

**Statistiques paramétriques et non  
paramétriques  
Informatique**

**E.C. PSRS73B et PSRS73C**

**Présentation du cours 2004/2005**

**Organisation matérielle**

Cours-TD de Statistiques : 24 heures.  
Jeudi 8h15-10h10 salle A111

Travaux dirigés d'informatique : 12 heures  
Mardi 8h15-10h10 sem B - salle A206 ou A208

Monitorat informatique

Contrôle des connaissances : (contrôle continu)  
EC PSRS73B : Examen écrit (2 heures)  
EC PSRS73C : Note de TD

## **Bibliographie**

- B. Cadet Méthodes statistiques en psychologie. P.U. de Caen
- B. Beaufile. 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 De Boeck Université
- P. Rateau, Méthode et statistique expérimentales en sciences humaines, Ellipses

## **Documents fournis :**

Transparents du cours de statistiques

Fiches de TD de statistiques et d'informatique

### *Adresses Web*

– accès en dehors des salles informatiques :

<http://geai.univ-brest.fr/~carpentier/>

<http://infolettres.univ-brest.fr/~carpentier/>

– accès depuis les salles informatiques :

<http://letsamba.univ-brest.fr/~carpentier/>

## **Contenu**

### *Statistiques :*

Aspects méthodologiques : quelle stratégie pour analyser les données ? Quelles sont les méthodes disponibles pour tester telle hypothèse ?

Compléments aux méthodes de statistiques descriptives et inférentielles vues en licence :

Statistiques paramétriques : effet calibré, taille d'un effet ; test de normalité des distributions parentes ; loi de Fisher Snedecor ; ANOVA.

Statistiques non paramétriques : test de Kolmogorov-Smirnov ; test de Kruskal-Wallis.

Compléments sur la corrélation et la régression linéaires à 2 ou plusieurs variables.

### *Informatique :*

Mise en oeuvre des traitements étudiés à l'aide d'un logiciel de traitement statistique professionnel (Statistica).

## Conditions d'application du test de Student

Le test de Student est un test paramétrique. Comme tous les tests de ce type, son utilisation est soumise à des conditions sur la distribution des variables sur les populations de référence.

Rappel : hypothèses *a priori* pour l'application du test de Student (égalité des moyennes) sur deux groupes indépendants :

- Normalité des distributions parentes
- Égalité des variances (homoscédasticité des résidus)

Problèmes :

- Comment étudier si ces conditions sont respectées ?
- Peut-on s'affranchir de ces conditions ?

## Tester la normalité d'une distribution

Variable numérique  $X$  définie sur une population  
( $x_i$ ) : valeurs observées sur un échantillon de taille  $n$   
Au vu de cet échantillon : est-il légitime de supposer  
que la distribution de  $X$  dans la population est une loi  
normale ?

Différents tests proposés : Test de Kolmogorov-Smirnov,  
test de Lilliefors, test de Shapiro-Wilk.

Etude du test de Kolmogorov-Smirnov (sur un exemple) :

Echantillon : 8, 9, 9, 10, 10, 10, 11, 13, 14, 14

$H_0$  :  $X$  est distribuée selon une loi normale dans la  
population

$H_1$  :  $X$  n'est pas distribuée selon une loi normale.

Construction de la statistique de test :

Moyenne observée :  $\bar{x} = 10.8$

Ecart type corrigé :  $s_c = 2.15$

Valeurs centrées réduites :  $z_i = \frac{x_i - \bar{x}}{s_c}$

$x_i$	8	9	10	11	13	14
$z_i$	-1.30	-0.84	-0.37	-0.09	1.02	1.49

Détermination de la distribution cumulative théorique et calcul des écarts entre distributions cumulatives observée et théorique

$z_i$	$S_n(z_i)$	$F(z_i)$	Ecart absolu
-1.30	0.1	0.0968	0.2443
-0.84	0.3	0.2005	
-0.37	0.6	0.3557	
-0.09	0.7	0.4641	
1.02	0.8	0.8461	
1.49	1	0.9319	

Maximum des écarts absolus :  $D_{obs} = 0.2443$ .

Taille de l'échantillon :  $n = 10$ .

Consultation de tables spécialisées :  $D_{crit} = 0.41$

Conclusion : L'hypothèse de normalité de  $X$  sur la population parente ne peut pas être rejetée.

Remarques :

- Méthode critiquable lorsque la moyenne et l'écart type sont estimés.
- Variante proposée par Lilliefors : même statistique de test, tables différentes.
- Autre test : Shapiro-Wilk
- Représentation graphique : droite de Henry

## Etude des conditions sur les variances

Pour le test t de Student, il existe des formules (approximatives) à utiliser lorsqu'on ne fait pas l'hypothèse d'égalité des variances.

– La statistique de test est alors :

$$t' = \frac{\bar{x}_1 - \bar{x}_2}{E} \quad \text{avec} \quad E^2 = \frac{s_{1c}^2}{n_1} + \frac{s_{2c}^2}{n_2}$$

– Le nombre de degrés de liberté à prendre en compte dépend des variances observées.

## Comparaison de deux variances

### Test de Fisher et loi de Fisher Snedecor

**Exemple.** Etude sur la boulimie. Deux groupes de sujets :

Variable dépendante : écart relatif par rapport au poids normal.

	Simple	Avec vom.
$\bar{x}_i$	4.61	-0.83
$s_{ic}^2$	219.04	79.21
$n_i$	49	32

### 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

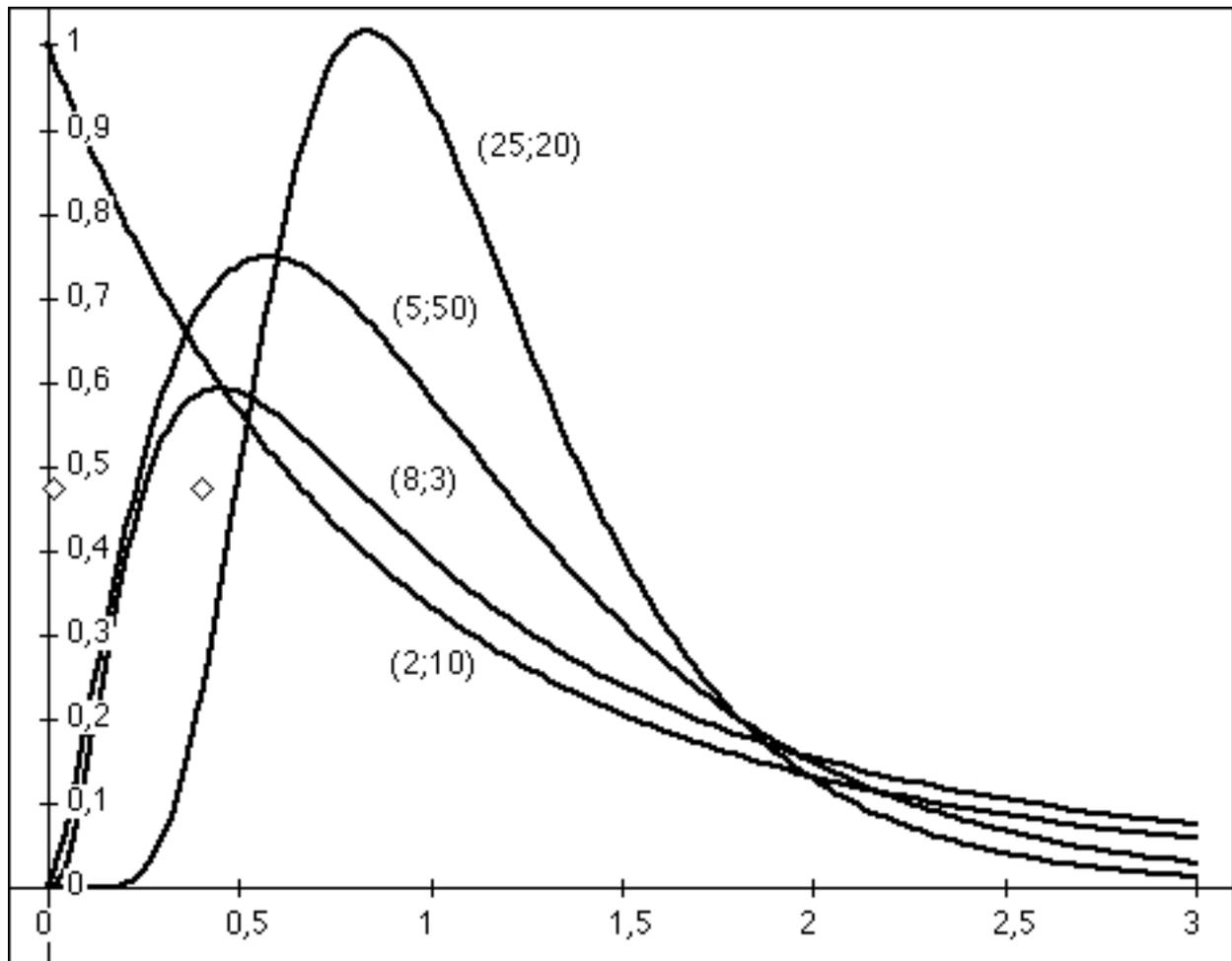
$H_1$  : La première variance est supérieure à la deuxième.

*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



Sur l'exemple considéré :

$$F_{obs} = \frac{219.04}{79.21} = 2.76$$

Pour  $\alpha = 5\%$ ,  $ddl_1 = 48$  et  $ddl_2 = 31$ ,  $F_{crit} = 1.79$ .

On rejette donc l'hypothèse  $H_0$ .

## Taille d'un effet – Puissance d'un test

Rappel concernant les test statistiques : erreurs de type I et II

		Hypothèse vraie	
		$H_0$	$H_1$
Hypothèse retenue	$H_0$	$1 - \alpha$	$\beta$
	$H_1$	$\alpha$	$1 - \beta$

- $\alpha$  : seuil de significativité. C'est aussi la probabilité de rejeter  $H_0$  alors que  $H_0$  est vraie (risque de première espèce ou risque de commettre une erreur de type I)
- $\beta$  : risque de seconde espèce. C'est la probabilité d'accepter  $H_0$  alors que  $H_0$  est fausse (risque de commettre une erreur de type II).
- $1 - \beta$  : probabilité de détecter correctement un cas où  $H_0$  doit être rejetée. **Puissance du test.**

Comment influencer sur la puissance d'un test ?

- Changement du seuil
- Choix d'une autre variable dépendante
- Taille des échantillons

## Taille d'un effet

Variable dépendante  $X$  définie sur deux populations.

Moyennes (sur les populations) :  $\mu_1$  et  $\mu_2$ .

Ecart type (le même pour les deux populations) :  $\sigma$ .

Taille de l'effet :

$$d = \frac{\mu_1 - \mu_2}{\sigma}$$

## Ecart calibré

Notion voisine de la précédente, mais calcul sur les échantillons observés.

*Dans une situation de groupes indépendants :*

Moyennes :  $\bar{x}_1, \bar{x}_2$

Écart types (non corrigés) :  $s_1$  et  $s_2$

Écart type corrigé pondéré :  $s$  défini par :  $s^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

*Ecart calibré :*  $EC = \frac{\bar{x}_1 - \bar{x}_2}{s}$

*Dans une situation de groupes appariés :*

Moyennes :  $\bar{x}_1, \bar{x}_2$

Écart type corrigé de la série des différences individuelles :  $s_c$

*Ecart calibré :*  $EC = \frac{\bar{x}_1 - \bar{x}_2}{s_c}$

## Estimer la puissance d'un test

*Recombinaison de la taille de l'effet et les tailles d'échantillons*

Pour estimer la puissance d'un test, on calcule :

$$\delta = d f(n)$$

$d$  est la taille de l'effet (ou son estimation, avec les données que nous connaissons)

$f(n)$  est une fonction de la taille du ou des échantillons, fonction qui dépend du test statistique utilisé.

**Exemple** : test de comparaison à une norme.

On a :  $\mu_2 = 100$  (norme),  $\sigma = 15$  et on pense que, dans la population considérée,  $\mu_1 = 105$ .

La statistique de test à utiliser est ici :  $Z = \frac{\bar{x} - \mu_2}{E}$

avec  $E^2 = \frac{\sigma^2}{n}$

D'où :  $\delta = d\sqrt{n}$

On a ici :  $d = \frac{105 - 100}{15} = 0.33$

Quelle taille d'échantillon faut-il choisir pour que la puissance du test soit au moins de 80% au seuil de 5% ?

Calcul de  $\delta$  :  $\delta = 0.33\sqrt{n}$

Valeur lue dans la table : il faut que  $\delta \geq 2.8$

D'où :  $0.33\sqrt{n} \geq 2.8$

$n \geq \left(\frac{2.8}{0.33}\right)^2$  c'est-à-dire  $n \geq 72$ .

## **Puissance d'un test d'égalité de moyennes sur 2 échantillons indépendants**

*Cas de 2 échantillons équilibrés*

$n$  effectif de chacun des deux échantillons.

$$\text{Dans ce cas : } \delta = d \sqrt{\frac{n}{2}}$$

*Cas d'échantillons non équilibrés*

$$\text{Dans ce cas : } \delta = d \sqrt{\frac{n_h}{2}}$$

où  $n_h$  est la *moyenne harmonique* des deux tailles d'échantillons définie par :

$$\frac{2}{n_h} = \frac{1}{n_1} + \frac{1}{n_2} \quad \text{ou} \quad n_h = \frac{2n_1n_2}{n_1 + n_2}$$

## **Puissance d'un test d'égalité de moyennes sur 2 échantillons appariés**

*Estimation de la taille de l'effet :*

$d = \frac{\mu_1 - \mu_2}{\sigma}$  où  $\sigma$  est l'écart type de la série des différences individuelles.

*Estimation de la puissance du test :*

On prend  $\delta = d \sqrt{n}$

## **Puissance d'un test d'égalité de deux proportions sur des groupes indépendants**

*Estimation de la taille de l'effet :*

On a des estimations  $p_1$  et  $p_2$  des proportions sur les deux populations. On peut estimer l'effet par :

$$d = \frac{p_1 - p_2}{\sqrt{p(1-p)}}$$

où  $p$  est la proportion pondérée sur l'ensemble des deux populations.

On peut aussi remarquer que  $p(1-p) \leq 0.25$  et estimer  $d$  par :  $d = 2(p_1 - p_2)$ .

*Détermination de la puissance du test :*

On prend alors :  $\delta = d\sqrt{\frac{n_h}{2}}$  où  $n_h$  est, comme précédemment, la moyenne harmonique des effectifs des deux échantillons.

## 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 :*

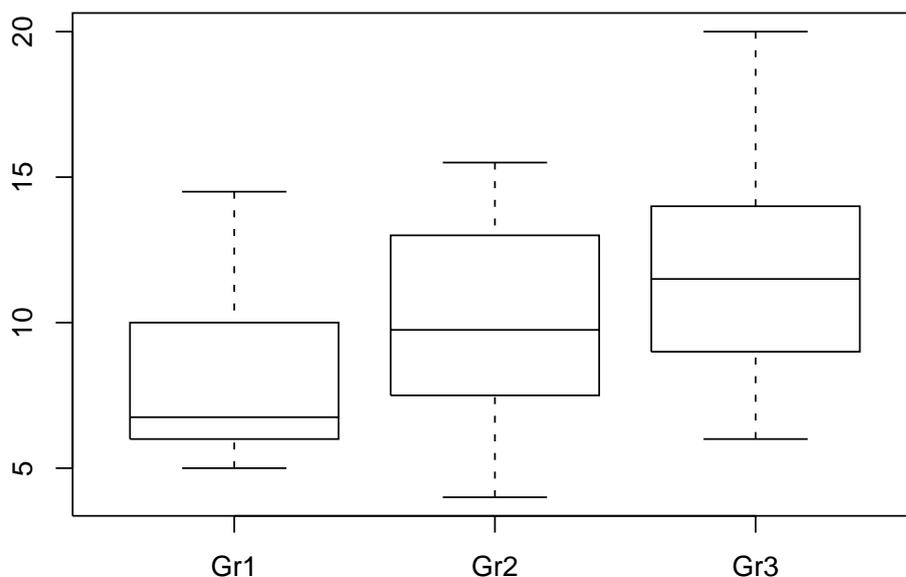
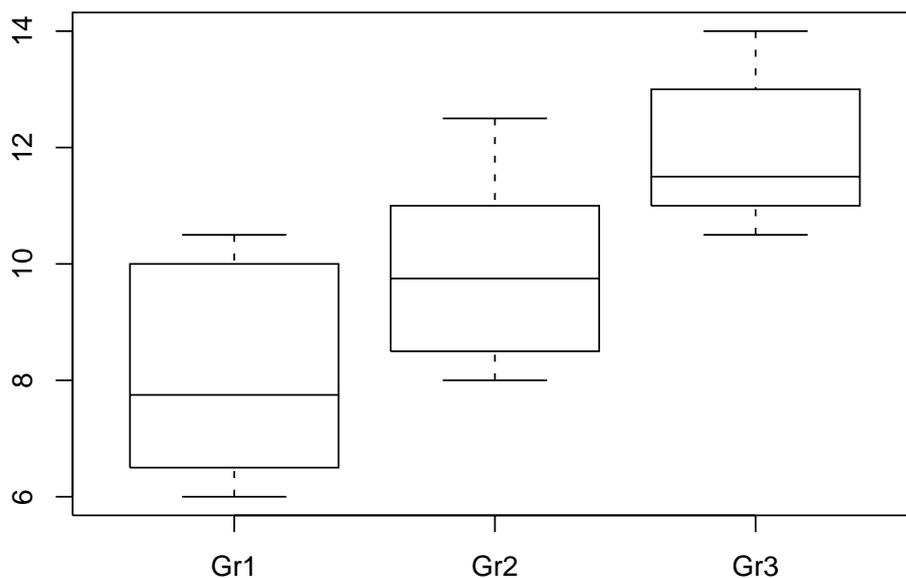
	Gr1	Gr2	Gr3
	6	8	10.5
	6.5	8.5	10.5
	6.5	8.5	11
	7	9	11
	7.5	9.5	11
	8	10	12
	8	11	13
	10	11	13
	10	12	14
	10.5	12.5	14
$\bar{x}_i$	8	10	12

*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

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

Boîtes à moustaches pour les deux situations proposées



## Comparer $a$ moyennes sur des groupes indépendants

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

Une variable  $\mathcal{A}$ , de modalités  $A_1, A_2, \dots, A_a$  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 ?

*Hypothèses "a priori" :*

- distribution normale de  $X$  dans chacun des groupes
- Egalité des variances dans les populations.

*Hypothèses du test :*

$H_0$  :  $\mu_1 = \mu_2 = \dots = \mu_a$

$H_1$  : Les moyennes ne sont pas toutes égales.

**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	
$\bar{x}_i$	10	12	14	12

*Variation (ou somme des carrés) totale :*

$$SC_T = (14 - 12)^2 + (6 - 12)^2 + \dots + (12 - 12)^2 = 144$$

*Décomposition de la variation totale :*

Score d'un sujet = Moyenne de son groupe + Ecart

C1	C2	C3	C1	C2	C3
10	12	14	4	4	0
10	12	14	-4	2	2
10	12	14	2	-4	0
10	12	14	0	-4	0
10	12	14	-2	2	-2

*Variation (ou somme des carrés) inter-groupes :*

$$SC_{inter} = (10 - 12)^2 + (10 - 12)^2 + \dots + (14 - 12)^2 = 40$$

*Variation (ou somme des carrés) intra-groupes :*

$$SC_{intra} = 4^2 + (-4)^2 + \dots + (-2)^2 = 104$$

Calcul des carrés moyens :

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

Statistique de test :

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

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

## Résultats

Source	Somme carrés	ddl	Carré Moyen	$F$
$C$	40	2	20	2.31
Résid.	104	12	8.67	
Total	144	14		

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

## Formules de calcul pour un calcul "à la main" efficace

*Construction de la statistique de test :*

*Notations :*

$n_1, n_2, \dots, n_a$  : effectifs des groupes.

$N$  : effectif total

$T_1, \dots, T_a$  : sommes des observations pour chacun des groupes.

$T.$  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 :*

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

*Statistique de test :*

$$F = \frac{CM_{inter}}{CM_{intra}}$$

$F$  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)	$SC_{inter}$	$a - 1$	$CM_{inter}$	$F_{obs}$
Résiduelle (intra-gr.)	$SC_{intra}$	$N - a$	$CM_{intra}$	
Total	$SC_T$	$N - 1$		

## Organisation des calculs

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

## Remarques

–  $SC_{inter}$  : c'est la somme des carrés (totale) que l'on obtiendrait si toutes les observations d'un groupe étaient égales à la moyenne de ce groupe.

$CM_{inter}$  : variance corrigée de cet ensemble de données.

–  $SC_{intra}$  : c'est la somme des carrés (totale) que l'on obtiendrait en "décalant" chaque observation de façon à avoir la même moyenne dans chaque groupe.

$CM_{intra}$  : "moyenne pondérée" des trois variances corrigées ainsi obtenues.

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

Pour les deux situations proposées en introduction :

### **Situation 1**

Analysis of Variance Table

Response : x1

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.000	40.000	17.008	1.659e-05 ***
Residuals	27	63.500	2.352		

### **Situation 2**

Analysis of Variance Table

Response : x2

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	2	80.00	40.00	2.7136	0.08436 .
Residuals	27	398.00	14.74		

## Tests non paramétriques

Paramètres : moyenne, variance, covariance, etc ;

Les tests tels que t de Student, ANOVA, etc sont des tests paramétriques :

- hypothèses relatives à un paramètre des populations parentes
- nécessité d'estimer un ou plusieurs paramètres de la distribution parente à l'aide des échantillons observés
- conditions d'application liées à la forme des distributions parentes

Il existe également des tests *non paramétriques* ou indépendants de toute distribution.

- pas de condition d'application
- peu affectés par la présence d'un ou plusieurs scores extrêmes
- ils ont en général une plus faible puissance que les tests paramétriques correspondants

## Tests non paramétriques sur deux groupes indépendants

Situation envisagée : un plan  $\mathcal{S} < \mathcal{A}_2 >$  avec un facteur  $\mathcal{A}$  à 2 niveaux définissant deux groupes indépendants et une variable dépendante  $X$  ordinale ou numérique

Effectifs des deux groupes :  $n_1$  et  $n_2$ .

### Test de la médiane

#### *Hypothèses*

$H_0$  : Les deux populations parentes ont même médiane.

$H_1$  : Les deux populations parentes ont des médianes différentes

#### *Construction de la statistique de test*

On détermine la médiane  $M$  de la série obtenue en réunissant les deux échantillons.

On constitue un tableau de contingence en croisant la variable indépendante et la variable dérivée "position par rapport à  $M$ "

	Gr 1	Gr 2	Ensemble
$\leq M$	$N_{11}$	$N_{12}$	$N_{1.}$
$> M$	$N_{21}$	$N_{22}$	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

On fait un test du  $\chi^2$  sur le tableau obtenu.

Condition d'application (selon Siegel et Castellan) : le nombre total d'observations doit être supérieur à 20.

## Test de Kolmogorov-Smirnov

$H_0$  : La distribution de la VD est la même dans les deux populations parentes

$H_1$  : Les distributions sont différentes (hypothèse bilatérale) ou l'intensité de la VD est plus forte dans le groupe  $j$  (hypothèse unilatérale)

### *Construction de la statistique de test*

On choisit (si besoin) un regroupement en classes :  $b_1, b_2, \dots, b_k$

On construit le tableau des fréquences cumulées dans les deux groupes :

	Gr 1	Gr 2
$X \leq b_1$	$F_{11}$	$F_{12}$
$X \leq b_2$	$F_{21}$	$F_{22}$
$\dots$	$\dots$	$\dots$
$X \leq b_k$	$F_{k1}$	$F_{k2}$

On calcule (cas d'un test bilatéral) :

$$D = \max |F_{i1} - F_{i2}|$$

Dans le cas d'un test unilatéral : on calcule le maximum des différences, ordonnées en fonction du sens du test.

Pour  $n_1 \leq 40$  ou  $n_2 \leq 40$ , on utilise des tables spécialisées.

Pour de grands échantillons ( $n_1 > 40$  et  $n_2 > 40$ ), on utilise l'approximation suivante : sous  $H_0$ , la statistique :

$$\chi^2 = 4D^2 \frac{n_1 n_2}{n_1 + n_2}$$

suit une loi du  $\chi^2$  à 2 ddl.

## Test de Wald-Wolfowitz

$H_0$  : La distribution de la VD est la même dans les deux populations parentes

$H_1$  : Les distributions sont différentes (hypothèse bilatérale) ou l'intensité de la VD est plus forte dans le groupe  $j$  (hypothèse unilatérale)

Méthode : on classe toutes les observations par ordre croissant. On construit un compteur démarrant à 1, et qui augmente d'une unité chaque fois que l'on change de groupe en parcourant la liste ordonnée. On obtient ainsi le nombre de "runs"  $u$ .

Exemple : On a fait passer une épreuve à 31 sujets, 14 hommes et 17 femmes. Le protocole des rangs observé est le suivant :

Hommes : 1 2 3 7 8 9 10 13 14 15 23 24 26 27

Femmes : 4 5 6 11 12 16 17 18 19 20 21 22 25 28 29 30 31

Détermination du nombre de "runs" :

MMM FFF MMMM FF MMM FFFFFFFF MM F MM FFFF

111 222 3333 44 555 6666666 77 8 99 0000

Ici :  $u = 10$ .

Pour  $n_1 \leq 10$  ou  $n_2 \leq 10$ , on utilise des tables spécialisées.

Pour  $n_1 > 10$  et  $n_2 > 10$ , on utilise l'approximation par une loi normale :

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1 \quad \sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}$$

$$Z = \frac{u - \mu \pm 0.5}{\sigma}$$

Ici :  $\mu = 16.35$   $\sigma^2 = 7.35$   $Z = -2.16$

Remarque : Ce test suppose l'absence d'ex-aequos.

## Test U de Mann-Whitney - Test de Wilcoxon Mann Whitney

$H_0$  : La probabilité qu'un score provenant de la première population soit supérieur à un score provenant de la seconde est 0.5

$H_1$  : Cette probabilité est différente de 0.5 (hypothèse bilatérale), inférieure à 0.5, supérieure à 0.5 (hypothèses unilatérales)

Méthode : On construit le protocole des rangs pour l'ensemble des  $(n_1 + n_2)$  observations (avec la convention du rang moyen pour les ex-aequos).

$W_1$  : somme des rangs du premier échantillon

$W_2$  : somme des rangs du deuxième échantillon.

Pour  $n_1 \leq 10$  ou  $n_2 \leq 10$ , on utilise des tables spécialisées.

Si  $n_1 > 10$  et  $n_2 > 10$ , ou si l'un des deux effectifs est supérieur à 12, on utilise l'approximation par une loi normale :

Test U de Mann-Whitney proprement dit (cf. Statistica) :

On calcule :

$$U_1 = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - W_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2 + 1)}{2} - W_2$$

$$U = \min(U_1, U_2)$$

$$Z = \frac{U - \frac{n_1 n_2}{2}}{E} \quad \text{avec} \quad E^2 = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Variante : On calcule les rangs moyens dans les deux groupes  $\bar{R}_1$  et  $\bar{R}_2$  puis la statistique :

$$Z = \frac{\bar{R}_1 - \bar{R}_2}{E} \quad \text{avec} \quad E^2 = \frac{(n_1 + n_2 + 1) (n_1 + n_2)^2}{12 n_1 n_2}$$

Remarques.

1. Dans le test de Wilcoxon, la variable  $X$  est supposée continue. La probabilité d'observer des ex-aequos est donc *en théorie* nulle. Cependant, certains auteurs ont proposé des formules correctives pour tenir compte des ex-aequos.

2. Un test analogue : le test de permutation des rangs de Wilcoxon.

## Tests non paramétriques sur $k$ groupes indépendants

Situation envisagée : un plan  $\mathcal{S} < \mathcal{A}_k >$  avec un facteur  $\mathcal{A}$  à  $k$  niveaux définissant  $k$  groupes indépendants et une variable dépendante  $X$  ordinale ou numérique

Effectifs des groupes :  $n_1, n_2, \dots, n_k$ .

### Test de la médiane

Le test de la médiane peut encore être utilisé dans cette situation.

### Test de Kruskal-Wallis

$H_0$  : La probabilité qu'un score provenant de l'une des populations soit supérieur à un score provenant d'une autre population est 0.5

$H_1$  : Cette probabilité est différente de 0.5

Méthode : On construit le protocole des rangs pour l'ensemble des observations.

Soit  $\bar{R}_j$  la moyenne des rangs dans le groupe  $j$ ,  $N$  le nombre total d'observations et  $\bar{R} = \frac{N+1}{2}$  le rang moyen général.

Statistique de test :

$$K = \frac{12}{N(N+1)} \sum n_j (\bar{R}_j - \bar{R})^2$$

ou

$$K = \left[ \frac{12}{N(N+1)} \sum n_j \bar{R}_j^2 \right] - 3(N+1)$$

Si le nombre de groupes est supérieur à 3 et le nombre d'observations dans chaque groupe est supérieur à 5,  $K$  suit approximativement une loi du  $\chi^2$  à  $(k-1)$ ddl.

## Tests non paramétriques sur 2 groupes appariés

Situation envisagée : un plan  $\mathcal{S} * \mathcal{A}_2$  avec un facteur  $\mathcal{A}$  à 2 niveaux définissant deux groupes appariés et une variable dépendante  $X$  ordinale ou numérique.

Effectif de l'échantillon de sujets :  $n$ .

### Test du $\chi^2$ de Mac Nemar

Il s'applique au cas où la variable  $X$  est dichotomique.

Situation générale :

		$A_1$	
		$X = 1$	$X = 0$
$A_2$	$X = 1$	$a$	$c$
	$X = 0$	$b$	$d$

Les paires discordantes sont les observations ( $X = 1$  en  $A_1$ ,  $X = 0$  en  $A_2$ ) et ( $X = 0$  en  $A_1$ ,  $X = 1$  en  $A_2$ )  
L'information utile est alors fournie par les effectifs "de discordance"  $b$  et  $c$ .

#### *Notations*

$p_1$  : fréquence de la combinaison ( $X = 1$  en  $A_1$ ,  $X = 0$  en  $A_2$ ) par rapport à la discordance totale dans la population.

$p_2$  : fréquence de la combinaison ( $X = 0$  en  $A_1$ ,  $X = 1$  en  $A_2$ ) par rapport à la discordance totale dans la population.

*Hypothèses du test*

$$H_0 : p_1 = p_2 (= 50\%)$$

$$H_1 : p_1 \neq p_2$$

*Statistique de test*

$$\chi^2 = \frac{(b - c)^2}{b + c}, \quad ddl = 1$$

ou, avec la correction de Yates (petits effectifs) :

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}, \quad ddl = 1$$

Condition d'application :  $b + c > 10$ .

**Remarques.**

1. Cette statistique est la distance du  $\chi^2$  calculée entre le tableau d'effectifs observés et le tableau d'effectifs théoriques suivant :

		$A_1$	
		$X = 1$	$X = 0$
$A_2$	$X = 1$	$a$	$\frac{b+c}{2}$
	$X = 0$	$\frac{b+c}{2}$	$d$

2. On peut aussi utiliser la statistique suivante, qui permet éventuellement un test unilatéral :

$$Z = \frac{b - c \pm 1}{\sqrt{b + c}}$$

Z suit la loi normale centrée réduite.

Correction de continuité ( $\pm 1$ ) : choisir le signe de façon à diminuer la valeur absolue de la statistique.

## Tests des permutations

Principe : on observe un protocole de différences individuelles  $d_i$ . On observe  $D_+$  différences positives et  $D_-$  différences négatives. On élimine les différences nulles. On imagine tous les protocoles obtenus en affectant arbitrairement les  $D_+$  signes “+” et les  $D_-$  signes “-” aux différences absolues  $|d_i|$ . Pour chaque protocole ainsi obtenu, on calcule  $\sum d_i$ . La zone d'acceptation de  $H_0$  est formée des protocoles conduisant à des sommes  $\sum d_i$  “proches de 0”.

Deux tests basés sur ce principe : test des signes, test de Wilcoxon.

### Test des signes

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appariés.

Variable dépendante : ordinale ou numérique.

- protocole du signe des différences individuelles ; modalités :  $-1, 0, 1$
- on élimine les différences nulles

$D_+$  : nombre de différences positives

$D_-$  : nombre de différences négatives

$N = D_+ + D_-$  : nombre total d'observations après élimination des différences nulles.

*Hypothèses du test :*

$H_0$  : les différences sont dues au hasard : dans la population parente, la fréquence des différences positives est 50%.

$H_1$  : Cette fréquence n'est pas 50% (test bilatéral)  
ou (tests unilatéraux)

Cette fréquence est inférieure à 50%

Cette fréquence est supérieure à 50%

- *Cas des petits échantillons*

Sous  $H_0$ ,  $D_+$  suit une *loi binomiale* de paramètres  $N$  et  $p = 0.5$ .

On raisonne en termes de "niveau de significativité".

Par exemple, dans le cas d'un test unilatéral tel que  $H_1$  : fréquence inférieure à 50%

on calcule la fréquence cumulée  $P(X \leq D_+)$  de  $D_+$  pour la loi binomiale  $B(N, 0.5)$ .

Pour un seuil  $\alpha$  donné :

Si  $P(X \leq D_+) < \alpha$  on retient  $H_1$

Si  $P(X \leq D_+) \geq \alpha$  on retient  $H_0$

- *Cas des grands échantillons : approximation par une loi normale*

$$Z = \frac{2D_+ \pm 1 - N}{\sqrt{N}}$$

$Z$  suit une loi normale centrée réduite.

Correction de continuité ( $\pm 1$ ) : choisir le signe de façon à diminuer la valeur absolue de la statistique.

## **Test de Wilcoxon sur des groupes appariés Test T, ou test des rangs signés**

C'est le test des permutations appliqué au protocole des rangs signés.

Un échantillon de sujets, placés dans deux conditions expérimentales différentes : groupes appariés.

Variable dépendante : numérique.

On construit :

- le protocole des effets individuels  $d_i$
- le protocole des valeurs absolues de ces effets  $|d_i|$
- le protocole des rangs appliqués aux valeurs absolues, en éliminant les valeurs nulles.

$T_+$  : somme des rangs des observations tq  $d_i > 0$

$T_-$  : somme des rangs des observations tq  $d_i < 0$

$N$  = nombre de différences non nulles

$T_m = \min(T_+, T_-)$  ;

$T_M = \max(T_+, T_-)$

### *Hypothèses*

$H_0$  : Dans la population parente, les effets individuels positifs et les effets individuels négatifs s'interclassent de manière homogène

$H_1$  : Les deux classements sont différents (test bilatéral) ou les effets individuels positifs apparaissent plus fréquemment dans les rangs les moins élevés (resp. les plus élevés) (test unilatéral).

- *Cas des petits échantillons*

$N \leq 15$  : utilisation de tables spécialisées

On compare  $T_m$  aux valeurs critiques indiquées par la table.

- *Cas des grands échantillons*

$N > 15$  : approximation par une loi normale

$$Z = \frac{T_+ \pm 0.5 - \frac{N(N+1)}{4}}{E}$$

avec

$$E^2 = \frac{N(N+1)(2N+1)}{24}$$

Z suit une loi normale centrée réduite.

Remarques.

Dans le test de Wilcoxon, la variable  $X$  est supposée continue. La probabilité d'observer des ex-aequos est donc *en théorie* nulle. Cependant, certains auteurs ont proposé des formules correctives pour tenir compte des ex-aequos.

## Tests non paramétriques sur $k$ groupes appariés

Situation envisagée : un plan  $\mathcal{S} * \mathcal{A}_k$  avec un facteur  $\mathcal{A}$  à  $k$  niveaux définissant des groupes appariés et une variable dépendante  $X$  ordinale ou numérique.

Effectif de l'échantillon de sujets :  $n$ .

### Test $Q$ de Cochran

Il s'applique au cas où la variable  $X$  est dichotomique.

$H_0$  : Dans la population, la probabilité de la modalité "1" est la même dans toutes les conditions.

$H_1$  : Dans la population, la probabilité de la modalité "1" n'est pas la même dans toutes les conditions.

Protocole observé :

Suj.	$A_1$	$A_2$	...	$A_k$	$L_i$	$L_i^2$
$s_1$	1	1	...	0	$L_1$	$L_1^2$
$s_2$	1	0	...	0	$L_2$	$L_2^2$
...						
$s_n$					$L_n$	$L_n^2$
$G_j$	$G_1$	$G_2$	...	$G_k$	$G$	

La statistique :

$$Q = \frac{(k-1) (k \sum G_j^2 - G^2)}{k \sum L_i - \sum L_i^2}$$

suit approximativement une loi du  $\chi^2$  à  $(k-1)$  ddl.

Calcul équivalent :

$$Q = \frac{k(k-1) \sum (G_j - \bar{G})^2}{k \sum L_i - \sum L_i^2}$$

où  $\bar{G}$  est la moyenne des  $G_j$ .

## Test de Friedman

Il s'applique au cas où la variable  $X$  est ordinale ou numérique.

$H_0$  : Dans les différentes conditions, les médianes sont égales :  $M_1 = M_2 = \dots = M_k$ .

$H_1$  : Les  $k$  médianes ne sont pas toutes égales.

*Statistique de test :*

On calcule un protocole de rangs *par sujet*.

Soit  $R_j$  la somme des rangs de la colonne  $j$  (dans la condition  $A_j$ ).

La statistique de Friedman est donnée par :

$$F = \left[ \frac{12}{nk(k+1)} \sum R_j^2 \right] - 3n(k+1)$$

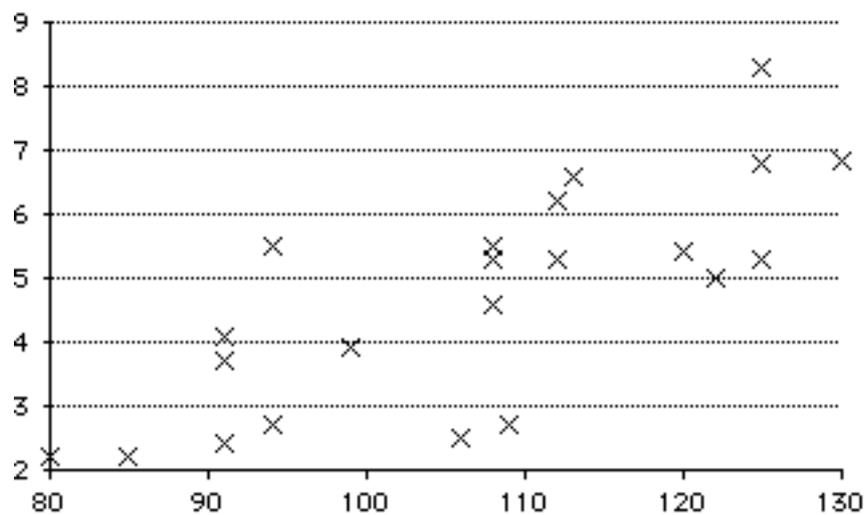
$F$  est tabulée pour les petites valeurs de  $n$  et  $k$ . Au delà,  $F$  suit approximativement une loi du  $\chi^2$  à  $(k-1)$  ddl.

## 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

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

## Covariance corrigée des variables $X$ et $Y$

$$Cov_c(X, Y) = \frac{n}{n-1} Cov(X, Y)$$

## Coefficient de corrélation de Bravais Pearson

On désigne par  $s(X)$  et  $s(Y)$  les écarts types de  $X$  et  $Y$  et par  $s_c(X)$ ,  $s_c(Y)$  leurs écarts types corrigés.

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

## Remarques

- $r$  n'est pas une estimation correcte du coefficient de corrélation dans la population. Certains logiciels de statistiques donnent comme estimation :

$$r_{aj} = \sqrt{1 - \frac{(1 - r^2)(N - 1)}{N - 2}}$$

- 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.

### *Conditions d'application*

Dans la population parente, le couple  $(X, Y)$  suit une *loi normale bivariée*, ce qui implique notamment :

- la normalité des distributions marginales de  $X$  et  $Y$  ;
- la normalité de la distribution de l'une des variables lorsque l'autre variable est fixée ;
- l'égalité des variances des distributions de l'une des variables pour deux valeurs distinctes de l'autre variable.

## Corrélation et statistiques non paramétriques : corrélation des rangs de Spearman

Si les données ne vérifient pas les conditions d'application précédentes, ou si les données observées sont elles-mêmes des classements, on pourra travailler sur les protocoles des rangs définis séparément pour chacune des deux variables.

	<i>Rangs X</i>	Rangs Y
$s_1$	$r_1$	$r'_1$
$s_2$	$r_2$	$r'_2$
...	...	...

Le coefficient de corrélation des deux protocoles de rangs est appelé coefficient de corrélation de Spearman, et noté  $R_s$ .

*Calcul de  $R_s$*  (en l'absence d'ex-aequos)

On calcule les différences individuelles  $d_i = r_i - r'_i$ , puis

$$R_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}$$

## Significativité du coefficient de corrélation de Spearman

- Pour  $N \leq 30$ , on utilise en général des tables spécialisées.
- Lorsque  $N > 30$  :

Certains auteurs (et Statistica) utilisent :

$$t = \sqrt{N - 2} \frac{R_s}{\sqrt{1 - R_s^2}}$$

et une loi de Student à  $N - 2$  ddl.

D'autres auteurs utilisent la statistique :

$$Z = \sqrt{N - 1} R_s$$

et une loi normale centrée réduite.

## Le coefficient $\tau$ de Kendall

Avec un protocole comportant  $N$  sujets, on a  $\frac{N(N-1)}{2}$  paires de sujets.

On examine chaque paire de sujets, et on note si les deux classements comportent une inversion ou non.

$s_3$	3	6
$s_4$	5	2

Désaccord, Inversion

$s_3$	3	4
$s_4$	5	6

Accord, Pas d'inversion

Le coefficient  $\tau$  est alors défini par :

$$\tau = \frac{\text{Nb d'accords} - \text{Nb de désaccords}}{\text{Nb de paires}}$$

ou

$$\tau = 1 - \frac{2 \times \text{Nombre d'inversions}}{\text{Nombre de paires}}$$

Pour  $N > 10$ , la statistique

$$Z = 3\tau \sqrt{\frac{N(N-1)}{2(2N+5)}}$$

suit une loi normale centrée réduite.

## 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  $X$

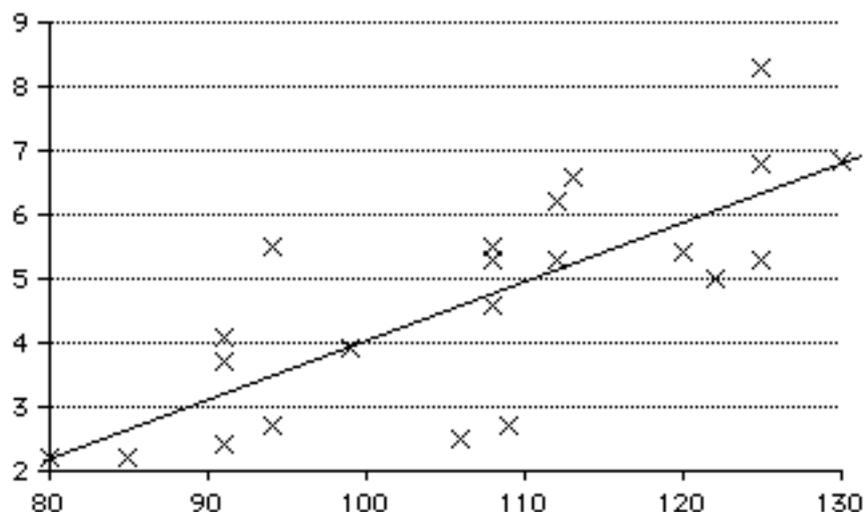
*Droite de régression de  $Y$  par rapport à  $X$  :*

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

$$y = ax + b$$

avec :

$$a = \frac{Cov(X, Y)}{s^2(X)} \quad ; \quad b = \bar{Y} - a\bar{X}$$



*Comparaison des valeurs observées et des valeurs estimées*

Valeurs estimées :  $\hat{y}_i = b_0 + b_1x_i$  : variable  $\hat{Y}$   
Erreur (ou résidu) :  $e_i = y_i - \hat{y}_i$  : variable  $E$

Les variables  $\hat{Y}$  et  $E$  sont indépendantes et 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)

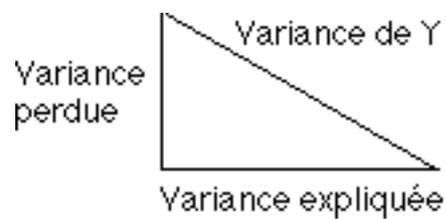
$s^2(E)$  : variance *perdue* ou *résiduelle*

$r^2$  : 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  $Y$ .



## Test du coefficient de corrélation à l'aide du $F$ de Fisher

Valeurs estimées :  $\hat{y}_i = b_0 + b_1x_i$

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

On introduit les sommes de carrés suivantes :

$$SC_{\text{Totale}} = \sum (y_i - \bar{y})^2$$

$$SC_{\text{Régression}} = \sum (\hat{y}_i - \bar{y})^2$$

$$SC_{\text{Résidus}} = \sum (\hat{y}_i - y_i)^2$$

*Lien avec le coefficient de corrélation*

$r^2 = \frac{SC_{\text{Régression}}}{SC_{\text{Totale}}}$  est le *coefficient de détermination*

*Tableau d'analyse de variance*

Source	SC	ddl	CM	$F$
Régression	$SC_{\text{Régression}}$	1	$CM_{\text{Reg}}$	$F_{\text{obs}}$
Résiduelle	$SC_{\text{Résidus}}$	$n - 2$	$CM_{\text{Res}}$	
Total	$SC_{\text{Totale}}$	$n - 1$		

$F_{\text{obs}} = \frac{CM_{\text{Reg}}}{CM_{\text{Res}}} = (n - 2) \frac{r^2}{1 - r^2}$  suit une loi de Fisher à 1 et  $n - 2$  ddl.

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

## Estimations de $\hat{y}$ et de $y$ par des intervalles de confiance

Dans la population, le lien entre  $X$  et  $Y$  suit le modèle mathématique :

$$Y = \beta_0 + \beta_1 X + \epsilon$$

où  $\epsilon$  est une variable statistique centrée et indépendante de  $X$

A partir de l'échantillon, nous avons calculé  $b_0, b_1, e_1, \dots, e_n$  tels que :

$$y_i = b_0 + b_1 x_i + e_i$$

$$\hat{y}_i = b_0 + b_1 x_i$$

Mais un autre échantillon amènerait d'autres valeurs de ces paramètres :  $b_0, b_1$  et  $e_i$  ne sont que des estimations de  $\beta_0, \beta_1$  et  $\epsilon_i$ .

*Questions que l'on peut se poser :*

- Quelle estimation peut-on donner de la variance de  $\epsilon$  ?
- On peut voir  $\hat{y}_i$  comme une estimation ponctuelle de la moyenne des valeurs de  $Y$  sur la population lorsque  $X = x_i$ . Peut-on déterminer un intervalle de confiance pour cette moyenne ?
- Etant donné une valeur  $x_i$  de  $X$ , quel intervalle de confiance peut-on donner pour les valeurs de  $Y$  correspondantes ?

*Estimation de  $Var(\epsilon)$  :*

$$s^2 = \frac{SC_{Res}}{n - 2}$$

*Estimation par un intervalle de confiance de la moyenne de  $Y$  pour  $X$  fixé :*

Pour une valeur  $x_p$  fixée de  $X$ , la variance des valeurs estimées  $\hat{y}_p$  est estimée par :

$$s_{\hat{y}_p}^2 = s^2 \left( \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x_i^2 - n\bar{x}} \right)$$

Un intervalle de confiance de  $Moy(Y|X = x_p)$  avec un degré de confiance  $\beta = 1 - \alpha$  est donné par :

$$\hat{y}_p - t_\alpha s_{\hat{y}_p} \leq Moy(Y|X = x_p) \leq \hat{y}_p + t_\alpha s_{\hat{y}_p}$$

où  $t_\alpha$  est la valeur du  $T$  de Student à  $n - 2$  ddl telle que  $P(|T| > t_\alpha) = \alpha$

*Détermination d'un intervalle de confiance pour les valeurs de Y : intervalle de prévision*

$Moy(Y|X = x_p)$  est connue par une estimation ponctuelle ( $\hat{y}_p$ ) et un intervalle de confiance.

La différence  $Y - Moy(Y|X = x_p)$  est le résidu  $\epsilon$ , dont on peut également donner un intervalle de confiance.

Finalement, on pourra écrire l'intervalle de confiance :

$$\hat{y}_p - t_\alpha s_{ind} \leq y \leq \hat{y}_p + t_\alpha s_{ind}$$

avec :

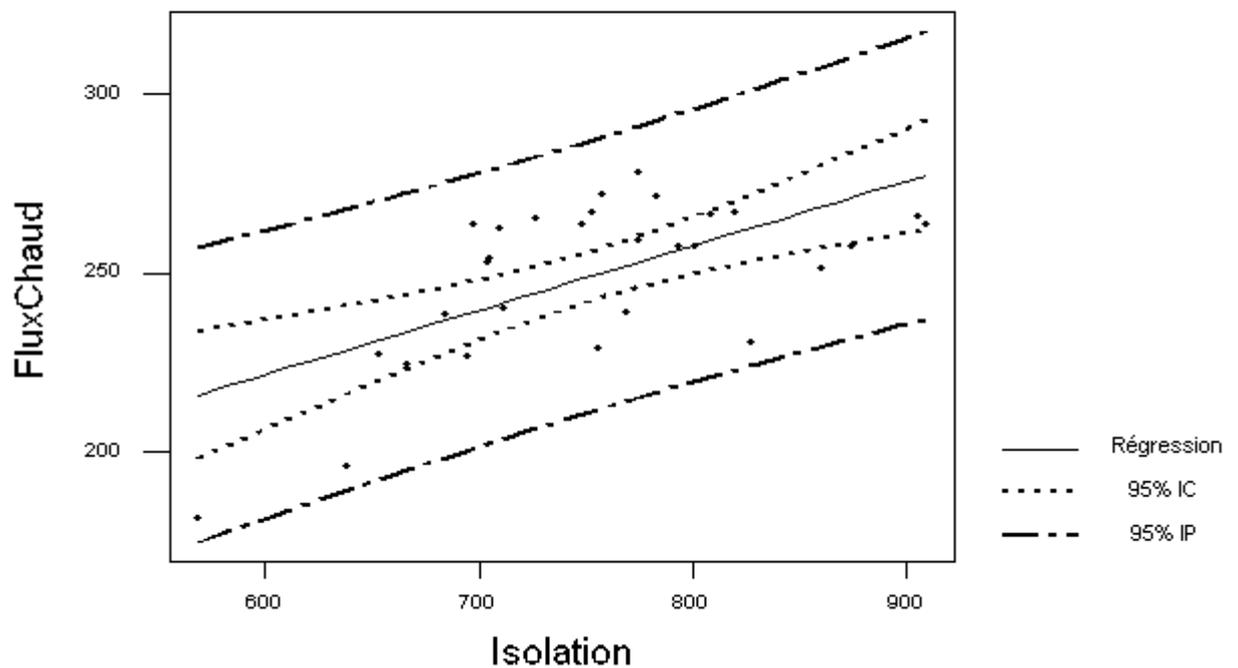
$$s_{ind}^2 = s^2 \left( 1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x_i^2 - n\bar{x}} \right)$$

Intervalle de confiance et intervalle de prévision apparaissent dans les graphiques réalisés par les logiciels sous forme de “bandes” :

### Graphique de la régression

$$Y = 113,749 + 0,180111X$$

R carré = 39,4 %



### Intervalle de prévision

### Graphique de la régression

$$Y = 113,749 + 0,180111X$$

R carré = 39,4 %

