

Examen 2020/2021

#Exercice 1

Chargement des données

```
data = read.table("C:/Users/ibnou/OneDrive/Bureau/TP TraitStat2/Annales/vins.data",  
header = TRUE)  
attach(data)
```

#1.1.1. Nombre de vins

```
nbvins = nrow(data)  
nbvins  
#Le nombre de vins est 681
```

#1.1.2. Visualisation des distributions des variables avec des boxplots

```
summary(data)  
x11()  
par(mfrow=c(2,3)) # Pour organiser les graphiques en 2 lignes, 3 colonnes
```

Boxplot pour Acides :

```
boxplot(data$acidite, main="Acides")
```

Boxplot pour Chlorides :

```
boxplot(data$chlorides, main="Chlorides")
```

Boxplot pour Sucre :

```
boxplot(data$sucre, main="Sucre")
```

Boxplot pour Sulphates :

```
boxplot(data$sulphates, main="Sulphates")
```

Boxplot pour Alcool :

```
boxplot(data$alcool, main="Degré d'alcool")
```

#D'après les boxplots, le composant "Sulphates" semble être présent en beaucoup plus grandes quantités que les autres composants.

```
summary(data$acidite)  
summary(data$chlorides)  
summary(data$cucure)  
summary(data$sulphates)  
summary(data$alcool)
```

#en regardant les valeurs minimales et les quartiles, les variables qui présentent des valeurs extrêmes du côté des petites valeur sont : "Chlorides" et "Sulphates"

1.1.3. Représentation de la densité continue estimée pour "Sulphates"

```
plot(density(data$sulphates), main="Densité continue estimée pour Sulphates")
```

#D'après le graphe, la distribution n'est pas multimodale, car il n'y a qu'un seul pic visible sur la densité estimée.

1.1.4. Vérification de la normalité avec un graphique QQ pour les "Sucre"

```
x11()  
qqnorm(chlorides)  
qqline(chlorides, col = "red") #pour ajouter la droite de référence
```

Les points dévient fortement la ligne rouge droite, ce qui indique une distribution non-normale.

#1.1.5.

Moyenne de Sucre

mean(sucre)

Variance non corrigée de Sucre

var(sucre)

#1.1.6. Compter le nombre de vins avec un degré de plus de 13%

sum(alcool > 13)

Il existe 2 vins avec un degré d'alcool supérieur à 13

#1.1.7.

/*

La formule pour un intervalle de confiance sur la moyenne est :

$$\bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

où :

- \bar{x} est la moyenne de l'échantillon,
- $t_{\alpha/2}$ est le quantile de la distribution de Student pour un niveau de confiance de $1 - \alpha$,
- s est l'écart-type de l'échantillon,
- n est la taille de l'échantillon.

*/

Calcul de la moyenne et de l'écart-type de Sulphates

mean_sulphates <- mean(sulphates)

sd_sulphates <- sd(sulphates)

Taille de l'échantillon

n <- length(sulphates)

Quantile de la distribution de Student pour le niveau de confiance

t_value <- qt(1 - (1 - 0.90) / 2, df = n - 1)

Calcul de l'intervalle de confiance

sup <- mean_sulphates - t_value * (sd_sulphates / sqrt(n))

inf <- mean_sulphates + t_value * (sd_sulphates / sqrt(n))

sup

inf

#l'intervalle de confiance obtenu est [44.34937, 48.49645]

Utilisation de t.test pour vérifier l'intervalle de confiance

t.test(sulphates, conf.level = 0.90)\$conf.int

#On obtient bien le même intervalle (44.34937 48.49645)

#1.1.8.

#Note : Si la p-valeur est inférieure au seuil de risque de 1% (0.01), alors nous rejetons l'hypothèse nulle.(les données remettent en cause l'hypothèse d'une quantité moyenne de sucre dans la population égale à 15g.). Si la p-valeur est supérieure à 0.01, alors nous ne rejetons pas l'hypothèse nulle

```
# Moyenne et écart-type du sucre
mean_sucre <- mean(sucre)
sd_sucre <- sd(sucre)
# Taille de l'échantillon
n <- length(sucre)
# Calcul de la statistique de test t
t_stat <- (mean_sucre - 15) / (sd_sucre / sqrt(n))
# Calcul de la p-valeur
p_value <- 2 * pt(-abs(t_stat), df=n - 1)
p_value
```

#on obtient une p-valeur de 0.03207212, qui est supérieure à 1%, alors nous ne rejetons pas l'hypothèse nulle H0 d'une quantité moyenne de sucre dans la population égale à 15g.

#1.2.1.

```
#graphiquement, on peut regarder à l'aide de heatmap
matrice_correlation = cor(data[, c("acidite", "chlorides", "sucre", "sulphates", "alcool")])
> heatmap(matrice_correlation,
           Rowv = NA, Colv = NA,
           col = colorRampPalette(c("blue", "white", "red"))(50),
           symm = TRUE,
           margins = c(5, 5),
           main = "Heatmap de Corrélation")
#D'après le graphe, les variables les plus corrélées sont "sucre" et "sulphates"
```

matrice_correlation

#On obtient la matrice de correlation suivante

```
#      acidite  chlorides   sucre sulphates  alcool
#acidite  1.0000000  0.102982708 -0.167732385 -0.11263942 -0.68627910
#chlorides 0.1029827  1.000000000  0.004880273  0.06922847 -0.27726289
#sucre    -0.1677324  0.004880273  1.000000000  0.67956819  0.07627748
#sulphates -0.1126394  0.069228475  0.679568190  1.000000000 -0.11030863
#alcool   -0.6862791 -0.277262892  0.076277478 -0.11030863  1.00000000
```

#D'après ces valeurs : les variables les plus corrélées sont "sucre" et "sulphates"

#1.2.2.

```
#Calculons d'abord la corrélation de Pearson
cor(sucre, sulphates)
#Puis, faisons un test de significativité
cor.test(sucre, sulphates)$p.value
#Si la p-valeur est inférieure à ce seuil, on rejette l'hypothèse nulle et on conclut qu'il y a une
corrélation significative entre les variables.
#Ce qui est le cas ici; on a obtenu une p-valeur de 2.017044e-93 qui est inférieur à 1%,
alors il y a bien une corrélation significative entre "sucre" et "sulphates"
```

#1.2.3.

```
#Calcul de la corrélation de Pearson entre acidite et chlorides
```

```
cor_ac = cor(acidite, chlorides)
# Transformation de la corrélation en statistique de test t
t_stat = cor_ac * sqrt((length(acidite) - 2) / (1 - corac^2))
t_stat
#Note : Si la valeur absolue de la statistique de test est grande, cela indique une forte
corrélation, et si elle est proche de zéro, cela indique une faible corrélation.
#On obtient 2.697829
```

```
#1.2.4.
cor(sucre, acidite)
```

```
#1.3.
```

```
/*
```

Pour effectuer un test sur la médiane de la quantité de sucre dans la population, nous allons utiliser un test de médiane basé sur le nombre de valeurs observées supérieures à 15g. Voici les étapes pour effectuer ce test :

Étape 1 : Calcul de la statistique de test T

La statistique de test T est définie comme le nombre de valeurs observées supérieures à 15g.

Étape 2 : Calcul de la moyenne et de l'écart-type de T sous l'hypothèse nulle

Sous l'hypothèse nulle : "quantité médiane de sucre = 15g dans la population", la moyenne de T est $n/2$ et l'écart-type est $\sqrt{n/4}$, où n est la taille de l'échantillon.

Étape 3 : Calcul de la z-statistique. La z-statistique est calculée comme :

$$z = \frac{T - \frac{n}{2}}{\sqrt{\frac{n}{4}}}$$

Étape 4 : Calcul de la p-valeur

La p-valeur est calculée en utilisant la loi normale standard (distribution normale centrée réduite).

Étape 5 : Test d'hypothèse au seuil de 1%

Nous allons comparer la z-statistique à la valeur critique de la distribution normale standard au seuil de 1% pour déterminer si nous rejetons ou non l'hypothèse nulle.

```
*/
```

```
# Calcul de T : nombre de valeurs supérieures à 15g
T <- sum(sucre > 15)
# Taille de l'échantillon
n <- length(data$Sucre)
# Calcul de la moyenne et de l'écart-type de T sous H0
mean_T <- n / 2
sd_T <- sqrt(n / 4)
# Calcul de la z-statistique
z_stat <- (T - mean_T) / sd_T
# Calcul de la p-valeur
p_value <- 2 * pnorm(-abs(z_stat)) # test bilatéral
# Seuil de risque de 1%
alpha <- 0.01
```

```
# Calcul de la région critique
```

```
z_critique <- qnorm(alpha/2)
```

```
# Comparaison avec la région critique
```

```
if (z_stat < -z_critique | z_stat > z_critique) {  
  cat("On rejette H0 au niveau de 1% de signification.\n")  
} else {  
  cat("On ne rejette pas H0 au niveau de 1% de signification.\n")  
}
```

```
#Exercice 2
```

```
# Chargement des données
```

```
data = read.table("C:/Users/ibnou/OneDrive/Bureau/TP TraitStat2/Annales/films.data",  
header = TRUE)  
attach(data)
```

```
# 2.1.1. Nombre de films dans le jeu de données
```

```
summary(data)
```

```
nrow(data)
```

```
#On a 98 films
```

```
#2.1.2. Le nombre de réalisateurs uniques
```

```
length(unique(director))
```

```
# Il y a 73 réalisateurs différents
```

```
#2.1.3. Visualisation de la distribution des catégories de films (genre)
```

```
#Pour le faire on crée le diagramme en barres
```

```
barplot(table(genre))
```

```
#2.1.4.
```

```
# Calcul de la proportion de comédies dans l'échantillon
```

```
nb_comedies <- sum(genre == "Comedy")
```

```
nb_films <- nrow(data)
```

```
prop_comedies <- nb_comedies / nb_films
```

```
prop_comedies
```

```
#On obtient environ 20,41%
```

```
#Test d'hypothèse sur la proportion :
```

```
# Calcul de la statistique de test z
```

```
z_stat <- (prop_comedies - 0.3) / sqrt(0.3 * (1 - 0.3) / nb_films)
```

```
z_stat
```

```
#On obtient -2.072074, Une z-statistique négative signifie que la proportion observée est inférieure à l'hypothèse
```

```
# Calcul de la p-valeur
```

```
p_value <- 2 * (1 - pnorm(abs(z_stat)))
```

```
#On obtient 0.03825858, on a une p-valeur inférieure à un seuil choisi (0.05), ce qui indique que nous rejetons l'hypothèse nulle. Cela signifie que nos données fournissent suffisamment de preuves pour dire que la proportion de comédies dans la population n'est pas égale à 30%.
```

#2.2.1. Table de contingence croisant les variables "genre" et "starring"

table(genre, starring)

#	starring				
#genre	Ben Affleck	Bradley Cooper	Johnny Depp	Liam Neeson	Matt Damon
# Action-Adventure	10	2	13	10	8
# Biography	0	0	4	3	0
# Comedy	5	5	6	1	3
# Drama	7	2	6	5	8

#2.2.2.

#La table de contingence attendue sous l'hypothèse d'indépendance peut être calculée en multipliant les totaux marginaux des lignes et des colonnes, puis en divisant par le total général.

#Si $O_{i,j}$ est l'observation à l'intersection de la ligne i et de la colonne j , la formule pour la valeur attendue $E_{i,j}$ est :

$E_{i,j} = R_i * C_j / N$

#Où :

R_i est le total marginal de la ligne i ,

C_j est le total marginal de la colonne j ,

N est le total général.

#2.2.3.

```
barplot(table(genre, starring), beside = TRUE, legend.text = TRUE, col =  
rainbow(nrow(table(genre, starring))), main = "Genres Cinématographiques par Acteur", xlab =  
"Acteur", ylab = "Nombre de Films", args.legend = list(x = "topright", bty = "n"))
```

#2.2.4.

Le lien entre les deux variables est-il fort ?

/* Le coefficient de Cramer est une mesure de l'association entre deux variables catégorielles. Il varie de 0 à 1, où 0 indique aucune association et 1 indique une association parfaite. La force de la corrélation peut être interprétée approximativement comme suit :

0 : Aucune corrélation.

0.1 à 0.3 : Corrélation faible.

0.3 à 0.5 : Corrélation modérée.

0.5 à 1 : Corrélation forte.

Exemple avec deux variables catégorielles 'Variable1' et 'Variable2'

```
cross_table <- table(Variable1, Variable2)
```

```
chisq_test <- chisq.test(cross_table)
```

```
cramer_coef <- sqrt(chisq_test$statistic / (sum(cross_table) * (min(dim(cross_table)) - 1)))
```

```
cramer_coef */
```

```
chisq_test <- chisq.test(table(genre, starring))
```

```
cramer_coef <- sqrt(chisq_test$statistic / (sum(table(genre, starring)) * (min(dim(table(genre,  
starring))) - 1)))
```

```
cramer_coef
```

#On obtient 0.2516143, donc on a un lien faible entre les deux variables

#? Est-il significatif ?

#Note : Le test de chi-deux nous fournira une statistique de test et une p-valeur. Si la p-valeur est inférieure à un seuil de signification (par exemple 0.05), nous rejetons l'hypothèse nulle d'indépendance, ce qui signifierait qu'il existe un lien significatif entre les genres de films et les acteurs.

Test de Chi-deux d'indépendance

```
chi2_test = chisq.test(table(genre, starring))
```

```
chi2_test
```

#On obtient p-value de 0.0983, ce qui est supérieur à notre seuil, donc il n'existe pas de lien significatif entre les deux variables

Examen 2022/2023

#Exercice 1

Chargement des données

```
data = read.table("C:/Users/ibnou/OneDrive/Bureau/TP  
TraitStat2/Annales/circonscriptions.data", header = TRUE)  
attach(data)  
# Résumé du jeu de données  
summary(data)
```

#1.1.1. Nombre de circonscriptions

```
nbcirconscriptions = nrow(data)  
nbcirconscriptions  
#Le nombre de circonscriptions est ...
```

#1.1.2. Visualisation des distributions des variables avec des boxplots

```
par(mfrow=c(3,4)) # Pour organiser les graphiques  
j <- 1  
for (i in data) {  
  hist(i, main=colnames(data)[j]) #boxplots des variables  
  j <- j + 1  
}
```

#1.1.3. Comparaison graphique les distributions des quatre catégories socioprofessionnelles

```
x11()  
par(mfrow=c(2,2))  
boxplot(pChomage, main="Chômeurs")  
boxplot(pCadre, main="Cadre")  
boxplot(pPI, main="intermédiaires")  
boxplot(pOuvrier, main="Ouvriers")  
#Les boxplots permettent de visualiser la médiane, la dispersion, et les valeurs aberrantes  
pour chaque catégorie socioprofessionnelle  
#.....
```

#1.1.4. Formation politique recevant le plus de voix en moyenne

```
# Comparaison des moyennes et des médianes  
mean_values <- c(mean(NUPES), mean(MAJORITE), mean(RN))  
median_values <- c(median(NUPES), median(MAJORITE), median(RN))  
partis <- c("NUPES", "Majorité", "RN")  
comparaison_df <- data.frame(Parti = partis, Moyenne = mean_values, Médiane =  
median_values)  
comparaison_df
```

#1.1.5. Parti politique avec la part d'électeurs la plus dispersée

```
# Comparaison des variances et des écart-type  
variance_values <- c(var(NUPES), var(MAJORITE), var(RN))  
sd_values <- c(sd(NUPES), sd(MAJORITE), sd(RN))  
comparaison_dispersion_df <- data.frame(Parti = partis, Variance = variance_values,  
EcartType = sd_values)  
comparaison_dispersion_df
```


#Le parti politique avec la plus grande variance ou le plus grand écart-type est celui avec la part d'électeurs la plus dispersée.

#1.1.6. Calcul de l'intervalle inter-quartile (IQR) du taux de chômage

```
iqr_chomage <- IQR(pChomage)
iqr_chomage
```

```
# Calcul de Q1 (premier quartile)
Q1 <- quantile(pChomage, 0.25)
# Calcul de Q3 (troisième quartile)
Q3 <- quantile(pChomage, 0.75)
# Calcul de l'IQR en soustrayant Q1 de Q3
IQR_chomage <- Q3 - Q1
# Calcul de l'intervalle interquartile
intervalle_interquartile <- c(Q1, Q3)
```

```
intervalle_interquartile
```

#1.1.7. Calcul de la moyenne et de la variance corrigée de la variable procheVille

```
mean(procheVille)
var(procheVille)
```

#1.1.8.

Graphique de densité de la variable ageMedian

```
plot(density(ageMedian))
```

Test de Shapiro-Wilk pour la normalité de ageMedian

```
shapiro.test(ageMedian)
```

#Si la p-valeur du test de Shapiro-Wilk est supérieure à un seuil de 0.05, on ne rejette pas l'hypothèse nulle selon laquelle la distribution est normale.

#1.1.9. Calcul de la région critique :

Nombre moyen d'inscrits

```
mean_nbInscrits <- mean(nblInscrits)
```

Taille de l'échantillon

```
n <- length(nblInscrits)
```

Écart-type de l'échantillon

```
sd_nbInscrits <- sd(nblInscrits)
```

Moyenne sous H0

```
mean_H0 <- 85
```

Calcul de la statistique de test t

```
t_stat <- (mean_nbInscrits - mean_H0) / (sd_nbInscrits / sqrt(n))
```

Calcul de la p-valeur

```
p_value <- pt(t_stat, df = n - 1)
```

Seuil de risque

```
alpha <- 0.01
```

```
# Région critique (à deux côtés pour un seuil de 1%)
```

```
critical_value <- qt(1 - alpha/2, df = n - 1)
```

```
inf <- mean_H0 - critical_value * (sd_nblnscrits / sqrt(n))
```

```
sup <- mean_H0 + critical_value * (sd_nblnscrits / sqrt(n))
```

```
t_stat
```

```
inf
```

```
sup
```

```
#Si la statistique de test t_stat se situe en dehors de la région critique, nous rejetterons  
l'hypothèse nulle.
```

```
#1.2.1. les 4 corrélations les plus fortes:
```

Pour identifier les corrélations linéaires fortes, nous pouvons calculer les coefficients de corrélation de Pearson entre toutes les paires de variables. Les corrélations proches de -1 ou 1 indiquent une forte corrélation linéaire, tandis que celles proches de 0 indiquent une faible corrélation.

```
# Calcul des corrélations de Pearson entre toutes les paires de variables et Affichage de la  
matrice de corrélation
```

```
cor(data[, c("nblnscrits", "procheVille", "ageMedian", "pChomage", "pCadre", "pPI",  
"pOuvrier", "ABSTENTION", "NUPES", "MAJORITE", "RN")])
```

```
#...
```

```
#1.2.2.
```

```
x11()
```

```
par(mfrow=c(2,2))
```

```
#Nuage de points pour (nblnscrits, pChomage)
```

```
plot(nblnscrits, pChomage, xlab = "Nombre d'inscrits", ylab = "Pourcentage de chômage",  
main = "Nuage de points")
```

```
#Nuage de points pour (pCadre, pPI)
```

```
plot(data$pCadre, data$pPI, xlab = "Pourcentage de cadres", ylab = "Pourcentage de  
professions intermédiaires", main = "Nuage de points")
```

```
#...
```

```
#1.2.3. Corrélation entre pCadre et pChomage significative?
```

```
cor.test(pCadre, pChomage, method = "pearson")
```

```
#Si la p-valeur (cor_test$p.value) est inférieure à 0.05 (ou un seuil de 5%), alors nous  
pouvons conclure que la corrélation est significative.
```

```
#1.2.4. Spearman
```

```
# Calcul du coefficient de corrélation de Spearman pour (pCadre, pChomage)
```

```
spearman_corr <- cor(data$pCadre, data$pChomage, method = "spearman")
```

```
# Calcul de la statistique de test Z
```

```
n <- nrow(data)
```

```
Z_stat <- sqrt(n - 2) * (1 - spearman_corr) / sqrt(1 - spearman_corr^2)
```

```
# Calcul de la p-valeur
p_value_spearman <- 2 * (1 - pt(abs(Z_stat), df = n - 2))
```

```
# Seuil de risque
alpha <- 0.05
```

```
# Région critique (bilatéral)
critical_value_spearman <- qt(1 - alpha/2, df = n - 2)
```

```
Z_stat
p_value_spearman
critical_value_spearman
```

#1.3.1.

#Supposons que nous choisissons la formation politique "NUPES" :

Création de la variable binaire pour NUPES

```
Voted_NUPES <- ifelse(data$NUPES > mean(data$NUPES), TRUE, FALSE)
```

#1.3.2. Comparaison des âges médians

```
boxplot(ageMedian ~ Voted_NUPES, data = data, xlab = "Vote pour NUPES", ylab = "Âge Médian", main = "Comparaison des âges médians")
```

#1.3.3. Test de comparaison des moyennes d'abstentionnistes

#Nous pouvons utiliser un test de Student pour comparer les moyennes des deux groupes pour la variable "ABSTENTION"

#La p-valeur du test nous permettra de savoir si les deux catégories ont en moyenne la même part d'abstentionnistes.

```
test = t.test(ABSTENTION ~ Voted_NUPES)
```

Conclusion

```
if (test$p.value < alpha) {
  cat("Il est possible que les deux catégories aient en moyenne la même part d'abstentionnistes.\n")
} else {
  cat("Il est peu probable que les deux catégories aient en moyenne la même part d'abstentionnistes.\n")
}
```

#1.3.4. Valeur observée de la statistique de test

/*

Pour le test de comparaison des moyennes d'abstentionnistes, la statistique de test t est donnée par :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Où \bar{x}_1, \bar{x}_2 sont les moyennes des deux groupes, s_1, s_2 sont les écarts-types et n_1, n_2 sont les tailles des échantillons.

*/

Calcul de la statistique de test

```

group1 <- ABSTENTION[Voted_NUPES == TRUE]
group2 <- ABSTENTION[Voted_NUPES == FALSE]
n1 <- length(group1)
n2 <- length(group2)
mean1 <- mean(group1)
mean2 <- mean(group2)
sd1 <- sd(group1)
sd2 <- sd(group2)
t_stat_observed <- (mean1 - mean2) / sqrt((sd1^2 / n1) + (sd2^2 / n2))
t_stat_observed

```

#La valeur observée est donc : ...

#Exercice 2

Chargement des données

```

data = read.table("C:/Users/ibnou/OneDrive/Bureau/TP TraitStat2/Annales/cinema.data",
header = TRUE)
attach(data)

```

#2.1.1. Le nombre de réalisateurs uniques

```
length(unique(director))
```

Il y a ... réalisateurs différents

#2.1.2. Visualisation de la distribution des catégories de films (genre)

#Pour le faire on crée le diagramme en barres

```
barplot(table(genre))
```

#2.1.3.

Calcul de la proportion de comédies dans l'échantillon

```
nb_comedies <- sum(genre == "Comedy")
```

```
nb_films <- nrow(data)
```

```
prop_comedies <- nb_comedies / nb_films
```

```
prop_comedies
```

#On obtient environ ...%

#2.1.4.

#Test d'hypothèse sur la proportion :

Calcul de la statistique de test z

```
z_stat <- (prop_comedies - 0.3) / sqrt(0.3 * (1 - 0.3) / nb_films)
```

```
z_stat
```

#On obtient -2.072074, Une z-statistique négative signifie que la proportion observée est inférieure à l'hypothèse

Calcul de la p-valeur

```
p_value <- 2 * (1 - pnorm(abs(z_stat)))
```

#On obtient 0.03825858, on a une p-valeur inférieure à un seuil choisi (0.05), ce qui indique que nous rejetons l'hypothèse nulle. Cela signifie que nos données fournissent suffisamment

de preuves pour dire que la proportion de comédies dans la population n'est pas égale à 30%.

#2.2.1. Table de contingence croisant les variables "genre" et "starring"

```
table(genre, starring)
```

#2.2.2.

#La table de contingence attendue sous l'hypothèse d'indépendance peut être calculée en multipliant les totaux marginaux des lignes et des colonnes, puis en divisant par le total général.

#Si $O_{i,j}$ est l'observation à l'intersection de la ligne i et de la colonne j , la formule pour la valeur attendue $E_{i,j}$ est :

$E_{i,j} = R_i * C_j / N$

#Où :

R_i est le total marginal de la ligne i ,

C_j est le total marginal de la colonne j ,

N est le total général.

#2.2.3. Graphe : pour chaque actrice son genre cinématographique de prédilection.

```
barplot(table(genre, starring), beside = TRUE, legend.text = TRUE, col =  
rainbow(nrow(table(genre, starring))), main = "Genres Cinématographiques par Acteur", xlab  
= "Acteur", ylab = "Nombre de Films", args.legend = list(x = "topright", bty = "n"))
```

#2.2.4.

Le lien entre les deux variables est-il fort ?

/* Le coefficient de Cramer est une mesure de l'association entre deux variables catégorielles. Il varie de 0 à 1, où 0 indique aucune association et 1 indique une association parfaite. La force de la corrélation peut être interprétée approximativement comme suit :

0 : Aucune corrélation.

0.1 à 0.3 : Corrélation faible.

0.3 à 0.5 : Corrélation modérée.

0.5 à 1 : Corrélation forte.

Exemple avec deux variables catégorielles 'Variable1' et 'Variable2'

```
cross_table <- table(Variable1, Variable2)
```

```
chisq_test <- chisq.test(cross_table)
```

```
cramer_coef <- sqrt(chisq_test$statistic / (sum(cross_table) * (min(dim(cross_table)) - 1)))
```

```
cramer_coef */
```

```
chisq_test <- chisq.test(table(genre, starring))
```

```
cramer_coef <- sqrt(chisq_test$statistic / (sum(table(genre, starring)) * (min(dim(table(genre,  
starring))) - 1)))
```

```
cramer_coef
```

#On obtient 0.2516143, donc on a un lien faible entre les deux variables

#? Est-il significatif ?

#Note : Le test de chi-deux nous fournira une statistique de test et une p-valeur. Si la p-valeur est inférieure à un seuil de signification (par exemple 0.05), nous rejetons l'hypothèse nulle d'indépendance, ce qui signifierait qu'il existe un lien significatif entre les genres de films et les acteurs.

```
# Test de Chi-deux d'indépendance
```

```
chi2_test = chisq.test(table(genre, starring))
```

```
chi2_test
```

```
#On obtient p-value de 0.0983, ce qui est supérieur à notre seuil, donc il n'existe pas de lien  
significatif entre les deux variables
```