Devoir Statistiques Avancées

Thomas Husson, Groupe 52

Table des matières

! Important

Utilisation de l'IA
Des LLMs ont été utilisés à plusieurs reprises dans ce devoir, pour deux tâches principales :

- En cas de problème d'éxécution du code R (pour suggérer correction et amélioration)
- Pour amélioration du rendu depuis un fichier Quarto Markdown vers PDF. Notamment certaines fonctions dont l'output R n'est pas compatible avec le rendu pdf (par exemple, `factanal()` pour l'analyse factorielle ou le rendu des tableaux automatisé des coefficients alpha de Cronbach et leur IC)

Utilisation des modèles Open Source disponibles sur Hugging Face ou ollama.

1 Énoncé et présentation des échelles

1.A Énoncé du devoir

Consigne :

- Étude d'épidémiologie clinique avec mesures répétées
- Données :
  - 146 patients déprimés
  - Évaluations à J0, J4, J7, J14, J21, J28, J42, J56
  - Autoévaluation (SCL90) et hétéroévaluation (échelle de dépression de Hamilton)
- Questions :
  1. Validation de l'échelle de dépression de Hamilton aux temps J0 et J56
  2. Comparaison de la réponse au traitement entre deux groupes de patients (groupe=0 et groupe=1) en utilisant le score brut de Hamilton avec une approche LOCF puis un modèle mixte
  3. Réponse à la question 2 en utilisant un critère binaire censuré « réponse au traitement » défini par une chute de 50% à l'échelle de Hamilton par rapport à J0
- Fichiers :
  - Fichier groupe (`outil groupe.xlsx`) (2 sous-groupes de patients)
  - Fichier autoévaluation (`outil autoeval.xlsx`) (SCL 90)
  - Fichier hdrs (`outil hdrs.xlsx`) (échelle de Hamilton)

1.B Échelles

Table 1: Présentation des échelles utilisées dans le devoir

	Échelle de Hamilton (HDRS)	Échelle SCL90
Objectif	Mesure l'intensité de la symptomatologie dépressive	"Inconfort psychopathologique" selon plusieurs dimensions.
Type	Hétéro-évaluation	Autoévaluation
Méthode	17 items codés de 2 à 4- Score 7 : pas de dépression clinique- Score 8–15 : dépression mineure- Score > 15 : dépression majeure	10 dimensions : somatisation, symptômes obsessionnels, sensibilité interpersonnelle, dépression, anxiété, hostilité, phobies, traits paranoïaques, traits psychotiques et symptômes divers.

2 Gestion des données

2.A Présentation des fichiers de données

Les 3 fichiers sont en format "large" : chaque ligne correspond à une visite d'un patient et une colonne par item de l'échelle (sauf l'item 16 = PERTE DE POIDS qui est codé en deux variables HAMD16A et HAMD16B dans l'échelle de Hamilton, selon que la perte de poids est déclarée par le patient ou appréciée par le médecin)

On créera donc une colonne `hdrs$HAMD16` qui prendra la valeur de `hdrs$HAMD16A` si elle est remplie, sinon la valeur de `hdrs$HAMD16B`.

2.A.1 Fichier Hamilton

- 1053 observations, 20 variables pour 146 patients
- On ajoute une colonne `score` qui contient le score total de l'échelle de Hamilton (somme des items)
- Les données d'une ligne (J7 du 128ème patient) sont manquantes → on supprime cette ligne.

2.A.2 Fichier scl90

- 1034 observations, 92 variables, 146 patients.
- On crée 10 nouvelles variables représentant les scores moyen des 10 dimensions de l'échelle SCL90.
- Données aberrantes parfois, qui sont recodées en données manquantes et représentent ainsi 0.6% des données totales.

2.A.3 Fichier groupe

- Répartit les 146 patients en 2 groupes (1 et 0)
- Pas de NA

2.B Import des données et data management

Les données sont importées à partir de fichiers Excel.

2.B.1 SCL90

Le jeu de données `scl90` est traité de la manière suivante :

- Visites ordonnées chronologiquement
- Identification des doublons
- Visualisation et gestion des données aberrantes
- Imputation des données manquantes par le mode pour chaque question
- Création des scores moyens par dimension (10 dimensions)

- Nouveau dataframe `scl90_dim` avec uniquement les 10 dimensions

2.B.2 HDRS

Le jeu de données `hdrs` est traité de la manière suivante :

- Visites ordonnées chronologiquement
- Identification des doublons
- Fusion des variables HAMD16A et HAMD16B en une seule variable HAMD16
- Création du score total HDRS (ajouté dans la colonne `hdrs$score`)

2.B.3 Groupes

2.B.4 Fusion des 3 fichiers

Convertir `hdrs_groupe`, `scl90_groupe` et `df_total_wide` de format “large” à format “long”

3 Question 1 : Validation de l'échelle Hamilton

> **Note**
>
> **Consigne de la question 1** : Lorsque l'on utilise un instrument de mesure subjective dans une étude clinique, il est toujours bon de le (re)valider rapidement. Procédez ici à cette **vérification** sur l'échelle de dépression de Hamilton, aux temps J0 et J56.

- Vérification d'une échelle de mesure subjective = 1/ Que mesure l'instrument ? 2/ Que vaut la mesure ?
- Premier temps : Évaluation préliminaire des réponses aux items, puis chercher une corrélation entre eux par une matrice de corrélation 2 à 2
- Second temps : Analyse de la structure dimensionnelle = **que mesure l'instrument ?**
  - Exploration de la structure par analyse en composante principale : visualiser les relations entre les items
  - Détermination du nombre de dimensions : diagramme des valeurs propres (*scree plot*) permet de déterminer le nombre de dimensions (composantes principales)
  - Si structure dimensionnelle identifiée : **analyse factorielle** permet de déterminer quels items se regroupent dans chaque dimension
- Troisième temps : Évaluation de la fiabilité interne = **que vaut la mesure ?**
  - La consistance interne des items (évalue si les items sont cohérents entre eux) sera évaluée par le calcul de l'alpha de Cronbach, calculé sur l'échelle totale et sur chaque dimension identifiée précédemment.
- Quatrième temps : Évaluation de la validité = **l'instrument mesure-t-il ce qu'il est censé mesurer ?** (similaire à la question "que mesure l'instrument ?")
  - Validité interne : déjà évaluée au cours du second temps (structure dimensionnelle)
  - Validité externe : corrélation avec d'autres instruments de mesure (ici les dimensions de l'échelle SCL90)

3.A Valdidation.à J0

3.A.1 Description

Les réponses sont représentées :

- par des histogrammes pour chaque item de l'échelle de Hamilton à J0
- par une matrice de corrélation 2 à 2 entre les items

NB : le code R utilise une fonction pour faciliter la création des histogrammes pour chaque item.

La fonction crée un histogramme pour chaque item listés dans un vecteur crée précédemment (`hdrs_items`).

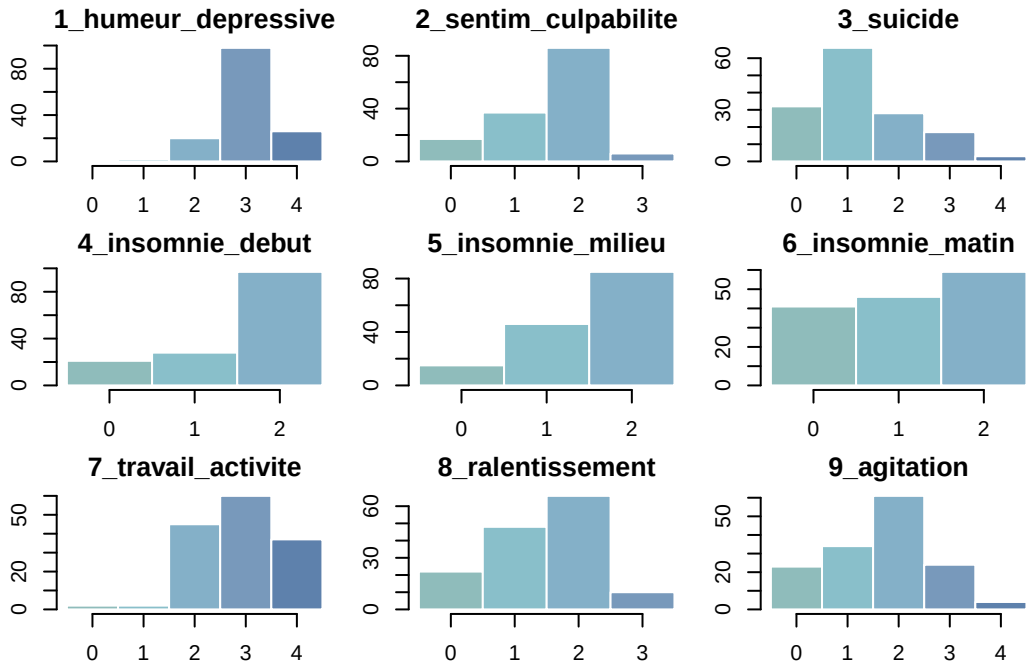

Figure 1: Histogrammes des scores des items de l'échelle de Hamilton à J0

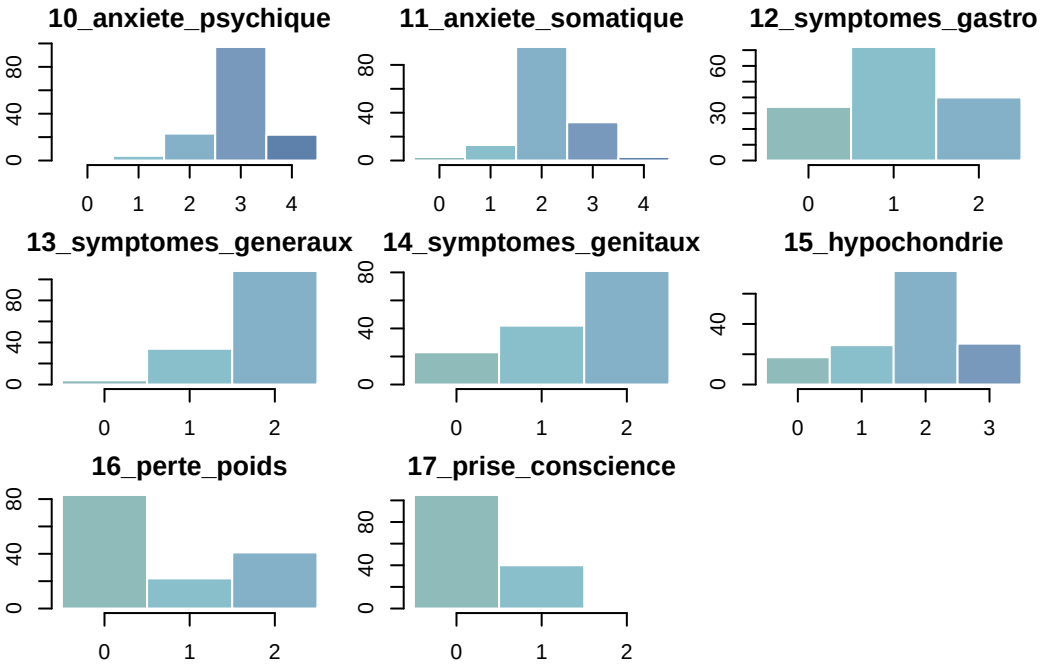

Figure 2: Histogrammes des scores des items de l'échelle de Hamilton à J0

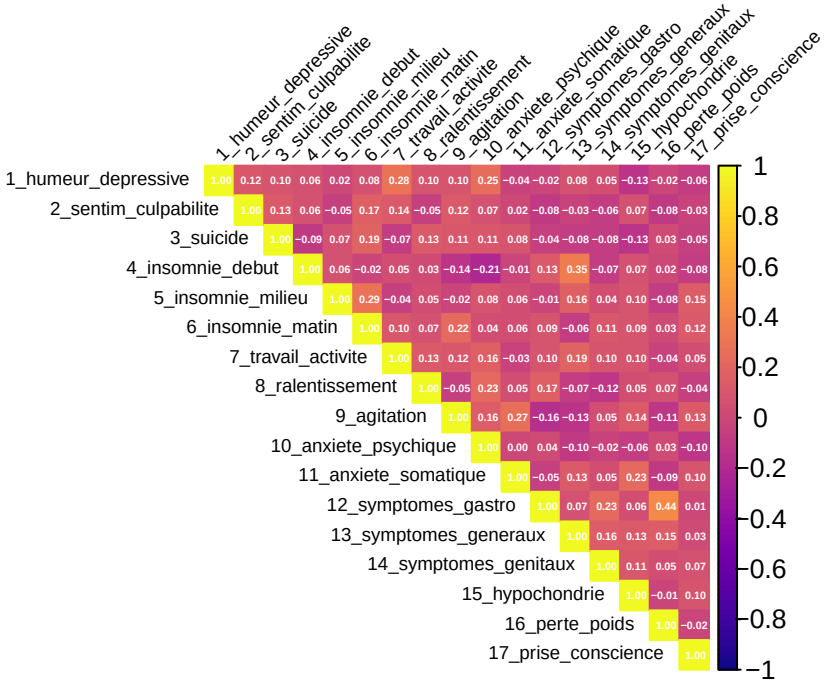

Figure 3: Matrice de corrélation entre les items de l'échelle de Hamilton à J0

- Il n'y a pas de données manquantes.
- Les histogrammes montrent que certains items ont une distribution asymétrique (ex : insomnie quelque soit le moment de la nuit, symptômes généraux, perte de poids…)
- La matrice de corrélation des items 2 à 2 ne retrouve pas de coefficient de corrélation supérieure à 0,50 en valeur absolu, il n'existe pas de redondance entre les items de l'échelle Hamilton.

3.A.2 Validité interne : structure dimensionnelle, analyse factorielle

3.A.2.1 Exploration de la structure dimensionnelle : analyse en composantes principales

- On peut réaliser une analyse en composantes principales (ACP) pour visualiser les relations entre les items.

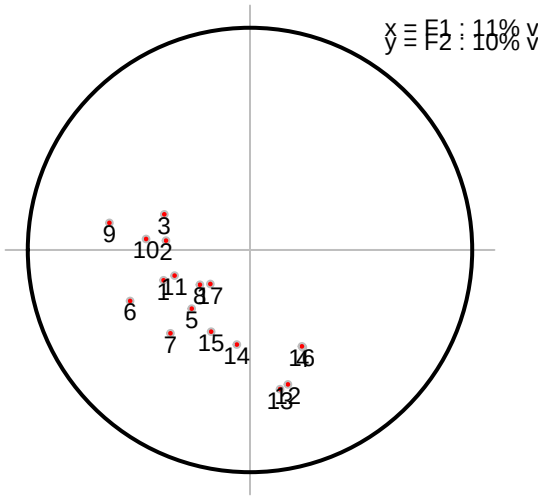

Figure 4: Analyse en composantes principales des items de l'échelle de Hamilton à J0

- Chaque point représente un item de l'échelle de Hamilton.
- Deux axes principaux :
  - l'axe horizontale `x` représente la première composante principale (PC1) qui explique 11% de la variance totale, l'axe verticale `y` représente la deuxième composante principale (PC2) qui explique 10% de la variance totale.
  - Ensemble, les deux premières composantes principales expliquent 21% de la variance totale, ce qui est relativement faible.
- La majorité des variables sont proches du centre, ce qui indique qu'elles ne contribuent pas fortement aux premières composantes principales.
- Au total, cette ACP ne révèle pas de structure dimensionnelle claire parmi les items de l'échelle de Hamilton à J0.

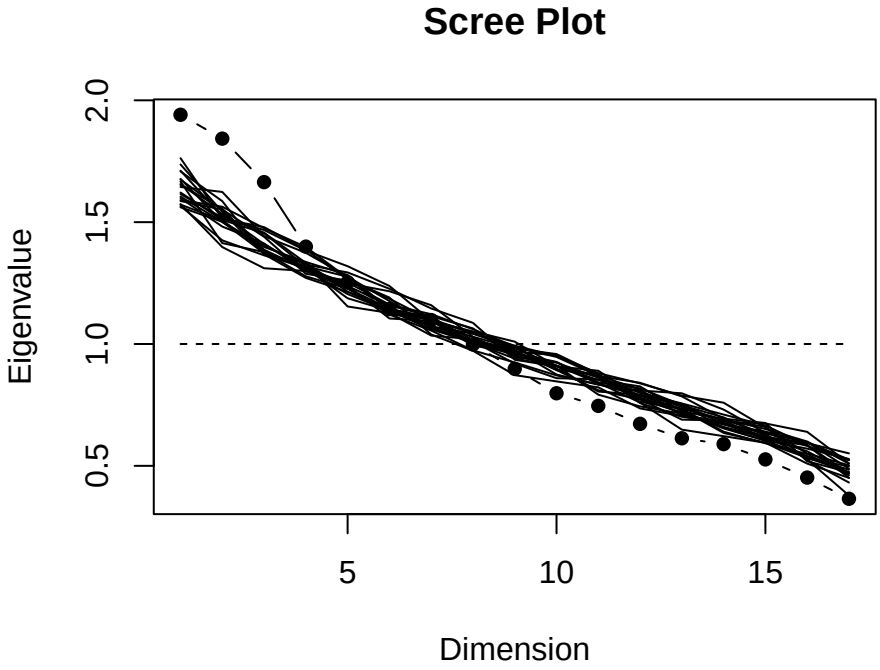

Figure 5: Diagramme des valeurs propres (scree plot) des items de l'échelle de Hamilton à J0 avec représentation de données simulées (analyse parallèle)

3.A.2.2 Exploration de la structure dimensionnelle : analyse factorielle

- À J0, le *scree plot* ne permet pas d'identifier un nombre clair de facteurs : les valeurs propres décroissent progressivement sans "coude" net.
- En analyse parallèle, on observe au moins 3 dimensions ayant une valeur propre supérieure à celle obtenue sur des données simulées.
- On pourrait réaliser des tests statistiques qui permettraient de déterminer le nombre optimal de dimensions, mais ces tests sont sujets à plusieurs biais :
  - on calculerait une p-value pour l'hypothèse "n facteurs sont suffisants"
  - mais ces tests sont difficiles à interpréter et sensibles à la taille de l'échantillon
  - On retient donc 3 facteurs principaux pour l'analyse factorielle.

Table 2: Contribution des facteurs à la variance de la réponse à chaque item du score de Hamilton évalué à J0 (analyse factorielle avec rotation varimax à 3 facteurs)

Variable	Factor1	Factor2	Factor3
1_humeur_depressive	0.006	0.028	0.263
2_sentim_culpabilite	-0.057	-0.076	0.240
3_suicide	-0.021	-0.120	0.207
4_insomnie_debut	0.110	0.359	-0.050
5_insomnie_milieu	0.004	0.118	0.240
6_insomnie_matin	0.132	-0.155	0.427
7_travail_activite	0.124	0.121	0.320
8_ralentissement	0.183	-0.105	0.101

Variable	Factor1	Factor2	Factor3
9_agitation	-0.107	-0.228	0.501
10_anxiete_psychique	0.066	-0.152	0.255
11_anxiete_somatique	-0.029	0.070	0.316
12_symptomes_gastro	0.992	0.044	-0.091
13_symptomes_generaux	0.046	0.977	0.194
14_symptomes_genitaux	0.238	0.120	0.147
15_hypochondrie	0.076	0.089	0.226
16_perte_poids	0.424	0.155	-0.100
17_prise_conscience	0.022	-0.004	0.176

- À J0, l'analyse factorielle exploratoire avec rotation varimax met en évidence 3 facteurs latents expliquant cumulativement 21,9 % de la variance des réponses aux items du score de Hamilton.
- Concernant chacun des 3 facteurs :
  - Facteur 1 : 8, 12, 14, 16 (principalement des symptômes somatiques).
  - Facteur 2 : 4, 13 (relatifs à l'asthénie).
  - Facteur 3 : les items restants (symptômes dépressifs psychiatriques proprement dits).

On peut rajouter 3 "sous-scores" au score total de Hamilton à J0, correspondant aux scores moyens des items chargés sur chaque facteur.

- A titre exploratoire, on peut refaire une ACP sur ces 3 sous-scores pour visualiser leur relation.

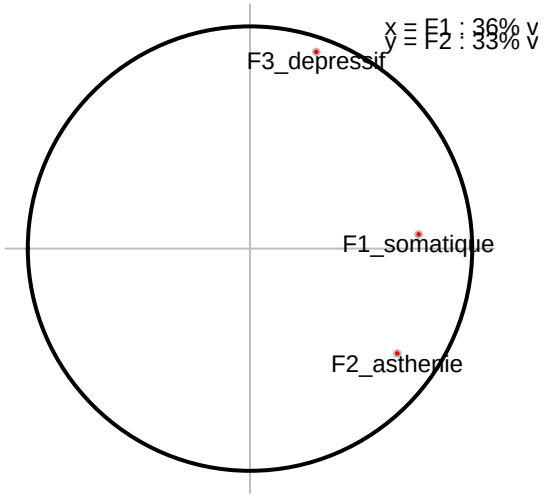

Figure 6: Analyse en composantes principales des sous-scores de l'échelle de Hamilton à J0

- Les 3 sous-scores sont bien représentés (proches du cercle). Le facteur "symptômes dépressifs" semble orthogonal aux deux autres facteurs.
- La variance totale expliquée par ces 3 sous-score est de 69%.

Une ACP focalisée sur ces 3 sous-scores et le score total de Hamilton permet de visualiser la relation entre le score total et les sous-scores.

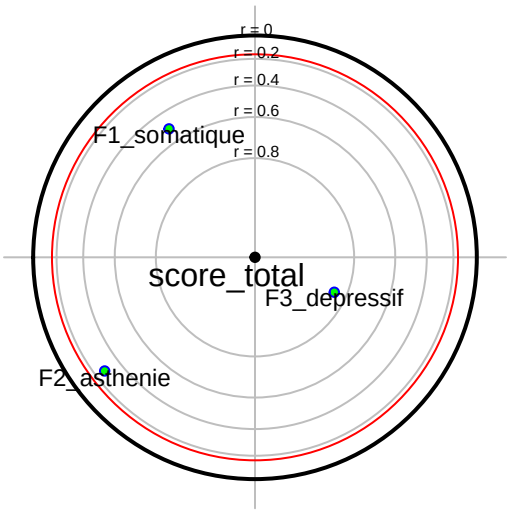

Figure 7: Analyse en composantes principales focalisée des sous-scores de l'échelle de Hamilton à J0

Le score total de Hamilton semble plus corrélé aux symptômes dépressifs (F3) qu'aux deux autres sous-scores.

3.A.3 Fiabilité interne = que vaut la mesure ?

3.A.3.1 Consistance interne : alpha de Cronbach La consistance interne des items de l'échelle de Hamilton à J0 est évaluée par le calcul de l'alpha de Cronbach, qui correspond globalement au **pourcentage de « variance partagée »** entre le score vrai (hypothétique) et la mesure obtenue.

Il permet ainsi de mesurer la cohérence entre les items d'une échelle de mesure, et est élevé lorsque les items sont fortement corrélés entre eux.

On peut donc calculer dans un premier temps l'alpha de Cronbach sur l'ensemble des items de l'échelle de Hamilton à J0, puis sur chacun des 3 facteurs identifiés précédemment.

Les intervalles de confiance (IC) à 95% des alpha de Cronbach sont estimés par la méthode du bootstrap avec 1000 rééchantillonnages. Le bootstrap est possible ici car il y a > 100 observations.

Table 3: Alpha de Cronbach et intervalles de confiance à 95% pour l'échelle de Hamilton à J0 et ses sous-échelles

Scale	Alpha	CI_lower	CI_upper
Global	0.456	0.262	0.624
F1_somatique	0.381	0.188	0.530
F2_asthenie	0.490	0.266	0.653
F3_depressif	0.486	0.349	0.622

Au total, quelque soit le niveau d'analyse (global ou par facteur), les alpha de Cronbach sont < 0.5, indiquant une faible consistance interne des items de l'échelle de Hamilton à J0.

3.A.4 Validité externe = l'instrument mesure-t-il ce qu'il est censé mesurer ?

- Validité externe d'un instrument cherche à démontrer que l'instrument se comporte logiquement par rapport au réseau théorique qui lui est associé.
- Selon la théorie nomologique (c'est à dire selon les relations postulées entre les différents concepts d'une même discipline), la dépression mesurée par l'échelle de Hamilton doit être fortement liée à d'autres manifestations de la détresse psychologique générale (mesurée par le SCL-90), mais distincte de certains autres concepts.
  - Ici, la validité du construit peut être évaluée en évaluant la validité convergente (corrélation forte entre des concepts proches).
  - Il est plus difficile d'évaluer la validité divergente (corrélation faible entre des concepts différents) car le SCL-90 mesure principalement des dimensions de la détresse psychologique ; de même pour la validité concurrente (corrélation forte avec un *gold-standard*, car nous ne disposons pas d'un instrument de mesure de la dépression reconnu comme un *gold-standard* ici).

3.A.4.1 Corrélation entre le score de Hamilton et les dimensions du SCL-90 à J0

- On peut représenter une matrice de corrélation entre le score total de Hamilton et les 10 dimensions du SCL-90 à J0.

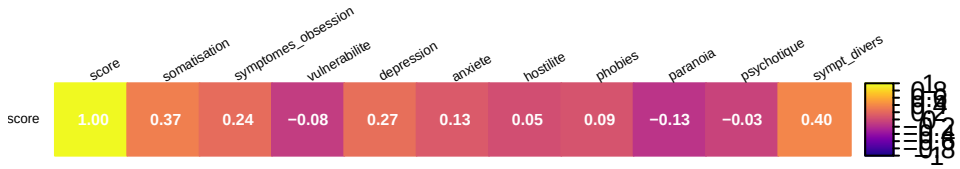

Figure 8: Corrélation entre le score total de l'échelle de Hamilton et les dimensions du SCL-90 à J0

La corrélation est au maximum à 0.37 avec la composante "somatisation" du SCL-90 et de 0.40 avec la composante "symptômes divers" du SCL-90, indiquant une validité convergente modérée entre le score total de Hamilton et cette dimension du SCL-90 à J0.

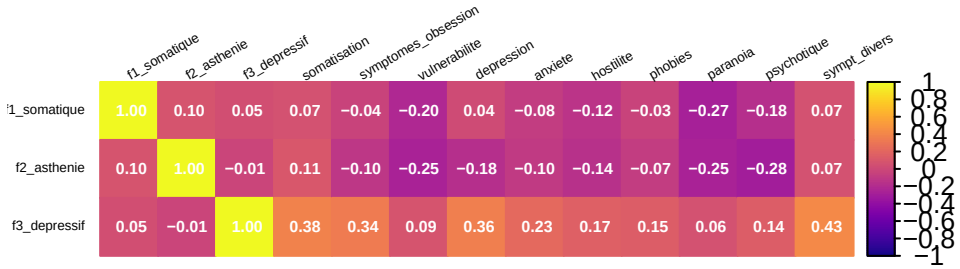

Figure 9: Corrélation entre les sous-scores de l'échelle de Hamilton et les dimensions SCL-90 à J0

- Il n'y a pas non plus de corrélation forte entre les sous-scores de Hamilton et les dimensions du SCL-90 à J0.
- Par exemple, le sous-score "symptômes somatiques" de Hamilton est faiblement corrélé avec la dimension "somatisation" du SCL-90 (r = 0.07 !!).

3.A.4.2 Conclusion validité externe à J0 La validité convergente entre le score total de Hamilton et les dimensions du SCL-90 à J0 est faible à modérée, suggérant que l'échelle de Hamilton mesure partiellement des aspects de la détresse psychologique générale, mais pas de manière très forte.

3.A.5 Conclusion globale à J0

À J0, l'échelle de dépression de Hamilton présente une structure dimensionnelle peu claire, avec une faible consistance interne des items (alpha de Cronbach < 0.5) et une validité convergente modérée avec les dimensions du SCL-90.

Ces résultats suggèrent que l'échelle de Hamilton pourrait ne pas être un instrument optimal pour mesurer la dépression dans cette population à ce moment précis.

3.B Validation à J56

3.B.1 Description

Comme à J0, les réponses sont représentées :

- par des histogrammes pour chaque item de l'échelle de Hamilton à J56
- par une matrice de corrélation 2 à 2 entre les items

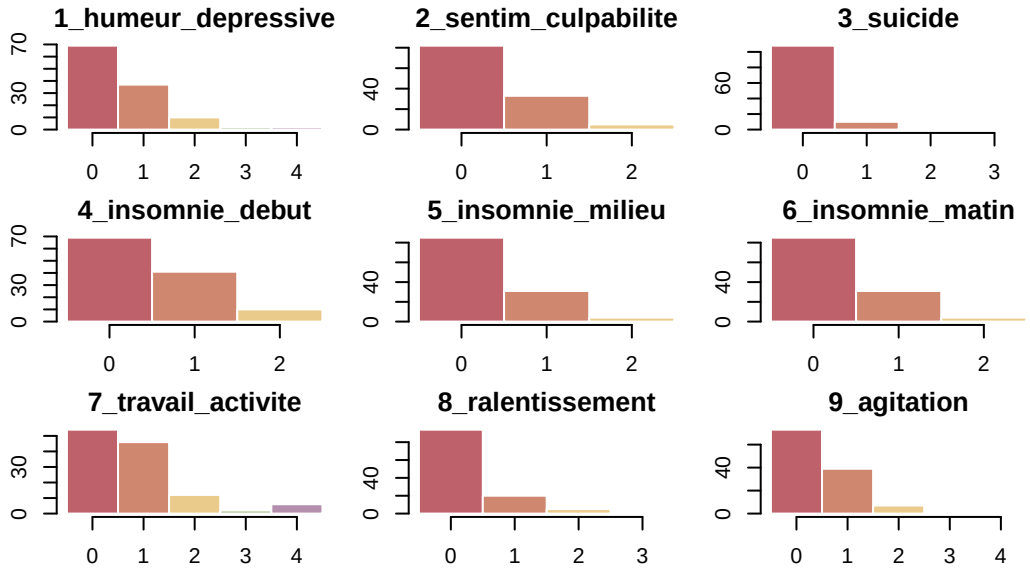

Figure 10: Histogrammes des scores des items de l'échelle de Hamilton à J56

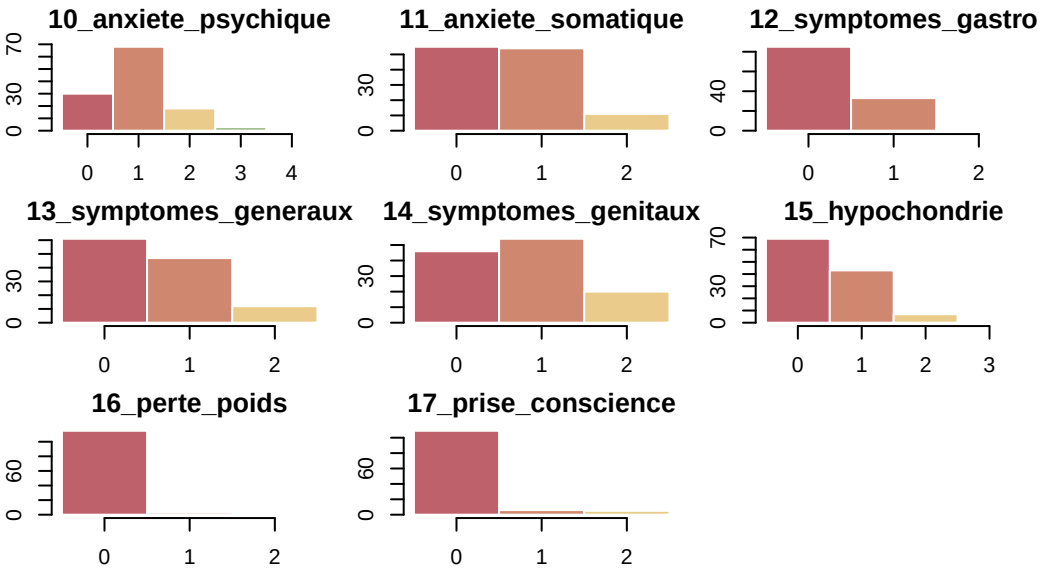

Figure 11: Histogrammes des scores des items de l'échelle de Hamilton à J56

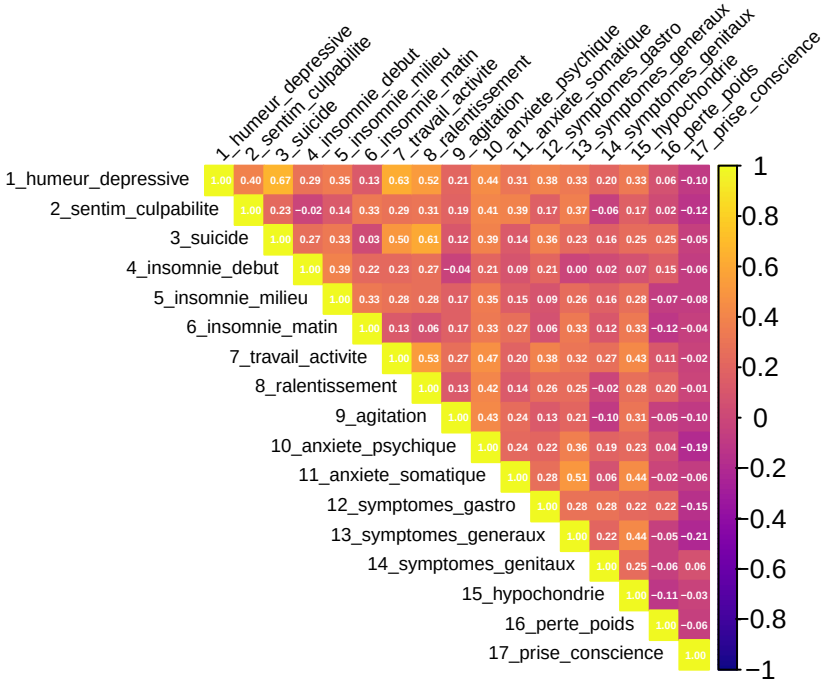

Figure 12: Matrice de corrélation entre les items de l'échelle de Hamilton à J56

3.B.2 Description

3.B.3 Validité interne : structure dimensionnelle, analyse factorielle

3.B.4 Validité externe

4 Question 2 : Comparaison de la réponse au traitement entre deux groupes de patients

5 Annexe – Code R de tous les chunks

```
library(forecast)
library(plotrix)
library(randomForest)
library(tidyr)
library(epiR)
library(viridisLite)
library(ggplot2)
library(binom)
library(survminer)
library(pROC)
library(treemap)
library(psy)
library(MASS)
```

```r
library(rpart)
library(rpart.plot)
library(plotly)
library(lmerTest)
library(psych)
library(lme4)
library(prettyR)
library(kableExtra)
library(gtsummary)
library(dplyr)
library(lattice)
library(survey)
library(corrplot)
library(mice)
library(paletteer)
library(nord)
library(wesanderson)
library(qgraph)
library(nlme)
library(pwr)
library(ape)
library(survival)
library(gmodels)
library(httpgd)
library(e1071)
library(psy)
library(reshape2)
knitr::opts_chunk$set(echo = TRUE)

library(readxl)
scl90 <-
 ↪  read_excel("/Users/thomashusson/Documents/Projets/M2biostatistiques/devoir_stats_avancees/d
 ↪  autoeval.xls")
groupe <-
 ↪  read_excel("/Users/thomashusson/Documents/Projets/M2biostatistiques/devoir_stats_avancees/d
 ↪  groupe.xls")
hdrs <-
 ↪  read_excel("/Users/thomashusson/Documents/Projets/M2biostatistiques/devoir_stats_avancees/d
 ↪  hdrs.xls")

describe(scl90)
summary(scl90)

# ordonne chronologiquement les visites pour chaque patient
scl90$VISIT <- factor(scl90$VISIT,
                      levels = c("J0", "J4", "J7", "J14", "J21", "J28", "J42",
                       ↪  "J56"),
                      ordered = TRUE)
```

```r
# ordonner les visites en fonction du numéro de patient
scl90 <- scl90[order(scl90$NUMERO, scl90$VISIT), ]

# identification des doublons
scl90$NUMERO[duplicated(scl90)]

# nombre de patients uniques
length(unique(scl90$NUMERO))

# visualisation des données aberrantes
apply(scl90, 2, table, useNA = "always")
scl90[, 3:92][scl90[, 3:92] > 4] <- NA
apply(scl90, 2, table, useNA = "always")

# données manquantes
sum(is.na(scl90))
# proportion :
sum(is.na(scl90)) / (nrow(scl90) * ncol(scl90))*100

#imputation par le mode pour chaque question
scl_questions <- c(names(select(scl90,starts_with("Q"))))
for (question in scl_questions) {
    original <- scl90[[question]]
    factorized <- as.factor(original)
    mode_value <- as.integer(names(which.max(table(factorized))))
    imputed <- original
    imputed[is.na(imputed)] <- mode_value
    scl90[[question]] <- imputed
}
# vérifier qu'il n'y a plus de NA
sum(is.na(scl90))

# création des scores moyens par dimension
scl90$somatisation <-
 ↪  apply(scl90[,c(3,6,14,29,44,50,51,54,55,58,60,42)],1,mean,na.rm=TRUE)
scl90$symptomes_obsession <-
 ↪  apply(scl90[,c(11,12,30,40,5,47,48,53,57,67)],1,mean,na.rm=TRUE)
scl90$vulnerabilite <-
 ↪  apply(scl90[,c(8,23,36,38,39,43,63,71,75)],1,mean,na.rm=TRUE)
scl90$depression <-
 ↪  apply(scl90[,c(7,16,17,22,24,28,31,32,33,34,56,73,81)],1,mean,na.rm=TRUE)
scl90$anxiete <- apply(scl90[,c(4,19,25,35,41,59,74,80,82,88)],1,mean,na.rm=TRUE)
scl90$hostilite <- apply(scl90[,c(13,26,65,69,76,83)],1,mean,na.rm=TRUE)
scl90$phobies <- apply(scl90[,c(15,27,49,72,77,84,52)],1,mean,na.rm=TRUE)
scl90$paranoia <- apply(scl90[,c(10,20,45,70,78,85)],1,mean,na.rm=TRUE)
scl90$psychotique <-
 ↪  apply(scl90[,c(9,18,37,64,79,86,87,89,92,90)],1,mean,na.rm=TRUE)
```

```r
scl90$sympt_divers <- apply(scl90[,c(21,46,61,62,66,68,91)],1,mean,na.rm=TRUE)

# création d'un nouveau dataframe avec uniquement les 10 dimensions
dimensions <- c("somatisation", "symptomes_obsession", "vulnerabilite",
 ↪  "depression", "anxiete", "hostilite", "phobies", "paranoia", "psychotique",
 ↪  "sympt_divers")
scl90_dim <- scl90[, c("NUMERO", "VISIT", dimensions)]

hdrs$VISIT <- factor(hdrs$VISIT,
                        levels = c("J0", "J4", "J7", "J14", "J21", "J28", "J42",
                         ↪  "J56"),
                        ordered = TRUE)

# ordonner les visites en fonction du numéro de patient
hdrs <- hdrs[order(hdrs$NUMERO, hdrs$VISIT), ]

# identification des doublons
hdrs$NUMERO[duplicated(hdrs)]

# nombre de patients uniques
length(unique(hdrs$NUMERO))

# fusion des variables HAMD16A et HAMD16B en une seule variable HAMD16
hdrs$HAMD16 <- ifelse(!is.na(hdrs$HAMD16A), hdrs$HAMD16A, hdrs$HAMD16B)
table(hdrs$HAMD16, useNA = "ifany")

# calcul du score total HDRS
items <- c("HAMD1","HAMD2","HAMD3","HAMD4","HAMD5","HAMD6",
            "HAMD7","HAMD8","HAMD9","HAMD10","HAMD11","HAMD12",
            "HAMD13","HAMD14","HAMD15", "HAMD16", "HAMD17")

#renommer la colonne hdrs$HAMD1 en hdrs$humeur_depressive
colnames(hdrs)[colnames(hdrs) == "HAMD1"] <- "1_humeur_depressive"
colnames(hdrs)[colnames(hdrs) == "HAMD2"] <- "2_sentim_culpabilite"
colnames(hdrs)[colnames(hdrs) == "HAMD3"] <- "3_suicide"
colnames(hdrs)[colnames(hdrs) == "HAMD4"] <- "4_insomnie_debut"
colnames(hdrs)[colnames(hdrs) == "HAMD5"] <- "5_insomnie_milieu"
colnames(hdrs)[colnames(hdrs) == "HAMD6"] <- "6_insomnie_matin"
colnames(hdrs)[colnames(hdrs) == "HAMD7"] <- "7_travail_activite"
colnames(hdrs)[colnames(hdrs) == "HAMD8"] <- "8_ralentissement"
colnames(hdrs)[colnames(hdrs) == "HAMD9"] <- "9_agitation"
colnames(hdrs)[colnames(hdrs) == "HAMD10"] <- "10_anxiete_psychique"
colnames(hdrs)[colnames(hdrs) == "HAMD11"] <- "11_anxiete_somatique"
colnames(hdrs)[colnames(hdrs) == "HAMD12"] <- "12_symptomes_gastro"
colnames(hdrs)[colnames(hdrs) == "HAMD13"] <- "13_symptomes_generaux"
colnames(hdrs)[colnames(hdrs) == "HAMD14"] <- "14_symptomes_genitaux"
colnames(hdrs)[colnames(hdrs) == "HAMD15"] <- "15_hypochondrie"
colnames(hdrs)[colnames(hdrs) == "HAMD16"] <- "16_perte_poids"
```

```r
colnames(hdrs)[colnames(hdrs) == "HAMD17"] <- "17_prise_conscience"

# calcul du score total HDRS
hdrs_items <-
 ↪  c("1_humeur_depressive","2_sentim_culpabilite","3_suicide","4_insomnie_debut","5_insomnie_m

hdrs$score <- rowSums(hdrs[, hdrs_items], na.rm = TRUE)

# supprimer la ligne 741 de hdrs
hdrs <- hdrs[-741, ]

# supprimer HAMD16A et HAMD16B si présentes
hdrs <- hdrs[, setdiff(names(hdrs), c("HAMD16A", "HAMD16B"))]

summary(groupe)
describe(groupe)
groupe$NUMERO[duplicated(groupe)]
length(unique(groupe$NUMERO))

# ordonner en fonction du numéro de patient
groupe <- groupe[order(groupe$NUMERO), ] # Questions 1 et 2

# fusion des 3 dataframes
hdrs_groupe <- merge(hdrs, groupe, by = "NUMERO", all.x = TRUE)
scl90_groupe <- merge(scl90, groupe, by = "NUMERO", all.x = TRUE)
df_total_wide <- merge(hdrs_groupe, scl90, by = c("NUMERO", "VISIT"), all.x =
 ↪  TRUE)

library(reshape2)

## HDRS -> long
hdrs_long <- melt(
    hdrs_groupe,
    id.vars = c("NUMERO", "VISIT", "GROUPE"),
    variable.name = "item",
    value.name = "value"
)

## SCL90 -> long
scl90_long <- melt(
    scl90_groupe,
    id.vars = c("NUMERO", "VISIT", "GROUPE"),
    variable.name = "item",
    value.name = "value"
)

## TOTAL (HDRS + SCL90) -> long
df_total_long <- melt(
```

```
    df_total_wide,
    id.vars = c("NUMERO", "VISIT", "GROUPE"),
    variable.name = "item",
    value.name = "value"
)


hdrs_J0 <- subset(hdrs_groupe, VISIT == "J0")

# Vraie palette Nord (package nord)
# On prend une palette qualitative (frost) et on l'étend à 5 couleurs
cols_nord <- nord::nord("frost", 5)

# Première série de graphiques (9 au maximum)
par(mfrow = c(3, 3), mar = c(2, 2, 2, 1))
items1 <- hdrs_items[1:min(9, length(hdrs_items))]
for (item in items1) {
    val <- na.omit(hdrs_J0[[item]])
    if (length(val) > 0) {
        m <- max(val)
        hist(val,
            main = item,
            xlab = "Score",
            col = cols_nord[1:(m + 1)],
            border = "white",
            breaks = seq(-0.5, m + 0.5, 1),
            xaxt = "n")
        axis(1, at = 0:m)
    } else {
        plot.new()
        title(main = paste(item, "(pas de données)"))
    }
    }
par(mfrow = c(1, 1))

# Deuxième série de graphiques (8 au maximum, de 10 à 17)
if (length(hdrs_items) > 9) {
    par(mfrow = c(3, 3), mar = c(2, 2, 2, 1))
    items2 <- hdrs_items[10:min(17, length(hdrs_items))]
    for (item in items2) {
        val <- na.omit(hdrs_J0[[item]])
        if (length(val) > 0) {
        m <- max(val)
        hist(val,
            main = item,
            xlab = "Score",
            col = cols_nord[1:(m + 1)],
            border = "white",
```

```
            breaks = seq(-0.5, m + 0.5, 1),
            xaxt = "n")
        axis(1, at = 0:m)
        } else {
        plot.new()
        title(main = paste(item, "(pas de données)"))
        }
    }
    par(mfrow = c(1, 1))
}

hdrs_J0_matrix <- hdrs_J0[, hdrs_items]
corr_matrix_J0 <- cor(hdrs_J0_matrix, use = "pairwise.complete.obs")

corrplot(corr_matrix_J0,
            method = "color",
            type = "upper",
            tl.col = "black",
            addCoef.col = "white",
            number.cex = 0.35,
            tl.cex = 0.6,
            tl.srt = 45,
            col = viridis::plasma(100)
            )

#renommer les noms des variables dans un nouveau df copié pour éviter la
 ↪  superposition
hdrs_J0_PCA <- hdrs_J0[,c(hdrs_items)]
colnames(hdrs_J0_PCA) <- c("1","2","3","4","5","6",
                           "7","8","9","10","11","12",
                           "13","14","15", "16", "17")
mdspca(hdrs_J0_PCA)

head(hdrs_J0[,c(hdrs_items)])
scree.plot(hdrs_J0[,c(hdrs_items)], simu=20, use = "P")

af_J0 <- factanal(
    na.omit(hdrs_J0[hdrs_items]),
    factors = 3,
    rotation = "varimax"
)

# Extraction propre des loadings
loadings_df <- as.data.frame(unclass(af_J0$loadings))

# Création explicite de la colonne Variable à partir des rownames
loadings_df$Variable <- rownames(loadings_df)
```

```r
# Suppression des rownames pour éviter toute ambiguïté
rownames(loadings_df) <- NULL

# Réorganisation : Variable en première colonne
loadings_df <- loadings_df[, c("Variable", colnames(loadings_df)[1:3])]

# Arrondi
loadings_df[, -1] <- round(loadings_df[, -1], 3)

# Affichage LaTeX
knitr::kable(
    loadings_df,
    caption = "Contribution des facteurs à la variance de la réponse à chaque item
     ↪  du score de Hamilton évalué à J0 (analyse factorielle avec rotation
     ↪  varimax à 3 facteurs)",
    booktabs = TRUE,
    align = "lccc"
)

hdrs_J0$f1_somatique <-
 ↪  rowMeans(hdrs_J0[,c("8_ralentissement","12_symptomes_gastro","14_symptomes_genitaux","16_pe
 ↪  na.rm=TRUE)
hdrs_J0$f2_asthenie <-
 ↪  rowMeans(hdrs_J0[,c("4_insomnie_debut","13_symptomes_generaux")], na.rm=TRUE)
hdrs_J0$f3_depressif <-
 ↪  rowMeans(hdrs_J0[,c("1_humeur_depressive","2_sentim_culpabilite","3_suicide","5_insomnie_mi
 ↪  na.rm=TRUE)

hdrs_J0_subscores <- hdrs_J0[,c("f1_somatique","f2_asthenie","f3_depressif")]
colnames(hdrs_J0_subscores) <- c("F1_somatique","F2_asthenie","F3_depressif")
mdspca(hdrs_J0_subscores)

df_fpca <- data.frame(
    score_total = hdrs_J0$score,
    hdrs_J0_subscores
)

fpca(
    score_total ~ .,
    data = df_fpca
)

cronbach(hdrs_J0[,hdrs_items])
cronbach(hdrs_J0[,c("8_ralentissement","12_symptomes_gastro","14_symptomes_genitaux","16_perte_
cronbach(hdrs_J0[,c("4_insomnie_debut","13_symptomes_generaux")])
cronbach(hdrs_J0[,c("1_humeur_depressive","2_sentim_culpabilite","3_suicide","5_insomnie_milieu

#estimation des IC par bootstrap
```

```r
set.seed(123)
alpha_bootstrap <- function(data, indices) {
    d <- data[indices, ]
    return(cronbach(d)$alpha)
}
library(boot)
# Alpha global
boot_alpha_global <- boot(hdrs_J0[,hdrs_items], alpha_bootstrap, R = 1000)
boot.ci(boot_alpha_global, type = "bca")
# Facteur 1
boot_alpha_f1 <-
 ↪  boot(hdrs_J0[,c("8_ralentissement","12_symptomes_gastro","14_symptomes_genitaux","16_perte_
 ↪  alpha_bootstrap, R = 1000)
boot.ci(boot_alpha_f1, type = "bca")
# Facteur 2
boot_alpha_f2 <- boot(hdrs_J0[,c("4_insomnie_debut","13_symptomes_generaux")],
 ↪  alpha_bootstrap, R = 1000)
boot.ci(boot_alpha_f2, type = "bca")
# Facteur 3
boot_alpha_f3 <-
 ↪  boot(hdrs_J0[,c("1_humeur_depressive","2_sentim_culpabilite","3_suicide","5_insomnie_milieu
 ↪  alpha_bootstrap, R = 1000)
boot.ci(boot_alpha_f3, type = "bca")

#représentation des alpha de Cronbach avec IC en tableau
alpha_df <- data.frame(
    Scale = c("Global", "F1_somatique", "F2_asthenie", "F3_depressif"),
    Alpha = c(
        round(cronbach(hdrs_J0[,hdrs_items])$alpha, 3),

         ↪  round(cronbach(hdrs_J0[,c("8_ralentissement","12_symptomes_gastro","14_symptomes_ge
         ↪  3),

         ↪  round(cronbach(hdrs_J0[,c("4_insomnie_debut","13_symptomes_generaux")])$alpha,
         ↪  3),

         ↪  round(cronbach(hdrs_J0[,c("1_humeur_depressive","2_sentim_culpabilite","3_suicide",
         ↪  3)
    ),
    CI_lower = c(
        round(boot.ci(boot_alpha_global, type = "bca")$bca[4], 3),
        round(boot.ci(boot_alpha_f1, type = "bca")$bca[4], 3),
        round(boot.ci(boot_alpha_f2, type = "bca")$bca[4], 3),
        round(boot.ci(boot_alpha_f3, type = "bca")$bca[4], 3)
    ),
    CI_upper = c(
        round(boot.ci(boot_alpha_global, type = "bca")$bca[5], 3),
        round(boot.ci(boot_alpha_f1, type = "bca")$bca[5], 3),
```

```
        round(boot.ci(boot_alpha_f2, type = "bca")$bca[5], 3),
        round(boot.ci(boot_alpha_f3, type = "bca")$bca[5], 3)
    )
)
knitr::kable(
    alpha_df,
    caption = "Alpha de Cronbach et intervalles de confiance à 95% pour l'échelle
    ↪  de Hamilton à J0 et ses sous-échelles",
    booktabs = TRUE,
    align = "lccc"
)

#création fichier large avec hdrs_J0 et scl90_dim à J0
scl90_J0 <- subset(scl90_dim, VISIT == "J0")
scl90_J0 <- scl90_J0[,c("NUMERO","VISIT",dimensions)]
scl90_J0 <- scl90_J0[order(scl90_J0$NUMERO), ]
hdrs_J0 <- hdrs_J0[order(hdrs_J0$NUMERO), ]
hdrs_scl90_J0 <- merge(hdrs_J0, scl90_J0, by = c("NUMERO", "VISIT"), all.x = TRUE)

# Matrice de corrélation complète
corr_validite_J0 <- cor(
    hdrs_scl90_J0[, c("score", dimensions)],
    use = "pairwise.complete.obs"
)

# Ne garder que la première ligne (par exemple "score")
corr_validite_J0_ligne1 <- corr_validite_J0["score", , drop = FALSE]

# Corrplot
corrplot(
    corr_validite_J0_ligne1,
    method = "color",
    type = "full",
    tl.col = "black",
    addCoef.col = "white",
    number.cex = 0.5,
    tl.cex = 0.4,
    tl.srt = 30,
    col = viridis::plasma(100)
)

corr_validité_sousscores_J0 <-
 ↪  cor(hdrs_scl90_J0[,c("f1_somatique","f2_asthenie","f3_depressif",
 ↪  dimensions)], use = "pairwise.complete.obs")

#ne garder que les 3 lignes (sous-scores)
corr_validité_sousscores_J0 <-
 ↪  corr_validité_sousscores_J0[c("f1_somatique","f2_asthenie","f3_depressif"), ,
 ↪  drop = FALSE]
```

```r
# Corrplot
corrplot(
    corr_validité_sousscores_J0,
    method = "color",
    type = "full",
    tl.col = "black",
    addCoef.col = "white",
    number.cex = 0.5,
    tl.cex = 0.4,
    tl.srt = 30,
    col = viridis::plasma(100)
)


hdrs_J56 <- subset(hdrs_groupe, VISIT == "J56")
# Vraie palette Nord (package nord)
# On prend une palette qualitative (aurora) et on l'étend à 5 couleurs
cols_nord <- nord::nord("aurora", 5)
# Première série de graphiques (9 au maximum)
par(mfrow = c(3, 3), mar = c(2, 2, 2, 1))
items1 <- hdrs_items[1:min(9, length(hdrs_items))]
for (item in items1) {
    val <- na.omit(hdrs_J56[[item]])
    if (length(val) > 0) {
        m <- max(val)
        hist(val,
            main = item,
            xlab = "Score",
            col = cols_nord[1:(m + 1)],
            border = "white",
            breaks = seq(-0.5, m + 0.5, 1),
            xaxt = "n")
        axis(1, at = 0:m)
    } else {
        plot.new()
        title(main = paste(item, "(pas de données)"))
    }
    }
par(mfrow = c(1, 1))
# Deuxième série de graphiques (8 au maximum, de 10 à 17)
if (length(hdrs_items) > 9) {
    par(mfrow = c(3, 3), mar = c(2, 2, 2, 1))
    items2 <- hdrs_items[10:min(17, length(hdrs_items))]
    for (item in items2) {
        val <- na.omit(hdrs_J56[[item]])
        if (length(val) > 0) {
        m <- max(val)
```

```
        hist(val,
            main = item,
            xlab = "Score",
            col = cols_nord[1:(m + 1)],
            border = "white",
            breaks = seq(-0.5, m + 0.5, 1),
            xaxt = "n")
        axis(1, at = 0:m)
        } else {
        plot.new()
        title(main = paste(item, "(pas de données)"))
        }
    }
    par(mfrow = c(1, 1))
}

hdrs_J56_matrix <- hdrs_J56[, hdrs_items]
corr_matrix_J56 <- cor(hdrs_J56_matrix, use = "pairwise.complete.obs")

corrplot(corr_matrix_J56,
            method = "color",
            type = "upper",
            tl.col = "black",
            addCoef.col = "white",
            number.cex = 0.35,
            tl.cex = 0.6,
            tl.srt = 45,
            col = viridis::plasma(100)
            )


# lire le fichier code généré
code <-
 ↪  readLines("/Users/thomashusson/Documents/Projets/M2biostatistiques/devoir_stats_avancees/al
 ↪  warn = FALSE)

cat("```r\n")
cat(code, sep = "\n")
cat("\n```")
```