

Examen de statistiques et analyse de données

Antoine Géré

22 janvier 2026

-
- 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 pollen du maïs (*Zea mays*) constitue une source de nourriture pour les larves de moustiques de l'espèce *Anopheles arabiensis*, principal vecteur de la malaria en Éthiopie. La production de maïs a augmenté de manière substantielle dans certaines régions d'Éthiopie ces dernières années et, sur la même période la malaria s'est étendu à de nouvelles zones où la maladie était auparavant rare. Cela soulève la question suivante :

L'augmentation de la culture du maïs est-elle en partie responsable de l'augmentation de la malaria en Éthiopie ?



Image by Flickr/CIMMYT

Une approche consiste à rechercher une association entre la production de maïs et l'incidence de la malaria dans différents sites géographiquement dispersés (Kebede et al. (2005)).

Le jeu de données « `malaria vs maize.csv` » contient des informations sur plusieurs variables :

- `village` : nom du village où les mesures ont été réalisées (ils sont tous différents)
- `maize_yield` : niveau de production de maïs (Low, Medium, High)

- `mean_elevation_m` : altitude du village
- `person.years` : nombre d'habitants dans le village recensé
- `number_malaria_cases` : nombre de cas de malaria recensé dans le village
- `incidence_rate_per_ten_thousand` : taux d'incidence de la malaria pour 10 000 habitants

On cherche à déterminer s'il existe une association statistique entre la culture du maïs et le taux d'incidence de la malaria.

Analyse descriptive

On importe la base de données sur **R** afin de réaliser quelques analyses descriptives.

```
data_ethiopia = read.csv("data_raw/malaria_vs_maize.csv")
head(data_ethiopia)
```

	village	maize_yield	mean_elevation_m	person.years
1	Gulim	High	2050	23458
2	Denbun	High	2050	20630
3	Wadra	High	2050	12805
4	Zalimashebekuma	High	2050	16857
5	Adel Agata	High	2050	12805
6	Fetamsantom	High	2050	18333

	number_malaria_cases	incidence_rate_per_ten_thousand
1	683	291
2	88	43
3	170	133
4	370	219
5	226	176
6	552	301


```
str(data_ethiopia)
```

```
'data.frame': 19 obs. of 6 variables:
 $ village : chr "Gulim" "Denbun" "Wadra"
 "Zalimashebekuma" ...
 $ maize_yield : chr "High" "High" "High" "High" ...
 $ mean_elevation_m : int 2050 2050 2050 2050 2050 2050 2050 2050
 2050 2200 2200 ...
 $ person.years : int 23458 20630 12805 16857 12805
 18333 18384 16063 12079 26223 ...
 $ number_malaria_cases : int 683 88 170 370 226 552 857 395 71
 218 ...
 $ incidence_rate_per_ten_thousand: int 291 43 133 219 176 301 466 246 59
 83 ...
```



```
summary(data_ethiopia)
```

village	maize_yield	mean_elevation_m	person.years
Length:19	Length:19	Min. :2050	Min. :10960
Class :character	Class :character	1st Qu.:2050	1st Qu.:12975
Mode :character	Mode :character	Median :2200	Median :16857
		Mean :2155	Mean :16590

	3rd Qu.:2200	3rd Qu.:18359
	Max. :2450	Max. :26223
number_malaria_cases	incidence_rate_per_ten_thousand	
Min. : 5.0	Min. : 3.0	
1st Qu.: 68.5	1st Qu.: 39.0	
Median :170.0	Median : 83.0	
Mean :224.6	Mean :129.4	
3rd Qu.:298.0	3rd Qu.:197.5	
Max. :857.0	Max. :466.0	

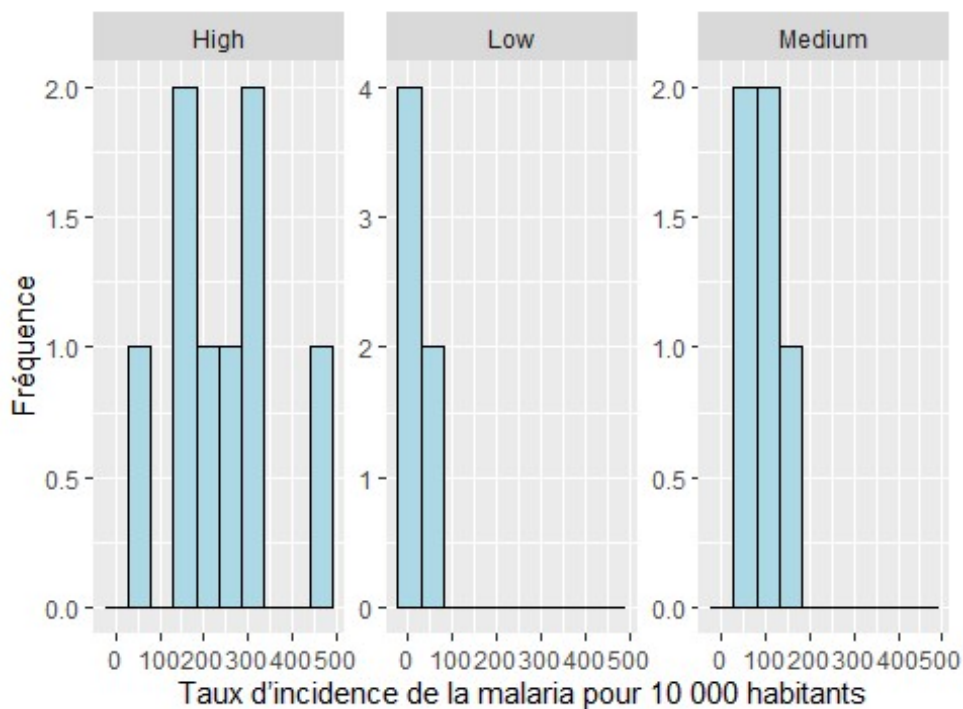
```
table(data_ethiopia$maize_yield)
```

High	Low	Medium
8	6	5

1. Le fichier de données

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

```
ggplot(data_ethiopia, aes(x = incidence_rate_per_ten_thousand)) +
  geom_histogram(bins = 10, fill = "lightblue", color = "black") +
  facet_wrap(~ maize_yield, scales = "free_y") +
  labs(
    title = "",
    x = "Taux d'incidence de la malaria pour 10 000 habitants",
    y = "Fréquence"
  )
```



2. Interprétation des histogrammes ci-dessus

- Les variances semblent-elles comparables ? Justifiez votre réponse.
- Les distributions paraissent-elles approximativement normales ? Justifiez votre réponse.
- Les moyennes semblent-elles clairement différentes ? Justifiez votre réponse

```
stats <- data_ethiopia %>%
  group_by(maize_yield) %>%
  summarise(
    Moyenne = mean(incidence_rate_per_ten_thousand),
    Ecart_type = sd(incidence_rate_per_ten_thousand),
    n = n()
  )
stats
```

A tibble: 3 × 4

	maize_yield	Moyenne	Ecart_type	n
	<chr>	<dbl>	<dbl>	<int>
1	High	234.	126.	8
2	Low	17.7	17.1	6
3	Medium	95.4	41.4	5

3. Interprétation du tableau descriptif

- Quel rendement a en moyenne le plus haut taux d'incidence de malaria pour 10 000 habitants ?
- Quel rendement a en moyenne le plus bas taux d'incidence de malaria pour 10 000 habitants ?
- Les variances semblent-elles comparables ? Justifier votre réponse.

Tests statistiques

Nous nous proposons de réaliser des tests statistiques afin de déterminer quel rendement a en moyenne le plus d'impact sur le taux d'incidence de la malaria pour 10 000 habitants.

Avant de se lancer, il nous est proposé de vérifier l'indépendance de la variable `maize_yield`. On réalise pour cela un test du χ^2 d'indépendance.

```
chisq.test(table(data_ethiopia$maize_yield))
```

Chi-squared test for given probabilities

```
data: table(data_ethiopia$maize_yield)
X-squared = 0.73684, df = 2, p-value = 0.6918
```

4. Indépendance

- Donnez l'hypothèse nulle pour le test du χ^2 d'indépendance.
- Interprétez le résultat de ce test du χ^2 .

On se propose de réaliser un `t.test` afin de déterminer s'il existe une association statistique entre la culture du maïs et le taux d'incidence de la malaria.

```
t.test(incidence_rate_per_ten_thousand ~ maize_yield, data = data_ethiopia)
```

5. T-test

- Donnez l'hypothèse nulle pour le `t.test`.
- Expliquez la raison pour laquelle ce `t.test` ne fonctionne pas dans ce cas précis.

L'ANOVA est un choix logique de méthode pour tester les différences de taux moyens de malaria pour des niveaux de production de maïs différents.

6. ANOVA

- Donnez l'hypothèse nulle pour l'ANOVA.
- Donnez les conditions d'application pour l'ANOVA.

```
leveneTest(incidence_rate_per_ten_thousand ~ maize_yield, data = data_ethiopia)
```

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

	Df	F value	Pr(>F)
group	2	3.9662	0.03991 *
	16		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```
anova_model <- aov(incidence_rate_per_ten_thousand ~ maize_yield, data_ethiopia)
shapiro.test(residuals(anova_model))
```

Shapiro-Wilk normality test

data: residuals(anova_model)
W = 0.88211, p-value = 0.02337

```
summary(anova_model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
maize_yield	2	168844	84422	11.23	0.000898 ***
Residuals	16	120294	7518		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

7. Violation des hypothèses

- Interprétez le `leveneTest`.
- Interprétez le `shapiro.test`.
- Ces données semblent-elles conformes aux hypothèses de l'ANOVA ? Justifiez votre réponse.
- Quelle information nous donne l'ANOVA ? Est-ce que cette information est exploitable ?

Afin de pouvoir utiliser l'ANOVA, on se propose de calculer le logarithme du taux d'incidence et d'effectuer le test sur ces données transformées.

```
data_ethiopia$log_incidence <-
log(data_ethiopia$incidence_rate_per_ten_thousand)

leveneTest(log_incidence ~ maize_yield, data = data_ethiopia)

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  2  3.4689 0.05604 .
      16
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova_log <- aov(log_incidence ~ maize_yield, data_ethiopia)
shapiro.test(residuals(anova_log))

      Shapiro-Wilk normality test

data:  residuals(anova_log)
W = 0.97606, p-value = 0.8875

summary(anova_log)

      Df Sum Sq Mean Sq F value    Pr(>F)
maize_yield  2   29.49   14.744    22.22 2.41e-05 ***
Residuals   16   10.62    0.663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

8. Transformation des données

- Les données transformées par le logarithme respectent-elles mieux les hypothèses de l'ANOVA que les données non transformées ?
- Interprétez le résultat de cette ANOVA.
- L'ANOVA permet-elle de déterminer exactement quelles productions de maïs, ont des effets différents sur le taux d'incidence ?

```
tukey <- TukeyHSD(anova_log)
tukey

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

Fit: aov(formula = log_incidence ~ maize_yield, data = data_ethiopia)

$maize_yield
      diff      lwr      upr    p adj
Low-High -2.8949350 -4.0300404 -1.7598295 0.0000180
Medium-High -0.7881599 -1.9863742  0.4100543 0.2365763
Medium-Low  2.1067751  0.8340678  3.3794823 0.0016001
```

9. Transformation des données

- Rappelez l'objectif d'un test post-hoc après une ANOVA.
- Quelles conclusions peut-on tirer de ces résultats ?

Analyse factorielle

Dans cette partie, on considère l'ensemble des variables quantitatives du jeu de données (y compris le taux d'incidence transformé).

```
var.quant = sapply(data_ethiopia, is.numeric)
data.quant = data_ethiopia[, var.quant]
```

10. Question théorique

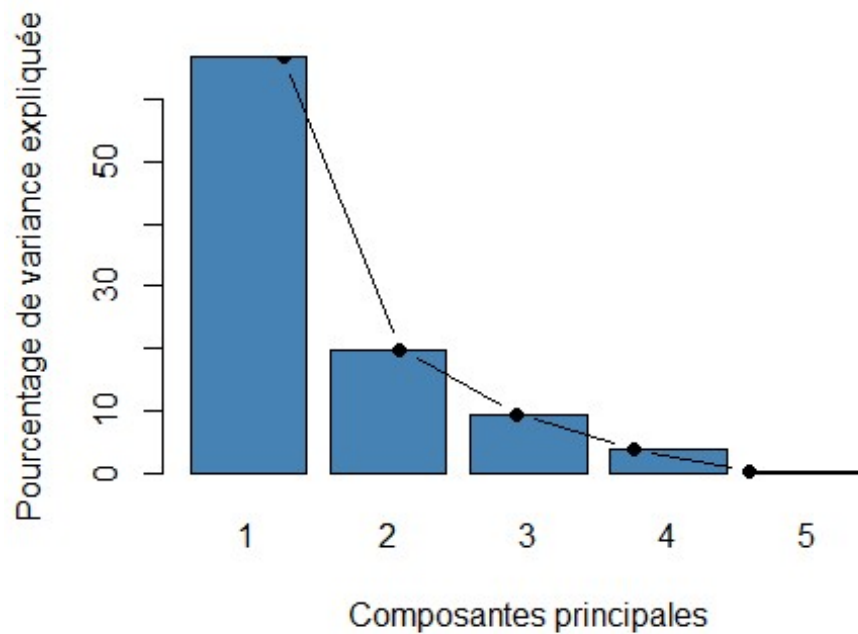
- Pourquoi est-il important de distinguer les variables quantitatives et qualitatives avant l'analyse ?
- Expliquez en quelques phrases l'objectif de l'ACP. Pourquoi cette méthode est-elle utilisée ?

```
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
comp 1	3.339376753	66.7875351	66.78754
comp 2	0.992079469	19.8415894	86.62912
comp 3	0.471894697	9.4378939	96.06702
comp 4	0.189876374	3.7975275	99.86455
comp 5	0.006772707	0.1354541	100.00000

```
barplot(vp[, 2],
        names.arg=1:nrow(vp),
        main = "",
        xlab = "Composantes principales",
        ylab = "Pourcentage de variance expliquée",
        col = "steelblue")

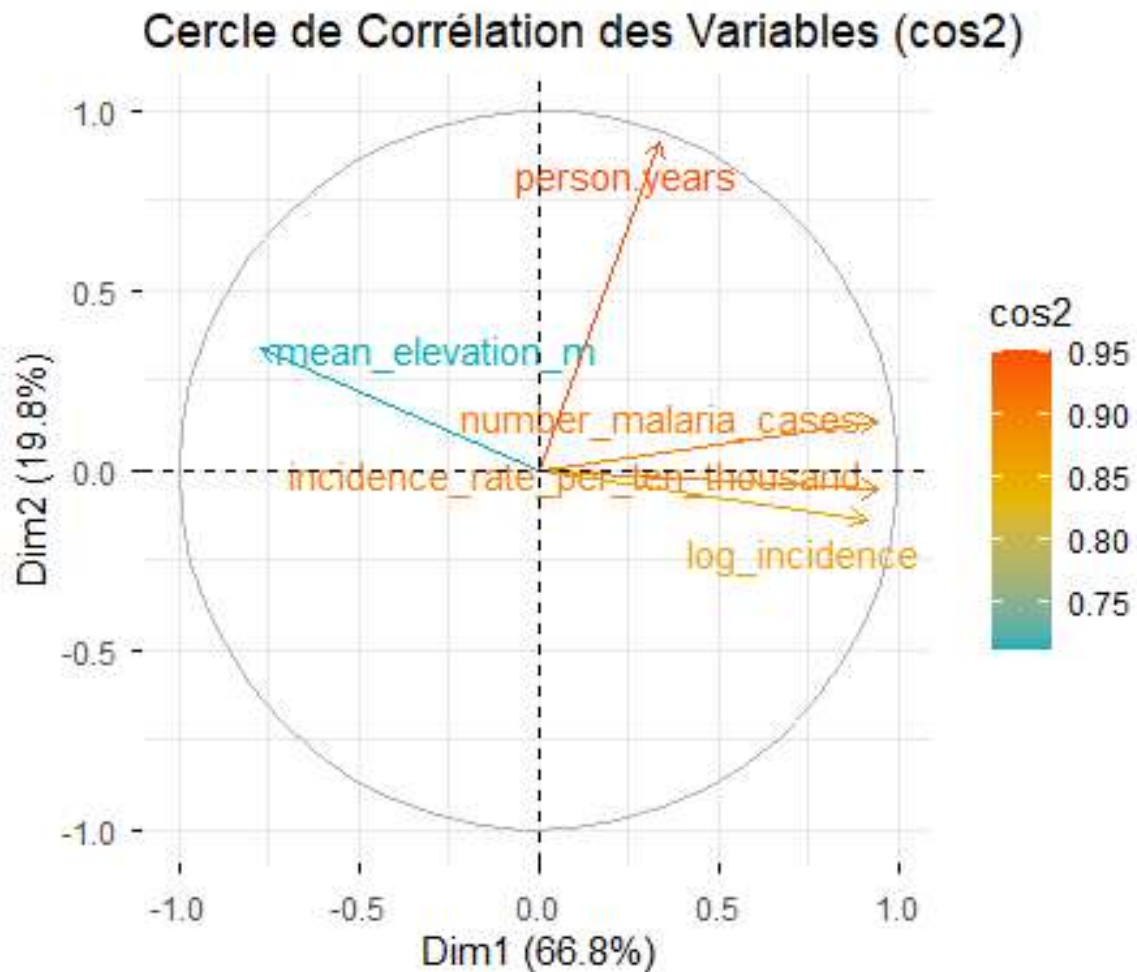
lines(x = 1:nrow(vp),
      vp[, 2],
      type="b",
      pch=19,
      col = "black")
```



11. Variance expliquée

Combien de composantes principales retiendriez-vous pour représenter l'essentiel de l'information ? Justifiez votre réponse.

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

12. Cercle de corrélation

- Identifiez deux variables fortement corrélées positivement.
- Identifiez deux variables fortement corrélées négativement.
- Identifiez deux variables faiblement corrélées.
- Que signifie la longueur et l'orientation des flèches ?
- Quelles variables ont une contribution importante sur l'axe 1 ?
- Interprétez ses résultats.

Références

Kebede, A., McCANN, J. C., Kiszewski, A. E., et Ye-Ebiyo, Y. (2005). New evidence of the effects of agro-ecologic change on malaria transmission. *The American journal of tropical medicine and hygiene* 73: 676-680.