

Examen de statistiques et analyse de données

Antoine Géré

10 décembre 2025

Table des matières

Exercice 1.....	1
Exercice 2.....	8
Références.....	14

- Document(s) autorisé(s) : **Non**
- Calculatrice autorisée : **Oui**
- Remarques :
 - Les exercices sont indépendants.
 - Il sera tenu compte de la propreté de votre copie, ainsi que de la clarté et de la qualité de la rédaction et du raisonnement.
 - Ne pas écrire avec un crayon papier, sauf pour dessiner et/ou annoter des croquis, le cas échéant.
 - Utiliser les notations indiquées dans le texte et justifier toutes vos réponses.

Exercice 1

Le coucou européen ne couve pas ses propres oeufs, mais les dépose dans les nids d'oiseaux d'autres espèces. Des études antérieures ont montré que les coucous ont parfois évolué pour pondre des oeufs dont la couleur est similaire à celle des oeufs de l'espèce hôte.

Est-ce également le cas pour la taille des oeufs ? Les coucous pondent-ils des oeufs de taille similaire à ceux de leurs hôtes ?



Le fichier de données est `cuckooeggs.csv`

```
cuckooeggs = read.csv("data_raw/cuckooeggs.csv")
head(cuckooeggs)

  host_species egg_length
1 Hedge Sparrow      20.85
2 Hedge Sparrow      21.65
3 Hedge Sparrow      22.05
4 Hedge Sparrow      22.85
5 Hedge Sparrow      23.05
6 Hedge Sparrow      23.05

summary(cuckooeggs)

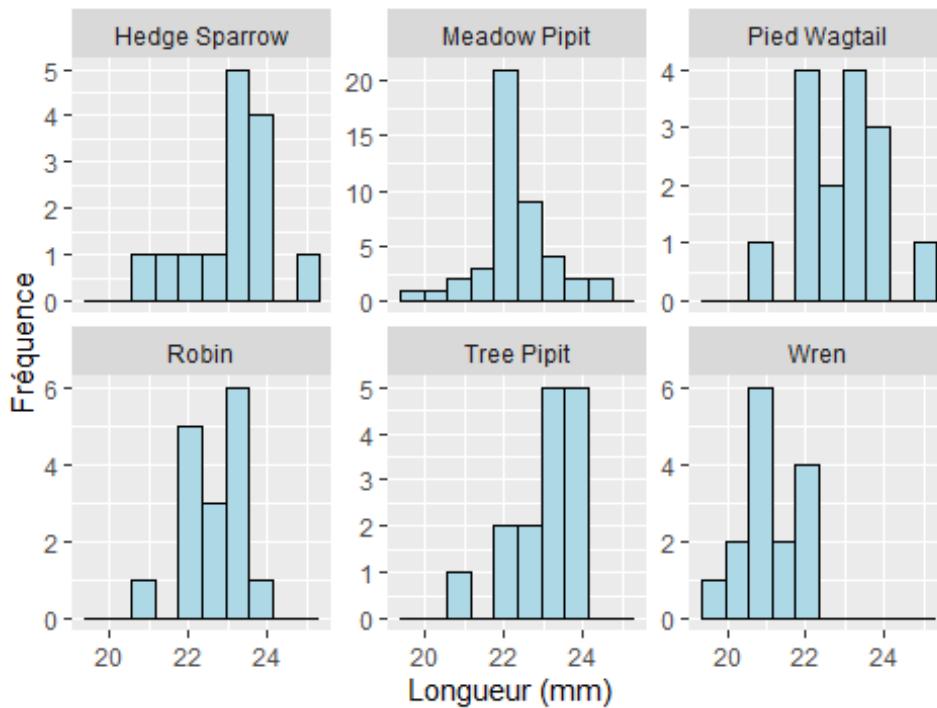
host_species      egg_length
Length:120      Min.   :19.65
Class :character 1st Qu.:21.85
Mode  :character Median :22.35
                  Mean   :22.46
                  3rd Qu.:23.25
                  Max.   :25.05
```

1. Le fichier de données

Présentez le fichier de données en indiquant le nombre de variables, leur type ainsi que leurs principales caractéristiques.

Nous traçons quelques graphiques ci-dessous.

```
library(ggplot2)
ggplot(cuckooeggs, aes(x = egg_length)) +
  geom_histogram(bins = 10, fill = "lightblue", color = "black") +
  facet_wrap(~ host_species, scales = "free_y") +
  labs(
    title = "",
    x = "Longueur (mm)", y = "Fréquence"
  )
```



2. Interprétation des graphes ci-dessus

- Comment se nomme ce type de graphe ?
- Décrivez les différences visibles entre les distributions des longueurs d'oeufs selon les espèces hôtes.
- Les variances semblent-elles comparables ?
- Les distributions paraissent-elles approximativement normales ?
- Les moyennes semblent-elles clairement différentes ?

On demande à R de nous donner les moyennes et les écarts types de la longueur des oeufs pour chacune des espèces hôtes.

```
stats <- cuckooeggs %>%
  group_by(host_species) %>%
  summarise(
    Moyenne = mean(egg_length),
    Ecart_type = sd(egg_length),
    n = n()
  )
stats

# A tibble: 6 × 4
  host_species    Moyenne   Ecart_type     n
  <chr>           <dbl>        <dbl> <int>
1 Hedge Sparrow  23.1        1.07      14
2 Meadow Pipit   22.3        0.921     45
3 Pied Wagtail   22.9        1.07      15
```

4 Robin	22.6	0.685	16
5 Tree Pipit	23.1	0.901	15
6 Wren	21.1	0.744	15

3. Interprétation du tableau descriptif

- a. Quelle espèce hôte présente les œufs les plus grands en moyenne ?
- b. Quelle espèce présente les œufs les plus petits ?
- c. Les différences de moyennes vous semblent-elles importantes ?
- d. Les écarts-types semblent-ils homogènes ?

On se propose, afin de savoir si les coucous pondent des oeufs de tailles différentes selon l'espèce hôte, de réaliser un **T-test**.

```
t.test(egg_length ~ host_species, data = cuckooeggs)
```

4. Interprétation d'une erreur

Expliquez la raison pour laquelle ce `t.test` ne fonctionne pas dans ce cas précis.

On tente alors de réaliser une ANOVA.

5. Conditions d'application de l'ANOVA

- a. Rappelez les conditions nécessaires à l'ANOVA à un facteur.
- b. En vous appuyant sur les histogrammes et le tableau des moyennes et écarts types, indiquez si elles sont raisonnablement satisfaites.
- c. Rappelez l'hypothèse nulle pour l'ANOVA

Sur conseil d'un collaborateur on réalise un `leveneTest`.

```
leveneTest(egg_length ~ host_species, data = cuckooeggs)
```

```
Warning in leveneTest.default(y = y, group = group, ...): group coerced to factor.
```

```
Levene's Test for Homogeneity of Variance (center = median)
```

```
    Df F value Pr(>F)
```

```
group   5  0.6397 0.6698
```

```
114
```

6. Levene test

- a. Donner l'hypothèse nulle pour ce test.
- b. Interpréter ce test, que pouvez vous en conclure ?

Ce collaborateur nous recommande vivement d'effectuer un `shapiro.test`. Nous en réalisons donc plusieurs.

```
by(cuckooeggs$egg_length, cuckooeggs$host_species, shapiro.test)
```

```
cuckooeggs$host_species: Hedge Sparrow
```

```
Shapiro-Wilk normality test
```

```
data: dd[x, ]
W = 0.94843, p-value = 0.5366

-----
cuckooeggs$host_species: Meadow Pipit

  Shapiro-Wilk normality test

data: dd[x, ]
W = 0.93006, p-value = 0.009424

-----
cuckooeggs$host_species: Pied Wagtail

  Shapiro-Wilk normality test

data: dd[x, ]
W = 0.96471, p-value = 0.7736

-----
cuckooeggs$host_species: Robin

  Shapiro-Wilk normality test

data: dd[x, ]
W = 0.95212, p-value = 0.5239

-----
cuckooeggs$host_species: Tree Pipit

  Shapiro-Wilk normality test

data: dd[x, ]
W = 0.89772, p-value = 0.08786

-----
cuckooeggs$host_species: Wren

  Shapiro-Wilk normality test

data: dd[x, ]
W = 0.93295, p-value = 0.3019

anova_res <- aov(egg_length ~ host_species, data = cuckooeggs)
shapiro.test(residuals(anova_res))
```

Shapiro-Wilk normality test

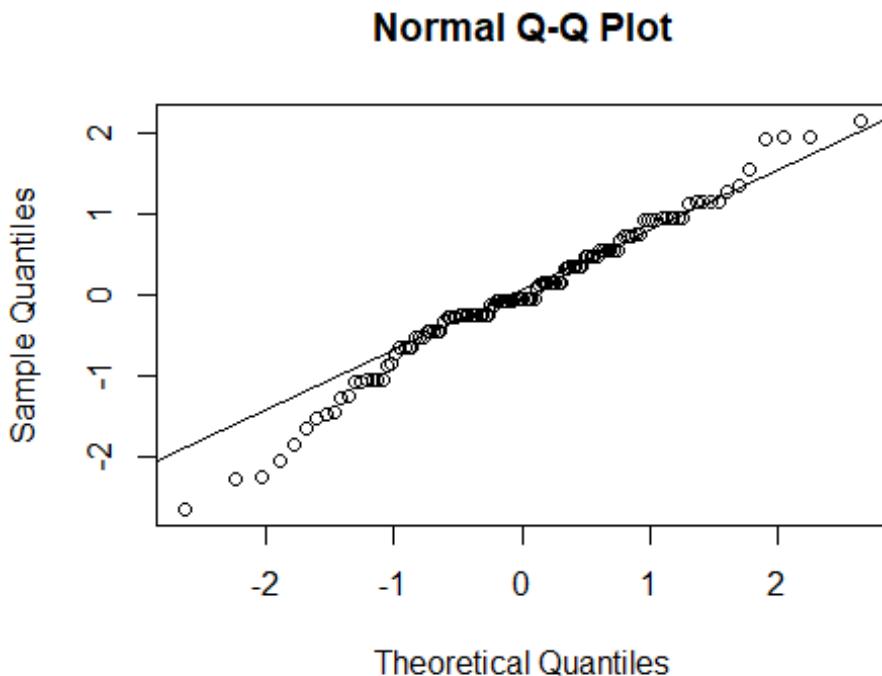
```
data: residuals(anova_res)
W = 0.9804, p-value = 0.07761
```

7. Shapiro test

- Donnez l'hypothèse nulle pour le shapiro.test.
- Interprétez chacun de ces shapiro.test.
- Etaient ils tous nécessaire ?
- Que pouvez vous en conclure pour l'ANOVA ?

Une fois nos shapiro.test réalisés, un autre collaborateur examine notre travail et nous indique qu'il aurait été pertinent de produire un qqplot. Nous tentons alors de tracer ce graphique.

```
qqnorm(residuals(anova_res))
qqline(residuals(anova_res))
```

**8. Q-Q plot**

- Rappelez la définition d'un q-q plot et expliquez en quoi l'observation de ce collaborateur était pertinente.
- Analysez et interprétez ce graphique.
- Ces résultats concordent-ils avec le test de Shapiro précédemment effectué ?

```
summary(anova_res)
```

```
Df Sum Sq Mean Sq F value    Pr(>F)
host_species   5  42.94   8.588   10.39 3.15e-08 ***
Residuals     114  94.25   0.827
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

9. Résultat de l'ANOVA

- D'après les résultats de l'ANOVA, peut-on conclure que l'effet de l'espèce hôte sur la taille des œufs est statistiquement significatif ?
- L'ANOVA permet-elle de déterminer exactement quelles espèces présentent des effets différents ?

Toujours le même collaborateur nous parle d'effectuer un test post hoc. On le réalise sans trop savoir ce que l'on est en train de tester.

```
tukey <- TukeyHSD(anova_res)
tukey

Tukey multiple comparisons of means
 95% family-wise confidence level

Fit: aov(formula = egg_length ~ host_species, data = cuckooeggs)
```

\$host_species	diff	lwr	upr	p	adj
Meadow Pipit-Hedge Sparrow	-0.82253968	-1.629133605	-0.01594576	0.0428621	
Pied Wagtail-Hedge Sparrow	-0.21809524	-1.197559436	0.76136896	0.9872190	
Robin-Hedge Sparrow	-0.54642857	-1.511003196	0.41814605	0.5726153	
Tree Pipit-Hedge Sparrow	-0.03142857	-1.010892769	0.94803563	0.9999990	
Wren-Hedge Sparrow	-1.99142857	-2.970892769	-1.01196437	0.0000006	
Pied Wagtail-Meadow Pipit	0.60444444	-0.181375330	1.39026422	0.2324603	
Robin-Meadow Pipit	0.27611111	-0.491069969	1.04329219	0.9021876	
Tree Pipit-Meadow Pipit	0.79111111	0.005291337	1.57693089	0.0474619	
Wren-Meadow Pipit	-1.16888889	-1.954708663	-0.38306911	0.0004861	
Robin-Pied Wagtail	-0.32833333	-1.275604766	0.61893810	0.9155004	
Tree Pipit-Pied Wagtail	0.18666667	-0.775762072	1.14909541	0.9932186	
Wren-Pied Wagtail	-1.77333333	-2.735762072	-0.81090459	0.0000070	
Tree Pipit-Robin	0.51500000	-0.432271433	1.46227143	0.6159630	
Wren-Robin	-1.44500000	-2.392271433	-0.49772857	0.0003183	
Wren-Tree Pipit	-1.96000000	-2.922428738	-0.99757126	0.0000006	

10. Test post-hoc - Tukey test

- Rappelez l'objectif d'un test post-hoc après une ANOVA.
- D'après les résultats du test de Tukey, quelles paires d'espèces présentent des différences significatives ?
- Quelles conclusions peut-on tirer de ces résultats ?

Exercice 2

Contexte et description du fichier de données

Le jeu de données utilisé dans cet examen est inspiré d'une étude publiée dans la revue Cell, Petropoulos et al. (2016), où des chercheurs ont mesuré l'activité de nombreux gènes dans des cellules provenant d'embryons humains très précoce. Leur objectif était d'observer comment les cellules évoluent au cours du développement et comment elles se différencient en plusieurs types cellulaires.

Pour notre exercice, nous utilisons une version simplifiée de ce type de données. Le fichier contient :

- 300 gènes (chaque ligne correspond à un gène différent)
- 12 échantillons (chaque colonne correspond à une condition biologique), représentant différents jours de développement et différents types de cellules :
 - Exemples : Day_5_EPI, Day_5_TE, Day_6_PE, etc.
- Les variables du fichier sont donc :
 - Gene : nom du gène
 - 12 variables quantitatives (une par condition), correspondant à l'expression du gène dans chaque échantillon (mesure numérique)

L'objectif de l'Analyse en composantes principales est d'identifier les grands axes qui résument les principales différences d'expression entre ces 12 conditions, et de visualiser si certaines conditions se regroupent selon des profils similaires.

Le fichier de données est mmc3.xlsx.

```
data = read_excel("mmc3.xlsx")
head(data)

# A tibble: 6 × 13
  Gene    Day_3 Day_4 Pre_Day_5 Day_5_EPI Day_5_PE Day_5_TE Day_6_EPI
  <chr>   <dbl> <dbl>     <dbl>      <dbl>     <dbl>      <dbl>      <dbl>
1 SEPT6    1.5    0       0.2       2.2       0.6       0.3       9.8
2 AARS2    2.3   13.8     17.3      44.1      19.5      8.5      21.8
3 ABCG2    0.2    0.3      7.6       3.3      10.4      35.1      23       54
4 ABHD12B   5.8    11      16.7      33.3      17       12.3      57.4
5 ABHD6    0.8    4.2      9.7       7.5      29.2      8.1      11.8
6 ACE      0.3   17.5     10.3      15.4      6.7       3.9      11.1
# i 4 more variables: Day_6_TE <dbl>, Day_7_EPI <dbl>, Day_7_PE <dbl>,
#   Day_7_TE <dbl>
```

```
var.quanti = sapply(data, is.numeric)
data.quanti = data[, var.quanti]
head(data.quanti)

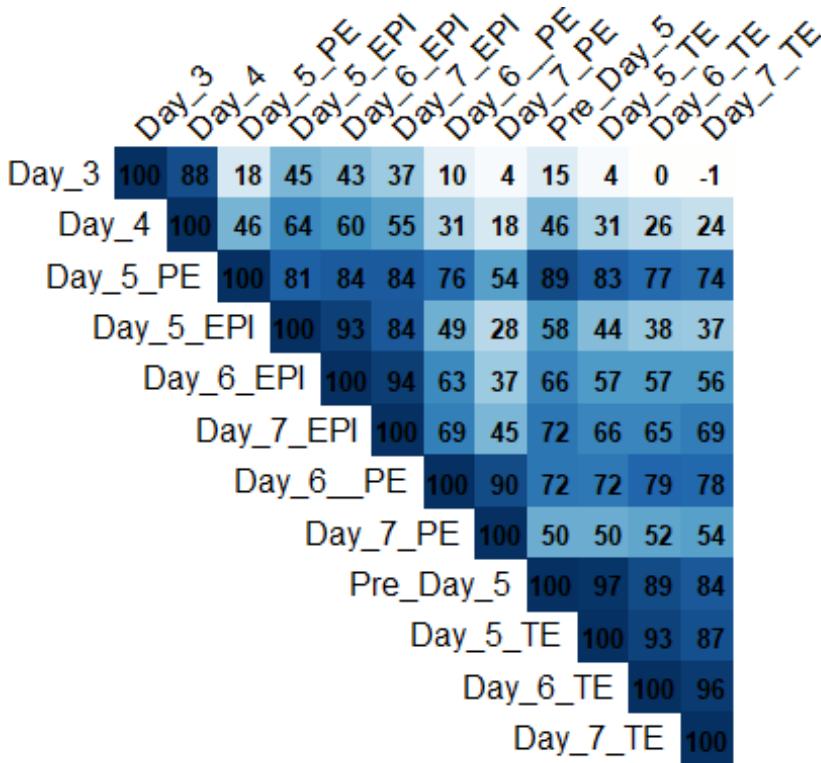
# A tibble: 6 × 12
  Day_3 Day_4 Pre_Day_5 Day_5_EPI Day_5_PE Day_5_TE Day_6_EPI Day_6__PE
  <dbl> <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>     <dbl>
1    1.5    0      0.2      2.2      0.6      0.3      9.8      0.7
2    2.3   13.8    17.3    44.1     19.5      8.5     21.8      7.5
3    0.2    0.3     7.6      3.3     10.4     35.1      23      54
4    5.8    11     16.7     33.3      17     12.3     57.4     18.3
5    0.8    4.2     9.7      7.5     29.2      8.1     11.8     14.8
6    0.3   17.5    10.3     15.4      6.7      3.9     11.1      1.3
# i 3 more variables: Day_7_EPI <dbl>, Day_7_PE <dbl>, Day_7_TE <dbl>
```

1. Question théorique

- a. Combien de gènes et de conditions contient le jeu de données ?
- b. Quelles sont les variables quantitatives ?
- c. Pourquoi est-il important de distinguer les variables quantitatives et qualitatives avant l'analyse ?
- d. Expliquez en quelques phrases l'objectif de l'ACP. Pourquoi cette méthode est-elle utilisée ?

```
Matrice.Correlation <- cor(data.quanti, use = "complete.obs")
```

```
corrplot(Matrice.Correlation,
         method = "color",
         type = "upper",
         order = "hclust",
         tl.col = "black",
         tl.srt = 45,
         addCoef.col = "black",
         cl.pos = "n",
         cl.cex = 1.2,
         addCoefasPercent = TRUE,
         number.cex = 0.8)
```



```
data.quant = data[, var.quant]
```

2. Matrice de corrélation

- Quelles conditions semblent fortement corrélées ? Quelles conditions semblent peu corrélées ?
- Que peut-on déduire de ces corrélations pour l'analyse en composantes principales ?

```
data.CR <- scale(data.quant, center = TRUE, scale=TRUE)
```

```
pca.data <- PCA(data.CR, graph = FALSE)
```

```
vp = pca.data$eig
```

```
vp
```

	eigenvalue	percentage of variance	cumulative percentage of variance	percentage of variance
comp 1	7.585892122	63.21576768		63.21577
comp 2	2.217238058	18.47698382		81.69275
comp 3	0.873147907	7.27623255		88.96898
comp 4	0.776224795	6.46853996		95.43752
comp 5	0.293172253	2.44310211		97.88063
comp 6	0.097438066	0.81198388		98.69261
comp 7	0.072100242	0.60083535		99.29345
comp 8	0.046325556	0.38604630		99.67949
comp 9	0.016444017	0.13703347		99.81653
comp 10	0.013633402	0.11361169		99.93014
comp 11	0.005016309	0.04180257		99.97194
comp 12	0.003367273	0.02806061		100.00000

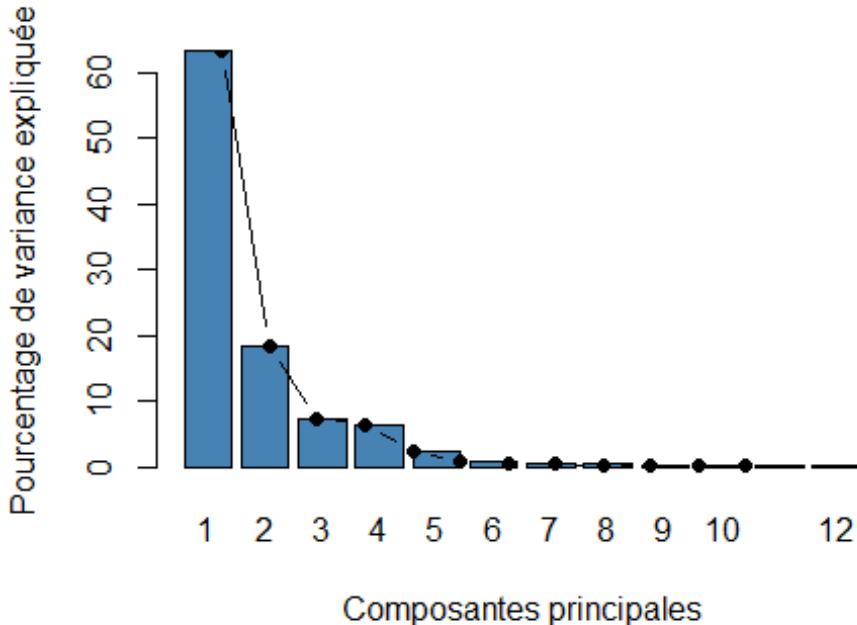
```
barplot(vp[, 2],  
       names.arg=1:nrow(vp),
```

```

main = "",
xlab = "Composantes principales",
ylab = "Pourcentage de variance expliquée",
col ="steelblue")

lines(x = 1:nrow(vp),
      vp[, 2],
      type="b",
      pch=19,
      col = "black")

```



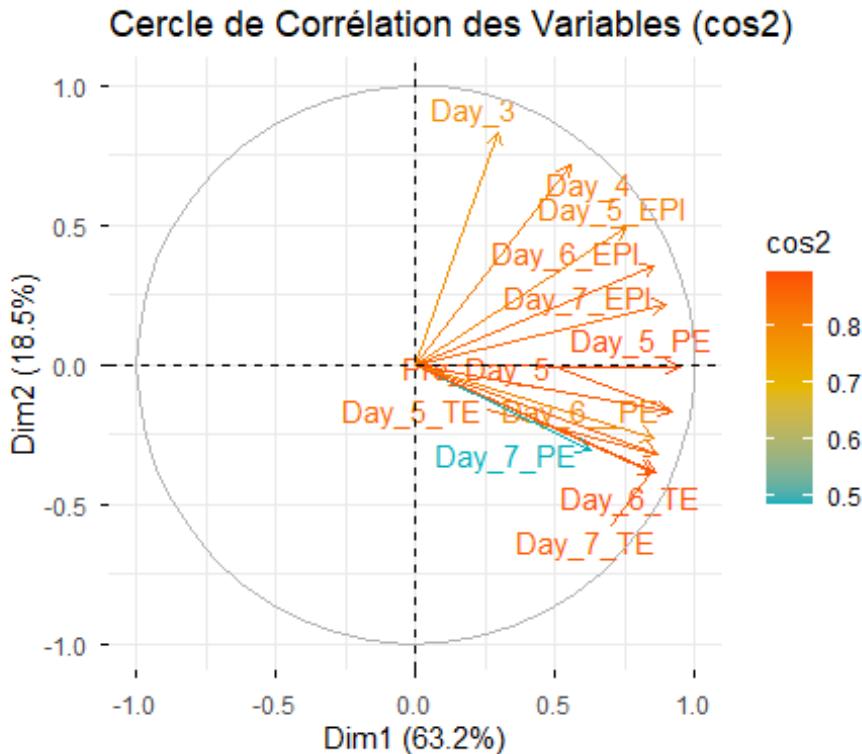
3. Variance expliquée

- Que représente la variance expliquée par chaque composante principale ?
- Combien de composantes principales retiendriez-vous pour représenter l'essentiel de l'information ? Justifiez votre réponse.
- Quelle proportion de variance est expliquée par les deux premières composantes ?

```

fviz_pca_var(pca.data,
              col.var = "cos2",
              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
              repel = TRUE,
              title = "Cercle de Corrélation des Variables (cos2)")

```



4. Cercle de corrélation

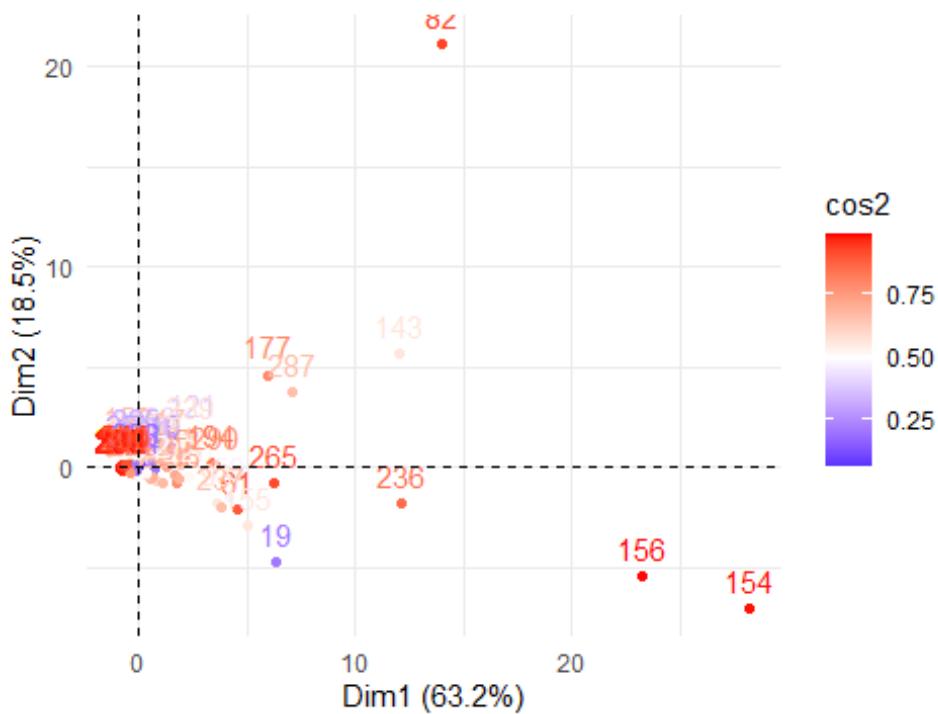
- Identifiez deux variables fortement corrélées et deux variables faiblement corrélées.
- Que signifie la longueur et l'orientation des flèches ?
- Quelles variables ont une contribution importante sur l'axe 2?

```
head(pca.data$ind$coord)
```

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
1	-0.8860388	-0.05653948	-0.05987579	0.0326202352	0.03833300
2	-0.6057873	0.10852636	-0.12243416	-0.0690008780	-0.10503767
3	-0.3699260	-0.31957924	0.03212768	0.1043982993	0.31567397
4	-0.5063049	0.11052344	-0.11409912	-0.1403434496	0.02666491
5	-0.6891486	-0.07722587	-0.02877404	0.0005839167	-0.07581851
6	-0.7792027	0.06145232	-0.06881027	0.0769645114	-0.05851516

```
fviz_pca_ind(pca.data, col.ind="cos2") +
  scale_color_gradient2(low="blue", mid="white", high="red", midpoint=0.50) +
  theme_minimal()
```

Individuals - PCA

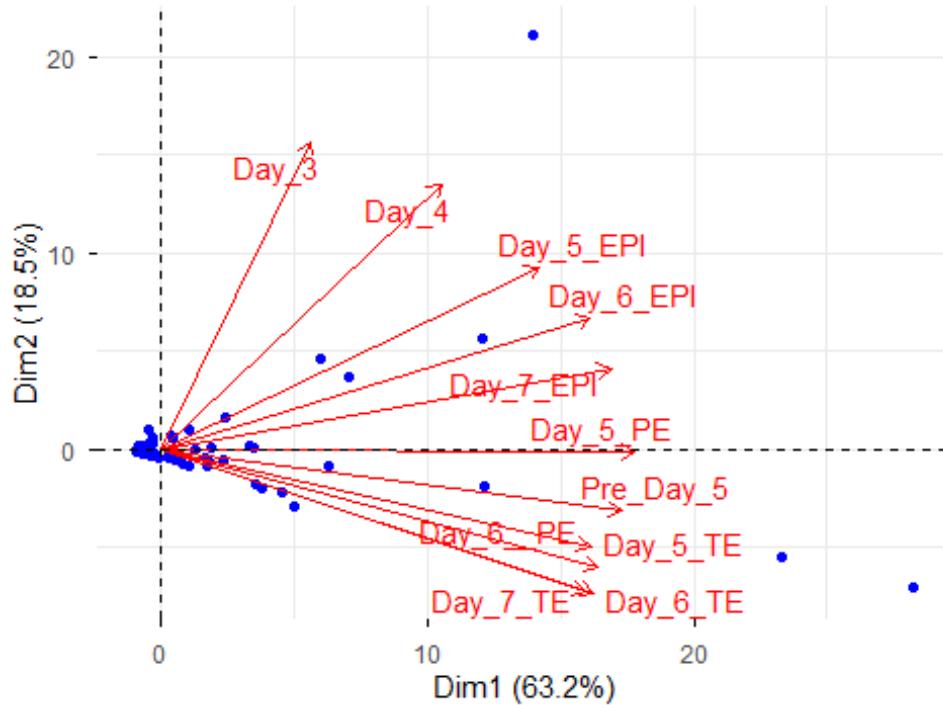


```
# Filtrer les individus avec cos² > 50%
ind_cos2 <- apply(pca.data$ind$cos2, 1, max) > 0.5

# Filtrer les variables avec cos² > 50%
var_cos2 <- apply(pca.data$var$cos2, 1, max) > 0.5

# Créer un graphique combiné des individus et des variables
fviz_pca_biplot(pca.data,
                  select.ind = list(cos2 = 0.5), # Sélectionner les individus
avec cos² > 50%
                  select.var = list(cos2 = 0.5), # Sélectionner les variables
avec cos² > 50%
                  repel = TRUE, # Éviter le chevauchement des étiquettes
title = "Biplot des Individus et des Variables (cos² > 50%)",
col.ind = "blue", # Couleur des individus
col.var = "red", # Couleur des variables
geom.ind = "point"
)
```

Biplot des Individus et des Variables ($\cos^2 > 50\%$)



5. Représentation des individus

Que révèle la position des individus sur le plan des deux premiers axes (proximité, similarité, groupes, outliers) ?

Références

Petropoulos, S., Edsgård, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Reyes, A. P., Linnarsson, S., Sandberg, R., et Lanner, F. (2016). Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. *Cell* 165: 1012-1026.