Statistiques et informatique

UV PSY38X2

Présentation du cours 2000/2001

Organisation matérielle

Cours magistral : 12 heures. Traitement de données en Psychologie

A peu près 7 heures de statistiques et
5 heures d'informatique
lundi 16h-17h - Salle A214

Travaux dirigés :

2 groupes de TD de statistiques.
lundi 17h-18h - A214
jeudi 16h-17h - A219

3 sous-groupes de TD en Informatique.
salles A204, 2 h. par quinzaine
1 sous-groupe - mercredi 8h15-10h15
2 sous-groupes - mercredi 10h30-12h30

Monitorat informatique

Contrôle des connaissances : (contrôle continu)

70 % Examen écrit (3 heures)
30 % Evaluation de TD (épreuve machine)

Bibliographie

- B. Cadet Méthodes statistiques en psychologie. P.U. de Caen
- 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
- M. Reuchlin. Précis de Statistiques. PUF Coll. Le Psychologue.
- 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

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

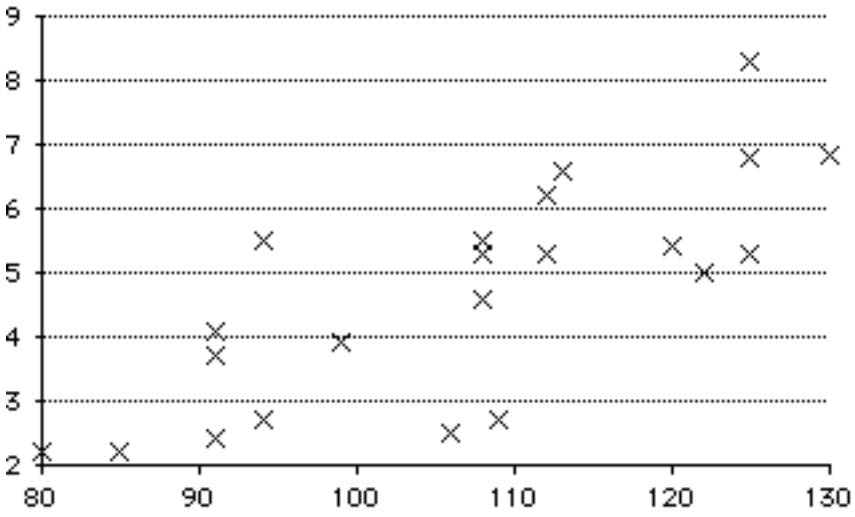
Informatique :
Modèles conceptuels de données
Bases de données

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

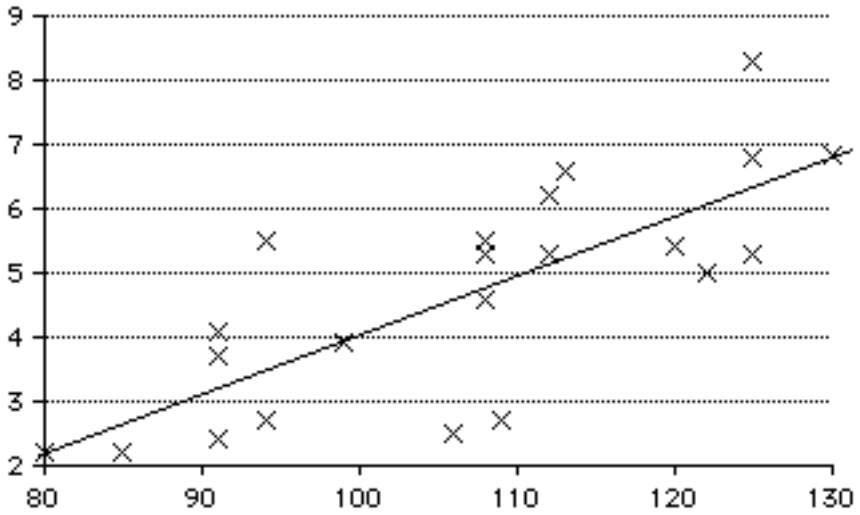
Droite de régression de Y par rapport à X :

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*

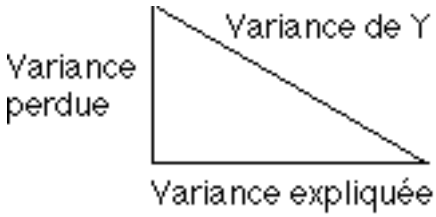
r^2 : part de la variance de Y qui est expliquée par la variance de X.

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

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

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

Coefficients de corrélation partiels

Corrélations obtenues en controlant 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.

Equation du plan de régression

Guère pertinent ici

$$Z = aX + bY + c$$

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

Si les variables sont réduites, a et b sont les coefficients de régression partiels.

Coefficient de corrélation multiple

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

$$R = \frac{Cov(Z, \hat{Z})}{s(Z)s(\hat{Z})} \quad (= 0.29)$$

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 ?

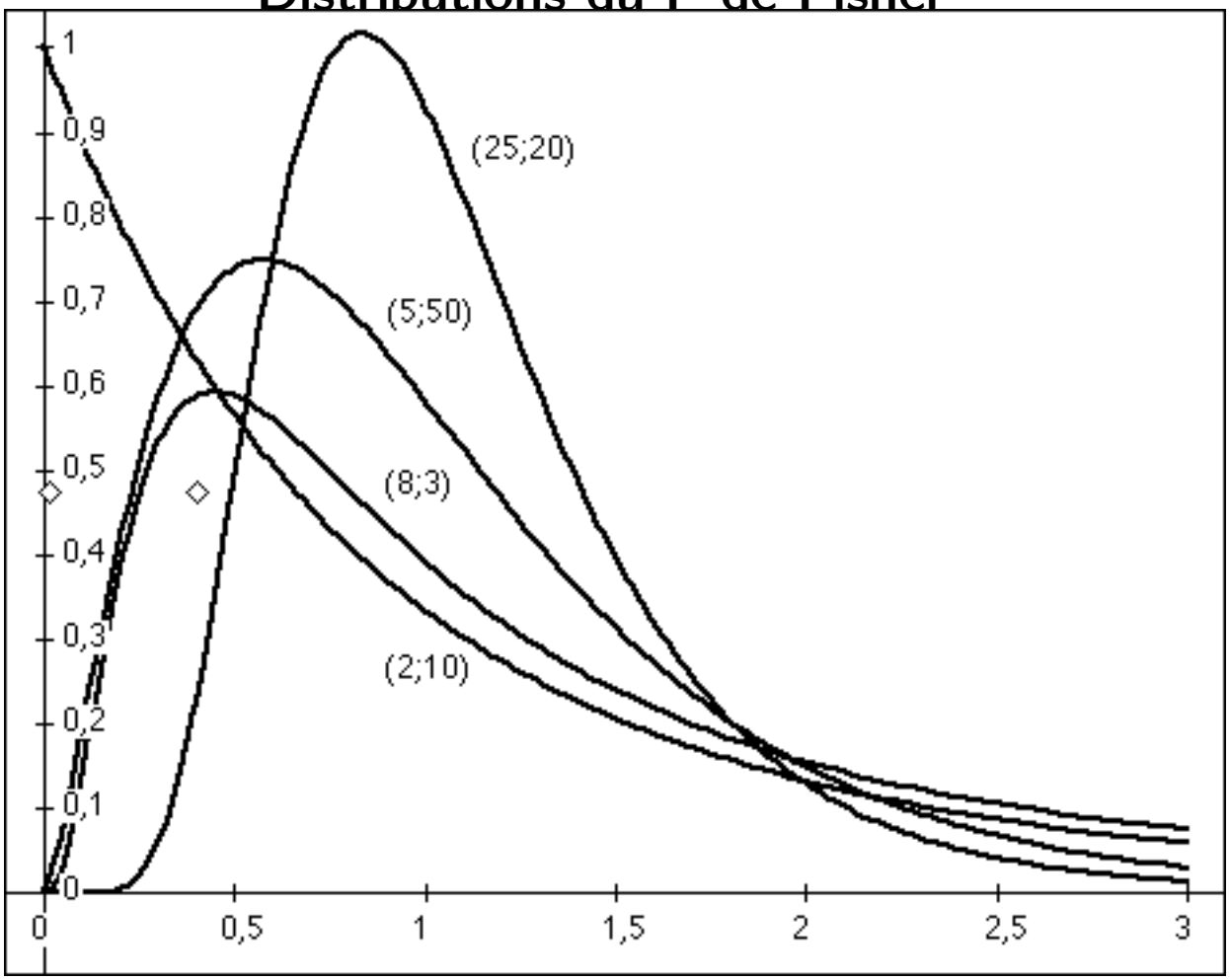
H_0 : Les variances sont égales
H_1 : La première variance est supérieure à la deuxième.

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 :

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

Elle se décompose en une variation "intra-groupes" et une variation "inter-groupes" :

$SC_T = SC_{inter} + SC_{intra}$ avec :

$$SC_{inter} = \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j} - \frac{T_G^2}{N}$$

$$SC_{intra} = \sum_{i,j} x_{ij}^2 - \sum_{j=1}^{a} \frac{T_{.j}^2}{n_j}$$

Carrés moyens :

$$CM_{inter} = \frac{SC_{inter}}{a-1} \quad ; \quad CM_{intra} = \frac{SC_{intra}}{N-a}$$

Statistique de test : $F = \dfrac{CM_{inter}}{CM_{intra}}$

F suit une loi de Fisher à $(a-1)$ et $(N-a)$ ddl.

Présentation des résultats

Source de variation	SC	ddl	CM	F
$\mathcal{A}$ (inter-groupes)	SC_1	$a-1$	CM_1	F_{obs}
Résiduelle (intra-groupes)	SC_2	$N-a$	CM_2	
Total	SC_T	$N-1$		

Organisation des calculs

i j	1	2	3	Total
1	x_{11}			
...	...	...	...	
$T_{\cdot j}$	$T_{\cdot 1}$			T_G
$T_{\cdot j}^2$				
n_j				N
$\frac{T_{\cdot j}^2}{n_j}$				
$\sum x_{ij}^2$				

Exemple :

15 sujets évaluent 3 couvertures de magazine. Sont-elles équivalentes ?

	C1	C2	C3
	14	16	14
	6	14	16
	12	8	14
	10	8	14
	8	14	12
$\overline{x}_i$	10	12	14

Calculs

i j	1	2	3	Total
1	$x_{11} = 14$	16	14	
2	$x_{21} = 6$	14	16	
...	...	...	...	
$T_{\cdot j}$	50	60	70	$T_G = 180$
$T_{\cdot j}^2$	2500	3600	4900	
n_j	5	5	5	$N = 15$
$\frac{T_{\cdot j}^2}{n_j}$	500	720	980	2200
$\sum x_{ij}^2$	540	776	988	2304

$$SC_{inter} = \sum_{j=1}^{a} \frac{T_{\cdot j}^2}{n_j} - \frac{T_G^2}{N} = 2200 - \frac{180^2}{15} = 40$$

$$SC_{intra} = \sum_{i,j} x_{ij}^2 - \sum_{j=1}^{a} \frac{T_{\cdot j}^2}{n_j} = 2304 - 2200 = 104$$

$$SC_T = \sum_{i,j} x_{ij}^2 - \frac{T_G^2}{N} = 144$$

$$CM_{inter} = \frac{SC_{inter}}{a-1} = 20 \ ; \ CM_{intra} = \frac{SC_{intra}}{N-a} = 8.67$$

$$F_{obs} = \frac{CM_{inter}}{CM_{intra}} = 2.31$$

F suit une loi de Fisher avec $ddl_1 = a - 1 = 2$ et $ddl_2 = N - a = 12$.

Résultats

Source	Somme carrés	ddl	Carré Moyen	F
$\mathcal{A}$	40	2	20	2.31
Résid.	104	12	8.67	
Total	144	14		

Pour α=5%, $F_{crit} = 3.88$: H_0 est acceptée

Remarques

– Dans le cas de groupes équilibrés d'effectif n :

CM_{inter} représente n fois la variance corrigée de la série des moyennes des groupes.

CM_{intra} est la moyenne des variances corrigées des groupes.

– X : distribution normale dans chacun des groupes

– Hypothèse d'égalité des variances

– Si 2 groupes, équivaut à un T de Student. $F = T^2$

Vocabulaire des plans d'expérience

Variable dépendante

– On formule une hypothèse : "telle variable a tel effet sur le comportement des sujets"

– On choisit une **variable dépendante**

Définir une variable mesurable (numérique) caractérisant le comportement du sujet.

Qualités d'une bonne variable dépendante : pertinence, sensibilité.

Variables indépendantes ou facteurs

– Recherche des **variables indépendantes** ou **facteurs de variation**

Indépendant : indépendant du sujet, manipulé ou contrôlé par l'expérimentateur

Causes susceptibles d'entraîner une variation de la variable dépendante

Les *Facteurs principaux* sont ceux dont on désire étudier l'effet. Ils sont aussi appelés *facteurs d'intérêt*.

Les *Facteurs secondaires* sont les autres causes susceptibles d'influer sur le comportement des sujets. Deux manières de les prendre en compte :

– Contrôle

– Neutralisation.

Les facteurs sont des variables nominales ou ordinales. Les valeurs prises par un facteur sont ses *modalités* ou *niveaux*.

Facteur systématique : l'ensemble des modalités possibles est fini (et petit). Toutes les modalités sont présentes dans l'expérience.

Facteur aléatoire : l'ensemble des modalités est grand (infini). On choisit alors (par tirage au sort) un ensemble de modalités.

En Psychologie, le facteur *sujet* est pratiquement toujours présent. Généralement, c'est un facteur aléatoire et secondaire. Il est souvent assimilé à une incertitude sur une mesure.

Facteur étiquette : par exemple, le sexe, ou le milieu socio-culturel.

Définition et écriture d'un plan d'expérience

En général, plusieurs facteurs, avec interaction. Donc : étude simultanée.

Plan factoriel :

Un plan factoriel est un plan dans lequel chaque modalité d'un facteur est combinée avec chaque combinaison de modalités des autres facteurs.

Plan en carré latin

Exemple : 3 facteurs comportant le même nombre de modalités.

On croise les deux premiers facteurs. Les modalités du troisième sont distribuées de façon à réaliser des permutations sur les lignes et les colonnes.

Exemple.

$a_1b_1c_1$	$a_1b_2c_2$	$a_1b_3c_3$
$a_2b_1c_2$	$a_2b_2c_3$	$a_2b_3c_1$
$a_3b_1c_3$	$a_3b_2c_1$	$a_3b_3c_2$

Plans quasi-complets (Rouanet - Lépine 1976)

Croisement : deux ou plusieurs facteurs sont croisés si chaque niveau de l'un des facteurs est combiné avec chaque niveau de chacun des autres facteurs.

Notation : $\mathcal{A}_3 * \mathcal{B}_5$ par exemple.

Emboitement : Un facteur $\mathcal{A}$ est emboité dans un facteur $\mathcal{B}$ si chaque niveau de $\mathcal{A}$ est combiné avec un seul niveau de $\mathcal{B}$.

Notation : $\mathcal{A} < \mathcal{B} >$

Emboitement équilibré : pour chaque niveau du facteur emboitant, on a le même nombre de niveaux du facteur emboité.

Les plans *quasi-complets* sont les plans qui peuvent être décrits à l'aide de relations de croisement et d'emboitement.

Définition :

Un plan est dit *quasi-complet* s'il possède les deux propriétés suivantes :

– Tous les facteurs croisés deux à deux sont croisés ou emboités

– Les facteurs croisés deux à deux sont croisés dans leur ensemble.

Exemple : $\mathcal{S}_4 < \mathcal{A}_2 > * \mathcal{B}_2$

Lorsque le facteur *sujet* est croisé avec d'autres facteurs : *plan à mesures répétées*

Modèle de score

Hypothèse : additivité des effets

Pour un plan $\mathcal{S} < \mathcal{A}_a >$:

$$x_{ij} = \mu + \alpha_j + e_{ij}$$

Pour un plan $\mathcal{S} * \mathcal{A}_a$

$$x_{ij} = \mu + \alpha_j + \epsilon_i + e_{ij}$$

Pour un plan $\mathcal{S} < \mathcal{A}_a * \mathcal{B}_b >$:

$$x_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + e_{ijk}$$

$\alpha\beta_{jk}$: effet de l'interaction entre $\mathcal{A}$ et $\mathcal{B}$.

Effet simple d'un facteur $\mathcal{A}$: effet observé de $\mathcal{A}$ lorsque les autres facteurs sont fixés.

Effet principal d'un facteur $\mathcal{A}$: effet observé de $\mathcal{A}$, sans tenir compte des autres conditions.

Interaction entre facteurs

Exemple : un plan $\mathcal{S} < \mathcal{A}_2 * \mathcal{B}_2 >$
Mémorisation d'une liste de mots. VD : nombre de mots mémorisés

Facteur $\mathcal{A}$: a_1 normal, a_2 déficit mnésique
Facteur $\mathcal{B}$: b_1 12 mots, b_2 30 mots

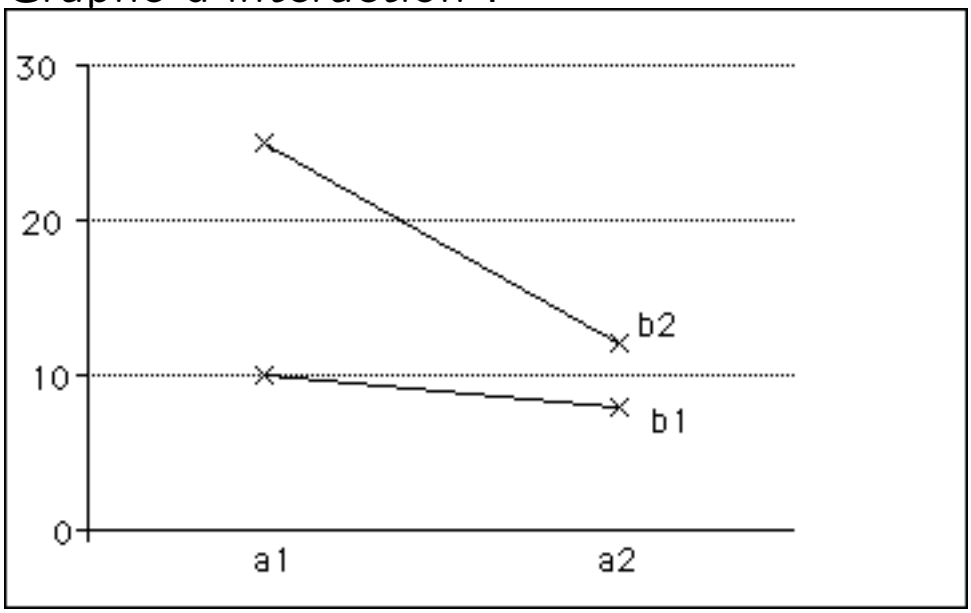
Moyennes observées sur 4 groupes indépendants :

	b_1	b_2
a_1	10	25
a_2	8	12

Modèle de score :

$$x_{ijk} = \mu + \alpha_j + \beta_k + \alpha\beta_{jk} + e_{ijk}$$

Graphe d'interaction :



Analyse de variance à plusieurs facteurs

Plan $\mathcal{S} < \mathcal{A}_a >$: déjà traité

Plan $\mathcal{S}_n * \mathcal{A}_a$

Notations
a : nombre de conditions expérimentales.
n : nombre de sujets.
x_{ij} : valeur de la VD pour le i-ième individu dans la condition expérimentale j.

Hypothèses du test
H_0 : Dans la population parente, les moyennes correspondant aux a conditions expérimentales sont égales.
H_1 : Les moyennes sont différentes.

Présentation des résultats

Source	S. carrés	ddl	C. moyen	F
$\mathcal{A}$	SC_1	$a-1$	CM_1	$\frac{CM_1}{CM_3}$
$\mathcal{S}$	SC_2	$n-1$	CM_2	
Résid.	SC_3	$(n-1)(a-1)$	CM_3	
Total	SC_T	$N-1$		

F suit une loi de Fisher-Snedecor à $a-1$ et $(n-1)(a-1)$ degrés de liberté.

Exemple : Effet du bruit sur la discrimination perceptive
Facteur : bruit (3 niveaux)
VD : nombre d'erreurs commises.

Sujets	Absence	Intermittent	Continu
1	117	119	127
2	130	126	131
3	122	118	129
4	123	117	134
5	126	120	137
6	116	120	128

Source	S. carrés	ddl	C. moyen	F
Bruit	403.11	2	201.56	19.98 **
Sujets	164.44	5	32.89	
Résid.	100.89	10	10.09	
Total	668.44	17		

Plan $\mathcal{S} < \mathcal{A}_a * \mathcal{B}_b >$

$\mathcal{A}$ et $\mathcal{B}$: facteurs fixes.

Notations
a, b, n, x_{ijk}

Interaction entre les facteurs $\mathcal{A}$ et $\mathcal{B}$

Tableau d'analyse de variance

Source	S. carrés	ddl	C. moyen	F
$\mathcal{A}$	SC_1	$a-1$	CM_1	$\frac{CM_1}{CM_4}$
$\mathcal{B}$	SC_2	$b-1$	CM_2	$\frac{CM_2}{CM_4}$
Inter. $\mathcal{AB}$	SC_3	$(a-1)(b-1)$	CM_3	$\frac{CM_3}{CM_4}$
Résid.	SC_4	$ab(n-1)$	CM_4	
Total	SC_T	$N-1$		

Exemple : facteurs : sexe, statut socio-économique.
VD : mesure du "locus of control"

	statut socio-économique		
	Bas	Moyen	Elevé
Hommes	10	16	18
	12	12	14
	8	19	17
	14	17	13
	10	15	19
	16	11	15
	15	14	22
	13	10	20
Femmes	8	14	12
	10	10	18
	7	13	14
	9	9	21
	12	17	19
	5	15	17
	8	12	13
	7	8	16

Sources de var.	ddl	SC	CM	F
Sexe	1	65.33	65.33	7.73**
Statut soc-éco	2	338.67	169.33	20.03**
$X \times C$	2	18.67	9.33	1.10 NS
Résidu	42	355.0	8.45	
Total	47	777.67		

Plan $\mathcal{S} < \mathcal{A}_a > * \mathcal{B}_b$

Plan à mesures partiellement répétées ou plan split-plot

$\mathcal{A}$ et $\mathcal{B}$: facteurs fixes.

Notations
a, b, n, x_{ijk}

Présentation des résultats

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{A}$	SC_1	$a-1$	CM_1	$\frac{CM_1}{CM_2}$
$\mathcal{S}(\mathcal{A})$	SC_2	$a(n-1)$	CM_2	
	Dans les sujets			
$\mathcal{B}$	SC_3	$b-1$	CM_3	$\frac{CM_3}{CM_5}$
Int. $\mathcal{AB}$	SC_4	$(a-1)(b-1)$	CM_4	$\frac{CM_4}{CM_5}$
Résid.	SC_5	$a(n-1)(b-1)$	CM_5	
Total	SC_T	$N-1$		

Exemple : Expérimentation de Bahrick (reconnaissance de portraits)

Facteurs : sexe du sujet, sexe du portrait
VD : nombre de portraits reconnus

Nom du sujet	Portrait masculin	Portrait féminin
Albert	6	6
Henri	6	6
Jules	5	5
Paul	5	5
Octave	5	6
Albertine	6	8
Henriette	7	8
Julie	6	6
Paule	7	7
Octavie	6	6

Source	S. carrés	ddl	C. moyen	F
	Entre les sujets			
$\mathcal{X}$	7.2	1	7.2	10.28*
$\mathcal{S}(\mathcal{X})$	5.6	8	0.7	
	Dans les sujets			
$\mathcal{P}$	0.8	1	0.8	3.2 NS
Int. $\mathcal{XP}$	0.2	1	0.2	0.8 NS
Résid.	2	8	0.25	
Total	15	19		

Remarques et conclusion

Modèle basé sur l'hypothèse d'additivité des effets

Conditions théoriques d'application de la méthode :

– Normalité de la VD dans les populations parentes

– Egalité des variances dans les populations parentes

Tests permettant de vérifier que les conditions sont remplies :

– Test de normalité de Lilliefors

– Test de O'Brien sur les variances

La méthode est robuste : elle fournit des résultats corrects, même si les conditions ne sont qu'approximativement vérifiées.

Il existe également des méthodes non paramétriques : travail sur des rangs

Retour sur la corrélation

Test du coefficient de corrélation à l'aide d'une analyse de variance

Source	SC	ddl	CM	F
Modèle	$ns_c^2(\widehat{Y})$	1	CM_1	F_{obs}
Résiduelle	$ns_c^2(E)$	$n-2$	CM_2	
Total	$ns_c^2(Y)$	$n-1$		

On retrouve : $F_{obs} = T_{obs}^2$