

8. Analyse de deux variables qualitatives

Tableau de contingence, profils et V de Cramer

Statistique Descriptive

Anicet Ebou

Institut National Polytechnique Félix Houphouët-Boigny

2026-2027



1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

Questions agronomiques typiques

- ▶ La **variété** d'igname influence-t-elle la **qualité sanitaire** du tubercule ?
- ▶ Le **type de sol** est-il associé à la **catégorie de rendement** ?
- ▶ La **pratique culturale** (labour / non-labour) change-t-elle la **catégorie d'infestation** ?
- ▶ La **région** influe-t-elle sur la **couleur du grain** de maïs ?

Ce qu'on cherche

Savoir si deux attributs qualitatifs varient **ensemble** ou de façon **indépendante**.

Variables qualitatives

Une variable est **qualitative** (ou catégorielle) si ses modalités sont des **catégories** sans ordre naturel imposé.

Exemples : variété, région, couleur, stade phénologique, pratique culturale.

Le module 7 croisait deux variables *quantitatives*.

Ce module croise deux variables *qualitatives*.

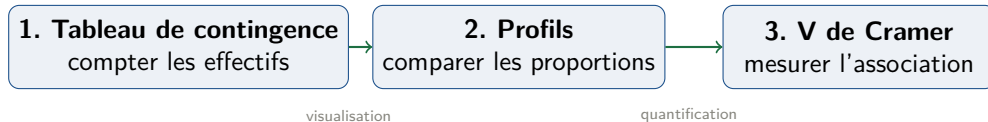


Tableau de contingence

Structure brute : effectifs
 n_{ij} pour chaque
combinaison de modalités.

Profils lignes / colonnes

Convertit les effectifs en
fréquences conditionnelles
pour comparer.

V de Cramer

Indice unique entre 0 et 1 : force
de l'association entre les deux
variables.

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

Tableau de contingence

Tableau à double entrée croisant les modalités de la variable X (lignes) et de la variable Y (colonnes).

La cellule (i, j) contient l'effectif n_{ij} : nombre d'individus ayant simultanément la modalité i de X et la modalité j de Y .

Construction du tableau de contingence

Exemple - 120 parcelles d'igname

$X = \text{Variété} \{ \text{Florido, Gnidrou, Krenglè} \}$ Y

$= \text{Qualité} \{ \text{Bonne, Moyenne, Mauvaise} \}$

	Bonne	Moyenne	Mauvaise	Total
Florido	28	12	5	45
Gnidrou	18	15	12	45
Krenglè	10	9	11	30
Total	56	36	28	120

Lecture directe

- ▶ $n_{11} = 28$: 28 parcelles Florido à bonne qualité
- ▶ $n_{i\bullet} = 45, 45, 30$: **marges lignes** (total par variété)
- ▶ $n_{\bullet j} = 56, 36, 28$: **marges colonnes** (total par qualité)
- ▶ $n = 120$: effectif total

Attention

Les effectifs bruts ne permettent pas de comparer directement les variétés si leurs totaux diffèrent.

Fréquence globale

$$f_{ij} = \frac{n_{ij}}{n}$$

Proportion de l'ensemble des individus ayant les deux modalités simultanément.

	Bonne	Moy.	Mauv.
Florido	23,3%	10,0%	4,2%
Gnidrou	15,0%	12,5%	10,0%
Krenglè	8,3%	7,5%	9,2%

Lecture

23,3 % de toutes les parcelles sont à la fois de variété Florido et de bonne qualité.

⇒ Vue d'ensemble mais difficile à interpréter seule : les variétés n'ont pas la même fréquence dans l'échantillon.

Les **profils conditionnels** (lignes ou colonnes) sont plus utiles pour comparer.

Tableau de contingence en R

```
# Données brutes
variete ← c(rep("Florido",45), rep("Gnidrou",45), rep("Krengle",30))
qualite ← c(
  rep("Bonne",28), rep("Moyenne",12), rep("Mauvaise",5), # Florido
  rep("Bonne",18), rep("Moyenne",15), rep("Mauvaise",12), # Gnidrou
  rep("Bonne",10), rep("Moyenne",9), rep("Mauvaise",11) # Krengle
)

# Tableau de contingence
tab ← table(variete, qualite)
print(tab)

# Avec marges
addmargins(tab)

# Fréquences globales (en %)
round(prop.table(tab) * 100, 1)
```

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

Profil ligne de la modalité i

Distribution de Y **parmi les individus** ayant la modalité i de X :

$$f_{j|i} = \frac{n_{ij}}{n_{i\bullet}} \quad \text{avec} \quad \sum_j f_{j|i} = 1$$

On **conditionne** sur la ligne i : chaque ligne totalise 100 %.

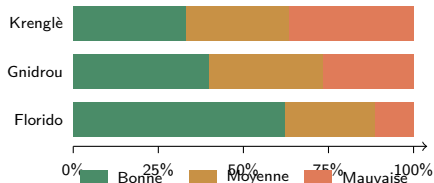
Profils lignes : fréquences conditionnelles par variété

Variété	Bonne	Moyenne	Mauvaise	Total
Florido	$28/45 = 62,2 \%$	$12/45 = 26,7 \%$	$5/45 = 11,1 \%$	100 %
Gnidrou	$18/45 = 40,0 \%$	$15/45 = 33,3 \%$	$12/45 = 26,7 \%$	100 %
Krenglè	$10/30 = 33,3 \%$	$9/30 = 30,0 \%$	$11/30 = 36,7 \%$	100 %

Lecture

62,2 % des parcelles de Florido ont une bonne qualité, contre **33,3 %** pour Krenglè.
Le profil ligne révèle les **différences entre variétés**.

Graphique en barres empilées à 100 %



Ce que révèle le graphique

- ▶ Les **barres sont comparables** car chacune somme à 100 %
- ▶ Florido : profil très favorable (vert dominant)
- ▶ Krenglè : profil défavorable (rouge important)
- ▶ Gnidrou : position intermédiaire

Profil de référence

Si les trois variétés avaient la **même distribution** de qualité, les barres seraient identiques $\Rightarrow$ **indépendance**.

Profils lignes en R

```
# Profils lignes : prop.table avec margin=1
profil_lignes <- prop.table(tab, margin = 1)
round(profil_lignes * 100, 1)

#           qualite
# variete  Bonne Mauvaise Moyenne
# Florido  62.2    11.1    26.7
# Gnidrou  40.0    26.7    33.3
# Krengle  33.3    36.7    30.0

# Graphique barres empilées à 100%
library(ggplot2)
df <- as.data.frame(profil_lignes)
ggplot(df, aes(x=variete, y=Freq, fill=qualite)) +
  geom_bar(stat="identity") +
  labs(title="Profils lignes : qualite selon la variete",
       x="Variete", y="Proportion") +
  scale_y_continuous(labels=scales::percent)
```

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

Profil colonne de la modalité j

Distribution de X **parmi les individus** ayant la modalité j de Y :

$$f_{i|j} = \frac{n_{ij}}{n_{\bullet j}} \quad \text{avec} \quad \sum_i f_{i|j} = 1$$

On **conditionne** sur la colonne j : chaque colonne totalise 100 %.

Profils colonnes : fréquences conditionnelles par qualité

Variété	Bonne	Moyenne	Mauvaise	Marge (%)
Florido	$28/56 = 50,0 \%$	$12/36 = 33,3 \%$	$5/28 = 17,9 \%$	37,5 %
Gnidrou	$18/56 = 32,1 \%$	$15/36 = 41,7 \%$	$12/28 = 42,9 \%$	37,5 %
Krenglè	$10/56 = 17,9 \%$	$9/36 = 25,0 \%$	$11/28 = 39,3 \%$	25,0 %
Total	100 %	100 %	100 %	

Lecture

Parmi les parcelles à **bonne qualité**, 50 % viennent de Florido. Parmi les parcelles à **mauvaise qualité**, seulement 17,9 % viennent de Florido.

Profil ligne - Question

*“Comment se répartit la qualité **au sein de chaque variété** ?”*

On fixe la **variable en ligne** (X).
Utile quand X est la variable **explicative** (cause supposée).

Exemple : Florido produit-elle de meilleures qualités que Krenglè ?

Profil colonne - Question

“D’où proviennent les parcelles de chaque catégorie de qualité ?”

On fixe la **variable en colonne** (Y).
Utile quand Y est la variable **d’intérêt** (résultat).

Exemple : Quelle variété est la plus représentée dans les bonnes qualités ?

À retenir

Profil ligne $\neq$ profil colonne. Choisir celui qui répond à la question posée.
Si les profils lignes **sont tous identiques**, les variables sont **indépendantes**.

Profils colonnes en R

```
# Profils colonnes : prop.table avec margin=2
profil_colonnes ← prop.table(tab, margin = 2)
round(profil_colonnes * 100, 1)

#           qualite
# variete  Bonne Mauvaise Moyenne
# Florido  50.0    17.9    33.3
# Gnidrou  32.1    42.9    41.7
# Krengle  17.9    39.3    25.0

# Comparer avec le profil marginal (structure globale)
profil_marginal ← prop.table(margin.table(tab, 1))
round(profil_marginal * 100, 1)
# Florido 37.5% Gnidrou 37.5% Krengle 25.0%

# Si profil colonne = profil marginal => independance
```

Indépendance statistique

Deux variables qualitatives X et Y sont **indépendantes** si la distribution de Y ne dépend pas de la modalité de X :

$$f_{j|i} = f_{\bullet j} \quad \text{pour tout } i, j$$

Autrement dit : tous les profils lignes sont identiques et égaux au **profil marginal** de Y .

Effectifs théoriques sous indépendance :

$$\hat{n}_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

Effectifs théoriques (si indépendance parfaite)

	Bonne	Moy.	Mauv.
Florida	21,0	13,5	10,5
Gnidrou	21,0	13,5	10,5
Krenglè	14,0	9,0	7,0

Interprétation

Les effectifs **observés** s'écartent des effectifs **théoriques** : preuve d'une association. Le V de Cramer mesure cet écart.

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

V de Cramer

$$V = \sqrt{\frac{\chi^2}{n \times \min(r-1, c-1)}}$$

où :

- ▶ $\chi^2 = \sum_{i,j} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$ mesure l'écart global aux effectifs théoriques
- ▶ r = nombre de lignes, c = nombre de colonnes
- ▶ n = effectif total

Propriétés

- ▶ $0 \leq V \leq 1$
- ▶ $V = 0$: indépendance parfaite
- ▶ $V = 1$: association parfaite (déterministe)
- ▶ **Sans unité** : comparaison possible entre études
- ▶ Symétrique : $V(X, Y) = V(Y, X)$

Note pédagogique

χ^2 est ici un outil de calcul intermédiaire, pas une mesure d'association. C'est le **V** qui interprète.

Étape 1 : effectifs théoriques

$$\hat{n}_{ij} = n_{i\bullet} \times n_{\bullet j} / n$$

	Bonne	Moy.	Mauv.
Florido	21,0	13,5	10,5
Gnidrou	21,0	13,5	10,5
Krenglè	14,0	9,0	7,0

Étape 2 : calcul de χ^2

$$\chi^2 = \frac{(28 - 21)^2}{21} + \frac{(12 - 13,5)^2}{13,5} + \frac{(5 - 10,5)^2}{10,5} + \dots$$

En additionnant les 9 termes : $\chi^2 \approx 14,2$

Étape 3 : calcul du V

$r = 3$ lignes, $c = 3$ colonnes

$$\min(r - 1, c - 1) = \min(2, 2) = 2$$

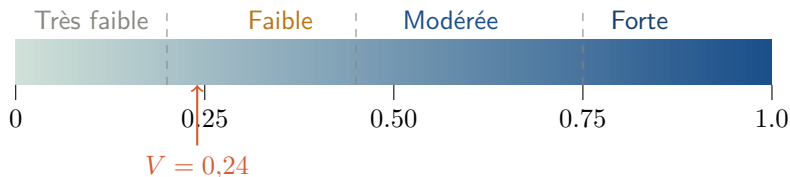
$$n = 120$$

$$V = \sqrt{\frac{14,2}{120 \times 2}} = \sqrt{\frac{14,2}{240}} = \sqrt{0,0592} \approx 0,24$$

Interprétation

$V \approx 0,24$: association **modérée** entre la variété et la qualité sanitaire. La variété influence la qualité, mais d'autres facteurs jouent aussi.

Interprétation du V de Cramer



V	Force de l'association	Interprétation agronomique
$< 0,10$	Négligeable	Les deux variables varient de façon quasi-indépendante
$0,10-0,30$	Faible à modérée	Association perceptible, mais partielle
$0,30-0,60$	Modérée à forte	La variable X prédit bien la distribution de Y
$> 0,60$	Forte à très forte	Association quasi-déterministe

Ces seuils sont des **conventions** (inspirées de Cohen, 1988), pas des règles absolues. L'interprétation doit toujours tenir compte du contexte.

```
# Chi-deux et V de Cramer
test ← chisq.test(tab, correct=FALSE)
chi2 ← test$statistic      # chi-deux = 14.2 (environ)
n ← sum(tab)               # 120
k ← min(nrow(tab)-1, ncol(tab)-1) # min(2,2) = 2
V_cramer ← sqrt(chi2 / (n * k))
cat("V de Cramer =", round(V_cramer, 3), "\n")
# V de Cramer = 0.243
# Avec le package vcd (plus direct)
# install.packages("vcd")
library(vcd)
assocstats(tab) # affiche V de Cramer et autres mesures
```

Important

La fonction `chisq.test()` est utilisée ici uniquement pour **calculer** le χ^2 intermédiaire. On n'interprète pas la p-valeur dans ce cours.

1. Pourquoi croiser deux variables qualitatives ?
2. Le tableau de contingence
3. Profils lignes
4. Profils colonnes
5. Le V de Cramer
6. Synthèse et boîte à outils

Tableau de contingence	Profils	V de Cramer
<i>Je veux...</i> voir les effectifs bruts de chaque combinaison <i>Limite</i> : ne tient pas compte des tailles de groupe	<i>Je veux...</i> comparer la distribution d'une variable selon les modalités de l'autre <i>Profil ligne</i> si X est la cause supposée ; <i>profil colonne</i> si Y est le résultat d'intérêt	<i>Je veux...</i> résumer en un seul chiffre la force du lien entre les deux variables <i>Limite</i> : ne dit pas dans quel sens va le lien, ni pourquoi

Démarche recommandée

1. Construire le tableau de contingence
2. Calculer et visualiser les profils lignes
3. Calculer le V de Cramer
4. Interpréter en contexte

Limites à garder en tête

- **Association $\neq$ causalité** (comme en bivarié quantitatif)
- Le V de Cramer mesure la force, pas la direction du lien
- Avec de très petits effectifs dans certaines cellules, les calculs

Compétences acquises - Module 8

- ▶ Construire et lire un **tableau de contingence**
- ▶ Calculer et interpréter des **profils lignes** et **profils colonnes**
- ▶ Distinguer les deux types de profils selon la question posée
- ▶ Identifier le **profil d'indépendance** (profil marginal)
- ▶ Calculer et interpréter le **V de Cramer**
- ▶ Produire l'analyse complète en **R**

Vue d'ensemble du cours

M1	Introduction
M2	Échantillonnage
M3	Tableaux & graphiques
M4	Tendance centrale
M5	Dispersion
M6	Forme de la distribution
M7	Bivarié quant. $\times$ quant.
M8	Bivarié qual. $\times$ qual.

“Les données agricoles sont riches en variables catégorielles : sol, variété, région, stade. Savoir les croiser, c'est savoir lire le terrain.”

1. Une étude croise le **type de semence** (certifiée / fermière / sauvage) et la **résistance à la sécheresse** (forte / moyenne / faible) sur 150 champs.
 - ▶ Quel outil utilisez-vous pour comparer la distribution de la résistance selon le type de semence ?
 - ▶ Que signifie $V = 0,05$ dans ce contexte ?
 - ▶ Si les profils lignes sont tous identiques, que peut-on conclure ?
2. Comparez les profils **lignes** et **colonnes** : dans quelle situation préférez-vous l'un à l'autre ?
3. Le V de Cramer vaut 0,45 entre la région de production et la catégorie de teneur en sucre du manioc. **Que dit cela, et que ne dit-il pas ?**

Merci de votre attention

Statistiques Descriptives · Module 8

Equipe Bioinformatique et Biostatistique

Laboratoire de Microbiologie, Biotechnologie et Bioinformatique

Institut National Polytechnique Félix Houphouët-Boigny

