Analyse bivariée (partie 2)

Vincent Audigier

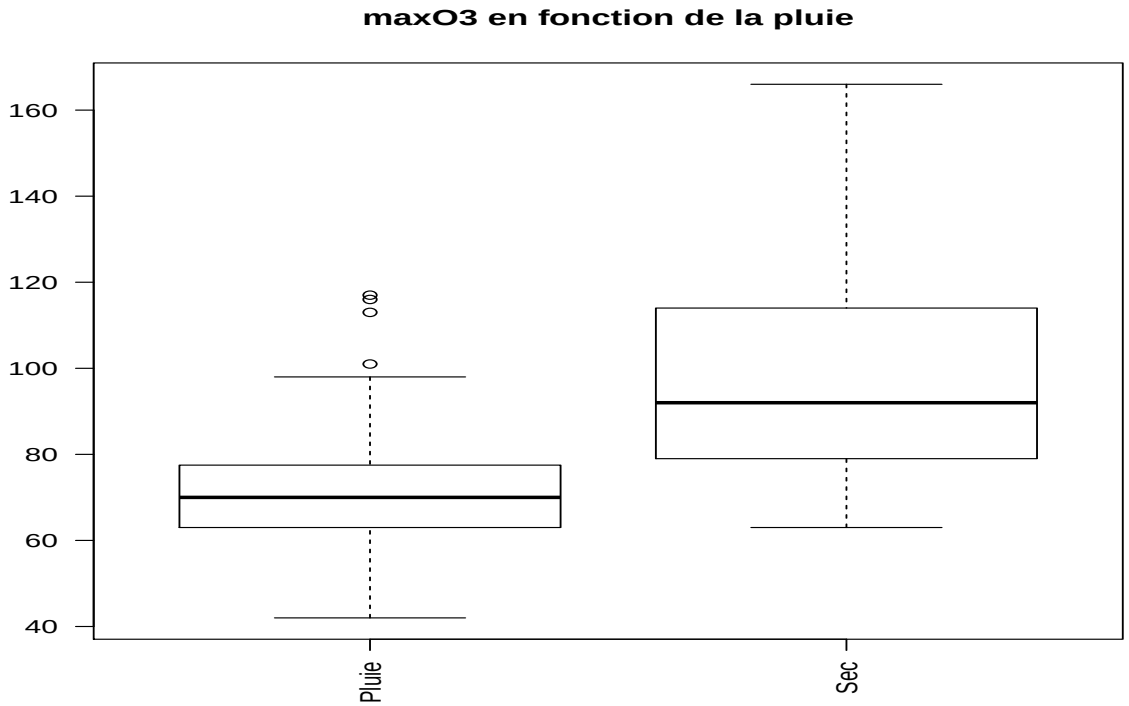
CNAM, Paris

STA101

Plan

Liaison entre une variable quantitative et qualitative

- Lien entre la valeur du pic d'O3 (quantitative) et la présence de pluie (qualitative)?



- Peut-on le quantifier ?
- Ce lien apparent est-il significatif ?

Formules de décomposition

- On note Y la variable quantitative (Ex : maxO3), X la variable qualitative (Ex : présence de pluie) à K modalités

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{k=1}^{K} n_k \bar{y}_k \\ s_Y^2 &= \underbrace{\frac{1}{n}\sum_{k=1}^{K} n_k (\bar{y}_k - \bar{y})^2}_{\text{var inter}} + \underbrace{\frac{1}{n}\sum_{k=1}^{K} n_k s_k^2}_{\text{var intra}} \end{aligned}$$

avec

- K le nombre de groupes (i.e. nb de modalités de la variable quali)
- n_k l'effectif associé au groupe k
- $\bar{y}_k$ moyenne de y dans le groupe k
- s_k^2 la variance de y dans le groupe k

Démonstration

$$s_Y^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$$

On note y_{kj} la valeur de Y pour l'individu j du groupe k.

$$(y_{kj} - \bar{y}) = (\bar{y}_k - \bar{y}) + (y_{kj} - \bar{y}_k)$$

Démonstration

$$s_Y^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$$

On note y_{kj} la valeur de Y pour l'individu j du groupe k.

$$\begin{aligned}
(y_{kj} - \bar{y}) &= (\bar{y}_k - \bar{y}) + (y_{kj} - \bar{y}_k) \\
(y_{kj} - \bar{y})^2 &= (\bar{y}_k - \bar{y})^2 + (y_{kj} - \bar{y}_k)^2 + 2(y_k - \bar{y})(y_{kj} - \bar{y}_k)
\end{aligned}$$

Démonstration

$$s_Y^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$$

On note y_{kj} la valeur de Y pour l'individu j du groupe k.

$$\begin{aligned}
(y_{kj} - \bar{y}) &= (\bar{y}_k - \bar{y}) + (y_{kj} - \bar{y}_k) \\
(y_{kj} - \bar{y})^2 &= (\bar{y}_k - \bar{y})^2 + (y_{kj} - \bar{y}_k)^2 + 2(y_k - \bar{y})(y_{kj} - \bar{y}_k) \\
\sum_k \sum_j (y_{kj} - \bar{y})^2 &= \sum_k \sum_j (\bar{y}_k - \bar{y})^2 + \sum_k \sum_j (y_{kj} - \bar{y}_k)^2 \\
&\quad + 2\sum_k \sum_j (\bar{y}_k - \bar{y})(y_{kj} - \bar{y}_k)
\end{aligned}$$

Démonstration

$$s_Y^2 = \frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \frac{1}{n}\sum_{k=1}^{K} n_k s_k^2$$

On note y_{kj} la valeur de Y pour l'individu j du groupe k.

$$\begin{aligned}
(y_{kj} - \bar{y}) &= (\bar{y}_k - \bar{y}) + (y_{kj} - \bar{y}_k) \\
(y_{kj} - \bar{y})^2 &= (\bar{y}_k - \bar{y})^2 + (y_{kj} - \bar{y}_k)^2 + 2(y_k - \bar{y})(y_{kj} - \bar{y}_k) \\
\sum_k \sum_j (y_{kj} - \bar{y})^2 &= \sum_k \sum_j (\bar{y}_k - \bar{y})^2 + \sum_k \sum_j (y_{kj} - \bar{y}_k)^2 \\
&\quad + 2\sum_k \sum_j (\bar{y}_k - \bar{y})(y_{kj} - \bar{y}_k) \\
ns_Y^2 &= \sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 + \sum_{k=1}^{K} n_k s_k^2 \\
&\quad + 2\sum_k (\bar{y}_k - \bar{y}) \sum_j (y_{kj} - \bar{y}_k)
\end{aligned}$$

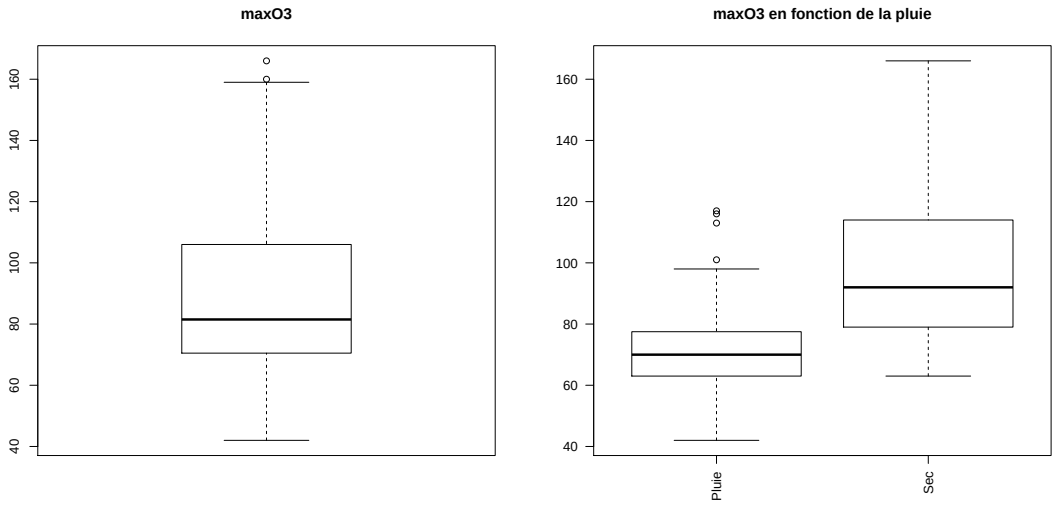
Rapport de corrélation

- Principe : s'il existe un lien, alors la valeur du pic d'O3 n'est pas la même selon qu'il pleuve ou non. On compare les variations des valeurs des pics d'O3 d'un groupe à l'autre à la variation totale des pics d'O3.
- On calcule

$$\eta^2(Y,X) = \frac{\textit{var inter}}{\textit{var totale}} = \frac{\frac{1}{n}\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Propriétés :
  - $0 \leq \eta^2 \leq 1$
  - $\eta^2 = 0$ traduit une absence de lien
  - $\eta^2 = 1$ traduit un lien parfait

Exemple



	$\bar{y}$	n
	90.30	112

Table: maxO3

	$\bar{y}_k$	n_k
Pluie	73.40	43
Sec	100.84	69

Table: maxO3 selon pluie

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 88191.68$$

$$\sum_{k=1}^{K} n_k(\bar{y}_k - \bar{y})^2 = 43 \times (73.40 - 90.30)^2 + 69 \times (100.84 - 90.30)^2$$

$$= 19954.15$$

$$\eta^2 = \frac{\textit{var inter}}{\textit{var totale}} = \frac{19954.15}{88191.68} = 0.226$$

Test de nullité

- le rapport de corrélation varie selon l'échantillon, mais la liaison entre deux variables ne varie que par les variables considérées
- le rapport de corrélation calculé à partir des données est une version empirique d'un coefficient théorique
- Test de Fisher

NB : par la suite on note η^2 le coefficient théorique, et $\hat{\eta}^2$ son estimation

Test de Fisher

- H_0: $\eta^2 = 0$ contre H_1: $\eta^2 \neq 0$

$$F = \frac{\hat{\eta}^2 \times (n-K)}{(K-1)(1-\hat{\eta}^2)} \sim_{H_0} Fisher^{\nu_1=K-1}_{\nu_2=n-K}$$

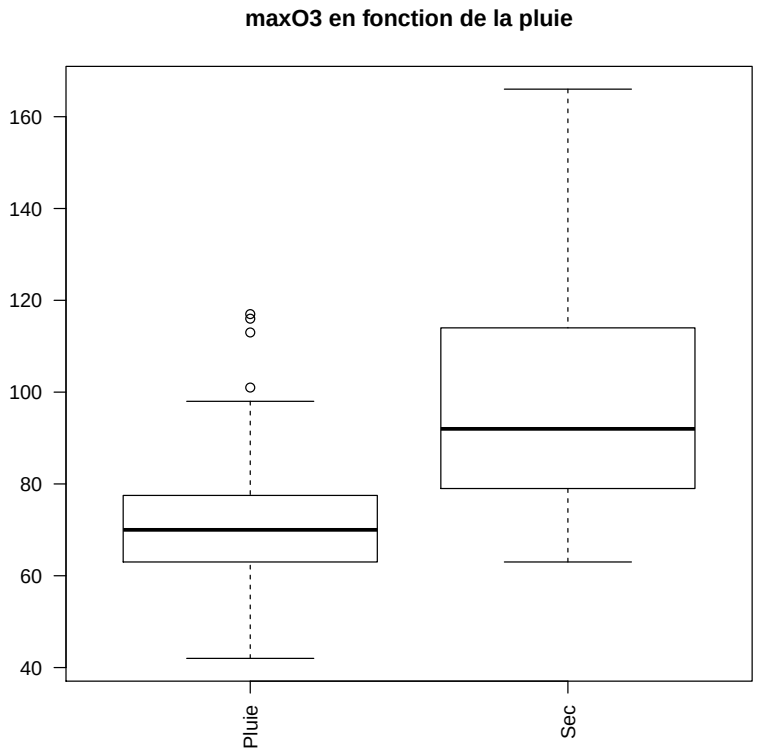
- Test unilatéral
  - on calcule F^{obs} que l'on compare à $q^{K-1}_{n-K}(1-\alpha)$
  - si $F^{obs} < q^{K-1}_{n-K}(1-\alpha)$, on ne rejette pas H_0 (au risque α)
    On ne peut pas dire qu'il y ait une liaison entre X et Y
  - si $F^{obs} \geq q^{K-1}_{n-K}(1-\alpha)$, on rejette H_0 (au risque α)
    On conclut que X et Y sont liées

- **Attention !** Un test significatif ne signifie pas une association forte.

Limites

- La distribution sous l'hypothèse nulle n'est vraie que si les variances dans les différents groupes sont identiques
- En pratique, on compare graphiquement les distributions des données dans chaque groupe
- Quand cette hypothèse n'est pas raisonnable, il faut utiliser d'autres tests (e.g. Test de Welch si $K = 2$)

Exemple



- On observe $\eta^2 = 0.226$
- Mais les variances au sein des groupes ne sont pas identiques
- Test de Welch (p.val$< 10^{-6}$)

Plan

Analyse de la liaison entre deux variables qualitatives

- X_1 et X_2 deux variables à K et K' modalités
- Présentation sous la forme d'une **table de contingence**
- Exemple : direction du vent et présence de pluie

	Est	Nord	Ouest	Sud	$n_{i.}$
Pluie	2	10	26	5	43
Sec	8	21	24	16	69
$n_{.j}$	10	31	50	21	112

Table: Table de contingence

- n_{ij} est l'effectif des individus tels que $X_1 = m_i$ et $X_2 = m_j$
- $n_{i.} = \sum_j n_{ij}$ **total marginal colonne** $n_{.j} = \sum_i n_{ij}$ **total marginal ligne**
- Comment mesurer une liaison entre ces deux variables ?

Profils-lignes, profils-colonnes

On appelle **profils-lignes** (resp. prof.-**colonnes**) le tableau des fréquences conditionnelles $\frac{n_{ij}}{n_{i.}}$ (resp. $\frac{n_{ij}}{n_{.j}}$)

	Est	Nord	Ouest	Sud	$n_{i.}$
Pluie	2	10	26	5	43
Sec	8	21	24	16	69
$n_{.j}$	10	31	50	21	112

Table: Table de contingence

	Est	Nord	Ouest	Sud
Pluie	0.2	0.3	0.5	0.2
Sec	0.8	0.7	0.5	0.8
Somme	1.0	1.0	1.0	1.0

Table: Profils colonnes

	Est	Nord	Ouest	Sud	
Pluie	0.0	0.2	0.6	0.1	1
Sec	0.1	0.3	0.3	0.2	1

Table: Profils lignes

- Absence de lien $\Rightarrow$ profils lignes (resp. colonnes) identiques (égales aux fréquences marginales)

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{n} \text{ pour tout } i$$

Représentation graphique

- visualiser la liaison = visualiser les différences des profils
- graphique en mosaïque (*mosaic plot*)

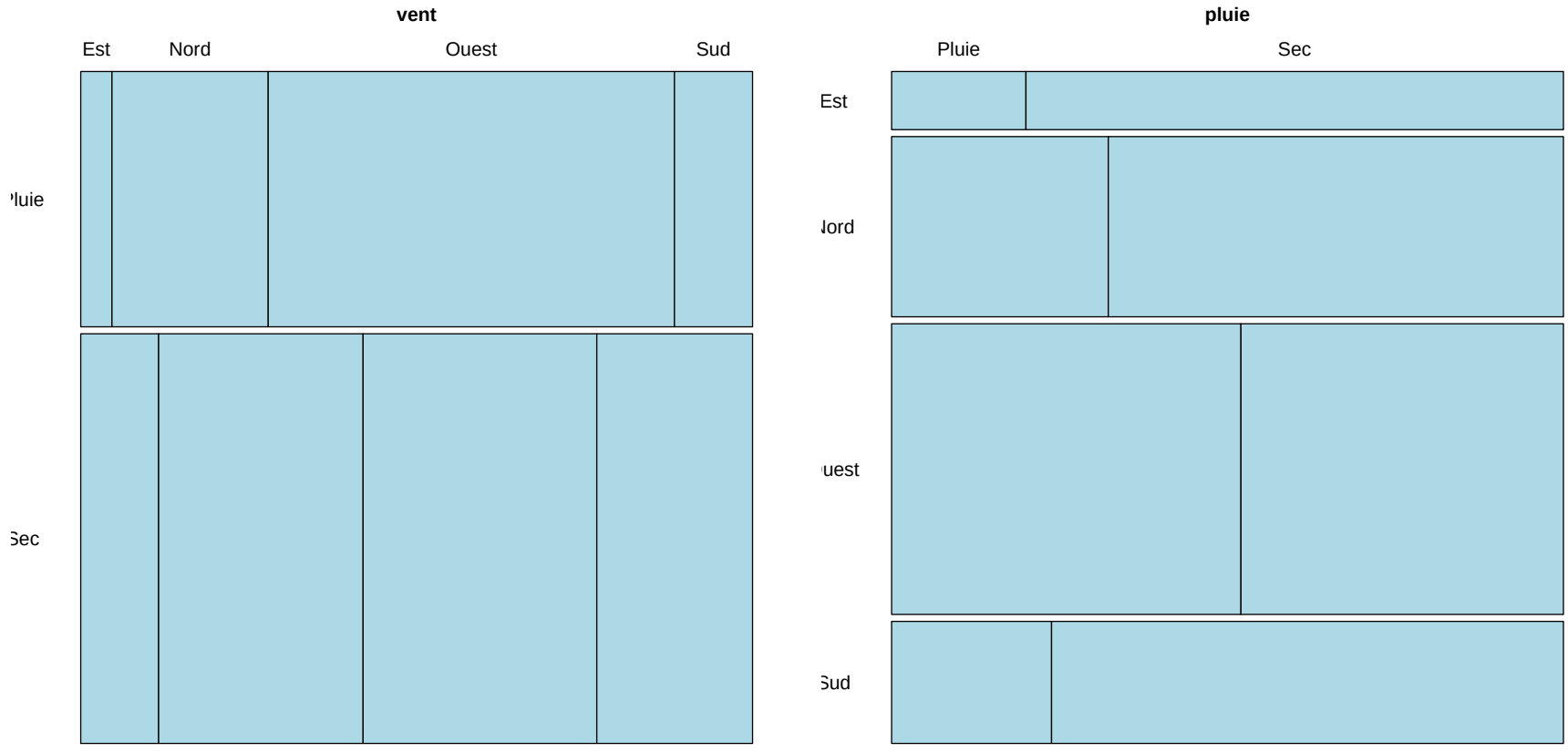


Figure: mosaic plots des profils-lignes et profils-colonnes

Khi-deux

- Mesurer la liaison entre deux variables qualitatives, en regardant la différence entre les effectifs conjoints observés et ceux attendus sous l'hypothèse d'indépendance

$$
\begin{aligned}
\chi^2 &= \sum_{(i,j)} \frac{(\text{effectifs observés} - \text{effectifs attendus})^2}{\text{effectifs attendus}} \\
&= \sum_{(i,j)} \frac{\left(n_{ij} - \left(\frac{n_{.j}}{n} \times \frac{n_{i.}}{n} \times n\right)\right)^2}{\left(\frac{n_{.j}}{n} \times \frac{n_{i.}}{n} \times n\right)} \\
&= \sum_{(i,j)} \frac{(n_{ij} - (n_{.j} \times n_{i.}/n))^2}{(n_{.j} \times n_{i.}/n)}
\end{aligned}
$$

- Différence faible $\rightarrow$ absence de liaison
- Différence grande $\rightarrow$ liaison forte

Exemple

	Est	Nord	Ouest	Sud	Somme
Pluie	2	10	26	5	43
Sec	8	21	24	16	69
Somme	10	31	50	21	112

Table: Effectifs observés

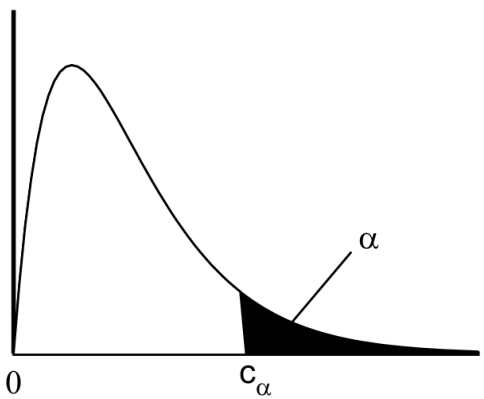
	Est	Nord	Ouest	Sud	Somme
Pluie	3.84	11.90	19.20	8.06	43
Sec	6.16	19.10	30.80	12.94	69
Somme	10	31	50	21	112

Table: Effectifs attendus sous l'hypothèse d'indépendance

$$\begin{aligned}\chi^2 &= \frac{(2-3.84)^2}{3.84} \\ &+ \frac{(10-11.90)^2}{11.90} \\ &+ \dots \\ &+ \frac{(16-12.94)^2}{12.94} \\ &= 7.73\end{aligned}$$

Interprétation

- Le critère du chi-deux traduit l'écart à l'indépendance
- Toujours supérieur à 0, mais n'est pas toujours inférieur à 1
- Pour déterminer si la valeur observée est grande ou non, on utilise des arguments statistiques
  - sous l'hypothèse d'indépendance, on s'attend à ce que la distribution du critère soit une loi du χ^2 à $(K-1)\times(K'-1)$ degrés de liberté (si effectifs attendus ≥ 5)
  - calculer la probabilité (sous l'hypothèse d'indépendance) d'observer une distance au moins aussi grande que celle observée sur les données



α / ν	0,9	0,5	0,3	0,2	0,1	0,05
1	0,016	0,455	1,074	1,642	2,706	3,841
2	0,211	1,386	2,408	3,219	4,605	5,991
3	0,584	2,366	3,665	4,642	6,251	7,815
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Autres coefficients

- Il existe d'autres coefficients basés sur le χ^2 qui visent à le normaliser entre 0 et 1

- Par exemple : le coefficient C de Cramer et le T de Tschuprow

$$C = \sqrt{\frac{\chi^2/n}{(min(K, K') - 1)}}$$

$$T = \sqrt{\frac{\chi^2/n}{\sqrt{(K-1)(K'-1)}}}$$

- $0 \leq T \leq C \leq 1$

Ex :

$$C = \sqrt{\frac{7.73/112}{(min(2,4)-1)}} = 0.26 \; T = \sqrt{\frac{7.73/112}{\sqrt{1 \times 3}}} = 0.20$$

Test d'indépendance

- H_0: indépendance vs H_1: non-indépendance
- Statistique de test : χ^2
- Loi sous H_0 : loi du chi-deux à $(K-1)(K'-1)$ ddl (si effectifs attendus ≥ 5)
- zone de rejet : $[c_{1-\alpha}; +\infty[$ avec $c_{1-\alpha}$ le quantile d'ordre $1-\alpha$ de la loi du chi-deux

Plan

Matrices des covariances

Pour des variables **quantitatives** $p \geq 3$, on peut calculer

- toutes les variances des variables
- toutes les covariances entre les couples de variables
- l'ensemble peut se présenter sous la forme d'une matrice $V_{p \times p}$
  - avec les variances de chaque variable sur la diagonale
  - les covariances sur les élément hors diagonale
- V est dite matrice de variance-covariance

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
maxO3	794.5	61.6	89.4	98.9	-45.5	-41.2	-31.4	39.2	33.9	31.0	545.6
T9	61.6	9.8	11.1	12.0	-3.9	-3.4	-2.4	2.1	1.9	1.5	51.4
T12	89.4	11.1	16.3	17.3	-6.1	-6.1	-4.3	4.6	3.5	3.1	64.4
T15	98.9	12.0	17.3	20.5	-6.9	-6.7	-6.1	5.4	4.4	3.6	72.8
Ne9	-45.5	-3.9	-6.1	-6.9	6.7	4.7	3.3	-3.4	-3.8	-3.6	-20.3
Ne12	-41.2	-3.4	-6.1	-6.7	4.7	5.2	3.8	-3.0	-3.3	-2.8	-23.4
Ne15	-31.4	-2.4	-4.3	-6.1	3.3	3.8	5.4	-2.5	-2.8	-2.5	-20.3
Vx9	39.2	2.1	4.6	5.4	-3.4	-3.0	-2.5	6.9	5.5	5.0	25.3
Vx12	33.9	1.9	3.5	4.4	-3.8	-3.3	-2.8	5.5	7.8	6.6	17.7
Vx15	31.0	1.5	3.1	3.6	-3.6	-2.8	-2.5	5.0	6.6	7.9	15.1
maxO3v	545.6	51.4	64.4	72.8	-20.3	-23.4	-20.3	25.3	17.7	15.1	799.6

Table: matrice de variance-covariance pour le jeu ozone

Matrices des corrélations

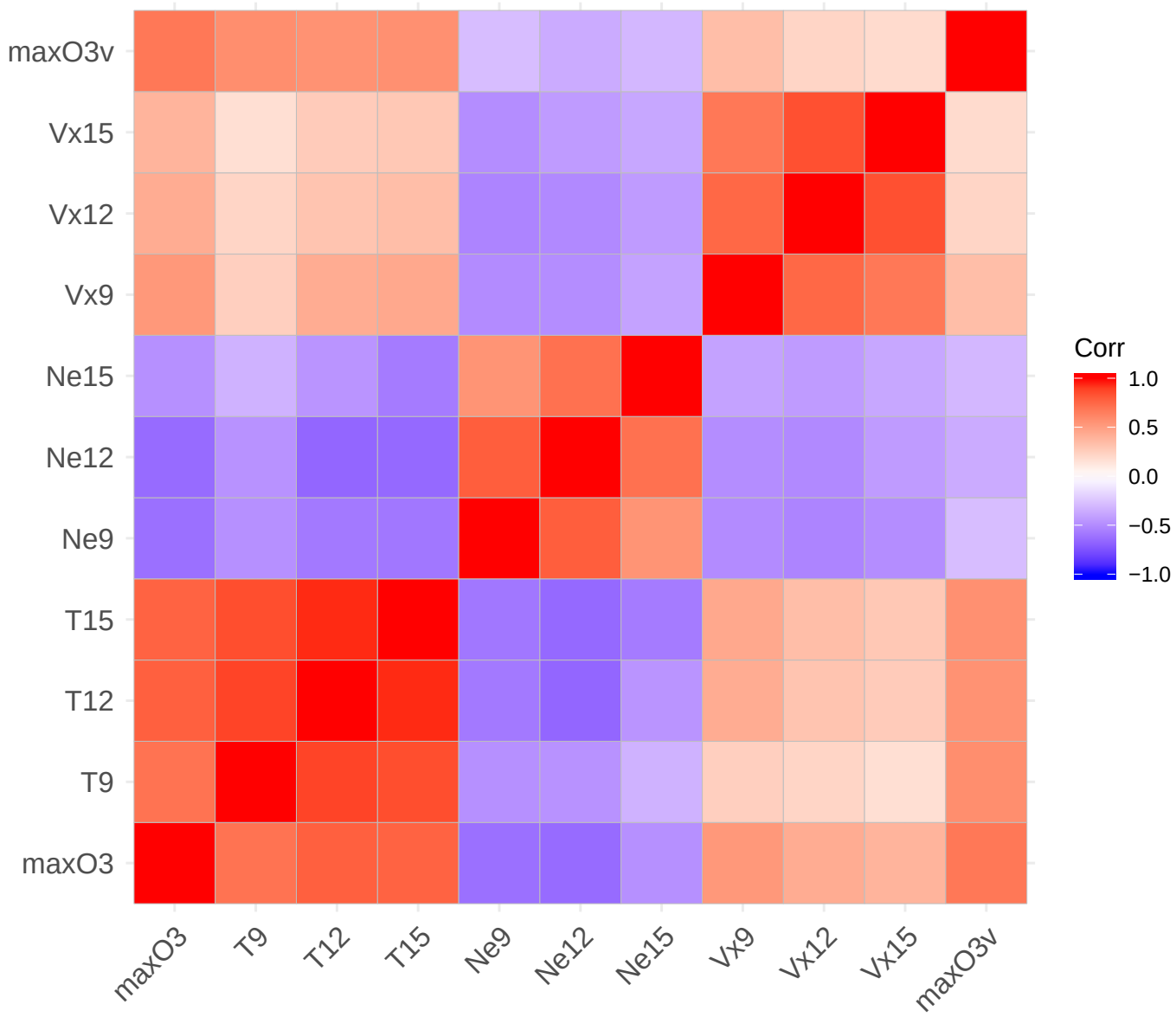
Pour des variables **quantitatives** $p \geq 3$, on peut calculer

- toutes les corrélations entre les couples de variables
- l'ensemble peut se présenter sous la forme d'une matrice $R_{p \times p}$ appelée matrice des corrélations
- utile pour l'interprétation

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15	maxO3v
maxO3	1.0	0.7	0.8	0.8	-0.6	-0.6	-0.5	0.5	0.4	0.4	0.7
T9	0.7	1.0	0.9	0.8	-0.5	-0.5	-0.3	0.3	0.2	0.2	0.6
T12	0.8	0.9	1.0	0.9	-0.6	-0.7	-0.5	0.4	0.3	0.3	0.6
T15	0.8	0.8	0.9	1.0	-0.6	-0.6	-0.6	0.5	0.3	0.3	0.6
Ne9	-0.6	-0.5	-0.6	-0.6	1.0	0.8	0.6	-0.5	-0.5	-0.5	-0.3
Ne12	-0.6	-0.5	-0.7	-0.6	0.8	1.0	0.7	-0.5	-0.5	-0.4	-0.4
Ne15	-0.5	-0.3	-0.5	-0.6	0.6	0.7	1.0	-0.4	-0.4	-0.4	-0.3
Vx9	0.5	0.3	0.4	0.5	-0.5	-0.5	-0.4	1.0	0.8	0.7	0.3
Vx12	0.4	0.2	0.3	0.3	-0.5	-0.5	-0.4	0.8	1.0	0.8	0.2
Vx15	0.4	0.2	0.3	0.3	-0.5	-0.4	-0.4	0.7	0.8	1.0	0.2
maxO3v	0.7	0.6	0.6	0.6	-0.3	-0.4	-0.3	0.3	0.2	0.2	1.0

Table: matrice des corrélations pour le jeu ozone

Représentation graphique



Matrice des coefficients de Cramer

Pour des variables **qualitatives** $p \geq 3$, on peut calculer

- les coefficients de Cramer entre ces différentes variables
- l'ensemble peut se présenter sous la forme d'une matrice
  - symétrique
  - avec des 1 sur la diagonale
  - les coefficients de Cramer pour chaque couple hors de la diagonale
- on peut faire de même avec les coefficients de Tschuprow

Tableau de Burt

Pour *p* variables **qualitatives**

- Le tableau de Burt généralise le tableau de contingence
- C'est la concaténation des différents tableaux de contingence entre chaque couple de variables
- Exemple à 4 variables

Tableau de Burt :

	Q1-1	Q1-2	Q1-3	Q2-1	Q2-2	Q3-1	Q3-2	Q4-1	Q4-2	Q4-3
Q1-1	8	0	0	4	4	3	5	5	2	1
Q1-2	0	8	0	6	2	7	1	3	5	0
Q1-3	0	0	10	4	6	2	8	1	4	5
Q2-1	4	6	4	14	0	10	4	6	6	2
Q2-2	4	2	6	0	12	2	10	3	5	4
Q3-1	3	7	2	10	2	12	0	5	6	1
Q3-2	5	1	8	4	10	0	14	4	5	5
Q4-1	5	3	1	6	3	5	4	9	0	0
Q4-2	2	5	4	6	5	6	5	0	11	0
Q4-3	1	0	5	2	4	1	5	0	0	6

Conclusion

- L'analyse bivariée renseigne sur les liaisons entre variables deux à deux
- Elle peut être représentée graphiquement ou sous forme d'indicateurs
- Le choix de ces indicateurs est propre à la nature des variables considérées
- Elle permet également d'identifier des valeurs aberrantes qui n'auraient pas pu être identifiées par une analyse univariée
- Limite de l'approche : le nombre de couples de variables explose quand le nombre de variables augmente