Rappels (II) : Analyse statistique pour des variables quantitatives et qualitatives

Master 2 Recherche SES-IES Analyse de données

Ana Karina Fermin

Université Paris Nanterre

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

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 : Employés dans une entreprise canadienne.
- Variables : salaire brut actuel par an (X_1), salaire de départ par an (X_2), sexe (X_3),nombre d'années d'étude (X_4), ...
- On dispose d'un échantillon de $\mathbf{X} = (X_1, \ldots, X_d)$ de taille $n = 474$

$$D_n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$$

avec $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$ le i-ème individu $(i = 1, \ldots, n)$.

Étude de deux variables

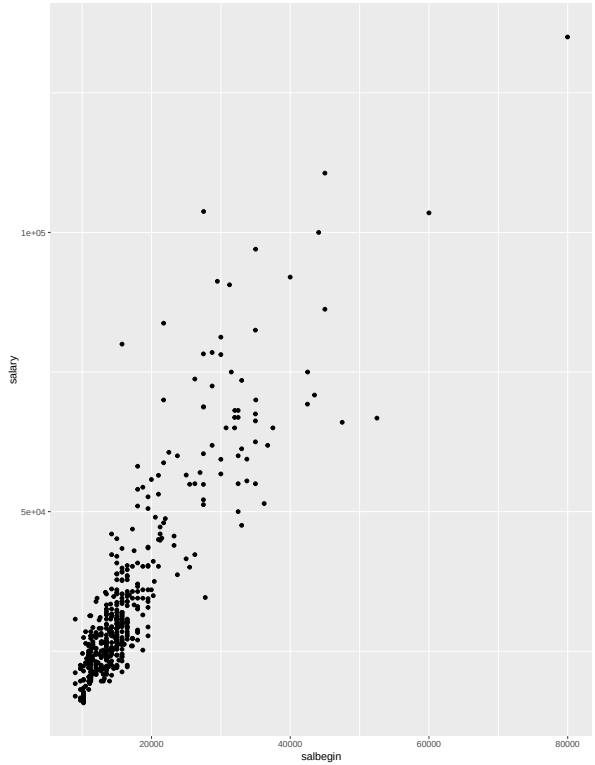
L'étude simultanée de deux variables X et Y définies sur une même population $\mathcal{P}$ a pour but de mettre en évidence une éventuelle liaison (relation, dépendance) entre les variables.

Exemple : Données salaires (étudiées dans l'atelier) : 6 observations du tableau

	salary	salbegin	jobtime	prevexp	educ	minority	sex
1	57000	27000	98	144	15	Non	H
2	40200	18750	98	36	16	Non	H
3	21450	12000	98	381	12	Non	F
4	21900	13200	98	190	8	Non	F
5	45000	21000	98	138	15	Non	H
6	32100	13500	98	67	15	Non	H

Deux variables quantitatives X et Y

- X : salaire de départ (salbegin)
- Y : salaire actuel (salary)



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

Démarche du test d'indépendance pour deux variables quantitatives

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 coefficient de corrélation linéaire observée (sur l'échantillon)
3. Calculer la valeur réalisée de la statistique du test de corrélation linéaire (sur l'échantillon)
4. Prendre une décision basée sur la p-valeur. Conclure.

Test d'indépendance pour deux variables quantitatives

- Résultat du test de Pearson (obtenue avec le logiciel R)

```
> cor.test(Salaire$salary,Salaire$salbegin)

Pearson's product-moment correlation
data:  Salaire$salary and Salaire$salbegin
t = 40.276, df = 472, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.8580696 0.8989267
sample estimates:
cor
0.8801175
```

- On rejette H0 !

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)

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

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

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

- Valeur observée de la statistique du test (sous H0)

$$t_n = \frac{\mathrm{r}}{\sqrt{(1 - \mathrm{r}^2)/(n-2)}}$$

Rem : La statistique du test de Pearson suit une loi de Student à $n - 2$ degrés de liberté

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 salaires

- X : minority (appartenance à une minorité), variable qualitative à 2 modalités $\{\text{Oui}, \text{Non}\}$
- Y : sex, variable qualitative à 2 modalités $\{F, H\}$

```
minority   sex
Non:370    H:258
Oui:104    F:216
```

- On va croiser les deux variables !

Test d'indépendance

- On veut tester l'existence d'une liaison entre le sexe et l'appartenance à une minorité chez les salariés de l'entreprise.
- Effectifs observées, n_{ij}

	H	F
Non	194	176
Oui	64	40

- On teste :
  H_0: Les variables X et Y sont indépendantes
  H_1: Les variables X et Y ne sont pas indépendantes
- Que peut-on conclure au risque $\alpha = 1\%$?

1. Résultat du test (obtenue avec le logiciel R)

```
Pearson's Chi-squared test with
Yates' continuity correction

data:  nij
X-squared = 2.3592, df = 1,
p-value = 0.1245
```

2. On ne rejette pas H0 !

Deux variables qualitatives

- Effectifs observées, n_{ij}

	H	F
Non	194	176
Oui	64	40

- Effectifs espérées, e_{ij}

	H	F
Non	201.39241	168.60759
Oui	56.60759	47.39241

- Comparer ces deux tableau ! On a besoin de définir une distance mesurant l'écart entre les tableaux qu'on appelle distance du chi-2.

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.

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

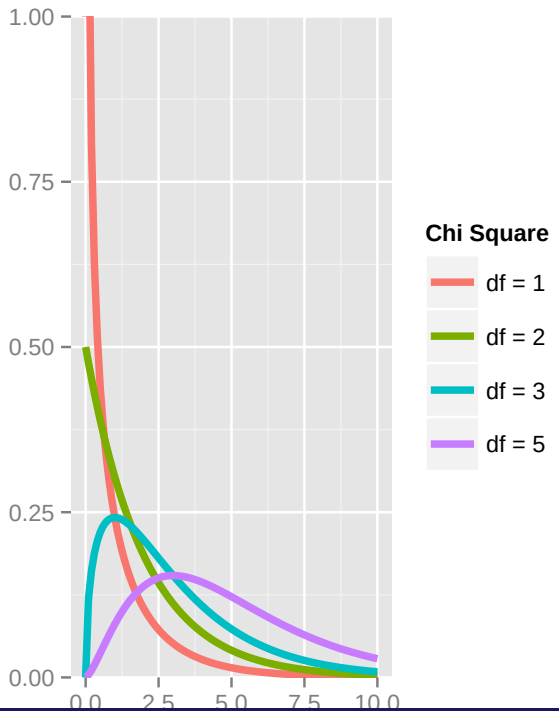
Sous l'hypothèse nulle H_0, la v.a. Q_n^2 suit approximativement une loi $\chi^2((k-1) \times (l-1))$ **dès que** $n \geqslant 30$ **et que les effectifs théoriques sont supérieurs ou égaux à** 5.

Distance du chi-2 (χ^2)

- Statistique du test et loi

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

dès que $n \geqslant 30$ et $n_{ij} \geqslant 5$ pour tout i, j



- Sous H0, la statistique de test s'approche à une loi chi-2 avec 1 degré de liberté.
- Valeur observée de la statistique du test

$$q_n^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - e_{ij})^2}{e_{ij}}$$

- Valeur observée de la statistique de test : 2.3592

Modéliser pour analyser et prédire ...

- Mettre en relation une variable expliquée (cible) et une ou plusieurs variables explicatives.
- Exemple : Élections Municipales à Marseille par bureau de vote (2014)
  - Données : https://www.data.gouv.fr/fr/datasets/marseille-elections-municipales/
  - Stat. descriptives de quelques variables :

```
     Ravier            Gaudin          Mennucci       UnionMarseille
 1st Qu.: 8.657   1st Qu.:12.719   1st Qu.: 8.228   1st Qu.: 0.00000
 Median :11.711   Median :19.680   Median :10.263   Median : 0.07428
 Mean   :12.025   Mean   :19.540   Mean   :10.783   Mean   : 1.04482
 3rd Qu.:14.918   3rd Qu.:25.340   3rd Qu.:12.741   3rd Qu.: 0.67431

 chomage         etrangers           CS1              CS2
 1st Qu.: 7.035   1st Qu.: 2.6397   1st Qu.:0.00000   1st Qu.:2.69266
 Median :10.216   Median : 4.7354   Median :0.00000   Median :3.36538
 Mean   :10.900   Mean   : 6.7774   Mean   :0.05326   Mean   :3.52885
 3rd Qu.:12.918   3rd Qu.: 8.4764   3rd Qu.:0.00000   3rd Qu.:4.19708
```

- Faire une analyse descriptive des données....Par exemple :
  - Le candidat UMP, Jean-Claude Gaudin a réuni en moyenne 19,54% des suffrages exprimés dans chaque bureau de vote.
  - Le pourcentage de vote n'était pas homogène dans l'ensemble de la ville, et les résultats dans chaque bureau de vote ont connu des écarts assez importants.
  - Le vote pour la liste de Jean-Claude Gaudin a varié entre les extremums 5,108 et 49,063, avec une médiane (autant de bureaux au dessus que de bureaux en dessous de cette valeur) de 19,764

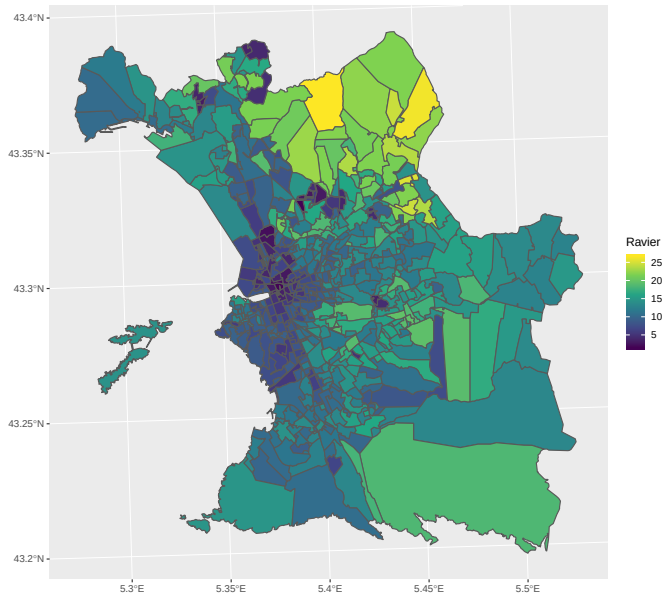
  .
- On se pose des question et on essaye d'y répondre.
- Peut-on expliquer la proportion de vote en fonction des variables sociologiques ?

- But : Estimer la proportion de vote FN (Ravier) à l'aide des variables sociologiques des bureaux de vote.

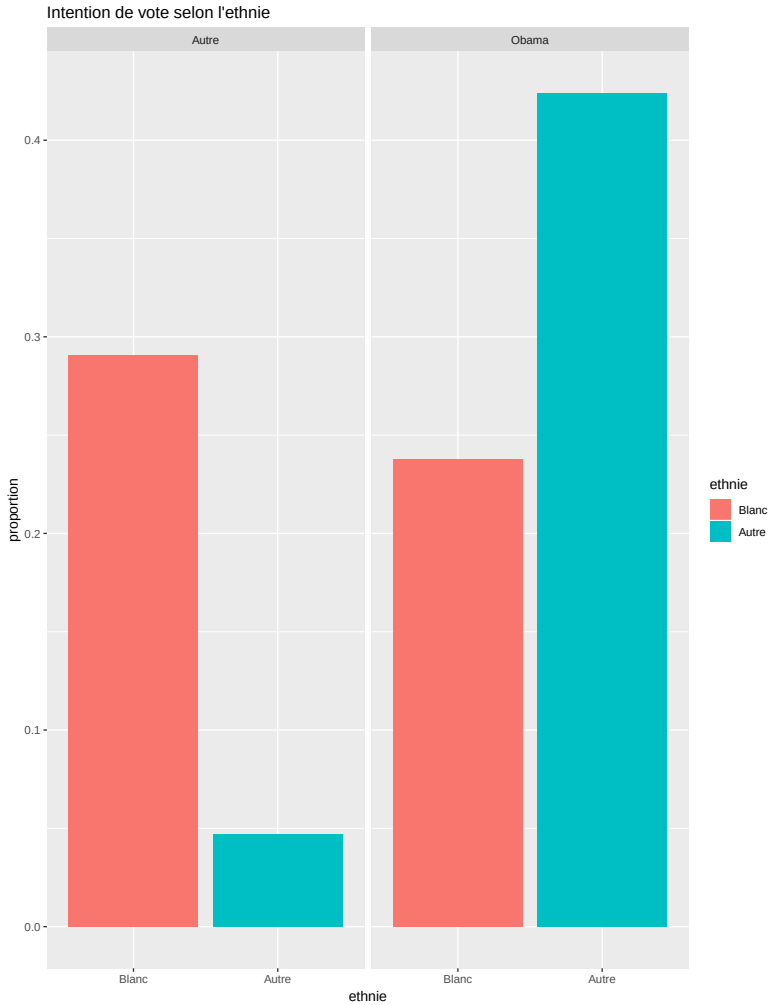
  FN = f(CSP, population étrangère, taux de chômage, ...)

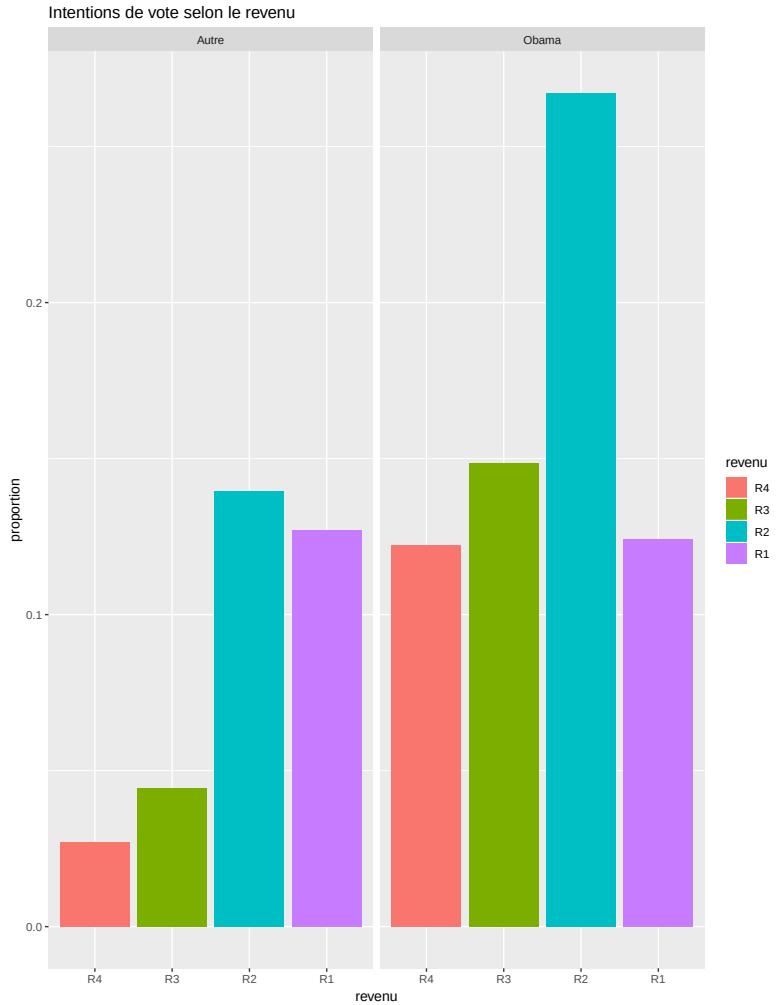
```
lm(formula = FN ~ ., data = Elections2014)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.82631    2.53500   6.638 1.12e-10 ***
CS1          0.87337    1.02290   0.854 0.393750
CS2         -0.08502    0.18098  -0.470 0.638792
CS3         -0.32557    0.06626  -4.913 1.34e-06 ***
CS4          0.12842    0.06446   1.992 0.047065 *
CS5          0.03866    0.05775   0.670 0.503575
CS6          0.33080    0.09857   3.356 0.000871 ***
etrangers   -0.28120    0.06342  -4.434 1.22e-05 ***
chomage     -0.48382    0.08299  -5.830 1.20e-08 ***
HLM         -0.03361    0.01243  -2.703 0.007181 **
---
Residual standard error: 3.397 on 374 degrees of freedom
Multiple R-squared:  0.5151,Adjusted R-squared:  0.5035
F-statistic: 44.15 on 9 and 374 DF,  p-value: < 2.2e-16
```

- Plusieurs modèles peuvent s'appliquer.
  - Sélection d'un bon modèle.
  - Comment faire ? Quelles méthodes connaissiez vous ?
- Modélisation spatial, régression spatial, ...(voir cours Mme Hardouin)
  - Le vote FN semble plus élevé dans les quartiers du nord de Marseille, probablement au nord sont installées les classes populaires; les ouvriers en particulier y sont sur-représentés.
  - Il semble donc y avoir corrélation spatiale.



- Election américane de 2008 (Source : Jean-Herman Guay)
- Une question se pose avec ces données :
- Est-ce que les électeurs de minorités ethniques (vs les Blancs), les gens moins favorisés économiquement (vs les plus favorisés) ont plus grande probabilité d'appuyer le candidat démocrate Barack Obama ?
- Peut-on estimer la probabilité de voter pour Obama à partir des variables explicatives ?


Intention de vote selon l'ethnie
Autre
Obama
proportion
0.0
0.1
0.2
0.3
0.4
Blanc
Autre
ethnie
ethnie
Blanc
Autre


Intentions de vote selon le revenu
Autre
Obama
proportion
0.0
0.1
0.2
R4
R3
R2
R1
revenu
revenu
R4
R3
R2
R1

- Estimer la probabilité de voter pour Obama à partir des variables explicatives.

$$\mathbb{P}(\text{Vote} = 1|\text{Sexe}, \text{ethnie}, \text{Revenu}) = f(\text{Sexe}, \text{ethnie}, \text{Revenu})$$

```
glm(formula = vote ~ sexe + ethnie + revenu, data = ElectionUSA2008)

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3161     0.2147   1.472  0.14098
sexeH         0.2036     0.1307   1.558  0.11933
ethnieAutre   2.2842     0.1490  15.326  < 2e-16 ***
revenuR3     -0.2809     0.2512  -1.118  0.26342
revenuR2     -0.5742     0.2178  -2.637  0.00838 **
revenuR1     -1.0881     0.2284  -4.764  1.9e-06 ***
---

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1843.5  on 1440  degrees of freedom
Residual deviance: 1455.3  on 1435  degrees of freedom
AIC: 1467.3

Number of Fisher Scoring iterations: 4
```