Notes S1 M2 MSR

Thomas Husson

Table des matières

1 Mes notes de cours **15**

2 Résumé statistique **16**
1 Position et dispersion . . . 16
1.A Position . . . 16
1.A.1 Moyenne . . . 16
1.A.2 Médiane . . . 16
1.A.3 Mode . . . 16
1.B Dispersion . . . 16
1.B.1 Étendue = empan . . . 17
1.B.2 Écart interquartile (IQR) . . . 17
1.B.3 Écart-type . . . 17
1.B.4 Variance . . . 18
1.C Exemple sur R . . . 18
2 Analyses en sous groupe . . . 19
2.A Principe . . . 19
2.B Dans R . . . 19
3 Dépendance, liaison et association . . . 23
3.A Variables quantitatives . . . 23
3.A.1 Dépendance . . . 24
3.A.2 Dépendance monotone ou linéaire . . . 24
3.A.2.1 Exemple sur R . . . 25
3.A.2.2 Données simulées . . . 27
3.A.3 Concordance . . . 31
3.A.3.1 Principle . . . 31
3.A.3.2 Exemple sur R . . . 32
3.A.4 Résumé des paramètres de dépendance entre deux variables quantitatives . . . 35
3.B Variables catégorielles . . . 36
3.B.1 Dépendance . . . 36
3.B.1.1 Chi2 et associés . . . 36
3.B.1.2 Odds-ratio et risque relatif . . . 37
3.B.1.3 Exemple sur R . . . 37
3.B.2 Dépendance monotone . . . 38
3.B.2.1 Exemple sur R . . . 39
3.B.3 Concordance . . . 40
3.B.3.1 Coefficient kappa de Cohen . . . 40
3.B.3.2 Sensibilité, spécificité, VPP, VPN . . . 40

3 Tests statistiques **44**
1 Introduction : en pratique, comment utiliser les tests statistiques ? . . . 45
1.A À quoi servent vraiment les tests statistiques ? . . . 45
1.B Quand l'usage est légitime . . . 45
1.C Mésusages fréquents (et pourquoi ils posent problème) . . . 45
1.D Fil conducteur à retenir . . . 47
2 Quels tests pour quelles hypohèses ? . . . 47
2.A Comparaison de 2 pourcentages . . . 47
2.A.1 Test du chi2 de Pearson . . . 47
2.A.1.1 Chi2 R . . . 47
2.A.2 Test exact de Fisher . . . 51
2.A.2.1 Fisher R . . . 51
2.B Comparaison d'un pourcentage a un pourcentage théorique . . . 53
2.C Comparaison de 3 pourcentages . . . 53
2.D Comparaison de 2 moyennes . . . 56
2.D.0.1 Comparaison de 2 moyennes avec R . . . 57

2.E Comparaison de 3 moyennes ou plus . . . 59
2.E.1 Exemple R . . . 59
2.F Test de nullité d'un coefficient de corrélation . . . 62
2.F.1 Test de nullité du coefficient de corrélation avec R . . . 62
2.G Tests appariés / non appariés . . . 63
2.G.1 Exemple R données appariées . . . 63
2.G.1.1 Exemple 1 . . . 63
2.G.1.2 Exemple 2 . . . 65
2.G.1.3 tbl_summary apparié . . . 66
3 Calcul de β et du nombre de sujets à inclure . . . 69
3.A Exemple R . . . 70
4 Tests uni- ou bilatéraux ? . . . 72
4.A Définitions . . . 72
4.B Exemple . . . 72
5 Intervalles de confiance . . . 73
5.A Définitions . . . 73
5.B Exemple R . . . 73

4 Modèles **75**
1 Modèle linéaire . . . 76
1.A Principe et mise en œuvre . . . 76
1.B Exemple avec le jeu de données `smp` . . . 76
2 Corrélation et modèle linéaire . . . 80
3 Le test t : un cas particulier du modèle linéaire . . . 84
3.A Exemple R . . . 84
3.A.1 Test t . . . 84
3.A.2 Régression linéaire . . . 85
3.B Relation entre test t et régression linéaire . . . 85
4 Introduction aux GLM . . . 87
4.A Comment ça marche les GLM ? . . . 87
4.B Le prédicteur linéaire . . . 87
4.C La fonction de lien . . . 88
4.C.1 Tableau synthétique des modèles linéaires généralisés (GLM) . . . 89
4.D La structure d'erreur . . . 90
4.E Maximum de vraisemblance et déviance . . . 90
5 Modèle logistique . . . 91
5.A Modèle linéaire inadapté pour cas témoins . . . 91
5.B Principe et mise en œuvre . . . 91
5.C Transformation de Y . . . 92
5.C.1 Odds . . . 92
5.C.2 Log-odds . . . 92
5.D Maximum de vraisemblance . . . 92
5.E Conditions de validité . . . 93
5.F Exemple R . . . 93
5.F.1 Description des variables . . . 93
5.F.2 Fonction glm . . . 93
5.F.3 Tester l'absence d'effet d'une variable . . . 95
6 Modèle log-binomial . . . 97
6.A Principe et mise en œuvre . . . 97
6.B Exemple R . . . 97
6.C Problème du modèle log-binomial . . . 99
6.D Limitations du modèle log-binomial et alternative pratique . . . 99
6.E Alternative : calculer un RR marginal à partir d'un modèle logistique . . . 100
7 Modèle logistique pour *odds* proportionnels . . . 101
8 Modèle de Poisson et binomial négatif pour taux d'incidence . . . 102
8.A Pourquoi les modèles linéaires classiques ne sont pas adaptés . . . 102
8.B La distribution de Poisson . . . 102
8.C Caractéristique du GLM de Poisson . . . 103

8.D Conditions de validité . . . 104
8.D.1 Indépendance des réponses . . . 104
8.D.2 Distribution des réponses . . . 104
8.D.3 Absence de surdispersion . . . 104
8.D.3.1 Comment mesurer la surdispersion . . . 104
8.D.3.2 Causes fréquentes de surdispersion . . . 105
8.D.3.3 Que faire en cas de surdispersion ? . . . 105
8.E Exemple R . . . 105
8.E.1 Modèle de Poisson . . . 106
8.E.2 Modèle quasi-Poisson . . . 110
8.E.3 Modèle binomial négatif . . . 111
8.E.4 Comment choisir directement le meilleur modèle . . . 114
9 Modèles de survie . . . 115
9.A Modèles de survie paramétrique : Weibull etc . . . 115
9.B Modèles de survie semi-paramétrique = Modèle de Cox . . . 115
9.B.1 Hypothèse de proportionnalité des risques . . . 115
9.B.2 Hypothèse de log-linéarité . . . 116
9.B.3 Exemple R . . . 116

5 Autour_de_la_modélisation **119**
1 Introduction . . . 120
2 Codages des variables explicatives quantitatives . . . 120
2.A Problématique . . . 120
2.B Exemple R sur les modèles additifs généralisés (GAM) . . . 120
2.B.1 Modèle additif généralisé (GAM) avec splines pour une variable explicative quantitative 121
2.B.2 Modèle additif généralisé (GAM) avec splines pour une régression logistique . . . 123
2.B.2.1 Découpage en classes . . . 124
2.B.2.2 Polynômes orthogonaux . . . 126
3 Codage des variables explicatives catégorielles . . . 128
3.A Exemple : ABO . . . 129
3.B Exemple R . . . 130
3.B.1 Avec une variable codée 0/1 . . . 130
3.B.2 Avec une variable codée -1/1 . . . 132
4 Choix des variables explicatives . . . 135
4.A Principe . . . 135
5 À propos des termes d'interaction . . . 136
6 Données manquantes . . . 137
6.A Types de données manquantes . . . 137
6.B Gestion des données manquantes . . . 137
6.C Exemple R avec le package `mice` . . . 137
7 Bootstrap . . . 140
7.A Principe . . . 140
7.B Conditions à respecter pour utiliser le bootstrap : . . . 141
7.C Diagnostics et avantages . . . 142
7.D Exemple R . . . 143

6 GLM 1 : Modèles linéaires généralisés **147**
1 Introduction . . . 147
1.A Retour sur le modèle linéaire simple . . . 147
1.B Modèles linéaires généralisés (GLM) . . . 148
1.B.1 Modèle logistique . . . 149
2 Définition des risques . . . 149
2.A Exemple . . . 149
2.B Mesures d'associations . . . 151
2.C Interprétation des risques . . . 151
2.C.1 Exemple . . . 152
2.D Calcul des OR et RR dans un tableau de contingence . . . 152
2.D.1 Exemple numérique avec R . . . 153
2.D.2 Intervalles de confiance des OR et RR . . . 154

3 Régression logistique univariable . . . 154
3.A Introduction . . . 154
3.B Transformation logistique . . . 155
3.C Modèle logistique . . . 155
3.C.1 Si X est une variable binaire (0/1) : . . . 155
3.C.2 Si X est une variable catégorielle à 3 valeurs (0, 1, 2) : . . . 156
3.C.3 Si X est une variable catégorielle à K valeurs . . . 156
3.C.4 Si X est une variable continue / quantitative . . . 157
3.D Fonction de vraisemblance . . . 157
3.E Test de Wald . . . 157
3.F Exemple avec R . . . 157
4 Régression logistique multivariable . . . 159
4.A Principe . . . 159
4.A.1 Principe général . . . 159
4.A.2 Hypothèse de multiplicativité des effets . . . 159
4.A.3 Notion d'interaction . . . 160

7 GLM 1 TP : Modèles linéaires généralisés **165**
1 Consigne . . . 165
2 Question 1 = Table descriptive . . . 165
2.A Recode de la variable "Group" . . . 166
2.B Construction du tableau . . . 167
2.B.1 Version nullach !!! . . . 167
2.C Autres façons de faire . . . 170
2.C.1 tableone . . . 170
2.C.2 gtsummary : je vois pas pourquoi on utilise pas ce banger ! . . . 171

8 Survie **173**
1 Plan du cours . . . 173
2 Introduction . . . 173
2.A Données de survie . . . 173
2.B Problème de la censure . . . 174
3 Données de survie . . . 175
3.A Modélisation satistique . . . 175
3.B Applications . . . 176
3.C Terminologie . . . 176
3.D Ce qui est nécessaire . . . 176
3.D.1 Exercice . . . 179
3.E Loi de probabilité de T . . . 180
3.E.1 **Densité de probabilité** . . . 180
3.E.2 Fonction de répartition . . . 181
3.E.3 Fonction de survie . . . 183
3.E.4 Fonction de risque instantané . . . 184
3.E.5 Fonction de risque cumulée . . . 185
4 #### Relations entre les fonctions . . . 185
5 Estimation d'une courbe de survie : estimateur de Kaplan-Meier . . . 186
5.A Objectif de l'analyse de survie . . . 186
5.B Approches . . . 186
5.B.1 Estimation en absence de censure . . . 186
5.B.1.1 Exemple . . . 187
5.B.1.2 2e exemple avec temps simulés . . . 189
5.B.2 Estimation en présence de censure : probabilités conditionnelles . . . 192
5.B.2.1 Exemple 1 . . . 193
5.B.2.2 Exemple 2 avec R . . . 194
6 Comparaison de courbes de survie : test de Log-Rank . . . 203
6.A Principe . . . 203
6.B Test de Log-Rank dans R . . . 203
7 Modèle de Cox : analyse multivariée de survie . . . 204
7.A Hypothèse de proportionnalité des risques . . . 205

7.B Hypothèse de log-linéarité . . . 205
7.C Interprétation des coefficients du modèle de Cox . . . 206
7.C.1 Variable discrète . . . 206
7.C.2 Variable continue . . . 207
7.C.3 Interprétation du HR . . . 208
7.C.4 Résumé . . . 208
7.C.5 Tableau de synthèse : interprétation des coefficients du modèle de Cox . . . 208
7.D Tests statistiques pour les coefficients du modèle de Cox . . . 209
7.D.1 Hypothèses . . . 209
7.D.2 Tests disponibles : Wald et Rapport de vraisemblance (Likelihood Ratio Test, LRT) . 209
7.D.2.1 Test de Wald . . . 209
7.D.2.2 Test du rapport de vraisemblance (Likelihood Ratio Test, LRT) . . . 210
7.D.2.3 Comparaison des deux tests . . . 211
8 TP . . . 212
8.A Modèle de Cox avec R . . . 212
8.A.1 Exemple avec le jeu de données `lung` . . . 212
8.A.1.1 Charger et afficher les premières lignes du jeu de données `lung` . . . 212
8.A.1.2 Créer un objet de type `Surv` avec les colonnes `time` et `status` . . . 213
8.A.1.3 Tester l'effet du sexe (`sex`) sur la survie avec un log-rank test . . . 213
8.A.1.4 Ajuster un modèle de Cox avec `sex` comme covariable . . . 214
8.A.1.5 Obtenir l'intervalle de confiance à 95 % pour le hazard ratio . . . 217
9 TP 2 . . . 218
9.A Réponse . . . 218
9.A.1 Nombre d'évènements et de censures dans chaque groupe . . . 219
9.A.2 Nombre de sujets exposés au risque de rechute à 12 et 18 semaines dans chaque groupe 220
9.A.3 Chi^2 de Pearson approprié ? . . . 221
9.A.4 Survie sans rechute à 6 et 12 semaines dans les deux groupes de traitement . . . 221
9.A.5 Médiane de survie sans rechute dans chaque groupe de traitement . . . 222
9.A.6 Estimer et tracer les courbes de survie de Kaplan-Meier pour les deux groupes de traitement . . . 222
9.A.7 Quel test pour comparer les courbes de survie entre les deux groupes de traitement ? Justifiez votre choix. . . . 224
10 TP 3 . . . 225
10.A Questions . . . 225
10.B Réponse . . . 225
10.B.1 Description des données . . . 225
10.B.2 Estimation d'une courbe de survie de Kaplan-Meier globale . . . 226
10.B.3 Influence du sexe sur la survie . . . 227
10.B.4 Test de Log-Rank et Modèle de Cox pour comparer les courbes de survie entre les groupes . . . 229
10.B.5 Si on regarde en ajustant sur l'épaisseur de la tumeur . . . 230
10.B.6 Interaction entre sexe et ulcération . . . 233

9 GLM Poisson **236**
1 Plan du cours . . . 236
2 Introduction : données de comptage, données longtiduinales, données en cluster . . . 236
2.A Données de comptage . . . 236
2.A.1 Exemple de données de comptage . . . 237
2.A.2 Approximation binomiale / Poisson . . . 238
2.A.3 Surdispersion . . . 239
2.A.4 Excès de 0 . . . 239
3 Modèle de poisson . . . 240
3.A Modèle . . . 240
3.B Interprétation des coefficients . . . 241
3.C Tests statistiques . . . 243
3.D Exemple avec R . . . 243
3.D.1 1 . . . 243
3.D.2 2 . . . 244
3.D.3 3 . . . 246

3.E Variance “sandwich” . . . 248
3.E.1 Exemple avec R . . . 249
3.F Traitement de la surdispersion . . . 251
3.G Traitement des excès de 0 . . . 252
4 Résumé . . . 252
4.A Données de comptage et histoire binomiale Poisson . . . 252
4.B Modèle de Poisson et interprétation des coefficients . . . 253
4.C Surdispersion : définition, conséquences et illustration R . . . 254
4.D Variance robuste (« variance sandwich ») . . . 255
4.E Excès de zéros : modèles Hurdle et Zero-Inflated . . . 256
4.E.1 Modèle Hurdle . . . 256
4.E.2 Modèle Zero-Inflated . . . 257
4.F En pratique : . . . 257
5 Application . . . 258
6 Références . . . 261

10 Modèles linéaires mixtes **262**
1 Introduction aux modèles linéaires mixtes . . . 263
1.A Quelques stratégies pour gérer la non-indépendance des observations . . . 263
1.B Le modèle linéaire mixte (formulation mathématique) . . . 264
1.C Estimation . . . 264
1.D Types d’estimateurs . . . 265
2 Utilisation de R pour les modèles linéaires mixtes . . . 266
2.A Représentations . . . 266
2.A.1 Version Wide (large) . . . 266
2.A.2 Version Long (longue) . . . 266
2.B Modèle linéaire mixte avec R . . . 270
2.B.1 Utilisation de la fonction `lmer()` du package `lme4` . . . 270
2.B.2 Utilisation de la fonction `lme()` du package `nlme` . . . 272
2.B.3 Exemple 3 : ‘lmerTest’ : obtention de p-values pour les effets fixes . . . 274
2.B.4 Pertinence de la variable `time` globale dans le modèle . . . 275
2.B.5 Pertinence de la variable `sex` dans le modèle . . . 276
3 Applications (études publiées) . . . 276
3.A Déclin cognitif des personnes âgées aux États-Unis pendant et après la Covid !! . . . 276
3.B Analyse statistique . . . 277
3.C Résultats . . . 278
4 Modèles linéaires mixtes : extension . . . 278
4.A Modèles non linéaires mixtes (NLMM) . . . 278
4.B Modèles linéaires généralisées mixtes (GLMM) . . . 279

11 Résumé Modèle de Poisson et MLL **280**

12 Modèle de Poisson **281**

13 Modèle linéaire mixte **282**

14 Modèles linéaires mixtes TP **283**
1 Impact du sexe sur la distance, ajusté sur le temps et un effet aléatoire patient. . . . 283
1.A Modèle avec temps en tant que variable catégorielle . . . 283
1.A.1 Si on veut regarder s’il y a un effet-temps différent selon le sexe . . . 285
1.B Modèle avec temps en tant que variable quantitative . . . 286

15 Wrap up modèles linéaires mixtes **289**
1 1. Régression Linéaire Multiple (RLM) . . . 289
2 2. Régression Logistique . . . 290
3 3. Régression de Poisson (comptages / GLM Poisson) . . . 292
4 4. Modèles Linéaires Mixtes (effets fixes + aléatoires) . . . 293

16 Variabilités interindividuelles **297**
1 Contexte : données hiérarchiques . . . 297

2 Variabilité inter-individuelle . . . 297
2.A Définition intuitive . . . 297
2.B Interprétation . . . 298
3 Variabilité inter-centre . . . 298
3.A Idée de base . . . 298
3.B Coefficient de corrélation intraclasse (ICC) . . . 299
4 Modèle linéaire mixte : séparer les deux variabilités . . . 299
4.A Modèle avec intercept aléatoire de centre . . . 299
4.B Implications pratiques . . . 299
5 Pourquoi on ne retrouve pas ces variances avec des modèles séparés . . . 300
5.A Modèles “par centre” . . . 300
5.B Modèles “par patient” (données longitudinales) . . . 301
6 Illustration rapide en R . . . 301
6.A Simulation avec centres . . . 301
6.B Modèle mixte . . . 303
6.C Modèles séparés par centre . . . 304
6.D Résumé . . . 305

17 Effets fixes et aléatoires **306**
1 Contexte : pourquoi parler d’effets fixes et aléatoires ? . . . 306
2 Effets fixes : covariables dont on veut estimer l’effet . . . 306
2.A Définition intuitive . . . 306
2.B Ce que signifie « covariables dont on veut estimer l’effet » . . . 307
3 Effets aléatoires : modéliser la structure de corrélation . . . 307
3.A Idée de base . . . 307
3.B « Variables qui modélisent la structure de corrélation » . . . 308
3.B.1 Exemple: intercept aléatoire de centre $(1 \mid \text{centre})$ . . . 308
3.B.2 Exemple: Intercept aléatoire de patient $(1 \mid \text{patient})$ dans des données longitudinales 308
3.B.3 Exemple: Pente aléatoire $(\text{temps} \mid \text{patient})$ . . . 309
4 Résumé conceptuel : ce qui distingue effets fixes et aléatoires . . . 309
5 Exemple simple en R . . . 310
6 En une phrase . . . 312

18 Intercept variances effets_aleatoires **313**
1 Contexte et objectif . . . 313
2 Rappel : droite de régression et « intercept » . . . 313
2.A Modèle linéaire simple . . . 313
2.B Qu’est-ce que l’« intercept » ? . . . 313
2.C Qu’est-ce que la pente ? . . . 316
3 Ajouter des groupes : modèle linéaire mixte simple . . . 316
4 Les deux variances : entre groupes et à l’intérieur des groupes . . . 317
4.A Variance d’une observation Y_{ij} . . . 317
4.B Corrélation entre deux observations d’un même groupe . . . 318
5 Matrice de variance–covariance : expliquer simplement . . . 319
5.A Exemple : intercept et pente aléatoires . . . 319
6 Effets aléatoires emboîtés vs effets aléatoires croisés . . . 320
6.A Effets aléatoires emboîtés (hierarchiques) . . . 320
6.B Effets aléatoires croisés . . . 320
7 Petit exemple R pour illustrer les deux variances . . . 321
8 Résumé . . . 324

19 Wrap Up Régressions **325**
1 Régressions linéaires multiples . . . 325
1.A Conditions de validité : . . . 325
1.B Corrélatons entre variables explicatives . . . 326
1.B.1 Exemple d’interprétation avec interaction . . . 326
1.B.2 Si X1 était catégorielle à plus de 2 modalités ? . . . 327
1.C Réponse aux questions . . . 328

2 Régression logistique . . . 329
2.A Interactions . . . 329
2.B Conditions de validité de la régression logistique . . . 329
2.C Autres trucs : . . . 329
3 Références . . . 329

20 Effet centre **330**
1 Introduction . . . 331
2 Modèle linéaire . . . 331
2.A Analyse naïve sans prise en compte du centre . . . 331
2.A.1 Problèmes posés par cette analyse naïve . . . 332
2.A.2 Exemple R avec jeu de données fictif . . . 334
2.B Analyse avec prise en compte de l'effet centre . . . 336
2.B.1 Exemple R . . . 336
2.C Analyse intra-centres (modèles conditionnels, mixtes ou non) . . . 337
2.C.1 Modèle linéaire avec effet fixe par centre . . . 337
2.C.2 Modèle linéaire avec pente commune (intercept variable par centre) . . . 338
2.C.3 Modèle mixte avec effet aléatoire de centre . . . 339
2.C.4 Exemple R . . . 341
2.C.5 Conditions de validités des modèles mixtes . . . 343
2.D Modèle marginal = GEE (Generalized Estimating Equations) . . . 343
2.D.1 Définition . . . 343
2.D.2 Principe des GEE . . . 344
3 Modèle linéaire généralisé . . . 347
3.A Principes généraux . . . 347
3.B Exemple R avec données simulées . . . 347
4 Introduction aux séries chronologiques . . . 366
4.A Modèle auto-régressif d'ordre 1 : . . . 366
4.A.1 Qu'est-ce qu'une régression (dans ce contexte) ? . . . 366
4.A.2 Pourquoi dit-on souvent que « $a = 0$ » ? . . . 367
4.A.3 Variance : définition générale . . . 367
4.A.4 Variance du bruit : ce que c'est exactement . . . 367
4.A.5 Propagation de la variance dans un AR(1) . . . 369
4.A.6 Variance stationnaire : définition explicite . . . 369
4.A.7 Corrélation temporelle : définition explicite . . . 370
4.A.8 Exemple chiffré complet . . . 370
4.B Simulation courte . . . 370
4.C Corrélation temporelle . . . 371
5 Modèle MA = Moving Average . . . 371
6 Autocorrélogramme partiel . . . 372
6.A Exemple 1 . . . 372
6.B Exemple 2 . . . 373
6.B.1 Modèle autorégressif AR(1) . . . 373
6.B.2 Modèle moyenne mobile MA(1) . . . 373
6.B.3 Autocorrélogramme . . . 374
6.B.4 Autocorrélogramme partiel . . . 374
6.B.5 Processus stationnaire . . . 374
7 Différents modèles : AR, MA, ARMA, SARMA, SARIMA, SARIMAX . . . 374
8 Exemple d'analyse de données avec R . . . 375
8.A Méthode graphique . . . 375
8.B Méthode automatique avec fonction `auto.arima` . . . 379
9 Fonction `stl()` . . . 382

21 Series chronologiques (vidéo) **385**
1 Introduction aux séries chronologiques . . . 385
1.A Modèle auto-régressif d'ordre 1 : . . . 385
1.A.1 Qu'est-ce qu'une régression (dans ce contexte) ? . . . 385
1.A.2 Pourquoi dit-on souvent que « $a = 0$ » ? . . . 385
1.A.3 Variance : définition générale . . . 386

1.A.4 Variance du bruit : ce que c'est exactement . . . 386
1.A.5 Propagation de la variance dans un AR(1) . . . 388
1.A.6 Variance stationnaire : définition explicite . . . 388
1.A.7 Corrélation temporelle : définition explicite . . . 389
1.A.8 Exemple chiffré complet . . . 389
1.B Simulation courte . . . 389
1.C Corrélation temporelle . . . 390
2 Modèle MA = Moving Average . . . 390
3 Autocorrélogramme partiel . . . 391
3.A Exemple 1 . . . 391
3.B Exemple 2 . . . 392
3.B.1 Modèle autorégressif AR(1) . . . 392
3.B.2 Modèle moyenne mobile MA(1) . . . 392
3.B.3 Autocorrélogramme . . . 392
3.B.4 Autocorrélogramme partiel . . . 393
3.B.5 Processus stationnaire . . . 393
4 Différents modèles : AR, MA, ARMA, SARMA, SARIMA, SARIMAX . . . 393
5 Exemple d'analyse de données avec R . . . 393
5.A Méthode graphique . . . 393
5.B Méthode automatique avec fonction `auto.arima` . . . 397
6 Fonction `stl()` . . . 400

22 Comparaisons Multiples **402**
1 Introduction . . . 402
2 Quand ajuster ? . . . 402
3 Comment ajuster ? . . . 402
3.A Exemple . . . 402
3.B Méthode de Bonferroni = répartition du risque alpha . . . 403
3.C Méthode de Holm = répartition du risque alpha . . . 403
3.D Méthode par hiérarchie . . . 403
4 En dehors des essais thérapeutiques . . . 403
4.A Génomique . . . 403
5 Questions . . . 404

23 Donnees manquantes **405**
1 Introduction aux données manquantes . . . 405
2 Types de données manquantes . . . 405
2.A MCAR : Missing Completely At Random . . . 405
2.B MAR : Missing At Random . . . 405
2.C MNAR : Missing Not At Random . . . 406
3 Imputation multiple . . . 406
3.A Principe de l'imputation multiple . . . 406
3.A.1 1ère étape : tirage au sort . . . 407
3.A.2 2ème étape : régressions . . . 407
3.A.3 3ème étape : itérations . . . 407
3.A.4 4ème étape : répétition . . . 407
3.A.5 5ème étape : analyse . . . 407
3.B Gérer les analyses après imputation multiple . . . 408

24 Bootstrap **409**
1 Introduction . . . 409
2 Principe du bootstrap . . . 409
2.A Tirage avec remise . . . 409
2.B Histograme des 10 000 valeurs de b . . . 410
2.C Faisabilité du bootstrap . . . 411
3 Exemple en R sur expsy . . . 411
4 Second exemple en R sur SMP (CUSM) . . . 415

25 Méthodes non supervisées **419**
1 Introduction aux méthodes non supervisées . . . 420
1.A Groupes homogènes de sujets . . . 420
1.A.1 Analyse en composantes principales . . . 420
1.A.1.1 Principe . . . 420
1.A.1.2 Application pratique avec R . . . 421
1.A.2 Classifications hiérarchiques . . . 428
1.A.2.1 Principe . . . 428
1.A.2.2 Application pratique avec R . . . 428
1.A.3 Nuées dynamiques . . . 433

26 Méthodes non supervisées 2 **434**
1 Introduction . . . 434
2 Principe des composantes principales . . . 434
3 Utilisations en statistiques . . . 443
4 Plan sur lequel les points sont des variables plutôt que des sujets . . . 443
5 Représentation en sphère . . . 445
6 En d'autres termes : . . . 447
7 Conclusion . . . 447
8 Références . . . 448

27 Methodes multidimensionnelles 3 **449**
1 Exemple 1 . . . 450
1.A Représentation des variables . . . 451
1.B Interprétation sociologique . . . 451
1.C Représentations des individus par groupes . . . 451
2 Exemple 2 : Échelle de dépression "Depressive mood scale" . . . 453
2.A Variances de l'ACP . . . 454
2.B Inteprétation : . . . 454
2.C Représentation sphérique : gagne une dimension . . . 456
3 Analyse en composante principale focalisée . . . 456
4 Références . . . 458

28 Analyse Factorielle **459**
1 Principes de l'analyse factorielle . . . 459
2 Analyse factorielle à un facteur . . . 459
2.A Modélisation . . . 459
2.B Exemple . . . 459
3 Utilisation : surtout analyse factorielle à > 1 facteur . . . 460
3.A Indétermination . . . 460
3.B Rotation . . . 460
5 A nouveau l'exemple sur la dépression : . . . 462
5.A Analyse factorielle à 4 facteurs avec rotation varimax. . . . 462
6 Relation ACP / Analyse factorielle . . . 462
7 Comment choisir le nb de facteurs ? . . . 463
8 Références . . . 465

29 Clinimetrie **466**
1 Définition de la clinimétrie . . . 466
2 Accord inter-juge . . . 466
2.A Accord inter-juge d'une variable binaire . . . 466
2.A.1 **Exemple : diagnostic d'une appendicite (OUI/NON)** . . . 466
2.A.2 **Concordance** . . . 466
2.A.3 **Concordance corrigée** . . . 467
2.A.4 Exemple diagnostic à plusieurs classes . . . 468
2.B Acord inter-juge d'une variable ordonnée . . . 469
2.C Avec 3 juges . . . 469
2.D Interprétation du Kappa de Cohen . . . 470
2.E Intervalle de confiance du Kappa . . . 470
2.F Différence avec PABAK . . . 470

2.G Exemples sur R . . . 472
2.G.1 2 catégories . . . 472
2.G.2 3 catégories . . . 473
2.G.3 kappa pondéré . . . 474
2.G.4 kappa de light (3 juges) . . . 474
3 Accord inter-juge pour variable quantitative . . . 476
3.A Coefficient de corrélation intraclasse (ICC) . . . 476
3.A.1 Si Y est une variable ordonnée . . . 476
3.B Exemple sur R . . . 476
4 Blant et Altman . . . 478
4.A Principe . . . 479
5 Exemple . . . 479
6 Alpha de Cronbach . . . 481
6.A Exemple . . . 481
6.B Calcul . . . 481
6.C Exemple sur R . . . 482
6.D Interprétation . . . 483
6.D.1 Ce que mesure l'alpha . . . 483
6.D.2 Comment interpréter l'alpha . . . 483
6.D.3 Limite fondamentale . . . 483
7 2. Pourquoi l'alpha nécessite un questionnaire unidimensionnel ? . . . 484
8 3. Comment vérifier l'unidimensionnalité ? . . . 484
8.A Par l'analyse des valeurs propres (scree plot) . . . 484
8.A.1 Rappel : qu'est-ce qu'une valeur propre ? . . . 484
8.A.2 Principe . . . 484
9 4. Exemple complet . . . 485
9.A Questionnaire de 6 items censés mesurer l'anxiété . . . 485
9.A.1 Étape A : valeurs propres obtenues . . . 485
9.A.2 Interprétation . . . 485
9.A.3 Étape B : alpha de Cronbach . . . 485
10 5. Exemple inverse (multidimensionnel) . . . 485
11 6. Synthèse finale . . . 486
11..1 Alpha de Cronbach . . . 486
11..2 Valeurs propres . . . 486
11..3 Exemple . . . 486
12 Questions à l'examen sur ce cours . . . 486
13 1. Ce que mesure Kappa (qualitatif) . . . 486
13.A Tableau 2×2 (2 juges, variable binaire) . . . 486
13.A.1 Concordance brute . . . 487
13.A.2 Concordance liée au hasard . . . 487
13.A.3 Kappa de Cohen . . . 487
13.A.4 Extensions . . . 487
13.A.5 Limites . . . 487
14 2. Ce que mesure l'ICC (quantitatif) . . . 487
14..1 Cours, « Tip » . . . 488
14..2 Cours, 3.A.1. . . . 488
15 3. En quoi κ et ICC sont proches . . . 488
16 4. En quoi ils sont différents . . . 488
16.A Type de données . . . 488
16.B Méthode de calcul . . . 489
16.C Sensibilité aux marges . . . 489
16.D Structure d'erreur . . . 489
17 5. Exemple concret du cours (rb1 / rb2) . . . 489
17.A Kappa de Cohen . . . 489
17.A.1 Concordance brute . . . 489
17.A.2 Marges . . . 489
17.A.3 Concordance au hasard . . . 490
17.A.4 Kappa . . . 490
17.B ICC (cours) . . . 490

18 6. Synthèse courte . . . 490
18.A Proches car : . . . 490
18.B Différents car : . . . 490

30 Wrap Up Clinimétrie **491**
1 Clinimétrie . . . 491
2 Mesure catégorielle (diagnostic clinique) . . . 491
2..1 Kappa de Cohen = concordance corrigée pour le hasard . . . 491
2.A Si plus de 2 juges : . . . 491
3 Mesure continue (ex : pression artérielle) . . . 491
3.A Coefficient de corrélation intraclasse (ICC) . . . 492
4 Alpha de Cronbach . . . 492
4.A Principe du coefficient . . . 492
4.B Interprétation . . . 492

31 Mesures Subjectives **493**
1 Définitions . . . 493
2 Importance de la subjectivité en médecine . . . 493
3 Comment mesurer la subjectivité ? . . . 493
4 Valider un instrument . . . 493
4.A Que vaut la mesure ? = reproductibilité (fiabilité) . . . 494

32 Mesures Subjectives bonus **498**
1 Module 1 : Les Fondamentaux et la Définition de la Mesure . . . 498
1.A 1. Le contexte médical . . . 498
1.B 2. Qu'est-ce que la subjectivité ? . . . 498
1.C 3. Les trois approches de la mesure . . . 498
1.D 4. Subjectif vs Objectif : une fausse opposition . . . 498
1.E 5. Modes de recueil . . . 499
1.F 3. Les Dispositifs de Réponse (Les échelles) . . . 499
1.G 4. La Sélection des Items . . . 499
2 Module 2 : La Construction de l'Instrument . . . 499
2.A 1. Le Cahier des Charges (La fondation) . . . 499
2.B 2. La Chasse aux Items (D'où viennent les questions ?) . . . 500
2.C 3. La Sélection : La méthode du Consensus . . . 500
2.D 4. La Rédaction et les Réponses . . . 500
2.D.1 Module 3 : La Validation (Partie 1) — Que mesure l'instrument ? . . . 500
2.D.1.1 1. La Validité de Contenu ("Est-ce qu'on a tout couvert ?") . . . 501
2.D.1.2 2. La Validité de Structure ("Est-ce cohérent ?") . . . 501
2.D.1.3 3. La Validité de Construit ("Est-ce logique ?") . . . 501
2.D.2 Module 4 : La Validation (Partie 2) — Que vaut la mesure ? . . . 502
2.D.2.1 1. La Précision (ou Fidélité / Reliability) . . . 502
2.D.2.2 2. La Sensibilité au Changement . . . 503
2.D.2.3 3. Les Modèles de Mesure (La "boîte noire" mathématique) . . . 503

33 Méthodes Qualitatives **504**
1 C'est quoi la recherche qualitative ? . . . 504
2 Historique de la recherche qualitative . . . 505
3 6 méthodes prédominantes . . . 506
4 Induction / déduction . . . 507
5 Itération . . . 508
6 Présupposés du chercheur . . . 508
7 Atelier *Posture* . . . 508
8 Journal de bord . . . 509
9 Quel prisme . . . 509
10 Théorisation ancrée (social) . . . 510
11 Méthode par phénoménologie interprétative (individu) = IPA . . . 511
12 Échantillonnage . . . 512

34 Entretiens **514**
1 Critères de scientificité / validité . . . 514

35 Méthodes mixtes **519**
1 Les Fondamentaux de la Recherche Qualitative . . . 519
1.A Diversité des Approches et Origines . . . 519
1.B Exigence de Rigueur et Notions Clés du Protocole . . . 519
1.C Concepts Méthodologiques Essentiels . . . 520
1.C.1 Positionnement du Chercheur : Neutralité et Distance . . . 520
1.C.2 La Question de Recherche et le Corpus . . . 520
1.C.3 L'Échantillonnage et le Codage . . . 521
1.C.3.1 Validation des Résultats . . . 521
2 Le Clivage Quantitatif/Qualitatif et les Méthodes Mixtes . . . 521
2.A Opposition Historique et Philosophique . . . 521
2.A.1 Le Conflit Académique et ses Enjeux . . . 522
2.B La Différence Méthodologique Fondamentale . . . 522
2.C Stratégies des Méthodes Mixtes . . . 523
2.D Vers une Typologie Épistémologique Élargie . . . 523
3 Vers une typologie plus riche : la matrice 3×3 de Falissard et al. (2013) . . . 525
4 Les méthodes mixtes : fondements, usages, typologies . . . 525
4.A Pourquoi mélanger les approches ? . . . 525
4.B Trois stratégies classiques de méthodes mixtes . . . 526
4.B.1 Quali Quanti (exploratoire) . . . 526
4.B.2 Quanti Quali (explicatif) . . . 526
4.B.3 Quanti + Quali en parallèle (triangulation) . . . 526
4.C Pourquoi cela fonctionne-t-il scientifiquement ? . . . 527
5 Applications en psychiatrie, médecine et neurosciences . . . 527
6 Enjeux philosophiques, épistémologiques et politiques . . . 527
7 Conclusion générale : pourquoi dépasser le clivage quali/quanti ? . . . 527

36 Bonus residus **529**
1 Contexte général : ce que fait un modèle . . . 529
2 Observé, prédit, résidu : bien distinguer . . . 529
3 Terme d'erreur ε_i vs résidu e_i . . . 530
3.A Terme d'erreur ε_i (théorique) . . . 530
3.B Résidu e_i (observé) . . . 530
4 Rôle des résidus dans la régression linéaire . . . 531
4.A Méthode des moindres carrés . . . 531
4.B Hypothèses du modèle = hypothèses sur les résidus . . . 531
4.B.1 Moyenne nulle . . . 531
4.B.2 Variance constante (homoscédasticité) . . . 531
4.B.3 Indépendance . . . 532
4.B.4 Normalité (dans le cadre gaussien) . . . 532
5 Exemple concret en R avec un jeu de données fictif . . . 532
5.A Ajuster un modèle linéaire et récupérer les résidus . . . 533
5.A.1 Graphe des résidus vs valeurs prédites . . . 534
5.B Histogramme et QQ-plot des résidus . . . 535
6 Lien avec les modèles plus complexes . . . 537
7 Résumé . . . 537

1Mes notes de cours

Toutes mes notes !

Certains issues du livre CUSM, d’autres des vidéos.

Il y a peut être des doublons : ce n’est pas relu !

2Résumé statistique

1 Position et dispersion

1.A Position

Paramètres de position = valeurs qui résument la tendance centrale d'une distribution.

- Moyenne
- Médiane
- Mode

1.A.1 Moyenne

Moyenne = somme des valeurs divisée par le nombre de valeurs.

$$\frac{1}{n}\sum_{i=1}^{n} x_i$$

Correspond au centre de gravité des points si on les représente sur une droite.
Hypothèses :

- Les valeurs sont indépendantes
- Équivalence de la quantité (1 euro vaut 1 euro quelque soit sa position sur la droite des réels) : donc les notes c'est pas top en vrai !!
- Les valeurs sont continues

1.A.2 Médiane

Signification plus directe : valeur qui partage la distribution en deux parties égales.
Si la distribution est symétrique, la moyenne et la médiane sont égales.

1.A.3 Mode

Mode = valeur la plus fréquente.

1.B Dispersion

Mesures de dispersion = valeurs qui résument la variabilité d'une distribution.

- Étendue = empan

1.B.1 Étendue = empan

Correspond à la différence entre la valeur maximale et la valeur minimale.

1.B.2 Écart interquartile (IQR)

= Q3 - Q1

1.B.3 Écart-type

écart type = écart "le plus typique" par rapport à la moyenne.

Si m est la moyenne des observations x_i, l'écart type s est défini par la racine carrée de la variance :

$$s = \sqrt{[(x_1 - m^2) + \cdots + (x_n - m)^2]/(n-1)}$$

La variance correspond à la moyenne des carrés des écarts par rapport à la moyenne.

Donc en gros : écart type = racine carrée des carrés des écarts par rapport à la moyenne.

Pourquoi ajouter des carrés ?

- Positive les écarts négatifs
- Accentue les écarts importants
- Et pour une super propriété de l'écart-type :
  - La variance de deux variables indépendantes est égale à la somme de leurs variances.

Comment l'interpréter ?

- Dans le cas d'une distribution normale,
  - environ 2/3 des observations se situent à moins d'un écart-type de la moyenne.
  - environ la moitié des observations se situent à $[m - (2/3)s; m + (2/3)s]$

Avantages et inconvénients :

- Avantage :
  - utilise toutes les valeurs de la distribution
  - propriétés mathématiques intéressantes : la variance de la somme de deux variables indépendantes est égale à la somme de leurs variances.
  - s'utilise dans de nombreux tests statistiques (t-test, ANOVA, régression linéaire, etc.)
  - Facile à manipuler mathématiquement
  - Interprétation claire dans le cas de distributions normales
- Inconvénient :
  - sensible aux valeurs extrêmes (outliers)
  - ne s'interprète pas facilement dans le cas de distributions asymétriques
  - Pas adapté aux variables ordinales ou catégorielles

1.B.4 Variance

Variance = écart type au carré
ou moyenne des carrés des valeurs moins le carré de la moyenne.

1.C Exemple sur R

Utilisation du jeu de données `smp.d` (version réduite de `smp`) et de la fonction `summary()`

```
summary(smp.d)
```

```
      age                profession    nb.enfants       depression
 Min.   :19.00   ouvrier      :228   Min.   : 0.000   Min.   :0.0000
 1st Qu.:28.00   sans.emploi  :221   1st Qu.: 0.000   1st Qu.:0.0000
 Median :37.00   employé      :136   Median : 1.000   Median :0.0000
 Mean   :38.94   commerçant   : 91   Mean   : 1.572   Mean   :0.3917
 3rd Qu.:48.00   intermédiaire: 57   3rd Qu.: 2.000   3rd Qu.:1.0000
 Max.   :84.00   (Other)      : 60   Max.   :14.000   Max.   :1.0000
 NA's   :2       NA's         :  6   NA's   :26
 schizophrenie       gravite      recherche.nouv   evit.danger
 Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000
 Median :0.0000   Median :4.000   Median :2.000   Median :2.000
 Mean   :0.0801   Mean   :3.635   Mean   :2.058   Mean   :1.865
 3rd Qu.:0.0000   3rd Qu.:5.000   3rd Qu.:3.000   3rd Qu.:3.000
 Max.   :1.0000   Max.   :7.000   Max.   :3.000   Max.   :3.000
                  NA's   :4       NA's   :104     NA's   :108
 dep.recompense
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.152
 3rd Qu.:3.000
 Max.   :3.000
 NA's   :114
```

Deux inconvénients à la fonction `summary()`:

- Ne donne pas l'écart-type
- La disposition des résultats n'est pas très claire.

On peut utiliser la fonction `describe()` du package `prettyR` pour un résumé plus complet.

```
describe(smp.d)
```

```
Description of smp.d
```

```
 Numeric
                mean median    var    sd valid.n
age            38.94     37 175.72 13.26     797
nb.enfants      1.57      1   3.42  1.85     773
depression      0.39      0   0.24  0.49     799
schizophrenie   0.08      0   0.07  0.27     799
gravite         3.64      4   2.72  1.65     795
recherche.nouv  2.06      2   0.77  0.88     695
evit.danger     1.87      2   0.76  0.87     691
dep.recompense  2.15      2   0.69  0.83     685

 Factor

profession ouvrier sans.emploi employé commerçant intermédiaire autre cadre
   Count    228.00      221.00  136.00      91.00         57.00 31.00    24
   Percent   28.54       27.66   17.02      11.39          7.13  3.88     3

profession <NA> agriculteur
   Count   6.00        5.00
   Percent 0.75        0.63
Mode ouvrier
```

2 Analyses en sous groupe

2.A Principe

Dans un essai thérapeutique, il faut décrire es caractéristiques des patients inclus dans chaque groupe de traitement.

Il faut donc les décrire en fonction de différentes modalités (groupes de traitement, sexe, âge, etc.)

2.B Dans R

On peut aussi utiliser la fonction `table()` pour faire des tableaux de contingence.

```r
table(
    smp.d$profession,
    smp.d$depression,
    deparse.level=2, ## deparse.level fait apparaître les noms des variables dans
     ↪  le tableau
    useNA="ifany")
```

```
                smp.d$depression
smp.d$profession   0   1
```

```
   agriculteur     3   2
   commerçant     65  26
   cadre          16   8
   intermédiaire  31  26
   employé        81  55
   ouvrier       132  96
   autre          22   9
   sans.emploi   132  89
   <NA>            4   2
```

Si on voulait les pourcentages plutôt :

La fonction `prop.table()` permet de calculer des pourcentages à partir d'un tableau de contingence.

L'option `margin` permet de choisir si on veut les pourcentages par ligne (margin=1) ou par colonne (margin=2).

```r
options(digits=3) # pour afficher 3 décimales
prop.table(
    table(
        smp.d$profession,
        smp.d$depression,
        deparse.level=2, # deparse.level fait apparaître les noms des variables
         ↪  dans le tableau
        useNA="ifany"),
    margin=1) # margin=1 pourcentage par ligne ; margin=2 pourcentage par colonne
```

```
                smp.d$depression
smp.d$profession     0     1
   agriculteur   0.600 0.400
   commerçant    0.714 0.286
   cadre         0.667 0.333
   intermédiaire 0.544 0.456
   employé       0.596 0.404
   ouvrier       0.579 0.421
   autre         0.710 0.290
   sans.emploi   0.597 0.403
   <NA>          0.667 0.333
```

Mais encore une fois, je trouve personnellement que le top est d'utiliser `tbl_summary` du package `gtsummary`.

Il va falloir me convaincre de ne pas utiliser ce banger absolu : je ne vois pas pourquoi.

A la limite, pourquoi pas `tableone` aussi.

avec `tableone` : (fait des tests t pour les variables continues et Chi2 / Fisher pour les catégorielles par défaut)

```r
library(tableone)
vars <- c("age","profession","nb.enfants", "gravite","recherche.nouv",
↪  "evit.danger","dep.recompense")
catVars <- c("profession","gravite","recherche.nouv",
↪  "evit.danger","dep.recompense")
table1 <- CreateTableOne(vars = vars, data = smp.d, factorVars = catVars, strata =
↪  "depression")
print(table1, showAllLevels = TRUE, formatOptions = list(digits = 2))
```

```
                        Stratified by depression
                         level         0             1             p      test
  n                                      486           313
  age (mean (SD))                      39.93 (13.69) 37.41 (12.42)  0.009
  profession (%)         agriculteur       3 ( 0.6)      2 ( 0.6)   0.333
                         commerçant       65 (13.5)     26 ( 8.4)
                         cadre            16 ( 3.3)      8 ( 2.6)
                         intermédiaire    31 ( 6.4)     26 ( 8.4)
                         employé          81 (16.8)     55 (17.7)
                         ouvrier         132 (27.4)     96 (30.9)
                         autre            22 ( 4.6)      9 ( 2.9)
                         sans.emploi     132 (27.4)     89 (28.6)
  nb.enfants (mean (SD))                1.57 (1.92)   1.58 (1.73)   0.936
  gravite (%)            1               100 (20.7)      6 ( 1.9)  <0.001
                         2               111 (23.0)     18 ( 5.8)
                         3                82 (17.0)     33 (10.5)
                         4                89 (18.5)     74 (23.6)
                         5                68 (14.1)    114 (36.4)
                         6                26 ( 5.4)     55 (17.6)
                         7                 6 ( 1.2)     13 ( 4.2)
  recherche.nouv (%)     1               168 (38.5)     81 (31.3)   0.005
                         2               107 (24.5)     50 (19.3)
                         3               161 (36.9)    128 (49.4)
  evit.danger (%)        1               229 (52.8)     86 (33.5)  <0.001
                         2               109 (25.1)     45 (17.5)
                         3                96 (22.1)    126 (49.0)
  dep.recompense (%)     1               121 (28.1)     71 (28.0)  <0.001
                         2               148 (34.3)     49 (19.3)
                         3               162 (37.6)    134 (52.8)
```

avec `gtsummary` :

- (attention il a tendance à faire des tests de Wilcoxon par défaut pour les variables continues, il faut lui dire de faire des t-tests si on veut ça)
- Pour les variables catégorielles, il fait par défaut des tests du Chi2 (ou Fisher si effectifs petits) donc autant ne pas lui donner d'instructions

```r
# utiliser smp.d.bis avec la variable depression en facteur recodé en "Dépressif"
↪  / "Non dépressif"
smp.d.bis <- smp.d
smp.d.bis$depression <- factor(smp.d.bis$depression, levels=c(0,1), labels=c("Non
↪  dépressif","Dépressif"))

tableau <- smp.d.bis %>%
    tbl_summary(
        by = depression,
        statistic = list(
            all_continuous() ~ "{mean} ({sd})",
            # all_continuous() ~ "{median} [{p25}, {p75}]",
            all_categorical() ~ "{n} / {N} ({p}%)"
            ),
        digits = all_continuous() ~ 2,
        missing = "no"
        ) %>%
        modify_header(label = "**Caractéristiques**") %>%
        bold_labels() %>%
        add_overall() %>%
        add_p(
            # test = list( (rajouter si tests spécifiques pr les 2)
            all_continuous()  ~ "t.test" # ou "wilcox.test" pour test non
             ↪  paramétrique
            # all_categorical() ~ "chisq.test" # ou "fisher.test" si effectifs
             ↪  petits
            )
# ajouter les p-values pour les comparaisons entre groupes
```

Juste un peu chiant pour avoir un bel affichage en pdf mais franchement…

```r
tableau %>%
    # Conversion en objet kable (LaTeX standard)
    as_kable_extra(booktabs = TRUE, longtable = TRUE) %>%
    kableExtra::column_spec(1, width = "6cm") %>%
    # (Optionnel) Ajuste la taille de la police si le tableau est encore trop
     ↪  large
    kableExtra::kable_styling(latex_options = c("repeat_header"), font_size = 9)
```

Caractéristiques	**Overall** N = 799	**Non dépressif** N = 486	**Dépressif** N = 313	**p-value**
age	38.94 (13.26)	39.93 (13.69)	37.41 (12.42)	0.007
profession				
agriculteur	5 / 793 (0.6%)	3 / 482 (0.6%)	2 / 311 (0.6%)	
commerçant	91 / 793 (11%)	65 / 482 (13%)	26 / 311 (8.4%)	
cadre	24 / 793 (3.0%)	16 / 482 (3.3%)	8 / 311 (2.6%)	
intermédiaire	57 / 793 (7.2%)	31 / 482 (6.4%)	26 / 311 (8.4%)	

(continued)

Caractéristiques	Overall N = 799	Non dépressif N = 486	Dépressif N = 313	p-value
employé	136 / 793 (17%)	81 / 482 (17%)	55 / 311 (18%)	
ouvrier	228 / 793 (29%)	132 / 482 (27%)	96 / 311 (31%)	
autre	31 / 793 (3.9%)	22 / 482 (4.6%)	9 / 311 (2.9%)	
sans.emploi	221 / 793 (28%)	132 / 482 (27%)	89 / 311 (29%)	
nb.enfants	1.57 (1.85)	1.57 (1.92)	1.58 (1.73)	>0.9
schizophrenie	64 / 799 (8.0%)	29 / 486 (6.0%)	35 / 313 (11%)	0.008
gravite				<0.001
1	106 / 795 (13%)	100 / 482 (21%)	6 / 313 (1.9%)	
2	129 / 795 (16%)	111 / 482 (23%)	18 / 313 (5.8%)	
3	115 / 795 (14%)	82 / 482 (17%)	33 / 313 (11%)	
4	163 / 795 (21%)	89 / 482 (18%)	74 / 313 (24%)	
5	182 / 795 (23%)	68 / 482 (14%)	114 / 313 (36%)	
6	81 / 795 (10%)	26 / 482 (5.4%)	55 / 313 (18%)	
7	19 / 795 (2.4%)	6 / 482 (1.2%)	13 / 313 (4.2%)	
recherche.nouv				0.005
1	249 / 695 (36%)	168 / 436 (39%)	81 / 259 (31%)	
2	157 / 695 (23%)	107 / 436 (25%)	50 / 259 (19%)	
3	289 / 695 (42%)	161 / 436 (37%)	128 / 259 (49%)	
evit.danger				<0.001
1	315 / 691 (46%)	229 / 434 (53%)	86 / 257 (33%)	
2	154 / 691 (22%)	109 / 434 (25%)	45 / 257 (18%)	
3	222 / 691 (32%)	96 / 434 (22%)	126 / 257 (49%)	
dep.recompense				<0.001
1	192 / 685 (28%)	121 / 431 (28%)	71 / 254 (28%)	
2	197 / 685 (29%)	148 / 431 (34%)	49 / 254 (19%)	
3	296 / 685 (43%)	162 / 431 (38%)	134 / 254 (53%)	

[1] Mean (SD); n / N (%)
[2] Welch Two Sample t-test; NA; Pearson's Chi-squared test

3 Dépendance, liaison et association

Deux variables sont dépendantes si une valeur donne une information sur l'autre.

Par exemple, le poids et la taille sont dépendantes : connaître la taille d'une personne permet d'avoir une idée de son poids.

3.A Variables quantitatives

3 types de liaisons entre variables quantitatives :

- Dépendance : connaître X permet de mieux estimer Y (par exemple tabac et maladie respiratoire, taille et poids, etc.)
- Dépendance monotone : X et Y varient dans le même sens (par exemple âge et pression artérielle)
  - Dépendance linéaire : relation linéaire entre X et Y (par exemple taille et poids chez les adultes)

  - Corrélation
  - Variance partagée = proportion de la variance de Y expliquée par X dans une relation linéaire entre les deux variables
- Concordance : si X est plus grand pour un individu que pour un autre, alors Y est aussi plus grand pour le premier individu (par exemple taille et poids)
  - pour une variable quantitative : coefficient de corrélation intraclasse (ICC)

3.A.1 Dépendance

Il n'existe pas de paramètre estimant parfaitement la dépendance ou l'indépendance entre deux variables quantitatives.

L'idéal serait d'avoir un paramètre $\delta(X, Y)$ valant 0 quand X et Y sont indépendantes et 1 quand elles sont parfaitement dépendantes.

3.A.2 Dépendance monotone ou linéaire

Le coefficient de corrélation de Pearson r mesure la dépendance linéaire entre deux variables quantitatives X et Y.

Il est désigné par les lettres r ou ρ (rho), en référence à Karl Pearson qui l'a introduit en 1895.

Il varie entre -1 (les deux variables X et Y sont parfaitement linéairement dépendantes de façon négative) et +1 (les deux variables X et Y sont parfaitement linéairement dépendantes de façon positive).

Corrélaton nulle ($r = 0$) signifie que les deux variables sont linéairement indépendantes.

NB : le coefficient ne matérialise pas la **force** de la dépendance, mais seulement son **type** !

Pour matérialiser la force de la relation, on utilise le **coefficient de détermination** r^2.

Le paramètre r^2 représente approximativement la **proportion de la variance de** Y expliquée par la variance de X dans une relation linéaire entre les deux variables.

C'est à dire :

- X est le nombre d'heures de révision
- Y est la note obtenue à un examen

On trouve un coefficent de corrélation $r = 0.8$ entre X et Y, alors $r^2 = 0.64$ soit 64%.

Donc : 64% de la variabilité (de la variance) des notes Y s'expliquer par le le modèle linéaire basé sur le nombre d'heures de révision X.

Ce n'est pas la même chose que "r^2% des notes s'expliquent par le nombre d'heures de révision".

> **!** Important
>
> **Coefficient de corrélation** r :
>
> - mesure la **direction** de la relation linéaire entre deux variables quantitatives.

- varie entre -1 et +1 (avec 0 = pas de relation linéaire).
- en pratique : mesure un peu la force quand même… mais il n'y a pas vraiment d'unité pour l'exprimer

---

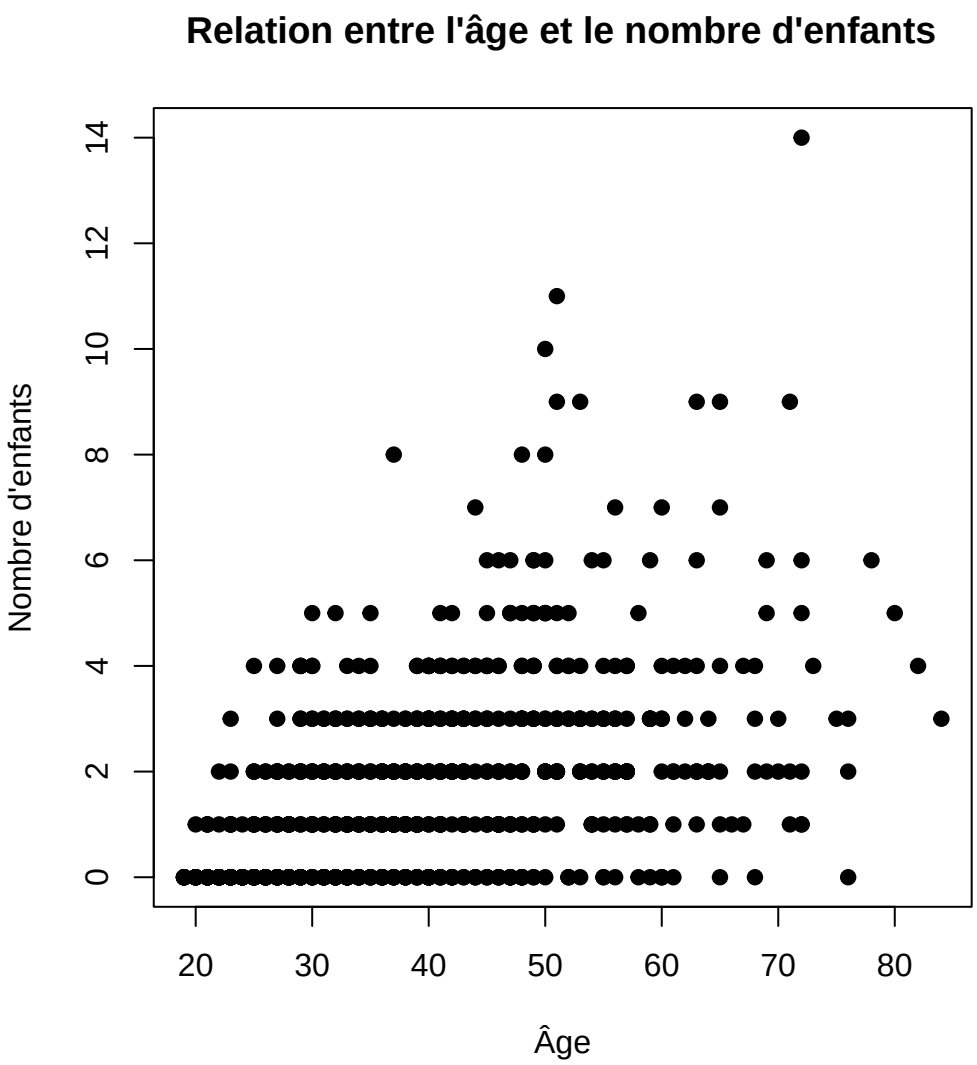
Coefficient de détermination r^2 :*

- mesure la **force** de la relation linéaire entre deux variables quantitatives avec une unité (en % de variance expliquée)
- varie entre 0 et 1 (avec 0 = pas de relation linéaire).
- représente la proportion de la variance de Y expliquée par la variance de X (et non pas le pourcentage de valeurs de Y expliquées par X).

3.A.2.1 Exemple sur R

3.A.2.1.1 SMP Représentation graphique de la relation entre l'âge et le nombre d'enfants dans le jeu de données `smp`.

```
plot(smp$age, smp$nb.enfants,
        xlab="Âge",
        ylab="Nombre d'enfants",
        main="Relation entre l'âge et le nombre d'enfants",
        pch=19) # pch = sert à choisir le type de point
```



Calcul du coefficient de corrélation de Pearson entre l'âge et le nombre d'enfants.

```
cor(
    smp$age,
    smp$nb.enfants,
    use="complete.obs") # ignore les valeurs manquantes
```

```
[1] 0.498
```

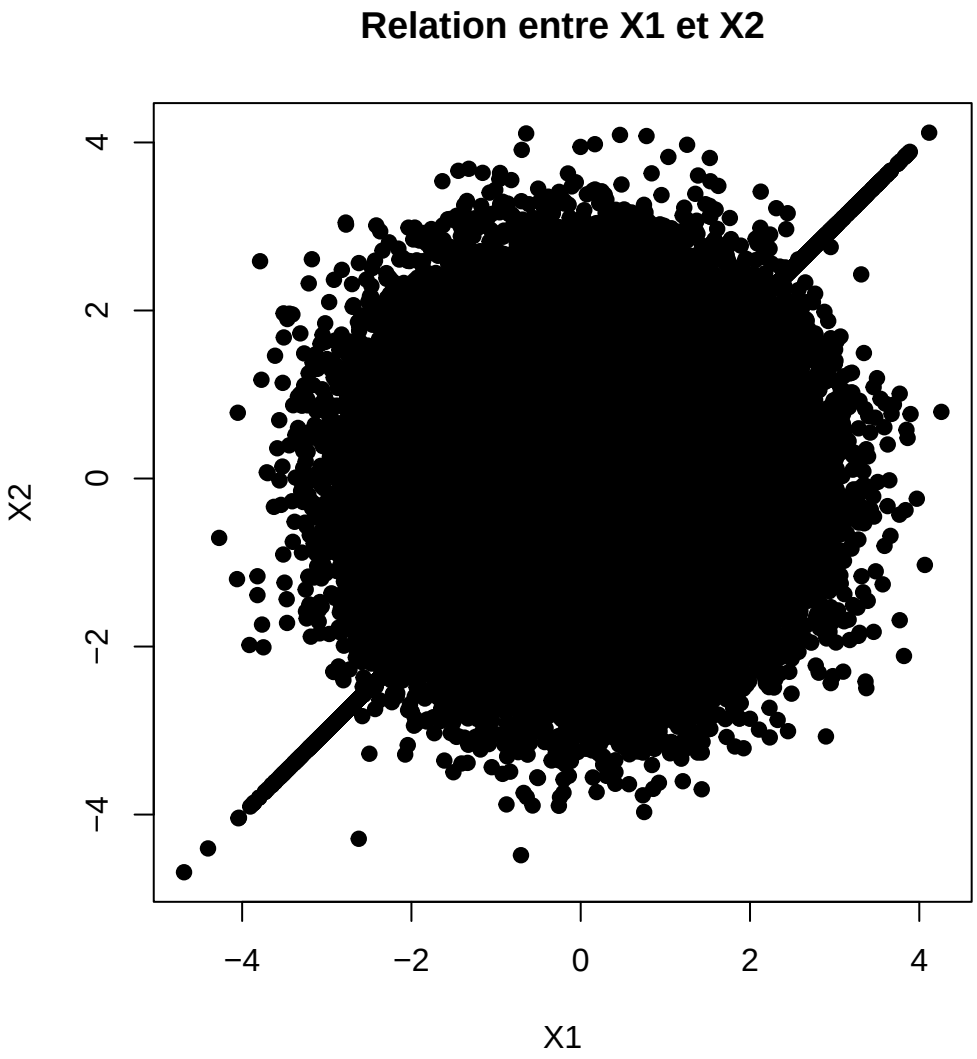
Le coefficient de corrélation est de 0.498, ce qui indique une dépendance linéaire positive entre l'âge et le nombre d'enfants.

```
set.seed(20230430)
x <- rnorm(100000) # génère 100000 valeurs aléatoires suivant une loi normale
y <- rnorm(100000)
```

```
z <- rnorm(100000)
X1 <- c(x,y) # concatène les deux vecteurs x et y
X2 <- c(x,z)
```

3.A.2.2 Données simulées Représentation graphique de la relation entre X1 et X2.

```
plot(X1, X2,
        xlab="X1",
        ylab="X2",
        main="Relation entre X1 et X2",
        pch=19)
```



Coorélation entre X1 et X2.

```
round(cor(X1, X2),3)
```

```
[1] 0.498
```

Variance de X1 :

```
round(var(X1),3)
```

```
[1] 1
```

logique ce soit = 1 car X1 est la concaténation de deux variables indépendantes de variance 1 (car générées par `rnorm` donc suivent une loi normale standard).

Pour calculer la variance partagée entre X1 et X2, on utilise la formule de la variance de la somme de deux variables aléatoires indépendantes :

$$Var(X1 + X2) = Var(X1) + Var(X2) + 2Cov(X1, X2)$$

Sur R :

```
# corrélation de Pearson mise au carrée donne la part de variance partagée
rho2  <- (cor(X1, X2)^2)
rho2
```

```
[1] 0.248
```

Une autre manière d'obtenir ça :

1. Construire un modèle linéaire de Y en fonction de X
2. Extraire la part de variance résiduelle (non expliquée par X) du modèle
3. Variance expliquée par X = **1 - variance résiduelle**

(mais `summary(lm())` donne directement le R² dans la partie "Multiple R-squared")

```
res <- lm(X1 ~ X2)
# summary donne l'info dans "Multiple R-squared"
summary(res)
```

```
Call:
lm(formula = X1 ~ X2)

Residuals:
   Min     1Q Median     3Q    Max 
-5.079 -0.487 -0.001  0.486  4.867 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.00361    0.00194    1.86    0.063 .  
X2           0.49770    0.00194  256.59   <2e-16 ***
---
```

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

Residual standard error: 0.868 on 199998 degrees of freedom
Multiple R-squared:  0.248, Adjusted R-squared:  0.248
F-statistic: 6.58e+04 on 1 and 2e+05 DF,  p-value: <2e-16
```

```r
# variance résiduelle
round(var(residuals(res)),3)
```

```
[1] 0.754
```

```r
# variance expliquée par X2
1 - round(var(residuals(res)),3)
```

```
[1] 0.246
```

Important

NB : la variance partagée n'est pas la même chose que la covariance !!
Covariance = mesure comment deux variables varient ensemble, positive si les deux variables augmentent ensemble, négative si l'une augmente quand l'autre diminue.

Paramètre	Symbole	Interprétation	Formule / calcul R
Coefficient de corrélation	r ou ρ	Direction et force de la relation linéaire entre deux variables quantitatives	`cor(X, Y)`
Variance partagée	r^2	Proportion de la variance de Y expliquée par X (force de la relation linéaire)	`rho2 <- cor(X, Y)^2`
Covariance	$\text{Cov}(X, Y)$	Mesure comment deux variables varient ensemble (positive : ensemble, négative : sens inverse)	`cov(X, Y)`

3.A.2.2.1 Matrice de corrélation Pour calculer la matrice de corrélation entre plusieurs variables quantitatives, on peut utiliser la fonction `cor()` en lui passant un data frame ou une matrice.

```r
quanti <- c("age","nb.enfants","depression","schizophrenie",
 ↪  "gravite","recherche.nouv","evit.danger","dep.recompense")
round(cor(smp.d[,quanti],use="pairwise.complete.obs"),digits=3)
```

```
age nb.enfants depression schizophrenie gravite
```

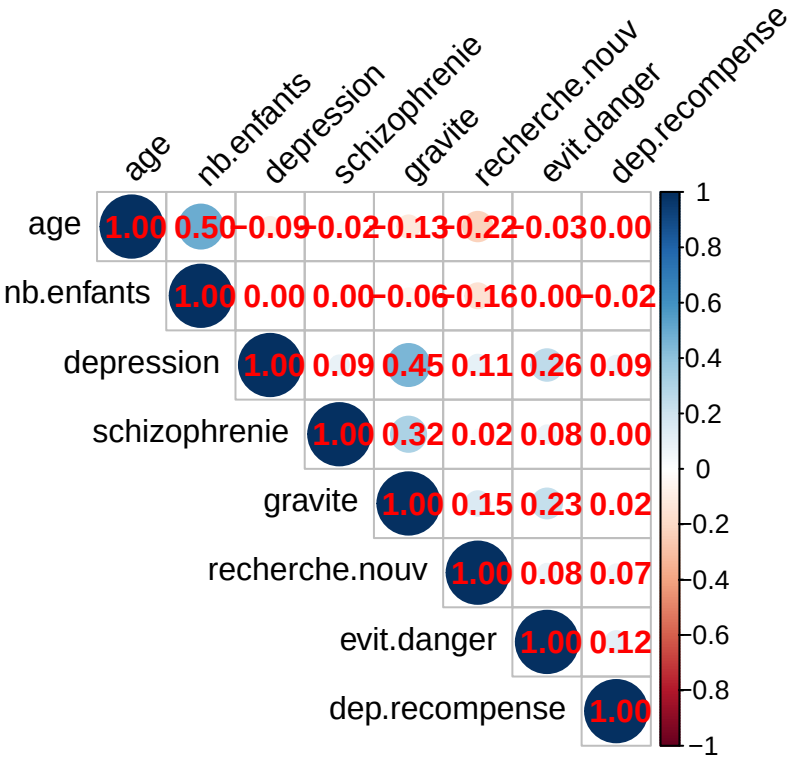
```
age             1.000      0.498     -0.093        -0.021  -0.127
nb.enfants      0.498      1.000      0.003        -0.003  -0.057
depression     -0.093      0.003      1.000         0.094   0.454
schizophrenie  -0.021     -0.003      0.094         1.000   0.318
gravite        -0.127     -0.057      0.454         0.318   1.000
recherche.nouv -0.223     -0.159      0.109         0.022   0.154
evit.danger    -0.027      0.004      0.256         0.081   0.230
dep.recompense -0.001     -0.023      0.089        -0.004   0.019
               recherche.nouv evit.danger dep.recompense
age                    -0.223      -0.027         -0.001
nb.enfants             -0.159       0.004         -0.023
depression              0.109       0.256          0.089
schizophrenie           0.022       0.081         -0.004
gravite                 0.154       0.230          0.019
recherche.nouv          1.000       0.081          0.071
evit.danger             0.081       1.000          0.119
dep.recompense          0.071       0.119          1.000
```

On peut représenter ça graphiquement avec la fonction `corrplot()` du package `corrplot`.

```r
library(corrplot)
```

```
corrplot 0.95 loaded
```

```r
corrplot(
    round(cor(smp.d[,quanti],use="pairwise.complete.obs"),digits=3),
    method="circle",
    addCoef.col = "red",
    type="upper",
    tl.col="black",
    tl.srt=45)
```



Une matrice de corrélation est symétrique (mêmes valeurs de part et d'autre de la diagonale), car la corrélation entre X et Y est la même que celle entre Y et X.

> **Note**
>
> **A quoi ça peut servir dans une étude rétropective ?**
> Dans une étude qui compare 2 techniques chirurgicales (A vs B), la matrice de corrélation sert surtout à explorer et comprendre les relations entre les nombreuses variables mesurées autour de l'intervention.
>
> - Explorer les facteurs pré-opératoires entre eux (âge corrélé au score ASA, etc.)
> - Repérer la colinéarité entre co-variables avant une régression ajustée
> - Relier facteurs pré-op et outcomes post-op (âge, IMC, ASA vs durée d'hospitalisation, perte sanguine, etc.)
> - Notamment selon le type de chirurgie (groupe A vs groupe B) : est ce que les plus vieux ont plus de perte sanguine avec la technique A que B ?

3.A.3 Concordance

3.A.3.1 Principle Par exemple, "concordance" entre deux échographies identiques fait par deux médecins différents.

Variables quantitiatives : concordance mesurée par le coefficient de corrélation intraclasse (ICC).

- ICC varie entre 0 (pas de concordance) et 1 (concordance parfaite).
- Dans l'exemple du score échographique, le coefficient de corrélation intraclasse =

$$\frac{\text{variance inter-patients}}{\text{variance inter-patients} + \text{variance inter-radiologues} + \text{variance résiduelle}}$$

Donc en gros : ICC = "**vrai signal**" / ("**vrai signal**" + "**bruit**")

Vaut 1 quand il n'y a aucun bruit (variance inter-radiologues et résiduelle = 0).

Vaut 0 quand il n'y a que de bruit, donc les mesures sont totalement indépendantes entre elles.

3.A.3.2 Exemple sur R Dans l'étude `smp`, 2 cliniciens sont présentés lors des entretiens : un junior et un senior.

A la fin de chaque entretien, ils remplissaient plusieurs questionnaires, dont un comportait l'échelle *CGI = Clinical Global Impression* (échelle de 1 à 7, 1 = pas malade, 7 = très malade).

Il est possible de quantifier le niveau de concordance entre les notes CGI données par le junior et le senior à l'aide du coefficient de corrélation intraclasse (ICC).

Dans ce cas : il s'agit d'un ICC de type "2-way random effects, absolute agreement, single rater/measurement" (notation ICC(2,1) de Shrout & Fleiss, 1979).

- 2-way random effects : les deux évaluateurs (junior et senior) sont considérés comme des échantillons aléatoires d'une population plus large d'évaluateurs possibles.
- Absolute agreement : on s'intéresse à l'accord absolu entre les évaluateurs, pas seulement à la corrélation.
- Single rater/measurement : on considère les notes individuelles de chaque évaluateur, pas une moyenne.

C'est le type d'ICC le plus couramment utilisé en pratique clinique.

```r
psy::icc(
    smp.aij[,c("gravite.jun","gravite.sen")]
)
```

```
$nb.subjects
[1] 796

$nb.raters
[1] 2

$subject.variance
[1] 2.4

$rater.variance
[1] -0.000228

$residual
```

```
[1] 0.257

$icc.consistency
[1] 0.903

$icc.agreement
[1] 0.903
```

- `$subject.variance` : variance sujets = variance signal
- `$rater.variance` : variance évaluateurs = variance bruit due aux différences entre évaluateurs
- `$residual` : variance résiduelle = variance bruit due aux autres sources d'erreur
- `$icc.agreement` : coefficient de corrélation intraclasse ICC vaut 0,9

Mais la librarie `irr` propose aussi une fonction `icc()` pour calculer le coefficient de corrélation intraclasse, il faut juste la paramétrer un peu plus mais l'output est plus clair.

```r
library(irr)
```

```
Loading required package: lpSolve


Attaching package: 'irr'

The following object is masked from 'package:psy':

    icc
```

```r
irr::icc(
    smp.aij[,c("gravite.jun","gravite.sen")],
    model="twoway", # sinon oneway si un seul évaluateur par sujet
    type="agreement", # sinon consistency = uniformité des réponses
    unit="single" # single ou average
)
```

```
 Single Score Intraclass Correlation

   Model: twoway
   Type : agreement

   Subjects = 796
     Raters = 2
   ICC(A,1) = 0.903

 F-Test, H0: r0 = 0 ; H1: r0 > 0
 F(795,795) = 19.7 , p = 5.62e-295
```

```
95%-Confidence Interval for ICC Population Values:
 0.89 < ICC < 0.915
```

3.A.4 Résumé des paramètres de dépendance entre deux variables quantitatives

Résumé des principaux paramètres de dépendance entre deux variables quantitatives :

- Dépendance
- Dépendance monotone ou linéaire
  - Coefficient de corrélation de Pearson
  - Coefficient de détermination = variance partagée (≠ covariance)
- Concordance = coefficient de corrélation intraclasse (ICC)

Paramètre	Symbole	Interprétation	Formule / calcul R
Coefficient de corrélation	r ou ρ	Direction et force de la relation linéaire entre deux variables quantitatives	`cor(X, Y)`
Variance partagée	r^2	Proportion de la variance de Y expliquée par X (force de la relation linéaire)	`rho2 <- cor(X, Y)^2`
Covariance	Cov(X, Y)	Mesure comment deux variables varient ensemble (positive : ensemble, négative : sens inverse)	`cov(X, Y)`
Coefficient de corrélation intraclasse	ICC	Mesure la concordance entre plusieurs mesures quantitatives	`psy::icc(dataframe)` ou `irr::icc(dataframe, model=..., type=..., unit=...)`

3.B Variables catégorielles

3.B.1 Dépendance

3.B.1.1 Chi2 et associés

- Existe-t-il une relation entre les deux variables catégorielles ?
- Si oui, quelle est la force de cette relation ?

Pour ces questions : on utilise

- **Le test du Chi2 d'indépendance** sert à évaluer l'existence d'une relation entre deux variables catégorielles.

Et des transformations normalisées du Chi2 pour évaluer la force de cette relation :

(pour normaliser le Chi2 : racine carré de chi2/nombre d'observations)

- **Le V de Cramer** pour la force de l'association entre deux variables catégorielles **non ordonnées** (type de chirurgie) (= on pourrait utiliser le coefficient de Pearson pour des variables binaires).
  - Varie entre 0 (pas d'association) et 1 (association parfaite)
- Le coefficient de Pearson : plutôt pour variables quantitatives, mais utilisable pour des variables binaires (0/1)
  - Varie entre -1 et +1
- **Le coefficient de Spearman** : pour variables **ordinales** ou quantitatives non linéaires (rangées)
  - Coefficient de Spearman = corrélation de Pearson calculée sur les rangs des données
  - Varie entre -1 et +1

> **Note**
>
> **ANOVA ou V de Cramer pour variables catégorielles non ordonnées ?**
>
> - **ANOVA** :
> - 1 variable quantitative, 1 ou plusieurs variables catégorielles non ordonnées
> - compare les moyennes de la variable quantitative entre les différentes modalités de la variable catégorielle
> - test statistique (p-value)
> - outcome quantitatif
> - **V de Cramer** :
> - Variables catégorielles non ordonnées
> - mesure la force de l'association entre les variables
> - valeur entre 0 et 1
> - outcome catégoriel

3.B.1.2 Odds-ratio et risque relatif Pour des X et Y binaires (0/1) :

Exemple : association entre décès en USI et existence d'une infection ou non

	Décès (Y=1)	Pas de décès (Y=0)
Infection (X=1)	a	b
Pas d'infection(X=0)	c	d

- **Risque relatif (RR)** = risque de décès chez les patients infectés / risque de décès chez les patients non infectés = [a/(a+b)] / [c/(c+d)]
  - On a RR fois plus de risque de décès si on est infecté
  - $RR = \frac{\text{\% de deces chez les infectes}}{\text{\% de deces chez les non infectes}} = \frac{a}{a+b} / \frac{c}{c+d}$
- **Odds-ratio (OR)** = rapports des côtes = interprétation plus subtile
  - Odds de décès chez les patients infectés = a/b (côte)
  - Odds de décès chez les patients non infectés = c/d
  - $OR = \frac{\frac{\text{morts infectes}}{\text{vivants infectes}}}{\frac{\text{morts non infectes}}{\text{vivants non infectes}}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$
  - Il y a OR fois plus de "morts par rapport aux vivants" si on est infecté que de "morts par rapport aux vivants" si on n'est pas infecté.

RR et OR sont positifs et varient de 0 à l'infini.

S'ils valent 1 : pas d'association entre X et Y, les variables sont indépendantes.

S'ils valent 0 ou sont très grands : forte association entre X et Y.

Rapport entre RR et OR :

- Si l'événement étudié est rare (<10%) : RR ≈ OR
- Si l'événement est fréquent (>10%) : **OR surestime le RR** (OR > RR si RR > 1 ; OR < RR si RR < 1)

3.B.1.3 Exemple sur R Sur fichier `smp.d` : force de l'association entre la variable "dépression" (`smp.d$depression`) et le FDR "le prisonnier a un niveau élevé d'évitement du danger" (`smp.d$evit.danger`).

La variable dépression est binaire (0 = non dépressif, 1 = dépressif).

La variable evit.danger n'est pas binaire, mais codée 1, 2 ou 3 pour "faible", "moyen" ou "élevé".

Il faut donc la recoder en binaire (0 = faible ou moyen, 1 = élevé).

```
smp.d$evit.danger.b <- smp.d$evit.danger > 2
smp.d$depression.b <- smp.d$depression == 1
tb <- table(
        smp.d$depression.b,
        smp.d$evit.danger.b,
```

```r
        deparse.level=2
        )
tb
```

```
                  smp.d$evit.danger.b
smp.d$depression.b FALSE TRUE
             FALSE   338   96
             TRUE    131  126
```

Pour obtenir facilement le RR et l'OR, on peut utiliser la fonction `epi.2by2()` du package `epiR`.

```r
epi.2by2(
    tb,
    method="cohort.count", # sinon case.control
    conf.level=0.95
    )
```

```
             Outcome+    Outcome-      Total                 Inc risk *
Exposure+         338          96        434     77.88 (73.68 to 81.70)
Exposure-         131         126        257     50.97 (44.69 to 57.24)
Total             469         222        691     67.87 (64.25 to 71.34)

Point estimates and 95% CIs:
---------------------------------------------------------------------
Inc risk ratio                                   1.53 (1.34, 1.74)
Inc odds ratio                                   3.39 (2.43, 4.73)
Attrib risk in the exposed *                     26.91 (19.65, 34.16)
Attrib fraction in the exposed (%)              34.55 (25.88, 42.85)
Attrib risk in the population *                  16.90 (9.87, 23.93)
Attrib fraction in the population (%)           24.90 (19.77, 30.45)
---------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 53.594 Pr>chi2 = <0.001
Fisher exact test that OR = 1: Pr>chi2 = <0.001
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

3.B.2 Dépendance monotone

Une dépendance monotone ne peut s'envisager qu'entre des variables ordinales (rangées) ou entre une variable ordinale et une variable quantitative.

- **Coefficient de Spearman** : corrélation de Pearson calculée sur les rangs des données.
  - Varie entre -1 et +1
  - Utilisable pour des variables ordinales (rangées) ou quantitatives non linéaires

Problème : donner du sens à une corrélation basée sur des rangs !! dépend ++ du codage

3.B.2.1 Exemple sur R En pratique : dans l'étude santé mentale en prison, les deux variables de tempérament : « recherche de nouveauté » et « évitement du danger » sont codées en 1, 2 et 3 pour, respectivement, « bas », « moyen » et « élevé ». Si l'on souhaite apprécier dans quelle mesure un niveau élevé de recherche de nouveauté est associé à un niveau bas d'évitement du danger, il est possible d'estimer un coefficient de corrélation de Spearman ou de Pearson :

Dans l'étude `smp` : variable "recherche de nouveauté" (`smp.d$recherche.nouv`) et variable "évitement du danger" (`smp.d$evit.danger`).

- les deux variables `recherche.nouv` et `evit.danger` sont codées 1, 2 ou 3 pour "faible", "moyen" ou "élevé".
- objectif : apprécier dans quelle mesure un niveau élevé de recherche de nouveauté est associé à un niveau bas d'évitement du danger.

```
table(smp$recherche.nouv)
```

```
  1   2   3 
249 157 289 
```

```
table(smp$evit.danger)
```

```
  1   2   3 
315 154 222 
```

Calcul du coefficient de corrélation de Spearman entre les deux variables.

```
cor(
    smp.d$recherche.nouv,
    smp.d$evit.danger,
    method="spearman",
    use="complete.obs") # ignore les valeurs manquantes
```

```
[1] 0.0785
```

Le coefficient de corrélation de Spearman est de 0.078, ce qui indique une très faible dépendance monotone positive entre la recherche de nouveauté et l'évitement du danger.

NB : si on avait utilisé le coefficient de Pearson :

```
cor(
    smp.d$recherche.nouv,
    smp.d$evit.danger,
    method="pearson",
    use="complete.obs") # ignore les valeurs manquantes
```

```
[1] 0.0807
```

Le coefficient de corrélation de Pearson est de 0.081 : les deux coefficients sont très proches (c'est assez fréquent quand les variables sont ordinales avec peu de modalités).

3.B.3 Concordance

3.B.3.1 Coefficient kappa de Cohen Pour les variables catégorielles, on pourrait se dire que mesurer à quel point 2 variables s'accordent reviendraient à compter la proportion de fois ou elles ont la même valeur.

Problème : cette proportion de concordance peut être due au hasard !!

On corrige ça avec le **kappa de Cohen**.

$$\text{kappa} = \frac{\text{concordance observee - concordance due au hasard}}{1 - \text{concordance due au hasard}}$$

- kappa $= 0$: concordance observée = concordance due au hasard
- kappa $= 1$: concordance parfaite

3.B.3.2 Sensibilité, spécificité, VPP, VPN Le kappa de Cohen mesure une **concordance globale et symétrique** entre deux variables catégorielles.

Mais parfois, on s'intéresse à la capacité d'une variable à s'approcher d'ue variable de référence

= mesurer à quel point Y prédit correctement X.

Dans ce cas, il vaut mieux utiliser des paramètres asymétriques :

- Sensibilité = proportion de vrais positifs parmi les positifs réels = $P(Y = 1|X = 1)$
- Spécificité = proportion de vrais négatifs parmi les négatifs réels = $P(Y = 0|X = 0)$

Le problème avec la sensibilité et la spécificité : elles ne tiennent pas compte de la prévalence de la condition réelle X (c'est à dire la proportion de $X = 1$ dans la population).

- Valeur prédictive positive (VPP) = proportion de vrais positifs parmi les positifs prédits = $P(X = 1|Y = 1)$
- Valeur prédictive négative (VPN) = proportion de vrais négatifs parmi les négatifs prédits = $P(X = 0|Y = 0)$

Dans un tableau de contingence :

	Y=1 (test positif)	Y=0 (test négatif)
X=1 (condition réelle présente)	a (vrais positifs)	b (faux négatifs)
X=0 (condition réelle absente)	c (faux positifs)	d (vrais négatifs)

- Sensibilité = a / (a + b)
- Spécificité = d / (c + d)

- Valeur prédictive positive (VPP) = a / (a + c)
- Valeur prédictive négative (VPN) = d / (b + d)

On peut ainsi représenter une courbe ROC (Receiver Operating Characteristic) qui trace la sensibilité en fonction de 1 - spécificité pour différents seuils de décision.

3.B.3.2.1 Exemple 1 sur R

- Les deux cliniciens (junior et senior) posent un diagnostic de schizophrénie (1 = oui, 0 = non) pour chaque patient.
- Niveau d'accord inter-juges pour une variable catégorielle : kappa de Cohen.

```
psy::ckappa(
    smp.aij[,c("scz.jun","scz.sen")]
)
```

```
$table
    0  1
0 715 11
1  30 43

$kappa
[1] 0.65
```

Autre méthode avec le package `irr` :

```
irr::kappa2(
    smp.aij[,c("scz.jun","scz.sen")],
    weight="unweighted" # ou "equal" ou "squared" pour kappa pondéré
)
```

```
 Cohen's Kappa for 2 Raters (Weights: unweighted)

 Subjects = 799
   Raters = 2
    Kappa = 0.65

        z = 18.6
  p-value = 0
```

Le clinicien junior a posé le diagnostic de schizophrénie chez 30 + 43 = 73 détenus alors que le clinicien senior l'a fait pour seulement 11 + 43 = 54.

Au total, le coefficient kappa vaut 0,65.

3.B.3.2.2 Exemple 2 sur R

- Étude sur 244 patients déprimés hospitalisés : tâche de lecture de texte puis comptage de 1 à 10, enregistrement voix.
- Extraction de la fréquence fondamentale (hauteur de voix), connue pour être ~75–140 Hz chez les hommes et ~170–250 Hz chez les femmes.
- Objectif : voir dans quelle mesure la hauteur de voix prédit le sexe en testant un seuil de 155 Hz.
- Méthode : **calcul sensibilité et spécificité du seuil 155 Hz** pour discriminer hommes et femmes.

```
sexe.f <- vox$sexe == 2
voix.aigue <- vox$moyf0>155
# moyenne des femmes avec voix aiguë = sensibilité
# = proportion de femmes avec test positif
mean(voix.aigue[sexe.f], na.rm=TRUE)
```

```
[1] 0.917
```

```
# moyenne des hommes sans voix aiguë = spécificité
# = proportion d'hommes avec test négatif
mean(!voix.aigue[!sexe.f], na.rm=TRUE)
```

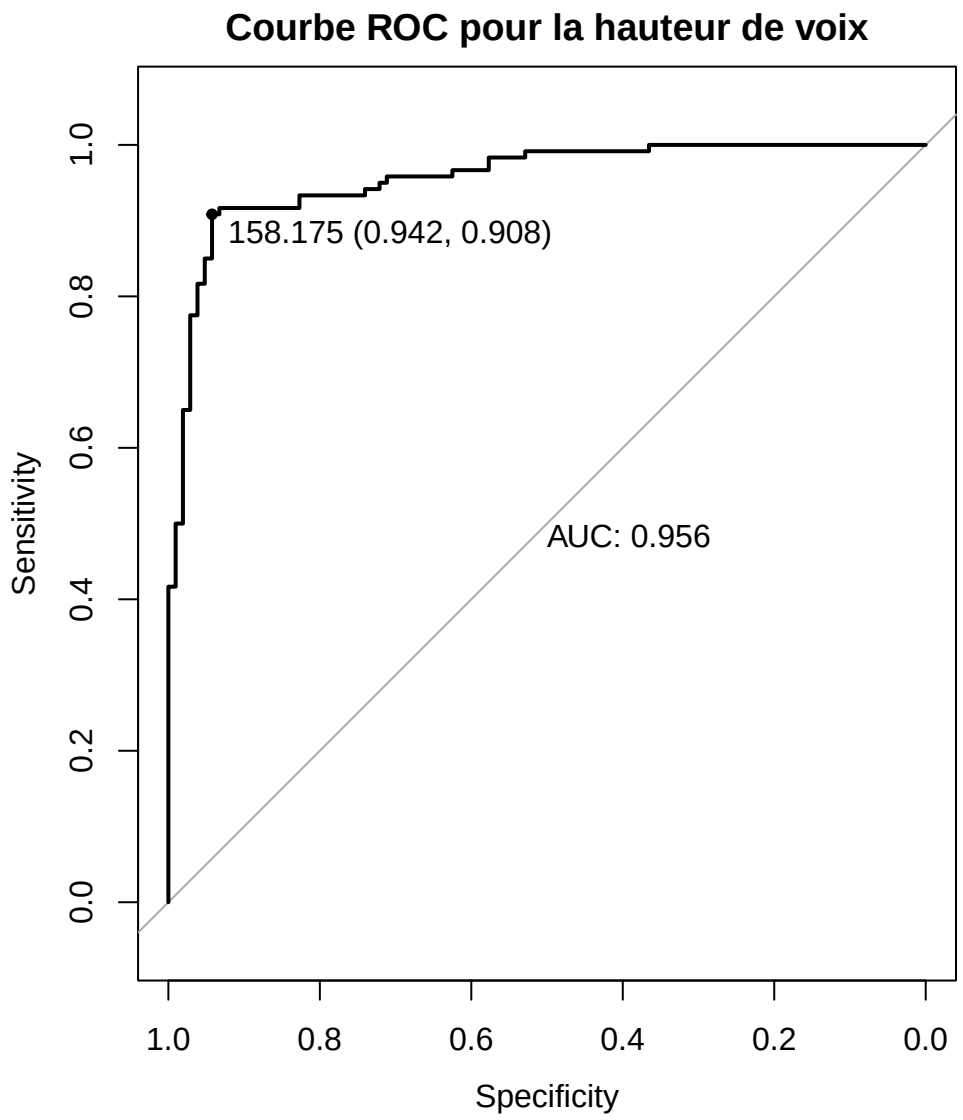
```
[1] 0.933
```

Vérifier le seuil de 155 Hz avec une courbe ROC :

```
rocf0 <- roc(sexe.f~vox$moyf0)
```

```
Setting levels: control = FALSE, case = TRUE

Setting direction: controls < cases
```

```
plot(rocf0,
    main="Courbe ROC pour la hauteur de voix",
    print.thres="best",
    print.thres.best.method="youden",
    print.auc=TRUE)
```



Seuil optimal calculé par indice de Youden (maximise la somme de la sensibilité et de la spécificité) = 158,17 Hz

AUC : calcule la qualité globale du test

- Aire comprise entre 0 et 1
- = probabilité que la hauteur de voix d'une femme soit plus élevée que celle d'un homme pris au hasard

3Tests statistiques

1 Introduction : en pratique, comment utiliser les tests statistiques ?

1.A À quoi servent vraiment les tests statistiques ?

Les tests statistiques ont deux fonctions simultanées :

- **Outils d'inférence**

Ils permettent de passer de ce qui est observé sur un échantillon à une affirmation sur une population plus large (effet d'un traitement, différence entre groupes, association entre variables).

- **Langage commun de la communauté scientifique**

Ils fournissent un cadre standardisé (H0/H1, p-value, intervalles de confiance, seuils "significatifs") qui permet de présenter des résultats de manière compréhensible et acceptable dans les revues.

Mais ce sont des outils inductifs, pas des démonstrations logiques :

- Rejeter ou "accepter" H0/H1 ne prouve jamais qu'une hypothèse est vraie.
- La p-value exprime à quel point les données sont compatibles avec H0, pas la probabilité que H0 soit vraie ou fausse.
- La crédibilité scientifique d'un résultat dépend aussi de la théorie, de la littérature, du plan d'étude, et du bon sens clinique.

1.B Quand l'usage est légitime

Un test statistique est pertinent quand il répond à une question scientifique centrale, formulée à l'avance :

- Comparer un critère principal entre deux traitements dans un essai randomisé.
- Tester l'association entre une exposition majeure (tabac, intervention chirurgicale) et un outcome clinique important (mortalité, complication grave).
- Évaluer un effet que le chercheur est capable d'interpréter en termes cliniques ou de santé publique.

Dans ces situations :

- L'hypothèse nulle H0 et l'hypothèse alternative H1 sont définies clairement.
- Le test vient au service d'une question, et non l'inverse.

1.C Mésusages fréquents (et pourquoi ils posent problème)

C'est surtout l'usage disproportionné des tests qui crée les dérives : on teste tout, partout, pour justifier des phrases qui n'ont pas besoin de test.

1. Test de normalité de Shapiro-Wilk
    - H0 : la variable suit une loi normale dans la population.

- Si p > 0,05, la tentation est de conclure : « La variable est normale, donc on peut utiliser des tests paramétriques. »
- Problèmes :
  - Avec un petit effectif, le test a peu de puissance : il ne rejette H0 que si la distribution est très clairement non normale.
  - Interpréter p > 0,05 comme “preuve de normalité” est faux : cela signifie juste qu’on n’a pas détecté d’écart flagrant avec la normale.
  - Avec un très gros effectif, le test peut donner p < 0,05 pour des écarts minimes, sans impact pratique.
- Conclusion raisonnable :
  - p > 0,05 “aucune incompatibilité majeure détectée avec la normalité, compte tenu de la taille de l’échantillon”.
  - Cela ne dispense pas de regarder un histogramme, un QQ-plot, et d’évaluer si un modèle gaussien est sensé dans le contexte.

Dire “Shapiro est non significatif, donc la variable est normale” est un mésusage.

2. Tester les caractéristiques des perdus de vue
   - Usage fréquent :
     - Comparer les patients perdus de vue à ceux qui restent dans l’étude (âge, sexe, gravité, etc.).
     - Si p > 0,05 partout “les perdus de vue ne sont pas différents, donc ils ne posent pas de problème”.
   - Problèmes :
     - Là encore, p > 0,05 n’est pas une preuve d’égalité :
     - avec peu de perdus de vue, les tests manquent de puissance ;
     - des différences cliniquement importantes peuvent passer sous le radar.
     - Le mécanisme de perte de vue est souvent non testable (motifs non observés, biais de sélection).
   - Conclusion raisonnable :
     - Ces tests peuvent donner un indice (pas de différence massive détectée),
     - mais ne suffisent pas à conclure que les perdus de vue sont “neutres” pour l’analyse.
3. Batteries massives de tests exploratoires
   - Tester toutes les variables de base entre les bras d’un essai randomisé (alors que la randomisation rend ces tests conceptuellement inutiles).
   - Tester tous les sous-groupes, toutes les interactions, toutes les variables disponibles “pour voir ce qui sort”.
   - Retenir ensuite les résultats avec p < 0,05 comme s’il s’agissait de preuves robustes.
   - Problèmes :
     - Inflation du risque d’erreur de type I : plus il y a de tests, plus on obtient de “significatifs” par hasard.

- Risque élevé de p-hacking : on sélectionne après coup ce qui donne un beau p.

1.D Fil conducteur à retenir

- Un test statistique doit être utilisé :
  - pour éclairer une question scientifique précise,
  - avec une hypothèse claire et une interprétation clinique possible.
- Il ne doit pas servir :
  - à “prouver” que quelque chose est vrai ou normal (ex : Shapiro-Wilk),
  - à valider a posteriori des affirmations faibles ou secondaires,
  - ni à compenser l’absence de réflexion sur le plan d’étude.

Les tests sont un outil chiffré parmi d’autres au service du raisonnement scientifique, pas un substitut au raisonnement.

2 Quels tests pour quelles hypohèses ?

2.A Comparaison de 2 pourcentages

2.A.1 Test du chi2 de Pearson

Le test du chi2 de Pearson compare la **répartition** de deux variables qualitatives (ex : présence/absence d’un événement dans deux groupes).

En fait : il **teste l’indépendance** entre deux variables qualitatives dans un tableau de contingence.

$\chi^2 = \sum \frac{(O-E)^2}{E}$

Plus le χ^2 est grand, plus les observations (O) s’éloignent des effectifs attendus sous H0 (E).

Conditions de validité :

- Effectifs théoriques ≥ 5 dans chaque case du tableau de contingence.
- Indépendance des observations = pas de mesures répétées, pas de données appariées.
- Distribution appromiximativement chi2 sous H0 (c’est à dire une somme de carrés de lois normales centrées réduites).
- Approximation de la loi binomiale par la loi normale

2.A.1.1 Chi2 R dans l’étude smp : on teste l’hypothèse que le niveau de dépression est fonction de l’âge, avec âge binarisé pour > ou < 50 ans

Il est nécessaire de désactiver la correction de continuité de Yates (par défaut dans R) pour un test du chi2 classique

La correction du chi2 de Yates est une correction appliquée pour réduire le biais dans l’estimation du chi2 lorsque les effectifs sont petits.

Cependant, elle peut être trop conservatrice et diminuer la puissance du test.

```
smp$agebin <- ifelse(smp$age>50, ">50", "<=50")
table_depression_age <- table(smp$depression, smp$agebin)
chisq.test(table_depression_age, correct=FALSE)
```

```
    Pearson's Chi-squared test

data:  table_depression_age
X-squared = 6.0611, df = 1, p-value = 0.01382
```

L'avntage de cette approche est qu'elle est très simple à mettre en œuvre et à interpréter, et surtout elle affiche un message d'erreur si les conditions de validité ne sont pas remplies.

Avec `tbl_summary` du package `gtsummary` :

```
# | message : false
# | warning: false
tableau <- smp %>%
    tbl_summary(
        by = agebin,
        include = depression,
        statistic = all_categorical() ~ "{n} / {N} ({p}%)",
        percent = "column",
        missing = "no"
    ) %>%
    add_p(
    test = all_categorical() ~ "chisq.test",
    test.args = all_categorical() ~ list(correct = FALSE)   # sans correction de
     ↪  continuité
    )
```

```
2 missing rows in the "agebin" column have been removed.
```

Characteristic	**<=50** N = 646	**>50** N = 151	**p-value**
depression [1] n / N (%) [2] Pearson's Chi-squared test	267 / 646 (41%)	46 / 151 (30%)	0.014

Avec la fonction `CrossTable` du package `gmodels` (mais qui est un peu verbeuse) mais qui donne les résultats complets avec et sans correction de Yates :

```r
CrossTable(
    smp$depression, smp$agebin,
    fisher=TRUE,
    chisq=TRUE,
    expected=TRUE,
    sresid=TRUE,
    format="SPSS")
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  797

               | smp$agebin
smp$depression |     <=50  |      >50  | Row Total |
---------------|-----------|-----------|-----------|
             0 |      379  |      105  |      484  |
               |  392.301  |   91.699  |           |
               |    0.451  |    1.929  |           |
               |   78.306% |   21.694% |   60.728% |
               |   58.669% |   69.536% |           |
               |   47.553% |   13.174% |           |
               |   -0.672  |    1.389  |           |
---------------|-----------|-----------|-----------|
             1 |      267  |       46  |      313  |
               |  253.699  |   59.301  |           |
               |    0.697  |    2.983  |           |
               |   85.304% |   14.696% |   39.272% |
               |   41.331% |   30.464% |           |
               |   33.501% |    5.772% |           |
               |    0.835  |   -1.727  |           |
---------------|-----------|-----------|-----------|
  Column Total |      646  |      151  |      797  |
               |   81.054% |   18.946% |           |
---------------|-----------|-----------|-----------|


Statistics for All Table Factors
```

```
Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  6.061119     d.f. =  1     p =  0.01381901

Pearson's Chi-squared test with Yates' continuity correction
------------------------------------------------------------
Chi^2 =  5.613999     d.f. =  1     p =  0.01781755


Fisher's Exact Test for Count Data
------------------------------------------------------------
Sample estimate odds ratio:  0.6222266

Alternative hypothesis: true odds ratio is not equal to 1
p =  0.0159447
95% confidence interval:  0.4152049 0.9219831

Alternative hypothesis: true odds ratio is less than 1
p =  0.008343204
95% confidence interval:  0 0.8689116

Alternative hypothesis: true odds ratio is greater than 1
p =  0.995116
95% confidence interval:  0.4423434 Inf



       Minimum expected frequency: 59.30113
```

En calculant le test “OR = 1” avec la fonction `epi.2by2` du package `epiR` :

```
library(epiR)
epi.2by2(table_depression_age, method="cohort.count", conf.level=0.95)
```

```
             Outcome+    Outcome-      Total                 Inc risk *
Exposure+         379         105        484     78.31 (74.36 to 81.90)
Exposure-         267          46        313     85.30 (80.89 to 89.04)
Total             646         151        797     81.05 (78.16 to 83.72)

Point estimates and 95% CIs:
---------------------------------------------------------------------
Inc risk ratio                                   0.92 (0.86, 0.98)
Inc odds ratio                                   0.62 (0.43, 0.91)
Attrib risk in the exposed *                     -7.00 (-12.37, -1.62)
Attrib fraction in the exposed (%)              -8.94 (-16.30, -1.83)
Attrib risk in the population *                  -4.25 (-9.02, 0.52)
Attrib fraction in the population (%)           -5.24 (-6.35, -3.49)
---------------------------------------------------------------------
Uncorrected chi2 test that OR = 1: chi2(1) = 6.061 Pr>chi2 = 0.014
Fisher exact test that OR = 1: Pr>chi2 = 0.016
 Wald confidence limits
 CI: confidence interval
 * Outcomes per 100 population units
```

> **❗ Important**
>
> Le test du chi2 teste l’hypothèse d’indépendance entre deux variables qualitatives.
> Ca revient exactement au même de tester l’hypothèse que l’OR = 1 dans un tableau de contingence 2x2.
> Donc la p-value du chi2 et celle du test “OR = 1” sont identiques.

2.A.2 Test exact de Fisher

Le test exact de Fisher compare la **répartition** de deux variables qualitatives dans des petits échantillons (ex : présence/absence d’un événement dans deux groupes).

Conditions de validité :

- Effectifs théoriques < 5 dans au moins une case du tableau de contingence.
- Indépendance des observations = pas de mesures répétées, pas de données appariées.

```
fisher.test(smp$age>75, smp$depression)
```

2.A.2.1 Fisher R

```
    Fisher's Exact Test for Count Data

data:  smp$age > 75 and smp$depression
p-value = 0.7104
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.05833312 3.79332074
sample estimates:
odds ratio
 0.6164281
```

La p value du test vaut 0,71.

On aurait pu aussi faire ça avec d'autres foncions :

Avec `tbl_summary` du package `gtsummary` :

```r
smp$age75 <- ifelse(smp$age>75, ">75", "<=75")
tableau_fisher <- smp %>%
    tbl_summary(
        by = age75,
        include = depression,
        statistic = all_categorical() ~ "{n} / {N} ({p}%)",
        percent = "column",
        missing = "no"
    ) %>%
    add_p(
    test = all_categorical() ~ "fisher.test"
    )
```

```
2 missing rows in the "age75" column have been removed.
```

Characteristic	**<=75** N = 790	**>75** N = 7	**p-value**
depression [1] n / N (%) [2] Fisher's exact test	311 / 790 (39%)	2 / 7 (29%)	0.7

2.B Comparaison d'un pourcentage a un pourcentage théorique

Situation peu fréquente !

Par exemple : équiprobabilité de naissance entre garçons et filles (50%)

```r
binom.test(538, 1000, p = 0.5)
```

```

    Exact binomial test

data:  538 and 1000
number of successes = 538, number of trials = 1000, p-value = 0.01766
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5065264 0.5692493
sample estimates:
probability of success
                 0.538
```

p-value = 0.01766 : on rejette l'hypothèse d'équiprobabilité.

2.C Comparaison de 3 pourcentages

Peu fréquente non plus !

utiliser aussi le test du chi2 de Pearson

```r
smp$age.t <- cut(smp$age, breaks = c(-Inf, 25, 50, Inf))
tb <- table(smp$age.t, smp$depression, deparse.level=2)
prop.table(tb, 1)
```

```
           smp$depression
smp$age.t           0         1
  (-Inf,25] 0.5357143 0.4642857
  (25,50]   0.6007905 0.3992095
  (50, Inf] 0.6953642 0.3046358
```

```r
chisq.test(tb, correct=FALSE)
```

```

    Pearson's Chi-squared test

data:  tb
X-squared = 8.0084, df = 2, p-value = 0.01824
```

Et encore avec `tbl_summary` du package `gtsummary` :

```
tableau_age3 <- smp %>%
    tbl_summary(
        by = age.t,
        include = depression,
        statistic = all_categorical() ~ "{n} / {N} ({p}%)",
        percent = "column",
        missing = "no"
    ) %>%
    add_p(
    test = all_categorical() ~ "chisq.test",
    test.args = all_categorical() ~ list(correct = FALSE)   # sans correction de
     ↪  continuité
    )
```

```
2 missing rows in the "age.t" column have been removed.
```

Characteristic	(-Inf,25] N = 140	(25,50] N = 506	(50, Inf] N = 151	p-value
depression [1] n / N (%) [2] Pearson's Chi-squared test	65 / 140 (46%)	202 / 506 (40%)	46 / 151 (30%)	0.018

Le problème pour interpréter cette p-value :

- On sait qu'il y a une différence entre au moins deux groupes d'âge, mais on ne sait pas lesquels.
- On sait en gros que l'hypothèse $p_1 = p_2 = p_3$ est fausse, mais on ne sait pas quelles sont les paires pour lesquelles $p_i \neq p_j$.

Dans cette situation : il faut utiliser un test Chi2 de tendance :

```
num <- tb[ ,2]
den <- tb[ ,1] + tb[ ,2] # on aurait pu faire rowSums(tb)
prop.trend.test(num, den)
```

```
    Chi-squared Test for Trend in Proportions

data:  num out of den ,
 using scores: 1 2 3
X-squared = 7.84, df = 1, p-value = 0.00511
```

La p value est < 0,05 : on rejette l'hypothèse d'égalité des proportions en fonction de l'âge.

Il y a donc une tendance significative de la proportion de dépression en fonction de l'âge.

On peut aussi faire ça avec le package `gtsummary` mais il faut définir une fonction de test personnalisée ce qui n'est pas dingo quand même !

```
trend_test <- function(data, variable, by, ...) {
    tab <- table(data[[by]], data[[variable]])  # lignes = âge, colonnes =
↪   depression
    # on teste la tendance des proportions de "depression = 1" selon l'âge
    res <- prop.trend.test(x = tab[, "1"], n = rowSums(tab))
    dplyr::tibble(p.value = res$p.value)
}

tableau_trend <- smp %>%
    tbl_summary(
        by = age.t,
        include = depression,
        statistic = all_categorical() ~ "{n} / {N} ({p}%)",
        percent = "column",
        missing = "no"
    ) %>%
    add_p(
        test = depression ~ "trend_test"
    )
```

```
2 missing rows in the "age.t" column have been removed.
```

Characteristic	**(-Inf,25]** N = 140	**(25,50]** N = 506	**(50, Inf]** N = 151	**p-value**
depression [1] n / N (%)	65 / 140 (46%)	202 / 506 (40%)	46 / 151 (30%)	0.005

2.D Comparaison de 2 moyennes

Repose sur :

- le test t de Student pour échantillons indépendants (ou appariés)
- le test de Welch (var.equal=FALSE dans R)
- le test t de Student pour échantillons appariés (paired=TRUE dans R)
- le test non paramétrique de Wilcoxon :
  - test de Mann-Whitney pour échantillons indépendants
  - test de Wilcoxon pour échantillons appariés

Test t de student pour échantillons indépendants

Conditions de validité :

- Variable quantitative continue approximativement normale dans chaque groupe (mais en vrai s'il y a plus de 30 sujets par groupe, on peut s'en fiche un peu)
- Homogénéité des variances dans chaque groupe (c'est à dire que les variances sont à peu près similaires)

Pour apprécier la comparabilité des variances :

- diagramme de normalité (= QQ plot) (voir `qqnorm` et `qqline` en R) dans chaque groupe ou courbe de densité superposée à la loi normale (voir `density` et `lines` en R)
- calcul des écarts-types dans chaque groupe (`by(var1, var2, sd, na.rm=TRUE)` en R)
- test de Levene : bof bof parce que c'est comme Shapiro : si p > 0,05 on conclut que les variances sont égales (ce qui n'est pas une preuve !) et si p < 0,05 on conclut que les variances sont différentes (ce qui peut être dû à un petit échantillon

Si les variances ne sont vraiment pas égales : utiliser le test de Welch (var.equal=FALSE dans R)

Conditions de validité :

- déjà : avec n>30 par groupe, on s'en fiche un peu !! (sauf si différence ÉNORME de taille entre les groupes)
- si < 30 mais les groupes ont à peu près les mêmes effectifs + les mêmes distributions (11 et 13 sujets, écarts types similaires) : on peut s'en fiche aussi
- la validité est douteuse si taille inégale ou distribution très différentes (30 vs 300 sujets, CRP négative vs 250 mg/L)

Si c'est vraiment très éloigné d'une distribution normale (comme le nombre d'enfants) : utiliser un test non paramétrique (Wilcoxon)

> **Note**
>
> **Wilcoxon / Mann Withney**
> **Wilcoxon (`wilcox.test` en R)**
>
> - peut être apparié `paired = TRUE` (Wilcoxon signed rank)
> - ou non apparié `paired = FALSE` (Mann Whitney / Wilcoxon rank sum)

Mann Whitney au sens strict

- c'est la version non appariée seulement
- pour deux groupes indépendants.
- sur R : `wilcox.test(var1 ~ var2, data=..., paired=FALSE)`

2.D.0.1 Comparaison de 2 moyennes avec R

On teste l'

hypothèse : moyenne d'âge des détenus déprimés est différente de la moyenne d'âge des détenus non déprimés

```
t.test(smp$age ~ smp$depression, var.equal=TRUE)
```

```

    Two Sample t-test

data:  smp$age by smp$depression
t = 2.6337, df = 795, p-value = 0.008611
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 0.6425038 4.4032412
sample estimates:
mean in group 0 mean in group 1
       39.93182        37.40895
```

La p value est de 0,0086 : on rejette l'hypothèse d'égalité des moyennes d'âge entre les deux groupes.

maginons maintenant que nous souhaitions comparer le nombre d'enfants des détenus selon qu'ils aient ou non été abusés dans leur enfance. Malgré la taille importante de l'échantillon, le fait que la distribution du nombre d'enfants soit très éloignée d'une distribution normale peut nous rendre réticent à l'idée d'utiliser un test t. L'alternative est alors de recourir à un test non paramétrique comme le test de Wilcoxon ou de Mann-Whitney

Si on veut comparer le nombre d'enfants entre les deux groupes de détenus (abusés ou non dans leur enfance) :

La distribution du nombre d'enfants est très asymétrique (beaucoup de zéros, quelques valeurs élevées)

$\longrightarrow$ on utilise un test non paramétrique = test de Wilcoxon / Mann-Whitney (en l'occurence Mann Whitney car les deux groupes sont indépendants)

```
wilcox.test(smp$nb.enfants~smp$abus.enfant, correct=FALSE)
```

```

    Wilcoxon rank sum test
```

```
data:  smp$nb.enfants by smp$abus.enfant
W = 62664, p-value = 0.1938
alternative hypothesis: true location shift is not equal to 0
```

NB : si on voulait en faisant paired = TRUE, il faut passer la syntaxe en mode vecteur (mais ça ne marcherait pas ici car les deux groupes ne sont pas appariés et donc n'ont pas le même effectif) :

```r
x <- smp$nb.enfants[smp$abus.enfant == 0]  # groupe 0
y <- smp$nb.enfants[smp$abus.enfant == 1]  # groupe 1

wilcox.test(x, y, paired = TRUE, correct = FALSE)
```

! Important

Résumé des tests de comparaison de moyennes
Comparaison de moyennes avec test t de Student :

- Effectif : > 30 par groupe ou à peu près égaux (11 et 12 par exemple)
- Distribution : symétrique (approximativement normale) et variances à peu près égales

Si échantillons appariés : test t apparié, ou Wilcoxon apparié (= signed rank)
Si distribution très asymétrique : test non paramétrique de Mann-Whitney (= Wilcoxon non apparié = rank sum)
Si effectifs très inégaux ou variances très différentes : test de Welch

2.E Comparaison de 3 moyennes ou plus

Comparaison pas si fréquente (idem que la comparaison de 3 pourcentages).

Faire un test pour rejeter l'hypothèse $\mu_1 = \mu_2 = \mu_3$ a peu de sens !

Le test ANOVA (Analysis of Variance) permet de comparer les moyennes de plusieurs groupes (3 ou plus).

Si une variable à une distribution égale dans les 3 sous groupes, alors la variacnce intra-groupe doit être proche de la variance inter-groupe.

On passe par la comparaison des variances car on ne peut pas faire de comparaison directe des moyennes quand il y a plus de 2 groupes.

Le test ANOVA calcule le ratio de ces variances (F-statistique) et détermine la p-value associée en fonction de la distribution de F.

$$F = \frac{\text{Variance inter-groupe}}{\text{Variance intra-groupe}}$$

Exemple imagé :

- 3 plantes (A, B, C) reçoivent 3 traitements différents (T1, T2, T3).
- On mesure la hauteur des plantes après 1 mois..
- On calcule la variance des hauteurs entre les groupes (inter-groupe) et la variance des hauteurs à l'intérieur de chaque groupe (intra-groupe).
- Si les traitements n'ont pas d'effet, les variances inter et intra-groupe sont similaires.
- Si un traitement a un effet, la variance inter-groupe sera plus grande que la variance intra-groupe.

Conditions de validité :

- Indépendance des observations (pas de mesures répétées, pas de données appariées).
- Variance de la variable étudiée similaire dans chaque groupe (homoscédasticité).
- Normalité des résidus du modèle (avec la même résistance que pour le test t : si n>30 par groupe et faible différences des deux groupes, on peut s'en fiche un peu).

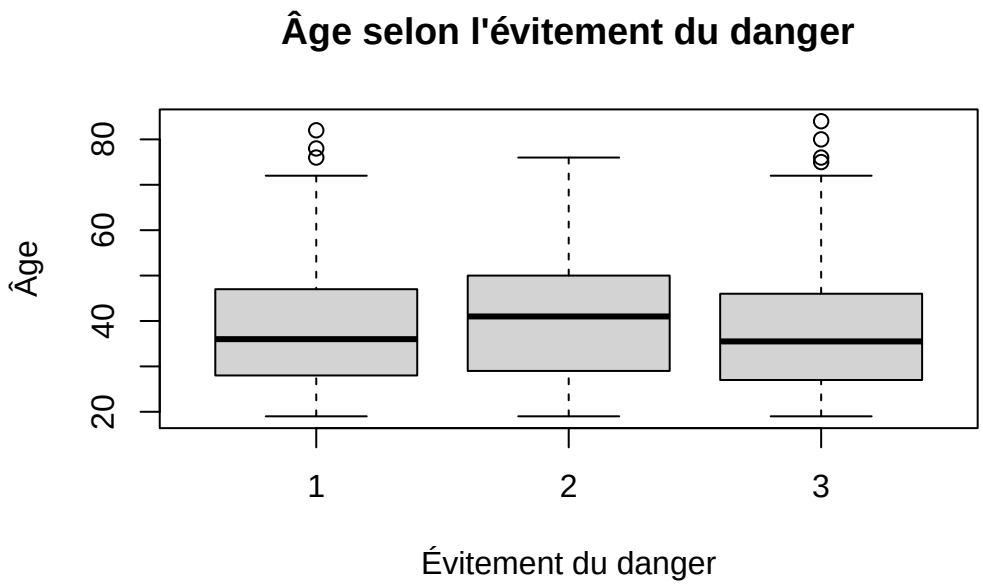
2.E.1 Exemple R

Exemple : niveau d'évitement du danger mesuré avec un questionnaire codée en 1, 2 et 3 pour des niveaux respectivement « faible », « modéré » ou « élevé ».

Il faut réfuter l'hypothèse que la moyenne d'âge est la même dans les trois groupes d'évitement du danger.

D'abord visualisation :

```
boxplot(age ~ evit.danger, data = smp,
        xlab = "Évitement du danger",
        ylab = "Âge",
        main = "Âge selon l'évitement du danger")
```



Les moyennes et écart-types d'âge semblent visuellement proches dans les trois groupes

Sur R, il faut passer par la fonction `lm` (linear model) pour faire une ANOVA :

Attention, il faut bien transformer la variable catégorielle en facteur avec `as.factor()` (sinon R la traite comme une variable quantitative continue !)

Syntaxe :

- `lm(var_quantitative ~ var_catégorielle, data=...)`
- test : fonction `anova(...)` ou `drop1(..., test="F")`

La fonction `drop1` permet de tester l'effet d'une variable dans un modèle linéaire en comparant le modèle complet avec un modèle réduit (sans la variable testée).

```
res <- lm(age~as.factor(evit.danger), data=smp)
drop1(res, test="F")
```

```
Single term deletions

Model:
age ~ as.factor(evit.danger)
                       Df Sum of Sq    RSS    AIC F value  Pr(>F)
<none>                               122212 3582.2
as.factor(evit.danger)  2    1062.1 123274 3584.2  2.9895 0.05097 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ou anova(res)
anova(res)
```

```
Analysis of Variance Table
```

```
Response: age
                        Df Sum Sq Mean Sq F value  Pr(>F)
as.factor(evit.danger)   2   1062  531.04  2.9895 0.05097 .
Residuals              688 122212  177.63
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

> **Note**
>
> **Lien entre ANOVA et régression linéaire**
> Fondamentalement, l'ANOVA et la régression linéaire sont la même méthode mathématique (elles appartiennent toutes deux au "Modèle Linéaire Général").
> La différence est dans leur application :
>
> - L'ANOVA est un cas particulier de régression utilisée quand les variables explicatives sont uniquement des catégories (des groupes comme "Engrais A, B, C"). Elle regarde si les moyennes de ces catégories diffèrent.
> - La Régression Linéaire est plus souple et permet d'utiliser des variables explicatives continues (des nombres comme la température ou l'âge) pour prédire une valeur.
>
> En résumé : faire une ANOVA revient exactement à faire une régression linéaire sur des variables catégorielles (transformées en variables "muettes" ou dummy variables).
> Elles utilisent toutes les deux le même moteur : analyser la variance pour séparer le signal du bruit.

2.F Test de nullité d'un coefficient de corrélation

Les coefficients de corrélation mesurent la force et la direction d'une association linéaire entre deux variables quantitatives.

Ils peuvent être :

- Coefficient de corrélation de Pearson (r) : mesure la corrélation linéaire entre deux variables quantitatives continues. Varie entre -1 et +1.
- Coefficient de corrélation de Spearman (ρ) : mesure la corrélation monotone entre deux variables, basé sur les rangs. Varie entre -1 et +1.
- Coefficient kappa de Cohen (κ) : mesure l'accord entre deux évaluateurs pour des variables qualitatives. Varie entre -1 et +1.
- Coefficient de corrélation intraclasse (ICC) : mesure la fiabilité entre plusieurs mesures quantitatives. Varie entre 0 et 1.

Tester la nullité de corrélation revient à tester l'absence d'association linéaire entre les variables, c'est à dire leur indépendance linéaire.

(2 variables indépendantes ont nécessairement une absence de corrélation linéaire)

Attention : tester la nullité d'une corrélation n'implique pas que le coefficient de corrélation soit = 0 !

Attention aussi aux conditions de validité des coefficients

Attention aussi aux conditions de validité du test de nullité

- Dans l'idéal : que X et Y soient distribuées normalement dans la population
- Faisable quand même si X est normale et Y est binaire
- Et même si les deux sont binaires !

2.F.1 Test de nullité du coefficient de corrélation avec R

Utilisation de la fonction `cor.test(...)` :

```
cor.test(smp$age, smp$recherche.nouv, method="pearson")
```

```
    Pearson's product-moment correlation

data:  smp$age and smp$recherche.nouv
t = -6.009, df = 693, p-value = 3.016e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2920740 -0.1506631
sample estimates:
       cor
-0.2225387
```

- Corrélation négative à -0,22 entre âge et recherche de nouveauté.

- p-value < 0,0001 : on rejette l'hypothèse d'absence de corrélation linéaire entre âge et recherche de nouveauté : **il est très peu probable que les variables soient indépendantes linéairement**.

Tableau corrélation

2.G Tests appariés / non appariés

Exemple de données appariées :

- Mesures répétées sur les mêmes sujets (avant/après traitement).
- Paires de sujets liés (jumeaux, couples).
- Score de propension

> **!** Important
>
> **Tests possibles sur données appariées :**
>
> - Sur des moyennes :
> - Test t apparié (paired=TRUE dans R) : avec conditions de validité du test t.
> - Test de Wilcoxon apparié (paired=TRUE dans R) : si conditions de validité du test t non remplies (surtout asymétrie évidente)
> - Sur des pourcentages :
> - Test du Chi2 de McNemar : pour comparer deux proportions appariées (ex : succès/échec avant/après traitement sur les mêmes sujets).
> - Test exact du Chi2 de McNemar : version exacte du test de McNemar, utilisé lorsque les effectifs sont petits.

2.G.1 Exemple R données appariées

2.G.1.1 Exemple 1 jeu de données `scl` : évolution au cours du temps du score SCL90 de patients déprimés hospitalisés

Objectif : tester l'hypothèse que le score SCL90 au jour 0 est différent du score SCL90 au jour 4

1. Visualisation des données

```r
head(scl)
```

```
  NUMERO   BIRTHDAT SEXE HEIGHT WEIGHT Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9 Q10 Q11 Q12
1      0 0073-06-18    2    157     56 NA NA NA NA NA NA NA NA NA  NA  NA  NA
3      3 0077-03-22    2    173     60  3  3  2  1  1  2  0  0  2   2   3   3
2      3 0077-03-22    2    173     60  2  4  2  1  3  0  0  0  2   2   3   4
4      3 0077-03-22    2    173     60  1  3  3  0  2  1  0  0  2   3   3   3
5      4       <NA>   NA     NA     NA  3  3  4  2  3  2  3  1  2   3   4   4
```

```
6       4        <NA>   NA      NA      NA  2  1  1  0  0  0  0  0  2   2   1   3
  Q13 Q14 Q15 Q16 Q17 Q18 Q19 Q20 Q21 Q22 Q23 Q24 Q25 Q26 Q27 Q28 Q29 Q30 Q31
1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
3   0   3   1   0   1   3   0   4   1   3   3   3   0   3   2   3   3   3   3
2   0   4   1   0   2   2   3   4   1   3   3   3   0   4   1   4   3   4   3
4   0   4   1   0   1   1   1   4   1   4   3   4   0   3   1   4   4   4   4
5   1   3   2   0   1   2   1   2   2   0   2   3   0   4   2   3   3   4   4
6   0   3   1   0   1   0   0   0   1   0   1   2   0   3   1   3   3   4   4
  Q32 Q33 Q34 Q35 Q36 Q37 Q38 Q39 Q40 Q41 Q42 Q43 Q44 Q45 Q46 Q47 Q48 Q49 Q50
1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
3   0   2   3   2   2   1   1   3   3   2   2   1   2   3   2   0   2   2   3
2   0   3   2   3   1   0   2   4   4   2   1   1   3   1   3   0   2   2   3
4   0   3   3   2   1   0   2   3   0   1   1   1   1   2   3   0   1   1   3
5   4   1   1   2   1   0   3   4   4   4   3   3   0   3   3   0   3   3   3
6   3   0   1   1   1   0   2   3   1   3   0   0   0   1   3   0   4   4   2
  Q51 Q52 Q53 Q54 Q55 Q56 Q57 Q58 Q59 Q60 Q61 Q62 Q63 Q64 Q65 Q66 Q67 Q68 Q69
1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
3   2   1   3   2   3   2   4   2   2   3   3   3   0   0   3   1   0   2   3
2   1   1   3   2   4   2   4   1   1   0   1   0   0   2   2   3   0   0   2
4   1   1   3   1   4   1   4   1   1   1   2   0   1   1   2   1   1   0   1
5   2   3   2   4   4   3   4   3   2   1   2   3   4   0   2   3   0   0   3
6   2   1   0   4   4   4   3   1   1   0   0   2   3   0   2   3   3   1   2
  Q70 Q71 Q72 Q73 Q74 Q75 Q76 Q77 Q78 Q79 Q80 Q81 Q82 Q83 Q84 Q85 Q86 Q87 Q88
1  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA  NA
3   2   3   3   0   1   3   1   3   3   1   1   0   1   1   0   0   2   3   0
2   0   4   4   0   1   3   2   2   3   0   1   0   0   1   0   0   3   1   0
4   2   3   3   0   1   4   3   3   3   1   0   1   0   1   0   0   3   1   1
5   2   3   3   1   1   0   0   3   3   4   3   2   2   2   1   0   2   3   3
6   1   3   2   1   0   0   1   2   3   4   1   0   1   0   1   0   2   3   2
  Q89 Q90 age VISIT scl JOUR.VISIT somatisation obsession sensitivite
1  NA  NA  25  <NA>  NA          1           NA         NA          NA
3   1   3  21    J0 168          1           26         23          17
2   2   3  21    J4 164          5           24         23           9
4   3   1  21    J7 154          7           14         26          10
5   3   4  NA    J0 208          1           35         29          16
6   3   4  NA    J4 138          4           21         22           9
  depression anxiete hostilite phobie paranoia psychose divers
1         NA      NA        NA     NA       NA       NA     NA
3         30      25         7      9        8       14      9
2         35      31         7      6        6        9     14
4         35      26        11      9        6        8      9
5         40      26        14      8        8       22     10
6         32      17         9      4        2       15      7
```

Transformer en format long avec `dcast` du package `reshape2` :

```r
scl.w <- dcast(
        scl[,c("NUMERO","VISIT","depression")],
        NUMERO~VISIT,
        value.var="depression")
```

t-test apparié :

```r
t.test(scl.w$J0, scl.w$J4, paired=TRUE)
```

```

    Paired t-test

data:  scl.w$J0 and scl.w$J4
t = 6.2146, df = 99, p-value = 1.223e-08
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 3.172146 6.147854
sample estimates:
mean difference
           4.66
```

p-value = 1.223e-08 : on rejette l'hypothèse d'égalité des moyennes de score SCL90 entre le jour 0 et le jour 4.

2.G.1.2 Exemple 2 Comparaison de pourcentages = test du Chi2 de McNemar

Ici, on teste l'hypothèse que la proportion de patients déprimés (score SCL90 > 20) est la même au jour 0 et au jour 4.

```r
tab <- table(scl.w$J0>20, scl.w$J4>20, deparse.level=2)
tab
```

```
             scl.w$J4 > 20
scl.w$J0 > 20 FALSE TRUE
        FALSE     9    0
        TRUE     16   75
```

Le tableau de contingence montre que :

- 9 patients ne sont pas déprimés aux jours 0 et 4.
- 75 patients sont déprimés aux jours 0 et 4.
- 16 patients sont déprimés au jour 0 mais pas au jour 4.
- 0 patients sont déprimés au jour 4 mais pas au jour 0.

```r
mcnemar.test(tab, correct=FALSE)
```

```
    McNemar's Chi-squared test

data:  tab
McNemar's chi-squared = 16, df = 1, p-value = 6.334e-05
```

mcnemar.test(tab, correct=FALSE)

```
    McNemar's Chi-squared test
```

data: tab McNemar's chi-squared = 16, df = 1, p-value = 6.334e-05

p-value = 6.334e-05 : on rejette l'hypothèse d'égalité des proportions de patients déprimés entre le jour 0 et le jour 4.

2.G.1.3 tbl_summary apparié On met les deux : comparaison de moyennes et comparaison de pourcentages

Après PSM :

```r
table3adjusted <- m.final %>%
    tbl_summary(
        include = all_of(cols_to_include_in_table3),
        missing = "ifany",
        by = I_geste_comparaison_1HartmannCoelio,  # Comparaison entre groupes
        percent = "column",
        statistic = list(
        all_categorical() ~ "{n} ({p}%)",
        all_continuous() ~ "{median} [{p25}, {p75}]"
        )
    ) %>%
    add_p(
        test = list(
        all_continuous() ~ "paired.wilcox.test",  # Test de Wilcoxon pour les
         ↪  donnees continues
        all_categorical() ~ "mcnemar.test"  # Test de McNemar pour les donnees
         ↪  categoriques
        ),
        group = Pair_ID  # Utilisation de l'ID de paire
    )
```

Mais selon chat, utilisation de `group = Pair_ID` n'est pas très safe et conseille de faire comme ça :

1. Fonction Wilcoxon apparié personnalisée

```r
paired_wilcox_gts <- function(data, variable, by, ...) {
    d <- data |>
        dplyr::select(Pair_ID, all_of(by), all_of(variable)) |>
        tidyr::pivot_wider(
        names_from = !!rlang::sym(by),
        values_from = !!rlang::sym(variable)
        )

    res <- wilcox.test(d[[1]], d[[2]], paired = TRUE)
    tibble::tibble(p.value = res$p.value)
    }
```

2. Fonction McNemar personnalisée

```r
mcnemar_gts <- function(data, variable, by, ...) {
    d <- data |>
        dplyr::select(Pair_ID, all_of(by), all_of(variable)) |>
        tidyr::pivot_wider(
        names_from = !!rlang::sym(by),
        values_from = !!rlang::sym(variable)
        )

    tab <- table(d[[1]], d[[2]])
    res <- mcnemar.test(tab, correct = FALSE)

    tibble::tibble(p.value = res$p.value)
    }
```

3. Fonction t de Student apparié personnalisée

```r
paired_ttest_gts <- function(data, variable, by, ...) {

    d <- data %>%
        dplyr::select(Pair_ID, !!rlang::sym(by), !!rlang::sym(variable)) %>%
        tidyr::pivot_wider(
        names_from  = !!rlang::sym(by),
        values_from = !!rlang::sym(variable)
        )

    res <- t.test(d[[1]], d[[2]], paired = TRUE)

    tibble::tibble(p.value = res$p.value)
    }
```

4. Tableau final

```r
table3adjusted <- m.final %>%
    tbl_summary(
        include = all_of(cols_to_include),
        by = colonne_de_comparaison,  # Comparaison entre groupes
        percent = "column",
        statistic = list(
        all_categorical() ~ "{n} ({p}%)",
        all_continuous()  ~ "{median} [{p25}, {p75}]"
        )
    ) %>%
    add_p(
        test = list(
        all_continuous()  ~ paired_wilcox_gts,   # ou paired_ttest_gts
        all_categorical() ~ mcnemar_gts
        )
    )
```

3 Calcul de β et du nombre de sujets à inclure

Dans la théorie de Neyman et Pearson, le calcul du nombre de sujets à inclure dans une étude est directement lié au contrôle des risques α et β.

Exemple : équiprobabilité des naissances de garçons (p_1) et de filles (p_2)

- $H_0 : p_1 = 0,5$ (et donc $p_2 = 0,5$ également)
- $H_1 : p_1 = 0,55$ (et donc $p_2 = 0,45$)

Pour α fixé (par exemple à 5%), on peut déterminer une règle de décision (seuil critique) qui permet de rejeter H_0 si le pourcentage de naissances de garçons est suffisamment éloigné de 50%.

β varie

- la différence vraie : Plus l'effet réel est grand, plus il est facile à détecter, donc β diminue (et la puissance augmente).
- le nombre de sujets inclus : Plus on inclut de sujets, plus on a d'informations, donc β diminue (et la puissance augmente).
- H_1 choisi : plus H_1 est proche de H_0, plus il est difficile de détecter la différence, donc β augmente (et la puissance diminue).

Exemple :

- hypothèse $p_1 = 0,55 \rightarrow$ pour $\alpha = 5\%$ et $\beta = 10\%$

```r
pwr.p.test(h=ES.h(0.5,0.55),power=0.9,sig.level=0.05,alternative="less")
```

```
     proportion power calculation for binomial distribution (arcsine transformation) 

              h = -0.1001674
              n = 853.5244
      sig.level = 0.05
          power = 0.9
    alternative = less
```

Pour $p_1 = 0,55$, il faut environ 850 naissances pour obtenir une puissance de 90% avec un risque α de 5%.

- hypothèse $p_1 = 0,60 \rightarrow$ pour $\alpha = 5\%$ et $\beta = 10\%$

```r
pwr.p.test(h=ES.h(0.5,0.60),power=0.9,sig.level=0.05,alternative="less")
```

```
     proportion power calculation for binomial distribution (arcsine transformation) 

              h = -0.2013579
```

```
              n = 211.2183
      sig.level = 0.05
          power = 0.9
    alternative = less
```

Pour $p_1 = 0,60$, il faut environ 200 naissances pour obtenir une puissance de 90% avec un risque α de 5%.

- hypothèse $p_1 = 0,52 \rightarrow$ pour $\alpha = 5\%$ et $\beta = 10\%$

```r
pwr.p.test(h=ES.h(0.5,0.52),power=0.9,sig.level=0.05,alternative="less")
```

```
     proportion power calculation for binomial distribution (arcsine transformation)

              h = -0.04001067
              n = 5349.549
      sig.level = 0.05
          power = 0.9
    alternative = less
```

Pour $p_1 = 0,52$, il faut environ 5000 naissances pour obtenir une puissance de 90% avec un risque α de 5%.

3.A Exemple R

enons l'exemple d'un essai évaluant l'efficacité d'une nouvelle molécule dans le sepsis sévère (étude comparative versus placebo). Compte tenu des critères d'inclusion, les données épidémiologiques font état d'une mortalité à 30 jours (critère d'efficacité) égale à 36%.

Exemple : efficacité d'une nouvelle molédcule.

- H_0 : mortalité = 36% (placebo)
- Investigateurs : "ne pas passer à côté d'une baisse de 10 points de la mortalité"
- H_1 : mortalité = 26% (nouvelle molécule)
- α = 5% (risque de conclure à tort que la molécule est efficace)
- β = 10% (risque de ne pas détecter l'efficacité de la molécule)

La fonction `pwr.2p.test` du package `pwr` permet de calculer le nombre de sujets nécessaires pour atteindre la puissance désirée dans une étude comparant deux proportions.

```r
pwr.2p.test(
    h = ES.h(p1 = 0.36, p2 = 0.26), # taille d'effet
    sig.level = 0.05,
    power= 0.90,
    alternative = "two.sided"
    )
```

```
     Difference of proportion power calculation for binomial distribution (arcsine transformati

              h = 0.2168606
              n = 446.8531
      sig.level = 0.05
          power = 0.9
    alternative = two.sided

NOTE: same sample sizes
```

Il faut donc inclure environ 447 patients par groupe (soit un total de 894 patients) pour détecter une différence de mortalité de 10 points entre les deux groupes avec une puissance de 90% et un risque α de 5%.

4 Tests uni- ou bilatéraux ?

4.A Définitions

- Selon Fisher :
  - inférence statistique bilatérale par nature
  - on cherche à apprécier l'adéquation entre un échantillon observé et une population hypothétique infinie sous-jacente.
  - donc on teste toujours si l'estimateur issu de l'échantillon est suffisamment différent du paramètre de la population sous-jacente (qu'il soit supérieur ou inférieur).
- Selon Neyman et Pearson :
  - inférence statistique unilatérale inférieure ou supérieure selon le choix des hypothèses H_0 et H_1.

4.B Exemple

1. Essai randomisé : on teste

- H_0 : la nouvelle thérapeutique n'est pas plus efficace que le placebo
- H_1 : la nouvelle thérapeutique est plus efficace que le placebo

Mais on teste pas l'infériorité de la nouvelle thérapeutique par rapport au placebo.

Donc en soit, c'est un test unilatéral supérieur, et le risque de première espèce α est à 2,5% (et non 5%).

2. Étude d'**inférence inductive** : on teste

- Formulation bilatéral
- Avec le risque de première espèce α égal à 5%.

5 Intervalles de confiance

5.A Définitions

"Vraie valeur" de la moyenne = espérance de la moyenne =

$$\mu \text{ +/- } \frac{\sigma}{\sqrt{n}}$$

Avec : σ = écart-type de la population et n = taille de l'échantillon

En fait :

- Si âge : 40 [37-43] ans dans un échantillon de 100 personnes
- C'est l'intervalle de confiance qui a 95% de chances de contenir la vraie valeur de la moyenne d'âge dans la population.
- et pas "il y a 95% des valeurs de l'échantillon qui sont entre 37 et 43 ans".

L'intervalle de confiance à 95% correspond à la grande majorité (95%) des valeurs **théoriques possibles** de ce paramètre, compatibles avec l'échantillon observé.

Pour observer les fluctuations de l'estimation d'un paramètre dans un échantillon donné, il faut utiliser les intervalles empiriques ou les intervalles de fluctuation.

En gardant à l'esprit que les observations sont considérées implicitement comme un échantillon aléatoire issu d'une population sous-jacente alors que ça n'est que rarement le cas dans les faits !

5.B Exemple R

Prévalence de la schizophrénie dans l'étude `smp`.

FOnction `binom.confint()` de la bibliothèque `binom` de R permet d'estimer la prévalence et un intervalle de confiance.

```
y <- na.omit(smp$schizophrenie)
binom.confint(x=sum(y), n=length(y), method="asymptotic")
```

```
      method  x   n       mean      lower      upper
1 asymptotic 64 799 0.08010013 0.06127831 0.09892194
```

```
cat("Prévalence de la schizophrénie : ",
    round(binom.confint(x=sum(y), n=length(y), method="asymptotic")$mean*100,2),
    "% [",
    round(binom.confint(x=sum(y), n=length(y), method="asymptotic")$lower*100,2),
    "% ; ",
    round(binom.confint(x=sum(y), n=length(y), method="asymptotic")$upper*100,2),
    "%]"
    )
```

```
Prévalence de la schizophrénie :  8.01 % [ 6.13 % ;  9.89 %]
```

Donc l'échantillon observé est compatible avec des populations infinies sous-jacentes pour lesquelles la « vraie » prévalence de schizophrénie serait comprise entre 6,1% et 9,9%.

Le truc c'est qu'on a fait un sondage en grappe, donc il faut utiliser la bibliothèque `survey` de R pour calculer un intervalle de confiance correct.

```r
smp.design <- svydesign(ids=~1, data=smp)
```

```
Warning in svydesign.default(ids = ~1, data = smp): No weights or probabilities
supplied, assuming equal probability
```

```r
svyciprop(~schizophrenie, smp.design, na.rm=TRUE, method="beta")
```

```
                       2.5%  97.5%
schizophrenie 0.0801 0.0622 0.1012
```

Sur une variable quantitative : fonction `t.test()` de R permet d'estimer la moyenne et un intervalle de confiance.

```r
cat("Moyenne d'âge : ",
    round(t.test(smp$age)$estimate,2),
    " ans [",
    round(t.test(smp$age)$conf.int[1],2),
    " ans ; ",
    round(t.test(smp$age)$conf.int[2],2),
    " ans]"
    )
```

```
Moyenne d'âge :  38.94  ans [ 38.02  ans ;  39.86  ans]
```

Et pour sondage en grappe :

```r
modsvy <- svyglm(age~1, design=smp.design)
confint(modsvy)
```

```
               2.5 %   97.5 %
(Intercept) 38.01932 39.86273
```

4Modèles

1 Modèle linéaire

1.A Principe et mise en œuvre

Modèle linéaire = modèle de régression linéaire = régression linéaire.

Utilisé quand une variable quantitative (continue) normalement distribuée Y doit être mise en relation avec une ou plusieurs variables $X_1, X_2, ..., X_p$ (quantitatives ou qualitatives, binaires ou catégorielles (dans le cas de variables catégorielles, elles sont automatiquement transformées en variables binaires)).

- Y = variable dépendante, expliquée, à prédire
- $X_1, X_2, ..., X_p$ = variables indépendantes, explicatives, prédictives, ou régresseurs

Équation du modèle linéaire :

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + ... + \alpha_p X_p + \epsilon$$

où :

- α_0 = ordonnée à l'origine (intercept)
- $\alpha_1, \alpha_2, ..., \alpha_p$ = coefficients de régression (pentes)
- ϵ = terme d'erreur (résidus), supposé suivre une loi normale centrée réduite (d'espérance nulle = 0, avec espérance = moyenne)

Objectif de l'algorithme = estimer les coefficients de régression ($\alpha_0, \alpha_1, \alpha_2, ..., \alpha_p$) de façon à minimiser la variance estimée de ϵ (méthode des moindres carrés).

En gros : l'objectif est de trouver le modèle pour lequel le bruit est minimal.

Conditions de validité

- Normalité de la variable résiduelle ϵ
- Homoscédasticité des résidus (variance constante de ϵ pour toutes les valeurs de X)
- Indépendance des résidus ε (les résidus doivent être indépendants les uns des autres, c'est à dire pas de structure d'autocorrélation en particulier pas de structure temporelle)

> **Note**
>
> **Relation régression linéaire et ANOVA**
> L'ANOVA (analyse de la variance) est en fait un cas particulier de la régression linéaire où toutes les variables explicatives sont catégorielles. Ainsi, une ANOVA à un facteur est équivalente à une régression linéaire avec une variable explicative binaire.

1.B Exemple avec le jeu de données `smp`

Objectif : modéliser la durée de l'entretien en fonction de différentes variables explicatives.

La variable à expliquer est : durée de l'entretien (quantitative, donc régression linéaire adaptée)

- âge (quantitative)
- trouble psychiatrique
- trouble de la personnalité
- traumatisme pendant l'enfance
- type de prison

1. Description des variables :

- Étude descriptive des variables
- Vérification des conditions de validité du modèle : notament la normalité de la variable à expliquer

2. Construction du modèle linéaire

```
smp2 <- smp[smp$duree.interv>15,] # exclut valeurs aberrantes
mod <- lm(
    duree.interv~
    schizophrenie + depression + abus.subst + gravite + charactere + trauma.enfant
 ↪  + age + factor(type.centre),
    data=smp2)
summary(mod)
```

```
Call:
lm(formula = duree.interv ~ schizophrenie + depression + abus.subst +
    gravite + charactere + trauma.enfant + age + factor(type.centre),
    data = smp2)

Residuals:
    Min      1Q  Median      3Q     Max
-38.578 -13.855  -1.769  10.922  64.405

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          42.48996    3.99998  10.623  < 2e-16 ***
schizophrenie         3.06420    2.80997   1.090 0.275911
depression            6.71269    1.64793   4.073 5.21e-05 ***
abus.subst            4.60037    1.79854   2.558 0.010760 *
gravite               1.06236    0.56548   1.879 0.060737 .
charactere            1.62547    0.93536   1.738 0.082723 .
trauma.enfant        -0.66805    1.66931  -0.400 0.689144
age                   0.20788    0.06066   3.427 0.000649 ***
factor(type.centre)2  4.09540    2.53401   1.616 0.106546
factor(type.centre)3 -1.29681    2.44159  -0.531 0.595509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 18.02 on 645 degrees of freedom
  (144 observations deleted due to missingness)
Multiple R-squared:  0.1086,    Adjusted R-squared:  0.09619
F-statistic: 8.734 on 9 and 645 DF,  p-value: 1.919e-12
```

- Schizophrénie : augmentation de la durrée moyenne de 3 minutes, non significatif (p=0,28)
- Dépression : augmentation de la durée moyenne de 6 minutes, significatif
- Gravité : augmentation de la durée moyenne de **1 minute par point de gravité**, significatif
- Type de centre : deux lignes, car recodée en deux variables binaires (centre 2 vs centre 3, centre 3 vs centre 1)

Mais la présentation ne permet pas de tester l'effet global de la variable « type de centre » , c'est-à-dire répondre à la question : la durée d'entretien est-elle indépendante du type de centre, à valeur égale des autres variables ?

Dans ce cas : fonction `drop1()` :

La fonction `drop1()` permet de tester l'effet global d'une variable explicative en comparant le modèle complet avec un modèle sans cette variable.

Ici : `drop1(mod, test="F")` permet de tester l'effet global de chaque variable explicative en comparant le modèle complet avec un modèle sans cette variable, en utilisant un test F (test F = test de comparaison de modèles (ANOVA)).

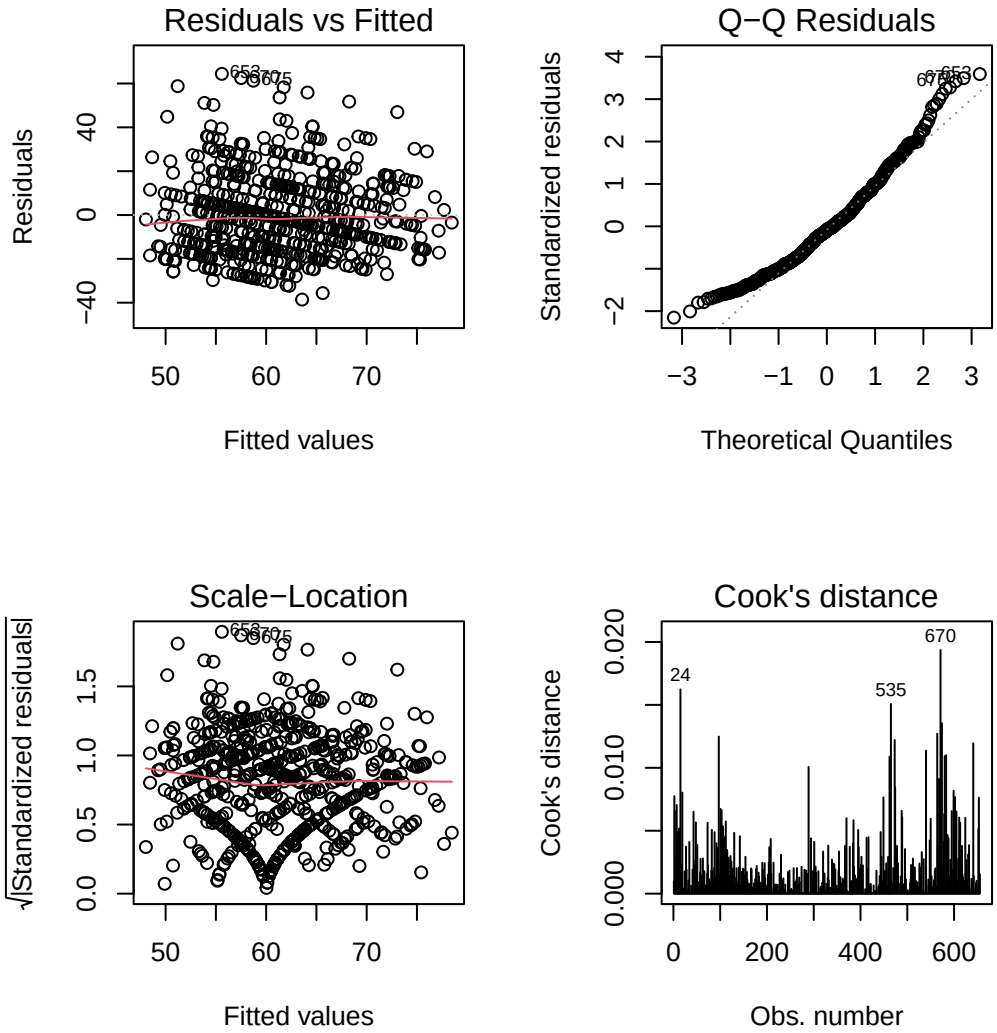
```r
drop1(mod, test="F")
```

```
Single term deletions

Model:
duree.interv ~ schizophrenie + depression + abus.subst + gravite +
    charactere + trauma.enfant + age + factor(type.centre)
                    Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                           209523 3798.0
schizophrenie        1     386.3 209910 3797.2  1.1891 0.2759114
depression           1    5390.0 214913 3812.6 16.5927  5.21e-05 ***
abus.subst           1    2125.3 211649 3802.6  6.5425 0.0107602 *
gravite              1    1146.5 210670 3799.6  3.5295 0.0607370 .
charactere           1     981.0 210504 3799.1  3.0199 0.0827234 .
trauma.enfant        1      52.0 209575 3796.2  0.1602 0.6891441
age                  1    3814.8 213338 3807.8 11.7436 0.0006493 ***
factor(type.centre)  2    3446.2 212970 3804.7  5.3044 0.0051888 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le type de centre a donc un effet global significatif (p=0,005) sur la durée de l'entretien.

3. Vérification des conditions de validité du modèle

```r
par(mfrow=c(2,2))
plot(mod, which=1:4)
```



Residuals vs Fitted : permet de vérifier que la variance des résidus est constante et globalement indépendante des valeurs prédites (homoscédasticité).

Normal Q-Q : permet de vérifier la normalité des résidus (les points doivent être proches de la diagonale).

Cook's distance : permet d'identifier les points influents (points avec une grande distance de Cook, au-dessus de la ligne pointillée).

2 Corrélation et modèle linéaire

Rappel sur le coefficient de corrélation de Pearson (r) :

- r varie entre -1 et 1
- $r = 1$: corrélation positive parfaite (lorsque X augmente, Y augmente de façon linéaire)
- $r = -1$: corrélation négative parfaite (lorsque X augmente, Y diminue de façon linéaire)
- $r = 0$: absence de corrélation linéaire entre X et Y

Donc quand 2 variables X et Y sont parfaitement corrélées (r = 1 ou -1), elles sont linéairement déterminées ($Y = a0 + a1X$).

Proximité entre corrélation et régression linéaire :

- Si X et et Y sont deux variables aléatoires, la régression linéaire de Y en fonction de X permet de prédire les valeurs de Y à partir des valeurs de X.
- Si X et Y sont linéairement liés, le modèle de régression linéaire permet de quantifier cette relation linéaire.
- La corrélation sert uniquement à dire s'il existe un lien entre deux variables et si ce lien va dans le même sens ou en sens inverse.
- La régression sert à utiliser une variable pour estimer la valeur attendue d'une autre variable.
- La dispersion d'une variable correspond au fait que ses valeurs sont plus ou moins éloignées les unes des autres.
- Dire qu'une partie de la dispersion est expliquée signifie qu'une partie des différences observées dans Y est liée aux valeurs de X.
- Le carré de la corrélation correspond à la fraction de la dispersion de Y qui est expliquée par X à travers la relation linéaire.
- Une fraction signifie simplement une partie du tout, par exemple 40 % de ce qui varie dans Y.
  - Si cette fraction est faible, alors connaître X aide très peu à estimer Y.
  - Si cette fraction est élevée, alors connaître X aide beaucoup à estimer Y.
- Même si une grande partie de la dispersion est expliquée, cela ne prouve jamais que X est la cause de Y.
- Corrélation et régression décrivent donc le même lien linéaire, mais la corrélation mesure la force du lien et la régression sert à faire des estimations chiffrées de Y. Si un exemple numérique concret doit accompagner ces phrases pour ancrer chaque notion, cela peut être ajouté.

```r
## 1. Simulation des données

set.seed(123)          # pour rendre l'exemple reproductible
n <- 100               # nombre d'observations

X <- rnorm(n, mean = 0, sd = 1)     # génération X
```

```r
bruit <- rnorm(n, mean = 0, sd = 1) # génération du bruit

Y <- 2 * X + bruit # création de Y en fonction de X avec du bruit
```

Corrélation entre X et Y :

```r
r <- cor(X, Y)
r
```

```
[1] 0.8786993
```

r = 0.8786 : forte corrélation positive entre X et Y.

3. Régression linéaire de Y sur X

Faire une régression linéaire de Y en fonction de X permet de construire une règle numérique qui permet de prédire Y à partir de X, à partir de données réelles (bruitées)

```r
mod <- lm(Y ~ X)
summary(mod)
```

```
Call:
lm(formula = Y ~ X)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.9073 -0.6835 -0.0875  0.5806  3.2904 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.10280    0.09755  -1.054    0.295    
X            1.94753    0.10688  18.222   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9707 on 98 degrees of freedom
Multiple R-squared:  0.7721,	Adjusted R-squared:  0.7698 
F-statistic:   332 on 1 and 98 DF,  p-value: < 2.2e-16
```

Affiche les coefficients, l'ordonnée à l'origine (intercept), la pente, le R², etc.
r^2 correspond à la proportion de la variance de Y expliquée par X dans le modèle linéaire, c'est à dire la fraction de la variance de Y qui est expliquée par X à travers la relation linéaire.

4. Valeurs ajustées (prédictions pour les X observés)

```r
Y_chapeau <- fitted(mod)
```

L'instruction `fitted(mod)` permet d'obtenir les valeurs prédites de Y (notées $Y_{chapeau}$) pour chaque valeur observée de X, en utilisant le modèle de régression linéaire.

5. Vérification r^2 = Var(Y_chapeau) / Var(Y)

On vérifie ici que le r^2 obtenu dans le résumé du modèle est bien égal à la variance des valeurs prédites ($Y_{chapeau}$) divisée par la variance de Y.

```r
var_Y <- var(Y)
var_Y_chapeau <- var(Y_chapeau)

r2_via_cor <- r^2
r2_via_var <- var_Y_chapeau / var_Y

r2_via_cor
```

```
[1] 0.7721124
```

```r
r2_via_var
```

```
[1] 0.7721124
```

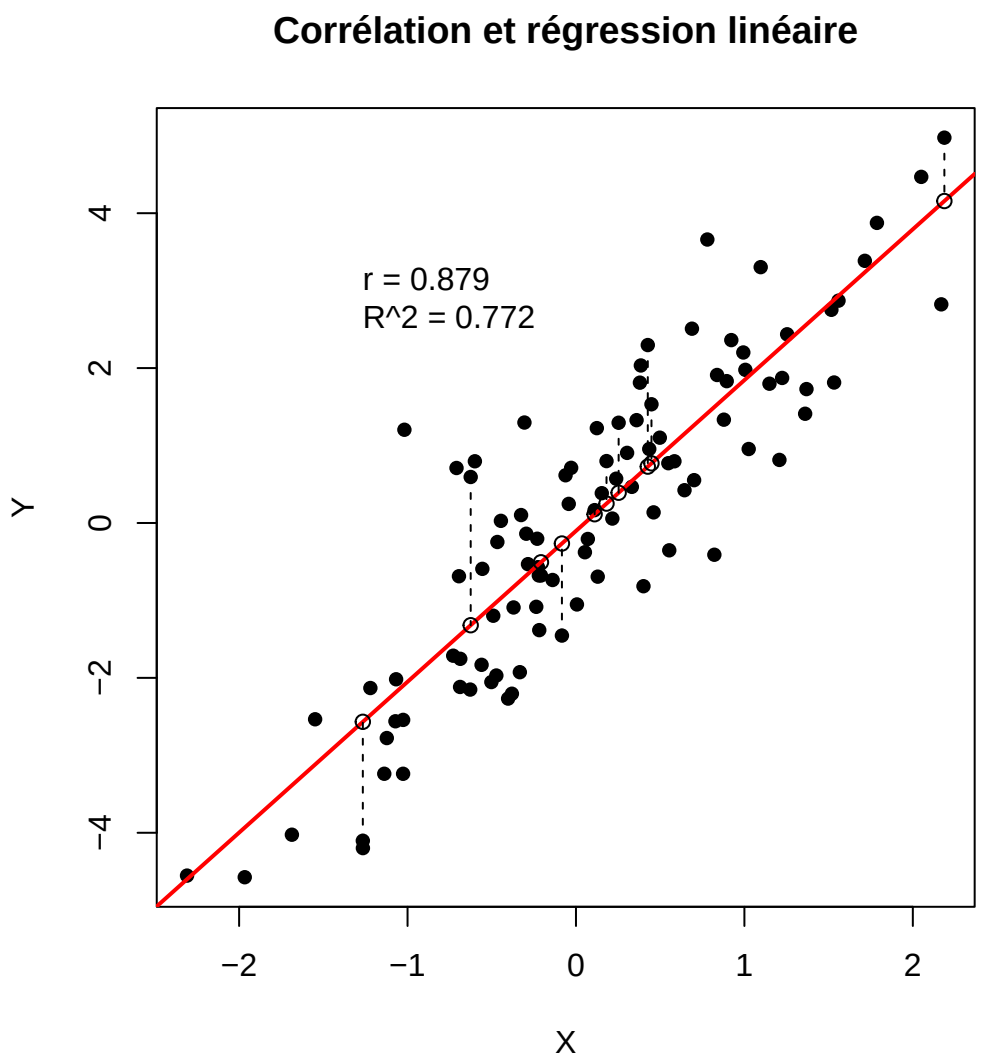
Si les deux valeurs sont égales, cela confirme que le r^2 du modèle linéaire correspond bien à la fraction de la variance de Y expliquée par X.

Cela revient à dire :

- La corrélation dit à quel point X et Y vont ensemble.
- La régression dit combien des différences de Y on peut retrouver grâce à X (quelle est la part de la variance de Y expliquée par X = quelle part des variations de Y est liée aux variations de X).
- Le pourcentage de la variance de Y expliquée par X est donné par le carré du coefficient de corrélation (r^2).
- Dans le cas d'une régression avec une seule variable X, ces deux informations sont numériquement équivalentes une fois mises au carré.

Avec une seule variable explicative X :

- r^2 = carré du coefficient de corrélation
- = part de la variabilité de Y prédite par la régression
- = % de la variance de Y expliquée par X.



> **❗ Important**
>
> **Résumé : relation entre corrélation et régression linéaire**
> La corrélation mesure à quel point deux variables X et Y varient ensemble de façon linéaire. La régression utilise cette relation linéaire pour prédire les valeurs de Y à partir des valeurs de X, dans le cas où il y a une **seule** variable explicative X.

3 Le test t : un cas particulier du modèle linéaire

Rappel : le test t permet de comparer **la moyenne** d'une variable quantitative entre deux groupes.

Dans le cadre d'un modèle linéaire, on peut considérer que le test t est un cas particulier où :

- la variable à expliquer Y est quantitative
- la variable explicative X est binaire (deux groupes)
- le modèle linéaire devient un modèle de régression linéaire avec une seule variable binaire explicative (un facteur à deux modalités).

$$Y = \alpha_0 + \alpha_1 \text{groupe}_i + \epsilon_i$$

où groupe_i est une variable binaire (0 ou 1) indiquant le groupe auquel appartient l'observation i.

Dans ce cas, le modèle linéaire permet de tester si la moyenne de Y est significativement différente entre les deux groupes.

3.A Exemple R

3.A.1 Test t

On avait vu un test t qui comprarait la moyenne d'âge des détenus selon qu'ils étaient ou non déprimés.

```r
t.test(smp$age ~ smp$depression, var.equal=TRUE)
```

```
    Two Sample t-test

data:  smp$age by smp$depression
t = 2.6337, df = 795, p-value = 0.008611
alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
95 percent confidence interval:
 0.6425038 4.4032412
sample estimates:
mean in group 0 mean in group 1
       39.93182        37.40895
```

t = 2.634 : la moyenne d'âge des détenus déprimés est significativement plus basse que celle des détenus non déprimés.

p-value = 0.008611 : la différence de moyenne est significative (p < 0,05).

3.A.2 Régression linéaire

On peut aussi faire une régression linéaire avec la variable binaire `depression` comme variable explicative :

```r
summary(lm(age ~ depression, data=smp))
```

```
Call:
lm(formula = age ~ depression, data = smp)

Residuals:
    Min      1Q  Median      3Q     Max 
-20.932 -10.932  -1.409   8.068  46.591 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  39.9318     0.6003  66.519  < 2e-16 ***
depression   -2.5229     0.9579  -2.634  0.00861 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.21 on 795 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.008649,	Adjusted R-squared:  0.007402 
F-statistic: 6.936 on 1 and 795 DF,  p-value: 0.008611
```

t value = - 2,634 et p value = 0,008611 : les résultats sont identiques à ceux du test t.

3.B Relation entre test t et régression linéaire

Le test t est donc un cas particulier de la régression linéaire où la variable explicative est binaire (un facteur à deux modalités).

Dans ce cas, le modèle linéaire permet de tester si la moyenne de la variable à expliquer est significativement différente entre les deux groupes.

Considérer le modèle llinéaire comme une généralisation du test t permet de transposer une partie des propriétés du test t à la régression linéaire :

- **Normalité des résidus** : **hypothèse faible**, pouvant être facilement compensée par un échantillon suffisamment grand
- **Indépendance des observations** : reste nécessaire
- **Homoscédasticité** : = égalité de la variance : peut être supprimée par un estimateur robuste
  - Estimateur “sandwich” : fonctionne en utilisant la matrice de variance-covariance des résidus pour ajuster les erreurs standards des coefficients du modèle.

  - Bootstrap : méthode de rééchantillonnage qui permet d'estimer la variance des coefficients du modèle en générant de multiples échantillons à partir des données originales.

- **Défaut de linéarité d'une variable quantitative** : peut poser problème, mais peut être compensé par une transformation de la variable (logarithmique, racine carrée, etc.) ou par l'utilisation de modèles non linéaires.

> **Note**
>
> **Méthode des moindres carrés**
> La régression linéaire utilise la méthode des moindres carrés pour estimer les coefficients du modèle, de manière à minimiser la somme des carrés des résidus (différences entre les valeurs observées et les valeurs prédites).
> L'objectif est de trouver les coefficients α_0 et α_1 qui minimisent la somme des carrés des erreurs (résidus) entre les valeurs observées de Y et les valeurs prédites par le modèle.

4 Introduction aux GLM

4.A Comment ça marche les GLM ?

Les modèles linéaires généralisés reposent sur 3 éléments:

1. Un prédicteur linéaire
2. Une fonction de lien
3. Une structure des erreurs

4.B Le prédicteur linéaire

« prédicteur linéaire« , c'est un terme un peu complexe pour dire, que comme dans les modèles linéaires classiques, les réponses prédites par les modèles vont l'être à partir d'une **combinaison linéaire des variables prédictives**

Le prédicteur linéaire est : η_i

$$\eta_i = \sum_{j=1}^{p} \beta_j X_{ij}$$

où :

- η_i : **prédicteur linéaire** pour l'individu i.
  - C'est la combinaison linéaire de toutes les variables explicatives.
  - Dans un GLM, **c'est la quantité que la fonction de lien va transformer** pour produire la moyenne du modèle (ex : logit, log, identité…).
- X_{ij} : **valeur de la variable explicative** j pour l'individu i.
  - Chaque individu a un ensemble de covariables (âge, sexe, exposition, etc.), notées $X_{i1}, X_{i2}, \dots, X_{ip}$.
- β_j : **coefficient associé à la variable explicative** j.
  - Il quantifie l'effet de X_{ij} sur la quantité modélisée.
  - Dans un GLM :
    * en régression linéaire : effet moyen sur Y
    * en logistique : effet sur les log-odds
    * en Poisson : effet sur le log de l'incidence
    * etc.
- p : **nombre total de variables explicatives** incluses dans le modèle.
- La somme $\sum_{j=1}^{p} \beta_j X_{ij}$

- signifie qu'on multiplie chaque covariable par son coefficient, puis qu'on additionne le tout pour obtenir le **résultat global pour l'individu** i.

En résumé :

Le prédicteur linéaire η_i est la "combinaison linéaire" de toutes les covariables.

C'est lui que le modèle va ensuite transformer (via la fonction de lien) en **probabilité**, **moyenne**, ou **taux**, selon le type de GLM.

4.C La fonction de lien

= étape délicate des GLM !

Contrairement aux modèles linéaires classiques, les valeurs prédites par le prédicteur linéaire ne correspondent pas à la prédiction moyenne d'une observation, mais à la transformation (par une fonction mathématique) de celle-ci.

En pratique, cela signifie que les valeurs du prédicteur linéaire sont obtenues en transformant préalablement les valeurs observées par la fonction de lien.

Autrement dit, les beta sont estimés après transformation des réponses selon la fonction de lien choisie.

Le prédicteur linéaire et la fonction de lien sont ainsi liés par une équation qui contraint les valeurs prédites par le modèle à être dans l'échelle des valeurs observées.

Les formules sont complexes et incompréhnsibles mais globalement :

4.C.1 Tableau synthétique des modèles linéaires généralisés (GLM)

Type de réponse	Domaine des valeurs possibles	Distribution des erreurs	Fonction de lien	Comment la moyenne est obtenue	Fonction de la variance
Quantitative continue	Toutes les valeurs réelles (négatives ou positives)	Gaussienne (normale) = modèle linéaire classique	Identité : combinaison linéaire des variables explicatives = **la moyenne des résultats**	La moyenne est la somme des effets des variables explicatives Donc effet **additif**	La variance est constante = **homoscédasticité**
Comptage (0,1,2,3,...)	Nombres entiers positifs	Poisson	Logarithme : transforme la moyenne avec un logarithme pour pouvoir modéliser des valeurs qui ne peuvent être que positives.	La moyenne est obtenue en appliquant l'exponentielle à la combinaison linéaire des variables. Donc effet **multiplicatif**	La variance est proportionnelle à la moyenne Plus la moyenne est grande, plus la variance l'est aussi
Binaire (oui/non)	0 ou 1	Binomiale	Logit	La moyenne (probabilité) est obtenue en appliquant une fonction « en S » à la combinaison linéaire des variables Fonction **sigmoïde**	La variance dépend à la fois de la probabilité et de sa complémentaire (1 - probabilité) Si probabilité est faible ou élevée : faible variance Si probabilité 50% : variance maximale

En gros : dans les GLM, les données sont d'abord transformées et que cette transformation permet ensuite aux prédictions d'avoir des contraintes identiques aux réponses observées (par exemple, d'être toujours positives ou nulles), autrement dit de fournir des prédictions cohérentes !

4.D La structure d'erreur

A une fonction de lien donnée, correspond généralement une structure d'erreur particulière.

Il s'agit d'une famille de distribution des erreurs.

Par exemple, pour les données de comptage, la fonction de lien est le log et la structure d'erreur correspondante est la distribution de Poisson.

Cette structure d'erreur, permet notamment de spécifier correctement la relation entre la moyenne et la variance.

Cette relation est utilisée par l'approche de maximum de vraisemblance pour estimer les coefficients des paramètres (les beta) du GLM.

4.E Maximum de vraisemblance et déviance

Les coefficients des paramètres d'un GLM sont estimés par la méthode du maximum de vraisemblance, qui fait appelle à la notion de déviance.

La déviance est en quelque sorte une généralisation de la variance.

Elle mesure l'écart entre le modèle ajusté et le modèle saturé (modèle qui s'ajuste parfaitement aux données).

L'objectif est de minimiser la déviance, c'est-à-dire de trouver les coefficients des paramètres qui rendent le modèle aussi proche que possible des données observées.

5 Modèle logistique

5.A Modèle linéaire inadapté pour cas témoins

En étude cas-témoins :

- on choisit un certain nombre de cas (Y=1) et un certain nombre de témoins (Y=0) de manière artificielle,
- donc la proportion de cas dans les données ne reflète pas le vrai risque dans la population.

On ne peut donc pas interpréter :

- ni la moyenne de Y comme un risque,
- ni la différence de moyennes comme une différence absolue de risque.

Or le modèle linéaire sur Y (0/1) raisonne justement en termes de moyennes et de différences absolues, donc ça ne marche pas en cas-témoins.

En revanche, en cas-témoins, l'odds-ratio reste interprétable d'où l'intérêt de la logistique.

5.B Principe et mise en œuvre

Objectif du modèle logistique : modéliser une variable binaire (0 ou 1) en fonction de variables explicatives quantitatives ou qualitatives.

Modèle logistique = régression logistique.

Équation du modèle logistique :

$$Y = \text{log}\left(\frac{p}{1-p}\right) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p + \epsilon$$

Il est très peu probable que ε suive une loi normale :

- Y est binaire (0 ou 1), donc la variance de Y n'est pas constante (elle dépend de la valeur de p)
- Les valeurs de à droites varient entre $-\infty$ et $+\infty$

NB : les résidus se calculent $\varepsilon = Y - (\alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p)$.

En fait l'équation n'a pas de sens !

- À gauche : Y = 0 ou 1
- À droite : une combinaison linéaire + une erreur ε ça peut être n'importe quel réel, positif ou négatif.
- Un modèle linéaire peut prédire des valeurs < 0 ou > 1, or on veut prédire quelque chose qui ressemble à une probabilité.
- Il n'y a ni log-linéarité, ni proportionnalité des risques.

L'objectif est donc de modéliser la "probabilité" que Y = 1 en fonction des variables explicatives $X_1, X_2, ..., X_p$.

C'est à dire $P(Y = 1|X_1, X_2, ..., X_p)$. (pour probabilité de Y sachant X1, X2, …, Xp)

5.C Transformation de Y

5.C.1 Odds

L'objectif est de modéliser la probabilité que Y = 1 en fonction des variables explicatives $X_1, X_2, ..., X_p$.

Donc on cherche à modéliser $p = P(Y = 1|X_1, X_2, ..., X_p)$.

- p est compris entre 0 et 1.
- on cherche en fait à calculer l'odds = la cote = la probabilité que Y = 1 / probabilité que Y = 0 = $p/(1-p)$.

Les odds varient entre 0 et +∞. Exemple :

- si p = 0.5 odds = 0.5 / 0.5 = 1
- si p = 0.8 odds = 0.8 / 0.2 = 4

5.C.2 Log-odds

On applique la transformation logarithmique aux odds afin d'obtenir une variable qui varie entre -∞ et +∞ et de la rendre compatible avec une régression linéaire.

Attention, on supprime ε car :

- on ne peut pas avoir de terme d'erreur dans une régression logistique
- en fait, comme on modélise la probabilité que $Y = 1$ à partir des données observées, l'erreur est "contenue" dans les données elles-mêmes.

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \alpha_0 + \alpha_1 X_1 + ... + \alpha_p X_p$$

5.D Maximum de vraisemblance

L'estimation des coefficients du modèle logistique se fait par la méthode du maximum de vraisemblance.

Vraisemblance = probabilité d'observer les données réelles, en fonction des paramètres du modèle.

L'idée est de trouver les coefficients $\alpha_0, \alpha_1, ..., \alpha_p$ qui maximisent la probabilité d'observer les données réelles, en supposant que le modèle est correct.

Quand on dit "maximiser la log-vraisemblance", c'est équivalent à maximiser la vraisemblance, mais c'est plus facile à calculer avec des logarithmes.

L'utilisation des logarithmes facilite les calculs car :

- la log-vraisemblance transforme les produits en sommes, ce qui est plus simple à manipuler mathématiquement.
- elle permet d'éviter des problèmes numériques liés à la manipulation de très petites probabilités.

- Elle est souvent utilisée en statistique pour simplifier les calculs d'optimisation.

5.E Conditions de validité

La condition de validité la plus importante : au moins 10 évènements par variable explicative dans le modèle (10 événements = 10 cas où Y=1).

Cela permet d'assurer que les estimations des coefficients du modèle sont stables et fiables.

5.F Exemple R

La fonction `glm()` permet de faire une régression logistique en R.

Objectif : modéliser l'existence d'un risque suicidaire élevé à l'aide des variables :

- abus dans l'enfance (oui/non),
- procédure disciplinaire pendant l'incarcération (oui/non),
- durée de la peine (<1 mois, 1-6 mois, 6-12 mois, 1-5 ans, >5 ans),
- âge (continue)
- type de prison (« 1 » pour maison centrale, « 2 » pour centre de détention et « 3 » pour maison d'arrêt).

5.F.1 Description des variables

```r
describe(
    smp[, c("hr.suicide","abus.enfant","discipline", "duree.peine","age")],
    num.desc=c("mean","sd","median", "min","max","valid.n"))
```

```
Description of smp[, c("hr.suicide", "abus.enfant", "discipline", "duree.peine",      "age")]


 Numeric
             mean    sd median min max valid.n
hr.suicide   0.20  0.40      0   0   1     760
abus.enfant  0.28  0.45      0   0   1     792
discipline   0.23  0.42      0   0   1     793
duree.peine  4.32  0.85      5   1   5     575
age         38.94 13.26     37  19  84     797
```

5.F.2 Fonction glm

```r
mod <- glm(
    hr.suicide ~
```

```r
        abus.enfant + discipline + duree.peine + age + factor(type.centre),
    data=smp,
    family = "binomial")
summary(mod)
```

```
Call:
glm(formula = hr.suicide ~ abus.enfant + discipline + duree.peine +
    age + factor(type.centre), family = "binomial", data = smp)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.329750   0.787283  -0.419  0.67533
abus.enfant           0.633421   0.228495   2.772  0.00557 **
discipline            0.454877   0.254880   1.785  0.07431 .
duree.peine          -0.295252   0.149316  -1.977  0.04800 *
age                  -0.005002   0.009468  -0.528  0.59729
factor(type.centre)2 -0.045704   0.333694  -0.137  0.89106
factor(type.centre)3  0.273555   0.371157   0.737  0.46110
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 555.49  on 548  degrees of freedom
Residual deviance: 531.28  on 542  degrees of freedom
  (250 observations deleted due to missingness)
AIC: 545.28

Number of Fisher Scoring iterations: 4
```

```r
exp(coefficients(mod))
```

```
         (Intercept)          abus.enfant           discipline
           0.7191038            1.8840458            1.5759788
         duree.peine                  age factor(type.centre)2
           0.7443441            0.9950108            0.9553243
factor(type.centre)3
           1.3146295
```

```r
exp(confint(mod))
```

```
Waiting for profiling to be done...

                          2.5 %    97.5 %
```

```
(Intercept)          0.1526791 3.3673241
abus.enfant          1.1996784 2.9433962
discipline           0.9514471 2.5901545
duree.peine          0.5549039 0.9978941
age                  0.9764391 1.0134442
factor(type.centre)2 0.5042208 1.8782340
factor(type.centre)3 0.6406323 2.7632295
```

Pour afficher tous les résultats en un seul tableau facilement exportable en latex (tableau gtsummary)

```
res_table1 <- tbl_regression(
    mod,
    exponentiate = TRUE
    ) |>
    modify_table_styling(
        columns = estimate,
        footnote = "Estimates are odds ratios from a logistic regression model."
    )

res_table1
```

Characteristic	OR[1]	95% CI	p-value
abus.enfant	1.88	1.20, 2.94	0.006
discipline	1.58	0.95, 2.59	0.074
duree.peine	0.74	0.55, 1.00	0.048
age	1.00	0.98, 1.01	0.6
factor(type.centre)			
1	—	—	
2	0.96	0.50, 1.88	0.9
3	1.31	0.64, 2.76	0.5

[1]Estimates are odds ratios from a logistic regression model.
Abbreviations: CI = Confidence Interval, OR = Odds Ratio

5.F.3 Tester l'absence d'effet d'une variable

Fonction `drop1` permet de tester si "globalement", l'effet d'une variable catégorielle a un effet sur `hr.suicide`

La fonction `drop1` permet de "supprimer" la variable du modèle.

```
drop1(mod,.~.,test="Chisq")
```

```
Single term deletions
```

```
Model:
hr.suicide ~ abus.enfant + discipline + duree.peine + age + factor(type.centre)
                    Df Deviance    AIC    LRT Pr(>Chi)
<none>                   531.28 545.28
abus.enfant          1   538.78 550.78 7.5010 0.006166 **
discipline           1   534.41 546.41 3.1287 0.076924 .
duree.peine          1   535.18 547.18 3.8968 0.048379 *
age                  1   531.56 543.56 0.2810 0.596049
factor(type.centre)  2   532.72 542.72 1.4368 0.487522
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On utilise ici un test du Chi2 car on utilise le Maximum de Vraisemblance, qui suit une loi du Chi2 asymptotiquement.

Pour un modèle linéaire, on utilisait un test F car on utilisait la méthode des moindres carrés, qui suit une loi F (pour Fisher) asymptotiquement.

> **Note**
>
> **Différence entre test F et test du Chi2 dans le contexte des modèles linéaires et logistiques**
>
> - Test F : utilisé dans les modèles linéaires (régression linéaire) basés sur la méthode des moindres carrés. Il compare la variance expliquée par le modèle à la variance résiduelle pour évaluer la signification des variables explicatives.
> - Test du Chi2 : utilisé dans les modèles logistiques basés sur le maximum de vraisemblance. Il compare la log-vraisemblance du modèle complet avec celle d'un modèle réduit (sans la variable testée) pour évaluer l'effet global de la variable explicative.

6 Modèle log-binomial

Problème du modèle logistique : les odds-ratios sont ajustés et donc difficiles à interpréter.

Comment traduire un OR en truc du quotidien (si la maladie est fréquente) ?

Solution : modèle log-binomial.

6.A Principe et mise en œuvre

Objectif du modèle log-binomial : modéliser une variable binaire (0 ou 1) en fonction de variables explicatives quantitatives ou qualitatives, **en estimant directement les risques relatifs (RR) au lieu des odds-ratios (OR).**

Modèle log-binomial = régression log-binomial.

Équation du modèle log-binomial :

$$\log[p(Y=1)] = a_0 + a_1 X_1 + ... + a_p X_p$$

où $p(Y=1)$ est la probabilité que Y = 1 (risque).

Problème mathématique :

- à gauche : $\log[p(Y=1)]$ varie entre $-\infty$ et 0 (car $p(Y=1)$ varie entre 0 et 1) donc $\log[p(Y=1)]$ ne peut pas prendre toutes les valeurs réelles.
- à droite : une combinaison linéaire de variables explicatives qui peut prendre toutes les valeurs réelles (entre $-\infty$ et $+\infty$).

On a donc un problème de convergence (on dit que l'algorithme converge quand il trouve une solution stable) :

- une probabilité doit être ≤ 1,
- mais $\exp(\cdot)$ peut dépasser 1 très facilement.

Le modèle log-binomial fonctionne donc avec des contraintes :

- $\exp(a_0 + a_1 X_1 + ...) \leq 1$: le modèle interdit des combinaisons de variables explicatives qui donneraient des probabilités > 1.
- donc les coefficients sont fortement contraints à se rapprocher de 0.

6.B Exemple R

On utilise la fonction `logbin()` du package `logbin`.

Objectif : idem que précédemment, modéliser l'existence d'un risque suicidaire élevé.

```
mod2 <- logbin(
    hr.suicide ~
        abus.enfant + discipline + duree.peine + age + factor(type.centre),
    data = smp
```

```r
)
summary(mod2)
```

```
Call:
logbin(formula = hr.suicide ~ abus.enfant + discipline + duree.peine +
    age + factor(type.centre), data = smp)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4029  -0.6821  -0.5819  -0.4928   2.0684

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.928551   0.585717  -1.585  0.11289
abus.enfant           0.480690   0.167112   2.876  0.00402 **
discipline            0.275071   0.187525   1.467  0.14242
duree.peine          -0.191619   0.107833  -1.777  0.07557 .
age                  -0.004893   0.007487  -0.654  0.51343
factor(type.centre)2  0.011596   0.270326   0.043  0.96578
factor(type.centre)3  0.224932   0.294407   0.764  0.44486
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

    Null deviance: 555.49  on 548  degrees of freedom
Residual deviance: 532.19  on 542  degrees of freedom
  (250 observations deleted due to missingness)

  AIC: 546.19
AIC_c: 546.40

Number of iterations: 59719 (best: 2929)
```

```r
exp(coefficients(mod2))
```

```
         (Intercept)          abus.enfant           discipline
           0.3951260            1.6171894            1.3166236
         duree.peine                  age factor(type.centre)2
           0.8256215            0.9951193            1.0116637
factor(type.centre)3
           1.2522375
```

```r
exp(confint(mod2))
```

```
                          2.5 %   97.5 %
```

```
(Intercept)          0.1253640 1.245370
abus.enfant          1.1655076 2.243916
discipline           0.9116757 1.901441
duree.peine          0.6683343 1.019925
age                  0.9806236 1.009829
factor(type.centre)2 0.5955717 1.718455
factor(type.centre)3 0.7032138 2.229903
```

Et représentation en tableau :

```
res_table2 <- tbl_regression(mod2, exponentiate = TRUE) |>
    modify_table_styling(
        columns = estimate,
        footnote = "Estimates are risk ratios from a log-binomial model."
    )
```

```
Warning: The `tidy()` method for objects of class `logbin` is not maintained by the broom team,

This warning is displayed once per session.
```

```
res_table2
```

Characteristic	RR[1]	95% CI	p-value
abus.enfant	1.62	1.17, 2.24	0.004
discipline	1.32	0.91, 1.90	0.14
duree.peine	0.83	0.67, 1.02	0.076
age	1.00	0.98, 1.01	0.5
factor(type.centre)			
1	—	—	
2	1.01	0.60, 1.72	>0.9
3	1.25	0.70, 2.23	0.4

[1]Estimates are risk ratios from a log-binomial model.
Abbreviations: CI = Confidence Interval, RR = Relative Risk

6.C Problème du modèle log-binomial

6.D Limitations du modèle log-binomial et alternative pratique

Le modèle **log-binomial**, utilisé pour estimer directement un **risque relatif (RR)** ajusté lorsque la maladie est fréquente, présente plusieurs limites majeures :

- méthodes d'estimation instables lorsque la prévalence est élevée ou que de nombreuses covariables sont incluses ;

- problèmes fréquents de **convergence**, car les probabilités prédites doivent rester ≤ 1 alors que la forme exponentielle peut facilement dépasser cette limite ;
- les coefficients sont contraints vers 0, ce qui produit des **risques relatifs artificiellement proches de 1** ;
- le modèle estime des **RR conditionnels**, difficiles à interpréter en présence de nombreuses covariables ou de covariables corrélées ;
- pour des maladies fréquentes, l'hypothèse d'absence d'interaction est souvent intenable, menant soit à des probabilités > 100 %, soit à un aplatissement général des RR.

6.E Alternative : calculer un RR marginal à partir d'un modèle logistique

Une approche robuste consiste à :

1. ajuster une **régression logistique** ;
2. créer deux jeux de données contrefactuels :
   - un où l'exposition vaut 0 pour tous les sujets ;
   - un où l'exposition vaut 1 pour tous les sujets ;
3. prédire les probabilités dans chaque scénario ;
4. calculer :
   - le risque moyen non exposé : p_0 ;
   - le risque moyen exposé : p_1 ;
   - le risque relatif marginal : $RR = \dfrac{p_1}{p_0}$;
   - l'odds-ratio marginal : $OR = \dfrac{p_1/(1-p_1)}{p_0/(1-p_0)}$;
   - la différence absolue de risque : $DAR = p_1 - p_0$.

> **Note**
>
> **Régression logistique selon Firth**
> Firth modifie la régression logistique classique en ajoutant au critère à optimiser un terme supplémentaire qui devient très pénalisant quand un coefficient essaie de partir vers des valeurs extrêmes.
> Concrètement, ce terme repose sur la « courbure » de la vraisemblance : plus les données poussent un coefficient à devenir énorme (cas de séparation ou quasi-séparation), plus ce terme augmente et force la solution à rester finie.
> Mathématiquement, c'est exactement ce qu'on obtiendrait si on imposait une loi a priori très spécifique (la prior de Jeffreys) sur les coefficients, ce qui revient à ajouter une petite quantité d'« information artificielle » pour casser la séparation parfaite.
> Cette correction est construite pour compenser l'erreur systématique des estimateurs classiques, de sorte que, en situation de petits effectifs ou d'événement rare, les coefficients issus de Firth sont moins systématiquement trop grands en valeur absolue.

7 Modèle logistique pour *odds* proportionnels

Ces modèles sont adaptés quand la variable à expliquer est qualitative ordonnée. p151 du pdf

8 Modèle de Poisson et binomial négatif pour taux d'incidence

Variable de type "compte" (= entier positif)

= nombre d'occurences d'un évènement dans un intervalle de temps donné (= nombre de buts marqués, d'œufs pondus…)

Les modèles de Poisson et binomial négatif permettent de modéliser une variable de type "compte" en fonction de variables explicatives quantitatives ou qualitatives.

Définition :

- Ce sont des GLM (= modèles linéaires généralisés)
- Fonction lien log = logarithme naturel
- Structure d'erreur de type Poisson.

8.A Pourquoi les modèles linéaires classiques ne sont pas adaptés

Données de compte : ne remplissent pas les conditions de validité des modèles linéaires classiques.

- Ne suivent pas une loi normale (mais une loi de Poisson)
- Leur variance n'est pas constante (mais proportionnelle à la moyenne)

8.B La distribution de Poisson

La distribution de Poisson est une distribution de probabilité discrète qui décrit le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe, lorsque ces événements se produisent avec une moyenne constante et indépendamment les uns des autres.

Équation :

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Avec :

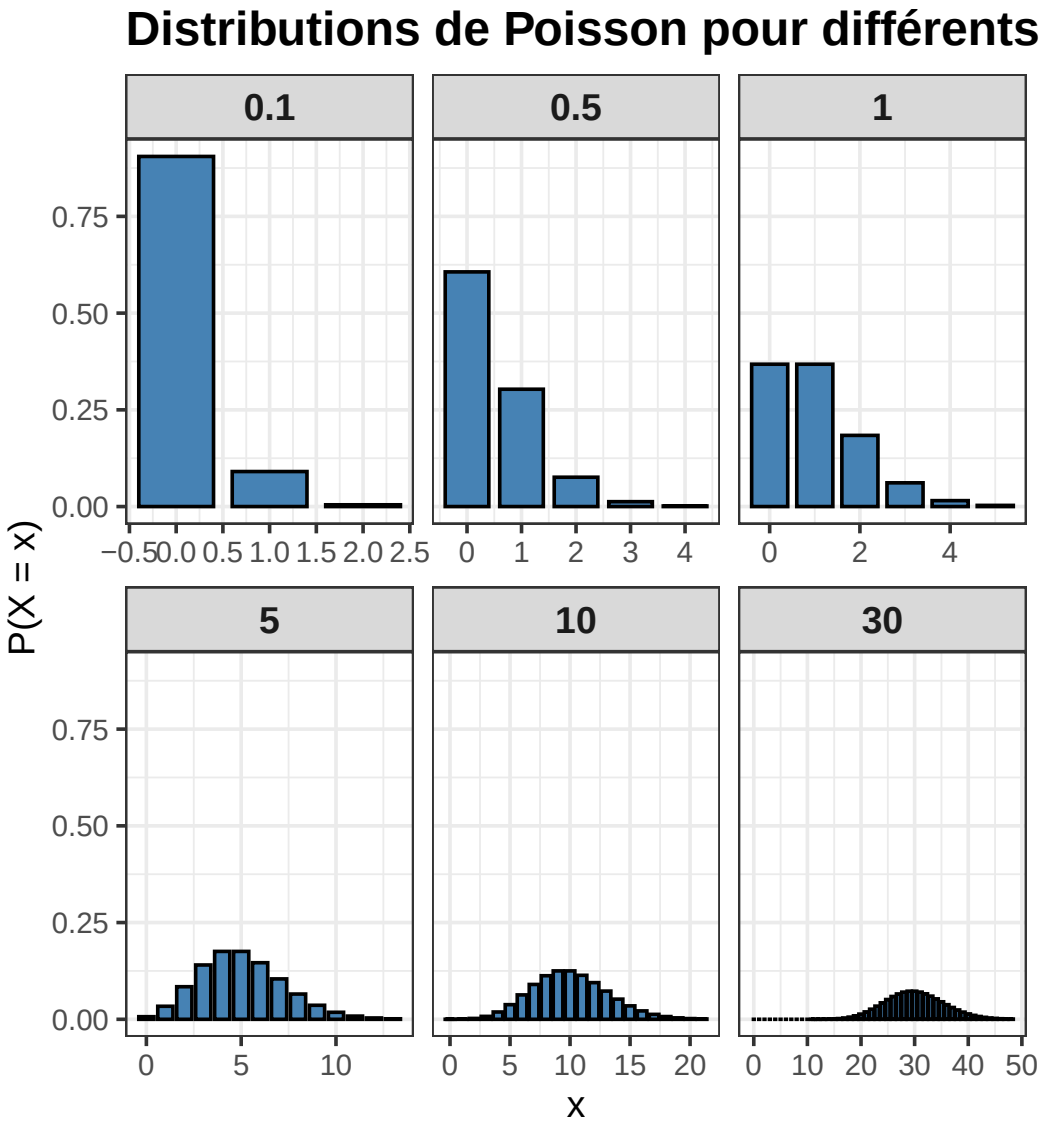
- $P(X = k)$: probabilité d'observer exactement k événements dans l'intervalle
- λ : moyenne (et variance) du nombre d'événements dans l'intervalle
- e : base du logarithme naturel (environ 2,71828)
- $k!$: factorielle de k (produit des entiers de 1 à k)

Donc la distribution de Poisson est ainsi définie par un seul paramètre : λ !

Exemples de distributions de poisson avec différentes valeurs de λ :

Plus λ augmente, plus la distribution de Poisson se rapproche d'une loi Normale.

On dit qu'une loi de Poisson peur être approximée par une loi Normale quand λ = égal 20 ou 30



La distribution de Poisson possède deux éléments remarquables :

- L'espérance (ou moyenne) d'une variable aléatoire distribuée selon une loi de poisson est égale à Lambda :$E(y) = \lambda$
- La variance d'une variable aléatoire distribuée selon une loi de poisson est aussi égale à Lambda : $Var(y) = \lambda$

8.C Caractéristique du GLM de Poisson

1. Prédicteur linéaire : combinaison linéaire des variables explicatives

- $\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$
- où η_i est le prédicteur linéaire pour l'observation i.

2. Fonction de lien dans le modèle de Poisson : logarithme naturel

- log$(\mu_i) = \eta_i$ = prédicteur linéaire

En gros : les valeurs prédites par le prédicteur linéaire du GLM ne correspondent pas à la prédiction moyenne d'une observation, **mais à la transformation log de celle-ci**.

Pour obtenir la prédiction moyenne, on applique l'exponentielle au prédicteur linéaire :

- $\mu_i = \exp(\eta_i)$

3. Structure d'erreur : distribution de Poisson

- La variable à expliquer suit une loi de Poisson.
- C'est à dire que son espérance (sa moyenne) et sa variance sont égales à λ.

8.D Conditions de validité

1. les réponses sont indépendantes.
2. les réponses sont distribuées selon une loi de Poisson, de paramètre Lambda.
3. il n'existe pas de surdispersion

8.D.1 Indépendance des réponses

- Pas de structures de corrélation entre les données
- Par exemple pas de données répétées
- Si données répétées : utiliser un modèle linéaire généralisé à effet mixte (Generalized Linear Mixed Models, ou GLMM).

8.D.2 Distribution des réponses

- Généralement supposée
- On peut comparer la distribution des données *vs* une distribution théorique de Poisson

8.D.3 Absence de surdispersion

- Selon la loi de Poisson, la variance des réponses est égale à la moyenne des réponses.
- Surdispersion = variance réelle > variance théorique.
  - Dans ce cas : risque de sous-estimation de l'erreur standard des paramètres du modèle.
  - et donc de p-value excessivement faible.

8.D.3.1 Comment mesurer la surdispersion On calcule un paramètre appelé ϕ :

$$\phi = \frac{\text{variance observee}}{\text{variance theorique}} = \frac{\text{variance observee}}{\text{moyenne}}.$$

En pratique, comme la variance théorique d'un modèle de Poisson vaut la moyenne, ϕ indique directement à quel point les données sont plus dispersées que ce que le modèle prévoit.

Dans un modèle ajusté, on n'a pas accès à la vraie variance, donc ϕ est estimé par :

$$\hat{\phi} = \frac{\text{deviance residuelle}}{\text{nombre de degres de liberte}}.$$

Avec la déviance résiduelle = mesure de l'écart entre le modèle ajusté et les données observées.

- Si $\hat{\phi} \approx 1$: les données sont compatibles avec la loi de Poisson.
- Si $\hat{\phi} > 1$: il y a surdispersion.
- Il n'existe **pas de seuil universel** (1.5 ? 2 ?) : le seuil dépend aussi de la taille des données.

Certains packages (comme `AER`) proposent un test dédié pour aider à prendre la décision.

8.D.3.2 Causes fréquentes de surdispersion La surdispersion n'est pas un "bug", mais un symptôme d'un problème réel dans les données :

- corrélation entre les réponses (par exemple des mesures répétées),
- variable explicative importante manquante,
- excès de zéros par rapport à ce que prévoit une loi de Poisson.

8.D.3.3 Que faire en cas de surdispersion ? Si $\hat{\phi}$ est clairement supérieur à 1, il faut changer le modèle, car la régression de Poisson n'est plus adaptée.

Deux alternatives classiques :

- **quasi-Poisson** :
  - même structure que Poisson, mais avec variance ajustée ;
  - lève la contrainte d'égalité entre moyenne et variance.
- **binomiale négative** :
  - variance plus flexible, souvent adaptée aux données avec beaucoup de dispersion.
  - modèle plus complexe, mais souvent plus robuste.

Ces modèles corrigent l'erreur standard en la multipliant par :

$$\sqrt{\hat{\phi}}.$$

Ce qui remet les p-values à des niveaux plus réalistes.

8.E Exemple R

Modèle expliquant le nombre d'antécédents d'incarcération des détenus du df `smp`.

- Données de comptage d'événements non-indépendants car survenant chez un même sujet !

- Modèle de Poisson : inapproprié car adapté au comptage d'événements indépendant est manifestement inapproprié,
- Modèles quasi-Poisson ou binomial négatifs restent possibles car ne reposent pas sur l'hypothèse d'indépendance.

Pour mesurer le degré d'invalidité du modèle de Poisson, on compare la moyenne et la variance du nombre d'antécédents d'incarcération.

```
mean(smp$n.prison, na.rm=TRUE)
```

```
[1] 1.839949
```

```
var(smp$n.prison, na.rm=TRUE)
```

```
[1] 11.94743
```

```
var(smp$n.prison, na.rm=TRUE) / mean(smp$n.prison, na.rm=TRUE)
```

```
[1] 6.493348
```

8.E.1 Modèle de Poisson

Essayons de calculer un modèle de Poisson malgré tout incluant :

- `charactere` (intensité d'un trouble de la personnalité),
- `recherche.nouv` (dimension de recherche de la nouveauté),
- `famille.prison` (antécédents familiaux d'incarcération),
- `abus.enfant` (antécédents d'abus pendant l'enfance),
- `age` (l'âge)
- et `type.centre` (le type d'établissement pénitentiaire).

```
mod_pois <- glm(
    n.prison ~ charactere + recherche.nouv + famille.prison + abus.enfant + age +
↪   factor(type.centre),
    data = smp,
    family = poisson()
)
summary(mod_pois)
```

```
Call:
glm(formula = n.prison ~ charactere + recherche.nouv + famille.prison +
    abus.enfant + age + factor(type.centre), family = poisson(),
```

```
    data = smp)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.232314   0.178078  -6.920 4.51e-12 ***
charactere            0.300950   0.029450  10.219  < 2e-16 ***
recherche.nouv        0.236566   0.037707   6.274 3.52e-10 ***
famille.prison        0.536230   0.059981   8.940  < 2e-16 ***
abus.enfant           0.402111   0.059269   6.785 1.16e-11 ***
age                   0.006819   0.002381   2.864  0.00419 **
factor(type.centre)2  0.322015   0.111315   2.893  0.00382 **
factor(type.centre)3  0.276732   0.109034   2.538  0.01115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2834.4  on 633  degrees of freedom
Residual deviance: 2360.4  on 626  degrees of freedom
  (165 observations deleted due to missingness)
AIC: 3289.6

Number of Fisher Scoring iterations: 6
```

Affichage des résultats en tableau :

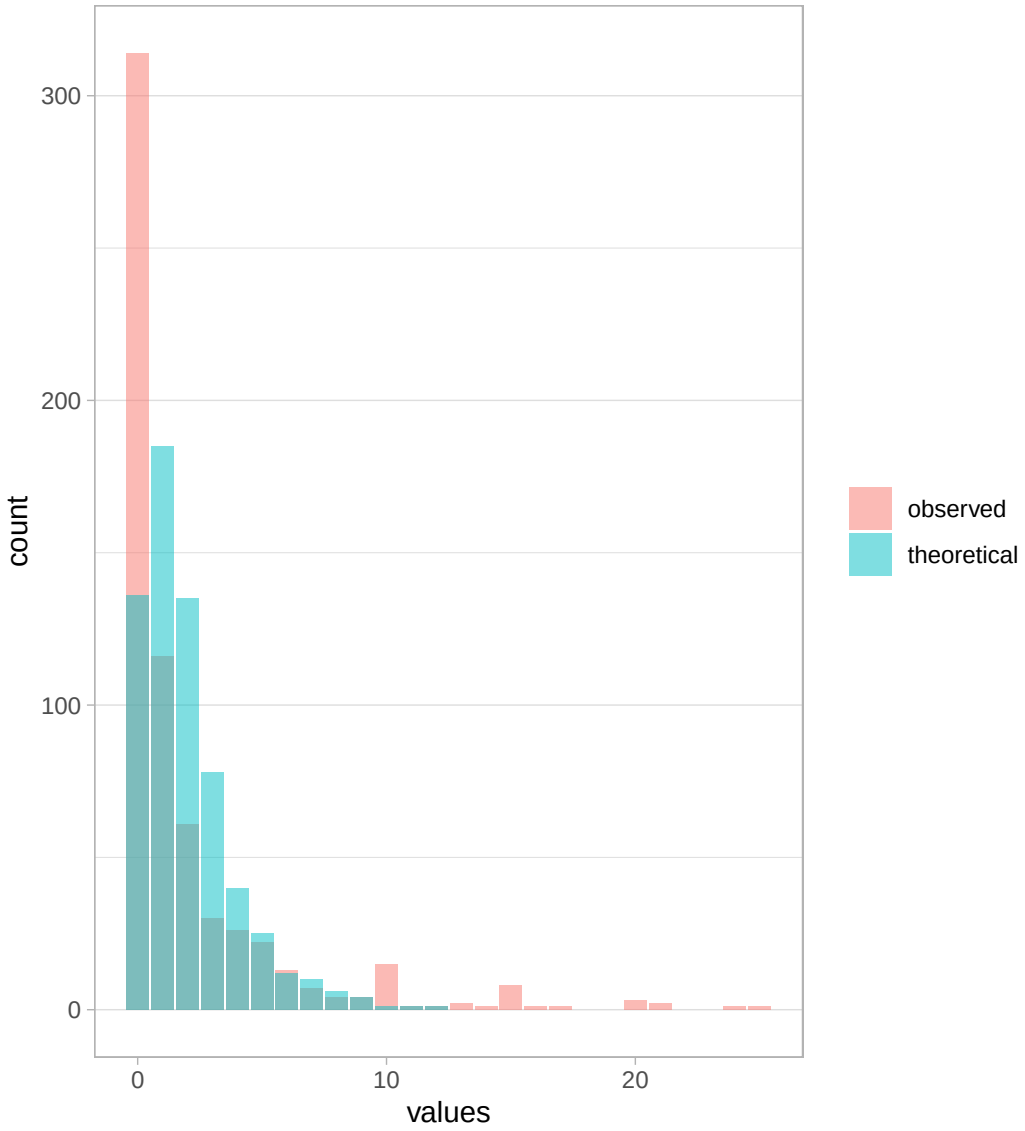
```r
res_table_pois <- tbl_regression(
    mod_pois,
    exponentiate = TRUE
    ) |>
    modify_table_styling(
        columns = estimate,
        footnote = "Estimates are incidence rate ratios from a Poisson regression
        ↪  model."
    )
res_table_pois
```

Characteristic	IRR[1]	95% CI	p-value
charactere	1.35	1.28, 1.43	<0.001
recherche.nouv	1.27	1.18, 1.36	<0.001
famille.prison	1.71	1.52, 1.92	<0.001
abus.enfant	1.49	1.33, 1.68	<0.001
age	1.01	1.00, 1.01	0.004
factor(type.centre)			
1	—	—	
2	1.38	1.11, 1.73	0.004
3	1.32	1.07, 1.64	0.011

[1]Estimates are incidence rate ratios from a Poisson regression model.
Abbreviations: CI = Confidence Interval, IRR = Incidence Rate Ratio

Le package `guideR` de Larmarange propose une fonction `guideR::observed_vs_theoretical()` qui permet justement de comparer la distribution observée avec la distribution théorique d'un modèle.

```
mod_pois |>
    guideR::observed_vs_theoretical()
```



Il existe même une fonction automatique évaluent la surdispersion dans le package `DHARMa`, qui fait un test de surdispersion.

```
mod_pois |>
    performance::check_overdispersion()
```

```
# Overdispersion test

       dispersion ratio =    5.512
  Pearson's Chi-Squared = 3450.403
                p-value =  < 0.001

Overdispersion detected.
```

8.E.2 Modèle quasi-Poisson

```r
mod_quasi <- glm(
    n.prison ~ charactere + recherche.nouv + famille.prison + abus.enfant + age +
↪   factor(type.centre),
    data = smp,
    family = quasipoisson()
)
summary(mod_quasi)
```

```
Call:
glm(formula = n.prison ~ charactere + recherche.nouv + famille.prison +
    abus.enfant + age + factor(type.centre), family = quasipoisson(),
    data = smp)

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          -1.232314   0.418080  -2.948 0.003322 **
charactere            0.300950   0.069140   4.353 1.57e-05 ***
recherche.nouv        0.236566   0.088526   2.672 0.007730 **
famille.prison        0.536230   0.140820   3.808 0.000154 ***
abus.enfant           0.402111   0.139148   2.890 0.003988 **
age                   0.006819   0.005590   1.220 0.223032
factor(type.centre)2  0.322015   0.261338   1.232 0.218345
factor(type.centre)3  0.276732   0.255984   1.081 0.280090
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 5.511868)

    Null deviance: 2834.4  on 633  degrees of freedom
Residual deviance: 2360.4  on 626  degrees of freedom
  (165 observations deleted due to missingness)
AIC: NA

Number of Fisher Scoring iterations: 6
```

Affichage des résultats en tableau :

Characteristic	IRR[1]	95% CI	p-value
charactere	1.35	1.18, 1.55	<0.001
recherche.nouv	1.27	1.07, 1.51	0.008
famille.prison	1.71	1.30, 2.25	<0.001
abus.enfant	1.49	1.14, 1.96	0.004
age	1.01	1.00, 1.02	0.2
factor(type.centre)			
1	—	—	
2	1.38	0.85, 2.38	0.2
3	1.32	0.82, 2.25	0.3

[1]Estimates are incidence rate ratios from a quasi-Poisson regression model.
Abbreviations: CI = Confidence Interval, IRR = Incidence Rate Ratio

Comparaison avec un modèle théorique de quasi-Poisson : pas possible avec `guideR::observed_vs_theoretical` car ne supporte pas le quasi-Poisson.

On peut quand même vérifier la surdispersion avec `performance::check_overdispersion()` :

```
mod_quasi |>
    performance::check_overdispersion()
```

```
# Overdispersion test

       dispersion ratio =    5.512
  Pearson's Chi-Squared = 3450.403
                p-value =  < 0.001

Overdispersion detected.
```

8.E.3 Modèle binomial négatif

Library `MASS` propose la fonction `glm.nb()` pour ajuster un modèle binomial négatif.

```
mod_nb <- glm.nb(
    n.prison ~ charactere + recherche.nouv + famille.prison + abus.enfant + age +
↪   factor(type.centre),
    data = smp
)
summary(mod_nb)
```

```
Call:
glm.nb(formula = n.prison ~ charactere + recherche.nouv + famille.prison +
    abus.enfant + age + factor(type.centre), data = smp, init.theta = 0.4746238145,
```

```
    link = log)

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.330018   0.395188  -3.366 0.000764 ***
charactere            0.304019   0.081209   3.744 0.000181 ***
recherche.nouv        0.240279   0.081137   2.961 0.003063 **
famille.prison        0.604064   0.146477   4.124 3.72e-05 ***
abus.enfant           0.462282   0.147036   3.144 0.001667 **
age                   0.008619   0.005423   1.589 0.111979
factor(type.centre)2  0.216223   0.246183   0.878 0.379780
factor(type.centre)3  0.286117   0.240209   1.191 0.233607
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.4746) family taken to be 1)

    Null deviance: 680.90  on 633  degrees of freedom
Residual deviance: 593.39  on 626  degrees of freedom
  (165 observations deleted due to missingness)
AIC: 2212

Number of Fisher Scoring iterations: 1

              Theta:  0.4746
          Std. Err.:  0.0430

 2 x log-likelihood:  -2193.9550
```

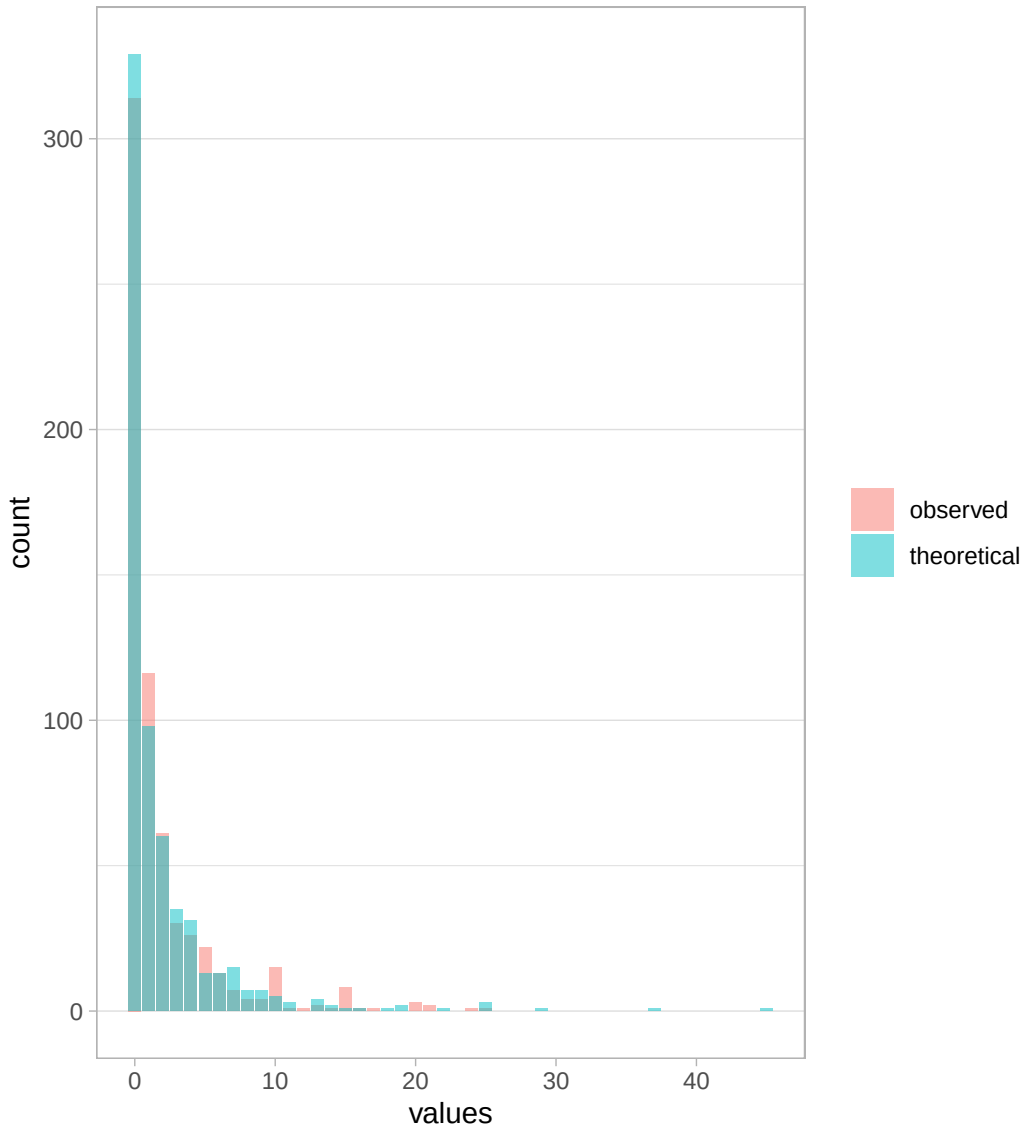
Affichage des résultats en tableau :

Characteristic	IRR[1]	95% CI	p-value
charactere	1.36	1.16, 1.59	<0.001
recherche.nouv	1.27	1.08, 1.49	0.003
famille.prison	1.83	1.39, 2.43	<0.001
abus.enfant	1.59	1.19, 2.14	0.002
age	1.01	1.00, 1.02	0.11
factor(type.centre)			
1	—	—	
2	1.24	0.75, 2.00	0.4
3	1.33	0.82, 2.12	0.2

[1]Estimates are incidence rate ratios from a negative binomial regression model.
Abbreviations: CI = Confidence Interval, IRR = Incidence Rate Ratio

Comparaison avec un modèle théorique de binomial négatif :

```r
mod_nb |>
    guideR::observed_vs_theoretical()
```



Détection de surdispersion :

```r
mod_nb |>
    performance::check_overdispersion()
```

```
# Overdispersion test

 dispersion ratio = 0.768
          p-value = 0.352

No overdispersion detected.
```

8.E.4 Comment choisir directement le meilleur modèle

Pour comparer objectivement les trois modèles ajustés ci-dessus (Poisson, quasi-Poisson, binomial négatif), on utilise :

- l'AIC (quand disponible),
- la surdispersion,
- et l'adéquation modèle/données.

9 Modèles de survie

9.A Modèles de survie paramétrique : Weibull etc

9.B Modèles de survie semi-paramétrique = Modèle de Cox

Modèle de Cox = régression de Cox = modèle des risques proportionnels de Cox.

Modèle semi-paramétrique similaire à la régression linéaire multiple ou à la régression logistique multiple, mais il est spécialement conçu pour les données de survie.

≠ Test du log-rank qui compare deux courbes de survie de manière univariée (non ajustée)

Équation du modèle de Cox :

$$h(t|Z_1, ..., Z_p) = h_0(t) \times \text{exp}(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$$

- $h_0(t)$: fonction de risque de base (baseline hazard function) = fonction de risque instantané lorsque toutes les covariables Z_i sont égales à 0.
- exp$(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$: effet multiplicatif constant des P covariables sur la fonction de risque instantané.
- β_i : coefficient associé à la covariable Z_i, représentant l'effet de cette covariable sur le risque instantané.
  - Relation entre β_i et le hazard ratio (HR) : $HR =$ exp(β_i) (donc $\beta_i =$ log(HR))

2 hypothèses principales : **proportionnalité des risques** et **log-linéarité**.

9.B.1 Hypothèse de proportionnalité des risques

Hypothèse de proportionnalité des risques : les rapports des risques entre les individus restent constants dans le temps, c'est à dire que l'effet des covariables sur le risque instantané est multiplicatif et ne dépend pas du temps.

$\frac{h(t|Z_k=1)}{h(t|Z_k=0)} =$ exp$(\beta k) = constante$

Quand on compare 2 groupes (exemple : fumeur vs non-fumeur), le rapport de leurs risques instantanés reste le même tout au long du suivi, et ce rapport de risques = hazard ratio (HR).

Dire que les risques sont proportionnels signifie :

- à chaque instant t, le groupe A a par exemple 1,5 fois plus de risque de l'événement que le groupe B ;
- ce facteur 1,5 ne change pas au cours du temps.

Exemple : On suit 100 patients opérés.

Variable explicative : fumeur (1) vs non-fumeur (0).

Supposons : HR = 2.

Cela veut dire :

- à tout moment du suivi, un fumeur a le double du risque instantané d'avoir une complication,
- même si le risque global diminue avec le temps (fin de la période postopératoire aiguë), le ratio reste stable entre fumeur et non fumeur.

Si cette hypothèse n'est pas vraie (ex : au début les fumeurs sont à très haut risque, mais plus tard le risque redevient identique), alors le modèle de Cox classique n'est plus adapté. (*Dans ce cas, on peut utiliser des modèles de Cox avec effets temporels*).

9.B.2 Hypothèse de log-linéarité

Hypothèse de log-linéarité : **concerne les variables quantitatives**.

L'effet des covariables est linéaire sur le logarithme du risque instantané et pas directement sur le risque.

C'est à dire que chaque unité d'augmentation de la covariable Z_i entraîne une augmentation constante du logarithme du risque instantané.

$\rightarrow$ Chaque augmentation d'une unité de la variable produit le même pourcentage de variation du risque.

log$(h(t|Z_1, ..., Z_p)) =$ log$(h_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p$

Exemple :

Variable explicative : âge (en années).

On suppose que $\beta = 0,05$ (avec β le coefficient associé à l'âge dans le modèle de Cox).

Cela veut dire :

- chaque année supplémentaire multiplie le risque instantané par $exp(0,05) \approx 1,05$
- donc +5 % de risque par année d'âge.
  - Si on passe de 50 à 51 ans : +5 %.
  - Si on passe de 70 à 71 ans : encore +5 %.
- L'effet est proportionnel, constant.

9.B.3 Exemple R

295 patientes du jeu de données `ks` :

- `survival` : temps de survie en mois
- `eventdeath` : indicateur d'événement (1 = décès, 0 = censuré)
- `age` : âge en années
- `grade` : grade de la tumeur du sein (1, 2 ou 3)
- `hormonal` : statut hormonal (1 = positif, 0 = négatif)

Objectif : modéliser le risque de décès en fonction de l'âge (> 40 ans), en prenant en compte le grade de la tumeur et le statut hormonal.

```r
mod <- coxph(
    Surv(survival,eventdeath)
    ~(age>40)+grade+ hormonal,
    data=ks)
summary(mod)
```

```
Call:
coxph(formula = Surv(survival, eventdeath) ~ (age > 40) + grade +
    hormonal, data = ks)

  n= 272, number of events= 77

                coef exp(coef) se(coef)      z Pr(>|z|)
age > 40TRUE -0.4279    0.6519   0.2439 -1.754   0.0794 .
grade         0.9528    2.5931   0.1822  5.229  1.7e-07 ***
hormonal     -0.1867    0.8297   0.4357 -0.428   0.6683
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
age > 40TRUE    0.6519     1.5340    0.4041     1.051
grade           2.5931     0.3856    1.8143     3.706
hormonal        0.8297     1.2053    0.3532     1.949

Concordance= 0.73  (se = 0.026 )
Likelihood ratio test= 41.41  on 3 df,   p=5e-09
Wald test            = 34.63  on 3 df,   p=1e-07
Score (logrank) test = 39.58  on 3 df,   p=1e-08
```

Affichage des résultats en tableau :

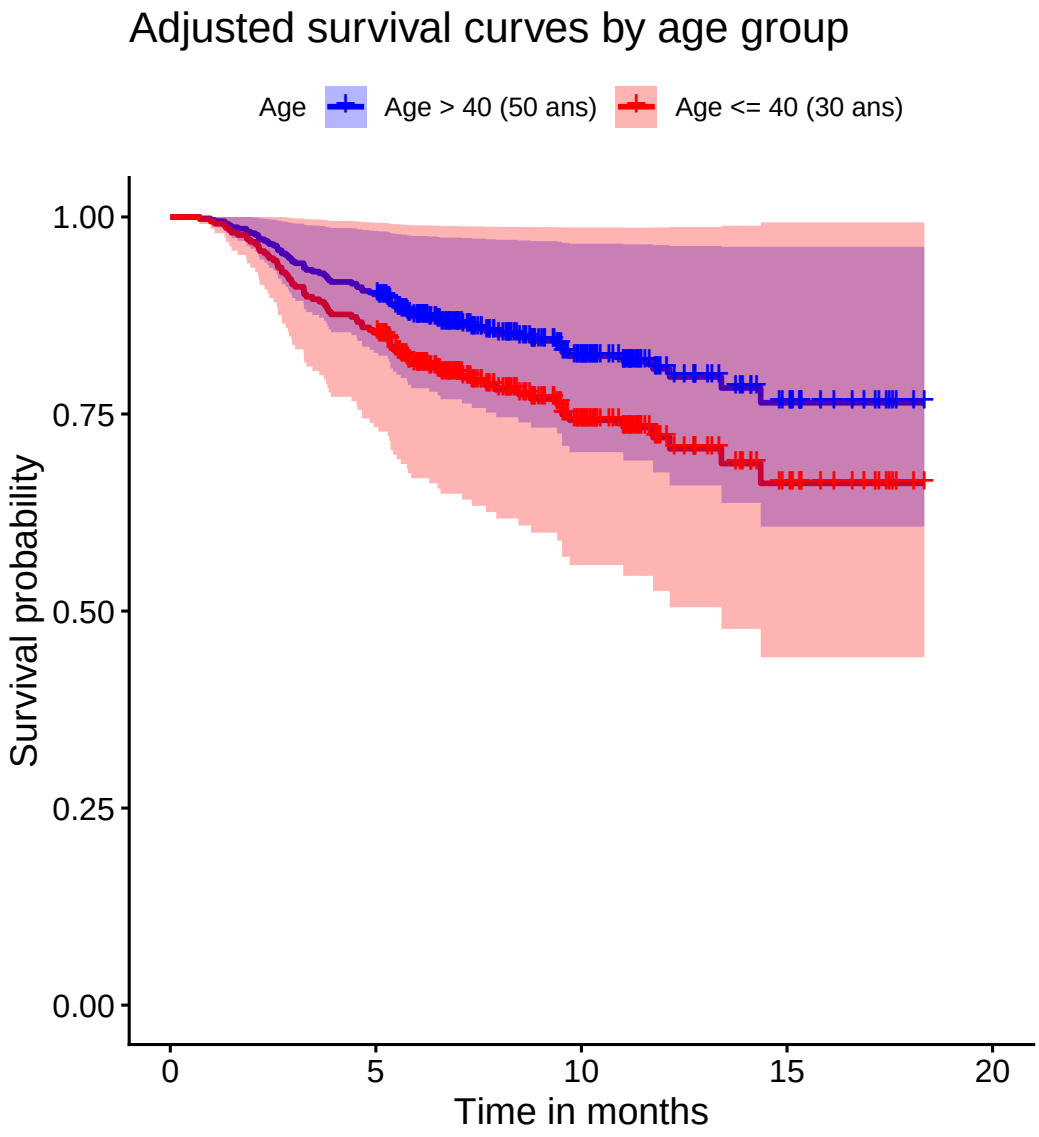
Characteristic	HR[1]	95% CI	p-value
age > 40			
FALSE	—	—	
TRUE	0.65	0.40, 1.05	0.079
grade	2.59	1.81, 3.71	<0.001
hormonal	0.83	0.35, 1.95	0.7

[1]Estimates are hazard ratios from a Cox proportional hazards model.
Abbreviations: CI = Confidence Interval, HR = Hazard Ratio

Courbes de survie ajustées :

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

```
i The deprecated feature was likely used in the ggpubr package.
  Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
```



5Autour_de_la_modélisation

1 Introduction

Modèle : “simplifie” la réalité.

1. comment coder les variables explicatives (quantitatives et catégorielles) ?
2. interactions entre variables explicatives.
3. sélection automatique de variables explicatives.
4. données manquantes.
5. techniques de ré-échantillonnage.

2 Codages des variables explicatives quantitatives

2.A Problématique

On considère le modèle linéaire suivant :

$$Y = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p + \epsilon$$

- Y = mesure des capacités cognitives d’un sujet
- X1 = âge du sujet (en années)

Comme ce modèle fait l’hypothèse d’une relation linéaire, il fait l’hypothèse que la variation entre 60 et 80 ans est la même qu’entre 20 et 40 ans.

On peut apprécier la forme de la relation entre la variable a expliquer Y (NB : dans le cas d’une régression logistique, il s’agit de logit$(P(Y = 1))$) et la variable explicative X_1 en utilisant des méthodes non paramétriques (loess, splines, etc.).

Si la relation n’est pas linéaire, on peut envisager plusieurs solutions :

1. Convert X_i en variable catégorielle (ex. quartiles, quintiles, etc.) : perte d’information, perte de puissance mais interprétation simple des résultats.
2. Ajouter des modèles polynomiaux en X_i (ex. X_i et X_i^2) : permet de modéliser des relations non linéaires simples (ex. relations quadratiques) mais interprétation complexe
3. Recourir à un “modèle additif généralisé” (GAM) : permet de modéliser des relations non linéaires complexes (ex. splines, loess, etc.) mais interprétation complexe

2.B Exemple R sur les modèles additifs généralisés (GAM)

Les modèles additifs généralisés (ou GAM, pour generalized additive models) offrent une méthode flexible pour décrire une relation non-linéaire entre des prédicteurs et une variable réponse.

La moyene de Y dépend de la somme de fonctions des variables explicatives, mais il n’y a pas d’hypothèse de linéarité.

Dans ce cadre, on laisse des "degrés de liberté" aux courbes de régression (et à la relation entre Y et X_i) pour qu'elles puissent s'adapter aux données.

Par exemple : si un seul degré de liberté, la relation est linéaire.

Avec deux degrés de liberté, la relation est quadratique, etc.

En gros :

- modèle linéaire généralisé = somme des lignes = sommes des courbes à degrés de liberté 1
- modèle additif généralisé = somme des courbes = sommes des courbes à degrés de liberté k

2.B.1 Modèle additif généralisé (GAM) avec splines pour une variable explicative quantitative

On considère la durée de l'entretien `smp$duree.interv` comme variable à expliquer et une liste de variables explicatives parmi lesquelles l'âge `smp$age`.

On trace un *spline* représentant la forme de la relation entre ces deux variables.

```
mod <- gam(duree.interv~s(age, k=20, fx=TRUE),data=smp)
plot(mod)
```

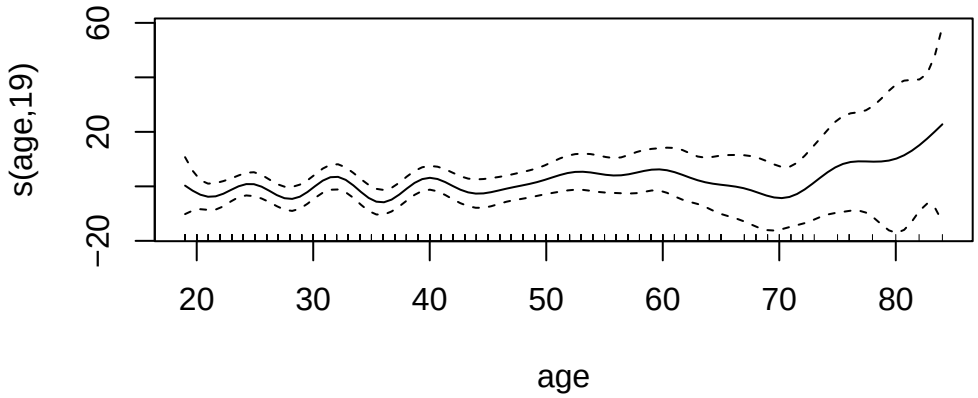


Figure 5.1: Spline représentant la forme de la relation entre la durée de l'entretien et l'âge des détenus.

On peut augmenter le nombre de paramètre du *spline* en augmentant le paramètre `k` (nombre de noeuds).

Si on laisse par défaut, `mgcv` choisit automatiquement une valeur de `k`

Pour savoir ce qu'il a choisi :

```
mod <- gam(duree.interv~s(age),data=smp)
summary(mod)
```

```
Family: gaussian
Link function: identity

Formula:
duree.interv ~ s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.2016     0.7066   88.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
       edf Ref.df     F p-value
s(age)   1      1 5.911  0.0153 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.00657   Deviance explained = 0.79%
GCV = 372.52  Scale est. = 371.51    n = 744
```

interprétation :

- edf = 1 : le *spline* est linéaire (1 degré de liberté)
- p-value = 0.0153 : la relation entre âge et durée de l'entretien est significative
- ça signifie que `mgcv` a choisi `k=2` (car pour un *spline* cubique, le degré de liberté est égal à `k-1`)
- Donc le *spline* est linéaire (car 2-1=1)
- donc la relation entre âge et durée de l'entretien est modélisée comme linéaire
- On peut vérifier en forçant `k=2` :

```r
mod1 <- gam(duree.interv~s(age,k=2,fx=TRUE),data=smp)
```

```
Warning in smooth.construct.tp.smooth.spec(object, dk$data, dk$knots): basis dimension, k, incr
```

```r
plot(mod1)
```

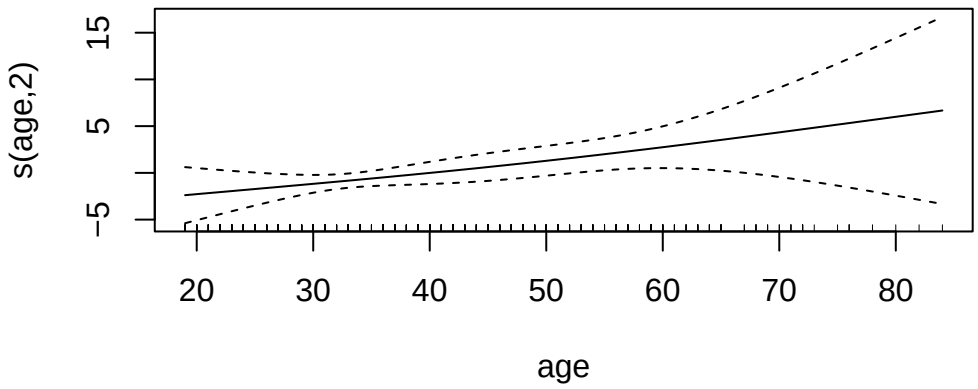


Figure 5.2: Spline représentant la forme de la relation entre la durée de l'entretien et l'âge des détenus, avec k=2.

Mais on peut forcer une valeur de `k` différente : par exemple ici on force `k=4`

```
mod2 <- gam(duree.interv~s(age,k=4,fx=TRUE),data=smp)
plot(mod2)
```

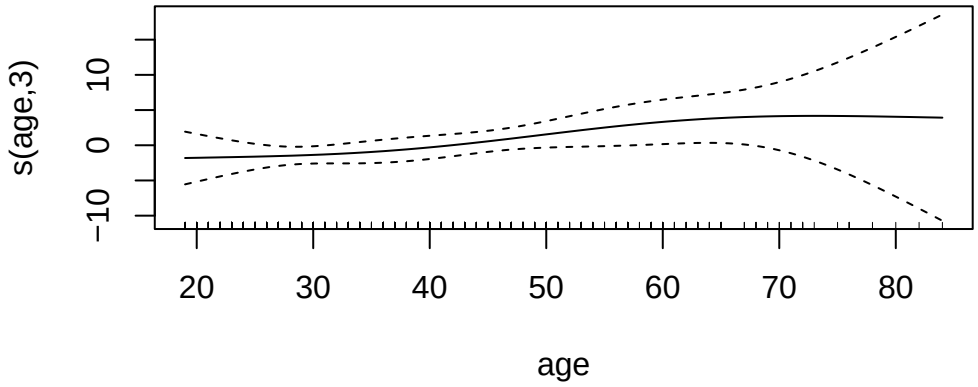


Figure 5.3: Spline représentant la forme de la relation entre la durée de l'entretien et l'âge des détenus, avec k=4.

En gros : plus `k` est grand, plus le *spline* est flexible.

2.B.2 Modèle additif généralisé (GAM) avec splines pour une régression logistique

On essaie d'expliquer l'abus de substance `smp$abus.subst` (0/1) en fonction de l'âge `smp$age`.

```
mod_logit <- gam(abus.subst~s(age), data=smp, family=binomial)
plot(mod_logit)
```

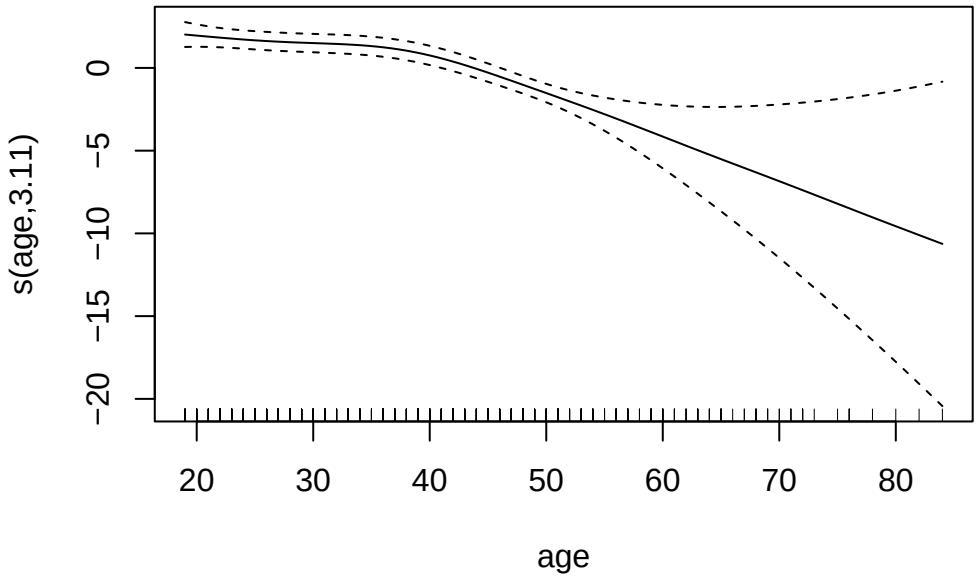


Figure 5.4: Spline représentant la forme de la relation entre le logit de la probabilité de présenter un abus de substance et l'âge des détenus.

La spline pointe un écart statistiquement significatif à la linéarité l'évolution est stable jusqu'à 40 ans puis semble s'infléchir.

Le nombre de degrés de liberté effectif est de 3.114 donc supérieur à 1. Il est codé "edf" dans le tableau et est écrit sur l'axe des y du graphique.

Et donc là : que faire ?

- Si l'âge est un facteur de confusion accessoire d'intérêt marginal : on le laisse comme ça
- Si l'âge est un facteur de confusion important : il faut adopter une autre stratégie
  - On peut le recouper en classes : plus l'échantillon est grand, plus on peut faire de classes

2.B.2.1 Découpage en classes La fonction `cut()` permet de découper une variable quantitative en classes.

```
smp$age.4f <- cut(
    smp$age,
    breaks=c(-Inf, 25, 35, 45, Inf),
    labels=c("<25","25-35","35-45",">45"))
table(smp$age.4f, useNA="ifany")
```

```
  <25 25-35 35-45   >45  <NA>
  140   217   202   238     2
```

Les effectifs par classe sont suffisants pour inclure cette variable dans un modèle de régression logistique.

On peut ensuite l'inclure dans un modèle de régression logistique classique :

```r
mod_logit2 <- glm(
    abus.subst ~
    age.4f +
    factor(type.centre),
    data=smp,
    family="binomial")
summary(mod_logit2)
```

```

Call:
glm(formula = abus.subst ~ age.4f + factor(type.centre), family = "binomial",
    data = smp)

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -1.0570     0.3886  -2.720 0.006525 **
age.4f25-35           -0.2397     0.2274  -1.054 0.291920
age.4f35-45           -0.8364     0.2405  -3.478 0.000505 ***
age.4f>45             -3.3742     0.4537  -7.436 1.04e-13 ***
factor(type.centre)2   0.6805     0.3781   1.800 0.071901 .
factor(type.centre)3   1.0808     0.3554   3.041 0.002355 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 921.27  on 796  degrees of freedom
Residual deviance: 757.98  on 791  degrees of freedom
  (2 observations deleted due to missingness)
AIC: 769.98

Number of Fisher Scoring iterations: 6
```

Représentation en table des résultats :

```r
library(gtsummary)
tbl <-
    mod_logit2 %>%
    tbl_regression(exponentiate = TRUE) %>%
    as_gt()
tbl
```

Characteristic	OR	95% CI	p-value
age.4f			
<25	—	—	
25-35	0.79	0.50, 1.23	0.3
35-45	0.43	0.27, 0.69	<0.001
>45	0.03	0.01, 0.08	<0.001
factor(type.centre)			
1	—	—	
2	1.97	0.97, 4.31	0.072
3	2.95	1.52, 6.20	0.002

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

2.B.2.2 Polynômes orthogonaux Le problème avec les polynômes "classiques" (ex. X_i et X_i^2) est que les termes sont corrélés entre eux.

On peut utiliser des polynômes orthogonaux pour éviter ce problème.

La fonction `poly()` permet de créer des polynômes orthogonaux en choisissant des degrés de polynome pour lesquels les termes sont orthogonaux donc non corrélés.

```
dt <- na.omit(smp[,c("abus.subst","age","type.centre")])
mod3 <- glm(
    abus.subst ~
    poly(age, degree = 3) + factor(type.centre),
    data=dt,
    family="binomial")
```

```
Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

Warning : le processus de convergence est fragile, car les polynômes de degré élevé peuvent causer des problèmes numériques.

Avec un degré 2, ça passe mieux :

```
dt <- na.omit(smp[,c("abus.subst","age","type.centre")])
mod3 <- glm(
    abus.subst ~
    poly(age, degree = 2) + factor(type.centre),
    data=dt, family="binomial")
summary(mod3)
```

```
Call:
glm(formula = abus.subst ~ poly(age, degree = 2) + factor(type.centre),
    family = "binomial", data = dt)
```

```
Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -2.8028     0.4042  -6.934 4.10e-12 ***
poly(age, degree = 2)1 -68.6238    10.7444  -6.387 1.69e-10 ***
poly(age, degree = 2)2 -32.6108     7.9009  -4.127 3.67e-05 ***
factor(type.centre)2     0.6178     0.3824   1.616  0.10616
factor(type.centre)3     1.0559     0.3591   2.941  0.00328 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 921.27  on 796  degrees of freedom
Residual deviance: 740.34  on 792  degrees of freedom
AIC: 750.34

Number of Fisher Scoring iterations: 7
```

Le seul truc interprétable sont les p-values associées aux termes polynomiaux (pas les coefficients).

On visualise la relation proposée par le modèle en utilisant la fonction `predict()`.

```r
ages=seq(20,50,1)
avgpred <- sapply(
    ages,
    function(age) {
        dt$age <- age
        mean(predict(mod3, newdata=dt, type="response"))}
        )
plot(ages, avgpred,
    type="l",
    xlim=c(20,50), ylim=c(0,1),
    las=1,
    xlab="Âge", ylab="Prédiction moyenne")
```

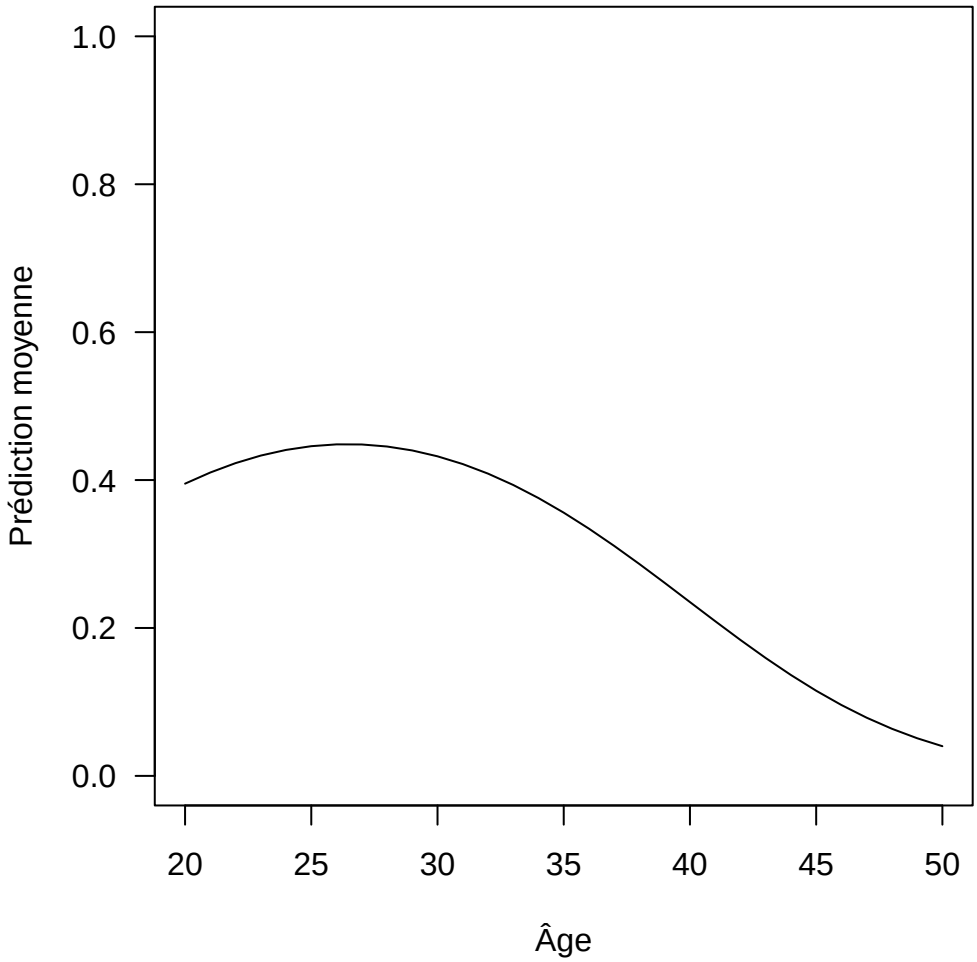


Figure 5.5: Prédiction moyenne du risque d'abus de substance en fonction de l'âge des détenus, standardisée sur le type de centre.

Cette approche de prédiction moyenne consiste à évaluer le pourcentage de risque d'abus de substance selon le modèle mod3, pour chaque âge et après « standardisation » sur le type de centre (car le type de centre est inclus dans `mod3`).

3 Codage des variables explicatives catégorielles

Dans le modèle linéaire, les variables explicatives catégorielles sont classiquement

- Transformées en variables binaires
- Ou en variables quantitatives discrètes (le problème est que l'écart entre les modalités n'est pas forcément le même)

Dans le cas de variables catégorielles comptant \2 classes : il faut recoder par une série de variables binaires.

3.A Exemple : ABO

4 groupes sanguins : A, B, AB, O

- Approche classique : codage 0/1 pour chaque groupe sanguin (3 variables binaires, une de référence)
  - Groupe A : $X_1 = 1$ et $X_2 = 0$ et $X_3 = 0$
  - Groupe B : $X_1 = 0$ et $X_2 = 1$ et $X_3 = 0$
  - Groupe AB : $X_1 = 0$ et $X_2 = 0$ et $X_3 = 1$
  - Groupe O (référence) : $X_1 = 0$ et $X_2 = 0$ et $X_3 = 0$

En tableau :

Groupe sanguin	X_1	X_2	X_3
A	1	0	0
B	0	1	0
AB	0	0	1
O (référence)	0	0	0

Équation du modèle linéaire :

$$Y = a_0 + a_1 Z_1 + a_2 Z_2 + ... + a_p Z_p + b_1 X_1 + b_2 X_2 + b_3 X_3 + \epsilon$$

Dans ce modèle : le coefficient b_1 sera interprété comme la différence moyenne de Y entre le groupe A et le groupe O (référence) à $Z_1, Z_2, ..., Z_p$ fixés.

- D'autres codages sont possibles : notamment "-1 / 1"
  - Groupe A : $X_1 = 1$ et $X_2 = X_3 = 0$
  - Groupe B : $X_2 = 1$ et $X_1 = X_3 = 0$
  - Groupe AB : $X_3 = 1$ et $X_1 = X_2 = 0$
  - Groupe O (référence) : $X_1 = X_2 = X_3 = -1$

Équation du modèle linéaire :

$$Y = a_0 + a_1 Z_1 + a_2 Z_2 + ... + a_p Z_p + b_1 X_1 + b_2 X_2 + b_3 X_3 + \epsilon$$

Dans ce modèle : le coefficient b_1 sera interprété comme la différence entre l'effet observé dans le groupe "A" et la moyenne non pondérée des effets observés dans les autres groupes (B, AB et O) à $Z_1, Z_2, ..., Z_p$ fixés.

Moyenne non pondérée : il s'agit de la moyenne des moyennes de chaque groupe, calculée sans tenir compte des effectifs respectifs de ces groupes (chaque groupe a le même poids dans le calcul de la moyenne globale).

3.B Exemple R

3.B.1 Avec une variable codée 0/1

variable catégorielle : `smp$profession` (8 modalités)

```
str(smp$profession)
```

```
 Factor w/ 8 levels "agriculteur",..: 7 NA 4 6 8 6 7 2 6 6 ...
```

Pour savoir comment R code cette variable dans un modèle de régression linéaire : on utilise la fonction `contrasts()` :

```
contrasts(smp$profession)
```

```
              commerçant cadre intermédiaire employé ouvrier autre sans.emploi
agriculteur            0     0             0       0       0     0           0
commerçant             1     0             0       0       0     0           0
cadre                  0     1             0       0       0     0           0
intermédiaire          0     0             1       0       0     0           0
employé                0     0             0       1       0     0           0
ouvrier                0     0             0       0       1     0           0
autre                  0     0             0       0       0     1           0
sans.emploi            0     0             0       0       0     0           1
```

Par défaut, R utilise un codage “0/1” avec la première modalité comme référence (ici “Agriculteurs”).

Pour changer la modalité de référence, on peut utiliser la fonction `relevel()` :

```
smp$profession <- relevel(smp$profession, ref="ouvrier")
contrasts(smp$profession)
```

```
              agriculteur commerçant cadre intermédiaire employé autre
ouvrier                 0          0     0             0       0     0
agriculteur             1          0     0             0       0     0
commerçant              0          1     0             0       0     0
cadre                   0          0     1             0       0     0
intermédiaire           0          0     0             1       0     0
employé                 0          0     0             0       1     0
autre                   0          0     0             0       0     1
sans.emploi             0          0     0             0       0     0
              sans.emploi
ouvrier                 0
agriculteur             0
commerçant              0
```

```
cadre                   0
intermédiaire           0
employé                 0
autre                   0
sans.emploi             1
```

Ici, la modalité de référence est “ouvrier”.

On cherche à expliquer la variable “haut risque suicidaire” `smp$hr.suicide` (0/1) en fonction de la profession `smp$profession` + autres variables de confusion.

```r
mod <- glm(
    hr.suicide~
    abus.enfant+discipline+duree.peine+profession+factor(type.centre),
    data=smp,
    family="binomial")
summary(mod)
```

```
Call:
glm(formula = hr.suicide ~ abus.enfant + discipline + duree.peine +
    profession + factor(type.centre), family = "binomial", data = smp)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.28420    0.78455  -0.362  0.71717
abus.enfant              0.62004    0.23521   2.636  0.00839 **
discipline               0.39596    0.24881   1.591  0.11151
duree.peine             -0.32932    0.14853  -2.217  0.02661 *
professionagriculteur    2.13259    1.25662   1.697  0.08968 .
professioncommerçant    -0.17238    0.40182  -0.429  0.66793
professioncadre         -0.77484    0.77783  -0.996  0.31917
professionintermédiaire -0.92553    0.64236  -1.441  0.14963
professionemployé       -0.41118    0.36301  -1.133  0.25734
professionautre         -2.03833    1.05821  -1.926  0.05408 .
professionsans.emploi    0.27877    0.26911   1.036  0.30024
factor(type.centre)2    -0.06536    0.34254  -0.191  0.84868
factor(type.centre)3     0.31189    0.37935   0.822  0.41099
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 554.57  on 546  degrees of freedom
Residual deviance: 512.92  on 534  degrees of freedom
  (252 observations deleted due to missingness)
AIC: 538.92
```

```
Number of Fisher Scoring iterations: 5
```

Il y a bien une estimation pour 7 modalité de la variable `profession`, sauf pour la modalité de référence “ouvrier”.

Interprétation : chaque coefficient de la variable `profession` est interprété par rapport à la modalité de référence “ouvrier”.

Si on veut tester l’effet global de la variable `profession` dans le modèle, on utilise `drop1`:

```
drop1(mod,.~.,test="Chisq")
```

```
Single term deletions

Model:
hr.suicide ~ abus.enfant + discipline + duree.peine + profession +
    factor(type.centre)
                    Df Deviance    AIC     LRT Pr(>Chi)
<none>                   512.92 538.92
abus.enfant          1   519.72 543.72  6.7974 0.009129 **
discipline           1   515.40 539.40  2.4807 0.115250
duree.peine          1   517.84 541.84  4.9206 0.026538 *
profession           7   530.83 542.83 17.9130 0.012369 *
factor(type.centre)  2   514.88 536.88  1.9568 0.375911
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.B.2 Avec une variable codée -1/1

On peut utiliser la fonction `contrasts<-` pour changer le codage par défaut de R.

```
contrasts(smp$profession) <- contr.sum
contrasts(smp$profession)
```

```
              [,1] [,2] [,3] [,4] [,5] [,6] [,7]
ouvrier          1    0    0    0    0    0    0
agriculteur      0    1    0    0    0    0    0
commerçant       0    0    1    0    0    0    0
cadre            0    0    0    1    0    0    0
intermédiaire    0    0    0    0    1    0    0
employé          0    0    0    0    0    1    0
autre            0    0    0    0    0    0    1
sans.emploi     -1   -1   -1   -1   -1   -1   -1
```

Problème : le codage en matrice [,1] [,2].. n’est pas très lisible.

Renommer :

```r
colnames(contrasts(smp$profession)) <-
↪  c("ouvrier","agriculteur","commerçant","cadre","intermédiaire","employé","autre")
contrasts(smp$profession)
```

```
              ouvrier agriculteur commerçant cadre intermédiaire employé autre
ouvrier             1           0          0     0             0       0     0
agriculteur         0           1          0     0             0       0     0
commerçant          0           0          1     0             0       0     0
cadre               0           0          0     1             0       0     0
intermédiaire       0           0          0     0             1       0     0
employé             0           0          0     0             0       1     0
autre               0           0          0     0             0       0     1
sans.emploi        -1          -1         -1    -1            -1      -1    -1
```

Modèle de régression logistique avec ce nouveau codage :

```r
mod <- glm(hr.suicide ~
        abus.enfant + discipline + duree.peine + profession + factor(type.centre),
↪
        data=smp,
        family="binomial")
summary(mod)
```

```
Call:
glm(formula = hr.suicide ~ abus.enfant + discipline + duree.peine +
    profession + factor(type.centre), family = "binomial", data = smp)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)             -0.52306    0.82744  -0.632  0.52729
abus.enfant              0.62004    0.23521   2.636  0.00839 **
discipline               0.39596    0.24881   1.591  0.11151
duree.peine             -0.32932    0.14853  -2.217  0.02661 *
professionouvrier        0.23886    0.29659   0.805  0.42060
professionagriculteur    2.37146    1.10197   2.152  0.03140 *
professioncommerçant     0.06649    0.39303   0.169  0.86567
professioncadre         -0.53598    0.69640  -0.770  0.44151
professionintermédiaire -0.68667    0.58349  -1.177  0.23926
professionemployé       -0.17232    0.36126  -0.477  0.63336
professionautre         -1.79947    0.93733  -1.920  0.05489 .
factor(type.centre)2    -0.06536    0.34254  -0.191  0.84868
factor(type.centre)3     0.31189    0.37935   0.822  0.41099
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 554.57  on 546  degrees of freedom
Residual deviance: 512.92  on 534  degrees of freedom
  (252 observations deleted due to missingness)
AIC: 538.92

Number of Fisher Scoring iterations: 5
```

Les log odds-ratios correspondants aux modalités de la variable « profession » sont maintenant relatifs à la moyenne des effets des 8 modalités

Si on veut voir “sans emploi” aussi :

```r
estimable(mod,c(0,0,0,0,-1,-1,-1,-1,-1,-1,-1,0,0),conf.int=0.95)
```

```
                                   Estimate Std. Error X^2 value DF Pr(>|X^2|)
(0 0 0 0 -1 -1 -1 -1 -1 -1 -1 0 0) 0.517635  0.2952967  3.072771  1 0.07961365
                                      Lower.CI Upper.CI
(0 0 0 0 -1 -1 -1 -1 -1 -1 -1 0 0) -0.06822514 1.103495
```

4 Choix des variables explicatives

4.A Principe

Quelles variables explicatives doit-on inclure dans le modèle ?

2 types de modèles :

- Prédictifs : estimer la probabilité de survenue d'une pathologie
- Explicatif : comprendre les principaux déterminants de la survenue d'une maladie

Pour les modèles prédictifs : c'est du machine learning.

Choix des variables pour les modèles explicatifs :

- Approche hypothético-déductive : réfléchir en amont aux variables à inclure
- Graphes acycliques orientés : article
- Analyse bivariée initiale :
  - inclure les variables significatives en analyse bivariée
  - problème : risque de confusion entre variables explicatives
  - et il faut inclure dans tous les cas quelque chose de cliniquement pertinent

NB : L'analyse pas à pas ("stepwise") est fortement déconseillée pour la sélection de variables, en particulier pour les modèles explicatifs. Voici pourquoi :

1. **Biais de sélection et sur-ajustement (Overfitting)** : Le stepwise teste de nombreuses combinaisons et ne garde que celles qui "marchent" le mieux sur l'échantillon donné. Cela a tendance à capturer le bruit aléatoire (le hasard) spécifique à cet échantillon plutôt que la vraie relation biologique ou clinique. Le modèle final performe souvent très bien sur les données d'apprentissage mais mal sur de nouvelles données.
2. **P-values et intervalles de confiance invalides** : Les calculs classiques des p-values et des intervalles de confiance supposent que le modèle a été spécifié *a priori*. Quand on utilise les données pour choisir le modèle, on effectue des tests multiples implicites sans correction. En conséquence, les p-values affichées sont artificiellement trop petites (on trouve trop de résultats "significatifs") et les intervalles de confiance sont trop étroits (donnant une fausse impression de précision).
3. **Instabilité du modèle** : Le retrait ou l'ajout de quelques observations peut changer radicalement la liste des variables sélectionnées par l'algorithme.
4. **Biais des coefficients** : Les coefficients des variables retenues sont souvent surestimés (biais loin de zéro), car seules les variables ayant par hasard un effet fort dans cet échantillon passent le filtre de sélection.

Recommandation : Privilégier une sélection basée sur la connaissance du domaine (littérature, plausibilité biologique, DAGs) plutôt que sur des algorithmes automatiques basés uniquement sur la significativité statistique.

5 À propos des termes d'interaction

Dans un modèle linéaire / linéaire généralisé / modèle de Cox, les variables explicatives ont des effets qui **s'additionnent** avant transformation par la réciproque de la fonction de lien.

- Dans le cas du modèle linéaire : les effets s'additionnent directement
- Dans le cas du modèle logistique : les log-odds s'additionnent, donc les odds-ratios se multiplient
- Dans le cas du modèle de Cox : les log-hazards s'additionnent, donc les hazard-ratios se multiplient

Dans le cas d'une intercation entre deux variables explicatives, l'effet d'une variable dépend de la valeur de l'autre variable.

Par exemple, l'effet de l'alcool sur le cancer du larynx est potentialisé par le tabac.

Dans un modèle : on peut inclure un terme d'interaction entre deux variables explicatives.

La formule serait :

$$Y = a_0 + a_1 X_1 + a_2 X_2 + a_3 (X_1 * X_2) + \epsilon$$

où $X_1 * X_2$ est le terme d'interaction entre les variables explicatives X_1 et X_2.

L'interprétation des coefficients devient plus complexe en présence d'interactions.

Par exemple, dans un modèle de régression linéaire avec interaction entre deux variables explicatives X_1 et X_2, le coefficient a_1 représente l'effet de X_1 sur Y lorsque X_2 est égal à zéro, et vice versa pour a_2.

L'effet combiné des deux variables sur Y est donné par la somme des coefficients a_1, a_2 et a_3 multiplié par les valeurs respectives de X_1 et X_2.

Pour analyser l'effet de X_1 sur Y, il peut être suggéré de ne pas inclure le terme d'interaction si l'objectif principal est d'estimer l'effet moyen de X_1 sur Y.

6 Données manquantes

6.A Types de données manquantes

- Manquantes complètement aléatoires (MCAR) : la probabilité qu'une donnée soit manquante est indépendante de la valeur de la donnée elle-même et des autres variables observées.
  - Exemple : un questionnaire perdu par la poste, ou une erreur de saisie aléatoire.
- Manquantes aléatoires (MAR) : la probabilité qu'une donnée soit manquante dépend des autres variables observées, mais pas de la valeur de la donnée elle-même.
  - Exemple : les patients plus âgés sont moins susceptibles de répondre à certaines questions, mais parmi les patients du même âge, la probabilité de réponse ne dépend pas de la valeur de la donnée manquante.
- Manquantes non aléatoires (MNAR) : la probabilité qu'une donnée soit manquante dépend de la valeur de la donnée elle-même, même après avoir pris en compte les autres variables observées.
  - Exemple : les patients avec des symptômes plus graves sont moins susceptibles de répondre à une question sur leur état de santé.

6.B Gestion des données manquantes

1. Les décrire !
2. Les imputer : par la médiane, la moyenne, le mode… Ou faire de l'imputation multiple avec le package `mice`.

6.C Exemple R avec le package `mice`

```
# sélection du jeu de données avec les variables d'intérêt
smp.imp <- smp[,c(2,7:93)]
# imputation multiple avec le package mice. par défaut : 5 imputations
smp.mice <- mice(smp.imp, seed=1)
```

```
 iter imp variable
  1   1  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  1   2  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  1   3  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  1   4  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  1   5  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  2   1  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  2   2  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  2   3  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
```

```
  2   4  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  2   5  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  3   1  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  3   2  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  3   3  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  3   4  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  3   5  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  4   1  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  4   2  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  4   3  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  4   4  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  4   5  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  5   1  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  5   2  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  5   3  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  5   4  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
  5   5  gravite  nb.enfants  scolarite  profession.b  annees.prison  preventive  crime  duree.
```

```r
# extraction de la première imputation
smp.imp.1 <- complete(smp.mice,1)
table(smp$determination,smp.imp.1$determination, useNA="ifany")
```

```
         0   1
  0    567   0
  1      0 147
  <NA>  71  14
```

Si on veut estimer le modèle avec les données imputées, il faut utiliser la fonction `pool()` :

```r
mod.imp <- with(
        smp.mice,
                glm(
                    hr.suicide ~
                    gravite + atcd.suicide + depression+ atcd.prison
                    + contre.pers + abus.subst + determination +
                     ↪  factor(smp$type.centre),
                    family="binomial"
                    )
                )
summary(pool(mod.imp),type="all")[,c(1,3,4,7,10)]
```

```
                       term    estimate std.error      p.value        fmi
1               (Intercept) -6.91754466 0.6601498 5.418862e-24 0.02341323
2                   gravite  0.68760625 0.1086462 1.447867e-09 0.12281306
3              atcd.suicide  1.62561164 0.2623188 7.823907e-09 0.17519303
```

```
4                 depression  1.38815347 0.2754109 2.062645e-06 0.19918619
5                atcd.prison  0.58632019 0.2583099 2.398243e-02 0.10003463
6                contre.pers -0.08098422 0.2616474 7.570826e-01 0.06867361
7                 abus.subst -0.13431550 0.2780758 6.294464e-01 0.09650375
8             determination  0.92304111 0.2930070 2.559647e-03 0.27109243
9  factor(smp$type.centre)2  0.65096200 0.4071465 1.102619e-01 0.00780281
10 factor(smp$type.centre)3  0.84577731 0.3817517 2.720780e-02 0.06065212
```

La fonction `with(glm(...))` estime le même modèle logistique sur chacun des 5 jeux de données imputées contenus dans smp.mice.

La fonction `pool()` fait une synthèse des 5 séries de résultats.

Parmi les résutats disponibles il y a : les log odds-ratios moyens, leurs erreurs types obtenues à partir de la somme des variances intra-modèles et des variances inter-modèles.

La FMI (Fraction of Missing Information) quantifie, pour chaque variable, la part de la variance dûe aux données manquantes.

7 Bootstrap

7.A Principe

Dans un modèle logistique, les p value et les IC des OR sont calculées à partir des « dérivées partielles secondes de la log vraisemblance », calculs dont la fiabilité est en principe garantie pour des tailles d'échantillons tendant vers l'infini.

Pour des échantillons de taille modérée, on peut utiliser le bootstrap pour estimer empiriquement la distribution des paramètres du modèle.

En gros, le bootstrap consiste à :

- Tirer au hasard avec remise des échantillons de taille n (taille de l'échantillon initial) parmi les n observations disponibles.
- Estimer le modèle sur chacun de ces échantillons bootstrap.
- Répéter l'opération un grand nombre de fois (ex: $k = 1000$ ou 10000).
- Appliquer la méthode des percentiles pour construire des intervalles de confiance empiriques à partir des k estimations obtenues.
  - On classe les k estimations obtenues (ex: les 1000 Odds-Ratios) du plus petit au plus grand.
  - On retire les 2,5% des valeurs les plus basses et les 2,5% des valeurs les plus élevées (les extrêmes).
  - L'intervalle entre la plus petite et la plus grande valeur restante constitue l'intervalle de confiance à 95%.

Cela revient à dire : *"Si je répétais mon étude un grand nombre de fois, dans 95% des cas, mon paramètre tomberait dans cet intervalle"*.

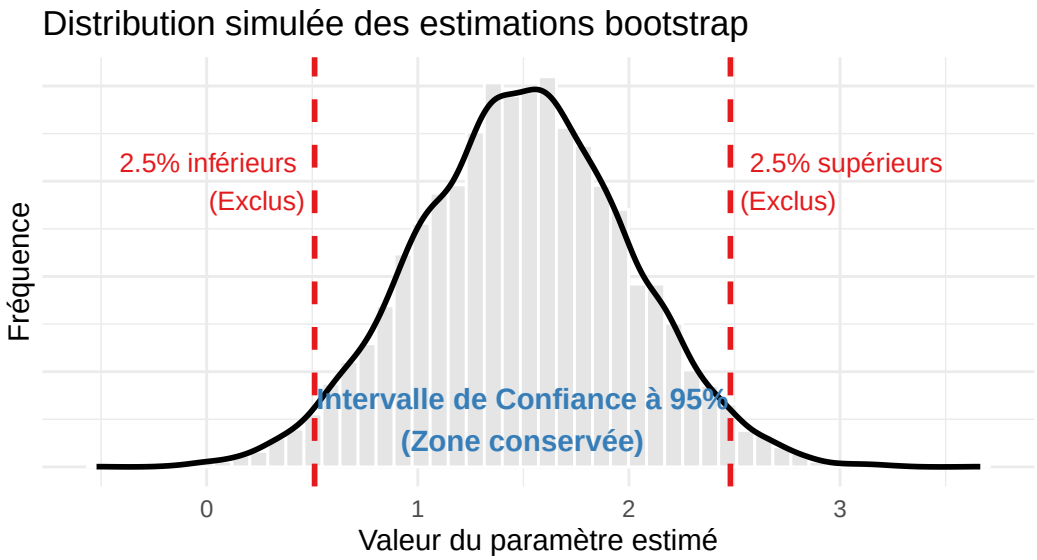


Figure 5.6: Illustration du principe du Bootstrap (Méthode des percentiles)

Le bootstrap présente l'avantage de ne pas reposer sur des trucs mathématiques impénétrables

!

Il repose toutefois sur l'hypothèse que l'échantillon initial est représentatif de la population cible et que les observations sont indépendantes.

Les observations ré-échantillonnées peuvent correspondre à un grappe (*cluster*) de données corrélées entre elles, telles que l'ensemble des mesures répétées chez un même patient ou dans un même centre.

7.B Conditions à respecter pour utiliser le bootstrap :

1. Indépendance des observations ré-échantillonnées
2. Nombre suffisant d'observations dans l'échantillon initial pour que les fluctuations de ré-échantillonnages soient proches des fluctuations d'échantillonnages de la population
3. Fluctuations d'échantillonnages continues (non discrète) : La statistique calculée ne doit pas faire de "sauts" (être discrète) mais varier de façon fluide.

Exemple de fluctuation non continue (discrète) : La Médiane.

Si on calcule la médiane sur des données discrètes (ex: des notes entières de 0 à 10) avec un petit échantillon, la médiane ne pourra prendre que quelques valeurs spécifiques (ex: 2, 2.5, 3).

La distribution bootstrap sera "en peigne" (avec des trous) et non une courbe lisse. Cela rend les intervalles de confiance peu fiables.

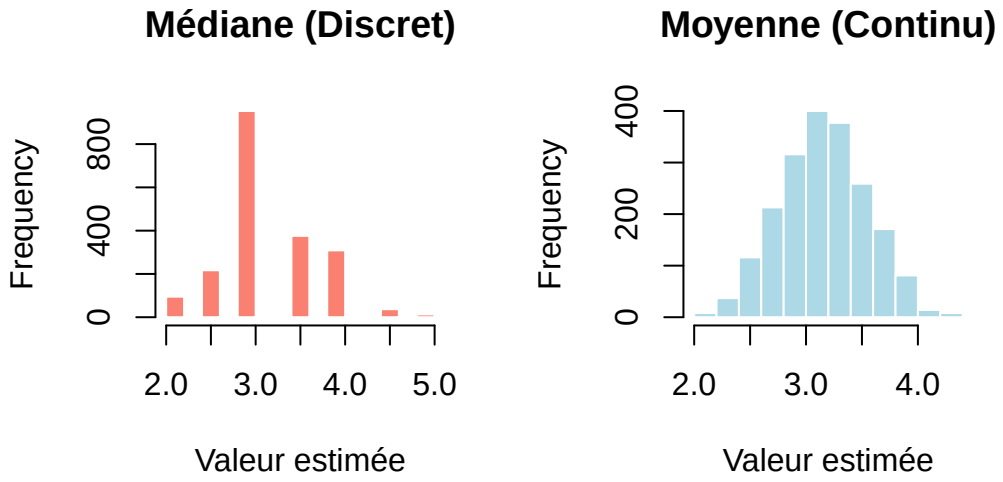


Figure 5.7: Comparaison : La médiane d'une variable discrète (gauche) produit une distribution bootstrap discontinue (en peigne), contrairement à la moyenne (droite) qui est plus continue.

4. Stabilisation asymptotique des fluctuations d'échantillonnages (c'est à dire que la distribution des fluctuations d'échantillonnages converge vers une distribution limite lorsque la taille de l'échantillon tend vers l'infini, souvent une distribution normale)

Exemple : Convergence vers la normalité.

Si la variable d'origine suit une distribution très asymétrique (ex: loi exponentielle), la distribution bootstrap de la moyenne sera elle-même asymétrique pour de petits échantillons. Elle ne deviendra "Normale" (en cloche) que si la taille de l'échantillon est suffisante.

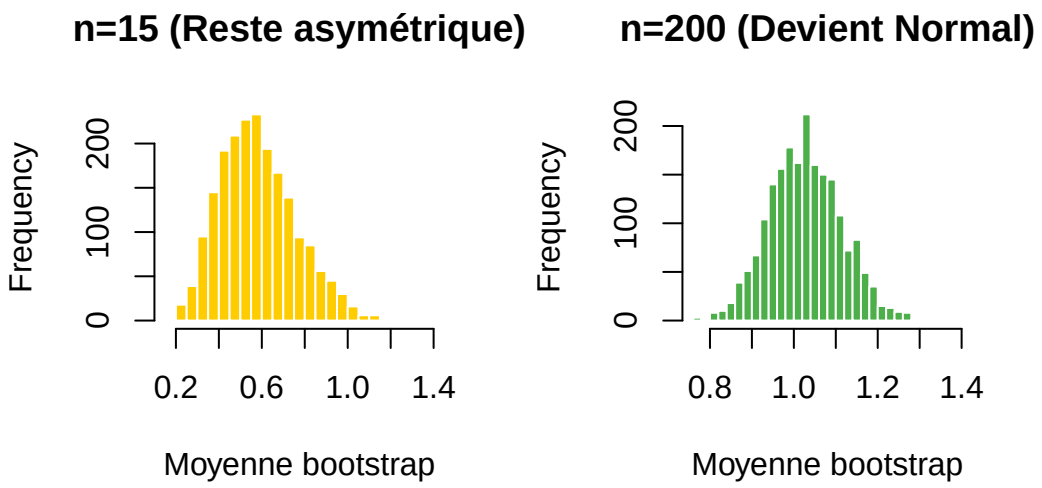


Figure 5.8: Illustration de la stabilisation asymptotique : Pour une variable asymétrique, la distribution bootstrap de la moyenne ne devient normale que si n est assez grand.

7.C Diagnostics et avantages

L'analyse graphique de la distribution des estimations bootstrap (histogramme, Q-Q plot) est essentielle.

1. Intérêt des diagnostics graphiques :

- **Vérifier la normalité** : Si la distribution est proche d'une courbe en cloche (Normale), les intervalles de confiance classiques sont fiables.
- **Sanity Check** : Permet de repérer des anomalies (bimodalité, valeurs aberrantes) invisibles avec un simple calcul.

2. Avantages généraux de la méthode :

- **Pédagogique** : Permet de visualiser concrètement l'incertitude et les fluctuations d'échantillonnage.
- **Priorité à la clinique** : On choisit la statistique la plus pertinente cliniquement (ex: médiane, ratio) sans être limité par la complexité mathématique de son intervalle de confiance.
- **Intelligibilité** : Offre souvent une alternative plus compréhensible aux modèles complexes (comme les modèles mixtes) pour gérer des données difficiles.
- **Interprétation facilitée de l'Intervalle de Confiance** : L'intervalle de confiance classique est souvent mal compris car sa définition est théorique et abstraite (*"si on répétait l'étude une infinité de fois…"*). Avec le bootstrap, cette répétition devient concrète et visible grâce à l'histogramme des simulations. On comprend intuitivement que l'IC correspond simplement à la zone où se concentrent la majorité (95%) des résultats simulés.

Exemple de diagnostics graphiques :

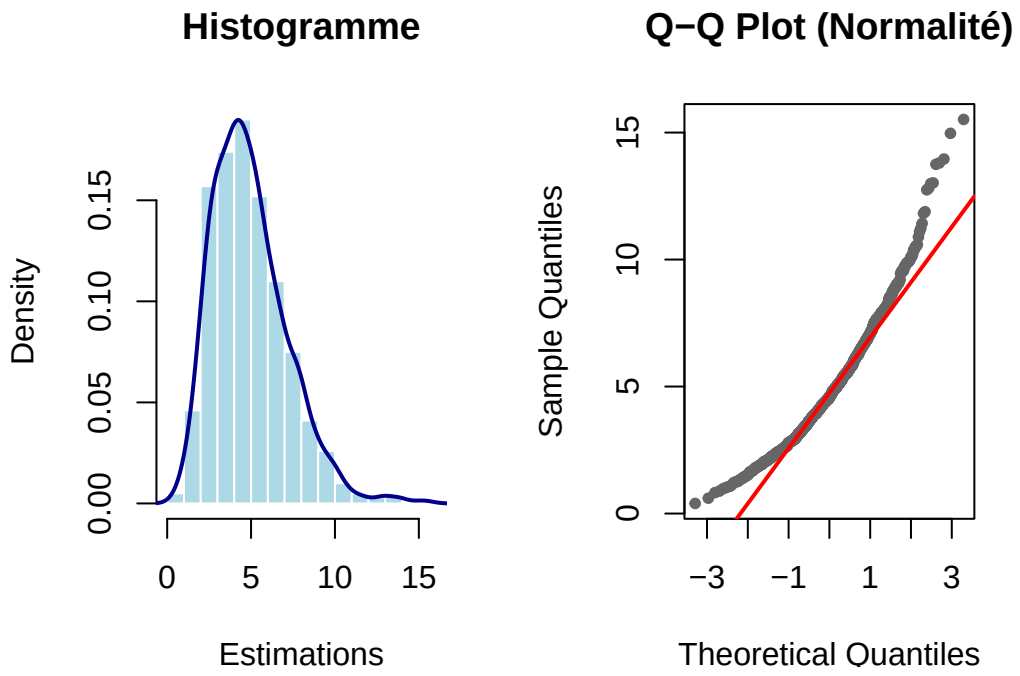


Figure 5.9: Diagnostics : L'histogramme et le Q-Q plot permettent de juger de la normalité de la distribution bootstrap.

7.D Exemple R

Modèle linéaire expliquant la variable « durée de l'entretien » réalisée lors de l'étude santé mentale en prison.

En vérifiant les conditions de validité du modèle, il y a un doute sur l'absence de normalité des résidus.

```
mod <- lm(
    duree.interv~

 ↪  schizophrenie+depression+abus.subst+gravite+charactere+trauma.enfant+age+factor(type.centre
    data=smp
    )
summary(mod)
```

```
Call:
lm(formula = duree.interv ~ schizophrenie + depression + abus.subst +
    gravite + charactere + trauma.enfant + age + factor(type.centre),
    data = smp)

Residuals:
    Min      1Q  Median      3Q     Max
-38.578 -13.855  -1.769  10.922  64.405
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           42.48996    3.99998  10.623  < 2e-16 ***
schizophrenie          3.06420    2.80997   1.090 0.275911
depression             6.71269    1.64793   4.073 5.21e-05 ***
abus.subst             4.60037    1.79854   2.558 0.010760 *
gravite                1.06236    0.56548   1.879 0.060737 .
charactere             1.62547    0.93536   1.738 0.082723 .
trauma.enfant         -0.66805    1.66931  -0.400 0.689144
age                    0.20788    0.06066   3.427 0.000649 ***
factor(type.centre)2   4.09540    2.53401   1.616 0.106546
factor(type.centre)3  -1.29681    2.44159  -0.531 0.595509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.02 on 645 degrees of freedom
  (144 observations deleted due to missingness)
Multiple R-squared:  0.1086,    Adjusted R-squared:  0.09619
F-statistic: 8.734 on 9 and 645 DF,  p-value: 1.919e-12
```

```r
par(mfrow=c(1,2))
hist(residuals(mod), main="Histogramme des résidus", xlab="Résidus")
qqnorm(residuals(mod), main="QQ-plot des résidus")
qqline(residuals(mod), col="red")
```

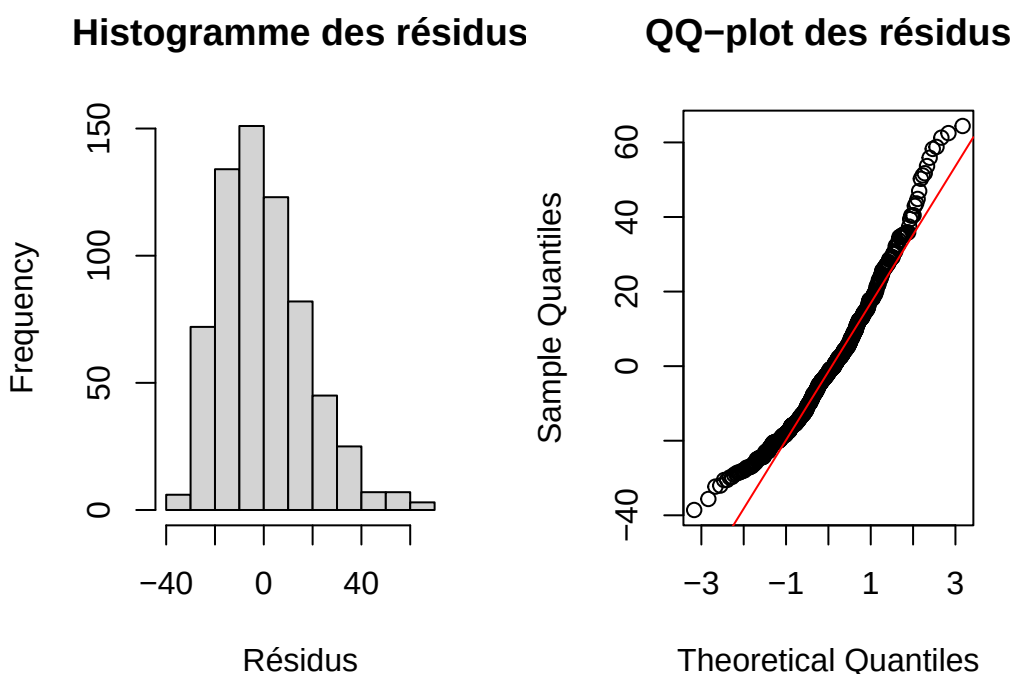


Figure 5.10: Histogramme et QQ-plot des résidus du modèle linéaire expliquant la durée de l'entretien. La distribution des résidus semble s'écarter de la normalité.

```r
par(mfrow=c(1,1))
```

On fait une estimation par *bootstrap* en négligeant la dépendance intra-centre.

```r
lm.boot <- function(data, index) {
        smp.boot <- data[index,]
        mod <- lm(
            duree.interv ~
            schizophrenie + depression + abus.subst + gravite + charactere +
↪  trauma.enfant +age + factor(type.centre),
            data = smp.boot)

    coefficients(mod)
}
```

Concernant la fonction R :

Il faut définir une fonction `lm.boot` :

- entrée : jeu de données `data`
- vecteur d'indices : ìndex'
- Estime les paramètres du modèle linéaire pour chaque échantillon bootstrap.
- Les coeficients sont stockés dans `coefficients(mod)`.

Puis utilisation de la fonction `boot()` du package `boot` pour réaliser le bootstrap.

- `smp` : jeu de données original
- `lm.boot` : fonction définie précédemment à "bootstrapper", c'est à dire à appliquer sur chaque échantillon bootstrap

Si l'on est plus spécialement intéressé par la variable « schizophrénie » il faut choisir le deuxième paramètre du tableau de résultats `[,2]` car la première colonne correspond à l'ordonnée à l'origine = intercept.

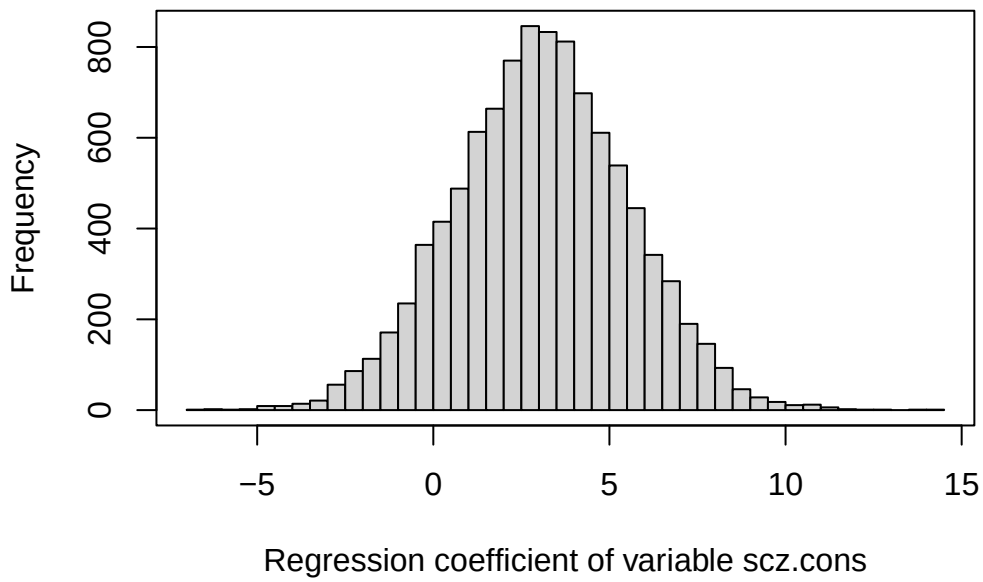


Figure 5.11: Histogramme des 10000 coefficients de régression de la variable « schizophrénie » estimés via bootstrap, c'est-à-dire issus de 10000 jeux de données obtenus par tirage au sort avec remise à partir du jeu de données original.

Pour obtenir l'IC à 95% de la variable « schizophrénie » et le p-value associée :

```
# index = 2 car on s'intéresse au 2ème coefficient (schizophrénie)
# type = "bca" pour méthode des percentiles corrigée (plus robuste sur petits
↪  échantillons)
boot.ci(resboot, index=2, type= "bca")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = resboot, type = "bca", index = 2)

Intervals :
Level       BCa
95%   (-1.737,  7.871 )
Calculations and Intervals on Original Scale
```

```
# p-value bilatérale
2*sum(resboot$t[,2]<=0)/10000
```

```
[1] 0.2168
```

6GLM 1 : Modèles linéaires généralisés

1 Introduction

1.A Retour sur le modèle linéaire simple

BAS : score quantitatif de dévoleppement cognitif

Dans ce tableau :

Durée de l'allaitement maternel	Moyennes (n)	Coefficients bruts (IC 95%) (n=10 944)	Coefficients partiellement ajustés (IC 95%) (n=10 929)	Coefficients complètement ajustés (IC 95%) (n=10 416)
Jamais	106.5 (3825)	Reference	Reference	Reference
< 2,0 mois	110.2 (2901)	3.7 (2.8–4.6)	0.7 (–0.1–1.5)	0.7 (–0.1–1.5)
2,0–3,9 mois	111.8 (1047)	5.4 (4.1–6.6)	1.2 (0–2.4)	1.2 (0–2.4)
4,0–5,9 mois	113.0 (889)	6.5 (5.1–7.9)	1.2 (0–2.5)	1.0 (–0.3–2.3)
6,0–11,9 mois	114.1 (1392)	7.7 (6.5–8.8)	2.2 (1.2–3.2)	2.0 (1.0–3.0)
≥ 12,0 mois	114.2 (890)	7.7 (6.6–8.9)	2.4 (1.3–3.5)	2.4 (1.3–3.6)

- Durée d'allaitement : variable explicative **catégorielle ordinale** en 6 niveaux : Jamais ; < 2 mois ; 2-3,9 mois ; 4-5,9 mois ; 6-11,9 mois ; ≥ 12 mois
- Estimation de l'effet de l'allaitement sur le score BAS
- Modèle avec coefficients "bruts" + partiellement ajustés + ajustés
- "Brut" : analyse univariable = régression linéaire simple
- "Ajustement" : analyse multivariable = régression linéaire multiple

Questions :

- Pourquoi le modèle brut est sur 10 944 et le modèle ajusté complet sur 10 416 ?
  - Probablement en raison des données manquantes pour les variables de confusion incluses dans le modèle ajusté complet.
- Pourquoi l'intervalle de confiance est diminué dans le modèle ajusté complet par rapport au modèle brut ?
  - L'ajustement pour les variables de confusion peut réduire la variabilité non expliquée, ce qui peut conduire à des intervalles de confiance plus étroits.
  - Même s'il ne s'agit pas d'une variable de confusion, l'ajout de variables explicatives pertinentes peut améliorer la précision des estimations.

- Pourquoi pas de p-value ?
  - Dans un modèle linéaire, les coefficients s'interprètent comme des différences de moyennes entre les groupes par rapport à la catégorie de référence.
  - C'est différent d'un modèle de régression logistique où les coefficients sont des log-odds ratios, et où les p-values sont souvent utilisées pour tester l'association.
  - 3,7 : différence de moyenne de score de BAS entre le groupe "< 2 mois" et le groupe "Jamais".
- Ce ne sont pas des odds ratios ?
  - Non, dans un modèle linéaire, les coefficients représentent des différences de moyennes, pas des rapports de cotes.
  - Dans un modèle linéaire, la valeur neutre = 0 (pas de différence de moyenne).
  - **Mesures d'associations (OR, RR, HR) sont des rapports dont la valeur neutre = 1.**
- Pourquoi des catégories d'âge maternelle et pas l'âge en continu ?
  - Utilisation de catégories peut faciliter l'interprétation des résultats.
  - Utilisation de l'âge en continue reviendrait à faire l'hypothèse d'une relation linéaire entre l'âge maternel et le score BAS, ce qui peut ne pas être approprié.
- Si on devait avoir une variable quantitative dont il est difficile d'avoir un seuil
  - Binarisation en optimisant un seuil (Youden)
  - Binarisation par la médiane
  - Catégories d'âge

1.B Modèles linéaires généralisés (GLM)

- Extension des modèles linéaires pour permettre de modéliser des variables dépendantes qui ne suivent pas une distribution normale.
- Permettent de modéliser des relations entre une variable dépendante et une ou plusieurs variables indépendantes, tout en tenant compte de la nature spécifique de la variable dépendante.
- Composants principaux des GLM :
  - **Fonction de lien** : établit une relation entre la moyenne de la variable dépendante et les variables indépendantes.
  - **Distribution de la famille exponentielle** : spécifie la distribution de la variable dépendante (ex. binomiale, Poisson, etc.).
- Exemples courants de GLM :
  - Régression logistique : pour les variables dépendantes binaires (oui/non).
  - Régression de Poisson : pour les variables dépendantes de comptage.
  - Régression gamma : pour les variables dépendantes continues et positives.

1.B.1 Modèle logistique

Reprend presque toutes les caractéristiques de la régression linéaire, mais est adapté pour les variables dépendantes binaires.

- possibilité d'obtenir des coefficients bruts ou ajustés
- codage des variables explicatives catégorielles
- interprétation des résultats

SAUF

- La variable expliquée (critère de jugement) est **BINAIRE** (oui/non, succès/échec, malade/-pas malade)
- La force de l'association est mesurée par des **odds ratios (OR)** au lieu de différences de moyennes

> ❗ Important
>
> **PLAN DU COURS**
>
> 1. Mesures d'association : RR et OR
> 2. Modèle logistique univariable
> - Interprétation des coefficients
> - Test statistiques sur les coefficients
> 3. Modèle logistique multivariable
> - Notion d'interaction
> - Construction de modèles multivariables

2 Définition des risques

2.A Exemple

- 256 patients avec transplantation de moelle osseuse
- Variable à expliquer Y : rejet de la greffe (oui/non) = variable binaire dichotomique
- Covariable X : dose faible (X = 1) vs dose élevée (X = 0) d'un médicament immunosuppresseur

Étude du lien entre dose de médicament et risque de rejet de la greffe.

Tableau de contingence :

	Rejet (Y=1)	Pas de rejet (Y=0)	Total
Dose faible (X=1)	37	71	108
Dose élevée (X=0)	31	117	148
Total	68	188	256

	Rejet (Y=1)	Pas de rejet (Y=0)	Total

- Pourcentage de rejet dans chaque groupe :
  - Dose faible : 37/108 ≈ 34%
  - Dose élevée : 31/148 ≈ 21%
- Comparaison de pourcentages = test du chi2

```
# Tableau de contingence
tableau <- matrix(c(37, 71, 31, 117), ncol = 2)
colnames(tableau) <- c("Rejet", "Pas de rejet")
rownames(tableau) <- c("Dose faible", "Dose élevée")
tableau
```

```
            Rejet Pas de rejet
Dose faible    37           31
Dose élevée    71          117
```

```
# Test du chi2
chisq.test(tableau, correct = FALSE)
```

```
    Pearson's Chi-squared test

data:  tableau
X-squared = 5.6732, df = 1, p-value = 0.01723
```

Syntaxe :

- `matrix(c(a, b, c, d), ncol = 2)` : création du tableau de contingence
- `chisq.test(tableau, correct = FALSE)` : test du chi2
  - correct = FALSE : pas de correction de Yates pour les petits effectifs

Interprétation des résultats :

- LIEN significatif
- mais pas de mesure de la force
- pour mesurer la force : on pourrait faire 34/21 = 1,62 = RR

Questions :

- Est-ce une mesure de la force de l'association entre dose et rejet ?
  - Non, le test du chi2 indique s'il y a une association statistiquement significative, mais ne quantifie pas la force de cette association.
- Est-ce une mesure de la signification statistique ?

  - Oui, le test du chi2 fournit une p-value qui indique si l'association observée est statistiquement significative.
- Est-ce que ça dépend de l'effectif ?
  - Oui, la puissance du test du chi2 dépend de la taille de l'échantillon. Des échantillons plus grands peuvent détecter des associations plus faibles.

2.B Mesures d'associations

Risque : soit un facteur de risque X (Oui = 1, Non = 0) et un événement Y

- R1 = P(Y=1 | X=1) = probabilité que l'événement se produise chez les sujets exposés au facteur de risque X=1
- Dans l'exemple, R1 = 37/108 = 0,3426 (risque que le rejet se produise dans le groupe dose faible)
- R0 = P(Y=1 | X=0) = probabilité que l'événement se produise chez les sujets non exposés au facteur de risque X=0
- Dans l'exemple, R0 = 31/148 = 0,2095 (risque que le rejet se produise dans le groupe dose élevée)

Ratio de risque = Risque Relatif (RR)

- $RR = R_1/R_0$.
- Dans le cas de l'exemple au dessus :
  - R_1 = 37/108 = 0,3426 (risque de rejet dans le groupe dose faible)
  - R_0 = 31/148 = 0,2095 (risque de rejet dans le groupe dose élevée)
  - RR = R_1 / R_0 = 0,3426 / 0,2095 = 1,61 / R0 = (37/108) / (31/148) = 1,62

Cote ou Odds :

- $Odds = R/(1 - R)$ = probabilité que l'événement se produise / probabilité que l'événement ne se produise pas
- Par exemple si R vaut 0,75 (risque = 75%)
- O = 0,75 / (1-0,75) = 3 donc la cote est de 3 (3 fois plus de chances que l'événement se produise que de ne pas se produire)

Rapport de cotes = Odds Ratio (OR)

- $OR = \frac{R_1/1-R_1}{R_0/1-R_0}$

RR et OR = 2 manières différentes de mesurer l'association entre variables binaires

2.C Interprétation des risques

RR	OR
• Facilement interprétable – Facteur par lequel le risque de Y est multiplié si le facteur X est présent • Pas estimable pour tous les types d'études épidémiologiques (notamment pas pour cas témoins) • Moins bonnes propriétés statistiques et mathématiques que l'OR	• Interpétation moins immédiate (rapport des cotes souvent confondus avec RR) • Obtenus dans analyses statistiques multivariées • Estimable dans les enquêtes cas-témoin (maladie à prévalence non estimable)

2.C.1 Exemple

- 100 hommes : 90 ont bu du vin la semaine précédente
  - Odds d'un homme d'avoir bu du vin : 9:1
- 100 femmes : 20 ont bu
  - Odds : 20/80 : 1:4
- OR = 9/0,25 = 36 donc peu interprétable
- RR = 0.9/0.2 = 4,5

NB : l'OR est symétrique mais pas le RR !

- OR de ne pas boire de vin : 0.11/4 = 1/36 = 1/OR d'avoir bu
- RR de ne pas avoir bu : 0.1/0.8 = 1/8 = 0.125 ≠ de 1/RR !!

2.D Calcul des OR et RR dans un tableau de contingence

	Rejet (Y=1)	Pas de rejet (Y=0)	Total
Dose faible (X=1)	a	b	a+b
Dose élevée (X=0)	c	d	c+d
Total	a+c	b+d	n

Risques :

- Risque dans le groupe exposé (X=1) : $R_1 = \frac{a}{a+b}$
- Risque dans le groupe non exposé (X=0) : $R_0 = \frac{c}{c+d}$
- Risque relatif (RR) : $RR = \frac{R_1}{R_0} = \frac{a/(a+b)}{c/(c+d)}$

Côtes :

- Cote dans le groupe exposé (X=1) : $O_1 = \frac{a}{b}$

- Cote dans le groupe non exposé (X=0) : $O_0 = \frac{c}{d}$
- Odds ratio (OR) : $OR = \frac{R_1/(1-R_1)}{R_0/(1-R_0)} = \frac{a/b}{c/d} = \frac{ad}{bc}$

> **Tip**
>
> Imaginons qu'on multiplie les effectifs des réponse = oui
>
> - OR : $\frac{(2a)(d)}{(b)(2c)} = \frac{2a*d}{b*2c}$ donc OR ne change pas (un 2 de chaque côté se simplifie)
> - Donc OR beaucoup plus utilisé ! (peut être calculé dans bcp + de cas)
> - RR : $\frac{(2a)/(2a+b)}{(2c)/(2c+d)}$ donc RR change (le 2 ne se simplifie pas)
> - Le RR dépend de la proportion rejet / non rejet !!

2.D.1 Exemple numérique avec R

```r
# Effectifs
a <- 37  # Rejet, Dose faible
b <- 31  # Pas de rejet, Dose faible
c <- 71  # Rejet, Dose élevée
d <- 117 # Pas de rejet, Dose élevée

#création et affichage tableau de contingence
tableau <- matrix(c(a, b, c, d), ncol = 2)
colnames(tableau) <- c("Rejet", "Pas de rejet")
rownames(tableau) <- c("Dose faible", "Dose élevée")
tableau
```

```
            Rejet Pas de rejet
Dose faible    37           71
Dose élevée    31          117
```

```r
# Calcul des OR et RR
RR <- a/(a+c)/(b/(b+d))
OR <- (a*d)/(b*c)

# Affichage des résultats
OR
```

```
[1] 1.966833
```

```
RR
```

```
[1] 1.635603
```

OR = 1.966833

RR = 1.635603

> **Tip**
>
> - A-t-on toujours OR > RR ?
> - Oui, sauf lorsque l'événement est rare (risque faible), où OR et RR peuvent être proches.
> - Dans tous les cas **plus éloigné de 1** que le RR.

2.D.2 Intervalles de confiance des OR et RR

- Calcul des intervalles de confiance (IC) pour les OR et RR

$$SE = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

3 Régression logistique univariable

3.A Introduction

Exemple.1 : on veut étudier le relation entre pathologie et plusieurs FDR

Exemple 2 : on veut étudier le lien entre survenue d'un évènement et plusieurs évènements intercurrents

- Variable dépendante Y :
  - binaire dichotomique (0/1)
  - Loi de Bernoulli = success/failure
- Variables explicatives x1, x2, …, xp :
  - quantitatives et/ou qualitatives
  - peuvent être continues ou discrètes
- On a des données chez n patients numérotés i = 1, 2, …, n
  - Y_i : réponse binaire pour le patient i
  - $x_{i1}, x_{i2}, \ldots, x_{ip}$: covariables explicatives pour le patient i

3.B Transformation logistique

- Au lieu de modéliser le lien entre Y_i et les x_i directement, on veut un modèle pour $p_i = E(Y = 1/X = x_i)$
  - C'est à dire la probabilité que $Y_i = 1$ sachant les valeurs des covariables explicatives pour le patient i
  - Exemple : dose plus élevée associée à une plus grande probabilité de réponse
- p_i n'est pas directement observée
  - Peut être estimée chez tous les patients avec $X = x_i$ si variable catégorielle ou groupements
- Problème : $p_i = P(Y_i = 1 \| X = x_i)$ est une probabilité
  - Doit être comprise entre 0 et 1 (donc les modèles linéaires ne sont pas adaptés car supposent la normalité des résidus)
  - Relation linéaire entre p_i et les x_i peut donner des valeurs en dehors de [0, 1]
- Solution : transformation logistique (fonction logit)
  - $q = logit(p) =$ ln $\left(\frac{p}{1-p}\right)$
  - même transformation que R/(1-R) = Odds
  - prendre le log permet d'avoir quelque chose entre -∞ et +∞
- Fonction expit (inverse du logit) :
  - $p = expit(q) = \frac{e^q}{1+e^q} = \frac{1}{1+e^{-q}}$
  - permet de revenir à une probabilité entre 0 et 1

3.C Modèle logistique

On modélise logit de $p = P(Y = 1/X = x)$ comme une combinaison linéaire des variables explicatives x1, x2, …, xp

$logit(p) = \alpha + \beta x$

Transformation inverse (expit) :

$p = \frac{1}{1+exp(-\alpha+\beta x)} = \frac{exp(\alpha+\beta x)}{1+exp(\alpha+\beta x)}$

3.C.1 Si X est une variable binaire (0/1) :

- quand X = 0, $logit(p_0) = alpha = logit(R_O$
- quand X = 1, $logit(p_1) = alpha + beta = logit(R_1)$
- OR =

- $= \frac{R_1/(1-R_1)}{R_0/(1-R_0)}$ ou $ln(OR) = logit(R_1) - logit(R_0) = beta$)
- $= \frac{exp(alpha+beta)}{exp(\quad alpha)}$
- $= exp(beta)$

Valeur neutre pour l'OR = 1 donc valeur neutre pour β = 0

3.C.2 Si X est une variable catégorielle à 3 valeurs (0, 1, 2) :

- Équivalent à 2 variables binaires indicatrices :
  - $x_1 = 1$ si X = 1, 0 sinon
  - $x_2 = 1$ si X = 2, 0 sinon
- Modèle à 2 variables :
  - $logit(p) = alpha + beta_1 x_1 + beta_2 x_2$
  - quand X = 0, $logit(p_0) = alpha = logit(R_0)$
  - quand X = 1, $logit(p_1) = alpha + beta_1 = logit(R_1)$
  - quand X = 2, $logit(p_2) = alpha + beta_2 = logit(R_2)$
- $OR_1 = exp(beta_1) vs 0$
- $OR_2 = exp(beta_2) vs 0$
- OR de 2 vs 1 = $exp(beta_2)/exp(beta_1) = OR2/OR1$

3.C.3 Si X est une variable catégorielle à K valeurs

- Définir une classe de référence (X=0)
- Il y aura K-1 OR par rapport à la classe de référence

Attention au choix de la classe de référence !

Atention aux regroupements de catégories (si faibles effectifs)

3.C.4 Si X est une variable continue / quantitative

Hypothèse : relation linéaire de logit(p) avec x (sur le logOR)
$f(x) = P(Y = 1/X = x = \frac{1}{1+exp(-(\alpha+\beta x))}$

- Risque que Y = 1 sachant que X prend la valeur x

Entre deux valeurs x1 et x2 :

- $ln(OR) = \beta(x_1 - x_0)$; OR = $exp(\beta(x_1 - x_0))$

NB : si pour des variables QUANTITATIVES, l'OR a une UNITÉ

- Pour 1 an d'âge
- pour 10mg de morphine.

3.D Fonction de vraisemblance

- Estimation des paramètres du modèle (α, β1, β2, …, βp) par la méthode du maximum de vraisemblance
- Vraisemblance = probabilité d'observer les données telles qu'elles ont été observées, en fonction des paramètres du modèle
- On cherche les valeurs des paramètres qui maximisent cette probabilité
- Ordinateur cherche les meilleur coefficient α et β qui maximisent la vraisemblance

3.E Test de Wald

Test de Wald = Basé sur le logOR

- logciels fournissent β et l'erreur d'estimation SE(β) (erreur standard)
- H0 = β = 0 (OR = 1)
  - β/se(β) suit une loi normale centrée réduite N(0,1)
  - comparaison de β/se(β) à 1,96 pour un test bilatéral au seuil 5%
  - calcul du IC à partir de β et se(β) = β ± 1,96*se(β)

Intervalle de confiance sur OR

- on prend l'exponentielle des bornes de l'IC pour b : [exp(b1) ; exp(b2)]
- intervalle non symétrique autour de exp(b)
- Si IC pour β contient 0, IC pour OR contient 1 alors rejet de H0

3.F Exemple avec R

Exemple du rejet de greffe

```r
# Données
beta <- log(OR)
beta
```

```
[1] 0.6764248
```

```r
# Estimation de l'erreur standard
se_beta <- sqrt(1/37 + 1/71 + 1/31 + 1/117)
se_beta
```

```
[1] 0.2862108
```

```r
# Test de Wald
beta/se_beta
```

```
[1] 2.36338
```

```r
# ici > 1,96 donc on rejette H0

# pvalue
2 * (1 - pnorm(abs(beta/se_beta)))
```

```
[1] 0.0181091
```

- OR = 1.96
- β = log(OR) = 0,68 (b = coefficient)
- se(β) = 0,29 (écart-type de l'estimateur)
- Rapport coeffcient/écart-type : b/se(b) = 2,36 > 1,96 donc on rejette H0

Il y a donc un lien significatif entre infection et décès

Calcul de l'IC à 95% pour OR

```r
# Calcul des bornes de l'IC pour b
beta1 <- b - 1.96 * se_beta
beta2 <- b + 1.96 * se_beta
beta1
```

```
[1] 30.43903
```

```r
beta2
```

```
[1] 31.56097
```

```r
#exponentielle des bornes pour obtenir l'IC pour OR
ic_or <- c(exp(beta1), exp(beta2))
ic_or
```

```
[1] 1.657683e+13 5.090453e+13
```

Donc IC à 95%

- pour β : 0,68 ± 1,96 * 0.29 = [0,11 ; 1,24]
- pour OR : exponentielle de l'IC de β [1,12 ; 3,45].

4 Régression logistique multivariable

4.A Principe

4.A.1 Principe général

Variable dichotomique Y dont on veut étudier le lien avec p variables continues ou discrètes : x_1, x_2, ..., x_p
On dit que Y est

- la variable dépendante
- la variable à expliquer

On dit que les X_k sont

- les variables indépendantes
- les variables explicatives
- les covariables

Extension de la régression logistique univariable :
$logit(P(Y = 1 \| x_1, ..., x_p)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$

4.A.2 Hypothèse de multiplicativité des effets

Le modèle logistique multivarié suppose la **multiplicativité** des effets des facteurs de risque
Cas particulier : deux facteurs de risque X_1 et X_2 prenant deux valeurs (0/1) (ex tabac et alcool et cancer du poumon)

- OR pour X_1 = 1 vs 0 : $OR_1 = exp(\beta_1)$
- OR pour X_2 = 1 vs 0 : $OR_2 = exp(\beta_2)$
- OR pour X_1 = 1 et X_2 = 1 = $OR_{1,2} = exp(\beta_1 + \beta_2) = exp(\beta_1) * exp(\beta_2) = OR_1 * OR_2$

Si OR de cancer de poumon de 3 pour tabac et de 2 pour alcool ; OR pour tabac + alcool = 3 * 2 = 6

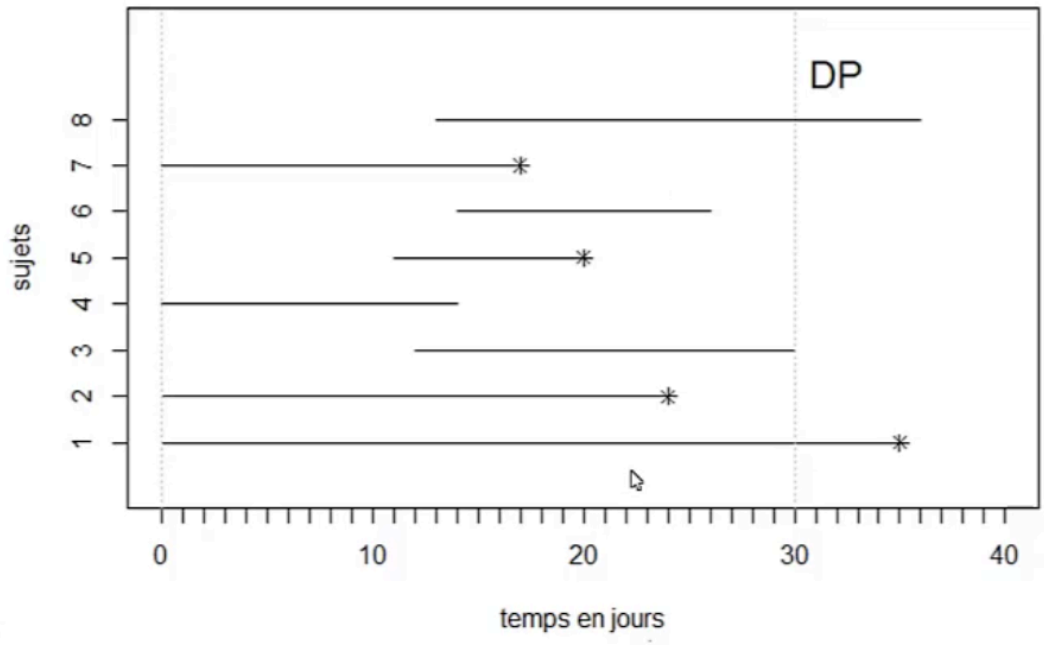
Mais il est possible que l'effet combiné soit différent de la multiplication des effets individuels (interaction) !!!

Dans ce cas, il faut ajouter un terme d'interaction dans le modèle

On peut très bien faire un tableau avec les β et les $se(\beta)$!

- Dans ce cas : les **coefficients** sont comparés à 0
- Pour les variables continues exprimer les coefficients et les OR **DANS L'UNITÉ** d'origine !
- **IC de l'OR :** $exp(\beta \pm 1,96 * se(\beta))$

A partir du graphique déterminer le temps de participation T_k pour chaque patient et l'état I_k en T_k.



16/11/20 13

4.A.3 Notion d'interaction

Interaction entre deux variables explicatives X_1 et X_2 :

$logit(P(Y = 1 \| x_1, x_2)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 * x_2)$

Rajoute un coefficient et lève l'hypothèse de multiplicativité des OR

> Tip
>
> **Définitions essentielles : β, estimateur, estimation**
> Dans un modèle logistique, on étudie l'association entre une variable explicative X (par exemple, dose d'un traitement) et la probabilité d'un événement Y (par exemple, rejet de greffe). Le modèle s'écrit :

logit$(p) = \alpha + \beta X$
$p = P(Y = 1 \mid X)$
Dans cette équation :

- α est une constante appelée intercept ;
- β est un paramètre du modèle, qui mesure l'effet de X sur la probabilité de l'événement.

Un paramètre est une quantité théorique, supposée fixe (c'est à dire constante), qui décrit la relation vraie dans la population étudiée. Dans le modèle logistique, on montre que :
$\beta =$ log(OR),
OR $= e^{\beta}$.
Ainsi :

- si $\beta = 0$, alors OR = 1 : il n'existe pas d'association entre X et Y ;
- si $\beta > 0$, alors OR > 1 : présence d'un effet augmentant le risque ;
- si $\beta < 0$, alors OR < 1 : présence d'un effet protecteur.

En pratique, on ne connaît jamais la vraie valeur de β.
On cherche donc à l'estimer à partir d'un échantillon.
Un estimateur est une méthode de calcul permettant d'approcher un paramètre à partir des données.
La valeur obtenue par cet estimateur dans un échantillon donné est appelée estimation.
Dans la régression logistique, l'estimateur utilisé pour β est l'estimateur du maximum de vraisemblance, noté $\hat{\beta}$ ("bêta chapeau").
La valeur numérique affichée par le logiciel (par exemple, 0.68) est l'estimation de β dans l'échantillon étudié.
□
Principe du maximum de vraisemblance : comment obtient-on $\hat{\beta}$?
Pour chaque individu i, le modèle logistique calcule une probabilité prédite de l'événement :
$p_i = P(Y_i = 1 \mid X_i) = \frac{e^{\alpha+\beta X_i}}{1+e^{\alpha+\beta X_i}}$.
Si $Y_i = 1$, la probabilité d'observer cette valeur est p_i. Si $Y_i = 0$, cette probabilité est $1 - p_i$.
On définit alors la vraisemblance totale des paramètres α et β comme la probabilité d'observer l'ensemble des données :
$L(\alpha, \beta) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}$.
Le principe du maximum de vraisemblance consiste à choisir les valeurs $\hat{\alpha}$ et $\hat{\beta}$ qui maximisent cette probabilité.
Autrement dit, on sélectionne les paramètres qui rendent les données observées les plus plausibles selon le modèle.
Pour des raisons numériques, on maximise généralement la log-vraisemblance :
$\ell(\alpha, \beta) = \sum_{i=1}^{n} \left[Y_i \log(p_i) + (1 - Y_i) \log(1 - p_i)\right]$,
ce qui est équivalent à maximiser la vraisemblance elle-même.

Le logiciel recherche alors les valeurs de α et β qui maximisent $\ell(\alpha, \beta)$. La valeur obtenue pour β est l'estimation $\widehat{\beta}$.
□
Erreur standard : quantifier l'incertitude sur l'estimation $\widehat{\beta}$
Si l'on refaisait l'étude plusieurs fois sur des échantillons différents, on n'obtiendrait pas toujours la même estimation $\widehat{\beta}$. L'estimateur $\widehat{\beta}$ varie d'un échantillon à l'autre : il s'agit d'une variable aléatoire.
L'erreur standard SE$(\widehat{\beta})$ mesure l'ampleur typique de cette variabilité. Elle correspond à la racine carrée de la variance théorique de l'estimateur. Plus SE$(\widehat{\beta})$ est petit, plus l'estimateur est précis.
Cette erreur standard est calculée (ou approximée) automatiquement par le logiciel à partir de la dérivée seconde de la log-vraisemblance (matrice d'information de Fisher).
□
Standardisation de l'estimateur et loi normale centrée réduite
Dans les modèles logistiques, et plus largement dans les modèles estimés par maximum de vraisemblance, un résultat fondamental est que, sous certaines conditions (échantillon suffisamment grand, modèle correctement spécifié), l'estimateur $\widehat{\beta}$ suit approximativement une loi normale de moyenne β et de variance SE$(\widehat{\beta})^2$.
Sous l'hypothèse nulle $H_0 : \beta = 0$, on a donc :
$\widehat{\beta} \approx \mathcal{N}(0, \text{SE}(\widehat{\beta})^2)$.
Pour exploiter ce résultat, on standardise l'estimateur, c'est-à-dire qu'on le transforme en une variable sans unité, mesurant la distance à 0 en nombre d'erreurs standard :
$Z = \frac{\widehat{\beta}-0}{\text{SE}(\widehat{\beta})} = \frac{\widehat{\beta}}{\text{SE}(\widehat{\beta})}$. Cette variable Z suit alors approximativement une loi normale centrée réduite $\mathcal{N}(0, 1)$, c'est-à-dire une loi normale de moyenne 0 et de variance 1.
Cette propriété est à la base du test de Wald.
□
Test de Wald : définition, interprétation, limites
Principe du test
On souhaite tester :
$H_0 : \beta = 0$
vs
$H_1 : \beta \neq 0.$
Le test de Wald utilise la statistique :
$W = \frac{\widehat{\beta}}{\text{SE}(\widehat{\beta})}$.
Sous $H_0 : \beta = 0$, W suit approximativement une loi normale centrée réduite. Pour un test bilatéral au seuil de 5 %, on rejette H_0 si :
$|W| > 1.96.$
La p-value se déduit de la loi normale. L'intervalle de confiance de β est :
$\widehat{\beta} \pm 1.96 \times \text{SE}(\widehat{\beta})$.
L'intervalle de confiance de l'odds ratio s'obtient en exponentiant les bornes.
Limites et précautions sur le test de Wald*
Le test de Wald repose sur des approximations asymptotiques. Il peut être peu fiable dans plusieurs situations : 1. Échantillons de petite taille

```
-   L'approximation normale est mauvaise et le test peut être trop libéral ou trop conservate
```

2. Événements rares
   - Si l'événement étudié est très rare, $\widehat{\beta}$ peut avoir une distribution très asymétrique. Les intervalles de confiance peuvent être trop étroits ou décalés.
3. Effets très importants (odds ratios très élevés)
   - Dans ces cas, la distribution de $\widehat{\beta}$ est fortement asymétrique. Les intervalles basés sur Wald peuvent être inadaptés.
4. Séparation complète ou quasi-complète
   - Si une covariable prédit parfaitement l'événement (par exemple, tous les événements surviennent dans un seul groupe), $\widehat{\beta}$ "diverge" et le test de Wald n'est plus utilisable.

Dans ces situations, il est souvent préférable d'utiliser des méthodes alternatives comme la régression logistique pénalisée (méthode de Firth), des tests exacts, ou le test du rapport de vraisemblance.
□

Test du rapport de vraisemblance (Likelihood Ratio Test)
Le test du rapport de vraisemblance (LRT) compare deux modèles emboîtés : un modèle réduit (sans la covariable testée) et un modèle complet (avec la covariable).
Pour tester $\beta = 0$, on compare :

- modèle réduit : logit$(p) = \alpha$;
- modèle complet : logit$(p) = \alpha + \beta X$.

Le logiciel calcule la log-vraisemblance maximale des deux modèles :

- $\ell_{\text{réduit}}$: $modlesansX$;
- ℓ_{complet} : $modleavecX$.

La statistique de test est :
$\Lambda = 2(\ell_{\text{complet}} - \ell_{\text{réduit}})$.
Sous $H_0 : \beta = 0$ et sous des conditions générales, Λ suit une loi du χ^2 avec un degré de liberté égal au nombre de paramètres testés (ici, 1). On en déduit une p-value.
Le LRT présente plusieurs avantages :

- il repose sur la comparaison globale des modèles et non sur l'approximation locale autour de $\widehat{\beta}$;
- il est plus robuste que le test de Wald lorsque l'échantillon est petit, lorsque les événements sont rares, ou lorsque l'effet étudié est important ;
- il permet de tester simultanément plusieurs coefficients (par exemple, pour des variables qualitatives à plusieurs modalités).

Résumé

- β est le paramètre du modèle logistique correspondant au log(odds ratio).
- $\hat{\beta}$ est son estimation obtenue par maximum de vraisemblance, en choisissant les valeurs qui rendent les données observées les plus probables.
- L'erreur standard de $\hat{\beta}$ mesure la variabilité de l'estimation d'un échantillon à l'autre.
- Le test de Wald repose sur la standardisation de $\hat{\beta}$, qui suit approximativement une loi normale centrée réduite sous l'hypothèse nulle.
- Ce test peut être peu fiable en cas de petits échantillons, d'événements rares ou d'effets extrêmes.
- Le test du rapport de vraisemblance compare un modèle avec et sans la covariable, en s'appuyant sur la log-vraisemblance. Il est souvent plus robuste et peut être utilisé pour tester plusieurs coefficients simultanément.

7GLM 1 TP : Modèles linéaires généralisés

1 Consigne

1. Construire une table descriptive de cet échantillon
2. Proposer un ou plusieurs modèles de régression logistique pour étudier les facteurs associés à la démence et rapportez les résultats sous forme de table ou figure pouvant apparaître dans un article scientifique.

Base de données de 373 patients atteints ou non d'une démence de type Alzheimer.

La colonne "Group" contient cette information sous forme d'une chaîne de caractères.

Les autres variables importantes sont :

- Age : contient l'âge en année
- M.F. : contient le sexe (M/F)
- EDUC : contient le niveau d'éducation en année
- SES : contient le statut socio-économique (1 à 5)
- MMSE : contient le score au Mini-Mental State Examination (0 à 30)
- CDR : contient le Clinical Dementia Rating (0 à 5)
- eTIV : contient le volume intracrânien estimé en mL
- nWBV : contient la proportion du volume intracrânien occupée par le cerveau

2 Question 1 = Table descriptive

```r
head(alzh)
```

```
        Group M.F Age EDUC SES MMSE CDR eTIV  nWBV   ASF
1 Nondemented   M  87   14   2   27 0.0 1987 0.696 0.883
2 Nondemented   M  88   14   2   30 0.0 2004 0.681 0.876
3    Demented   M  75   12  NA   23 0.5 1678 0.736 1.046
4    Demented   M  76   12  NA   28 0.5 1738 0.713 1.010
5    Demented   M  80   12  NA   22 0.5 1698 0.701 1.034
6 Nondemented   F  88   18   3   28 0.0 1215 0.710 1.444
```

```r
levels(alzh$Group)
```

```
NULL
```

```r
str(alzh)
```

```
'data.frame':   373 obs. of  10 variables:
 $ Group: chr  "Nondemented" "Nondemented" "Demented" "Demented" ...
 $ M.F  : chr  "M" "M" "M" "M" ...
 $ Age  : int  87 88 75 76 80 88 90 80 83 85 ...
 $ EDUC : int  14 14 12 12 12 18 18 12 12 12 ...
 $ SES  : int  2 2 NA NA NA 3 3 4 4 4 ...
 $ MMSE : int  27 30 23 28 22 28 27 28 29 30 ...
 $ CDR  : num  0 0 0.5 0.5 0.5 0 0 0 0.5 0 ...
 $ eTIV : int  1987 2004 1678 1738 1698 1215 1200 1689 1701 1699 ...
 $ nWBV : num  0.696 0.681 0.736 0.713 0.701 0.71 0.718 0.712 0.711 0.705 ...
 $ ASF  : num  0.883 0.876 1.046 1.01 1.034 ...
```

```r
unique(alzh$Group)
```

```
[1] "Nondemented" "Demented"    "Converted"
```

```r
table(alzh$Group)
```

```
  Converted    Demented Nondemented
         37         146         190
```

2.A Recode de la variable "Group"

On recode la variable "Group" pour qu'elle devienne binaire et prenne la valeur 1 si le patient est atteint de démence et 0 sinon.

```r
alzh = read.csv("~/Documents/Projets/M2biostatistiques/Cours/alzheimer.csv")
alzh$Group = 1*(alzh$Group %in% c("Demented","Converted"))
table(alzh$Group)
```

```
  0   1
190 183
```

Syntaxe :

- `1*(condition)` : crée une variable binaire qui prend la valeur 1 si la condition est vraie et 0 sinon. (conversion logique en numérique)
- `condition` : `df$var %in% c("val1","val2")` vérifie si les valeurs de la variable `var` dans le dataframe `df` appartiennent au vecteur `c("val1","val2")`. Contient TRUE ou FALSE.

- l'opérateur `%in%` est utilisé pour vérifier l'appartenance d'une valeur à un ensemble de valeurs.

On aurait pu faire :

```
alzh = read.csv("~/Documents/Projets/M2biostatistiques/Cours/alzheimer.csv")
condition_validation <- c("Demented","Converted")
alzh$Group = 1*(alzh$Group %in% condition_validation)
table(alzh$Group)
```

```
  0   1 
190 183 
```

ou utiliser la fonction `ifelse()` avec :

- `ifelse(test, yes, no)` : retourne `yes` si `test` est vrai et `no` sinon.

```
alzh = read.csv("~/Documents/Projets/M2biostatistiques/Cours/alzheimer.csv")
alzh$Group = ifelse(alzh$Group %in% c("Demented","Converted"), 1, 0)
table(alzh$Group)
```

```
  0   1 
190 183 
```

2.B Construction du tableau

2.B.1 Version nullach !!!

```
median(alzh$Age)
```

```
[1] 77
```

```
quantile(alzh$Age, probs = c(0.25, 0.75))
```

```
25% 75% 
 71  82 
```

```
median(alzh$Age[alzh$Group==1])
```

```
[1] 76
```

```r
mm = median(alzh$Age[alzh$Group==0])
```

Si on veut afficher entre tirets

```r
iqr = paste(quantile(alzh$Age, probs = c(0.25, 0.75)), collapse = " - ")
```

Si on veut faire stylé :

```r
paste(mm," (", iqr, ")", sep = "")
```

```
[1] "77 (71 - 82)"
```

Et mettre ça dans une fonction :

- function(gr,v) : gr = groupe (0 ou 1), v = variable (Age, EDUC, etc..)
- median(alzh$Age[alzh$Group==gr]) : calcule la médiane de la variable Age pour le groupe gr
- quantile(alzh$Age, probs = c(0.25, 0.75)) : calcule le 1er et 3ème quartile de la variable Age
- paste(…, collapse = " - ") : concatène les deux quartiles avec un tiret entre eux
- paste(mm," (", iqr,")", sep =" ") : crée une chaîne de caractères avec la médiane et l'IQR formatée

```r
#ajouter la fonction à l'environnement
report_quanti = function(v, gr) {
    x = alzh[[v]][alzh$Group == gr] #groupe = gr car on veut séparer les deux
 ↪  groupes
    mm  = median(x, na.rm = TRUE)
    iqr = paste(quantile(x, probs = c(0.25, 0.75), na.rm = TRUE),
                collapse = " - ")
    paste(mm, " (", iqr, ")", sep = "")
}
# il faut mettre v entre double crochet car on veut accéder à la variable dont le
 ↪  nom est stocké dans v
# si on met qu'un seul crochet, R cherche une variable nommée "v" dans le
 ↪  dataframe alzh
```

Tip

Pourquoi il faut mettre un double crochets ?
Les dataframe se comportent à la fois comme des listes, à la fois comme des matrices

```r
class(alzh)
```

```
[1] "data.frame"
```

comportement de matrice :

```
#avec un seul crochet, accès par nom de colonne ou par indice
alzh[1,3]
```

```
[1] 87
```

comportement de liste :

```
#avec double crochet, accès du contenu
alzh[[1]] #contenu de la première variable
```

```
  [1] 0 0 1 1 1 0 0 0 0 0 1 1 1 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 0 0 0 0 1 1 1 1
 [38] 1 1 1 1 0 0 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 1 1
 [75] 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 0 1 1 0 0 0 0
[112] 0 0 0 1 1 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 0 0 0
[149] 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 0 0 0 0 1 1
[186] 1 1 1 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 1 1 1 1
[223] 1 0 0 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 1 0
[260] 0 0 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 0 0 0 0 0 0 0 1
[297] 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 0 1 1 1 1 0 0 0 1 1 1 1 1 1
[334] 0 0 0 0 0 1 1 0 0 0 1 1 1 1 1 1 0 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 1 1 1 1 1
[371] 0 0 0
```

Pour appliquer à toutes les variables quantitatives de la base : Age, EDUC, SES, MMSE, CDR, eTIV, nWBV

```
vars = c("Age","EDUC","SES","MMSE","CDR","eTIV","nWBV")
```

Utilisation de la fonction `sapply()` pour appliquer `report_quanti` à chaque variable dans `vars` pour les deux groupes (0 et 1)

- Syntaxe : 2 arguments : `sapply(X, FUN)`
  - `X` : vecteur d'éléments à itérer (ici `vars`)
  - `FUN` : fonction à appliquer (ici une fonction anonyme qui prend `v` et appelle `report_quanti` pour les deux groupes)

```
sapply(vars, report_quanti, gr=0) #groupe 0 (il faut le donner parce que la
↪  fonction report_quanti en a besoin)
```

```
                Age                          EDUC
"77 (71 - 82)"                "16 (13 - 18)"
                SES                          MMSE
   "2 (2 - 3)"                "29 (29 - 30)"
                CDR                          eTIV
```

```
                 "0 (0 - 0)" "1474.5 (1358.25 - 1634.75)"
                       nWBV
   "0.739 (0.71725 - 0.769)"
```

```r
sapply(vars, report_quanti, gr=1) #groupe 1
```

```
                    Age                     EDUC                      SES
         "76 (72 - 82)"          "14 (12 - 16)"             "3 (1 - 4)"
                   MMSE                      CDR                     eTIV
         "27 (23 - 29)"       "0.5 (0.5 - 0.5)"  "1463 (1357 - 1568.5)"
                   nWBV
"0.713 (0.695 - 0.738)"
```

2.C Autres façons de faire

2.C.1 tableone

```r
library(tableone)
vars = c("Age","M.F","EDUC","SES","MMSE","CDR","eTIV","nWBV")
table1 = CreateTableOne(vars = vars, strata = "Group", data = alzh, factorVars =
↪  c("M.F.","SES","CDR"))
```

```
Warning in ModuleReturnVarsExist(factorVars, data): The data frame does not
have: M.F.  Dropped
```

```r
print(table1, showAllLevels = TRUE, formatOptions = list(big.mark = ","))
```

```
                  Stratified by Group
                   level 0                 1                 p      test
  n                          190               183
  Age (mean (SD))          77.06 (8.10)      76.97 (7.16)    0.909
  M.F (%)          F         129 (67.9)         84 (45.9)   <0.001
                   M          61 (32.1)         99 (54.1)
  EDUC (mean (SD))         15.14 (2.74)      14.03 (2.91)   <0.001
  SES (%)          1          41 (21.6)         47 (28.7)    0.004
                   2          71 (37.4)         32 (19.5)
                   3          42 (22.1)         40 (24.4)
                   4          34 (17.9)         40 (24.4)
                   5           2 ( 1.1)          5 ( 3.0)
  MMSE (mean (SD))         29.23 (0.88)      25.36 (4.40)   <0.001
  CDR (%)          0         188 (98.9)         18 ( 9.8)   <0.001
                   0.5         2 ( 1.1)        121 (66.1)
                   1           0 ( 0.0)         41 (22.4)
                   2           0 ( 0.0)          3 ( 1.6)
  eTIV (mean (SD))       1,495.50 (184.89) 1,480.48 (166.73)  0.411
  nWBV (mean (SD))           0.74 (0.04)       0.72 (0.03)   <0.001
```

2.C.2 gtsummary : je vois pas pourquoi on utilise pas ce banger !

```r
library(gtsummary)
```

```
Attaching package: 'gtsummary'

The following object is masked from 'package:ape':

    where
```

```r
colnames(alzh)
```

```
 [1] "Group" "M.F"   "Age"   "EDUC"  "SES"   "MMSE"  "CDR"   "eTIV"  "nWBV"
[10] "ASF"
```

```r
cols_to_include = c("Group", "Age","M.F","EDUC","SES","MMSE","CDR","eTIV","nWBV")
alzh$Group = factor(alzh$Group, labels = c("Non-Demented","Demented"))

table2 = alzh %>%
    select(all_of(cols_to_include)) %>%
    tbl_summary(
        by = Group,
        # Change 'labels' to 'label' below
        label = list(
        Age ~ "Age (years)",
        M.F ~ "Sex",
        EDUC ~ "Education (years)",
        SES ~ "Socioeconomic Status",
        MMSE ~ "Mini-Mental State Examination",
        CDR ~ "Clinical Dementia Rating",
        eTIV ~ "Estimated Total Intracranial Volume",
        nWBV ~ "Normalized Whole Brain Volume"
        ),
        statistic = list(all_continuous() ~ "{median} ({p25} - {p75})",
                         all_categorical() ~ "{n} / {N} ({p}%)"),
        digits = all_continuous() ~ 2
    ) %>%
    modify_header(label = "**Variable**") %>%
    bold_labels() %>%
    add_n() %>%
    add_p() %>%
    modify_caption("**Table 1. Descriptive Statistics by Dementia Status**") %>%
    as_kable()
```

```
# View the table
table2
```

Table 7.1: **Table 1. Descriptive Statistics by Dementia Status**

Variable	N	**Non-Demented** N = 190	**Demented** N = 183	p-value
Age (years)	373	77.00 (71.00 - 82.00)	76.00 (72.00 - 82.00)	0.9
Sex	373			<0.001
F		129 / 190 (68%)	84 / 183 (46%)	
M		61 / 190 (32%)	99 / 183 (54%)	
Education (years)	373	16.00 (13.00 - 18.00)	14.00 (12.00 - 16.00)	<0.001
Socioeconomic Status	354			0.003
1		41 / 190 (22%)	47 / 164 (29%)	
2		71 / 190 (37%)	32 / 164 (20%)	
3		42 / 190 (22%)	40 / 164 (24%)	
4		34 / 190 (18%)	40 / 164 (24%)	
5		2 / 190 (1.1%)	5 / 164 (3.0%)	
Unknown		0	19	
Mini-Mental State Examination	371	29.00 (29.00 - 30.00)	27.00 (23.00 - 29.00)	<0.001
Unknown		0	2	
Clinical Dementia Rating	373			<0.001
0		188 / 190 (99%)	18 / 183 (9.8%)	
0.5		2 / 190 (1.1%)	121 / 183 (66%)	
1		0 / 190 (0%)	41 / 183 (22%)	
2		0 / 190 (0%)	3 / 183 (1.6%)	
Estimated Total Intracranial Volume	373	1,474.50 (1,358.00 - 1,636.00)	1,463.00 (1,357.00 - 1,569.00)	0.5
Normalized Whole Brain Volume	373	0.74 (0.72 - 0.77)	0.71 (0.70 - 0.74)	<0.001

8Survie

1 Plan du cours

1. Introduction : données de survie et censure
2. Estimation de la fonction de survie : estimateur de Kaplan-Meier
3. Comparaison de fonctions de survie : test du log-rank
4. Modélisation de la survie : modèle de Cox

2 Introduction

Donnée censuré ≠ donnée manquante :

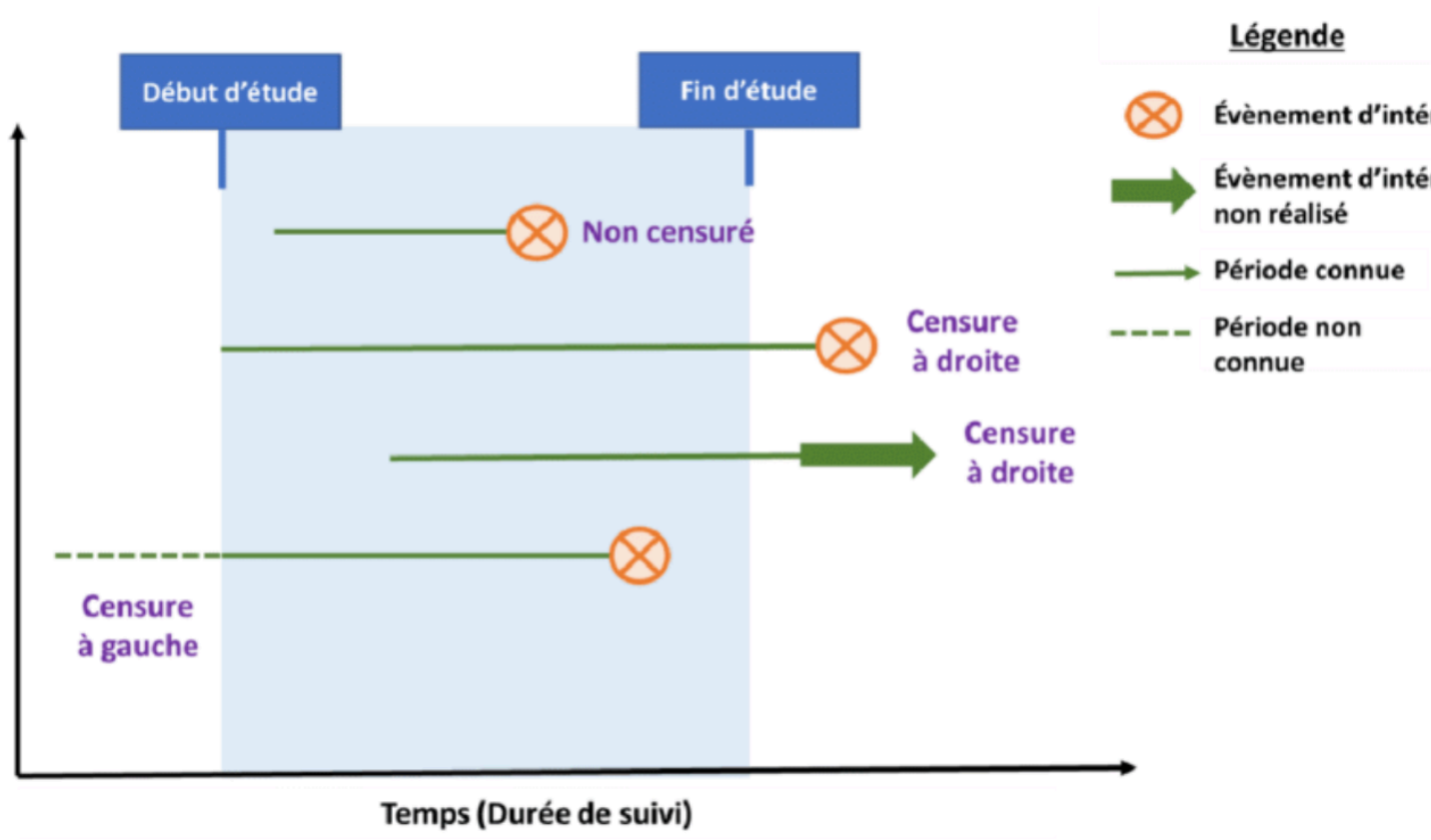
- **Censurée** : on sait que l'événement d'intérêt n'est pas survenu avant un certain temps (par ex : pas de récidive jusqu'à la perte de vue)
- **Manquante** : on ne sait pas si l'événement d'intérêt est survenu ou non

NB : toutes les variables correspondant à une durée ne sont pas nécessairement censurées !

2.A Données de survie

Donnée de survie = *survival time* ou *time to event data*

- Définition : délai de survenue d'un événement d'intérêt (*endpoint event*) à partir d'un temps de départ (souvent le temps 0)
- Correspond à tout délai entre deux dates d'intérêt =
  - Censure à gauche : point de départ !
  - Censure à droite : événement non survenu avant la fin de l'étude



2.B Problème de la censure

- certains ne présentent pas l'événement d'intérêt pendant la période d'étude
- tous les patients n'ont pas le même temps d'observation

On aimerait que le délai T_i soit le délai jusqu'à l'événement d'intérêt pour chaque individu i.
Mais on observe en fait :

- $min(T_i, C)$ et $1_{T_i > C}$ c'est à dire que :
  - $min(T_i, C)$: on cherche la plus petite valeur entre le délai jusqu'à l'événement d'intérêt T_i et une durée d'observation maximale fixe C
  - $1_{T_i > C}$: 1 à chaque fois que T_i est supérieur à C (censure à droite) = c'est à dire que l'événement n'est pas survenu avant la fin de l'étude
  - 0 sinon : si l'évènemet est survenu avant la censure
- ou $min(T_i, C_i)$, $1_{T_i > C_i}$

C : durée d'observation maximale fixe

C_i : durée d'observation variable selon les individus = **censure aléatoire**

Information partielle = censure à droite.

Par exemple :

- C = 3 ans (censure à 3 ans pour tous les individus)
- T_i = délai jusqu'à la récidive pour le patient i
- Si le patient i récidive à 2 ans : on observe $min(2, 3) = 2$ et $1_{2>3} = 0$ (événement observé avant la censure)
- Si le patient i ne récidive pas avant 3 ans : on observe $min(T_i, 3) = 3$ et $1_{T_i>3} = 1$ (événement non observé avant la censure)

Tip

En gros : il faut différencier

- les individus pour lesquels on observe l'événement d'intérêt avant la date de censure (on connaît leur temps de survie exact, parce qu'ils "n'ont pas survécu" jusqu'à la censure)
- les individus pour lesquels on ne sait pas si l'événement d'intérêt est survenu ou non avant la date de censure (on sait juste qu'ils ont "survécu" jusqu'à la censure)

Donc colonnes dans la BDD pour le critère de survie :

- L'ÉVÈNEMENT :
  - Survenue ou non (0/1)
  - Date et délai par rapport au temps de départ
- DURÉE DE SUIVI
  - Indépendante de la survenue ou non de l'événement d'intérêt
  - Depuis de le temps de départ

---

3 Données de survie

3.A Modélisation satistique

Analyse statistique dépend de la question de recherche :

- Est-ce que l'évènement s'est produit (pendant la période d'étude) ? = **modèle binaire** = régression logistique
- Quand l'évènement s'est-il produit ? = **modèle de survie** = régression de Cox

3.B Applications

- **Essai thérapeutique** : comparer l'efficacité de deux interventions revient à comparer les durées de survie après intervention dans les deux groupes
- **Étude épidémiologique** : estimation de l'association entre un facteur de risque et la durée de survie ou le temps de survenue d'une maladie

3.C Terminologie

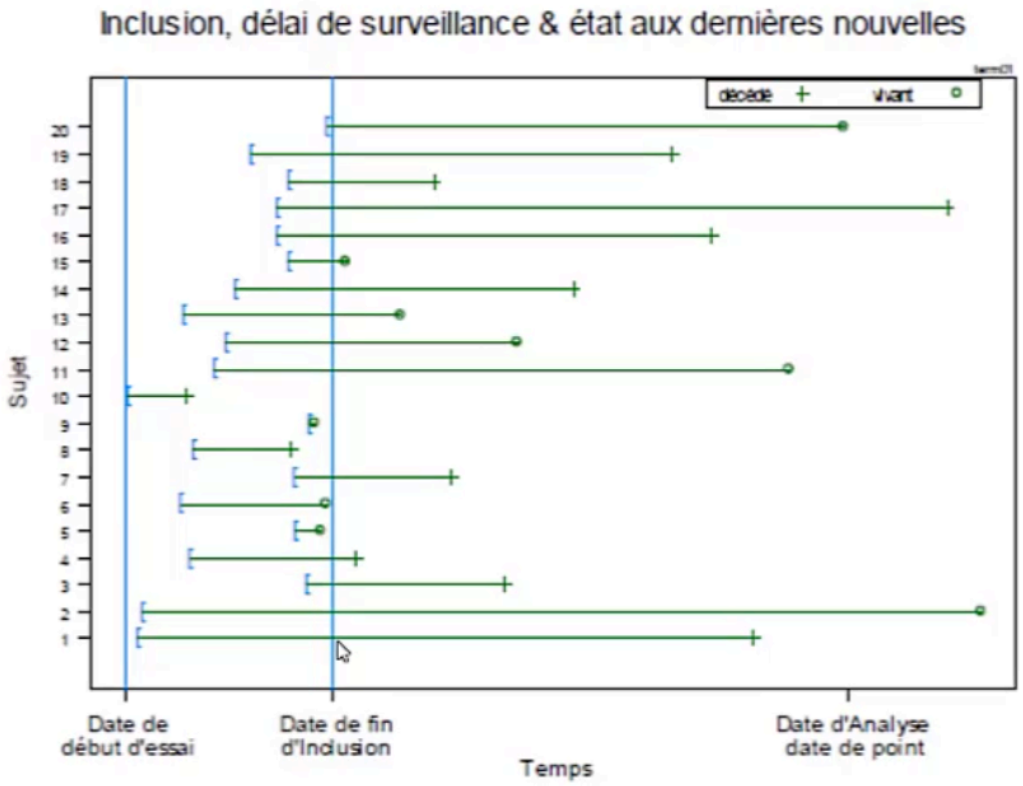
- **Date d'origine** : point de départ, varie selon patient, pour le calcul des durées de survie (ex : date de diagnostic, date de traitement, date d'inclusion dans l'étude)
- **Date des dernières nouvelles** : date de la dernière information connue sur le patient (ex : date de décès, date de la dernière consultation, date de la fin de l'étude)
- **Date de point** : commune à tous les patients, pour le calcul des durées de survie (ex : date de fin de l'étude)
- **Censure** : information incomplète, l'événement d'intérêt n'est pas survenu avant la date de point
- **Temps de participation** : variable d'étude
  - Évènement avant date de point (décès, récidive, etc.) : temps de participation = délai entre date d'origine et date de l'événement
  - Pas d'évènement avant date de point (censure) :
    - La date des dernières nouvelles est antérieure à la date de point : perdus de vue
    - La date des dernières nouvelles est égale à la date de point : censure administrative

3.D Ce qui est nécessaire

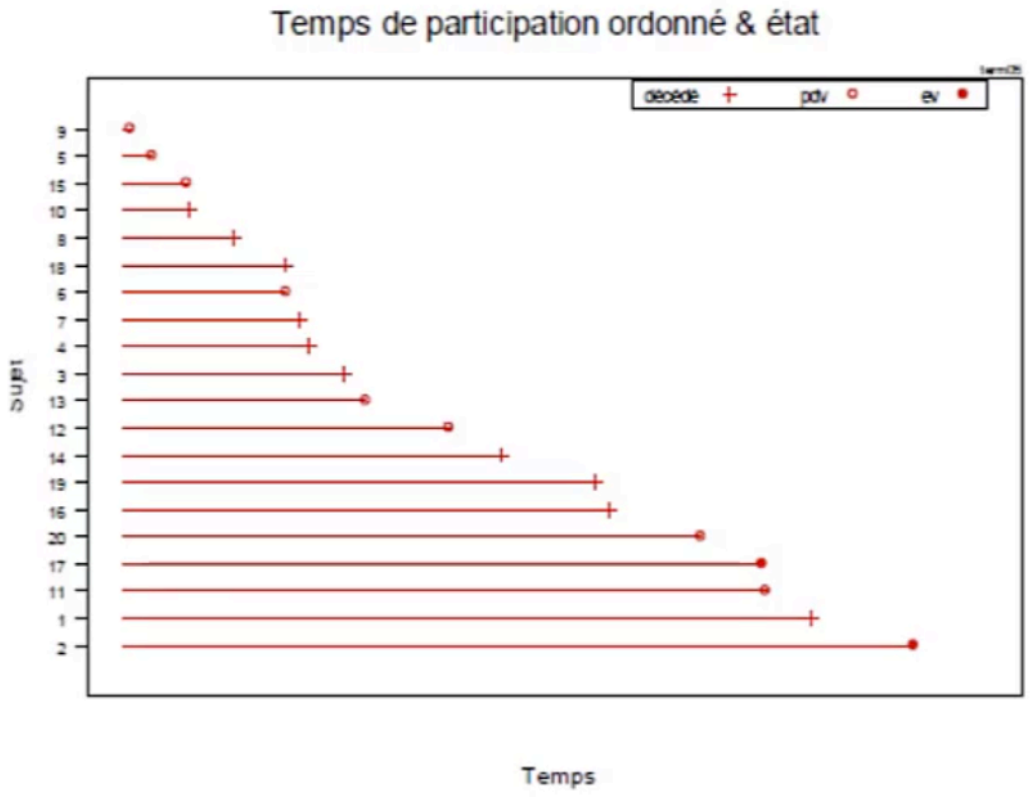
Pour chaque sujet, on doit disposer de :

- **Temps de participation** (délai entre date d'origine et date de l'événement ou date des dernières nouvelles)
- **État de l'événement** (1 = événement survenu, 0 = censuré) à la fin du temps de participation

Exemple : données simulées

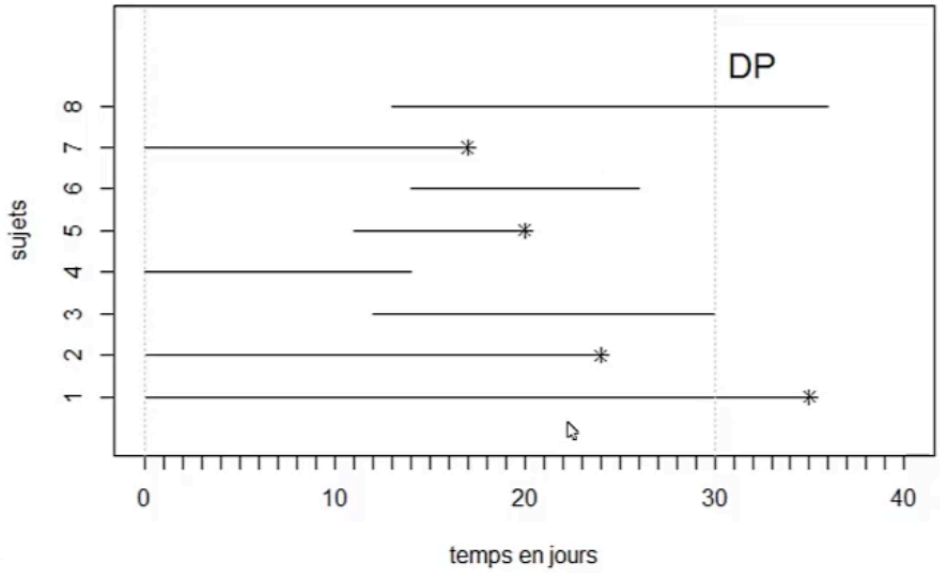


Exemple : données simulées



3.D.1 Exercice

A partir du graphique déterminer le temps de participation T_k pour chaque patient et l'état I_k en T_k.



16/11/20 13

Suivi :

- Patient 1 : 30 jours (évènement à 35 jours)
- Patient 2 : 24 jours (évènement à 24 jours)

Numéro du malade	Date d'origine (DO)	Date : Etat aux dernières nouvelles	Etat à la date de point (DP)	T_k jours	I_k
1	J0	J35 DCD	Vivant	30	0
2	J0	J24 DCD	DCD	24	1
3	J12	J30 Vivant	Vivant	18	0
4	J0	J14 Vivant	Perdu de vue	14	0
5	J11	J20 DCD	DCD	9	1
6	J14	J26 Vivant	Perdu de vue	12	0
7	J0	J17 DCD	DCD	17	1
8	J13	J36 Vivant	Vivant	12	0

3.E Loi de probabilité de T

Loi de probabilité de T = délai jusqu'à l'évenement (non observée)
Décrite par l'une de ces fonctions :

- Densité de probabilité, $f(t)$
- Fonction de répartition, $F(t)$
- Fonction de survie, $S(t)$
- Fonction de risque instantané, $h(t)$
- Fonction de risque cumulée, $H(t)$

3.E.1 Densité de probabilité

$f(t)$ = $\lim_{\Delta t \to 0} \frac{P(T<t+\Delta t)-P(T<t)}{\Delta t}$

- $lim_{\Delta t \to 0}$: signifie que l'on regarde un intervalle de temps de plus en plus petit autour de t (on réduit l'intervalle Δt jusqu'à ce qu'il tende vers 0).
- T : variable aléatoire représentant le délai jusqu'à l'événement d'intérêt = moment où l'événement se produit.
- t : moment spécifique dans le temps où l'on évalue la densité de probabilité.
- $P(T < t)$: probabilité que l'événement d'intérêt T se produise avant le temps t.

On calcule la densité de probabilité $f(t)$ en prenant la limite lorsque l'intervalle de temps Δt

autour de t devient très petit. Cela nous permet d'estimer la probabilité que l'événement T se produise précisément à ce moment t.

En gros :

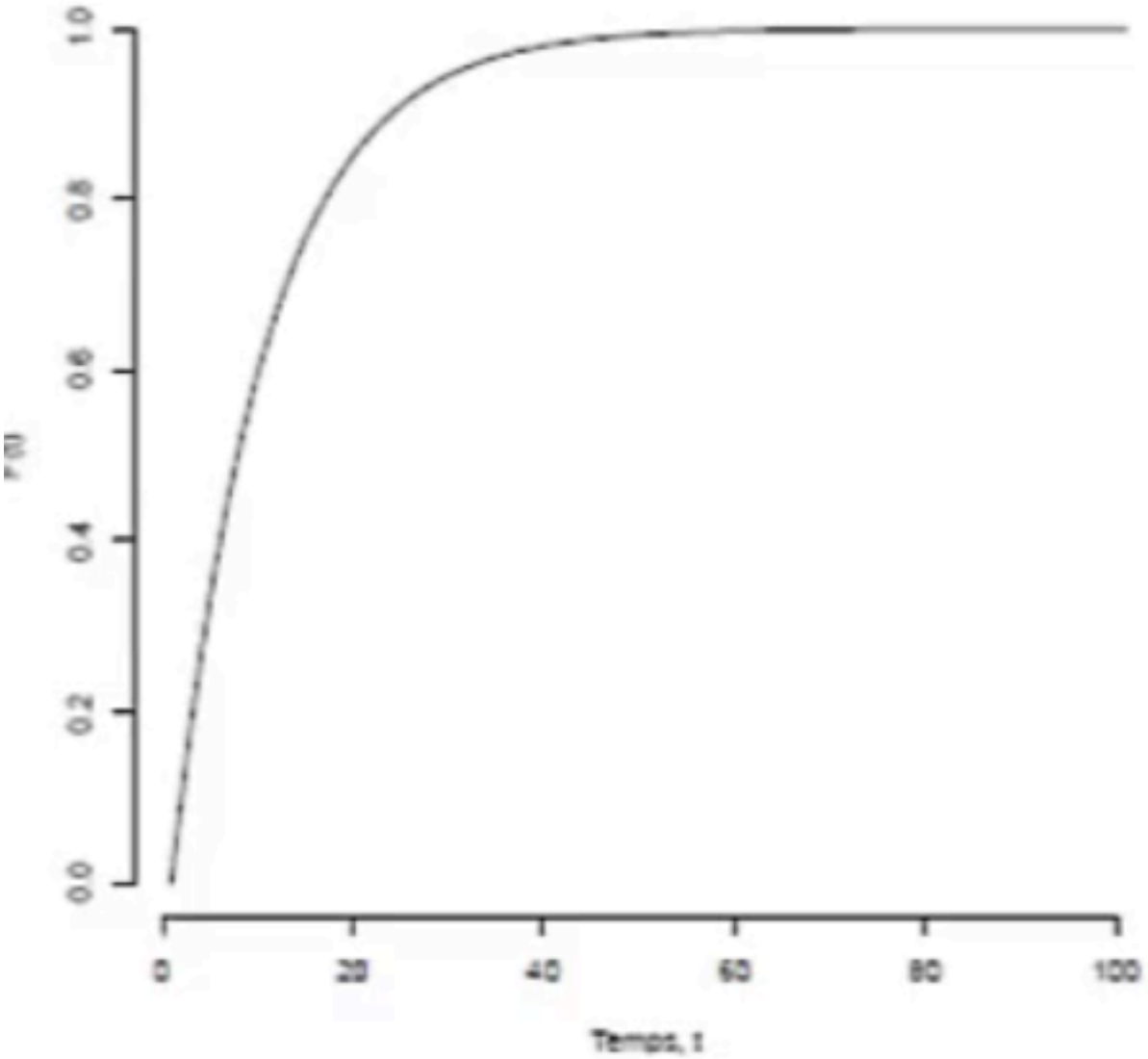
1. Différence entre :
   - La probabilité que l'évènement T se produise avant un certain temps $t + \Delta t$ (un intervalle de temps très petit autour de t)
   - et la probabilité que l'évènement T se produise avant un certain temps t
2. Diviser le résultat par la taille de l'intervalle de temps Δt
3. On obtient ainsi un "taux moyen" de probabilité par unité de temps dans cet intervalle très petit autour de t
4. En prenant la limite lorsque Δt tend vers 0, on obtient la **densité de probabilité instantanée**

3.E.2 Fonction de répartition

Probabilité que la variable aléatoire T prenne une valeur inférieure ou égale à une quantité t :

$F(t) = P(T \leq t) = \int_0^t f(u)du$

- $F(t)$: fonction de répartition, qui donne la probabilité que l'événement d'intérêt T se produise avant ou à un moment spécifique t.
- $P(T \leq t)$: probabilité que l'événement T se produise avant ou à temps t.
- $\int_0^t f(u)du$: intégrale de la densité de probabilité $f(u)$ de 0 à t, qui calcule la probabilité cumulative jusqu'à ce moment.

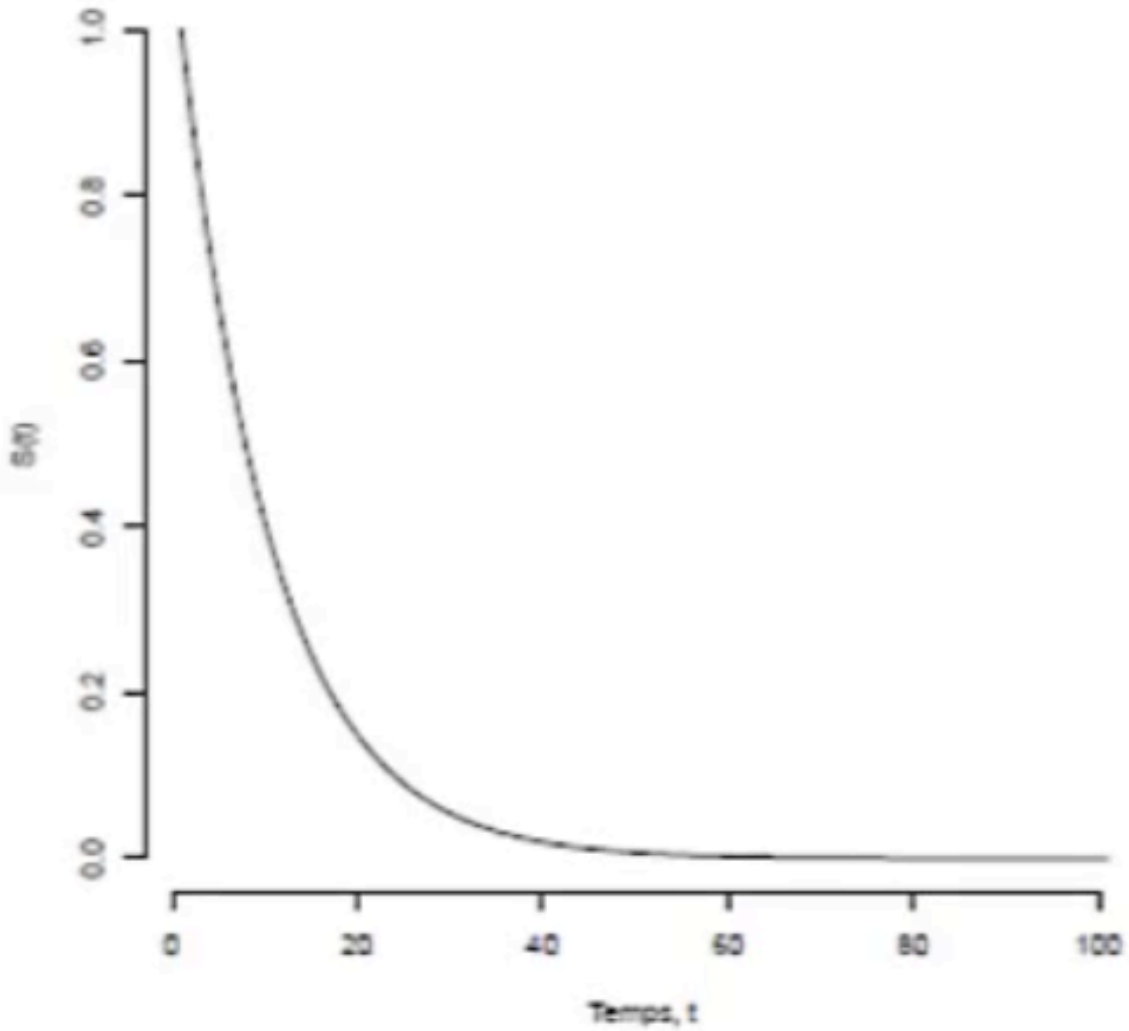


Propriétés :

- Fonction croissante (plus on attend, plus la probabilité que l'événement se soit produit augmente)
- La vitesse de croissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $F(0) = 0$: si on n'attend pas, la probabilité de voir l'évènement = 0
- $\lim_{t\to\infty} F(t) = 1$: si on attend indéfiniment, la probabilité de voir l'évènement = 1
- F dérivable et $F' = f$ avec f la densité de probabilité (la dérivée de la fonction de répartition = la fonction de densité).

3.E.3 Fonction de survie

Représente la fraction d'individus encore en vie en t.
$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du$



Propriétés :

- Fonction décroissante (plus on attend, plus la probabilité de survie diminue)
- La vitesse de décroissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $S(0) = 1$: si on n'attend pas, la probabilité de survie = 1
- $\lim_{t\to\infty} S(t) = 0$: si on attend indéfiniment, la probabilité de survie = 0
- S dérivable et $S' = -f$ avec f la densité de probabilité (la dérivée de la fonction de survie = l'opposé de la fonction de densité).

3.E.4 Fonction de risque instantané

Définition : fonction de “hazard” ou “hasard”. Représente le risque instantané de survenue de l’événement à l’instant t, conditionnellement au fait que l’individu ait survécu jusqu’à ce temps t.

Densité “**conditionnelle**” de l’événement à l’instant t sachant que l’individu est encore en vie à ce moment.

$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$

NB : la barre verticale “|” signifie “conditionnellement à”.

- $h(t)$: fonction de risque instantané, qui mesure le risque de survenue de l’événement d’intérêt à un moment spécifique t, conditionnellement au fait que l’individu ait survécu jusqu’à ce temps t.
- $P(t \leq T < t + \Delta t \mid T \geq t)$:
  - probabilité que l’événement T se produise dans l’intervalle de temps entre t et $t + \Delta t$,
  - **conditionnellement** au fait que l’individu ait survécu jusqu’à temps t.
- Δt : intervalle de temps très petit autour de t.
- $\lim_{\Delta t \to 0}$: signifie que l’on regarde un intervalle de temps de plus en plus petit autour de t (on réduit l’intervalle Δt jusqu’à ce qu’il tende vers 0).
  - Permet d’obtenir un taux instantané de risque à ce moment précis t.
  - Sinon, on obtiendrait une moyenne sur un intervalle de temps plus large.

En résumé : c’est une **densité conditionnelle** qui mesure le risque instantané de survenue de l’événement à un moment spécifique t, en tenant compte du fait que l’individu a déjà survécu jusqu’à ce temps t ET NE L’A PAS ENCORE PRÉSENTÉ !

On peut représenter plus simplement la fonction en :

$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$

C’est à dire que la fonction de risque instantané est le rapport entre la densité de probabilité et la fonction de survie.

C’est logique ! parce que :

- Fonction de densité $f(t)$: probabilité que l’événement se produise précisément à l’instant t
- Fonction de survie $S(t)$: probabilité que l’individu soit encore en vie à temps t (donc n’ait pas encore présenté l’événement)
- Donc le rapport $f(t)/S(t)$: probabilité que l’événement se produise à l’instant t sachant que l’individu est encore en vie à ce moment.
- Le signe négatif dans $-\frac{S'(t)}{S(t)}$ vient du fait que la dérivée de la fonction de survie $S'(t)$ est négative (puisque $S(t)$ est décroissante). Donc en prenant l’opposé, on obtient une valeur positive pour la fonction de risque instantané.

3.E.5 Fonction de risque cumulée

Représente le risque cumulé de survenue de l'événement jusqu'au temps t.

$H(t) = \int_0^t h(u)du$

Propriétés :

- Fonction croissante (plus on attend, plus le risque cumulé augmente)
- La vitesse de croissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $H(0) = 0$: si on n'attend pas, le risque cumulé = 0
- $\lim_{t\to\infty} H(t) = \infty$: si on attend indéfiniment, le risque cumulé = infini

4 #### Relations entre les fonctions

Fonction	Notation	Définition
Densité de probabilité	$f(t)$	$\lim_{\Delta t\to 0} \frac{P(t<T<t+\Delta t)}{\Delta t}$
Fonction de répartition	$F(t)$	$P(T \leq t) = \int_0^t f(u)du$
Fonction de survie	$S(t)$	$P(T > t) = 1 - F(t) = \int_t^\infty f(u)du$
Fonction de risque instantané	$h(t)$	$\lim_{\Delta t\to 0} \frac{P(t\leq T<t+\Delta t \mid T\geq t)}{\Delta t}$
Fonction de risque cumulée	$H(t)$	$\int_0^t h(u)du$

$P(t < T < t + \Delta t) = P(t < T < t + \Delta t \mid T > t) \times P(T > t)$

$P(t < T < t + \Delta t)$ = la probabilité que l'événement se produise dans l'intervalle de temps entre t et $t + \Delta t$.

Mais en fait 2 probabilités en une :

1. **probabilité** que l'individu soit encore en vie au temps t : $P(T > t)$ = la probabilité que l'individu ait survécu jusqu'à temps t (donc n'ait pas encore présenté l'événement avant t).
   NB : c'est la fonction de survie $S(t)$
2. **probabilité** que l'événement se produise dans l'intervalle de temps entre t et $t+\Delta t$ PARMI LES INDIVIDUS ayant survécu jusqu'à temps t : $P(t \leq T < t + \Delta t \mid T \geq t)$.
   NB : c'est la fonction de risque instantané $h(t)$ multipliée par la taille de l'intervalle Δt

Et sachant que :

f = densité de probabilité = $\lim_{\Delta t\to 0} \frac{P(t<T<t+\Delta t)}{\Delta t}$

h = fonction de risque instantané = $\lim_{\Delta t\to 0} \frac{P(t\leq T<t+\Delta t|T\geq t)}{\Delta t}$

S = fonction de survie = $P(T > t)$

donc on en déduit :

$f(t) = h(t) \times S(t)$

Et sachant que :

f = densité de probabilité = $-S'(t)$

S = fonction de survie = $1 - F(t)$ = $e^{-H(t)}$ car $H(t) = -\ln(S(t))$

En résumé :

Fonction	Notation	Relation avec les autres fonctions
Densité de probabilité	$f(t)$	$f(t) = h(t) \times S(t)$
Fonction de répartition	$F(t)$	$F(t) = 1 - S(t)$
Fonction de survie	$S(t)$	$S(t) = e^{-H(t)}$
Fonction de risque instantané	$h(t)$	$h(t) = \frac{f(t)}{S(t)}$
Fonction de risque cumulée	$H(t)$	$H(t) = -\ln(S(t))$

5 Estimation d'une courbe de survie : estimateur de Kaplan-Meier

5.A Objectif de l'analyse de survie

- Estimer le délai médian avant la survenue de l'événement d'intérêt
- Comparer ce délai entre plusieurs groupes de patients
- Étudier l'effet de variables explicatives sur ce délai

5.B Approches

- **Approche paramétrique** : modélisation de la fonction de risque = on fait une hypothèse sur la forme de la fonction de risque (ex : loi exponentielle)
- **Approche non paramétrique** : estimation de la fonction de survie sans faire d'hypothèse sur la forme de la fonction de risque = estimateur de Kaplan-Meier
- **Approche semi-paramétrique** : modèle multivarié : modélisation de l'effet des variables explicatives sur la fonction de risque sans faire d'hypothèse sur la forme de la fonction de risque = modèle de Cox

5.B.1 Estimation en absence de censure

En l'absence de censure (c'est-à-dire quand le délai de survenue de l'événement est connu pour tous les individus), l'approche non paramétrique consiste à **remplacer la probabilité théorique par une simple proportion** basée sur les données observées.

L'estimateur empirique de la fonction de répartition $\widehat{F}(t)$ représente la **proportion d'individus** dans la population étudiée qui ont présenté l'événement d'intérêt avant ou à un moment donné t.

- **Calcul :** Cette proportion est calculée en sommant les indicateurs (qui valent 1 si l'événement T_i a eu lieu avant t, et 0 sinon) pour tous les individus, puis en divisant par le nombre total d'individus n,.
- **Propriété :** Cet estimateur est une **fonction en escalier, monotone, croissante de 0 à 1**.
- **Estimateur de la fonction de répartition** $F(t)$ à partir des données observées

  $\widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(T_i \leq t)}$.

  - $1_{(T_i \leq t)}$ = indicatrice qui vaut 1 si $T_i \leq t$ (si l'évènement T a eu lieu avant t) et 0 sinon.
  - $\sum_{i=1}^{n} 1_{(T_i \leq t)}$ = somme des 0 et des 1 pour tous les individus = nombre d'individus ayant présenté l'évènement avant t.
  - diviser le tout par le nombre total d'individus n permet d'obtenir la proportion d'individus ayant présenté l'évènement avant t.
  - Donc $\widehat{F}(t)$ = estimateur empirique = proportion d'individus ayant présenté l'évènement avant t dans la population étudiée
  - $\widehat{F}(t)$ est un estimateur de la fonction de répartition $F(t)$, qui serait la proportion "vraie" d'individus ayant présenté l'évènement avant t dans la population générale / idéale
- **Estimateur de la fonction de survie** $S(t)$ = $1 - \widehat{F}(t)$

L'estimateur empirique de la fonction de survie ($\widehat{S}(t)$) est le complément direct de $\widehat{F}(t)$, soit $\widehat{S}(t) = 1 - \widehat{F}(t)$.

- **Signification :** Il représente la **proportion d'individus n'ayant pas présenté l'événement avant** t, ou ayant « survécu au-delà de t » dans la population étudiée. Il est calculé comme le nombre d'individus survivant au-delà de t, divisé par le nombre total d'individus.
- **Propriété :** C'est une **fonction en escalier, monotone, décroissante de 1 à 0**.

  $\widehat{S}(t) = 1 - \widehat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(T_i > t)}$

  - $1_{(T_i > t)}$ = indicatrice qui vaut 1 si $T_i > t$ (si l'évènement T a eu lieu après t) et 0 sinon.
  - $\sum_{i=1}^{n} 1_{(T_i > t)}$ = somme des 0 et des 1 pour tous les individus = nombre d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t).
  - diviser le tout par le nombre total d'individus n permet d'obtenir la proportion d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t).
  - Donc $\widehat{S}(t)$ = estimateur empirique = proportion d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t) dans la population étudiée

  = $\widehat{S}(t) = \frac{\text{Nombre d'individus survivant au dela de } t}{\text{Nombre total d'individus}}$

5.B.1.1 Exemple

1. Données = temps de décès (en mois) de 10 patients :

   13, 13, 14, 13, 15, 11, 17, 13, 14, 15

2. Estimation de la fonction de survie

Estimateur de $S(t) = 1 - F(t)$

$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(T_i > t)} = \frac{\text{Nombre d'individus survivant au dela de } t}{\text{Nombre total d'individus}}$

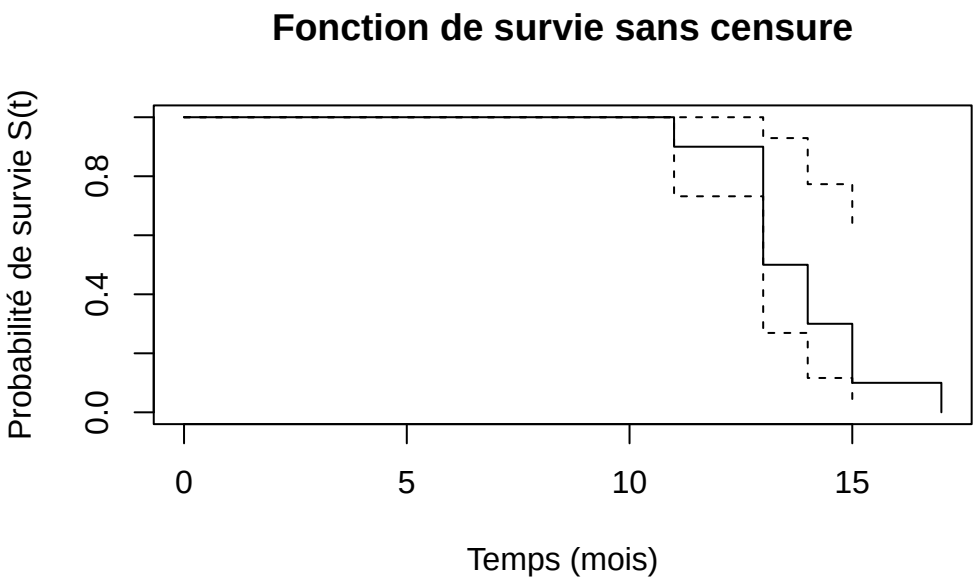
en tableau :

- Premier décès à 11 mois, compté dans l'intervalle 11 - 12 par convention.

t	Nb décès	Nb décès cumulé	$\hat{S}(t)$
0	0	0	10/10 = 1.0
11	1	1	9/10 = 0.9
13	4	5	5/10 = 0.5
14	2	7	3/10 = 0.3
15	2	9	1/10 = 0.1
17	1	10	0/10 = 0.0

3. Tracé de la fonction de survie sans Kaplan Meier

```r
library(survival)

# 1) Données
temps <- c(13, 13, 14, 13, 15, 11, 17, 13, 14, 15)

# Comme il n'y a PAS de censure : event = 1 pour tout le monde
event <- rep(1, length(temps))

# Objet Surv
surv_object <- Surv(time = temps, event = event)

# 2) Estimation de la fonction de survie
surv_fit <- survfit(surv_object ~ 1)

# 3) Tracé de la fonction de survie
plot(
    surv_fit,
    xlab = "Temps (mois)",
    ylab = "Probabilité de survie S(t)",
    main = "Fonction de survie sans censure")
```



5.B.1.2 2e exemple avec temps simulés On utilise `runif` qui génère des nombres aléatoires suivant une loi uniforme (uniforme = tous les intervalles de même longueur ont la même probabilité d'être choisis).

`floor` arrondit à l'entier inférieur.

```
T = floor(runif(80,0,50)) # runif génère 1000 temps de survie entre 0 et 50 qui
 ↪  sont arrondis à l'entier inférieur avec floor
head(T)
```

```
[1] 46 28 21 41  4 23
```

```
table(T)
```

```
T
 1  2  3  4  5  6  7  8  9 10 12 13 14 16 17 19 20 21 22 23 24 25 26 27 28 29
 3  2  1  4  3  1  2  1  1  2  3  2  2  1  1  2  2  2  1  2  2  1  2  2  3  1
31 32 33 34 35 36 37 38 39 41 43 44 45 46 47 48 49
 2  2  3  2  1  1  2  1  2  3  3  2  1  3  1  1  1
```

Le jeu de données est généré mais on a pas les valeurs uniques.

```
tt <- unique(T)
length(tt) # 39 valeurs uniques
```

```
[1] 43
```

Il faut aussi ordonner les valeurs uniques.

```r
tt <- sort(tt)
tt
```

```
 [1]  1  2  3  4  5  6  7  8  9 10 12 13 14 16 17 19 20 21 22 23 24 25 26 27 28
[26] 29 31 32 33 34 35 36 37 38 39 41 43 44 45 46 47 48 49
```

Pour regarder à la date 10 :

- avec `head()` : R prend chacune des valeurs de T et regarde si elle est supérieure à 10 (TRUE) ou non (FALSE).
- avec `mean()` : R calcule la proportion de TRUE (donc la proportion de survivants au delà de 10), en convertissant TRUE en 1 et FALSE en 0.

```r
head(T > 10)
```

```
[1]  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
```

```r
mean(T > 10) # proportion de TRUE = proportion de survivants au delà de 10
```

```
[1] 0.75
```

On peut faire ça pour toutes les valeurs uniques de T.

```r
S <- function(t) mean(T > t)
```

Syntaxe : `function(t) mean(T > t)`

- `function(t)` : on définit une fonction qui prend un argument `t`
- `mean(T > t)` : la fonction calcule la proportion de survivants au delà de `t`

Pas de séparation par des virgules car une seule instruction dans le corps de la fonction = R comprend que tout ce qui suit `function(t)` fait partie du corps de la fonction.

```r
S(10) # proportion de survivants au delà de 10
```

```
[1] 0.75
```

```r
S(20) # proportion de survivants au delà de 20
```

```
[1] 0.5875
```

On applique cette fonction à toutes les valeurs uniques de T avec `sapply`.

`sapply` applique une fonction (ici une fonction anonyme) à chaque élément d'un vecteur (ici `tt`).

≠ `lapply` qui renvoie une liste, `sapply` renvoie un vecteur ou une matrice.

```r
S_values <- sapply(tt, S)
S_values
```

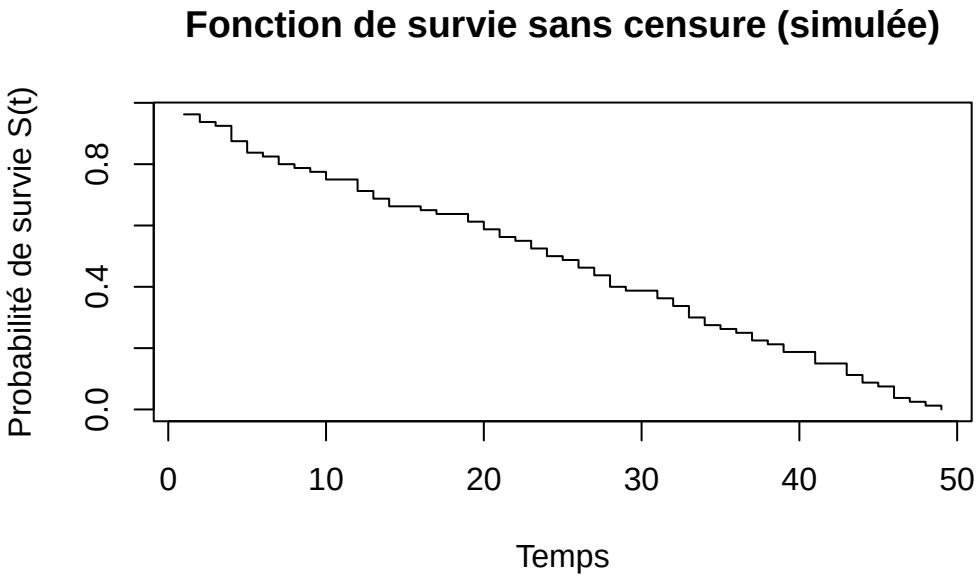
```
 [1] 0.9625 0.9375 0.9250 0.8750 0.8375 0.8250 0.8000 0.7875 0.7750 0.7500
[11] 0.7125 0.6875 0.6625 0.6500 0.6375 0.6125 0.5875 0.5625 0.5500 0.5250
[21] 0.5000 0.4875 0.4625 0.4375 0.4000 0.3875 0.3625 0.3375 0.3000 0.2750
[31] 0.2625 0.2500 0.2250 0.2125 0.1875 0.1500 0.1125 0.0875 0.0750 0.0375
[41] 0.0250 0.0125 0.0000
```

On peut tracer la fonction de survie estimée.

Syntaxe de `plot` : `plot(x, y, type, xlab, ylab, main)`

- `x` : valeurs sur l'axe des x (ici `tt`)
- `y` : valeurs sur l'axe des y (ici `S_values`)
- `type="s"` : pour une fonction en escalier

```r
plot(
    tt, # x values = temps
    S_values, # y values = probabilité de survie S(t) calculée
    type="s", # type = "s" pour une fonction en escalier
    xlab="Temps",
    ylab="Probabilité de survie S(t)",
    main="Fonction de survie sans censure (simulée)")
```



5.B.2 Estimation en présence de censure : probabilités conditionnelles

Probabilités conditionnelles = on ne considère que les individus “à risque” à chaque instant.

- À chaque temps t_j où un événement est observé, on calcule la probabilité de survie conditionnelle **sachant que l'individu a survécu jusqu'à ce temps** t_j.
- On multiplie ces probabilités conditionnelles pour obtenir la probabilité de survie jusqu'à un temps t donné.

Par exemple :

Probabilité d'être en vie à 2 et 3 ans : probabilité décomposée en deux parties :

$S(2) = P(T > 2) = P(T > 2 \mid T > 1) \times P(T > 1)$

$S(3) = P(T > 3) = P(T > 3 \mid T > 2) \times P(T > 2)$

$S(3) = P(T > 3 \mid T > 2) \times S(2)$

= Probabilité de survivre entre 1 et 2 ans sachant qu'on a survécu jusqu'à 1 an multipliée par la probabilité de survivre jusqu'à 1 an.

On généralise à un ensemble de K temps ordonnés définis arbitraitement et aléatoirement

On “découpe” la période d'étude pour obtenir des petits intervalles $[t_{k-1}, t_k)$.

$S(t_k) = P(T > t_k \mid T > t_{k-1}) \times P(T > t_{k-1}) = P(T > t_k \mid T > t_{k-1}) \times S(t_{k-1})$

Parce que $S(t_{k-1}) = P(T > t_{k-1})$

Tableau de données Kaplan Meier :

N_k = nombre d'individus à risque au temps t_k (ayant survécu jusqu'à t_k)

Q_k = probabilité conditionnelle de survie au temps t_k = $P(T > t_k \mid T > t_{k-1})$

S_k = probabilité cumulée de survie jusqu'au temps t_k = $S(t_k)$

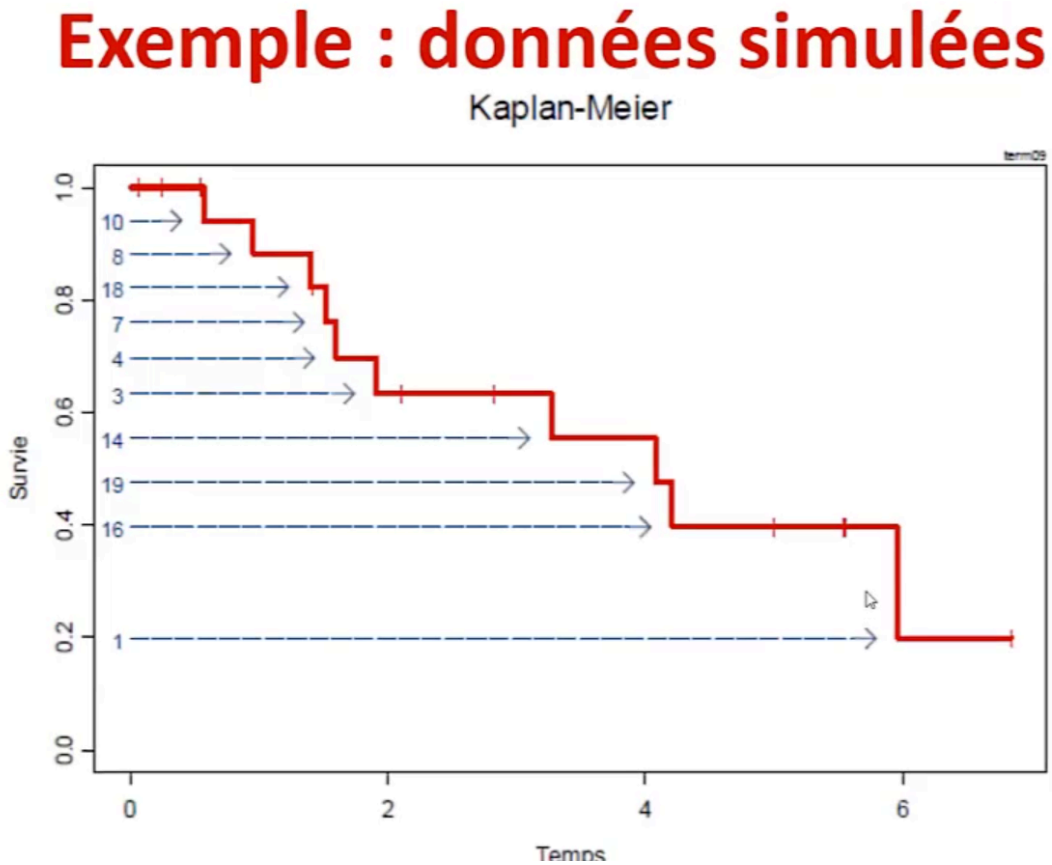
t_k (temps)	N_k (nombre à risque en pendant la période)	D_k (décès en début de période)	C_k (censurés)	Q_k (Probabilité condition-nelle de survie)	S_k (Probabilité cumulée de survie)
t_0 = 0	N_0 = n	0	0	1	1
t_1	N_1 = N	d_1	c_1	$Q_1 = \frac{1-d_1}{N_1}$	$S_1 = q_1$
t_2	N_2 = $N_1 - D_1 - C_1$	d_2	c_2	$Q_2 = \frac{N_2-d_2}{N_2}$	$S_2 = q_1 q_2$
...	...	...	...	...	...
t_k	$N_k = N_{k-1} - D_{k-1} - C_{k-1}$	d_k	c_k	$Q_k = \frac{N_k-d_k}{N_k}$	$S_k = q_k q_{k-1} .. q_2 q_1$

Calculs à faire :

- N_k : nombre de sujet à risques (à chaque teps diminuée du nombre de décès et de censure durant la période précédante)

- Q_k : quantités de survie conditionnelle

$$\left(\frac{\text{Nombre de personne sur la periode} - \text{Nombre deces sur la periode}}{\text{Nombre de personne sur la periode}}\right)$$



5.B.2.1 Exemple 1 Données de survie avec censure :

1, 1, 1+, 1+, 1+, 2, 2, 2, 2+, 3, 3, 3+, 4+, 5+, avec “+” représente une censure.

Donc 14 individus au total.

Par défaut, un évènement qui a lieu au temps 1 a lieu dans l’intervalle 1 à 2, donc représenté sur la deuxième ligne.

Tableau :

Temps t_k	Nombre à risque N_k	Décès D_k	Censurés C_k	Probabilité condition-nelle de survie Q_k	Probabilité cumulée de survie S_k
0	14	0	0	1	1
1	14	2	3	1 - 2/14 = 6/7	6/7
2	9	3	1	1 - 3/9 = 6/9	6/7 * 6/9 = 4/7
3	5	2	1	1 - 2/5 = 3/5	4/7 * 3/5 = 12/35

Temps t_k	Nombre à risque N_k	Décès D_k	Censurés C_k	Probabilité condition-nelle de survie Q_k	Probabilité cumulée de survie S_k
4	2	0	1	1 - 0/2 = 1	12/35 * 1 = 12/35
5	1	0	1	1 - 0/1 = 1	12/35 * 1 = 12/35

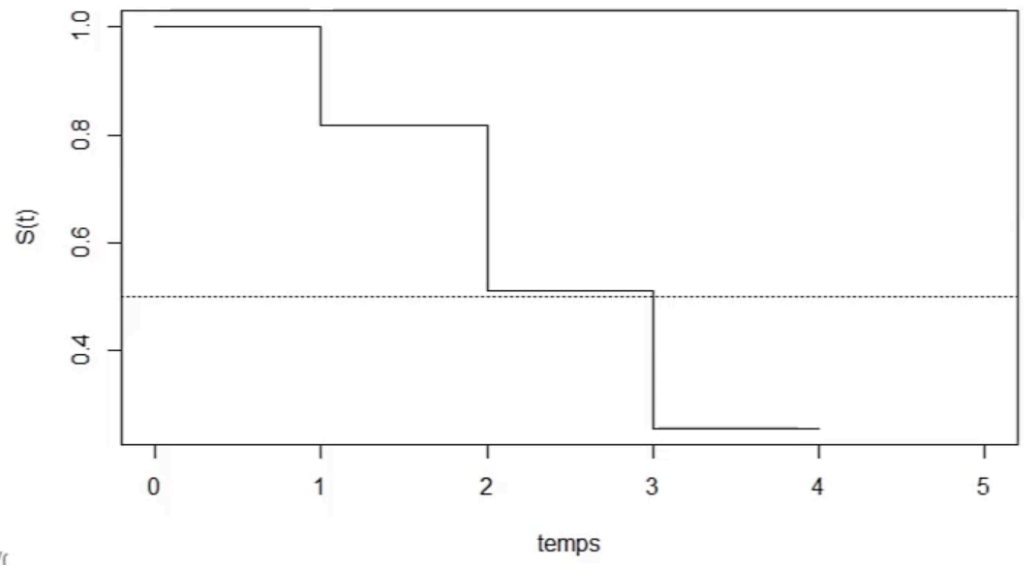
Survie médiane à 3 semaines

TP / Question 4 :

Fonction en escalier, monotone décroissante de 1 à 0 :

- Temps en abscisse
- Taux de survie cumulatif en ordonnée

Marche d'escalier à chaque production d'événement



Sur une courbe de survie, décrire la médiane de survie (la où la barre horizontale coupe la barre verticale à 0.5).

5.B.2.2 Exemple 2 avec R Faire un objet de type surv reprenant les données de survie.

D'abord charger la librairie `survival`.

```
library(survival)
```

Utiliser un jeu de données avec censure (ex : `lung` dans la librairie `survival`).

```r
head(lung) # afficher les premières lignes
```

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
```

Les colonnes importantes :

- `lung$time` : temps de SUIVI (en jours) = follow-up time
- `lung$status` : état de l'événement (1 = censuré, 2 = décès) = censoring status

Ensuite : on crée un objet de type `Surv`.

Un objet de type `Surv` combine les informations de temps de suivi et d'état de l'événement.

En gros, dans une équation contenant des données de survie, il faut mettre un objet `Surv` pour représenter la variable dépendante, qui contient à la fois le temps de suivi et l'état de l'événement.

```r
Surv(lung$time, lung$status)
```

```
  [1]  306   455   1010+ 210   883   1022+ 310   361   218   166   170   654
 [13]  728    71   567   144   613   707    61    88   301    81   624   371
 [25]  394   520   574   118   390    12   473    26   533   107    53   122
 [37]  814   965+   93   731   460   153   433   145   583    95   303   519
 [49]  643   765   735   189    53   246   689    65     5   132   687   345
 [61]  444   223   175    60   163    65   208   821+  428   230   840+  305
 [73]   11   132   226   426   705   363    11   176   791    95   196+  167
 [85]  806+  284   641   147   740+  163   655   239    88   245   588+   30
 [97]  179   310   477   166   559+  450   364   107   177   156   529+   11
[109]  429   351    15   181   283   201   524    13   212   524   288   363
[121]  442   199   550    54   558   207    92    60   551+  543+  293   202
[133]  353   511+  267   511+  371   387   457   337   201   404+  222    62
[145]  458+  356+  353   163    31   340   229   444+  315+  182   156   329
[157]  364+  291   179   376+  384+  268   292+  142   413+  266+  194   320
[169]  181   285   301+  348   197   382+  303+  296+  180   186   145   269+
[181]  300+  284+  350   272+  292+  332+  285   259+  110   286   270    81
[193]  131   225+  269   225+  243+  279+  276+  135    79    59   240+  202+
[205]  235+  105   224+  239   237+  173+  252+  221+  185+   92+   13   222+
[217]  192+  183   211+  175+  197+  203+  116   188+  191+  105+  174+  177+
```

Note

```
Surv(
    time,
    time2,
    event,
    type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'),
    origin=0)
is.Surv(x)
```

- `time`: temps de suivi pour les données à censure à droite (pour les données à censure par intervalle, le premier argument est le temps de début de l'intervalle).
- `time2`: temps de fin de l'intervalle pour les données à censure par intervalle ou les données de processus de comptage uniquement. Les intervalles sont supposés être ouverts à gauche et fermés à droite, (début, fin].
- `event`: indicateur de statut, normalement 0=vivant, 1=décédé. D'autres choix sont TRUE/FAUX (TRUE = décès) ou 1/2 (2=décès).
- `type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate')`: type de censure
- `origin=0`: pour les données de processus de comptage, l'origine de la fonction de risque, c'est à dire le temps à partir duquel on commence à compter les événements.
- `is.Surv(x)`: fonction pour vérifier si un objet `x` est de type `Surv`.

Pour faire la table de Kaplan Meier, on utilise la fonction `survfit`.

```
fit <- survfit(Surv(lung$time, lung$status)~ 1)
fit
```

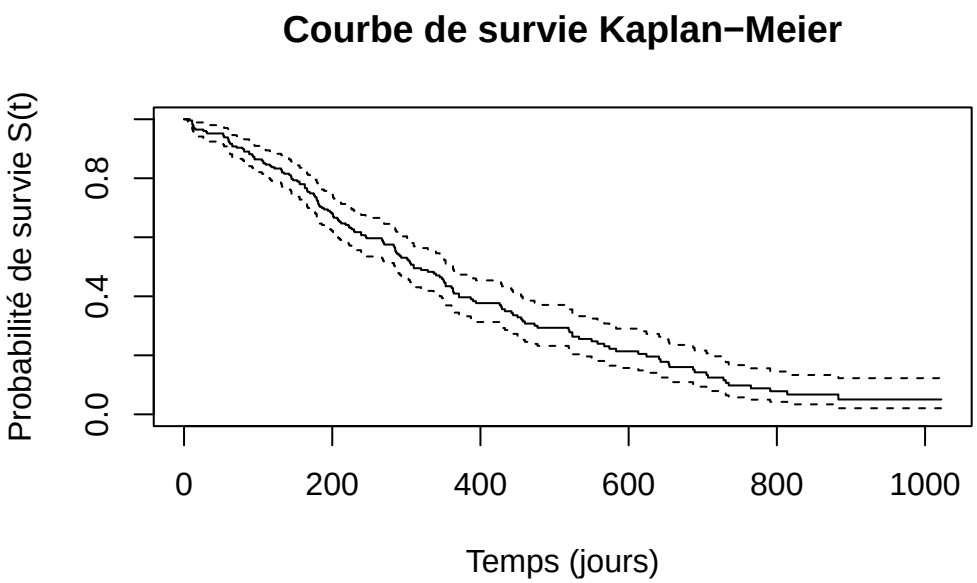
```
Call: survfit(formula = Surv(lung$time, lung$status) ~ 1)

       n events median 0.95LCL 0.95UCL
[1,] 228    165    310     285     363
```

Syntaxe :

- `survfit(formula, data, ...)`
- `formula` : formule de modélisation, où la variable dépendante est un objet `Surv` et le côté droit de la formule spécifie les variables explicatives
- `data` : jeu de données contenant les variables utilisées dans la formule.
- `~ 1` : signifie qu'il n'y a pas de variable explicative, on estime la survie globale.

Pour tracer la courbe de survie Kaplan Meier :

```r
plot(
    fit,
    xlab = "Temps (jours)",
    ylab = "Probabilité de survie S(t)",
    main = "Courbe de survie Kaplan-Meier")
```



> **Note**
>
> Pour obtenir de l'aide sur les fonctions :
>
> ```r
> ?survfit
> ?plot.survfit
> ```
>
> C'est `plot.survfit` qui est la méthode de traçage pour les objets de type `survfit`, l'aide ne se trouve pas dans `plot` seul car `plot` est une fonction générique qui peut être utilisée pour différents types d'objets.

Pour obtenir le résumé (la table) de l'estimation Kaplan Meier :

```r
summary(fit)
```

```
Call: survfit(formula = Surv(lung$time, lung$status) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    228       1   0.9956 0.00438       0.9871        1.000
   11    227       3   0.9825 0.00869       0.9656        1.000
   12    224       1   0.9781 0.00970       0.9592        0.997
   13    223       2   0.9693 0.01142       0.9472        0.992
   15    221       1   0.9649 0.01219       0.9413        0.989
   26    220       1   0.9605 0.01290       0.9356        0.986
```

30	219	1	0.9561	0.01356	0.9299	0.983
31	218	1	0.9518	0.01419	0.9243	0.980
53	217	2	0.9430	0.01536	0.9134	0.974
54	215	1	0.9386	0.01590	0.9079	0.970
59	214	1	0.9342	0.01642	0.9026	0.967
60	213	2	0.9254	0.01740	0.8920	0.960
61	211	1	0.9211	0.01786	0.8867	0.957
62	210	1	0.9167	0.01830	0.8815	0.953
65	209	2	0.9079	0.01915	0.8711	0.946
71	207	1	0.9035	0.01955	0.8660	0.943
79	206	1	0.8991	0.01995	0.8609	0.939
81	205	2	0.8904	0.02069	0.8507	0.932
88	203	2	0.8816	0.02140	0.8406	0.925
92	201	1	0.8772	0.02174	0.8356	0.921
93	199	1	0.8728	0.02207	0.8306	0.917
95	198	2	0.8640	0.02271	0.8206	0.910
105	196	1	0.8596	0.02302	0.8156	0.906
107	194	2	0.8507	0.02362	0.8056	0.898
110	192	1	0.8463	0.02391	0.8007	0.894
116	191	1	0.8418	0.02419	0.7957	0.891
118	190	1	0.8374	0.02446	0.7908	0.887
122	189	1	0.8330	0.02473	0.7859	0.883
131	188	1	0.8285	0.02500	0.7810	0.879
132	187	2	0.8197	0.02550	0.7712	0.871
135	185	1	0.8153	0.02575	0.7663	0.867
142	184	1	0.8108	0.02598	0.7615	0.863
144	183	1	0.8064	0.02622	0.7566	0.859
145	182	2	0.7975	0.02667	0.7469	0.852
147	180	1	0.7931	0.02688	0.7421	0.848
153	179	1	0.7887	0.02710	0.7373	0.844
156	178	2	0.7798	0.02751	0.7277	0.836
163	176	3	0.7665	0.02809	0.7134	0.824
166	173	2	0.7577	0.02845	0.7039	0.816
167	171	1	0.7532	0.02863	0.6991	0.811
170	170	1	0.7488	0.02880	0.6944	0.807
175	167	1	0.7443	0.02898	0.6896	0.803
176	165	1	0.7398	0.02915	0.6848	0.799
177	164	1	0.7353	0.02932	0.6800	0.795
179	162	2	0.7262	0.02965	0.6704	0.787
180	160	1	0.7217	0.02981	0.6655	0.783
181	159	2	0.7126	0.03012	0.6559	0.774
182	157	1	0.7081	0.03027	0.6511	0.770
183	156	1	0.7035	0.03041	0.6464	0.766
186	154	1	0.6989	0.03056	0.6416	0.761
189	152	1	0.6943	0.03070	0.6367	0.757
194	149	1	0.6897	0.03085	0.6318	0.753
197	147	1	0.6850	0.03099	0.6269	0.749
199	145	1	0.6803	0.03113	0.6219	0.744

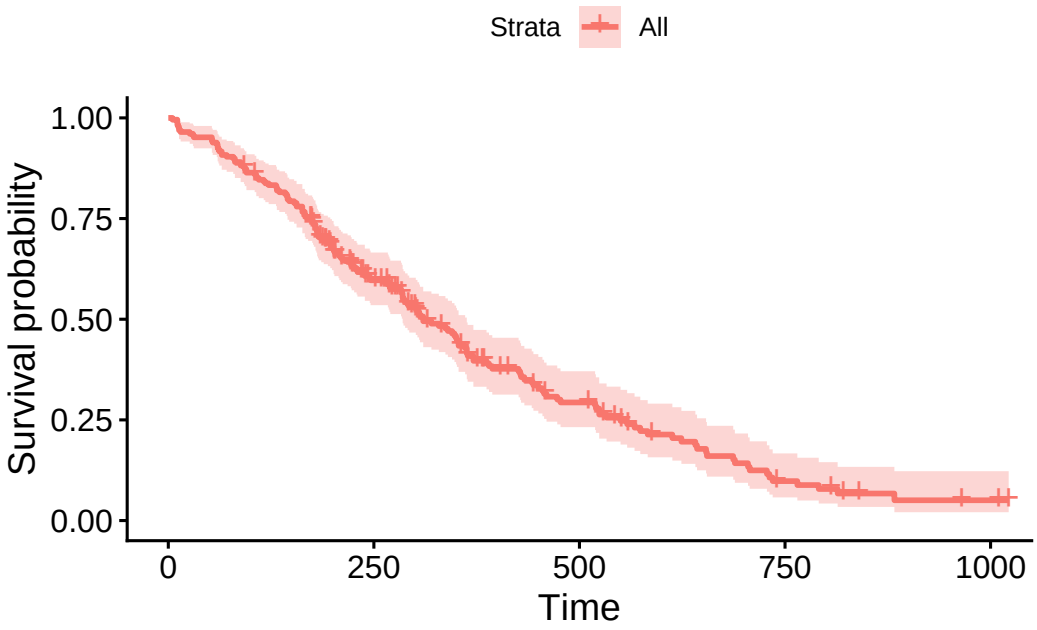
201	144	2	0.6708	0.03141	0.6120	0.735
202	142	1	0.6661	0.03154	0.6071	0.731
207	139	1	0.6613	0.03168	0.6020	0.726
208	138	1	0.6565	0.03181	0.5970	0.722
210	137	1	0.6517	0.03194	0.5920	0.717
212	135	1	0.6469	0.03206	0.5870	0.713
218	134	1	0.6421	0.03218	0.5820	0.708
222	132	1	0.6372	0.03231	0.5769	0.704
223	130	1	0.6323	0.03243	0.5718	0.699
226	126	1	0.6273	0.03256	0.5666	0.694
229	125	1	0.6223	0.03268	0.5614	0.690
230	124	1	0.6172	0.03280	0.5562	0.685
239	121	2	0.6070	0.03304	0.5456	0.675
245	117	1	0.6019	0.03316	0.5402	0.670
246	116	1	0.5967	0.03328	0.5349	0.666
267	112	1	0.5913	0.03341	0.5294	0.661
268	111	1	0.5860	0.03353	0.5239	0.656
269	110	1	0.5807	0.03364	0.5184	0.651
270	108	1	0.5753	0.03376	0.5128	0.645
283	104	1	0.5698	0.03388	0.5071	0.640
284	103	1	0.5642	0.03400	0.5014	0.635
285	101	2	0.5531	0.03424	0.4899	0.624
286	99	1	0.5475	0.03434	0.4841	0.619
288	98	1	0.5419	0.03444	0.4784	0.614
291	97	1	0.5363	0.03454	0.4727	0.608
293	94	1	0.5306	0.03464	0.4669	0.603
301	91	1	0.5248	0.03475	0.4609	0.597
303	89	1	0.5189	0.03485	0.4549	0.592
305	87	1	0.5129	0.03496	0.4488	0.586
306	86	1	0.5070	0.03506	0.4427	0.581
310	85	2	0.4950	0.03523	0.4306	0.569
320	82	1	0.4890	0.03532	0.4244	0.563
329	81	1	0.4830	0.03539	0.4183	0.558
337	79	1	0.4768	0.03547	0.4121	0.552
340	78	1	0.4707	0.03554	0.4060	0.546
345	77	1	0.4646	0.03560	0.3998	0.540
348	76	1	0.4585	0.03565	0.3937	0.534
350	75	1	0.4524	0.03569	0.3876	0.528
351	74	1	0.4463	0.03573	0.3815	0.522
353	73	2	0.4340	0.03578	0.3693	0.510
361	70	1	0.4278	0.03581	0.3631	0.504
363	69	2	0.4154	0.03583	0.3508	0.492
364	67	1	0.4092	0.03582	0.3447	0.486
371	65	2	0.3966	0.03581	0.3323	0.473
387	60	1	0.3900	0.03582	0.3258	0.467
390	59	1	0.3834	0.03582	0.3193	0.460
394	58	1	0.3768	0.03580	0.3128	0.454
426	55	1	0.3700	0.03580	0.3060	0.447

```
  428      54        1    0.3631 0.03579        0.2993         0.440
  429      53        1    0.3563 0.03576        0.2926         0.434
  433      52        1    0.3494 0.03573        0.2860         0.427
  442      51        1    0.3426 0.03568        0.2793         0.420
  444      50        1    0.3357 0.03561        0.2727         0.413
  450      48        1    0.3287 0.03555        0.2659         0.406
  455      47        1    0.3217 0.03548        0.2592         0.399
  457      46        1    0.3147 0.03539        0.2525         0.392
  460      44        1    0.3076 0.03530        0.2456         0.385
  473      43        1    0.3004 0.03520        0.2388         0.378
  477      42        1    0.2933 0.03508        0.2320         0.371
  519      39        1    0.2857 0.03498        0.2248         0.363
  520      38        1    0.2782 0.03485        0.2177         0.356
  524      37        2    0.2632 0.03455        0.2035         0.340
  533      34        1    0.2554 0.03439        0.1962         0.333
  550      32        1    0.2475 0.03423        0.1887         0.325
  558      30        1    0.2392 0.03407        0.1810         0.316
  567      28        1    0.2307 0.03391        0.1729         0.308
  574      27        1    0.2221 0.03371        0.1650         0.299
  583      26        1    0.2136 0.03348        0.1571         0.290
  613      24        1    0.2047 0.03325        0.1489         0.281
  624      23        1    0.1958 0.03297        0.1407         0.272
  641      22        1    0.1869 0.03265        0.1327         0.263
  643      21        1    0.1780 0.03229        0.1247         0.254
  654      20        1    0.1691 0.03188        0.1169         0.245
  655      19        1    0.1602 0.03142        0.1091         0.235
  687      18        1    0.1513 0.03090        0.1014         0.226
  689      17        1    0.1424 0.03034        0.0938         0.216
  705      16        1    0.1335 0.02972        0.0863         0.207
  707      15        1    0.1246 0.02904        0.0789         0.197
  728      14        1    0.1157 0.02830        0.0716         0.187
  731      13        1    0.1068 0.02749        0.0645         0.177
  735      12        1    0.0979 0.02660        0.0575         0.167
  765      10        1    0.0881 0.02568        0.0498         0.156
  791       9        1    0.0783 0.02462        0.0423         0.145
  814       7        1    0.0671 0.02351        0.0338         0.133
  883       4        1    0.0503 0.02285        0.0207         0.123
```

Pour faire un autre tracé avec “ggsurvplot” de la librairie “survminer” :

```r
library(survminer)
ggsurvplot(fit, data = lung)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
i The deprecated feature was likely used in the ggpubr package.
  Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
```



On va essayer de faire la même chose en comparant les groupes de sexe (variable `sex` dans le jeu de données `lung`).

```
fit_sex <- survfit(Surv(time, status) ~ sex, data = lung)
fit_sex
```

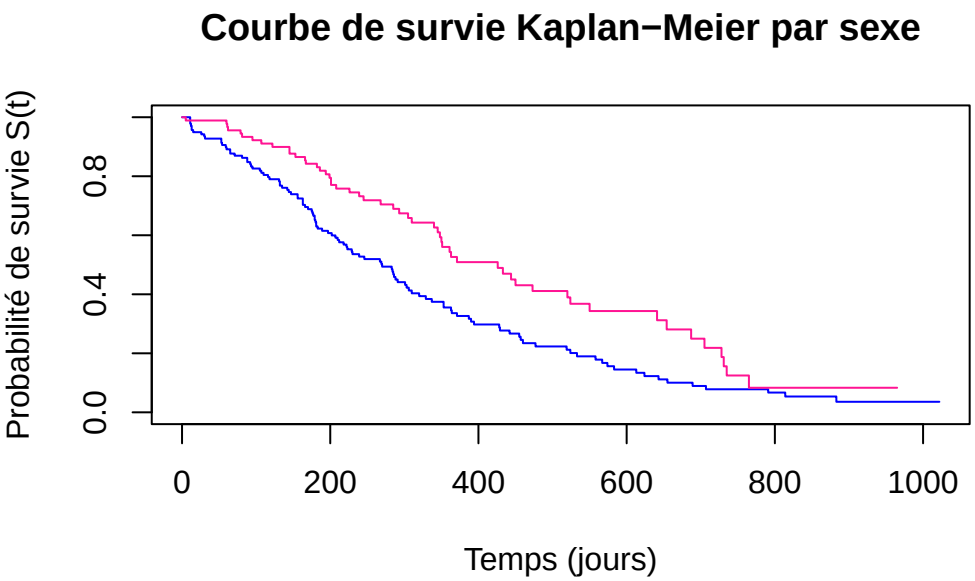
```
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     212     310
sex=2  90     53    426     348     550
```

Syntaxe :

- `Surv(time, status) ~ sex` : on modélise la survie en fonction de la variable explicative `sex`.

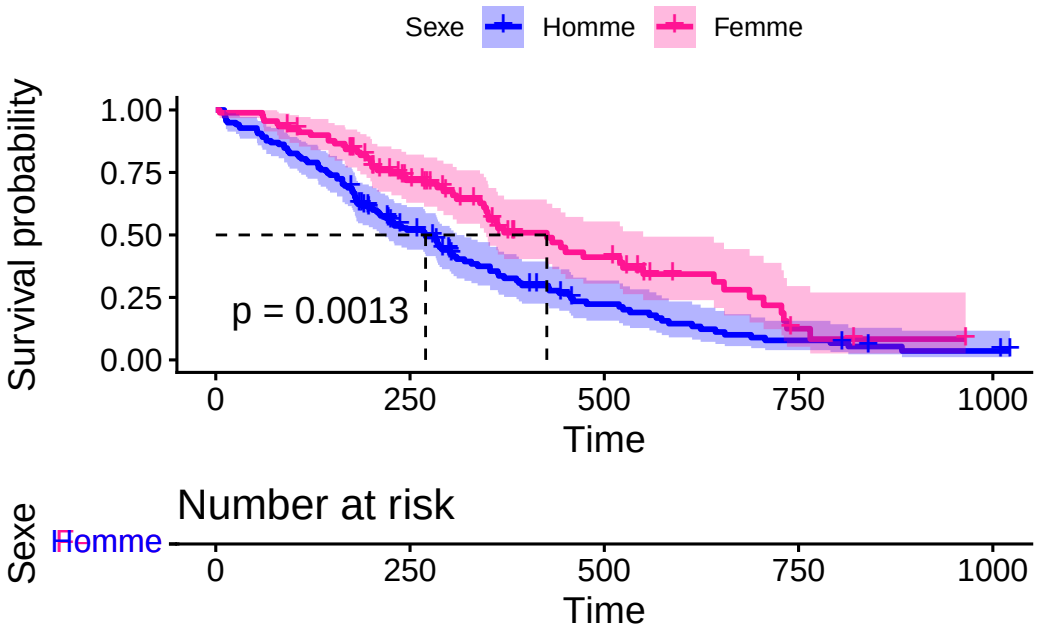
Pour tracer la courbe de survie Kaplan Meier par sexe :

```
plot(
    fit_sex,
    xlab = "Temps (jours)",
    ylab = "Probabilité de survie S(t)",
    main = "Courbe de survie Kaplan-Meier par sexe",
    col = c("blue", "deeppink"),
    conf.int = FALSE # pas d'intervalle de confiance
)
```



On le refait avec `ggsurvplot` + ajout de l'intervalle de confiance + test de log-rank.

```
ggsurvplot(
    fit_sex,
    data = lung,
    conf.int = TRUE,
    pval = TRUE,
    risk.table = TRUE,
    risk.table.col = "strata",
    legend.labs = c("Homme", "Femme"),
    legend.title = "Sexe",
    palette = c("blue", "deeppink"),
    surv.median.line = "hv"
)
```



6 Comparaison de courbes de survie : test de Log-Rank

2 cadres principaux :

- Essai comparatif : différence de délais de survie entre 2 groupes
- Étude épidémiologique (cohorte) : impact de facteurs de risque sur la survie

Comparaison des fonctions de survie dans leur ensemble.

6.A Principe

Comparaison de deux (p) fonctions de survie à partir de deux (p) échantillons indépendants.
Comparaison de deux fonctions de survie / totalité des courbes
Hypothèse nulle H_0 : les deux fonctions de survie ne sont pas différentes
Hypothèse alternative H_1 : les deux fonctions de survie sont différentes

6.B Test de Log-Rank dans R

Utilisation de la fonction `survdiff` de la librairie `survival`.

```r
# Charger les librairies nécessaires
library(survival)
library(survminer)
# Créer un objet Surv
surv_object <- Surv(time = lung$time, event = lung$status)
# Ajuster le modèle de survie par sexe
surv_fit <- survfit(surv_object ~ lung$sex)
# Effectuer le test de Log-Rank : fonction survdiff
logrank_test <- survdiff(surv_object ~ lung$sex)
# Afficher les résultats du test
print(logrank_test)
```

```
Call:
survdiff(formula = surv_object ~ lung$sex)

             N Observed Expected (O-E)^2/E (O-E)^2/V
lung$sex=1 138      112     91.6      4.55      10.3
lung$sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Résultats :

- **N** : nombre d'individus dans chaque groupe (138 hommes, 90 femmes).
- **Observed** : nombre d'événements observés (décès) dans chaque groupe (112 hommes, 53 femmes).

- **Expected** : nombre d'événements attendus dans chaque groupe sous l'hypothèse nulle (91.6 hommes, 73.4 femmes).
- **(O-E)^2/E** : contribution de chaque groupe au test du chi carré basé sur la différence entre les événements observés et attendus.
- **(O-E)^2/V** : contribution de chaque groupe au test du chi carré basé sur la variance des différences observées-attendues.
- **Chisq** : statistique du test du chi carré (10.3).
- **degrees of freedom** : degrés de liberté du test (1, car deux groupes).
- **p** : valeur p associée au test (0.001).

Interprétation :

- La valeur p (0.001) est inférieure au seuil de signification habituel (0.05), ce qui indique une différence statistiquement significative entre les fonctions de survie des hommes et des femmes.

Conclusion : on rejette l'hypothèse nulle et on conclut qu'il y a une différence significative entre les fonctions de survie des hommes et des femmes dans cette étude.

> **!** Important
>
> Donc les étapes sont:
>
> 1. Créer un objet de survie `Surv` comprenant les temps de suivi et l'état de l'événement.
> 2. Ajuster un modèle de survie avec `survfit` en fonction de la variable explicative (ici le sexe).
> 3. Effectuer le test de Log-Rank avec la fonction `survdiff`.

---

7 Modèle de Cox : analyse multivariée de survie

Modèle de Cox = modèle semi-paramétrique = modélisation de l'effet des variables explicatives sur la fonction de risque sans faire d'hypothèse sur la forme de la fonction de risque.

- Décrire et tester la dépendance
  - Entre une réponse (fonction de risque instantané $h(t)$ ou $\lambda(t)$)
  - et un vecteur de variables explicatives (covariables qualitatives ou quantitatives) $Z = (Z_1, Z_2, ..., Z_p)$

$h(t|Z_1, ..., Z_p) = h_0(t) \times \text{exp}(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$

- $h_0(t)$: fonction de risque de base (baseline hazard function) = fonction de risque instantané lorsque toutes les covariables Z_i sont égales à 0.
- exp$(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$: effet multiplicatif des P covariables sur la fonction de risque instantané.

- β_i : coefficient associé à la covariable Z_i, représentant l'effet de cette covariable sur le risque instantané.
  - Relation entre β_i et le hazard ratio (HR) : $HR =$ exp(β_i) (donc $\beta_i =$ log(HR))

2 hypothèses principales : **proportionnalité des risques** et **log-linéarité**.

7.A Hypothèse de proportionnalité des risques

Hypothèse de proportionnalité des risques : les rapports des risques entre les individus restent constants dans le temps, c'est à dire que l'effet des covariables sur le risque instantané est multiplicatif et ne dépend pas du temps.

$\frac{h(t|Z_k=1)}{h(t|Z_k=0)} =$ exp$(\beta k) = constante$

Quand on compare 2 groupes (exemple : fumeur vs non-fumeur), le rapport de leurs risques instantanés reste le même tout au long du suivi, et ce rapport de risques = hazard ratio (HR).

Dire que les risques sont proportionnels signifie :

- à chaque instant t, le groupe A a par exemple 1,5 fois plus de risque de l'événement que le groupe B ;
- ce facteur 1,5 ne change pas au cours du temps.

Exemple : On suit 100 patients opérés.

Variable explicative : fumeur (1) vs non-fumeur (0).

Supposons : HR = 2.

Cela veut dire :

- à tout moment du suivi, un fumeur a le double du risque instantané d'avoir une complication,
- même si le risque global diminue avec le temps (fin de la période postopératoire aiguë), le ratio reste stable entre fumeur et non fumeur.

Si cette hypothèse n'est pas vraie (ex : au début les fumeurs sont à très haut risque, mais plus tard le risque redevient identique), alors le modèle de Cox classique n'est plus adapté. (*Dans ce cas, on peut utiliser des modèles de Cox avec effets temporels*).

> **Note**

7.B Hypothèse de log-linéarité

Hypothèse de log-linéarité : **concerne les variables quantitatives**.

L'effet des covariables est linéaire sur le logarithme du risque instantané et pas directement sur le risque.

C'est à dire que chaque unité d'augmentation de la covariable Z_i entraîne une augmentation constante du logarithme du risque instantané.

$\rightarrow$ Chaque augmentation d'une unité de la variable produit le même pourcentage de variation du risque.

$\log(h(t|Z_1, ..., Z_p)) = \log(h_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p$

Exemple :

Variable explicative : âge (en années).

On suppose que $\beta = 0,05$ (avec β le coefficient associé à l'âge dans le modèle de Cox).

Cela veut dire :

- chaque année supplémentaire multiplie le risque instantané par $exp(0,05) \approx 1,05$
- donc +5 % de risque par année d'âge.
  - Si on passe de 50 à 51 ans : +5 %.
  - Si on passe de 70 à 71 ans : encore +5 %.
- L'effet est proportionnel, constant.

> **Note**

7.C Interprétation des coefficients du modèle de Cox

Modèle de Cox : $h(t \mid Z) = h_0(t) \times \exp(\beta Z)$

où :

- $h_0(t)$ est le risque instantané de base,
- Z est la covariable (discrète ou continue),
- β est le coefficient associé,
- $\exp(\beta)$ est le hazard ratio (HR).

7.C.1 Variable discrète

Variable discrète (ex : sexe, fumeur/non-fumeur)** :

Groupe de référence A avec $Z_A = 0$ et groupe comparé B avec $Z_B = 1$.

- $Z_A = 0 : h_A(t) = h_0(t)$ (risque instantané du groupe de référence))
- $Z_B = 1 : h_B(t) = h_0(t) \times \exp(\beta)$ (risque instantané du groupe comparé))

Le hazard ratio est :

$HR = \frac{h_B(t)}{h_A(t)} = \frac{h_0(t) \times \exp(\beta)}{h_0(t)} = \exp(\beta)$

Points importants :

- Le HR ne dépend pas du risque de base $h_0(t)$ (mais seulement de β)
- Il est supposé constant et multiplicatif dans le temps (proportionnalité des risques).
- Il s'interprète comme un rapport d'incidence instantané (ce n'est pas une probabilité).

Exemple : On étudie la mortalité après chirurgie.

Variable : fumeur (1) vs non-fumeur (0).

Le modèle donne $\beta = 0{,}69$.

Alors :

$HR = \exp(0{,}69) \approx 2$

Interprétation :

- À tout instant du suivi, un fumeur a deux fois plus de risque instantané de mourir qu'un non-fumeur.
- L'effet est multiplicatif et constant au cours du temps.
- Ce n'est pas une probabilité ; c'est un rapport de taux instantanés.

7.C.2 Variable continue

Variable continue (ex : âge, pression artérielle)** :

Le modèle utilise :

$h(t \mid X) = h_0(t) \times \exp(\beta X)$

Ici, $\exp(\beta)$ correspond au HR pour une unité d'augmentation de la variable.

- Le Hazard Ratio est exprimé en unité de x (ex HR pour 1 an d'âge)
- Hypothèse de **modélisation** :
  - Chaque unité d'augmentation de la variable même pourcentage de variation du risque instantané
  - L'effet d'un an de plus sur le risque instanté ne dépend pas de l'âge (il est indépendant du niveau de la variable)
  - L'effet est constant dans le temps (proportionnalité des risques)

Exemple : Variable : âge (en années).

Supposons $\beta = 0{,}05$.

Alors : $HR = \exp(0{,}05) \approx 1{,}051$

Interprétation :

- Pour chaque année supplémentaire, le risque instantané augmente de 5,1 %.
- Passer de 50 51 ans : +5,1 %.
- Passer de 70 71 ans : encore +5,1 %.
- Le modèle considère donc une pente constante sur le logarithme du risque.

Cela découle directement de l'écriture :

$\log h(t) = \log h_0(t) + \beta X$

qui impose une relation linéaire sur le log-risque.

7.C.3 Interprétation du HR

- **HR = 1** : pas d'effet de la covariable sur le risque instantané.
- **HR > 1** : la covariable est associée à une augmentation du risque instantané.
  - Ex : HR = 2 doublement du risque instantané.
  - Ex : HR = 1,5 augmentation de 50 % du risque instantané.
- **HR < 1** : la covariable est associée à une diminution du risque instantané.

Modèle univariable : HR brut

Modèle multivariable : HR ajusté (pour les autres covariables du modèle)

> ❗ Important
>
> **Ne pas confondre** :
>
> - Valeur vraie dans la population (interprétable directoement)
> - Estimation dans l'échantillon (avec intervalle de confiance et test statistique)

7.C.4 Résumé

7.C.5 Tableau de synthèse : interprétation des coefficients du modèle de Cox

Variable	Forme du modèle	HR = $exp(\beta)$	Interprétation	Exemple
Discrète binaire (0/1)	`$h(t\|Z) = h0(t) * exp(\beta Z)$`	$exp(\beta)$	Rapport de risque instantané entre Z=1 et Z=0, constant dans le temps	$\beta = 0.69$ HR = 2 : risque doublé
Discrète multi-catégorielle	Référence + indicateurs	$exp(\beta_k)$	Comparaison de chaque catégorie à la référence	Réf = non-fumeur ; fumeur HR=2 ; ex-fumeur HR=1.3
Continue (par unité)	`$h(t\|X) = h0(t) * exp(\beta X)$`	$exp(\beta)$	Variation du risque instantané pour +1 unité de X	$\beta = 0.05$ HR = 1.051 : +5.1 % par unité
Continue (pour ΔX unités)	`$h(t\|X) = h0(t) * exp(\beta X)$`	$exp(\beta \cdot \Delta X)$	Variation du risque pour ΔX unités	$\Delta X = 10$ ans : HR = exp$(0.05 \times 10) = 1.65$

Variable	Forme du modèle	HR = $exp(\beta)$	Interprétation	Exemple
Log-linéarité	Linéarité sur log(h)	—	Chaque unité □ même pourcentage de variation du risque	Effet constant : 50□51 = 70□71
Proportionnalité des risques	HR constant dans le temps	—	L'effet ne change pas avec le temps	Courbes "parallèles" en risque instantané

7.D Tests statistiques pour les coefficients du modèle de Cox

Le modèle de Cox estime un ensemble de coefficients β . Pour tester si une covariable a un effet significatif sur le risque instantané, deux tests principaux existent :

1. Le test de Wald
2. Le test du rapport de vraisemblance (Likelihood Ratio Test, LRT)

7.D.1 Hypothèses

- **Hypothèse nulle** H_0 : la covariable n'a pas d'effet sur le risque instantané ($\beta = 0$ □ $HR = 1$)
- **Hypothèse alternative** H_1 : la covariable a un effet sur le risque instantané ($\beta \neq 0$ □ $HR \neq 1$)

7.D.2 Tests disponibles : Wald et Rapport de vraisemblance (Likelihood Ratio Test, LRT)

7.D.2.1 Test de Wald Le modèle de Cox fournit une estimation de l'écart type du coefficient $\hat{\beta}$.

Le test de Wald utilise l'estimation du coefficient $\hat{\beta}$ et son écart type (son incertitude) pour calculer une statistique de test.

Calcul de la statistique de Wald :

$Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$

Avec :

- $\hat{\beta}$: estimateur du coefficient de la covariable
- $SE(\hat{\beta})$: écart type de l'estimateur

Sous l'hypothèse nulle, Z suit une loi normale centrée réduite, parce que pour de grands échantillons, l'estimateur $\hat{\beta}$ est asymptotiquement normal, c'est à dire que sa distribution approche une loi normale lorsque la taille de l'échantillon augmente.

On compare donc la valeur absolue de Z à la table de la loi normale pour obtenir une p-value.
ex : si $|Z| > 1.96$, p-value < 0.05

- Si $|Z|$ est grand la variable est significative.
- Si $|Z| < 1.96$ p-value > 0.05 variable non significative.

Exemple : Supposons que le coefficient pour fumeur soit :

- $\hat{\beta}$ = 0.69
- $SE(\hat{\beta})$ = 0.30

$Z = 0.69/0.30 = 2.30$
 p-value < 0.05 effet significatif.

> **!** Important
>
> Le test de Wald donne une p-value pour chaque coefficient du modèle, parce qu'il teste chaque coefficient séparément.

Pourquoi le test de Wald donne une p-value par coefficient ?

- Parce qu'il teste chaque coefficient séparément :
- $H_0 : \beta_i = 0$
- et chaque coefficient β_i est testé individuellement avec son propre estimateur $\hat{\beta}_i$ et son propre écart type $SE(\hat{\beta}_i)$.
- Donc, si le modèle contient :
  - 1 variable quantitative 1 coefficient 1 p-value
  - 1 variable qualitative à 3 modalités 2 coefficients 2 p-values
  - p variables p p-values

7.D.2.2 Test du rapport de vraisemblance (Likelihood Ratio Test, LRT) = modèles emboîtés que l'on compare

= **modèle complet** = avec la covariable *vs* **modèle réduit** = sans la covariable

On compare la log-vraisemblance des deux modèles (log-vraisemblance = mesure de l'adéquation du modèle aux données).

Calcul de la statistique du rapport de vraisemblance :

$LR = 2 \times (\log L_2 - \log L_1)\ \chi^2(\text{p}\ ddl)$

Avec :

- L_1 : la log-vraisemblance du modèle réduit (le moins riche)
- L_2 : la log-vraisemblance du modèle complet (le plus riche)
- p : nombre de paramètres ajoutés entre les deux modèles (nombre de degrés de liberté)

Sous H_0, le test du rapport de vraisemblance suit une loi du chi carré avec p degrés de liberté.

Quand on dit que le test est fait sous H_0, cela signifie que l'on suppose que la covariable n'a pas d'effet sur le risque instantané.

On compare la valeur de LR à la table du chi carré pour obtenir une p-value.

Pourquoi le LRT donne une seule p-value par variable ?

Car il ne teste pas β_1, β_2, β_3 individuellement mais l'ensemble du bloc associé à la variable :

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Donc :

- Une variable binaire 1 coefficient 1 p-value
- Une variable qualitative à 4 niveaux 3 coefficients 1 p-value globale
- Un bloc de variables (ex : spline, interaction, polynôme) 1 p-value globale

7.D.2.3 Comparaison des deux tests Nombre de p-values

- Wald plusieurs p-values (une par coefficient)
- LRT une seule p-value par variable, même si cette variable comporte plusieurs coefficients.

> **Note**
>
> Variable : statut fumeur (oui/non)
>
> - Wald : 1 coefficient 1 p-value
> - LRT : 1 comparaison modèle complet vs réduit 1 p-value
>
> Variable : IMC en 3 catégories (normal, surpoids, obésité) codée en 2 indicatrices car pour un facteur à k modalités, un modèle de Cox crée $k-1$ variables indicatrices (dummy variables). Donc pour 3 catégories 2 coefficients β_1 et β_2.
>
> - Wald : 2 coefficients 2 p-values
> - LRT : 1 comparaison modèle complet vs réduit 1 p-value

Quel test utiliser ?

Le test du rapport de vraisemblance est plus utilisé pour tester un ensemble de covariables simultanément (modèle multivariable).

À la limité d'une taille d'échantillon très grande, ces deux tests donnent des résultats similaires.

> **Important**
>
> Le top : rapport de vraisemblance pour chaque variable, rapporté dans un tableau

Tableau comparatif

Caractéristique	Test de Wald	Test du rapport de vraisemblance (LRT)
Ce qui est testé	Chaque coefficient séparément	L'ensemble des coefficients d'une même variable
Nombre de p-values obtenues	Une p-value par coefficient	Une seule p-value par variable (ou par bloc de variables)
Ex : variable binaire	1 coefficient 1 p-value	1 degré de liberté 1 p-value
Ex : variable qualitative à 4 niveaux	3 coefficients 3 p-values	Test global sur les 3 coefficients 1 p-value
Interprétation	Effet individuel de chaque niveau / coefficient	Apport global de la variable au modèle
Type de test	Normal (approximation asymptotique)	Khi-deux
Fiabilité	Peut être instable si les SE sont mal estimés	Plus robuste en théorie
Usage typique	Lecture coefficient par coefficient	Comparaison modèles (réduit vs complet)

8 TP

8.A Modèle de Cox avec R

- Charger la librairie `survival`.
  - `Surv` : créer un objet de survie
  - `survfit` : ajuster un modèle de survie (Kaplan Meier)
  - `survdiff` : test de Log-Rank
  - `coxph` : ajuster un modèle de Cox
- Conditions du df : "survival object" faisable
  - Temps de suivi
  - État de l'événement

8.A.1 Exemple avec le jeu de données `lung`

```
# Charger la librairie survival
library(survival)
# Afficher les premières lignes du jeu de données
head(lung)
```

8.A.1.1 Charger et afficher les premières lignes du jeu de données `lung`

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
```

```
1     3  306       2  74    1        1        90        100      1175       NA
2     3  455       2  68    1        0        90         90      1225       15
3     3 1010       1  56    1        0        90         90        NA       15
4     5  210       2  57    1        1        90         60      1150       11
5     1  883       2  60    1        0       100         90        NA        0
6    12 1022       1  74    1        1        50         80       513        0
```

```r
# Créer un objet Surv
surv_obj <- Surv(time = lung$time, event = lung$status)
surv_obj
```

8.A.1.2 Créer un objet de type `Surv` avec les colonnes `time` et `status`

```
  [1]  306   455  1010+  210   883  1022+  310   361   218   166   170   654
 [13]  728    71   567   144   613   707    61    88   301    81   624   371
 [25]  394   520   574   118   390    12   473    26   533   107    53   122
 [37]  814   965+   93   731   460   153   433   145   583    95   303   519
 [49]  643   765   735   189    53   246   689    65     5   132   687   345
 [61]  444   223   175    60   163    65   208   821+  428   230   840+  305
 [73]   11   132   226   426   705   363    11   176   791    95   196+  167
 [85]  806+  284   641   147   740+  163   655   239    88   245   588+   30
 [97]  179   310   477   166   559+  450   364   107   177   156   529+   11
[109]  429   351    15   181   283   201   524    13   212   524   288   363
[121]  442   199   550    54   558   207    92    60   551+  543+  293   202
[133]  353   511+  267   511+  371   387   457   337   201   404+  222    62
[145]  458+  356+  353   163    31   340   229   444+  315+  182   156   329
[157]  364+  291   179   376+  384+  268   292+  142   413+  266+  194   320
[169]  181   285   301+  348   197   382+  303+  296+  180   186   145   269+
[181]  300+  284+  350   272+  292+  332+  285   259+  110   286   270    81
[193]  131   225+  269   225+  243+  279+  276+  135    79    59   240+  202+
[205]  235+  105   224+  239   237+  173+  252+  221+  185+   92+    13   222+
[217]  192+  183   211+  175+  197+  203+  116   188+  191+  105+  174+  177+
```

```r
# Test de Log-Rank pour le sexe
logrank_sex <- survdiff(surv_obj ~ lung$sex)
print(logrank_sex)
```

8.A.1.3 Tester l'effet du sexe (`sex`) sur la survie avec un log-rank test

```
Call:
survdiff(formula = surv_obj ~ lung$sex)

              N Observed Expected (O-E)^2/E (O-E)^2/V
```

```
lung$sex=1 138      112     91.6      4.55      10.3
lung$sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

8.A.1.4 Ajuster un modèle de Cox avec `sex` comme covariable

Fonction `coxph` (pour Cox Proportional Hazards) :

```r
# Ajuster un modèle de Cox avec le sexe comme covariable
cox_model_sex <- coxph(surv_obj ~ lung$sex)
summary(cox_model_sex)
```

```
Call:
coxph(formula = surv_obj ~ lung$sex)

  n= 228, number of events= 165

            coef exp(coef) se(coef)      z Pr(>|z|)
lung$sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
lung$sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

Résultats :

`coef` :

- estimation du coefficient $\hat{\beta}$ pour la variable `sex` (-0.5310) =
- = $h(t|\text{sex}) = h_0(t)\,\exp(\beta \times \text{sex})$
  - avec `sex` = 1 pour les femmes et 0 pour les hommes.
  - coefficient β correspond à l'impact du sexe (femme vs homme) sur le logarithme du risque instantané.
- `exp(coef)` : estimation du hazard ratio (HR) pour les femmes par rapport aux hommes (0.5880)
  - $\exp(\hat{\beta}) = HR$
  - = $\exp(-0.5310) \approx 0.588$

  - Interprétation : à tout instant du suivi, les femmes ont environ 41.2 % (1 - 0.588) de risque instantané en moins de mourir par rapport aux hommes.

- `se(coef)` : écart type de l'estimateur du coefficient (0.1672) =
  - incertitude sur l'estimation de $\hat{\beta}$.
  - Utilisé pour le test de Wald (formule : $Z = \hat{\beta}/SE(\hat{\beta})$).
  - pour obtenir l'IC de l'HR : $IC_{95\%} =$ exp$(\hat{\beta} \pm 1.96 \times SE(\hat{\beta}))$
- `z` : statistique de test de Wald (-3.176) =
  - calculée comme $z = \hat{\beta}/SE(\hat{\beta})$.
  - Utilisée pour obtenir la p-value associée au test de Wald.
  - Valeur absolue élevée de `z` indique un effet significatif.
  - Plus bas, il est écrit 10.09, qui est le carré de -3.176 (car le test du chi carré est le carré du test de Wald).
- `Pr(>|z|)` : p-value associée au test de Wald (0.00149) =
  - probabilité d'observer une statistique de test aussi extrême que `z` sous l'hypothèse nulle ($\beta = 0$).
  - p-value < 0.05 indique que l'effet du sexe sur le risque instantané est statistiquement significatif.
- `exp(-coef)` : inverse du hazard ratio (1.701) =
  - exp$(-\hat{\beta}) = 1/HR$
  - Interprétation : à tout instant du suivi, les hommes ont environ 70.1 % (1.701 - 1) de risque instantané en plus de mourir par rapport aux femmes.
- `lower .95` et `upper .95` : bornes inférieure et supérieure de l'intervalle de confiance à 95 % pour le hazard ratio (0.4237, 0.816) =
  - Calculé comme $IC_{95\%} =$ exp$(\hat{\beta} \pm 1.96 \times SE(\hat{\beta}))$.
  - Indique la précision de l'estimation du HR.
  - Si l'IC ne contient pas 1, l'effet est statistiquement significatif.
- `Concordance` : mesure de la capacité prédictive du modèle (0.579) =
  - Valeur entre 0.5 (aucune capacité prédictive) et 1 (prédiction parfaite).
  - Indique dans quelle mesure le modèle classe correctement les individus en fonction de leur risque.
- `Likelihood ratio test` : statistique du test du rapport de vraisemblance (10.63) avec sa p-value (0.001) =
  - Compare le modèle complet (avec `sex`) au modèle réduit (sans `sex`).
  - p-value < 0.05 indique que l'ajout de `sex` améliore significativement le modèle.
- `Wald test` : statistique du test de Wald (10.09) avec sa p-value (0.001) =
  - Teste l'effet individuel de `sex` sur le risque instantané.

- p-value < 0.05 indique que l'effet de `sex` est statistiquement significatif.
- Intérêt par rapport à `z` et `Pr(>|z|)` : c'est la même information mais exprimée en chi carré.
  * Exprimé en chi^2 pour faciliter la comparaison avec d'autres tests (comme le LRT).
  * Mais en vrai, c'est juste le carré de `z`.

> **Note**
>
> **Pourquoi le test de Wald global renvoie un chi-2 ?**
> Parce que le carré d'une variable normale centrée réduite suit une loi du chi-2 à 1 degré de liberté.
> C'est un théorème fondamental :
> $(\mathcal{N}(0,1))^2 \sim \chi^2(1)$
> Donc :
> $W = z^2$
> Dans l'output :
> $z = -3.176$
> $ Wald □^2 = (−3.176)^2 = 10.09$
> Donc :
>
> - Le test individuel affiche z (test indidividuel = test de Wald pour le coefficient)
> - Le test global affiche z² = chi-2 (= test de Wald global pour la variable)
>
> Même p-value, présentation différente.

> **Note**
>
> **Qu'est-ce que** N dans $N(0,1)$?
> $N(0,1)$ signifie loi Normale centrée réduite, c'est-à-dire :
>
> - moyenne = 0
> - variance = 1
> - forme en cloche parfaitement définie
>
> Quand on écrit :
> $z \sim \mathcal{N}(0,1)$
> cela veut dire : la statistique z suit une loi normale de moyenne 0 et de variance 1.
> Dans ce contexte N désigne la loi Normale : $N(\mu, \sigma^2)$ = loi normale de moyenne μ et variance σ^2.
> Donc N = normal distribution.
> **Pourquoi** z suit une loi normale $N(0,1)$?
> Dans un modèle de Cox, l'estimateur du coefficient $\hat{\beta}$ est asymptotiquement normal :
> $\hat{\beta} \approx \mathcal{N}(\beta, SE(\hat{\beta})^2)$
> Sous l'hypothèse nulle $H_0 : \beta = 0$, cela devient :

$\hat{\beta} \approx \mathcal{N}(0, SE^2)$
Si on divise par l'écart-type (on le fait pour standardiser la variable, c'est à dire pour lui donner une moyenne 0 et une variance 1) :
$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$
Alors z suit :
$z \sim \mathcal{N}(0, 1)$
Donc :

- z = coefficient standardisé
- z = suit une loi normale centrée réduite
- C'est pour ça qu'on calcule $Pr(> |z|)$ avec la loi normale

NB : diviser par l'écart-type est une opération classique en statistique pour standardiser une variable et obtenir une variable sans unité, ce qui permet de comparer des variables entre elles.
Ça permet aussi d'obtenir une statistique de test qui suit une loi connue (ici la loi normale centrée réduite car on divise par l'écart-type).

8.A.1.5 Obtenir l'intervalle de confiance à 95 % pour le hazard ratio

Coefficients β et leur exponentielle (HR)

```
# Calculer les coefficients ß
coefficients(cox_model_sex)
```

```
  lung$sex
-0.5310235
```

```
# Exponentielle (hazard ratio)
exp(coefficients(cox_model_sex))
```

```
 lung$sex
0.5880028
```

Pour l'intervalle de confiance à 95 % :

```
# Obtenir l'intervalle de confiance à 95 % pour les coefficients
confint(cox_model_sex)
```

```
              2.5 %     97.5 %
lung$sex -0.8586875 -0.2033595
```

```r
# Obtenir l'intervalle de confiance à 95 % pour le hazard ratio
exp(confint(cox_model_sex))
```

```
             2.5 %    97.5 %
lung$sex 0.4237178 0.8159848
```

Risque de base : fonction de risque de base estimée

```r
# Obtenir la fonction de risque de base estimée
base_hazard <- basehaz(cox_model_sex, centered = FALSE)
head(base_hazard)
```

```
       hazard time
1 0.008907762    5
2 0.035855686   11
3 0.044934060   12
4 0.063237497   13
5 0.072463620   15
6 0.081740068   26
```

9 TP 2

Base de données `gehan` (disponible dans la librairie `MASS`) : 42 patients leucémiques avec 2 groupes de traitement 6-MP vs controls avec les durées de rémission

1. Combien y a-t-il d'évènements et de censures dans chacun des groupes ?
2. Combien y-a-t-il de sujets exposés au risque de rechute à 12 et 18 semaines dans chaque groupe ?
3. Un test du Chi^2 de Pearson serait-il approprié pour comparer les proportions d'évènements entre les deux groupes ? Justifiez votre réponse.
4. Quelle est la survie sans rechute à 6 et 12 semaines dans les deux groupes de traitement ?
5. Quelle est la médiane de survie sans rechute dans chaque groupe de traitement ?
6. Estimez et tracez les courbes de survie de Kaplan-Meier pour les deux groupes de traitement.
7. Quel test pour comparer les courbes de survie entre les deux groupes de traitement ? Justifiez votre choix.

9.A Réponse

Import de la librairie et des données

```r
# Charger la librairie survival
library(MASS)
# Afficher les premières lignes des données
head(gehan,20)
```

```
   pair time cens   treat
1     1    1    1 control
2     1   10    1    6-MP
3     2   22    1 control
4     2    7    1    6-MP
5     3    3    1 control
6     3   32    0    6-MP
7     4   12    1 control
8     4   23    1    6-MP
9     5    8    1 control
10    5   22    1    6-MP
11    6   17    1 control
12    6    6    1    6-MP
13    7    2    1 control
14    7   16    1    6-MP
15    8   11    1 control
16    8   34    0    6-MP
17    9    8    1 control
18    9   32    0    6-MP
19   10   12    1 control
20   10   25    0    6-MP
```

Colonnes :

- `pair` : numéro de la paire de patients appariés
- `time` : temps de suivi
- `cens` : indicateur de censure (0 = censure, 1 = événement)
- `treat` : groupe de traitement (control ou 6-MP)

Donc c'est bizarre !

- Moins d'observations censurées (0) que d'observations non censurées (1) !
- Donc c'est probablement l'inverse ! :
  - `cens` = 1 censure
  - `cens` = 0 événement
- `pair` : numéro de la paire de patients appariés mais on va l'ignorer ici

9.A.1 Nombre d'évènements et de censures dans chaque groupe

Fonction `table` pour obtenir un tableau croisé des événements et censures par groupe de traitement.

```r
# Résumé des événements et censures par groupe de traitement
table(gehan$treat, gehan$cens, dnn = c("Traitement", "Censure"))
```

```
          Censure
Traitement  0  1
   6-MP    12  9
   control  0 21
```

9.A.2 Nombre de sujets exposés au risque de rechute à 12 et 18 semaines dans chaque groupe

Stratégie :

1. Créer un objet `Surv` qui contiendra les temps de suivi et l'état de censure en un seul objet
   - Utiliser la fonction `Surv` du package `survival`
   - Syntaxe : `Surv(data$time, data$event)`

```r
surv_object_gehan <- Surv(time = gehan$time, event = gehan$cens)
surv_object_gehan
```

```
 [1]  1  10  22   7   3  32+ 12  23   8  22  17   6   2  16  11  34+  8  32+ 12
[20] 25+  2  11+  5  20+  4  19+ 15   6   8  17+ 23  35+  5   6  11  13   4   9+
[39]  1   6+  8  10+
```

2. Utiliser la fonction `survfit` pour calculer la survie
   - Syntaxe : `survfit(Surv_object ~ group_variable, data = dataset)`

```r
surv_fit_gehan <- survfit(surv_object_gehan ~ gehan$treat, data = gehan)
surv_fit_gehan
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                     n events median 0.95LCL 0.95UCL
gehan$treat=6-MP    21      9     23      16      NA
gehan$treat=control 21     21      8       4      12
```

3. Extraire le nombre de sujets à risque à 12 et 18 semaines en utilisant la fonction `summary`
   - Syntaxe : `summary(survfit_object, times = c(times_of_interest))`

```r
summary_gehan_12_18 <- summary(surv_fit_gehan, times = c(12, 18))
summary_gehan_12_18
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                gehan$treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12     12       5    0.753  0.0963        0.586        0.968
   18      9       2    0.627  0.1141        0.439        0.896

                gehan$treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12      6      17   0.1905  0.0857       0.0789        0.460
   18      2       2   0.0952  0.0641       0.0255        0.356
```

9.A.3 Chi^2 de Pearson approprié ?

Non, le test du Chi^2 de Pearson n'est pas approprié pour comparer les proportions d'évènements entre les deux groupes dans ce contexte de données de survie.

Justification :

- Le test du Chi^2 de Pearson compare les proportions d'événements entre des groupes à un moment fixe, sans tenir compte du temps de suivi ni de la censure.
- Dans les données de survie, les temps de suivi varient entre les individus, et certains individus peuvent être censurés (c'est-à-dire que l'événement d'intérêt n'a pas été observé avant la fin du suivi).
- Le test du Chi^2 de Pearson ne prend pas en compte ces aspects temporels et de censure, ce qui peut biaiser les résultats.
- Pour comparer les courbes de survie entre les groupes, il est plus approprié d'utiliser le test de Log-Rank, qui tient compte des temps de suivi et de la censure.

9.A.4 Survie sans rechute à 6 et 12 semaines dans les deux groupes de traitement

Utiliser la fonction `summary` pour extraire les estimations de survie à 6 et 12 semaines.

```r
summary_gehan_6_12 <- summary(surv_fit_gehan, times = c(6, 12))
summary_gehan_6_12
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                gehan$treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
   12     12       2    0.753  0.0963        0.586        0.968

                gehan$treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     12       9    0.571  0.1080       0.3945        0.828
   12      6       8    0.190  0.0857       0.0789        0.460
```

Résultats :

- À 6 semaines :
  - Groupe 6-MP : survie sans rechute = 85.7 % (IC 95 % : 72.0 % - 100 %)
  - Groupe contrôle : survie sans rechute = 57.1 % (IC 95 % : 39.5 % - 82.8 %)
- À 12 semaines :
  - Groupe 6-MP : survie sans rechute = 75.3 % (IC 95 % : 58.6 % - 96.8 %)
  - Groupe contrôle : survie sans rechute = 19.0 % (IC 95 % : 7.9 % - 46.0 %)

9.A.5 Médiane de survie sans rechute dans chaque groupe de traitement

La médiane est contenue dans l'objet crée avec `Surv` : `surv_fit_gehan`.

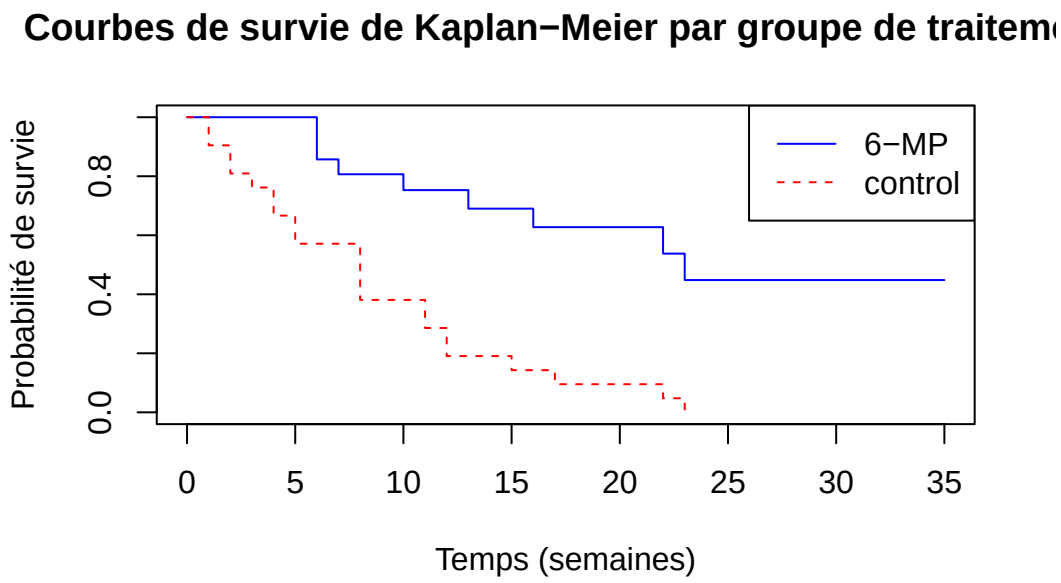
```
# Obtenir la médiane de survie pour chaque groupe
surv_fit_gehan
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                     n events median 0.95LCL 0.95UCL
gehan$treat=6-MP    21      9     23      16      NA
gehan$treat=control 21     21      8       4      12
```

9.A.6 Estimer et tracer les courbes de survie de Kaplan-Meier pour les deux groupes de traitement

Tracer les courbes de survie de Kaplan-Meier pour les deux groupes de traitement

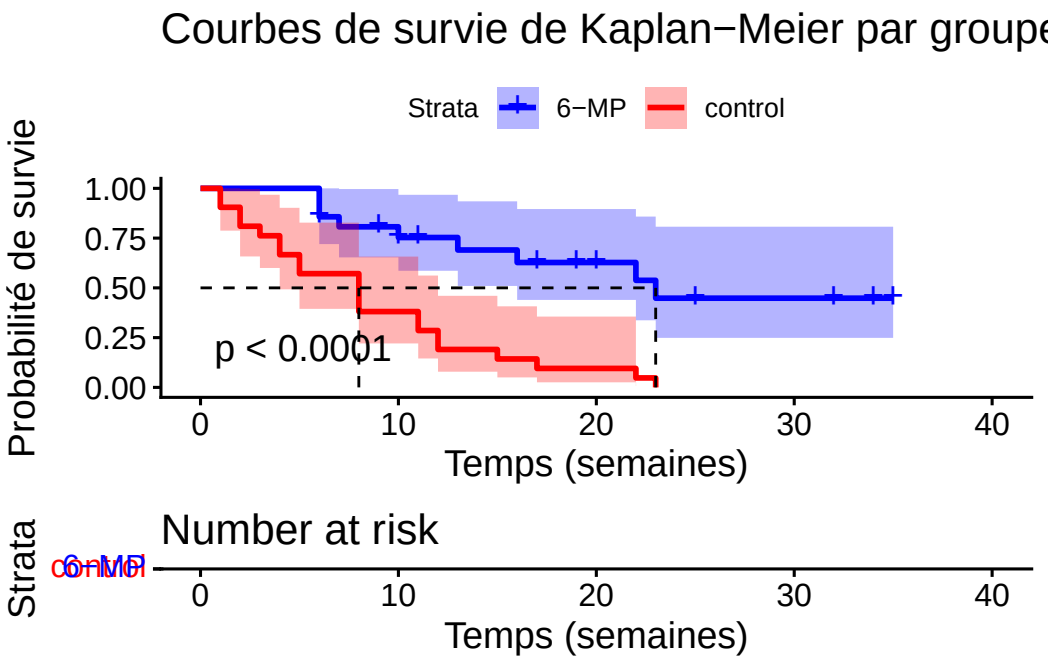
```
# Tracer les courbes de survie
plot(surv_fit_gehan, col = c("blue", "red"), lty = 1:2,
        xlab = "Temps (semaines)", ylab = "Probabilité de survie",
        main = "Courbes de survie de Kaplan-Meier par groupe de traitement")
legend("topright", legend = levels(gehan$treat), col = c("blue", "red"), lty =
 ↪  1:2)
```



Utiliser `ggsurvplot` pour une visualisation améliorée

```
# Charger la librairie survminer pour une meilleure visualisation
library(survminer)
# Tracer les courbes de survie avec ggsurvplot
ggsurvplot(surv_fit_gehan, data = gehan,
            pval = TRUE, conf.int = TRUE,
            risk.table = TRUE,
            xlab = "Temps (semaines)",
            ylab = "Probabilité de survie",
            title = "Courbes de survie de Kaplan-Meier par groupe de traitement",
            legend.labs = levels(gehan$treat),
            palette = c("blue", "red"),
            surv.median.line = "hv")
```

```
Ignoring unknown labels:
* colour : "Strata"
```



9.A.7 Quel test pour comparer les courbes de survie entre les deux groupes de traitement ? Justifiez votre choix.

Le test approprié pour comparer les courbes de survie entre les deux groupes de traitement est le **test de Log-Rank**.

Justification :

- Le test de Log-Rank est spécifiquement conçu pour comparer les courbes de survie entre deux ou plusieurs groupes dans le contexte des données de survie.
- Il prend en compte les temps de suivi variables et la censure des données, ce qui est essentiel dans les analyses de survie.
- Le test de Log-Rank évalue si les différences observées dans les courbes de survie sont statistiquement significatives en comparant le nombre d'événements observés à celui attendu dans chaque groupe à chaque instant de temps.

```r
# Effectuer le test de Log-Rank
logrank_test <- survdiff(surv_object_gehan ~ gehan$treat)
print(logrank_test)
```

```
Call:
survdiff(formula = surv_object_gehan ~ gehan$treat)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
gehan$treat=6-MP    21        9     19.3      5.46      16.8
gehan$treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

10 TP 3

Base de données `Melanoma` dans le package `MASS` : étude de survie de patients atteints de mélanome avec différentes variables explicatives.

10.A Questions

1. Description des données
2. Estimation d'une courbe de survie de Kaplan-Meier globale
3. Influence du sexe sur la survie
4. Test de Log-Rank pour tester la différence de survie en foncton du sexe, en stratifiant sur le niveau d'ulcération

10.B Réponse

10.B.1 Description des données

Importer les données et la librairie nécessaire

```
# Charger la librairie MASS
library(MASS)
# Afficher les premières lignes des données Melanoma
head(Melanoma)
```

```
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

Colonnes :

- `time` : temps de suivi en mois
- `status` : état de l'événement (1 = décès dû au mélanome, 2 = alive, 3 = décès dû à d'autres causes)
- `sex` : sexe (0 = femme, 1 = homme)
- `age` : âge au moment du diagnostic
- `year` : année du diagnostic
- `thickness` : épaisseur de la tumeur en mm
- `ulcer` : présence d'ulcération (0 = non, 1 = oui)

Truc : pour l'analyse de survie, on va recoder `status` en binaire (1 et 3 = décès toute cause, 2 = censuré)

```r
Melanoma$status2 <- ifelse(Melanoma$status == 2, 0, 1)
```

10.B.2 Estimation d'une courbe de survie de Kaplan-Meier globale

1. Créer un objet `Surv` avec les temps de suivi et l'état de l'événement

```
Surv(time = data$time, event = data$status == 1)
```

```r
surv_object_melanoma <- Surv(time = Melanoma$time, event = Melanoma$status2 == 1)
surv_object_melanoma
```

```
  [1]   10     30     35+    99    185    204    210    232    232    279    295    355
 [13]  386    426    469    493    529    621    629    659    667    718    752    779
 [25]  793    817    826    833    858    869    872    967    977    982   1041   1055
 [37] 1062   1075   1156   1228   1252   1271   1312   1427   1435   1499+ 1506   1508+
 [49] 1510+ 1512+ 1516   1525   1542+ 1548   1557+ 1560   1563+ 1584   1605+ 1621
 [61] 1627+ 1634+ 1641+ 1641+ 1648+ 1652+ 1654+ 1654+ 1667   1678+ 1685+ 1690
 [73] 1710+ 1710+ 1726   1745+ 1762+ 1779+ 1787+ 1787+ 1793+ 1804+ 1812+ 1836+
 [85] 1839+ 1839+ 1854+ 1856+ 1860   1864+ 1899+ 1914+ 1919+ 1920+ 1927+ 1933
 [97] 1942+ 1955+ 1956+ 1958+ 1963+ 1970+ 2005+ 2007+ 2011+ 2024+ 2028+ 2038+
[109] 2056+ 2059+ 2061   2062   2075+ 2085   2102+ 2103   2104+ 2108   2112+ 2150+
[121] 2156+ 2165+ 2209+ 2227+ 2227+ 2256   2264+ 2339+ 2361+ 2387+ 2388   2403+
[133] 2426+ 2426+ 2431+ 2460+ 2467   2492+ 2493+ 2521+ 2542+ 2559+ 2565   2570+
[145] 2660+ 2666+ 2676+ 2738+ 2782   2787+ 2984+ 3032+ 3040+ 3042   3067+ 3079+
[157] 3101+ 3144+ 3152+ 3154   3180+ 3182   3185+ 3199+ 3228+ 3229+ 3278+ 3297+
[169] 3328+ 3330+ 3338   3383+ 3384+ 3385+ 3388+ 3402+ 3441+ 3458   3459+ 3459+
[181] 3476+ 3523+ 3667+ 3695+ 3695+ 3776+ 3776+ 3830+ 3856+ 3872+ 3909+ 3968+
[193] 4001+ 4103+ 4119+ 4124+ 4207+ 4310+ 4390+ 4479+ 4492+ 4668+ 4688+ 4926+
[205] 5565+
```

2. Calculer la courbe de survie de Kaplan-Meier globale avec `survfit`

`survfit`permet d'ajuster un modèle de survie de Kaplan-Meier.

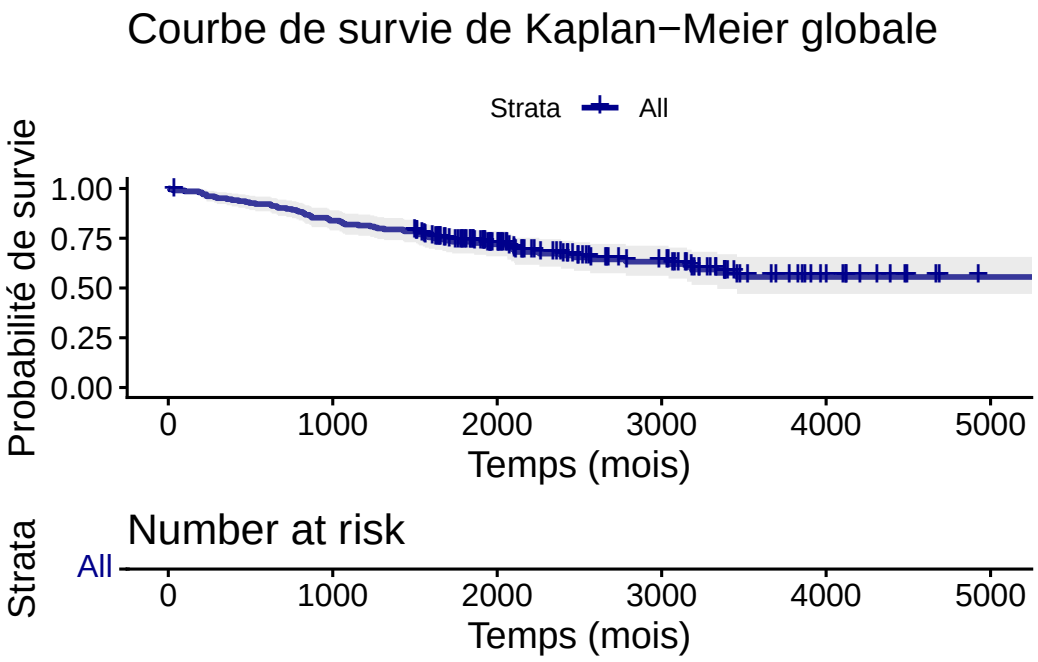
```r
surv_fit_melanoma <- survfit(surv_object_melanoma ~ 1, data = Melanoma)
surv_fit_melanoma
```

```
Call: survfit(formula = surv_object_melanoma ~ 1, data = Melanoma)

       n events median 0.95LCL 0.95UCL
[1,] 205     71     NA    3338      NA
```

3. Tracer la courbe de survie avec 'ggsurvplot' pour une meilleure visualisation

```r
# Charger la librairie survminer pour une meilleure visualisation
library(survminer)
# Tracer la courbe de survie globale
ggsurvplot(surv_fit_melanoma,
            data = Melanoma,
            conf.int = TRUE,
            risk.table = TRUE,
            xlab = "Temps (mois)",
            ylab = "Probabilité de survie",
            title = "Courbe de survie de Kaplan-Meier globale",
            palette = "darkblue"
        )
```

```
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* colour : "Strata"
```



10.B.3 Influence du sexe sur la survie

1. Calculer les courbes de survie de Kaplan-Meier par sexe

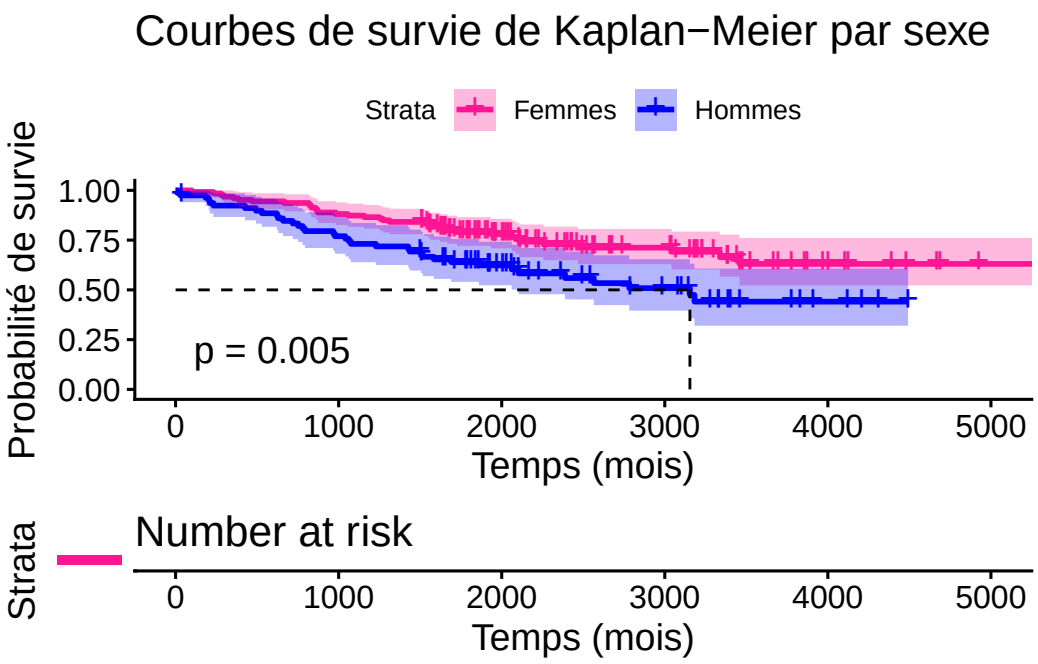
```r
surv_fit_sex <- survfit(surv_object_melanoma ~ Melanoma$sex, data = Melanoma)
surv_fit_sex
```

```
Call: survfit(formula = surv_object_melanoma ~ Melanoma$sex, data = Melanoma)

                 n events median 0.95LCL 0.95UCL
Melanoma$sex=0 126     35     NA      NA      NA
Melanoma$sex=1  79     36   3154    2103      NA
```

2. Tracer les courbes de survie par sexe

```r
# Tracer les courbes de survie par sexe
ggsurvplot(
        surv_fit_sex,
        data = Melanoma,
        pval = TRUE,
        conf.int = TRUE,
        risk.table = TRUE,
        xlab = "Temps (mois)",
        ylab = "Probabilité de survie",
        title = "Courbes de survie de Kaplan-Meier par sexe",
        legend.labs = c("Femmes", "Hommes"),
        palette = c("deeppink", "blue"),
        surv.median.line = "hv",
        risk.table.height = 0.25,
        risk.table.y.text.col = TRUE,
        risk.table.y.text = FALSE,
        censor = TRUE
        )
```

```
Ignoring unknown labels:
* colour : "Strata"
```



10.B.4 Test de Log-Rank et Modèle de Cox pour comparer les courbes de survie entre les groupes

1. Effectuer le test de Log-Rank pour comparer les courbes de survie entre les groupes

```
# Effectuer le test de Log-Rank
logrank_test_sex <- survdiff(surv_object_melanoma ~ Melanoma$sex)
print(logrank_test_sex)
```

```
Call:
survdiff(formula = surv_object_melanoma ~ Melanoma$sex)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Melanoma$sex=0 126       35     46.3      2.75       7.9
Melanoma$sex=1  79       36     24.7      5.14       7.9

 Chisq= 7.9  on 1 degrees of freedom, p= 0.005
```

2. Ajuster un modèle de Cox avec le sexe comme covariable

```
# Ajuster un modèle de Cox avec le sexe comme covariable
cox_model_sex <- coxph(surv_object_melanoma ~ Melanoma$sex)
summary(cox_model_sex)
```

```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex)
```

```
  n= 205, number of events= 71

               coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex 0.6559    1.9269   0.2376 2.761  0.00577 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex     1.927      0.519      1.21      3.07

Concordance= 0.585  (se = 0.031 )
Likelihood ratio test= 7.51  on 1 df,   p=0.006
Wald test            = 7.62  on 1 df,   p=0.006
Score (logrank) test = 7.9  on 1 df,   p=0.005
```

10.B.5 Si on regarde en ajustant sur l'épaisseur de la tumeur

```r
# Ajuster un modèle de Cox avec le sexe et l'épaisseur de la tumeur comme
↪ covariables
cox_model_sex_thickness <- coxph(surv_object_melanoma ~ Melanoma$sex +
↪ Melanoma$thickness)
summary(cox_model_sex_thickness)
```

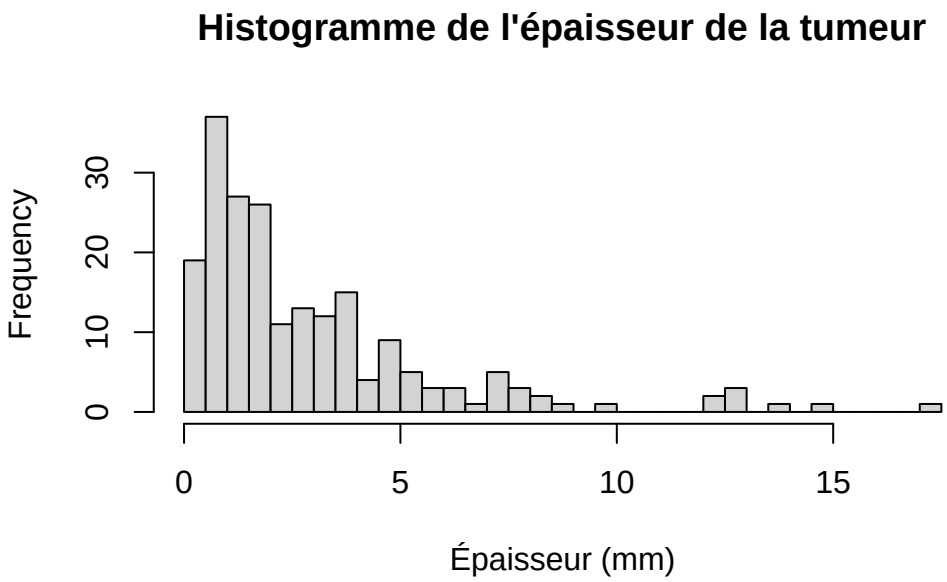
```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness)

  n= 205, number of events= 71

                      coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex       0.56284   1.75565  0.23787 2.366    0.018 *
Melanoma$thickness 0.14883   1.16048  0.03011 4.942 7.72e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex           1.756     0.5696     1.101     2.798
Melanoma$thickness     1.160     0.8617     1.094     1.231

Concordance= 0.696  (se = 0.034 )
Likelihood ratio test= 26.11  on 2 df,   p=2e-06
Wald test            = 30.98  on 2 df,   p=2e-07
Score (logrank) test = 34.19  on 2 df,   p=4e-08
```

Résultats :

- Le coefficient pour `Melanoma$sex` est 0.56284, ce qui signifie que les hommes ont un risque instantané de décès dû au mélanome environ 75.6 % plus élevé que les femmes, après ajustement pour l'épaisseur de la tumeur.
  - HR = exp(0.56284) = 1.75565
- Le coefficient pour `Melanoma$thickness` est 0.14883, indiquant que pour chaque augmentation de 1 mm de l'épaisseur de la tumeur, le risque instantané de décès dû au mélanome augmente d'environ 16.0 %.
  - HR = exp(0.14883) = 1.16048 en plus par mm (variable quantitative donc HR par unité)

Mais le truc c'est que l'épaisseur de la tumeur est très variable !!

```
hist(Melanoma$thickness, breaks = 30, main = "Histogramme de l'épaisseur de la
↪  tumeur", xlab = "Épaisseur (mm)")
```



Mais pour les dermatos : ce qui compte c'est si l'épaisseur est > ou < 1 mm (critère pronostique important)

Donc on peut créer une variable binaire `thickness_bin` :

```
Melanoma$thickness_bin <- ifelse(Melanoma$thickness > 1, 1, 0)
head(Melanoma, 3)
```

```
  time status sex age year thickness ulcer status2 thickness_bin
1   10      3   1  76 1972      6.76     1       1             1
2   30      3   1  56 1968      0.65     0       1             0
3   35      2   1  41 1977      1.34     0       0             1
```

```
table(Melanoma$thickness_bin)
```

```
  0   1
 56 149
```

Puis ajuster le modèle de Cox avec cette variable binaire :

```r
# Ajuster un modèle de Cox avec le sexe et l'épaisseur binaire de la tumeur comme
↪  covariables
cox_model_sex_thickness_bin <- coxph(surv_object_melanoma ~ Melanoma$sex +
↪  Melanoma$thickness_bin)
summary(cox_model_sex_thickness_bin)
```

```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness_bin)

  n= 205, number of events= 71

                         coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex           0.6813    1.9764   0.2379 2.864  0.00419 **
Melanoma$thickness_bin 1.1275    3.0879   0.3571 3.157  0.00159 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex               1.976     0.5060     1.240     3.151
Melanoma$thickness_bin     3.088     0.3238     1.533     6.218

Concordance= 0.651  (se = 0.031 )
Likelihood ratio test= 20.51  on 2 df,   p=4e-05
Wald test            = 17.75  on 2 df,   p=1e-04
Score (logrank) test = 19.02  on 2 df,   p=7e-05
```

Résultats :

- Le coefficient pour `Melanoma$sex` est 0.6813 et HR = exp(0.6813) = 1.9764 donc risque instantané de décès dû au mélanome environ 97.6 % plus élevé pour les hommes par rapport aux femmes.
- Le coefficient pour `Melanoma$thickness_bin` est 1.1275 et HR = exp(1.1275) = 3.0879 donc risque instantané de décès dû au mélanome environ 208.8 % plus élevé pour les patients avec une épaisseur de tumeur > 1 mm par rapport à ceux avec une épaisseur ≤ 1 mm.

Pour savoir quel est le meilleur modèle entre `thickness` en quantitatif et binaire, on va comparer par rapport au modèle avec `sex` seul

1. On redéfinit les modèles

```r
surv_object_melanoma <- Surv(time = Melanoma$time, event = Melanoma$status2 == 1)
cox_model_sex <- coxph(surv_object_melanoma ~ Melanoma$sex)
cox_model_sex_thickness <- coxph(surv_object_melanoma ~ Melanoma$sex +
↪  Melanoma$thickness)
cox_model_sex_thickness_bin <- coxph(surv_object_melanoma ~ Melanoma$sex +
↪  Melanoma$thickness_bin)
```

2. Comparaison avec le test du rapport de vraisemblance

```r
# Comparaison des modèles avec le test du rapport de vraisemblance
anova(cox_model_sex, cox_model_sex_thickness)
```

```
Analysis of Deviance Table
 Cox model: response is  surv_object_melanoma
 Model 1: ~ Melanoma$sex
 Model 2: ~ Melanoma$sex + Melanoma$thickness
   loglik  Chisq Df Pr(>|Chi|)
1 -346.73
2 -337.43 18.598  1  1.614e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(cox_model_sex, cox_model_sex_thickness_bin)
```

```
Analysis of Deviance Table
 Cox model: response is  surv_object_melanoma
 Model 1: ~ Melanoma$sex
 Model 2: ~ Melanoma$sex + Melanoma$thickness_bin
   loglik  Chisq Df Pr(>|Chi|)
1 -346.73
2 -340.23 12.997  1   0.000312 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On a l'impression qu'en mettant la variable en QUANTITATIF , c'est mieux !

- Chi^2 plus grand (18.598 > 12.997)
- p-value plus petite

Mais en vrai : ce sont des considérations uniquement statistiques !

Si la variable qui compte c'est si l'épaisseur est > ou < 1 mm, on prendra le modèle avec la variable binaire

10.B.6 Interaction entre sexe et ulcération

Modèle de Cox avec interaction entre sexe et ulcération

```r
# Ajuster un modèle de Cox avec interaction entre sexe et ulcération
surv_cox_interaction <- Surv(Melanoma$time, Melanoma$status2)
cox_model_interaction <- coxph(surv_cox_interaction ~ Melanoma$sex +
↪  strata(Melanoma$ulcer))
summary(cox_model_interaction)
```

```
Call:
coxph(formula = surv_cox_interaction ~ Melanoma$sex + strata(Melanoma$ulcer))

  n= 205, number of events= 71

               coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex 0.4995    1.6479   0.2396 2.084   0.0371 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex     1.648     0.6068      1.03     2.636

Concordance= 0.568  (se = 0.031 )
Likelihood ratio test= 4.31  on 1 df,   p=0.04
Wald test            = 4.34  on 1 df,   p=0.04
Score (logrank) test = 4.43  on 1 df,   p=0.04
```

Log rank et courbe de Kaplan-Meier stratifiée sur le niveau d'ulcération

1. Effectuer le test de Log-Rank stratifié sur le niveau d'ulcération

Syntaxe : `survdiff(Surv_object ~ group_variable + strata(stratification_variable))`

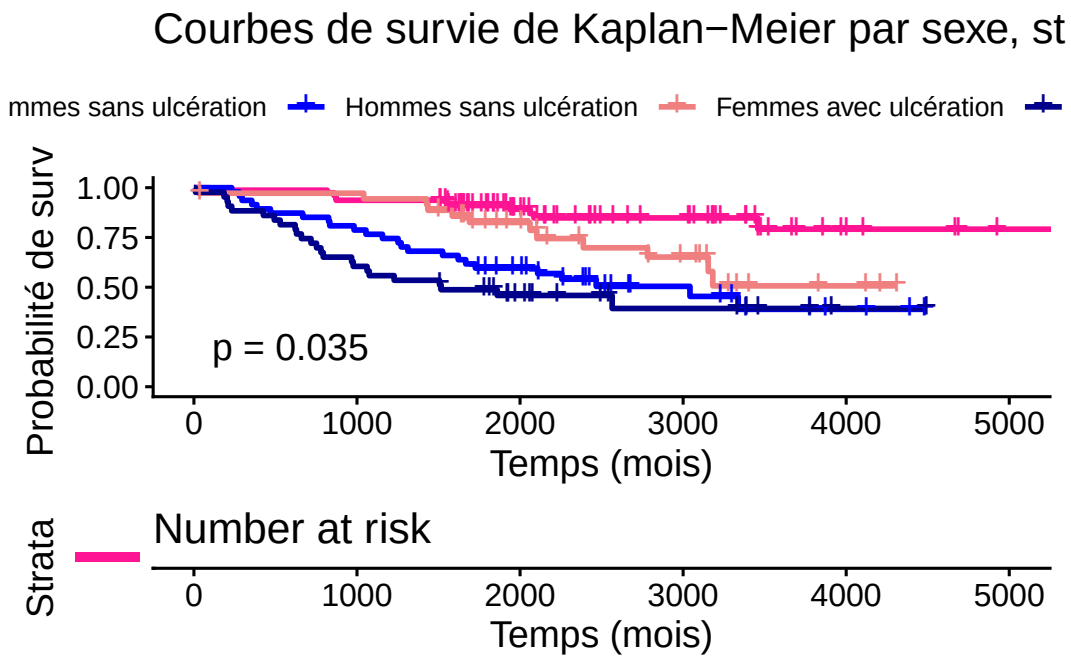
```r
# Effectuer le test de Log-Rank stratifié sur le niveau d'ulcération
logrank_test_stratified <- survdiff(surv_object_melanoma ~ Melanoma$sex +
↪  strata(Melanoma$ulcer))
print(logrank_test_stratified)
```

```
Call:
survdiff(formula = surv_object_melanoma ~ Melanoma$sex + strata(Melanoma$ulcer))

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Melanoma$sex=0 126       35     43.6      1.68      4.43
Melanoma$sex=1  79       36     27.4      2.67      4.43

 Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

2. Tracer les courbes de survie de Kaplan-Meier par sexe, stratifiées sur le niveau d'ulcération

```r
# Tracer les courbes de survie par sexe, stratifiées sur le niveau d'ulcération
ggsurvplot(
        survfit(surv_object_melanoma ~ Melanoma$sex + strata(Melanoma$ulcer), data
         ↪  = Melanoma),
        data = Melanoma,
        pval = TRUE,
        conf.int = FALSE,
        risk.table = TRUE,
        xlab = "Temps (mois)",
        ylab = "Probabilité de survie",
        title = "Courbes de survie de Kaplan-Meier par sexe, stratifiées sur le
         ↪  niveau d'ulcération",
        legend.labs = c("Femmes sans ulcération", "Hommes sans ulcération",
         ↪  "Femmes avec ulcération", "Hommes avec ulcération"),
        palette = c("deeppink", "blue", "lightcoral", "darkblue"),
        risk.table.y.text.col = TRUE,
        risk.table.y.text = FALSE,
        censor = TRUE
        )
```

```
Ignoring unknown labels:
* colour : "Strata"
```



9GLM Poisson

1 Plan du cours

1. Introduction : données de comptage, données longtiduinales, données en cluster
2. Modèles de Poisson pour les données de comptage
    - Loi de Poisson
    - Interprétation des coefficients
    - Cas des données sur-dispersées (variance > moyenne)

2 Introduction : données de comptage, données longtiduinales, données en cluster

2.A Données de comptage

Les données de comptage sont des données qui représentent le nombre d'occurrences d'un événement dans un intervalle de temps ou d'espace donné.

- **Variable aléatoire discrète**
  - Valeur dans les entiers naturels positifs (0, 1, 2, …)
  - Exemples : nombre de visites à l'hôpital, nombre d'accidents de la route, nombre de naissances, etc.
- **Ordre de grandeur des comptages**
  - Souvent, les comptages sont de taille modérée (par exemple, de 0 à quelques dizaines).
  - Cependant, ils peuvent aussi être très élevés dans certains contextes (par exemple, le nombre de visiteurs sur un site web).
- **Caractéristiques des données de comptage**
  - Beaucoup de zéros (événements rares)
  - **Pas d'unité !** : les comptages sont des nombres absolus
  - Distribution asymétrique : la distribution est souvent biaisée à droite (peu de grandes valeurs)
  - Non normale : les données de comptage ne suivent pas une distribution normale
  - Relation moyenne-variance : dans une distribution de Poisson, **la variance est souvent égale voire supérieure à la moyenne (sur-dispersion)**
  - On peut comparer les deux car pas d'unité
- **Lien avec la distribution de Poisson**

  - La distribution de Poisson est souvent utilisée pour modéliser les données de comptage.
  - Elle est caractérisée par un paramètre λ qui représente à la fois la moyenne et la variance des comptages.
  - λ correspond au taux moyen d'occurrence de l'événement par unité de temps ou d'espace.
- **Notion d'exposition**
  - Dans certains cas, les données de comptage sont associées à une notion d'exposition, c'est-à-dire la durée ou la surface pendant laquelle les événements peuvent se produire.
  - Par exemple, le nombre d'accidents de la route peut être rapporté au nombre de kilomètres parcourus.
  - L'exposition est souvent utilisée comme un offset dans les modèles de régression pour ajuster les comptages en fonction de la durée ou de la surface d'observation.

Loi de Poisson

- Paramètre $\lambda > 0$ (ne peut pas être négatif)
- Espérance : $\mathbb{E}(X) = \lambda$
- Variance : Var$(X) = \lambda$

= distribution de probabilité discrète

- décrivant le nombre d'événements se produisant dans un intervalle de temps ou d'espace fixe,
- lorsque ces événements se produisent avec une moyenne constante et indépendamment du temps écoulé depuis le dernier événement.

$P(X = k) = \frac{\lambda^k}{k!} \times e^{-\lambda}$, pour $k = 0, 1, 2, ...$

2.A.1 Exemple de données de comptage

10 000 tirages d'une variable aléatoire

- avec loi de Poisson de paramètre $\lambda = 1$ (donc moyenne et variance égales à 1)
- avec loi de Poisson de paramètre $\lambda = 5$ (donc moyenne et variance égales à 5)
- avec loi de Poisson de paramètre $\lambda = 20$ (donc moyenne et variance égales à 20)

	Pandemic year 1		Pandemic year 2	
	Cohen's d effect size (95% CI)	p value	Cohen's d effect size (95% CI)	p value
Whole cohort				
Executive function				
Verbal Reasoning	0·15 (0·12 to 0·17)	<0·0001	0·02 (0·00 to 0·04)	0·12
Working memory				
Paired Associate Learning	0·77 (0·30 to 0·37)	<0·0001	0·74 (0·72 to 0·76)	<0·0001
Self-Ordered Search	0·15 (0·13 to 0·18)	<0·0001	0·15 (0·13 to 0·18)	<0·0001
Digit Span	0·19 (0·17 to 0·21)	<0·0001	0·14 (0·12 to 0·16)	<0·0001
Composite	0·51 (0·49 to 0·53)	<0·0001	0·47 (0·44 to 0·49)	<0·0001

Plus le paramètre λ est grand, plus la distribution ressemble à une distribution normale.
Pour $\lambda = 1$ ou 5, la distribution est asymétrique et concentrée sur les petites valeurs.
Donc cette loi a moins d'intérêt pour des comptages élevés.

2.A.2 Approximation binomiale / Poisson

- Loi binomiale $\mathfrak{B}(n, p)$ = nombre de succès dans n essais indépendants, avec probabilité p de succès à chaque essai.
  - Exemple : nombre d'appels au SAMU dans une journée parmi 10 000 habitants, avec une probabilité p d'appel faible.

Démarche de modélisation :

- La ville contient un nombre n d'habitants (très grand)
- Chaque habitant a une probabilité p (très petite) d'appeler le SAMU dans la journée
- Les appels surviennent de manière indépendante entre les habitants
- Le nombre total d'appels X suit une loi binomiale $B(n, p)$, d'espérance $\lambda = n \times p$

Cognitive decline in older adults in the UK during and after the COVID-19 pandemic: a longitudinal analysis of PROTECT study data

Anne Corbett, Gareth Williams, Byron Creese, Adam Hampshire, Vincent Hayman, Abbie Palmer, Akos Filakovzsky, Kathryn Mills, Jeffrey Cummings, Dag Aarsland, Zunera Khan, Clive Ballard

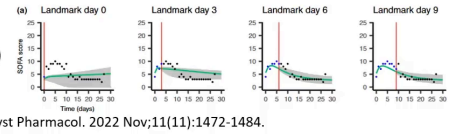
Summary

Background Although the long-term health effects of COVID-19 are increasingly recognised, the societal restrictions during the COVID-19 pandemic hold the potential for considerable detriment to cognitive and mental health, particularly because major dementia risk factors—such as those related to exercise and dietary habits—were affected during this period. We used longitudinal data from the PROTECT study to evaluate the effect of the pandemic on cognition in older adults in the UK.

En augmentant n et en diminuant p de manière à ce que le produit $n \times p$ reste constant, la loi binomiale $\mathfrak{B}(n, p)$ se rapproche de la loi de Poisson (de paramètre $\lambda = n \times p$).

2.A.3 Surdispersion

Modèles non linéaires mixtes (NLMM)
Ex : Evolution d'un biomarqueur



Lavalley-Morelle A et al. CPT Pharmacometrics Syst Pharmacol. 2022 Nov;11(11):1472-1484.
doi: 10.1002/psp4.12856

Exemple : données de transcriptomique Bowtie

- Chaque gène a un niveau d'expression (entier positif) sur chaque échantillon
- On représente, pour chaque gène, la variance en fonction de la moyenne
- Sous l'hypothèse d'une loi de Poisson, les points devraient tomber proche de la diagonale $y = x$

Or : ce n'est pas la cas !

Il faut donc une loi plus générale (ex. loi binomiale négative) pour modéliser ces données.

2.A.4 Excès de 0

Modèles linéaires généralisés mixtes (GLMM)
Ex : logistique mixte

Bouzid D et al. BMC Med Educ. 2022 Dec 13;22(1):861. d
doi: 10.1186/s12909-022-03919-1

Table 1 Summary of the factors influencing students' scores variability

	Station 1				Sta
	Score median- IQR (Q1-Q3) or % success	Agreement N (%)	Students' ICC (%)	Raters'ICC (%)	Sco me or
score (/100)	60 (50-70)	21 (25)	60.2	23.0	60 (
item 1	65	76 (89)	97	3	61
item 2	93	75 (88)	82	18	64
item 3	94	74 (87)	*	*	99
item 4	95	76 (89)	*	*	39
item 5	33	67 (79)	31	48	38
item 6	48	80 (94)	92	0	79

Exemple : nombre d'épisode de migraine durant la dernière année

- Les nons malades n'ont pas d'épisode (0) = zéros "structurels" ;
- Les malades peuvent aussi ne pas avoir d'épisode = zéros "aléatoires", mais aussi 1, 2, …

- Il en résulte un excès de 0 par rapport à une distribution “standard”

Solution : un modèle de mélange (on modélise à part les “faux” 0 : “zero-inflated model”)

- 1ère partie du modèle estime la probabilité d’être toujours à 0 (par exemple “non malade”) ;
- 2nde partie du modèle estime le comptage conditionnel (0, 1, 2, …) chez les individus susceptibles d’avoir des événements (par exemple les malades), souvent avec une Poisson ou une binomiale négative.
- Ce type de modèle permet de mieux tenir compte de la structure des données et d’obtenir des estimations plus réalistes des effets des covariables.

3 Modèle de poisson

3.A Modèle

On observe

- une réponse Y de type comptage (0, 1, 2, …)
- des covariables explicatives X1, X2, …, Xp, dont on souhaite examiner l’association avec la moyenne de Y.

On modélise le paramètre λ de la loi de Poisson (moyenne et variance de Y) via une fonction de lien exponentielle :
$Y \sim Poisson(\lambda)$

$$\lambda = \mathbb{E}(Y|X) = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Donc λ est l’exponentielle d’un prédicteur linéaire = fonction log

Le lien assure que λ est toujours positif.

C’est différent d’une régression logistique où la moyenne est un “expit” du prédicteur linéaire (donc fonction logit inverse)

> **Note**
>
> **Modèle linéaire** : suppose que chaque variable explicative a un effet additif sur la moyenne de la variable réponse, avec une relation linéaire directe.
>
> - Linéaire = proportionnel et additif : chaque variable explicative influence la moyenne de la variable réponse de manière proportionnelle à son coefficient.
> - Additif car les effets des différentes variables explicatives s’additionnent pour déterminer la moyenne de la variable réponse.

Liens en régressions = fonction de lien = façon de transformer la variable qu’on veut expliquer en quelque chose qu’on peut modéliser linéairement.

La fonction de lien doit être adaptée à la variable

Exprimer une probabilité p (entre 0 et 1) : on la transforme entre -infini et +infini avec la fonction logit ou probit

- logit : fonctionne en faisant le logarithme du rapport entre la probabilité d'un événement et la probabilité de son complément (1 - p)
- probit : utilise la fonction normale cumulative inverse pour transformer la probabilité en une valeur sur l'axe des réels.
  - on prend une probabilité p
  - on trouve la valeur x telle que la probabilité d'obtenir une valeur ≤ à x dans une distribution normale standard soit égale à p.
  - Cela permet de modéliser des données binaires en utilisant une approche basée sur la distribution normale.
- Exprimer une moyenne μ positive (comptage) : on la transforme entre 0 et +infini avec la fonction log

! Important

Régression linéaire : moyenne est égale au prédicteur linéaire

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Régression logistique : moyenne est une fonction "expit" du prédicteur linéaire

$$\mathbb{E}(Y|X) = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Régression de Poisson : moyenne est une fonction exponentielle du prédicteur linéaire

$$\mathbb{E}(Y|X) = exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

Bonus : tableau comparatif liens GLM

3.B Interprétation des coefficients

1. Si la variable X est binaire

$\lambda_0 = \mathbb{E}(Y \| X = 0) = exp(\beta_0)$
$\lambda_1 = \mathbb{E}(Y \| X = 1) = exp(\beta_0 + \beta_1) = exp(\beta_0) \times exp(\beta_1)$
Donc le ratio des moyennes est
$\frac{\lambda_1}{\lambda_0} = \frac{\mathbb{E}(Y \| X=1)}{\mathbb{E}(Y \| X=0)} = \frac{exp(\beta_0+\beta_1)}{exp(\beta_0)} = exp(\beta_1)$
Le coefficient s'inteprète comme un **log d'un risque relatif**, c'est à dire le logarithme du ratio des moyennes entre les deux groupes.

2. Si la variable X est catégorielle
   - On choisit une catégorie de référence (par exemple la première catégorie)

- Comme das les autres modèles linéaires ou GLM
- Chaque coefficient β_k correspond au log du ratio des moyennes entre la catégorie k et la catégorie de référence.

3. Si la variable X est continue
   - exp(β_1) correspond au RR associé à une augmentation de 1 unité de X.

Dans un modèle multivariable, l'hypothèse est un effet additif** des variables explicatives sur le log(λ), donc un effet **multiplicatif** sur λ.**

Tip

Exemple
Imaginons que
$\lambda =$ moyenne du nombre de visites aux urgences par mois.
On ajuste un modèle de Poisson :

$$\text{log}(\lambda) = \beta_0 + 0.4\,(\text{fumeur}) + 0.7\,(\text{diabétique})$$

Cas 1 — Personne non fumeuse et non diabétique
Les deux indicateurs valent 0.

$$\lambda = e^{\beta_0}$$

Cas 2 — Personne fumeuse uniquement

$$\lambda = e^{\beta_0} \times e^{0.4}$$

Comme $e^{0.4} \approx 1.49$, cela correspond à **+49%** sur la moyenne du comptage.

Cas 3 — Personne fumeuse + diabétique

$$\lambda = e^{\beta_0} \times e^{0.4} \times e^{0.7}$$

Comme $e^{0.7} \approx 2.01$, cela correspond à **+101%** supplémentaires.
Ainsi, l'augmentation est :

- 1.49 (effet fumeur)
- $\times 2.01$ (effet diabète)
- = **multiplication des deux effets**

Les effets sont donc **additifs sur** $\log(\lambda)$,
mais ils deviennent **multiplicatifs sur** λ après exponentiation.
C'est pourquoi, dans une régression de Poisson,
les coefficients exponentiés e^{β} s'interprètent comme des **risques relatifs**.

3.C Tests statistiques

Comme pour les autres GLM, l'ordinateur estime, par la méthode du maximum de vraisemblance :

- les coefficients β.
- les erreurs standards associées.

On dispose donc :

- du test de Wald pour chaque coefficient (H0 : $\beta_k = 0$)
- du test du rapport de vraisemblance pour le modèle global (H0 : tous les $\beta_k = 0$) si l'on compare 2 modèles emboités.

❗ Important

Attention à la surdispersion !

- p-value du test de Wald peut être trop petite (risque de faux positifs augmenté).
- 95% IC peut être trop étroit = diminution de la couverture réelle.

3.D Exemple avec R

3.D.1 1

On crée une variable X simulant 10 000 tirages et suivant une loi de Poisson de paramètre 3.
On crée ensuite une variable $Y = 4 \times X$.
Lorsqu'on multiplie une variable par une constante :

- Sa moyenne est multipliée par cette constante
- Sa variance est multipliée par le carré de cette constante

Donc la variance de Y est 16 fois celle de X = sur-dispersion (variance = 4* la moyenne)

```
X <- rpois(10000,3) #simule 10000 tirages d'une loi de poisson de paramètre 3
Y <- 4*X
fit <- glm(Y ~ 1, family = poisson()) #Y ~ 1 veut dire qu'il n'y a pas de
 ↪  covariable explicative
summary(fit)
```

```
Call:
glm(formula = Y ~ 1, family = poisson())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 2.483205   0.002889   859.5   <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: 43375  on 9999  degrees of freedom
Residual deviance: 43375  on 9999  degrees of freedom
AIC: 83773

Number of Fisher Scoring iterations: 5
```

- Coefficient estimé : 2.48
- Erreur standard : 0.00298
- z value = rapport coefficient estimé / erreur standard = 858
- p-value < 2e-16 (p value petite car grand échantillon)

Dispersion parameter for poisson family taken to be 1 : indique que l'on a supposé une variance égale à la moyenne (pas de sur-dispersion prise en compte)

Pour trouver le paramètre λ de la loi de Poisson, on exponentie le coefficient estimé :

```r
exp(coef(fit))
```

```
(Intercept)
    11.9796
```

On trouve une moyenne de 11.96 (≈ 12), ce qui est cohérent avec la moyenne de Y (4*3 = 12).

Donc 2 messages :

1. Loi de Poisson : moyenne = variance et données peuvent être sur-dispersées
2. Même en présence de sur-dispersion, l'estimation de la moyenne par le modèle de Poisson reste correcte.

3.D.2 2

2 vecteurs :

- $Y1$ contient 100 tirages d'une loi de Poisson de paramètre 3 (moyenne théorique = 3)
- $Y2$ contient 100 tirages d'une loi de Poisson de paramètre 4 (moyenne théorique = 4)

Donc RR théorique = $\text{RR}_{\text{theorique}} = \frac{\lambda_1}{\lambda_0} = \frac{4}{3} \approx 1{,}33$.

Objectif en ajustant un modèle de Poisson sur la variable Y est d'estimer ce RR à partir des données simulées.

Variable de comptage Y définie par :

$$Y = 4 \times (Y1 + Y2)$$

Vecteur explicatif X :

- Indicateur binaire
- prenant la valeur 0 pour les 100 premières observations et la valeur 1 pour les 100 suivantes.

Puis ajuster un modèle de Poisson :

$$\log(\lambda) = \beta_0 + \beta_1 X$$

et afficher le résumé du modèle.

```r
set.seed(123)

# Deux vecteurs de Poisson
Y1 <- rpois(100,3)
Y2 <- rpois(100,4)

# Variable de comptage finale
Y <- 4 * c(Y1, Y2)

# Variable explicative binaire
X <- c(rep(0, 100), rep(1, 100))
```

Estimation du modèle de Poisson :

```r
# Modèle de Poisson
fit <- glm(Y ~ X, family = poisson())
summary(fit)
```

```
Call:
glm(formula = Y ~ X, family = poisson())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.46470    0.02916  84.522   <2e-16 ***
X            0.33258    0.03821   8.704   <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: 813.15  on 199  degrees of freedom
Residual deviance: 736.34  on 198  degrees of freedom
AIC: 1594.4

Number of Fisher Scoring iterations: 5
```

Intercept : 2,46 (correspond à la strate X=0)

Variable X : 0,33 (correspond à la différence entre les strates X=1 et X=0)

On exponentie le coefficient de X pour l'interpréter comme un risque relatif :

```r
exp(coef(fit))
```

```
(Intercept)           X
  11.760000    1.394558
```

Donc finalement, malgré la sur-dispersion, on trouve un RR de 1,39, ce qui est cohérent avec le ratio des moyennes théoriques (4/3 = 1,33).

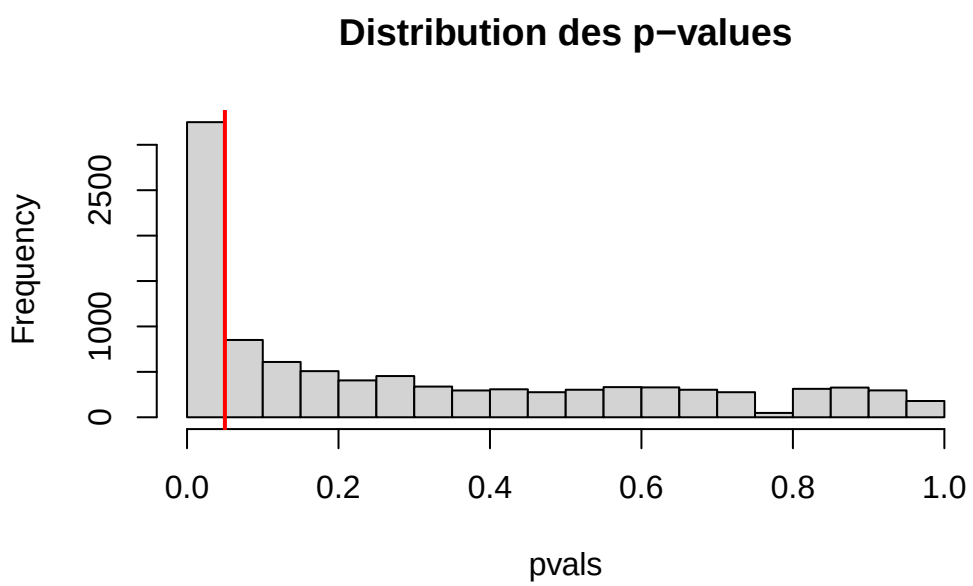
3.D.3 3

On souhaite étudier la distribution des p-values obtenues lorsqu'on répète 10 000 fois la simulation suivante :

1. On simule deux vecteurs :
   - $Y1$: 100 tirages d'une loi de Poisson de paramètre 3
   - $Y2$: 100 tirages d'une loi de Poisson de paramètre 3
2. On construit ensuite la variable de comptage :
   - $Y = 4 \times (Y1 + Y2)$
3. On définit une variable explicative binaire :
   - $X = \{0, 0, \dots, 0, 1, 1, \dots, 1\}$
   - avec 100 zéros suivis de 100 1.
4. On ajuste un modèle de Poisson :
   - log$(\lambda) = \beta_0 + \beta_1 X$
5. On extrait la p-value associée à β_1.
6. On répète cette simulation 10 000 fois et on visualise la distribution empirique des p-values.

```r
set.seed(123)

# Variable explicative (fixe)
```

```r
X <- c(rep(0, 100), rep(1, 100))

# Fonction qui renvoie la p-value du coefficient de X
test <- function() {

    # Simulations Poisson indépendantes
    Y1 <- rpois(100, lambda = 3) #Y1 suit une loi de Poisson de paramètre 3
    Y2 <- rpois(100, lambda = 3) #Y2 suit une loi de Poisson de paramètre 3 aussi
 ↪  !

    # Variable de réponse : concaténer Y1 et Y2, puis multiplier par 4 pour
     ↪  générer de la sur-dispersion
    Y <- 4 * c(Y1, Y2)

    # Modèle de Poisson
    fit <- glm(Y ~ X, family = poisson)

    # p-value du coefficient de X (ligne 2, colonne 4) = pvalue du test de Wald
    pval <- summary(fit)$coefficients[2, 4]

    return(pval)
}

# 10 000 répétitions
pvals <- replicate(10000, test())

# Histogramme des p-values
hist(pvals, breaks = 30, main = "Distribution des p-values")
abline(v = 0.05, col = "red", lwd = 2)
```



Combien de p-values sont inférieures à 0,05 en valeur absolue ?

```r
sum_05 <- sum(pvals < 0.05)
prop_05 <- sum_05 / length(pvals) * 100
cat("Il y a", sum_05, "p-values < 0.05 soit", round(prop_05, 1), "%\n")
```

```
Il y a 3249 p-values < 0.05 soit 32.5 %
```

On constate une augmentation du nombre de p-values faibles (inférieures à 0,05) par rapport à une distribution uniforme.

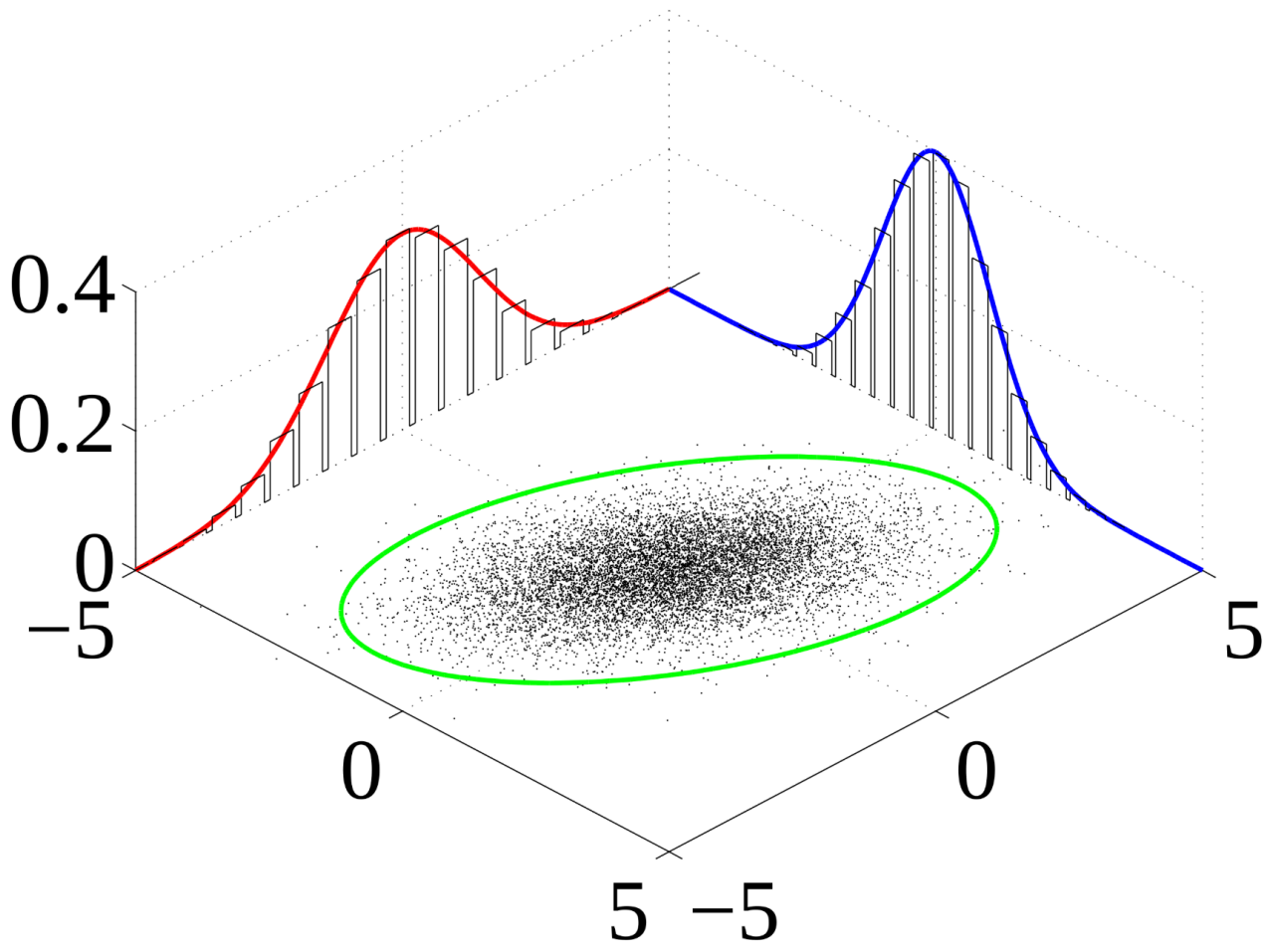
Du fait de la sur-dispersion, le test de Wald produit trop souvent des p-values faibles, ce qui augmente le risque de faux positifs.

Le problème, c'est pas l'estimation du RR (de la moyenne), mais l'inférence statistique (tests, IC) qui est biaisée par la sur-dispersion (= la variance importante de la variable).

3.E Variance "sandwich"

Dans un modèle de Poisson, on cherche à "maximiser" la log-vraisemblance :

- **vraisemblance** : se fait dans tous les modèles statistiques
  - objectif : trouver les paramètres qui rendent les données observées les plus probables
  - fonction de vraisemblance : fonction qui mesure à quel point un certain choix de paramètres rend les données observées probable. C'est une mesure de la plausibilité des données en fonction des paramètres du modèle.
- **log-vraisemblance** : logarithme de la fonction de vraisemblance. permet de rendre plus simple à manipuler (produit devient somme)
- **maximum de vraisemblance** : méthode d'estimation des paramètres du modèle en choisissant les valeurs qui maximisent la log-vraisemblance.
- **courbure de la log-vraisemblance** :
  - plus la courbe est pointue autour de son maximum, plus l'estimation est précise (erreur standard faible)
  - plus la courbe est plate autour de son maximum, moins l'estimation est précise (erreur standard élevée)
  - la matrice de variance/covariance des paramètres est liée à la courbure de la log-vraisemblance.



Donc la matrice de variance/covariance des paramètres peut être estimé :

1. Directement par l'inverse de la courbure de la log-vraisemblance au voisinage du maximum de vraisemblance (estimation classique)
    - Mais biaisé si sur-dispersion ++
2. Avec l'estimateur "sandwich" (robuste aux violations des hypothèses du modèle, notamment la sur-dispersion)
    - Dans lequel la matrice de la courbure est multipliée gauche et à droite par une autre matrice
    - Obtient la variance "sandwich"
    - Méthode très robuste à une mauvaise "spécification" du modèle

NB : "spécification" = fait de choisir un modèle statistique particulier pour représenter les données, en supposant que ce modèle capture correctement la structure sous-jacente des données.

3.E.1 Exemple avec R

1. On génère

- $Y1 \sim$ Pois(3)

- $Y2 \sim$ Pois(3)

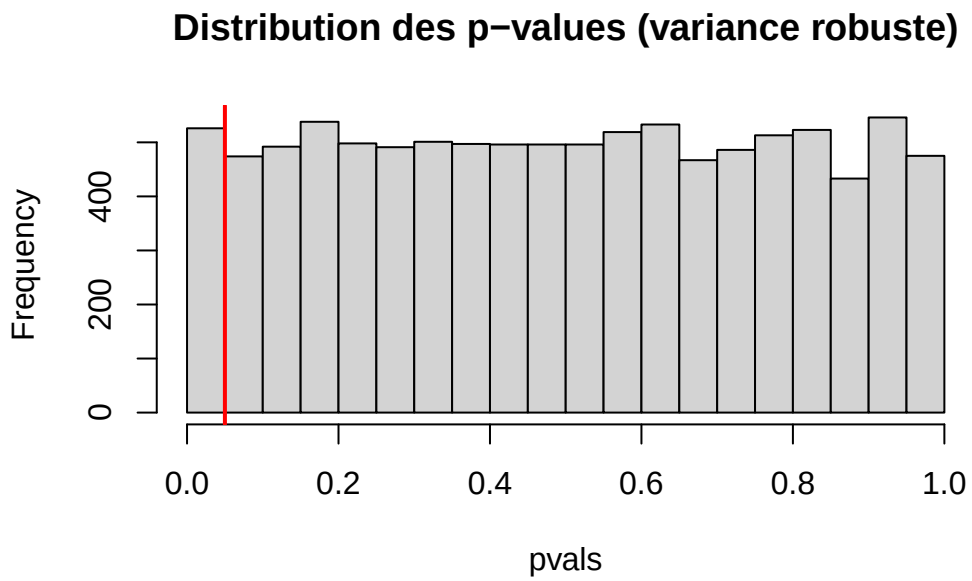
2. On définit :
    - $Y = 4 \times (Y1 + Y2)$
3. On ajuste un modèle de Poisson :
    - log$(\lambda) = \beta_0 + \beta_1 X$
4. Mais cette fois, on utilise un estimateur de variance robuste (sandwich)
    - donc la p-value provient de : `coeftest(fit, vcov. = sandwich)`
5. On extrait la p-value du coefficient de X.
6. On répète 10 000 fois et on trace l'histogramme des p-values.

```r
set.seed(123)

library(sandwich)
library(lmtest)
```

```
Loading required package: zoo


Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

```r
# Variable explicative fixe
X <- c(rep(0, 100), rep(1, 100))

# Fonction de simulation
test <- function() {

    Y1 <- rpois(100, 3)
    Y2 <- rpois(100, 3)
    Y  <- 4 * c(Y1, Y2)

    fit <- glm(Y ~ X, family = poisson)

    # p-value robuste (sandwich)
    pv <- coeftest(fit, vcov. = sandwich)[2, 4] #P-value du test de Wald avec
 ↪  variance robuste

    return(pv)
}
```

```r
# 10 000 répétitions
pvals <- replicate(10000, test())

# Histogramme des p-values
hist(pvals, breaks = 30, main = "Distribution des p-values (variance robuste)")
abline(v = 0.05, col = "red", lwd = 2)
```



Combien de p-values sont inférieures à 0,05 en valeur absolue ?

```r
sum_05 <- sum(pvals < 0.05)
prop_05 <- sum_05 / length(pvals) * 100
cat("Il y a", sum_05, "p-values < 0.05 soit", round(prop_05, 1), "%\n")
```

```
Il y a 526 p-values < 0.05 soit 5.3 %
```

L'estimation robuste de la variance permet de corriger le problème de sur-dispersion

3.F Traitement de la surdispersion

1. Correction des variances des estimateurs (estimateur robuste de la variance, bootstrap)
    1. **Estimateur robuste de la variance (sandwich)**
        - Simple à mettre en œuvre
        - Permet d'obtenir des tests et IC corrects
        - Ne modifie pas l'estimation des coefficients R
    2. **Bootstrap**
        - Permet d'obtenir des intervalles de confiance et des tests basés sur la distribution empirique des estimateurs
        - Plus coûteux en calcul
        - Ne modifie pas l'estimation des coefficients R

2. Utilisation d'une autre loi pour modéliser les données
    1. **Quasi-Poisson** : (`family = quasipoisson()` dans R) - Modifie l'estimation de la variance en ajoutant un paramètre de dispersion estimé à partir des données - Permet d'obtenir des tests et IC corrects
    2. **Binomiale négative avec** θ fixé : (`family = negative.binomial(theta = 1)` dans R, avec `theta` paramètre de dispersion fixé)
        - Modifie à la fois l'estimation des coefficients et de la variance
        - Permet d'obtenir des tests et IC corrects
    3. **Binomiale négative avec** θ estimé : (`family = negative.binomial()` dans R, avec `theta` paramètre de dispersion estimé à partir des données) ou fonction `glm.nb()` du package `MASS`

    - Modifie à la fois l'estimation des coefficients et de la variance
    - Permet d'obtenir des tests et IC corrects

3.G Traitement des excès de 0

- Modèles « hurdle » et « zero-inflated » : mélange entre
  - Un modèle logistique pour les zéros
  - Et un modèle de compte pour le reste
- Différences (ténues) dans la formulation mathématique
  - Fonction `zeroinfl()` ou `hurdle()` du package `pscl`
  - Les covariables peuvent être différentes pour les deux modèles (problème de sélection)

Modèle `hurdle` :

- Modélise la probabilité d'avoir un comptage > 0 avec un modèle logistique
- Puis modélise la distribution des comptages conditionnels à être > 0 avec un modèle de Poisson ou binomiale négative tronquée

Modèle `zero-inflated` :

- Modélise la probabilité d'être dans le groupe des zéros structurels avec un modèle logistique
- Puis modélise la distribution des comptages avec un modèle de Poisson ou binomiale négative pour les individus susceptibles d'avoir des événements (y compris les zéros aléatoires)

4 Résumé

4.A Données de comptage et histoire binomiale Poisson

Les données de comptage correspondent au **nombre d'occurrences** d'un événement dans un intervalle (temps, espace, patients, etc.).

Exemples : nombre d'appels au SAMU par jour, nombre de visites aux urgences par mois, nombre d'épisodes de migraine sur un an.

On peut raconter l'histoire suivante :

- Il y a un nombre très élevé d'unités (habitants, patients, jours à risque, etc.), noté n.
- Chaque unité a une **très petite probabilité** p de présenter l'événement sur la période.
- Les événements entre unités sont supposés indépendants.

Le nombre total d'événements X suit alors **une loi binomiale** $\mathcal{B}(n, p)$, d'espérance

$$\mathbb{E}(X) = np.$$

Quand : - n est **très grand**, - p est **très petit**, - et le produit np reste **constant** (noté λ),

alors la binomiale se **rapproche d'une loi de Poisson** de paramètre λ :

- $X \sim$ Poisson(λ),
- $\mathbb{E}(X) = \lambda$,
- Var$(X) = \lambda$.

C'est ce qui justifie l'usage de la loi de Poisson pour de nombreux phénomènes de comptage.

4.B Modèle de Poisson et interprétation des coefficients

Dans un **GLM de Poisson**, on modélise la moyenne conditionnelle

$$\lambda_i = \mathbb{E}(Y_i \mid X_i)$$

d'un comptage Y_i par un lien logarithme :

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}.$$

Ainsi :

- $\lambda_i = \exp(\beta_0 + \cdots + \beta_p X_{ip}) > 0$ (moyenne toujours positive),
- le modèle est **linéaire** sur $\log(\lambda)$.

Interprétation des coefficients :

- Si X est **binaire** (0 / 1), alors
  - $\mathbb{E}(Y \mid X = 0) = \lambda_0$,
  - $\mathbb{E}(Y \mid X = 1) = \lambda_1$,
  - et
    - $\frac{\lambda_1}{\lambda_0} = \exp(\beta_1)$.
    - $\exp(\beta_1)$ est donc un **rapport de moyennes** (souvent interprété comme un **risque relatif**).

- Si X est **catégorielle**, chaque coefficient correspond au **log-rapport de moyenne** par rapport à une catégorie de référence (comme en régression linéaire ou logistique).
- Si X est **continue**, $\exp(\beta_1)$ est le **facteur multiplicatif** sur la moyenne associé à une augmentation d'une unité de X.

En multivariable :

- les effets sont **additifs** sur $\log(\lambda)$,
- donc **multiplicatifs** sur λ après exponentiation,
- les $\exp(\beta)$ s'interprètent comme des **risques relatifs / rapports de moyennes ajustés**.

4.C Surdispersion : définition, conséquences et illustration R

Le modèle de Poisson impose

$$\text{Var}(Y \mid X) = \mathbb{E}(Y \mid X) = \lambda.$$

En pratique, on observe fréquemment :

- une **variance empirique** nettement **supérieure** à la moyenne,
- c'est la **surdispersion** :
  - $\text{Var}(Y) > \mathbb{E}(Y)$.

Dans le code, plusieurs exemples montrent ce phénomène :

- Création d'une variable $X \sim \text{Poisson}(3)$,
- puis définition de $Y = 4X$.

Alors :

- $\mathbb{E}(Y) = 4\,\mathbb{E}(X) = 4 \times 3 = 12$,
- $\text{Var}(Y) = 4^2\,\text{Var}(X) = 16 \times 3 = 48$,

ce qui donne une variance **4 fois plus grande** que la moyenne.

Même constat dans la simulation plus complexe :

- $Y_1 \sim \text{Poisson}(3)$, $Y_2 \sim \text{Poisson}(3)$,
- $Y = 4(Y_1 + Y_2)$.

Là encore, la variance est **très supérieure** à la moyenne, donc il y a surdispersion.

Effets sur l'inférence :

- Les **coefficients** $\hat{\beta}$ (et donc les rapports de moyennes) restent souvent **correctement estimés**.
- En revanche, les **erreurs standards** calculées sous le modèle de Poisson supposent à tort $\text{Var}(Y) = \mathbb{E}(Y)$.
- En présence de surdispersion :
- les erreurs standards sont **trop petites**,

- les tests de Wald donnent **trop de p-values petites**,
- les intervalles de confiance sont **trop étroits**,
- l'**erreur de type I** (faux positifs) est **inflée**.

C'est ce que montrent les simulations R :

- on répète 10 000 fois la génération de données surdispersées,
- on ajuste un modèle de Poisson standard,
- on récupère la p-value du coefficient de X à chaque fois,
- l'histogramme des p-values montre **un excès de valeurs < 0,05**,
- la proportion de p-values < 0,05 dépasse largement les 5 % attendus sous H_0.

Conclusion :
le problème n'est pas l'estimation des moyennes / RR, mais la validité des tests et des IC si l'on ne corrige pas la surdispersion.

4.D Variance robuste (« variance sandwich »)

L'idée générale est la suivante :

- On garde le **modèle de Poisson** pour estimer les $\widehat{\beta}$.
- On change **la façon d'estimer la variance de** $\widehat{\beta}$.
- On remplace la variance « classique » (basée sur la courbure de la log-vraisemblance sous Poisson) par une **variance robuste**, appelée **variance sandwich**.

Cela a pour effet :

- de **ne pas modifier** les estimations $\widehat{\beta}$,
- mais d'**augmenter** les erreurs standards lorsque la surdispersion est présente,
- donc de **corriger** les p-values et les intervalles de confiance.

Dans le code :

- 1 ajustement d'abord d'un `glm(..., family = poisson)`,
- puis

```r
library(sandwich)
library(lmtest)
coeftest(fit, vcov. = sandwich)
```

```
z test of coefficients:

            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 2.464704   0.054807 44.9704 < 2.2e-16 ***
X           0.332577   0.070939  4.6882 2.756e-06 ***
```

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

- `sandwich` calcule la matrice de variance robuste,
- `coeftest()` recalcule les tests en utilisant cette variance.

En répétant la simulation 10 000 fois avec la variance robuste :

- l'histogramme des p-values devient beaucoup plus proche de l'uniforme,
- la proportion de p-values < 0,05 redevient ≈ 5 %,
- le contrôle de l'erreur de type I est rétabli.

Message clé :

- En présence de surdispersion, on doit obligatoirement l'adresser,
- La variance sandwich est une solution simple et robuste pour obtenir des tests valides sans changer les coefficients.

□

4.E Excès de zéros : modèles Hurdle et Zero-Inflated

En plus de la surdispersion, les données de questionnaire ou de pratique clinique présentent souvent un excès de zéros :

- certains individus sont structurellement à 0 (non concernés) : zéros « structurels »,
- d'autres pourraient avoir des événements mais en ont 0 sur la période : zéros « aléatoires ».

Une Poisson simple ne peut pas produire autant de zéros, même en étant surdispersée. On utilise alors des modèles à deux composantes.

4.E.1 Modèle Hurdle

Principe : 1. Partie 1 : franchir la marche (hurdle)

```
-   On modélise la probabilité d'avoir un comptage strictement positif :

-   $\mathbb{P}(Y > 0 \mid X)$.

-   Typiquement avec une régression logistique.
```

2. Partie 2 : distribution conditionnelle de Y sachant $Y > 0$
   - On travaille uniquement sur les observations avec $Y > 0$.
   - On ajuste une Poisson tronquée en 0 ou une binomiale négative tronquée.

Les zéros sont donc complètement sortis de la partie « compte » et expliqués à part.

En R, on peut utiliser la fonction `hurdle()` du package `{pscl}`.

4.E.2 Modèle Zero-Inflated

Autre histoire possible :

- une proportion des individus est dans un état « toujours 0 » (non exposés, non malades, etc.),
- le reste suit un modèle de comptage classique (Poisson ou binomiale négative), qui peut lui-même produire des zéros.

On a alors : 1. Un modèle (souvent logistique) pour la probabilité d'être un zéro structurel.

2. Un modèle de comptage pour les individus susceptibles d'avoir des événements (incluant des zéros « aléatoires »).

Ce qu'on appelle :

- des « vrais zéros » de la loi de Poisson,
- et des « faux zéros » venant de personnes non concernées.

En R, cela se fait par exemple avec `zeroinfl()` du package `{pscl}`.

Difficultés méthodologiques

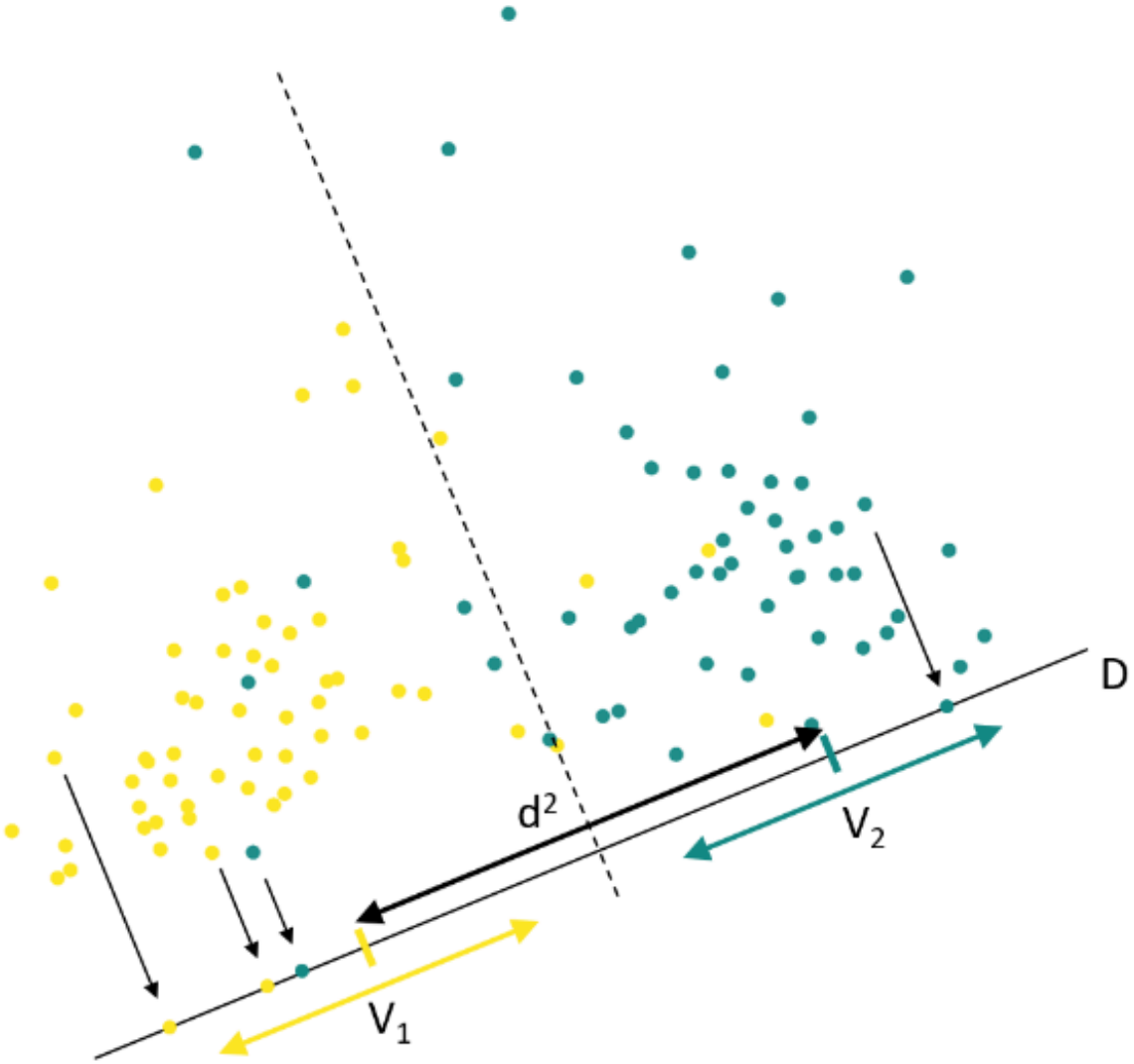
Ces modèles à deux parties posent des questions de spécification :

- Quelle covariable mettre dans la partie « zéro »,
- quelle covariable mettre dans la partie « compte »,
- lesquelles mettre dans les deux,
- avec peu de recommandations standardisées.

4.F En pratique :

- il existe toujours une possibilité de critique : « pourquoi cette covariable est-elle dans la partie zéro et pas dans l'autre ? » ;
- d'où la nécessité de bien réfléchir à la signification des zéros et de justifier le choix des covariables dans chaque sous-modèle.

5 Application



= Association entre l'utilisation d'AOD avec et sans médicaments concomitants et le risque de saignement majeur dans la fibrillation auriculaire non valvulaire

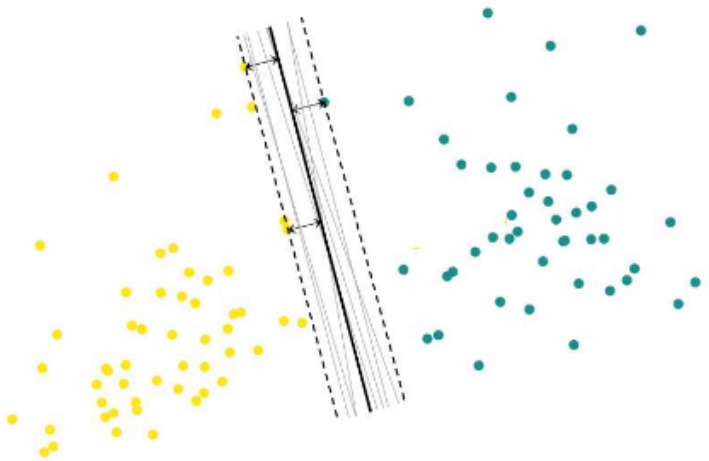


Fig. 3.4 — *Deux groupes de points sont parfaitement séparables par une infinité de droites discriminantes. Celle proposée par le SVM maximise l'écart minimal avec les points des deux groupes. Cette solution semble naturelle, elle est par ailleurs susceptible d'avoir de meilleures performances sur un échantillon de validation.*

Objectif : Estimer le risque relatif (RR) de saignement majeur associé à l'utilisation d'AOD **avec et sans médicaments concomitants**, en ajustant pour les facteurs de confusion potentiels.

Design :

- Étude de cohorte rétrospective
- Patients tous exposés à des AOD, ± médicaments concomitants : atorvastatine, amiodarone, digoxine, inhibiteurs calciques, fluconazole…
- Mesure de l'incidence des évènements (saignements majeurs) pendant la période d'exposition

Analyse :

- On regarde l'incidence d'évènements

Modèles possibles :

- **Modèle logistique et OR** : évènement Y/N à un temps donné
- **Modèle de Cox (survie)** : temps jusqu'à l'évènement : fait de co-prescrie une statine augmente le **risque instantané** de saignement majeur à chaque instant t ?
- **Modèle de Poisson** : nombre d'évènements pendant la période d'exposition : fait de co-prescrie une statine augmente le **taux moyen** de saignement majeur pendant la période d'exposition ? = permet d'obtenir un **taux d'incidence** / **rapport d'incidence**
  - Permet ainsi de calculer un rapport d'incidence ajusté (IRR) pour chaque médicament concomitant
  - le λ du modèle de Poisson correspond au **taux d'incidence** (nombre d'évènements / temps d'exposition)

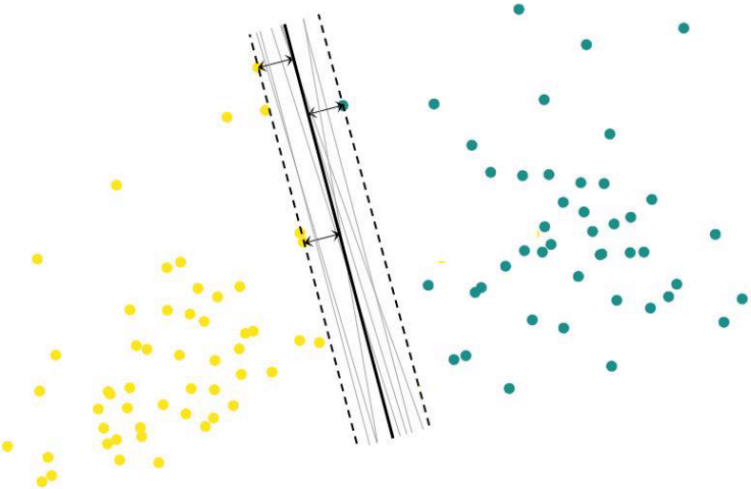


Fig. 3.4 — *Deux groupes de points sont parfaitement séparables par une infinité de droites discriminantes. Celle proposée par le SVM maximise l'écart minimal avec les points des deux groupes. Cette solution semble naturelle, elle est par ailleurs susceptible d'avoir de meilleures performances sur un échantillon de validation.*

Table 3. Major Bleeding Risk Among Patients Taking a NOAC for Nonvalvular Atrial Fibrillation With Concurrent Medications

Concurrent Medication	Person-Quarters With NOAC Use	No. of Bleeding Events	Crude Major Bleeding Incidence Rate (99% CI) per 1000 Person-Years	Adjusted Incidence Rate (99% CI) per 1000 Person-Years	Adjusted Incidence Rate Difference (99% CI) per 1000 Person-Years	Adjusted Rate Ratio (99% CI)
Atorvastatin						
With	123 420	1056	34.22 (31.51 to 36.94)	34.57 (31.87 to 37.50)	−14.38 (−17.76 to −10.99)	0.71 (0.64 to 0.78)
Without	323 617	3459	42.75 (40.88 to 44.63)	48.96 (46.48 to 51.57)	1 [Reference]	1 [Reference]
Digoxin						
With	100 513	1130	44.97 (41.52 to 48.42)	45.69 (42.23 to 49.43)	−4.46 (−8.45 to −0.47)	0.91 (0.83 to 0.99)
Without	346 524	3413	39.40 (37.66 to 41.13)	50.14 (47.34 to 53.11)	1 [Reference]	1 [Reference]

- Person-quarters with NOAC use : nombre de trimestres-personnes d'exposition aux AOD
- No. of bleeding events : nombre de saignements majeurs observés
- Crude major bleeding incidence rate : taux d'incidence brut (non ajusté) de saignement majeur par 1000 personnes-années (se fait en faisant par exemple 1056 / (123420 * 3 mois) * 12 mois * 1000)
- Adjusted incidence rate : taux d'incidence ajusté de saignement majeur par 1000 personnes-années (obtenu par le modèle de Poisson ajusté)
- Pour faire les IC : estimateur Sandiwch (décrit dans les méthodes)

Le problème est qu'ils ont

- à la fois ajusté pour des variables de confusion (âge, sexe, comorbidités, score de risque, etc.)
- et à la fois pondéré par le temps d'exposition (trimestres-personnes)

6 Références

- Zeileis, A., Kleiber, C., & Jackman, S. (2008). Regression models for count data in R. Journal of Statistical Software, 27(8), 1-25.

10Modèles linéaires mixtes

1 Introduction aux modèles linéaires mixtes

- Principale limitation du modèle linéaire : l'hypothèse des observations indépendantes = indépendance des résidus
  - Observations doivent rester indépendantes après avoir été ajustées pour les prédicteurs fixes (âge, sexe)
  - Par ex : "patients-trimestres" : ce sont les mêmes patients qui sont mesurés à plusieurs reprises
- Problème : en pratique, il existe bcp de situation ou c'est pas raisonnable
  - patients d'un essai clinique sont inclus dans plusieurs centres : deux patients d'un même centre se ressemblent plus que deux patients de centres différents
  - méta-analyse vise à synthétiser les résultats de plusieurs essais thérapeutiques
  - unité statistique est la "visite" d'un patient dans un service, mais le même patient peut revenir plusieurs fois
  - la même mesure biologique est répétée plusieurs fois chez un même individu (données longitudinales)
- Conséquence principale : le nombre de patients "effectif" est plus petit que la taille de la base de données $\rightarrow$ inflation de l'erreur de type 1

Bonus Résidus

1.A Quelques stratégies pour gérer la non-indépendance des observations

- Ignorer l'effet centre / le caractère longitudinal
  - Par ex : 4 visites pour 1 patient = 1 seul patient
  - Problème : perte d'information donc perte de puissance
- Estimer un modèle par centre / par patient et "combiner" les estimations
  - Problème : pas de quantification de la variabilité inter-centres / inter-patients
- Incorporer le centre / le patient comme effet fixe (c'est à dire comme une variable catégorielle / un facteur)
  - Problème : perte de puissance, pas de quantification de la variabilité inter-centres / inter-patients, impossible si bcp de centres avec peu de patients par centre

Bonus Variabilité

Stratégie privilégiée : Modèles linéaires mixtes (MLM) :

- Combine des effets **fixes** : covariables qui ont une influence sur la moyenne dont on veut estimer l'effet
- et des effets **aléatoires** : variables qui modélisent la structure de corrélation entre les observations (rendent compte de la non-indépendance des observations)

Effets aléatoires : "covariables latentes" qui ne sont pas observées mais qui influencent la variable réponse

(par exemple $\rightarrow$ le fait que les patients d'un même centre se ressemblent plus que des patients de centres différents)

Bonus effets

Exemple d'effet aléatoire : **niveau de programmation en R dans la salle**

Probabilité de réponse dépend de la difficulté du questionnaire : impossible de la modéliser = effet non mesuré et aléatoire = effet aléatoire

1.B Le modèle linéaire mixte (formulation mathématique)

Modèle :

$$Y_{ij} = \beta X_{ij} + \gamma_1 Z_{ij} + \epsilon_{ij}$$

- Y_{ij} : réponse (quantitative) pour l'observation du patient j issu du cluster i (par ex : mesure biologique au temps j pour le patient i)
- βX_{ij} : effet fixe (unique ici) dépendant de la variable X_{ij} (âge, sexe, traitement, etc.)
- $\gamma_1 Z_{ij}$: effet aléatoire dépendant de la variable Z_{ij} (par ex : centre, patient, etc.). Le coefficient γ_1 est commun à tous les patients du cluster i
- ϵ_{ij} : erreur aléatoire (résidu) pour l'observation du patient j du cluster i (dont on fait l'hypothèse qu'elle suit une loi normale centrée réduite $N(0, \sigma^2)$)

L'effet aléatoire γ_i n'est pas observé, il faut spécifier sa loi ($\gamma_i \sim N(0, \sigma_Z^2)$)

Ainsi, 2 patients du même cluster i se ressemblent plus que 2 patients de clusters différents car ils partagent la même valeur de l'effet aléatoire γ_i

$\rightarrow$ permet de rendre compte de la non-indépendance des observations sur les patients d'un même cluster i

2 applications possibles :

- Estimer les effets fixes et prendre en compte la non-indépendance des observations (plusieurs visites par patient, patients dans des centres différents, etc.)
- Estimer séparément les composantes de la variance (variabilité intra-cluster et inter-cluster)

1.C Estimation

$$Y_{ij} = \beta X_{ij} + \gamma_1 Z_{ij} + \epsilon_{ij}$$

L'ordinateur fournit une estimation :

- des effets fixes β et leur incertitude (binaire, catégorielle avec catégorie de référence…) (comme dans un modèle linéaire classique)
- des composantes de la variance :
  - variance de l'effet aléatoire : σ_Z^2 et son incertitude (variabilité inter-cluster)

  - variance résiduelle : σ^2 et son incertitude

Avec des modèles plus complexes, il peut y avoir plusieurs effets aléatoires avec une structure de corrélation plus complexe entre eux (matrice de variance-covariance)

> **Note**
>
> Exemple : questions à un examen (items) posées à des étudiants (individus)
>
> - Caractéristiques des questions (difficulté, type de question, etc.) = effets aléatoires
> - Caractéristiques des étudiants (niveau, sexe, etc.) = effets aléatoires
>
> **= Croisement des deux effets aléatoires**
>
> - Corrélation intra-questions (les réponses aux questions se ressemblent plus entre elles que les réponses à des questions différentes)
> - Corrélation intra-étudiants (les réponses d'un étudiant se ressemblent plus entre elles que les réponses de différents étudiants)
>
> C'est différent de patients inclus dans des centres (effet emboîté)
>
> - Patients 1 traité dans un centre A
> - Patient 2 traité dans un centre B
>
> Donc les effets sont emboîtés car un patient n'est traité que dans un seul centre

2 types d'estimateurs possibles :

- Maximum de vraisemblance (ML)
- Maximum de vraisemblance restreint (REML) = option par défaut dans R

Bonus : bonus intercept - variances - effets aléatoires

1.D Types d'estimateurs

- Maximum de vraisemblance (ML)
  - Variances estimées biaisées, surtout avec de petits échantillons
  - Possibilité de faire des tests de rapport de vraisemblance (tester la pertinence de l'ajout d'un effet fixe dans le modèle)
- Maximum de vraisemblance restreint (REML) = option par défaut dans "'{html}
  - Variances estimées non biaisées
  - Important, entre autres, pour le calcul de quantités (telles que l'ICC) dépendant des variances

Choix de l'estimateur dépend de l'objectif de l'analyse

- Si on veut quantifier un effet traitement en évacuant le côté "données en cluster" = ML

- Si on veut d'évaluer et d'interpréter les composantes de la variance (variabilité intra- et inter-cluster) = REML

2 Utilisation de R pour les modèles linéaires mixtes

Charger une base dental.RData contenant des mesures dentaires répétées chez des enfants (quantitatif)

Mesure à 8-10-12-14 ans

2 représentations possibles des données

2.A Représentations

2.A.1 Version Wide (large)

- 1 ligne par individu
- 1 colonne par mesure répétée
- Inadapté pour l'analyse avec R

```
head(dental)
```

```
# A tibble: 6 x 6
     id sex      y8   y10   y12   y14
  <dbl> <fct> <dbl> <dbl> <dbl> <dbl>
1     1 Girl   21    20    21.5  23
2     2 Girl   21    21.5  24    25.5
3     3 Girl   20.5  24    24.5  26
4     4 Girl   23.5  24.5  25    26.5
5     5 Girl   21.5  23    22.5  23.5
6     6 Girl   20    21    21    22.5
```

2.A.2 Version Long (longue)

- 1 ligne par mesure répétée
- 1 colonne pour l'âge (temps)
- 1 colonne pour la mesure d'intérêt (distance)
- Adapté pour l'analyse avec R

Utilisation de la fonction `melt()` du package `reshape2` pour transformer la base wide en base long.

Syntaxe : `melt(data, ..., na.rm = FALSE, value.name = "value")`

- `data` = Data set to melt
- `...` = further arguments passed to or from other methods.

- `na.rm` = Should NA values be removed from the data set? This will convert explicit missings to implicit missings.
- `value.name` = name of variable used to store values

```
dental_long = melt(dental, id.vars=c("id","sex"), variable.name="time",
↪  value.name="distance")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

```
dental_long
```

```
    id  sex time distance
1    1 Girl   y8     21.0
2    2 Girl   y8     21.0
3    3 Girl   y8     20.5
4    4 Girl   y8     23.5
5    5 Girl   y8     21.5
6    6 Girl   y8     20.0
7    7 Girl   y8     21.5
8    8 Girl   y8     23.0
9    9 Girl   y8     20.0
10  10 Girl   y8     16.5
11  11 Girl   y8     24.5
12  12  Boy   y8     26.0
13  13  Boy   y8     21.5
14  14  Boy   y8     23.0
15  15  Boy   y8     25.5
16  16  Boy   y8     20.0
17  17  Boy   y8     24.5
18  18  Boy   y8     22.0
19  19  Boy   y8     24.0
20  20  Boy   y8     23.0
21  21  Boy   y8     27.5
22  22  Boy   y8     23.0
23  23  Boy   y8     21.5
24  24  Boy   y8     17.0
25  25  Boy   y8     22.5
26  26  Boy   y8     23.0
27  27  Boy   y8     22.0
28   1 Girl  y10     20.0
29   2 Girl  y10     21.5
30   3 Girl  y10     24.0
31   4 Girl  y10     24.5
32   5 Girl  y10     23.0
33   6 Girl  y10     21.0
```

34	7	Girl	y10	22.5
35	8	Girl	y10	23.0
36	9	Girl	y10	21.0
37	10	Girl	y10	19.0
38	11	Girl	y10	25.0
39	12	Boy	y10	25.0
40	13	Boy	y10	22.5
41	14	Boy	y10	22.5
42	15	Boy	y10	27.5
43	16	Boy	y10	23.5
44	17	Boy	y10	25.5
45	18	Boy	y10	22.0
46	19	Boy	y10	21.5
47	20	Boy	y10	20.5
48	21	Boy	y10	28.0
49	22	Boy	y10	23.0
50	23	Boy	y10	23.5
51	24	Boy	y10	24.5
52	25	Boy	y10	25.5
53	26	Boy	y10	24.5
54	27	Boy	y10	21.5
55	1	Girl	y12	21.5
56	2	Girl	y12	24.0
57	3	Girl	y12	24.5
58	4	Girl	y12	25.0
59	5	Girl	y12	22.5
60	6	Girl	y12	21.0
61	7	Girl	y12	23.0
62	8	Girl	y12	23.5
63	9	Girl	y12	22.0
64	10	Girl	y12	19.0
65	11	Girl	y12	28.0
66	12	Boy	y12	29.0
67	13	Boy	y12	23.0
68	14	Boy	y12	24.0
69	15	Boy	y12	26.5
70	16	Boy	y12	22.5
71	17	Boy	y12	27.0
72	18	Boy	y12	24.5
73	19	Boy	y12	24.5
74	20	Boy	y12	31.0
75	21	Boy	y12	31.0
76	22	Boy	y12	23.5
77	23	Boy	y12	24.0
78	24	Boy	y12	26.0
79	25	Boy	y12	25.5
80	26	Boy	y12	26.0
81	27	Boy	y12	23.5

```
82   1 Girl  y14     23.0
83   2 Girl  y14     25.5
84   3 Girl  y14     26.0
85   4 Girl  y14     26.5
86   5 Girl  y14     23.5
87   6 Girl  y14     22.5
88   7 Girl  y14     25.0
89   8 Girl  y14     24.0
90   9 Girl  y14     21.5
91  10 Girl  y14     19.5
92  11 Girl  y14     28.0
93  12  Boy  y14     31.0
94  13  Boy  y14     26.5
95  14  Boy  y14     27.5
96  15  Boy  y14     27.0
97  16  Boy  y14     26.0
98  17  Boy  y14     28.5
99  18  Boy  y14     26.5
100 19  Boy  y14     25.5
101 20  Boy  y14     26.0
102 21  Boy  y14     31.5
103 22  Boy  y14     25.0
104 23  Boy  y14     28.0
105 24  Boy  y14     29.5
106 25  Boy  y14     26.0
107 26  Boy  y14     30.0
108 27  Boy  y14     25.0
```

On peut aussi faire avec `gather()` du package `tidyr` (tidyverse)

Syntaxe : `gather( data, key = "key", value = "value", ..., na.rm = FALSE, convert = FALSE, factor_key = FALSE )`

avec :

- `data` = data frame.
- `key`, `value` = Names of new key and value columns, as strings or symbols. This argument is passed by expression and supports quasiquotation (you can unquote strings and symbols).
- `...` = A selection of columns. If empty, all variables are selected. You can supply bare variable names, select all variables between x and z with x:z, exclude y with -y.
- `na.rm` = If TRUE, will remove rows from output where the value column is NA.
- `convert` = If TRUE will automatically run type.convert() on the key column.
- `factor_key` = If FALSE, the default, the key values will be stored as a character vector. If TRUE, will be stored as a factor, which preserves the original ordering of the columns.

```r
dental_long2 = gather(dental, key="time", value="distance", -id, -sex)
```

```
Warning: attributes are not identical across measure variables; they will be
```

```
dropped
```

```r
head(dental_long2)
```

```
# A tibble: 6 x 4
     id sex   time  distance
  <dbl> <fct> <chr>    <dbl>
1     1 Girl  y8        21
2     2 Girl  y8        21
3     3 Girl  y8        20.5
4     4 Girl  y8        23.5
5     5 Girl  y8        21.5
6     6 Girl  y8        20
```

2.B Modèle linéaire mixte avec R

Pour gérer les données manquantes : pas grave si elles sont MAR. R gère automatiquement les données manquantes en utilisant toutes les données disponibles pour chaque individu (approche "full information maximum likelihood" FIML)

2.B.1 Utilisation de la fonction `lmer()` du package `lme4`

Syntaxe : `lmer(response ~ fixed_effects + (random_intercept | grouping_factor_1) + (random_intercept | grouping_factor_2), data = data_frame, REML = TRUE )`

Avec :

- `response` = Variable dépendante numérique (ce que le modèle cherche à prédire).
- `fixed_effects` = Effets fixes, c'est-à-dire les variables explicatives dont on estime un coefficient unique pour l'ensemble de la population.
  - on met 1 s'il n'y a pas de covariable (ici pas d'effet fixe = intercept seul)
- `(random_intercept | grouping_factor_1)` = Effet aléatoire spécifiant un intercept propre à chaque niveau du facteur grouping_factor_1.
  - `grouping_factor_1` = variable sur laquelle on veut modéliser un effet aléatoire (par ex : patient, centre, etc.)
  - `random_intercept` = indique qu'on modélise un intercept qui varie selon les niveaux de grouping_factor_1 (c'est à dire un coefficient qui soit différent pour chaque niveau de grouping_factor_1, donc pour chaque patient, centre, etc..)
- `(random_intercept | grouping_factor_2)` = Deuxième effet aléatoire, indépendant du premier, avec un intercept variant selon grouping_factor_2.
- `data_frame` = Jeu de données contenant toutes les variables utilisées dans le modèle.
- `REML` = Argument logique indiquant si l'estimation doit utiliser REML (TRUE) ou ML (FALSE).
  - Par défaut, REML = TRUE donc pas besoin de le spécifier

```r
library(lme4)
fitl <- lmer(distance ~ 1 + (1 | id), data = dental_long, REML = TRUE)
summary(fitl)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: distance ~ 1 + (1 | id)
   Data: dental_long

REML criterion at convergence: 515.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2400 -0.5277 -0.1073  0.4732  2.7687

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.752    1.937
 Residual             4.930    2.220
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  24.0231     0.4297 26.0000   55.91   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Vraisemblance restreinte à l'estimation des paramètres : 515.4 (difficile à interpréter)

Information sur les résidus (Scaled residuals) : min, 1er quartile, médiane, 3ème quartile, max

Informations sur les effets aléatoires :

- Random effects :
  - Nombre d'observations : 108
  - Nombre de groupes (patients) : 27
  - 2 effets aléatoires pour lesquelles la variance est rapportée
    - Residual : variance résiduelle (intra-cluster) = σ^2 = 4.930
    - id (Intercept) : variance de l'effet aléatoire = patient = σ_Z^2 = 3.752
- Fixed effects :
  - Moyenne estimée : 24.0231
  - Erreur standard de la moyenne estimée : 0.4297
  - Pas de p-value associée (pas de test statistique par défaut dans R pour les effets fixes dans les MLM)

Les deux variances estimées (résiduelle et effet aléatoire) permettent de quantifier la variabilité intra- et inter-cluster.

La variabilié résiduelle est du même ordre de grandeur que la variabilité inter-patients.

On a donc raison d'utiliser ce modèle pour estimer la moyenne car il y a une forte variabilité entre patients, quasiment du même ordre de grandeur que la variabilité résiduelle !

2.B.2 Utilisation de la fonction `lme()` du package `nlme`

Syntaxe : `lme(fixed, data, random, correlation, weights, subset, method, na.action, control, contrasts, keep.data)`

Ici : `fit2 <- lme(fixed = distance ~ 1, data = dental_long, random = ~ 1 | id, method = "REML")`

2 formules séparées :

- Une pour la moyenne : `distance ~ 1` (1 = pas d'effet fixe = on estime juste la moyenne globale)
- Une pour les effets aléatoires : `random = ~ 1 | id` (effet aléatoire du patient sur l'intercept)

Avec :

- `fixed` = Formule des effets fixes : `response ~ covariables`.
  - Exemple : `distance ~ age + sexe` pour estimer des effets moyens de l'âge et du sexe.
- `data` = Jeu de données contenant toutes les variables du modèle.
  - Exemple : `data = dental_long`.
- `random` = Spécification des effets aléatoires.
  - Exemple : `~ 1 | patient` pour un intercept aléatoire par patient.
  - Exemple : `~ age | patient` pour un intercept **et** une pente aléatoire selon l'âge.
- `correlation` = Structure de corrélation intra-groupe (optionnelle).

  Exemple : `corAR1(form = ~ time | patient)` pour une autocorrélation AR(1) dans le temps au sein de chaque patient.
- `weights` = Structure d'hétéroscédasticité (optionnelle).
  - Exemple : `varIdent(form = ~ 1 | sexe)` pour autoriser une variance différente selon le sexe.
- `subset` = Sous-ensemble des lignes utilisées.
  - Exemple : `subset = age > 10` pour ne garder que certains individus.
- `method` =Méthode d'estimation : `"REML"` ou `"ML"`.
  - Exemple : utiliser `"ML"` si l'on compare des modèles à effets fixes différents.
- `na.action` = Comportement en présence de valeurs manquantes.
  - Par défaut : `na.fail` (le modèle s'arrête si NAs).

- `control` = Paramètres de contrôle de l'algorithme d'estimation.
  - Exemple : `control = lmeControl(msMaxIter = 200)` pour augmenter le nombre d'itérations.
- `contrasts` = Choix des contrastes pour les variables catégorielles.
  - Exemple : définir des contrastes spécifiques pour un facteur.
- `keep.data` = Indique si le jeu de données doit être conservé dans l'objet modèle. Utile pour réutiliser l'objet plus tard.

```r
fit2 <- lme(fixed = distance ~ 1, data = dental_long, random = ~ 1 | id, method =
↪  "REML")
summary(fit2)
```

```
Linear mixed-effects model fit by REML
  Data: dental_long
       AIC      BIC    logLik
  521.3618 529.3803 -257.6809

Random effects:
 Formula: ~1 | id
        (Intercept) Residual
StdDev:    1.937002 2.220312

Fixed effects:  distance ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 24.02315 0.4296606 81 55.91192       0

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.2400448 -0.5277439 -0.1072888  0.4731815  2.7687301

Number of Observations: 108
Number of Groups: 27
```

Formula : donne des écarts-types (StdDev) pour l'effet aléatoire (Intercept) et pour les résidus (Residual) (et pas des variances comme avec lmer) (pour passer de l'un à l'autre, il faut faire variance = StdDev²)

- StdDev de l'effet aléatoire (Intercept) = 1.937002 $\longrightarrow$ variance de l'effet aléatoire = 3.752 (même valeur que précédemment avec lmer)
- StdDev des résidus (Residual) = 2.220312 $\longrightarrow$ variance résiduelle = 4.930 (même valeur que précédemment avec lmer)

Distance moyenne estimée = 24.02315 (même valeur que précédemment avec lmer).

Interprétation = la distance moyenne dentaire chez les enfants est estimée à environ 24.02 unités.

2.B.3 Exemple 3 : 'lmerTest' : obtention de p-values pour les effets fixes

Utilisation de la fonction `lmer()` du package `lmerTest` (extension de `lme4`)

Le fait d'importer la library `lmerTest` remplace automatiquement la fonction `lmer()` de `lme4` par celle de `lmerTest` qui fournit des p-values pour les effets fixes.

Syntaxe : `lmer(response ~ fixed_effects + (random_intercept | grouping_factor_1) + (random_intercept | grouping_factor_2), data = data_frame, REML = TRUE )`

Ici : `fit3 <- lmer(distance ~ time + (1 | id), data = dental_long, REML = TRUE)`

En gros, on veut estimer l'effet moyen de l'âge (time) sur la distance dentaire en prenant en compte la non-indépendance des mesures répétées chez un même enfant (id)

```r
library(lmerTest)
fit3 <- lmer(distance ~ time + (1 | id), data = dental_long, REML = TRUE)
summary(fit3)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: distance ~ time + (1 | id)
   Data: dental_long

REML criterion at convergence: 443.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7376 -0.5248  0.0153  0.4027  3.7212

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 4.465    2.113
 Residual             2.078    1.442
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  22.1852     0.4923 43.3911  45.066  < 2e-16 ***
timey10       0.9815     0.3924 78.0000   2.501   0.0145 *
timey12       2.4630     0.3924 78.0000   6.277 1.80e-08 ***
timey14       3.9074     0.3924 78.0000   9.958 1.52e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
        (Intr) timy10 timy12
timey10 -0.399
```

```
timey12 -0.399  0.500
timey14 -0.399  0.500  0.500
```

- 3 coefficients estimés : la variable `time`a été exprimée comme une variable catégorielle avec 4 modalités (8, 10, 12 et 14 ans) avec 8 ans comme catégorie de référence
  - `timey10` = différence moyenne estimée entre 10 ans et 8 ans = 0.9815 (p-value = 0.0145)
  - `timey12` = différence moyenne estimée entre 12 ans et 8 ans = 2.4630 (p-value = 1.80e-08)
  - `timey14` = différence moyenne estimée entre 14 ans et 8 ans = 3.9074 (p-value = 1.52e-15)
- La library `lmerTest` fournit des p-values *relativement* valides y compris avec une estimation REML (mais pas parfaite non plus).
- Aussi !! La variance et l'écart-type ont ÉVOLUÉS : elles sont différentes de celles des modèles précédents car on a ajouté un effet fixe `time` dans le modèle.
  - C'est comme si on avait "ignoré" la variabilité expliquée par l'âge dans les modèles précédents (sans `time`), donc la variabilité résiduelle diminue.
  - On a donc diminué la variabilité individuelle (résiduelle) en expliquant une partie de cette variabilité par l'âge.
  - On pourrait calculer un ICC en faisant :
    * ICC = variance effet aléatoire / (variance effet aléatoire + variance résiduelle)
    * ICC = 4.465 / (4.465 + 2.078) = 0.682
    * Donc la variance expliquée par le patient compte à peu près pour 68.2% de la variance totale (après avoir ajusté pour l'âge)
  - Avoir le minimum de variance résiduelle possible est important pour avoir des intervalles de confiance les plus précis possibles sur les effets fixes.

2.B.4 Pertinence de la variable `time` globale dans le modèle

On pourrait faire un rapport de vraisemblance entre :

- Modèle complet avec `time` : `fit3 <- lmer(distance ~ time + (1 | id), data = dental_long, REML = TRUE)`
- Modèle sans `time` : `fitl <- lmer(distance ~ 1 + (1 | id), data = dental_long, REML = TRUE)`

Le problème c'est qu'on a fitté les deux modèles avec REML (par défaut dans R) donc on ne peut pas faire de test de rapport de vraisemblance, parce que les vraisemblances restreintes ne sont pas comparables entre deux modèles avec des effets fixes différents.

On peut donc faire avec `anova()` qui utilise une approximation de Satterthwaite pour calculer une p-value globale pour la variable `time`

```r
anova(fit3)
```

```
Type III Analysis of Variance Table with Satterthwaite's method
     Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)
time 237.19  79.064     3    78   38.04 2.986e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- p-value globale pour la variable `time` < 2.2e-16 $\longrightarrow$ la variable `time` est globalement significative dans le modèle

Il faut donc bien mettre un effet fixe `time` dans le modèle et il y a un effet significatif de l'âge sur la distance dentaire chez les enfants.

2.B.5 Pertinence de la variable `sex` dans le modèle

`sex` est une variable binaire donc c'est différent!

On peut faire un test de Wald car on ne compare pas des vraisemblances restreintes entre deux modèles différents.

3 Applications (études publiées)

3.A Déclin cognitif des personnes âgées aux États-Unis pendant et après la Covid !!

Cognitive decline in older adults in the UK during and after the COVID-19 pandemic: a longitudinal analysis of PROTECT study data

Anne Corbett, Gareth Williams, Byron Creese, Adam Hampshire, Vincent Hayman, Abbie Palmer, Akos Filakovzsky, Kathryn Mills, Jeffrey Cummings, Dag Aarsland, Zunera Khan, Clive Ballard

Summary

Background Although the long-term health effects of COVID-19 are increasingly recognised, the societal restrictions during the COVID-19 pandemic hold the potential for considerable detriment to cognitive and mental health, particularly because major dementia risk factors—such as those related to exercise and dietary habits—were affected during this period. We used longitudinal data from the PROTECT study to evaluate the effect of the pandemic on cognition in older adults in the UK.

= Déclin cognitif des personnes âgées aux États-Unis pendant et après la Covid

- Données longtitudinales
- Patients de 50 ans ou plus
- 3 périodes d'observation :

  - Avant 1ère vague (jusqu'au 29 février 2020)
  - Pendant 1ère vague (1er mars 2020 - 28 février 2021)
  - Seconde vague (1er mars 2021 - 28 février 2022)
- *NB : si on avait eu 2 temps : avant et après la 1ère vague, on aurait pu faire une analyse en différences de différences, c'est à dire en faisant juste la différence pour chaque patient, qui permet de retrouver des mesures indépendantes*
- Ont utilisé un modèle linéaire mixte pour modéliser le déclin cognitif en fonction du temps (avant, pendant, après la 1ère vague)

3.B Analyse statistique

- Scores cognitifs assignés à trois périodes : pré-pandémie, année 1, année 2.
- Modèle mixte linéaire :
  - score cognition = à gauche comme outcome (quantitatif) = Y_{ij}
  - Effets aléatoires individuels patients ajoutés (variation propre à chaque participant) : $(1|id)$ = $\gamma_1 Z_{ij}$
  - Covariables : âge (continu), sexe — facteurs connus pour influencer la cognition : βX_{ij}
  - année pandémie = variable explicative = variable d'intérêt : catégorielle à 3 niveaux (avant, année 1, année 2) = effet fixe = βX_{ij}
- Modèle utilisé pour estimer les différences entre moyennes selon les années pandémie.
- Analyses incluant les participants manquants à une ou deux périodes (tests t).
- Packages utilisés : `lmer`, `emmeans`.
- Résultats rapportés en Cohen's d effect sizes + IC 95 % + p-values.
  - Ils auraient pu rapporter des coefficients de régression (qui s'interprète comme une différence de moyenne dans un modèle linéaire)
  - Le problème : c'est que les scores cognitifs n'ont pas d'unité claire (ils sont composites) donc difficile d'interpréter une différence de moyenne en "unités de score cognitif"
    * Par exemple : une différence de moyenne de -3 unités de score cognitif entre avant et année 1 : est-ce que c'est beaucoup ? peu ? comment savoir ?
  - Effect size de Cohen = (moyenne groupe 1 - moyenne groupe 2) / écart-type pooled = coefficient rapporté à l'écart-type en population = pas d'unité
  - Objectif = **ADIMENSIONNER** le coefficient
  - **on met donc en rapport la différence de score avec la dispersion des scores entre les patients**

Si tu veux une version encore plus condensée ou orientée "méthodes stats", je peux le faire.

3.C Résultats

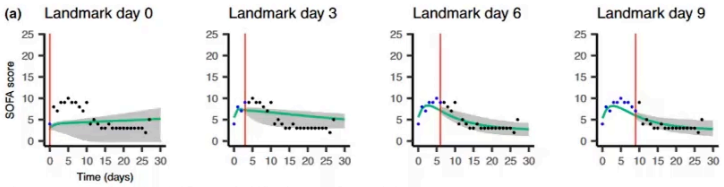
	Pandemic year 1		Pandemic year 2	
	Cohen's d effect size (95% CI)	p value	Cohen's d effect size (95% CI)	p value
Whole cohort				
Executive function				
Verbal Reasoning	0·15 (0·12 to 0·17)	<0·0001	0·02 (0·00 to 0·04)	0·12
Working memory				
Paired Associate Learning	0·77 (0·30 to 0·37)	<0·0001	0·74 (0·72 to 0·76)	<0·0001
Self-Ordered Search	0·15 (0·13 to 0·18)	<0·0001	0·15 (0·13 to 0·18)	<0·0001
Digit Span	0·19 (0·17 to 0·21)	<0·0001	0·14 (0·12 to 0·16)	<0·0001
Composite	0·51 (0·49 to 0·53)	<0·0001	0·47 (0·44 to 0·49)	<0·0001

- 0,15 : déclin (devrait être un signe négatif en vrai)
  - ne correspond pas à un nombre de points
  - correspond à "0,15 écarts-types en dessous à l'année 1 par rapport à l'année 0"

4 Modèles linéaires mixtes : extension

4.A Modèles non linéaires mixtes (NLMM)

Modèles non linéaires mixtes (NLMM)
Ex : Evolution d'un biomarqueur



Lavalley-Morelle A et al. CPT Pharmacometrics Syst Pharmacol. 2022 Nov;11(11):1472-1484.
doi: 10.1002/psp4.12856

Exemple : évolution d'un biomarqueur

- Modèle non linéaire (exponentiel, logistique, etc.)
- Incorporation d'effets aléatoires (patient, centre, etc.)
- Estimation des paramètres fixes et des composantes de la variance

Différence avec un modèle linéaire mixte : la relation entre la variable dépendante et les covariables n'est pas linéaire.

Donc ça ne s'interpète pas comme une différence de moyenne mais comme un paramètre de croissance, une constante de temps, etc.

4.B Modèles linéaires généralisées mixtes (GLMM)

Modèles linéaires généralisés mixtes (GLMM)
Ex : logistique mixte

Bouzid D et al. BMC Med Educ. 2022 Dec 13;22(1):861. d
doi: 10.1186/s12909-022-03919-1

Table 1 Summary of the factors influencing students'scores variability

	Station 1				Sta
	Score median- IQR (Q1-Q3) or % success	Agreement N (%)	Students' ICC (%)	Raters'ICC (%)	Sco me or
score (/100)	60 (50-70)	21 (25)	60.2	23.0	60 (
item 1	65	76 (89)	97	3	61
item 2	93	75 (88)	82	18	64
item 3	94	74 (87)	*	*	99
item 4	95	76 (89)	*	*	39
item 5	33	67 (79)	31	48	38
item 6	48	80 (94)	92	0	79

Exemple : logistique mixte

- Variable dépendante binaire (oui/non)
- Incorporation d'effets aléatoires (patient, centre, etc.)
- Estimation des paramètres fixes (odds ratios) et des composantes de la variance

Différence avec un modèle linéaire mixte : la variable dépendante n'est pas quantitative et suit une loi de probabilité d'une famille exponentielle (binomiale, Poisson, etc.)

Donc ça ne s'interpète pas comme une différence de moyenne mais comme un odds ratio, un risque relatif, etc.

11Résumé Modèle de Poisson et MLL

12Modèle de Poisson

- Modélisation de données de comptage (ex: nombre de visites aux urgences, nombre de crises d'asthme, nombre de suicides, etc.)
  - Application fréquente aux taux d'incidence et risque relatif dans les études épidémiologiques
  - Attention aux tests staistiques et intervalles de confiance en cas de surdispersion
  - Modéliser à part les zéros (hurdel ou zero-inflated models) si la situation le nécessite

13Modèle linéaire mixte

- Utile en cas de donées répétées longitudinales ou de données groupées (hiérarchiques)
- Permet de modéliser la corrélation entre mesures répétées chez un même individu ou entre individus d'un même groupe
  - La plupart du temps, on rapporte les effets fixes uniquement (et le MLL permet de prendre en compte la structure des données)
  - Il arrive que l'on s'intéresse aux composantes de la variance (reproductibilité)
- Attention dans l'utilisation des tests statistiques ou de toutes généralisation (GLMM, NLMEM) car les méthodes d'estimation sont plus complexes

14Modèles linéaires mixtes TP

Transformer le jeu de données dental en format long.

```
dental_long <- melt(dental, id.vars = c("id", "sex"), variable.name = "time",
↪  value.name = "distance")
```

```
Warning: attributes are not identical across measure variables; they will be
dropped
```

```
head(dental_long)
```

```
  id  sex time distance
1  1 Girl   y8     21.0
2  2 Girl   y8     21.0
3  3 Girl   y8     20.5
4  4 Girl   y8     23.5
5  5 Girl   y8     21.5
6  6 Girl   y8     20.0
```

1. Est-ce que la distance moyenne dépend du sexe, après ajustement sur le temps et d'un effet aléatoire "patient" ?
2. Codage catégoriel de la variable temps (time) ou quantitatif ? Arguments pour l'un ou l'autre ?

1 Impact du sexe sur la distance, ajusté sur le temps et un effet aléatoire patient.

1.A Modèle avec temps en tant que variable catégorielle

Stratégie : utilisation de la fonction lmer() du package lmerTest.

Syntaxe : `lmer(response ~ fixed_effects + (random_intercept | grouping_factor_1) + (random_intercept | grouping_factor_2), data = data_frame, REML = TRUE )`

On propose donc : `lmer(distance ~ sex + time + (1 | id), data = dental_long, REML = TRUE )`

Avec :

- `response = distance` = Variable dépendante numérique (ce que le modèle cherche à prédire).
- `fixed_effects` = Effets fixes, c'est-à-dire les variables explicatives dont on estime un coefficient unique pour l'ensemble de la population.
  - ici : `sex` (variable catégorielle) et `time` (variable catégorielle).

- `(random_intercept | grouping_factor_1)` = (1 | id) = Premier effet aléatoire, avec un intercept variant selon grouping_factor_1.
  - `random_intercept` = 1 (intercept variable) selon les niveaux de `grouping_factor_1` = `id` (identifiant unique de chaque enfant).
  - `grouping_factor_1` = `id` (identifiant unique de chaque enfant).
- `data_frame` = Jeu de données contenant toutes les variables utilisées dans le modèle.
- `REML` = Argument logique indiquant si l'estimation doit utiliser REML (TRUE) ou ML (FALSE).
  - Par défaut, REML = TRUE donc pas besoin de le spécifier

```r
model_dental_cat <- lmer(distance ~ sex + time + (1 | id), data = dental_long,
↪  REML = TRUE)
summary(model_dental_cat)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: distance ~ sex + time + (1 | id)
   Data: dental_long

REML criterion at convergence: 433.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8225 -0.5116 -0.0066  0.4169  3.6565

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.260    1.805
 Residual             2.078    1.442
Number of obs: 108, groups:  id, 27

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  20.8098     0.6335 33.8026  32.850  < 2e-16 ***
sexBoy        2.3210     0.7614 25.0000   3.048  0.00538 **
timey10       0.9815     0.3924 78.0000   2.501  0.01447 *
timey12       2.4630     0.3924 78.0000   6.277 1.80e-08 ***
timey14       3.9074     0.3924 78.0000   9.958 1.52e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
        (Intr) sexBoy timy10 timy12
sexBoy  -0.712
timey10 -0.310  0.000
timey12 -0.310  0.000  0.500
timey14 -0.310  0.000  0.500  0.500
```

Interprétation :

- Intercept (20.81) : Distance moyenne initiale pour les filles (référence de la variable sexe) à l'année 8 (référence de la variable temps).
- sexBoy (2.32) : Les garçons ont en moyenne une distance dentaire supérieure de 2.32mm par rapport aux filles, toutes choses égales par ailleurs (donc quelque soit le temps). Significatif (p = 0.00538).
- timey10 (0.98) : À l'année 10, la distance moyenne augmente de 0.98 mm par rapport à l'année 8, toutes choses égales par ailleurs. Significatif (p = 0.01447).
- timey12 (2.46) : À l'année 12, la distance moyenne augmente de 2.46 mm par rapport à l'année 8, toutes choses égales par ailleurs. Significatif (p = 1.80e-08).
- timey14 (3.91) : À l'année 14, la distance moyenne augmente de 3.91 mm par rapport à l'année 8, toutes choses égales par ailleurs. Significatif (p = 1.52e-15).

1.A.1 Si on veut regarder s'il y a un effet-temps différent selon le sexe

Stratégie : ajout d'un terme d'interaction entre sexe et temps.

```r
model_dental_cat_interaction <- lmer(distance ~ sex * time + (1 | id), data =
↪  dental_long, REML = TRUE)
summary(model_dental_cat_interaction)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: distance ~ sex * time + (1 | id)
   Data: dental_long

REML criterion at convergence: 423.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.7358 -0.4761  0.0469  0.4691  3.6613

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.285    1.813
 Residual             1.975    1.405
Number of obs: 108, groups:  id, 27

Fixed effects:
               Estimate Std. Error      df t value Pr(>|t|)
(Intercept)     21.1818     0.6915 46.0791  30.630  < 2e-16 ***
sexBoy           1.6932     0.8983 46.0791   1.885   0.0658 .
timey10          1.0455     0.5992 75.0000   1.745   0.0851 .
timey12          1.9091     0.5992 75.0000   3.186   0.0021 **
timey14          2.9091     0.5992 75.0000   4.855 6.41e-06 ***
```

```
sexBoy:timey10  -0.1080     0.7784 75.0000  -0.139   0.8901
sexBoy:timey12   0.9347     0.7784 75.0000   1.201   0.2337
sexBoy:timey14   1.6847     0.7784 75.0000   2.164   0.0336 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) sexBoy timy10 timy12 timy14 sxB:10 sxB:12
sexBoy      -0.770
timey10     -0.433  0.334
timey12     -0.433  0.334  0.500
timey14     -0.433  0.334  0.500  0.500
sexBy:tmy10  0.334 -0.433 -0.770 -0.385 -0.385
sexBy:tmy12  0.334 -0.433 -0.385 -0.770 -0.385  0.500
sexBy:tmy14  0.334 -0.433 -0.385 -0.385 -0.770  0.500  0.500
```

Interprétation :

- Les interactions entre sexe et temps permettent d'estimer si l'effet du temps sur la distance dentaire diffère entre les garçons et les filles.
- Par exemple, l'effet du temps à l'année 14 pour les garçons est estimé en combinant l'effet principal de timey14 (2.9091) et l'effet d'interaction sexBoy:timey14 (1.6847), ce qui donne un effet total de 4.5938 mm pour les garçons par rapport aux filles à l'année 8. (le calcul se fait en additionnant les coefficients : 2.9091 + 1.6847 = 4.5938).
- L'effet d'interaction sexBoy:timey14 est significatif (p = 0.0336), ce qui suggère que l'augmentation de la distance dentaire à l'année 14 est plus prononcée chez les garçons par rapport aux filles.

1.B Modèle avec temps en tant que variable quantitative

Ici, `time` est une variable catégorielle.

```r
dental_long$time
```

```
  [1] y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8  y8
 [19] y8  y8  y8  y8  y8  y8  y8  y8  y8  y10 y10 y10 y10 y10 y10 y10 y10 y10
 [37] y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10 y10
 [55] y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12 y12
 [73] y12 y12 y12 y12 y12 y12 y12 y12 y12 y14 y14 y14 y14 y14 y14 y14 y14 y14
 [91] y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14 y14
Levels: y8 y10 y12 y14
```

On peut créer une nouvelle variable `time_numeric` qui code le temps en tant que variable quantitative.

1ère possibilité : codage numérique direct avec `as.numeric()`

```r
dental_long$time_numeric1 <- as.numeric(dental_long$time)
dental_long$time_numeric1
```

```
  [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
 [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [75] 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
```

2e possibilité avec `substr()` pour extraire la partie numérique

1. Extraire la partie numérique de la variable `time` en utilisant la fonction `substr()` pour enlever le "y" initial.
   - `substr(dental_long$time, 2, nchar(as.character(dental_long$time)))` extrait les caractères de la position 2 à la fin de la chaîne, supprimant ainsi le "y".
2. Convertir le résultat en numérique avec `as.numeric()`.
   - `as.numeric(...)` convertit la chaîne de caractères extraite en une variable numérique.

```r
dental_long$time_numeric2 <- as.numeric(substr(dental_long$time, 2,
↪  nchar(as.character(dental_long$time))))
dental_long$time_numeric2
```

```
  [1]  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8  8
 [26]  8  8 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10 10
 [51] 10 10 10 10 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12 12
 [76] 12 12 12 12 12 12 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14 14
[101] 14 14 14 14 14 14 14 14
```

Stratégie : utilisation de la fonction lmer() du package lmerTest.

Syntaxe : `lmer(response ~ fixed_effects + (random_intercept | grouping_factor_1) + (random_intercept | grouping_factor_2), data = data_frame, REML = TRUE )`

On propose donc : `lmer(distance ~ sex + time_numeric + (1 | id), data = dental_long, REML = TRUE )`

```r
model_dental_num <- lmer(distance ~ sex + time_numeric2 + (1 | id), data =
↪  dental_long, REML = TRUE)
summary(model_dental_num)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: distance ~ sex + time_numeric2 + (1 | id)
   Data: dental_long

REML criterion at convergence: 437.5
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.7489 -0.5503 -0.0252  0.4534  3.6575 

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 3.267    1.807   
 Residual             2.049    1.432   
Number of obs: 108, groups:  id, 27

Fixed effects:
              Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)   15.38569    0.89598 87.59408  17.172  < 2e-16 ***
sexBoy         2.32102    0.76142 25.00000   3.048  0.00538 ** 
time_numeric2  0.66019    0.06161 80.00000  10.716  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr) sexBoy
sexBoy      -0.504       
time_numrc2 -0.756  0.000
```

Interprétation :

- Intercept (15.39) : Distance moyenne initiale pour les filles (référence de la variable sexe) à l'année 0 (intercept extrapolé).
- sexBoy (2.32) : Les garçons ont en moyenne une distance dentaire supérieure de 2.32mm par rapport aux filles, toutes choses égales par ailleurs (donc quelque soit le temps). Significatif (p = 0.00538).
-

15Wrap up modèles linéaires mixtes

1 1. Régression Linéaire Multiple (RLM)

Sous-partie	Contenu
Intérêt (utilité du modèle)	La RLM est utilisée pour modéliser la relation entre une **variable dépendante quantitative** et plusieurs variables indépendantes (ou explicatives) X_k. Elle sert à l'analyse multivariable (ajustement) pour estimer l'effet d'un facteur (X_1) en maintenant les autres facteurs ($X_2, \ldots$) constants.
Nature de la variable dépendante Y	La variable Y est attendue comme **quantitative continue**. Exemple typique : la pression artérielle, ou un score quantitatif de développement cognitif (BAS).
Structure du modèle (formule mathématique)	$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$ **Définition des termes :** Y est la variable dépendante (continue). $X_1, X_2, \ldots, X_k$ sont les variables indépendantes (quantitatives ou catégorielles). β_0 est l'ordonnée à l'origine (intercept). β_i sont les coefficients de régression. ε est le terme d'erreur (résidu théorique).
Fonction de lien et justification	**Lien utilisé : Lien identité**. Le modèle suppose une relation linéaire directe et additive entre le prédicteur linéaire et la moyenne de Y. **Formule inverse :** $\mathbb{E}(Y\|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$.
Conditions de validité / hypothèses du modèle	Les hypothèses principales portent sur le terme d'erreur ε : 1. **Moyenne nulle :** $\mathbb{E}(\varepsilon) = 0$. 2. **Normalité :** ε suit une loi normale. 3. **Variance constante (Homoscédasticité) :** Var$(\varepsilon) = \sigma^2$ (variance non dépendante de X). 4. **Indépendance des résidus :** Les ε_i sont indépendants. **Vérification :** On utilise l'histogramme et le QQ-plot des résidus pour la normalité, et le graphique des résidus en fonction des valeurs prédites ($\widehat{Y}_i$) pour l'homoscédasticité (vérifier l'absence d'hétéroscédasticité).
Gestion des variables catégorielles et interactions	Les variables catégorielles à K modalités ($K > 2$) doivent être recodées en $K - 1$ variables binaires indicatrices (variables *dummy*) en choisissant une catégorie de référence. Le coefficient β_i d'une variable indicatrice représente alors la différence moyenne de Y entre cette modalité et la catégorie de référence.

Sous-partie	Contenu
	Interactions : L'ajout d'un terme multiplicatif, comme $\beta_3(X_1X_2)$, modélise un **effet multiplicatif** (synergie) entre les variables explicatives. Attention, l'ajout d'interactions rend l'interprétation des coefficients isolés β_1 et β_2 difficile, à moins d'utiliser un codage spécifique (-1/1).
Interprétation des coefficients β	**Si** X_i est continue : β_i est la variation moyenne de Y associée à une augmentation d'une unité de X_i, toutes les autres variables X_j étant maintenues constantes. **Si** X_i est binaire (0/1) : β_i est la **différence moyenne** de Y entre les sujets avec $X_i = 1$ et les sujets avec $X_i = 0$, toutes les autres variables X_j étant maintenues constantes.
Mesure d'association clé (issue de e^β ou des β)	La mesure pertinente est la **Différence de Moyenne** (β), qui exprime de combien la valeur de Y augmente ou diminue. **Valeur neutre : 0** (zéro) signifie l'absence de différence de moyenne ou d'association.
Tests statistiques utilisés	Les tests sont typiquement basés sur la statistique t (ou Z pour de grands échantillons) pour évaluer $H_0 : \beta_i = 0$. L'intervalle de confiance (IC) sur β est utilisé.
Spécificités / extensions importantes	Limitation majeure : l'hypothèse d'indépendance des observations est souvent violée avec les données longitudinales ou groupées. Les extensions incluent les **Modèles Linéaires Mixtes (MLM)** pour gérer la non-indépendance et décomposer les variances.
Applications possibles : Exemple	Étude de l'effet de la durée de l'allaitement maternel (variable catégorielle) sur le score cognitif BAS (variable quantitative) chez l'enfant, en ajustant pour d'autres facteurs (régression linéaire multiple).

---

2 2. Régression Logistique

Sous-partie	Contenu
Intérêt (utilité du modèle)	La régression logistique est un type de Modèle Linéaire Généralisé (GLM) utilisé pour modéliser une variable dépendante de nature **binaire (dichotomique)**. Elle permet d'étudier l'association entre des variables explicatives et la probabilité p qu'un événement survienne.
Nature de la variable dépendante Y	La variable Y est **binaire dichotomique** (0/1). Elle suit une loi de Bernoulli. Exemple typique : Maladie oui/non (ex. HTA oui/non, rejet de greffe oui/non).

Sous-partie	Contenu
Structure du modèle (formule mathématique)	On modélise le logit de p comme une combinaison linéaire des variables explicatives : log $\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$. | **Définition des termes :** $p = P(Y = 1\|X)$ est la probabilité que l'événement survienne. $\frac{p}{1-p}$ est la cote (Odds). log $\left(\frac{p}{1-p}\right)$ est le logit de p. β_k sont les coefficients de régression.
Fonction de lien et justification	**Lien utilisé : Logit (ou logistique)**. **Formule du lien :** $g(p) =$ log $\left(\frac{p}{1-p}\right)$. **Justification :** Le logit permet de transformer la probabilité p (qui est contrainte entre 0 et 1) en une variable définie entre $-\infty$ et $+\infty$, ce qui est adapté à un prédicteur linéaire. **Fonction inverse (Expit) :** $p = \frac{e^{\beta_0 + \cdots + \beta_k X_k}}{1 + e^{\beta_0 + \cdots + \beta_k X_k}}$.
Conditions de validité / hypothèses du modèle	La condition principale est que la **taille de l'échantillon soit suffisante** (au minimum 10 événements par variable explicative). Les coefficients sont estimés par la méthode du maximum de vraisemblance.
Gestion des variables catégorielles et interactions	**Variables catégorielles :** L'approche est la même qu'en RLM : on définit une classe de référence, et chaque coefficient β_k correspond au log-odds ratio entre la catégorie k et la référence. **Interactions :** Elles sont gérées exactement de la même manière qu'en régression linéaire, en ajoutant un terme multiplicatif $\beta_3(X_1 X_2)$. Le modèle multivarié suppose la **multiplicativité des effets** sur l'OR, sauf si une interaction est ajoutée.
Interprétation des coefficients β	Le coefficient β correspond au **logarithme de l'Odds Ratio** (ln(OR)). exp(β) est l'Odds Ratio (OR) ajusté sur les autres variables. **Si** X_i est continue : exp(β_i) est l'OR associé à une augmentation d'une unité de X_i. **Si** $\beta_i = 0$, alors OR $= 1$, signifiant l'absence d'association.
Mesure d'association clé (issue de e^β ou des β)	La mesure clé est l'**Odds Ratio (OR)**. **Valeur neutre : 1** (OR = 1 signifie qu'il n'y a pas d'association).
Tests statistiques utilisés	1. **Test de Wald :** Permet de tester $H_0 : \beta_k = 0$ (ou OR$_k = 1$) pour un coefficient individuel. Il repose sur le fait que $\hat{\beta}/$SE$(\hat{\beta})$ suit approximativement une loi normale centrée réduite $\mathcal{N}(0, 1)$ pour un grand échantillon. 2. **Test du Rapport de Vraisemblance (LRT) :** Compare la log-vraisemblance d'un modèle complet à celle d'un modèle réduit (sans la covariable testée). La statistique $\Lambda = 2(\ell_{\text{complet}} - \ell_{\text{réduit}})$ suit une loi du χ^2. Il est souvent plus robuste que le test de Wald, notamment pour les échantillons petits ou les événements rares.

Sous-partie	Contenu
Spécificités / extensions importantes	Le test de Wald peut être peu fiable en cas de petits échantillons, d'événements très rares ou d'effets extrêmes (séparation complète). L'OR est symétrique et peut être estimé dans les enquêtes cas-témoins (où le RR ne l'est pas). **Extension :** Les **Modèles Linéaires Généralisés Mixtes (GLMM)**, incluant la régression logistique mixte, permettent d'introduire des effets aléatoires pour les données corrélées (longitudinales ou groupées).
Applications possibles :	Modélisation du risque de rejet de greffe (Y binaire) en fonction de la dose d'un médicament (X binaire) et d'autres facteurs (multivariable).
Exemple	Étude des facteurs associés à la démence (Y binaire).

---

3 3. Régression de Poisson (comptages / GLM Poisson)

Sous-partie	Contenu
Intérêt (utilité du modèle)	La régression de Poisson est un GLM adapté à la modélisation des **données de comptage** ($Y = 0, 1, 2, ...$). Elle est fréquemment utilisée en épidémiologie pour modéliser les **taux d'incidence** et obtenir des Risques Relatifs (RR) ou Rapports de Taux d'Incidence (IRR).
Nature de la variable dépendante Y	La variable Y est une variable aléatoire **discrète** prenant des valeurs dans les entiers naturels positifs ($0, 1, 2, ...$). Elle suit une loi de Poisson, $Y \sim$ Poisson(λ), où $\lambda > 0$ est le paramètre qui représente la moyenne et la variance théorique. Exemple typique : nombre de visites à l'hôpital, nombre de crises d'asthme.
Structure du modèle (formule mathématique)	On modélise le logarithme du paramètre de moyenne $\lambda = \mathbb{E}(Y\|X)$ par un prédicteur linéaire : $\log(\lambda) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ **Définition des termes :** λ est la moyenne (et la variance théorique) du comptage. X_k sont les variables explicatives. β_k sont les coefficients de régression. Si l'on modélise un taux d'incidence, on inclut souvent $\log(\text{Temps d'exposition})$ comme un *offset* dans le modèle.
Fonction de lien et justification	**Lien utilisé : Logarithmique (Log)**. **Formule du lien :** $g(\mu) = \log(\mu)$ avec $\mu = \lambda$. **Justification :** L'exponentielle du prédicteur linéaire assure que $\lambda = \mathbb{E}(Y\|X)$ est toujours positif ($\lambda > 0$), ce qui est nécessaire pour une moyenne de comptage. **Fonction inverse :** $\lambda = e^{\beta_0 + \cdots + \beta_p X_p}$.

Sous-partie	Contenu
Conditions de validité / hypothèses du modèle	L'hypothèse fondamentale de la distribution de Poisson est l'**équidispersion** : **la moyenne est égale à la variance** ($\mathbb{E}(Y\|X) = \text{Var}(Y\|X) = \lambda$). En cas de violation de cette hypothèse (surdispersion : $\text{Var}(Y) > \mathbb{E}(Y)$), les erreurs standards et les tests statistiques deviennent incorrects.
Gestion des variables catégorielles et interactions	**Variables catégorielles :** Chaque β_k représente le log du ratio des moyennes par rapport à la catégorie de référence. **Interactions :** Les effets sont additifs sur le $\log(\lambda)$, ce qui signifie qu'ils sont **multiplicatifs** sur la moyenne λ après exponentiation. Par exemple, si l'effet Fumeur et l'effet Diabète sont ajustés, leur effet combiné sur λ est la multiplication de leurs $\exp(\beta)$.
Interprétation des coefficients β	β_i représente le **logarithme du Risque Relatif** ($\ln(\text{RR})$) ou du Rapport de Moyennes/Taux. $\exp(\beta_i)$ est le Risque Relatif (RR) ajusté pour les autres variables. **Si** X_i est continu : $\exp(\beta_i)$ est le facteur multiplicatif (RR) sur la moyenne λ associé à une augmentation d'une unité de X_i.
Mesure d'association clé (issue de e^β ou des β)	La mesure clé est le **Risque Relatif (RR)** ou le **Rapport de Taux d'Incidence (IRR)**. **Valeur neutre : 1** (RR = 1 signifie l'absence d'association).
Tests statistiques utilisés	**Test de Wald** et **Test du Rapport de Vraisemblance (LRT)** sont utilisés pour l'inférence. L'estimation se fait par le maximum de vraisemblance. **Attention :** En cas de surdispersion, le test de Wald est biaisé (trop de faux positifs). Il faut alors utiliser la **variance robuste (estimateur sandwich)** ou des modèles alternatifs (Quasi-Poisson, Binomiale Négative).
Spécificités / extensions importantes	1. **Surdispersion :** $\text{Var}(Y) > \mathbb{E}(Y)$. Corrigée par l'estimateur de variance **sandwich** (corrige l'inférence sans changer $\hat{\beta}$) ou la régression **Quasi-Poisson** / **Binomiale Négative**. 2. **Excès de zéros :** Lorsque le nombre de $Y = 0$ est supérieur à la prévision de Poisson. Traité par les modèles **Hurdle** ou **Zero-Inflated** (modèles à deux composantes).
Applications possibles : Exemple	Analyse du risque de saignement majeur (nombre d'événements) chez des patients exposés à des anticoagulants, en utilisant les trimestres-personnes comme exposition (pour calculer un taux d'incidence ajusté ou IRR).

---

4 4. Modèles Linéaires Mixtes (effets fixes + aléatoires)

Sous-partie	Contenu
Intérêt (utilité du modèle)	Les Modèles Linéaires Mixtes (MLM) sont essentiels pour gérer la **non-indépendance des observations**. Ils sont utilisés dans les situations de données répétées (longitudinales) ou de données groupées (hiérarchiques), comme des patients suivis dans différents centres. Ils permettent d'estimer des effets fixes (covariables d'intérêt) tout en modélisant explicitement la corrélation au sein des groupes.
Nature de la variable dépendante Y	La variable Y est attendue comme **quantitative continue**. Exemple typique : Mesures biologiques répétées chez un patient, score cognitif mesuré à différents temps.
Structure du modèle (formule mathématique)	Un MLM combine des effets fixes et des effets aléatoires. Exemple avec intercept aléatoire de groupe (cluster i, observation j) : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}$ **Définition des termes :** β_0, β_1 sont les **effets fixes** (effets moyens dans la population). X_{ij} est la covariable. u_i est l'**effet aléatoire** de groupe i (e.g., centre, patient). On suppose $u_i \sim \mathcal{N}(0, \tau^2)$. ε_{ij} est l'erreur résiduelle individuelle. On suppose $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$. τ^2 est la variance inter-groupe, σ^2 est la variance intra-groupe (résiduelle).
Fonction de lien et justification	**Lien utilisé : Lien identité** (car Y est quantitative et suppose une distribution Gaussienne, comme en RLM). La relation est linéaire sur la moyenne conditionnelle aux effets aléatoires.
Conditions de validité / hypothèses du modèle	Les hypothèses principales sont la **normalité des effets aléatoires** u_i et la **normalité des erreurs résiduelles** ε_{ij}. On suppose aussi l'homoscédasticité des erreurs résiduelles. L'indépendance est modélisée : les Y_{ij} du même groupe sont corrélés via u_i, mais ε_{ij} sont indépendants.
Gestion des variables catégorielles et interactions	**Effets fixes :** Les variables catégorielles sont traitées comme en RLM (variable indicatrice par rapport à une catégorie de référence). Les interactions sont ajoutées comme des termes multiplicatifs. **Effets aléatoires :** Ils modélisent la structure de corrélation. On utilise la notation (1 $\vert$ groupe) pour un **intercept aléatoire** (niveau de Y qui varie entre les groupes) et (temps $\vert$ patient) pour une **pente aléatoire** (l'effet du temps varie entre les patients). Les effets peuvent être **emboîtés** (patient dans centre) ou **croisés** (patient vu par plusieurs médecins).
Interprétation des coefficients β	Les β sont les **effets fixes** : ils représentent l'effet moyen d'une covariable sur Y **dans l'ensemble de la population**. L'interprétation est celle d'une **différence de moyenne** (si X est binaire) ou d'une variation de la moyenne (si X est continu), ajustée pour la corrélation interne des données.

Sous-partie	Contenu
Mesure d'association clé (issue de e^{β} ou des β)	La mesure principale est la **Différence de Moyenne** (β) pour les effets fixes. Une autre mesure clé est le **Coefficient de Corrélation Intraclasse (ICC)**, calculé à partir des composantes de variance. **Formule de l'ICC :** $\rho = \frac{\tau^2}{\tau^2+\sigma^2}$. **Valeur neutre : 0** pour β (pas d'effet fixe).
Tests statistiques utilisés	1. **Test de Wald :** Pour les coefficients des effets fixes (β_k). Utilise des approximations (e.g., méthode de Satterthwaite) pour calculer les degrés de liberté. 2. **Test du Rapport de Vraisemblance (LRT) :** Pour comparer la pertinence de l'ajout d'un effet fixe ou d'un effet aléatoire. Nécessite l'estimation par **Maximum de Vraisemblance (ML)** pour comparer des modèles avec des effets fixes différents. L'estimation par défaut est souvent le **REML** (Maximum de Vraisemblance Restreint), qui donne des variances non biaisées.
Spécificités / extensions importantes	1. **Décomposition de la variance :** Les MLM permettent de quantifier la variabilité inter-groupe (τ^2) et intra-groupe (σ^2), ce qui est impossible avec des modèles séparés. 2. **Extensions :** • **Modèles Linéaires Généralisés Mixtes (GLMM)** pour des Y non-Gaussiens (e.g., logistique mixte pour Y binaire, Poisson mixte pour les comptages). • **Modèles Non Linéaires Mixtes (NLMM)** pour les trajectoires de croissance complexes.
Applications possibles : Exemple	Analyse de la distance dentaire (mesure quantitative) mesurée à 8, 10, 12, 14 ans chez un même enfant. Le modèle inclut l'âge (effet fixe) et un intercept aléatoire par enfant (pour tenir compte de la non-indépendance des mesures répétées).

Sous-partie	Contenu
Analogie de cloture	Imaginez que vous essayez d'estimer l'effet d'une nouvelle méthode d'enseignement sur la performance des élèves. Le **Modèle Linéaire Multiple classique** ne verrait qu'une masse unique d'élèves, comme si tous étudiaient dans la même pièce et étaient totalement indépendants. Si cette méthode marche mieux pour certains professeurs ou dans certaines écoles, le modèle classique serait trompé, car il mélangerait l'effet de la méthode avec la qualité de l'école (l'erreur de type I serait gonflée). Le **Modèle Linéaire Mixte** agit comme un filtre sophistiqué : il considère les élèves comme étant **emboîtés** dans des classes, elles-mêmes emboîtées dans des écoles. Les **effets fixes** mesurent l'effet moyen de la méthode d'enseignement pour *tous* les élèves (le résultat qui vous intéresse), tandis que les **effets aléatoires** mesurent à quel point chaque école ou chaque classe s'écarte de cette moyenne globale. C'est comme déterminer la hauteur moyenne des vagues sur l'océan (effet fixe), tout en reconnaissant que la marée varie d'une côte à l'autre (effet aléatoire inter-côtes) et que les vagues locales sont différentes selon la plage (effet aléatoire intra-côte).

A

16Variabilités interindividuelles

1 Contexte : données hiérarchiques

On considère un schéma très courant :

- des patients indexés par $i = 1, \dots, n_j$,
- regroupés dans des centres (hôpitaux, services) indexés par $j = 1, \dots, J$,
- pour chaque patient on observe une réponse Y_{ij} (par exemple : score, biomarqueur, indicateur binaire),
- et des covariables X_{ij} (âge, sexe, gravité, traitement, etc.).

Si on ignore la structure en centres, on pourrait écrire un modèle linéaire simple :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}.$$

Mais en pratique :

- les patients diffèrent entre eux (même à X_{ij} égal),
- les centres diffèrent entre eux (même profil de patients, pratiques différentes, etc.).

Les modèles linéaires mixtes servent à séparer ces sources de variabilité.

Plan :

1. Définir la variabilité inter-individuelle.
2. Définir les variations inter-centre.
3. Voir comment un modèle mixte les représente.
4. Expliquer pourquoi on ne peut pas estimer ces variances avec des modèles séparés par centre / par patient.
5. Illustrer avec un petit exemple en R.

2 Variabilité inter-individuelle

2.A Définition intuitive

Même dans un centre donné et pour les mêmes covariables X_{ij}, deux patients n'auront pas exactement la même valeur de Y_{ij}.

Cette dispersion entre patients, une fois qu'on a pris en compte les covariables, est la variabilité inter-individuelle.

Dans un modèle linéaire simple :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij},$$

on suppose souvent :

$$\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

où :

- ε_{ij} est l'erreur individuelle,
- σ^2 est la variance inter-individuelle résiduelle : à covariables égales (et dans un même centre si on l'a inclus), elle mesure à quel point les patients diffèrent entre eux.

2.B Interprétation

On peut lire σ^2 comme :

- la "dispersion" des résultats individuels autour de ce que le modèle prédit,
- ce qui reste comme variabilité entre patients après avoir ajusté sur les covariables et, le cas échéant, les effets de centre.

□

3 Variabilité inter-centre

3.A Idée de base

Même si l'on ajuste sur les mêmes covariables X_{ij}, deux centres peuvent :

- prendre en charge les patients différemment,
- avoir des équipes ou des ressources différentes,
- avoir des populations de patients légèrement différentes.

Donc, à covariables égales, la moyenne de Y_{ij} peut différer d'un centre à l'autre.

On introduit alors un effet aléatoire de centre u_j :

$$Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \varepsilon_{ij}.$$

Avec un modèle à effets aléatoires, on suppose typiquement :

$$u_j \sim \mathcal{N}(0, \tau^2), \qquad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

et u_j indépendant de ε_{ij}.

- τ^2 : variance inter-centre (variabilité entre les centres),
- σ^2 : variance inter-individuelle résiduelle (variabilité entre patients à l'intérieur d'un centre).

Au total :

$$\text{Var}(Y_{ij} \mid X_{ij}) = \tau^2 + \sigma^2.$$

3.B Coefficient de corrélation intraclasse (ICC)

Une quantité clé est l'ICC (intra-class correlation) :

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}.$$

- Si ρ est proche de 0 : la variabilité est surtout inter-individuelle, les centres se ressemblent.
- Si ρ est élevé : les différences entre centres expliquent une proportion importante de la variance totale.

□

4 Modèle linéaire mixte : séparer les deux variabilités

4.A Modèle avec intercept aléatoire de centre

On résume :

$$Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \varepsilon_{ij},$$

avec :

- $u_j \sim \mathcal{N}(0, \tau^2)$: effet aléatoire de centre,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: erreur résiduelle individuelle.

Le modèle donne :

- une estimation de β_1 (effet moyen de la covariable),
- une estimation de τ^2 (variabilité inter-centre),
- une estimation de σ^2 (variabilité inter-individuelle).

On obtient donc deux niveaux de variabilité clairement séparés.

4.B Implications pratiques

- Meilleure estimation des erreurs standards de β_1 (on ne fait pas "comme si" tout était indépendant).
- Possibilité de quantifier :

  - à quel point les centres diffèrent entre eux (via τ^2 et ρ),
  - à quel point les patients diffèrent au sein d'un centre (via σ^2).
  - On peut généraliser à d'autres centres (on modélise une distribution des centres, pas seulement ceux observés).

□

5 Pourquoi on ne retrouve pas ces variances avec des modèles séparés

5.A Modèles "par centre"

Supposons qu'au lieu d'un modèle mixte unique, on ajuste un modèle linéaire séparé dans chaque centre :

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}^{(j)}, \quad j = 1, \dots, J.$$

On obtient alors une collection de paramètres :

- $\widehat{\beta}_{0j}$,
- $\widehat{\beta}_{1j}$,

mais il n'y a pas de paramètre τ^2 dans un modèle global.

On pourrait regarder la variance empirique des $\widehat{\beta}_{0j}$ entre centres, mais cette variance mélange :

- la vraie variabilité inter-centre,
- la variabilité d'estimation (certains centres ont peu de patients, donc $\widehat{\beta}_{0j}$ est très instable).

Sans modèle commun qui impose :

$$\beta_{0j} = \beta_0 + u_j, \quad u_j \sim \mathcal{N}(0, \tau^2),$$

il est impossible de séparer proprement :

- "ce qui vient vraiment de différences entre centres" (τ^2),
- de "ce qui vient du bruit d'échantillonnage" sur chaque $\widehat{\beta}_{0j}$.

Donc :

- dans des modèles séparés par centre, on n'a pas de paramètre explicite pour la variance inter-centre,
- on ne peut pas estimer τ^2 ni un ICC propre,
- on perd la structure hiérarchique globale.

5.B Modèles “par patient” (données longitudinales)

Même idée si l’unité d’analyse est la visite et que l’on a plusieurs visites par patient.
Un modèle mixte pour des données longitudinales pourrait être :

$$Y_{ij} = \beta_0 + b_i + \beta_1 t_{ij} + \varepsilon_{ij},$$

avec :

- $b_i \sim \mathcal{N}(0, \omega^2)$: effet aléatoire patient (niveau moyen propre à chaque patient),
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: bruit intra-patient.

Alors :

- ω^2 = variabilité inter-individuelle (entre patients),
- σ^2 = variabilité intra-individuelle (au sein d’un même patient).

Si, au lieu de ça, on ajuste un modèle séparé pour chaque patient :

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}^{(i)},$$

on obtient une collection de pentes et d’intercepts :

- $\widehat{\beta}0i$, $\widehat{\beta}1i$,

mais :

- pas de paramètre ω^2 pour décrire la distribution des effets patients,
- la dispersion des $\widehat{\beta}0i$ et $\widehat{\beta}1i$ est un mélange de :
  - vraie variabilité entre patients,
  - bruit d’estimation (surtout quand chaque patient a peu de mesures).

On ne peut donc pas :

- décomposer proprement la variance entre et au sein des patients,
- prédire pour un nouveau patient à partir d’une distribution $b_i \sim \mathcal{N}(0, \omega^2)$,
- bénéficier du shrinkage (effets des petits échantillons ramenés vers la moyenne).

□

6 Illustration rapide en R

6.A Simulation avec centres

On simule des données avec :

- $J = 10$ centres,

- $n_j = 50$ patients par centre,
- un effet centre u_j de variance τ^2,
- un bruit individuel ε_{ij} de variance σ^2.

```r
J <- 10
n_per_center <- 50

center <- rep(1:J, each = n_per_center)

# Covariable individuelle
x <- rnorm(J * n_per_center, mean = 0, sd = 1)

# Paramètres "vrais"
beta0 <- 2
beta1 <- 1
tau   <- 1    # écart-type inter-centre
sigma <- 2    # écart-type inter-individuel

# Effets aléatoires de centre
u <- rnorm(J, mean = 0, sd = tau)
u_center <- u[center]

# Bruits individuels
eps <- rnorm(J * n_per_center, mean = 0, sd = sigma)

# Réponse
y <- beta0 + u_center + beta1 * x + eps

dat <- data.frame(
    y = y,
    x = x,
    center = factor(center)
)

head(dat)
```

```
          y           x center
1 0.6428930 -0.56047565      1
2 1.6002347 -0.23017749      1
3 4.7217458  1.55870831      1
4 1.8798106  0.07050839      1
5 0.2945232  0.12928774      1
6 1.6435736  1.71506499      1
```

6.B Modèle mixte

On ajuste un modèle mixte avec intercept aléatoire de centre : (c'est à dire un effet aléatoire sur l'intercept, qui correspond aux différences de niveau moyen entre centres)

```r
mod_mixed <- lmer(y ~ x + (1 | center), data = dat)
summary(mod_mixed)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | center)
   Data: dat

REML criterion at convergence: 2148.5

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.64638 -0.63464  0.02181  0.67665  2.72271

Random effects:
 Groups   Name        Variance Std.Dev.
 center   (Intercept) 0.5698   0.7548
 Residual             4.1129   2.0280
Number of obs: 500, groups:  center, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.57420    0.25537   6.164
x            1.10000    0.09432  11.663

Correlation of Fixed Effects:
  (Intr)
x -0.013
```

Dans la partie Random effects, on retrouve :

- une estimation de τ^2 (variance inter-centre),
- une estimation de σ^2 (variance résiduelle).

On peut donc en déduire un ICC estimé :

```r
var_comp <- as.data.frame(VarCorr(mod_mixed))
tau2_hat <- var_comp$vcov[var_comp$grp == "center"]
sigma2_hat <- var_comp$vcov[var_comp$grp == "Residual"]

icc_hat <- tau2_hat / (tau2_hat + sigma2_hat)
icc_hat
```

```
[1] 0.1216758
```

6.C Modèles séparés par centre

Si on fait un modèle séparé par centre :

```r
coefs_by_center <- lapply(split(dat, dat$center), function(dd) {
    coef(lm(y ~ x, data = dd))
})

coefs_by_center
```

```
$`1`
(Intercept)           x
   1.626168    1.200121

$`2`
(Intercept)           x
  0.6134143   1.2447135

$`3`
(Intercept)           x
   2.811325    1.029217

$`4`
(Intercept)           x
   2.360334    1.266255

$`5`
(Intercept)           x
   1.064226    1.149654

$`6`
(Intercept)           x
  2.0856707   0.8961968

$`7`
(Intercept)           x
  0.9985427   0.9285005

$`8`
(Intercept)           x
  0.3389266   1.3435537

$`9`
(Intercept)           x
  2.1766631   0.7207045

$`10`
(Intercept)           x
```

```
1.598613     1.330416
```

On obtient :

- une liste d'intercepts et de pentes par centre,
- mais aucun paramètre global pour la variance inter-centre.

La variance empirique des intercepts estimés :

```r
betas0 <- sapply(coefs_by_center, function(b) b[1])
var(betas0)
```

```
[1] 0.6462605
```

mélange :

- la vraie dispersion entre centres,
- et la variabilité d'estimation due au fait que chaque centre a une taille finie (n_j).

Il n'y a pas, ici, d'équivalent direct de τ^2 et de l'ICC estimés proprement par le modèle mixte.
□

6.D Résumé

- La variabilité inter-individuelle correspond à la dispersion entre patients après ajustement sur les covariables (variance σ^2).
- Les variations inter-centre correspondent aux différences de niveau moyen (ou de pente) entre centres (variance τ^2).
- Un modèle linéaire mixte permet de :
  - séparer ces deux composantes τ^2 et σ^2,
  - calculer un ICC $\rho = \tau^2/(\tau^2 + \sigma^2)$,
  - ajuster correctement les erreurs standards des effets fixes.
- Si l'on ajuste des modèles séparés par centre ou par patient :
  - on n'a plus de paramètre global τ^2 (ou ω^2),
  - la variabilité entre coefficients estimés mélange vraie variabilité et bruit d'estimation,
  - on ne peut pas quantifier proprement la variabilité inter-centre / inter-individuelle ni la réinjecter dans un cadre prédictif.

Les modèles mixtes sont justement construits pour modéliser explicitement ces variabilités, au lieu de les laisser "perdues" dans des modèles séparés.

17Effets fixes et aléatoires

---

1 Contexte : pourquoi parler d'effets fixes et aléatoires ?

On travaille avec des données corrélées :

- patients vus plusieurs fois (données longitudinales),
- patients inclus dans différents centres,
- élèves dans des classes, classes dans des écoles, etc.

Schéma classique :

- i indexe les individus (patients) à l'intérieur d'un groupe,
- j indexe les groupes (centres, services, écoles).

On observe une réponse Y_{ij} (par exemple un score, une mesure biologique) et des covariables X_{ij} (âge, sexe, traitement, temps, etc.).

Dans un modèle linéaire simple (sans structure de corrélation explicite) :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}.$$

Problème : les observations ne sont pas indépendantes (mêmes patients, mêmes centres, etc.).

Idée des modèles linéaires mixtes (MLM) :

Combiner des effets fixes (covariables dont on veut estimer l'effet) et des effets aléatoires (variables qui modélisent la structure de corrélation entre les observations).

2 Effets fixes : covariables dont on veut estimer l'effet

2.A Définition intuitive

Un effet fixe, c'est :

- une covariable pour laquelle on veut estimer un effet moyen dans la population,
- avec un coefficient unique valable pour tous les individus (et pour tous les centres) dans le modèle.

Exemples typiques :

- traitement (A vs B),
- âge, sexe,

- dose d'un médicament,
- temps (en jours depuis l'inclusion).

Dans un modèle mixte, on écrit par exemple :

$$Y_{ij} = \beta_0 + \beta_1, \text{traitement}ij + \beta_2, \text{âge}ij + \cdots + \text{(effets aléatoires)} + \varepsilon_{ij}.$$

Les β_k sont des effets fixes :

- β_1 = effet moyen du traitement dans l'ensemble de la population,
- β_2 = effet moyen de l'âge, etc.

On interprète β_1 comme dans une régression classique :

- si Y_{ij} est continu : différence moyenne de Y associée au traitement, après ajustement,
- si on est dans un modèle logistique ou Poisson : log-odds ratio, log-rate ratio, etc.

2.B Ce que signifie « covariables dont on veut estimer l'effet »

Quand on dit :

Effets fixes : covariables dont on veut estimer l'effet,

on veut dire :

- ce sont les variables pour lesquelles l'effet lui-même est l'objet du résultat scientifique,
- on va rapporter ces effets dans un tableau de résultats (estimateur, intervalle de confiance, p-value),
- on suppose que cet effet est le même dans tous les centres / tous les patients, à l'intérieur du modèle.

Exemple :

- on veut savoir si le nouveau traitement réduit la mortalité, en moyenne,
- on met le traitement en effet fixe,
- on interprète $\widehat{\beta}_{\text{traitement}}$ comme l'effet moyen du traitement.

□

3 Effets aléatoires : modéliser la structure de corrélation

3.A Idée de base

Les effets aléatoires servent à modéliser :

- le fait que des observations appartenant au même groupe se ressemblent plus que des observations de groupes différents,
- la variabilité entre groupes (centres, patients, écoles, etc.),

- la corrélation entre observations d'un même groupe.

On écrit par exemple, pour des patients i dans des centres j :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + \varepsilon_{ij},$$

avec :

- u_j = effet aléatoire de centre (spécifique au centre j),
- souvent on suppose $u_j \sim \mathcal{N}(0, \tau^2)$,
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ = erreur résiduelle individuelle.

Ici :

- β_0 et β_1 sont des effets fixes,
- u_j est un effet aléatoire,
- u_j introduit de la corrélation entre les Y_{ij} du même centre.

3.B « Variables qui modélisent la structure de corrélation »

La phrase :

Effets aléatoires : variables qui modélisent la structure de corrélation entre les observations

se traduit par : - on ajoute des termes aléatoires (comme u_j) qui sont partagés par plusieurs observations, - deux observations qui partagent le même u_j sont corrélées, - la manière dont on spécifie ces effets aléatoires (intercepts, pentes aléatoires, structure) détermine la forme de la corrélation.

3.B.1 Exemple: intercept aléatoire de centre $(1 \mid \text{centre})$

Modèle :

$$Y_{ij} = \beta_0 + u_j + \beta_1 X_{ij} + \varepsilon_{ij}.$$

- toutes les observations d'un même centre partagent le même u_j,
- cela crée une corrélation intra-centre,
- $\tau^2 = \text{Var}(u_j)$ mesure la variabilité inter-centre.

3.B.2 Exemple: Intercept aléatoire de patient $(1 \mid \text{patient})$ dans des données longitudinales

Modèle :

$$Y_{ij} = \beta_0 + b_i + \beta_1 t_{ij} + \varepsilon_{ij},$$

où t_{ij} est le temps et b_i un effet aléatoire patient.

- toutes les mesures du même patient partagent le même b_i,
- cela modélise le fait que chaque patient a un niveau moyen propre,
- Var(b_i) donne la variabilité inter-patient.

3.B.3 Exemple: Pente aléatoire $(\textbf{temps} \mid \textbf{patient})$

Modèle :

$$Y_{ij} = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \varepsilon_{ij}.$$

- chaque patient a son intercept propre b_{0i} et sa pente propre b_{1i},
- cela permet de modéliser des trajectoires individuelles différentes dans le temps.

□

4 Résumé conceptuel : ce qui distingue effets fixes et aléatoires

On peut résumer ainsi :

- Effets fixes
  - On estime des coefficients uniques (par exemple β_1 pour le traitement).
  - On s'intéresse directement à la valeur de ces coefficients.
  - On les interprète comme des effets moyens (différence moyenne, log-odds ratio moyen, etc.).
- Effets aléatoires
  - On modélise des variables latentes comme u_j, b_i, propres aux centres ou aux patients.
  - On ne s'intéresse pas à la valeur spécifique de chaque u_j en tant que résultat principal, mais à :
  - la variance de ces effets (par exemple τ^2),
  - la structure de corrélation qu'ils induisent.
- Ils permettent de dire : "Les centres/patients diffèrent entre eux selon une distribution (par exemple normale) avec telle variance".

Les modèles linéaires mixtes :

- combinent des effets fixes (pour les covariables dont on veut l'effet estimé)
- et des effets aléatoires (pour représenter la structure de corrélation et la variabilité entre groupes).

□

5 Exemple simple en R

On simule des données avec :

- $J = 8$ centres,
- $n_j = 40$ patients par centre,
- un effet fixe x (par exemple une covariable continue),
- un effet aléatoire de centre u_j,
- une erreur individuelle ε_{ij}.

```r
J <- 8
n_per_center <- 40
center <- factor(rep(1:J, each = n_per_center))

# Covariable individuelle
x <- rnorm(J * n_per_center, mean = 0, sd = 1)

# Paramètres "vrais"
beta0 <- 2      # effet fixe intercept
beta1 <- 1.5    # effet fixe pour x
tau   <- 1      # écart-type inter-centre
sigma <- 2      # écart-type inter-individuel

# Effet aléatoire de centre
u <- rnorm(J, mean = 0, sd = tau)
u_center <- u[center]

# Bruit individuel
eps <- rnorm(J * n_per_center, mean = 0, sd = sigma)

# Réponse
y <- beta0 + u_center + beta1 * x + eps

dat <- data.frame(
    y = y,
    x = x,
    center = center
)

head(dat)
```

```
            y           x center
1  4.06775203 -0.56047565      1
2  1.90674422 -0.23017749      1
3  6.61061002  1.55870831      1
4 -0.03367521  0.07050839      1
5  0.98815299  0.12928774      1
```

```
6  7.84846318  1.71506499      1
```

On ajuste alors un modèle mixte :

- effet fixe : x,
- effet aléatoire : intercept aléatoire pour le centre.

```r
mod <- lmer(y ~ x + (1 | center), data = dat)
summary(mod)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: y ~ x + (1 | center)
   Data: dat

REML criterion at convergence: 1363.6

Scaled residuals:
     Min       1Q   Median       3Q      Max
-2.85395 -0.64960  0.02815  0.68295  2.71681

Random effects:
 Groups   Name        Variance Std.Dev.
 center   (Intercept) 0.5729   0.7569
 Residual             3.9426   1.9856
Number of obs: 320, groups:  center, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   2.1368     0.2897   7.375
x             1.5949     0.1165  13.692

Correlation of Fixed Effects:
  (Intr)
x -0.012
```

Dans la sortie :

- partie Fixed effects :
  - (Intercept) : estimation de β_0,
  - x : estimation de β_1, effet fixe de la covariable.
- partie Random effects :
  - center : variance estimée de u_j (approximation de τ^2),
  - Residual : variance résiduelle (approximation de σ^2).

On peut extraire les deux types d'effets :

```r
# Effets fixes
fixef(mod)
```

```
(Intercept)           x
   2.136808    1.594914
```

```r
# Effets aléatoires de centre (BLUPs)
ranef(mod)$center
```

```
  (Intercept)
1  0.07338514
2  0.64048205
3 -1.13642609
4  0.81385579
5 -0.59847350
6  0.64985586
7 -0.51545981
8  0.07278056
```

Interprétation :

- fixef(mod) donne les estimations des effets fixes (ce qu'on met dans le tableau de résultats).
- ranef(mod)$center donne, pour chaque centre, l'estimation de u_j (effet aléatoire centre) :
  - ces valeurs traduisent le fait que certains centres sont au-dessus ou en dessous de la moyenne,
  - mais ce n'est pas leur valeur exacte isolée qui est le résultat principal,
- ce qui nous intéresse surtout, c'est la variance de ces effets (la variabilité inter-centre) et la corrélation qu'ils induisent.

□

6 En une phrase

- **Effets fixes** : covariables dont on veut estimer et interpréter l'effet moyen (β), commun à toute la population.
- **Effets aléatoires** : composantes aléatoires (comme u_j, b_i) ajoutées pour représenter la variabilité entre groupes et la corrélation entre observations, via leurs variances (τ^2, etc.).

Les modèles linéaires mixtes sont précisément construits pour combiner les deux, afin de : - obtenir des effets fixes cohérents, - et modéliser correctement la structure de dépendance des données.

18Intercept variances effets_aleatoires

1 Contexte et objectif

On veut expliquer :

1. ce qu'est un intercept (et une pente) dans une régression,
2. ce que fait un modèle linéaire mixte :
    - deux composantes de variance : entre groupes et à l'intérieur des groupes,
3. ce qu'est une matrice de variance–covariance (sans jargon inutile),
4. la différence entre effets aléatoires emboîtés et effets aléatoires croisés.

2 Rappel : droite de régression et « intercept »

2.A Modèle linéaire simple

On commence par un modèle de base avec une seule covariable X :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

- Y_i : résultat pour l'individu i (par exemple : tension artérielle).
- X_i : covariable (par exemple : âge).
- β_0 et β_1 : paramètres du modèle.
- ε_i : part non expliquée (résidu) pour l'individu i.

Graphiquement, c'est une droite sur un nuage de points (X_i, Y_i).

2.B Qu'est-ce que l'« intercept » ?

Dans la formule :

- β_0 est appelé intercept.

Interprétation :

- si on prolonge la droite jusqu'à $X = 0$, la valeur de Y lue sur l'axe vertical est β_0 ;
- donc l'intercept est la valeur moyenne de Y quand $X = 0$ (dans le modèle).

En français : c'est le « point de départ » de la droite.

Représentation sur R :

```r
set.seed(123)
n <- 100
X <- rnorm(n, mean = 50, sd = 10)
beta0 <- 100
beta1 <- 2
epsilon <- rnorm(n, mean = 0, sd = 15)
Y <- beta0 + beta1 * X + epsilon

## 1) Ajuster les deux modèles
mod_avec_int <- lm(Y ~ X)       # avec intercept
mod_sans_int <- lm(Y ~ X - 1)   # sans intercept

coef(mod_avec_int)
```

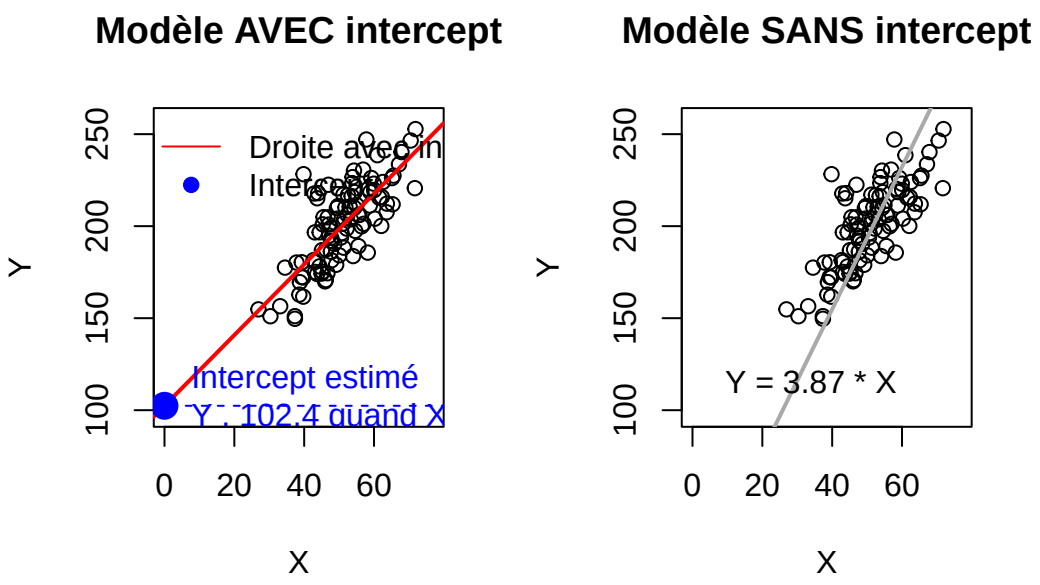
```
(Intercept)           X
 102.393325    1.921293
```

```r
coef(mod_sans_int)
```

```
      X
3.87073
```

```r
# Intercept estimé du modèle avec intercept
beta0_hat <- coef(mod_avec_int)[1]
beta1_hat <- coef(mod_avec_int)[2]

## 2) Limites des axes : on inclut les données ET l'intercept
xlim <- c(0, max(X) + 5)
ylim <- c(min(c(Y, beta0_hat)) - 5, max(Y) + 5)

par(mfrow = c(1, 2))

## --- PLOT 1 : Modèle AVEC intercept (on montre clairement l'intercept) --- ##
plot(
    X, Y,
    xlim = xlim, ylim = ylim,
    main = "Modèle AVEC intercept",
    xlab = "X", ylab = "Y"
)

# Droite de régression ajustée (avec intercept)
abline(mod_avec_int, col = "red", lwd = 2)

# Point de l'intercept estimé : (X = 0, Y = beta0_hat)
points(0, beta0_hat, pch = 19, col = "blue", cex = 1.8)
```

```r
# Ligne horizontale au niveau de l'intercept
abline(h = beta0_hat, col = "blue", lty = 2)

# Petit texte explicatif à côté du point
text(
    x = 0,
    y = beta0_hat,
    labels = paste0("Intercept estimé\nY   ", round(beta0_hat, 1), " quand X =
     ↪  0"),
    pos = 4, offset = 0.7, col = "blue"
)

legend(
    "topleft",
    legend = c("Droite avec intercept", "Intercept"),
    col    = c("red", "blue"),
    lty    = c(1, NA),
    pch    = c(NA, 19),
    bty    = "n"
)

## --- PLOT 2 : Modèle SANS intercept (comparaison) --- ##
plot(
    X, Y,
    xlim = xlim, ylim = ylim,
    main = "Modèle SANS intercept",
    xlab = "X", ylab = "Y"
)

# Droite ajustée SANS intercept (forcée à passer par (0,0))
abline(mod_sans_int, col = "darkgray", lwd = 2)

text(
  x = min(xlim) + diff(xlim) * 0.05,
  y = min(ylim) + diff(ylim) * 0.1,
  labels = paste0("Y = ", round(coef(mod_sans_int)[1], 2), " * X"),
    pos = 4
)
```




```r
par(mfrow = c(1, 1))
```

2.C Qu'est-ce que la pente ?

- β_1 est la pente de la droite.

Interprétation :

- si X augmente de 1 unité, la valeur moyenne de Y augmente (ou diminue) en moyenne de β_1 unités.

C'est l'effet moyen de X sur Y.

□

3 Ajouter des groupes : modèle linéaire mixte simple

Maintenant, les individus sont regroupés :

- par exemple, des patients dans des centres,
- ou des mesures répétées dans des patients.

On indexe :

- i = le groupe (centre),
- j = l'observation dans le groupe i (patient ou visite).

On propose un modèle linéaire mixte simple (intercept aléatoire de groupe) :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}.$$

- Y_{ij} : réponse pour l'observation j dans le groupe i.

- X_{ij} : covariable (âge, temps, etc.).
- β_0, β_1 : effets fixes (comme dans la régression simple).
- u_i : effet aléatoire de groupe i.
- ε_{ij} : erreur résiduelle pour l'observation j du groupe i.

Hypothèses sur les termes aléatoires :

- $u_i \sim \mathcal{N}(0, \sigma_u^2)$: les groupes ont des niveaux différents, mais ces niveaux suivent une loi normale autour de 0.
- $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$: bruit individuel.
- u_i est indépendant des ε_{ij}.

Intuition :

- β_0 : niveau moyen global.
- u_i : décalage spécifique du groupe i par rapport au niveau moyen (certains centres sont « un peu plus hauts », d'autres « un peu plus bas »).
- $\beta_1 X_{ij}$: effet de la covariable.
- ε_{ij} : bruit individuel autour de tout ça.

□

4 Les deux variances : entre groupes et à l'intérieur des groupes

On veut expliciter les 2 variances importantes :

1. variance entre groupes : σ_u^2,
2. variance à l'intérieur des groupes (résiduelle) : σ^2.

4.A Variance d'une observation Y_{ij}

On part du modèle :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}.$$

On considère X_{ij} comme donné (fixe). La partie aléatoire est $u_i + \varepsilon_{ij}$.
On cherche :

$$\text{Var}(Y_{ij} \mid X_{ij}) = \text{Var}(u_i + \varepsilon_{ij}).$$

Comme u_i et ε_{ij} sont indépendants :

$$\text{Var}(u_i + \varepsilon_{ij}) = \text{Var}(u_i) + \text{Var}(\varepsilon_{ij}) = \sigma_u^2 + \sigma^2.$$

Donc :

- σ_u^2 est la part de la variance venant des différences entre groupes,
- σ^2 est la part de la variance venant des différences entre individus à l'intérieur d'un groupe.

La variance totale conditionnelle est :

$$\text{Var}(Y_{ij} \mid X_{ij}) = \sigma_u^2 + \sigma^2.$$

Variance totale conditionelle = variance entre groupes + variance à l'intérieur des groupes.

4.B Corrélation entre deux observations d'un même groupe

Prenons deux observations différentes j et k dans le même groupe i :

- $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}$,
- $Y_{ik} = \beta_0 + \beta_1 X_{ik} + u_i + \varepsilon_{ik}$.

On s'intéresse à :

$$\text{Cov}(Y_{ij}, Y_{ik} \mid X_{ij}, X_{ik}).$$

Les termes déterministes (β_0, $\beta_1 X_{ij}$, $\beta_1 X_{ik}$) n'apportent pas de variance parce que ce sont des constantes.

Il reste :

$$\text{Cov}(u_i + \varepsilon_{ij}, u_i + \varepsilon_{ik}).$$

En développant et en utilisant les indépendances :

- $\text{Cov}(u_i, u_i) = \text{Var}(u_i) = \sigma_u^2$,
- les autres covariances (entre u_i et ε, entre deux ε différents) sont nulles.

Donc :

$$\text{Cov}(Y_{ij}, Y_{ik} \mid X) = \sigma_u^2.$$

La corrélation entre deux observations du même groupe est alors :

$$\text{Corr}(Y_{ij}, Y_{ik} \mid X) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}.$$

Cette quantité est le coefficient de corrélation intraclasse (ICC).
□

5 Matrice de variance–covariance : expliquer simplement

Jusqu'ici, chaque groupe i avait un seul effet aléatoire : u_i (un « niveau » propre au groupe).
Maintenant, imaginons qu'on autorise :

- un niveau propre au groupe,
- et une pente propre au groupe pour une variable (par exemple le temps).

5.A Exemple : intercept et pente aléatoires

On écrit :

$$Y_{ij} = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}), t_{ij} + \varepsilon_{ij}.$$

- t_{ij} : temps (ou autre covariable).
- b_{0i} : décalage du groupe i sur l'axe vertical (niveau propre).
- b_{1i} : décalage de la pente du groupe i (vitesse d'évolution propre).

On suppose :

- b_{0i} et b_{1i} sont aléatoires,
- ils ont une certaine variance chacun,
- et ils peuvent être corrélés (par exemple, les groupes qui partent plus haut vont aussi plus vite).

On peut résumer ça dans un petit tableau 2×2 qu'on appelle matrice de variance–covariance :

$$\Sigma_b = \begin{pmatrix} \text{Var}(b_{0i}) & \text{Cov}(b_{0i}, b_{1i}) \\ \text{Cov}(b_{0i}, b_{1i}) & \text{Var}(b_{1i}) \end{pmatrix}.$$

Concrètement :

- Var(b_{0i}) : à quel point les niveaux des groupes sont dispersés,
- Var(b_{1i}) : à quel point les pentes des groupes sont dispersées,
- Cov(b_{0i}, b_{1i}) :
  - positive : les groupes qui partent plus haut ont aussi des pentes plus grandes,
  - négative : les groupes qui partent plus haut ont des pentes plus faibles.

La « matrice de variance–covariance », c'est juste une façon compacte de stocker :

- plusieurs variances,

- et les covariances entre les effets aléatoires.

□

6 Effets aléatoires emboîtés vs effets aléatoires croisés

6.A Effets aléatoires emboîtés (hierarchiques)

On dit que des effets sont emboîtés (ou imbriqués) quand chaque unité de niveau inférieur appartient à un seul niveau supérieur.
Exemple typique :

- des patients dans des centres,
- chaque patient est suivi dans un seul centre.

Notation :

- i : centre,
- j : patient dans le centre i,
- k : mesure répétée dans le patient j du centre i.

On pourrait écrire un modèle :

$$Y_{ijk} = \beta_0 + u_i + v_{ij} + \varepsilon_{ijk}.$$

- u_i : effet aléatoire de centre (tous les patients du centre i partagent ce u_i),
- v_{ij} : effet aléatoire de patient emboîté dans le centre i,
- ε_{ijk} : bruit résiduel.

Structure :

- patient emboîté dans centre,
- on note parfois « patient dans centre »,
- en R : (1 | centre/patient) ou (1 | centre) + (1 | centre:patient).

Idée :

- on a une première source de variabilité entre centres,
- puis, à l'intérieur d'un centre donné, une variabilité entre patients.

6.B Effets aléatoires croisés

Effets croisés : une unité de niveau inférieur peut être associée à plusieurs modalités de deux facteurs.
Exemples :

1. Patients et médecins :

```
-   un patient peut être vu par plusieurs médecins,

-   un médecin voit plusieurs patients.

-   Patients et médecins sont « croisés », pas hiérarchiques.
```

2. Élèves × enseignants dans un système où les élèves ont plusieurs profs et les profs ont plusieurs classes.

Modèle croisé simple :

- i : patient,
- j : médecin,
- k : consultation.

On pourrait écrire :

$$Y_{ijk} = \beta_0 + u_i + v_j + \varepsilon_{ijk}.$$

- u_i : effet aléatoire du patient i,
- v_j : effet aléatoire du médecin j,
- chaque observation Y_{ijk} reçoit un bout de variabilité propre au patient et un bout propre au médecin.

En R, on écrirait par exemple :

```
lmer(Y ~ 1 + (1 | patient) + (1 | medecin), data = ...)
```

Différence clé :

- emboîté : chaque unité de niveau inférieur n'appartient qu'à un seul niveau supérieur (patient dans un centre).
- croisé : les deux facteurs se combinent librement (un patient peut être vu par plusieurs médecins, un médecin peut voir plusieurs patients).

□

7 Petit exemple R pour illustrer les deux variances

On simule des données avec :

- 10 centres,
- 40 patients par centre,
- un intercept aléatoire de centre,
- une variance résiduelle.

```r
J <- 10
n_per_center <- 40

centre <- factor(rep(1:J, each = n_per_center))
x <- rnorm(J * n_per_center, mean = 0, sd = 1)

beta0 <- 2
beta1 <- 1.5
sigma_u <- 1    # écart-type inter-centre
sigma_e <- 2    # écart-type intra-centre

u <- rnorm(J, mean = 0, sd = sigma_u)
u_centre <- u[centre]

eps <- rnorm(J * n_per_center, mean = 0, sd = sigma_e)

y <- beta0 + beta1 * x + u_centre + eps

dat <- data.frame(
    y = y,
    x = x,
    centre = centre
)

head(dat)
```

```
          y          x centre
1  5.578167  2.1988103      1
2  2.607432  1.3124130      1
3  6.051474 -0.2651451      1
4  3.856798  0.5431941      1
5  4.602392 -0.4143399      1
6 -2.843757 -0.4762469      1
```

On ajuste un modèle mixte :

```r
mod <- lmer(y ~ x + (1 | centre), data = dat, REML = TRUE)
summary(mod)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: y ~ x + (1 | centre)
   Data: dat

REML criterion at convergence: 1731.1

Scaled residuals:
```

```
     Min       1Q   Median       3Q      Max 
-2.69721 -0.65353  0.03108  0.62633  2.71704 

Random effects:
 Groups   Name        Variance Std.Dev.
 centre   (Intercept) 0.3296   0.5741  
 Residual             4.2620   2.0645  
Number of obs: 400, groups:  centre, 10

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept)   1.7175     0.2089   9.0058   8.222 1.77e-05 ***
x             1.3779     0.1060 390.9037  13.004  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
  (Intr)
x -0.019
```

Dans la sortie :

- partie Random effects :
  - centre : estimation de σ_u^2 (variance entre centres),
  - Residual : estimation de σ^2 (variance à l'intérieur des centres).

On peut extraire les variances :

```r
vc <- as.data.frame(VarCorr(mod))
vc
```

```
       grp        var1 var2      vcov     sdcor
1   centre (Intercept) <NA> 0.3296151 0.5741211
2 Residual        <NA> <NA> 4.2619585 2.0644511
```

Et l'ICC :

```r
sigma_u2_hat <- vc$vcov[vc$grp == "centre"]
sigma2_hat   <- vc$vcov[vc$grp == "Residual"]

icc_hat <- sigma_u2_hat / (sigma_u2_hat + sigma2_hat)
icc_hat
```

```
[1] 0.07178695
```

Cela correspond à :

$$\rho = \frac{\sigma_u^2}{\sigma_u^2 + \sigma^2}.$$

□

8 Résumé

1. Intercept :

```
-   paramètre $\beta_0$,

-   valeur moyenne de $Y$ quand $X = 0$,

-   point où la droite coupe l'axe vertical.
```

2. Modèle linéaire mixte simple :

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_i + \varepsilon_{ij}.$$

   - Y_{ij} : observation j dans le groupe i,
   - X_{ij} : covariable,
   - β_0, β_1 : effets fixes,
   - u_i : effet aléatoire de groupe, variance σ_u^2 (entre groupes),
   - ε_{ij} : bruit résiduel, variance σ^2 (dans les groupes).
3. Deux variances :
   - variance entre groupes : σ_u^2,
   - variance à l'intérieur des groupes : σ^2,
   - variance totale : $\sigma_u^2 + \sigma^2$,
   - ICC : $\rho = \sigma_u^2/(\sigma_u^2 + \sigma^2)$.
4. Matrice variance–covariance :
   - sert à stocker les variances et covariances de plusieurs effets aléatoires (par exemple niveau et pente d'un même groupe),
   - c'est juste un petit tableau 2×2 avec Var et Cov.
5. Effets aléatoires :
   - emboîtés : patients dans centres (chaque patient dans un seul centre),
   - croisés : patient et médecin (un patient peut voir plusieurs médecins, un médecin voit plusieurs patients).

19Wrap Up Régressions

1 Régressions linéaires multiples

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$

- Y : variable **dépendante** (quantitative continue) : par exemple la pression artérielle
- X1, X2, …, Xk : variables **indépendantes** :
  - peuvent être quantitatives continues (âge, IMC, etc.)
    - dans le cas de l'âge : ça s'exprime "de combien augmente la valeur de Y pour une augmentation d'une unité d'âge (1 an)"
  - peuvent être binaires
  - pour les variables catégorielles ordonnées : elles sont transformées en variables quantitatives !
- ε : terme d'erreur

Et donc : β1 correspond à la différence du facteur X1 (par ex âge) entre sujets de facteur X2 constant (par ex sexe)

On peut tester :

- H0 : β1 = 0 (pas d'association entre X1 et Y)
- H1 : β1 ≠ 0 (association entre X1 et Y)

Interprétation d'un coefficient βi de régression :

- Si Xi est une variable quantitative continue : βi correspond à la variation moyenne de Y associée à une augmentation d'une unité de Xi, toutes les autres variables Xj (j ≠ i) étant maintenues constantes.
- Si Xi est une variable binaire (0/1) : βi correspond à la différence moyenne de Y entre les sujets avec Xi = 1 et les sujets avec Xi = 0, toutes les autres variables Xj (j ≠ i) étant maintenues constantes.

1.A Conditions de validité :

Du test β1 = 0 :

- ε soit un bruit d**e variance constante** et **suivant une loi normale**
- donc idem qu'un test t !!
- pour vérifier : histogramme des résidus !

Définition des résidus :

$$\varepsilon_i = Y_i - \hat{Y}_i$$

- pour chaque individu i, le modèle prédit une valeur Ŷi (ce que le modèle estime pour Yi)
- la valeur Yi correspond à la valeur réelle observée pour l'individu i

- Le résidu pour l'individu i est : $\varepsilon_i = Y_i - \hat{Y}_i$
- Le résidu mesure **l'erreur du modèle** pour l'individu i.
- Si proche de 0 : le modèle prédit bien la valeur observée
- Si grand (positif ou négatif) : le modèle prédit mal la valeur observée
- Pour vérifier les conditions de validité :
  - tracer les résidus en fonction des valeurs prédites $\hat{Y}_i$
  - vérifier la normalité des résidus (histogramme, Q-Q plot)

hétéroscédasticité : variance des résidus non constante

1.B Corrélatons entre variables explicatives

$$Y = \beta_0 + \beta_1(X_1X_2) + \beta_3X_3 + \cdots + \beta_kX_k + \varepsilon$$

- Par ex β1 : âge, β2 : IMC.
- Effet **multiplicatif** (par interaction : encore + que additif)
- Recherche une **synergie** entre les variables explicatives

Synergie = non linéarité, donc c'est rare de mettre en évidence une non-linéarité en médecine.

Mais en vrai, les FDR multipliés se potentialisent !

ATTENTION : si on met des interactions dans le modèle, on ne peut plus interpréter les coefficients βi de façon isolée en dehors du facteur β1 !

Le mieux :

- D'abord faire un modèle sans interaction
- Puis faire un modèle avec interaction

Mais pour les esthètes de la statistiques : le modèle avec interaction ne peut pas vraiment être utilisé car les résidus sont multipliés, donc c'est pas des vrais résidus. Bruno dit que ce n'est pas grave.

1.B.1 Exemple d'interprétation avec interaction

$$Y = \beta_0 + \beta_1 \text{ retraite} + \beta_2 \text{ surpoids} + \beta_3(\text{âge} \times \text{IMC}) + \varepsilon$$

	OR et p value	Interprétation
β1		**Ininterprétable**
β2		**Ininterprétable**
β3		OUI

Mais si c'est codé en -1 / 1 (pour FALSE - TRUE) : Alors **β1 et β2 pourraient être interprétable** !

- **β1** est un effet de l'âge dans une population où il y aurait autant de surpoids que de pas de surpoids

- **β2** est un effet du surpoids dans une population où il y aurait autant de vieux que de jeunes
- **β3** ne change pas que ce soit en -1 / 1 ou 0 / 1 : reste l'effet d'interaction entre âge et IMC
- NB : si variable quantitative : ne pas le faire

Pour ça dans R, il faut changer le contraste avec la fonction `contr.sum()` Documentation R

1.B.2 Si X1 était catégorielle à plus de 2 modalités ?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

- Y : Tension artérielle
- X1 : surpoids
- X2 : profession (catégorielle à 4 modalités : cadre, intermédiaire, employé, ouvrier) : variable pas quantitative, pas ordonnée, pas binaire.

Il faut recoder X2 en variables binaires (variables indicatrices) (besoin d'en faire que 3 variables binaires pour 4 modalités) :

Profession	X2_cadre	X2_intermédiaire	X2_employé
Cadre	1	0	0
Intermédiaire	0	1	0
Employé	0	0	1
Ouvrier	0	0	0

Donc le modèle devient :

$$Y = \beta_0 + \beta_1 \text{ surpoids} + \beta_2\, X2_{\text{cadre}} + \beta_3\, X2_{\text{intermédiaire}} + \beta_4\, X2_{\text{employé}} + \varepsilon$$

Et on peut même résumer : β2 est la différence moyenne de tension artérielle entre les cadres et les ouvriers (catégorie de référence), toutes les autres variables Xj (j ≠ 2) étant maintenues constantes.

Pour choisir la profession de référence : (fonction `relevel()` dans R)

- soit la catégorie la plus fréquente
- soit la catégorie la plus "neutre"
- mais le logiciel en choisit une par défaut (souvent la première par ordre alphabétique)

Pour éviter d'avoir une catégorie de référence :

- Exemple : comparaison inter-hôpitaux
- Utiliser des **contrastes**

Par exemple comparaison Lariboisière, Pompidou, Bichat, Beaujon

tableau :

Hôpital	X2_Lariboisière	X2_Pompidou	X2_Bichat
Lariboisière	1	0	0
Pompidou	0	1	0
Bichat	0	0	1
Beaujon	0	0	0

On crée des variables indicatrices mais on les code différemment (avec des -1 et des 1) :

Hôpital	X2_Lariboisière	X2_Pompidou	X2_Bichat	X2_Beaujon
Lariboisière	-1	0	0	1
Pompidou	0	-1	0	1
Bichat	0	0	-1	1
Beaujon	1	1	1	1

Dans ce cas, X2 est la différence de TAmoyenne en comparaison à l'effet moyen de tous les hôpitaux.

Si nécessaire : vidéo à ≈ 30 minutes

1.C Réponse aux questions

Modèle : Y = a0 + a1(X1*X2)

Dans ce cas : interprétation difficile !!

Le mieux est de faire Y = a0 + a1X1 + a2X2 + a3(X1*X2)

2 Régression logistique

Modélisation d'une variable dépendante binaire (0/1).

$$\text{log}\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

avec :

- p : probabilité que maladie = 1 (événement)
- (1-p) : probabilité que maladie = 0 (non-événement)
- log(p/(1-p)) : logit de p

Sinon idem que pour la régression linéaire multiple mais pas de bruit ε (car variable dépendante binaire et on modélise la probabilité p).

Ici :

- Y = maladie (0/1) pour HTA oui / non
- β_1 : effet du surpoids sur l'HTA ou non
  - et $exp(\beta_1)$ = OR ajusté sur β_2 de β_1 par rapport à l'HTA.
  - c'est à dire : si le surpoids était rare (donc OR ≈ RR) et $exp(\beta_1) = 2$, la probabilité d'avoir une HTA si l'on est en surpoids = 2.

2.A Interactions

Exactement de la même manière qu'en régression linéaire ! et possible de recoder en -1 / 1 pour pas avoir de variable de référence.

2.B Conditions de validité de la régression logistique

- Taille de l'échantillon suffisante (10 événements par variable explicative au minimum)

2.C Autres trucs :

- Test de calibration / Hosmer-Lemeshow : accepter l'hypothèse nulle alors qu'on a pas la puissance !! Donc en théorie on ne peut pas vraiment faire ce test

3 Références

Vidéo Falissard :

- Blog larmarange
- Webin-R

20Effet centre

1 Introduction

Les mesures répétées désignent des situations où plusieurs observations sont recueillies sur une même unité statistique (souvent un patient) au cours du temps, ou lorsque les observations sont regroupées au sein de clusters (par exemple, des patients au sein d'un même hôpital ou des membres d'une même famille).

Le défi majeur posé par ces données réside dans la non-indépendance des observations.

En effet, les mesures effectuées sur un même patient à différents moments sont généralement corrélées entre elles.

Ignorer cette corrélation en utilisant des méthodes statistiques classiques (comme la régression linéaire standard ou l'ANOVA classique) viole l'hypothèse d'indépendance des résidus, conduisant à des estimations biaisées de la précision (intervalles de confiance trop étroits) et à une augmentation du risque d'erreur de type I.

Si l'analyse est relativement simple avec seulement deux points de mesure (analyse de l'évolution $Y_{aprs} - Y_{avant}$ ou ajustement sur la valeur basale), elle devient plus complexe avec trois mesures ou plus.

2 Modèle linéaire

Imaginons qu'on corrèle la durée d'entretien (en minutes) Y avec le niveau de dépression X.

Y est supposé normalement distribué.

X est supposé quantitatif.

Corrélation : peut être intra- ou inter-centre

- Intra-centre : rechercher si un sujet avec un niveau de dépression X élevé a une durée d'entretien Y supérieure à celle d'un autre sujet du même centre présentant un niveau de dépression plus bas.
- Inter-centre : la durée moyenne des entretiens $\widehat{Y}$ réalisés dans les prisons avec un haut niveau de dépression est-elle supérieure ou inférieure à la durée moyenne des entretiens réalisés dans les autres prisons ?

2.A Analyse naïve sans prise en compte du centre

Imaginons qu'on néglige l'effet centre et qu'on réalise une régression linéaire simple de Y en fonction de X..

On prend 3 centres comprenanant chacun 9 détenus.

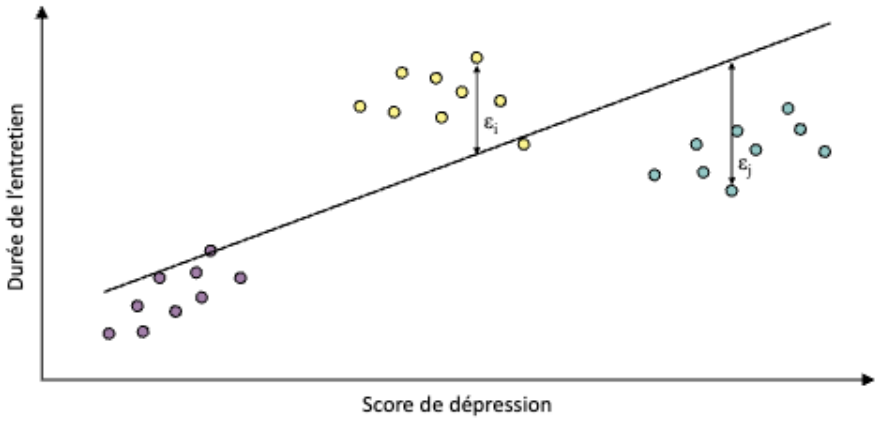


Figure 20.1: Estimation d'une corrélation entre « durée de l'entretien » et «score de dépression » sans tenir compte de l'effet centre (ici un centre par couleur). La droite de régression minimise la somme des ε_i^2 et néglige l'effet centre.

Ici :

- Effet centre particulièrement important
- Chaque centre semble associé à un niveau spécifique de dépression $\hat{X}$
- Et à un niveau moyen spécifique de durée d'entretien $\hat{Y}$

2.A.1 Problèmes posés par cette analyse naïve

= biais de confusion par effet de groupe + non-indépendance des résidus

Problème 1 : Biais de confusion par effet de groupe.

- Si les centres pour lesquels le niveau de dépression est le plus élevé sont aussi ceux où la durée moyenne des entretiens est la plus longue, la pente de régression est **biaisée**.
- Elle mélange l'effet individuel (la relation durée-dépression pour un détenu) et l'effet du groupe. Le modèle ne peut distinguer si la dépression est liée à la durée de l'entretien ou à d'autres facteurs propres au centre (personnel, type de population, etc.).
- *Conséquence :* Le coefficient de régression estimé est faussé et ne représente pas la véritable association au niveau individuel.

Problème 2 : Non-indépendance des résidus.

- Les résidus ε_i ne sont pas indépendants. Les observations au sein d'un même groupe (centre) se ressemblent plus qu'avec celles d'autres groupes.
- C'est la **corrélation intra-classe**.
- *Conséquence :* Ce problème ne biaise pas l'estimation de la pente, mais il conduit à une **sous-estimation de son erreur-standard**.
- *Impact :* L'intervalle de confiance est artificiellement étroit et la p-value trop petite. Le risque est de conclure à tort à un effet significatif (Erreur de Type I).

- **Solutions : Bootstrap par grappe ou estimateur sandwich**
  - **Le Bootstrap par grappe (Cluster Bootstrap) :**
    - *Principe :*
      - Puisque les individus d'un groupe ne sont pas indépendants, on ré-échantillonne les **groupes** (les centres) avec remise.
      - Pour chaque groupe tiré, on inclut tous les individus qu'il contient.
    - *Résultat :*
      - On obtient une distribution de 1000 coefficients qui reflète l'incertitude liée à la variabilité *entre les groupes*. L'IC à 95% est construit à partir de cette distribution (ex: via les percentiles).
      - **Cet intervalle sera presque toujours plus large** que l'IC naïf, reflétant une estimation plus honnête de l'incertitude.

  - **L'estimateur "Sandwich" (Estimateur Robuste de la Variance) :**
    - **Le problème :**
      - Le modèle classique est "trop confiant". Il pense que chaque ligne de données apporte une information unique.
      - Or, si les patients d'un même centre se ressemblent, on a moins d'information réelle qu'on ne le croit.
      - L'erreur-standard calculée classiquement est donc trop petite.
    - **Solution** :
      - Au lieu de se fier uniquement à la théorie (qui suppose l'indépendance), l'estimateur Sandwich regarde les résidus réels (les erreurs du modèle).
      - Si, dans un centre, tous les résidus vont dans le même sens (ex: le modèle se trompe toujours par excès pour ce centre), l'estimateur détecte cette corrélation. Il utilise cette "réalité du terrain" pour corriger mathématiquement la variance à la hausse.
    - **Principe :** La formule mathématique de la variance robuste se compose de trois blocs multipliés entre eux : $A \times B \times A$.
    - On l'appelle "Sandwich" uniquement parce que la correction (B) est coincée entre deux blocs identiques (A), comme une tranche de jambon entre deux tranches de pain.
      1. **Le bloc A (La Théorie) :** C'est la variance calculée par le modèle classique. Elle suppose que tout est parfait (indépendance).
      2. **Le bloc B (La Réalité) :** C'est une correction calculée directement à partir des **données brutes** (les résidus). Si les erreurs sont corrélées dans les groupes, ce bloc B va "gonfler" la valeur.
      3. **Le calcul :** On multiplie $\texttt{Théorie} \times \texttt{Correction} \times \texttt{Théorie}$.
    - ***Résultat :*** Les estimations des coefficients (la pente) ne changent pas, mais les

intervalles de confiance s'élargissent et les p-values augmentent, reflétant une incertitude plus honnête.

2.A.2 Exemple R avec jeu de données fictif

1. Génération des données

```
set.seed(1)
# génération des Xi et Yi pour 5 temps de mesure dans 200 centres
x1 <- runif(200)*0.3 # Génération des xi pour le centre j
x2 <- runif(200)*0.3
x3 <- runif(200)*0.3
x4 <- runif(200)*0.3
x5 <- runif(200)*0.3
a <- rnorm(200)*100+300 # Les coefficiens de la régression
b <- rnorm(200)*100+150 # sont fonction du centre
y1 <- a+b*x1+rnorm(200)*72.5
y2 <- a+b*x2+rnorm(200)*72.5
y3 <- a+b*x3+rnorm(200)*72.5
y4 <- a+b*x4+rnorm(200)*72.5
y5 <- a+b*x5+rnorm(200)*72.5

# dt est le fichier "large" (une ligne par centre)
dt <- data.frame(1:200,x1,y1,x2,y2,x3,y3,x4,y4,x5,y5)
names(dt)[1] <- "centre"

# conversion du fichier "large" en fichier "long"
dtl <- reshape(dt,idvar="centre",varying=2:11,v.names=c("x","y"),timevar =
↪  "temps",direction="long")
dtl$centre <- as.factor(dtl$centre)
```

2. Régression linéaire simple sans prise en compte de l'effet centre

```
mod.lm <- lm(y ~ x, data=dtl)
summary(mod.lm)$coefficients
```

```
            Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 286.6649   8.232679 34.820363 1.591602e-174
x           153.6712  47.571204  3.230341  1.276828e-03
```

Ici, le calcul de l'erreur-standard (l'écart-type) et de la p-value repose sur l'hypothèse que les résidus sont indépendants, ce qui n'est pas le cas.

Impossible de le savoir si on n'a pas d'information sur les données !!

Attention parce que les résidus semblent "normaux" (cf. graphique ci-dessous), cela ne garantit pas leur indépendance.

```
par(mfrow=c(1,2))
hist(residuals(mod.lm),main="Histogramme des résidus",xlab="Résidus")
qqnorm(residuals(mod.lm),main="Q-Q plot des résidus")
qqline(residuals(mod.lm))
```

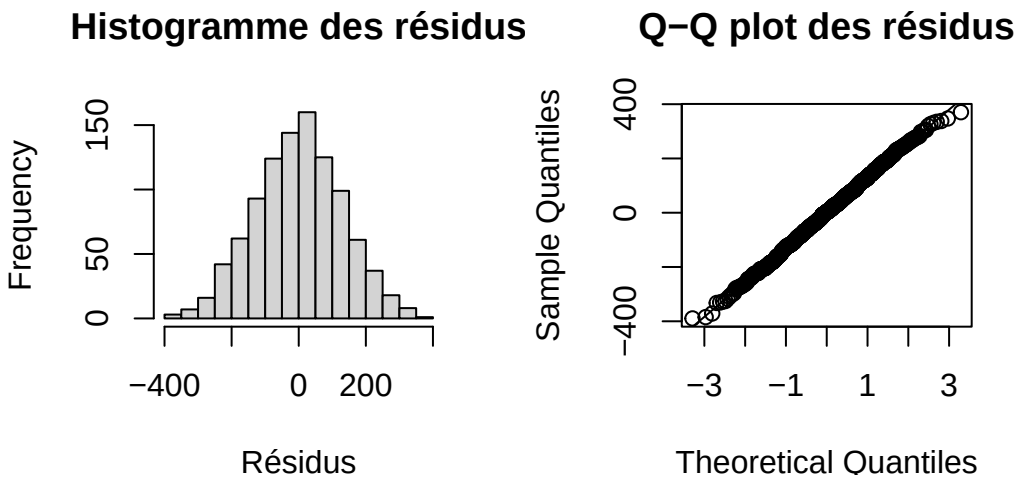


Figure 20.2: Graphique des résidus du modèle linéaire simple sans prise en compte de l'effet centre

Bootsrap

Il est possible de recourir à un bootstrap pour contourner cette limite, en utilisant la fonction `clusbootglm()` de la bibliothèque `ClusterBootstrap`.

```
set.seed(1)
library("ClusterBootstrap")
mod.clusboot <- clusbootglm(y~x,data=dtl,clusterid=centre)
```

```
$boot.coefs :
(Intercept)         x
286.6787            153.6561
$boot.sds
(Intercept)         x
11.03841            50.14656
```

L'écart type de b = 50,14 (par rapport à 47,57 naïf). Les coefficients ne changent pas.

Estimateur sandwich

```
library("sandwich")
sqrt(diag(vcovCL(lm(y~x,data=dtl),cluster=~centre))["x"])
```

```
        x
50.73999
```

L'estimation de l'erreur type (50.74) est très proche de celle du bootstrap, prenant toujours en compte la dépendance des observations d'un même centre.

2.B Analyse avec prise en compte de l'effet centre

il s'agit ici de calculer les moyennes de X et Y pour chaque centre puis d'en faire une régression linéaire

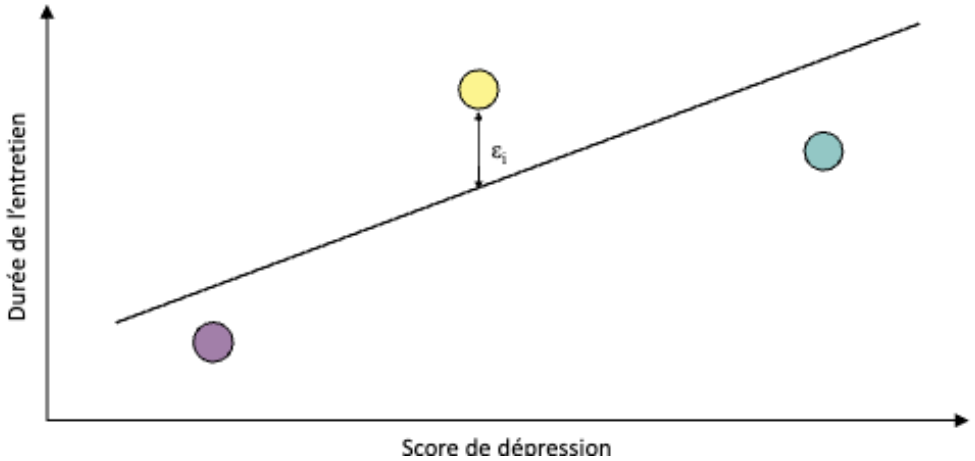


Figure 20.3: Estimation d'une pente inter-centres à partir des moyennes de X et Y pour chaque centre.

2.B.1 Exemple R

```
#moyenne de Y et X par centre
ymeans = tapply(dtl$y, dtl$centre, mean)
xmeans = tapply(dtl$x, dtl$centre, mean)

#régression linéaire des moyennes
summary(lm(ymeans ~ xmeans))
```

```
Call:
lm(formula = ymeans ~ xmeans)

Residuals:
    Min      1Q  Median      3Q     Max 
-309.22  -75.94   -4.72   73.62  308.89 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    268.8       29.6   9.080   <2e-16 ***
```

```
xmeans         272.9      190.4   1.434    0.153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 111.4 on 198 degrees of freedom
Multiple R-squared:  0.01027,   Adjusted R-squared:  0.005276
F-statistic: 2.055 on 1 and 198 DF,  p-value: 0.1532
```

`xmeans` : fournit une estimation de la pente inter-centres.

Fournit une estimation de la pente inter-centres, avec une erreur type estimée en prenant en compte l'indépendance des centres.

Cela indique comment la moyenne de Y varie d'un centre à l'autre en fonction de la moyenne de X du centre. L'erreur type est ici estimée correctement car on travaille sur les centres (qui sont indépendants entre eux), et non sur les observations répétées.

2.C Analyse intra-centres (modèles conditionnels, mixtes ou non)

3 méthodes :

1. Modèle linéaire avec effet fixe par centre (chaque centre)
2. Modèle linéaire avec pente commune (intercept variable par centre)
3. Modèle mixte avec effet aléatoire de centre (c'est à dire que les centres sont vus comme un échantillon aléatoire d'une population plus large de centres possibles)

2.C.1 Modèle linéaire avec effet fixe par centre

Pour évaluer la corrélation intra-centre entre X et Y, on peut faire une régression linéaire dans chaque centre puis calculer la moyenne des pentes obtenues.

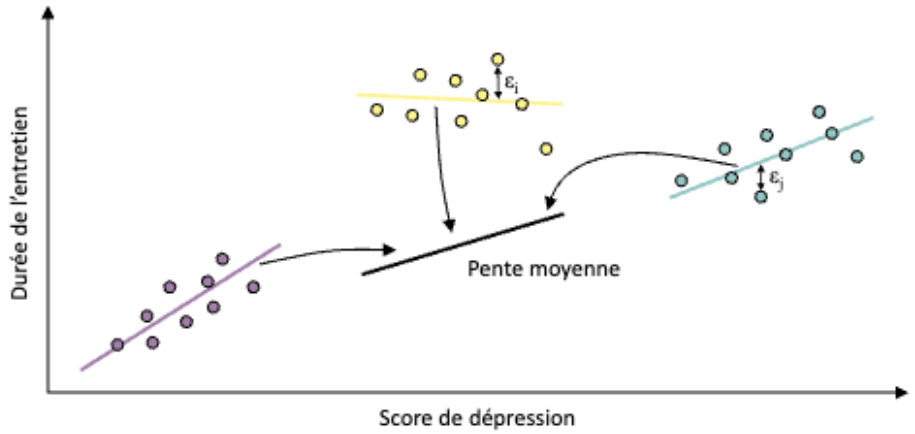


Figure 20.4: Approche conditionnelle artisanale pour laquelle une régression est réalisée dans chaque centre, une pente moyenne étant calculée dans un second temps.

C'est une approche "conditionnelle" car chaque pente est calculée "conditionnellement" à un centre donné.

L'équation serait :

$$Y_{ij} = a + [a.centre_j] + bx_{ij} + [b.centre_j]x_{ij} + \varepsilon_{ij}$$

- i : individu
- j : centre
- a : intercept global
- b : pente globale
- $[a.centre_j]$: déviation de l'intercept pour le centre j
- $[b.centre_j]$: déviation de la pente pour le centre j

Globalement : ça revient à faire une régression linéaire avec des interactions entre X et le centre.

Donc chaque centre a sa propre droite de régression (intercept et pente différents).

En R simplement : `lm(y ~ x * centre)`.

2.C.2 Modèle linéaire avec pente commune (intercept variable par centre)

Variante plus simple : pente commune

Modèle proposé :

$$y_{ij} = a + [a.\text{centre}_j] + b\,x_{ij} + \varepsilon_{ij}$$

Ici :

- chaque centre j a son intercept propre : $a + [a.\text{centre}_j]$
- mais la pente b est identique dans tous les centres

Graphiquement : toutes les droites sont parallèles (même pente) mais décalées verticalement (intercepts différents).

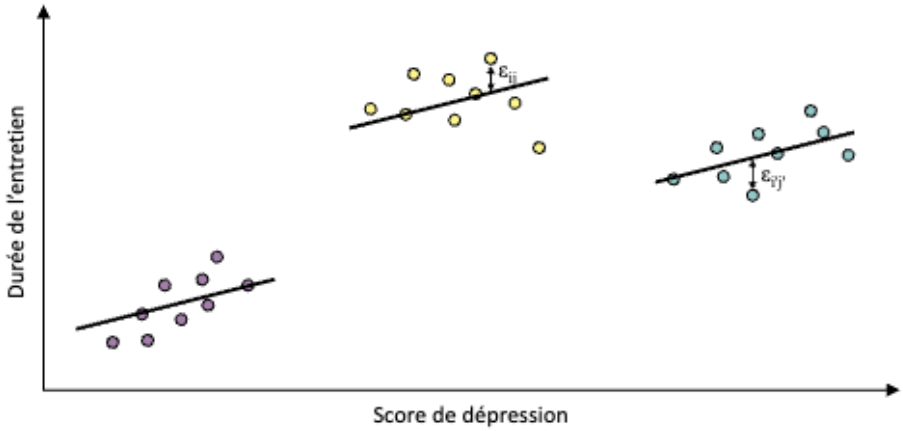


Figure 20.5: Approche conditionnelle avec pente commune, elle correspond au modèle : $y_{ij} = a + [a.\text{centre}_j] + b\,x_{ij} + \varepsilon_{ij}$

En R : `lm(y \~ x + centre)`

Avec 2 trucs importants à savoir sur l'interprétation de b :

- b : sorte de moyenne des pentes centre par centre, pondérée par la variance de x dans chaque centre.
  - Donc lié aux pentes que l'on obtiendrait si on faisait une régression séparée dans chaque centre, puis qu'on faisait une moyenne pondérée.
- erreur type de b : correcte uniquement si vraiment toutes les pentes sont égales entre centres.
  - Si en réalité les pentes diffèrent un peu entre centres, mais que le modèle force une pente commune, alors :
    * b reste un estimateur moyen
    * mais l'incertitude autour de b est mal évaluée d'où la suggestion d'utiliser un estimateur sandwich (variance robuste) ou le bootstrap.

Résumé :

- Modèle plus simple, mais repose sur l'hypothèse forte : "même pente partout".
- Si cette hypothèse est fausse, le b affiché reste "une pente moyenne", mais son écart-type est trop optimiste.

2.C.3 Modèle mixte avec effet aléatoire de centre

Utile surtout si beaucoup de centres.

Passage aux modèles "mixtes" : centre comme effet aléatoire

Idée : plutôt que mettre une variable catégorielle [centre] avec 10 000 modalités, on introduit une variable aléatoire (centre).

Donc la variable catégorielle [centre] devient une variable aléatoire gaussienne (centre).

Nouveau modèle :

$$y_ij = a + (a_\text{centre} * j) + b, x * ij + \varepsilon_ij$$

Ou :

- $(a_\text{centre} * j)$: effet aléatoire pour le centre j, c'est à dire une variable aléatoire qui suit une distribution normale et qui modélise la variabilité des intercepts entre centres.
- b : pente commune à tous les centres.
- ε_ij : erreur résiduelle pour l'individu i dans le centre j.

Les effets sont "mixtes" car il y a à la fois des effets fixes (a, b) et des effets aléatoires (a_{centre*j}).

Ca aide car :

- Pas besoin de 9999 variables binaire (dummy) pour les centres.
- Processus aléatoire pour les centres.

Il faut le faire surtout s'il y a plus de 5 centres.

2.C.4 Exemple R

Calcul des pentes de régressions de Y en fonction de X dans chaque centre

```
# fonction calculant la pente pour un centre donné
pente_intra=function(centre) {coef(lm(data=dtl[dtl$centre==centre,], y~x))["x"]}
# application de la fonction à tous les centres
pentes = sapply(levels(dtl$centre), pente_intra)

# affichage de la pente moyenne et de son erreur standard
cat("La pente moyenne est de", round(mean(pentes), 2),
    "avec une erreur standard de", round(sd(pentes) / sqrt(length(pentes)), 2),
    ↪  "\n")
```

```
La pente moyenne est de 131.66 avec une erreur standard de 41.6
```

1. Modèle avec effet fixe par centre $Y_{ij} = a + [a.centre_j] + bx_{ij} + [b.centre_j]x_{ij} + \varepsilon_{ij}$

C'est surtout le paramètre b qui nous intéresse.

On utilise `contr.sum` pour que le coefficient `x` corresponde à la pente MOYENNE de tous les centres (sinon ce serait la pente du centre de référence). Le "point zéro" n'est plus le centre 1 mais la moyenne des centres.

```
dtlbis <- dtl
contrasts(dtlbis$centre) <- contr.sum
# estimation du modèle avec effet fixe par centre
mod.lmci <- lm(y~x*centre, data=dtlbis)
# affichage dans un cat()
cat("La pente intra-centres estimée est de",
 ↪  round(summary(mod.lmci)$coefficients["x","Estimate"],2),
    "avec une erreur standard de", round(summary(mod.lmci)$coefficients["x","Std.
    ↪  Error"],2), "\n")
```

```
La pente intra-centres estimée est de 131.66 avec une erreur standard de 39.47
```

2. Modèle avec pente commune $Y_{ij} = a + [a.\text{centre}_j] + b\,x_{ij} + \varepsilon_{ij}$

```
mod.lmc <- lm(y~x+centre, data=dtl)
cat("La pente intra-centres estimée est de",
 ↪  round(summary(mod.lmc)$coefficients["x","Estimate"],2),
    "avec une erreur standard de", round(summary(mod.lmc)$coefficients["x","Std.
    ↪  Error"],2), "\n")
```

```
La pente intra-centres estimée est de 118.26 avec une erreur standard de 31.51
```

3. Modèle mixte avec effet aléatoire de centre $Y_{ij} = a + (a_{\text{centre}*j}) + b, x * ij + \varepsilon_{ij}$

Il faut utiliser la fonction `lmer()` de la bibliothèque `lme4` ou `nlme`.

Syntaxe : `(1|centre)` signifie qu'on modélise un intercept aléatoire par centre.

```r
library(nlme)
mod.lmer1 <- lmer(y~x+(1|centre), data=dtl)
summary(mod.lmer1)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: y ~ x + (1 | centre)
   Data: dtl

REML criterion at convergence: 11949.7

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.1811 -0.6001  0.0126  0.6147  2.7906

Random effects:
 Groups   Name        Variance Std.Dev.
 centre   (Intercept) 11242    106.03
 Residual              5723     75.65
Number of obs: 1000, groups:  centre, 200

Fixed effects:
            Estimate Std. Error      df t value Pr(>|t|)
(Intercept)  291.354      9.146 346.352  31.857  < 2e-16 ***
x            122.391     31.081 840.929   3.938 8.91e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
  (Intr)
x -0.509
```

La méthode ici utilisée est la **vraisemblance restreinte (REML)**, dont le principe est d'estimer les paramètres de variance en maximisant la vraisemblance des résidus (c'est à dire leur cohérence avec les données observées), ce qui est préférable pour estimer les paramètres de variance des effets aléatoires.

L'effet centre et le résidus sont donc caractérisés par des variances estimées (du fait qu'on les considère comme des variables aléatoires).

La variance de l'effet centre est donc estimée sensiblement plus importante que le bruit présent dans le modèle (résidus).

L'estimation de la pente vaut 122,39, légèrement différente des modèles précédents.

Pour interpréter la pente d'un modèle à effets mixtes :

- Si la variance intra-centre de X est très faible, la pente fixe du modèle mixte s'interprète comme une pente inter-centre.
- Si la variance inter-centre de X est faible (donc faible effet centre), la pente fixe s'interprète comme une pente intra-centre.

4. Modèle mixte avec pente aléatoire par centre $Y_{ij} = a + (a_{\text{centre}*j}) + (b_{\text{centre}*j})x_{ij} + \varepsilon_{ij}$

2.C.5 Conditions de validités des modèles mixtes

- La pente inter-centres est égale à la moyenne des pentes intra-centres (c'est à dire que les pentes des centres ne diffèrent pas trop entre elles) ;
- L'indépendance des résidus, ce qui peut poser des problèmes lorsque la structurede corrélation intra-centre est complexe, notamment lorsqu'elle est susceptible d'être négative, car les modèles mixtes classiques ne permettent pas de modéliser des corrélations négatives entre les observations d'un même centre ;
- Contribution des plus gros centres supérieure à celle des petits centres. Ca pose problème uniquement si la taille du centre est corrélée à la variable Y.
- Indépendance, normalité, homoscédasticité et nullité moyenne des effets aléatoires ; qu'il s'agisse des ordonnées à l'origine spécifiques à chacun des centres ou de leurs pentes ;

2.D Modèle marginal = GEE (Generalized Estimating Equations)

Y : score de dépression, X durée de l'entretien, patients regroupés par centre et il existe une corrélation intra-centre.

Approche conditionnelle (modèles mixtes, etc.) : "Quel est l'effet de X dans un centre donné ?"

Approche marginale : "Quel est l'effet de X en moyenne dans la population, tous centres confondus ?"

$\rightarrow$ on ne modélise pas les centres un par un ; on s'intéresse à la moyenne globale et on corrige seulement la corrélation.

2.D.1 Définition

Dans l'approche conditionnelle : on réalise une série de modèles au sein même de chaque centre (on regarde ce qui se passe "à l'intérieur" des centres).

Dans l'approche marginale : on écrit un modèle très simple : en gros, comme lm(y~x), mais on remplace l'estimateur classique par un autre estimateur, qui corrige pour la corrélation intra-centre.

Il faut faire l'hypothèse d'égalité des pentes inter-centres et intra-centre : la pente de la relation X - Y est la même entre les centres (si on compare des centres entre eux) et à l'intérieur des centres (si on regarde les patients d'un même centre).

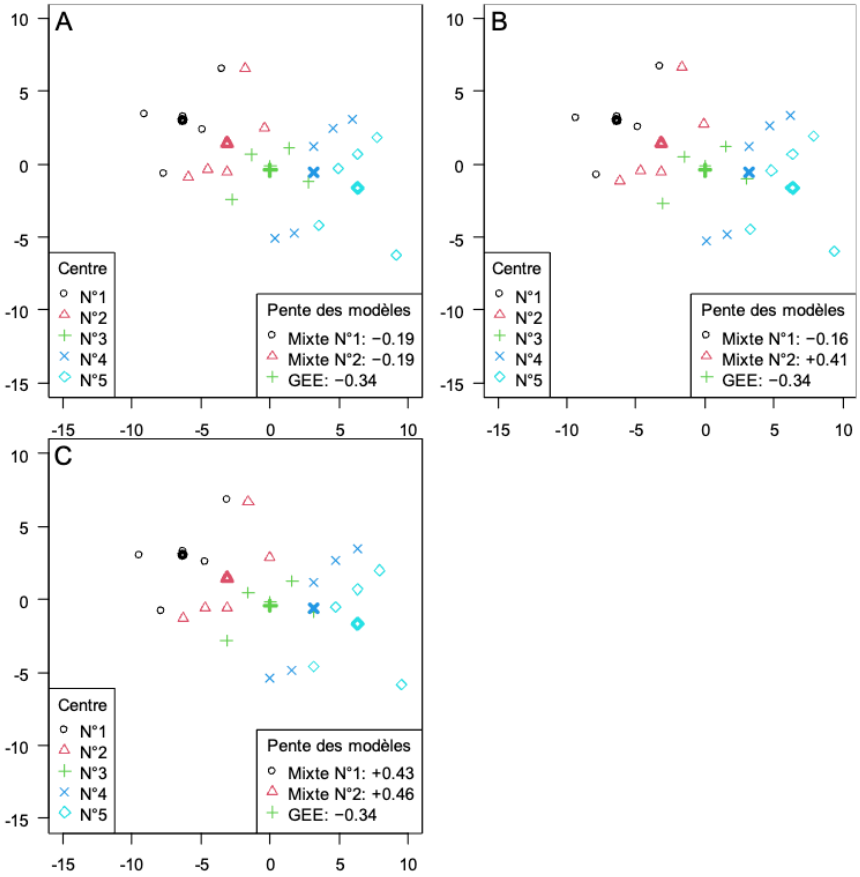
Sous cette hypothèse, l'estimation GEE de la pente est plus efficace (variance plus faible) que l'estimation naïve de `lm(y ~ x)`, c'est à dire même moyenne mais moins de variance.

2.D.2 Principe des GEE

1. D'abord estimation d'à quel point les observations d'un même centre sont corrélées entre elles (corrélation intra-centre).
2. Ensuite, GEE utilise cette information pour ajuster la manière de calculer les coefficients du modèle linéaire

Si la corrélation intra-centre est faible, GEE donne des résultats très proches de `lm(y ~ x)`.

Si la corrélation intra-centre est forte, GEE fournit une estimation plus précise de la pente, mais l'interprétation de cette pente devient plus complexe.



Simulations de jeux de données multicentriques de taille modeste (4 observations par centre dans 5 centres).

Les panneaux A, B et C correspondent à des jeux légèrement différents.

Dans les encadrés sont données :

- N°1 : les pentes fixes d'un modèle à effets mixtes à intercept aléatoire $y_{ij} = a + [a.centre_j] + bx_{ij} + \varepsilon_{ij}$;
- N°2 : les pentes d'un modèle à effets mixtes à intercept et pente aléatoires ($y_{ij} = a + [a.centre_j] + bx_{ij} + [b.centre_j]x_{ij} + \varepsilon_{ij}$).
- Les pentes d'un modèle linéaire estimé par GEE sont également estimées.

Le symbole en gras représente la moyenne de chaque centre (barycentre des points du centre).

Une modification minime des données conduit les modèles à effets mixtes à estimer tantôt la pente inter-centres (décroissante donc négative), et tantôt la pente intra-centre (positive)

##is# Exemple R

Il faut utiliser la library `gee`.

Dans la syntaxe, l'utilisation de `order` est indispensable pour que les observations soient regroupées par centre.

L'option `corstr="exchangeable"` indique que la corrélation intra-centre est supposée identique entre toutes les paires d'observations d'un même centre (structure de corrélation dite "échangeable"), c'est à dire que la corrélation entre les observations 1 et 2 d'un centre est la même que celle entre les observations 1 et 3, etc.

```
dtlgee <- dtl[order(dtl$centre,dtl$temps),]
mod.gee <- gee(
                y~x,
                data=dtlgee,
                id=centre,
                corstr="exchangeable"
                )
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)           x
   286.6649    153.6712
```

```
summary(mod.gee)
```

```
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Identity
 Variance to Mean Relation: Gaussian
 Correlation Structure:     Exchangeable

Call:
gee(formula = y ~ x, id = centre, data = dtlgee, corstr = "exchangeable")

Summary of Residuals:
       Min         1Q     Median         3Q        Max
-391.96041  -86.83861    2.71576   83.68505  368.27139
```

```
Coefficients:
            Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) 291.3500   9.135526 31.891982    9.410923 30.958709
x           122.4175  31.121353  3.933554   34.743959  3.523419

Estimated Scale Parameter:  16929.16
Number of Iterations:  2

Working Correlation
          [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 1.0000000 0.6610401 0.6610401 0.6610401 0.6610401
[2,] 0.6610401 1.0000000 0.6610401 0.6610401 0.6610401
[3,] 0.6610401 0.6610401 1.0000000 0.6610401 0.6610401
[4,] 0.6610401 0.6610401 0.6610401 1.0000000 0.6610401
[5,] 0.6610401 0.6610401 0.6610401 0.6610401 1.0000000
```

3 Modèle linéaire généralisé

3.A Principes généraux

Si Y n'est plus une variable quantitative normalement distribuée, mais une variable binaire (ex: succès/échec), on peut utiliser un modèle linéaire généralisé (GLM) avec une fonction de lien logit (régression logistique).

NB : condition de validité de l'estimateur du maximum de vraisemblance de la régression logistique = au moins 5 à 10 évènements par variable explicative incluse dans le modèle.

On peut faire une "régression logistique conditionnelle" :

- vraisemblance partielle au sein de chaque centre puis agréger ces vraisemblances en une seule valeur
  - permet d'obtenir des OR intra-centres, mais empêche l'étude de l'effet centre lui-même.
- "modèle linéaire généralisé à effets mixtes = GLMM" (Generalized Linear Mixed Model) :
  - la variable (centre) est vue comme un effet aléatoire.

On peut aussi faire une approche "marginale" :

- Ignorer l'effet centre en premier lieu (modèle logistique simple), avec si possible un bootstrap par grappe ou un estimateur sandwich pour corriger l'erreur standard pour prendre en compte la corrélation intra-centre.
- Utiliser un modèle GEE avec une fonction de lien logit pour obtenir des OR "marginales" corrigées de la corrélation intra-centre.

3.B Exemple R avec données simulées

On reprend le jeu de données simulées mais transforme Y en variable binaire selon un seuil de 300.

```
dtl$y.b <- ifelse(dtl$y>300,1,0)
dtlbis$y.b <- ifelse(dtlbis$y>300,1,0)
dtlgee$y.b <- ifelse(dtlgee$y>300,1,0)
```

1. Modèle non ajusté : régression logistique simple sans prise en compte de l'effet centre

```
modglm <- glm(y.b~x,data=dtl,family="binomial")
library(gtsummary)
# tableau gt summary avec modify footnote pour dire comment les OR ont été obtenus
↪  (en anglais)
tbl_glm <- tbl_regression(
    modglm,
    exponentiate = TRUE,
    label = list(x ~ "Duration of interview (X)")
    ) %>%
```

```
    modify_footnote(
        estimate ~ "Odds Ratios (OR) calculated from logistic regression model
↪   without accounting for center effect."
    )
tbl_glm
```

Characteristic	OR[1]	95% CI	p-value
Duration of interview (X)	4.33	1.03, 18.4	0.047

[1]Odds Ratios (OR) calculated from logistic regression model without accounting for center effect.
Abbreviations: CI = Confidence Interval, OR = Odds Ratio

OR calculé caractérise la relation entre X et Y en ignorant l'effet centre et en supposant que toutes les observations sont indépendantes.

2. Estimation par bootstrap par grappe

Pour corriger l'erreur standard résultant de la corrélation intra-centre, on peut utiliser un bootstrap par grappe ou un estimateur sandwich.

```
# Bootstrap par grappe
modboot <- clusbootglm(
                    y.b~x,
                    data=dtl,
                    clusterid=centre,
                    family="binomial")
modboot$boot.coefs
```

```
(Intercept)           x
-0.07908927  1.47229199
```

```
modboot$boot.sds
```

```
(Intercept)           x
  0.1520077   0.7915578
```

```
exp(modboot$boot.coefs)[2]
```

```
       x
4.359215
```

L'OR estimé après bootstrap est légèrement différent, et l'erreur standard est plus grande, reflétant l'incertitude accrue due à la corrélation intra-centre.

Le rapport entre le coefficient non exponentié et l'erreur standard permet de tester la présence d'une corrélation :

```
cat("Le rapport coefficient/erreur standard est de",
    round((modboot$boot.coefs[2]) / (modboot$boot.sds[2]), 2),
    "\n")
```

```
Le rapport coefficient/erreur standard est de 1.86
```

Le rapport entre le coefficient (non exponentié) et son erreur standard correspond au z-score du test de Wald.

Il sert à tester l'hypothèse nulle H0 : le coefficient de X est nul (OR = 1), c'est-à-dire « pas d'association entre X et Y » après correction de la corrélation intra-centre.

Si ce z-score est inférieur en valeur absolue à 1,96, on ne met pas en évidence d'association significative au seuil de 5 %.

Ici, le coefficient vaut 1,87 donc < 1,96 pas de preuve d'une association significative entre X et Y après correction de la corrélation intra-centre.

3. Modèle ajusté à effet fixe : régression logistique avec ajustement sur la variable catégorielle [centre]

```
modglmaj <- glm(y.b~x+centre,data=dtl,family="binomial")
summary(modglmaj)$coefficients["x", ]
```

```
  Estimate Std. Error    z value   Pr(>|z|)
2.19692245 1.24982624 1.75778230 0.07878456
```

```
exp(summary(modglmaj)$coefficients["x", "Estimate"])
```

```
[1] 8.997281
```

Le problème est que le modèle compte 199 variables indicatrices (dummy) pour les centres, ce qui est lourd.

Il faudrait au mieux 2000 observations pour respecter la règle des 10 évènements par variable, or il y en a 535.

4. Modèle logistique conditionnel :

La fonction `clogit()` permet de faire une régression logistique conditionnelle en utilisant la vraisemblance partielle.

Vraisemblance partielle = méthode d'estimation qui permet d'éliminer les paramètres de nuisance (ici, les intercepts spécifiques à chaque centre) en conditionnant sur le nombre d'événements observés dans chaque groupe.

Au lieu d'estimer la probabilité absolue de l'événement, on estime la probabilité qu'un individu ait l'événement *sachant* le nombre total d'événements observés dans son centre.

Cela permet d'estimer l'association intra-centre sans avoir à estimer les coefficients de chaque centre.

C'est mathématiquement équivalent à un modèle de Cox stratifié.

```
library(survival)
summary(clogit(y.b~x+strata(centre),dtl))
```

```
Call:
coxph(formula = Surv(rep(1, 1000L), y.b) ~ x + strata(centre),
    data = dtl, method = "exact")

  n= 1000, number of events= 535

   coef exp(coef) se(coef)     z Pr(>|z|)
x 1.754     5.779    1.115 1.573    0.116

  exp(coef) exp(-coef) lower .95 upper .95
x     5.779      0.173    0.6494     51.43

Concordance= 0.551  (se = 0.037 )
Likelihood ratio test= 2.48  on 1 df,   p=0.1
Wald test            = 2.47  on 1 df,   p=0.1
Score (logrank) test = 2.49  on 1 df,   p=0.1
```

Le coefficient obtenu correspond au Log-Odds Ratio intra-centre.

5. Modèle mixte à pente commune

Ici, la pente commune correspond à une pente commune au sein des différents centres.

```
moglmer1 <- glmer(y.b~x+(1|centre),data=dtl,family="binomial")
summary(moglmer1)$coefficients["x", ]
```

```
  Estimate Std. Error    z value   Pr(>|z|)
1.92515054 1.01661177 1.89369294 0.05826578
```

```
exp(summary(moglmer1)$coefficients["x", "Estimate"])
```

```
[1] 6.856181
```

6. Modèle mixte autorisant des pentes propres à chaque centre

```
modglmer2 <- glmer(y.b~x+(1+x|centre),data=dtl,family="binomial")
```

```
boundary (singular) fit: see help('isSingular')
```

```r
summary(modglmer2)$coefficients["x", ]
```

```
  Estimate Std. Error    z value   Pr(>|z|)
 2.3676541  1.0784455  2.1954323  0.0281326
```

```r
exp(summary(modglmer2)$coefficients["x", "Estimate"])
```

```
[1] 10.67233
```

`boundary (singular) fit: see help('isSingular')` nous alerte sur un possible problème numérique : cela signifie que le modèle a du mal à estimer certains paramètres en raison d'une matrice de variance-covariance singulière, souvent causée par un manque de variabilité dans les données ou une sur-paramétrisation du modèle.

7. Modèle marginal GEE avec fonction de lien logit

```r
modgee <- gee(
                y.b~x,
                data=dtlgee,
                id=centre,
                corstr="exchangeable",
                family="binomial"
                )
```

```
Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

running glm to get initial regression estimate

(Intercept)           x
-0.07879536  1.46485165
```

```r
summary(modgee)$coefficients["x", ]
```

```
   Estimate  Naive S.E.     Naive z Robust S.E.    Robust z
  1.1122609   0.6119977   1.8174265   0.6426709   1.7306851
```

```r
exp(summary(modgee)$coefficients["x", "Estimate"])
```

```
[1] 3.041227
```

Au final, on a 7 OR différents avec des IC et des pvalue différentes :

title: Mesures repetees

```
\newpage

### Problème posé
Ici, c'est un effet "sujet" (  d'un effet "centre").

Mais ça ne se traite pas pareil :

-   Les sujets d’un même centre (une classe, un service hospitalier, etc.) sont en général cons

    -   Mis à part les variables explicatives du modèle (âge, sexe, etc.), rien ne permet de di

    -   On peut donc supposer que la corrélation de la variable étudiée est la même pour toutes

-   Avec des mesures répétées, la situation change.

    -   Si l’on mesure chaque jour pendant un mois, la valeur observée un jour donné sera en gé

### Exemple R 1
-   Jeu de données `scl90`  : mesures répétées du score SCL90 à 8 reprises sur 56 jours chez 14


::: {.cell}
```

```
::: {.cell-output-display}
![Évolution du score de dépression en fonction du temps et de l'âge](11_Mesures_repetees_files/
:::
:::


Question : étude de l'évolution du score SCL90 au cours du temps.

1.   Data management :


::: {.cell}

```{.r .cell-code}
scl$VISIT.NUM <- recode(scl$VISIT,J0=1,J4=2,J7=3,J14=4,J21=5,J28=6,J42=7,J56=8)
```
:::

Approche naïve : régression linéaire classique entre score et temps.

::: {.cell}

```{.r .cell-code}
mod1 <- lm(depression ~ VISIT.NUM + age, data=scl, na.action=na.exclude)
tbl_mod1 <- tbl_regression(
    mod1,
    label = list(VISIT.NUM ~ "Visit Number", age ~ "Age (years)")
    ) %>%
    modify_footnote(
        estimate ~ "Coefficients calculated from linear regression model without accounting for
    )
tbl_mod1
```

::: {.cell-output-display}
\begin{table}
\fontsize{12.0pt}{14.0pt}\selectfont
\begin{tabular*}{\linewidth}{@{\extracolsep{\fill}}lccc}
\toprule
\textbf{Characteristic} & \textbf{Beta}\textsuperscript{\textit{1}} & \textbf{95\% CI} & \textb
\midrule\addlinespace[2.5pt]
Visit Number & -3.2 & -3.5, -2.8 & <0.001 \\
Age (years) & 0.11 & 0.06, 0.17 & <0.001 \\
\bottomrule
\end{tabular*}
\begin{minipage}{\linewidth}
\textsuperscript{\textit{1}}Coefficients calculated from linear regression model without accour
```

```
Abbreviation: CI = Confidence Interval\\
\end{minipage}
\end{table}

:::
:::


Résultats :

-   Interprétables pour le calcul du coefficient de régression (beta),

-   mais est à proscrire pour le calcul de son incertitude (erreur type) car il ignore la struc

On pourrait vérifier en regardant les résidus du modèle avec QQplot et histogramme :

::: {.cell}
::: {.cell-output-display}
![Diagnostic des résidus du modèle linéaire classique](11_Mesures_repetees_files/figure-pdf/unr
:::
:::


C'est pas terrible !!

Examen dans le détail de la structure de corrélation :


::: {.cell}

```{.r .cell-code}
# création d'un tableau des résidus
dtresid <- na.omit(data.frame(scl$NUMERO, scl$VISIT.NUM, residuals(mod1)))

# reshaping pour avoir une colonne par visite (transformation en "large")
names(dtresid) <- c("NUMERO","VISIT.NUM","RESID")
dtresid.w <- reshape(dtresid, v.names="RESID", idvar="NUMERO", timevar="VISIT.NUM", direction='

# calcul de la matrice de corrélation des résidus
round(cor(dtresid.w[,2:9], use="pairwise.complete.obs"),2)
```

::: {.cell-output .cell-output-stdout}

```
        RESID.1 RESID.2 RESID.3 RESID.4 RESID.5 RESID.6 RESID.7 RESID.8
RESID.1    1.00    0.66    0.57    0.49    0.42    0.42    0.32    0.28
RESID.2    0.66    1.00    0.77    0.72    0.64    0.58    0.49    0.41
```
```

```
RESID.3    0.57    0.77    1.00    0.89    0.73    0.68    0.57    0.50
RESID.4    0.49    0.72    0.89    1.00    0.90    0.86    0.75    0.63
RESID.5    0.42    0.64    0.73    0.90    1.00    0.88    0.78    0.71
RESID.6    0.42    0.58    0.68    0.86    0.88    1.00    0.87    0.81
RESID.7    0.32    0.49    0.57    0.75    0.78    0.87    1.00    0.87
RESID.8    0.28    0.41    0.50    0.63    0.71    0.81    0.87    1.00
```

:::
:::

Pour mieux illustrer la matrice de corrélation des résidus : correlation plot

::: {.cell}
::: {.cell-output-display}
![Matrice de corrélation des résidus](11_Mesures_repetees_files/figure-pdf/unnamed-chunk-7-1.pd
:::
:::

Il y a :

- Une forte corrélation des résidus consécutifs

- Décroissance des résidus si on passe des résidus mesurés à des temps consécutifs aux résidu

Illustration par une ACP :

::: {.cell}
::: {.cell-output-display}
![ACP des résidus : les corrélations décroissent avec l'éloignement temporel](11_Mesures_repete
:::
:::

Pour contourner : on peut faire un bootstrap (car le bootstrap ne fait pas d'hypothèse d'indépe

Au lieu de piocher des mesures individuelles au hasard (ce qui casserait la dépendance), on pio

Cela fonctionne car cette méthode préserve la structure de corrélation temporelle à l'intérieur

::: {.cell}

```{.r .cell-code}
set.seed(1)
mod.clusboot <- clusbootglm(depression~VISIT.NUM,data=scl,clusterid=NUMERO)
mod.clusboot$lm.coefs
```

::: {.cell-output .cell-output-stdout}

```
(Intercept)   VISIT.NUM
  34.676641   -3.198859
```

:::

```{.r .cell-code}
mod.clusboot$boot.sds
```

::: {.cell-output .cell-output-stdout}

```
(Intercept)   VISIT.NUM
  0.9057995   0.1519799
```

:::
:::

Les résultats sont "bruts" : calculés sur l'échantillon dans son ensemble.

Une approche conditionnelle rechercherait une association entre `depression` et `VISIT.NUM` au

Exemples d'approches conditionnelles pour les mesures répétées :

- Régression logistique conditionnelle (effets fixes sujet)
    - On conditionne sur chaque sujet (on lui donne un intercept propre, non estimé explicite
    - On compare les mesures d'un même sujet entre elles, ce qui contrôle parfaitement pour t
- Modèles linéaires mixtes / GLMM (effets aléatoires sujet)
    - On introduit des effets aléatoires (souvent un intercept aléatoire par sujet, parfois c

    - On obtient aussi un effet « conditionnel » (effet du temps ou du traitement pour un suje

- GEE (Generalized Estimating Equations)

    - On ne modélise pas explicitement les effets aléatoires, mais directement la corrélation

    - L'effet estimé est plutôt « marginal/population-averaged » (effet moyen dans la populati

Modèle conditionnel à effets mixtes (LMM)
Inutilisable ici car la structure de la matrice de corrélation des résidus n'est pas homogène (

Pour un LMM : il faudrait une corrélation égale entre toutes les paires de mesures d'un sujet,

On peut tout de même faire un bootstrap qui portera sur un modèle linéaire ajusté sur la va

- On ne suppose plus que les mesures répétées sont indépendantes, on suppose seulement que le

- En ré-échantillonnant des sujets entiers (NUMERO) avec toutes leurs visites, on préserve la

- Le modèle linéaire est alors ajusté sur chaque échantillon bootstrap « par sujet », et la v

::: {.cell}

```{.r .cell-code}
set.seed(1)
mod.clusbootcond <- clusbootglm(depression ~ VISIT.NUM + factor(NUMERO),data=scl,clusterid=NUMB
mod.clusbootcond$lm.coefs[1:2]
```

::: {.cell-output .cell-output-stdout}

```
(Intercept)   VISIT.NUM
  39.710388   -3.188528
```

:::

```{.r .cell-code}
mod.clusbootcond$boot.sds[1:2]
```

::: {.cell-output .cell-output-stdout}

```
```

```
(Intercept)   VISIT.NUM
  5.2422655   0.1517664
```

:::
:::

Modèle linéaire avec effets mixtes (intercept et pentes fixes + aléatoires)
Équation du modèle :

$$
y_{ij} = a + (a_{sujet_j}) + b x_{ij} + (b_{sujet_j}) x_{ij} + \varepsilon_{ij}
$$

Où :

- a : intercept fixe (moyen dans la population) et b : pente fixe (moyenne dans la popula

- a_{sujet_j} : intercept aléatoire du sujet j (écart du sujet j par rapport à l'intercept

On fournit ici l'argument ε_{ij} à la fonction `lme` pour modéliser la structure de

ε_{ij} = `random=~VISIT.NUM|NUMERO` : correspond à une corrélation linéaire décrois

`random = ~ VISIT.NUM | NUMERO` : pour chaque sujet (NUMERO), on ajoute un intercept aléatoire

::: {.cell}

```{.r .cell-code}
modCS <- lme(
            depression~VISIT.NUM,
            random=~VISIT.NUM|NUMERO
            ,data=scl,
            na.action=na.omit)
summary(modCS)
```

::: {.cell-output .cell-output-stdout}

```
Linear mixed-effects model fit by REML
  Data: scl
       AIC      BIC    logLik
  6528.841 6558.086 -3258.421
```

```
Random effects:
 Formula: ~VISIT.NUM | NUMERO
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev   Corr
(Intercept) 9.590448 (Intr)
VISIT.NUM   1.476833 -0.447
Residual    5.132011

Fixed effects:  depression ~ VISIT.NUM
               Value Std.Error  DF   t-value p-value
(Intercept) 34.23482 0.8874113 825  38.57830       0
VISIT.NUM   -3.08111 0.1501042 825 -20.52646       0
 Correlation:
          (Intr)
VISIT.NUM -0.538

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.134106317 -0.506968777 -0.007139085  0.502921836  3.629924688

Number of Observations: 969
Number of Groups: 143
```

:::
:::

Structure auto-régressive (AR(1))
Les corrélations des résidués mesurés à des temps immédiatement successifs sont tous identitque

Entre t et t+2 : les corrélations sont toutes égales à ρ^2

Entre t et t+3 : les corrélations sont toutes égales à ρ^3

Par défaut :

- Variance résiduelle constante σ^2

- Matrice de variance-covariance des résidus = $\sigma^2 \begin{pmatrix}1 & \rho & \rho^2 & .

::: {.cell}

```{.r .cell-code}
modAR1 <- lme(
            depression~VISIT.NUM,
```

```
            random=~VISIT.NUM|NUMERO,
            correlation=corAR1(form=~VISIT.NUM),
            data=scl,
            na.action=na.omit)
```

::: {.cell-output .cell-output-error}

```
Error in lme.formula(depression ~ VISIT.NUM, random = ~VISIT.NUM | NUMERO, : nlminb problem, co
  message = iteration limit reached without convergence (10)
```

:::
:::

NB : `VISIT.NUM` doit être numérique entier croissant (1,2,3,...).

Ici, l'algorithme ne converge pas car le modèle est complexe, et le nombre d'itérations maximum

On change des paramètres :

::: {.cell}

```{.r .cell-code}
modAR1 <- lme(
            depression~VISIT.NUM,
            random=~VISIT.NUM|NUMERO,
            correlation=corAR1(form=~VISIT.NUM),
            data=scl,
            na.action=na.omit,
            control=list(msMaxIter=1000,msMaxEval=1000,
            opt="optim", optimMethod="Nelder-Mead"))
summary(modAR1)
```

::: {.cell-output .cell-output-stdout}

```
Linear mixed-effects model fit by REML
  Data: scl
       AIC      BIC    logLik
  6354.534 6388.654 -3170.267

Random effects:
```

```
 Formula: ~VISIT.NUM | NUMERO
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 2.8747631 (Intr)
VISIT.NUM   0.5483729 0.989
Residual    8.8811504

Correlation Structure: ARMA(1,0)
 Formula: ~VISIT.NUM | NUMERO
 Parameter estimate(s):
     Phi1
0.7676115
Fixed effects:  depression ~ VISIT.NUM
               Value Std.Error  DF   t-value p-value
(Intercept) 34.96530 0.8807073 825  39.70138       0
VISIT.NUM   -3.11243 0.1564600 825 -19.89280       0
 Correlation:
          (Intr)
VISIT.NUM -0.611

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-2.47367228 -0.76647525 -0.08857657  0.59546791  2.43449131

Number of Observations: 969
Number of Groups: 143
```

:::
:::

AIC et BIC : indices de parcimonie.

Variance et covariance des effets aléatoires dans la partie "Random effects" ; des effets fixes

Le paramètre de corrélation ρ est estimé dans la partie "Correlation Structure" et appelé

Pour retrouver le paramètre de corrélation entre résidus consécutifs :

::: {.cell}

```{.r .cell-code}
# 1ère façon
modAR1$modelStruct$corStruct
```

::: {.cell-output .cell-output-stdout}

```
Correlation structure of class corARMA representing
     Phi1
0.7676115
```

:::

```{.r .cell-code}
# 2ème façon
round(cov2cor(getVarCov(modAR1,type="conditional", individual=9)[[1]]),2)
```

::: {.cell-output .cell-output-stdout}

```
     1    2    3    4    5    6    7    8
1 1.00 0.77 0.59 0.45 0.35 0.27 0.20 0.16
2 0.77 1.00 0.77 0.59 0.45 0.35 0.27 0.20
3 0.59 0.77 1.00 0.77 0.59 0.45 0.35 0.27
4 0.45 0.59 0.77 1.00 0.77 0.59 0.45 0.35
5 0.35 0.45 0.59 0.77 1.00 0.77 0.59 0.45
6 0.27 0.35 0.45 0.59 0.77 1.00 0.77 0.59
7 0.20 0.27 0.35 0.45 0.59 0.77 1.00 0.77
8 0.16 0.20 0.27 0.35 0.45 0.59 0.77 1.00
```

:::
:::

Comme ρ est inférieur à 1, les corrélations entre résidus décroissent avec l'éloignement t

Modèle à structure de covariance non structurée (unstructured)
Les blocs de corrélation 8x8 ont une structure libre, avec 24 paramètres de corrélation à estim

::: {.cell}

```{.r .cell-code}
modSymm <- lme(depression~VISIT.NUM,
                random=~VISIT.NUM|NUMERO,
                data=scl,
```

```
                correlation=corSymm(form=~VISIT.NUM),
                na.action=na.omit,
                control=list(msMaxIter=1000,msMaxEval=1000,
                opt="optim",
                optimMethod="BFGS")
                )
summary(modSymm)
```

::: {.cell-output .cell-output-stdout}

```
Linear mixed-effects model fit by REML
  Data: scl
       AIC     BIC    logLik
  6348.787 6514.51 -3140.394

Random effects:
 Formula: ~VISIT.NUM | NUMERO
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev    Corr
(Intercept) 6.3042211 (Intr)
VISIT.NUM   0.4542091 0.9
Residual    6.5210451

Correlation Structure: General
 Formula: ~VISIT.NUM | NUMERO
 Parameter estimate(s):
 Correlation:
  1      2      3      4      5      6      7
2  0.310
3  0.028  0.422
4 -0.267  0.090  0.634
5 -0.407 -0.084  0.240  0.661
6 -0.350 -0.251  0.098  0.505  0.622
7 -0.554 -0.544 -0.239  0.138  0.223  0.644
8 -0.613 -0.763 -0.464 -0.169 -0.025  0.423  0.664
Fixed effects:  depression ~ VISIT.NUM
               Value Std.Error  DF   t-value p-value
(Intercept) 34.77922 0.8371978 825  41.54242       0
VISIT.NUM   -3.04206 0.1563385 825 -19.45818       0
 Correlation:
          (Intr)
VISIT.NUM -0.591

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-3.1429944 -0.8093216 -0.1225221  0.5561505  2.8984123
```

```
Number of Observations: 969
Number of Groups: 143
```

:::
:::

Les estimations des effets aléatoires ont changé : l'écart type de l'intercept est passé de 2,8

Comme c'est pas structuré, les effets aléatoires et les résidus participent tous deux à la modé

- Plusieurs briques décrivent la même covariance, donc elles se « partagent le travail ».
    - La covariance mesure comment 2 mesures varient ensemble : si, pour un même sujet, quanc
    - Dans un modèle à mesures répétées, la matrice de covariance regroupe toutes ces covaria
- Les effets aléatoires (intercept/pente aléatoires par sujet) génèrent déjà une partie de la
- La structure corSymm non structurée ajoute, en plus, une matrice de covariance très flexibl
- Au total : la covariance observée entre deux mesures = covariance due aux effets aléatoires

Comment choisir ?

- Les effets fixes sont proches entre les modèles, mais ce ne sera pas toujours vrai, et on n
- On peut modéliser la corrélation des résidus, càd tenir compte du fait que deux erreurs pou
    - En théorie, si on gère bien cette corrélation, on donne moins de poids à des observatio
- Dans notre exemple, ce gain théorique ne se voit presque pas : les erreurs-types sont très
- Si le modèle est mal spécifié (non-linéarité, non-additivité…), on ne sait pas bien comment
- Avantage possible de `lme` : on peut estimer la variance des pentes entre sujets ou le $\rh
- Problème pratique : il faut choisir une structure de covariance des résidus (AR(1), non str

`<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4ifQ== -->`{=html}

```{=html}
<!-- quarto-file-metadata: eyJyZXNvdXJjZURpciI6Ii4iLCJib29rSXRlbVR5cGUiOiJjaGFwdGVyIiwiYm9vaOlO
```

Series chronologiques

```{.quarto-title-block template='/Applications/quarto/share/projects/book/pandoc/title-blo
---
title: Series chronologiques

---
```

4 Introduction aux séries chronologiques

L'objectif est de modéliser une variable à expliquer qui est

- un "processus stationnaire",
- aléatoire, avec une moyenne dans le temps
- et pour lequel la corrélation de la mesure au temps t_1 avec la mesure au temps t_2 dépend uniquement de la distance entre t_1 et t_2.

Méthodes SARIMAX / ARMA / ARIMA

- SARIMAX : Seasonal Autoregressive Integrated Moving Average with eXogenous regressors
- ARMA : Autoregressive Moving Average
- ARIMA : Autoregressive Integrated Moving Average

4.A Modèle auto-régressif d'ordre 1 :

Soit une variable Y mesurée au temps i

- Régression de Y sur ce qui se passe au temps $i-1$: $Y_(i) = (a) + bY_(i-1) + \epsilon$
  - Ce qui se passe au temps i = une fonction linéaire de ce qui se passe au temps $i-1$ + un bruit aléatoire
  - Processus auto-régressif d'ordre 1 (AR(1))

On observe une variable Y dans le temps : Y_1, Y_2, Y_3.

Le modèle AR(1) écrit :

$Y_i = a + bY_{i-1} + \varepsilon_i$

où ε_i est un bruit aléatoire indépendant d'un instant à l'autre, c'est à dire qu'il n'y a pas de corrélation entre les ε_i entre eux

4.A.1 Qu'est-ce qu'une régression (dans ce contexte) ?

Une régression est un modèle qui cherche à expliquer une variable par une autre, en supposant une relation linéaire.

Dans un AR(1), la variable explicative n'est pas une autre variable, mais l**a même variable observée juste avant**, c'est-à-dire Y_{i-1}.

C'est donc une régression temporelle, ou auto-régression :

- Y_{i-1} est l'information disponible au temps précédent,
- Y_i est ce qu'on cherche à prédire,
- ε_i représente la part de Y_i qui n'est pas expliquée par Y_{i-1}.

Autrement dit : le présent dépend linéairement du passé (= *c'est une fonction linéaire*), plus un terme aléatoire.

4.A.2 Pourquoi dit-on souvent que « $a = 0$ » ?

Le coefficient a est une constante, comme l'intercept dans une régression classique.
Cependant, dans la majorité des cours et logiciels :

1. La série est centrée avant modélisation.
   - Centrer consiste à soustraire la moyenne : $Y_i \leftarrow Y_i - \bar{Y}$.
   - Une série centrée a une moyenne égale à 0.
2. Si la série est centrée, alors le terme constant devient inutile.
   - En effet, **la constante sert justement à ajuster une moyenne non nulle**.
   - Donc, dans une série centrée : $Y_i = bY_{i-1} + \varepsilon_i$
   - et on peut prendre $a = 0$ par convention.

4.A.3 Variance : définition générale

La variance est une mesure de la dispersion d'une variable autour de sa moyenne.

- Une grande variance = la variable varie beaucoup.
- Une petite variance = la variable varie peu.

Mathématiquement :
Var$(Y) = E[(Y - E[Y])^2]$
Dans un AR(1), il y a deux sources de variance :

1. la variance du bruit ε_i
2. la propagation de ce bruit à travers le coefficient b

4.A.4 Variance du bruit : ce que c'est exactement

Le bruit ε_i est un terme aléatoire supposé :

- d'espérance 0 (= moyenne, c'està dire que le bruit n'a pas de tendance moyenne à être positif ou négatif),
- indépendant d'un instant à l'autre,
- de variance constante σ^2.

Cette variance σ^2 mesure l'amplitude des fluctuations imprévisibles.
C'est la "source" de variabilité du modèle.
Si σ^2 est grande : chaque instant reçoit un choc important.
Si σ^2 est petite : les variations imprévisibles sont faibles.

Note

“Choc” lié à une variance importante :
Le Rôle de l’Erreur (Le Choc Ponctuel)
Dans le modèle AR(1) ($_i = a + bY_{i-1} + \varepsilon_i$), l’erreur ε_i est appelée le **bruit aléatoire**.

- Le bruit ε_i représente la part de la valeur actuelle Y_i qui **n’est pas expliquée** par la valeur précédente Y_{i-1}.
- Ce bruit est la **source de variabilité** du modèle. Sa variance (σ^2) mesure l’amplitude des fluctuations imprévisibles à un instant donné.
- Dans une **régression ordinaire (non temporelle)**, cette erreur ε_i affecterait uniquement la valeur courante Y_i, puis son effet s’arrêterait là.

La Propagation Temporelle (La Mémoire)
Dans l’AR(1), l’erreur ε_i ne disparaît pas immédiatement, car la valeur Y_i qu’elle vient de modifier devient la **variable explicative** pour la période suivante.
Ainsi, l’erreur ε_i influence non seulement Y_i, mais aussi toutes les valeurs futures $Y_{i+1}, Y_{i+2}, \ldots$ à travers la relation autorégressive.

1. **L’effet sur Y_i:** Si ε_i est grand (un “choc” important), il fait monter ou descendre la valeur Y_i.
2. **L’influence sur Y_{i+1} (via b):** La valeur future Y_{i+1} dépend de Y_i via le coefficient b ($Y_{i+1} = a + b_i + \varepsilon_{i+1}$). Comme Y_i a été modifié par ε_i, l’erreur initiale ε_i se répercute sur Y_{i+1} à travers Y_i et l’influence est mesurée par b.
3. **L’influence sur Y_{i+2} (via b^2):** De même, Y_{i+2} dépend de Y_{i+1}, qui lui-même contenait déjà l’influence du choc ε_i. L’effet de l’erreur initiale ε_i est donc transmis à Y_{i+2} via b^2.
4. **L’influence sur Y_{i+k}:** L’influence de l’erreur ε_i sur une valeur lointaine Y_{i+k} est mesurée par b^k. En résumé, l’erreur ε_i se propage dans le futur, mais son impact **décroît exponentiellement** avec la distance temporelle (puisque k augmente). C’est ce que l’on appelle la **corrélation temporelle**.

L’Équilibre de la Variance Totale
La **variance** mesure la dispersion de la variable.
Puisque l’erreur ε_i ne s’arrête jamais d’influencer les valeurs futures tant que le coefficient b n’est pas nul, la variabilité totale observée dans la série Y_i est la somme de toutes les influences passées. (si b devient nul, l’effet de l’erreur s’arrête immédiatement car Y_i ne dépend plus de Y_{i-1} mais dépend uniquement de l’erreur actuelle ε_i : la nouvelle équation sera $Y_i = a + (0 \times Y_{i-1}) + \varepsilon_i$)
La **variance totale** de la série Y_i est le résultat d’un équilibre entre deux facteurs : la variance du bruit σ^2 et la mémoire du système (le coefficient b).

1. **La variabilité injectée à chaque instant** : C’est la variance du bruit σ^2 qui mesure l’amplitude des fluctuations imprévisibles de ε_i. Si σ^2 est grande, cela signifie qu’il y a une grande quantité de variabilité nouvelle et imprévisible introduite à chaque pas de temps : chaque instant reçoit un “choc” important.

2. **La quantité de mémoire du système = la propagation de cette variabilité** Ceci est donné par le coefficient b. Plus b est grand, plus les erreurs passées continuent à influencer fortement les valeurs futures parce que

Si la série est **stationnaire** (c'est-à-dire que $|b| < 1$), la variance totale est stable et finie. Elle est calculée par la formule : $Var(Y_i) = \frac{\sigma^2}{1-b^2}$ Cette formule montre clairement que la variance totale du processus est **plus grande que la variance du bruit seul** (σ2), car l'effet de la propagation des erreurs passées s'accumule (via le dénominateur $1 - b^2$).
Imaginons l'AR(1) comme un verre d'eau posé sur une table : le bruit (ε_i) est une petite goutte qui tombe dans l'eau à chaque instant (i). Dans une régression classique, l'ondulation de la goutte s'arrêterait immédiatement. Mais dans l'AR(1), si l'ondulation d'une goutte (l'erreur ε_i) est assez forte (si b est grand), elle crée un mouvement qui influence l'emplacement des gouttes suivantes (Y_{i+1}, Y_{i+2}), ce qui fait que le niveau total d'agitation de l'eau (la variance totale) est bien supérieur à la simple force d'une seule goutte (σ^2).

4.A.5 Propagation de la variance dans un AR(1)

C'est ici que l'AR(1) devient différent d'une régression classique.

Dans une régression ordinaire, l'erreur affecte uniquement la valeur courante.

Dans un AR(1), l'erreur affecte toute la suite, car :

$Y_{i+1} = a + bY_i + \varepsilon_{i+1}$

Donc si ε_i est grand :

- il fait monter ou descendre Y_i,
- ce Y_i influence Y_{i+1} via le coefficient b,
- et influence Y_{i+2} via b^2, etc.

Ainsi, les erreurs se propagent dans le futur, et la variance totale est un équilibre entre :

- la variabilité injectée à chaque instant (σ^2),
- la quantité de mémoire du système (le coefficient b).

4.A.6 Variance stationnaire : définition explicite

Une variance stationnaire signifie :

1. La variance de Y_i reste la même pour tous les instants : elle ne diverge pas, ne monte pas, ne descend pas.
2. Le processus garde une variabilité "stable" dans le temps.

Autrement dit : la série ne devient ni explosive, ni de plus en plus variable, ni de moins en moins variable.

Cette situation n'est possible que si $|b| < 1$.

- Si $|b| < 1$, les erreurs se propagent mais finissent par s'atténuer variance stable.

- Si $|b| = 1$, les erreurs se cumulent sans limite variance infinie (marche aléatoire).
- Si $|b| > 1$, la série explose variance diverge.

Dans le cas stationnaire ($|b| < 1$) :

Var$(Y_i) = \frac{\sigma^2}{1-b^2}$

Cette formule signifie :

- la variance totale est plus grande que la variance du bruit,
- plus b est grand, plus la variance de Y_i augmente,
- car plus b est grand, plus les erreurs anciennes continuent à influencer les valeurs futures.

4.A.7 Corrélation temporelle : définition explicite

La corrélation mesure l'intensité du lien linéaire entre deux valeurs.

Dans un AR(1) stationnaire:

Corr$(Y_i, Y_{i-k}) = b^k$

Donc :

- Corr$(Y_i, Y_{i-1}) = b$
- Corr$(Y_i, Y_{i-2}) = b^2$
- Corr$(Y_i, Y_{i-3}) = b^3$
- etc.

C'est une corrélation **qui décroît exponentiellement avec la distance temporelle.**

4.A.8 Exemple chiffré complet

Prenons un AR(1) :

$Y_i = 0.8Y_{i-1} + \varepsilon_i, \quad$ Var$(\varepsilon_i) = 1$

Variance stationnaire

Var$(Y_i) = \frac{1}{1-0.8^2} = \frac{1}{1-0.64} = \frac{1}{0.36} \approx 2.78$

Donc :

- la variance du processus est presque 3 fois la variance du bruit,
- car b=0.8 propage beaucoup les erreurs passées.

4.B Simulation courte

Supposons Y_0 = 10.

- $\varepsilon_1 = 0.5$

$Y_1 = 0.8 \times 10 + 0.5 = 8.5$

```
-   $\varepsilon_2 = -0.2$
```

$Y_2 = 0.8 ×8.5 - 0.2 = 6.6

- $\varepsilon_3 = 0.1$

$Y_3 = 0.8 \times 6.6 + 0.1 = 5.38$

La série décroît progressivement vers 0 (car $|b| < 1$) et reste influencée par chaque erreur qui se propage.

4.C Corrélation temporelle

Corr$(Y_3, Y_2) = 0.8$

Corr$(Y_3, Y_1) = 0.8^2 = 0.64$

Corr$(Y_3, Y_0) = 0.8^3 = 0.512$

Résumé

Un AR(1) est un modèle où :

- le présent est expliqué par le passé (régression),
- le coefficient b crée une corrélation décroissante avec le passé,
- la variance totale résulte de l'équilibre entre :
  - la variance du bruit,
  - la propagation de cette variance via b,
- la variance stationnaire est une variance stable, finie, identique à travers le temps,
- la constante a disparaît si on travaille sur une série centrée.

---

5 Modèle MA = Moving Average

Un modèle MA(q) exprime la valeur actuelle comme une **MOYENNE MOBILE** des **ERREURS PASSÉES**.

Les **BRUITS** (ou erreurs) sont liés les uns avec les autres.

Modèle MA(1) :

$Y_j = \varepsilon_i + \varepsilon_{i-1}$

$Y_{j-1} = \varepsilon_{i-1} + \varepsilon_{i-2}$

La corrélation entre Y_j et Y_{j-1} est donc due à la présence commune de ε_{i-1} dans les deux équations. Donc corrélation ≠ 0 car il y a un terme commun.

Alors que si on fait la corrélation entre Y_j et Y_{j-2} :

$Y_j = \varepsilon_i + \varepsilon_{i-1}$

$Y_{j-2} = \varepsilon_{i-2} + \varepsilon_{i-3}$

Il n'y a pas de terme commun entre les deux équations, donc corrélation = 0.

Donc avec :

- Un modèle AR(1) : la corrélation diminue exponentiellement avec le temps = autocorrélogramme.
- Un modèle MA(1) : la corrélation est nulle au-delà du lag 1.

Donc à partir de la réprésentation de la corrélation en fonction du lag (autocorrélogramme), on peut deviner si le modèle est AR ou MA.

6 Autocorrélogramme partiel

Un autocorrélogramme partiel permet de voir la corrélation entre

- Y_i et Y_{i-p}
- ajusté sur $Y_{i-1}, Y_{i-2}, ..., Y_{i-(p-1)}$.

Corrélation spécifique ajustée entre le passé ± lointain et le présent, en ajustant sur tout ce qui se passe entre les deux.
Si on fait avec $p = 2$$:

- On regarde la corrélation entre Y_i et Y_{i-2}

On a Y_1, Y_2, Y_3 (car p=2 donc on a besoin de 3 observations pour avoir un lag de 2 entre Y_3 et Y_1)

- Processus autorégressif d'ordre 1: Y_1 agit sur Y_2 (lag 1) et Y_2 agit sur Y_3 (lag 1)
- Corrélation entre Y_1 et Y_3 vaut la corrélation entre chaque Y exponentiée par le lag : b^2
- Mais si on ajuste sur Y_2, on enlève l'effet de Y_2 sur Y_3, donc Y_1 n'a plus d'effet sur Y_3 !
- La corrélation entre Y_1 et Y_3 ajustée sur Y_2 vaut donc 0.

Donc l'autocorrélogramme partiel va donner 0 au-delà du lag 1 pour un modèle MA(1).

6.A Exemple 1

$Yi = bYi - 1 + cYi - 2 + \varepsilon_i + \varepsilon_{i-1}$
Modèle ARMA(2,1)
Dans un modèle ARMA(p, q),

- p = ordre autorégressif (AR) □ nombre de termes Y retardés,
- q = ordre moyenne mobile (MA) □ nombre de termes ε retardés.

1. Partie AR (autorégressive)
   - Y_{i-1}
   - Y_{i-2}

Donc 2 termes AR □ p = 2.

2. Partie MA (moyenne mobile)

- ε_i toujours présent, ce n'est pas compté dans q
- ε_{i-1}

Le terme ε_i est l'erreur courante : on ne le compte pas dans l'ordre.

On compte seulement les retards.

Donc il y a 1 terme MA retardé q = 1.

6.B Exemple 2

Mesure de la TA tous les matins

Y_i = TA du jour i : il y en a donc 365

6.B.1 Modèle autorégressif AR(1)

Les Y_i ont une variance = 1 c'est à dire que la TA varie peu d'un jour à l'autre.

$\rightarrow Y_i = bY_{i-1} + \varepsilon_i$

- On peut pas faire une régression linéaire habituelle car Y_i est à la fois à gauche et à droite de l'équation
- si variance = 1 / variance stationnaire :
  - b = corrélation entre Y_i et Y_{i-1} quelque soit i
  - avec le temps : corrélation entre Y_i et Y_{i-2} = b^2
  - corrélation entre Y_i et Y_{i-k} = b^k
  - avec le temps, la corrélation tend vers 0 de façon exponentielle

6.B.2 Modèle moyenne mobile MA(1)

$Y_i = \varepsilon_i + \varepsilon_{i-1}$

- 2 bruits : régression linéaire impossible
- Le temps i et le temps $i-1$ n'ont rien à voir l'un avec l'autre et sont liés au bruit
  - Appareil de mesure a aussi un bruit et a un défaut ponctuel à un moment donné
  - $*$ stress du jour qui influe sur le bruit du jour aussi

Globalement : dans la mesure du jour, le bruit observé vient d'un bruit issu de la mesure du jour + du bruit d'hier qui vient d'un contexte, influençant l'erreur de mesure qu'il y a dans Y_i aujourd'hui.

- Corrélation entre Y_i et Y_{i-1} = ϱ et est due au fait qu'il y a un terme commun dans les 2 équations : ε_{i-1}
- Corrélation entre Y_i et Y_{i-2} = 0 car pas de terme commun dans l'équation

6.B.3 Autocorrélogramme

Permet assez facilement de visualiser si le modèle est AR ou MA

6.B.4 Autocorrélogramme partiel

Corrélation entre Y_i et Y_{i-k} ajustée sur $Y_{i-1}, Y_{i-2}, ..., Y_{i-(k-1)}$ (ajusté sur toutes les mesures qu'on a entre les 2)
Si modèle AR(1) :

- Y_1 agit sur Y_2 et Y_2 agit sur Y_3
- Mais Y_1 agit sur Y_3 uniquement via Y_2
- Donc si on ajuste sur Y_2, la corrélation entre Y_1 et Y_3 ajustée sur Y_2 = 0

Donc modèle AR(1) dans autocorrélogramme partiel ressemble au modèle MA(1) dans autocorrélogramme classique.

6.B.5 Processus stationnaire

= Processus dont les propriétés statistiques (moyenne, variance, autocorrélation) ne changent pas au cours du temps.

7 Différents modèles : AR, MA, ARMA, SARMA, SARIMA, SARIMAX

On part d'un processus stationnaire.

1. **Description** : évolution au cours du temps => MOYENNE STABLE ? VARIANCE STABLE ?
2. **Modélisation** : autocorrélogramme et autoccorrélogramme partiel => modèle AR, MA ou ARMA ?
    - voire SARMA : saisonnier (avec saisonnalité)
    - voire SARIMA : avec dérive au cours du temps (écart à la stationnarité : I = Integrated c'est à dire qu'on intègre une tendance linéaire au modèle)
    - voire SARIMAX : avec des variables explicatives (exogènes)
        - exogènes = variables explicatives externes au modèle de la série chronologique
        - $Y_i = bY_{i-1} + cY_{i-2} + \varepsilon_i + \varepsilon_{i-1} + d\vartheta$

---

8 Exemple d'analyse de données avec R

8.A Méthode graphique

Le fichier `gs` contient des données sur les tendances de recherche Google (sur recherche = suicide entre Juin 2017 et Juin 2022).

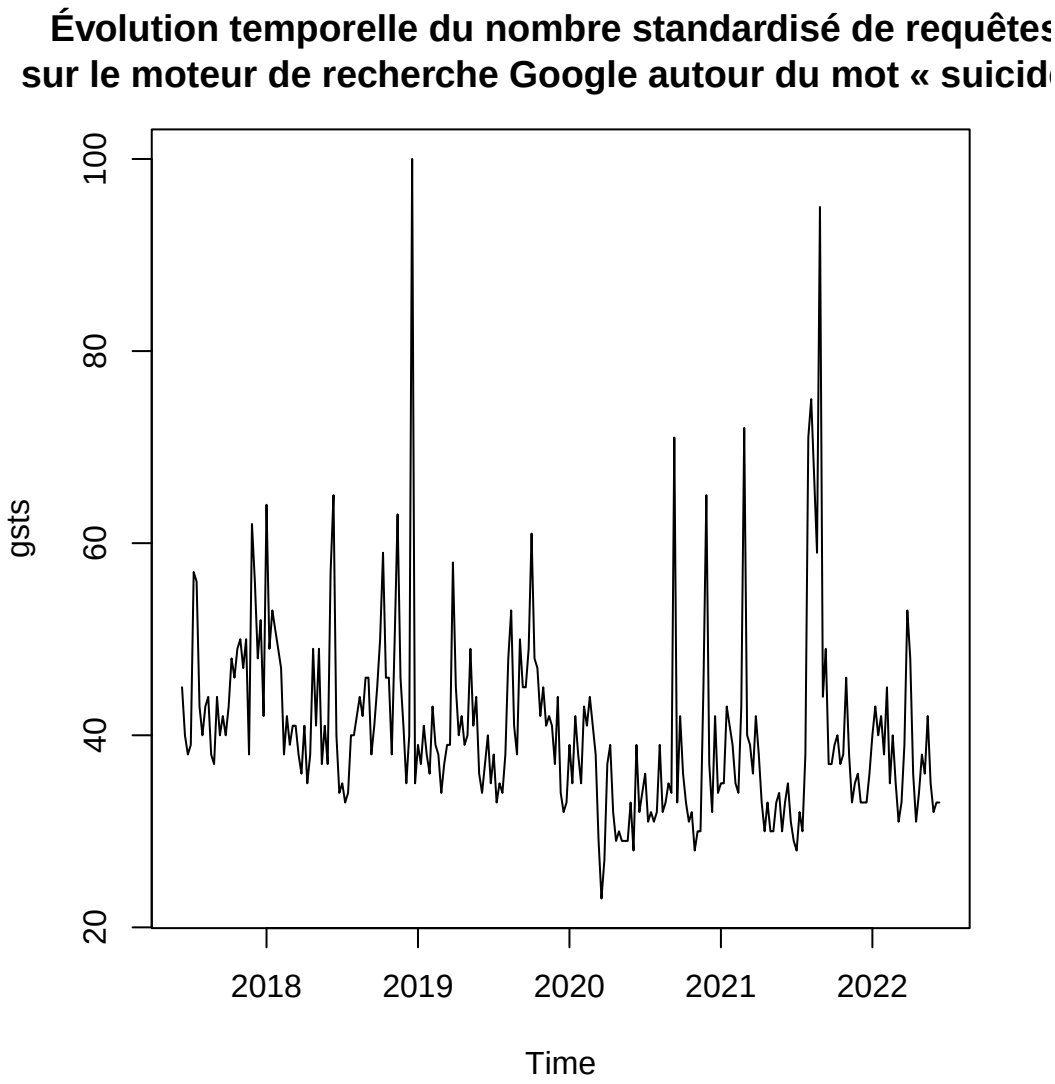
Voir si indice X = survenue du confinement modifie la recherche du mot "suicide"

Modèle = ARMAX (avec X = facteur exogène = confinement)

- avec X

1. Transformation de la variable `gs$suicide` en une série chronologique grâce à la fonction `ts` = timeserie
   - si pas de fonction `ts`: on peut pas faire d'analyses sur séries chronologiques avec R.
   - syntaxe de la fonction `ts` :
     - `ts(data, frequency, start)`
     - `data` = données,
     - `frequency` = fréquence (ex: 52 semaines dans une année),
       * permet de dire que la semaine i et la semaine $i+52$ sont comparables (même semaine de l'année)
     - `start` = date de début (ex: c(2017,24))
2. Affichage de la série chronologique avec la fonction `plot`

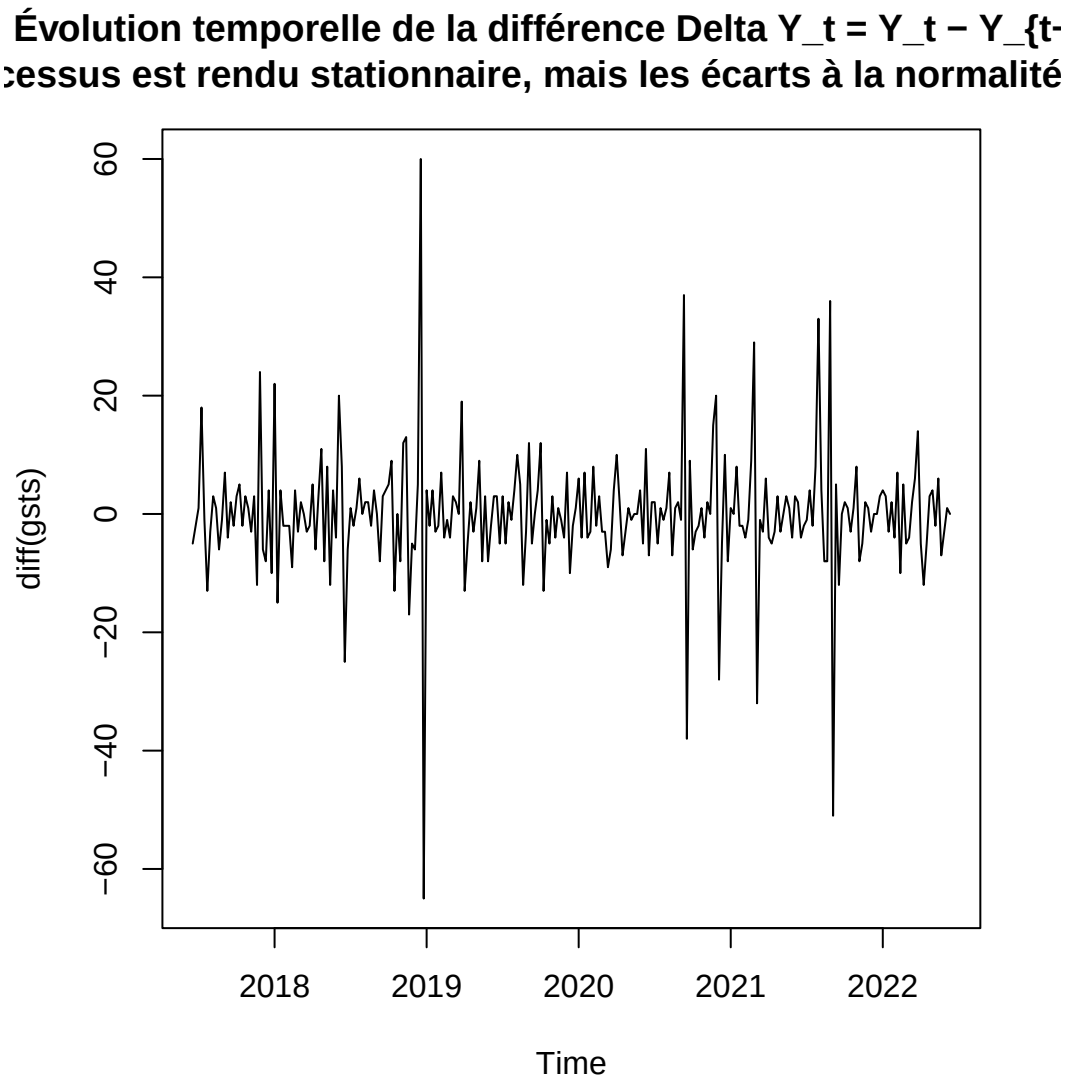
```
#construction du datasat pour série chronologique
gsts <- ts(data=gs$suicide,frequency=52,start=c(2017,24))
#affichage du graphique qui montre l'évolution de la série chronologique
plot(
    gsts,
    main = "Évolution temporelle du nombre standardisé de requêtes \nsur le moteur
     ↪  de recherche Google autour du mot « suicide »"
)
```



- On considère “à la louche” que la variance est stationnaire au cours du temps.
- Il y a peut-être une tendance à la baisse (le processus ne serait pas strictement stationnaire)
- 2 pics d’incidence sont observés, respectivement vers la fin des années 2018 et 2021 : cela peut faire craindre des problèmes d’écart à l’hypothèse de normalité dans les analyses à venir.

Pour obtenir un processus stationnaire : une solution classique est de modéliser la différence : $\Delta Y_t = Y_t - Y_{t-1}$

```
plot.ts(diff(gsts), main = "Évolution temporelle de la différence Delta Y_t = Y_t
 ↪  - Y_{t-1}. \nLe processus est rendu stationnaire, mais les écarts à la
 ↪  normalité persistent.")
```



- Le processus est beaucoup plus stationnaire
- Il persiste un pic en 2019, qui fait que la variance est pas complètement constante, et la normalité des résidues εi ne sera pas parfaite.

3. Autocorrélogramme

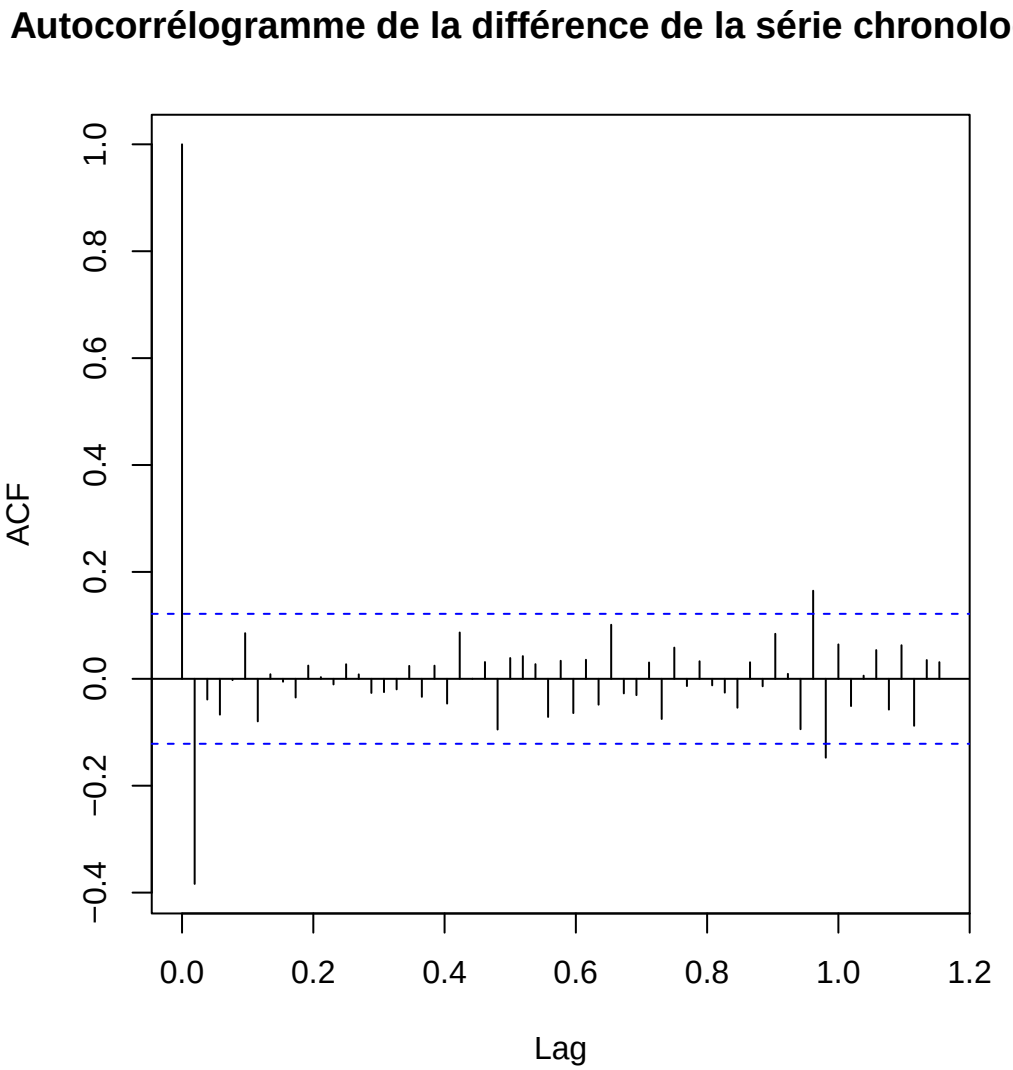
Fonction `acf` (auto correlation function) permet de tracer l'autocorrélogramme

Faite sur la différence de la série chronologique pour rendre le processus stationnaire

Syntaxe :

- `acf(x, lag.max, main)`
  - `x` = série chronologique
  - `lag.max` = nombre maximum de lags à afficher
  - `main` = titre du graphique

```
acf(diff(gsts), lag.max= 60, main="Autocorrélogramme de la différence de la série
↪  chronologique")
```



- 1ère mesure est fortement corrélée avec la mesure précédente (lag 1) (corrélation entre Y_i et Y_i donc on s'en fiche un peu)
- 2e mesure : corrélation négative Y_i et Y_{i-1} largement au dessus du seuil de significativité (pointillés bleus)
- 3e mesure : aucune corrélation

Il y a un décrochage assez net pour k > 1 (lag > 1), ce qui est en faveur d'un modèle MA(1).

Donc au moins une compensante MA(1) !

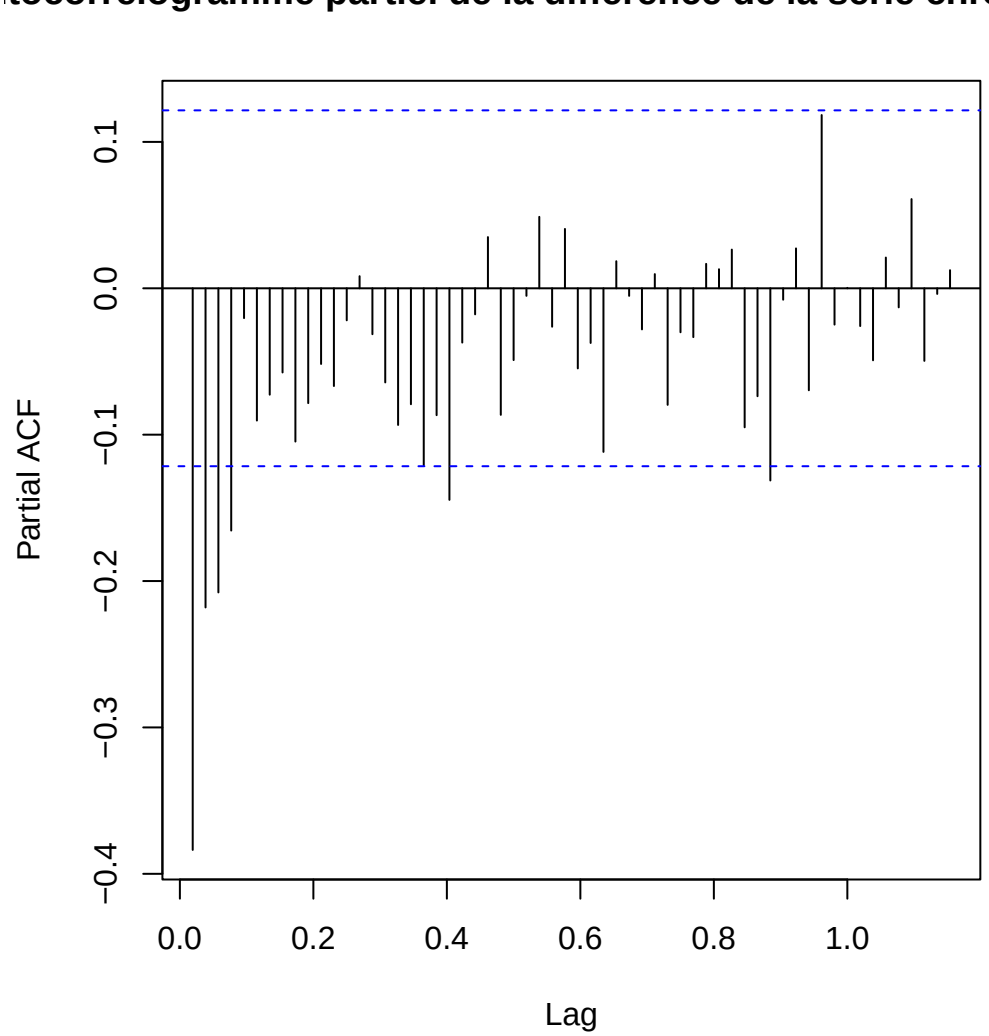
4. Autocorrélogramme partiel

Fonction `pacf` (partial auto correlation function) permet de tracer l'autocorrélogramme partiel

Syntaxe :

- `pacf(x, lag.max, main)`
  - `x` = série chronologique
  - `lag.max` = nombre maximum de lags à afficher
  - `main` = titre du graphique

```r
pacf(diff(gsts), lag.max= 60, main="Autocorrélogramme partiel de la différence de
↪  la série chronologique")
```



- 1ère corrélation partielle est élevée (lag 1)
- 2e corrélation partielle est statistiquement significative (lag 2)
- Mais chute énorme de corrélation entre 1ère, 2e…

Donc probablement une composante AR(1)

Au total : probablement un modèle ARMA(1,1)

8.B Méthode automatique avec fonction `auto.arima`

1. Sans variable exogène

Fonction `auto.arima` du package `forecast` permet de faire une sélection automatique du modèle ARIMA le plus adapté aux données.

Pour Falissard : c'est un piège car automatique et n'oblige pas à réfléchir !

Syntaxe :

- `auto.arima(y, xreg, seasonal, trace)`
  - `y` = série chronologique
  - `xreg` = matrice de variables explicatives exogènes (optionnel)
  - `seasonal` = si TRUE, recherche de composantes saisonnières (optionnel)
  - `trace` = si TRUE, affiche les modèles testés (optionnel)

```r
auto.arima(diff(gsts), trace=TRUE)
```

```
 Fitting models using approximations to speed things up...

 ARIMA(2,0,2)(1,0,1)[52] with non-zero mean : Inf
 ARIMA(0,0,0)            with non-zero mean : 1969.555
 ARIMA(1,0,0)(1,0,0)[52] with non-zero mean : 1953.522
 ARIMA(0,0,1)(0,0,1)[52] with non-zero mean : 1901.823
 ARIMA(0,0,0)            with zero mean     : 1967.529
 ARIMA(0,0,1)            with non-zero mean : 1899.763
 ARIMA(0,0,1)(1,0,0)[52] with non-zero mean : 1937.627
 ARIMA(0,0,1)(1,0,1)[52] with non-zero mean : 1939.683
 ARIMA(1,0,1)            with non-zero mean : 1889.213
 ARIMA(1,0,1)(1,0,0)[52] with non-zero mean : 1915.516
 ARIMA(1,0,1)(0,0,1)[52] with non-zero mean : 1890.644
 ARIMA(1,0,1)(1,0,1)[52] with non-zero mean : 1917.15
 ARIMA(1,0,0)            with non-zero mean : 1930.944
 ARIMA(2,0,1)            with non-zero mean : 1890.456
 ARIMA(1,0,2)            with non-zero mean : 1890.673
 ARIMA(0,0,2)            with non-zero mean : 1892.568
 ARIMA(2,0,0)            with non-zero mean : 1921.135
 ARIMA(2,0,2)            with non-zero mean : 1892.543
 ARIMA(1,0,1)            with zero mean     : 1887.24
 ARIMA(1,0,1)(1,0,0)[52] with zero mean     : 1913.441
 ARIMA(1,0,1)(0,0,1)[52] with zero mean     : 1888.642
 ARIMA(1,0,1)(1,0,1)[52] with zero mean     : 1915.07
 ARIMA(0,0,1)            with zero mean     : 1897.757
 ARIMA(1,0,0)            with zero mean     : 1928.903
 ARIMA(2,0,1)            with zero mean     : 1888.431
 ARIMA(1,0,2)            with zero mean     : 1888.639
 ARIMA(0,0,2)            with zero mean     : 1891.134
 ARIMA(2,0,0)            with zero mean     : 1919.077
 ARIMA(2,0,2)            with zero mean     : 1890.504

 Now re-fitting the best model(s) without approximations...

 ARIMA(1,0,1)            with zero mean     : 1884.453
```

```
 Best model: ARIMA(1,0,1)            with zero mean

Series: diff(gsts)
ARIMA(1,0,1) with zero mean

Coefficients:
         ar1      ma1
      0.3803  -0.9698
s.e.  0.0622   0.0196

sigma^2 = 80.35:  log likelihood = -939.18
AIC=1884.36   AICc=1884.45   BIC=1895.04
```

Ici : Modèle ARIMA(1,0,1) est sélectionné donc équivalent à ARMA(1,1)

- Avec I pour Integrated = 0 (= dérive au cours du temps) (car on a différencié la série chronologique initiale)
- (1,0,1) car :
  - p = 1 (AR(1))
  - d = 0 (Integrated)
    - d=1 : ça veut dire qu'il y a une tendance linéaire à la baisse, ce qui aurait été le cas si on avait pris 'gsts"
    - d=2 : tendance quadratique, c'est à dire courbe en U ou en cloche
  - q = 1 (MA(1))
- sans composante saisonnière (car pas de saisonnalité dans les données) (s'il y en avait une, ça aurait été noté ARIMA(p,d,q)(P,D,Q)[s] avec s = saisonnalité)

2. Avec variable exogène

On regarde s'il y a une rupture le 17/03/20 = confinement

```
1.  Création de la variable exogène
```

Variable `x` construite :

- Elle vaut –1 pour les semaines avant la semaine 145 (date du 17 Mars 2020)
- Et vaut 0 jusqu'en Juin 2022.
- Utiliser la fonction `rep` pour répéter les valeurs *(c'est à dire que `rep(-1,145)` `crée un vecteur de 145 valeurs égales à -1 et` rep(0,115)' crée un vecteur de 115 valeurs égales à 0)*

```r
x <- c(rep(-1,145), rep(0,115))
```

Puis même modèle `res` mais en y intégrant la variable exogène `x` via l'argument `xreg`

Fonction `arima` permet de faire une modélisation ARIMA avec des variables exogènes

Syntaxe :

- `arima(x, order, xreg)`
  - `x` = série chronologique
  - `order` = ordre du modèle ARIMA (p,d,q) (ici, on sait que c'est (1,0,1) d'après l'analyse précédente)
  - `xreg` = matrice de variables explicatives exogènes

```
res <- arima(diff(gsts), order = c(1,0,1), xreg = x)
res
```

```
Call:
arima(x = diff(gsts), order = c(1, 0, 1), xreg = x)

Coefficients:
         ar1      ma1  intercept       x
      0.3793  -0.9999     0.0185  0.0855
s.e.  0.0576   0.0110     0.0299  0.0472

sigma^2 estimated as 77.35:  log likelihood = -936.58,  aic = 1883.17
```

Estimation de l'effet de `x` avec son écart-type.

Si on fait le rapport de `x` sur son écart-type, on obtient une statistique de test de Wald

On le compare à une loi normale centrée réduite pour obtenir une p-value : c'est à dire qu'on le compare à 1,96 (seuil de significativité à 5%).

Le paramètre de `x`= 0.0855

L'écart-type de `x` = 0.0472

Rapport : 0.0855 / 0.0472 = 1.81 < 1.96 donc p-value > 0.05

Donc pas d'effet significatif du confinement sur les recherches Google autour du mot "suicide".

---

9 Fonction `stl()`

Fonction graphique purement exploratoire qui permet de décomposer une série chronologique en 3 composantes :

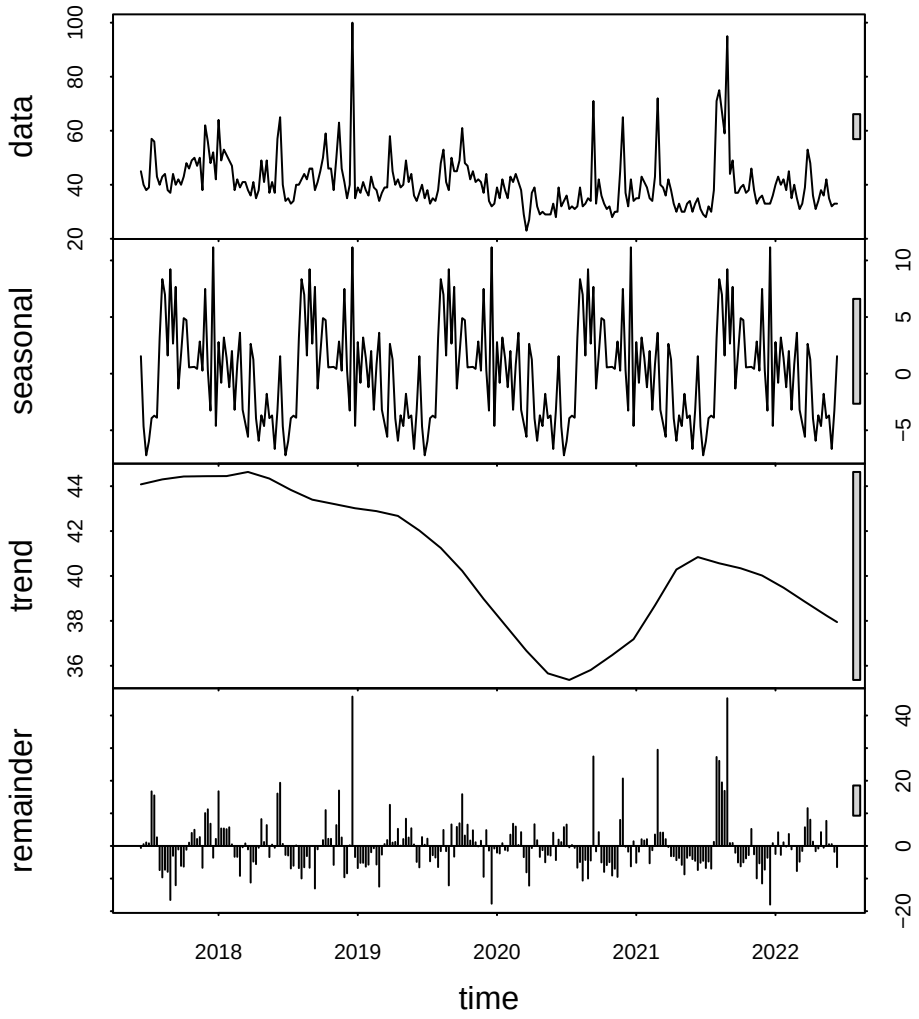
- données originales
- saisonnalité
- tendance
- résidus

Syntaxe :

- `stl(x, s.window, t.window)`
  - `x` = série chronologique
  - `s.window` = fenêtre de lissage pour la saisonnalité (ex: “periodic”)
  - `t.window` = fenêtre de lissage pour la tendance (ex: 13)

```
plot(
    stl(
        gsts,
        s.window="periodic"),
        main="Analyse graphique exploratoire d'une série chronologique à l'aide de
         ↪  la fonction `stl()`proposant une \ndécomposition en trois termes : une
         ↪  composante périodique, une composante de tendance et un résidu."
    )
```

graphique exploratoire d'une série chronologique à l'aide de la fonction ‘stl()‘propos ition en trois termes : une composante périodique, une composante de tendance et



- On observe une saisonnalité annuelle avec un pic fin Août (avant la rentrée scolaire) :::

21Series chronologiques (vidéo)

1 Introduction aux séries chronologiques

Méthodes SARIMAX / ARMA / ARIMA

- SARIMAX : Seasonal Autoregressive Integrated Moving Average with eXogenous regressors
- ARMA : Autoregressive Moving Average
- ARIMA : Autoregressive Integrated Moving Average

1.A Modèle auto-régressif d'ordre 1 :

Soit une variable Y mesurée au temps i

- Régression de Y sur ce qui se passe au temps $i-1$: $Y_(i) = (a) + bY_(i-1) + \epsilon$
  - Ce qui se passe au temps i = une fonction linéaire de ce qui se passe au temps $i-1$ + un bruit aléatoire
  - Processus auto-régressif d'ordre 1 (AR(1))

On observe une variable Y dans le temps : Y_1, Y_2, Y_3.

Le modèle AR(1) écrit :

$Y_i = a + bY_{i-1} + \varepsilon_i$

où ε_i est un bruit aléatoire indépendant d'un instant à l'autre, c'est à dire qu'il n'y a pas de corrélation entre les ε_i entre eux

1.A.1 Qu'est-ce qu'une régression (dans ce contexte) ?

Une régression est un modèle qui cherche à expliquer une variable par une autre, en supposant une relation linéaire.

Dans un AR(1), la variable explicative n'est pas une autre variable, mais l**a même variable observée juste avant**, c'est-à-dire Y_{i-1}.

C'est donc une régression temporelle, ou auto-régression :

- Y_{i-1} est l'information disponible au temps précédent,
- Y_i est ce qu'on cherche à prédire,
- ε_i représente la part de Y_i qui n'est pas expliquée par Y_{i-1}.

Autrement dit : le présent dépend linéairement du passé (= *c'est une fonction linéaire*), plus un terme aléatoire.

1.A.2 Pourquoi dit-on souvent que « $a = 0$ » ?

Le coefficient a est une constante, comme l'intercept dans une régression classique.

Cependant, dans la majorité des cours et logiciels :

1. La série est centrée avant modélisation.
   - Centrer consiste à soustraire la moyenne : $Y_i \leftarrow Y_i - \bar{Y}$.
   - Une série centrée a une moyenne égale à 0.
2. Si la série est centrée, alors le terme constant devient inutile.
   - En effet, **la constante sert justement à ajuster une moyenne non nulle**.
   - Donc, dans une série centrée : $Y_i = bY_{i-1} + \varepsilon_i$
   - et on peut prendre $a = 0$ par convention.

1.A.3 Variance : définition générale

La variance est une mesure de la dispersion d'une variable autour de sa moyenne.

- Une grande variance = la variable varie beaucoup.
- Une petite variance = la variable varie peu.

Mathématiquement :
Var$(Y) = E[(Y - E[Y])^2]$
Dans un AR(1), il y a deux sources de variance :

1. la variance du bruit ε_i
2. la propagation de ce bruit à travers le coefficient b

1.A.4 Variance du bruit : ce que c'est exactement

Le bruit ε_i est un terme aléatoire supposé :

- d'espérance 0 (= moyenne, c'està dire que le bruit n'a pas de tendance moyenne à être positif ou négatif),
- indépendant d'un instant à l'autre,
- de variance constante σ^2.

Cette variance σ^2 mesure l'amplitude des fluctuations imprévisibles.
C'est la "source" de variabilité du modèle.
Si σ^2 est grande : chaque instant reçoit un choc important.
Si σ^2 est petite : les variations imprévisibles sont faibles.

> **Note**
>
> **"Choc" lié à une variance importante :**
> **Le Rôle de l'Erreur (Le Choc Ponctuel)**
> Dans le modèle AR(1) ($_i = a + bY_{i-1} + \varepsilon_i$), l'erreur ε_i est appelée le **bruit aléatoire**.
>
> - Le bruit ε_i représente la part de la valeur actuelle Y_i qui **n'est pas expliquée** par la

valeur précédente Y_{i-1}.

- Ce bruit est la **source de variabilité** du modèle. Sa variance (σ^2) mesure l'amplitude des fluctuations imprévisibles à un instant donné.
- Dans une **régression ordinaire (non temporelle)**, cette erreur ε_i affecterait uniquement la valeur courante Y_i, puis son effet s'arrêterait là.

La Propagation Temporelle (La Mémoire)
Dans l'AR(1), l'erreur ε_i ne disparaît pas immédiatement, car la valeur Y_i qu'elle vient de modifier devient la **variable explicative** pour la période suivante.
Ainsi, l'erreur ε_i influence non seulement Y_i, mais aussi toutes les valeurs futures $Y_{i+1}, Y_{i+2}, \ldots$ à travers la relation autorégressive.

1. **L'effet sur Y_i:** Si ε_i est grand (un "choc" important), il fait monter ou descendre la valeur Y_i.
2. **L'influence sur Y_{i+1} (via b):** La valeur future Y_{i+1} dépend de Y_i via le coefficient b ($Y_{i+1} = a + b_i + \varepsilon_{i+1}$). Comme Y_i a été modifié par ε_i, l'erreur initiale ε_i se répercute sur Y_{i+1} à travers Y_i et l'influence est mesurée par b.
3. **L'influence sur Y_{i+2} (via b^2):** De même, Y_{i+2} dépend de Y_{i+1}, qui lui-même contenait déjà l'influence du choc ε_i. L'effet de l'erreur initiale ε_i est donc transmis à Y_{i+2} via b^2.
4. **L'influence sur Y_{i+k}:** L'influence de l'erreur ε_i sur une valeur lointaine Y_{i+k} est mesurée par b^k. En résumé, l'erreur ε_i se propage dans le futur, mais son impact **décroît exponentiellement** avec la distance temporelle (puisque k augmente). C'est ce que l'on appelle la **corrélation temporelle**.

L'Équilibre de la Variance Totale
La **variance** mesure la dispersion de la variable.
Puisque l'erreur ε_i ne s'arrête jamais d'influencer les valeurs futures tant que le coefficient b n'est pas nul, la variabilité totale observée dans la série Y_i est la somme de toutes les influences passées. (si b devient nul, l'effet de l'erreur s'arrête immédiatement car Y_i ne dépend plus de Y_{i-1} mais dépend uniquement de l'erreur actuelle ε_i : la nouvelle équation sera $Y_i = a + (0 \times Y_{i-1}) + \varepsilon_i$)
La **variance totale** de la série Y_i est le résultat d'un équilibre entre deux facteurs : la variance du bruit σ^2 et la mémoire du système (le coefficient b).

1. **La variabilité injectée à chaque instant** : C'est la variance du bruit σ^2 qui mesure l'amplitude des fluctuations imprévisibles de ε_i. Si σ^2 est grande, cela signifie qu'il y a une grande quantité de variabilité nouvelle et imprévisible introduite à chaque pas de temps : chaque instant reçoit un "choc" important.
2. **La quantité de mémoire du système = la propagation de cette variabilité** Ceci est donné par le coefficient b. Plus b est grand, plus les erreurs passées continuent à influencer fortement les valeurs futures parce que

Si la série est **stationnaire** (c'est-à-dire que $|b| < 1$), la variance totale est stable et finie. Elle est calculée par la formule : $Var(Y_i) = \frac{\sigma^2}{1-b^2}$ Cette formule montre clairement que la variance totale du processus est **plus grande que la variance du bruit seul** (σ2), car l'effet de la propagation des erreurs passées s'accumule (via le dénominateur $1 - b^2$).
Imaginons l'AR(1) comme un verre d'eau posé sur une table : le bruit (ε_i) est une petite goutte qui tombe dans l'eau à chaque instant (i). Dans une régression classique, l'ondulation de la goutte s'arrêterait immédiatement. Mais dans l'AR(1), si l'ondulation d'une goutte (l'erreur ε_i) est assez forte (si b est grand), elle crée un mouvement qui influence l'emplacement des gouttes suivantes (Y_{i+1}, Y_{i+2}), ce qui fait que le niveau total d'agitation de l'eau (la variance totale) est bien supérieur à la simple force d'une seule goutte (σ^2).

1.A.5 Propagation de la variance dans un AR(1)

C'est ici que l'AR(1) devient différent d'une régression classique.

Dans une régression ordinaire, l'erreur affecte uniquement la valeur courante.

Dans un AR(1), l'erreur affecte toute la suite, car :

$Y_{i+1} = a + bY_i + \varepsilon_{i+1}$

Donc si ε_i est grand :

- il fait monter ou descendre Y_i,
- ce Y_i influence Y_{i+1} via le coefficient b,
- et influence Y_{i+2} via b^2, etc.

Ainsi, les erreurs se propagent dans le futur, et la variance totale est un équilibre entre :

- la variabilité injectée à chaque instant (σ^2),
- la quantité de mémoire du système (le coefficient b).

1.A.6 Variance stationnaire : définition explicite

Une variance stationnaire signifie :

1. La variance de Y_i reste la même pour tous les instants : elle ne diverge pas, ne monte pas, ne descend pas.
2. Le processus garde une variabilité "stable" dans le temps.

Autrement dit : la série ne devient ni explosive, ni de plus en plus variable, ni de moins en moins variable.

Cette situation n'est possible que si $|b| < 1$.

- Si $|b| < 1$, les erreurs se propagent mais finissent par s'atténuer variance stable.
- Si $|b| = 1$, les erreurs se cumulent sans limite variance infinie (marche aléatoire).
- Si $|b| > 1$, la série explose variance diverge.

Dans le cas stationnaire ($|b| < 1$) :

Var$(Y_i) = \frac{\sigma^2}{1-b^2}$

Cette formule signifie :

- la variance totale est plus grande que la variance du bruit,
- plus b est grand, plus la variance de Y_i augmente,
- car plus b est grand, plus les erreurs anciennes continuent à influencer les valeurs futures.

1.A.7 Corrélation temporelle : définition explicite

La corrélation mesure l'intensité du lien linéaire entre deux valeurs.

Dans un AR(1) stationnaire:

Corr$(Y_i, Y_{i-k}) = b^k$

Donc :

- Corr$(Y_i, Y_{i-1}) = b$
- Corr$(Y_i, Y_{i-2}) = b^2$
- Corr$(Y_i, Y_{i-3}) = b^3$
- etc.

C'est une corrélation **qui décroît exponentiellement avec la distance temporelle.**

1.A.8 Exemple chiffré complet

Prenons un AR(1) :

$Y_i = 0.8Y_{i-1} + \varepsilon_i, \quad$ Var$(\varepsilon_i) = 1$

Variance stationnaire

Var$(Y_i) = \frac{1}{1-0.8^2} = \frac{1}{1-0.64} = \frac{1}{0.36} \approx 2.78$

Donc :

- la variance du processus est presque 3 fois la variance du bruit,
- car b=0.8 propage beaucoup les erreurs passées.

1.B Simulation courte

Supposons Y_0 = 10.

- $\varepsilon_1 = 0.5$

$Y_1 = 0.8 \times 10 + 0.5 = 8.5$

- `$\varepsilon_2 = -0.2$`

$Y_2 = 0.8 ×8.5 - 0.2 = 6.6

- $\varepsilon_3 = 0.1$

$Y_3 = 0.8 \times 6.6 + 0.1 = 5.38$

La série décroît progressivement vers 0 (car $|b| < 1$) et reste influencée par chaque erreur qui se propage.

1.C Corrélation temporelle

Corr$(Y_3, Y_2) = 0.8$

Corr$(Y_3, Y_1) = 0.8^2 = 0.64$

Corr$(Y_3, Y_0) = 0.8^3 = 0.512$

Résumé

Un AR(1) est un modèle où :

- le présent est expliqué par le passé (régression),
- le coefficient b crée une corrélation décroissante avec le passé,
- la variance totale résulte de l'équilibre entre :
  - la variance du bruit,
  - la propagation de cette variance via b,
- la variance stationnaire est une variance stable, finie, identique à travers le temps,
- la constante a disparaît si on travaille sur une série centrée.

---

2 Modèle MA = Moving Average

Un modèle MA(q) exprime la valeur actuelle comme une **MOYENNE MOBILE** des **ERREURS PASSÉES**.

Les **BRUITS** (ou erreurs) sont liés les uns avec les autres.

Modèle MA(1) :

$Y_j = \varepsilon_i + \varepsilon_{i-1}$

$Y_{j-1} = \varepsilon_{i-1} + \varepsilon_{i-2}$

La corrélation entre Y_j et Y_{j-1} est donc due à la présence commune de ε_{i-1} dans les deux équations. Donc corrélation ≠ 0 car il y a un terme commun.

Alors que si on fait la corrélation entre Y_j et Y_{j-2} :

$Y_j = \varepsilon_i + \varepsilon_{i-1}$

$Y_{j-2} = \varepsilon_{i-2} + \varepsilon_{i-3}$

Il n'y a pas de terme commun entre les deux équations, donc corrélation = 0.

Donc avec :

- Un modèle AR(1) : la corrélation diminue exponentiellement avec le temps = autocorrélogramme.
- Un modèle MA(1) : la corrélation est nulle au-delà du lag 1.

Donc à partir de la réprésentation de la corrélation en fonction du lag (autocorrélogramme), on peut deviner si le modèle est AR ou MA.

3 Autocorrélogramme partiel

Un autocorrélogramme partiel permet de voir la corrélation entre

- Y_i et Y_{i-p}
- ajusté sur $Y_{i-1}, Y_{i-2}, ..., Y_{i-(p-1)}$.

Corrélation spécifique ajustée entre le passé ± lointain et le présent, en ajustant sur tout ce qui se passe entre les deux.
Si on fait avec $p = 2$$:

- On regarde la corrélation entre Y_i et Y_{i-2}

On a Y_1, Y_2, Y_3 (car p=2 donc on a besoin de 3 observations pour avoir un lag de 2 entre Y_3 et Y_1)

- Processus autorégressif d'ordre 1: Y_1 agit sur Y_2 (lag 1) et Y_2 agit sur Y_3 (lag 1)
- Corrélation entre Y_1 et Y_3 vaut la corrélation entre chaque Y exponentiée par le lag : b^2
- Mais si on ajuste sur Y_2, on enlève l'effet de Y_2 sur Y_3, donc Y_1 n'a plus d'effet sur Y_3 !
- La corrélation entre Y_1 et Y_3 ajustée sur Y_2 vaut donc 0.

Donc l'autocorrélogramme partiel va donner 0 au-delà du lag 1 pour un modèle MA(1).

3.A Exemple 1

$Yi = bYi - 1 + cYi - 2 + \varepsilon_i + \varepsilon_{i-1}$
Modèle ARMA(2,1)
Dans un modèle ARMA(p, q),

- p = ordre autorégressif (AR) → nombre de termes Y retardés,
- q = ordre moyenne mobile (MA) → nombre de termes ε retardés.

1. Partie AR (autorégressive)
   - Y_{i-1}
   - Y_{i-2}

Donc 2 termes AR → p = 2.

2. Partie MA (moyenne mobile)
   - ε_i → toujours présent, ce n'est pas compté dans q

- ε_{i-1}

Le terme ε_i est l'erreur courante : on ne le compte pas dans l'ordre.

On compte seulement les retards.

Donc il y a 1 terme MA retardé q = 1.

3.B Exemple 2

Mesure de la TA tous les matins

Y_i = TA du jour i : il y en a donc 365

3.B.1 Modèle autorégressif AR(1)

Les Y_i ont une variance = 1 c'est à dire que la TA varie peu d'un jour à l'autre.

$\rightarrow Y_i = bY_{i-1} + \varepsilon_i$

- On peut pas faire une régression linéaire habituelle car Y_i est à la fois à gauche et à droite de l'équation
- si variance = 1 / variance stationnaire :
  - b = corrélation entre Y_i et Y_{i-1} quelque soit i
  - avec le temps : corrélation entre Y_i et Y_{i-2} = b^2
  - corrélation entre Y_i et Y_{i-k} = b^k
  - avec le temps, la corrélation tend vers 0 de façon exponentielle

3.B.2 Modèle moyenne mobile MA(1)

$Y_i = \varepsilon_i + \varepsilon_{i-1}$

- 2 bruits : régression linéaire impossible
- Le temps i et le temps $i-1$ n'ont rien à voir l'un avec l'autre et sont liés au bruit
  - Appareil de mesure a aussi un bruit et a un défaut ponctuel à un moment donné
  - $*$ stress du jour qui influe sur le bruit du jour aussi

Globalement : dans la mesure du jour, le bruit observé vient d'un bruit issu de la mesure du jour + du bruit d'hier qui vient d'un contexte, influençant l'erreur de mesure qu'il y a dans Y_i aujourd'hui.

- Corrélation entre Y_i et Y_{i-1} = ϱ et est due au fait qu'il y a un terme commun dans les 2 équations : ε_{i-1}
- Corrélation entre Y_i et Y_{i-2} = 0 car pas de terme commun dans l'équation

3.B.3 Autocorrélogramme

Permet assez facilement de visualiser si le modèle est AR ou MA

3.B.4 Autocorrélogramme partiel

Corrélation entre Y_i et Y_{i-k} ajustée sur $Y_{i-1}, Y_{i-2}, ..., Y_{i-(k-1)}$ (ajusté sur toutes les mesures qu'on a entre les 2)

Si modèle AR(1) :

- Y_1 agit sur Y_2 et Y_2 agit sur Y_3
- Mais Y_1 agit sur Y_3 uniquement via Y_2
- Donc si on ajuste sur Y_2, la corrélation entre Y_1 et Y_3 ajustée sur Y_2 = 0

Donc modèle AR(1) dans autocorrélogramme partiel ressemble au modèle MA(1) dans autocorrélogramme classique.

3.B.5 Processus stationnaire

= Processus dont les propriétés statistiques (moyenne, variance, autocorrélation) ne changent pas au cours du temps.

4 Différents modèles : AR, MA, ARMA, SARMA, SARIMA, SARIMAX

On part d'un processus stationnaire.

1. **Description** : évolution au cours du temps => MOYENNE STABLE ? VARIANCE STABLE ?
2. **Modélisation** : autocorrélogramme et autoccorrélogramme partiel => modèle AR, MA ou ARMA ?
    - voire SARMA : saisonnier (avec saisonnalité)
    - voire SARIMA : avec dérive au cours du temps (écart à la stationnarité : I = Integrated c'est à dire qu'on intègre une tendance linéaire au modèle)
    - voire SARIMAX : avec des variables explicatives (exogènes)
        - exogènes = variables explicatives externes au modèle de la série chronologique
        - $Y_i = bY_{i-1} + cY_{i-2} + \varepsilon_i + \varepsilon_{i-1} + d\vartheta$

---

5 Exemple d'analyse de données avec R

5.A Méthode graphique

Le fichier gs contient des données sur les tendances de recherche Google (sur recherche = suicide entre Juin 2017 et Juin 2022).

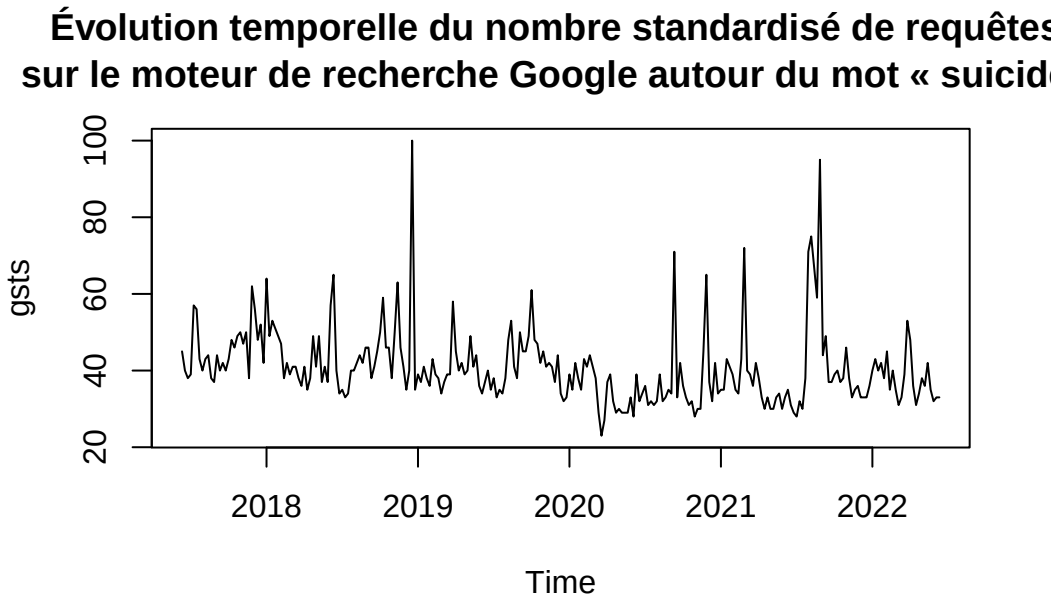
Voir si indice X = survenue du confinement modifie la recherche du mot "suicide"

Modèle = ARMAX (avec X = facteur exogène = confinement)

- avec X

1. Transformation de la variable `gs$suicide` en une série chronologique grâce à la fonction `ts` = timeserie
   - si pas de fonction `ts`: on peut pas faire d'analyses sur séries chronologiques avec R.
   - syntaxe de la fonction `ts` :
     - `ts(data, frequency, start)`
     - `data` = données,
     - `frequency` = fréquence (ex: 52 semaines dans une année),
       * permet de dire que la semaine i et la semaine $i + 52$ sont comparables (même semaine de l'année)
     - `start` = date de début (ex: c(2017,24))
2. Affichage de la série chronologique avec la fonction `plot`

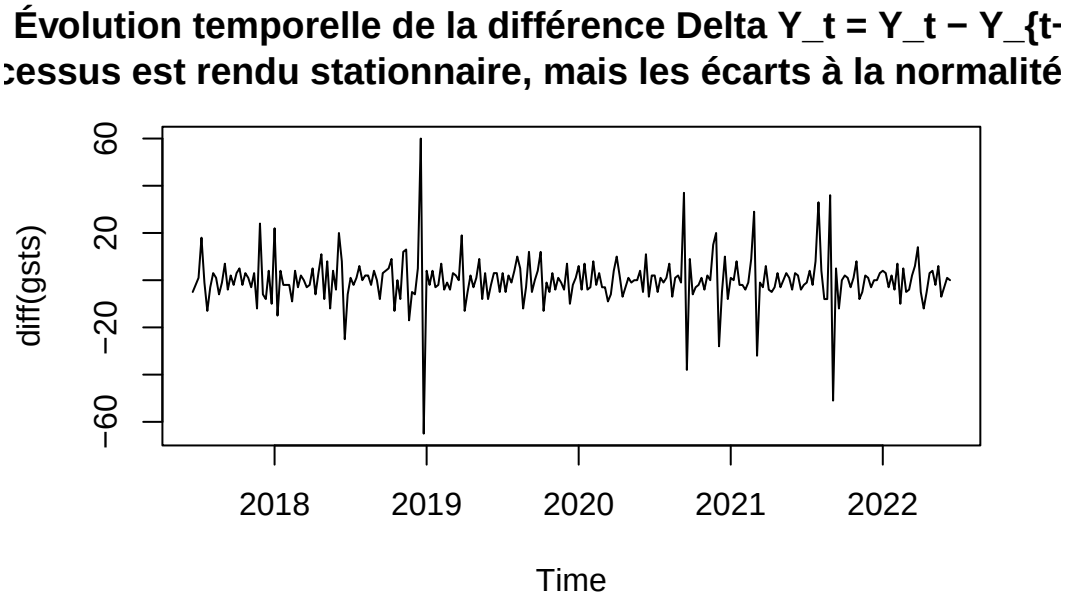
```
#construction du datasat pour série chronologique
gsts <- ts(data=gs$suicide,frequency=52,start=c(2017,24))
#affichage du graphique qui montre l'évolution de la série chronologique
plot(
    gsts,
    main = "Évolution temporelle du nombre standardisé de requêtes \nsur le moteur
     ↪  de recherche Google autour du mot « suicide »"
)
```



- On considère "à la louche" que la variance est stationnaire au cours du temps.
- Il y a peut-être une tendance à la baisse (le processus ne serait pas strictement stationnaire)
- 2 pics d'incidence sont observés, respectivement vers la fin des années 2018 et 2021 : cela peut faire craindre des problèmes d'écart à l'hypothèse de normalité dans les analyses à venir.

Pour obtenir un processus stationnaire : une solution classique est de modéliser la différence : $\Delta Y_t = Y_t - Y_{t-1}$

```
plot.ts(diff(gsts), main = "Évolution temporelle de la différence Delta Y_t = Y_t
↪  - Y_{t-1}. \nLe processus est rendu stationnaire, mais les écarts à la
↪  normalité persistent.")
```



- Le processus est beaucoup plus stationnaire
- Il persiste un pic en 2019, qui fait que la variance est pas complètement constante, et la normalité des résidues εi ne sera pas parfaite.

3. Autocorrélogramme

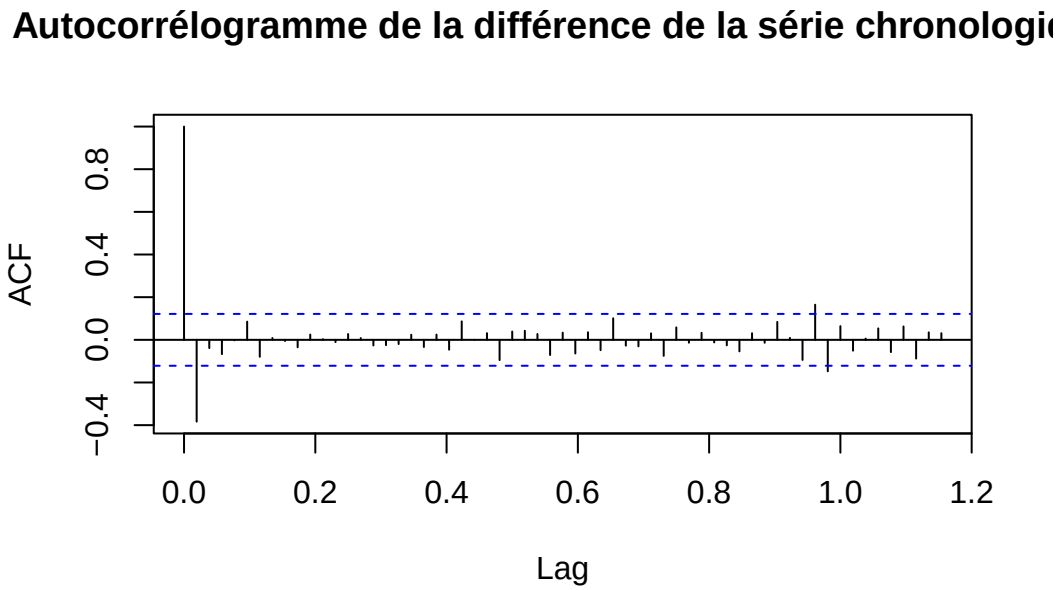
Fonction `acf` (auto correlation function) permet de tracer l'autocorrélogramme

Faite sur la différence de la série chronologique pour rendre le processus stationnaire

Syntaxe :

- `acf(x, lag.max, main)`
  - `x` = série chronologique
  - `lag.max` = nombre maximum de lags à afficher
  - `main` = titre du graphique

```
acf(diff(gsts), lag.max= 60, main="Autocorrélogramme de la différence de la série
↪  chronologique")
```



- 1ère mesure est fortement corrélée avec la mesure précédente (lag 1) (corrélation entre Y_i et Y_i donc on s'en fiche un peu)
- 2e mesure : corrélation négative Y_i et Y_{i-1} largement au dessus du seuil de significativité (pointillés bleus)
- 3e mesure : aucune corrélation

Il y a un décrochage assez net pour k > 1 (lag > 1), ce qui est en faveur d'un modèle MA(1).

Donc au moins une compensante MA(1) !

4. Autocorrélogramme partiel

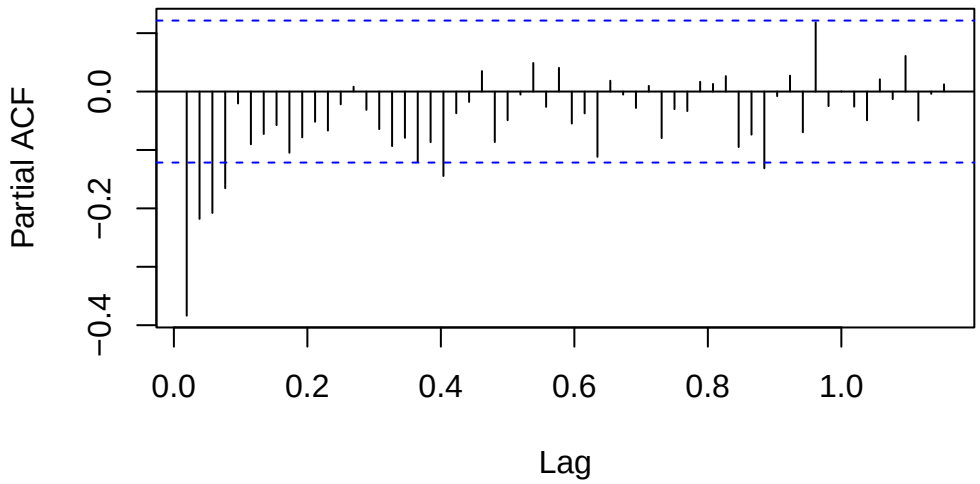
Fonction `pacf` (partial auto correlation function) permet de tracer l'autocorrélogramme partiel

Syntaxe :

- `pacf(x, lag.max, main)`
  - `x` = série chronologique
  - `lag.max` = nombre maximum de lags à afficher
  - `main` = titre du graphique

```
pacf(diff(gsts), lag.max= 60, main="Autocorrélogramme partiel de la différence de
↪  la série chronologique")
```

utocorrélogramme partiel de la différence de la série chronol



- 1ère corrélation partielle est élevée (lag 1)
- 2e corrélation partielle est statistiquement significative (lag 2)
- Mais chute énorme de corrélation entre 1ère, 2e…

Donc probablement une composante AR(1)

Au total : probablement un modèle ARMA(1,1)

5.B Méthode automatique avec fonction `auto.arima`

1. Sans variable exogène

Fonction `auto.arima` du package `forecast` permet de faire une sélection automatique du modèle ARIMA le plus adapté aux données.

Pour Falissard : c'est un piège car automatique et n'oblige pas à réfléchir !

Syntaxe :

- `auto.arima(y, xreg, seasonal, trace)`
  - `y` = série chronologique
  - `xreg` = matrice de variables explicatives exogènes (optionnel)
  - `seasonal` = si TRUE, recherche de composantes saisonnières (optionnel)
  - `trace` = si TRUE, affiche les modèles testés (optionnel)

```
auto.arima(diff(gsts), trace=TRUE)
```

```
Fitting models using approximations to speed things up...

ARIMA(2,0,2)(1,0,1)[52] with non-zero mean : Inf
```

```
 ARIMA(0,0,0)            with non-zero mean : 1969.555
 ARIMA(1,0,0)(1,0,0)[52] with non-zero mean : 1953.522
 ARIMA(0,0,1)(0,0,1)[52] with non-zero mean : 1901.823
 ARIMA(0,0,0)            with zero mean     : 1967.529
 ARIMA(0,0,1)            with non-zero mean : 1899.763
 ARIMA(0,0,1)(1,0,0)[52] with non-zero mean : 1937.627
 ARIMA(0,0,1)(1,0,1)[52] with non-zero mean : 1939.683
 ARIMA(1,0,1)            with non-zero mean : 1889.213
 ARIMA(1,0,1)(1,0,0)[52] with non-zero mean : 1915.516
 ARIMA(1,0,1)(0,0,1)[52] with non-zero mean : 1890.644
 ARIMA(1,0,1)(1,0,1)[52] with non-zero mean : 1917.15
 ARIMA(1,0,0)            with non-zero mean : 1930.944
 ARIMA(2,0,1)            with non-zero mean : 1890.456
 ARIMA(1,0,2)            with non-zero mean : 1890.673
 ARIMA(0,0,2)            with non-zero mean : 1892.568
 ARIMA(2,0,0)            with non-zero mean : 1921.135
 ARIMA(2,0,2)            with non-zero mean : 1892.543
 ARIMA(1,0,1)            with zero mean     : 1887.24
 ARIMA(1,0,1)(1,0,0)[52] with zero mean     : 1913.441
 ARIMA(1,0,1)(0,0,1)[52] with zero mean     : 1888.642
 ARIMA(1,0,1)(1,0,1)[52] with zero mean     : 1915.07
 ARIMA(0,0,1)            with zero mean     : 1897.757
 ARIMA(1,0,0)            with zero mean     : 1928.903
 ARIMA(2,0,1)            with zero mean     : 1888.431
 ARIMA(1,0,2)            with zero mean     : 1888.639
 ARIMA(0,0,2)            with zero mean     : 1891.134
 ARIMA(2,0,0)            with zero mean     : 1919.077
 ARIMA(2,0,2)            with zero mean     : 1890.504

 Now re-fitting the best model(s) without approximations...

 ARIMA(1,0,1)            with zero mean     : 1884.453

 Best model: ARIMA(1,0,1)            with zero mean

Series: diff(gsts)
ARIMA(1,0,1) with zero mean

Coefficients:
         ar1      ma1
      0.3803  -0.9698
s.e.  0.0622   0.0196

sigma^2 = 80.35:  log likelihood = -939.18
AIC=1884.36   AICc=1884.45   BIC=1895.04
```

Ici : Modèle ARIMA(1,0,1) est sélectionné donc équivalent à ARMA(1,1)

- Avec I pour Integrated = 0 (= dérive au cours du temps) (car on a différencié la série chronologique initiale)
- (1,0,1) car :
  - p = 1 (AR(1))
  - d = 0 (Integrated)
    - d=1 : ça veut dire qu'il y a une tendance linéaire à la baisse, ce qui aurait été le cas si on avait pris 'gsts"
    - d=2 : tendance quadratique, c'est à dire courbe en U ou en cloche
  - q = 1 (MA(1))
- sans composante saisonnière (car pas de saisonnalité dans les données) (s'il y en avait une, ça aurait été noté ARIMA(p,d,q)(P,D,Q)[s] avec s = saisonnalité)

2. Avec variable exogène

On regarde s'il y a une rupture le 17/03/20 = confinement

`1. Création de la variable exogène`

Variable `x` construite :

- Elle vaut –1 pour les semaines avant la semaine 145 (date du 17 Mars 2020)
- Et vaut 0 jusqu'en Juin 2022.
- Utiliser la fonction `rep` pour répéter les valeurs *(c'est à dire que `rep(-1,145)` crée un vecteur de 145 valeurs égales à -1 et `rep(0,115)`' crée un vecteur de 115 valeurs égales à 0)*

```
x <- c(rep(-1,145), rep(0,115))
```

Puis même modèle `res` mais en y intégrant la variable exogène `x` via l'argument `xreg`

Fonction `arima` permet de faire une modélisation ARIMA avec des variables exogènes

Syntaxe :

- `arima(x, order, xreg)`
  - `x` = série chronologique
  - `order` = ordre du modèle ARIMA (p,d,q) (ici, on sait que c'est (1,0,1) d'après l'analyse précédente)
  - `xreg` = matrice de variables explicatives exogènes

```
res <- arima(diff(gsts), order = c(1,0,1), xreg = x)
res
```

```
Call:
```

```
arima(x = diff(gsts), order = c(1, 0, 1), xreg = x)

Coefficients:
         ar1      ma1  intercept       x
      0.3793  -0.9999     0.0185  0.0855
s.e.  0.0576   0.0110     0.0299  0.0472

sigma^2 estimated as 77.35:  log likelihood = -936.58,  aic = 1883.17
```

Estimation de l'effet de `x` avec son écart-type.

Si on fait le rapport de `x` sur son écart-type, on obtient une statistique de test de Wald

On le compare à une loi normale centrée réduite pour obtenir une p-value : c'est à dire qu'on le compare à 1,96 (seuil de significativité à 5%).

Le paramètre de `x`= 0.0855

L'écart-type de `x` = 0.0472

Rapport : 0.0855 / 0.0472 = 1.81 < 1.96 donc p-value > 0.05

Donc pas d'effet significatif du confinement sur les recherches Google autour du mot "suicide".

---

6 Fonction `stl()`

Fonction graphique purement exploratoire qui permet de décomposer une série chronologique en 3 composantes :

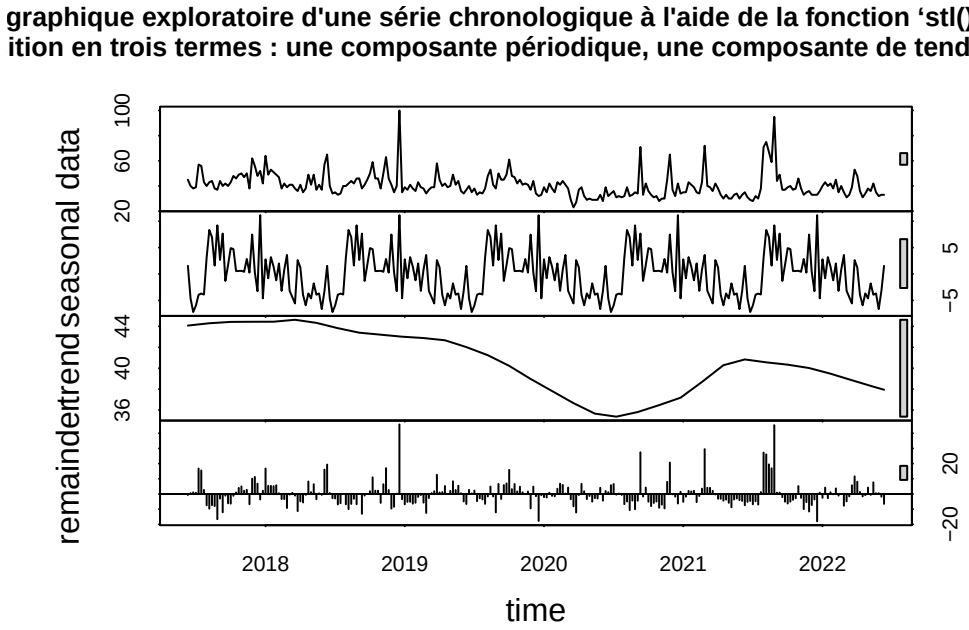
- données originales
- saisonnalité
- tendance
- résidus

Syntaxe :

- `stl(x, s.window, t.window)`
  - `x` = série chronologique
  - `s.window` = fenêtre de lissage pour la saisonnalité (ex: "periodic")
  - `t.window` = fenêtre de lissage pour la tendance (ex: 13)

```
plot(
    stl(
        gsts,
        s.window="periodic"),
```

```
        main="Analyse graphique exploratoire d'une série chronologique à l'aide de
         ↪  la fonction `stl()`proposant une \ndécomposition en trois termes : une
         ↪  composante périodique, une composante de tendance et un résidu."
    )
```



- On observe une saisonnalité annuelle avec un pic fin Août (avant la rentrée scolaire) :::

22Comparaisons Multiples

1 Introduction

- **Fisher** : calcule les petits *p* pour chaque comparaison.
  - Il fait ainsi de l'**inférence inductive**, c'est à dire qu'il utilise chacune des *p-values* pour en déduire des nouvelles connaissances.
  - Les observations de départ sont vraies, mais les conclusions sont probables
  - Génomique : 1 million de petits *p* : ça ne rentrera jamais dans la tête !
- **Neyman-Pearson** : calcule une **seule** *p-value* pour l'ensemble des comparaisons.
  - Hypothèse nulle H0 et hypothèse alternative H1
  - Risque de première espèce (alpha) et risque de deuxième espèce (beta)
    * α : probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie
    * β : probabilité de ne pas rejeter l'hypothèse nulle alors qu'elle est fausse
  - Si on augmente le nombre d'hypothèses, on augmente le risque alpha (parce que 5% à chaque fois !)

> ❗ Important
>
> Donc risque avec les comparaisons multiples dans la théorie de Neyman et Pearson, mais pas vraiment dans la théorie de Fisher.

Adjusting for multiple testing—when and how?

2 Quand ajuster ?

- Dans Neyman et Pearson = dans les essais randomisés contrôlés

3 Comment ajuster ?

3.A Exemple

Dans l'asthme, on évalue le VEMS et la QoL.

- 1ère situation : Médicament bon si VEMS significatif **ET** QoL significatif
- 2e situation : Médicament bon si VEMS significatif **OU** QoL significatif
  - Problème : 2 tests indépendants, chacun avec un alpha de 5%

3.B Méthode de Bonferroni = répartition du risque alpha

Méthode de Bonferonni = la plus classique.

Principe : diviser le risque alpha par le nombre de tests effectués.

Dans le cas de l'exemple : 2 tests, donc alpha = 0.05/2 = 0.025

Problème :

- perte de puissance !
- Critères d'efficacité souvent corrélés

3.C Méthode de Holm = répartition du risque alpha

Méthode de Holm = plus puissante que Bonferroni.

Principe : ajuster le seuil alpha en fonction du rang des p-values.

1. Ordonner les p-values de la plus petite à la plus grande
2. Comparer chaque p-value au seuil alpha divisé par le nombre de tests restants

Exemple : 4 tests, alpha = 0.05

Rang	p-value	Seuil alpha ajusté	Décision
1	0.01	0.05/4 = 0.0125	Rejeter H0
2	0.03	0.05/3 = 0.0167	
3	0.04	0.05/2 = 0.025	Ne pas rejeter H0
4	0.06	0.05/1 = 0.05	Ne pas rejeter H0

3.D Méthode par hiérarchie

Si critère principal (Survie sans progression) et critères secondaires (Taux CCR, Survie Brute, Qualité de vie)

Hiérarchisation :

1. Tester le critère principal au seuil alpha
2. Si significatif, tester le premier secondaire au seuil alpha ; sinon stop !
3. Si significatif, tester le second secondaire au seuil alpha ; sinon stop !

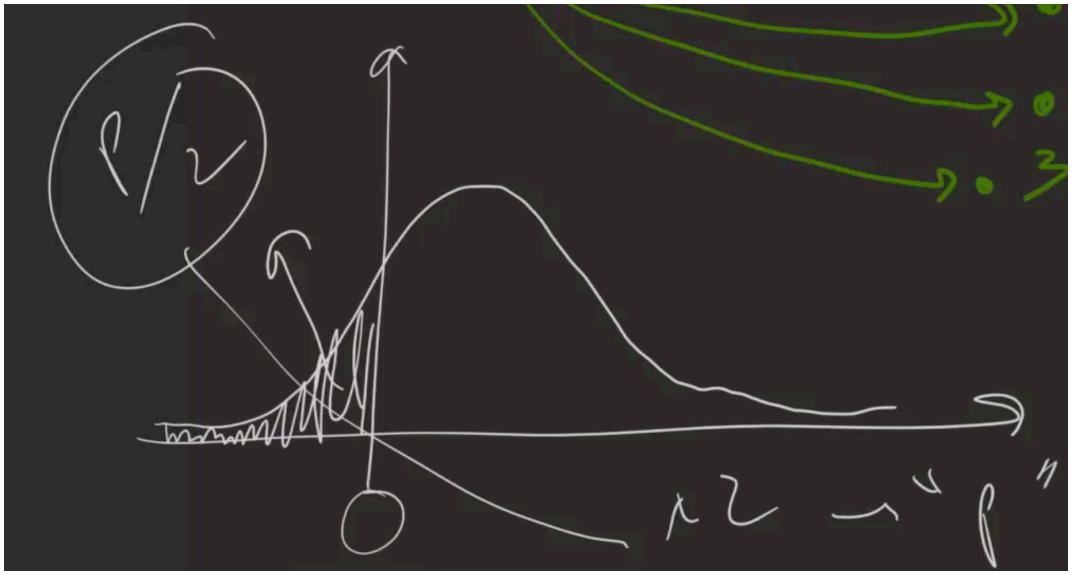
4 En dehors des essais thérapeutiques

4.A Génomique

- Très grand nombre de tests
- Correction nécessaire !
  - Bonferroni : possible mais trop conservateur

  - FDR : False Discovery Rate (Benjamini-Hochberg)
    * Contrôle le taux de fausses découvertes parmi les résultats significatifs
    * Par exemple 20 marqueurs significatifs avec FDR à 15%
    * Signifie que 3 marqueurs (15% de 20) sont susceptibles d'être des faux positifs
    * On maîtrise pas le risque de première espèce sur chaque test individuel, mais le pourcentage de faux positifs parmi les tests déclarés significatifs
- Ou référence inductive ! $ *"ce sont des pistes intéressantes"*

5 Questions

1. Quand on recherche des interactions
   - Par exemple 70 interactions
   - Si on veut **affirmer** qu'il y en a une significative, garder 0,05 n'a pas de sens
   - Solution :
     - Classer les p-values
     - Et les mettre sur une courbe
     - certaines p values sont /plus petites que ne le voudrait le hasard
     - Cox et Wermuth
     - 

23Donnees manquantes

1 Introduction aux données manquantes

1. Considérer le problème = **décrire les données manquantes** : combien ? où ? pourquoi ?
2. Gestion des données manquantes
   - **Virer les sujets avec données manquantes** (listwise deletion) = `na.rm=TRUE` : bof notamment si rajout de variables dans un modèle -> perte de sujets car modèle ajusté se fait sur les sujets avec données complètes
   - **Imputation des données manquantes** (remplir les valeurs manquantes avec des estimations)
     - **Imputation simple** : mode (valeur la plus fréquente) pour les variables catégorielles, moyenne/médiane pour les variables continues (Library `Hmisc` : fonction `impute()` (imputation simple))
     - **Imputation multiple** : créer plusieurs jeux de données imputées, analyser chaque jeu de données, combiner les résultats (Library `mice` : fonction `mice()` (imputation multiple))

2 Types de données manquantes

2.A MCAR : Missing Completely At Random

Les données manquantes sont indépendantes des valeurs observées et non observées. Exemple : perte de questionnaire.

= Données manquantes **complètement au hasard**

2.B MAR : Missing At Random

Les données manquantes **dépendent des valeurs observées** mais **pas des valeurs non observées**.

Exemple :

- les personnes âgées sont moins susceptibles de remplir un questionnaire en ligne.
- les patients très musulmans répondent moins sur leur consommation d'alcool, mais on connaît leur âge et sexe.
  - donc religion TRUE –> + de données manquantes et moins d'alcool
  - religion FALSE –> peu de données manquantes, + d'alcool
  - si on a religion dans le questionnaire : c'est ok ! sinon on sait pas pourquoi alcool est manquant (biais dans la donnée manquante)

Une fois qu'on enlève l'effet de toutes les variables mesurées susceptibles d'influencer à la fois

la réponse d'une variable mesurée et à la fois la probabilité que cette variable soit manquante, ce qui reste dans la valeur manquante de la donnée : c'est du bruit

Autrement dit : une fois ajusté sur les variables mesurées pertinentes, le caractère manquant d'une observation ne dépend plus de sa valeur non observée.

> ❗ Important
>
> **MAR** : le fait qu'une donnée soit manquante dépend uniquement des autres variables observées dans le jeu de données, pas de sa propre valeur non observée.
> Une fois ces variables observées prises en compte, le caractère manquant devient aléatoire.

2.C MNAR : Missing Not At Random

Les données manquantes **dépendent des valeurs non observées**.

Exemple : les patients avec une dépression sévère sont moins susceptibles de remplir un questionnaire sur la dépression.

= Données manquantes **non au hasard**

Pour faire la différence entre MCAR et MNAR :

- Les tests statistiques ne marchent pas terrible
- Souvent, on fait des hypothèses basées sur la connaissance du domaine

3 Imputation multiple

Surtout si bcp de données ou manquantes ou si données manquantes sur données importantes

3.A Principe de l'imputation multiple

Dans un modèle :

$Y = a + bX_1 + cX_2 + \epsilon$

On s'intéresse au paramètre b.

Jeu de données :

ID	Y	X1	X2
1	2.3	1.2	NA
2	3.1	NA	0.5
3	NA	0.8	1.1
4	4.0	1.5	0.9
5	2.8	1.0	0.7

Pour faire une imputation multiple :

3.A.1 1ère étape : tirage au sort

- Tirer au sort la valeur manquante de Y par une autre valeur de Y (`df[3, "Y"]` est manquant, on le remplace par `df[5, "Y"]` = 2.8)
- Idem pour X1 : `df[2, "X1"]` est manquant, on le remplace par `df[1, "X1"]` = 1.2
- Idem pour X2 : `df[1, "X2"]` est manquant, on le remplace par `df[5, "X2"]` = 0.7

On obtient un jeu de données **INTERMÉDIAIRE** qui permet de réaliser une régression complète

3.A.2 2ème étape : régressions

3 régressions sur le jeu de données intermédiaire :

- $Y[3] = a + bX_1[3] + cX_2[3] + \epsilon$ permet d'obtenir la valeur estimée de Y3
- $X_1[2] = a' + b'Y[2] + c'X_2[2] + \epsilon'$ permet d'obtenir la valeur estimée de X12
- $X_2[1] = a'' + b''Y[1] + c''X_1[1] + \epsilon''$ permet d'obtenir la valeur estimée de X21

Ainsi on obtient un jeu premier jeu données **IMPUTÉ** mais pas encore **convergent**

Convergent : plus on impute, plus les jeux de données se ressemblent

3.A.3 3ème étape : itérations

On recommence les étapes 1 et 2 plusieurs fois (10 à 20 fois) pour obtenir un jeu de données **IMPUTÉ CONVERGENT**

Quand le jeu de données a convergé, il est parfaitement bien imputé car toutes les données manquantes sont construites à partir de toutes les données disponibles sur les autres variables.

3.A.4 4ème étape : répétition

- Il faut **REFAIRE LE TIRAGE AU SORT** de départ pour obtenir un nouveau jeu de données intermédiaire
- Puis refaire les étapes 2 et 3 pour obtenir un **NOUVEAU JEU DE DONNÉES IMPUTÉ CONVERGENT**

On a donc un **DEUXIÈME JEU DE DONNÉES IMPUTÉ CONVERGENT**

Puis répétition N fois (en général 5 à 10 fois) pour obtenir N jeux de données imputés convergents.

3.A.5 5ème étape : analyse

On réalise l'analyse statistique (régression linéaire, logistique, etc.) sur chaque jeu de données imputé convergent.

jeu ' : $Y^{(')} = a^{(')} + b^{(')}X_1 + c^{(')}X_2 + \epsilon^{(')}$

jeu 1 : $Y^{(1)} = a^{(1)} + b^{(1)}X_1 + c^{(1)}X_2 + \epsilon^{(1)}$
jeu 20 : $Y^{(20)} = a^{(20)} + b^{(20)}X_1 + c^{(20)}X_2 + \epsilon^{(20)}$
On obtient donc

- 20 estimations différentes de b.
- avec 20 variances de b différentes. (variance inter-modèle)
- et une variance de b1 à b20 = variance intra-modèle

Prendre :

- La moyenne des 20 estimations de b comme estimation finale de b.
- – variance inter-modèle + variance intra-modèle pour obtenir la variance finale de b.

b moyen = estimation non biaisée de b
somme des variances intra + inter-modèle = variance non biaisée de b

3.B Gérer les analyses après imputation multiple

20 jeu de données imputés !

24Bootstrap

1 Introduction

Définition du bootstrap = technique de rééchantillonnage avec remise à partir d'un échantillon initial pour estimer la distribution d'un estimateur.

Utile pour petits effectifs ou lorsque les hypothèses classiques ne sont pas respectées.

En recherche clinique : souvent faible nombre de sujets et variance importantes.

- Si 2*10 sujets :
  - ok pour tests non paramétriques : Wilcoxon et Mann-Whitney
  - bof pour tests paramétriques (car difficile de dire que ça suit une loi normale alors qu'on a peu de données) : t de Student, chi2, Fisher…
- Exemple si n = 100 et distribution très asymétrique :
  - test t pas terrible !!!
  - analyse de variance ou régression linéaire : il faudrait aussi une distribution normale !
- Solution : bootstrap ! = rééchantillonner avec remise à partir de l'échantillon initial pour estimer la distribution de l'estimateur.

2 Principe du bootstrap

Imaginons qu'on veuille faire la régression :

$$Y = a + bX1 + cX2 + \epsilon$$

Sauf que Y ne suit pas du tout une loi normale.

On reprend le jeu de données

ID	Y	X1	X2
1			
2			
3			
…			
100			

2.A Tirage avec remise

On tire au hasard avec remise 100 observations parmi les 100 initiales, c'est à dire :

- On en tire au sort 1 : **le numéro 28**

- On le remet dans le sac
- On en tire au sort un deuxième : **le numéro 77**
- On le remet dans le sac
- On en tire au sort un troisième : **le numéro 28**

ça donnera un tableau :

ID	Y	X1	X2
28			
77			
28			
...			
13			

Donc nouveau jeu de données !!
Sur lequel on fait à nouveau la régression linéaire :

$$Y = a + bX1 + cX2 + \epsilon$$

$\rightarrow$ on met le b en mémoire !
On refait la même chose !!

ID	Y	X1	X2
6			
82			
15			
...			
6			

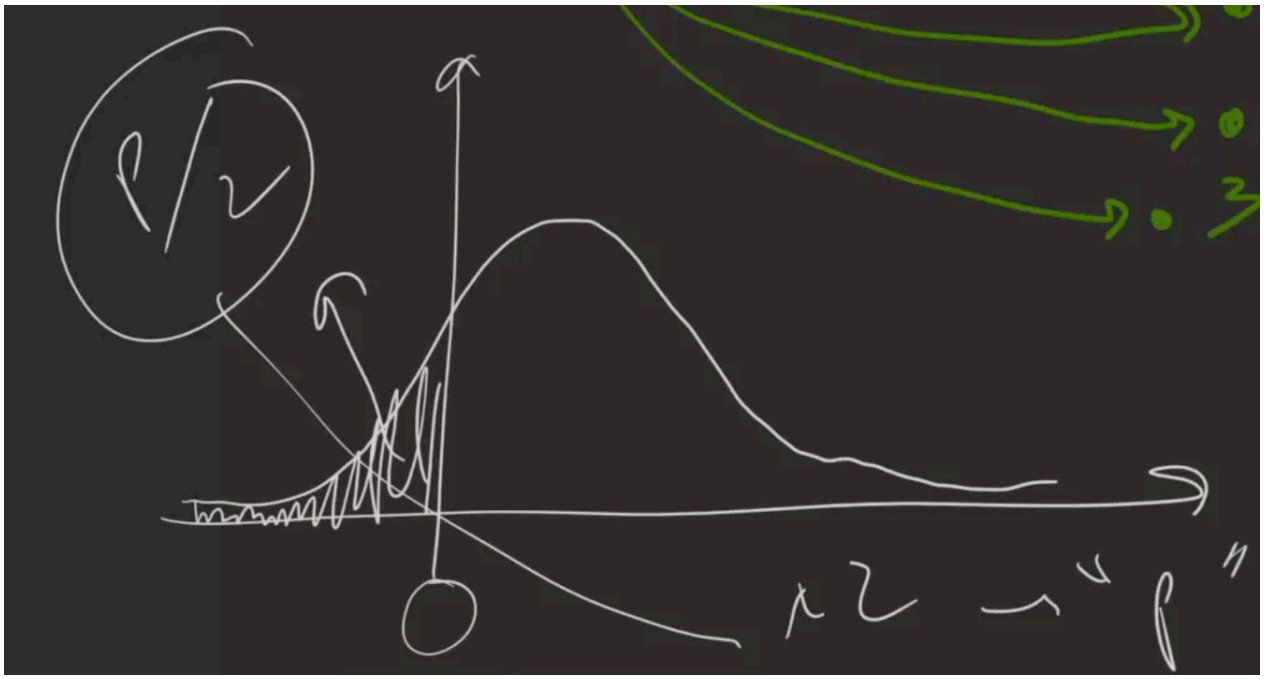
Nouveau modèle de régression linéaire :

$$Y = a + bX1 + cX2 + \epsilon$$

$\rightarrow$ on obtient un nouveau b qu'on garde aussi
Et ainsi de suite !!!
On fait ça 10 000 fois (au delà de 10 000, pas vraiment interessant)

2.B Histograme des 10 000 valeurs de b

Objectif : test de b = 0 (c'est à dire pas d'effet de X1 sur Y)
On trace l'histogramme des 10 000 valeurs de b obtenues



- On prend les deux extrémités (2.5% et 97.5%) pour obtenir un intervalle de confiance à 95%
- L'intervalle de confiance n'a pas besoin de la normalité de la distribution car fait à la main directement à partir des données !
- Combien de b < 0 donne la p value divisée par 2 multipliée par 2 donne le vrai petit p du test b = 0 (c'est à dire pas d'effet de X1 sur Y)

QUESTIONS :

- Échantillons sont indépendants ?
  - Oui ! car on tire au hasard avec remise donc chaque tirage est indépendant des autres
  - C'est un tirage de 100 sujets parmi la population infinie !
- Machine learning ?
  - Non : c'est plutôt une simulation statistique
- Nombre de régressions ?
  - 10 000 régressions

2.C Faisabilité du bootstrap

A partir de 30 - 50 sujets, le bootstrap est fiable.

100 : on est serein !

3 Exemple en R sur expsy

1. Librairies nécessaires

```r
library(boot)
```

```
Attaching package: 'boot'

The following object is masked from 'package:survival':

    aml
```

```r
library(psy)

# fabriquer une fonction pour faire le bootstrap de kappa
# fonction entre accolades
# x, tirer au sort à chaque ligne
# 2, c'est pour récupérer kappa
ckappa.boot <- function(data,x) {ckappa(data[x,])[[2]]}
```

2. Faire le bootstrap

```r
library(boot)
library(psy)

# fabriquer une fonction pour faire le bootstrap de kappa
# fonction entre accolades
# x, tirer au sort à chaque ligne
# 2, c'est pour récupérer kappa
ckappa.boot <- function(data,x) {ckappa(data[x,])[[2]]}
```

Syntaxe :

- créer une fonction `ckappa.boot` qui prend en entrée data (jeu de données) et x (tirage au sort des lignes)
- à l'intérieur de la fonction, on calcule kappa sur les données rééchantillonnées data[x,] et on récupère la 2ème valeur (kappa)
  - pour savoir que c'est la 2ème valeur, on a fait `ckappa(expsy[,c(12,14)])` pour voir la sortie

3. Lancer le bootstrap

```r
res <- boot(expsy[,c(12,14)],ckappa.boot,1000)
```

Syntaxe :

- `res` : objet qui va contenir le résultat du bootstrap
- `boot()` : fonction de la librairie boot pour faire le bootstrap
- `expsy[,c(12,14)]` : jeu de données

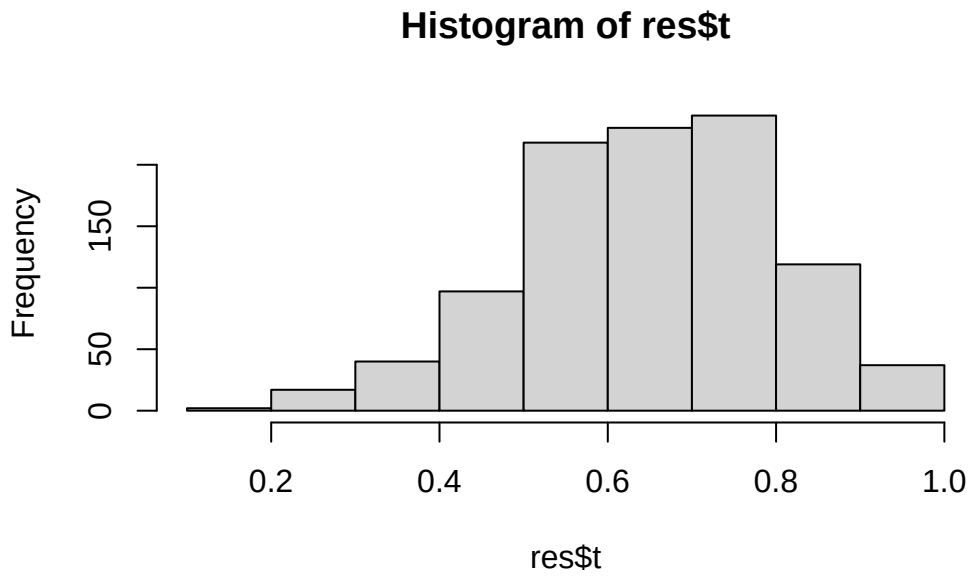
  - on ne prend que les colonnes 12 et 14 (2 évaluateurs)
- `ckappa.boot` : fonction créée précédemment pour calculer kappa sur les données rééchantillonnées
- `1000` : nombre de rééchantillons

4. Résultats

```r
res
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = expsy[, c(12, 14)], statistic = ckappa.boot, R = 1000)


Bootstrap Statistics :
     original      bias    std. error
t1* 0.6627907 -0.01481519    0.148357
```

```r
hist(res$t)
```



Syntaxe :

- `res` : affiche le résultat du bootstrap
  - `$t0` : kappa calculé sur les données initiales
  - `$t` : vecteur des 1000 kappa calculés sur les échantillons rééchantillonnés
- `hist(res$t)` : histogramme des 1000 kappa calculés sur les échantillons rééchantillonnés

5. Intervalles de confiance

```r
quantile(res$t,c(0.025,0.975))
```

```
     2.5%     97.5%
0.3494058 0.9156627
```

```r
boot.ci(res,type="bca")
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = res, type = "bca")

Intervals :
Level       BCa
95%   ( 0.2927,  0.9091 )
Calculations and Intervals on Original Scale
```

```r
ckappa(expsy[,c(11,13)])
```

```
$table
  1 2 3 4
1 1 0 0 0
2 1 4 0 0
3 0 4 9 2
4 0 0 2 6

$kappa
[1] 0.5421053
```

Syntaxe :

- `quantile(res$t,c(0.025,0.975))` : calcul des quantiles 2.5% et 97.5% des kappa calculés sur les échantillons rééchantillonnés
- `boot.ci(res,type="bca")` : calcul de l'intervalle de confiance à 95%
  - type = "bca" : méthode BCa (bias-corrected and accelerated) qui est meilleure que la méthode des quantiles

6. Test statistique : exemple pour kappa > 0.2

Si on veut tester :

- hypothèse nulle **H0: kappa ≤ 0.2**
- hypothèse alternative **H1: kappa > 0.2**.

```r
# On calcule le nombre de fois où le kappa des échantillons bootstrap est
↪  inférieur ou égal à 0.2
n_inf <- sum(res$t <= 0.2)
n_inf
```

```
[1] 2
```

```r
p_value <- (n_inf / 1000)
p_value
```

```
[1] 0.002
```

```r
#phrase pour résultat :
cat("La p-value unilatérale pour le test kappa < 0.2 est de ", round(p_value, 4),
↪  "\n")
```

```
La p-value unilatérale pour le test kappa < 0.2 est de  0.002
```

Avec une p-value < 0.025, on rejette l'hypothèse nulle au niveau de 2.5% et on conclut que le kappa est significativement supérieur à 0.2.

4 Second exemple en R sur SMP (CUSM)

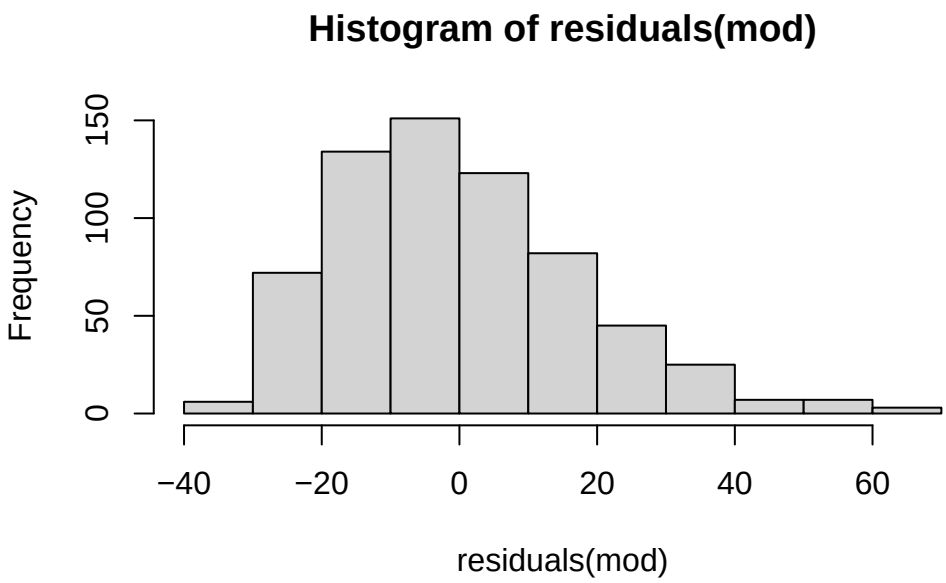
Modèle de régression linéaire expliquant la variable "durée de l'entretien" (`duree.interv`)

```r
mod <- lm(duree.interv~schizophrenie+depression+abus.subst+
↪  gravite+charactere+trauma.enfant+age+ factor(type.centre),data=smp)
summary(mod)
```

```
Call:
lm(formula = duree.interv ~ schizophrenie + depression + abus.subst +
    gravite + charactere + trauma.enfant + age + factor(type.centre),
    data = smp)

Residuals:
    Min      1Q  Median      3Q     Max
-38.578 -13.855  -1.769  10.922  64.405

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          42.48996    3.99998  10.623  < 2e-16 ***
schizophrenie         3.06420    2.80997   1.090 0.275911
```

```
depression            6.71269    1.64793   4.073 5.21e-05 ***
abus.subst            4.60037    1.79854   2.558 0.010760 *
gravite               1.06236    0.56548   1.879 0.060737 .
charactere            1.62547    0.93536   1.738 0.082723 .
trauma.enfant        -0.66805    1.66931  -0.400 0.689144
age                   0.20788    0.06066   3.427 0.000649 ***
factor(type.centre)2  4.09540    2.53401   1.616 0.106546
factor(type.centre)3 -1.29681    2.44159  -0.531 0.595509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.02 on 645 degrees of freedom
  (144 observations deleted due to missingness)
Multiple R-squared:  0.1086,    Adjusted R-squared:  0.09619
F-statistic: 8.734 on 9 and 645 DF,  p-value: 1.919e-12
```

```r
#histogramme des résidus
hist(residuals(mod))
```



On constate que les résidus ne suivent pas une loi normale.

On va donc faire un bootstrap pour obtenir des intervalles de confiance et des p-values pour les coefficients.

1. Créer la fonction de bootstrap

Syntaxe :

- créer une fonction `lm.boot` : c'est la recette de cuisine, qui prend comme arguments `data` et `index`
  - `data` (jeu de données = liste d'ingrédients) : recevra le jeu de données `smp` qui passera dans la fonction `boot()`

  - `index` (lignes qui seront tirées au sort avec remise par la fonction `boot()` = liste de course spécifiques) : vecteur des indices des lignes tirées au sort avec remise (peut ête ,x)
    * Cet argument `index` est **automatiquement généré et fourni par la fonction** `boot()` à chaque itération.

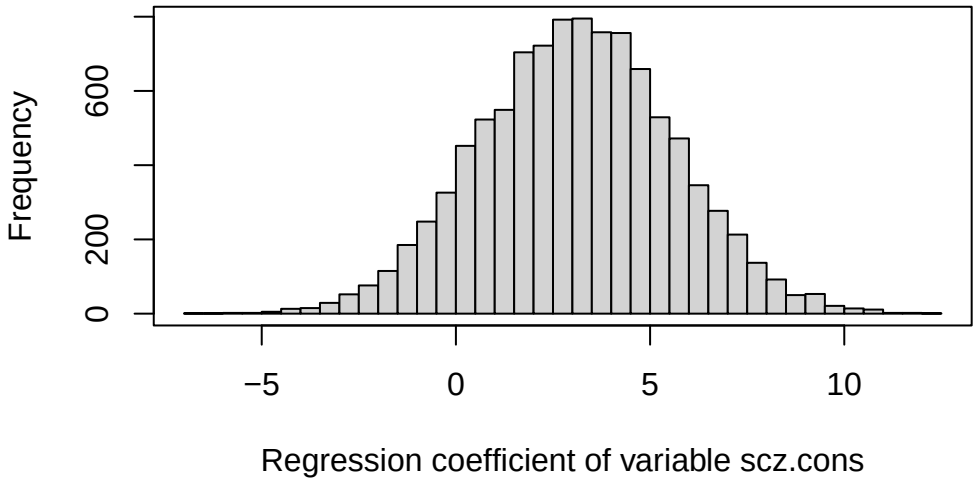
- La seconde partie : le chef cuisinier boot() qui a accès au garde-manger (jeu de données initial `smp`) et à la recette de cuisine (fonction `lm.boot`).
- Le chef :
  - crée lui même une liste de courses spécifiques (vecteur `index` des lignes tirées au sort avec remise)
  - prend la recette de cuisine (fonction `lm.boot`) et l'exécute avec les ingrédients (jeu de données `smp[index, ]`)
  - la fonction renvoie les coefficients du modèle ajusté sur l'échantillon bootstrap.
  - le chef récupère les coefficients et les stocke.
  - refait ça 10 000 fois.

1. Fonction `lm.boot` = recette de cuisine

```
# fonction pour bootstrap des coefficients de régression linéaire
lm.boot <- function(data, index) {
    # 1. Création de l'échantillon bootstrap
    # La fonction `boot()` appelle cette fonction en boucle.
    # À chaque appel, `boot()` fournit un nouveau vecteur `index` qu'elle a
    ↪ généré.
    smp.boot <- data[index,]

    # 2. Ajustement du modèle linéaire sur l'échantillon bootstrap
    mod <- lm(duree.interv ~ schizophrenie + depression + abus.subst + gravite +
↪   charactere + trauma.enfant + age + factor(type.centre), data = smp.boot)

    # 3. Extraction et renvoi des coefficients du modèle
    coefficients(mod)
}
```

2. Application au jeu de données `smp` et affichage de l'histogramme des coefficients de la variable `scz.cons` (schizophrénie)

```
resboot <- boot(smp, lm.boot, 10000)

hist(resboot$t[,2], breaks=40, main="", xlab="Regression coefficient of variable
↪   scz.cons")
box()
```



3. Intervalles de confiance pour tous les coefficients

```
ci_bca <- boot.ci(resboot, index = 2, type = "bca")  # on stocke l'IC une seule
↪  fois
p_val_bootstrap <- 2 * sum(resboot$t[, 2] <= 0) / 10000

# Afficher proprement les deux résultats sans chevauchement d'affichage
cat(
    sprintf(
    "L'intervalle de confiance à 95%% pour le coefficient de la variable scz.cons
     ↪  est [%.3f, %.3f].\n",
    ci_bca$bca[4], ci_bca$bca[5]
    ),
    sprintf(
    "La p-value bilatérale pour le test du coefficient de la variable scz.cons est
     ↪  de %.4f.\n",
    p_val_bootstrap
    ),
    sep = ""
)
```

```
L'intervalle de confiance à 95% pour le coefficient de la variable scz.cons est [-1.830, 7.839]
La p-value bilatérale pour le test du coefficient de la variable scz.cons est de 0.2140.
```

Syntaxe :

- `boot.ci(resboot, index=2, type="bca")` : calcul de l'intervalle de confiance à 95% pour le coefficient de la variable `scz.cons` (schizophrénie)
  - `resboot` : résultat du bootstrap
  - `index=2` : précise que c'est la variable « schizophrénie » qui nous intéresse. (schizophrénie)
  - `type="bca"` : méthode BCa (bias-corrected and accelerated) qui est meilleure que la méthode des quantiles sur les petits échantillons

25Méthodes non supervisées

1 Introduction aux méthodes non supervisées

- Objectif des méthodes non supervisées = mettre en évidence des groupes homogènes de **sujets** ou de **variables**.
- 2 options :
  - Réduction du nombre de dimensions pour rendre compatible avec nos yeux : **ACP**
  - Demander à l'rdinateur de se substituer à nous : nuées dynamiques, classification hierarchiques…

1.A Groupes homogènes de sujets

1.A.1 Analyse en composantes principales

1.A.1.1 Principe

- But = réduire le nombre de dimensions (sujets ou variables) tout en conservant un maximum d'information.

La première composante principale est la droite dans le nuage de points qui **explique le maximum de variance**, c'est à dire la direction suivant laquelle le nuage de points s'étire au maximum.

Il s'agit donc d'une combinaison linéaire des variables initiales (ici x, y et z) qui maximise la variance des projections des points sur cette droite.

Représentation R de cette droite sur le plot précédent (avec les distances de chaque point à cette droite) :

Pour obtenir la seconde composante principale :

- Il faut d'abord projeter les points sujets sur un plan perpandiculaire à la droite première composante

Pour obtenir la seconde composante principale : idem : direction suivant laquelle la projection des points s'étire au maximum

Et la troisième composante principale est représentée de la même manière

Code R qui montre les trois composantes principales dans un espace 3D avec une longueur proportionnelle à la variance expliquée :

- Relation entre les composantes principales et les valeurs propres / vecteurs propres
  - Chacun des vecteurs est appelé une **composante principale** = un **vecteur propre** de la matrice de covariance des variables initiales.
  - Sa longueur est proportionnelle à la variance expliquée par cette composante principale et correspond à une **valeur propre**.
- Propriétés mathématiques des composantes principales
  - Les composantes principales sont orthogonales entre elles.
  - Le plan formé par les deux premières composantes principales est appelé le premier plan principal.

- Si l'on projette les points sur ce plan, par une application du théorème de Pythagore, on montre que la variance des projections est maximale.
  * d correspond à la distance entre le centre de gravité des points (leur moyenne) et la projection du point sur le plan principal.
  * e correspond à la distance entre le point initial et sa projection sur le plan principal (c'est la partie "non expliquée" par le plan principal).
- Le premier plan principal maximise la variance des projections, donc maximise (au total) les distances d.
- Donc, par le théorème de Pythagore, il minimise (au total) les distances e.
- La somme des distances d et e est constante car elle correspond elle correspond à la "dispersion totale" du nuage autour du centre de gravité (inertie totale, proportionnelle à la variance totale).
- Autrement dit, le premier plan principal est aussi celui qui minimise la somme des carrés des distances entre les points et leurs projections :
  * c'est le plan qui passe au plus près des points (distance globale minimale au sens des moindres carrés).

1.A.1.2 Application pratique avec R Fichier `abd`: regroupe 24 signes et symptômes codés 0/1 chez des patients consultant pour des douleurs abdominales.

Une interprétation géométrique serait de dire que chaque patient est un point dans un espace à 24 dimensions (une dimension par symptôme).

On peut projeter ces points dans un espace à 2 dimensions (le premier plan principal) pour visualiser les similarités entre patients.

```r
abd.signes <- colnames(abd)[1:24]
pc <- princomp(abd[, abd.signes], cor = TRUE)
summary(pc)
```

```
Importance of components:
                          Comp.1     Comp.2    Comp.3     Comp.4     Comp.5
Standard deviation     2.1057156 1.41534793 1.3710081 1.16865372 1.10374651
Proportion of Variance 0.1847516 0.08346707 0.0783193 0.05690631 0.05076068
Cumulative Proportion  0.1847516 0.26821867 0.3465380 0.40344428 0.45420496
                           Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.09610205 0.99048784 0.97832112 0.95176293 0.92833149
Proportion of Variance 0.05005999 0.04087776 0.03987968 0.03774386 0.03590831
Cumulative Proportion  0.50426495 0.54514270 0.58502238 0.62276624 0.65867455
                         Comp.11    Comp.12   Comp.13   Comp.14   Comp.15
Standard deviation     0.9214850 0.91020266 0.8875715 0.8614295 0.8516506
Proportion of Variance 0.0353806 0.03451954 0.0328243 0.0309192 0.0302212
Cumulative Proportion  0.6940552 0.72857469 0.7613990 0.7923182 0.8225394
                          Comp.16    Comp.17    Comp.18   Comp.19    Comp.20
Standard deviation     0.80584750 0.78163569 0.75471070 0.7267781 0.67760329
Proportion of Variance 0.02705792 0.02545643 0.02373284 0.0220086 0.01913109
```

```
Cumulative Proportion  0.84959731 0.87505374 0.89878658 0.9207952 0.93992627
                          Comp.21    Comp.22    Comp.23    Comp.24
Standard deviation     0.67418766 0.59730211 0.58417513 0.53778254
Proportion of Variance 0.01893871 0.01486541 0.01421919 0.01205042
Cumulative Proportion  0.95886498 0.97373039 0.98794958 1.00000000
```

La fonction montre les variances expliquées par dimension. L'option `cor=TRUE` permet ici de normaliser les variables (on aurait pu s'en passer car les variables sont toutes binaires).

- Étapes :
  - Création d'un df avec les scores des deux premières composantes principales et le diagnostic
  - Calcul des écart-types des points projetés sur chaque composante principale
    * les écarts types correspondent géométriquement à la "longueur" de chaque axe = à la variance expliquée par chaque composante
    * le carré des écarts types correspond aux valeurs propres de chaque composante principale, c'est à dire à la variance expliquée par chaque composante
  -

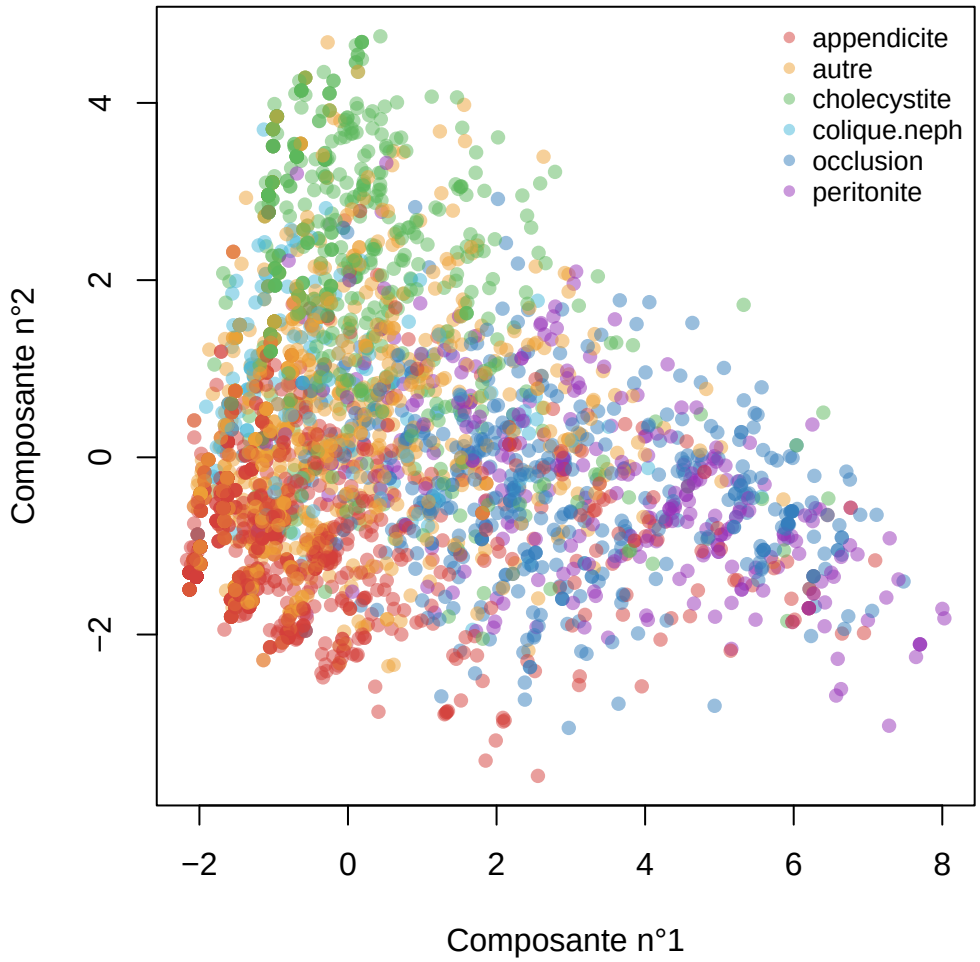
```r
# d'abord création d'un dataframe avec les scores des deux premières composantes
↪ principales et le diagnostic
dt.pc <- data.frame(pc$scores[,1],pc$scores[,2],abd$diagnostic)

# écart type des points projetés sur chaque composante principale (correspond
↪ géométriquement à la "longueur" de chaque axe = à la variance expliquée par
↪ chaque composante)
# le carré des écart-types correspond aux valeurs propres

#calcul des écarts types et des variances des deux premières composantes
↪ principales (variance = écart type au carré)
ecarts_types_12 <- pc$sdev[1:2]
variance_ecarts_types_12 <- ecarts_types_12^2
valeur_propres_12 <- variance_ecarts_types_12

valeur_propres_12
```

```
  Comp.1   Comp.2
4.434038 2.003210
```

```r
# pour calculer le pourcentage de variance expliquée par chaque composante
↪ principale

pourcentage_variance_expliquee_12 <- valeur_propres_12 / sum(pc$sdev^2) * 100
pourcentage_variance_expliquee_12
```

```
   Comp.1    Comp.2
18.475159  8.346707
```

```r
names(dt.pc) <- c("pc1","pc2","diagnostic")
min.x <- min(dt.pc$pc1)
max.x <- max(dt.pc$pc1)
min.y <- min(dt.pc$pc2)
max.y <- max(dt.pc$pc2)
couleur <- as.vector(c("#D43F3A80","#EEA23680","#5CB85C80",
"#46B8DA80","#357EBD80","#9632B880"))
plot(dt.pc$pc1, dt.pc$pc2,
    col=couleur[as.numeric((factor(dt.pc$diagnostic)))],
    xlim = c(min.x, max.x), ylim = c(min.y, max.y), pch=16,
    xlab="Composante n°1", ylab="Composante n°2")
legend( "topright",
        legend=levels(factor(dt.pc$diagnostic)),
        bty="n",
        text.col="black",
        col=couleur,
        pch=16,
        cex=0.8
)
```



Une façon beaucoup plus simple de calculer tout ça et d'avoir des jolis graphiques est d'utiliser la fonction `PCA()` du package `FactoMineR` :

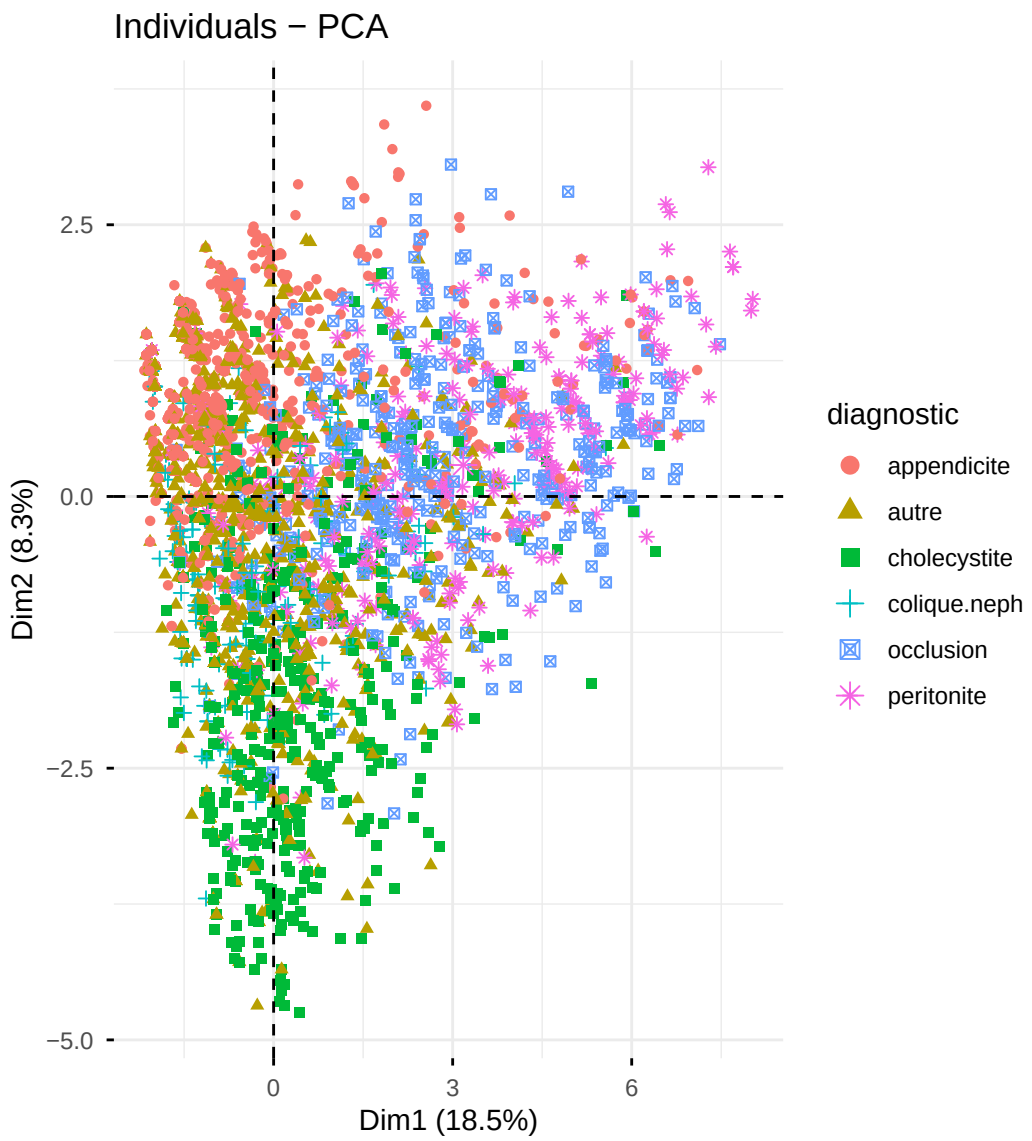
```
# Données : variables quantitatives + une variable qualitative "diagnostic"
X <- abd
X$diagnostic <- factor(X$diagnostic)

# PCA (diagnostic en qualitative supplémentaire)
quali_idx <- which(names(X) == "diagnostic")
res.pca <- PCA(X, quali.sup = quali_idx, scale.unit = TRUE, graph = FALSE)

# 1) Variance expliquée (équivalent de pc$sdev^2/sum(pc$sdev^2))
eig <- get_eigenvalue(res.pca)
eig[1:2, ]              # PC1 et PC2 (valeur propre, %, cumul)
```

```
      eigenvalue variance.percent cumulative.variance.percent
Dim.1   4.434038        18.475159                    18.47516
Dim.2   2.003210         8.346707                    26.82187
```

```r
# (alternative base FactoMineR)
res.pca$eig[1:2, ]       # eigenvalue, percentage of variance, cumulative
```

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1   4.434038              18.475159                          18.47516
comp 2   2.003210               8.346707                          26.82187
```

```r
# 2) Tableau équivalent à dt.pc
dt.pc <- data.frame(
    pc1 = res.pca$ind$coord[, 1],
    pc2 = res.pca$ind$coord[, 2],
    diagnostic = X$diagnostic
)

# 3) Graphe des individus colorés par diagnostic
fviz_pca_ind(
    res.pca,
    axes = c(1, 2),
    geom = "point",
    habillage = "diagnostic",
    addEllipses = FALSE
)
```




```
# Bonus (si besoin des cos2 / contributions sans calcul manuel)
ind <- get_pca_ind(res.pca)   # ind$coord, ind$cos2, ind$contrib
var <- get_pca_var(res.pca)   # var$coord, var$cos2, var$contrib
```

Note

Dans le cadre de l'ACP, les termes **valeur propre** et **variance expliquée** sont synonymes.

- **eigenvalue** = valeur propre de la matrice sur laquelle l'ACP est faite (matrice de covariance si données centrées, ou matrice de corrélations si données centrées-réduites).
- Cette valeur propre est égale à la variance des scores (projections des individus) sur la composante correspondante, donc à la "variance portée / extraite" par cette composante.
- La **proportion de variance expliquée** par une composante = **valeur propre de cette**

composante / somme des valeurs propres.

Note

Dans le cadre de l'ACP, un **scree plot** est un graphique qui trace les valeurs propres (ou leur pourcentage) en fonction du numéro de la composante principale.
Cas de l'ACP (PCA)

- Dans l'ACP, les valeurs propres correspondent bien à la variance portée par chaque composante (variance des scores/projections sur l'axe). Un scree plot d'ACP trace donc ces valeurs (ou parfois leur pourcentage).
- Dans R, plot(prcomp_obj) produit précisément ce scree plot, et `sdev^2` correspond aux variances des composantes.

Cas de l'analyse factorielle exploratoire (EFA)

- Un scree plot est aussi utilisé en analyse factorielle : il trace des valeurs propres pour aider à choisir le nombre de facteurs/composantes.
- Mais l'interprétation "variance expliquée" dépend de la méthode d'extraction : en PCA on parle de réduction de dimension (variance totale), en analyse factorielle "commune" on cherche plutôt des facteurs latents et toute la variance n'est pas traitée de la même façon (variance commune vs spécifique).

Cas de l'analyse des correspondances (CA)

- En analyse des correspondances, on trace aussi un scree plot, mais ce qui est décomposé n'est pas une variance de variables quantitatives : ce sont des inerties (souvent présentées via des valeurs propres / valeurs singulières au carré).
- Le scree plot montre alors la part d'inertie captée par chaque dimension.

Conclusion

- ACP : valeur propre = variance portée par une composante (et donc base du "% expliqué").
- Scree plot : graphique des valeurs propres (ou de leurs pourcentages), quelle que soit la méthode.
- Autres méthodes : les "valeurs propres" existent souvent, mais elles peuvent représenter de l'inertie ou une autre notion proche, pas forcément la même "variance expliquée" que dans l'ACP.

1.A.2 Classifications hiérarchiques

1.A.2.1 Principe

- Pas de réduction de dimension
- Regroupe les points (sujets) par ordre de proximité décroissante

Sur 5 points, 2 et 3 sont proches.

La hauteur $d1$ correspondant à la distance entre les points 2 et 3.

Un point 6 est placé au "centre" de 2 et 3. En réalité, il s'agit d'un cluster de points.

Idem pour les points 4 et 5, dont la distance donne une hauteur $d2$ et sont remplacés par un cluster correspondant au point 7.

Le regroupement entre les points 1 et 6 et fait, donnant un cluster 8.

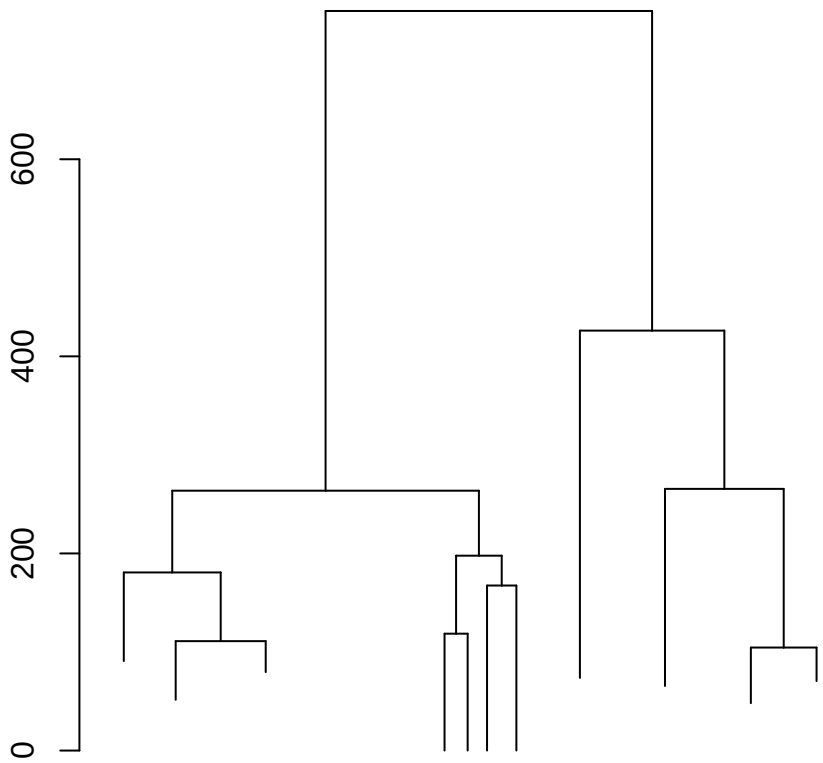
Puis finalement :

1.A.2.2 Application pratique avec R

Toujours sur les données `abd` :

- La fonction `hclust()` réalise la classification hiérarchique.
- Il faut calculer une matrice de distances entre les sujets avant, avec la fonction `dist()`.
- La méthode de regroupement choisie ici est la méthode de Ward (`method = "ward.D"`).
- La fonction `cut()` permet de couper l'arbre à une hauteur donnée pour obtenir des groupes. Ici, on coupe `upper`, ce qui correspond aux plus hautes branches de l'arbre.

```
abd.signes <- colnames(abd)[1:24]
cha <- hclust(dist(abd[, abd.signes]), method = "ward.D")
plot(cut(as.dendrogram(cha), h = 100)$upper, main = "", leaflab = "none")
```



On retient une solution à 5 clusters (choix subjectif et assumé).

La fonction `cutree()` permet d'obtenir l'appartenance de chaque sujet à un cluster.

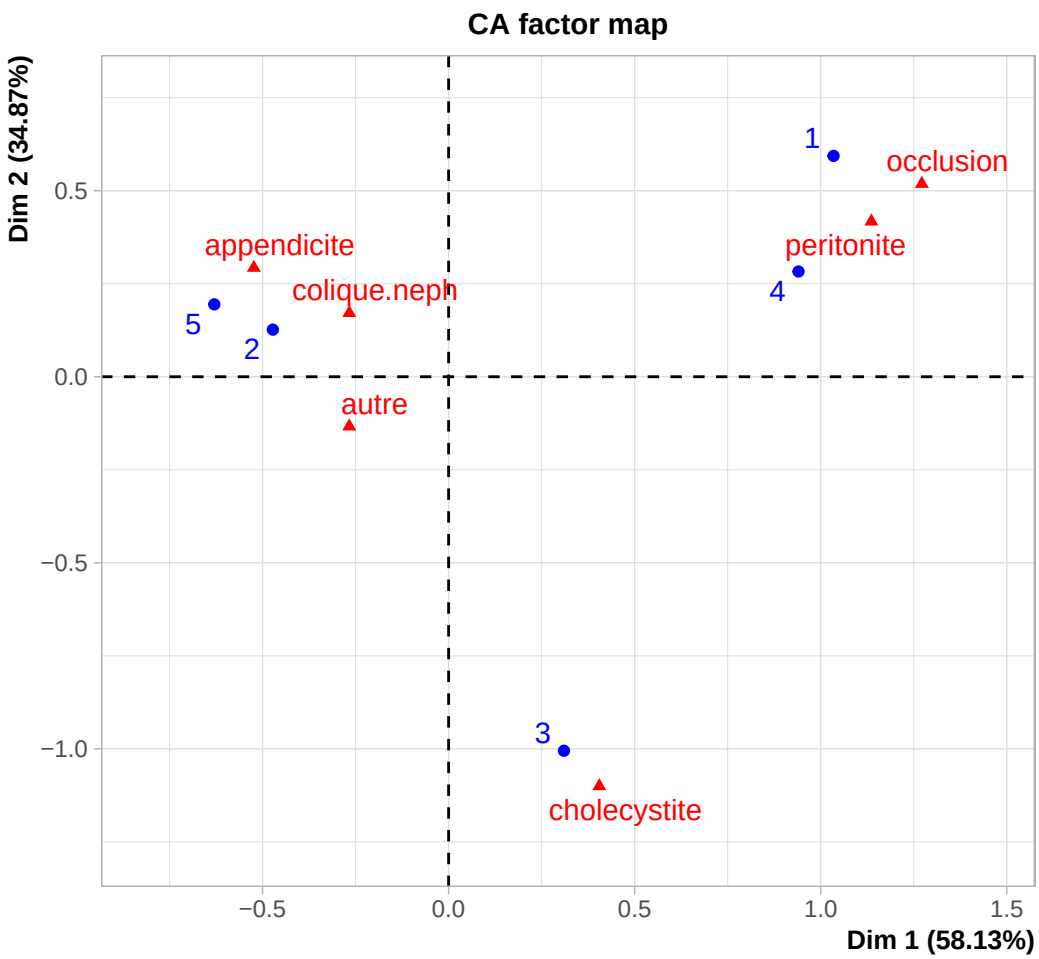
```
abd.5clust <- cutree(cha, k = 5)
abd.5clust[1:10]
```

```
 [1] 1 2 2 3 2 4 2 3 3 1
```

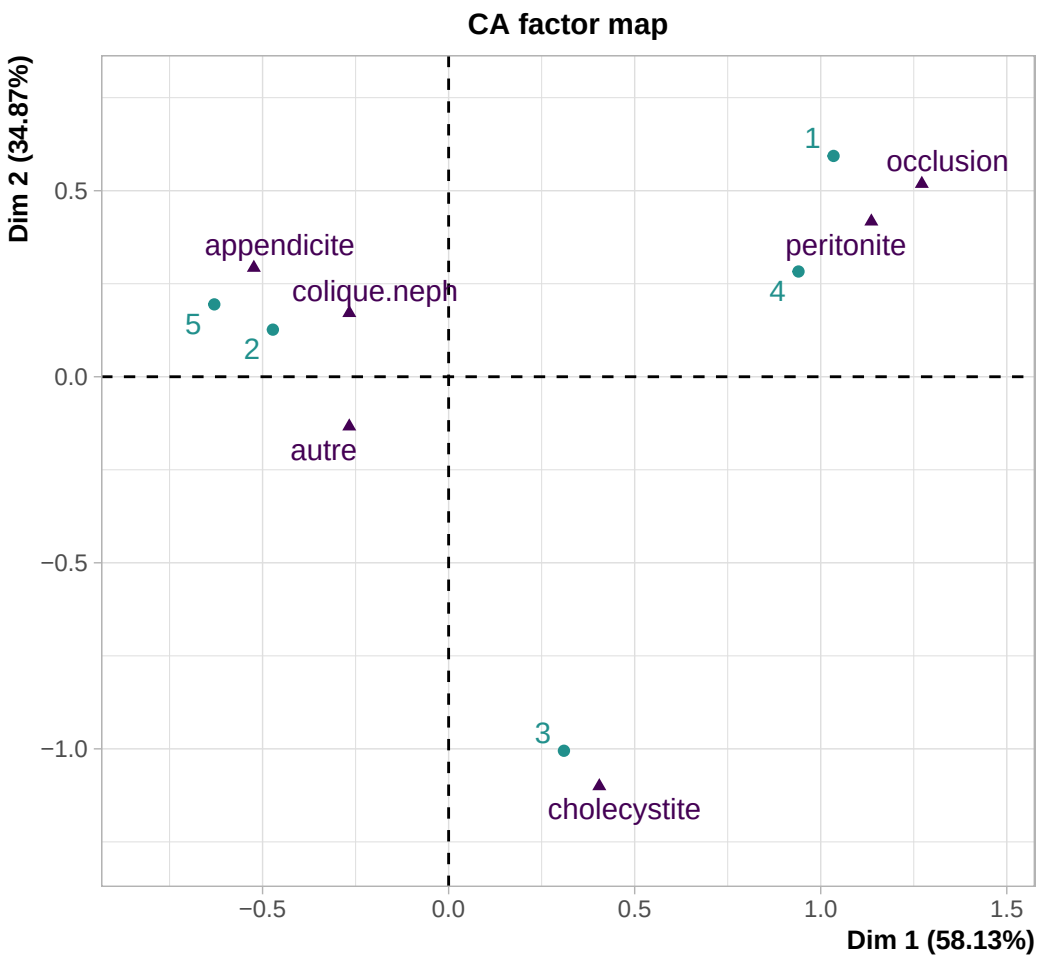
Le sujet 1 est affecté au cluster 1, le 2 au cluster 2, le 3 au cluster 2, etc.

- L'interprétation est un peu hasardeuse :
  - On peut faire une analyse des correspondances entre les clusters et les diagnostics (ça ressemble à une ACP mais c'est fait pour les variables catégorielles).

```
tbl <- table(abd.5clust, abd$diagnostic)
res.CA <- CA(tbl)
```




```
plot(res.CA,col.row=viridis(n=3)[2],col.col=viridis(n=3)[1])
```



- On voit bien que le cluster 3 est juste à côté du diagnostic "cholécystite"
- Mais si on regarde combien il y a de cholécystites dans le cluster 3 : on voit que c'est seulement 47% (même si elle est bcp plus rare dans les autres clusters)
- Syntaxe R : `prop.table(tbl, 1)` permet de calculer des pourcentages dans un tableau de contingence.
  - Pour calculer les pourcentages par ligne (i.e. par cluster), on met `1` en second argument.
  - Pour calculer les pourcentages par colonne (i.e. par diagnostic), on met `2` en second argument.

```r
round(prop.table(tbl, 1),2)
```

```
abd.5clust appendicite autre cholecystite colique.neph occlusion peritonite
         1        0.17  0.14         0.04         0.01      0.32       0.32
         2        0.50  0.33         0.06         0.07      0.02       0.02
         3        0.05  0.36         0.47         0.02      0.05       0.06
         4        0.13  0.15         0.14         0.05      0.40       0.13
         5        0.53  0.46         0.00         0.01      0.00       0.00
```

```
#ici, (tbl, 1) signifie qu'on calcule les pourcentages par ligne (i.e. par
 ↪  cluster)
```

1.A.3 Nuées dynamiques

26Méthodes non supervisées 2

1 Introduction

En Data Science : Y = la variable **à prédire**

$\longrightarrow$ On cherche + à expliquer qu'à prédire

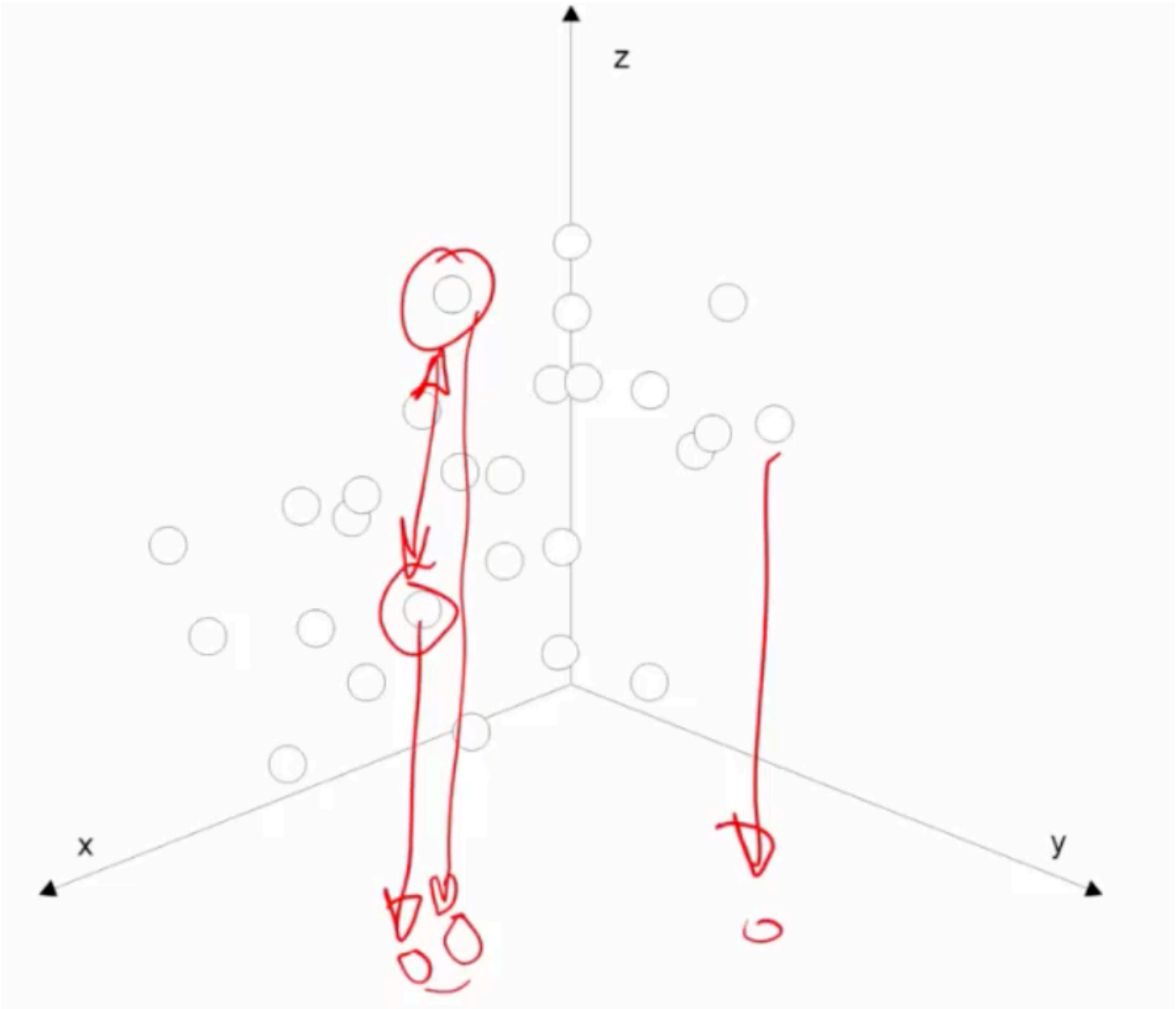
Objectif des méthodes multidimensionnelles = **réduire le nombre de dimensions**

- Dimensions = variables explicatives (X1, X2, …, Xk)
- Objectif : voir en "moins de dimensions" (2D, 3D) ce qui se passe dans un espace à k dimensions

Une méthode : projeter dans un espace de dimension inférieure !

2 Principe des composantes principales

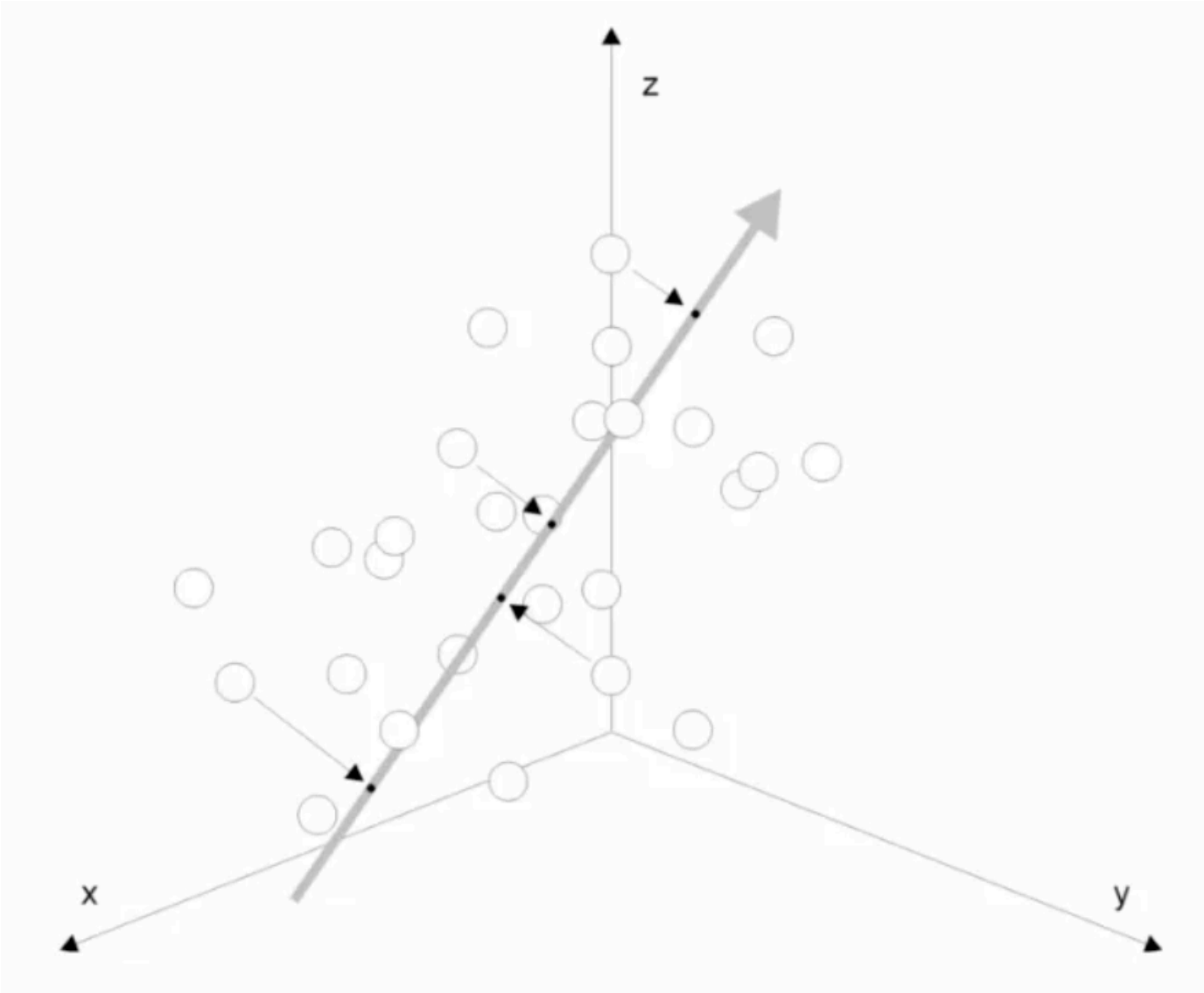
Mais réduction des dimensions = **distorsion** !



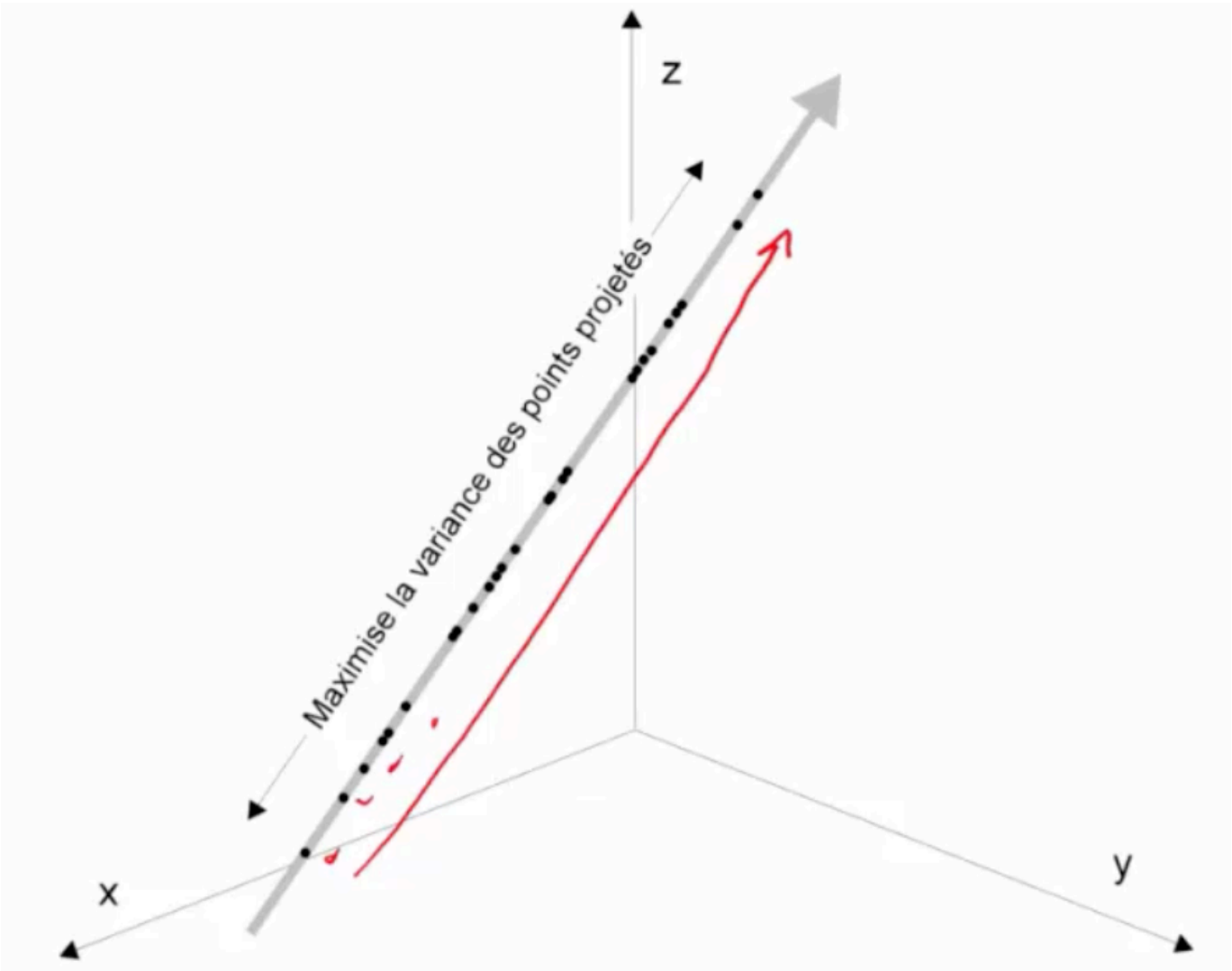
Une méthode : trouver une droite sur laquelle la variance des points projetés est la plus grande possible.

En gros : c'est la droite selon laquelle les points sont le plus étalés.

avec variance = dispersion des points autour de la moyenne

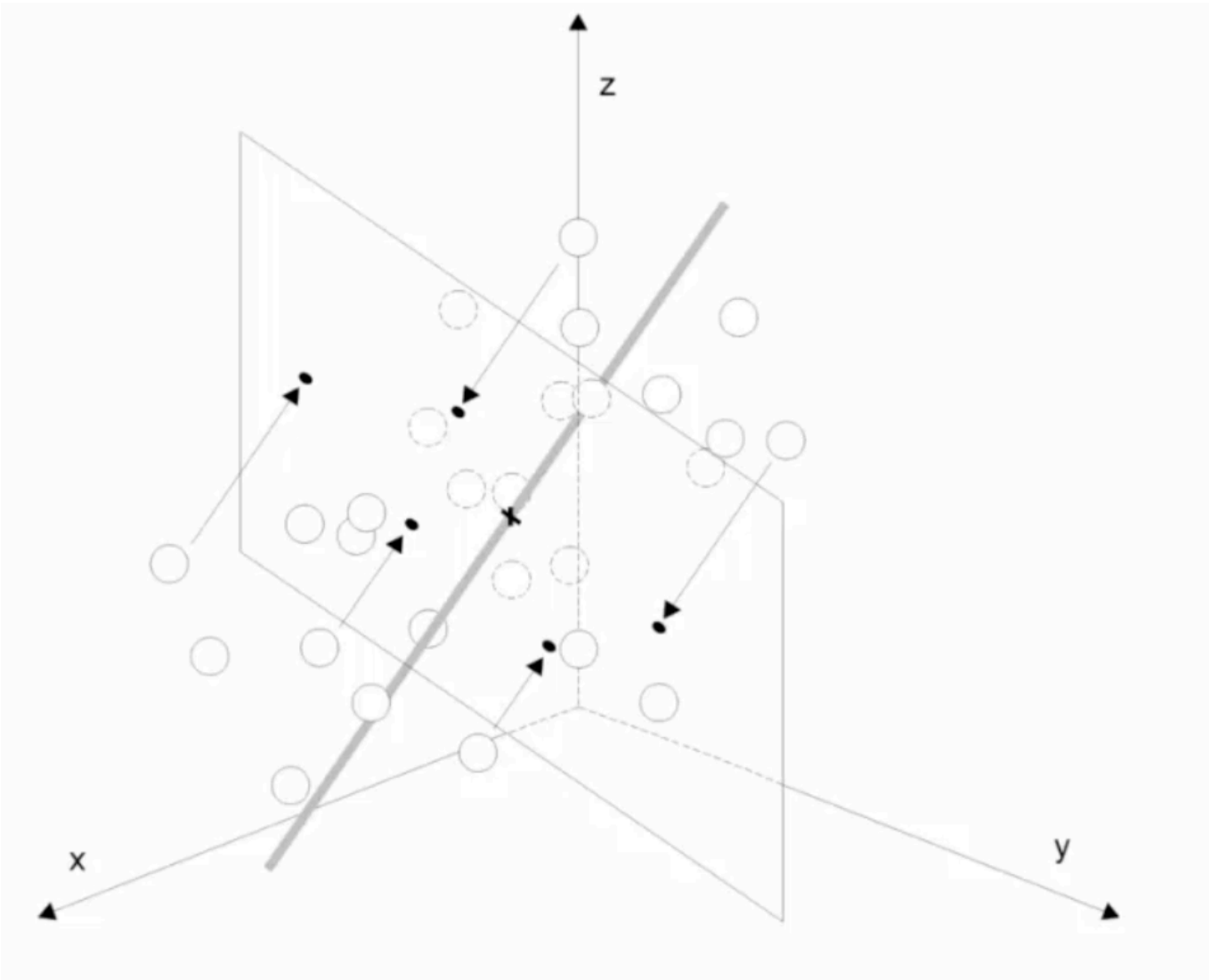


La droite qui maximise la variance des points projetés est la **1ère composante principale (PC1)**

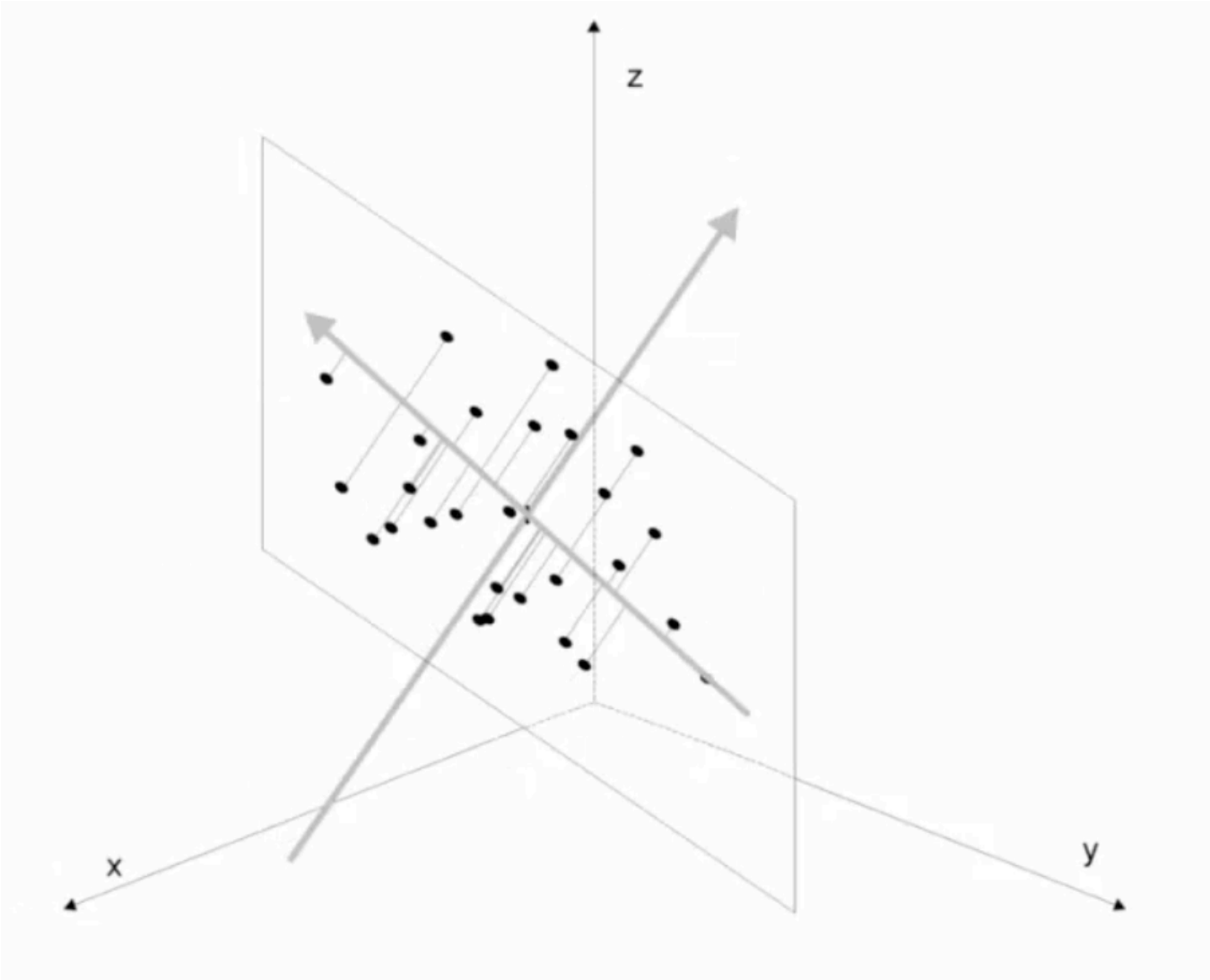


Ensuite :

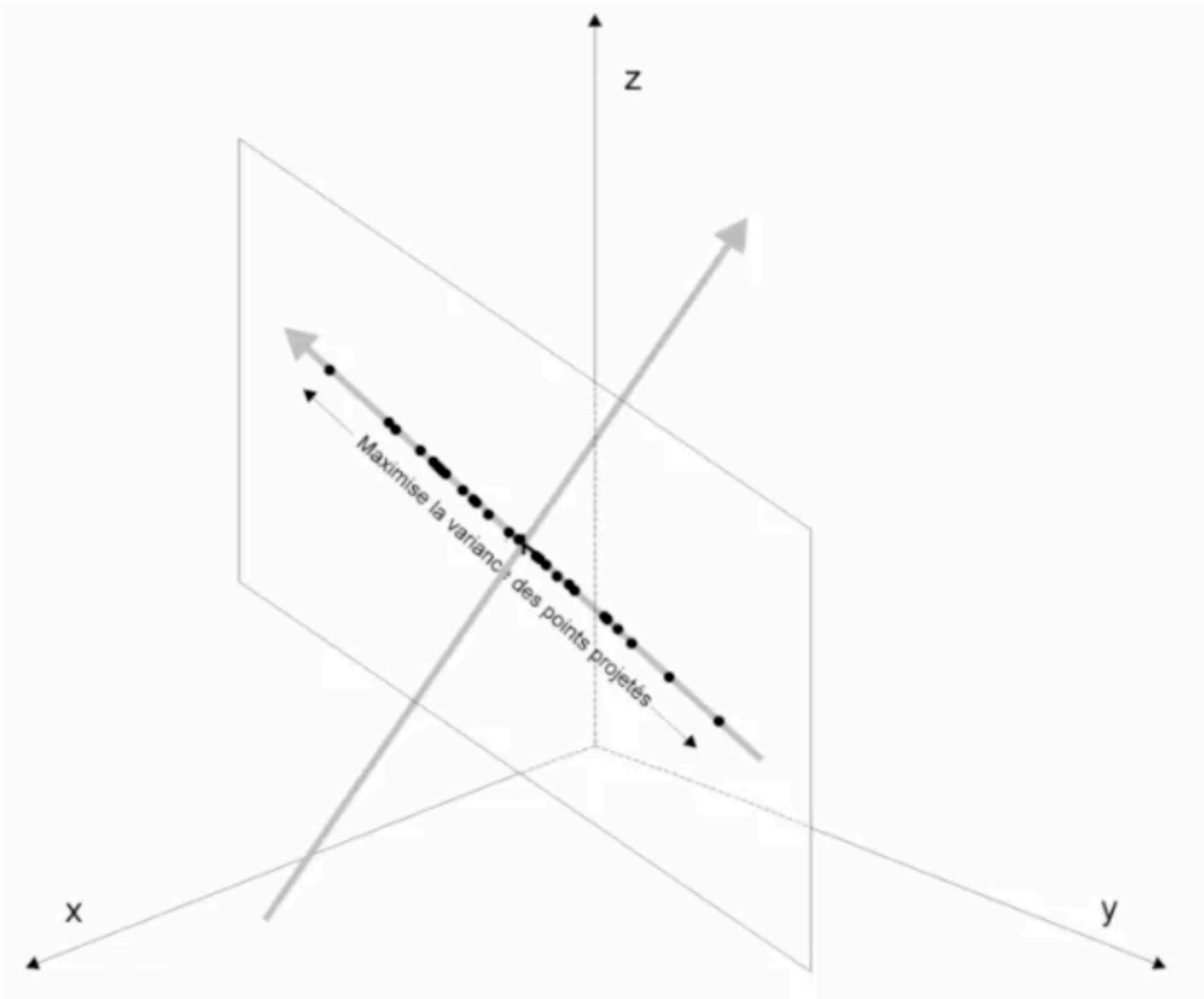
- prendre le plan perpandiculaire à la première composante principale
- et projection des points sur le plan
- On obtient un nouveau nuage de points sur ce plan



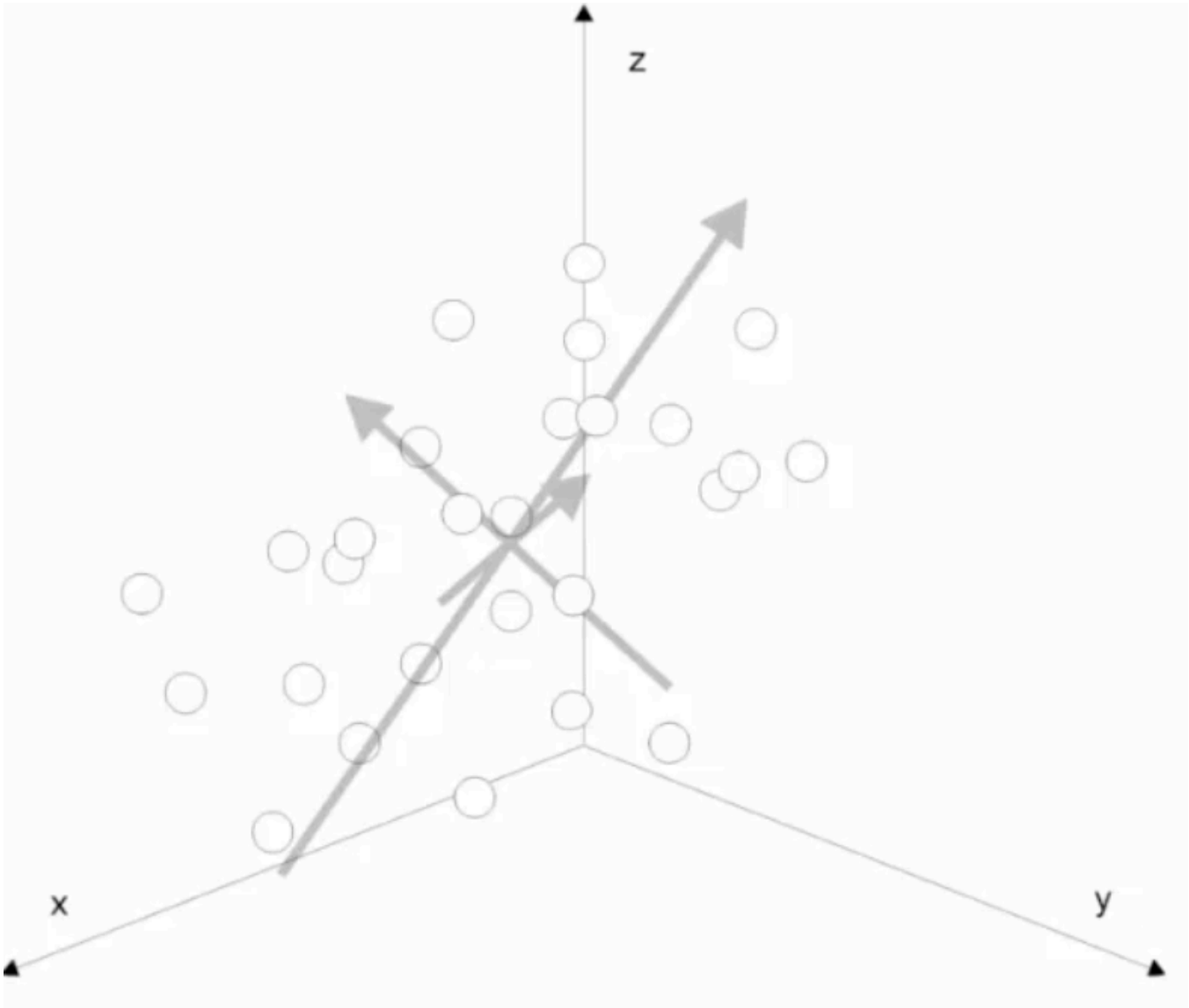
Ensuite on cherche la 1ère composante principale (ACP1) de ce nouveau plan : càd la droite par laquelle la variance des points projeté est maximale



On obtient les deux premières composantes principales.



Donc la 3e composante principale est celle qui est perpandiculaire aux 2 premières :

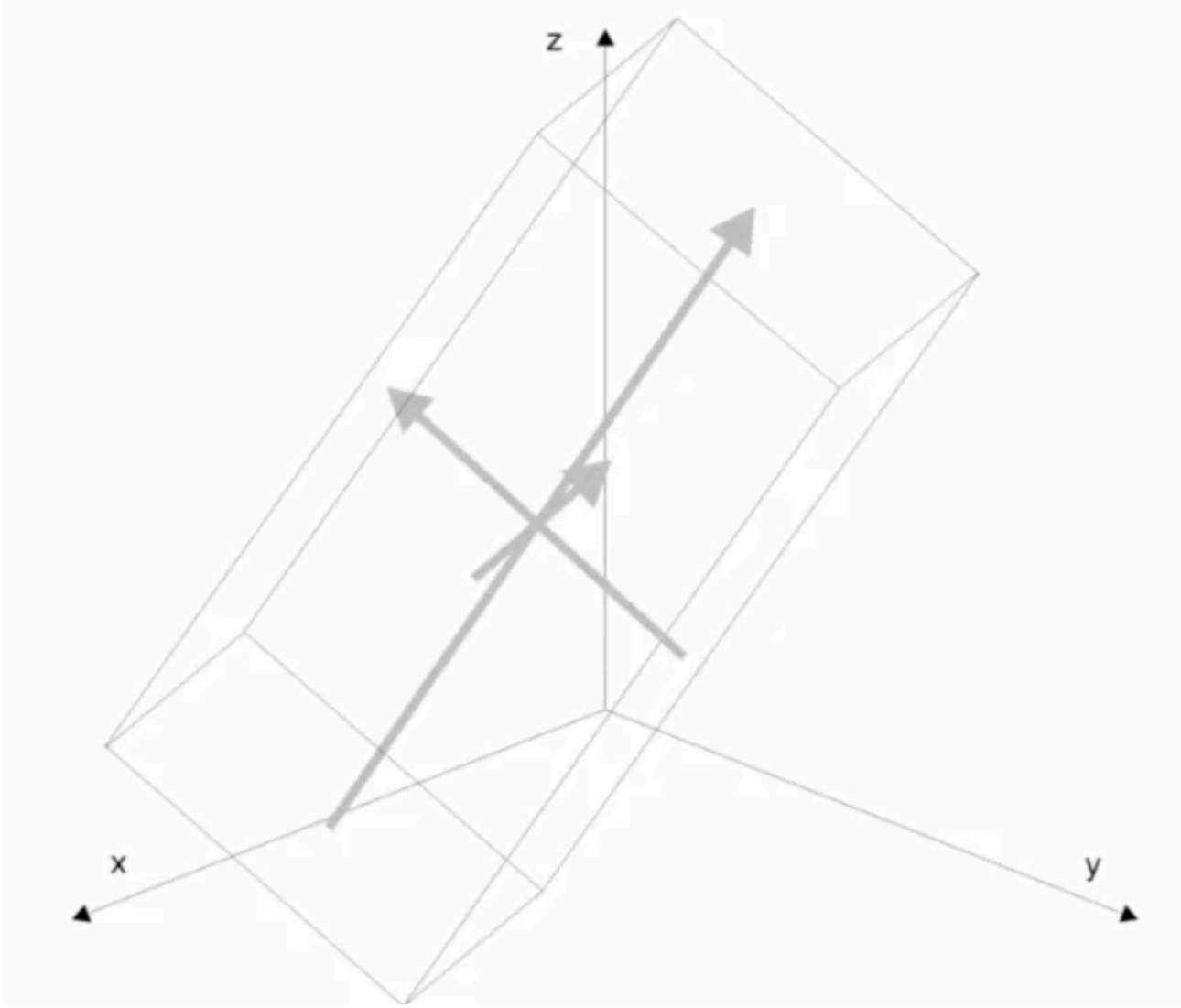


En résumé :

- **1ère composante principale (PC1)** :
  - droite qui maximise la variance des points projetés (le nuage de points s'étire le plus possible)
  - et plan sur lequel le "pied" ou le "résidu" des points projetés est le plus petit possible (distance entre le point et sa projection sur la droite)
  - Et comme les points projetées ont un résidu faible, ils sont les plus proches possible des points d'origine
  - C'est donc ce qui déforme le moins
- **2ème composante principale (PC2)** : droite perpendiculaire à PC1 qui maximise la variance des points projetés sur le plan perpendiculaire à PC1 (dans ce qui est perpandiculaire à PC1, PC2 est la droite selon laquelle les points sont le plus étirés)
- **3ème composante principale (PC3)** : droite perpendiculaire à PC1 et PC2 qui maximise la

variance des points projetés sur le plan perpendiculaire à PC1 et PC2

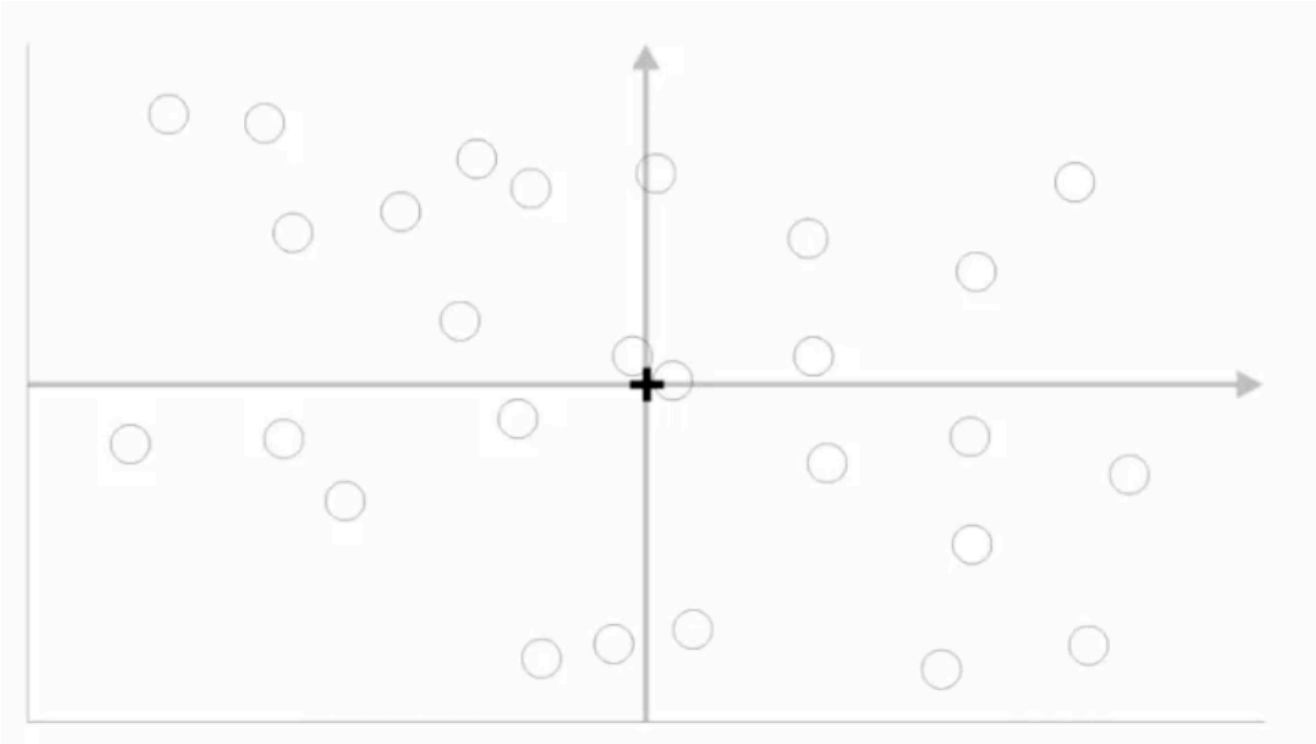
Si on représente les composantes principales dans une longueur qui correspond à la variance des points projetées, on obtient une boîte !



Questions :

- Si jamais il y a deux droites pour lesquelles on obtient la même variance $\longrightarrow$ ça n'arrive jamais !
- Algébriquement : on passe par des **matrices** à 3*3 car 3 dimensions qu'on diagonalise (diagonalisation = trouver les vecteurs propres et valeurs propres)
  - Vecteurs propres = directions des composantes principales
  - Valeurs propres = variances des points projetés sur les composantes principales

A la fin :



Retenir :

- On a des points dans un espace à k dimensions
- On cherche des droites (composantes principales) qui maximisent la variance des points projetés (directions selon lesquelles les points sont le plus étirés)

3 Utilisations en statistiques

- Visualisation de données multidimensionnelles
- Prendre les deux premières composantes principales (PC1 et PC2) et faire un graphique 2D

4 Plan sur lequel les points sont des variables plutôt que des sujets

En reprenant un dataframe classique :

X correspondent à des variables

Sujet	X1	X2	X3	X..	X40
S1	2	5	3		3
S2	4	3	6		6
S3	1	4	8		8
S…					
S1000	1	4	8		2

Ici :

- 1000 points
- Dans un espace à 40 dimensions

Si on projète sur le premier plan principal : projections sur 2 variables Z1 et Z2 (rajoutées dans le dataframe)

Si on renverse le jeu de données, ça donne un dataframe :

Variable	S1	S2	S3	…	S1000
X1	2	4	1		
X3	3	6	8		
X..					
X40	3		6	8	

Et donc ici :

- 40 points
- Dans un espace à 1000 dimensions

et X1 est un point dans un espace à 1000 dimensions.

En réalité :

- 2 points dans un espace à 3 dimensions : 2 points sur une droite donc en réalité, 2 points dans un espace à 1 dimension
- Idem dans 1000 dimensions : y a plein de dimensions qui ne servent à rien

NB : il faut "normaliser" = centrer et réduire les données avant de faire une ACP ! $\rightarrow$ C'est à dire soustraire la moyenne et diviser par l'écart-type

Comme on normalise les variables :

- Moyenne de Xi = 0
- Ecart-type de Xi = 1 et variance de Xi = 1 (car variance = écart-type au carré)

*Pourquoi la moyenne = 0 ?

- En soustrayant la moyenne de l'ensemble des valauers à chacune des valeurs, on centre les données autour de 0
  - Par exemple : 5 + 10 + 15 = 30 / 3 = 10 (moyenne) ;
  - et en soustrayant 10 à chaque valeur : -5 + 0 + 5 = 0

Mais pourquoi centrer et réduire ?

- Si on ne centre pas et réduit pas : les variables avec la plus grande variance vont dominer les composantes principales
- En centrant et réduisant : on met toutes les variables sur le même pied d'égalité

5 Représentation en sphère

Pourquoi, à la fin, les variables centrées-réduites forment une sphère dans l'ACP ?

- Imaginons p variables (par exemple p = 40).
- L'ACP cherche p composantes principales (PC1, PC2, …, PCp) : c'est à dire p nouvelles dimensions perpendiculaires entre elles.
- Pour chaque variable d'origine (par ex "âge", "créatinine", "CRP"…), on peut mesurer dans quelle direction la variable pointe dans ce nouvel espace (c'est à dire dans quelle direction la variable "âge" pointe dans l'espace défini par les composantes principales).
- Ces directions sont données par les corrélations entre la variable et chaque composante principale.

Exemple : - corr(age, PC1) = 0.8

- corr(age, PC2) = 0.5

Alors la variable "âge" est représentée par la flèche allant (0.8, 0.5).

Représentation en sphère car les longueurs des flèches = 1

- Variables centrées-réduites et avec longueur des flèches = variance
  - Moyenne = 0 (point d'origine des flèches)
  - Variance = 1 (longueur des flèches)

$\text{corr}^2(X, PC1) + \text{corr}^2(X, PC2) + \cdots + \text{corr}^2(X, PCp) = 1$

C'est un résultat direct du fait que l'ACP décompose entièrement la variance initiale.

Toutes les variables — sans exception — se transforment en flèches de longueur 1.

Elles ne peuvent donc exister que sur une surface dont tous les points sont à distance 1 du centre.

Pourquoi cette surface est une SPHÈRE (et un cercle quand en 2D) ?

Une sphère = ensemble des points dont la distance au centre = 1.

Exemples :

- en 1D deux points : –1 et +1
- en 2D un cercle
- en 3D une sphère
- en p dimensions une hypersphère

Comme toutes tes flèches ont une longueur = 1, leurs extrémités se trouvent toutes à distance 1 du centre.

Donc : dans l'espace complet (p dimensions) c'est une hypersphère

```
library(psy)
# Jeu de données fictif pour illustration
set.seed(123)
df_fake <- data.frame(
```

```
    age = rnorm(100, 50, 10),
    rs  = rnorm(100, 100, 15),
    ed  = rnorm(100, 12, 2),
    dr  = rnorm(100, 5, 1)
)

# ACP plane (cercle)
mdspca(df_fake)
```

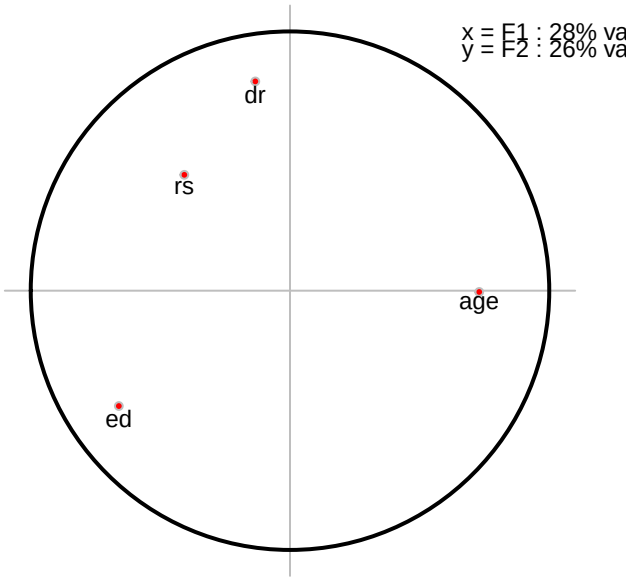


Figure 26.1: Représentation en sphère des variables dans une ACP

```
# ACP sphérique (variables sur la sphère unité)
sphpca(df_fake, v = 55)
```

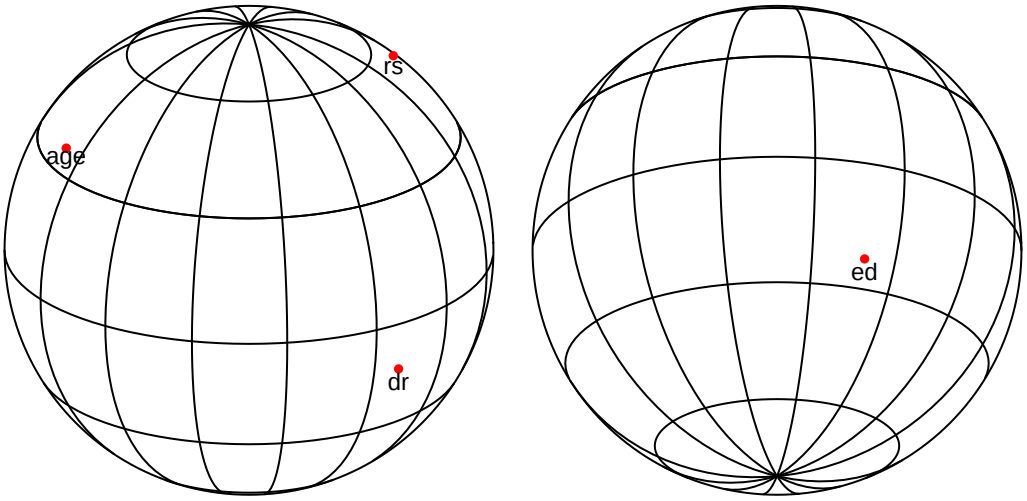


Figure 26.2: Représentation en sphère des variables dans une ACP

6 En d'autres termes :

1. Centration-réduction

Chaque variable est transformée pour avoir :

- une moyenne = 0 (centrée)
- une variance = 1 (réduction)

Cela garantit que toutes les variables ont le même poids dans l'ACP. Et surtout : la variance totale à "répartir" pour chaque variable est exactement 1.

2. L'ACP décompose cette variance

L'ACP fabrique des axes perpendiculaires entre eux (PC1, PC2, PC3…).

Chaque variable d'origine est représentée dans ce nouvel espace comme une flèche (un vecteur), dont les coordonnées sont les corrélations entre la variable et chaque composante principale.

- La longueur de la flèche est égale à 1 (car variance = 1)
- et les coordonnées de la flèche sont les corrélations entre la variable et chaque composante principale.
  - corr(X, PC1) la position de la variable sur l'axe PC1
  - corr(X, PC2) la position sur l'axe PC2

Relation entre variance et correlation :

- Variance(X) = somme des corrélations² entre X et chaque composante principale
- Si on représentait sur l'ensemble des composantes principales (mais c'est pas possible pour l'oeil humain), toutes corrélations seraient à 1.
- Ce qu'on voit en 2D, c'est une projection partielle de cette sphère dans un plan.

5. Interprétation :

- Variables proches entre elles = variables corrélées (deux variables proches corrélées positivement)
- Deux variables opposées par 180° corrélation négative
- Deux variables perpendiculaires corrélation ≈ 0
- Longueur = qualité de représentation dans le plan (c'est à dire qu'elle est bien représentée par les deux composantes principales)

7 Conclusion

- Centration-réduction : force chaque variable à avoir une variance = 1
- ACP : exprime exactement cette variance comme la somme des carrés de ses corrélations aux composantes principales
- La longueur de chaque vecteur-variable = 1 (car variance = 1)
- Chaque variable est projetée selon sa corrélation aux composantes principales

- La corrélation = coordonnée de la variable dans le plan des composantes principales
- Donc les variables proches entre elles sont corrélées

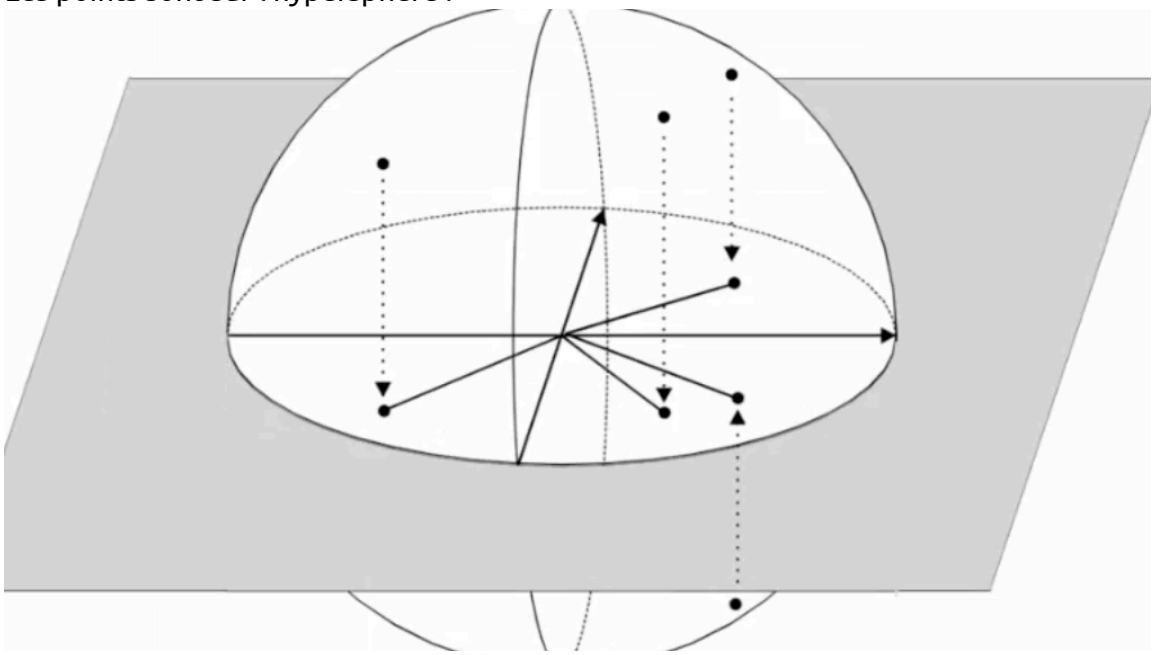
8 Références

Vidéo Falissard :

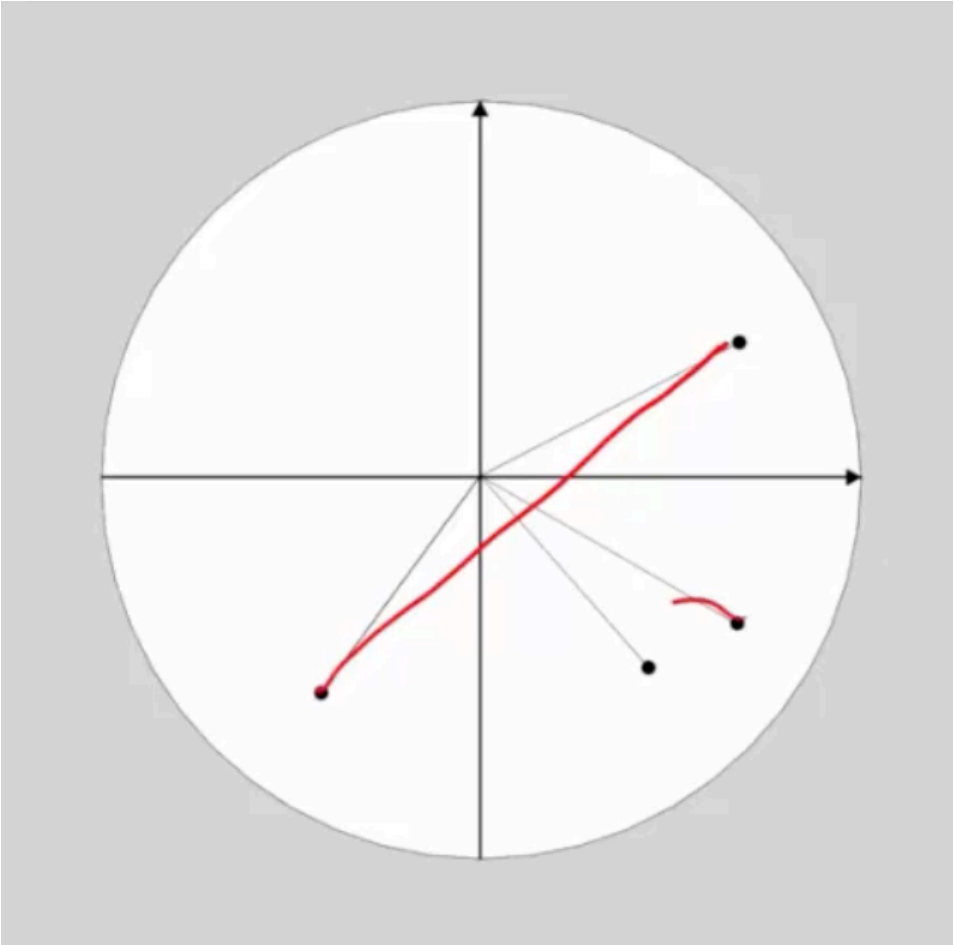
- Blog larmarange
- Webin-R
- Site ministère

27Methodes multidimensionnelles 3

Les points sont sur l'hypersphère :



On les projette sur le premier plan principal :

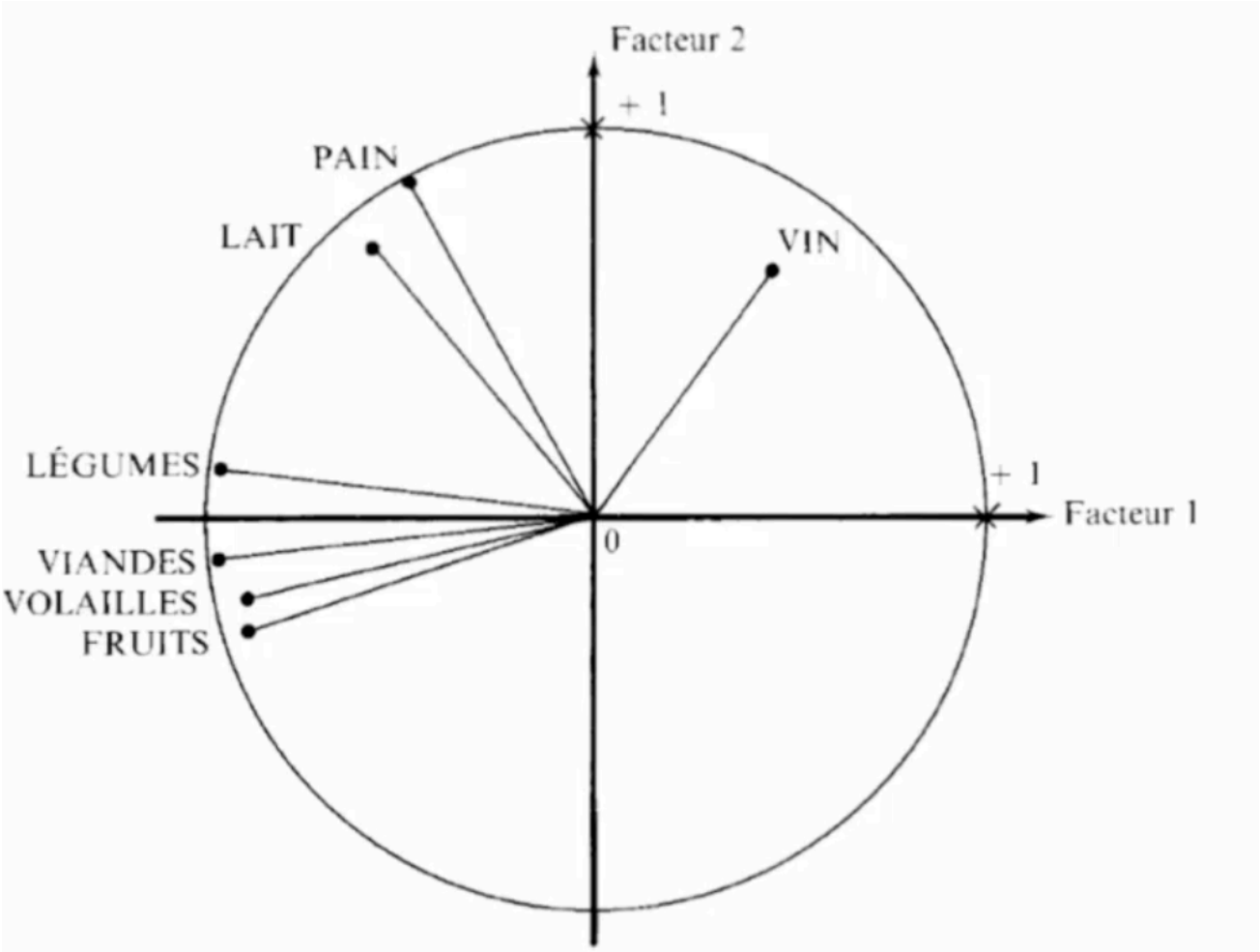


- Si diamétralement opposé : corrélés négativement
- Si proches : corrélés positivement
- Si perpendiculaires : pas corrélés (indépendants)

1 Exemple 1

7 variables ; 1 millions de données

1.A Représentation des variables



Points proches du cercle = bonne qualité de représentation en ACP

- Consommation de pain et de lait corrélés
- Viande : corrélé positivement à volailles et fruits
- Vin : indépendant de volailles / viandes / fruits

1.B Interprétation sociologique

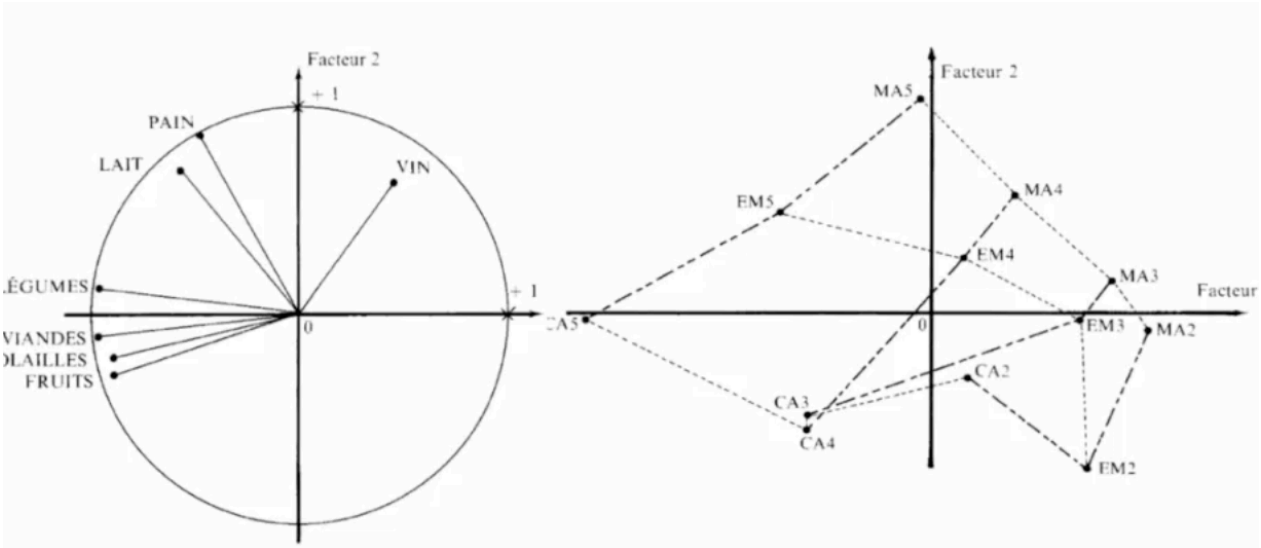
Population riche : Consomme plus de viande, fruits, vin
Population ayant des enfants : Consomme plus de pain et lait
Vin : classes populaires dans les années 50-60 !

1.C Représentations des individus par groupes

- On ne représente pas chaque individu (sinon 1 million de points)

- On fait des groupes (clustering) et on représente les centres des groupes (centre = moyenne des individus du groupe = barycentre)

$\longrightarrow$ On représente le point moyen de tous les cadres ayant 3 enfants par exemple



- MA5 : employé manuel avec 3 enfants (ménage = 5 personnes)
- EM5 : employé non manuel avec 3 enfants
- MA4 : travailleur manuel avec 2 enfants
- CA4 : cadre avec 2 enfants

Et en fait ça s'interprète pareil !!

- La où il y a le pain et le lait : c'est les ménages de 5 personnes !
- Et à l'opposé : les familles qui ont peu d'enfant
- Là ou il y a les légumes / fruits / viande : ce sont les cadres
- Vin : surtout employés manuels

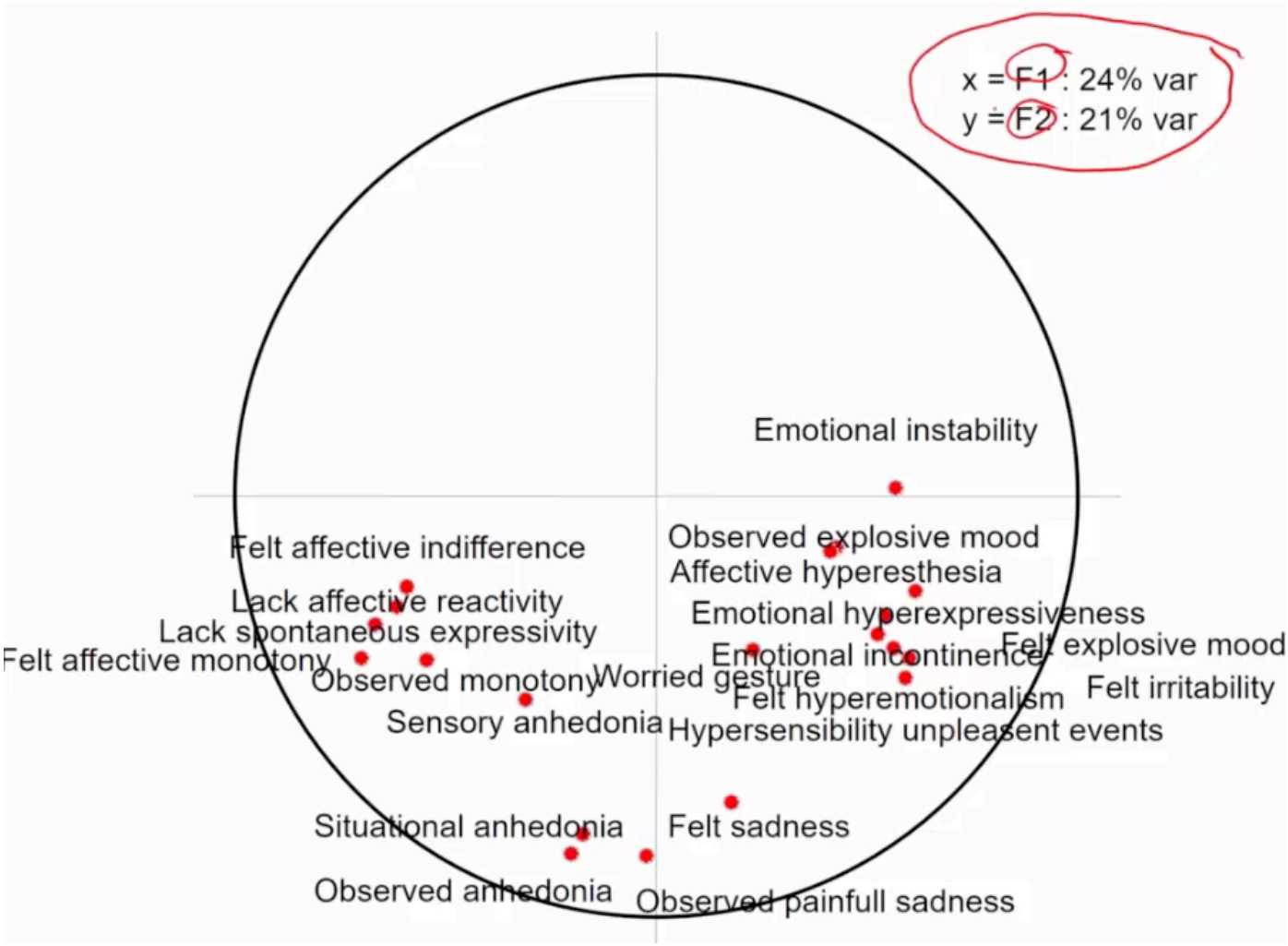
> **Note**
>
> **Les composantes principales des variables sont proches mathématiquement des composantes principales des sujets !**

2 Exemple 2 : Échelle de dépression “Depressive mood scale”

Beaucoup de critères qui se ressemblent :

- Observed painfull sadness
- Emotional hyperexpressiveness
- Emotional instability
- Observed monotony
- Lack spontaneous expressivity
- Lack affective reactivity
- Emotional incontinence
- Affective hyperesthesia
- Observed explosive mood
- Worried gesture
- Observed anhedonia
- Felt sadness
- Situational anhedonia
- Felt affective indifference
- Hypersensibility unplease
- Sensory anhedonia
- Felt affective monotony
- Felt hyperemotionalism
- Felt irritability
- Felt explosive mood

En ACP :

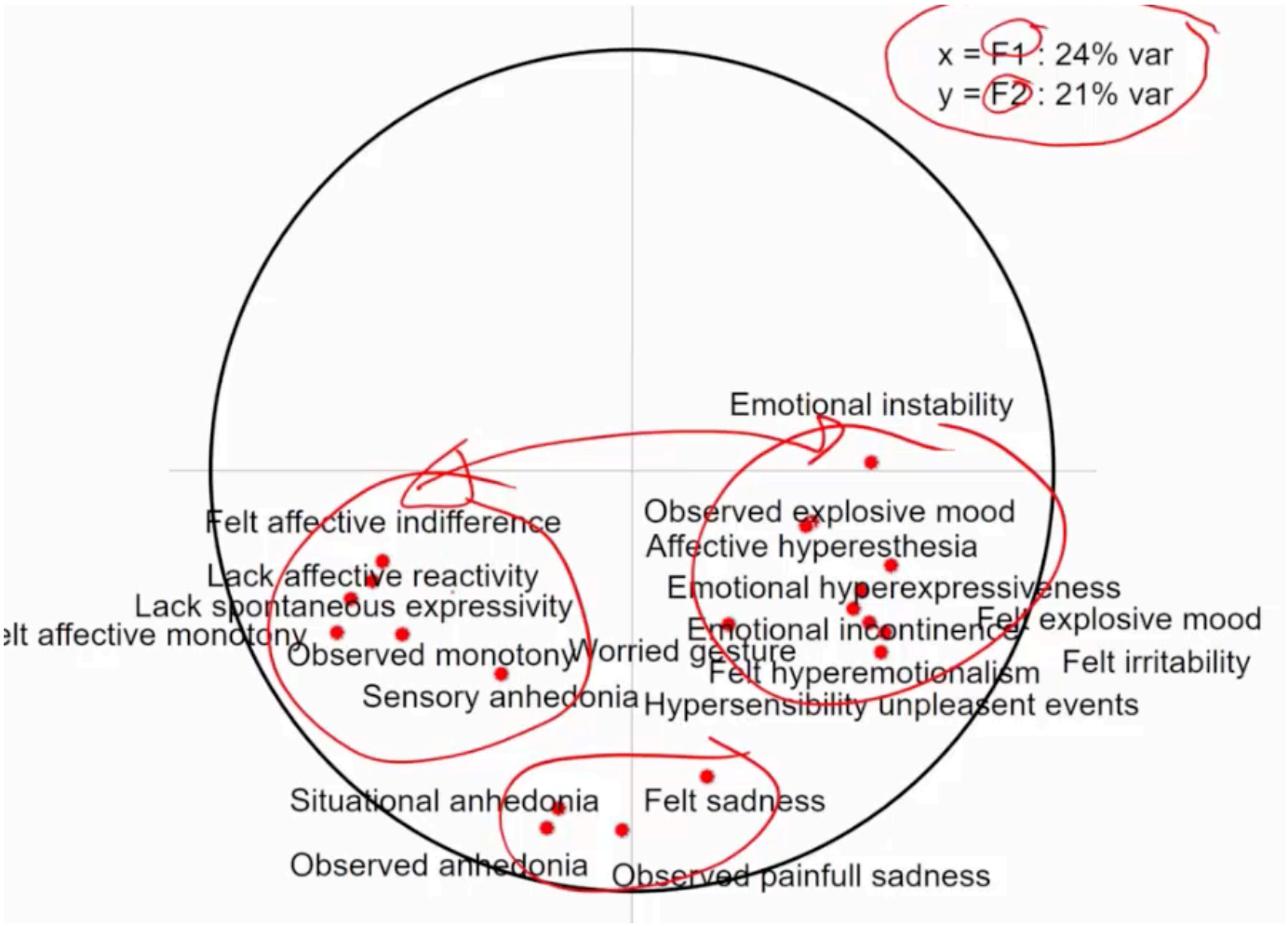


2.A Variances de l'ACP

- Variance de l'ACP1 = 24% et variance de l'ACP2 = 21%
- Donc les 2 premières composantes capturent 45% de la variance totale = c'est à dire qu'on explique pas mal de données

2.B Inteprétation :

- Distorsion + marquée que précédemment : les points sont plus éloignés du cercle (+ proches du centre)



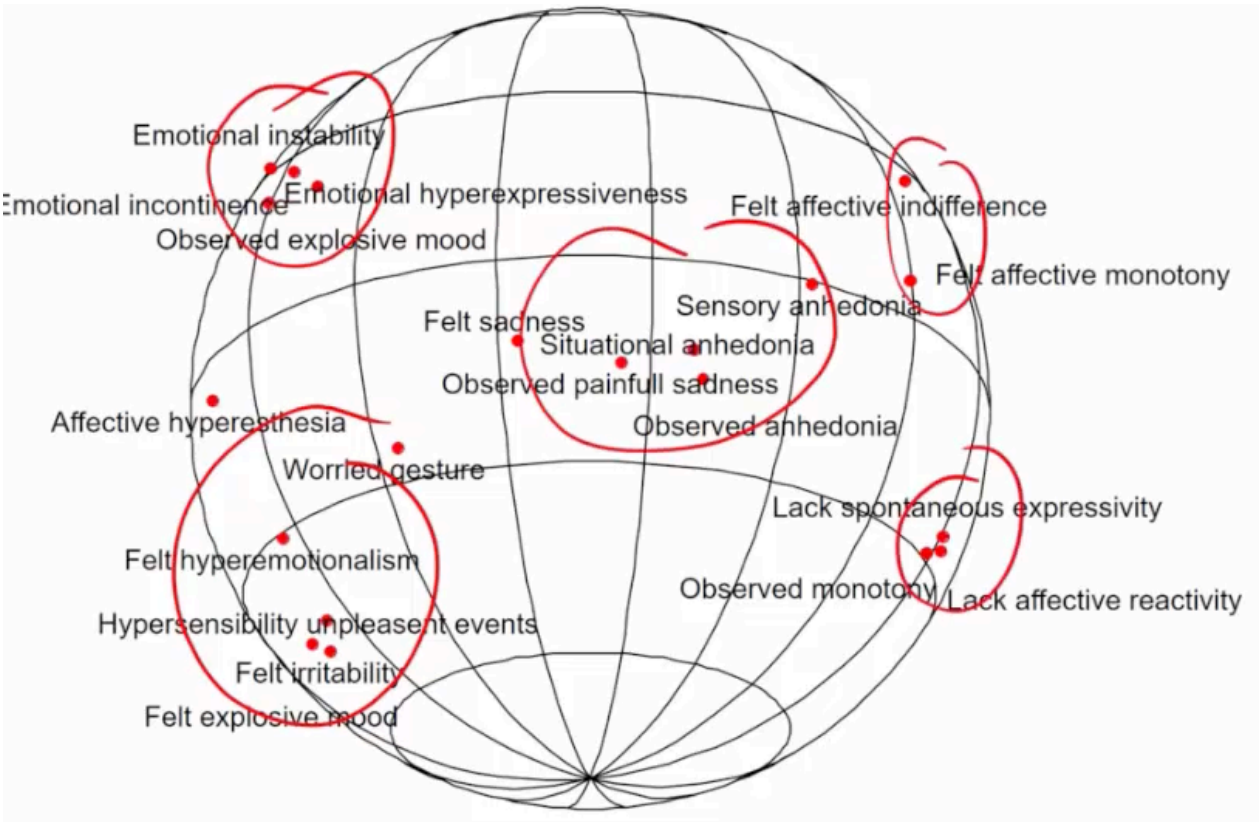
- 3 paquets de réponses :
  1. À gauche : déprimés émoussés
  2. À droite : déprimés irritables
  3. En bas : tristesse, anhédonie : “à part” = axe principal de la dépression

Et en fait on peut même interpréter la construction du truc !

- ACP1 : axe qui explique le plus la variance = sur lequel les points sont les plus éloignés
  - Correspond à l’axe des x ici (abscisse)
  - Sur cet axe : opposition émoussé vs irritable
  - Et donc on peut dire que cette opposition explique le plus de variabilité du phénotype dépressif
- ACP2 : axe qui explique le 2ème plus la variance = axe des y ici (ordonnée)
  - Ici : explique **l’intensité des symptômes dépressifs**

> **Note**
>
> Méthodes approximatives, mais c'est pas si mal !
>
> - alpha et beta et p : on pense avoir trouvé quelque chose de scientifique et d'incontournable
> - toujours un risque ! on sait pas exactement ce que dit le petit p
> - quand c'est pas vraiment tiré au sort : le petit p est la probabilité que le hasard puisse exprimer a lui seul un effet aussi grand que celui qu'on observe
> - mais le hasard, c'est qu'on imagine que l'échantillon est tiré d'une population initiale hypothétique infinie !

2.C Représentation sphérique : gagne une dimension

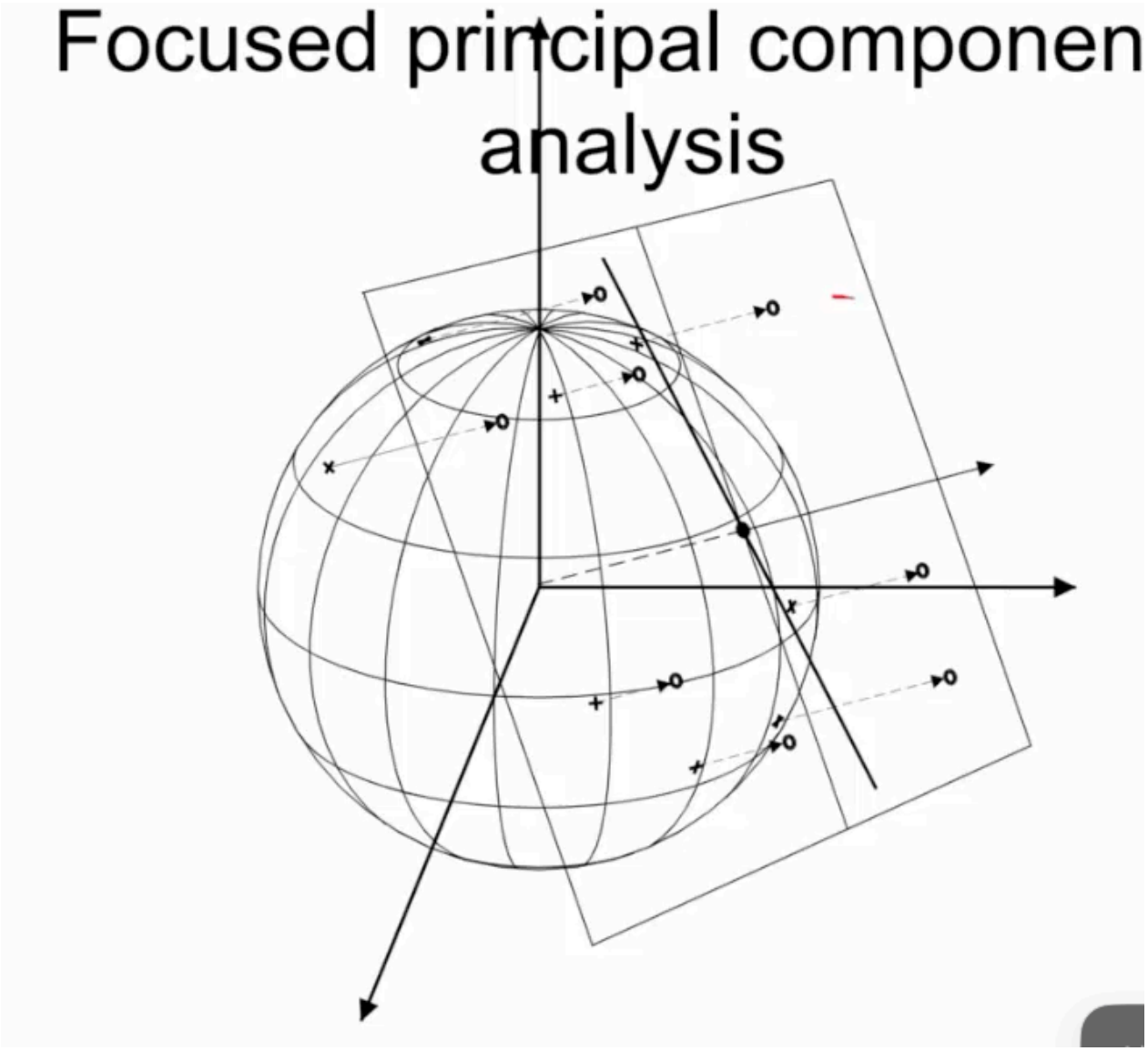


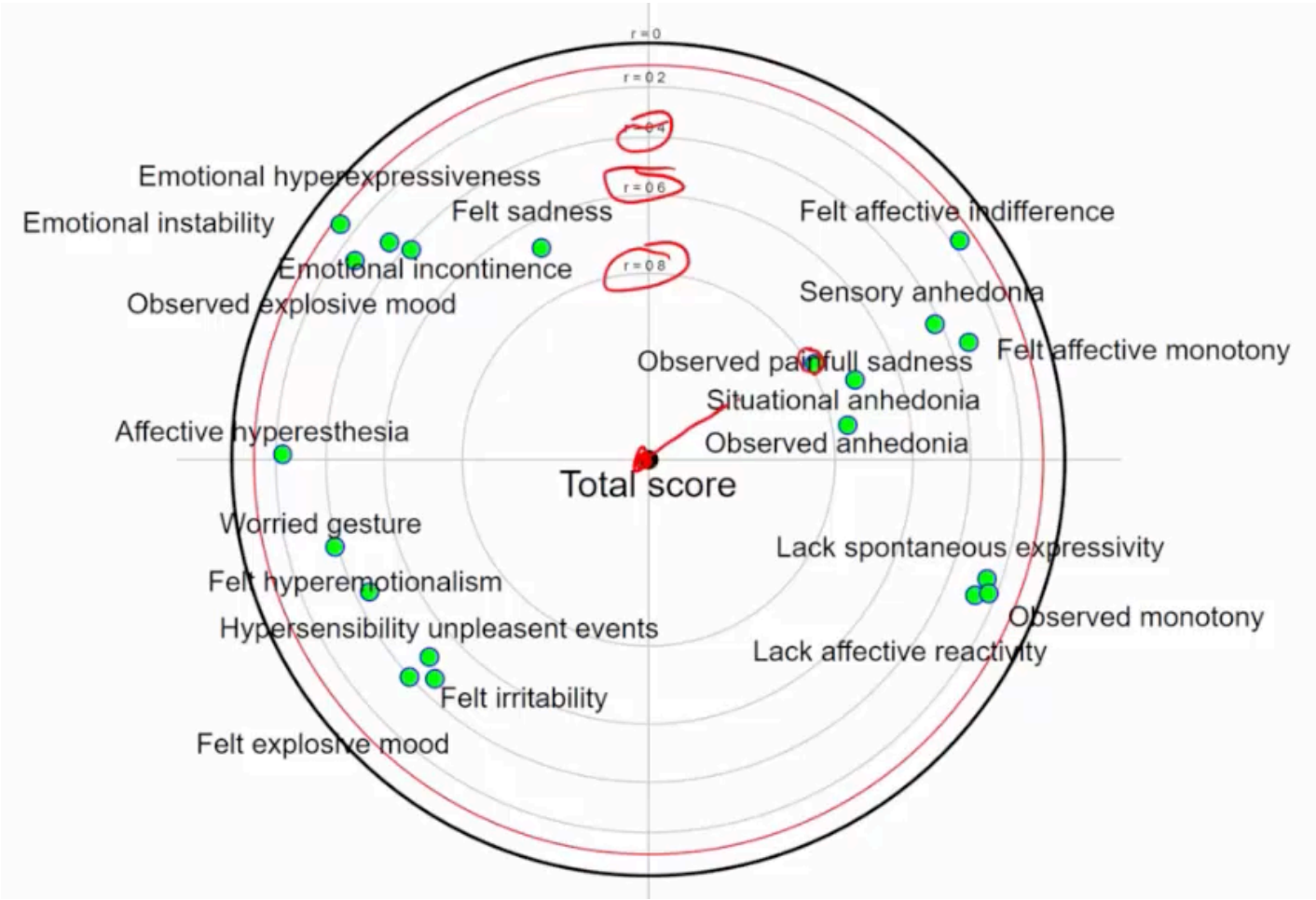
On a l'impression qu'il y a plus de paquets : parce que gagne une dimension !

Questions : ACP

3 Analyse en composante principale focalisée

Projeté sur un plan tangentiel à une des variables


Focused principal componen
analysis



Permet de voir les corrélations mais pas d'infos sur la qualité des variables

4 Références

Vidéo Falissard:

- Blog larmarange
- Webin-R
- Site ministère

28Analyse Factorielle

1 Principes de l'analyse factorielle

- Modèles : comme régression (avec résidus etc)
- Hyper utile pour données subjectives
- Ne permet pas de conclure

2 Analyse factorielle à un facteur

On imagine une réponse des sujets i à des variables 1, 2, 3, 4, 5
$I_1^i = a_1 + b_1\theta^i + e_1^i$
$I_2^i = a_2 + b_2\theta^i + e_2^i$
$I_3^i = a_3 + b_3\theta^i + e_3^i$
$I_4^i = a_4 + b_4\theta^i + e_4^i$
$I_5^i = a_5 + b_5\theta^i + e_5^i$
Avec

- θ_i : caractéristique commune aux variables
- e^i : terme regroupant l'erreur de mesure et la part spécifique à la variable

2.A Modélisation

On modélise ces variables sous la forme d'une **régression linéaire simple** (car analyse en un facteur)

Avec θ_i : variable dont on fait l'hypothèse qu'elle existe, mais n'est pas mesurée (elle n'est pas dans le jeu de données)

Pour calculer les paramètres, on combine les équations pour supprimer θ_i.

2.B Exemple

On fait un test de QI avec 5 "subtests" (variables)

$\rightarrow$ donc 5 tests

On fait l'hypothèse que la réponse à chacun de ces tests = une régression linéaire sur une seule variable = le QI global

$\rightarrow$ Donc c'est la même variable pour tous les 5 tests

Si ces 5 épreuves cognitives sont reliées à une seule et même variable : ils sont reliés à la **capacité cognitive générale**.

Donc si quelque soit l'épreuve, le répondant réussit bien, c'est qu'il existe une variable latente θ (capacité cognitive générale) qui explique la performance à chaque épreuve.

$\rightarrow$ Et ainsi on peut estimer θ = l'intelligence en la **quantifiant**.

C'est pourquoi on ne peut parfois pas faire de test de QI car les composantes sont trop différentes entre elles.

3 Utilisation : surtout analyse factorielle à > 1 facteur

On rajoute un facteur μ

$I_1^i = a_1 + b_1\theta^i + c_1\mu^i + e_1^i$ (calcul)

$I_2^i = a_2 + b_2\theta^i + c_2\mu^i + e_2^i$ (problèmes)

$I_3^i = a_3 + b_3\theta^i + c_3\mu^i + e_3^i$ (récitation)

$I_4^i = a_4 + b_4\theta^i + c_4\mu^i + e_4^i$ (histoire)

$I_5^i = a_5 + b_5\theta^i + c_5\mu^i + e_5^i$ (grammaire)

Imaginons : que θ soit dominant pour certains tests / et μ pour d'autres (pour lesquels θ est tout petit, et inversement)

Et donc on peut dire (avec un raisonnement inductif et exploratoire) que :

- θ = capacité verbale = QI verbal
- μ = capacité calculatoire = QI de performance

et plein d'autres dimensions

3.A Indétermination

On peut se poser la question si toutes ces variables dépendent d'une **méta-variable** (facteur supérieur).

Le pb : c'est que l'équation actuelle (le modèle des 5 réponses) est indéterminé, c'est à dire qu'il a un nombre infini de solutions.

Mais toutes ces solutions sont interprétées de la même façon : donc elles disent toutes la même chose.

Indétermination : quand on a une solution, on peut en trouver une autre grace à une **"rotation"** (c comme ça)

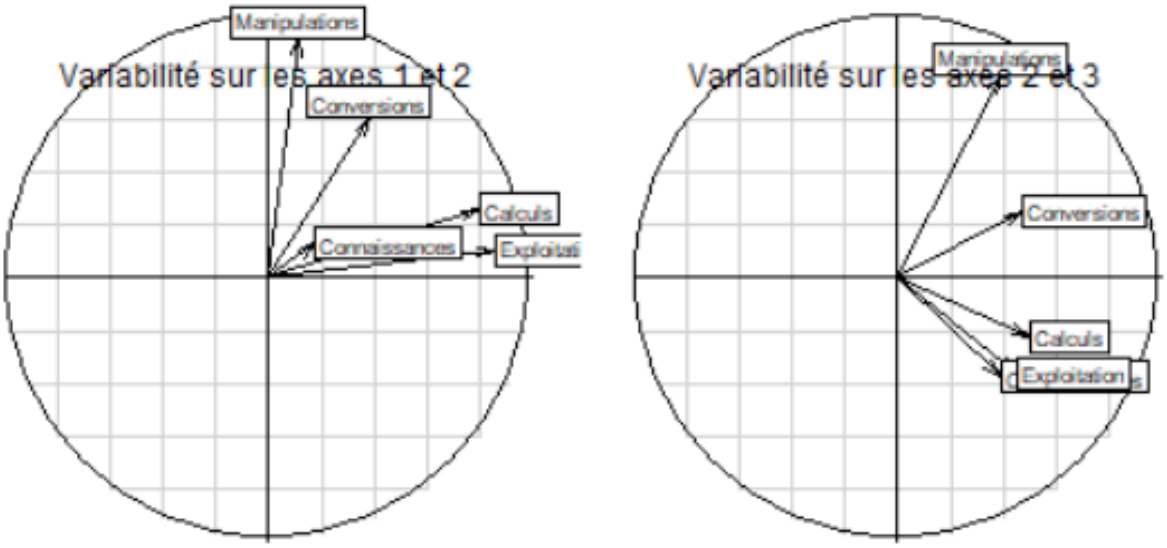
3.B Rotation

Tout le monde utilise la même rotation : la rotation varimax (orthogonale)

Donc en fait c'est plus vraiment indéterminé.

C'est comme si on faisait une ACP et qu'on réorganisait les axes pour mieux comprendre et mieux voir les choses.

A gauche : avant rotation varimax, à droite : après. On voit que c'est plus flagrant que les variables sont indépendantes.



On voit ainsi clairement par cette ACP que l'on peut décrire les étudiants en fonction de leur connaissances et de leurs aptitudes à la manipulation et la conversion d'unité.

4

5 A nouveau l'exemple sur la dépression :

5.A Analyse factorielle à 4 facteurs avec rotation varimax.

		Factor1	Factor2	Factor3	Factor4
1. Observed painfull sadness	EHD1	-0.111	0.050	0.301	**0.773**
2. Emotional hyperexpressiveness	EHD2	**0.663**	0.016	-0.104	**0.413**
3. Emotional instability	EHD3	**0.718**	0.036	-0.120	-0.076
4. Observed monotony	EHD4	**-0.829**	-0.174	0.174	-0.004
5. Lack spontaneous expressivity	EHD5	**-0.672**	-0.219	0.197	-0.094
6. Lack affective reactivity	EHD6	**-0.583**	-0.196	0.162	-0.153
7. Emotional incontinence	EHD7	**0.510**	0.053	0.002	0.385
8. Affective hyperesthesia	EHD8	0.290	0.199	-0.010	0.103
9. Observed explosive mood	EHD9	**0.411**	0.175	0.098	-0.169
10. Worried gesture	EHD10	0.083	0.117	-0.063	**0.419**
11. Observed anhedonia	EHD11	-0.133	0.077	**0.872**	0.130
12. Felt sadness	EHD12	0.122	0.112	0.114	**0.657**
13. Situational anhedonia	EHD13	-0.011	0.096	**0.873**	0.023
14. Felt affective indifference	EHD14	-0.089	**-0.548**	**0.449**	-0.194
15. Hypersensibility unpleasent events	EHD15	0.143	**0.793**	0.017	0.187
16. Sensory anhedonia	EHD16	-0.141	-0.157	**0.442**	0.087
17. Felt affective monotony	EHD17	-0.279	-0.347	0.362	0.008
18. Felt hyperemotionalism	EHD18	0.231	**0.581**	-0.172	0.320
19. Felt irritability	EHD19	0.146	**0.829**	0.027	0.016
20. Felt explosive mood	EHD20	0.174	**0.698**	0.047	0.004

- En gras : coefficients <u>élevés</u>
- Ici : Factor 1 (θ) correspond à un groupe de variables (instabilité émotionnelle : en + et en -) $\longrightarrow$ **dimension** car de -1 à +1
- Factor 2 (μ) : symptômes ressentis (*Felt symptoms*)
- Factor 3 (ν) : anhédonie
- Factor 4 (ξ) : tristesse + mimique inquiète + hyperexpressivité émotionnelle

6 Relation ACP / Analyse factorielle

Si on avait fait une analyse factorielle à 2 facteurs, on aurait retrouvé à peu près les mêmes résultats que l'ACP.

Mais c'est pas pareil :

- ACP : droites avec directions qui représentent l'étirement maximal des variables
- Analyse factorielle : modèle avec coefficients et résidus

Donc pendant longtemps, on a fait de l'ACP (plus simple) pour faire de l'analyse factorielle car l'un était une approximation de l'autre.

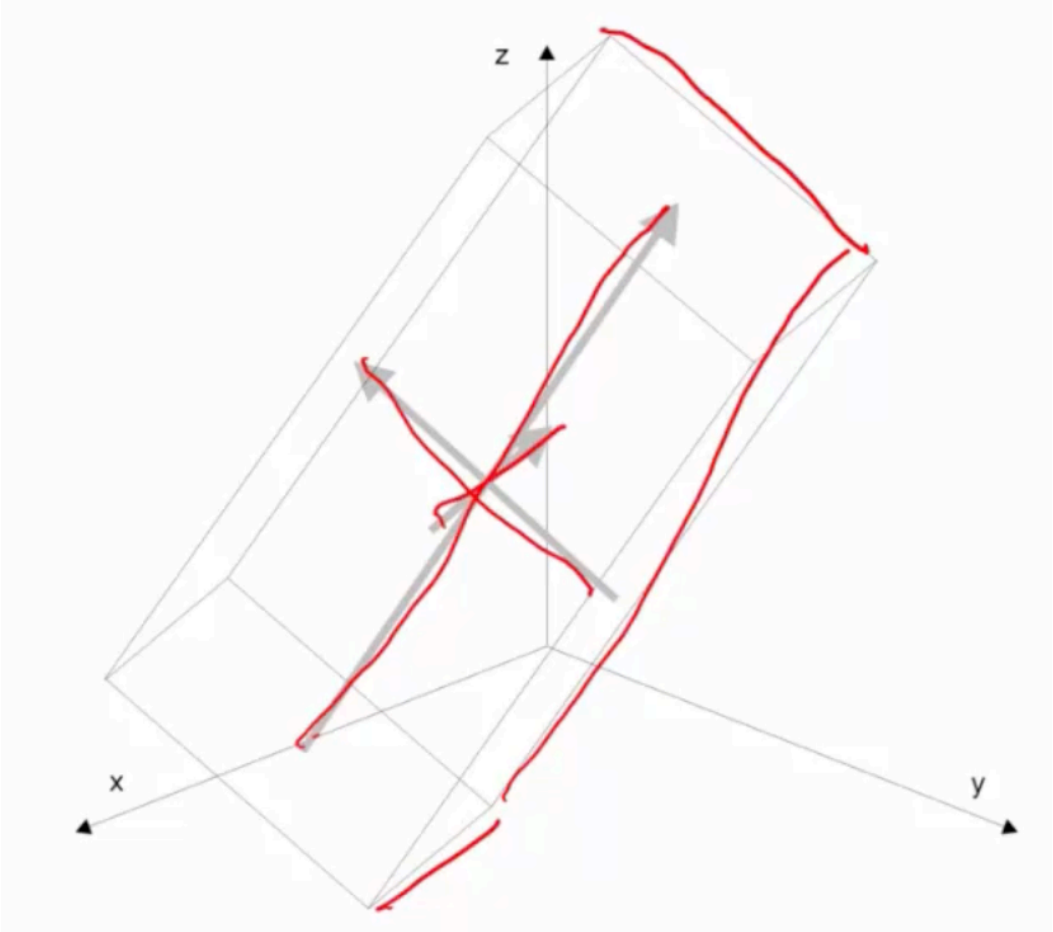
Sur R :

- `factanal()` : fonction pour faire de l'analyse factorielle (package : stats)
- `printcomp()` ou `mdspca` : fonction pour faire de l'ACP (packages : psych, psy)

7 Comment choisir le nb de facteurs ?

Il existe un "test d'adéquation" mais en vrai c'est pas terrible car on cherche toujours à rejeter l'hypothèse nulle (et on ne connait pas β)

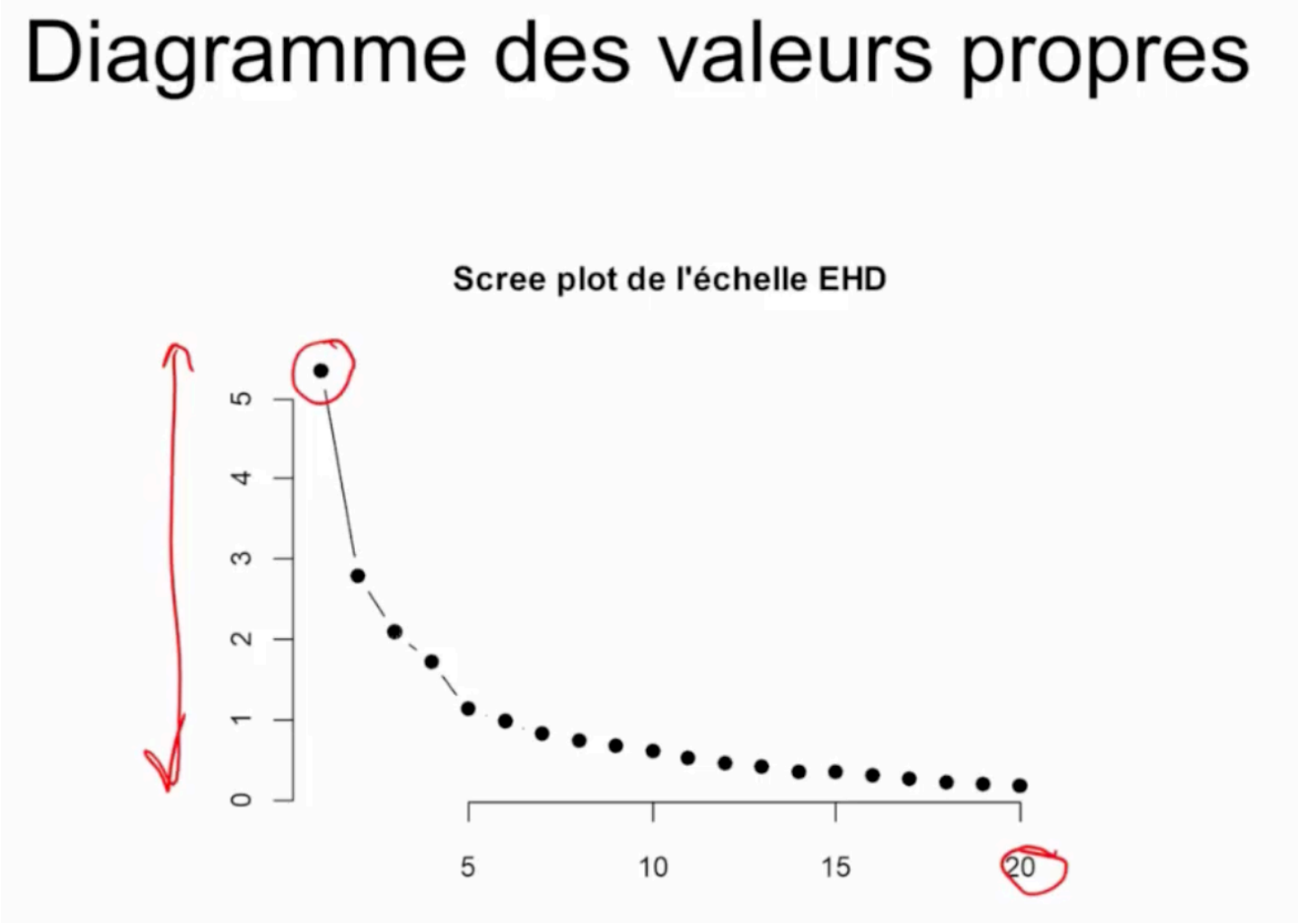
Autre méthode :

1. Faire une ACP (qui enferme le jeu de données dans une autre boîte)
2. Si une dimension enferme bcp de données (c'est à dire qu'elle a une grosse valeur propre) : on la garde
    - valeurs propres : dans la boite de l'ACP (la première dimension) : chaque valeur propre correspond à une "longueur" de la boîte
    - 1ère valeur propre = plus grande longueur de la boîte
    - Donc valeur propre = **variance des points sur les composantes principales**
    - 

    - Si la 1ère est très longue et les autres non : la boite rssemble à un stylo

- Dans ce cas on fera une analyse factorielle à 1 facteur

3. Pour faire une sélection : on fait une analyse parallèle
    1. On simule des données aléatoires au sein du jeu de données (même nb de variables et d'observations) : par ex 20
    2. On fait un scree plot (= diagramme des valeurs propres) pour voir où il y a un coude, et on garde les facteurs avant le coude.



Code R pour faire un scree plot : `fa.parallel()`

Questions :

Si les facteurs n'évoquent rien de tangible / connu $\rightarrow$: dans ce cas on essaie d'autres facteurs / en prendre +…

Bonus :

Qu'est-ce qu'une valeur propre (dans une ACP ou analyse factorielle) ?

On imagine un gros nuage de points (les données) avec une direction : un axe.

$\rightarrow$ La valeur propre = la quantité de variabilité (d'information) que cet axe arrive à capturer.

- Une grande valeur propre = l'axe explique beaucoup de choses dans les données.

- donc cet axe (= ce facteur) est utile, on le garde.
- Une petite valeur propre = l'axe n'explique pas grand-chose, il est presque du bruit = on ne le garde pas.

Il faut imaginer qu'on regarde les données dans un certain sens :

- Si on regarde dans le bon sens, on voit que les données sont très étirées : grande valeur propre.
- Si on regarde les données dans un sens ou les formes sont rondes : petite valeur propre.

8 Références

Vidéo Falissard:

- Blog larmarange
- Exercice CNRS
- Lyon

29Clinimetrie

1 Définition de la clinimétrie

La clinimétrie est une discipline qui s'intéresse à la mesure des phénomènes cliniques, en particulier à travers la création et l'évaluation d'instruments de mesure tels que les échelles, les questionnaires et les tests.

Elle vise à quantifier des aspects qualitatifs de la santé et de la maladie, tels que la douleur, la qualité de vie, la fonction physique ou mentale, afin de fournir des données objectives pour la prise de décision clinique et la recherche médicale.

En résumé : clinimétrie = science de la mesure clinique.

Exemple : variabilité inter-appareil de mesure de la TA

Mesure à prendre : **centralisation des mesures** (envoyer tous les échantillons à un même laboratoire pour analyse)

2 Accord inter-juge

2.A Accord inter-juge d'une variable binaire

2.A.1 Exemple : diagnostic d'une appendicite (OUI/NON)

	Juge1_reponse_positive	Juge1_reponse_negative
Juge2_reponse_positive	a	b
Juge2_reponse_negative	c	d

Comment quantifier / exprimer le fait qu'ils sont d'accord ?

2.A.2 Concordance

Concordance = % de fois où les juges sont d'accord

Formule :

$$\frac{a+d}{a+b+c+d}$$

- Limite : ne tient pas compte du fait que les juges peuvent être d'accord par hasard (s'ils tirent à pile ou face, ils seront d'accord 50% du temps en moyenne)
- Donc peut être élevée par hasard

2.A.3 Concordance corrigée

Concordance corrigée = Kappa de Cohen

- Kappa κ = $\frac{(Concordance - ConcordanceLieAuHasard)}{(1 - ConcordanceLieAuHasard)}$
  - Conc : concordance = $\frac{a+d}{a+b+c+d}$
  - ConcH : concordance liée au hasard

Calcul de la concordance liée au hasard :

$$\frac{(a+b)(a+c)+(c+d)(b+d)}{(a+b+c+d)^2}$$

- Si concordance parfaite : Kappa = 1
- Si concordance égale à celle du hasard : Kappa = 0 (ce n'est pas l'absence de concordance, c'est la concordance comme le voudrait le hasard)

Calculer la concordance liée au hasard :

Exemple :

- Diagnostic de SpA par deux rhumatologues
  - sur 100 patients : on prend dossier + examens complémentaires etc
  - les rhumatologues doivent dire OUI/NON au diagnostic de SpA
  - le truc c'est que les rhumatologues savent que les patients viennent d'une consult de rhumatologie
  - donc connaisse à peu près la prévalence de la SpA dans cette population (disons 10%)

	Juge1_reponse_positive	Juge1_reponse_negative	
Juge2_reponse_positive	a	b	0,1
Juge2_reponse_negative	c	d	0,9
	0,1	0,9	

Marges =- totaux sur les bords du tableau*

> **!** Important
>
> $\rightarrow$ probabilité d'être en a = 0,1 - 0,1 = 0,01
> $\rightarrow$ probabilité d'être en b = 0,1 - 0,9 = 0,09
> $\rightarrow$ probabilité d'être en c = 0,9 - 0,1 = 0,09
> $\rightarrow$ probabilité d'être en d = 0,9 - 0,9 = 0,81

Mais la démonstration :

- Probabilité que les deux juges répondent positivement : P(J1+) - P(J2+)
  - = $\frac{(a+b)}{(a+b+c+d)} - \frac{(a+c)}{(a+b+c+d)}$

    - Proportion de OUI du Juge 1 : $\frac{(a+b)}{N}$
    - Proportion de OUI du Juge 2 : $\frac{(a+c)}{N}$
    - N = (a + b + c + d) = total des patients
  - Probabilité de P(J1+) **UNION** P(J2+) = $\frac{(a+b)(a+c)}{N^2}$
  - = probabilité qu'un patient pris au hasard soit classé OUI par le Juge 1 ET OUI par le Juge 2, par hasard, si leurs réponses sont indépendantes.
- Calculer aussi la probabilité que les deux juges répondent négativement :
  - = Accord sur OUI + Accord sur NON
  - $\frac{(a+b)*(a+c)}{N^2} + \frac{(c+d)*(b+d)}{N^2}$
  - = $\frac{0{,}1*0{,}1}{N^2} + \frac{0{,}9*0{,}9}{N^2}$
  - = $\frac{0{,}01}{N^2} + \frac{0.81}{N^2}$
  - = $\frac{0{,}82}{N^2}$

En gros :

- \$$a = 0,1^2 = 0,01$
- \$$d = 0,9^2 = 0,81$
- $Concordance\ due\ au\ hasard\ conditionnelle\ aux\ marges = 0,82$

Limites :

- Kappa dépend des marges (prévalence de la condition)
- Kappa peut être faible même si la concordance est élevée (si marges très déséquilibrées)

≈ Resseble au coefficient de corrélation intraclasse (ICC) pour variables quantitatves

2.A.4 Exemple diagnostic à plusieurs classes

- Diagnostic de douleurs abdominales sur causes de douleurs abdominales (7 causes)
- 2 juges

	Juge1_X1	Juge1_X2	Juge1_X3	Juge1_X4	Juge1_X5	Juge1_X6	Juge1_X7
Juge2_X1	a	b	c	d	e	f	g
Juge2_X2	h	i	j	k	l	m	n
Juge2_X3	o	p	q	r	s	t	u
Juge2_X4	v	w	x	y	z		
Juge2_X5							
Juge2_X6							
Juge2_X7							

- Kappa κ = $\frac{(Concordance - ConcordanceLieAuHasard)}{(1 - ConcordanceLieAuHasard)}$
- Concordance = % sur la diagonale (a i j r..)

- Concordance liée au hasard
  - Somme des marges multipliées entre elles
  - Marge J1_X1 - Marge J2_X1 + Marge J1_X2 - Marge J2_X2 + …
  - Divisé par N^2
- Limite : Kappa dépend des marges (prévalence de chaque classe)

Dans R : fonction `Kappa.test()` du package `psy`

2.B Acord inter-juge d'une variable ordonnée

Exemple : frottis négatif - douteux - positif

Tableau :

	Juge1_negatif	Juge1_douteux	Juge1_positif
Juge2_negatif	a	b	c
Juge2_douteux	d	e	f
Juge2_positif	g	h	i

- Si on considère que c'est catégoriel : Kappa simple
  - Kappa κ = $\frac{(Concordance - ConcordanceLieAuHasard)}{(1 - ConcordanceLieAuHasard)}$
  - Problème :
    * en b, le Juge 1 a dit "douteux" et le Juge 2 "négatif"
    * et c'est pas pareil que si le Juge 1 avait dit "positif" et le Juge 2 "négatif" (c'est pire !)
    * valeurs en c : plus graves que valeurs en b
- Solution : Kappa pondéré (pondération des désaccords selon leur gravité)

Kappa pondéré : (Kappa weighted)

$$\kappa_w = \frac{(a + e + i) + (\frac{b+f+d+h}{2}) + 0}{Nombre\,de\,coefficients}$$

$$

2.C Avec 3 juges

- Utilisation de la méthode de Light
- Extension du Kappa de Cohen pour plus de 2 juges
- 3 calculs :
  - Kappa entre juge 1 et 2
  - Kappa entre juge 1 et 3

  - Kappa entre juge 2 et 3
  - Et calculer la moyenne des 3 kappas

Minimum 30 sujets si possible

2.D Interprétation du Kappa de Cohen

Plus le Kappa est proche de 1, meilleur est l'accord entre les juges et moins il est probable que cet accord soit dû au hasard.

Plus le Kappa est proche de 0, plus l'accord entre les juges est faible et plus il est probable que cet accord soit dû au hasard.

Kappa κ	Interprétation de l'accord entre juges
< 0	Accord moins que par hasard
0 - 0,20	Accord faible
0,21 - 0,40	Accord modéré
0,41 - 0,60	Accord substantiel
0,61 - 0,80	Accord fort
0,81 - 1,00	Accord presque parfait
> 1	Accord parfait

2.E Intervalle de confiance du Kappa

κ ne suit pas une loi normale et a une distribution asymétrique

Le plus simple : **bootstrap**

Bootstrap :

1. Tirer un échantillon bootstrap (avec remise) de la taille de l'échantillon initial
2. Calculer le Kappa sur cet échantillon bootstrap
3. Répéter les étapes 1 et 2 un grand nombre de fois (ex : 1000)
4. Utiliser les percentiles 2,5% et 97,5% des Kappas bootstrap pour construire l'IC à 95%

2.F Différence avec PABAK

- Kappa a des problèmes de **paradoxe** (dépendance aux marges)
  - Si prévalence très faible ou très élevée, Kappa peut être faible même si la concordance est élevée
  - Si biais entre juges (l'un est plus sévère que l'autre), Kappa peut être faible même si la concordance est élevée
  - par exemple : si un juge dit OUI 90% du temps et l'autre 10% du temps, Kappa sera faible même si ils sont d'accord 50% du temps
- **PABAK** (Prevalence-Adjusted Bias-Adjusted Kappa) ajuste le Kappa pour tenir compte de la prévalence et du biais

- Ça n’a pas d’intérêt selon Bruno Falissard

2.G Exemples sur R

2.G.1 2 catégories

```r
library(psy)
data(expsy)
expsy
```

```
   it1 it2 it3 it4 it5 it6 it7 it8 it9 it10 r1 rb1 r2 rb2 r3 rb3
1   NA   2   2   1   1   0   3   1   2    3 NA  NA  3   1  3   1
2    3   1   1   2   1   0   3   1   2    3  3   1  3   1  3   1
3    3   1   1   0   2   0   3   1   2    3  3   1  2   0  2   0
4    4   1   0   2   1   1   4   1   1    2  4   1  4   1  4   1
5    4   1   0   2   1   1   3   1   1    2  3   1  3   1  3   1
6    3   2   0   1   2   2   3   1   2    3  3   1  2   0  4   1
7    3   2   3   1   2   2   2   1   3    3  2   0  2   0  2   0
8    3   2   2   0   2   1   3   0   1    3  3   1  2   0  3   1
9    3   2   2  NA   2   0   3   0   1    2  3   1  3   1  3   1
10   3   2   1   0   2   2   3   2   2    3  3   1  3   1  2   0
11   4   2   2   0   2   2   3   1   1    4  3   1  4   1  4   1
12   4   2   1   2   2   2   3   1   2    1  3   1  3   1  3   1
13   4   1   3   2   2   2   3   1   2    3  3   1  3   1  3   1
14   3   2   2   1   1   2   3   0   3    4  3   1  2   0  2   0
15   4   0   3   1   2   0   3   3   1    4  3   1  3   1  3   1
16   4   0   1   2   2   2   4   2   3    3  4   1  4   1  4   1
17   4   0   1   2   2   2   4   2   3    3  4   1  4   1  4   1
18   4   1   1   2   2   1   3   1   0    4  3   1  3   1  3   1
19   4   1   1   2   0   0   4   2   1    3  4   1  3   1  3   1
20   3   1   1   2   1   1   2   1   1    3  2   0  2   0  2   0
21   4   2   3   2   1   1   4   2   2    3  4   1  4   1  4   1
22   3   2   1   2   1   1   3   2   1    3  3   1  4   1  3   1
23   3   1   1   2   1   1   3   2   1    3  3   1  3   1  2   0
24   3   2   1   2   2   2   4   1   1    3  4   1  4   1  4   1
25   3   1   2   2   2   2   4   0   3    3  4   1  4   1  3   1
26   3   2   3   2   2   2   2   2   3    3  2   0  2   0  2   0
27   3   2   2   2   2   2   1   1   2    3  1   0  1   0  2   0
28   3   2   1   1   1   1   2   2   2    2  2   0  1   0  1   0
29   3   1   0   2   2   2   2   2   1    3  2   0  2   0  2   0
30   3   2   1   2   2   2   4   2   2   NA  4   1  3   1  4   1
```

Ici :

- 30 patients
- 10 items (it1, it2, …, it10)
- 3 scores (r1, r2, r3) qui ont chacun été binarisés (rb1, rb2, rb3)

```r
# Calcul de Kappa entre les colonnes 12 et 14 (rb1 et rb3)
ckappa(expsy[, c("rb1", "rb3")])
```

```
$table
  0  1
0 6  0
1 4 19

$kappa
[1] 0.6627907
```

```r
#autre syntaxe possible
ckappa(data.frame(expsy$rb1, expsy$rb3))
```

```
$table
  0  1
0 6  0
1 4 19

$kappa
[1] 0.6627907
```

Interprétation

- Concordance : (6 + 19) / (6 + 4 + 19 + 0) = 25 / 29 = 0,862
- Concordance liée au hasard :
  - = ((6 + 0) - (6 + 4) + (4 + 19) - (0 + 19)) / (29^2)
  - = (6 - 10 + 23 - 19) / 841
  - = (60 + 437) / 841
  - = 497 / 841 = 0,591
- Kappa = (0,862 - 0,591) / (1 - 0,591) = 0,6627

2.G.2 3 catégories

```r
# Calcul de Kappa entre les colonnes r1 et r2 (3 catégories)
ckappa(expsy[, c("r1", "r2")])
```

```
$table
  1 2 3 4
1 1 0 0 0
2 1 4 0 0
3 0 4 9 2
```

```
4 0 0 2 6

$kappa
[1] 0.5421053
```

2.G.3 kappa pondéré

```r
# Calcul de Kappa pondéré entre les colonnes r1 et r2 (3 catégories)
wkappa(expsy[, c("r1", "r2")])
```

```
$table
  1 2 3 4
1 1 0 0 0
2 1 4 0 0
3 0 4 9 2
4 0 0 2 6

$weights
[1] "squared"

$kappa
[1] 0.7819549
```

2.G.4 kappa de light (3 juges)

```r
# Calcul de Kappa de Light entre les colonnes r1, r2 et r3 (3 juges)
lkappa(expsy[, c("it10", "rb1", "rb2")])
```

```
[1] 0.2276101
```

ou pour le retrouver soi même :

```r
# Calcul de Kappa de Light entre les colonnes r1, r2 et r3 (3 juges)
kappa_1 <- ckappa(expsy[, c("it10", "rb1")])$kappa #$kappa sert à extraire juste
 ↪  la valeur du kappa
kappa_2 <- ckappa(expsy[, c("it10", "rb2")])$kappa
kappa_3 <- ckappa(expsy[, c("rb1",  "rb2")])$kappa

kappa_light_manu <- (kappa_1 + kappa_2 + kappa_3) / 3
kappa_light_manu
```

```
[1] 0.2276101
```

> **Tip**
>
> `c.kappa()` gère les données manquantes en les ignorant (pairwise deletion)

3 Accord inter-juge pour variable quantitative

- 2 juges
- Score Y contenu sur 40 patients

$Y = a + b\,Juges + c\,Patients + \epsilon$

Puis analyse de variance à effets aléatoires : permet de savoir quelle part de la variance totale est due aux juges, aux patients, et à l'erreur résiduelle.

3.A Coefficient de corrélation intraclasse (ICC)

$$ICC = \frac{Variance\,entre\,Patients}{Variance\,totale}$$

- Variance totale = Variance entre patients + Variance entre juges + Variance résiduelle
  - Variance liée au juges : erreur systématique entre juges (si un juge est plus sévère que l'autre)
  - Variance résiduelle : erreur aléatoire (variabilité inexpliquée)
  - Variance entre patients : variabilité réelle entre les patients : **tout ce qui n'est pas une erreur**

En gros, le coefficient de corrélation intraclasse (ICC) est le rapport entre la **bonne variance** (variance entre patients) et la **variance totale** (toutes les sources de variance, y compris les erreurs).

- Si ICC proche de 1 : grande partie de la variance est due aux différences réelles entre les patients (bonne fiabilité entre juges)
- Si ICC proche de 0 : grande partie de la variance est due aux erreurs (mauvaise fiabilité entre juges)

> **Tip**
>
> Comme il ne s'agit pas d'un test, Y n'a pas besoin de suivre une loi normale.
> Donc Y pourrait être binaire !
> Et les juges se prononceraient OUI/NON donc on pourrait calculer un κ de Cohen !
> En pratique : c'est la même chose **asymptotiquement**, c'est à dire si le nombre de sujets est grand.

3.A.1 Si Y est une variable ordonnée

- $\kappa\,ponderee$ en $1/N^2 = ICC$

3.B Exemple sur R

```r
ckappa(expsy[, c("rb1", "rb2")])
```

```
$table
  0  1
0 6  0
1 4 19

$kappa
[1] 0.6627907
```

```r
icc(expsy[, c("rb1", "rb2")])
```

```
$nb.subjects
[1] 29

$nb.raters
[1] 2

$subject.variance
[1] 0.1403941

$rater.variance
[1] 0.007389163

$residual
[1] 0.06157635

$icc.consistency
[1] 0.695122

$icc.agreement
[1] 0.6705882
```

Kappa = 0,66 et ICC = 0,67

> **Tip**
>
> **Différence ICC agreement et ICC consistency**
>
> - **ICC agreement** :
> - prend en compte les différences systématiques entre juges (si un juge est plus sévère que l'autre)
> - formule :
>
> $$ICC_{agreement} = \frac{Variance\,entre\,Patients}{Variance\,entre\,Patients + Variance\,entre\,Juges + Variance\,rsid}$$

- **ICC consistency** :
  - ne prend pas en compte les différences systématiques entre juges (on s'en fiche si un juge est plus sévère que l'autre)
  - formule :

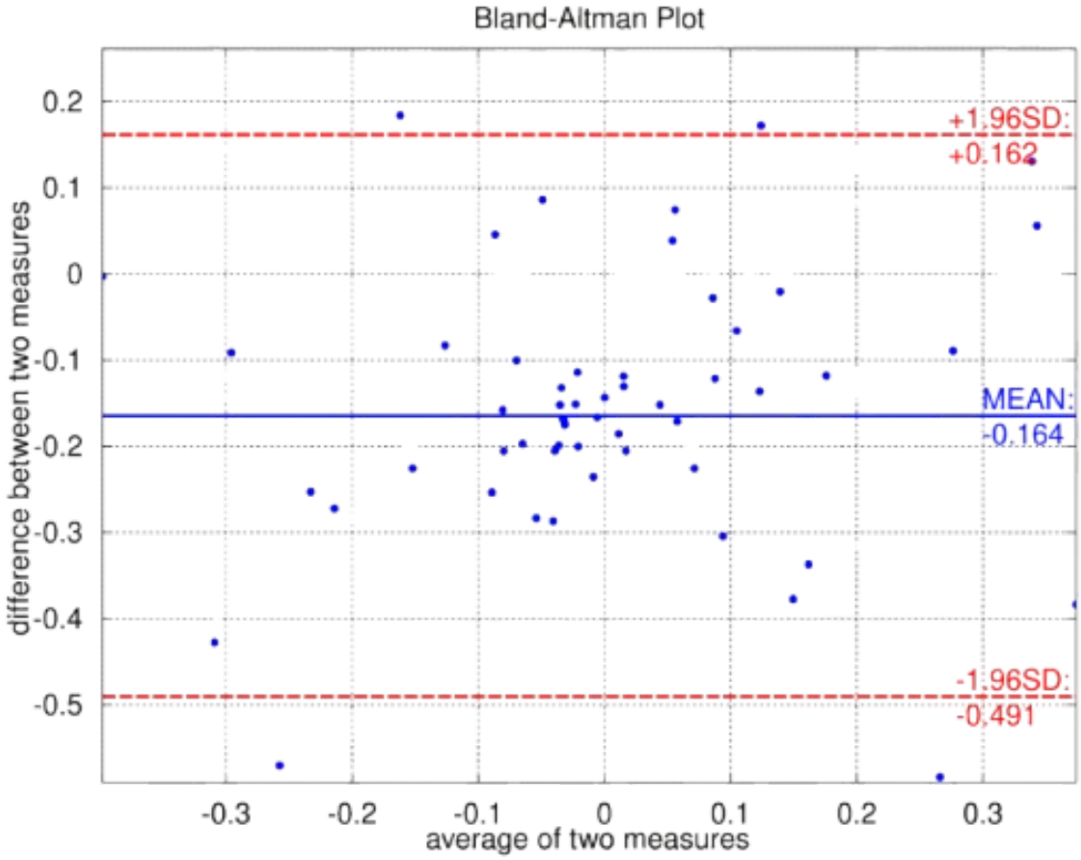
$$ICC_{consistency} = \frac{Variance\ entre\ Patients}{Variance\ entre\ Patients + Variance\ rsiduelle}$$

ICC consistency : utile par exemple dans un essai randomisé dans lequel on compare la fin *versus- le début de l'étude.

/newpage

4 Blant et Altman

- Variable QUANTITATIVE
- 2 juges (ou appareils de mesure)



4.A Principe

- Pour chaque sujet, on calcule la moyenne des deux mesures
- Pour chaque sujet, on calcule la différence entre les deux mesures
- On trace la différence en fonction de la moyenne

5 Exemple

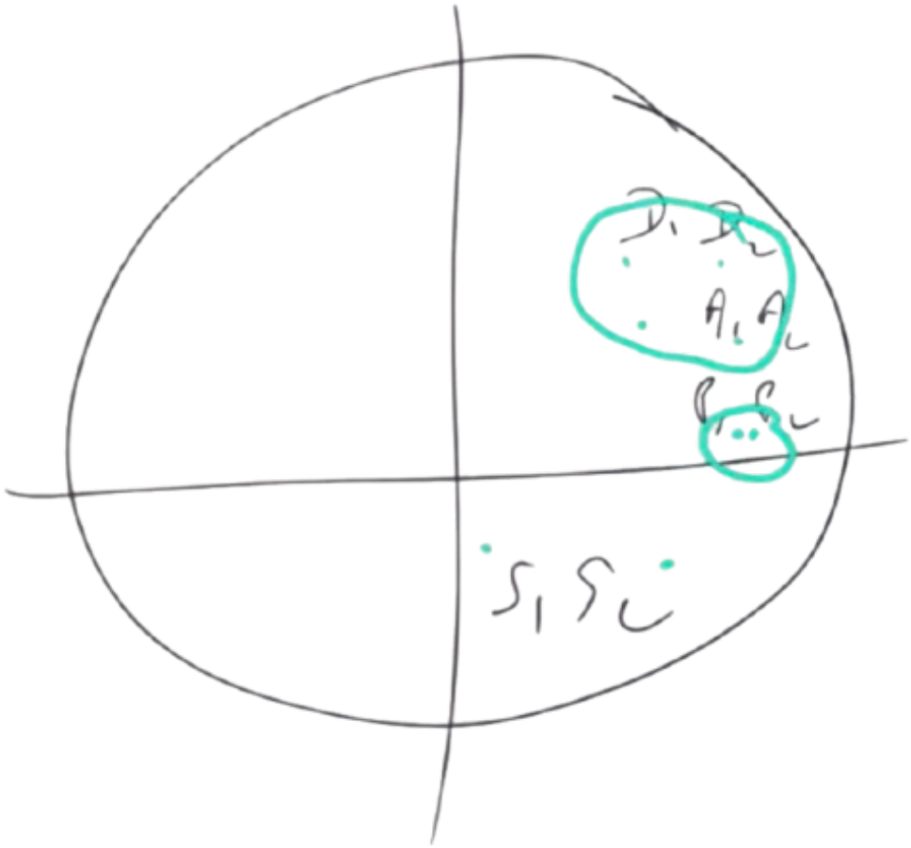
- 2 psychiatres
- Évaluent les patients avec des entretiens structurés
- donc calcul de coefficients kappa !

En tableau :

Patient	Depression_J1	Depression_J2	TAG_J1	TAG_J2	PTSD_J1	PTSD_J2	Schizo_J1	Schizo_J2
1	12	14	8	7	5	6	3	4
2	20	18	10	11	7	8	5	6
…	…	…	…	…	…	…	…	…

Panorama GLOBAL :

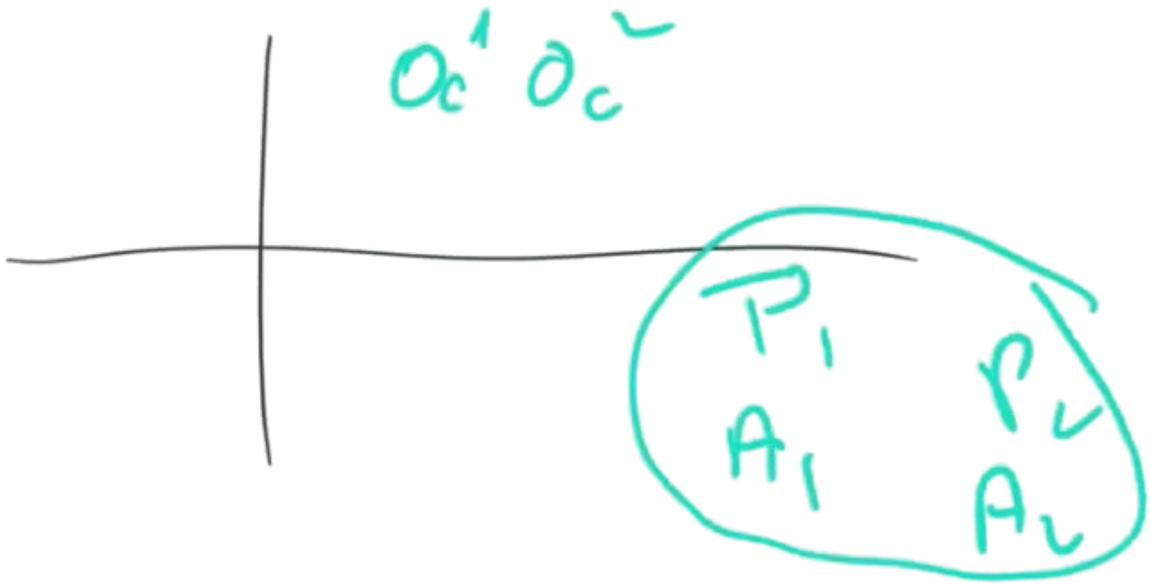
Analyse en composante principales (ACP) : faisable même si qualitatif ! et même si sur des variables binaires (faisable dans tous les cas)



- D1 et D2 un peu idem que A1 et A2
- PTSD1 et PTSD2 : très proches !
- Schizo1 et Schizo2 : bcp plus éloignés

Analyse des correspondances multiples (ACM) : pas vraiment applicable dans ce cas

- Même idée que ACP mais pour des variables qualitatives
- Si diagnostic avec 7 catégories



- Appendicite et péritonites : bcp plus proches, tendance à permuter les diagnostics !

---

6 Alpha de Cronbach

- = Accord “intra-juge” du patient avec lui-même (test + retest)
- Pas vraiment de la reproductibilité
- Mesure de la cohérence interne d’une échelle

6.A Exemple

Échelle unidimensionnelle (c’est à dire que tous les items mesurent la même chose)
H I1 I2 I3 I4 I5 I6

- Si on prend 2 groupes de 3 items H’ = (I1, I2, I3) et H’ ’ = (I4, I5, I6)
  - *mais on pourrait faire H’ = (I1, I4, I5) et H’ ’ = (I2, I3, I6) aussi*
- C’est comme 2 sous-échelles de H (qui mesurent chacune la même chose que H)
- et corréler une sous-échelle avec H ou l’autre : c’est mesurer la cohérence entre H et H
- Coefficients H’ et H’ ’ = *split-half correlation*

6.B Calcul

Pour calculer l’alpha de Cronbach, on fait la moyenne des corrélations entre **toutes les paires possibles** de sous-échelles.

(α de Cronbach = moyenne des split-half correlations) = corrélation de l'échelle avec elle-même.

6.C Exemple sur R

```r
library(psy)
data(expsy)
expsy
```

```
   it1 it2 it3 it4 it5 it6 it7 it8 it9 it10 r1 rb1 r2 rb2 r3 rb3
1   NA   2   2   1   1   0   3   1   2    3 NA  NA  3   1  3   1
2    3   1   1   2   1   0   3   1   2    3  3   1  3   1  3   1
3    3   1   1   0   2   0   3   1   2    3  3   1  2   0  2   0
4    4   1   0   2   1   1   4   1   1    2  4   1  4   1  4   1
5    4   1   0   2   1   1   3   1   1    2  3   1  3   1  3   1
6    3   2   0   1   2   2   3   1   2    3  3   1  2   0  4   1
7    3   2   3   1   2   2   2   1   3    3  2   0  2   0  2   0
8    3   2   2   0   2   1   3   0   1    3  3   1  2   0  3   1
9    3   2   2  NA   2   0   3   0   1    2  3   1  3   1  3   1
10   3   2   1   0   2   2   3   2   2    3  3   1  3   1  2   0
11   4   2   2   0   2   2   3   1   1    4  3   1  4   1  4   1
12   4   2   1   2   2   2   3   1   2    1  3   1  3   1  3   1
13   4   1   3   2   2   2   3   1   2    3  3   1  3   1  3   1
14   3   2   2   1   1   2   3   0   3    4  3   1  2   0  2   0
15   4   0   3   1   2   0   3   3   1    4  3   1  3   1  3   1
16   4   0   1   2   2   2   4   2   3    3  4   1  4   1  4   1
17   4   0   1   2   2   2   4   2   3    3  4   1  4   1  4   1
18   4   1   1   2   2   1   3   1   0    4  3   1  3   1  3   1
19   4   1   1   2   0   0   4   2   1    3  4   1  3   1  3   1
20   3   1   1   2   1   1   2   1   1    3  2   0  2   0  2   0
21   4   2   3   2   1   1   4   2   2    3  4   1  4   1  4   1
22   3   2   1   2   1   1   3   2   1    3  3   1  4   1  3   1
23   3   1   1   2   1   1   3   2   1    3  3   1  3   1  2   0
24   3   2   1   2   2   2   4   1   1    3  4   1  4   1  4   1
25   3   1   2   2   2   2   4   0   3    3  4   1  4   1  3   1
26   3   2   3   2   2   2   2   2   3    3  2   0  2   0  2   0
27   3   2   2   2   2   2   1   1   2    3  1   0  1   0  2   0
28   3   2   1   1   1   1   2   2   2    2  2   0  1   0  1   0
29   3   1   0   2   2   2   2   2   1    3  2   0  2   0  2   0
30   3   2   1   2   2   2   4   2   2   NA  4   1  3   1  4   1
```

```r
# Calcul de l'alpha de Cronbach sur les 10 items
cronbach(expsy[, 1:10])
```

```
$sample.size
```

```
[1] 27

$number.of.items
[1] 10

$alpha
[1] 0.1762655
```

Ici :

- On calcule l'alpha de Cronbach sur les 10 items (it1 à it10)
- α = 0,17 : faible cohérence interne entre les items

6.D Interprétation

- α < 0,6 : cohérence interne faible items peu liés
- 0,6 ≤ α < 0,9 : c'est bien !
- α ≥ 0,9 : cohérence interne très élevée items redondants (c'est presque trop)

PLUS LE NOMBRE D'ITEMS EST ÉLEVÉ, PLUS L'ALPHA A TENDANCE À ÊTRE ÉLEVÉ

---

1. Le coefficient alpha de Cronbach : rôle et interprétation

Le coefficient alpha de Cronbach sert à évaluer **la cohérence interne** d'un ensemble d'items censés mesurer **une seule et même dimension latente** (ex. anxiété, motivation, douleur). *Mesure latente = on ne peut pas la mesurer directement, on utilise des items comme indicateurs.*

6.D.1 Ce que mesure l'alpha

- Le degré auquel les items sont **corrélés entre eux**.
- Plus les items vont "dans le même sens", plus l'alpha est élevé.

6.D.2 Comment interpréter l'alpha

- α < 0,6 : cohérence interne faible items peu liés
- 0,6 ≤ α < 0,9 : c'est bien !
- α ≥ 0,9 : cohérence interne très élevée items redondants (c'est presque trop)

6.D.3 Limite fondamentale

L'alpha n'a de sens **que si le questionnaire est unidimensionnel**, c'est à dire que tous les items mesurent la même dimension latente. S'il existe plusieurs dimensions cachées, l'alpha peut être trompeusement élevé.

---

7 2. Pourquoi l'alpha nécessite un questionnaire unidimensionnel ?

Parce que l'alpha repose sur l'hypothèse que :

- toutes les corrélations entre items proviennent **d'un seul facteur latent**.

Si les items mesurent plusieurs dimensions différentes :

- certaines corrélations sont fortes au sein d'une dimension (dimension = groupe d'items liés),
- d'autres sont faibles entre dimensions,
- l'ensemble devient hétérogène,
- l'alpha ne reflète plus une seule cohérence interne.

D'où la nécessité de vérifier l'**unidimensionnalité** avant de rapporter l'alpha.

---

8 3. Comment vérifier l'unidimensionnalité ?

8.A Par l'analyse des valeurs propres (scree plot)

On réalise une **ACP** ou une **analyse factorielle exploratoire**, puis on examine les **valeurs propres**.

8.A.1 Rappel : qu'est-ce qu'une valeur propre ?

Une valeur propre représente :

- la **quantité de variance expliquée** par une composante (ou un facteur).

Repères :

- Valeur propre = 1 : autant de variance qu'un item
- Valeur propre > 1 : facteur important
- Une seule valeur propre nettement plus élevée que les autres = un seul facteur dominant

8.A.2 Principe

- **Unidimensionnel** = 1 seul facteur explique l'essentiel de la variance 1 valeur propre haute
- **Multidimensionnel** = plusieurs facteurs importants plusieurs valeurs propres hautes

---

9 4. Exemple complet

9.A Questionnaire de 6 items censés mesurer l'anxiété

(ex : tension, inquiétude, nervosité, irritabilité, ruminations, difficultés à se détendre)
Après avoir collecté les réponses de 200 personnes, on réalise une ACP.

9.A.1 Étape A : valeurs propres obtenues

```
Facteur 1 : 3,8
Facteur 2 : 0,7
Facteur 3 : 0,6
Facteur 4 : 0,4
Facteur 5 : 0,3
Facteur 6 : 0,2
```

9.A.2 Interprétation

- Le **facteur 1** explique **3,8 fois plus de variance qu'un item**.
- Les autres valeurs propres sont petites et décroissantes.
- Le **diagramme des valeurs propres** montre un **coude net après le facteur 1**.

 Le questionnaire est **unidimensionnel**. Calculer l'alpha de Cronbach est pertinent.

9.A.3 Étape B : alpha de Cronbach

On obtient par exemple : **α = 0,87**
Interprétation :

- Les 6 items sont homogènes.
- Ils mesurent bien **une seule dimension d'anxiété**.
- Le score total (la somme des 6 items) est cohérent.

---

10 5. Exemple inverse (multidimensionnel)

Si les valeurs propres étaient :

```
Facteur 1 : 2,6
Facteur 2 : 2,0
Facteur 3 : 0,7
Facteur 4 : 0,4
Facteur 5 : 0,2
Facteur 6 : 0,1
```

Interprétation :

- Deux facteurs importants (ex. anxiété + dépression).
- Le questionnaire est **bifactoriel**.
- Un alpha global ne serait **pas valide** (il mélangerait plusieurs dimensions).

---

11 6. Synthèse finale

11..1 Alpha de Cronbach

- Mesure la **cohérence interne** entre items
- Suppose **un seul facteur latent**
- Interprétation correcte seulement en cas d'unidimensionnalité

11..2 Valeurs propres

- Représentent la **variance expliquée** par chaque facteur
- Une seule valeur propre dominante **unidimensionnalité**
- Plusieurs valeurs propres élevées **multidimensionnalité**

11..3 Exemple

- Valeur propre 1 = 3,8 et les autres < 1 OK
- Alpha = 0,87 cohérence interne satisfaisante

---

12 Questions à l'examen sur ce cours

- Ca tombe assez souvent !
- Différence ICC agreement et ICC consistency
- Kappa de Cohen et ICC : proches et différents :

13 1. Ce que mesure Kappa (qualitatif)

13.A Tableau 2×2 (2 juges, variable binaire)

Juge 1 en ligne, Juge 2 en colonne :

	J2 +	J2 –
J1 +	a	b
J1 –	c	d

Taille totale : $N = a + b + c + d$

13.A.1 Concordance brute

Conc $= \frac{a+d}{a+b+c+d}$

13.A.2 Concordance liée au hasard

On utilise les marges (prévalences de OUI/NON) : Conc$_H = \frac{(a+b)(a+c)+(c+d)(b+d)}{(a+b+c+d)^2}$

13.A.3 Kappa de Cohen

$\kappa = \frac{\text{Conc} - \text{Conc}_H}{1 - \text{Conc}_H}$

Interprétation (cours)

- 0,41–0,60 : accord modéré
- 0,61–0,80 : accord substantiel
- 0,81–1 : accord (presque) parfait

13.A.4 Extensions

- **Plus de 2 classes** :
  - Conc = % diagonal
  - Conc_H = somme des produits des marges / N²
- **Variable ordonnée** : on utilise un **kappa pondéré**
- **Plus de 2 juges** : kappa de **Light** = moyenne des kappas par paires

13.A.5 Limites

- Très sensible aux marges (prévalence)
- Peut être faible même si la concordance brute est élevée
- D'où le PABAK (mais « inutile selon Falissard »)

---

14 2. Ce que mesure l'ICC (quantitatif)

$ICC = \frac{\text{Variance entre patients}}{\text{Variance totale}}$

Avec :

Var totale $=$ Var entre patients $+$ Var entre juges $+$ Var résiduelle

- Variance entre patients : variabilité réelle
- Variance entre juges : biais systématique
- Variance résiduelle : erreur aléatoire

Interprétation :

- ICC 1 : bonne fiabilité
- ICC 0 : mauvaise fiabilité

14..1 Cours, « Tip »

- Y n'a pas besoin d'être normale
- Même si Y est binaire on peut calculer κ ou ICC
- « Asymptotiquement, c'est la même chose »

14..2 Cours, 3.A.1.

Pour Y ordonnée : kappa pondéré (poids $1/N^2) = ICC$

---

15 3. En quoi κ et ICC sont proches

- Mesurent l'accord inter-juge
- κ pour qualitatif ; ICC surtout pour quantitatif
- Tous deux : accord réel vs hasard/erreurs
- Valeurs entre 0 et 1
- Sur binaire : κ ≈ ICC
- Sur ordinal : κ pondéré = ICC
- Le cours dit explicitement qu'ils mesurent la même idée de **fiabilité**

---

16 4. En quoi ils sont différents

16.A Type de données

- κ : qualitatif (binaire, nominal, ordinal)
- ICC : quantitatif (extensible au binaire/ordinal)

16.B Méthode de calcul

- κ : basé sur tableau de contingence et accord par hasard
- ICC : basé sur décomposition de variance (modèle ANOVA)

16.C Sensibilité aux marges

- κ très sensible
- ICC non formulé en termes de prévalence

16.D Structure d'erreur

- κ ne distingue pas sources d'erreur
- ICC sépare :
  - erreur entre juges
  - erreur résiduelle

---

17 5. Exemple concret du cours (rb1 / rb2)

Tableau :

	J2=0	J2=1
J1=0	6	0
J1=1	4	19

Donc : a = 6, b = 0, c = 4, d = 19, N = 29.

17.A Kappa de Cohen

17.A.1 Concordance brute

Conc $= \frac{25}{29} \approx 0{,}862$

17.A.2 Marges

- J1 : 0 = 6/29 ; 1 = 23/29
- J2 : 0 = 10/29 ; 1 = 19/29

17.A.3 Concordance au hasard

Conc$_H = \frac{6 \cdot 10 + 23 \cdot 19}{29^2} = \frac{497}{841} \approx 0{,}591$

17.A.4 Kappa

$\kappa \approx 0{,}66$

17.B ICC (cours)

- Var patients = 0,1404
- Var juges = 0,0074
- Var Résiduel = 0,0616

Variance totale : $0{,}2094$
$ICC = \frac{0{,}1404}{0{,}2094} \approx 0{,}67$

Conclusion de l'exemple :

- κ = 0,66
- ICC = 0,67

---

18 6. Synthèse courte

18.A Proches car :

- mesurent un accord au-delà du hasard/erreur
- valeurs entre 0 et 1
- sur binaire/ordinal très proches
- κ pondéré = ICC

18.B Différents car :

- κ basé contingence ; ICC basé variance
- κ sensible aux marges ; ICC non
- κ ne distingue pas types d'erreurs ; ICC si

30Wrap Up Clinimétrie

1 Clinimétrie

Objectif de la clinimétrie = reproductibilité de la mesure

2 Mesure catégorielle (diagnostic clinique)

- Mesure catégorielle : concordance entre les juges
- Problème : 50% liée au hasard si 2 juges

2..1 Kappa de Cohen = concordance corrigée pour le hasard

- Kappa de Cohen : mesure de la concordance entre 2 juges en tenant compte du hasard
- Formule : K = (Po - Pe) / (1 - Pe)
  - Po = proportion d'accord observée
  - Pe = proportion d'accord attendu par le hasard
- Exemple : 2 juges évaluent la présence ou l'absence d'une maladie chez 100 patients

	Juge 2 Malade	Juge 2 Non Malade	Total
Juge 1 Malade	40	10	50
Juge 1 Non Malade	20	30	50
Total	60	40	100

  - Po = (40 + 30) / 100 = 0.70
  - Pe = [(50/100) * (60/100)] + [(50/100) * (40/100)] = 0.30 + 0.20 = 0.50
  - K = (0.70 - 0.50) / (1 - 0.50) = 0.40
- Interprétation de Kappa :
  - 0 : concordance comme le hasard
  - 1 : concordance parfaite

2.A Si plus de 2 juges :

- Kappa de Light = extension du Kappa de Cohen pour plus de 2 juges

3 Mesure continue (ex : pression artérielle)

- Mesure continue : reproductibilité de la mesure entre 2 ou plusieurs évaluateurs

3.A Coefficient de corrélation intraclasse (ICC)

- ICC : mesure de la fiabilité entre plusieurs évaluateurs
- Méthode : analyse de variance à effet aléatoire

$$ICC = \frac{Variance\, entre\, Patients}{Variance\, totale}$$

4 Alpha de Cronbach

- = reproductibilité interne d'un questionnaire

Problème pour évaluer la reproductibilité d'un questionnaire :

- Soit le patient a changé
- Soit il répète la même chose
- Solution : évaluer la cohérence interne des items du questionnaire

4.A Principe du coefficient

- Calcul de la corrélation entre les items du questionnaire
- Pour calculer l'alpha de Cronbach, on fait la moyenne des corrélations entre **toutes les paires possibles** de sous-échelles.

(α de Cronbach = moyenne des split-half correlations) = corrélation de l'échelle avec elle-même.

4.B Interprétation

- α < 0,6 : cohérence interne faible items peu liés
- 0,6 ≤ α < 0,9 : c'est bien !
- α ≥ 0,9 : cohérence interne très élevée items redondants (c'est presque trop)

PLUS LE NOMBRE D'ITEMS EST ÉLEVÉ, PLUS L'ALPHA A TENDANCE À ÊTRE ÉLEVÉ

31Mesures Subjectives

1 Définitions

Objectif : attrait a un objet pensé (poids, TA, taille…)

Subjectif : attrait a un objet pensant (douleur, fatigue, qualité de vie…)

Par exemple :

- Thermomètre à mercure mesure objectivement la dilatation d'un liquide en fonction de la température
- Mais au final on sait pas vraiment ce qu'est la dilatation d'un liquide en fonction de la température

2 Importance de la subjectivité en médecine

- Mesures de qualité de vie : base de la médecine !
- Si nouveau traitement améliore la survie mais dégrade la qualité de vie : c'est bof !

3 Comment mesurer la subjectivité ?

1. Demander au patient = Patient Reported Outcomes (PRO)
2. Faire la mesure soi même (par hétéro-évaluation) pour éviter la **désirabilité sociale** (tendance à répondre ce que l'on pense que l'autre veut entendre)

> ❗ Important
>
> **Reproductivilité d'une mesure subjective**
>
> 1. Item bien écrit : mots simples
> 2. Concepts clairs et facilement accessibles par introspection
> 3. Pas de biais de désirabilité sociale (tendance à répondre ce que l'on pense que l'autre veut entendre)

Exemple pas reproductible : quel est votre niveau de qualité de vie ?

Pour mesure non reproductible : "*éclater*" en plein de petits items et en faire la somme

4 Valider un instrument

Valider = dit vrai (mais en vrai c'est impossible à prouver)

- Que vaut la mesure ? = reproductibilité (fiabilité)
- Que mesure l'instrument ? = validité (contenu, construit, critère)

4.A Que vaut la mesure ? = reproductibilité (fiabilité)

1. Reproductibilité = accord inter-juge
    - **Hétéro-évaluation** : accord inter-juge
        - Quantitatif : Coefficient de corrélation intraclasse (CCI)
        - Sinon : Kappa de Cohen
        - Si >= 3 juges : Kappa Light
    - **Auto-évaluation** : test-retest
        - Problème : les patients peuvent changer entre les deux tests
        - Coefficient alpha de Cronbach (consistance interne)
2. Que vaut la mesure = **structure dimensionnelle**
    - Une dimension = un construit = un concept
    - Si 20 items mais 2 dimensions 2 scores
    - Par ex QI : verbal + performance
    - **Pour explorer les dimensions :**
        - Analyse en composantes principales (ACP)
        - Puis analyse des valeurs propres (scree plot)
            * On tolère s'il y en a une qui est vraiment au dessus des autres
        - Pour explorer les autres dimensions : **analyse factorielle**
3. Que mesure l'instrument ? = validité
    1. Validité de contenu
        - Validité externe = comparaison au "gold standard" et corrélation entre les deux
        - Si pas de gold standard : validité concourante (= corrélation avec d'autres mesures) et divergente (absence de corrélation avec des mesures non liées)

Tip

Livre : Mesurer la Subjectivité en Santé
1. Subjectivité en médecine

- La médecine s'adresse à la fois au **corps malade** (sujet vivant) et au **vécu** de la maladie (sujet pensant).
- Le fait de se voir / se sentir / se savoir malade influence la **détresse**, la **prise en charge** et parfois l'**évolution** de la maladie.
- Est **subjectif** : ce qui concerne le sujet en tant qu'**être conscient** (qualité de vie, dépression, satisfaction).

- Est **objectif** : ce qui concerne le corps comme **objet mesuré** (TA, NFS, diamètres, etc.).
- « Subjectif » ne veut pas dire « arbitraire » : c'est simplement la dimension **consciente** de l'expérience.

---

2. Qu'est-ce qu'une mesure ?

- Définition courante : **attribuer un nombre** à un attribut d'un objet ou d'un sujet (ex. 10 kg pour la masse d'un sac).
- Cette définition interroge : quelle est la **réalité** de l'attribut (masse, tristesse) au-delà du mot et du chiffre ?
- Différentes conceptions de la mesure ont des **implications philosophiques** et méthodologiques importantes, surtout pour la subjectivité.

---

3. Trois approches du concept de mesure

- **Approche classique (physiciens)**
  - La mesure est un **nombre manipulable** dans des équations pour expliquer et prédire (gravitation, éclipses, marées…).
  - Elle fonctionne bien pour des grandeurs physiques, mais se heurte à des notions comme la **tristesse** ou la **souffrance**.
- **Approche opérationnelle**
  - La mesure est un **processus** donnant un nombre, sans chercher à comprendre la nature de ce qui est mesuré.
  - Utile en physique (temps, espace) mais problématique pour la subjectivité, car on perd de vue le **sens vécu** derrière les scores.
- **Approche représentationnelle**
  - La mesure est la **représentation numérique de faits empiriques**.
  - Le chiffre est un **outil de représentation** pour rendre intelligibles des phénomènes complexes.
  - C'est l'approche la plus adaptée aux **mesures subjectives** : le but n'est pas le score en soi, mais ce qu'il permet de **mieux comprendre**.

---

4. Mesures subjectives vs objectives : critiques et réponses

- **Critique 1 : "La subjectivité, ce ne sont que des mots flous"**

  - Exemple de la **tristesse** : difficile à définir, mais immédiatement comprise par tout le monde.
  - Comme pour le **temps**, on sait le mesurer sans savoir vraiment ce que c'est.
- **Critique 2 : "On ne peut pas vérifier ce que quelqu'un ressent vraiment"**
  - En pratique, dans un **climat de confiance**, le sujet rapporte globalement ce qu'il ressent.
  - L'**hétéro-évaluation** (évaluation par un clinicien empathique) permet parfois une description plus fiable, surtout en psychiatrie.
  - Elle permet aussi d'estimer la **variabilité inter-évaluateurs** et donc l'**erreur de mesure**, chose difficile avec l'auto-évaluation.
- **Critique 3 : "Les mesures subjectives ne seront jamais assez précises"**
  - De nombreuses mesures objectivées (ex. TA au brassard) sont aussi **peu précises**.
  - La précision des mesures subjectives peut **progresser**, et leurs paradigmes de mesure évoluent en parallèle de ceux des sciences « dures ».

---

5. Trois paradigmes de mesure : impression, étalon, théorie

- **Mesure "impression"**
  - On se prononce à partir d'une **impression globale** : « très lourd », « peu lourd ».
  - Parallèle : échelles globales comme la **CGI** (Clinical Global Impression).
- **Mesure "étalon"**
  - On compare à des **références** : « aussi lourd que deux dictionnaires ».
  - Parallèle subjectif : échelles type **Stanford Sleepiness Scale**, où chaque phrase décrit un niveau-repère de somnolence.
  - Limite : trop de niveaux difficulté à se situer clairement.
- **Mesure "théorie"**
  - La mesure est inséparable d'une **théorie générale** (ex. nouvelle définition du mètre via la vitesse de la lumière).
  - En subjectivité, les **instruments multi-items** relèvent de ce paradigme : ils reposent sur une théorie implicite de ce qu'ils mesurent.

6. Instruments multi-items et cadre théorique

- **Théorie définitoire**
  - Exemple : échelle de dépression de Hamilton (21 items : humeur, suicide, insomnie, anxiété, obsessions, etc.).

  - Les choix d'items (inclusions, omissions) traduisent une certaine **vision de la dépression** : c'est une théorie implicite de la dépression.

- **Théorie nomologique**
  - Chaque concept (dépression, anxiété, obsessions…) s'inscrit dans un **réseau de relations** avec d'autres concepts.
  - L'ensemble de ces relations constitue une **théorie nomologique** de la discipline (psychopathologie, qualité de vie, etc.).

---

7. Pourquoi les mesures subjectives paraissent plus fragiles

- On peut obtenir des **mesures subjectives très précises** (ex. dépression sur entretien filmé) et des **mesures objectives très grossières** (TA au brassard).
- Pourtant, la TA « paraît » plus solide, car elle s'appuie sur un **modèle physiologique clair** (cœur, artères, pression, bruits d'auscultation).
- Pour la dépression, le cadre théorique (définitions, liens avec autres symptômes) est **moins stabilisé**, d'où un sentiment de fragilité.
- Ce n'est donc pas seulement un problème de fiabilité, mais un problème de **théorie incomplète** de la subjectivité.

---

8. Interprétation des scores et nécessité de validation

- **Échelles globales** (type CGI)
  - Question simple : « niveau de gravité de l'état dépressif ».
  - Un score (ex. 3 = « légèrement déprimé ») garde une **signification intuitive** pour le clinicien.
- **Échelles multi-items** (Hamilton, HUI2 pour la qualité de vie)
  - Agrègent des contenus très variés (humeur, sommeil, douleur, mobilité, mémoire, fertilité…).
  - Un même score global peut correspondre à **de nombreux profils différents** → perte d'intuition.
  - L'expression « dépression » ou « qualité de vie » reste parlante, mais le score composite l'est beaucoup moins.
- **Conséquence**
  - Pour ces instruments, il faut **étayer le sens** des scores par des études :
  - montrer ce que le score reflète réellement,
    * préciser comment il se relie à d'autres variables cliniques.
    * Cette étape constitue la **validation** de l'instrument, indispensable avant d'interpréter les résultats ou de les utiliser en pratique.

32Mesures Subjectives bonus

1 Module 1 : Les Fondamentaux et la Définition de la Mesure

Ce premier module pose le cadre théorique. Il est crucial de comprendre *pourquoi* et *comment* on mesure l'impalpable.

1.A 1. Le contexte médical

- La médecine a évolué d'une approche purement physiopathologique (tension, glycémie) vers une prise en compte du sujet "pensant" .
- On mesure désormais des PROs (Patient Reported Outcomes) comme la qualité de vie, la douleur ou la satisfaction.
- Le postulat des PROs est que le patient est le mieux placé pour juger de son état, bien que des mécanismes de défense (comme le déni) puissent parfois biaiser ce jugement.

1.B 2. Qu'est-ce que la subjectivité ?

-Définition : Est subjectif ce qui concerne le sujet en tant qu'être conscient (le "sujet pensant"), par opposition à l'objectif qui concerne l'objet ou le corps (le "sujet vivant").

- **Idée reçue :** "Subjectif" ne veut pas dire "arbitraire" ou "imprécis". La tristesse est un sentiment universel et précis pour celui qui le vit.

1.C 3. Les trois approches de la mesure

Pour mesurer la subjectivité, le texte distingue trois visions philosophiques :

1. **L'approche classique (Physique) :** La mesure sert à découvrir des lois naturelles (ex: gravitation). Difficile à appliquer à la psychologie car il n'y a pas d'équations universelles de la tristesse.
2. **L'approche opérationnelle :** La mesure *est* l'opération réalisée (le score *est* la dépression). *Critique :* Elle vide la mesure de son sens. Si on ne fait que lire un chiffre sans l'interpréter, on perd la réalité du vécu du patient .
3. **L'approche représentationnelle (La solution retenue) :** La mesure est une "représentation numérique de faits empiriques". Les chiffres sont une commodité pour représenter une réalité complexe (la douleur, l'anxiété) afin de mieux la comprendre.

1.D 4. Subjectif vs Objectif : une fausse opposition

- On pense souvent que les mesures physiques (mètre, kilogramme) sont "vraies" et les mesures subjectives "floues".

- En réalité, même les mesures physiques reposent sur des **théories** (ex: la définition moderne du mètre repose sur la théorie de la relativité et la vitesse de la lumière).
- De même, pour mesurer la dépression, il faut une théorie de la dépression (**théorie définitoire**) et une théorie de ses liens avec le reste du monde (**théorie nomologique**).
- **Conclusion :** La mesure subjective n'est pas moins scientifique, elle a juste des théories sous-jacentes souvent moins abouties que la physique.

1.E 5. Modes de recueil

- **Auto-évaluation :** Le sujet se note lui-même. C'est la norme pour la qualité de vie.
- **Hétéro-évaluation :** Un tiers (médecin, proche) évalue le sujet.
  - `Avantage : Utile si le patient ne peut pas s'auto-évaluer (psychiatrie lourde, p`

1.F 3. Les Dispositifs de Réponse (Les échelles)

Comment le patient va-t-il répondre ?

Échelle de Likert : "Pas du tout", "Un peu", "Moyen", "Beaucoup". C'est le standard, souvent en 4, 5 ou 7 points.

Échelle Visuelle Analogique (EVA) : Une ligne continue de 10 cm où le patient place une croix.

Très utilisé pour la douleur, mais parfois difficile à conceptualiser pour certains patients .

1.G 4. La Sélection des Items

- On part souvent d'une liste très longue (trop longue).
- On utilise des méthodes de consensus comme la méthode Delphi (experts interrogés à distance en plusieurs tours) pour ne garder que les items les plus pertinents et consensuels .

2 Module 2 : La Construction de l'Instrument

Une fois que l'on sait *ce que* l'on veut mesurer (Module 1), il faut fabriquer l'outil pour le faire. Ce n'est pas juste une liste de questions ; c'est un processus rigoureux décrit dans le chapitre 3.

Voici les étapes clés pour construire un instrument de mesure subjective solide :

2.A 1. Le Cahier des Charges (La fondation)

Avant de rédiger, il faut définir le périmètre. C'est l'étape la plus souvent négligée. Il faut préciser :

- **Le concept :** Que mesure-t-on exactement ? (ex: l'anxiété, mais laquelle ? Anxiété-trait ou anxiété-état ?) .

- **La population :** Qui va répondre ? (ex: des enfants, des personnes âgées ?) Cela change le vocabulaire et la longueur.
- **Le mode d'administration :** Auto-évaluation (le patient remplit seul) ou hétéro-évaluation (un soignant remplit).

2.B 2. La Chasse aux Items (D'où viennent les questions ?)

On ne doit pas inventer les questions "dans son coin".

On utilise des **sources d'items**:

- **La littérature :** Ce qui existe déjà.
- **Les experts :** Cliniciens, psychologues.
- **Les patients :** Via des entretiens ou des **focus groups** (discussions de groupe). C'est crucial pour utiliser *leur* vocabulaire et comprendre ce qui compte vraiment pour eux.

2.C 3. La Sélection : La méthode du Consensus

On se retrouve souvent avec trop de questions.

Pour garder les meilleures, on utilise des méthodes de consensus comme la **méthode Delphi**:

1. On envoie les items à des experts.
2. Ils notent la pertinence de chaque item anonymement.
3. On leur renvoie les résultats globaux.
4. Ils re-notent en connaissance de cause jusqu'à ce qu'un consensus se dégage .

2.D 4. La Rédaction et les Réponses

-Principe du "coût minimum" : Répondre doit demander un effort minimal au patient.
Il faut éviter les phrases alambiquées, les doubles négations ("Je ne me sens pas malheureux"), et le jargon médical.

- **Le dispositif de réponse :** Comment le patient note-t-il son ressenti ?
  - **L'échelle de Likert** : Des cases à cocher (ex: Pas du tout, Un peu, Beaucoup, À la folie).
  - **L'Échelle Visuelle Analogique (EVA) :** Une ligne (souvent de 10cm) où le patient place une croix. C'est sensible mais parfois difficile à comprendre pour certains patients.

2.D.1 Module 3 : La Validation (Partie 1) — Que mesure l'instrument ?

La validation cherche à répondre à une question centrale : **Sommes-nous bien en train de mesurer ce que nous croyons mesurer ?**

Voici les trois piliers pour y répondre, selon le texte :

2.D.1.1 1. La Validité de Contenu (“Est-ce qu’on a tout couvert ?”)

Il faut vérifier que les items explorent toutes les facettes du concept, sans hors-sujet.

Exemple : Si on crée une échelle d’**anxiété**, il ne suffit pas de demander “Avez-vous peur ?”.

Il faut aussi explorer les symptômes physiques (mains moites), comportementaux (évitement), etc., pour couvrir tout le “contenu” du concept.

2.D.1.2 2. La Validité de Structure (“Est-ce cohérent ?”)

On vérifie ici l’architecture interne de l’instrument.

L’unidimensionnalité :

- Est-ce que tous mes items mesurent la même chose (une seule dimension) ?
- Ou est-ce que mon score total est un mélange hétérogène de plusieurs concepts (comme additionner des choux et des carottes) ?.

On utilise souvent des analyses statistiques (comme l’analyse en composante principale + analyse factorielle) pour voir si les items se regroupent bien ensemble.

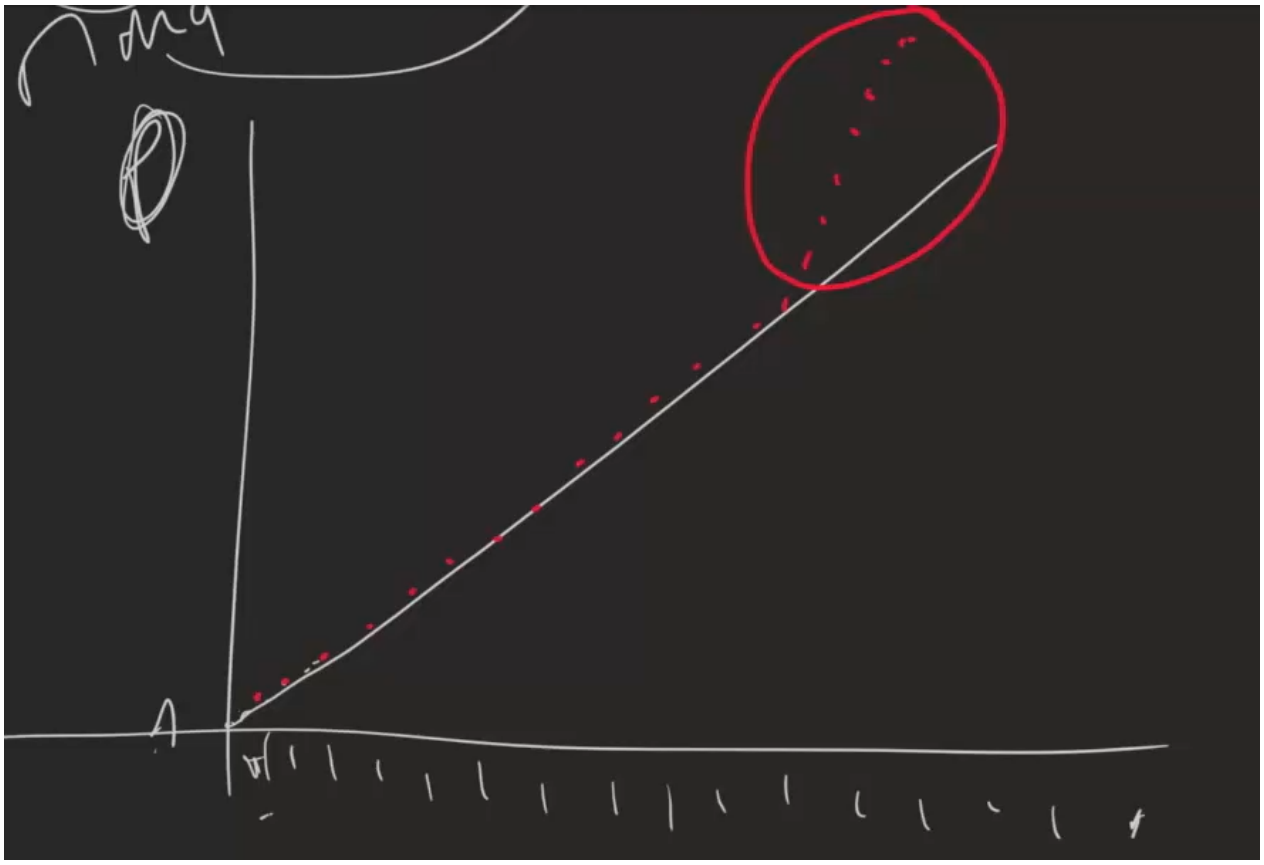
2.D.1.3 3. La Validité de Construit (“Est-ce logique ?”)

C’est l’étape la plus “scientifique”.

On formule des hypothèses sur la façon dont notre mesure devrait se comporter face à d’autres variables (c’est la **théorie nomologique**).

On distingue notamment :

- **Validité Convergente :** Notre instrument doit être corrélé (lié) avec d’autres instruments mesurant la même chose ou des choses proches.
- **Validité Divergente :** Notre instrument doit être *peu* corrélé avec des instruments mesurant des concepts différents.



2.D.2 Module 4 : La Validation (Partie 2) — Que vaut la mesure ?

Une fois qu'on sait *ce que* l'on mesure (validité), il faut s'assurer de la qualité technique de l'outil.

2.D.2.1 1. La Précision (ou Fidélité / Reliability) L'objectif est de réduire l'erreur de mesure au minimum.

On l'évalue sous trois angles principaux :

- **La consistance interne :** Les items d'un même questionnaire sont-ils cohérents entre eux ?
  - *L'outil star :* Le **coefficient Alpha de Cronbach**. Plus il est proche de 1, plus les items sont liés
- **La fidélité inter-juges :** Deux médecins évaluant le même patient obtiennent-ils le même résultat ?
  - Pour un diagnostic (Oui/Non) : On utilise le **coefficient Kappa** (qui corrige le hasard)
  - Pour un score chiffré (ex: 15/20) : On utilise le **Coefficient de Corrélation Intraclasse (CCI)**
- **La fidélité test-retest :** Si l'état du patient ne bouge pas, le score reste-t-il stable dans le temps ?

2.D.2.2 2. La Sensibilité au Changement C'est une propriété cruciale en médecine : l'instrument est-il capable de détecter une amélioration (ou une aggravation) de l'état de santé ?

C'est différent de la précision. Une balance peut être très précise (donner toujours le même poids) mais pas assez sensible pour détecter une variation de quelques grammes.

2.D.2.3 3. Les Modèles de Mesure (La "boîte noire" mathématique) Comment transforme-t-on les croix cochées par le patient en un score final ?

Modèles psychométriques :

- *Théorie Classique :* On fait la somme des scores. Simple et robuste.
- *Réponse à l'Item (IRT) :* Plus complexe, probabiliste. Elle permet de gérer les données manquantes et de comparer des patients ayant répondu à des questions différentes

Modèles économiques : Utilisent la notion d'**Utilité** (préférence pour un état de santé) pour calculer des **QALYs** (années de vie ajustées sur la qualité), essentiels pour les décideurs de santé publique

---

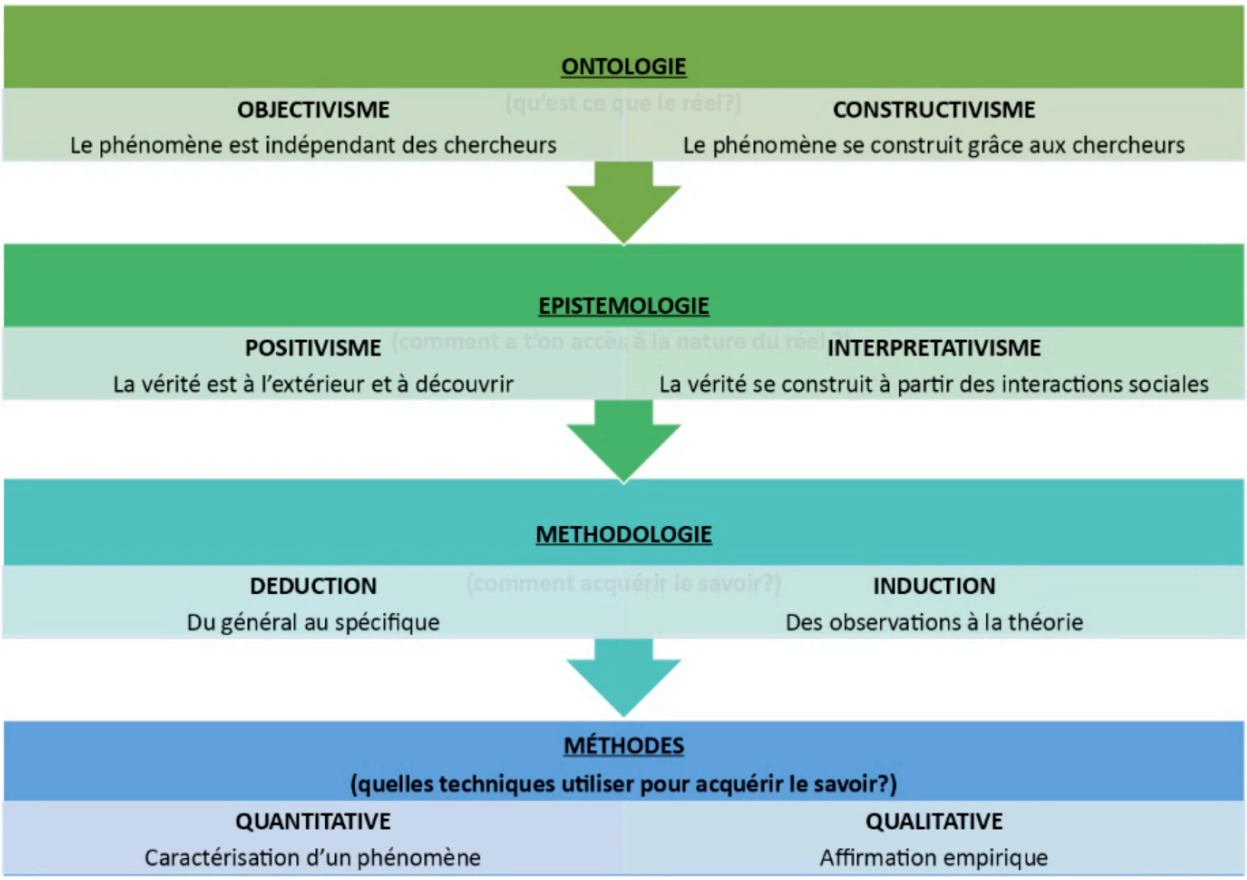
33Méthodes Qualitatives

1 C’est quoi la recherche qualitative ?

Canard ou lapin ? = les deux !

Représentation collective : un grand nombre de personnnes partagent la même vision (ça ressemble à un animal)

	Qualitative	Quantitative
Nature du sujet	Adaptée aux phénomènes sociaux / psychosociaux *Phénomènes : “expérience” témoignée par un “discours”*	Questions biomédicales
Objectif	Explore **l’existence** et la **signification** de ces phénomènes	**Quantifie** et **compare** des variables mesurables
Finalité	**Compréhension** d’un phénomène appréhendé dans son **contexte** (environnement, interactions relationnelles…)	**Explication** par des relations causales = corrélation statistiquement significative
Rapport aux hypothèses	**Construire** des hypothèses	Tester des hypothèse
Démarche logique	Plutôt **inductive (partir du terrain pour construire une théorie)**	Quasiment toujours **déductive**



Ontologie : qu'est ce qui constitue le réel ?

- Objectivisme : la réalité existe et il faut qu'on trouve les bons outils pour la mesurer
- **Constructivisme** : le phénomène se **construit** grâce aux chercheurs : le chercheur crée l'objet de recherche

Épistémologie :

- Positivisme : vérité à l'extérieur, à découvrir
- **Interprétativisme :** l'**interaction sociale** donne un sens ± différent à la réalité (visio vs présentiel)

Méthodologie :

- Déduction : du général au spécifique
- **Induction / interprétativisme :** observations +++ et la théorie est construite à partir de ça

Méthodes qualitatives : affirmation empirique, c'est à dire basées sur des données collectées sur le terrain (entretiens, observations, documents…)

Sociologie = décrire et interpréter.

2 Historique de la recherche qualitative

- Sciences sociales (1920-1930)

    - Décolonisation
    - Anthropologie
- Commerce/Marketing (1950-1960)
- Recherche en soins infirmiers (1980-1990)
- MeSH term (2003) : limite des sciences fondamentales

Méthodes mixtes : combiner méthodes qualitatives et quantitatives

3 6 méthodes prédominantes

1. Théorisation ancrée (Grounded Theory)
2. Phénoménologie interprétative (IPA : Interpretative Phenomenological Analysis)
3. Ethnographie
4. Analyse de discours (Discourse Analysis)
5. Analyse de contenu (Content Analysis)
6. Analyse de conversation ou conversationnelle (Conversation Analysis)

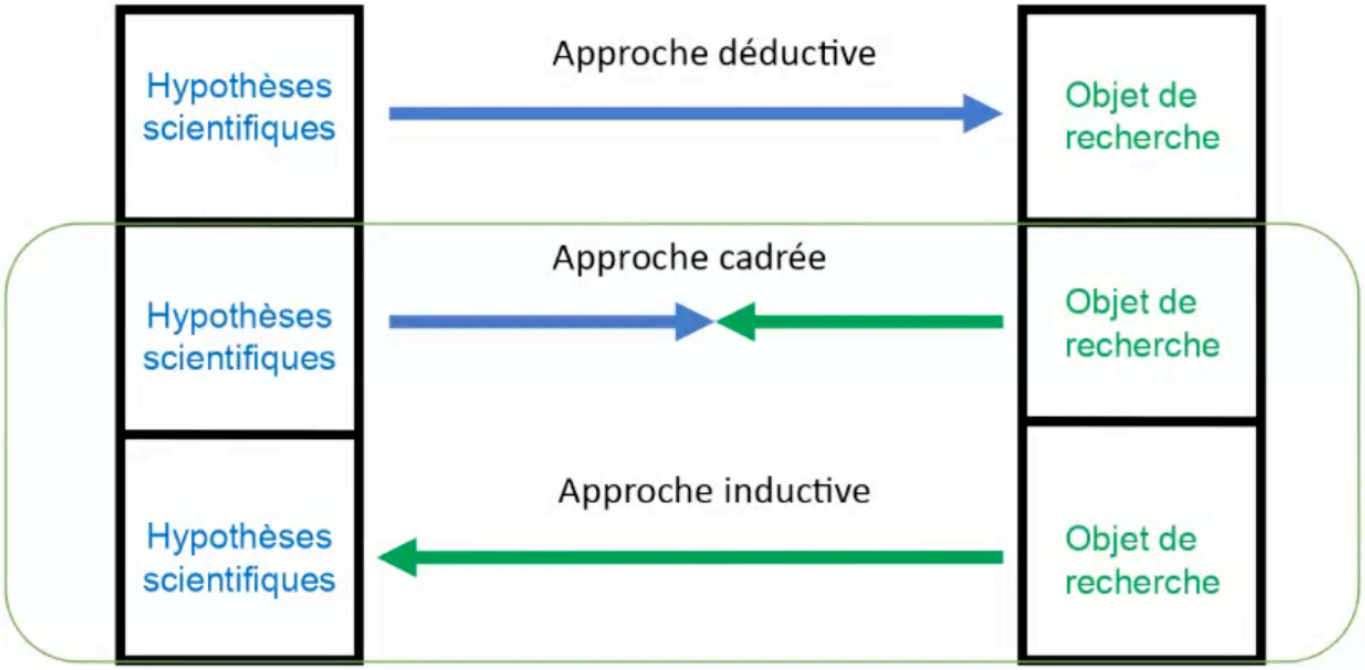
Base de données : Web of Science, Cinhal

> ❗ Important
>
> **80% sont**
>
> - Théorisation ancrée
> - Phénoménologie interprétative

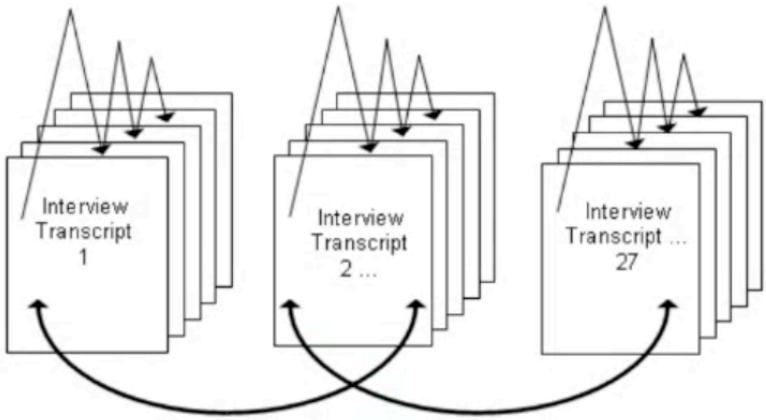
Points communs des méthodes : Le chercheur est l'outil !

- Contact prolongé avec le terrain - Imprégnation
- Compréhension holiste = compréhension globale du phénomène
- Retranscription fidèle = enregistrement puis ré-écriture mot à mot
- Techniques de recueil de données multiples : entretiens, observations, documents…
- Processus inductif (partir du terrain et aller à la théorie) et itératif
- Dé-contextualisation puis re-contextualisation (sortir le discours de son contexte puis le remettre dans son contexte)
- Description puis interprétation
- Modèle final issu de catégories et concepts émergents

4 Induction / déduction



5 Itération



= Analyse → terrain → recueil → analyse → terrain → recueil…
Donc le terrain influence **continuellement** l'analyse.

6 Présupposés du chercheur

Écrire et répondre à ces 7 questions sur le journal de bord :

1. Quelle est la **question initiale** que je me pose
2. Comment en suis-je venu à me poser cette question ?
3. Pourquoi cette question est-elle pertinente ?
4. Si j'étais moi-même interrogé, quelles seraient mes réponses, mon vécu ?
5. Quelles réponses j'attends des informateurs ?
6. Quelles réponses me surprendraient ?
7. Quelle est finalement ma **question de recherche** ?

7 Atelier *Posture*

Thématique de recheche : **La PREP et la population subsaharienne**

Question de recherche : **Quelle place pour la PREP dans une population africaine subsaharienne ?**

1. Expliciter ses présupposés
2. Identifier sa posture “spontanée”
3. Adopter la juste posture de chercheur (qui suis-je : médecin, civil ; quelle est ma posture : militant, sensibilisé..)

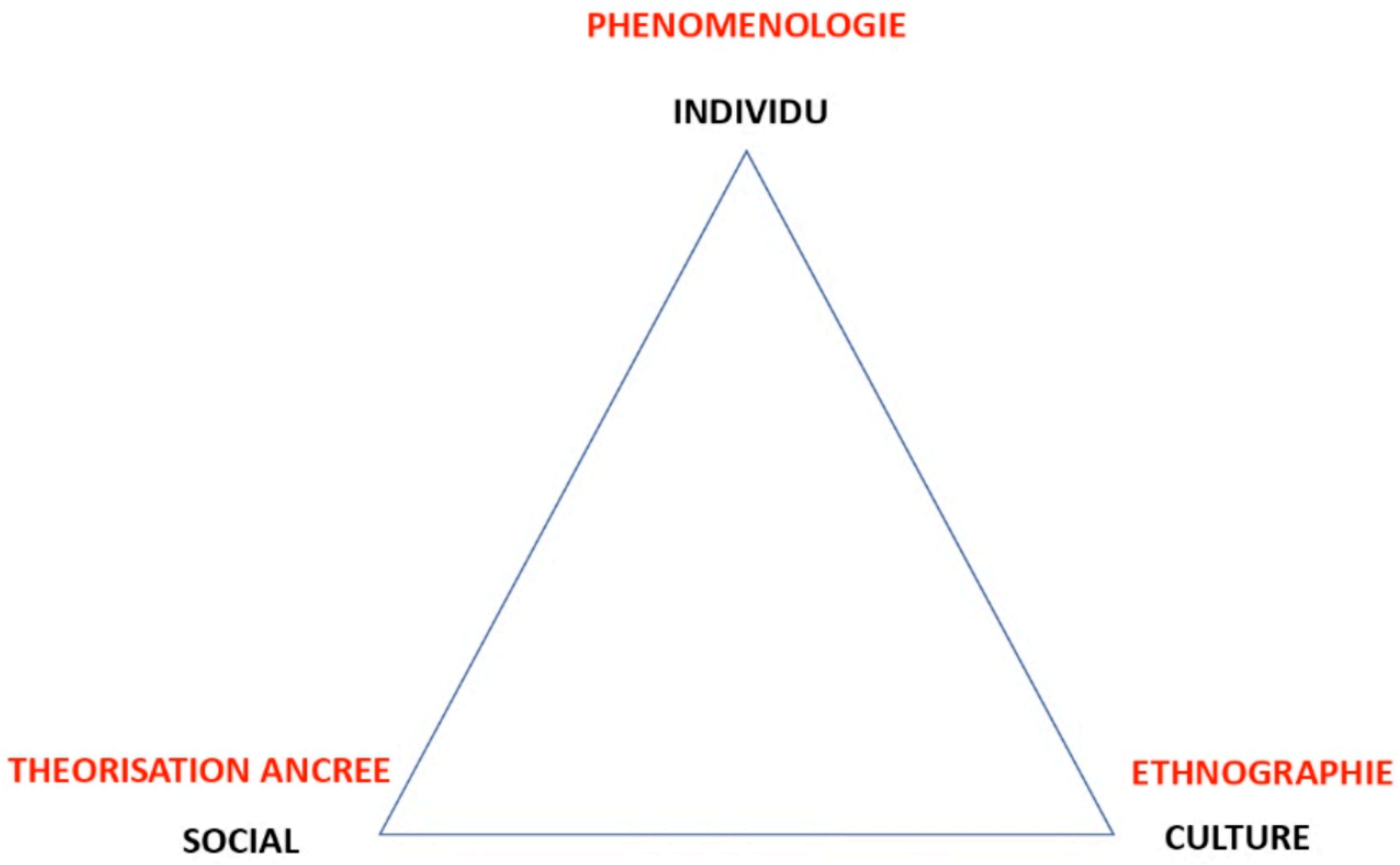
8 Journal de bord

Pourquoi : fondement du travail, cheminement du chercheur, résultats intermédiaires, maillage terrain-théorie

Quand : dès le début, tout au long du projet

Comment : n’importe!

9 Quel prisme



- Individu : phénoménologie
  - phénoménologie interprétative : définition = étude des vécus subjectifs
- Social : théorisation ancrée

  - théorisation ancrée : définition = construction de théorie à partir de données empiriques
- Culture : ethnographie
  - ethnographie : définition = étude des modes de vie et des comportements d'un groupe social

10 Théorisation ancrée (social)

Née dans les années 1970s

1. Interactionnisme symbolique
   - Comment un processus social est-il construit à partir des représentations qu'en ont différents groupes sociaux ?
2. Comparaison constante
   - Stratégies explicites d'analyse et de collecte des données
   - Courants constructivistes (c'est à dire que le chercheur construit la réalité) et objectivistes (la réalité existe en dehors du chercheur)
3. Modélisation
   - Théorie substantive (un domaine précis) = délinquance juvénile
   - Théorie formelle (plus générale) = violence

Les 3 niveaux d'analyse (comparaison constante)

Niveau 1	Niveau 2	Niveau 3
analyse ouverte	Analyse axiale	Analyse intégrative

3 niveaux d'analyse :

- Analyse ouverte = décrire
- Analyse axiale = relier les données de l'analyse ouverte
- Analyse intégrative = donner du sens en construisant une théorie

Exemples d'études par GTM : VÉCU - PERCEPTION

- Quelles sont les perceptions des MG de leur rôle dans la prise en charge des patients atteints de TCA?
- Quelles sont les représentations des acteurs de santé de la recherche en médecine générale?
- Quels sont les processus de décision et d'action du couple MG/patient dans la réalisation de l'Hemoccult?
- Comment la tarification à l'activité peut-elle avoir un impact sur l'identité professionnelle des médecins d'un service d'urgence ?

11 Méthode par phénoménologie interprétative (individu) = IPA

- Née dans les années 1990 - Psychologie clinique

= EXPÉRIENCE D'UN PHÉNOMÈNE VÉCU ET SON INTERPRÉTATION PAR L'INDIVIDU

1. Phénoménologie
    - **Comment les individus font-ils l'expérience et donnent-ils du sens à un phénomène ?**
2. Herméneutique (théorie de l'interprétation): est/interprété
    - Notion de double herméneutique = chercheur / patient (interprétation, empathie)

- Idiographie (détail et profondeur, phénomène et contexte particuliers):
  - Pénétrer un univers singulier

«Le plus profond vous allez dans votre analyse, le plus universel elle devient »> : approfondir le particulier nous amène au plus près de l'universel

Exemples :

- Comment le sens de l'identité d'une femme change pendant sa transition à la fonction maternelle?
  - *Grounded Theory serait : quelles sont les perceptions des mères sur la maternité ?*
- Comment les personnes atteintes de troubles des conduits alimentaires vivent leur prise en charge par le MG?
  - *Grounded Theory serait : quelles sont les perceptions des MG sur la prise en charge des TCA ?*
- Comment les familles qui participent à une thérapie multifamiliale des troubles alimentaires donnent sens à leur vécu du changement?

  - *Grounded Theory serait : quelles sont les perceptions des familles sur la thérapie multifamiliale des TCA ?*

- Comment les patientes vivent l'abord de leur poids en consultation de MG?

	Approche par théorisation ancrée (ATA)	Approche par analyse interprétative phénoménologique (AIP)
Objectif	Développer une théorie à propos de processus sociaux d'une population identifiée autour d'un phénomène	Explorer la manière dont une expérience de vie a été ressentie et comprise par la personne qui l'a vécue
Question de recherche	Comment un processus social est-il construit à partir des représentations qu'en ont différents groupes sociaux ?	Comment les individus font-ils l'expérience et donnent-ils du sens à un phénomène ?

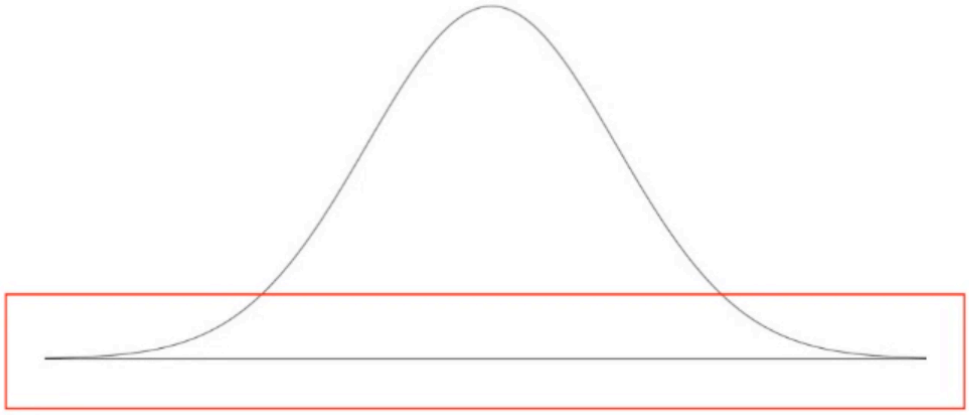
12 Échantillonnage

Objectif : diversité des points de vue !

≠ de la recherche quantitative : où l'objectif est la représentativité statistique

Objectif de l'échantillonnage

= maximiser l'expression de la diversité



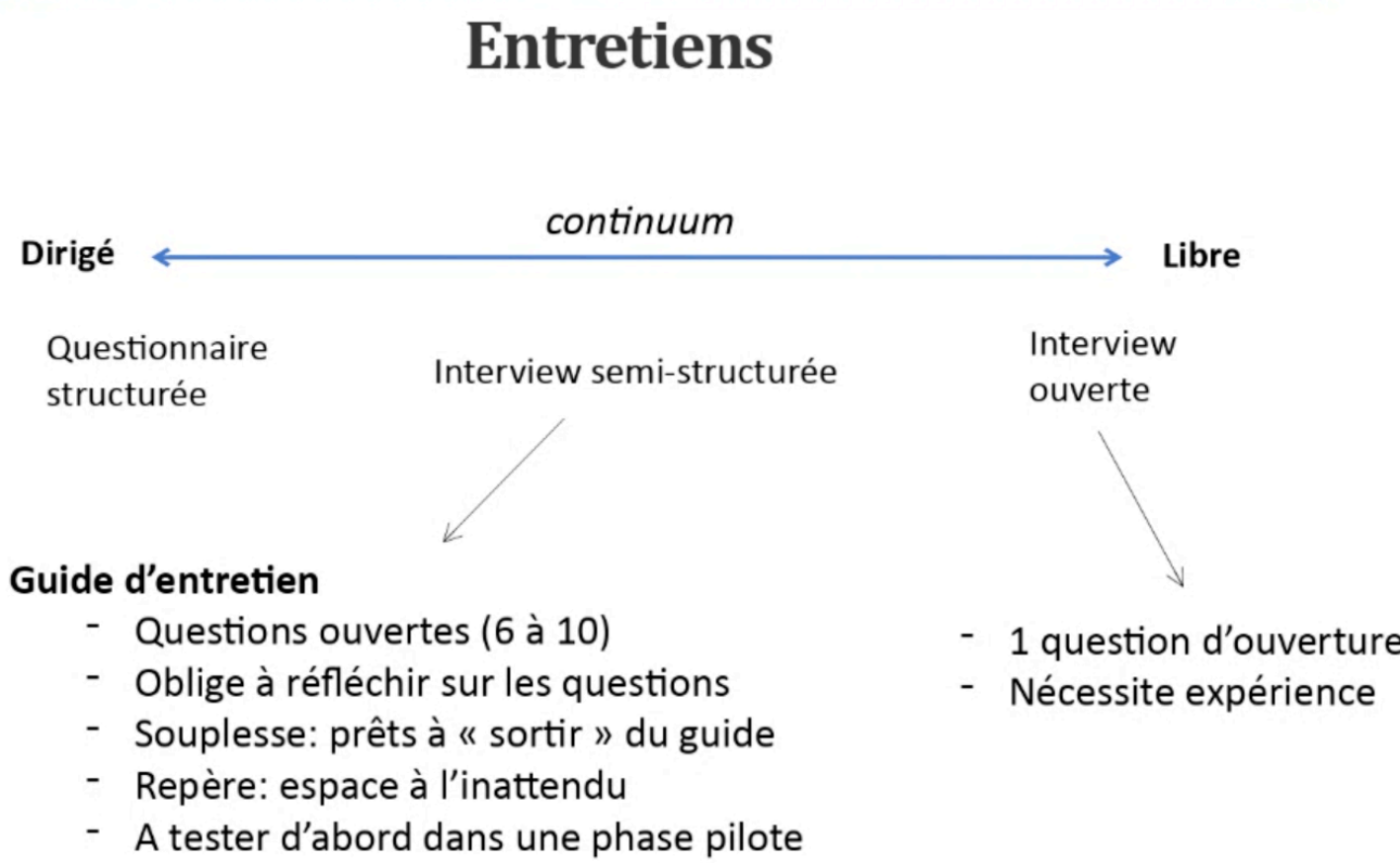
Donc :

- Petit nombre de sujets

- Échantillonage raisonné à variation maximal
  - Homogène : groupe qui partage la même expérience
    - Ex : femmes ayant vécu une IVG
  - Théorique : selon la théorie émérgente au fur et à mesure de l'analyse
    - Séléction du participant à l'entretien suivant dépend de l'analyse des données
    - La "théorie" / le schéma final doit suivre l'analyse des données
    - **Interroger les participants concernés par la théorie émergente**

34Entretiens

Phénoménologie : entretiens individuels (par exemple ressenti sur troubles de l'érection)



Théorisation ancrée : entretien semi-structurée : entretiens individuels ou compréhensifs

Entretiens collectifs : plutôt en théorisation ancrée (parce que phénoménologie = ressenti +++ donc pas entretien collectif)

- Dimension groupale
- Homogénéité / hétérogénéité

1 Critères de scientificité / validité

- Qualité des entretiens
  - Écoute active
  - Reformulation
- Suffisance et saturation
  - Suffisance : nombre d'entretiens suffisant pour répondre à la question de recherche

  - Saturation : plus d'informations nouvelles n'émergent des entretiens (surtout en théorisation ancrée)
- Triangulation
  - Multiples sources de données (entretiens, observations, documents…)
  - Multiples chercheurs
- Crédibilité, originalité, résonance, utilité

Élément	ATA – Explication détaillée	AIP – Explication détaillée
Objectif	**Ce que c'est :** L'ATA cherche à produire une théorie qui émerge directement des données recueillies sur le terrain.	

Concrètement : Le travail vise à comprendre un processus social (ex. trajectoires de patients, organisation d'un service…) et à générer un modèle explicatif.
Pourquoi ATA : L'ATA est le cadre méthodologique créé pour faire émerger des théories enracinées dans les données. | **Ce que c'est :** L'AIP vise à comprendre en profondeur comment une personne vit, ressent et interprète une expérience.
Concrètement : L'objectif n'est pas d'établir une théorie généralisable mais de saisir la singularité de l'expérience vécue.
Pourquoi AIP : L'AIP est une méthode phénoménologique et herméneutique centrée sur l'expérience subjective d'un individu. | | **Question de recherche** | **Ce que c'est :** Une question portant sur un processus social (comment un phénomène se déroule, se construit, se transforme dans un groupe).
Concrètement : L'intérêt porte sur la dynamique collective et les représentations sociales.
Pourquoi ATA : L'ATA explore les processus, les mécanismes et les conditions qui structurent les comportements sociaux. | **Ce que c'est :** Une question portant sur la manière dont une personne vit et donne sens à un événement.
Concrètement : La focale est sur l'expérience individuelle et sa signification.
Pourquoi AIP : L'AIP ne cherche jamais des processus collectifs : l'intérêt se porte sur l'intériorité et la compréhension du sens vécu. | | **Échantillonnage** | **Ce que c'est :** Un échantillonnage théorique évolutif.
Concrètement : Les participants sont choisis progressivement en fonction des besoins théoriques émergents (ex : sélectionner un profil spécifique pour tester ou enrichir une catégorie).
Pourquoi ATA : L'échantillonnage en ATA n'est jamais fixé à l'avance ; il accompagne le développement de la théorie. | **Ce que c'est :** Une sélection homogène de participants ayant vécu la même expérience.
Concrètement : L'objectif est de comparer des variations fines dans un même type d'expérience (ex : tous ont vécu un même diagnostic).
Pourquoi AIP : L'approche phénoménologique nécessite une homogénéité pour isoler la structure de l'expérience vécue. | | **Recueil de données** | **Ce que c'est :** Entretiens semi-dirigés, observations, parfois focus groups.
Concrètement : Les données doivent être riches et variées afin de permettre la construction progressive de catégories.
Pourquoi ATA : La multiplication des sources permet une saturation conceptuelle solide et un

croisement des perspectives. | **Ce que c'est :** Entretiens ouverts et approfondis.
Concrètement : Le participant raconte librement son expérience avec un minimum d'intervention du chercheur.
Pourquoi AIP : La priorité est la narration personnelle, subjective, intime. Les méthodes ouvertes favorisent cette profondeur phénoménologique. | | **Analyse** | **Ce que c'est :** La comparaison constante et trois niveaux d'analyse :
– ouverte : émergence des premières unités de sens
– axiale : articulation des catégories entre elles
– intégrative : construction du modèle théorique final
Concrètement : L'analyse est extrêmement systématique, progressive et recursive.
Pourquoi ATA : L'ATA repose strictement sur cette architecture analytique pour faire émerger une théorie. | **Ce que c'est :** Analyse focalisée sur chaque participant, puis comparaison secondaire entre participants :
– ouverte : premiers thèmes issus du discours
– axiale : structuration des thèmes
– intégrative : compréhension globale de l'expérience
Concrètement : L'analyse vise l'interprétation du vécu individuel avant la comparaison inter-individuelle.
Pourquoi AIP : L'accent est mis sur l'interprétation herméneutique du sens vécu, non sur la construction d'un modèle généralisable. | | **Modèle explicatif** | **Ce que c'est :** Production d'une **théorie substantielle** enracinée dans les données.
Concrètement : Un modèle processuel décrivant comment un phénomène social fonctionne.
Pourquoi ATA : L'ATA est spécifiquement conçue pour aboutir à une théorie nouvelle. | **Ce que c'est :** Production d'une **synthèse conceptuelle** des thèmes qui décrivent l'expérience.
Concrètement : Le résultat final est une compréhension raffinée de la manière dont un individu vit un phénomène.
Pourquoi AIP : L'objectif n'est pas de théoriser mais de clarifier les structures de l'expérience vécue. |

Voici une explication complète, point par point, pour chaque ligne du tableau, en explicitant : 1. Ce que signifie l'élément (objectif, échantillonnage, etc.) 2. Ce que cela veut dire concrètement 3. Pourquoi cela correspond à l'Approche par Théorisation Ancrée (ATA) 4. Pourquoi cela correspond à l'Analyse Interprétative Phénoménologique (AIP)

Le tout est présenté sous forme de tableau explicatif, sans tutoiement et sans spéculation.

Tableau explicatif complet (ATA vs AIP)

Élément	ATA – Explication détaillée	AIP – Explication détaillée
Objectif	**Ce que c'est :** L'ATA cherche à produire une théorie qui émerge directement des données recueillies sur le terrain.	

Concrètement : Le travail vise à comprendre un processus social (ex. trajectoires de patients, organisation d'un service…) et à générer un modèle explicatif.
Pourquoi ATA : L'ATA est le cadre méthodologique créé pour faire émerger des théories enracinées dans les données. | **Ce que c'est :** L'AIP vise à comprendre en profondeur comment une personne vit, ressent et interprète une expérience.

Concrètement : L'objectif n'est pas d'établir une théorie généralisable mais de saisir la singularité de l'expérience vécue.
Pourquoi AIP : L'AIP est une méthode phénoménologique et herméneutique centrée sur l'expérience subjective d'un individu. | | **Question de recherche** | **Ce que c'est :** Une question portant sur un processus social (comment un phénomène se déroule, se construit, se transforme dans un groupe).
Concrètement : L'intérêt porte sur la dynamique collective et les représentations sociales.
Pourquoi ATA : L'ATA explore les processus, les mécanismes et les conditions qui structurent les comportements sociaux. | **Ce que c'est :** Une question portant sur la manière dont une personne vit et donne sens à un événement.
Concrètement : La focale est sur l'expérience individuelle et sa signification.
Pourquoi AIP : L'AIP ne cherche jamais des processus collectifs : l'intérêt se porte sur l'intériorité et la compréhension du sens vécu. | | **Échantillonnage** | **Ce que c'est :** Un échantillonnage théorique évolutif.
Concrètement : Les participants sont choisis progressivement en fonction des besoins théoriques émergents (ex : sélectionner un profil spécifique pour tester ou enrichir une catégorie).
Pourquoi ATA : L'échantillonnage en ATA n'est jamais fixé à l'avance ; il accompagne le développement de la théorie. | **Ce que c'est :** Une sélection homogène de participants ayant vécu la même expérience.
Concrètement : L'objectif est de comparer des variations fines dans un même type d'expérience (ex : tous ont vécu un même diagnostic).
Pourquoi AIP : L'approche phénoménologique nécessite une homogénéité pour isoler la structure de l'expérience vécue. | | **Recueil de données** | **Ce que c'est :** Entretiens semi-dirigés, observations, parfois focus groups.
Concrètement : Les données doivent être riches et variées afin de permettre la construction progressive de catégories.
Pourquoi ATA : La multiplication des sources permet une saturation conceptuelle solide et un croisement des perspectives. | **Ce que c'est :** Entretiens ouverts et approfondis.
Concrètement : Le participant raconte librement son expérience avec un minimum d'intervention du chercheur.
Pourquoi AIP : La priorité est la narration personnelle, subjective, intime. Les méthodes ouvertes favorisent cette profondeur phénoménologique. | | **Analyse** | **Ce que c'est :** La comparaison constante et trois niveaux d'analyse :
– ouverte : émergence des premières unités de sens
– axiale : articulation des catégories entre elles
– intégrative : construction du modèle théorique final
Concrètement : L'analyse est extrêmement systématique, progressive et recursive.
Pourquoi ATA : L'ATA repose strictement sur cette architecture analytique pour faire émerger une théorie. | **Ce que c'est :** Analyse focalisée sur chaque participant, puis comparaison secondaire entre participants :
– ouverte : premiers thèmes issus du discours
– axiale : structuration des thèmes
– intégrative : compréhension globale de l'expérience
Concrètement : L'analyse vise l'interprétation du vécu individuel avant la comparaison interindividuelle.
Pourquoi AIP : L'accent est mis sur l'interprétation herméneutique du sens vécu, non sur la construction d'un modèle généralisable. | | **Modèle explicatif** | **Ce que c'est :** Production d'une

théorie substantielle enracinée dans les données.
Concrètement : Un modèle processuel décrivant comment un phénomène social fonctionne.
Pourquoi ATA : L'ATA est spécifiquement conçue pour aboutir à une théorie nouvelle. | **Ce que c'est :** Production d'une **synthèse conceptuelle** des thèmes qui décrivent l'expérience.
Concrètement : Le résultat final est une compréhension raffinée de la manière dont un individu vit un phénomène.
Pourquoi AIP : L'objectif n'est pas de théoriser mais de clarifier les structures de l'expérience vécue. |

ATA = comprendre un processus social

- Production d'une théorie
- Construction de catégories
- Analyse très systématique (ouverte, axiale, intégrative)
- Échantillonnage théorique
- Multiplicité des données
- Intérêt pour les interactions, les dynamiques, les mécanismes

AIP = comprendre une expérience vécue

- Focale sur la subjectivité
- Explication du sens donné à l'expérience
- Entretiens ouverts
- Échantillonnage homogène
- Analyse centrée sur chaque personne avant la comparaison
- Finalité : interpréter, pas théoriser

35Méthodes mixtes

1 Les Fondamentaux de la Recherche Qualitative

La recherche en sciences humaines, en médecine et en psychiatrie a longtemps été traversée par un clivage profond entre deux univers : celui des mots (qualitatif) et celui des nombres (quantitatif). Cette opposition façonne encore les représentations actuelles, bien qu'elle soit devenue largement insuffisante pour comprendre les méthodes réellement utilisées dans la recherche moderne.

1.A Diversité des Approches et Origines

Il n'existe pas *une* méthode qualitative, mais plusieurs, voire "plein".

Notions importantes : Exemples de méthodes qualitatives

1. **Théorie Ancérée (Grounded Theory) :** C'est la méthode que les généralistes utilisent le plus souvent. Elle a été proposée à l'origine par des sociologues.
2. **IPA (Interpretative Phenomenological Analysis)** : Très utilisée en psychologie, elle se concentre sur l'expérience subjective des individus.
3. **Phénoménologie** : Approche philosophique visant à décrire les expériences vécues sans les interpréter.
4. **Analyse Thématique**.
5. **Analyse Lexicale**.
6. **Analyse de Cas**.
7. **IPSEA :** Développée par des collègues psychiatres français, elle est proche de l'IPA.

La recherche qualitative existe depuis longtemps (ex: Freud avec l'introduction à la psychanalyse ou la psychopathologie de la vie quotidienne).

Cependant, les travaux historiques comme ceux de Freud sont souvent considérés comme discutables par les spécialistes modernes de la recherche qualitative, car ils n'étaient **pas du tout rigoureux** selon les canons actuels.

Leurs travaux n'incluaient ni **triangulation**, ni recueil d'un corpus *in extenso*, ni **codage**.

- Triangulation : codage en double (être 2 sur le même jeu de données)
- Saturation : on définit pas un nombre de sujets à inclure à l'avance : on arrête
- Purposive sampling : échantillonnage raisonné (= choix des sujets en fonction de critères précis, les plus hétérogènes possibles)

1.B Exigence de Rigueur et Notions Clés du Protocole

Historiquement, la **Théorie Ancérée (Glaser et Strauss)** est née d'une volonté de durcir la méthodologie des sciences humaines et sociales.

Face aux critiques des sciences dites "dures" qui considéraient leurs travaux comme de la "littérature" et entraînaient une réduction des budgets, des doctorats et des postes, les sociologues ont

décidé de rendre leur méthodologie plus rigoureuse.

Selon les principes de la science, pour qu'une recherche soit scientifique, un **protocole écrit** est nécessaire.

Le scientifique doit **choisir** (critères d'inclusion, d'exclusion, plan d'analyse) pour "bien regarder" et éviter de "mal regarder".

Ce choix et cette organisation sont cruciaux et s'opposent aux approches cliniques non structurées (comme les *case reports* sans protocole défini).

1.C Concepts Méthodologiques Essentiels

1.C.1 Positionnement du Chercheur : Neutralité et Distance

Le concept de **neutralité axiologique** est très fort en Théorie Ancérée.

- **Neutralité Axiologique :** L'effort maximal pour **se soustraire de la question de recherche et se mettre à distance**. L'objectif est que la théorie "jaillisse" des entretiens et non de la tête du chercheur.
- **Implication du Chercheur :** *Attention, point clé:* En réalité, la neutralité est un mythe. **Même en recherche quantitative biomédicale, les chercheurs ne sont jamais neutres** et sont toujours impliqués émotionnellement dans leur objet de recherche.
  - *Exemple :* Un chirurgien évaluant une chirurgie ou un chercheur évaluant un médicament. Les scientifiques en médecine (s'intéressant à l'humain et aux maladies) sont souvent encore plus impliqués que, par exemple, les physiciens.
- **Réflexologie (ou Réflectologie) :**
  - Fait partie de la recherche qualitative rigoureuse.
  - C'est le chapitre où le chercheur "se demande ce qu['il est] par rapport à [son] objet d'étude".
  - Il s'agit de reconnaître les biais potentiels qui peuvent influencer l'interprétation des résultats. Par exemple, une doctorante enceinte travaillant sur l'adoption internationale doit réfléchir à l'influence de son état sur son travail.

1.C.2 La Question de Recherche et le Corpus

- **Question Ouverte :** La question de recherche doit être fondamentalement **ouverte**. L'objectif est de comprendre une situation ou une expérience (ex : "Quel est le vécu subjectif des gens qui sont seuls, qui ont 25 ans et qui sont confinés ?").
- **Corpus :** L'ensemble des données recueillies, souvent du texte (entretiens), mais aussi des vidéos ou des photos.
- **Entretiens :** Souvent longs (autour d'une heure) et très ouverts. Ils sont dits **semi-structurés** car certaines thématiques doivent être abordées. Le soignant doit parler peu et laisser le sujet parler beaucoup, tout en le ramenant au questionnement principal s'il s'égare.
- **Photo-Élicitation :** Technique alternative (ou complémentaire) où l'on montre des photos

aux patients (ou on leur demande d'en prendre) pour les faire parler et "éliciter" une description de leur vécu subjectif.

1.C.3 L'Échantillonnage et le Codage

- **Purposive Sampling (Échantillonnage ciblé) :**
  - Contrairement au quantitatif, où l'on cherche l'aléatoire (le tirage au sort, bien que ce soit rarement le cas en pratique), le qualitativiste choisit des personnes avec un **maximum d'hétérogénéité** (hommes, femmes, vieux, jeunes, riches, pauvres, etc.) en fonction de la thématique étudiée.
  - L'objectif est d'éviter de saturer rapidement en incluant des sujets toujours similaires.
- **Corpus** : données recueillis *in extenso* (tout l'entretien, pas seulement des extraits) avec enregistrements et retranscription intégrale
- **Codage :**
  - L'étape de transformation du corpus. Le chercheur remplace une phrase par un mot ou une expression plus courte (**micro-codage**), puis remplace ces mots par des mots plus généraux.
    - L'objectif final est de réduire la complexité du corpus pour arriver à un nombre limité de **méga-catégories** (ou thèmes).
    - Outils : Des logiciels comme **NVivo, Atlas.ti** aident au codage, à l'enregistrement des codes, et à la récupération des phrases clés associées à un thème.

1.C.3.1 Validation des Résultats

- **Triangulation :**
  - C'est le fait d'utiliser plusieurs juges ou **codeurs** pour effectuer le codage. Cela permet de discuter et de se mettre d'accord sur les différences de catégorisation.
  - C'est l'équivalent qualitatif d'un accord inter-juge.
- **Saturation :**
  - C'est le **critère d'arrêt de l'étude**.
  - Contrairement au quantitatif où le nombre de sujets est défini à l'avance, en quali, on arrête d'inclure des sujets quand l'analyse ne produit **plus de nouveaux codes**.
  - Les nouveaux patients ne font que confirmer les thèmes déjà établis.

---

2 Le Clivage Quantitatif/Qualitatif et les Méthodes Mixtes

2.A Opposition Historique et Philosophique

L'opposition entre les mots (Quali) et les nombres (Quanti) est très ancienne :

- **Mythos vs. Logos :** Chez les philosophes grecs, l'opposition existait déjà entre le **mythe** (récit pour comprendre ce que la logique n'explique pas) et le **logos** (raison, logique, géométrie, qui structure le quantitatif).
- **Esprit de Finesse vs. Esprit de Géométrie (Pascal) :**
  - **Esprit de Géométrie :** Conceptualiser, définir clairement, organiser les concepts pour obtenir des conclusions rigoureuses et partageables (Quantitatif).
  - **Esprit de Finesse :** Compréhension rapide, intuitive (comme le poète qui condense une vérité profonde en quelques mots).
- **Systèmes de Traitement de l'Information (Kahneman) :** Deux processus de prise de décision chez l'humain : lent et argumenté (Géométrie/Rationnel) contre rapide et intuitif (Finesse/Intuitif).

2.A.1 Le Conflit Académique et ses Enjeux

Cette opposition se traduit par un **clivage** dans l'université (sciences dures vs. sciences molles).

- **Guerre Quanti/Quali :** Les quantitativistes peuvent percevoir les qualitativistes comme peu rigoureux ("charlots"), tandis que les qualitativistes peuvent voir les quantitativistes comme des "robots" manipulant des nombres sans comprendre l'intimité humaine.
- **Valorisation :**
  - Historiquement, la médecine a vu une évolution des profils littéraires (latin, grec) vers une forte pression sociale vers le quantitatif (mathématiques), accentuée par la digitalisation du monde.
  - Les grandes revues médicales restent frileuses envers le Quali, bien que les revues de spécialité publient de plus en plus d'articles qualitatifs et mixtes.

2.B La Différence Méthodologique Fondamentale

Les différences listées par des sociologues (comme Becker) entre quali et quanti (données ouvertes vs. fermées, étude du point de vue subjectif, étude en vraie vie, critères de validité différents) ne sont pas si claires ou tranchées, surtout avec l'avènement des nouvelles technologies (comme le Traitement Automatique du Langage/TAL).

Point clé : La différence intrinsèque actuelle réside dans le codeur humain.

- **Théorie de l'Esprit :**
  - Lors du codage qualitatif, le codeur **rajoute énormément d'informations** à l'information brute du corpus.
  - Il se demande ce que le sujet a voulu dire, utilisant sa **théorie de l'esprit** (sa connaissance de ce qu'est un être humain et ce qu'il est susceptible de penser) pour simplifier et synthétiser la réponse.
- **Conséquence :** C'est cette "magie" du codeur humain qui permet d'obtenir des résultats profonds et généralisables avec un petit nombre de sujets (ex: 20 sujets).
- **L'Avenir et l'IA :**
  - Un logiciel de statistiques n'a pas (encore) une théorie de l'esprit.

- Cependant, avec les outils de *deep learning* et l'analyse des requêtes (comme Google), on peut imaginer que l'Intelligence Artificielle pourrait acquérir une théorie de l'esprit, ce qui pourrait amenuiser cette différence méthodologique dans les prochaines années.

2.C Stratégies des Méthodes Mixtes

L'utilisation conjointe des méthodes qualitatives et quantitatives est appelée **méthodes mixtes** .

Le mélange apporte une plus-value car le quantitatif offre une vision macroscopique et populationnelle, tandis que le qualitatif explore la réalité clinique en détail.

Il existe plusieurs façons de les combiner :

1. **Qualitatif avant Quantitatif (Quali** $\rightarrow$ Quanti) :
   - *Objectif :* Le quali est exploratoire et aide à construire le quanti.
   - *Exemple :* Faire une étude qualitative pour comprendre les problèmes des patients atteints d'une maladie orpheline afin de créer ensuite un **critère d'efficacité** rigoureux pour un essai clinique quantitatif.
2. **Quantitatif avant Qualitatif (Quanti** $\rightarrow$ Quali) :
   - *Objectif :* Le quali est explicatif.
   - *Exemple :* Une étude quantitative montre un résultat surprenant (ex : les femmes héroïnomanes prennent moins de mesures de minimisation des risques que les hommes). Une étude qualitative subséquente (interviews des femmes) révèle que cela est dû au fait qu'elles sont en couple avec leur dealer, ce qui affecte leur partage de seringues.
   - *Avantage pour l'échantillonnage :* Les données quantitatives massives (âge, sexe, sévérité) permettent de réaliser un *purposive sampling* optimal pour le quali en assurant l'hétérogénéité maximale de l'échantillon qualitatif.
3. **Qualitatif et Quantitatif en Parallèle (Quali + Quanti) :**
   - *Objectif :* Triangulation.
   - *Exemple 1 :* Mener un essai randomisé (quanti) pour l'efficacité d'un anticancéreux, et en parallèle, une étude qualitative pour comprendre comment les patients tolèrent et vivent le médicament.
   - *Exemple 2 :* Appliquer une analyse de contenu qualitative (faite par des humains) et un **Traitement Automatique du Langage (TAL)** (statistique/quantitatif) au même corpus textuel. Les résultats du TAL sont souvent plus pauvres ou superficiels, mais la comparaison aide à trianguler et valider les thèmes, ou à comprendre les limites de l'outil textuel.

2.D Vers une Typologie Épistémologique Élargie

L'opposition simple quali/quanti n'est pas suffisante pour décrire tous les domaines scientifiques.

Par exemple, la biologie moléculaire (qui utilise des mots/concepts et des “équations” comme “l’ARN vient de l’ADN”) est considérée comme une science dure mais ne s’appuie pas sur des statistiques.

Une typologie 3x3 permet de dépasser ce clivage en utilisant deux dimensions : la nature de l’objet d’étude et la méthode de relation.

Typologie de la Recherche Scientifique:

Nature de l’Objet ↓ / Méthode de Relation →	Équations	Statistiques	Herméneutique
Globalement Quantitatif (Que des nombres)	**Physique** (Lois universelles)	Thermodynamique statistique	Théorie des catastrophes (René Thom)
Localement Quantitatif (Mesures locales)	Pharmacocinétique (Demi-vie d’un médicament)	**Recherche Clinique & Épidémiologie**	Études N-of-1 (Représentation graphique sans tests stat.)
Globalement Qualitatif (Que des mots)	**Biologie Moléculaire** (Équations avec des mots)	Traitement Automatique du Langage (TAL)	**Recherche Qualitative pure**

Cette grille montre qu’il existe d’autres manières de conceptualiser la recherche, où le Quantitatif et le Qualitatif ne sont que des aspects des méthodologies utilisées.

L’épidémiologie utilise par exemple des statistiques sur des mesures locales (localement quantitatif), tandis que la recherche qualitative utilise une interprétation humaine (**herméneutique**) sur un corpus globalement qualitatif.

Analogie pour le concept de Triangulation/Méthodes Mixtes :

Considérez la recherche sur un sujet complexe, comme une personne face à une maladie.

- La **méthode quantitative** est comme une **photographie satellite** de la personne et de sa population : elle donne la taille, la fréquence des symptômes, la corrélation avec d’autres facteurs (statistiques) ou même des lois universelles (équations). C’est précis mais macroscopique.
- La **méthode qualitative** est comme le **journal intime** de la personne, décrypté et interprété par un psychologue (le codeur humain) : elle révèle la signification profonde, le vécu subjectif et les bifurcations existentielles (herméneutique). C’est fin et détaillé, mais local.

La **méthode mixte** est l’association des deux : le scientifique dispose de la carte satellite (Quanti) pour bien cibler les zones importantes et des pages de journal (Quali) pour comprendre pourquoi les choses se passent ainsi sur le terrain, offrant une connaissance beaucoup plus complète que l’une ou l’autre méthode seule.

3 Vers une typologie plus riche : la matrice 3×3 de Falissard et al. (2013)

1. Dimension 1 : comment les objets sont représentés ?

- Littérairement : la réalité est décrite uniquement par des mots (psychanalyse classique, récits de cas, ethnographie).
- Mathématisation locale : Seules certaines facettes sont quantifiées (échelles de dépression, tests cognitifs, IRM fonctionnelle). Cas typique de la recherche clinique et épidémiologique.
- Mathématisation globale : La réalité est entièrement représentée par des structures mathématiques (physique classique, modèles formels complets).
- Dans le champ de la psychiatrie, seul un modèle marginal (spin glass) se rapproche de cette case.

2. Dimension 2 : comment les relations entre objets sont établies ?

- Herméneutique : Interprétation, induction, compréhension, exégèse. Typique de la psychanalyse ou de l'analyse de récits.
- Statistiques : Relations établies "en moyenne", pas universelles. Typique de l'épidémiologie et des essais cliniques.
- Équations : Relations déterministes, formelles, universelles. Exemple : loi de la gravitation.

3. La matrice 3×3 (reproduction structurée)

Objet / Relations	A. Herméneutique	B. Statistiques	C. Équations
1. Littéraire	Psychanalyse, ethnopsychiatrie, récits cliniques	Text mining, analyses linguistiques automatisées	Mécanismes neurobiologiques formulés en langage structuré (ex : "fluoxétine = ISRS")
2. Mathématisation locale	Designs AB/ABAB interprétés visuellement	Épidémiologie, IRM, neurosciences cognitives, essais cliniques	Pharmacocinétique (demi-vie), formalisme lacanien
3. Mathématisation globale	Modèles spin glass interprétés	Aucun exemple psychiatrique	Physique (mécanique classique, relativité)

4 Les méthodes mixtes : fondements, usages, typologies

4.A Pourquoi mélanger les approches ?

- Le qualitatif permet :
  - de comprendre,

  - d’explorer,
  - d’expliquer les mécanismes,
  - d’accéder à l’expérience subjective.
- Le quantitatif permet :
  - de mesurer,
  - de tester,
  - de généraliser.

Les méthodes mixtes utilisent les forces de chaque approche pour construire une validation croisée plus robuste.

4.B Trois stratégies classiques de méthodes mixtes

4.B.1 Quali Quanti (exploratoire)

Le qualitatif sert à :

- définir les concepts,
- comprendre les mécanismes,
- identifier les dimensions pertinentes,
- préparer les questionnaires,
- construire les critères de jugement pour un essai.

Exemple : comprendre le vécu des patients atteints d’une maladie orpheline pour élaborer un critère d’efficacité.

4.B.2 Quanti Quali (explicatif)

Le qualitatif sert à expliquer un résultat inattendu du quantitatif.

Exemple célèbre cité dans l’article : Des femmes héroïnomanes prennent moins de mesures de réduction des risques que les hommes ; les entretiens révèlent que cela est lié à la relation asymétrique avec leur dealer.

Le quantitatif aide aussi à sélectionner un échantillon qualitatif maximalement hétérogène.

4.B.3 Quanti + Quali en parallèle (triangulation)

Deux corpus indépendants, deux analyses complémentaires.

Peut inclure :

- essai clinique + entretiens patients,
- TAL (analyse automatique) + analyse humaine.

4.C Pourquoi cela fonctionne-t-il scientifiquement ?

Parce que l'analyse qualitative utilise la théorie de l'esprit du codage humain : le codeur ajoute énormément d'information implicite tirée de sa connaissance du monde, de la psychologie humaine, des interactions sociales.

Cette capacité dépasse encore largement celle des systèmes automatiques contemporains.

5 Applications en psychiatrie, médecine et neurosciences

L'article fournit de nombreux exemples concrets :

- Psychanalyse : littéraire × herméneutique.
- Texte automatique : littéraire × statistiques.
- Neurobiologie : littéraire × équations (langage structuré quasi formel).
- Épidémiologie : localement mathématique × statistiques.
- Pharmacocinétique : localement mathématique × équations.
- Modèles spin glass : globalement mathématique × herméneutique.

La matrice démontre que la psychiatrie mobilise pratiquement toutes les cases, ce qui rend illusoire l'idée d'une domination exclusive du quantitatif.

6 Enjeux philosophiques, épistémologiques et politiques

Le débat quali/quanti n'est pas seulement méthodologique : il est également idéologique et institutionnel.

Falissard et al. rappellent :

- certaines politiques publiques ont exigé "l'evidence-based" au détriment du qualitatif ;
- certaines positions inverses rejettent les approches quantitatives en psychothérapie ;
- la recherche clinique pourrait se couper du vécu des patients si elle se focalise trop sur les scores et les images cérébrales ;
- la clinique souffre lorsqu'elle se coupe trop des résultats d'essais contrôlés.

L'enjeu des méthodes mixtes est précisément de réconcilier ces pôles.

7 Conclusion générale : pourquoi dépasser le clivage quali/quanti ?

- Le clivage historique est trop pauvre.
- La réalité contemporaine de la recherche mobilise des approches multiples.
- La typologie 3×3 permet de comprendre cette diversité.
- Les méthodes mixtes tirent parti de cette richesse.

- Elles améliorent à la fois la compréhension et la généralisation. • Elles reconnectent recherche, clinique et compréhension du vécu.

Les méthodes mixtes ne sont donc pas un compromis, mais une architecture épistémologique complète, s'appuyant sur les apports uniques des mots, des nombres, de la théorie de l'esprit humaine et des modèles formels.

Elles constituent aujourd'hui un cadre méthodologique central pour la recherche psychiatrique, médicale et psychosociale, permettant d'articuler compréhension, mesure et action scientifique.

36Bonus residus

1 Contexte général : ce que fait un modèle

On observe, pour chaque individu i :

- une (ou plusieurs) variable(s) explicative(s) X_i
- une variable réponse Y_i

Exemple :

- $X_i = $ âge du patient (en années)
- $Y_i = $ pression artérielle systolique (en mmHg)

On veut résumer la relation entre X et Y par un modèle linéaire :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

avec :

- β_0 : intercept (valeur moyenne de Y quand $X = 0$)
- β_1 : pente (variation moyenne de Y pour une unité de X)
- ε_i : terme d'erreur théorique, la partie de Y_i non expliquée par le modèle

En pratique, on ne connaît pas β_0, β_1, on les estime à partir des données.
□

2 Observé, prédit, résidu : bien distinguer

Une fois le modèle ajusté, on obtient des estimations $\hat{\beta}_0$ et $\hat{\beta}_1$.
On peut alors calculer, pour chaque individu i :
Valeur observée :

$$y_i$$

Valeur prédite par le modèle :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

Résidu :

$$e_i = y_i - \hat{y}_i$$

Interprétation :

- si $e_i > 0$: l'observation est au-dessus de ce que le modèle prédit
- si $e_i < 0$: l'observation est en dessous de ce que le modèle prédit
- si $e_i = 0$: le modèle tombe pile sur la valeur observée

Les résidus :

- sont dans la même unité que Y (mmHg, g/L, kg, etc.)
- mesurent l'erreur de prédiction du modèle pour chaque observation

□

3 Terme d'erreur ε_i vs résidu e_i

3.A Terme d'erreur ε_i (théorique)

Dans le modèle :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

on suppose que ε_i vérifie (dans le modèle linéaire gaussien) :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

avec :

- $E(\varepsilon_i) = 0$
- Var$(\varepsilon_i) = \sigma^2$
- indépendance entre individus

ε_i est un objet théorique : on ne l'observe jamais directement.

3.B Résidu e_i (observé)

Le résidu se définit par :

$$e_i = y_i - \hat{y}_i$$

- il se calcule à partir des données et du modèle ajusté
- il dépend des estimations $\hat{\beta}_0$, $\hat{\beta}_1$
- c'est une estimation du terme d'erreur ε_i

Résumé :

- ε_i = erreur idéale dans le modèle théorique
- e_i = erreur observée (ce qu'on voit réellement dans les données)

4 Rôle des résidus dans la régression linéaire

4.A Méthode des moindres carrés

Dans la régression linéaire classique, les coefficients $\hat{\beta}_0, \hat{\beta}_1$ sont choisis pour minimiser la somme des carrés des résidus :

$$\text{SSR} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

L'algorithme cherche donc à rendre les résidus globalement aussi petits que possible (au sens des carrés).

4.B Hypothèses du modèle = hypothèses sur les résidus

Dans le modèle linéaire gaussien, on fait des hypothèses sur les ε_i.
En pratique, on vérifie ces hypothèses sur les résidus

$$e_i$$

:

4.B.1 Moyenne nulle

Théorique :

$$E(\varepsilon_i) = 0$$

En pratique : les résidus doivent osciller autour de 0.

4.B.2 Variance constante (homoscédasticité)

Théorique :

$$\text{Var}(\varepsilon_i) = \sigma^2 \quad \text{(indépendante de } X_i\text{)}$$

En pratique : la dispersion des résidus ne doit pas augmenter ou diminuer systématiquement avec les valeurs prédites.

4.B.3 Indépendance

Théorique : les ε_i sont indépendants.

En pratique : les résidus ne doivent pas montrer de structure dans le temps, par centre, par patient, etc.

4.B.4 Normalité (dans le cadre gaussien)

Théorique :

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

En pratique : l'histogramme et le QQ-plot des résidus doivent être compatibles avec une loi normale.

Toute l'analyse de diagnostic (graphiques de résidus, tests) repose sur le comportement des résidus.

□

5 Exemple concret en R avec un jeu de données fictif

On crée un jeu de données fictif simple :

- X = âge du patient, entre 20 et 80 ans
- Y = pression artérielle systolique (PAS, en mmHg)

On suppose que le « vrai » modèle (utilisé pour simuler les données) est :

$$Y_i = 90 + 0{,}6X_i + \varepsilon_i$$

avec :

$$\varepsilon_i \sim \mathcal{N}(0, 10^2)$$

```r
# Création d'un jeu de données fictif

n <- 100                              # nombre de patients
age <- runif(n, min = 20, max = 80)  # âges entre 20 et 80 ans

# Paramètres "vrais" du modèle de simulation
beta_0 <- 90      # intercept
beta_1 <- 0.6     # pente
sigma  <- 10      # écart-type de l'erreur
```

```r
# Terme d'erreur théorique simulé
epsilon <- rnorm(n, mean = 0, sd = sigma)

# Valeur observée de PAS (simulée)
PAS <- beta_0 + beta_1 * age + epsilon

# Data frame final
df <- data.frame(
    age = age,
    PAS = PAS
)

head(df)
```

```
       age      PAS
1 37.25465 114.8860
2 67.29831 130.0935
3 44.53862 116.2945
4 72.98104 147.4746
5 76.42804 133.5991
6 22.73339 118.8047
```

5.A Ajuster un modèle linéaire et récupérer les résidus

On ajuste le modèle linéaire :

$$\text{PAS}_i = \beta_0 + \beta_1 \cdot \text{age}_i + \varepsilon_i$$

avec `lm()`.

```r
# Ajustement du modèle linéaire
mod <- lm(PAS ~ age, data = df)

# Résumé du modèle
summary(mod)
```

```
Call:
lm(formula = PAS ~ age, data = df)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.3797  -6.1323  -0.1973   5.9633  22.1723 

Coefficients:
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 90.20984    3.00425   30.03   <2e-16 ***
age          0.58503    0.05697   10.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.693 on 98 degrees of freedom
Multiple R-squared:  0.5183,    Adjusted R-squared:  0.5134
F-statistic: 105.5 on 1 and 98 DF,  p-value: < 2.2e-16
```

On extrait ensuite :

- les valeurs prédites $\widehat{y}_i$
- les résidus $e_i = y_i - \widehat{y}_i$

```r
# Valeurs prédites et résidus
df$y_hat  <- fitted(mod)   # valeurs prédites
df$residu <- resid(mod)    # résidus

head(df)
```

```
       age      PAS    y_hat      residu
1 37.25465 114.8860 112.0049  2.88111844
2 67.29831 130.0935 129.5812  0.51227617
3 44.53862 116.2945 116.2662  0.02828346
4 72.98104 147.4746 132.9058 14.56884788
5 76.42804 133.5991 134.9224 -1.32327681
6 22.73339 118.8047 103.5095 15.29522857
```

Dans ce tableau :

- PAS = y_i (valeur observée)
- y_hat = $\widehat{y}_i$ (valeur prédite par le modèle)
- residu = $e_i = y_i - \widehat{y}_i$

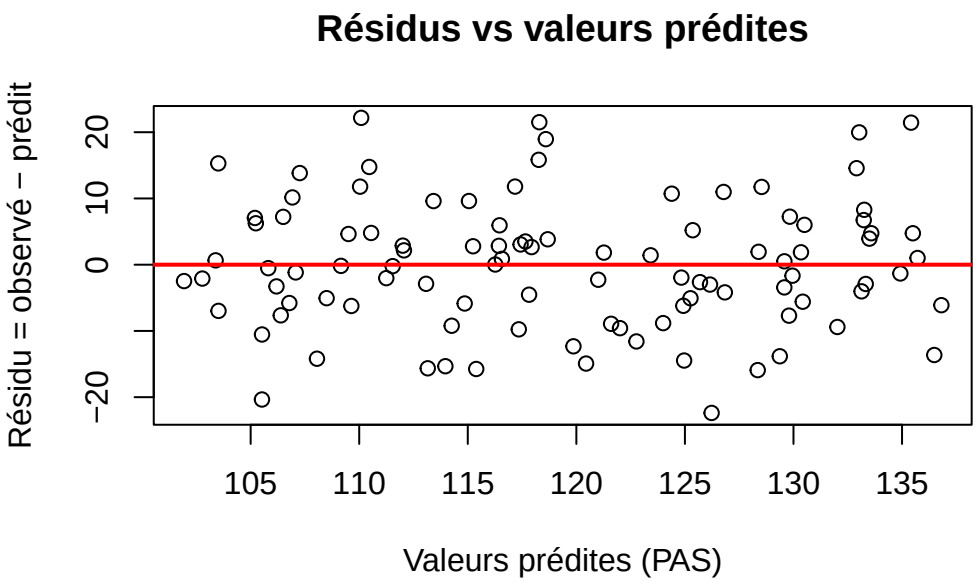
5.A.1 Graphe des résidus vs valeurs prédites

On trace les résidus en fonction des valeurs prédites :

- on souhaite les voir centrés autour de 0
- sans forme particulière (pas de « V », pas de structure claire)

```r
plot(
    x = df$y_hat,
    y = df$residu,
    xlab = "Valeurs prédites (PAS)",
```

```
    ylab = "Résidu = observé - prédit",
    main = "Résidus vs valeurs prédites"
)
abline(h = 0, col = "red", lwd = 2)
```



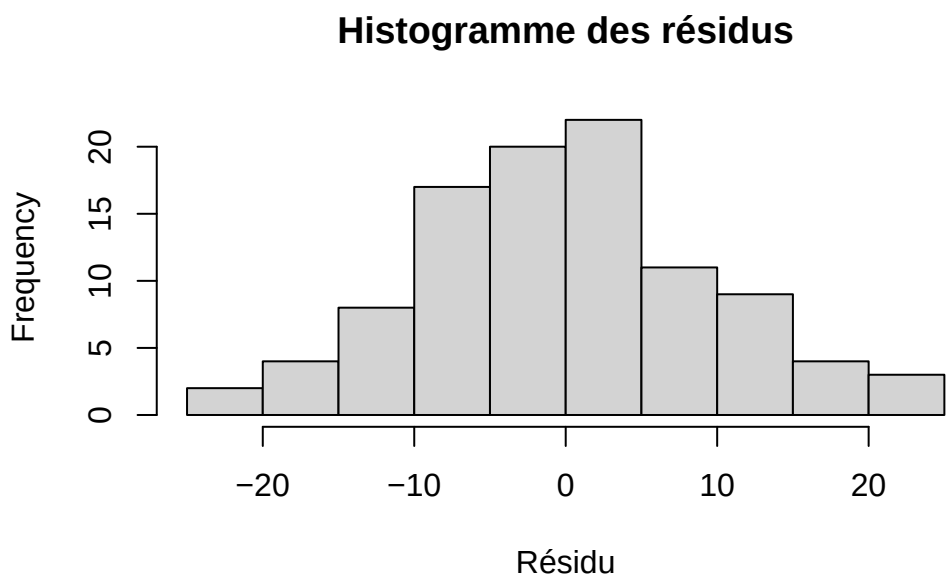
Interprétation :

- points au-dessus de la ligne rouge : PAS observée plus élevée que prévu
- points en dessous : PAS observée plus basse que prévu
- si le modèle est correct, le nuage doit être « diffus » autour de la ligne à 0, sans structure évidente

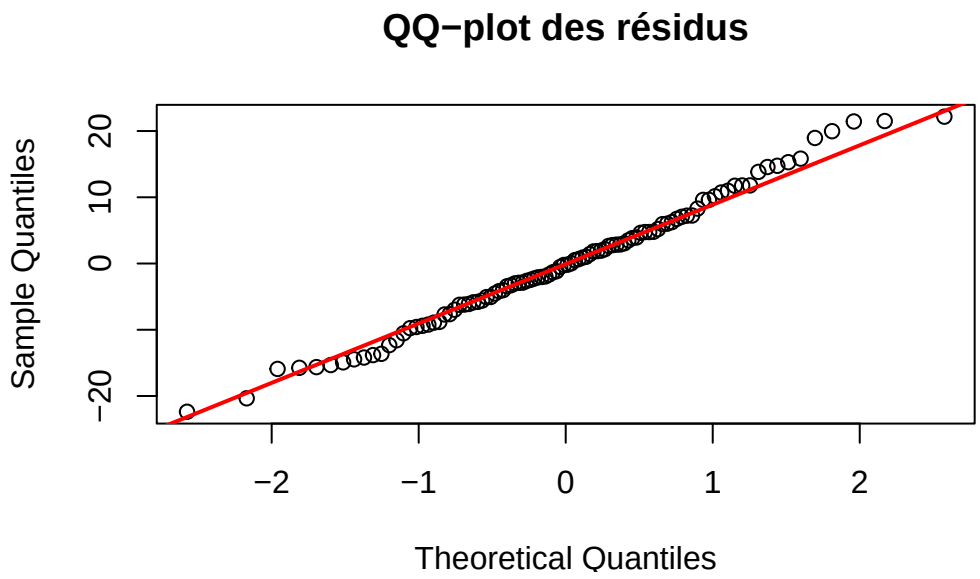
5.B Histogramme et QQ-plot des résidus

On regarde la distribution des résidus.

```
hist(
    df$residu,
    breaks = 15,
    main = "Histogramme des résidus",
    xlab  = "Résidu"
)
```




```
#| label: residuals-qqplot
#| echo: true
qqnorm(df$residu, main = "QQ-plot des résidus")
qqline(df$residu, col = "red", lwd = 2)
```



Idéalement :

- l'histogramme est approximativement symétrique autour de 0
- le QQ-plot montre les points proches de la droite : compatible avec une loi normale

6 Lien avec les modèles plus complexes

Dans des modèles plus complexes (GLM, modèles linéaires mixtes, etc.), on garde la même idée de base :

$$\text{résidu} = \text{valeur observée} - \text{valeur prédite (ou espérée) par le modèle}$$

- Dans un GLM (logistique, Poisson), la définition est adaptée pour tenir compte du fait que la variance dépend de la moyenne (résidus de Pearson, de déviance, etc.).
- Dans un modèle linéaire mixte, on définit les résidus après avoir pris en compte :
  - les effets fixes (âge, sexe, traitement…)
  - les effets aléatoires (patient, centre, etc.)

Mais le sens reste le même : les résidus mesurent ce qui n'est pas expliqué par le modèle pour chaque observation.

7 Résumé

Pour chaque observation i :

$$e_i = y_i - \hat{y}_i$$

- un résidu, c'est la différence entre la valeur observée et la valeur prédite
- les résidus sont au cœur :
  - de l'estimation (moindres carrés minimisent la somme des carrés des résidus)
  - du diagnostic du modèle (vérifier les hypothèses sur les erreurs)

Le terme d'erreur ε_i est théorique,
le résidu e_i est observable et c'est lui qu'on regarde dans les sorties et les graphiques.