Chapitre 2: Statistique bivariée

Arnaud Rousselle
arnaud.rousselle@iut-dijon.u-bourgogne.fr
http://arousselle.perso.math.cnrs.fr/

Année universitaire 2023-2024
Semestre 2

1. Objectifs
2. Présentation et traitement des données
3. Étude et mesure des liens entre les variables
4. Régression

Objectifs

- Introduire des outils pour étudier des relations entre variables statistiques sur une **même** population.

- Décider si deux variables X et Y définies sur une même population sont suffisamment liées pour pouvoir « expliquer » raisonnablement l'une grâce à l'autre ou, au contraire, ne dépendent pas l'une de l'autre.

Mots-clés :

- indépendance
- distance du Khi-2, coefficient de Cramér
- test du Khi-2 d'indépendance
- covariance, corrélation
- régression

Prérequis : Chapitre de statistique à une variable du S1

1 Objectifs

2 Présentation et traitement des données

3 Étude et mesure des liens entre les variables

4 Régression

Notations et données

(X, Y) : couple de variables statistiques définies sur la même population $\mathcal{P}$

- N : effectif (total) de la population
- $n_{i,j}$ effectif du couple de modalités (x_i, y_j)
- $f_{i,j} = \frac{n_{i,j}}{N}$: la fréquence de (x_i, y_j)

Tableau de contingence :

X \ Y	y_1	y_2	$\dots$	y_j	$\dots$	y_r
x_1	$n_{1,1}$	$n_{1,2}$	$\dots$	$n_{1,j}$	$\dots$	$n_{1,r}$
x_2	$n_{2,1}$	$n_{2,2}$	$\dots$	$n_{2,j}$	$\dots$	$n_{2,r}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_i	$n_{i,1}$	$n_{i,2}$	$\dots$	$n_{i,j}$	$\dots$	$n_{i,r}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
x_l	$n_{l,1}$	$n_{l,2}$	$\dots$	$n_{l,j}$	$\dots$	$n_{l,r}$

Effectifs marginaux, fréquences marginales

- Effectifs marginaux en X : $n_{i,.} = n_{i,1} + \cdots + n_{i,r} = \sum_{j=1}^{r} n_{i,j}$
- Effectifs marginaux en Y : $n_{.,j} = n_{1,j} + \cdots + n_{l,j} = \sum_{i=1}^{l} n_{i,j}$
- Fréquences marginales en X : $f_{i,.} = \frac{n_{i,.}}{N}$
- Fréquence marginale en Y : $f_{.,j} = \frac{n_{.,j}}{N}$

permettent de compléter le tableau de contingence :

X \ Y	y_1	y_2	$\ldots$	y_j	$\ldots$	y_r	$n_{i,.}$	$f_{i,.}$
x_1	$n_{1,1}$	$n_{1,2}$	$\ldots$	$n_{1,j}$	$\ldots$	$n_{1,r}$	$n_{1,.}$	$f_{1,.}$
x_2	$n_{2,1}$	$n_{2,2}$	$\ldots$	$n_{2,j}$	$\ldots$	$n_{2,r}$	$n_{2,.}$	$f_{2,.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_i	$n_{i,1}$	$n_{i,2}$	$\ldots$	$n_{i,j}$	$\ldots$	$n_{i,r}$	$n_{i,.}$	$f_{i,.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_l	$n_{l,1}$	$n_{l,2}$	$\ldots$	$n_{l,j}$	$\ldots$	$n_{l,r}$	$n_{l,.}$	$f_{l,.}$
$n_{.,j}$	$n_{.,1}$	$n_{.,2}$	$\ldots$	$n_{.,j}$	$\ldots$	$n_{.,r}$	N	
$f_{.,j}$	$f_{.,1}$	$f_{.,2}$	$\ldots$	$f_{.,j}$	$\ldots$	$f_{.,r}$		1

Représentations graphiques

- un stéréogramme (analogue dans le cas bivarié de l'histogramme du cas univarié en 3 dimensions)

- un nuage de points
  - peut être pondéré :
    - on attache l'étiquette $(n_{i,j})$ au point (x_i, y_j)
    - ou on représente le point de coordonnées (x_i, y_j) par un disque de rayon proportionnel à $n_{i,j}$

Un exemple

On a relevé les notes des étudiants d'un groupe de TP, durant les deux premiers TP notés d'informatique. Celle-ci sont présentées dans la liste suivante sous la forme (X, Y) où X est la note du premier TP et Y celle du deuxième :

$$(12, 13), (10, 9), (15, 16), (12, 13), (13, 12), (6, 7), (9, 10),$$
$$(6, 7), (16, 15), (11, 13), (12, 13), (18, 19), (1, 3), (11, 13).$$

Un exemple

On a relevé les notes des étudiants d'un groupe de TP, durant les deux premiers TP notés d'informatique. Celle-ci sont présentées dans la liste suivante sous la forme (X, Y) où X est la note du premier TP et Y celle du deuxième :

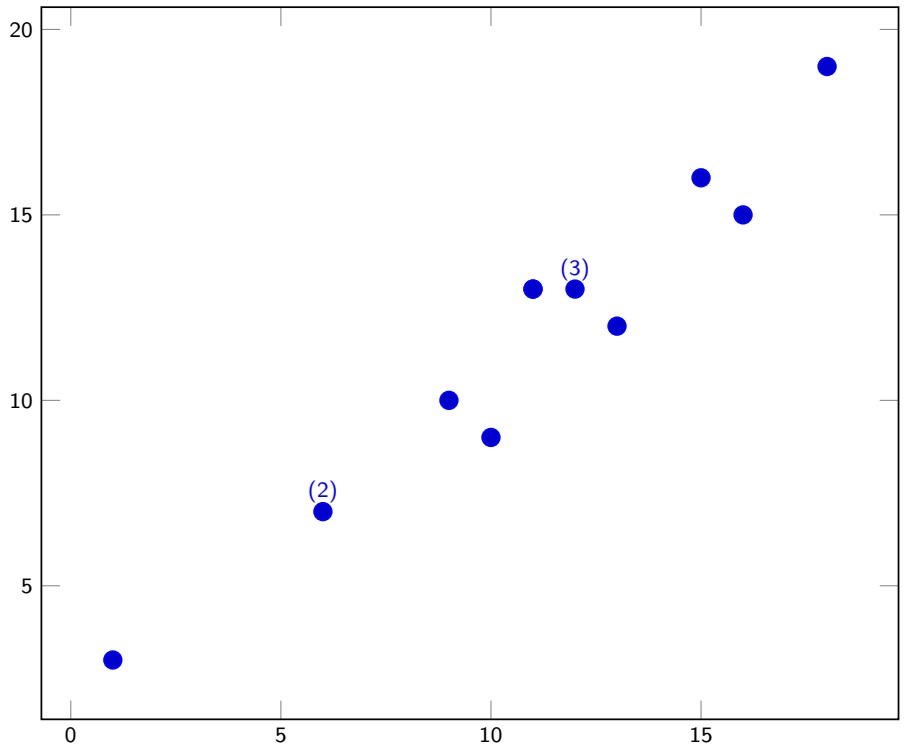
$$(12,13),\ (10,9),\ (15,16),\ (12,13),\ (13,12),\ (6,7),\ (9,10),$$
$$(6,7),\ (16,15),\ (11,13),\ (12,13),\ (18,19),\ (1,3),\ (11,13).$$

X \ Y	3	7	9	10	12	13	15	16	19	$n_{i,.}$	$f_{i,.}$
1	1	0	0	0	0	0	0	0	0	1	$\frac{1}{14} \simeq 0,07$
6	0	2	0	0	0	0	0	0	0	2	$\frac{1}{7} \simeq 0,14$
9	0	0	0	1	0	0	0	0	0	1	$\frac{1}{14} \simeq 0,07$
10	0	0	1	0	0	0	0	0	0	1	$\frac{1}{14} \simeq 0,07$
11	0	0	0	0	0	2	0	0	0	2	$\frac{1}{7} \simeq 0,14$
12	0	0	0	0	0	3	0	0	0	3	$\frac{3}{14} \simeq 0,21$
13	0	0	0	0	1	0	0	0	0	1	$\frac{1}{14} \simeq 0,07$
15	0	0	0	0	0	0	0	1	0	1	$\frac{1}{14} \simeq 0,07$
16	0	0	0	0	0	0	1	0	0	1	$\frac{1}{14} \simeq 0,07$
18	0	0	0	0	0	0	0	0	1	1	$\frac{1}{14} \simeq 0,07$
$n_{.,j}$	1	2	1	1	1	5	1	1	1	14	
$f_{.,j}$	$\frac{1}{14}$	$\frac{1}{7} \simeq 0,14$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{5}{14} \simeq 0,36$	$\frac{1}{14}$	$\frac{1}{14}$	$\frac{1}{14}$		1

Un exemple

On a relevé les notes des étudiants d'un groupe de TP, durant les deux premiers TP notés d'informatique. Celle-ci sont présentées dans la liste suivante sous la forme (X, Y) où X est la note du premier TP et Y celle du deuxième :

$$(12,13),\ (10,9),\ (15,16),\ (12,13),\ (13,12),\ (6,7),\ (9,10),$$
$$(6,7),\ (16,15),\ (11,13),\ (12,13),\ (18,19),\ (1,3),\ (11,13).$$



1 Objectifs

2 Présentation et traitement des données

3 Étude et mesure des liens entre les variables

4 Régression

Indépendance

Informellement, la notion d'indépendance exprime le fait que la variable X n'influence pas la variable Y et réciproquement.

Définition

On dit que X et Y sont indépendantes si :

$$f_{i,j} = f_{i,.} f_{.,j} \quad \text{pour tout } (i,j) \in \{1,\dots,l\} \times \{1,\dots,r\}.$$

Si X et Y sont indépendantes, on a

$$f_{i,j} = f_{i,.} f_{.,j}$$

donc

$$\frac{n_{i,j}}{N} = \frac{n_{i,.}}{N} \frac{n_{.,j}}{N}$$

donc

$$n_{i,j} = \frac{n_{i,.} n_{.,j}}{N}$$

Distance du Khi-2

But : Déterminer si X et Y sont indépendantes (ou proches de l'indépendance).
Observation : Si c'était le cas, on aurait théoriquement

$$T_{i,j} := n_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{N}.$$

Distance du Khi-2 de (X, Y) au cas indépendant

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{r} \frac{(n_{i,j} - T_{i,j})^2}{T_{i,j}}$$

où $T_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{N}$ est l'effectif théorique de la caractéristique (x_i, y_j) en faisant l'hypothèse que X et Ysont effectivement indépendantes.

Proposition

$$\chi^2 \in [0, \chi^2_{max}]$$

où $\chi^2_{max} = N \times \min(l-1, r-1)$

Coefficient de Cramér et interprétation

Observation :

- si X et Y sont indépendantes, $\chi^2 = 0$
- si $Y = X$ alors $\chi^2 = \chi^2_{\text{max}}$

Coefficient de Cramér

$$C = \sqrt{\frac{\chi^2}{\chi^2_{\text{max}}}} \in [0, 1]$$

Interpétation : On considérera que la dépendance entre les variables est :

1. très faible si $C \in [0; 0,045]$,
2. faible si $C \in]0,045; 0,09]$,
3. moyenne si $C \in]0,09; 0,18]$,
4. forte si $C \in]0,18; 0,36]$,
5. très forte si $C \in]0,36; 1]$.

Exemple

On a relevé les couleurs des yeux X et pointures Y des clientes ayant acheté une paire de chaussures dans un grand magasin un jour donné. Le résultats ont étés consignés dans le tableau suivant.

X \ Y	37	38	39	40
marrons	5	20	11	3
verts	2	10	10	3
bleus	5	11	15	5

Exemple

On complète ce tableau en calculant les effectifs marginaux et l'effectif total N.

X \ Y	37	38	39	40	$n_{i,.}$
marrons	5	20	11	3	39
verts	2	10	10	3	25
bleus	5	11	15	5	36
$n_{.,j}$	12	41	36	11	$N = 100$

On dresse le tableau des eff. théoriques $T_{i,j} = \frac{n_{i,.} n_{.,j}}{N}$ sous l'hypothèse d'indépendance.

X \ Y	37	38	39	40	$n_{i,.}$
marrons	$\frac{39 \times 12}{100} = 4,68$	15,99	14,04	4,29	39
verts	3	10,25	9	2,75	25
bleus	4,32	14,76	12,96	3,96	36
$n_{.,j}$	12	41	36	11	$N = 100$

$$\chi^2 = \sum_{i=1}^{l} \sum_{j=1}^{r} \frac{(n_{i,j} - T_{i,j})^2}{T_{i,j}} = \frac{(5-4,68)^2}{4,68} + \frac{(20-15,99)^2}{15,99} + \cdots + \frac{(5-3,96)^2}{3,96} \simeq 4,21$$

$$\chi^2_{\max} = 100 \times \min(3-1, 4-1) = 200 \text{ donc } C = \sqrt{\frac{\chi^2}{\chi^2_{\max}}} \simeq \sqrt{\frac{4,21}{200}} \simeq 0,15$$

C° : la couleur des yeux et la pointure sont des variables ayant une dépendance moyenne.

Test du χ^2 d'indépendance – la démarche

Étape 1 : Définition des hypothèses :
On définit les hypothèses :

H_0 (hypothèse nulle) : Les variables X et Y sont indépendantes.

H_1 (hypothèse alternative) : Les variables X et Y ne sont pas indépendantes.

Étape 2 : Calcul des effectifs théoriques sous H_0 :
On dresse le tableau des effectifs théoriques en utilisant que :

$$T_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{N},$$

où N est l'effectif total.

Note importante : Si certains effectifs théoriques sont inférieurs à 5, on doit regrouper des lignes ou des colonnes pour que ce ne soit plus le cas, sans quoi l'approximation suivante par la loi du χ^2 n'est pas valable. Si des regroupements sont effectués pour les effectifs théoriques, on réalise les mêmes regroupements pour les effectifs observés.

Test du χ^2 d'indépendance – la démarche

Étape 3 : Calcul de la distance du χ^2 observée :
On calcule :

$$\chi^2_{\text{obs}} = \sum_{i,j} \frac{(n_{i,j} - T_{i,j})^2}{T_{i,j}}.$$

Étape 4 : Degrés de liberté et loi de la statistique de test :
On détermine le nombre de degrés de liberté grâce à la formule :

$$\text{ddl} = (\text{nbc} - 1)(\text{nbl} - 1),$$

où nbc et nbl désignent respectivement le nombre de colonnes et de lignes dans le tableau après regroupements éventuels.
La loi de la statistique de test est alors simplement la loi du Khi-2 à ddl degrés de liberté, notée $\chi^2(\text{ddl})$.

Test du χ^2 d'indépendance – la démarche

Étape 5 : Risque d'erreur de première espèce et valeur critique :
On fixe un paramètre $\alpha \in [0, 1]$ appelé risque d'erreur de première espèce.
Celui-ci est la probabilité de rejeter à tord H_0, autrement dit :

$$\alpha = \mathbf{P}\left[\text{rejet de } H_0 | H_0 \text{ est vraie}\right]$$

est la probabilité de rejeter l'hypothèse d'indépendance H_0 à la fin du test sachant que H_0 est vraie.

Connaissant α et ddl, on détermine une valeur critique $\chi^2_c = \chi^2_{c,\alpha,\text{ddl}}$ telle que, si χ_2 suit la loi $\chi_2(\text{ddl})$, on a :

$$\mathbf{P}[\chi^2 > \chi^2_c] = \alpha.$$

Ceci est fait soit par lecture dans une table du χ^2, soit, à l'aide d'Excel, en appelant la fonction

$$\texttt{KHIDEUX.INVERSE}(\alpha;\texttt{ddl}).$$

Test du χ^2 d'indépendance – la démarche

Étape 6 : Règles de décision :

- Si $\chi^2_{\text{obs}} > \chi^2_c$, on rejette H_0 ;
- Si $\chi^2_{\text{obs}} \leq \chi^2_c$, on **ne** rejette **pas** H_0.

Ces règles de décisions ce comprennent comme suit. La valeur du Khi-2 observée χ^2_{obs} mesure une distance entre le tableau des effectifs observés et celui des effectifs théoriques sous l'hypothèse H_0. Ainsi, si les deux tableaux sont significativement éloignés pour cette distance, on rejettera le fait qu'il correspondent à une même situation et donc l'hypothèse d'indépendance.

Étape 7 : Conclusion : On conclue en utilisant les Étapes 3, 5 et 6.

Test du χ^2 d'indépendance – un exemple

Contexte :
Un mois après le lancement d'une campagne de publicité, le service marketing d'une entreprise souhaite savoir si la sa campagne a été perçue aussi bien sur l'ensemble du territoire de Bourgogne/Franche-Comté ou non. Pour cela, il a interrogé un panel de 155 habitants sur leur département de résidence X et le fait Y qu'ils aient vu la publicité ou non. Les résultats ont étés consignés dans le tableau suivant.

X \ Y	Oui	Non
Côte-d'Or	19	9
Doubs	20	9
Jura	8	6
Nièvre	5	6
Haute-Saône	4	8
Saône-et-Loire	5	23
Yonne	8	10
Territoire de Belfort	7	8

Test du χ^2 d'indépendance – un exemple

Étape 1 : Définition des hypothèses :

On définit les hypothèses :

H_0 (hypothèse nulle) : Les variables X et Y sont indépendantes :
la campagne a été perçue de la même façon sur l'ensemble du territoire

H_1 (hypothèse alternative) : Les variables X et Y ne sont pas indépendantes :
la perception de la campagne dépend de la localisation géographique

Test du χ^2 d'indépendance – un exemple

Étape 2 : Calcul des effectifs théoriques sous l'hypothèse H_0 :

En utilisant que

$$T_{i,j} = \frac{n_{i,\cdot} n_{\cdot,j}}{N},$$

on obtient les tableau des effectifs théoriques :

X \ Y	Oui	Non	$n_{i,\cdot}$
Côte-d'Or	$\frac{28\times 72}{155} = 13,73$	14,27	28
Doubs	14,22	14,78	29
Jura	6,862	7,14	14
Nièvre	5,40	5,60	11
Haute-Saône	5,88	6,12	12
Saône-et-Loire	13,73	14,27	28
Yonne	8,83	9,17	18
Territoire de Belfort	7,35	7,65	15
$n_{\cdot,j}$	76	79	$N = 155$

Test du χ^2 d'indépendance – un exemple

Étape 3 : Calcul de la distance du χ^2 observée :
On a :

$$\chi^2_{\text{obs}} = \sum_{i,j} \frac{(n_{i,j} - T_{i,j})^2}{T_{i,j}} \simeq 21,2640.$$

Étape 4 : Degrés de liberté et loi de la statistique de test :
On a :

$$\text{ddl} = (\text{nbc} - 1)(\text{nbl} - 1) = (8-1) \times (2-1) = 7,$$

et la statistique de test suit la loi $\chi^2(7)$.

Étape 5 : Risque d'erreur de première espèce et valeur critique :
En choisissant le risque $\alpha = 1\% = 0,01$ et en utilisant que ddl=7, on obtient :

$$\chi^2_c \simeq 18,4753.$$

Test du χ^2 d'indépendance – un exemple

Étape 6 : Règles de décision :

- Si $\chi^2_{\text{obs}} > \chi^2_c \simeq 18,4753$, on rejette H_0 ;
- Si $\chi^2_{\text{obs}} \leq \chi^2_c \simeq 18,4753$, on **ne** rejette **pas** H_0.

Étape 7 : Conclusion :

On a :

$$\chi^2_{\text{obs}} \simeq 21,2640 > \chi^2_c \simeq 18,4753$$

donc on rejette H_0 au risque d'erreur de première espèce de 1%.
Ainsi, on conclue que la campagne de publicité a été perçue de façon différente sur l'ensemble du territoire de Bourgogne/Franche-Comté au risque de première espèce de 1%.
Moins formellement, il y a moins d'un pourcent de chance d'avoir rejeter l'hypothèse d'indépendance alors qu'elle était vraie, ou encore de chercher à expliquer pourquoi la campagne de publicité aurait été perçue de façon différente sur le territoire alors que ce n'est pas le cas (et dons de risquer de fournir un travail supplémentaire inutile !).

Covariance

Covariance de (X, Y)

$$\mathrm{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{r} n_{i,j}(x_i - \overline{X})(y_j - \overline{Y})$$

où $\overline{X} = \frac{1}{N} \sum_i n_{i,.} x_i$, $\overline{Y} = \frac{1}{N} \sum_j n_{.,j} y_j$ sont les moyennes marginales en X et Y.

- Si des regroupements en classes sont utilisés, on remplace dans la formule précédente les x_i et/ou y_j par les centres des classes correspondantes.

Proposition

1. $\mathrm{Cov}(aX + b, Y) = a\,\mathrm{Cov}(X, Y)$, *pour tous* $a, b \in \mathbf{R}$
2. $\mathrm{Cov}(X, X) = \mathrm{V}[X]$
3. *Si X et Y sont indépendantes, alors* $\mathrm{Cov}(X, Y) = 0$.
4. *On a :*

$$\mathrm{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{r} n_{i,j} x_i y_j - \overline{XY}.$$

Coefficient de corrélation linéaire

Coefficient de corrélation linéaire de (X, Y)

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}[X]\,\mathrm{V}[Y]}}.$$

Proposition

1. $\mathrm{Cor}(X, Y) \in [-1, 1]$
2. $\mathrm{Cor}(X, X) = 1$, $\mathrm{Cor}(X, -X) = -1$
3. *Si X et Y sont indépendantes, alors* $\mathrm{Cor}(X, Y) = 0$.

Interprétation : On dira qu'il y a :

- une forte corrélation entre X et Y si $|\mathrm{Cor}(X, Y)| \geq 0,8$;
- une corrélation médiocre entre X et Y si $0,5 < |\mathrm{Cor}(X, Y)| < 0,8$;
- une mauvaise corrélation entre X et Y si $|\mathrm{Cor}(X, Y)| \leq 0,5$.

Mise en garde

Un lien fort entre deux variables entraîne a priori une forte corrélation de celles-ci mais une forte corrélation ne suffit pas pour établir un lien de cause à effet entre deux variables.

Il faut être attentif à ne pas mal interpréter ou surinterpréter les résultats d'une étude statistique et à garder un sens critique.

Exemple : Une étude a montré que les personnes résidant à proximité d'une centrale nucléaire sont significativement plus souvent malades que les autres.

On ne peut pour autant pas affirmer que des problèmes de fuites ou autres au niveau des centrales sont (seules) responsables ces maladies.

On peut remarquer que les terrains situés dans ces zones sont généralement très bon marché.

La santé et la pauvreté n'étant pas sans lién, l'étude ne permet pas, à elle seule, de conclure que les centrales nucléaires influent sur la santé.

Exemple (appartements)

Une agence immobilière a relevé les surfaces et prix des 10 appartements qu'elle a vendus une semaine donnée et les a consignés dans le tableau suivant :

n° i de l'appartement	Surface x_i en m^2	Prix y_i en K€
1	20	47
2	122	230
3	45	87
4	56	98
5	18	39
6	54	90
7	77	180
8	80	176
9	68	124
10	32	45

$$\overline{X} = \frac{20 + \cdots + 32}{10} = 57,2, \qquad \overline{Y} = \frac{47 + \cdots + 45}{10} = 111,6$$

$$\mathrm{V}[X] = \frac{(20 - 57,2)^2 + \cdots + (32 - 57,2)^2}{10} = 894,36$$

$$\mathrm{V}[Y] = \frac{(47 - 111,6)^2 + \cdots + (45 - 111,6)^2}{10} = 3813,44$$

$$\mathrm{Cov}(X, Y) = \frac{(20 - 57,2)(47 - 111,6) + \cdots + (32 - 57,2)(45 - 111,6)}{10} = 1794,18$$

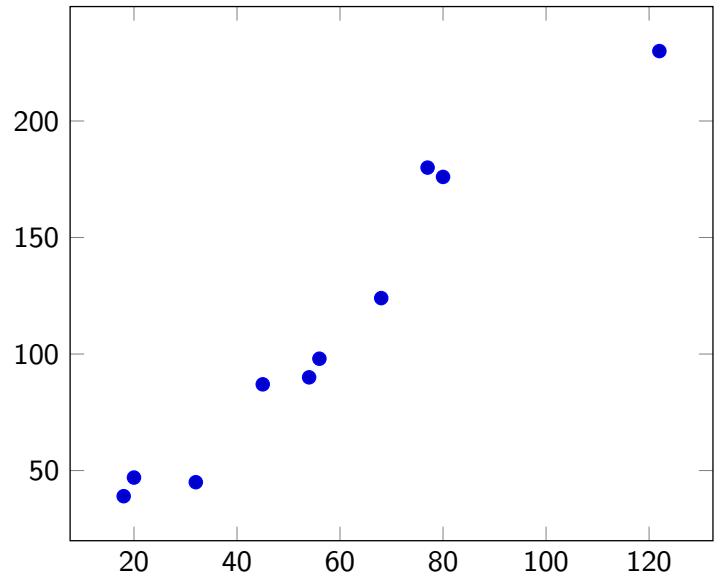
et donc

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{V}[X]\,\mathrm{V}[Y]}} \simeq 0,97.$$

Exemple (appartements)

Une agence immobilière a relevé les surfaces et prix des 10 appartements qu'elle a vendus une semaine donnée et les a consignés dans le tableau suivant :

n° i de l'appartement	Surface x_i en m^2	Prix y_i en K€
1	20	47
2	122	230
3	45	87
4	56	98
5	18	39
6	54	90
7	77	180
8	80	176
9	68	124
10	32	45



1 Objectifs

2 Présentation et traitement des données

3 Étude et mesure des liens entre les variables

4 Régression

Régression

But : Si X et Y sont fortement corrélées, chercher une fonction f telle que $Y = f(X)$ ou $X = f(Y)$.

- si on cherche à prédire Y en supposant X connue ($Y = f(X)$), on parle de regression de Y en X.
- si on cherche à prédire X en supposant Y connue ($X = f(Y)$), on parle de regression de X en Y
- on se limitera au cas f linéaire c-à-d que l'on cherchera une relation entre X et Y de la forme

$$Y = aX + b \qquad \text{(régression de } Y \text{ en } X\text{)}$$

ou

$$X = a'Y + b' \qquad \text{(régression de } X \text{ en } Y\text{)}.$$

Conditions d'application du modèle de régression linéaire

On effectuera une telle régression si :

- le nuage de points semble se répartir le long d'une droite,
- il est raisonnable de vouloir expliquer/prédire le caractère associé à la variable Y et par celui associé à X ou vice versa,
- le coefficient de corrélation $\text{Cor}(X, Y)$ est compris dans $[-1; -0,7] \cup [0,7; 1]$.

Trouver les meilleures droites en quel sens ?

Au sens des moindres carrés

1. Droite de régression de Y en X (au sens des moindres carrés) : droite $D_{Y|X} : y = ax + b$ avec a, b choisis de telle sorte que

$$E(a, b) = \sum_{i=1}^{N} (y_i - (ax_i + b))^2$$

soit minimale.

2. Droite de régression de X en Y (au sens des moindres carrés) : droite $D_{X|Y} : x = a'y + b'$ avec a', b' choisis de telle sorte que

$$E(a', b') = \sum_{i=1}^{N} (x_i - (a'y_i + b'))^2$$

soit minimale.

Équations des droites de régression dans la pratique ?

Proposition

1. *L'équation de la droite de régression de Y en X est :*

$$D_{Y|X} : y = ax + b,$$

avec $a = \frac{\mathrm{Cov}(X,Y)}{\mathrm{V}[X]}$ *et* $b = \overline{Y} - a\overline{X}$.

2. *L'équation de la droite de régression de X en Y est :*

$$D_{X|Y} : x = a'y + b',$$

avec $a' = \frac{\mathrm{Cov}(X,Y)}{\mathrm{V}[Y]}$ *et* $b' = \overline{X} - a'\overline{Y}$.

Exemple (appartements, suite)

- $|\operatorname{Cor}(X, Y)| \simeq 0,97 \geq 0,7$
- le nuage de points à une forme allongée, proche d'une droite
- il est sensé d'expliquer le prix d'un appartement par sa surface

On peut donc utiliser le modèle de régression linéaire.

Exemple (appartements, suite)

Droite de régression de Y en X : $y = ax + b$ avec

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}[X]} \simeq 2 \quad \text{et} \quad b = \overline{Y} - a\overline{X} \simeq -2,8.$$

Si un nouvel appartement dont la surface est de $x^* = 51m^2$ est proposé à la vente, cette droite de régression permet de prédire un prix de vente de :

$$y^* = ax^* + b \simeq 2 \times 51 - 2,8 = 99,2\text{K€}.$$

Exemple (appartements, suite)

Droite de régression de Y en X : $y = ax + b$ avec

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}[X]} \simeq 2 \quad \text{et} \quad b = \overline{Y} - a\overline{X} \simeq -2,8.$$

Droite de régression de X en Y : $x = a'y + b'$ avec

$$a' = \frac{\mathrm{Cov}(X, Y)}{\mathrm{V}[Y]} \simeq 0,47 \quad \text{et} \quad b' = \overline{X} - a'\overline{Y} \simeq 4,748.$$