

# S1\_CUSM\_RÉSUMÉ\_STATISTIQUE

## Table of contents

<b>1 Position et dispersion</b>	<b>2</b>
1.A Position . . . . .	2
1.A.1 Moyenne . . . . .	2
1.A.2 Médiane . . . . .	2
1.A.3 Mode . . . . .	2
1.B Dispersion . . . . .	2
1.B.1 Étendue = empan . . . . .	2
1.B.2 Écart interquartile (IQR) . . . . .	3
1.B.3 Écart-type . . . . .	3
1.B.4 Variance . . . . .	3
1.C Exemple sur R . . . . .	3
<b>2 Analyses en sous groupe</b>	<b>5</b>
2.A Principe . . . . .	5
2.B Dans R . . . . .	5
<b>3 Dépendance, liaison et association</b>	<b>9</b>
3.A Variables quantitatives . . . . .	9
3.A.1 Dépendance . . . . .	9
3.A.2 Dépendance monotone ou linéaire . . . . .	9
3.A.2.1 Exemple sur R . . . . .	10
3.A.2.1.1 SMP . . . . .	10
3.A.2.1.2 Données simulées . . . . .	12
3.A.2.1.3 Matrice de corrélation . . . . .	14
3.A.3 Concordance . . . . .	16
3.A.3.1 Principe . . . . .	16
3.A.3.2 Exemple sur R . . . . .	17
3.A.4 Résumé des paramètres de dépendance entre deux variables quantitatives . . . . .	19
3.B Variables catégorielles . . . . .	20
3.B.1 Dépendance . . . . .	20
3.B.1.1 Chi2 et associés . . . . .	20
3.B.1.2 Odds-ratio et risque relatif . . . . .	21
3.B.1.3 Exemple sur R . . . . .	21
3.B.2 Dépendance monotone . . . . .	22
3.B.2.1 Exemple sur R . . . . .	23
3.B.3 Concordance . . . . .	24
3.B.3.1 Coefficient kappa de Cohen . . . . .	24
3.B.3.2 Sensibilité, spécificité, VPP, VPN . . . . .	24
3.B.3.2.1 Exemple 1 sur R . . . . .	25
3.B.3.2.2 Exemple 2 sur R . . . . .	25

# 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 :

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

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  $[m - (2/3)s; m + (2/3)s]$

## 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 les 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.

```
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`).

```
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)

```

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

```
# 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...

```

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
<b>age</b>	38.94 (13.26)	39.93 (13.69)	37.41 (12.42)	0.007
<b>profession</b>				
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%)	
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%)	
<b>nb.enfants</b>	1.57 (1.85)	1.57 (1.92)	1.58 (1.73)	>0.9
<b>schizoprenie</b>	64 / 799 (8.0%)	29 / 486 (6.0%)	35 / 313 (11%)	0.008
<b>gravite</b>				<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%)	
<b>recherche.nouv</b>				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%)	
<b>evit.danger</b>				<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%)	
<b>dep.recompense</b>				<0.001
1	192 / 685 (28%)	121 / 431 (28%)	71 / 254 (28%)	
2	197 / 685 (29%)	148 / 431 (34%)	49 / 254 (19%)	

(continued)

Caractéristiques	Overall N = 799	Non dépressif N = 486	Dépressif N = 313	p-value
3	296 / 685 (43%)	162 / 431 (38%)	134 / 254 (53%)	

<sup>1</sup> Mean (SD); n / N (%)  
<sup>2</sup> 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  $\rho$  (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).

Corrélation 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 coefficient 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 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  $X$ ).

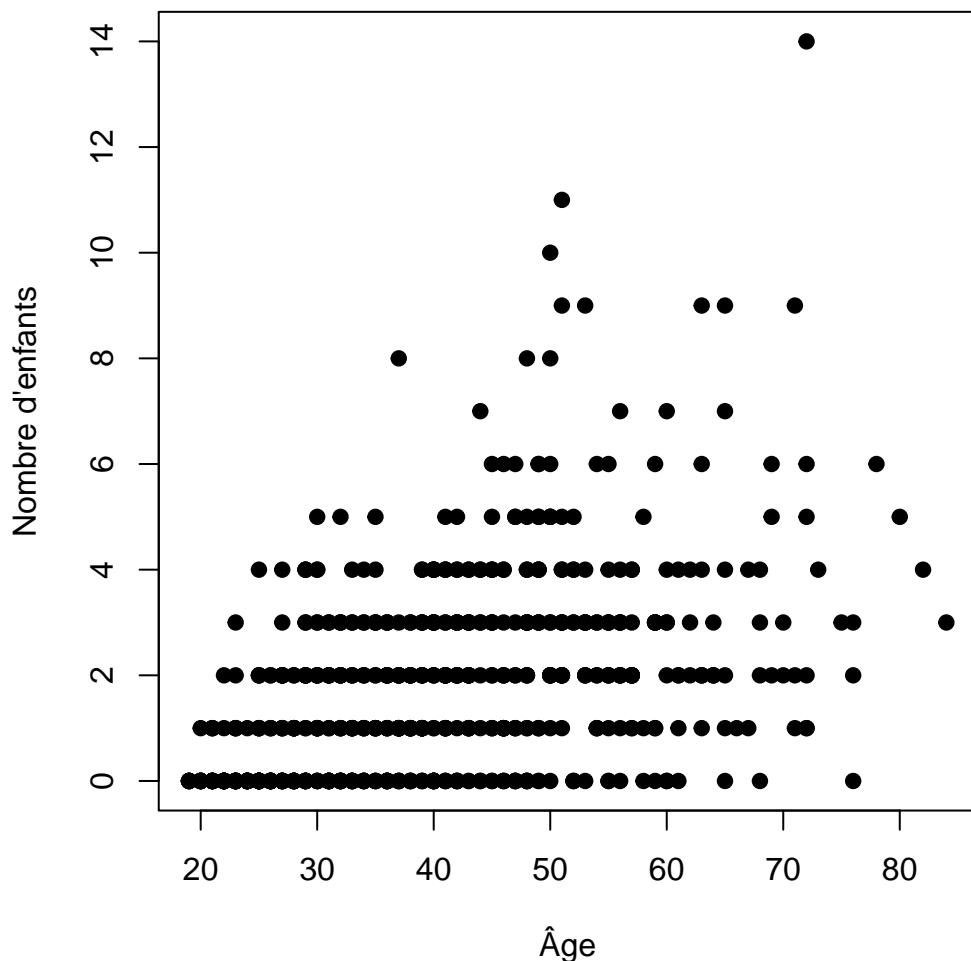
### **3.A.2.1 Exemple sur R**

#### **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
```

## Relation entre l'âge et le nombre d'enfants



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

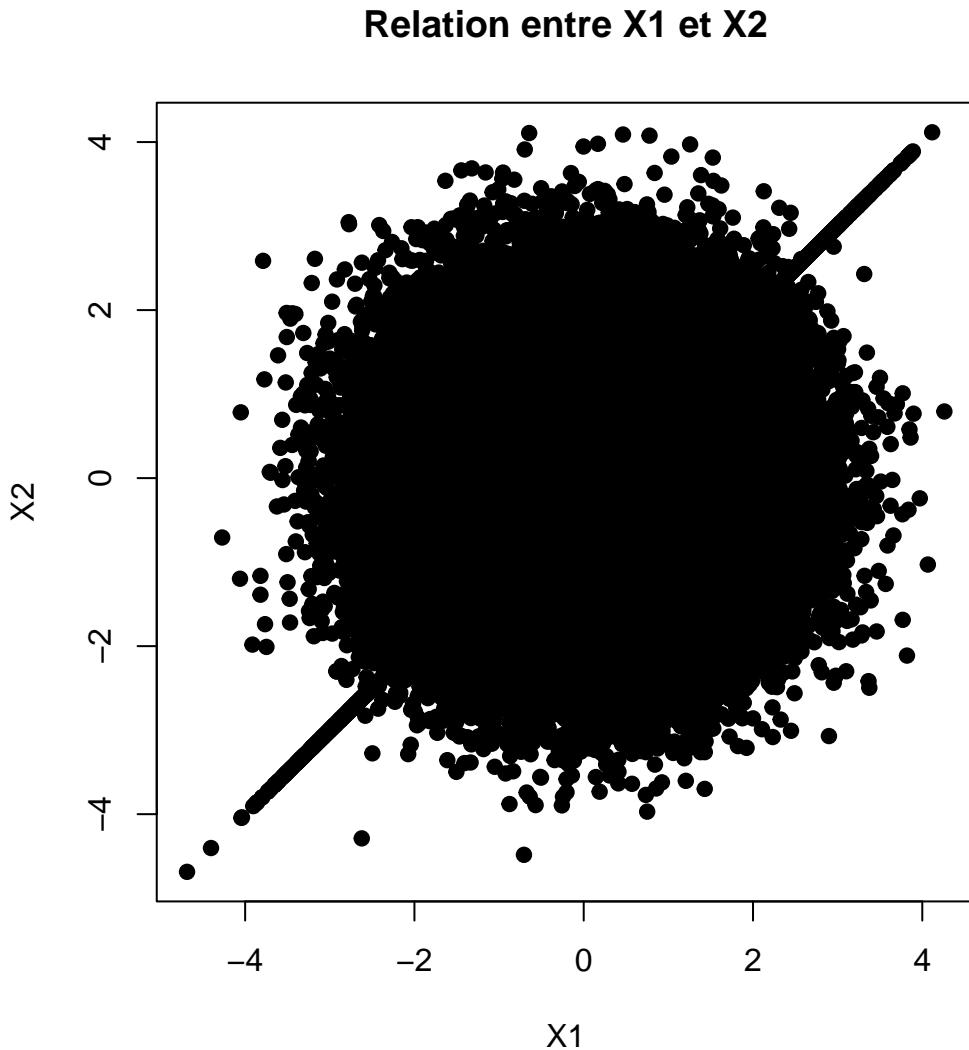
Le coefficient de corrélation est de 0.498, ce qui indique une dépendance linéaire positive entre l'âge et le nombre d'enfants.

### 3.A.2.1.2 Données simulées

```
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)
```

Représentation graphique de la relation entre X1 et X2.

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



Corrélation 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 carré 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  $X = 1 - \text{variance résiduelle}$

(mais `summary(lm())` donne directement le  $R^2$  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

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

[1] 0.754

```
# 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 $\rho$	Direction et force de la relation linéaire entre deux variables quantitatives	<code>cor(X, Y)</code>
Variance partagée	$r^2$	Proportion de la variance de $Y$ expliquée par $X$ (force de la relation linéaire)	<code>rho2 &lt;- cor(X, Y)^2</code>
Covariance	$\text{Cov}(X, Y)$	Mesure comment deux variables varient ensemble (positive : ensemble, négative : sens inverse)	<code>cov(X, Y)</code>

### 3.A.2.1.3 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.

```

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`.

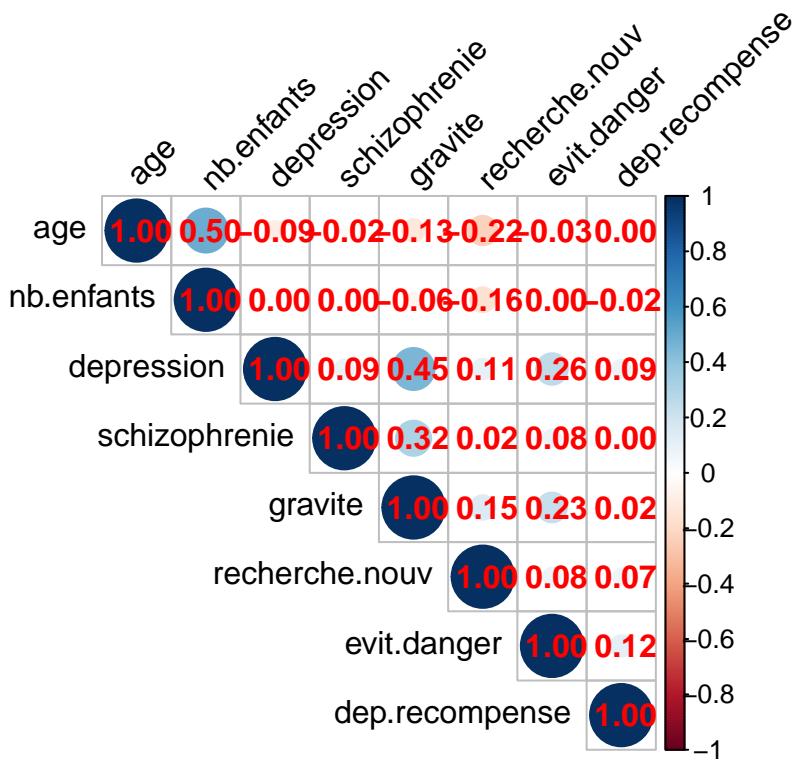
```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```

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  $X$ .

### **i Note**

#### **A quoi ça peut servir dans une étude rétrospective ?**

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 quantitatives : concordance mesurée par le coefficient de corrélation intra-classe (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 :  $\text{ICC} = \frac{\text{variance inter-patients}}{\text{variance inter-patients} + \text{variance inter-radiologues} + \text{variance résiduelle}}$

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.

```
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.

```
library(irr)
```

```
Loading required package: lpSolve

Attaching package: 'irr'

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

```

```
icc

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 intraclassé (ICC)
- 

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

## 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  $\chi^2/\text{nombre d'observations}$ )

- **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 **ordinaires** ou quantitatives non linéaires (rangées)
  - Coefficient de Spearman = corrélation de Pearson calculée sur les rangs des données
  - Varie entre -1 et +1

#### i 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
    - morts infectes
    - $OR = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \times d}{b \times c}$
    - vivants non infectes
  - 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,
    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`.

```

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:  $\text{chi2}(1) = 53.594$   $\text{Pr}>\text{chi2} = <0.001$

Fisher exact test that OR = 1:  $\text{Pr}>\text{chi2} = <0.001$

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 ordinaires (rangées) ou entre une variable ordinaire 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 ordinaires (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 ordinaires 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 reviendrait à compter la proportion de fois où elles ont la même valeur.

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

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

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

- $\text{kappa} = 0$  : concordance observée = concordance due au hasard
- $\text{kappa} = 1$  : 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'une 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  $1 - \text{spécificité}$  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  $\sim 75\text{--}140$  Hz chez les hommes et  $\sim 170\text{--}250$  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$sex == 2
voix.aigue <- vox$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 :

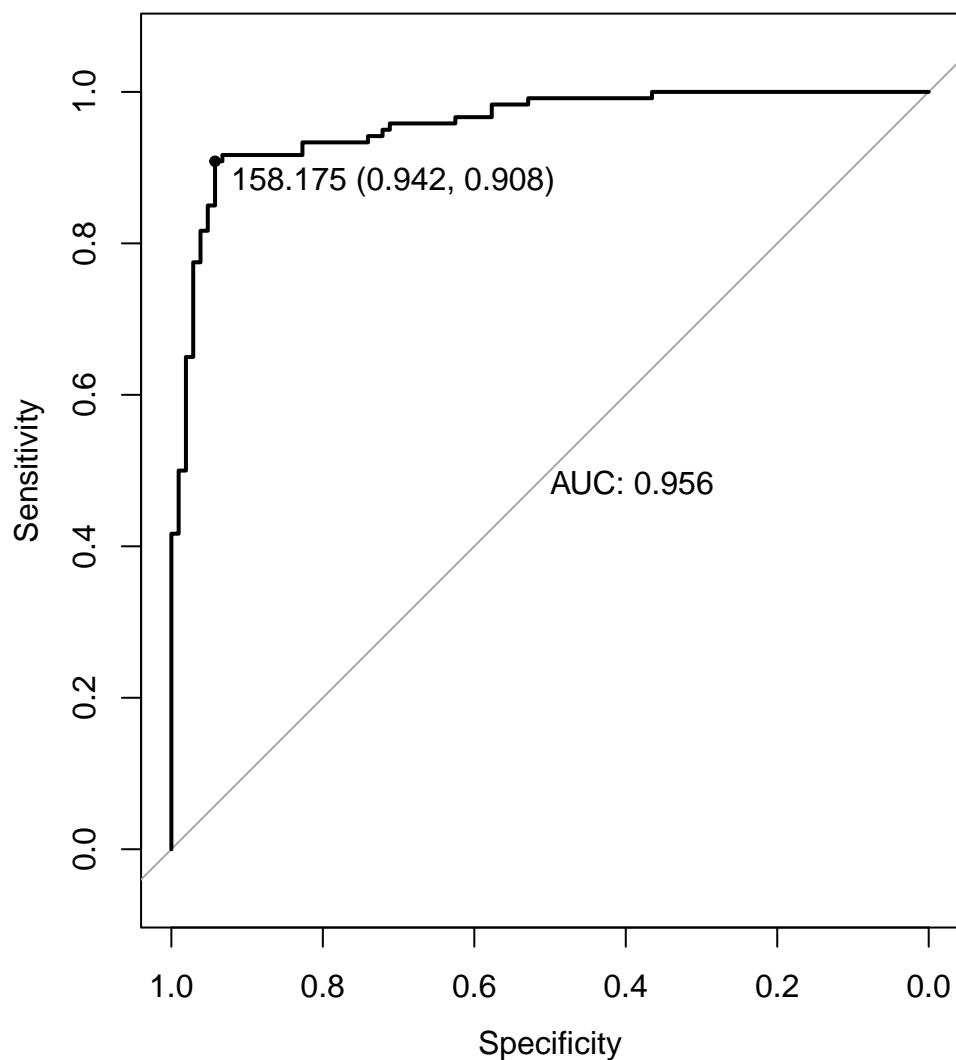
```
rocf0 <- roc(sexe.f~vox$moyf0)
```

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)
```

### Courbe ROC pour la hauteur de voix



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