

## Traitement des données en Psychologie

### UV PSY38X2

#### Présentation du cours 2002/2003

#### Organisation matérielle

Cours magistral: 12 heures

A peu près 8 heures de statistiques et  
4 heures d'informatique

lundi 16h-17h - Salle A009

Travaux dirigés:

2 groupes de TD de statistiques.

Gr. 1: lundi 17h-18h - Salle A009

Gr. 2: lundi 17h-18h - Salle A212

3 sous-groupes de TD en Informatique.

salle A204 ou A206, 2 h. par quinzaine

sous-gr. X - mercredi 17h-19h - Sem. A

sous-gr. Y - mercredi 17h-19h - Sem. A

sous-gr. Z - mercredi 17h-19h - 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/>

Voir aussi :

<http://cours.univ-brest.fr/Discipline/informatique/carpentier/>

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

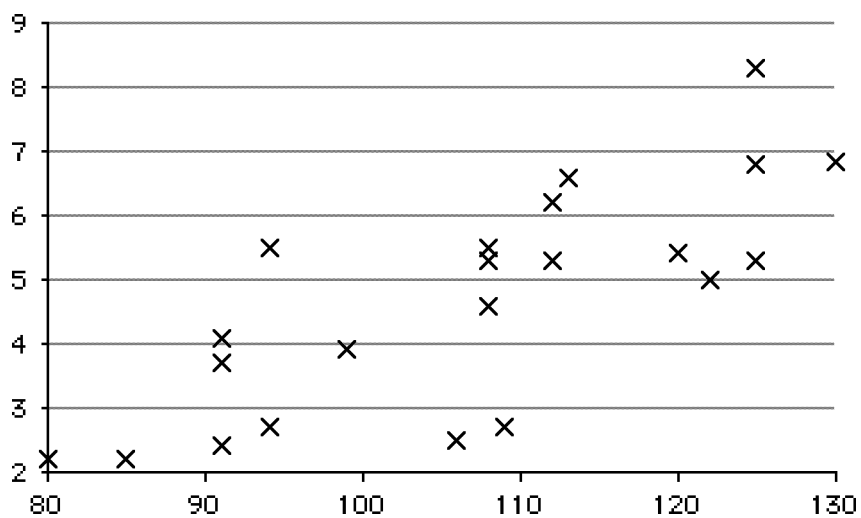
Bases de données

## 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 $Y$

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

ou

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

## 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
- $\rho$ : 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  $X$

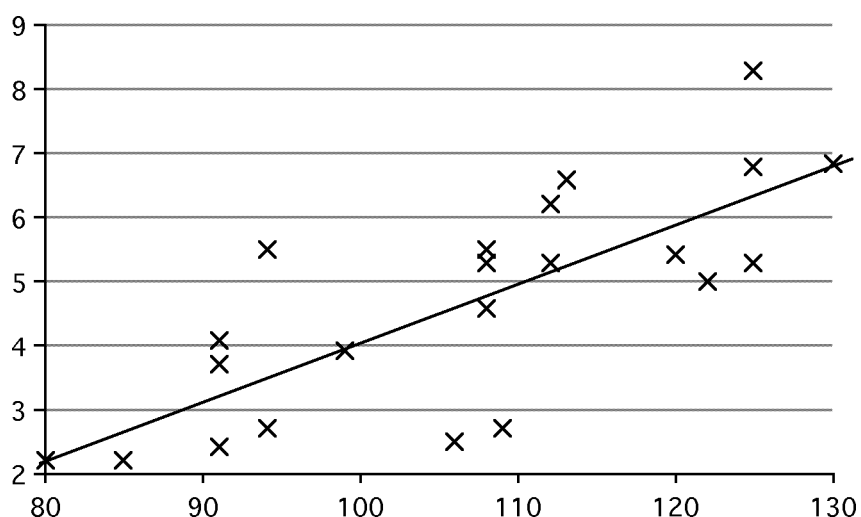
*Droite de régression de  $Y$  par rapport à  $X$  :*

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

$$y = ax + b$$

avec :

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



*Comparaison des valeurs observées et des valeurs estimées*

Valeurs estimées :  $\hat{y}_i = ax_i + b$

Erreur (ou résidu) :  $e_i = y_i - \hat{y}_i$

On montre que :

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

avec :

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

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

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

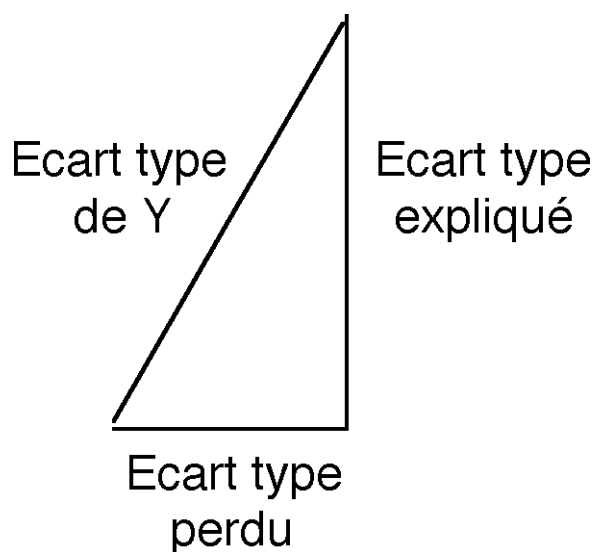
$r^2$  : part de la variance de  $Y$  qui est expliquée par la variance de  $X$ .  $r^2$  est appelé *coefficient de détermination*.



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 variables 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

$n$  : nombre d'observations (ici :  $n = 200$ )

	$X$	$Y$	$Z$
$s_1$	26.5	1	2100
...	...	...	...
$s_{200}$	34.5	2	4500

### 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 :*

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

$r_{xz} = 0.24$  \*\*: âge et poids sont corrélés

$r_{yz} = 0.28$  \*\*: rang et poids sont corrélés

$r_{xy} = 0.60$  \*\*: rang et âge sont fortement corrélés

### **“Hyperplan” de régression**

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

$$Z = a_0 + a_1X_1 + \dots + a_pX_p$$

Avec trois variables :

$$Z = c + aX + bY$$

Passé par le point moyen, c'est-à-dire :  $c = \bar{Z} - a\bar{X} - b\bar{Y}$

### *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".

### *Coefficients de corrélation partielle*

Corrélations obtenues en contrôlant la troisième variable. Pour calculer  $r_{yz.x}$ , par exemple :

- On calcule les résidus de la régression de  $Z$  par rapport à  $X$
- On calcule les résidus de la régression de  $Y$  par rapport à  $X$
- On calcule le coefficient de corrélation entre les deux séries obtenues.

$r_{yz.x} = 0.18$  \*\*: 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.

## Comparaison de deux variances F de Fisher

**Exemple.** Deux tests mesurant la même aptitude

– Groupe 1 : 25 sujets.  $\bar{x}_1 = 40$  ;  $s_{1c}^2 = 65$

– Groupe 2 : 30 sujets.  $\bar{x}_2 = 38$  ;  $s_{2c}^2 = 30$

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

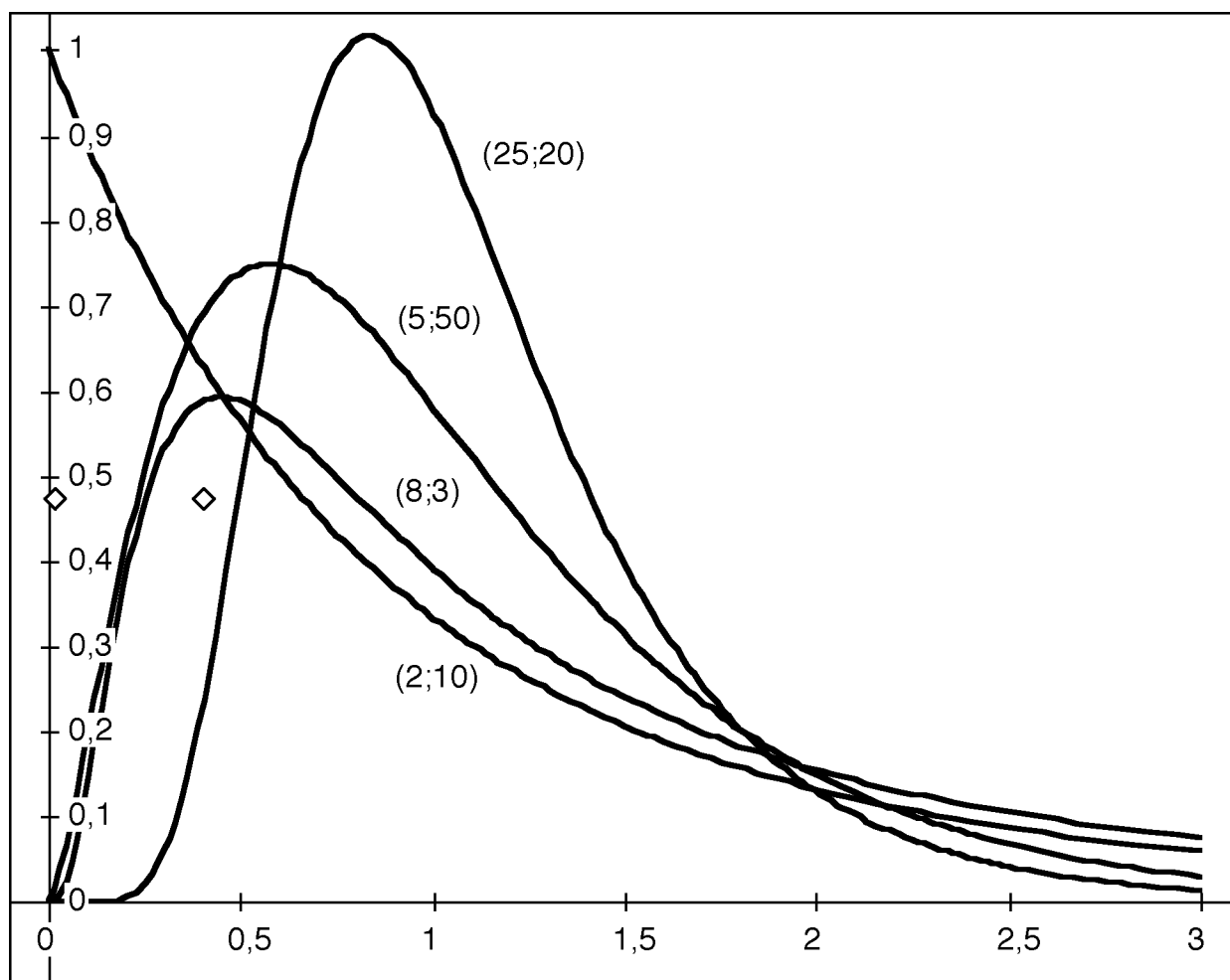
$H_1$  : La première variance est supérieure à la deuxième.

*Statistique de test*

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

$F$  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 :  $\bar{x}_1 = 8$ ,  $\bar{x}_2 = 10$ ,  $\bar{x}_3 = 12$ .

**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
$\bar{x}_i$	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
$\bar{x}_i$	8	10	12

Boîtes à moustaches pour les deux situations proposées

