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

Peut-on conclure, au risque $\alpha = 1\%$, qu'il existe une liaison entre le salaire de départ (salbegin : X) et le salaire actuel (salary : Y) ?

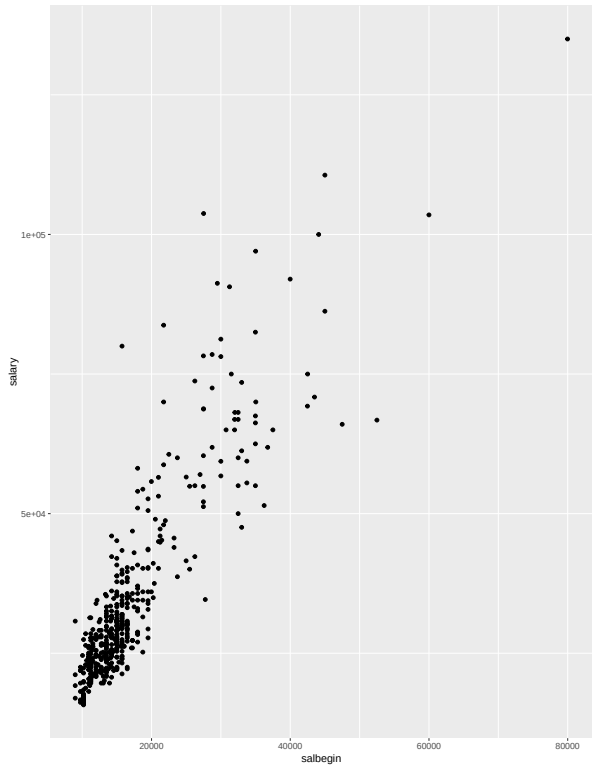
Test d'indépendance pour deux variables quantitatives

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.

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

- 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
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 !

Deux variables qualitatives X et Y

- On suppose que X peut prendre k modalités notées $A_1, ..., A_k$
- On suppose que Y peut prendre l modalités notées $B_1, ..., B_l$

Tableau de contingence

X/Y	B_1	...	B_j	...	B_l	$n_{i\bullet}$
A_1	n_{11}	...	n_{1j}	...	n_{1l}	$n_{1\bullet}$
...	...	...	...	...	...	
A_i	n_{i1}	...	n_{ij}	...	n_{il}	$n_{i\bullet}$
...	...	...	...	...	...	
A_k	n_{k1}	...	n_{kj}	...	n_{kl}	$n_{k\bullet}$
$n_{\bullet j}$	$n_{\bullet 1}$		$n_{\bullet j}$		$n_{\bullet l}$	n

$$n_{i\bullet} = n_{i1} + n_{i2} + \ldots + n_{il} \quad i = 1, \ldots, k$$

$$n_{\bullet j} = n_{1j} + n_{2j} + \ldots + n_{kj} \quad j = 1, \ldots, l$$

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.

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

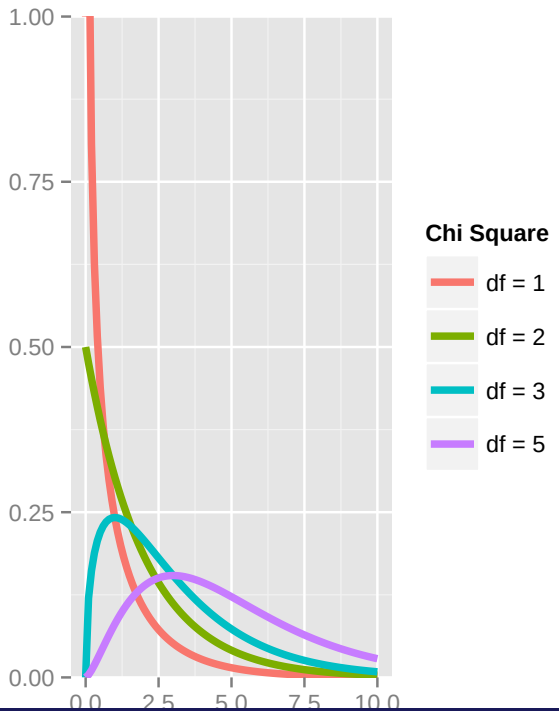
Sous l'hypothèse nulle H_0, la v.a. Q_n^2 suit approximativement une loi $\chi^2((k-1) \times (l-1))$ **dès que** $n \geqslant 30$ **et que les effectifs théoriques sont supérieurs ou égaux à** 5.

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