différence est positive, ce qui suggère que les moyennes sont plus élevées pour les aliments animaux par rapport aux aliments végétaux

2.c)

Dans le contexte du test t de Welch pour comparer les moyennes des indicateurs entre les aliments animaux (X) et végétaux (Y), les hypothèses nulles (H_0) et alternatives (H_1) sont formulées comme suit :

Hypothèse Nulle (H_0) : $\mu_X - \mu_Y \leq 0$

Cela signifie qu'il n'y a aucune différence significative ou que les moyennes des indicateurs pour les aliments animaux ne sont pas significativement plus grandes que celles des aliments végétaux.

Hypothèse Alternative (H_1) : $\mu_X - \mu_Y > 0$

Cela suggère qu'il y a une différence significative et positive entre les moyennes des indicateurs pour les aliments animaux par rapport aux aliments végétaux.

```
t.test(vectoranimal, vectorplante, alternative = "greater")
t.test
```

Sortie R :

Welch Two Sample t-test

```
data: vectoranimal and vectorplante
t = 9.191, df = 91.315, p-value = 6.116e-15
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 4.037914      Inf
sample estimates:
mean of x mean of y
5.6104899 0.6814053
```

En résumé, les moyennes des indicateurs pour les aliments animaux sont significativement plus élevées que celles des aliments végétaux. Cela suggère que, dans cette étude, les valeurs des indicateurs sont généralement plus élevées pour les aliments d'origine animale par rapport aux aliments d'origine végétale.

2.d)

Dans le contexte du test F pour comparer les variances entre les vecteurs vectorplante et vectoranimal, les hypothèses nulles (H_0) et alternatives (H_1) sont formulées comme suit :

Hypothèse Nulle (H_0) : $\sigma^2_{\text{plante}} = \sigma^2_{\text{animal}}$

Cela signifie qu'il n'y a aucune différence significative entre les variances des deux groupes.

Hypothèse Alternative (H_1) : $\sigma^2_{\text{plante}} < \sigma^2_{\text{animal}}$

Cela suggère qu'il existe une différence significative et que la variance des aliments végétaux est inférieure à celle des aliments animaux.

```
var.test(vectorplante, vectoranimal, alternative = "less")
var.test
```

Sortie R :

```
F test to compare two variances

data: vectorplante and vectoranimal
F = 0.047637, num df = 169, denom df = 87, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is less than 1
95 percent confidence interval:
 0.0000000 0.0642719
sample estimates:
ratio of variances
 0.04763711
```

Le test F indique une différence significative entre les variances des indicateurs pour les aliments animaux et végétaux ($p<2.2e-16$). La variance des indicateurs pour les aliments végétaux est significativement plus faible que celle des aliments animaux.

3)

Hypothèse :

Hypothèse nulle (H0) : La proportion d'aliments végétaux nécessitant plus d'un kilo d'équivalent CO2 par kilo de production est égale ou supérieure chez les végétaux que chez les animaux.

$H_0 : p_{végétaux} \geq p_{animal}$

Hypothèse alternative (H1) : La proportion d'aliments végétaux nécessitant plus d'un kilo d'équivalent CO2 par kilo de production est inférieure chez les végétaux que chez les animaux.

$H_1 : p_{végétaux} < p_{animal}$

```
# Séparez les données en deux groupes : végétaux et animaux
donnees_vegetaux <- subset(Agri_conv_TP, Categorie == "Plant production")
donnees_animaux <- subset(Agri_conv_TP, Categorie == "Animal production")

# Calculez la proportion pour chaque groupe
prop_vegetaux <- mean(donnees_vegetaux$Co2_eq > 1)
prop_animaux <- mean(donnees_animaux$Co2_eq > 1)

# Effectuez le test de proportion
test_proportion <- prop.test(c(sum(donnees_vegetaux$Co2_eq > 1), sum(donnees_animaux$Co2_eq > 1)),
                             c(length(donnees_vegetaux$Co2_eq), length(donnees_animaux$Co2_eq)),
                             alternative = "less")

# Affichez les résultats du test
print(test_proportion)

# Concluez en fonction des résultats
if (test_proportion$p.value < 0.05) {
  cat("\nLa proportion d'aliments végétaux nécessitant plus d'un kilo d'équivalent CO2 par kilo de production
est significativement inférieure chez les végétaux que chez les animaux.")
} else {
  cat("\nIl n'y a pas suffisamment de preuves pour conclure que la proportion d'aliments végétaux nécessitant
plus d'un kilo d'équivalent CO2 par kilo de production est inférieure chez les végétaux que chez les
animaux.")
}

...
```

Sortie R :

```
2-sample test for equality of proportions with continuity correction
```

```
data: c(sum(donnees_vegetaux$Co2_eq > 1), sum(donnees_animaux$Co2_eq > 1)) out of
c(length(donnees_vegetaux$Co2_eq), length(donnees_animaux$Co2_eq))
X-squared = 130.26, df = 1, p-value < 2.2e-16
alternative hypothesis: less
95 percent confidence interval:
-1.0000000 -0.6741401
sample estimates:
prop 1   prop 2
0.1705882 0.9204545
```

La proportion d'aliments végétaux nécessitant plus d'un kilo d'équivalent CO2 par kilo de production est significativement inférieure chez les végétaux que chez les animaux.

En conclusion, les résultats du test suggèrent fortement que la proportion d'aliments végétaux nécessitant plus d'un kilo d'équivalent CO2 par kilo de production est inférieure chez les végétaux que chez les animaux.

4.a)

```
# Installez le package corrplot si ce n'est pas déjà fait
# install.packages("corrplot")

# Chargez le package corrplot
library(corrplot)

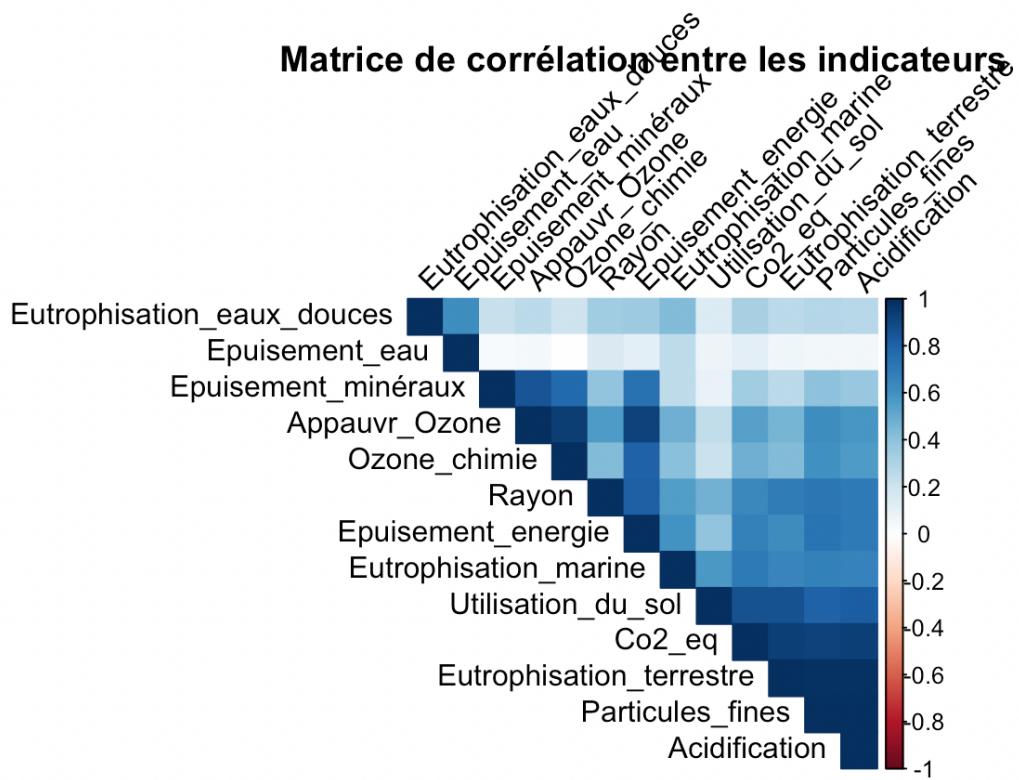
# Supposons que votre ensemble de données est stocké dans la variable "Agri_conv_TP"

# Sélectionnez les colonnes numériques pour la matrice de corrélation
colonnes_numeriques <- sapply(Agri_conv_TP, is.numeric)
matrice_correlation <- cor(Agri_conv_TP[, colonnes_numeriques])

# Personnalisez la représentation graphique avec corrplot
corrplot(matrice_correlation, method = "color", type = "upper", order = "hclust", tl.col = "black", tl.srt = 45)

# Ajoutez un titre à votre graphique
title("Matrice de corrélation entre les indicateurs")
```

Sortie R :



4.b)

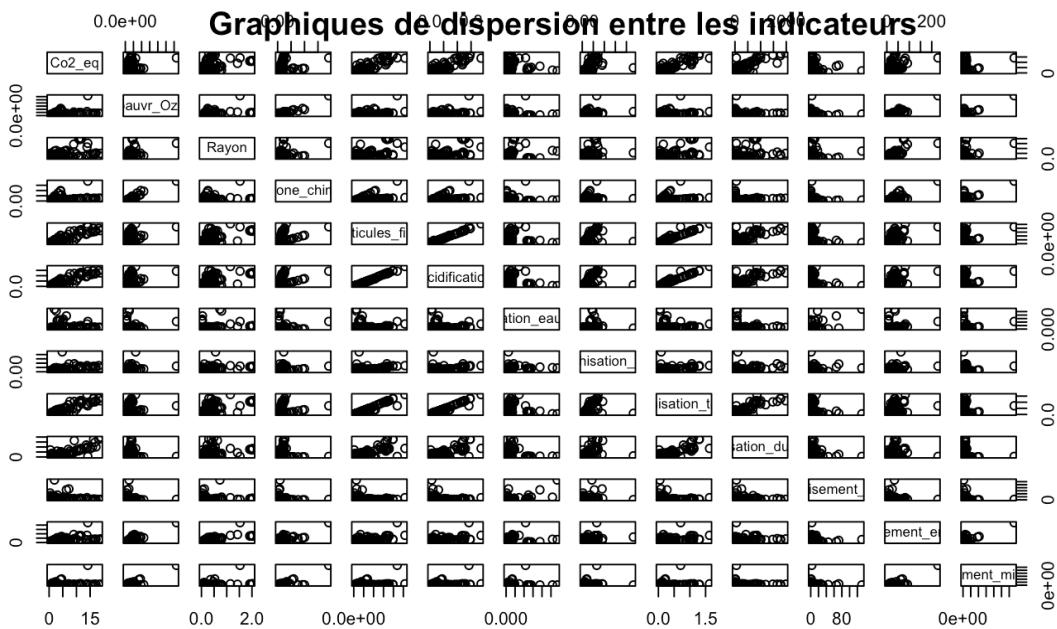
```
# Sélectionnez les colonnes numériques pour la matrice de dispersion
colonnes_numeriques <- sapply(Agri_conv_TP, is.numeric)
donnees_numeriques <- Agri_conv_TP[, colonnes_numeriques]

# Utilisez la fonction pairs pour créer la matrice de graphiques de dispersion
pairs(donnees_numeriques)

# Ajoutez un titre à votre graphique
title("Graphiques de dispersion entre les indicateurs")

# Commentez les relations visuelles entre les variables
```

Sortie R :



4.c)

```
# Supposons que votre ensemble de données est stocké dans la variable "Agri_conv_TP"

# Sélectionnez les colonnes numériques
colonnes_numeriques <- sapply(Agri_conv_TP, is.numeric)
donnees_numeriques <- Agri_conv_TP[, colonnes_numeriques]

# Sélectionnez la colonne "CO2_eq"
co2_eq <- Agri_conv_TP$Co2_eq

# Utilisez la fonction cor.test pour tester les corrélations avec CO2_eq
resultats_tests <- lapply(donnees_numeriques, function(variable) cor.test(variable, co2_eq))

# Affichez les résultats
resultats_tests
```

Sortie R :

```
$Co2_eq
Pearson's product-moment correlation

data: variable and co2_eq
t = Inf, df = 256, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 1 1
sample estimates:
cor
 1
```

En conclusion, les résultats du test suggèrent une corrélation parfaite positive entre les variables "variable" et "co2_eq". Cela implique une relation linéaire directe entre ces deux variables.

5)

La réalisation de nombreux tests dans ce TP nécessite une approche prudente en raison du risque d'erreurs statistique qui se produit lorsque l'on rejette à tort une hypothèse nulle qui est en réalité vraie. Il est crucial d'ajuster les valeurs-p pour contrôler la multiplicité des tests. De plus, la pertinence pratique des résultats doit toujours être considérée, et la connaissance du domaine est essentielle.

II) Exercice 2

On veut étudier l'influence du sexe sur la catégorie d'employé dans cette base de données.

Pour étudier le lien entre le sexe et la catégorie d'employés, on peut utiliser un test statistique appelé le test du chi carré pour l'indépendance. Ce test est approprié lorsque les variables sont catégorielles et que l'on souhaite déterminer s'il existe une association significative entre elles.

Code R :

```
# 2. Tableau de contingence
table_contingence <- table(employes$sexe, employes$catemp)

# Affichez le tableau de contingence
print(table_contingence)

# 3. Proportion d'hommes parmi les cadres et les secrétaires
# Utilisez la fonction prop.table pour calculer les proportions
prop_table <- prop.table(table_contingence, margin = 1) # margin = 1 pour les proportions par ligne (féminin, masculin)

# Affichez les proportions
prop_hommes_cadres <- prop_table["masculin", "Cadre"]
prop_hommes_secretaires <- prop_table["masculin", "Secr\xe9tariat"]

print(paste("Proportion d'hommes parmi les cadres :", prop_hommes_cadres))
print(paste("Proportion d'hommes parmi les secrétaires :", prop_hommes_secretaires))

# 4. Proportion de secrétaires parmi les hommes et les femmes
# Utilisez la fonction prop.table pour calculer les proportions
prop_table <- prop.table(table_contingence, margin = 2) # margin = 2 pour les proportions par colonne (Cadre, Responsable, Secr\xe9tariat)

# Affichez les proportions
prop_secretaires_hommes <- prop_table["masculin", "Secr\xe9tariat"]
prop_secretaires_femmes <- prop_table["f\xe9minin", "Secr\xe9tariat"]

print(paste("Proportion de secrétaires parmi les hommes :", prop_secretaires_hommes))
print(paste("Proportion de secrétaires parmi les femmes :", prop_secretaires_femmes))
```

Sortie R :

```

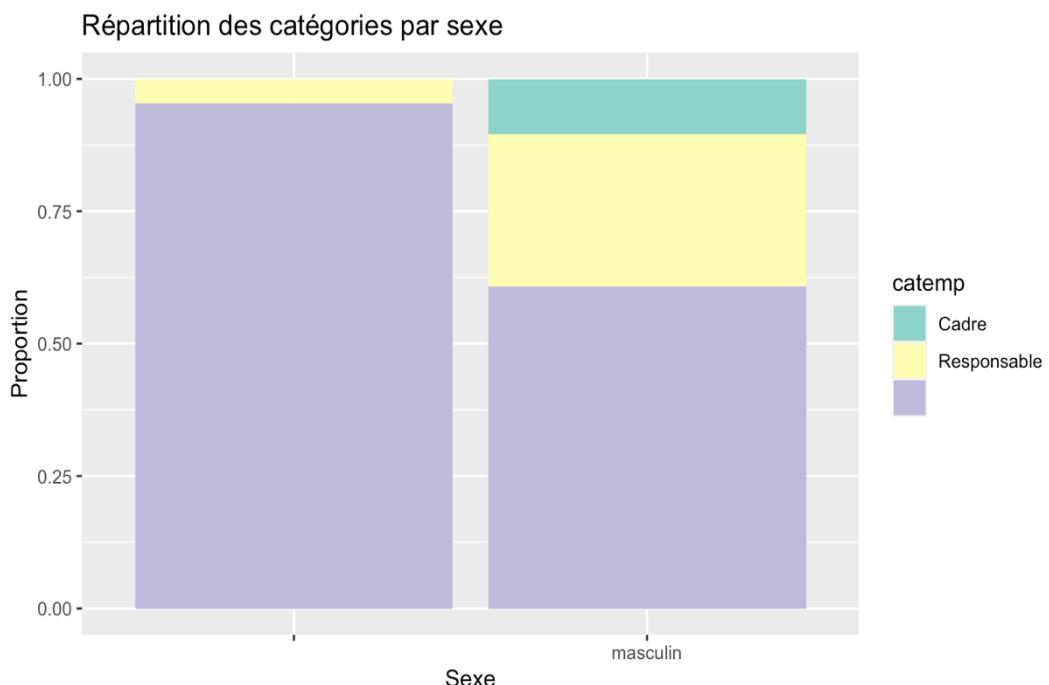
    Cadre Responsable Secrétaire
feminin      0       10      206
masculin     27       74      157
[1] "Proportion d'hommes parmi les cadres : 0.104651162790698"
[1] "Proportion d'hommes parmi les secrétaires : 0.608527131782946"
[1] "Proportion de secrétaires parmi les hommes : 0.432506887052342"
[1] "Proportion de secrétaires parmi les femmes : 0.567493112947658"

```

```
# 5. Graphique et test du chi-carré
library(ggplot2)
```

```
# Graphique en barres
ggplot(employes, aes(x = sexe, fill = catemp)) +
  geom_bar(position = "fill", stat = "count") +
  labs(title = "Répartition des catégories par sexe",
       x = "Sexe", y = "Proportion") +
  scale_fill_brewer(palette = "Set3")
```

Graphique en barre de la répartition des catégories par sexe :



Nous allons ici utiliser le test du chi-deux ou chi-carré comme expliquer auparavant.

Hypothèse :

H0 (hypothèse nulle) : Il n'y a pas de relation entre le sexe et la catégorie d'employés.

H1(hypothèse alternative): Il y'a une relation entre le sexe et la catégorie d'employés

```
# Test du chi-carré
#Hypothèse nulle (H0) : Il n'y a pas de relation entre le sexe et la catégorie d'employés.

##H0: phommes,categorie = pfemmes,categorie

#Hypothèse alternative (H1) : Il y a une relation entre le sexe et la catégorie d'employés.
##H1: phommes,categorie !=pfemmes,categorie
test_chi2 <- chisq.test(table_contingence)
print(test_chi2)
```

Sortie R :

```
Pearson's Chi-squared test

data: table_contingence
X-squared = 79.277, df = 2, p-value < 2.2e-16
```

Les résultats du test du chi carré indiquent une statistique du chi carré est de 79.277 avec 2 degrés de liberté et une p-valeur extrêmement faible ($pval < 2.2e-16$). Avec une p-valeur aussi faible, on rejette l'hypothèse nulle selon laquelle il n'y a pas de relation entre le sexe et la catégorie d'employés.

Conclusion :

Il existe une association significative entre le sexe et la catégorie d'employés. En d'autres termes, le test statistique suggère que la distribution des catégories d'employés diffère de manière significative en fonction du sexe.