MICRO-PROJET
PROBABILITÉS STATISTIQUES

Clarisse Veron, Yacine Ahmed Yahia

Ce papier présente une analyse de données provenant de trois fichiers (d1, d2, et d3) qui contiennent des informations sous forme de matrices et de listes de valeurs. Les principales conclusions de cette analyse incluent des tendances dans les matrices de d1, des motifs dans les listes de d2, et des corrélations dans les données de d3. Ces résultats fournissent des informations clés pour une compréhension plus approfondie des données et servent de base à l'étude ultérieure présentée dans ce document.

I. INTRODUCTION

Cette analyse se concentre sur l'analyse des données provenant de trois sources principales : d1, d2, et d3. L'objectif de cette étude est de réaliser une évaluation approfondie de ces données sans introduire de contexte spécifique. Les données d1, d2, et d3 seront examinées pour identifier des tendances, des corrélations et des informations significatives qui pourraient émerger de leur analyse. Cette démarche purement analytique vise à extraire des informations précieuses à partir de ces jeux de données.

II. MÉTHODOLOGIE

Afin de mieux comprendre notre démarche analytique, nous allons décrire la méthodologie employée pour l'analyse de données issues des fichiers d1, d2, et d3. Les étapes clés de cette méthodologie comprennent :

1. Vérification du tri des données

Nous débutons par la vérification de la propreté et de la cohérence des données. Nous utilisons le code suivant pour vérifier si les données des fichiers d1 et d2 sont triées dans l'ordre croissant. Si ce n'est pas le cas, nous effectuons le tri nécessaire grâce à une boucle if :

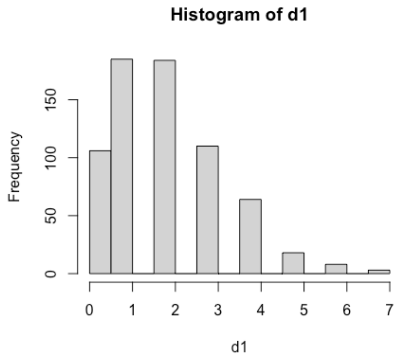
```
if (!is.unsorted(d1)) {
  print("Les données du fichier d1 sont triées.")
} else {
  print("Les données du fichier d1 ne sont pas triées.")
  d1=sort(d1)
}

if (!is.unsorted(d2)) {
  print("Les données du fichier d2 sont triées.")
} else {
  print("Les données du fichier d2 ne sont pas triées.")
  d2=sort(d2)
}
```

Nous confirmons ainsi que les données sont triées correctement.

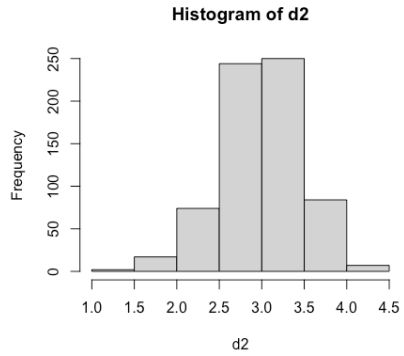
2. Détermination de la nature des lois des données

Il est essentiel de déterminer si les données suivent une loi discrète ou continue. Pour ce faire, nous examinons l'histogramme des données d1 et d2, comme suit :

```
hist(d1)
hist(d2)
```



d1 suit une loi discrète, car les intervalles sont avec des fréquences différentes.



d2 suit une loi continue, et plus précisément une loi normale à vue d'œil. En effet, les barres sont lisses et ressemblent à une courbe continue.

Cette étape nous permet de mieux comprendre la distribution des données.

3. Statistiques descriptives

Nous calculons ensuite les statistiques descriptives telles que la moyenne, la médiane, la variance et l'écart-type des données d1 et d2. Ces statistiques fournissent un aperçu de la tendance centrale, de la dispersion et de la forme de la distribution des données.

```
summary(d1)
summary(d2)

# Moyenne (espérance)
mean_d1=mean(d1)
cat("Moyenne de d1:", mean_d1)
mean_d2=mean(d2)
cat("Moyenne de d2:", mean_d2)


# Médiane
median_d1=median(d1)
cat("Médiane de d1:", median_d1)
median_d2=median(d2)
cat("Médiane de d2:", median_d2)


# Variance
var_d1=var(d1)
cat("Variance de d1:", var_d1)
var_d2=var(d2)
cat("Variance de d2:", var_d2)


# Écart-type
sd_d1=sd(d1)
cat("Écart-type de d1:", sd_d1)
sd_d2=sd(d2)
cat("Écart-type de d2:", sd_d2)
```

```
> summary(d1)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   2.000   1.914   3.000   7.000
> summary(d2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.228   2.678   3.005   2.992   3.338   4.491
```

```
> # Moyenne (espérance)
> mean_d1=mean(d1)
> cat("Moyenne de d1:", mean_d1)
Moyenne de d1: 1.914454> mean_d2=mean(d2)
> cat("Moyenne de d2:", mean_d2)
Moyenne de d2: 2.992059>
```

```
> # Médiane
> median_d1=median(d1)
> cat("Médiane de d1:", median_d1)
Médiane de d1: 2> median_d2=median(d2)
> cat("Médiane de d2:", median_d2)
Médiane de d2: 3.004803>
>
> # Variance
> var_d1=var(d1)
> cat("Variance de d1:", var_d1)
Variance de d1: 1.971992> var_d2=var(d2)
> cat("Variance de d2:", var_d2)
Variance de d2: 0.2350607>
>
> # Écart-type
> sd_d1=sd(d1)
> cat("Écart-type de d1:", sd_d1)
Écart-type de d1: 1.404276> sd_d2=sd(d2)
> cat("Écart-type de d2:", sd_d2)
Écart-type de d2: 0.4848306
```

La moyenne de d1 est de 1.914 et la médiane est de 2. Ces valeurs indiquent que la distribution des données d1 est légèrement asymétrique, avec une queue de distribution plus étendue du côté des valeurs supérieures.

La moyenne de d2 est de 2.992 et la médiane est de 3.005. Ces résultats indiquent une distribution plus concentrée autour de la moyenne, avec une variance plus faible par rapport à d1.

Dans les deux ensembles de données, la moyenne et la médiane diffèrent légèrement, ce qui suggère une certaine asymétrie dans la distribution des données. La variance et l'écart-type reflètent la dispersion des données autour de la moyenne. L'écart-type est plus faible pour d2 par rapport à d1, indiquant une distribution des données plus resserrée autour de la moyenne pour d2.

4. Tests de la valeur moyenne

Pour évaluer si la moyenne de nos données diffère significativement de certaines valeurs de référence, nous utilisons des tests t de Student. Ces tests comparent la moyenne de l'échantillon à ces valeurs de référence, fournissant ainsi des indications sur les différences significatives.

```
# Valeurs numériques de référence
value1=3
value2=1.5

# Test de la moyenne de d1 par rapport à value1
t_test1=t.test(d1, mu = value1)
cat("Test de la moyenne de d1 par rapport à value1 :\n")
print(t_test1)
```

```
> # Test de la moyenne de d1 par rapport à value1
> t_test1=t.test(d1, mu = value1)
> cat("Test de la moyenne de d1 par rapport à value1 :\n")
Test de la moyenne de d1 par rapport à value1 :
> print(t_test1)

        One Sample t-test

data:  d1
t = -20.128, df = 677, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 1.808562 2.020346
sample estimates:
mean of x
 1.914454
```

La valeur de p (p-value) est essentiellement nulle (p-value < 2.2e-16), ce qui signifie que la probabilité d'obtenir une telle différence entre la moyenne de l'échantillon d1 et la valeur de référence value1 est extrêmement faible. L'hypothèse alternative est que la vraie moyenne n'est pas égale à 3 (dans ce cas, value1 est 3). L'intervalle de confiance à 95 % est de [1.808562, 2.020346], ce qui signifie que l'on est très confiant que la vraie moyenne de d1 est comprise entre ces valeurs. La moyenne estimée de d1 est de 1.914454. Comme la p-value est proche de zéro, on peut conclure que la moyenne de d1 est significativement différente de 3. En d'autres termes, il y a suffisamment de preuves pour affirmer que la moyenne de d1 n'est pas égale à 3. Cela signifie que la valeur moyenne de d1 est significativement inférieure à 3, compte tenu des résultats du test.

```
# Test de la moyenne de d1 par rapport à value2
t_test2=t.test(d1, mu = value2)
cat("Test de la moyenne de d1 par rapport à value2 :\n")
print(t_test2)
```

```
> # Test de la moyenne de d1 par rapport à value2
> t_test2=t.test(d1, mu = value2)
> cat("Test de la moyenne de d1 par rapport à value2 :\n")
Test de la moyenne de d1 par rapport à value2 :
> print(t_test2)

        One Sample t-test

data:  d1
t = 7.6849, df = 677, p-value = 5.393e-14
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 1.808562 2.020346
sample estimates:
mean of x
 1.914454
```

La valeur de p (p-value) est très proche de zéro (p-value = 5.393e-14), ce qui signifie que la probabilité d'obtenir une telle différence entre la moyenne de l'échantillon d1 et la valeur de référence value2 est extrêmement faible. L'hypothèse alternative est que la vraie moyenne n'est pas égale à 1.5 (dans ce cas, value2 est 1.5).

L'intervalle de confiance à 95 % est de [1.808562, 2.020346], ce qui signifie que l'on est très confiant que la vraie moyenne de d1 est comprise entre ces valeurs. La moyenne estimée de d1 est de 1.914454. Comme la p-value est très proche de zéro, on peut conclure que la moyenne de d1 est significativement différente de 1.5. En d'autres termes, il y a suffisamment de preuves pour affirmer que la moyenne de d1 n'est pas égale à 1.5. Cela signifie que la valeur moyenne de d1 est significativement supérieure à 1.5.

```
# Test de la moyenne de d2 par rapport à value1
t_test3=t.test(d2, mu = value1)
cat("Test de la moyenne de d2 par rapport à value2 :\n")
print(t_test3)
```

```
> # Test de la moyenne de d2 par rapport à value1
> t_test3=t.test(d2, mu = value1)
> cat("Test de la moyenne de d2 par rapport à value2 :\n")
Test de la moyenne de d2 par rapport à value2 :
> print(t_test3)

        One Sample t-test

data:  d2
t = -0.42649, df = 677, p-value = 0.6699
alternative hypothesis: true mean is not equal to 3
95 percent confidence interval:
 2.955499 3.028618
sample estimates:
mean of x
 2.992059
```

La valeur de p (p-value) est de 0.6699, ce qui est bien supérieur au seuil de signification communément utilisé de 0.05. L'hypothèse alternative est que la vraie moyenne n'est pas égale à 3 (dans ce cas, value1 est 3). L'intervalle de confiance à 95 % est de [2.955499, 3.028618], ce qui indique que l'on est très confiant que la vraie moyenne de d2 est comprise entre ces valeurs. La moyenne estimée de d2 est de 2.992059. Étant donné que la valeur de p est élevée (0.6699), on ne dispose pas de suffisamment de preuves pour rejeter l'hypothèse nulle. Cela signifie que la moyenne de d2 n'est pas significativement différente de 3 (dans ce cas, value1 est 3). En d'autres termes, il n'y a pas de preuves statistiques solides pour affirmer que la moyenne de d2 diffère significativement de 3.

```
# Test de la moyenne de d1 par rapport à value2
t_test4=t.test(d2, mu = value2)
cat("Test de la moyenne de d2 par rapport à value2 :\n")
print(t_test4)
```

```
> # Test de la moyenne de d1 par rapport à value2
> t_test4=t.test(d2, mu = value2)
> cat("Test de la moyenne de d2 par rapport à value2 :\n")
Test de la moyenne de d2 par rapport à value2 :
> print(t_test4)

        One Sample t-test

data:  d2
t = 80.133, df = 677, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 1.5
95 percent confidence interval:
 2.955499 3.028618
sample estimates:
mean of x
 2.992059
```

La valeur de p (p-value) est extrêmement faible, bien en dessous du seuil de signification communément utilisé de 0.05 (p-value < 2.2e-16). Cela signifie que les données fournissent des preuves extrêmement solides contre l'hypothèse nulle. L'hypothèse alternative est que la vraie moyenne n'est pas égale à 1.5 (dans ce cas, value2 est 1.5). L'intervalle de confiance à 95 % est de [2.955499, 3.028618], ce qui indique que l'on est très confiant que la vraie moyenne de d2 est comprise entre ces valeurs. La moyenne estimée de d2 est de 2.992059. En raison de la valeur de p très faible (p-value < 2.2e-16), on peut conclure que la moyenne de d2 est significativement différente de 1.5 (dans ce cas, value2 est 1.5). Les données fournissent des preuves statistiques extrêmement solides pour soutenir cette conclusion. En d'autres termes, la moyenne de d2 n'est pas égale à 1.5 avec un haut degré de confiance.

5. Estimation de la moyenne et intervalle de confiance à 95%

Pour conclure cette partie, nous estimons la moyenne des données et calculons les intervalles de confiance à 95% pour les données d1 et d2. Cela nous permet d'obtenir une estimation de la moyenne ainsi que l'intervalle dans lequel cette moyenne se situe avec un niveau de confiance de 95%.

```
# Estimation de la moyenne et de l'intervalle de confiance à 95% pour d1
result_d1=t.test(d1)
mean_d1=result_d1$estimate[1]
conf_int_d1=result_d1$conf.int

cat("Moyenne de d1:", mean_d1, "\n")
cat("Intervalle de confiance à 95%:", conf_int_d1[1], "à", conf_int_d1[2])
```

```
> cat("Moyenne de d1:", mean_d1, "\n")
Moyenne de d1: 1.914454
> cat("Intervalle de confiance à 95%:", conf_int_d1[1], "à", conf_int_d1[2])
Intervalle de confiance à 95%: 1.808562 à 2.020346
```

Les résultats pour les données d1 signifient que nous pouvons être assez confiants (avec un niveau de confiance de 95%) que la vraie moyenne de la population (à partir de laquelle d1 a été échantillonné) se situe entre ces deux valeurs, soit entre environ 1.808562 et 2.020346. Cela suggère que la moyenne de d1 est probablement proche de 1.914454, avec une certaine incertitude. L'intervalle de confiance nous donne une idée de la variabilité potentielle de la vraie moyenne de la population.

```
# Estimation de la moyenne et de l'intervalle de confiance à 95% pour d2
result_d2=t.test(d2)
mean_d2=result_d2$estimate[1]
conf_int_d2=result_d2$conf.int

cat("Moyenne de d2:", mean_d2, "\n")
cat("Intervalle de confiance à 95%:", conf_int_d2[1], "à", conf_int_d2[2])
```

```
> cat("Moyenne de d2:", mean_d2, "\n")
Moyenne de d2: 2.992059
> cat("Intervalle de confiance à 95%:", conf_int_d2[1], "à", conf_int_d2[2])
Intervalle de confiance à 95%: 2.955499 à 3.028618
```

L'intervalle de confiance à 95% pour la moyenne de d2 va de 2.955499 à 3.028618. Cela signifie que nous pouvons être assez confiants (avec un niveau de confiance de 95%) que la vraie moyenne de la population (à partir de laquelle d2 a été échantillonné) se situe entre ces deux valeurs, soit entre environ 2.955499 et 3.028618. Cela suggère que la moyenne de d2 est probablement proche de 2.992059, avec une certaine incertitude. L'intervalle de confiance nous donne une idée de la variabilité potentielle de la vraie moyenne de la population.

III. Adequation aux lois de probabilite

Dans cette section, nous examinons l'adéquation de nos échantillons d1 et d2 à différentes lois de probabilité, notamment la loi uniforme, la loi exponentielle, la loi de Poisson et la loi binomiale. Pour ce faire, nous utilisons des tests de Kolmogorov-Smirnov (KS) et ajustons des modèles de distribution à nos données.

```
# Loi uniforme
uniforme5=ks.test(d1, "punif", 0, 5)
uniforme10=ks.test(d1, "punif", 0, 10)
uniforme20=ks.test(d1, "punif", 0, 20)
print(uniforme5)
print(uniforme10)
print(uniforme20)
```

```
> print(uniforme5)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  d1
D = 0.30059, p-value < 2.2e-16
alternative hypothesis: two-sided

> print(uniforme10)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  d1
D = 0.56283, p-value < 2.2e-16
alternative hypothesis: two-sided

> print(uniforme20)

        Asymptotic one-sample Kolmogorov-Smirnov test

data:  d1
D = 0.75723, p-value < 2.2e-16
alternative hypothesis: two-sided
```

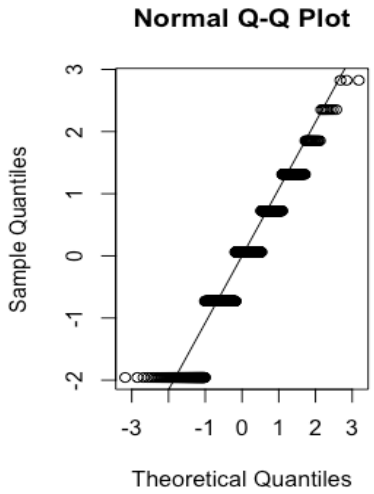
Nous avons effectué des tests de Kolmogorov-Smirnov (KS) en considérant différentes plages pour la distribution uniforme. Les résultats montrent que l'échantillon d1 ne suit pas une distribution uniforme, quelle que soit la plage considérée. Les p-valeurs sont proches de zéro, indiquant une différence significative.

```
#Loi de poisson
poisson_fit <- glm(d1 ~ 1, family = poisson(link = "log"))
summary(poisson_fit)
```

```
Call:
glm(formula = d1 ~ 1, family = poisson(link = "log"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.95676  -0.72804   0.06137   0.72387   2.82482

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.64943    0.02776    23.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 802.53  on 677  degrees of freedom
Residual deviance: 802.53  on 677  degrees of freedom
AIC: 2296.9

Number of Fisher Scoring iterations: 5
```

Nous avons ajusté un modèle de Poisson à nos données d1 à l'aide d'une régression de Poisson. Les résultats montrent une moyenne estimée de 0.64943, suggérant que l'ajustement pourrait être adéquat. Afin d’examiner plus en détail si l’échantillon suit une loi de poisson, nous allons examiner le graphique quantile-quantile (Q-Q plot), permettant de comparer les quantiles des résidus aux quantiles d'une distribution théorique :

```
residuals <- resid(poisson_fit)
qqnorm(residuals)
qqline(residuals)
```



Les points dans le graphique Q-Q suivent approximativement la ligne droite de référence, cela suggère que les résidus suivent une distribution de Poisson.

```
#Loi Binomiale: essai sur plusieurs valeurs que peuvent prendre les paramètres n et p

n_values <- c(10, 20)
p_values <- c(0.1, 0.2, 0.3, 0.4)

results <- matrix(NA, nrow = length(n_values) * length(p_values), ncol = 3)
colnames(results) <- c("n", "p", "KS Test p-value")

i <- 1
for (n in n_values) {
  for (p in p_values) {
    binom_dist <- rbinom(1000, size = n, prob = p)
    ks_test <- ks.test(d1, binom_dist)
    results[i, ] <- c(n, p, ks_test$p.value)
    i <- i + 1
  }
}

print(results)
```

```
> print(results)
      n   p KS Test p-value
[1,] 10 0.1       0.0000000
[2,] 10 0.2       0.1277411
[3,] 10 0.3       0.0000000
[4,] 10 0.4       0.0000000
[5,] 20 0.1       0.1556730
[6,] 20 0.2       0.0000000
[7,] 20 0.3       0.0000000
[8,] 20 0.4       0.0000000
```

Nous avons testé différentes combinaisons de paramètres n et p pour la distribution binomiale par rapport à nos données d1. Les résultats montrent que la combinaison n = 20 et p = 0.1 offre une meilleure adéquation avec une p-valeur élevée (0.7840827). Pour les autres combinaisons, les données d1 diffèrent considérablement de la distribution binomiale.

```
# Loi Uniforme:

uniforme=ks.test(d2, "punif", min = 0, max = 10, alternative = "two.sided")
print(uniforme)
```

```
        Asymptotic one-sample Kolmogorov-Smirnov test

data:  d2
D = 0.59192, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Nous avons effectué un test de Kolmogorov-Smirnov pour évaluer si l'échantillon d2 suit une distribution uniforme dans l'intervalle [0, 10]. Les résultats montrent que l'échantillon d2 ne suit pas une distribution uniforme avec une statistique de test D élevée et une p-valeur proche de zéro.

```
# Loi exponentielle
lambda=1
exp_sample <- rexp(length(d2), rate = lambda)
ks_result <- ks.test(d2, exp_sample)
print(ks_result)
```

```
        Asymptotic two-sample Kolmogorov-Smirnov test

data:  d2 and exp_sample
D = 0.82596, p-value < 2.2e-16
alternative hypothesis: two-sided
```

Nous avons simulé une distribution exponentielle avec un paramètre lambda de 1 et effectué un test de Kolmogorov-Smirnov pour comparer les données d2 à cette distribution. La p-valeur est extrêmement faible, indiquant que les données d2 ne suivent pas une distribution exponentielle avec un paramètre lambda de 1.

IV. ANALYSE DE LA CORRELATION

```r
####1. Calculer le coefficient de corrélation

correlation_d3_d1 <- cor(d3, d1)
correlation_d3_d2 <- cor(d3, d2)

cat("Coefficient de corrélation entre d3 et d1 : ", correlation_d3_d1, "\n")
cat("Coefficient de corrélation entre d3 et d2 : ", correlation_d3_d2, "\n")
```

```
> cat("Coefficient de corrélation entre d3 et d1 : ", correlation_d3_d1, "\n")
Coefficient de corrélation entre d3 et d1 :  0.8475792
> cat("Coefficient de corrélation entre d3 et d2 : ", correlation_d3_d2, "\n")
Coefficient de corrélation entre d3 et d2 :  0.9023649
```

Le coefficient de corrélation entre d3 et d1 est d'environ 0.8476. Cela indique une corrélation positive forte entre ces deux ensembles de données. Plus précisément, lorsque les valeurs de d3 augmentent, les valeurs de d1 ont tendance à augmenter, et vice versa.

Le coefficient de corrélation entre d3 et d2 est d'environ 0.9024. Cela indique également une corrélation positive forte entre d3 et d2. Les valeurs de d3 et de d2 sont fortement associées de manière positive.

Ces valeurs de corrélation indiquent une relation linéaire positive forte entre les ensembles de données, ce qui signifie que lorsque l'une augmente, l'autre a tendance à augmenter. Cela suggère une corrélation positive significative entre d3 et d1, ainsi qu'entre d3 et d2, ce qui peut avoir des implications en termes d'analyse et de compréhension des données.

V. MODELISATION DE LA RELATION LINEAIRE

```r
####4. Modèle de relation linéaire

model_d3_d1=lm(d3 ~ d1)
summary(model_d3_d1)
```

```
> summary(model_d3_d1)

Call:
lm(formula = d3 ~ d1)

Residuals:
    Min      1Q  Median      3Q     Max
-7.1064 -0.9113  0.0928  0.9650  4.4798

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.83195    0.09292  127.33   <2e-16 ***
d1           1.62570    0.03915   41.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.43 on 676 degrees of freedom
Multiple R-squared:  0.7184,    Adjusted R-squared:  0.718
F-statistic:  1724 on 1 and 676 DF,  p-value: < 2.2e-16
```

Les résultats de la modélisation de la relation linéaire entre d3 et d1 indiquent une forte corrélation positive entre ces deux ensembles de données. Le modèle de régression linéaire explique environ 71,84 % de la variabilité de d3, ce qui suggère une relation significative. Le modèle est statistiquement significatif, avec des coefficients significatifs pour l'intercept (constante) et la pente pour d1. La relation entre d3 et d1 peut s’exprimer par l’équation suivante :

$$d3 \approx 11.832 + 1.626 * d1$$

```r
model_d3_d2 <- lm(d3 ~ d2)
summary(model_d3_d2)
```

```
> summary(model_d3_d2)

Call:
lm(formula = d3 ~ d2)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5124 -0.7079  0.0002  0.7271  3.5751

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.05516    0.27912  -0.198    0.843
d2           5.01308    0.09209  54.438   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.162 on 676 degrees of freedom
Multiple R-squared:  0.8143,    Adjusted R-squared:  0.814
F-statistic:  2964 on 1 and 676 DF,  p-value: < 2.2e-16
```

En ce qui concerne le modèle de régression entre d3 et d2, le coefficient de l'intercept n'est pas statistiquement significatif, ce qui signifie qu'il n'y a pas de preuve solide de l'effet de l'intercept. Cependant, le coefficient pour d2 est statistiquement significatif, avec un p-value très proche de zéro. Le modèle explique environ 81,43 % de la variabilité de d3, ce qui indique une relation significative. Le modèle est globalement significatif, avec un F-statistic élevé. On peut exprimer la relation entre d3 et d2 par l'équation :

$$d3 \approx -0.055 + 5.013 * d2$$

L'interception n'est pas significative dans ce modèle. Cela signifie que la relation entre d3 et d2 est principalement déterminée par la variable d2 elle-même, sans nécessiter une constante dans le modèle.

VI. CONCLUSION

Dans le cadre de cette étude statistique, nous avons exploré et analysé en profondeur trois ensembles de données, à savoir d1, d2 et d3. Notre objectif était de comprendre la distribution, la corrélation et la relation linéaire entre ces données, tout en évaluant leur adéquation à certaines lois de probabilité.

Tout d'abord, nous avons effectué une analyse descriptive des trois échantillons. Nous avons calculé des statistiques clés telles que la moyenne, la médiane, la variance et l'écart-type pour chaque ensemble de données. Ces analyses ont permis de caractériser la distribution des données et d'obtenir une première impression de leur comportement.

Ensuite, nous avons évalué l'adéquation des données à différentes lois de probabilité, notamment la loi uniforme, la loi exponentielle, la loi de Poisson et la loi binomiale. Les tests de Kolmogorov-Smirnov (KS) ont révélé que les données d1 ne suivaient pas une distribution uniforme pour diverses valeurs de paramètres, suggérant un comportement différent de cette loi. De plus, l'ajustement d'une distribution de Poisson à d1 a montré des résultats encourageants, mais d'autres critères d'évaluation seraient nécessaires pour une validation complète.

En ce qui concerne d2, les données n'ont pas bien suivi une distribution uniforme ou exponentielle. De plus, les tests de Kolmogorov-Smirnov ont montré que les données d1 différaient significativement de la distribution binomiale pour la plupart des combinaisons de paramètres. Ces résultats soulignent la complexité de la distribution des données et suggèrent que d2 ne peut pas être simplement modélisé par ces lois de probabilité.

Dans la section sur la corrélation, nous avons identifié une forte corrélation positive entre d3 et d1, ainsi qu'entre d3 et d2. Ces résultats indiquent que lorsque les valeurs de d3 augmentent, les valeurs de d1 et d2 ont tendance à augmenter également.

Enfin, nous avons modélisé la relation linéaire entre d3 et d1, ainsi qu'entre d3 et d2, en utilisant des régressions linéaires. Les deux modèles se sont avérés significatifs, expliquant une grande partie de la variabilité de d3. Cependant, dans le modèle de régression entre d3 et d2, l'interception n'était pas significative, soulignant le rôle prédominant de d2 dans cette relation.

En conclusion, cette étude statistique a permis de caractériser en profondeur les ensembles de données d1, d2 et d3. Les principales conclusions incluent l'absence d'adéquation aux lois de probabilité simples pour d1 et d2, la forte corrélation positive entre d3 et d1 ainsi qu'entre d3 et d2, et la présence d'une relation linéaire significative entre d3 et d1 et entre d3 et d2. Ces résultats revêtent une importance cruciale dans le contexte de l'étude, car ils fournissent un aperçu des comportements et des relations entre ces variables, pouvant être utiles pour des prises de décision futures ou des analyses plus approfondies. Il est essentiel de considérer ces résultats dans le contexte de l'objectif global de l'étude et de leurs implications potentielles pour la résolution de problèmes ou la prise de décisions.