Traitement des données en Psychologie

UV PSY38X2

Présentation du cours 2001/2002

Organisation matérielle

Cours magistral : 12 heures
A peu près 7 heures de statistiques et
5 heures d'informatique
lundi 16h-17h - Salle B222

Travaux dirigés :
1 groupe de TD de statistiques.
lundi 17h-18h - B222
2 sous-groupes de TD en Informatique.
salle A204 ou A206, 2 h. par quinzaine
sous-gr. 1 - mercredi 8h15-10h15 - sem A
sous-gr. 2 - mercredi 8h15-10h15 - sem B
Monitorat informatique

Contrôle des connaissances : (contrôle continu)
70 % Examen écrit (3 heures)
30 % Evaluation de TD (épreuve machine)

Bibliographie

- G. Mialaret. Statistiques appliquées aux sciences humaines. PUF
- B. Beaufils. Statistiques appliquées à la psychologie. Tomes 1 et 2. LEXIFAC Bréal
- N. Guéguen. Manuel de statistiques pour psychologues
- D.C. Howell. Méthodes statistiques en sciences humaines
- D. Laveault, J. Grégoire. Introduction aux théories des tests en sciences humaines. De Boeck Université
- J.P. Rossi. La méthode expérimentale en Psychologie
- H. Abdi. Introduction au traitement statistique des données expérimentales. PUG
- P. Rateau. Méthode et statistique expérimentales en sciences humaines. Ellipses

Contenu

Documents fournis :

Copie des transparents en statistiques
Polycopié du cours en Informatique
Fiches de TD
Sites internet permettant de télécharger ces documents :
Depuis les salles de TD d'informatique :
`http://letsamba.univ-brest.fr/~carpenti/`
Depuis le reste de l'Université (domaine univ-brest.fr) :
`http://infolettres.univ-brest.fr/~carpenti/`
Depuis l'extérieur, ou depuis le domaine univ-brest.fr :
`http://geai.univ-brest.fr/~carpenti/`

N.B. Documents (autres que les fiches de TD d'informatique) au format .pdf lisible par Acrobat Reader

Programmes, contrôle des connaissances, documents plus anciens :
`http://geai.univ-brest.fr/enseignements.html`

Statistiques :

Corrélation linéaire. Régression linéaire
F de Fisher. Analyse de variance.
Introduction aux plans d'expériences

Informatique :

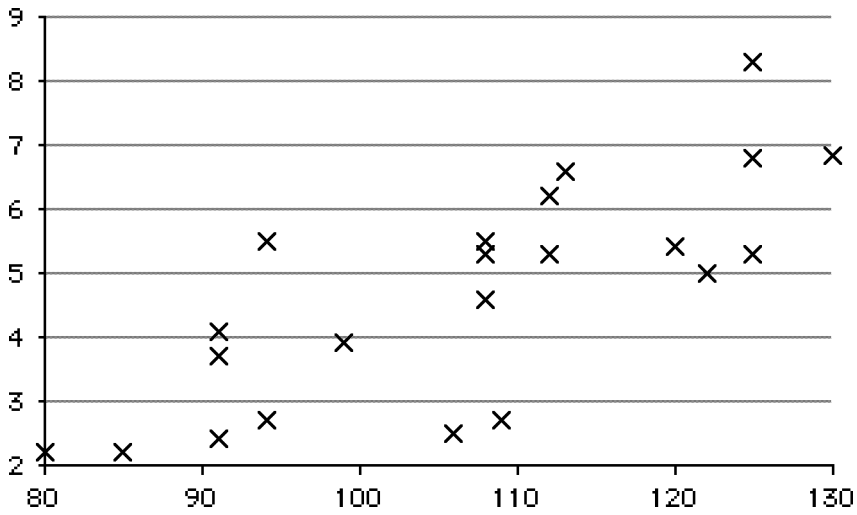
Modèles conceptuels de données
Bases de données

Corrélation linéaire

Données :

	X	Y
s_1	x_1	y_1
s_2	x_2	y_2
...	...	...

Nuage de points : points (x_i, y_i)



Covariance des variables X et Y

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

ou

$$Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \overline{x}\ \overline{y}$$

Coefficient de corrélation de Bravais Pearson

$$r = \frac{Cov(X,Y)}{s(X)s(Y)}$$

Remarques

– Formules analogues avec corrélation, écarts types, . . . corrigés

– Il existe des relations non linéaires

– Corrélation n'est pas causalité

Significativité du coefficient de corrélation

- Les données (x_i, y_i) constituent un échantillon
- r est une statistique
- ρ : coefficient de corrélation sur la population

H_0 : Indépendance sur la population ; $\rho = 0$
H_1 : $\rho \neq 0$ (bilatéral) ou $\rho > 0$ ou $\rho < 0$ (unilatéral)

Statistique de test

– Petits échantillons : tables spécifiques. $ddl = n - 2$
– Grands échantillons :

$$T = \sqrt{n-2}\ \frac{r}{\sqrt{1-r^2}}$$

T suit une loi de Student à $n - 2$ degrés de liberté.

Régression linéaire

Rôle "explicatif" de l'une des variables par rapport à l'autre. Les variations de Y peuvent-elles (au moins en partie) être expliquées par celles de X ? Peuvent-elles être prédites par celles de X ?

Modèle permettant d'estimer Y connaissant X

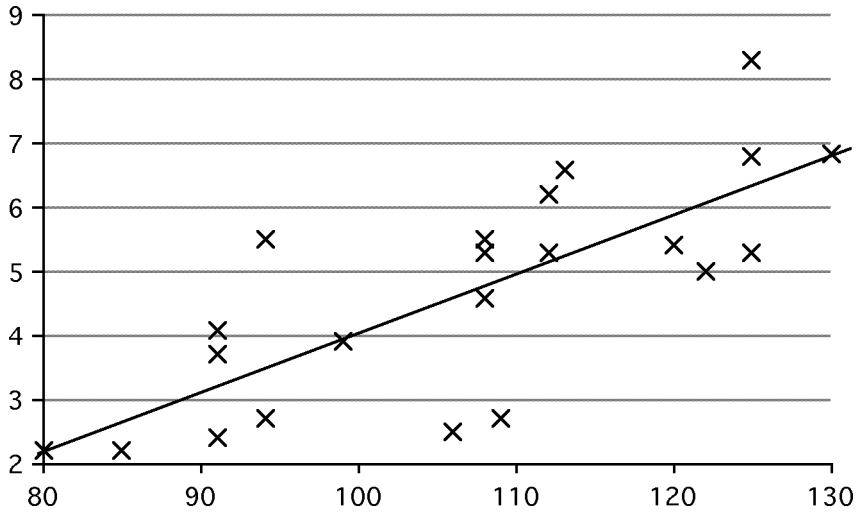
Droite de régression de Y par rapport à X :

La droite de régression de Y par rapport à X a pour équation:

$$y = ax + b$$

avec :

$$a = \frac{Cov(x_i, y_i)}{s^2(x_i)} \quad ; \quad b = \overline{y} - a\overline{x}$$



Comparaison des valeurs observées et des valeurs estimées

Valeurs estimées : $\widehat{y}_i = ax_i + b$
Erreur (ou résidu) : $e_i = y_i - \widehat{y}_i$

On montre que :

$$s^2(Y) = s^2(\widehat{Y}) + s^2(E)$$

avec :

$$\frac{s^2(E)}{s^2(Y)} = 1 - r^2 \quad ; \quad \frac{s^2(\widehat{Y})}{s^2(Y)} = r^2$$

$s^2(\widehat{Y})$: variance *expliquée* (par la variation de X, par le modèle)

$s^2(E)$: variance *perdue* ou *résiduelle*

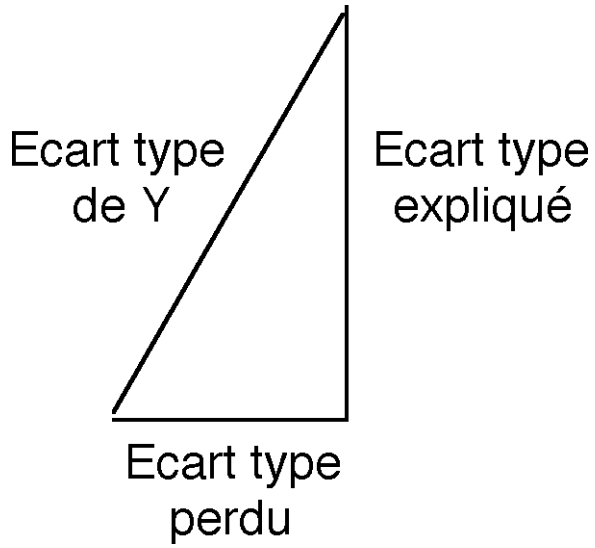
r^2 : part de la variance de Y qui est expliquée par la variance de X. *coefficient de détermination*.

Exemple : $r = 0.86$

$r^2 = 0.75$; $1 - r^2 = 0.25$; $\sqrt{1 - r^2} = 0.5$.

• La part de la variance de Y expliquée par la variation de X est de 75%.

• L'écart type des résidus est la moitié de l'écart type de Y.



Régression linéaire multiple

Position du problème

Une population (ou un échantillon) sur laquelle on a observé un ensemble de varaibles numériques.

	X_1	X_2	...	X_p
s_1	x_{11}	x_{12}	...	x_{1p}
...	...	...	...	...

Exemple avec trois variables

x_i : âge de la mère
y_i : rang de l'enfant dans la fratrie
z_i : poids de l'enfant à la naissance

Nuage de points

Pour trois variables : représentation dans l'espace.

Pour plus de trois variables, détermination des directions de "plus grande dispersion du nuage" : analyse en composantes principales.

Paramètres associés aux données

Matrice des covariances, matrice des corrélations.

Exemple : *Coefficients de corrélation des variables prises 2 à 2 :*

$r_{xz} = 0.24$ $**$: âge et poids sont corrélés
$r_{yz} = 0.28$ $**$: rang et poids sont corrélés
$r_{xy} = 0.60$ $**$: rang et âge sont fortement corrélés

$$\begin{array}{c} \\ X \\ Y \\ Z \end{array} \begin{bmatrix} X & Y & Z \\ 1 & 0.60 & 0.24 \\ 0.60 & 1 & 0.28 \\ 0.24 & 0.28 & 1 \end{bmatrix}$$

Coefficients de corrélation partielle

Corrélations obtenues en contrôlant la troisième variable

$r_{yz.x} = 0.18$ $**$: A âge constant, rangs et poids sont corrélés
$r_{xz,y} = 0.09$ NS : A rang constant, pas de corrélation entre âge et poids.

Seul le rang de naissance intervient. L'âge de la mère n'est lié au poids de l'enfant que par le rang de naissance.

"Hyperplan" de régression

L'une des variables (Z) est la variable "à prévoir". Les autres ($X_1, X_2, \ldots, X_p$) sont les variables "prédicatives".

$$Z = a_0 + a_1 X_1 + \ldots + a_p X_p$$

Si les variables sont réduites (écart type égal à 1), les coefficients a_i ($i \geq 1$) sont les coefficients de corrélation partielle $r_{ZX_i,X_1 \ldots X_p}$.

Avec trois variables :

$$Z = c + aX + bY$$

Passe par le point moyen : $c = \overline{Z} - a\overline{X} - b\overline{Y}$

Coefficient de corrélation multiple

$\hat{Z}$: valeurs estimées à l'aide de l'équation précédente.

$$R = r_{Z\hat{Z}} = \frac{Cov(Z, \hat{Z})}{s(Z)s(\hat{Z})}$$

Dans l'exemple proposé : $R = 0.29$

Comme précédemment, R^2 est la part de la variance "expliquée par le modèle".

Comparaison de deux variances
F de Fisher

Exemple. Deux tests mesurant la même aptitude
– Groupe 1 : 25 sujets. $\overline{x_1} = 40$; $s^2_{1c} = 65$
– Groupe 2 : 30 sujets. $\overline{x_2} = 38$; $s^2_{2c} = 30$

La précision est-elle la même pour les deux tests ?

Cas général

Deux échantillons de tailles n_1 et n_2 extraits de deux populations. Moyennes égales ou différentes. Distribution normale de la variable dans les populations parentes.

Problème : Les *variances* dans les populations parentes sont-elles égales ?

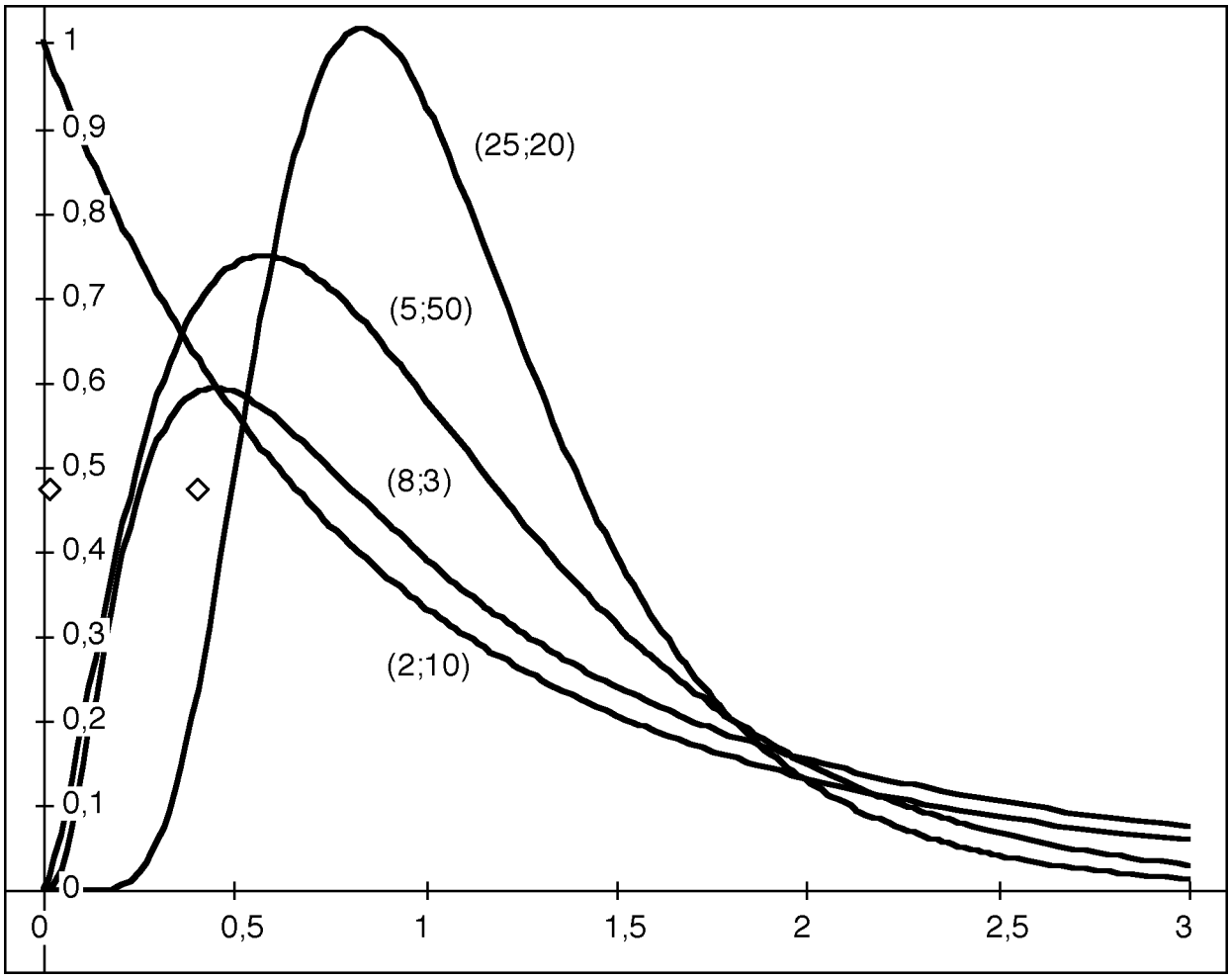
H_0 : Les variances sont égales
H_1 : La première variance est supérieure à la deuxième.

Statistique de test

$$F = \frac{s^2_{1,c}}{s^2_{2,c}}$$

F suit une **loi de Fisher à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.**

Distributions du F de Fisher



Analyse de Variance à un facteur

Exemple introductif : Test commun à trois groupes d'élèves. Moyennes observées dans les trois groupes : $\overline{x}_1 = 8$, $\overline{x}_2 = 10$,$\overline{x}_3 = 12$.

Question : s'agit-il d'élèves "tirés au hasard" ou de groupes de niveau ?

Première situation :

	Gr1	Gr2	Gr3
	6	8	10.5
	6.5	8.5	10.5
	6.5	8.5	11
	7	9	11
	7.5	9.5	11
	8	10	12
	8	11	13
	10	11	13
	10	12	14
	10.5	12.5	14
$\overline{x}_i$	8	10	12

Deuxième situation :

	Gr1	Gr2	Gr3
	5	4	6
	5.5	5.5	7
	6	7.5	9
	6	9	10
	6.5	9.5	11
	7	10	12
	7.5	11	13
	10	13	14
	12	15	18
	14.5	15.5	20
$\overline{x}_i$	8	10	12

Démarche utilisée : nous comparons la dispersion des moyennes (8, 10, 12) à la dispersion à l'intérieur de chaque groupe.

Comparer a moyennes sur des groupes indépendants

Plan d'expérience : $\mathcal{S} < \mathcal{A}_a >$

Une variable $\mathcal{A}$, de modalités $A_1, A_2, \ldots, A_a$ définit a groupes indépendants.

Variable dépendante X mesurée sur chaque sujet.
x_{ij} : valeur observée sur le i-ème sujet du groupe j.

Problème : La variable X a-t-elle la même moyenne dans chacune des sous-populations dont les groupes sont issus ?

H_0 : $\mu_1 = \mu_2 = \ldots = \mu_a$
H_1 : Les moyennes ne sont pas toutes égales.

Construction de la statistique de test :

Notations :
n_1, n_2, ..., n_a : effectifs des groupes.
N : effectif total
$T_{\cdot 1}$, ..., $T_{\cdot a}$: sommes des observations pour chacun des groupes.
$T_{\cdot\cdot}$ ou T_G : somme de toutes les observations.

Somme des carrés totale ou variation totale :

$$SC_T = \sum_{i,j} x_{ij}^2 - \frac{T_G^2}{N}$$