Rappels : Analyse statistique pour des variables quantitatives et qualitatives

Master 2 Recherche IES

Ana Karina Fermin

Université Paris-Ouest-Nanterre-La Défense

http://fermin.perso.math.cnrs.fr/

Objectifs du cours

- Présenter les méthodes statistiques de traitement des données comportant des variables qualitatives.
- Traiter à la fois des problèmes de régressions et des problèmes de classification avec un accent sur les méthodes dites linéaires.
- Proposer d'autres types de méthodes : classification supervisée et non supervisée.

Évaluation : Un examen final.

Remarques importantes

- Ce cours n'est pas un cours de statistiques.
- Nous supposons que vous avez déjà une connaissance de certaines méthodes présentées ici.
- Si vous souhaitez des précisions théoriques/méthodologiques à propos d'un certain type d'analyses, nous vous conseillons de voir la doc !

Thèmes abordés dans ce cours

- Introduction : rappels, analyse statistique pour des variables qualitatives et quantitatives.
- Régression: rappel, codage des variables qualitatives, diagnostique des résidus et validation du modèle.
- ANOVA. Sélection de modèles pour la régression.
- Régression *linéaire* pour des variables qualitatives à deux modalités: les modèles logit, probit.
- Régression logistique multiple, estimation des paramètres. Sélection de modèles.
- Classification supervisée à l'aide du modèle logistique.
- Autres méthodes de régression et de classification supervisée
- Classification non supervisée

Données (data, échantillon)

les données proviennent d'une ou plusieurs variables ou caractères qui sont mesurés simultanément sur un individu. Cet individu appartient à une population $\mathcal{P}$ de taille N (inconnue). On dispose d'un échantillon de taille n

Exemple

- Population : Étudiants de M2 IES de Paris Ouest et EHESS.
- Variables : Série du baccalauréat (X_1), Age (X_2), Sexe (X_3), Type de licence (X_4), Note de licence (X_5), Durée du trajet domicile-université (X_6).
- On dispose d'un échantillon de taille n noté

$$D_n = \{x_1, x_2, \ldots, x_n\}$$

avec $x_i = (x_{i1}, x_{i2}, \ldots, x_{i6})$ le i-ème individu ($i = 1, \ldots, n$).

Les données ozone

Les 13 variables observées sont :

- MaxO3 : Maximum de concentration d'ozone observé sur la journée (en gr/m3) mesurées chaque jour pendant 3 mois d'été à Rennes
- T9, T12, T15 : Température observée à 9, 12 et 15h
- Ne9, Ne12, Ne15 : Nébulosité observée à 9, 12 et 15h
- Vx9, Vx12, Vx15 : Composante E-O du vent à 9, 12 et 15h
- MaxO3v : Teneur maximum en ozone observée la veille
- vent: orientation du vent à 12h
- pluie : occurrence ou non de précipitations

On dispose d'un échantillon de taille $n = 112$.

Type de variable

Fichier ozone.txt: (disponibles sur ma page web)

```
Observations: 112
Variables:
$ maxO3  (int) 87, 82, 92, 114, 94, 80, 79, 79, 101, 106, 101, 90, 72, 70, 83,
$ T9     (dbl) 15.6, 17.0, 15.3, 16.2, 17.4, 17.7, 16.8, 14.9, 16.1, 18.3, 17.3
$ T12    (dbl) 18.5, 18.4, 17.6, 19.7, 20.5, 19.8, 15.6, 17.5, 19.6, 21.9, 19.3
$ T15    (dbl) 18.4, 17.7, 19.5, 22.5, 20.4, 18.3, 14.9, 18.9, 21.4, 22.9, 20.2
$ Ne9    (int) 4, 5, 2, 1, 8, 6, 7, 5, 2, 5, 7, 7, 7, 7, 8, 6, 0, 8, 2, 1, 1, 0
$ Ne12   (int) 4, 5, 5, 1, 8, 6, 8, 5, 4, 6, 7, 6, 5, 7, 7, 5, 1, 3, 1, 1, 0, 0
$ Ne15   (int) 8, 7, 4, 0, 7, 7, 8, 4, 4, 8, 3, 8, 6, 7, 7, 4, 1, 1, 0, 2, 0, 0
$ Vx9    (dbl) 0.6946, -4.3301, 2.9544, 0.9848, -0.5000, -5.6382, -4.3301, 0.00
$ Vx12   (dbl) -1.7101, -4.0000, 1.8794, 0.3473, -2.9544, -5.0000, -1.8794, -1.
$ Vx15   (dbl) -0.6946, -3.0000, 0.5209, -0.1736, -4.3301, -6.0000, -3.7588, -1
$ maxO3v (int) 84, 87, 82, 92, 114, 94, 80, 99, 79, 101, 106, 101, 90, 72, 70,
$ vent   (fctr) Nord, Nord, Est, Nord, Ouest, Ouest, Ouest, Nord, Nord, Ouest,
$ pluie  (fctr) Sec, Sec, Sec, Sec, Sec, Pluie, Sec, Sec, Sec, Sec, Sec, Sec, S
```

1. Statistique Descriptive (résumés numériques, méthodes exploratoires et représentation graphique)
   - Variable quantitative
     - Résumés numériques : moyenne empirique, variance et écart-type, min, max, quantiles, ...
     - Graphiques : Histogrammes, boite à moustache, ...
   - Variable qualitative
     - Résumés numériques : Tableaux de proportions, ...
     - Représentation graphique : Diagramme en tuyaux d'orgue, ...
2. Statistique Inférentielle : test d'hypothèses , estimation, modélisation statistique, ...
3. Étude des variables quantitatives et qualitatives

Analyser, interpréter et mettre en forme ses données

- La question ici est de comment exploiter l'ensemble des données recueillies au cours de la recherche ?
- Comment faire le lien entre l'ensemble de ces données ?
- Quel est le problème à traiter ?

Questions du jour (partie 1)

- Résumer les variables quantitatives du jeu de données
- Représenter la variable Ozone.
- Utiliser la variable ozone. Visualiser les QQ-plots, puis tester à l'aide des tests de Kolmogorov-Smirnov et de Shapiro-Wilks si il s'agit d'un échantillon Gaussien.
- Représenter le nuage de points de la variable Ozone en fonction de la Température à 12h. Un lien semble-t-il présent?
- Calculer les corrélations entre toutes les variables.

Analyser, interpréter et mettre en forme ses données

Questions du jour (partie 2)

- Résumer les variables qualitatives du jeu de données.
- Traiter la variable qualitative pluie.
- Croiser les variables qualitatives pluie et vent. Tableaux de contingence.
- Un lien semble-t-il présent entre ces deux variables ?
- Test chi-deux .

Test d'hypothèses (rappels)

Hypothèses : H_0 et H_1

- Un test statistique est une méthode statistique permettant de d'infirmer une hypothèse formulée sur la population.
- Un test oppose deux hypothèses : l'hypothèse nulle H_0 et l'hypothèse alternative H_1.
- A l'issue du test, on va décider de rejeter ou pas H_0. Quelle que soit la décision on peut se tromper.

Risques d'erreur α et β

- Dans un problème de décision, deux types d'erreur sont possibles :
  - Erreur de première espèce (α) : est l'erreur commise lorsqu'on décide de rejeter H_0 alors que celle-ci est vrai (la probabilité d'avoir un faux-positif).
  - Erreur de deuxième espèce (β) : est l'erreur commise lorsqu'on décide de ne pas rejeter H_0 alors que celle-ci est fausse (la probabilité d'avoir un faux-négatif).
- Ces deux risques varient en sens inverse: quand l'un diminue, l'autre augmente.
- On décide alors arbitrairement de privilégier l'hypothèse nulle H_0 en fixant α petit. En général $\alpha = 1\%,\ 5\%$ ou 10%.
- Quant au risque d'erreur β en général, il n'est pas calculable sauf dans des cas particuliers de l'expression de l'hypothèse H_1.

Valeur critique ou p-valeur

- L'usage ancien des tables statistiques donnant les quantiles des différentes lois usuelles n'a plus lieu d'être avec la pratique d'un logiciel statistique. En effet, ceux-ci fournissent directement la probabilité critique ou p-valeur (en anglais p-value) associée à un test donné.
- Il suffit de comparer la p-valeur fournit avec le seuil ou niveau de test α fixé.
- Plus la p-valeur est proche de 0, plus forte est la contradiction entre H_0 et le résultat observé avec l'échantillon.

Critère de décision basée sur la p-valeur:

On rejette l'hypothèse nulle H_0 si p-valeur $\leqslant \alpha$.

Choix du Test

Le choix du test et guidé par la question posée et la structure des données issues de l'expérience.

- Test paramétriques :
  - Souvent les observations sont supposées suivre un modèle gaussien
  - L'échantillon est de suffisamment grande taille pour accepter la normalité asymptotique par le théorème centrale limite.
- Test non paramétriques.
  - Petit échantillon.
  - Distribution non gaussienne. Pas d'hypothèse sur la forme des distributions !

Remarque : Lorsque les hypothèses d'un test paramétrique sont vérifiées, un test non-paramétrique est généralement moins puissant que un test paramétrique.

Une variables quantitative X

On dispose d'un échantillon de taille n de X issu de la population $\mathcal{P}$

$$\{x_1, x_2, \ldots, x_n\}$$

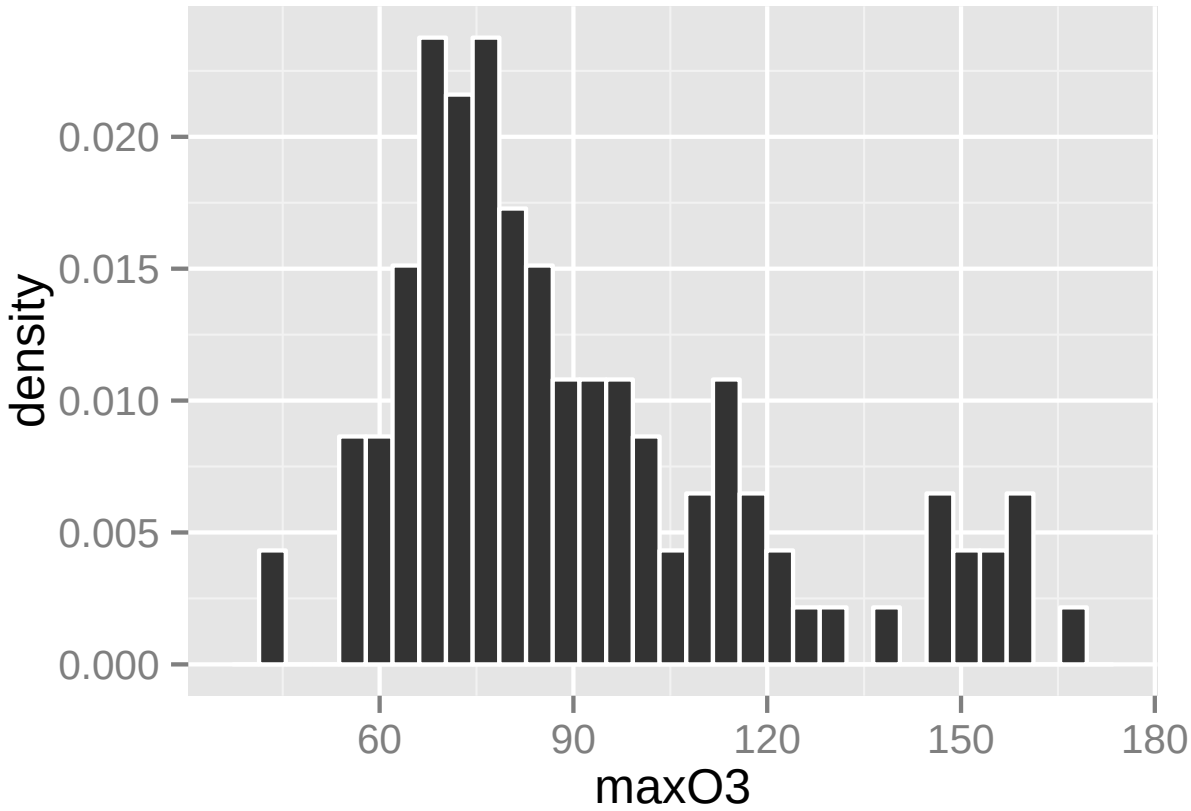
- Moyenne observée

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

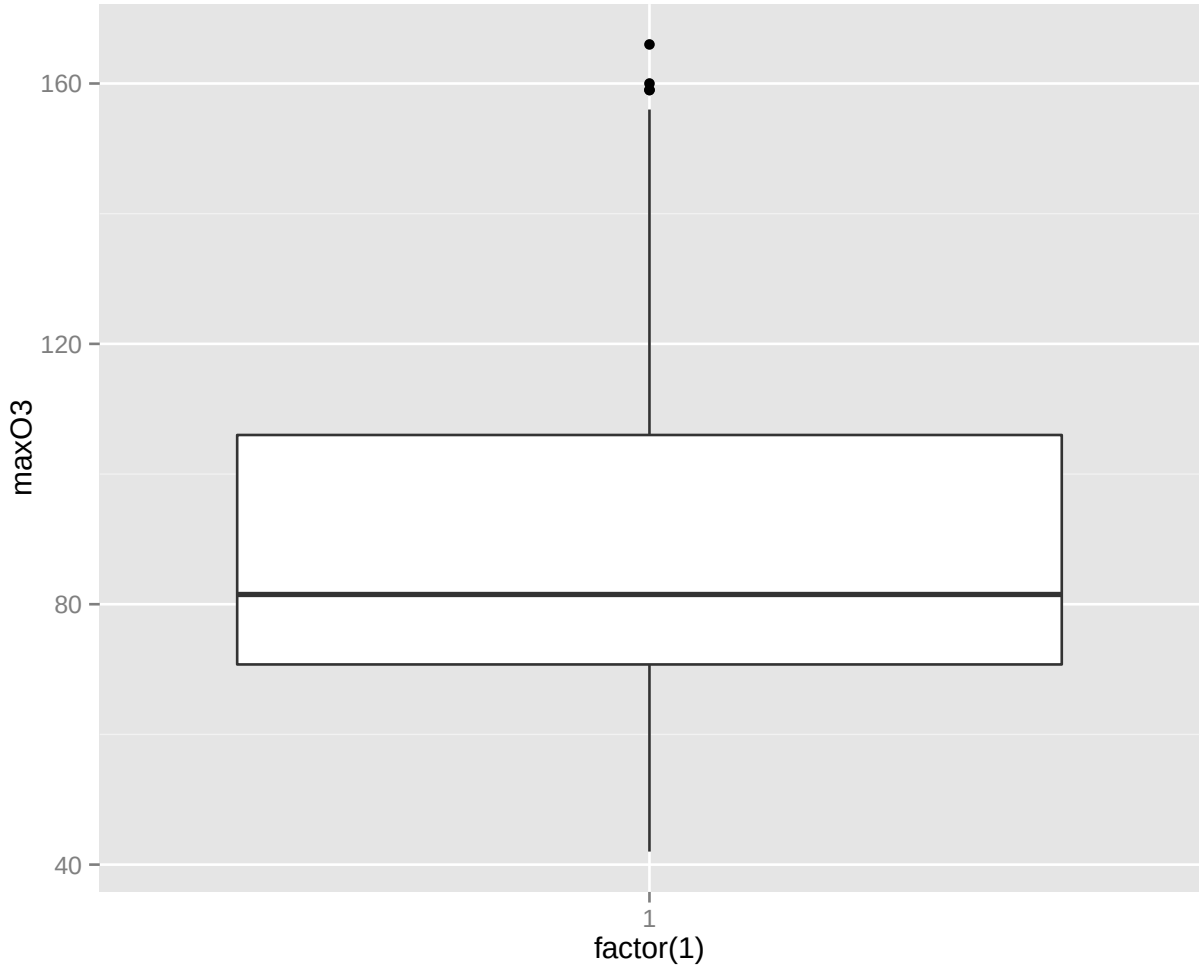
- Écart-type observé (corrigé)

$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Histogramme de maxO3



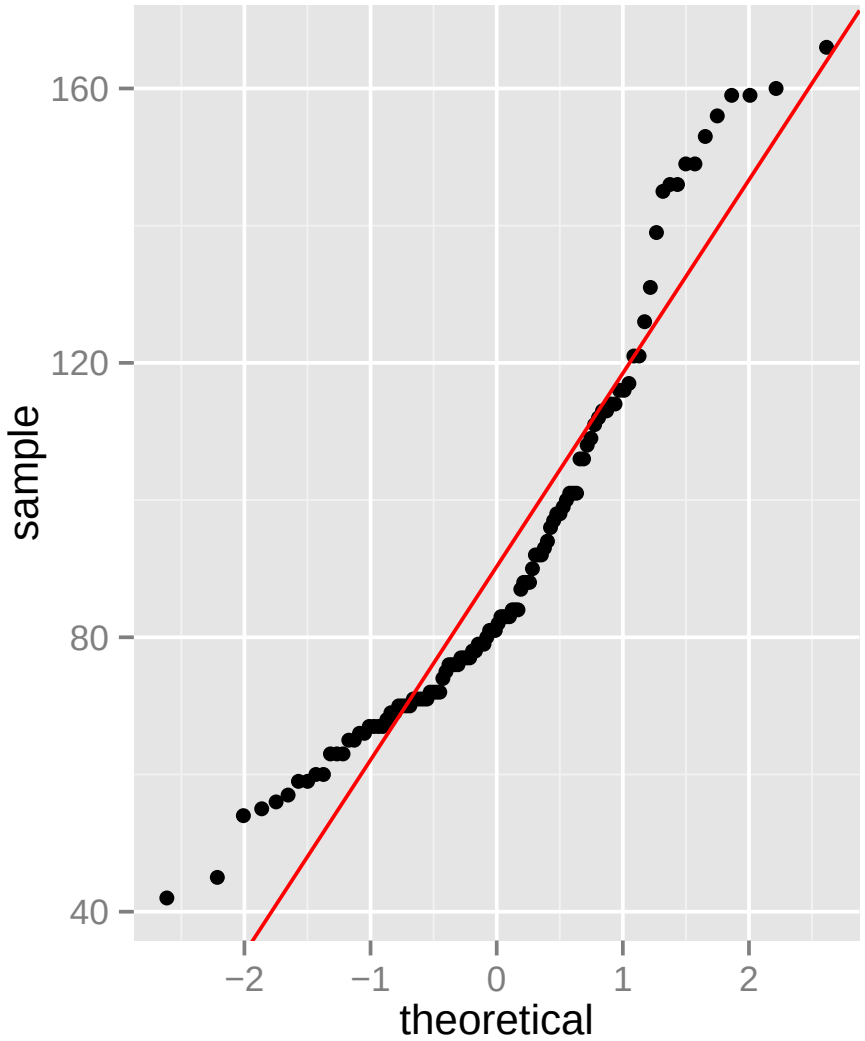
Boite à moustache de maxO3



Cas Gaussien

- Des nombreux outils statistiques nécessitent de vérifier le caractère gaussien ou non de la distribution.
- Un nombre important d'observations dans l'échantillon (par exemple ici $n = 112$) permet en partie de s'affranchir de cette hypothèse mais il est utile de savoir la vérifier et éventuellement de sélectionner la transformation la plus appropriée des données notamment pour les variables de concentration d'ozone.
- Outils : QQ plots (graphe de quantile-quantile), test de normalité.

QQ-Plots



Test de Normalité : Shapiro-Wilk et Kolmogorov-Smirnov

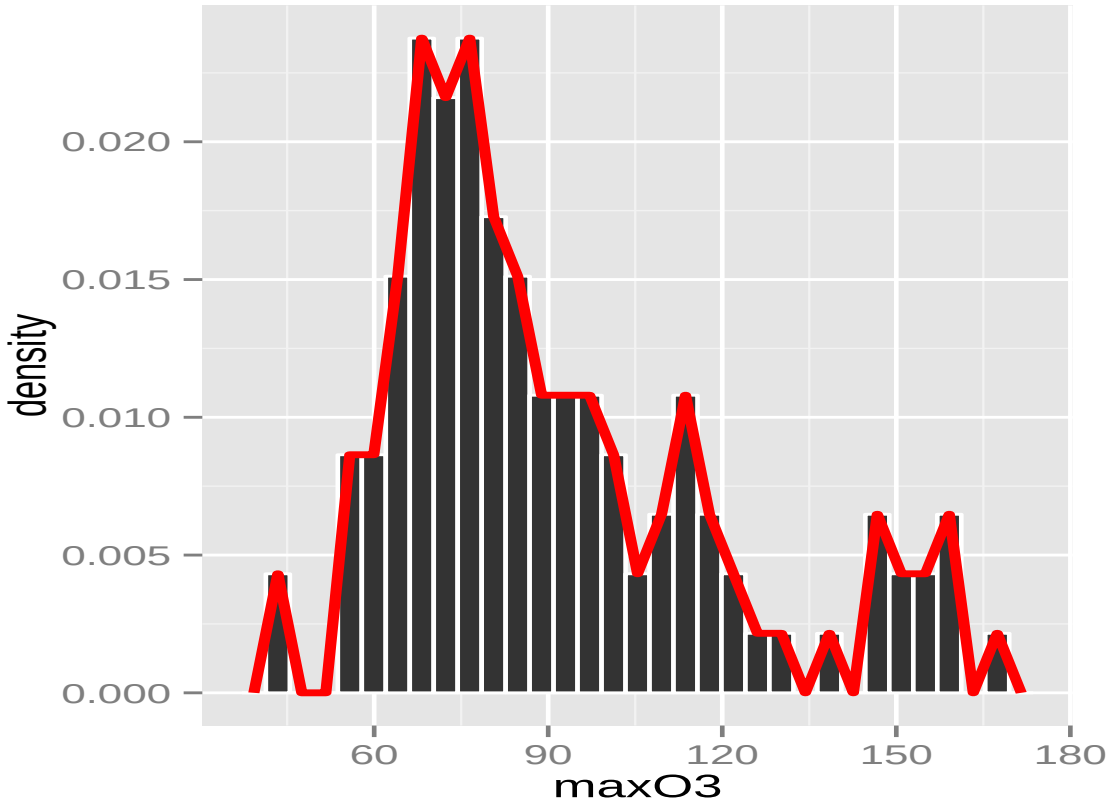
- Les résultats suivants permettent-ils de rejeter ou de conserver l'hypothèse que les mesures de maxO3 sont des réalisations i.i.d. d'une variable gaussienne ?
- Donner l'hypothèse nulle, l'hypothèse alternative et votre conclusion si le test est réalisé au niveau $\alpha = 5\%$.

```
Shapiro-Wilk normality test
W = 0.906, p-value = 8.516e-07
```

```
One-sample Kolmogorov-Smirnov test
D = 0.1599, p-value = 0.006509
```

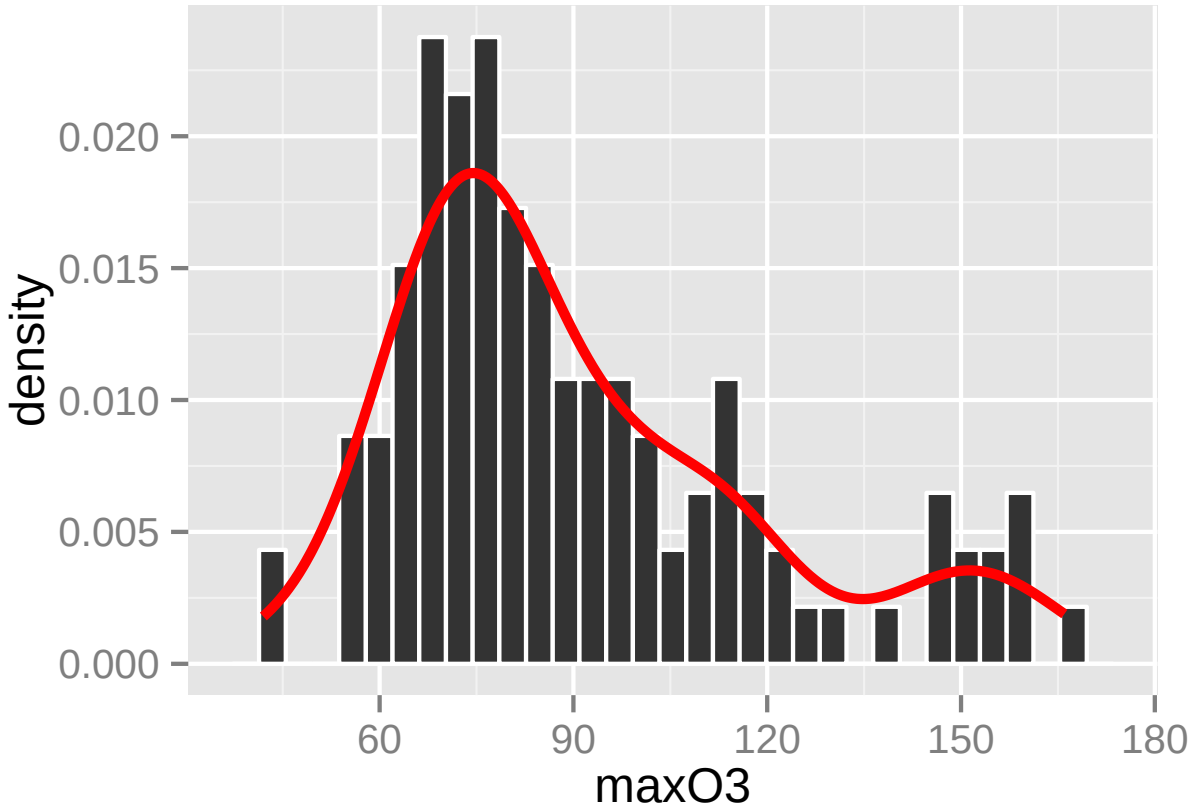
Histogramme de maxO3 et sa densité estimé

La loi des mesures de maxO3 est inconnue. On estime cette loi ! Supposant que cette loi possède une densité, on a représenté la densité estimé (estimation par histogramme).



Histogramme de maxO3 et sa densité estimé

La loi des mesures de maxO3 est inconnue. On estime cette loi ! Supposant que cette loi possède une densité, on a représenté la densité estimé (estimation par une méthode à noyau).



Étude de deux variables

L'étude simultanée de deux variables X et Y définies sur une même population $\mathcal{P}$ a pour but de mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables.

Deux variables quantitatives X et Y

On dispose d'un échantillon de taille n du couple (X, Y)

$$\{(x_1, y_1), \ldots, (x_n, y_n)\}.$$

- Moyennes observées

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- Écart-types observés (corrigés)

$$\mathrm{s}_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \mathrm{s}_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Coefficient de corrélation linéaire

- Covariance observée

$$\mathrm{cov}(x, y) = \frac{1}{n-1}\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)$$

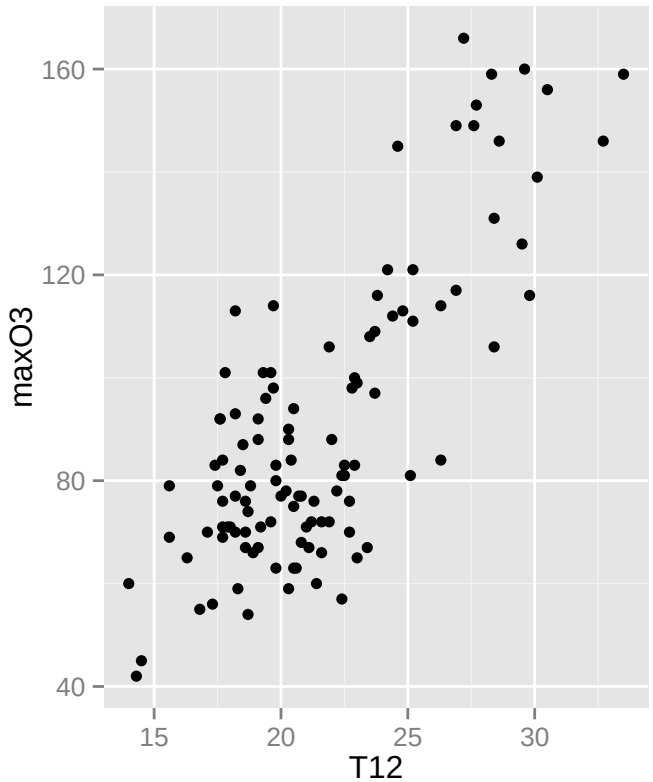
- Coefficient de corrélation linéaire observé

$$\mathrm{r}(x, y) = \frac{\mathrm{cov}(\mathrm{x}, \mathrm{y})}{\mathrm{s}_x \mathrm{s}_y}$$

Deux variables quantitatives X et Y

- X : température à midi
- Y : concentration maximale en ozone

Nuage de points



Exemple : Pollution à l'ozone

- X : température à midi
- Y : concentration maximale en ozone
- Moyennes

$$\bar{x} = 21.527 \quad \bar{y} = 90.304$$

- Écart-types et variances

$$\sigma_x = 4.042 \quad \sigma_y = 28.187 \quad \sigma_x^2 = 16.340 \quad \sigma_y^2 = 794.520$$

- Covariance et corrélation

$$cov(x, y) = 89.360 \quad r(x, y) = 0.784$$

Peut-on conclure, au risque d'erreur $\alpha = 1\%$, qu'il existe une liaison entre les variables X et Y ?

Test d'indépendance pour deux variables quantitatives

Test de Pearson et Test de Spearman

```
Pearson's product-moment correlation
t = 13.258, df = 110, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
sample estimates:
      cor
0.7842623

Spearman's rank correlation rho
S = 89097, p-value = 3.307e-13
alternative hypothesis: true rho is not equal to 0
  rho
0.6194629
Warning message:
Impossible de calculer la p-value exacte avec des ex-aequos
```

Résumé des variables quantitatives

```
      maxO3             T9              T12             T15
 Min.   : 42.00   Min.   :11.30   Min.   :14.00   Min.   :14.90
 1st Qu.: 70.75   1st Qu.:16.20   1st Qu.:18.60   1st Qu.:19.27
 Median : 81.50   Median :17.80   Median :20.55   Median :22.05
 Mean   : 90.30   Mean   :18.36   Mean   :21.53   Mean   :22.63
 3rd Qu.:106.00   3rd Qu.:19.93   3rd Qu.:23.55   3rd Qu.:25.40
 Max.   :166.00   Max.   :27.00   Max.   :33.50   Max.   :35.50

      Ne9             Ne12            Ne15            Vx9
 Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :-7.8785
 1st Qu.:3.000   1st Qu.:4.000   1st Qu.:3.00   1st Qu.:-3.2765
 Median :6.000   Median :5.000   Median :5.00   Median :-0.8660
 Mean   :4.929   Mean   :5.018   Mean   :4.83   Mean   :-1.2143
 3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.:7.00   3rd Qu.: 0.6946
 Max.   :8.000   Max.   :8.000   Max.   :8.00   Max.   : 5.1962

      Vx12            Vx15           maxO3v
 Min.   :-7.878   Min.   :-9.000   Min.   : 42.00
 1st Qu.:-3.565   1st Qu.:-3.939   1st Qu.: 71.00
 Median :-1.879   Median :-1.550   Median : 82.50
 Mean   :-1.611   Mean   :-1.691   Mean   : 90.57
 3rd Qu.: 0.000   3rd Qu.: 0.000   3rd Qu.:106.00
 Max.   : 6.578   Max.   : 5.000   Max.   :166.00
```

Matrice de corrélation

	maxO3	T9	T12	T15	Ne9	Ne12	Ne15	Vx9	Vx12	Vx15
maxO3	1.000	0.699	0.784	0.775	-0.622	-0.641	-0.478	0.528	0.431	0.392
T9	0.699	1.000	0.883	0.846	-0.484	-0.472	-0.325	0.251	0.222	0.170
T12	0.784	0.883	1.000	0.946	-0.584	-0.660	-0.458	0.430	0.313	0.271
T15	0.775	0.846	0.946	1.000	-0.586	-0.649	-0.575	0.453	0.344	0.287
Ne9	-0.622	-0.484	-0.584	-0.586	1.000	0.788	0.550	-0.498	-0.529	-0.494
Ne12	-0.641	-0.472	-0.660	-0.649	0.788	1.000	0.710	-0.493	-0.510	-0.432
Ne15	-0.478	-0.325	-0.458	-0.575	0.550	0.710	1.000	-0.401	-0.432	-0.378
Vx9	0.528	0.251	0.430	0.453	-0.498	-0.493	-0.401	1.000	0.750	0.682
Vx12	0.431	0.222	0.313	0.344	-0.529	-0.510	-0.432	0.750	1.000	0.837
Vx15	0.392	0.170	0.271	0.287	-0.494	-0.432	-0.378	0.682	0.837	1.000

Corrélation et causalité? Synonymes? NON!

- Quelle est la cause, quel est l'effet ?
- Un lien statistique n'est pas toujours le signe d'une causalité

Motivation : Un exemple intéressant

La consommation moyenne de chocolat par habitant est corrélée au nombre de lauréats du prix Nobel, d'après une étude de l'Américain Franz Messerli publiée en 2012 : en général, plus un pays présente une consommation élevée, plus nombreux sont ses ressortissants nobélisés. Mais on ne peut pas conclure qu'une consommation accrue de chocolat a pour effet d'augmenter le nombre de lauréats du prix Nobel. Il est même vraisemblable que la corrélation soit trompeuse : la richesse du pays pourrait être le facteur commun aux deux propriétés considérées, sans que ces propriétés soient causalement liées entre elles.

Corrélation et causalité? Synonymes? NON!

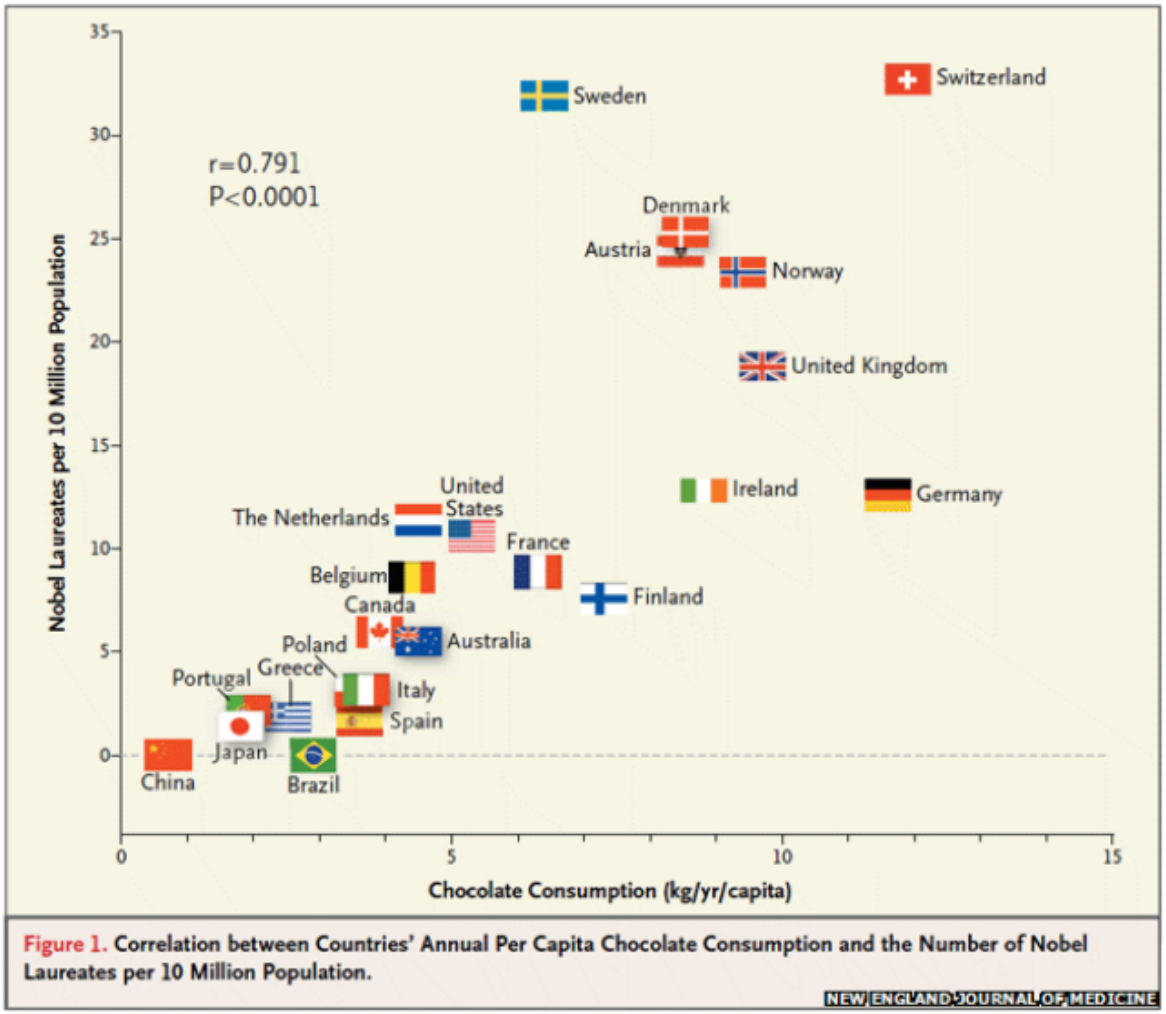


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

NEW ENGLAND JOURNAL OF MEDICINE

Nombre de prix Nobel par dix millions d'habitants en fonction de la consommation nationale de chocolat en kilogrammes par personne et par an.

Image : Franz H. Messerli, *The New England Journal of Medicine* **367**(16) (2012), p. 1562-1564

Une variable qualitative

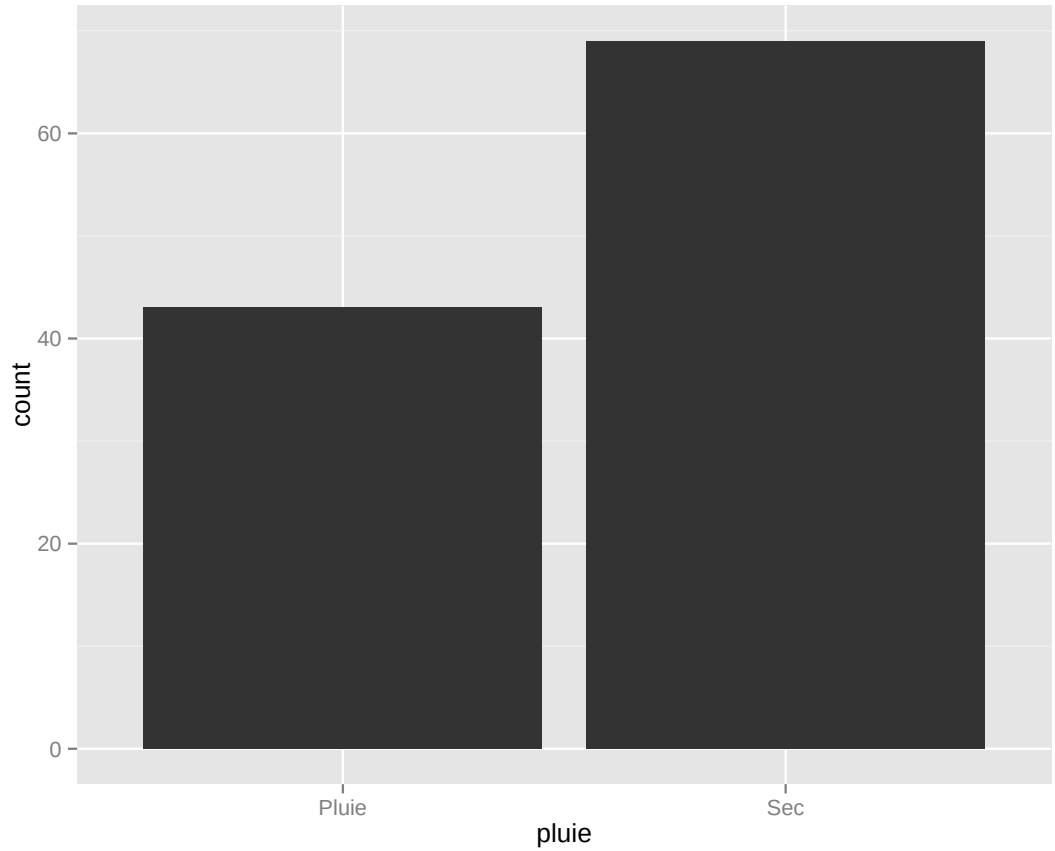
Variable pluie (type : qualitative à deux modalités)

Les tableaux des effectifs et fréquences

```
Tableau de effectifs (ni)
Pluie   Sec
43      69

Tableau de frequences (fi)
Pluie   Sec
0.384   0.616
```

Une variable qualitative

Diagramme en tuyaux d'orgue de la variable pluie



Étude de deux variables

L'étude simultanée de deux variables X et Y définies sur une même population $\mathcal{P}$ a pour but de mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables.

Exemple (ozone) : variables vent et pluie

```
vent           pluie
Est   :10      Pluie:43
Nord  :31      Sec  :69
Ouest:50
Sud   :21
```

Croiser les deux variables !

Deux variables qualitatives X et Y

- On suppose que X peut prendre k modalités notées $x_1, ..., x_k$
- On suppose que Y peut prendre l modalités notées $y_1, ..., y_l$

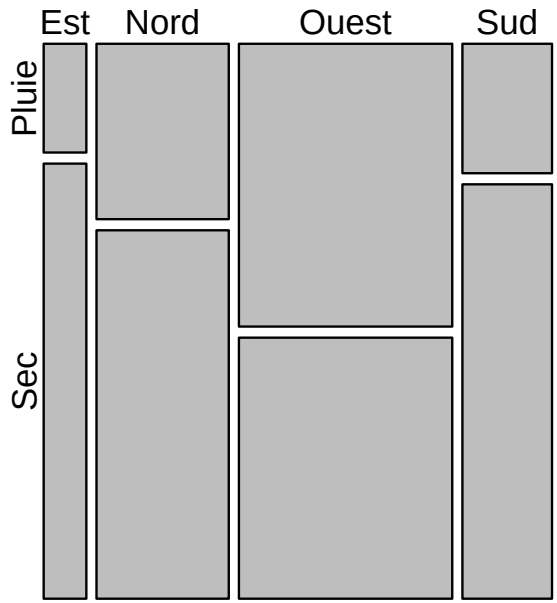
Tableau de contingence

X/Y	y_1	...	y_j	...	y_l	$n_{i\bullet}$
x_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1\bullet}$
...	...	...	...	...	...	
x_i	n_{i1}	...	n_{ij}	...	n_{il}	$n_{i\bullet}$
...	...	...	...	...	...	
x_k	n_{k1}	...	n_{kj}	...	n_{kl}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	n

Deux variables qualitatives : pluie et vent

Tableau de contingence et graphe de profil colonnes

	vent			
pluie	Est	Nord	Ouest	Sud
Pluie	2	10	26	5
Sec	8	21	24	16



Peut-on conclure, au risque d'erreur $\alpha = 1\%$, qu'il existe une liaison entre les variables vent et pluie ?

Test d'indépendance pour deux variables qualitatives

Démarche du test d'indépendance pour deux variables qualitatives

1. Poser les hypothèses nulle et alternative du test puis fixer le risque d'erreur α.
   Dans l'exemple on teste donc :
   H_0: Les variables X et Y sont indépendantes
   H_1: Les variables X et Y ne sont pas indépendantes
   au niveau $\alpha = 1\%$.
2. Calculer le tableau des effectifs théoriques attendus (qu'on devrait avoir) sous l'hypothèse H_0 d'indépendance
3. Comparer ce tableau correspondant à l'indépendance avec le tableau observé; pour cela on va définir une distance mesurant l'écart entre les tableaux qu'on appelle distance du chi-2.
4. Prendre une décision basée sur la p-valeur. Conclure.

Distance du chi-2 (χ^2)

- De manière générale, on calcule les effectifs théoriques sous l'hypothèse H_0 donnés par

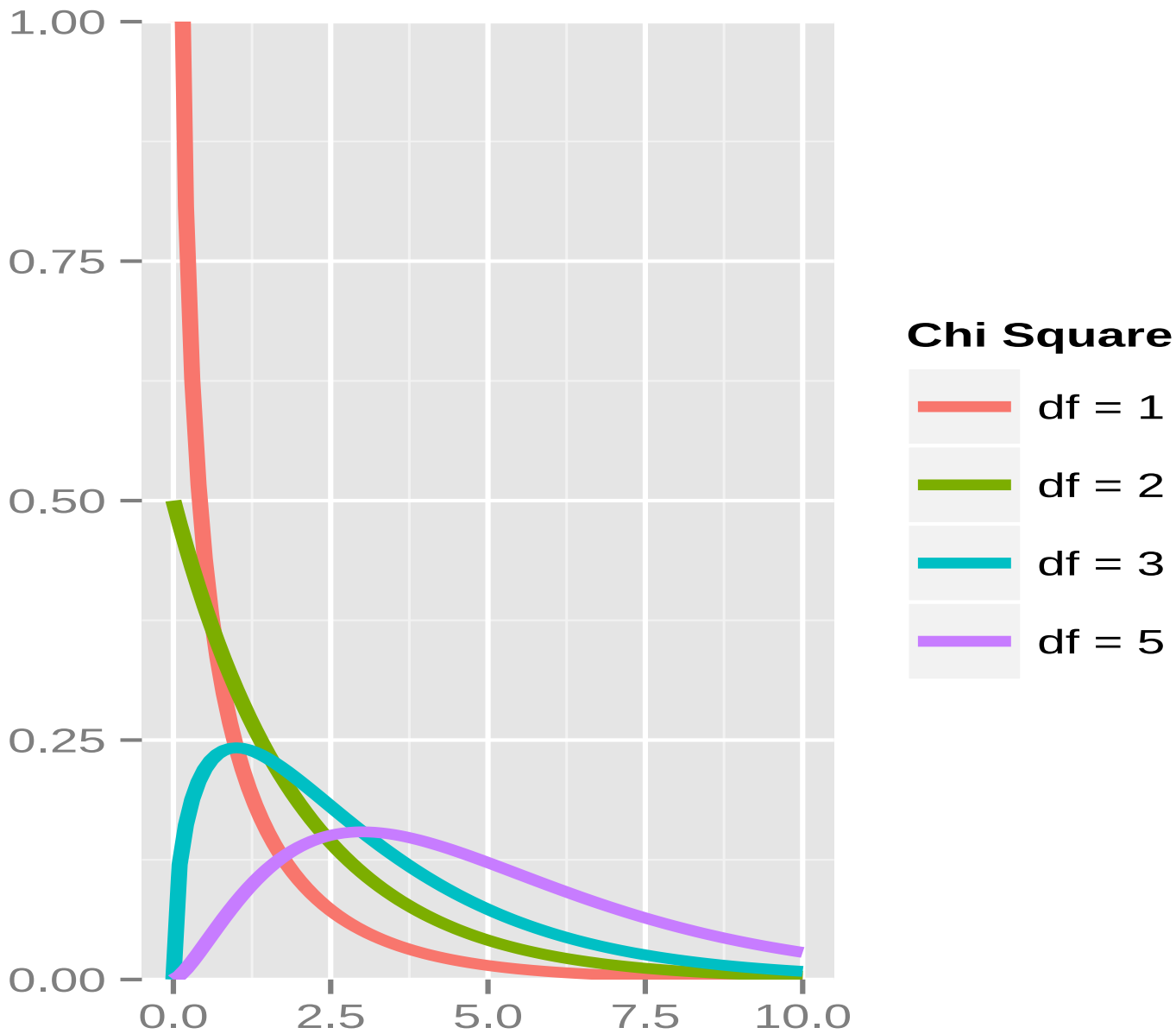
$$e_{ij} = \frac{n_{i\bullet} \times n_{\bullet j}}{n}$$

pour tous les i, j.

- On introduit la distance du chi-2 définie comme suit.

$$Q_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(N_{ij} - E_{ij})^2}{E_{ij}}$$

Sous l'hypothèse nulle H_0, la v.a. Q_n^2 suit approximativement une loi $\chi^2((k-1) \times (l-1))$ **dès que** $n \geqslant 30$ **et que les effectifs théoriques sont supérieurs ou égaux à** 5.


1.00
0.75
0.50
0.25
0.00
0.0
2.5
5.0
7.5
10.0
Chi Square
df = 1
df = 2
df = 3
df = 5

Tableu des effectifs observées (nij)

	Est	Nord	Ouest	Sud
Pluie	2	10	26	5
Sec	8	21	24	16

Tableau des effectifs esperés (eij)

	Est	Nord	Ouest	Sud
Pluie	3.839286	11.90179	19.19643	8.0625
Sec	6.160714	19.09821	30.80357	12.9375

Tableau des ecarts aux carrés (nij-eij)^2

	Est	Nord	Ouest	Sud
Pluie	0.8811462	0.3038862	2.4113123	1.1632752
Sec	0.5491201	0.1893784	1.5027019	0.7249396

Test d'indépendance pour deux variables qualitatives

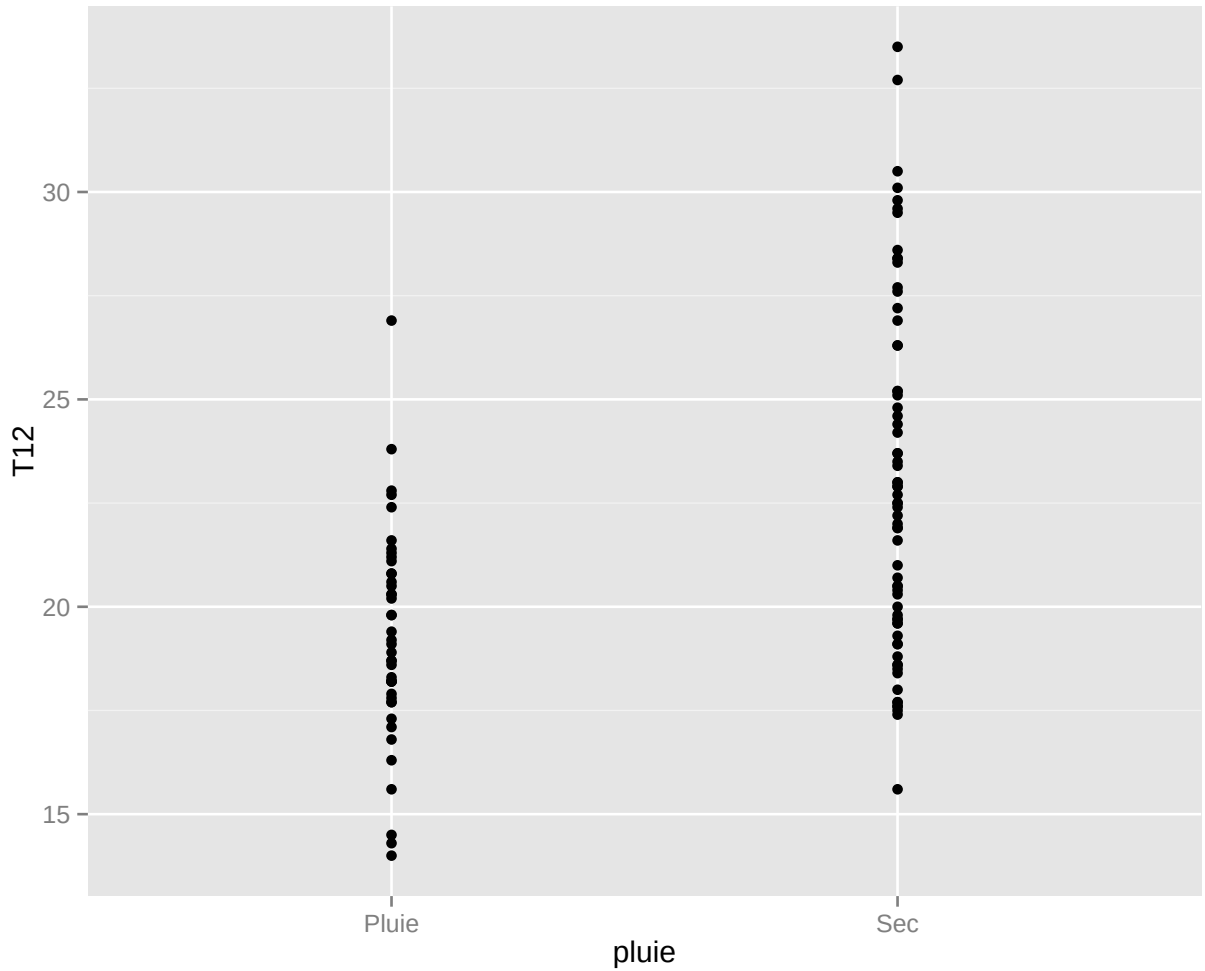
Peut-on conclure, au risque d'erreur $\alpha = 1\%$, qu'il existe une liaison entre les variables vent et pluie ?

```
Pearson's Chi-squared test

data:  table(ozone$pluie, ozone$vent)
X-squared = 7.7258, df = 3, p-value = 0.05203

Warning message:
In chisq.test(table(ozone$pluie, ozone$vent)) :
  l'approximation du Chi-2 est peut-être incorrecte
```

- Représentations des variables quantitatives et qualitatives
- Autres problèmes ....
- A vous de jouer !

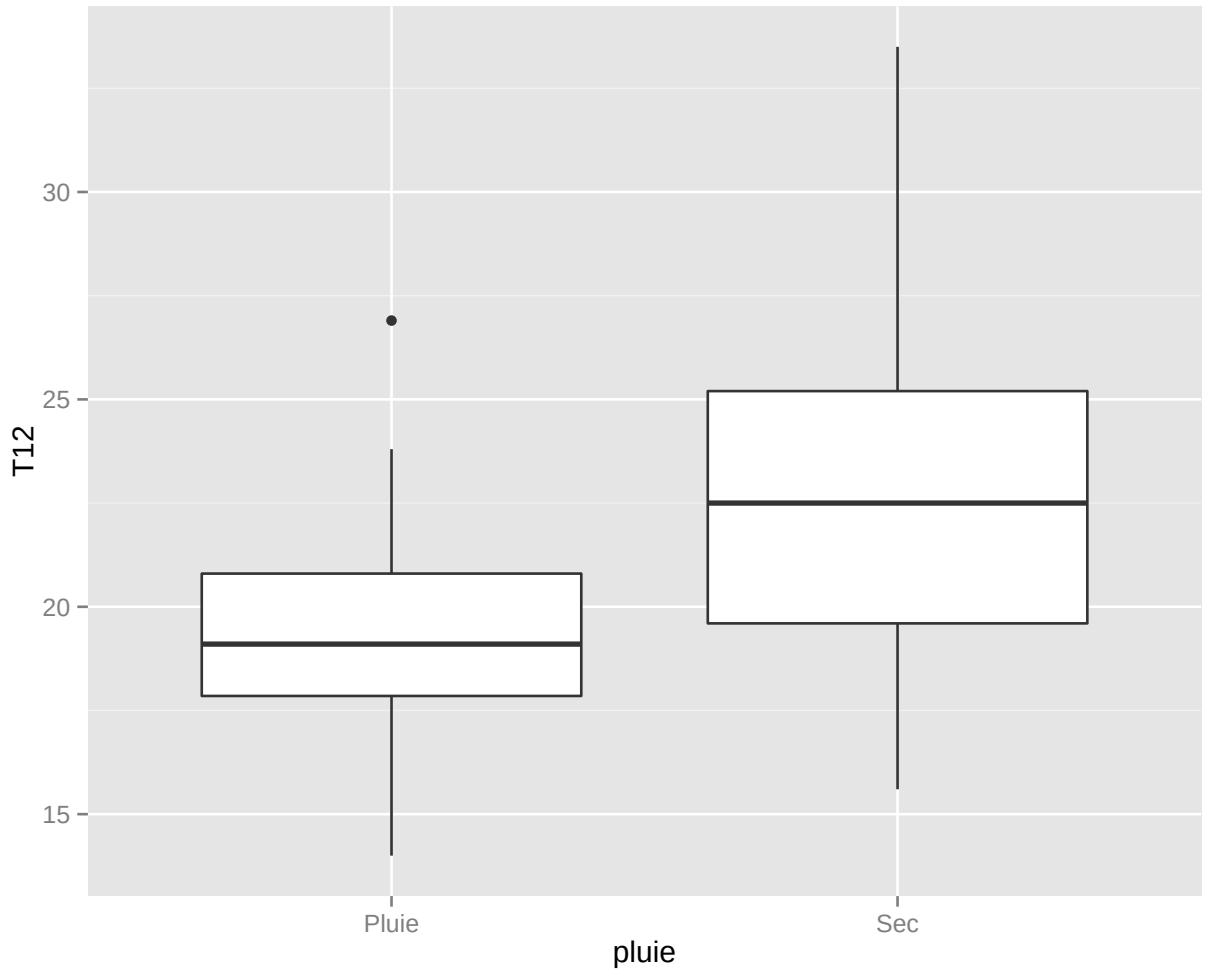
Nuage de points parallèle

Une variable quantitative et une variable qualitative (à 2 modalités)



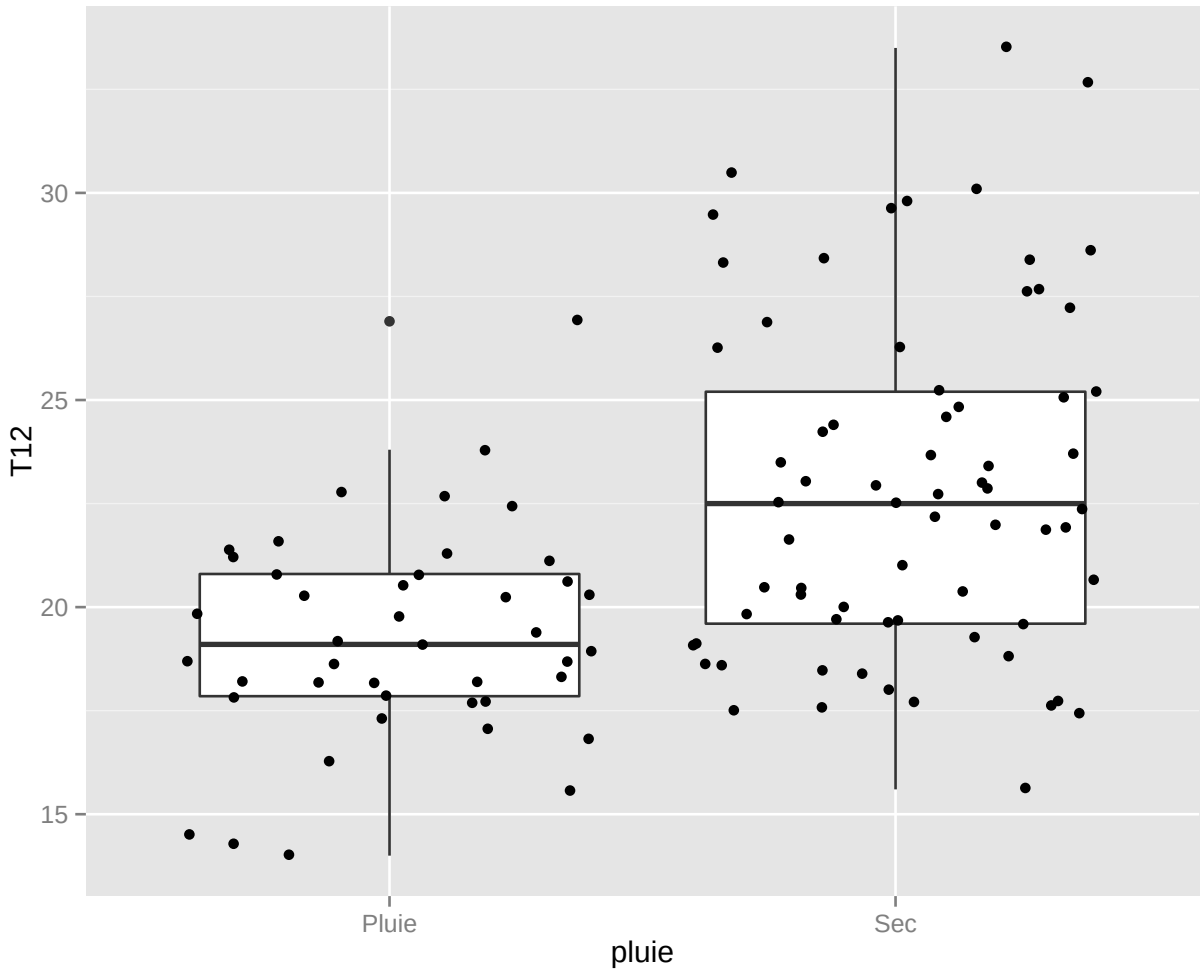
Boite à moustache parallèles

Une variable quantitative et une variable qualitative (à 2 modalités)



Nuage et boite à moustache parallèles

Une variable quantitative et une variable qualitative (à 2 modalités)



Nuage et boite à moustache parallèles

Deux variables quantitatives et une variable qualitative (à 2 modalités)