UE4 Évaluation des méthodes d'analyses appliquées aux sciences de la vie et de la santé

Liaison entre deux variables quantitatives - corrélation linéaire

Frédéric Mauny – 31 octobre 2013

Plan du cours

Relation entre 2 variables quantitatives

Rappel V. quantitatives

- Age, poids, taille d'une tumeur, dosage d'un paramètre biologique, durée d'exposition, concentration sérique d'un médicament

Question posée

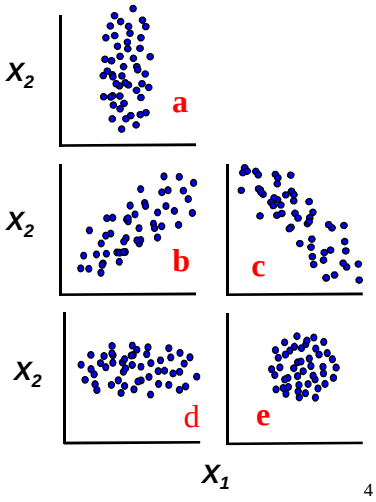
- Etudier la relation entre deux variables quantitatives
- Mesurer l'intensité avec laquelle deux variables évoluent ensemble
- Tester l'hypothèse d'un lien entre…

3

Représentation graphique : Le nuage de points

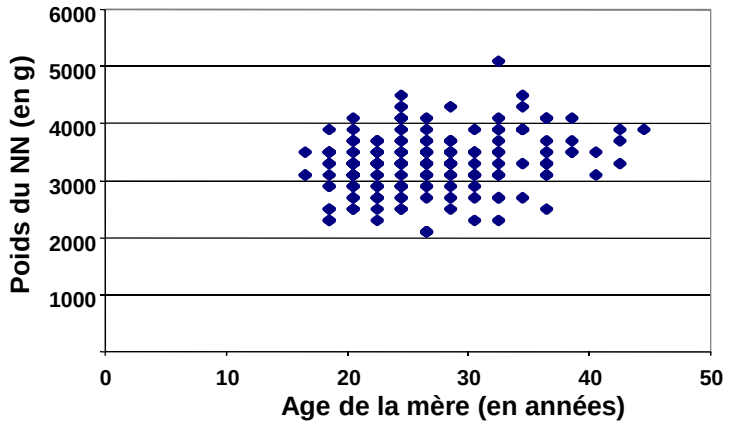
Principe :

- Soient X_1 et X_2 deux v. quantitatives
- A chaque sujet *i* est associé une valeur observée pour X_1 et une valeur observée pour X_2
- A chaque sujet *i* est associé un couple de valeurs (x_1, x_2)
- On représente chaque sujet par un point de coordonnées (x_1, x_2)



4

Relation entre 2 variables quantitatives



5

Plan du cours

1. Position du problème
2. La covariance
3. La corrélation linéaire
4. Conditions de validité

6

La Covariance de X et Y

- Notée $Cov(X, Y)$ ou σ_{XY}

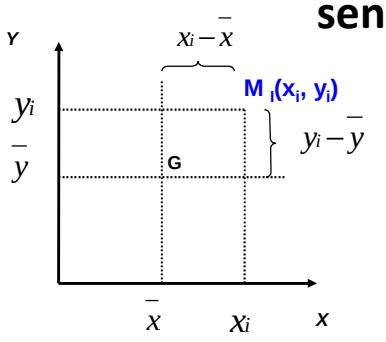
$$Cov(X, Y)=E[(X-\mu_X)(Y-\mu_Y)]$$

- Valeur moyenne des écarts de X et Y à leur moyenne
- $Cov(X, X)=var(X)$

- Covariance <u>observée</u> sur un échantillon :

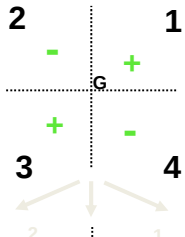
$$\text{cov}(X,Y)=\frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}=\frac{\sum xy-\Sigma x\Sigma y/n}{n-1}$$

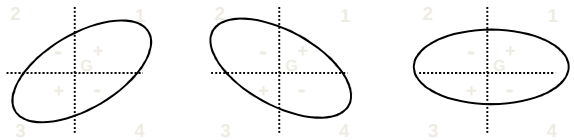
Numérateur de covariance et sens de la liaison entre X et Y



G, le centre de gravité du nuage

Signe du produit $(x_i-\bar{x})(y_i-\bar{y})$





Plan du cours

1. Position du problème
2. La covariance
3. La corrélation linéaire
   - Principe
   - Test d'hypothèse
   - Applications numériques
4. Conditions de validité

Corrélation linéaire entre X et Y

- Quantifiée par le coefficient de corrélation linéaire, noté $\rho(X, Y)$

$$\rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

- $-1 \leq \rho \leq 1$
- Si X et Y sont indépendantes, alors $\rho = 0$
- Si $\rho = 0$, et si X et Y sont distribuées normalement, alors X et Y sont indépendantes.

Test d'hypothèse

- Principe : le coefficient de corrélation diffère-t-il significativement de 0 ?
- Hypothèses
  H_0: $\rho(X, Y) = 0$
  H_1 : $\rho(X, Y) \neq 0$
- Seuil de rejet de H_0 = table statistique
- Calcul du coefficient estimé à partir des données de l'échantillon
- Interprétation

Coefficient de corrélation estimé

Coefficient de corrélation estimé, noté **r** :

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} \quad \Rightarrow \quad r = \frac{\sum xy - \Sigma x \Sigma y / n}{\sqrt{(\sum x^2 - \frac{(\Sigma x)^2}{n})(\Sigma y^2 - \frac{(\Sigma y)^2}{n})}}$$

- $r \in [-1 ; +1]$

TESTS

- Table $\rightarrow$ prob. que |r| dépasse une valeur seuil déterminée en fonction du nombre de degrés de liberté ($v = n-2$ *ddl*)
- Sous H_0, la quantité $t_{n-2ddl} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ suit une loi de *Student* à *n-2 ddl*

Règle de décision

- Table de coefficient de corrélation assez précise *(v = n-2 ddl),* on compare | *r* | à la valeur seuil lue dans la table du coefficient de corrélation linéaire
  - | *r* | dépasse la valeur seuil pour $\alpha = 0{,}05$
    - ➲ on rejette H_0, on accepte H_1
  - | *r* | ne dépasse pas la valeur seuil pour $\alpha = 0{,}05$
    - ➲ on ne rejette pas H_0
- Table de coefficient de corrélation pas assez précise, *on compare* $t_{n-2ddl} = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$ à la valeur seuil lue dans la table du *t* de *Student*
  - t_{n-2dd} dépasse la valeur seuil ➲ on rejette H_0, on accepte H_1
  - t_{n-2dd} ne dépasse pas la valeur seuil ➲ on ne rejette pas H_0

Exemple 1

Pour répondre à la question : « Existe-il un lien entre l'âge et le cholestérol plasmatique ? », on effectue chez 47 sujets le dosage plasmatique du cholestérol total et on note l'âge de ces patients.

1. Hypothèses

   H_0 : ρ(*Age*, *Cholestérol total*) = 0

   H_1 : ρ (*Age*, Cholestérol total) ≠ 0
2. On calcul le coefficient de corrélation linéaire : r=0,31
3. Que peut-on en conclure ?

Table du Coefficient de Corrélation Linéaire

Table r de Bravais Pearson

ν / α	0.10	0.05	0.02	ν / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	16	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	17	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	18	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	19	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	20	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	25	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	30	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	35	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	40	0.2573	0.3044	0.3578
10	0.4973	0.5760	0.6581	45	0.2428	0.2875	0.3384

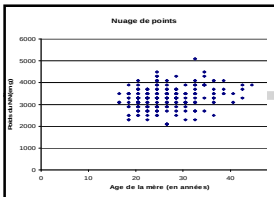
Exemple 1

v = n-2 ddl , v = 45 ddl r = 0,31

0,05 > p > 0,02

v / α	0.10	0.05	0.02	v / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	16	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	17	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	18	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	19	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	20	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	25	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	30	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	35	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	40	0.2573	0.3044	0.3578
10	0.4973	0.5760	0.6581	45	0.2428	0.2875	0.3384
11	0.4762	0.5529	0.6339	50	0.2306	0.2732	0.3218
12	0.4575	0.5324	0.6120	60	0.2108	0.2500	0.2948
13	0.4409	0.5139	0.5923	70	0.1954	0.2319	0.2737
14	0.4259	0.4973	0.5742	80	0.1829	0.2172	0.2565
15	0.4124	0.4821	0.5577	90	0.1726	0.2050	0.2422
				100	0.1638	0.1946	0.2301

Exemple 2



n=200 couples (X=âge mère, Y=poids de naissance)

Age de la mère

- Σx=5 298 Σx²=147 732, $\bar{x} = 26{,}5$ ans

Poids du NN

- Σy=661 800 Σy²=223 848 0000, $\bar{y} = 3309$ g
- Σxy=17 676 700

$$r = \frac{17676700 - 661800 * 5298/200}{\sqrt{(2238480000 - \frac{(661800)^2}{200})(147732 - \frac{(5298)^2}{200})}} =$$

Exemple 2

v = n-2 ddl , *v = 198 ddl* r = 0,24

v / α	0.10	0.05	0.02	v / α	0.10	0.05	0.02
1	0.9877	0.9969	0.9995	16	0.4000	0.4683	0.5425
2	0.9000	0.9500	0.980	17	0.3887	0.4555	0.5285
3	0.8054	0.8783	0.9343	18	0.3783	0.4438	0.5155
4	0.7293	0.8114	0.8822	19	0.3687	0.4329	0.5034
5	0.6694	0.7545	0.8329	20	0.3598	0.4227	0.4921
6	0.6215	0.7067	0.7887	25	0.3233	0.3809	0.4451
7	0.5822	0.6664	0.7498	30	0.2960	0.3494	0.4093
8	0.5494	0.6319	0.7155	35	0.2746	0.3246	0.3810
9	0.5214	0.6021	0.6851	40	0.2573	0.3044	0.3578
10	0.4973	0.5750	0.6581	45	0.2428	0.2875	0.3384
11	0.4762	0.5529	0.6339	50	0.2306	0.2732	0.3218
12	0.4575	0.5324	0.6120	60	0.2108	0.2500	0.2948
13	0.4409	0.5139	0.5923	70	0.1954	0.2319	0.2737
14	0.4259	0.4973	0.5742	80	0.1829	0.2172	0.2565
15	0.4124	0.4821	0.5577	90	0.1726	0.2050	0.2422
				100	0.1638	0.1946	0.2301

→ ?

$$t = \frac{0,24}{\sqrt{1-0,24^2}}\sqrt{198} = 3,5$$

v = n-2 ddl , *v = 198 ddl*

TABLE DU t DE STUDENT

ddl \ α	0,45	0,25	0,15	0,10	0,05	0,025	0,01	0,005	0,0005	TU
ddl \ α	0,90	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,001	TB
1	0,158	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619	
2	0,142	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598	
17	0,128	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965	
18	0,127	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922	
19	0,127	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883	
20	0,127	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850	
21	0,127	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819	
22	0,127	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792	
23	0,127	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767	
24	0,127	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745	
25	0,127	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725	
26	0,127	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707	
27	0,127	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690	
28	0,127	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674	
29	0,127	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659	
30	0,127	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646	
∞	0,126	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291	

→

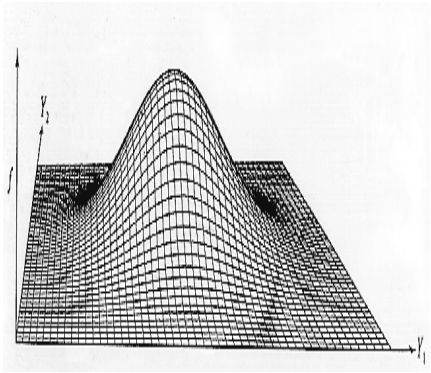
p<0,001

Plan du cours

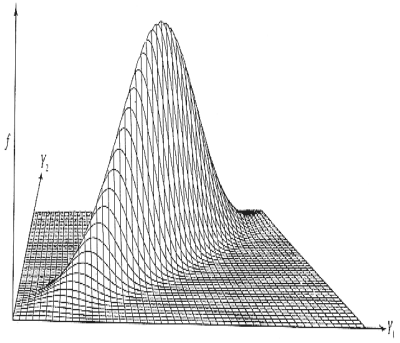
1. Position du problème
2. La covariance
3. La corrélation linaire
4. Conditions de validité
   - Normalité des distributions
   - Homoscédasticité
   - Relation linéaire entre les deux variables

Distribution des variables selon une loi normale

- Pour chaque valeur de X_1, les valeurs de X_2 sont normalement distribuées et *vice versa.*



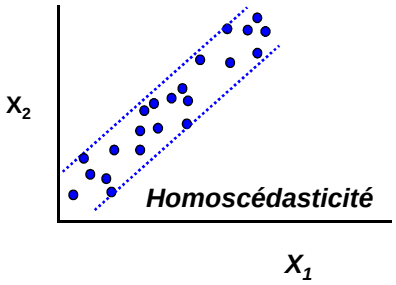
r = 0

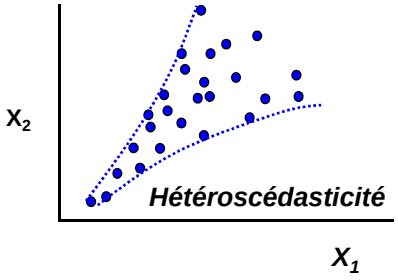


r = 0,8

Homoscédasticité

La variance de X_2 est constante pour toute valeur de X_1

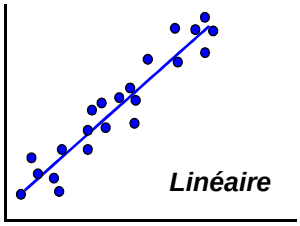


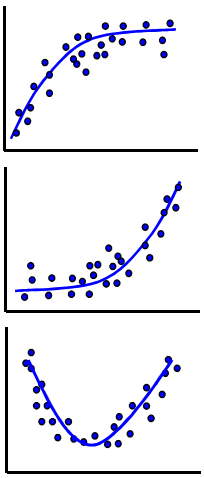


21

Linéarité de la relation

- La relation entre X_1 et X_2 est linéaire.

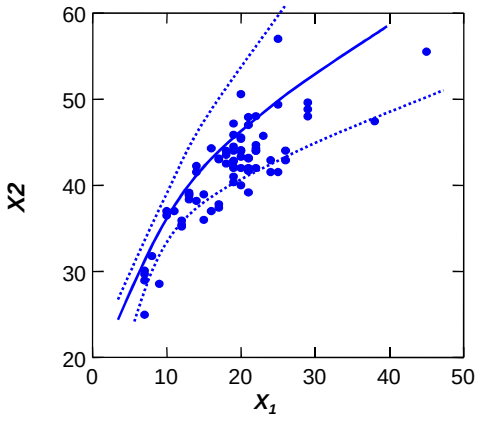




22

Conditions d'application non respectées…

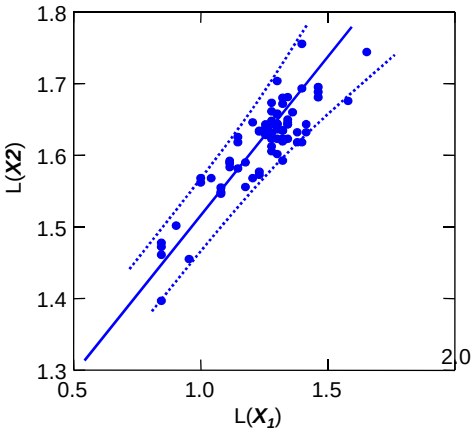
- La relation entre les variables X_1 et X_2 semble non-linéaire.
- La variance de X_2 semble augmenter quand X_1 augmente



23

… solutions

- Transformer les données (ex: Log).
- Test de corrélation non-paramétrique



24

Conclusion

- Identifier le problème statistique
  - Question posée
  - Nature des variables
  - Représentation graphique de données
  - Réalisation des différents phases du test
  - Respect des conditions d'application