Modèles d'Analyse de la Variance

- Ch11 - Anova 1: introduction
- Ch12 - Anova 2 : 1 facteur
- Ch13 - Anova 3 : plusieurs facteurs
- Ch14 - Anova 4 : mesures répétées
- Ch15 - Anova 5 : facteurs aléatoires

MODÈLES d'ANALYSE de la VARIANCE

Ch12 - Partie 2 : expériences avec 1 facteur

EXPÉRIENCES AVEC UN FACTEUR

facteur A modalités (niveaux) 1, 2, … , g

g = nombre de modalités (niveaux, groupes)

cas 1 : étude observationnelle / rétrospective

cas 2 : étude avec des unités expérimentales

protocole : complètement aléatoire (CRD), blocs, présence de facteurs secondaires, …

facteur fixé
- objectif : comparaison des moyennes
- modèle à effets fixés

facteur aléatoire
- modalités : au hasard d'une population de modalités
- conclusions s'appliquent à cette population
- modèle à effets aléatoires / composantes variances
- modèles mixtes : facteurs aléatoires + facteurs fixés
- estimation des composantes de la variance

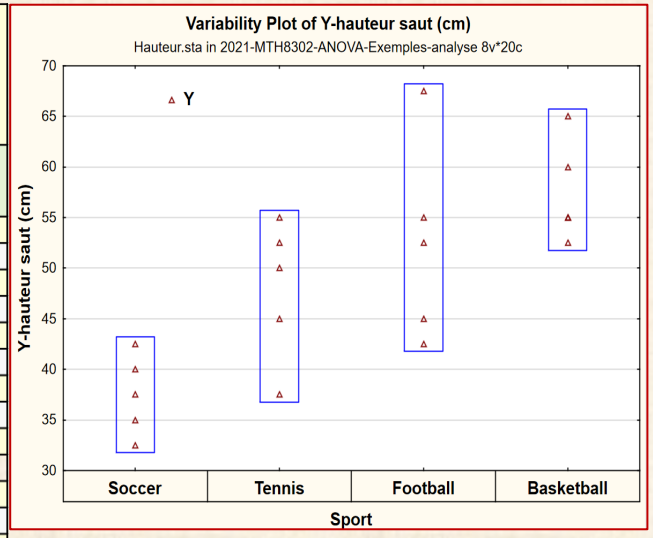
réponse Y :
- variable quantitative
- hypothèse : observations indépendantes
- modèles linéaires généralisés si Y transformée

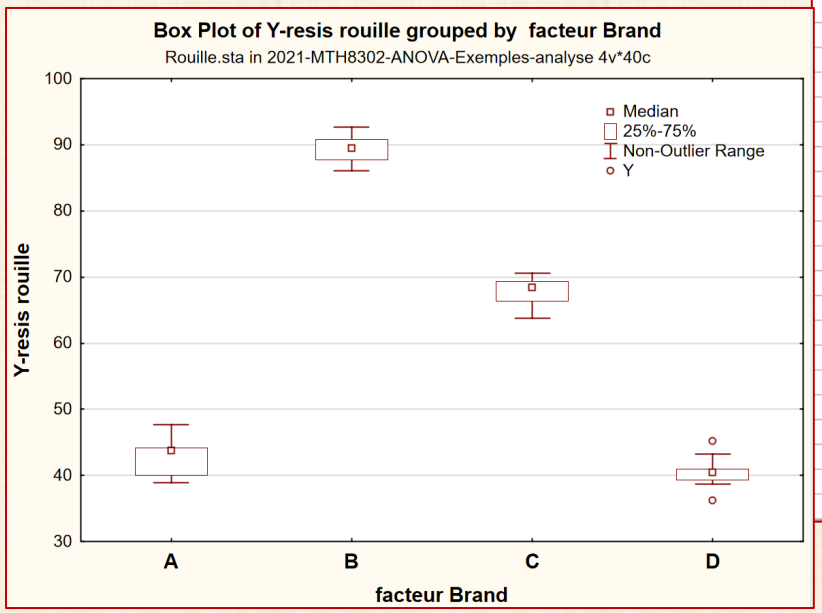
EXPÉRIENCES AVEC UN FACTEUR

Exemple : saut en hauteur

comparaison de 4 groupes d'athlètes : habilité à sauter en hauteur

id	Sport	Y-hauteur saut (cm)
1	Soccer	38
2	Soccer	43
3	Soccer	33
4	Soccer	40
5	Soccer	35
6	Tennis	45
7	Tennis	53
8	Tennis	38
9	Tennis	55
10	Tennis	50
11	Football	55
12	Football	68
13	Football	43
14	Football	45
15	Football	53
16	Basketball	60
17	Basketball	65
18	Basketball	55
19	Basketball	53
20	Basketball	55





Exemple : résistance rouille

1 ID	2 facteur Brand	3 rep	4 Y-resis rouille
1	A	1	43,9
2	A	2	39,0
3	A	3	46,7
4	A	4	43,8
5	A	5	44,2
6	A	6	47,7
7	A	7	43,6
8	A	8	38,9
9	A	9	43,6
10	A	10	40,0
11	B	1	89,8
12	B	2	87,1
13	B	3	92,7
14	B	4	90,6
15	B	5	87,7
16	B	6	92,4
17	B	7	86,1
18	B	8	88,1
19	B	9	90,8
20	B	10	89,1
21	C	1	68,4
22	C	2	69,3
23	C	3	68,5
24	C	4	66,4
25	C	5	70,0
26	C	6	68,1
27	C	7	70,6
28	C	8	65,2
29	C	9	63,8
30	C	10	69,2
31	D	1	36,2
32	D	2	45,2
33	D	3	40,7
34	D	4	40,5
35	D	5	39,3
36	D	6	40,3
37	D	7	43,2
38	D	8	38,7
39	D	9	40,9
40	D	10	39,7

EXPÉRIENCES AVEC UN FACTEUR

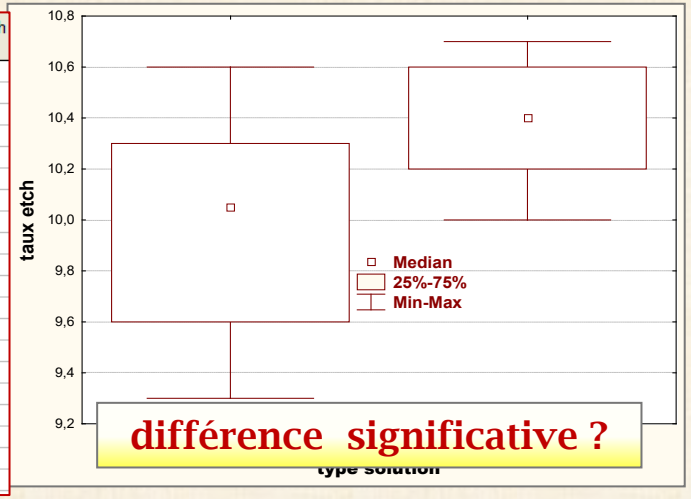
Exemple procédé de gravure (électronique)
(« wet etching ») enlèvement du silicium sur «puces »
variable de réponse Y : taux d'enlèvement du procédé
comparaison de solution1 et solution2

données : Y	n = 10				
solution 1 :	9.9	10.6	9.4	10.3	9.3
	10.0	9.6	10.3	10.2	10.1
solution 2 :	10.2	10.0	10.6	10.2	10.7
	10.7	10.4	10.4	10.5	10.3

moyenne : 9.97 (solution 1), 10.04 (solution 2)

Un facteur à 2 modalités

X-solution	rep	Y-taux etch
1	1	9,9
1	2	9,4
1	3	9,3
1	4	9,6
1	5	10,2
1	6	10,6
1	7	10,3
1	8	10,0
1	9	10,3
1	10	10,1
2	1	10,2
2	2	10,6
2	3	10,7
2	4	10,4
2	5	10,5
2	6	10,0
2	7	10,2
2	8	10,7
2	9	10,4
2	10	10,3

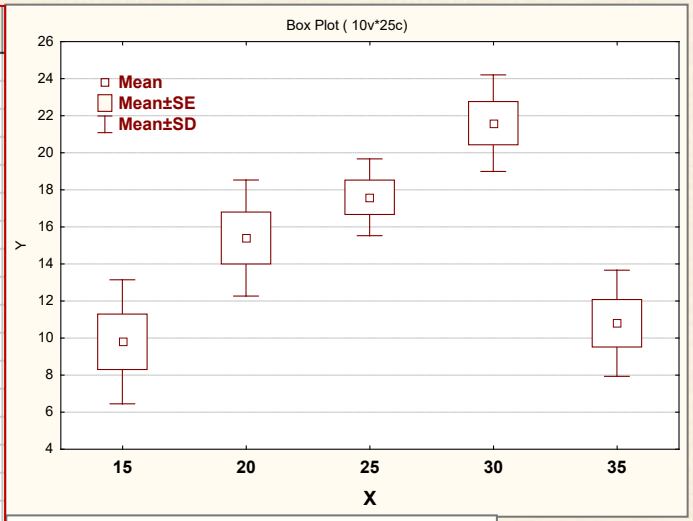


Exemple recherche nouvelle composition de fibres synthétique tissus

- facteur X : % coton varie 15 et 35
- réponse Y : force de tension tissu
- un facteur avec 5 modalités de X fixées : 15 20 25 30 35

données	n = 5 répétitions					
X	i	1	2	3	4	5	moyenne
15	1	7	7	15	11	9	9.8
20	2	12	17	12	18	18	15.4
25	3	14	18	18	19	19	17.6
30	4	19	25	22	19	23	21.6
35	5	7	10	11	15	11	10.8

X-coton	rep	Y-tension	X^2
15	1	7	225
15	2	7	225
15	3	15	225
15	4	11	225
15	5	9	225
20	1	12	400
20	2	17	400
20	3	12	400
20	4	18	400
20	5	18	400
25	1	14	625
25	2	18	625
25	3	18	625
25	4	19	625
25	5	19	625
30	1	19	900
30	2	25	900
30	3	22	900
30	4	19	900
30	5	23	900
35	1	7	1225
35	2	10	1225
35	3	11	1225
35	4	15	1225
35	5	11	1225



Influence du facteur X ?
différences significatives ?

EXPÉRIENCES AVEC UN FACTEUR

données

niveau	obs. y_{ij}				moyennes	# obs.	variances
1	y_{11}	y_{12}	...	y_{1n1}	$\overline{y_{1.}}$	n_1	s_1^2
2	y_{21}	y_{22}	....	y_{2n2}	$\overline{y_{2.}}$	n_2	s_2^2
.....							
i	y_{i1}	y_{i2}	...	$y_{i\,ni}$	$\overline{y_{i.}}$	n_i	s_i^2
.....							
g	y_{g1}	y_{g2}	...	y_{gng}	$\overline{y_{g.}}$	n_g	s_g^2
tous					$\overline{y_{..}}$	N	

$y_{i.} = \sum y_{ir}$

$y_{..} = \sum \sum y_{ir}$

$\overline{y_{i.}} = y_{i.} / n_i$

$N = \sum n_i$

$\overline{y_{..}} = y_{..} / N$

$SS_i = \sum (y_{ir} - \overline{y_{i.}})^2$

$s_i^2 = SS_i / (n_i - 1)$

Modèle à moyennes de cellules: effet général pas explicite

(1) $Y_{ir} = \mu_i + \varepsilon_{ir}$ $\quad i = 1, 2, \ldots, g \quad r = 1, 2, \ldots, n_i$

Y_{ij} : valeur de la variable de réponse r-ième essai modalité i

μ_r : moyenne de la cellule i - paramètre statistique à estimer

ε_{ir} : erreurs aléatoires indépendantes ~ $N(0, \sigma^2)$

EXPÉRIENCES AVEC UN FACTEUR

Modèle linéaire général : notation matricielle

$$Y = X\beta + \varepsilon$$

Y : vecteur N x 1 d'observations (données)
X : matrice du modèle N x p
fonction des k variables (facteurs) explicatives
β : vecteur p x 1 de paramètres (statistiques) à estimer
ε : vecteur N x 1 d'erreur ~ $N(0, \sigma^2)$

remarque

- la linéarité est relative à β
- $X_1, X_2, \ldots$ ne sont pas impliquées dans cette question de linéarité

QUESTIONS
- plan collecte données : n_i = ?
- estimation de β et σ^2
- décomposition variabilité : tableau ANOVA
- validation du modèle : analyse des résidus
- tests d'hypothèses
- comparaisons a posteriori
- etc

Exemple : données saut en hauteur

$g = 4 \quad n_i = n = 5 \quad N = 20 \quad Y : 20 \times 1 \quad X : 20 \times 4 \quad \varepsilon : 20 \times 1$

data

Soccer	38
Soccer	43
Soccer	33
Soccer	40
Soccer	35
Tennis	45
Tennis	53
Tennis	38
Tennis	55
Tennis	50
Football	55
Football	68
Football	43
Football	45
Football	53
Basketball	60
Basketball	65
Basketball	55
Basketball	53
Basketball	55

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ . \\ . \\ Y_{15} \\ Y_{21} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ Y_{41} \\ . \\ . \\ Y_{44} \\ Y_{45} \end{pmatrix} = \begin{pmatrix} 38 \\ 43 \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ 53 \\ 55 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ . \\ \varepsilon_{44} \\ \varepsilon_{45} \end{pmatrix}$$

$$Y = X \beta + \varepsilon$$

EXPÉRIENCES AVEC UN FACTEUR

Ajustement du modèle principe des moindres carrés

minimum $Q = \sum\sum (Y_{ij} - \mu_i)^2$

solution $\hat{\mu}_i = \overline{Y}_{i.}$ prédictions $\hat{Y}_{ij} = \overline{Y}_{i.}$

Tableau d'analyse de la variance - ANOVA

SOURCE	SS	df	MS = SS / df	F
facteur A	SS_A	$g - 1$	MS_A	$F_0 = MS_A / MSE$
erreur	SS_{err}	$N - g$	$MSE = \hat{\sigma}^2$	
total	SS_{tot}	$N - 1$		

$$SS_A = \sum n_i \sum (\overline{y}_{i.} - \overline{y}_{..})^2$$

$$SS_{err} = \sum\sum (y_{ij} - \overline{y}_{i.})^2$$

$$SS_{tot} = \sum\sum (y_{ij} - \overline{y}_{..})^2$$

$$N = \sum n_i$$

tous les calculs reposent sur Y uniquement

EXPÉRIENCES AVEC UN FACTEUR

décomposition de la variabilité

$$\sum\sum (y_{ij} - \overline{y_{..}})^2 = \sum n_i \sum (\overline{y_{i.}} - \overline{y_{..}})^2 + \sum\sum (y_{ij} - \overline{y_{i.}})^2$$

$SS_{tot} = SS_A + SS_{errr}$ $\qquad \overline{\mu} = \sum n_i \mu_i / (N - 1)$

$E(MSE) = \sigma^2$ $\qquad E(MS_A) = \sigma^2 + \sum n_i (\mu_i - \overline{\mu})^2 / (g - 1)$

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$ hypothèse nulle d'égalité des moyennes

Rejet de H_0 si $F_0 > F(g - 1, N - g, 1 - \alpha)$ loi F de Fisher-Snedecor

où $F(g - 1, N - g, 1 - \alpha) = (1 - \alpha)$ ième percentile loi F $(g - 1, N-g)$

avec $g - 1$ degrés de liberté au numérateur

et $N - g$ degrés de liberté au dénominateur

<u>logiciel statistique</u> donne p-value = Prob ($F > F_0$)

rejet de H_0 si p-value < α

α = seuil du test = risque de rejeter H_0 si vraie

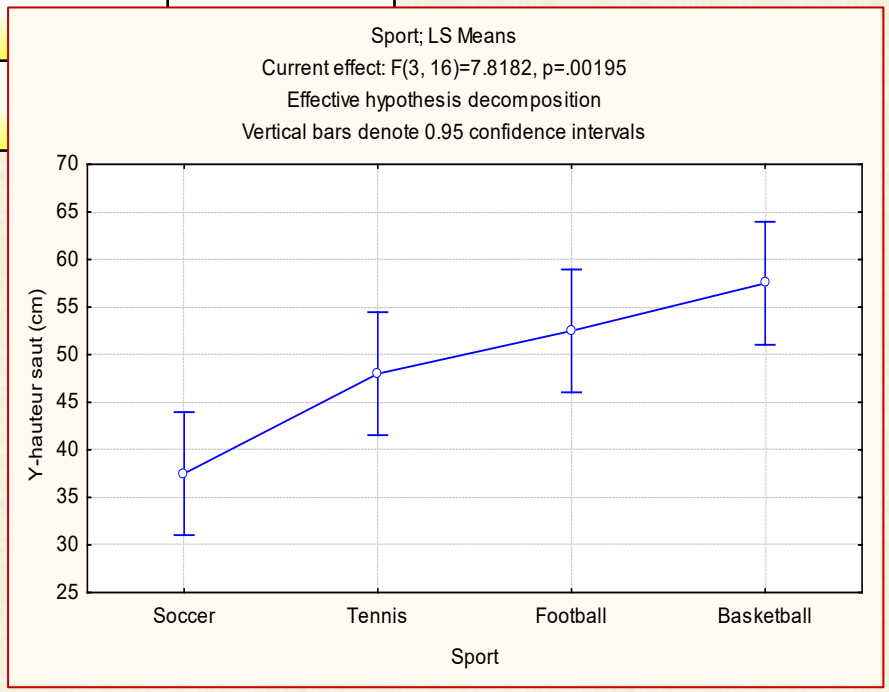
généralement $\alpha = 0{,}05$ ou moins (préférable)

EXPÉRIENCES AVEC UN FACTEUR

Exemple : données saut en hauteur

ANOVA

	df	Y-hauteur SS	Y-hauteur MS	Y-hauteur F	Y-hauteur p
Intercept	1	47775,31	47775,31	1029,502	0,000000
Sport	3	1088,44	362,81	7,818	0,001952
Erreur	16	742,50	46,41		
Total	19	1830,94			



EXPÉRIENCES AVEC UN FACTEUR

Modèle à type d'effets : effet général + effet différentiel

(1) $Y_{ir} = \mu_i + \varepsilon_{ir}$

$\mu_i = \mu + (\mu_i - \mu) = \mu + \tau_i \qquad \sum \tau_i = 0$

(2) $Y_{ir} = \mu + \tau_i + \varepsilon_{ir} \qquad i = 1, 2, \ldots, g \qquad r = 1, 2, \ldots, n_i$

Y_{ij} : valeur de la variable de réponse r-ème essai modalité i du facteur A

μ : effet général

τ_i : effet différentiel de la modalité i du facteur

ε_{ir} : erreurs aléatoires indépendantes $\sim N(0, \sigma^2)$

Définition de μ : 2 possibilités

définition 1 $\quad \mu = \sum \mu_i / g \qquad \sum \tau_i = 0$

définition 2 $\quad \mu = \sum \omega_i \mu_i \qquad \sum \omega_i = 1 \qquad \sum \omega_i \tau_i = 0$

exemple A définition 2 plus générale que définition 1

exemple: parc véhicules automobiles composée de
50 % compactes 30% berlines 20% VUS
Y : consommation essence

$E(Y) = \mu = 0{,}5 * \mu_1 + 0{,}3 * \mu_2 + 0{,}2 * \mu_3$

$\omega_1 = 0{,}5 \quad \omega_2 = 0{,}3 \quad \omega_4 = 0{,}2$

EXPÉRIENCES AVEC UN FACTEUR

exemple B $\omega_i = n_i / N$ taille des échantillons

si $n_i = n$ $\omega_i = 1/g$ définition 1

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$ $H_a : \mu_i$ pas tous égaux

devient

$H_0 : \tau_1 = \tau_2 = \ldots = \tau_g = 0$ $H_a : \tau_i \neq 0$ au moins un i

Approche par régression avec un codage à effet

$$Y_{ir} = \mu + \tau_i + \varepsilon_{ir} \qquad \sum \tau_i = 0$$

$\tau_g = - \tau_1 - \tau_2 - \ldots - \tau_{g-1}$ g = nombre de sous-groupes

variables de codage X_{irt} $t = 1, 2, 3, \ldots, g-1$ $r = 1, 2, \ldots, n_i$

$$X_{irt} = \begin{cases} 1 & \text{si observation groupe } i = 1, 2, \ldots, g-1 \\ -1 & \text{si observation provient groupe g} \\ 0 & \text{autrement} \end{cases}$$

exemple g = 4

(3) $$Y_{ir} = \mu + \tau_1 X_{ir1} + \tau_2 X_{ir2} + \tau_3 X_{ir3} + \varepsilon_{ir}$$

Exemple : saut en hauteur

Soccer	38
Soccer	43
Soccer	33
Soccer	40
Soccer	35
Tennis	45
Tennis	53
Tennis	38
Tennis	55
Tennis	50
Football	55
Football	68
Football	43
Football	45
Football	53
Basketball	60
Basketball	65
Basketball	55
Basketball	53
Basketball	55

38
43
33
40
35
45
53
38
55
50
55
68
43
45
53
60
65
55
53
55

=

X_{ir0}	X_{ir1}	X_{ir2}	X_{ir3}
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	1	0	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	1	0
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	0	0	1
1	-1	-1	-1
1	-1	-1	-1
1	-1	-1	-1
1	-1	-1	-1
1	-1	-1	-1

μ
τ_1
τ_2
τ_3

+

ε_{11}
ε_{12}
ε_{13}
ε_{14}
ε_{15}
.
.
.
.
.
.
.
.
.
.
ε_{41}
.
.
.
ε_{45}

$$Y = X \beta + \varepsilon$$

Exemple : saut en hauteur - analyse par régression

R = 0,77	R² = 0,594	Adjusted R² = 0,518	F(3,16) = 7,82	p = 0,00195		
	Beta	Std.Err.	b	Std.Err.	t(16)	p-level
Intercept			48.88	1.52	32.09	0.0000
X1	-0.841	0.195	-11.38	2.64	-4.31	0.0005
X2	-0.065	0.195	-0.88	2.64	-0.33	0.7445
X3	0.268	0.195	3.63	2.64	1.37	0.1884

Analysis of Variance; DV: Y-hauteur saut (cm)					
	SS	df	MS	F	p-level
Regress.	1088.44	3	362.81	7.82	0.00195
Residual	742.50	16	46.41		
Total	1830.94				

ANOVA identique page 9

EXPÉRIENCES AVEC UN FACTEUR

nombre d'observations n_i = ?

Puissance du test F = probabilité de rejeter H_0 si H_0 est fausse

Puissance = Prob ($F > F_{1-\alpha,\, g-1,\, N-g} \mid \Phi$) = $H(\Phi, g, N, \alpha)$

distribution F non centrale avec paramètre de non-centralité Φ

test avec seuil = α fixé (0,05 ou 0,01)

$$\Phi = (1/\sigma) \left[\sum n_i (\mu_i - \bar{\mu})^2 / g \right]^{0.5} \qquad \bar{\mu} = \Sigma \mu_i / g$$

$$n_i = n \qquad N = ng \qquad \Phi = (1/\sigma) (n/g)^{0,5} \left[\Sigma (\bar{\mu}_i - \mu)^2 \right]^{0.5}$$

puissance = 1 - β fixée cas fréquents 1 - β = 0,80 0,90 0,95

$\Delta = \max(\mu_i) - \min(\mu_i)$ n = fonction (g, α, $1 - \beta$, Δ/σ)

table: extrait Kunter & all 5 ed. p. 1343 page suivante

Δ/σ = 1,0 / 1,5 / 2,0 g = 2, 3,… , 10 α = 0,10 / 0,05 / 0,01

EXPÉRIENCES AVEC UN FACTEUR

puissance

= 1 - β

= 0,90

	Δ/σ = 1	Δ/σ = 1	Δ/σ = 1	Δ/σ = 1.5	Δ/σ = 1.5	Δ/σ = 1.5	Δ/σ = 2	Δ/σ = 2	Δ/σ = 2
g	α = 0,1	α = 0,05	α = 0,01	α = 0,10	α = 0,05	α = 0,01	α = 0,10	α = 0,05	α = 0,01
2	18	23	32	9	11	15	6	7	10
3	22	27	37	11	13	18	7	8	11
4	25	30	40	12	14	19	7	9	12
5	27	32	43	13	15	20	8	9	12
6	29	34	46	14	16	21	8	10	13
7	31	36	48	14	17	22	9	10	13
8	32	38	50	15	18	23	9	11	14
9	33	40	52	16	18	24	9	11	14
10	35	41	54	16	19	25	10	11	15

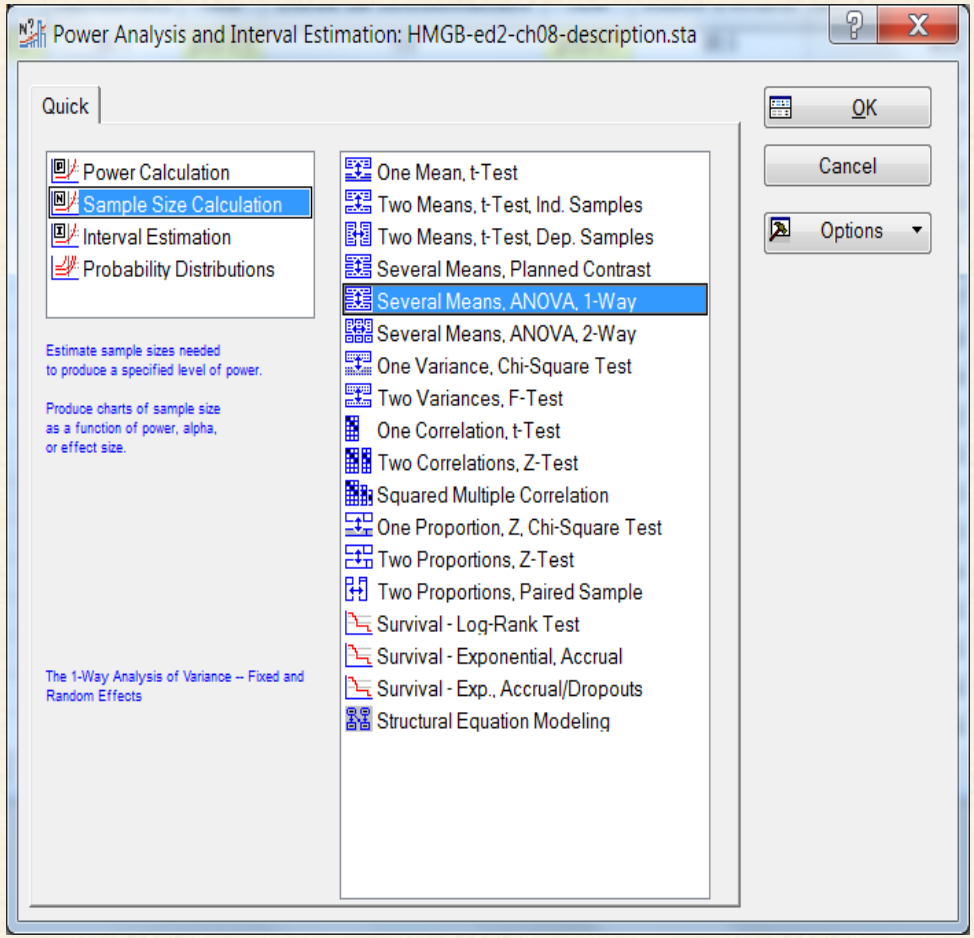
puissance

= 1 - β

= 0,95

g	α = 0,1	α = 0,05	α = 0,01	α = 0,1	α = 0,05	α = 0,01	α = 0,1	α = 0,05	α = 0,01
2	23	27	38	11	13	18	7	8	11
3	27	32	43	13	15	20	8	9	12
4	30	36	47	14	17	22	9	10	13
5	33	39	51	15	18	23	9	11	14
6	35	41	53	16	19	25	10	11	15
7	37	43	56	17	20	26	10	12	15
8	39	45	58	18	21	27	11	12	16
9	40	47	60	19	22	28	11	13	16
10	42	48	62	19	22	29	11	13	17

Utilisation de *STATISTICA* pour le calcul de la taille échantillonnale n



RMSSE

$$= \text{sqrt}\ \{\ [\Sigma\ (\ a_j\ /\ \sigma\ )^2\ ]\ /\ g - 1\ \}$$

$$= \text{sqrt}\ (\ s_a^2\ /\ \sigma^2\ )$$

$$= s_a\ /\ \sigma$$

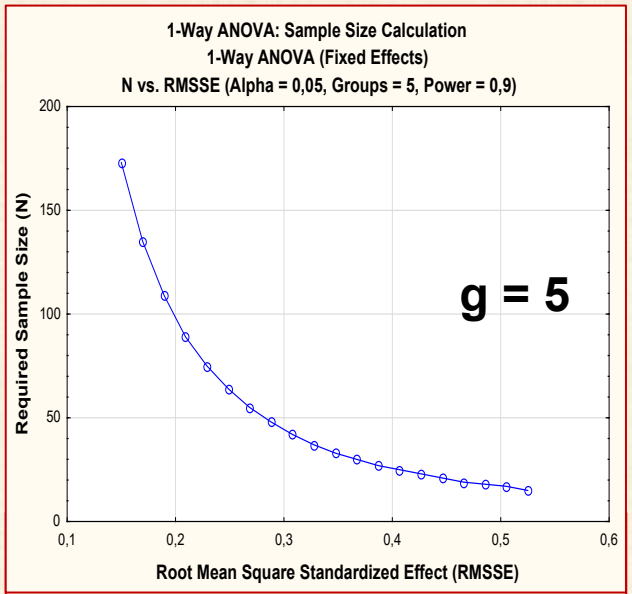
= sum of squared standardized effects, divided by the number of effects that are free to vary in the experiment

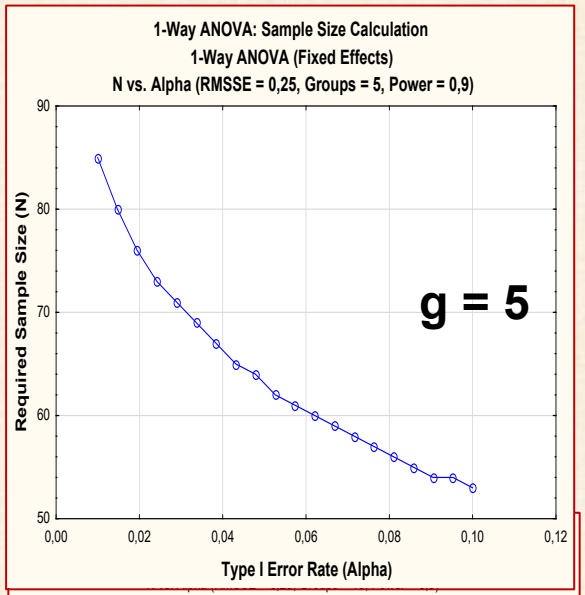
rough guidelines

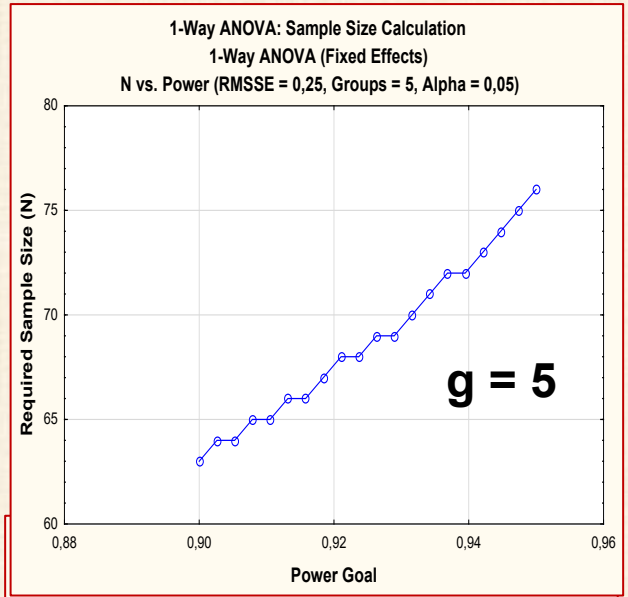
0,15 =	small	effects
0,30 =	medium	effects
0,50 =	large	effects

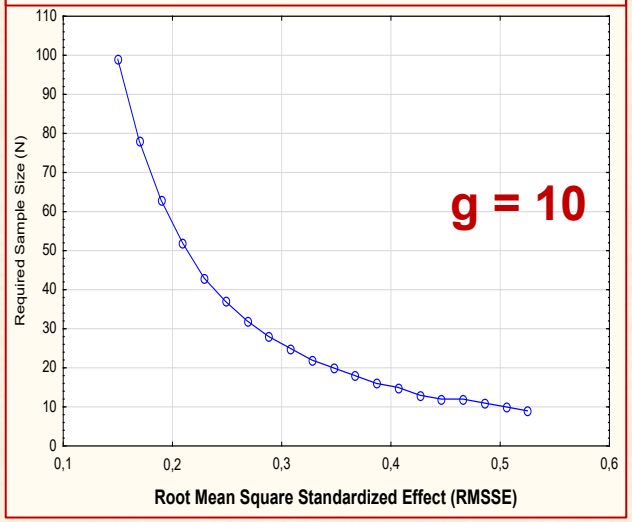
Cohen, J. (1983). *Statistical Power Analysis for the Behavioral Sciences. (2nd Ed.).* Mahwah, NJ: Lawrence Erlbaum Ass.

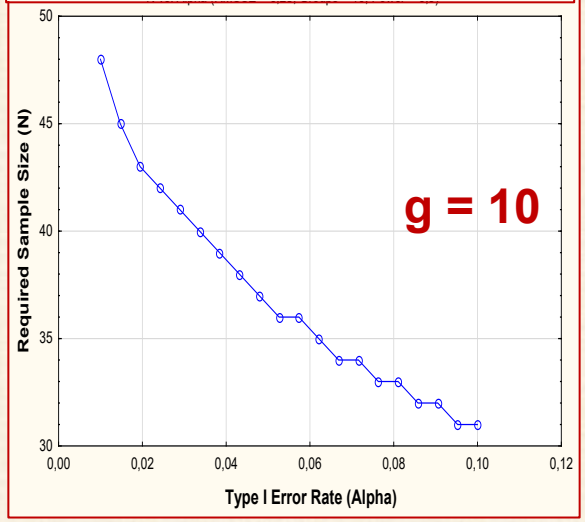
EXPÉRIENCES AVEC UN FACTEUR

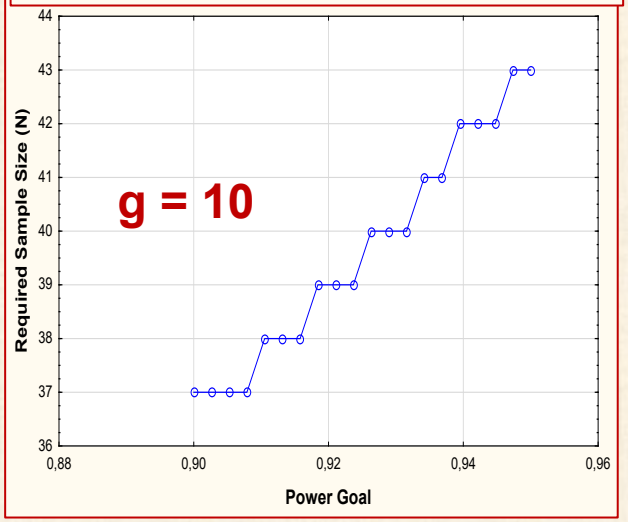








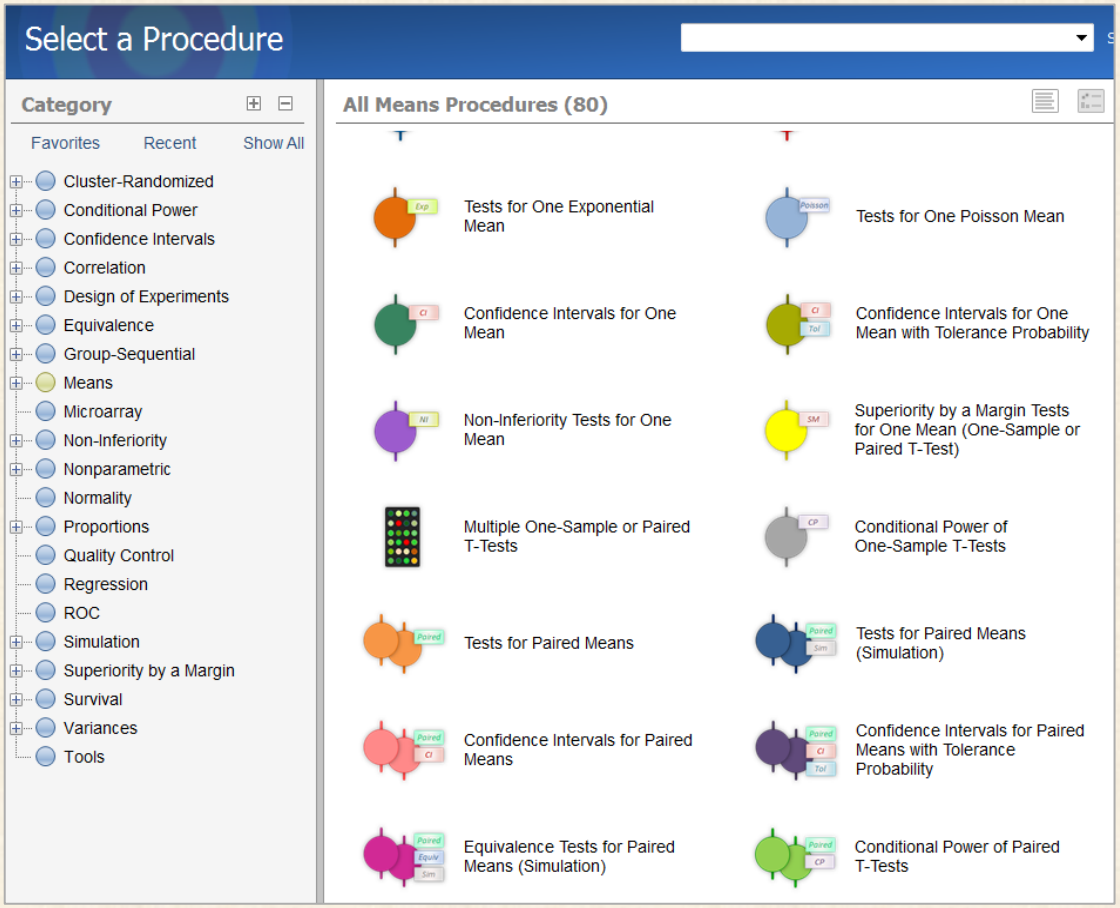




Utilisation de PASS : Power Analysis Sample Size

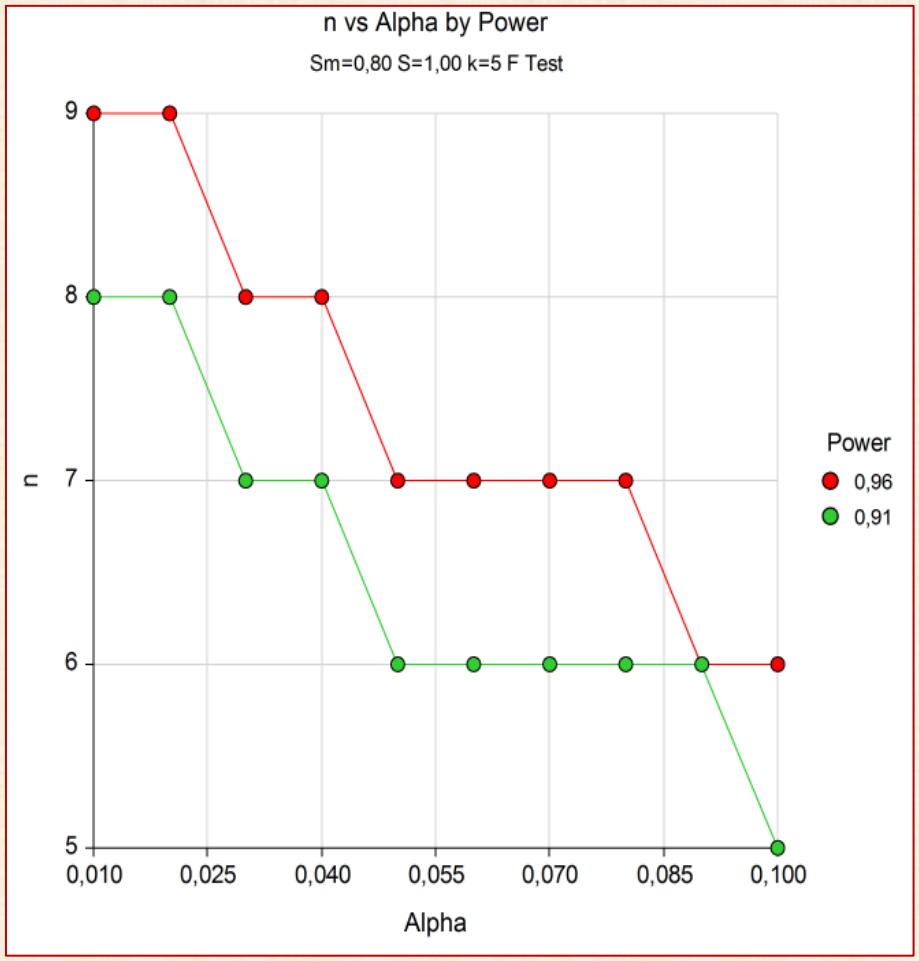
https://www.ncss.com/software/pass/

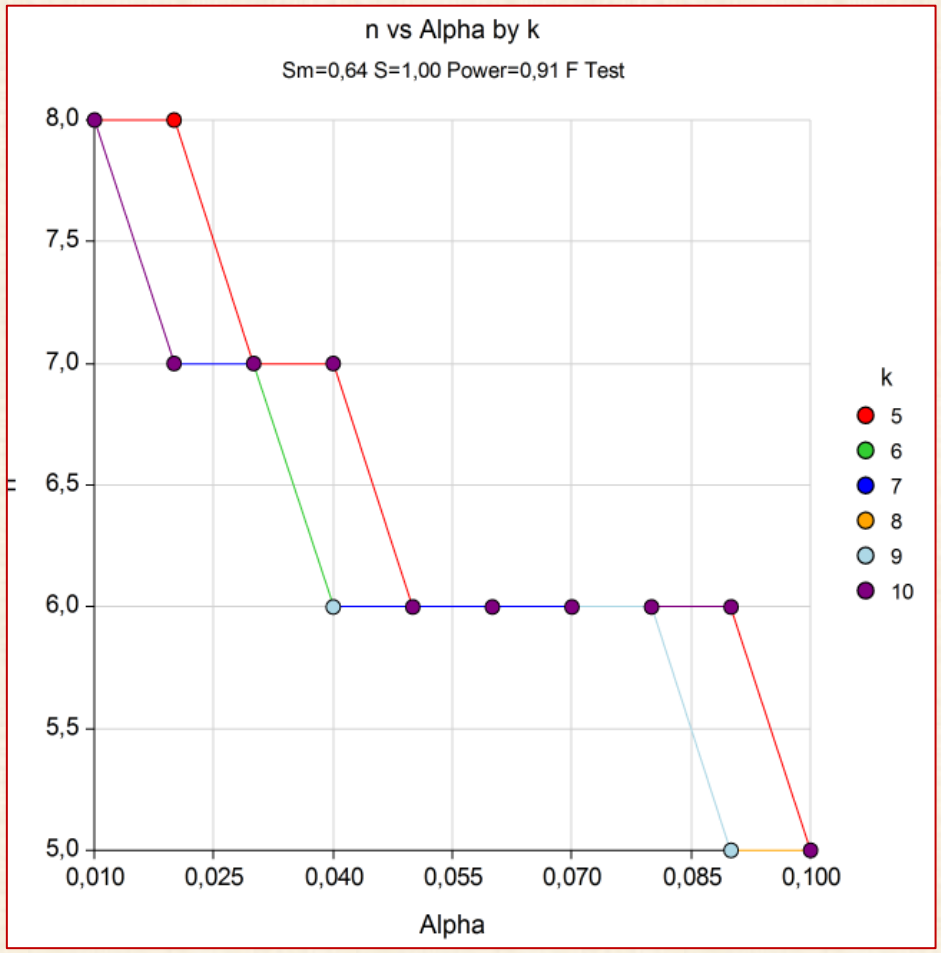
Le logiciel le plus complet sur le marché pour le calcul de la taille échantillonnale : plus de 1 000 procédures



EXPÉRIENCES AVEC UN FACTEUR

Utilisation de PASS : Power Analysis Sample Size





ANALYSE DES MOYENNES et COMPARAISONS MULTIPLES

Si test F est significatif : moyennes sont statistiquement différentes

Peut-on dire plus? Sont-elles toutes différentes ?

Sinon, quelle moyenne diffère de quelle autre?
Faire des comparaisons (contrastes) entre des groupes de moyennes?

Questions = l'analyse a posteriori (post-hoc) des moyennes.
Problème de comparaisons multiples sur les données.
Faut contrôler les risques associés à ces comparaisons multiples.

On veut contrôler le risque - avoir un coefficient de confiance global de $1 - \alpha$ sur l'ensemble des comparaisons (tests).

Si on fait un nombre de k comparaisons, chacune avec un coefficient de confiance de $1 - \alpha$, alors le coefficient de confiance global sur l'ensemble des k comparaisons diminue.

Plus on augmente le nombre de comparaisons (tests)

plus on augmente les chances de conclure à tort.

Le tableau p. 22 illustre le problème.

EXPÉRIENCES AVEC UN FACTEUR

nombre de modalités g	nombre de comparaisons k = g*(g-1)/2	coefficent de confiance global $(1 - \alpha)^k$	$1 - \alpha$ = 0,95
2	1	$1 - \alpha$	0,95
3	3	$(1 - \alpha)^3$	0,86
4	6	$(1 - \alpha)^6$	0,735
5	10	$(1 - \alpha)^{10}$	0,60
6	15	$(1 - \alpha)^{15}$	0,46
8	28	$(1 - \alpha)^{28}$	0,24
10	45	$(1 - \alpha)^{45}$	0,10

2 catégories de tests

- tests (comparaisons) planifiés **avant** l'exécution des calculs
- tests suggérés **après** l'analyse

post hoc ≈ a posteriori (« data snooping »)

EXPÉRIENCES AVEC UN FACTEUR

A - intervalle de confiance pour une moyenne particulière

$$\mu_i : \overline{Y_{i.}} \pm t(1 - \alpha/2, N - g) * MSE^{0.5} * [\, 1 / n_i \,]^{0,5}$$

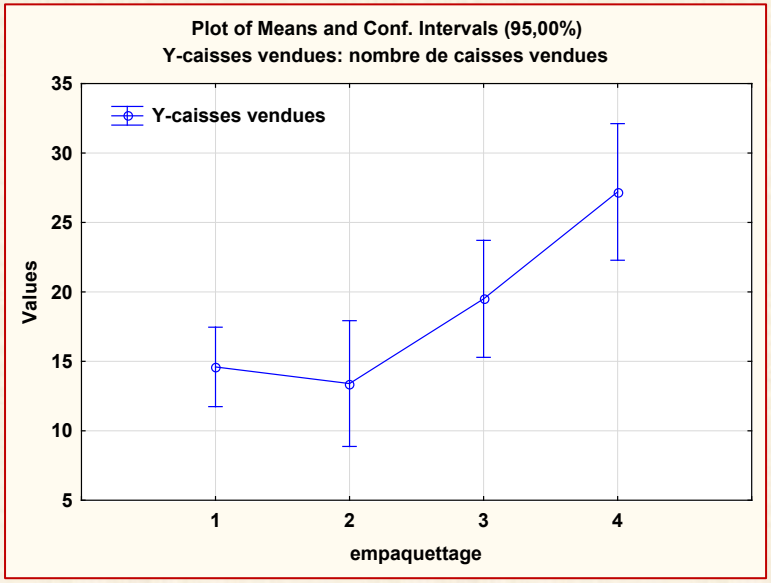
$t(1 - \alpha/2, N - g)$: $(1 - \alpha/2)^{ième}$ percentile loi T de Student

avec $(N - g)$ degrés de liberté $1 - \alpha$: coefficient de confiance

1 ID	2 empaquettage	3 magasin	4 Y-caisses vendues
1	e1	m1	11
2	e1	m2	17
3	e1	m3	16
4	e1	m4	14
5	e1	m5	15
6	e2	m1	12
7	e2	m2	10
8	e2	m3	15
9	e2	m4	19
10	e2	m5	11
11	e3	m1	23
12	e3	m2	20
13	e3	m3	18
14	e3	m4	17
15	e4	m1	27
16	e4	m2	33
17	e4	m3	22
18	e4	m4	26
19	e4	m5	28

Y-caisses vendues

Emp	Means	n	SD
1	14,6	5	2,30
2	13,4	5	3,65
3	19,5	4	2,65
4	27,2	5	3,96
all	18,63	19	6,44



exemple

$\overline{Y_{1.}} = 14,6$ MSE = 10,55 $1 - \alpha = 0.95$ $t(0.95, 15) = 2,13$

$$11,5 \leq \mu_1 \leq 17,7$$

EXPÉRIENCES AVEC UN FACTEUR

B - intervalle de confiance pour la différence entre 2 moyennes

$$\mu_i - \mu_{i'} : (\overline{Y}_{i.} - \overline{Y}_{i'.}) \pm t(1-\alpha/2, N-g) * MSE^{0.5} * [(1/n_i) + (1/n_{i'})]^{0.5}$$

exemple $\mu_3 - \mu_4 : (19{,}5 - 27{,}2) \pm 2{,}13 * 10{,}55^{0.5} * [(1/5) + (1/4)]^{0.5}$

$$-12{,}3 \leq \mu_3 - \mu_4 \leq -3{,}7$$

C - contraste = comparaison

$L = \Sigma c_i \mu_I \quad \Sigma c_i = 0 \quad L = \Sigma c_i \overline{Y}_{i.} \quad s(L) = MSE^{0.5} * [\Sigma c_i^2 / n_i]^{0.5}$

exemple 1&2 vs 3&4 $\quad L = 0{,}5*(\hat{\mu}_1 + \hat{\mu}_2) - 0{,}5*(\hat{\mu}_3 + \hat{\mu}_4)$

$L = -9{,}35 \quad \Sigma c_i^2 / n_i = 0{,}2125 \quad s(L) = 2{,}24$

$$-12{,}5 \leq L \leq -6{,}2$$

EXPÉRIENCES AVEC UN FACTEUR

Procédures d'inférences simultanées (= comparaisons multiples)

Les méthodes A – B – C ont deux limitations :

- le coefficient de confiance 1 – α et le seuil α d'un test s'applique à UN test seulement.
- Le test ou la comparaison ne doit pas être suggéré par les données (« data snooping »).

solution problème

utiliser une procédure de comparaison multiple qui inclut toutes les inférences possibles qui peuvent être anticipées et d'intérêt après que les données furent examinées.

Par exemple, on peut s'intéresser à toutes les comparaisons définies par les différences entre toutes les paires de moyennes.

3 procédures pour faire de l'inférence après avoir vu les données et en contrôlant le coefficient de confiance:

- méthode de Tukey («HSD = Honest Significant Differences »)
- méthode de Scheffé pour les contrastes
- méthode de Bonferronni pour les comparaisons prédéfinies

Méthode de Tukey

dédiée sur les comparaisons (contrastes spécifiques) définies par les différences entre toutes les moyennes prises 2 à 2

basée sur distribution « Studentized Range »

$Y_1, Y_2, \ldots, Y_g$: g observations indépendantes d'une population $N(\mu , \sigma^2)$

$W = \max(Y_1, Y_2, .., Y_g) - \min (Y_1, Y_2, .., Y_g)$: étendue ("range")

S^2 : estimation de σ^2 basée sur υ degrés de liberté

$Q (g, \upsilon) = W / S$ « studentized range »

valeurs $q(0,95; g; \upsilon)$ extrait table Kutner et all 5 ed. p. 1334

υ	g = 2	g = 3	g = 4	g = 5	g = 10	g = 15	g = 20
2	6,08	8.33	9.80	10.9	14.0	15.7	16.8
5	3,64	4.60	5.22	5.67	6.80	7.72	8.21
10	3,15	3.88	4.33	4.65	5.60	6.11	6.47
20	2,95	3.58	3.96	4.23	5.01	5.43	5.71
40	2,86	3.44	3.79	4.04	4.73	5.11	5.36
60	2,83	3.40	3.74	3.98	4.65	5.00	5.24
120	2,80	3.36	3.68	3.92	4.56	4.90	5.13
infini	2,77	3.31	3.63	3.86	4.47	4.80	5.01

EXPÉRIENCES AVEC UN FACTEUR

méthode de Tukey

$$D = \mu_i - \mu_{i'} \qquad \widehat{D} = \overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}$$

$$s^2(\widehat{D}) = MSE * [(1/n_i) + (1/n_{i'})]$$

$$D : \widehat{D} \pm 0.707 * q(1 - \alpha; g, N - g) * s(\widehat{D})$$

Tukey HSD test; variable Y-caisses vendues
probabilities for Post Hoc Tests Error:
Between MS = 10.547 df = 15

Empaquetage	{1} 14,6	{2} 13,4	{3} 19.5	{4} 27.2
1		0.9354	0.1550	0.0003
2	0.9354		0.0584	0.0002
3	0.1550	0.0584		0.0143
4	0.0003	0.0002	0.0143	

EXPÉRIENCES AVEC UN FACTEUR

Méthode de Scheffé — la plus générale

$$L = \Sigma c_i \mu_I \quad \Sigma c_i = 0 \quad L = \Sigma c_i Y_{i.} \quad s(\hat{L}) = MSE^{0.5} * [\Sigma c_i^2 / n_i]^{0.5}$$

$$L : \hat{L} \pm (g-1) * F(1-\alpha, g-1, N-g) * s(\hat{L})$$

empaquettage		Scheffe Test; Variable: Y-caisses vendues Marked differences are significant at p < ,05000			
		{1} M=14,600	{2} M=13,400	{3} M=19,500	{4} M=27,200
1	{1}		**0,950675**	**0,212530**	**0,000229**
2	{2}	**0,950675**		**0,089489**	**0,000086**
3	{3}	**0,212530**	**0,089489**		**0,024782**
4	{4}	**0,000229**	**0,000086**	**0,024782**	

Méthode de Bonferroni

$$L : \hat{L} \pm t(1 - (\alpha/2g), N-g) * s(\hat{L})$$

Cell No.	Bonferroni test; variable Y-caisses vendues Probabilities for Post Hoc Tests Error: Between MS = 10.547, df = 15.000				
	empaquetage	{1} 14.600	{2} 13.400	{3} 19.500	{4} 27.200
1	**1**		**1,0000**	**0,2397**	**0,0001**
2	**2**	**1,0000**		**0,0808**	**0,0000**
3	**3**	**0,2397**	**0,0808**		**0,0180**
4	**4**	**0,0001**	**0,0000**	**0,0180**	

EXPÉRIENCES AVEC UN FACTEUR

Comparaison des méthodes

- Si on veut seulement faire des comparaisons entre les paires, la procédure de Tukey est supérieure - méthode recommandée
- Si test F rejette l'égalité des moyennes : il existe au moins un contraste qui diffère de zéro parmi tous les contrastes.
- Procédure de Bonferroni est préférable à la procédure de Scheffé si le nombre de contrastes d'intérêt est à peu près le même que le nombre de modalités.
- Il existe d'autres procédures pour des fonctions spécialisées. Exemple : procédure de Dunnett pour comparer chaque traitement avec un traitement contrôle ;
- Procédure de Hsu : pour choisir le « meilleur » traitement.

ANOM : Analyis Of Means (Ott) méthode alternative au test F

Basée sur l'ensemble des tests de l'effet différentiel de chaque modalité.

avantage : représentation graphique semblable à une carte contrôle

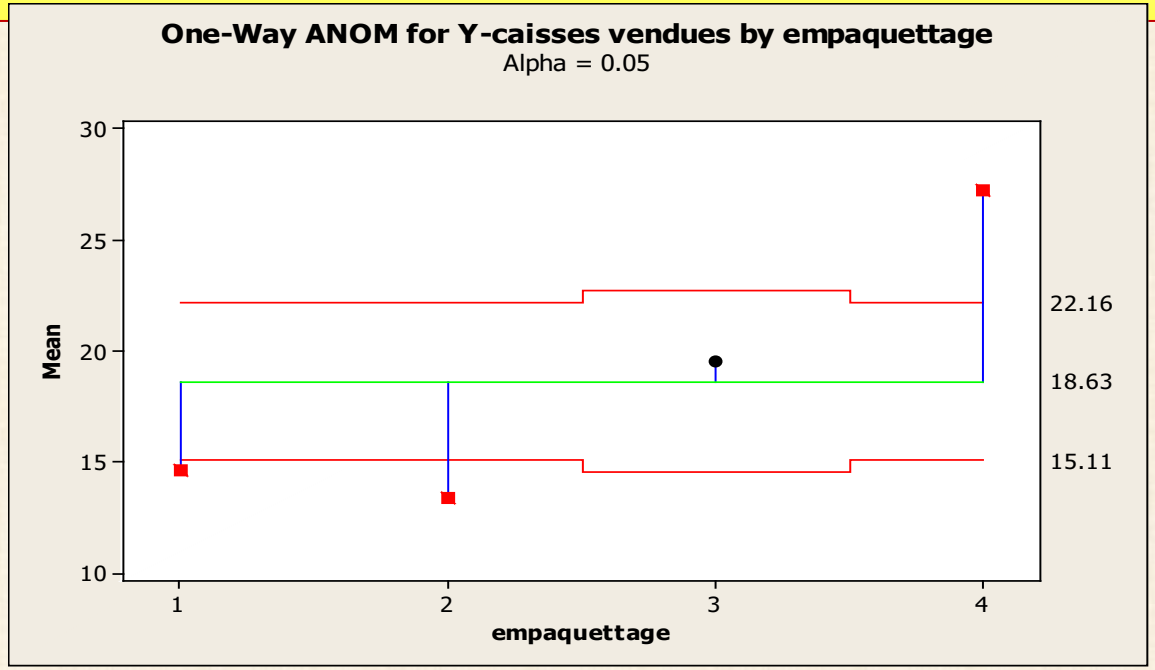
$$\hat{\tau}_i = \bar{Y}_{i.} - \bar{Y}_{..}$$

$$s^2(\hat{\tau}_i) = MSE\left[((g-1)/g)^2 (1/n_i) + (1/g^2) \left(\sum_{h \neq i} (1/n_h) \right) \right.$$

ANOM : test si les moyennes diffèrent de la moyenne globale

ANOVA : test si les moyennes sont différentes

avec Minitab



DIAGNOSTICS ET MESURES CORRECTIVES

- **Diagnostic** : écarts importants par rapport aux hypothèses de base
- **Si oui :** mesures correctives = ?

Analyse diagnostique : basée sur les résidus et des graphiques

…. comme l'analyse de régression

4 types de résidus

$e_{ij} = Y_{ij} - Y_{ij} = Y_{ij} - Y_{i.}$ résidu brut

$e_{ij}^* = e_{ij} / MSE^{0.5}$ résidu semi studentisé

$r_{ij} = e_{ij}^* / [(n_i - 1) / n_i]^{0.5}$ résidu studentisé

$t_{ij} = e_{ij}^* [(N - g - 1) / (SSE [1 - (1/n_i)] - e_{ij}^2)]^{0.5}$

résidu studentisé avec observation supprimée
studentized delete residual

Écarts du modèle d'ANOVA - ordre d'importance décroissante

1. variance non constante
2. erreurs (observations) non indépendantes
3. présence de valeurs aberrantes
4. normalité du terme d'erreur
5. omission de variables explicatives importantes

Analyse des résidus : tout modèle statistique

- vérifier les hypothèses de base après ajustement d'un modèle

HYPOTHESES de BASE (ordre d'importance) **COMMENT?**

H1 - variance constante? … tests de Hartley /Cochran / Bartlett / Levene

H2 - données aberrantes? ('outliers') …… résidus vs résidus 'deleted'

H3 - distribution normale? …… résidus sur échelle gaussienne (normale)

H4 - « bon » modèle? … R^2 et R^2 ajusté élevé: pas absolument nécessaire

H5 - observations indépendantes? ….. test Durbin-Watson

- Si hypothèses de base violées et / ou certaines formes réponse
  - transformation de la réponse Y : Box-Cox Y^{λ} $-2 < \lambda < 2$ $\lambda = ?$
  - test non paramétrique de Kruskall-Wallis

- H1 critique seulement si max var(Y) / min var(Y) > 10 (inter groupes)
- H2 on ne veut pas que l'analyse soit dominée par quelques observations
- H3 le test F est robuste vis-à-vis la non normalité
- H4 est utile en analyse de régression
  pas utile en modèle d' analyse de variance (variables X catégoriques)
- H5 surtout pour données observationnelles collectées à forte cadence
  pas le cas si les données proviennent d'expériences

DIAGNOSTICS ET MESURES CORRECTIVES

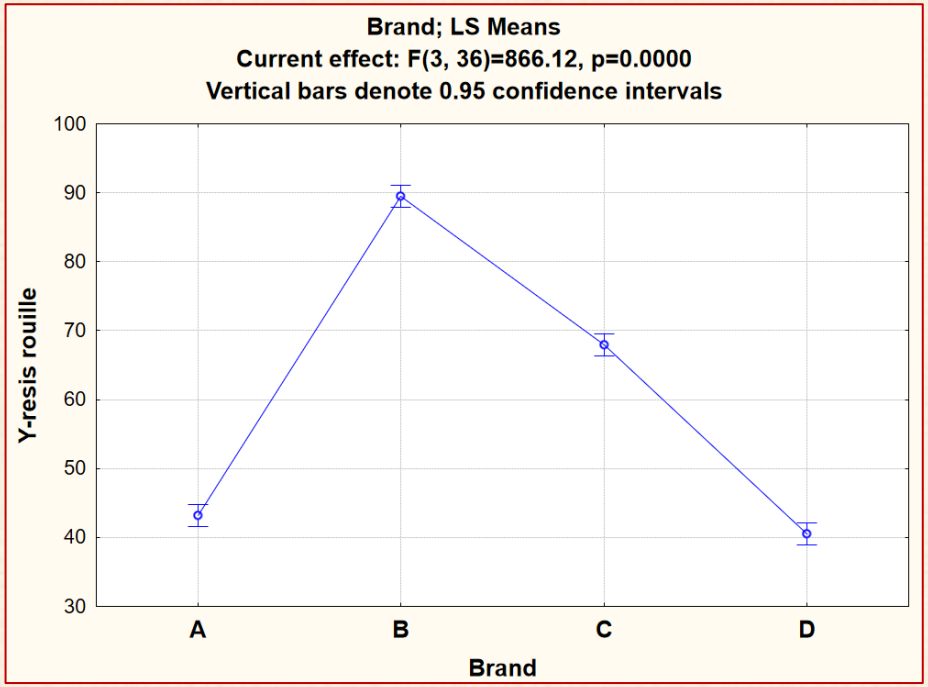
HYPOTHÈSES	VÉRIFICATION	DIAGNOSTIC
variance non constante	graphique de résidus studentisés VS valeurs prédites	- bande horizontale - tests : Hartley Brown-Forsythe
non indépendance	si l'ordre temporel est connu	- résidus VS temps - test d'indépendance sérielle
valeurs aberrantes	t_{ij} VS valeurs prédites	-
normalité	résidus sur échelle de probabilité gaussienne	écart par rapport à la droite-
omisssion	résidus VS valeurs prédites	résidus corrélés avec autres facteurs non tenu en compte

graphique des résidus (axe vertical) VS : Y prédits, Y observés, X
si allure = bande horizontale alors modèle OK

DIAGNOSTICS ET MESURES CORRECTIVES

Exemple : rouille

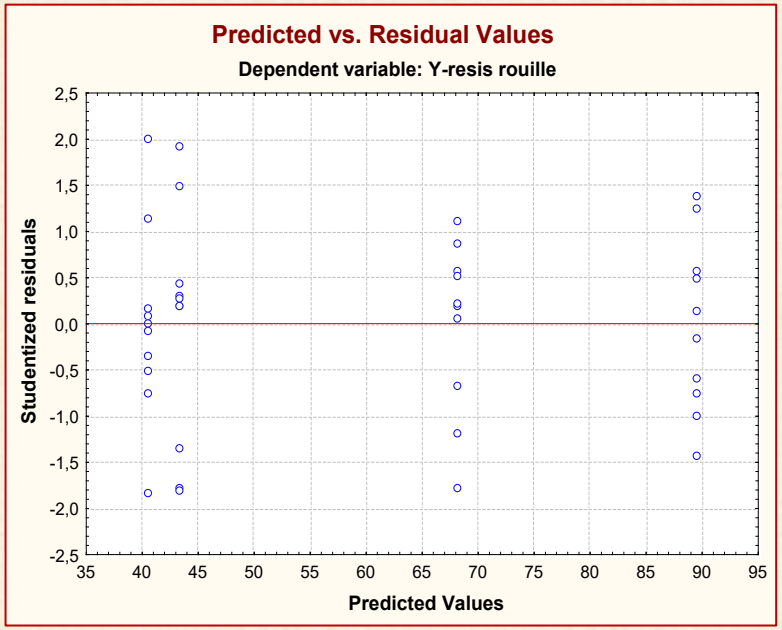
1 ID	2 facteur Brand	3 rep	4 Y-resis rouille
1	A	1	43,9
2	A	2	39,0
3	A	3	46,7
4	A	4	43,8
5	A	5	44,2
6	A	6	47,7
7	A	7	43,6
8	A	8	38,9
9	A	9	43,6
10	A	10	40,0
11	B	1	89,8
12	B	2	87,1
13	B	3	92,7
14	B	4	90,6
15	B	5	87,7
16	B	6	92,4
17	B	7	86,1
18	B	8	88,1
19	B	9	90,8
20	B	10	89,1
21	C	1	68,4
22	C	2	69,3
23	C	3	68,5
24	C	4	66,4
25	C	5	70,0
26	C	6	68,1
27	C	7	70,6
28	C	8	65,2
29	C	9	63,8
30	C	10	69,2
31	D	1	36,2
32	D	2	45,2
33	D	3	40,7
34	D	4	40,5
35	D	5	39,3
36	D	6	40,3
37	D	7	43,2
38	D	8	38,7
39	D	9	40,9
40	D	10	39,7

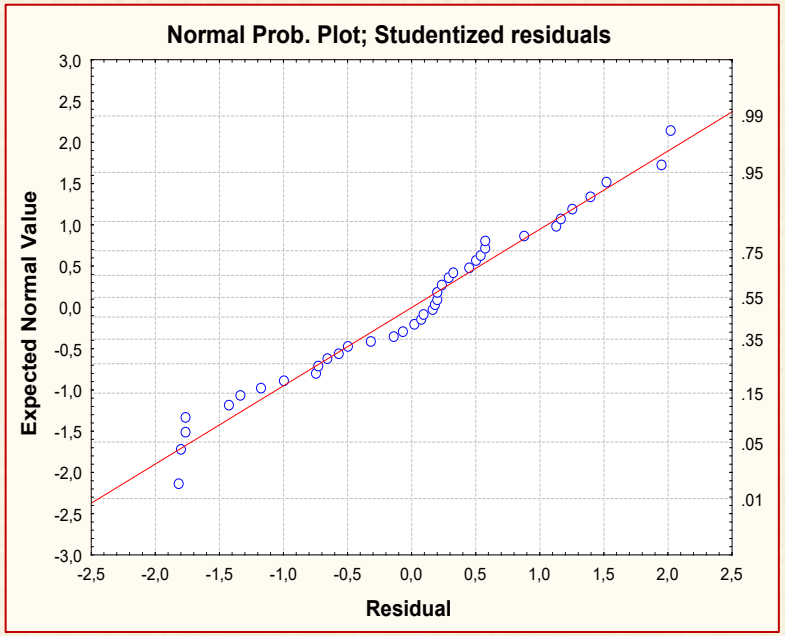


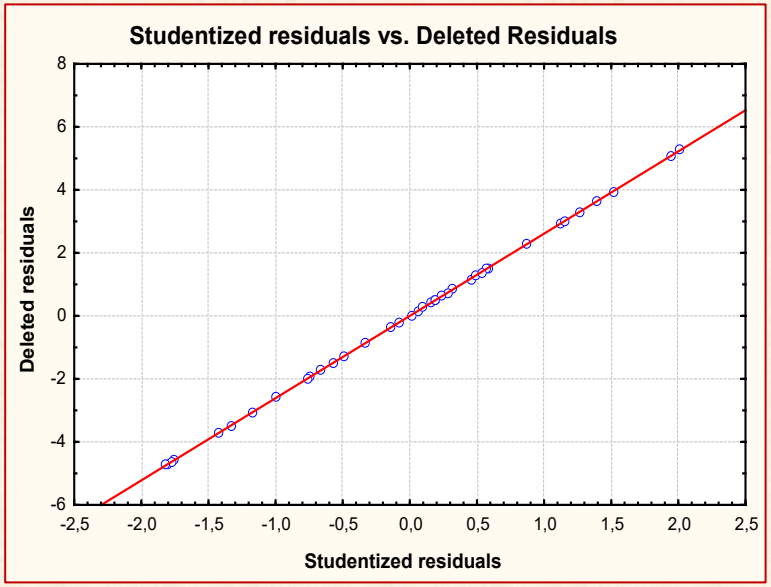
Source	Degr. of Freedom	Y-resis rouille SS	Y-resis rouille MS	Y-resis rouille F	Y-resis rouille p
Intercept	1	145202,5	145202,5	23649,26	0,0000
Brand	3	15953,5	5317,8	866,12	0,0000
Error	36	221,0	6,1		
Total	39	16174,5			

DIAGNOSTICS ET MESURES CORRECTIVES

Exemple : rouille







DIAGNOSTICS ET MESURES CORRECTIVES

Tests : homogénéité de la variance 5 tests

Hartley – Bartlett – Cochran - Brown-Forsythe, Levene

Test de Hartley exigence : $n_i = n$ + normalité

H_0 : $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2$

H_a : les variances ne sont pas toutes égales

Hartley : $H^* = \max(s_i^2) / \min(s_i^2)$

Rejet de H_0 si $H > H(1-\alpha, g, n-1)$

$H(1-\alpha, g, df)$: $(1-\alpha)$ percentile distribution de Hartley

Exemple : rouille

marque	# obs	moyenne	écart type	variance
tous	40	60.25	20.36	414.53
A	10	43.14	3.00	9.00
B	10	89.44	2.22	4.93
C	10	67.95	2.17	4.71
D	10	40.47	2.44	5.95

	Hartley F-max	Cochran C	Bartlett Chi-Sqr.	df	p
Y-resis rouille	1,913857	0,366478	1,198957	3	0,753255

on ne rejette pas H_0

DIAGNOSTICS ET MESURES CORRECTIVES

Test de Brown-Forsythe n_i peuvent être inégaux
test robuste à la non-normalité

$d_{ij} = | y_{ij} - \text{med}(y_{ij}) |$ med (y) = médiane (y)

FBF = MSTR / MSE

$MSTR = \Sigma n_i * (\overline{d}_{i.} - \overline{d}_{..})^2 / (g - 1)$

$MSE = \Sigma\Sigma(d_{ij} - \overline{d}_{i.})^2 / (N - g)$

$\overline{d}_{i.} = \Sigma d_{ij} / n_i$ $\overline{d}_{..} = \Sigma\Sigma d_{ij} / N$

FBF suit approximativement loi $F(g - 1, N - g)$

rejet H_0 si FBF $>$ $F(1- \alpha, g - 1, N - g)$

Test de Levene $d_{ij} = |y_{ij} - \text{moy}(y_{ij})|$

moy (y) = moyenne (y)

modification test de Levene = test de Brown-Forsythe

DIAGNOSTICS ET MESURES CORRECTIVES

Test de Cochran $n_i = n$

$C = \max(s_i^2) / \sum s_i^2$

loi d'échantillonage de C dépend de g et de n

rejet H_0 si $C > C(1 - \alpha ; n, g)$

tableau des percentiles de la distribution de C : $C(1 - \alpha ; n, g)$
Statistical Principles in Experimental Design, 2 ed.,
B.J. Winer, 1971, McGraww-Hill, p. 876)

n	percentile 1 - α	g = 2	g = 3	g = 4	g = 5	g = 8	g = 10
2	0.95	0.9985	0.9669	0.9065	0.8412	0.6798	0.6020
	0.99	0.9999	0.9933	0.9676	0.9279	0.7945	0.7175
3	0.95	0.9750	0.8709	0.7679	0.6838	0.5157	0.4450
	0.99	0.9950	0.9423	0.8643	0.7885	0.6152	0.5358
4	0.95	0.9392	0.7977	0.6841	0.5981	0.4377	0.3733
	0.99	0.9794	0.8831	0.7814	0.6957	0.5209	0.4469
5	0.95	0.9057	0.7457	0.6287	0.5441	0.3910	0.3311
	0.99	0.9586	0.8335	0.7212	0.6329	0.4627	0.3934
6	0.95	0.8772	0.7071	0.5895	0.5065	0.3595	0.3029
	0.99	0.9373	0.7933	0.6761	0.5875	0.4226	0.3572
8	0.95	0.8332	0.6530	0.5365	0.4564	0.3185	0.2666
	0.99	0.8988	0.7335	0.6129	0.5229	0.3704	0.3106
10	0.95	0.8010	0.6167	0.5017	0.4387	0.2926	0.2439
	0.99	0.8674	0.6912	0.5702	0.5037	0.3373	0.2813
17	0.95	0.7341	0.5466	0.4366	0.3645	0.2462	0.2032
	0.99	0.7949	0.6059	0.4884	0.4094	0.2779	0.2297
37	0.95	0.6602	0.4748	0.3720	0.3066	0.2022	0.1655
	0.99	0.7067	0.5153	0.4057	0.3351	0.2214	0.1811

DIAGNOSTICS ET MESURES CORRECTIVES

Test de Bartlett n_i inégaux

$$C = 1 + (1/3*(g - 1))*[\sum (1/(n_i-1) - (1/N)]$$

$$B = (2.303/C)*[(N - g)*\log(MSE) - \sum (n_i - 1)*\log(s_i^2)]$$

B suit approximativement loi χ^2

avec (g – 1) degrés de liberté

rejet H_0 : si $B > \chi 2 (1 - \alpha ; g - 1)$

Exemple

Hartley	Cochran	Bartlett	df	p
10.445	0.5865	12.985	4	0.0113

Test	SS Effect	df Effect	MS Effect	SS Error	Df Error	MS Error	F	p
Levene	8,69	4	2,17	24,8	35	0,71	3,07	0,029
Brown-Forsythe	9,35	4	2,34	27,9	35	0,80	2,94	0,034

tests concordent : variances inégales

DIAGNOSTICS ET MESURES CORRECTIVES

VARIANCES	NORMALITÉ	MESURE CORRECTIVE
hétérogènes	oui	régression pondérée
hétérogènes	non	transformation de Box-Cox
« gros » écarts	« gros » écarts	ANOVA non paramétrique Kruskall-Wallis

régression pondérée

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad \varepsilon_{ij} \sim N(0, \sigma_i^2)$$

modèle à moyennes de cellules

poids w $\qquad w_{ij} = 1 / s_i^2$

remplacement modèle d'ANOVA par un modèle de régression avec des variables indicatrices

+ ajustement de moindres carrés pondérés avec les poids w

concept : applications pour les facteurs aléatoires

Exemple : soudure variances inégales

1 ID	2 type flux	3 rep	4 Y-force soudure
1	A	1	14,87
2	A	2	16,81
3	A	3	15,83
4	A	4	15,47
5	A	5	13,60
6	A	6	14,76
7	A	7	17,40
8	A	8	14,62
9	B	1	18,43
10	B	2	18,76
11	B	3	20,12
12	B	4	19,11
13	B	5	19,81
14	B	6	18,43
15	B	7	17,16
16	B	8	16,40
17	C	1	16,95
18	C	2	12,28
19	C	3	12,00
20	C	4	13,18
21	C	5	14,99
22	C	6	15,76
23	C	7	19,35
24	C	8	15,52
25	D	1	8,59
26	D	2	10,90
27	D	3	8,60
28	D	4	10,13
29	D	5	10,28
30	D	6	9,98
31	D	7	9,41
32	D	8	10,04
33	E	1	11,55
34	E	2	13,36
35	E	3	13,64
36	E	4	12,16
37	E	5	11,62
38	E	6	12,39
39	E	7	12,05
40	E	8	11,95

groupe	i = 1 = A	i = 2 = B	i = 3 = C	i = 4 = D	i = 5 = E
s_i^2	1,531	1,570	6,185	0,667	0,592
w_{ij}	0.653	0.637	0.162	1.449	1.689

1 ID	2 flux	3 rep	4 poids	5 Y-soudure	6 indA	7 indB	8 indC	9 indD	10 indE
1	A	1	0,653	14,87	1	0	0	0	0
2	A	2	0,653	16,81	1	0	0	0	0
3	A	3	0,653	15,83	1	0	0	0	0
4	A	4	0,653	15,47	1	0	0	0	0
5	A	5	0,653	13,60	1	0	0	0	0
6	A	6	0,653	14,76	1	0	0	0	0
7	A	7	0,653	17,40	1	0	0	0	0
8	A	8	0,653	14,62	1	0	0	0	0
9	B	1	0,637	18,43	0	1	0	0	0
10	B	2	0,637	18,76	0	1	0	0	0
11	B	3	0,637	20,12	0	1	0	0	0
12	B	4	0,637	19,11	0	1	0	0	0
13	B	5	0,637	19,81	0	1	0	0	0
14	B	6	0,637	18,43	0	1	0	0	0
15	B	7	0,637	17,16	0	1	0	0	0
16	B	8	0,637	16,40	0	1	0	0	0
17	C	1	0,162	16,95	0	0	1	0	0
18	C	2	0,162	12,28	0	0	1	0	0
19	C	3	0,162	12,00	0	0	1	0	0
20	C	4	0,162	13,18	0	0	1	0	0
21	C	5	0,162	14,99	0	0	1	0	0
22	C	6	0,162	15,76	0	0	1	0	0
23	C	7	0,162	19,35	0	0	1	0	0
24	C	8	0,162	15,52	0	0	1	0	0
25	D	1	1,499	8,59	0	0	0	1	0
26	D	2	1,499	10,90	0	0	0	1	0
27	D	3	1,499	8,60	0	0	0	1	0
28	D	4	1,499	10,13	0	0	0	1	0
29	D	5	1,499	10,28	0	0	0	1	0
30	D	6	1,499	9,98	0	0	0	1	0
31	D	7	1,499	9,41	0	0	0	1	0
32	D	8	1,499	10,04	0	0	0	1	0
33	E	1	1,689	11,55	0	0	0	0	1
34	E	2	1,689	13,36	0	0	0	0	1
35	E	3	1,689	13,64	0	0	0	0	1
36	E	4	1,689	12,16	0	0	0	0	1
37	E	5	1,689	11,62	0	0	0	0	1
38	E	6	1,689	12,39	0	0	0	0	1
39	E	7	1,689	12,05	0	0	0	0	1
40	E	8	1,689	11,95	0	0	0	0	1

régression Y sur indA indB indC indD

Var	Beta	Std.Err. of Beta	B	Std.Err. of B	t(35)	p-level
indA	0,473	0,0157	15,420	0,513	30,036	0,0000
indB	0,568	0,0157	18,528	0,513	36,089	0,0000
indC	0,460	0,0157	15,004	0,513	29,225	0,0000
indD	0,299	0,0157	9,741	0,513	18,974	0,0000
indE	0,379	0,0157	12,340	0,513	24,036	0,0000

Exemple : soudure avec variances inégales

modèle complet (F) : $Y_{ij} = \mu_1 X_{ij1} + \mu_2 X_{ij2} + ... + \mu_g X_{ijg} + \varepsilon$

modèle réduit (R) : $Y_{ij} = \mu X_{ij1} + \mu X_{ij2} + \ldots + \mu X_{ijg} + \varepsilon$

$= \mu (X_{ij1} + X_{ij2} + \ldots + X_{ijg}) + \varepsilon$

$= \mu + \varepsilon$

avec l'hypothèse $H_0 : \mu_1 = \mu_2 = \ldots = \mu_g = \mu$ $g = 5$ $N = 40$

Test : $F_0 = [(SSE(R) - SSE(F)) / SSE(F)] * [(N - g) / (g - 1)]$ $N = n*g$

Modèle complet (F)

Y = 15.4* indA + 18.5 * indB + 15.0 * indC +9.7 * indD + 12.3* indE

SSE(F) = 73,8 avec N - g = 40 - 5 = 35 degrés de liberté (ddl)

Modèle réduit (R)

Y = 14.21 SSE(R) = 3,31*39 = 359,2 avec 39 degrés de liberté

Test F_0 = (359,2 - 73,8) / 73,8 * (35 / 4) = 27,07

F_1= F(0,99 ; 5, 35) = 3,59 $99^{ième}$ perc. dist. F avec (5 et 35) ddl

$F_0 > F_1$ rejet de H_0

conclusion : groupes sont de moyennes inégales

Transformations de la variable de réponse : cas de variances inégales

CONDITION	TRANSFORMATION
réponse Y est un comptage : distribution Poisson	$Y' = \sqrt{Y}$ ou $Y' = \sqrt{Y} + \sqrt{Y+1}$
réponse Y est une proportion : distribution binomiale	$Y' = 2 \arcsin(\sqrt{Y})$
s = écart type m = moyenne si (s^2 / m) quasi constant	$Y' = \sqrt{Y}$
s = écart type m = moyenne si (s / m) quasi constant	$Y' = \log(Y)$
s = écart type m = moyenne si (s / m^2) quasi constant	$Y' = 1 / Y$

Recommandation 1

examen quantités $s_i^2 / \overline{Y}_{i.}$ $s_i / \overline{Y}_{i.}$ $s_i / \overline{Y}_{i.}^2$
pour chaque niveau du facteur et choisir la transformation dont le coefficient de variation (CV) est le plus petit

Recommandation 2

transformation Box-Cox sur Y

$Y' = Y^{\lambda}$ $-2 < \lambda < 2$

choix de λ choisir λ tel que SSE(λ) minimum
arrondir la valeur : exemple -1,5 -1,0 -0,5 0 0,5 …
avec $\lambda = 0$ $Y' = \log(Y)$ cas fréquents

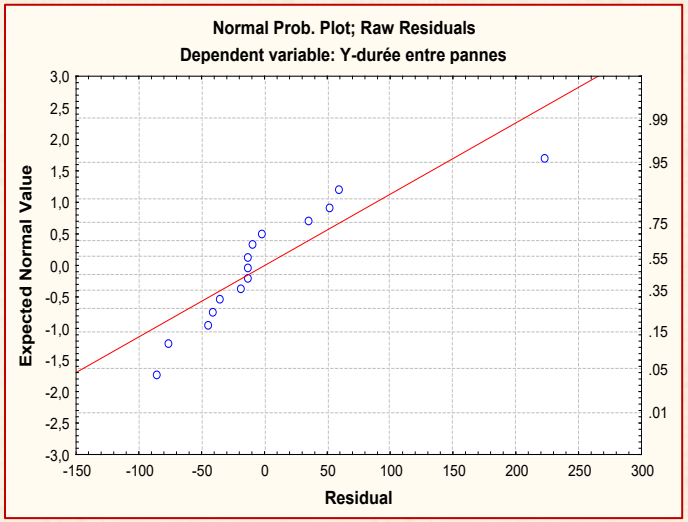
Exemple : temps panne

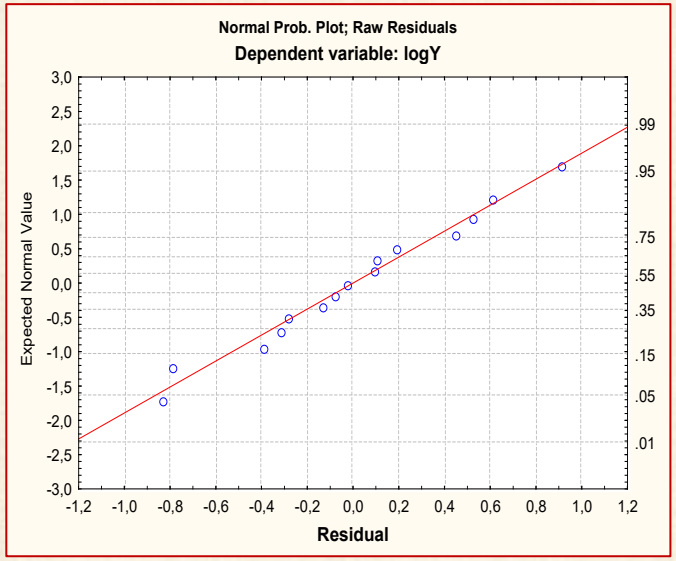
1 ID	2 ville	3 intervalle	4 Y_durée entre pannes	5 Y_Rang	6 logY
1	A	1	4,41	2	0,644
2	A	2	100,65	13	2,003
3	A	3	14,45	6	1,160
4	A	4	47,13	9	1,673
5	A	5	85,21	12	1,930
6	B	1	8,24	4	0,916
7	B	2	81,16	11	1,909
8	B	3	7,35	3	0,866
9	B	4	12,29	5	1,090
10	B	5	1,61	1	0,207
11	C	1	106,19	14	2,026
12	C	2	33,83	7	1,529
13	C	3	78,88	10	1,897
14	C	4	342,81	15	2,535
15	C	5	44,33	8	1,647

ville	obs	m moyenne	s écart type	s^2 / m	s / m	s / m^2
A	5	50.37	42.29	35.51	0.84	0.017
B	5	22.13	33.22	49.86	1.50	0.068
C	5	121.21	~~127.15~~	133.39	1.05	0.009
			moyenne	72,92	1,13	0,031
			std dev	52,854	0,338	0,032
			CV(%)	72	30	103

choix : s / m car CV min transformation logarithmique
confirmation : transformation Box-Cox

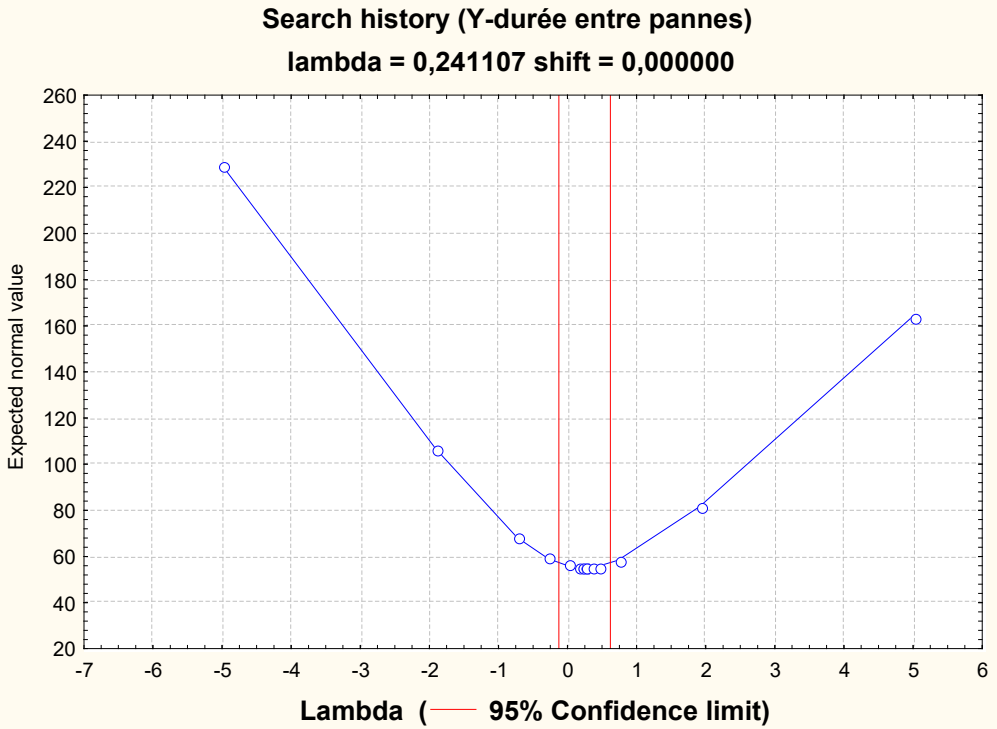
4 Y_durée entre pannes	5 Y_Rang
1,61	1
4,41	2
7,35	3
8,24	4
12,29	5
14,45	6
33,83	7
44,33	8
47,13	9
78,88	10
81,16	11
85,21	12
100,65	13
106,19	14
342,81	15





Transformation de Box-Cox

Exemple : temps panne



$\lambda = 0,24 \qquad -0,125 < \lambda < 0,618$

choix $\lambda = 0$ $\qquad Y' = \log(Y)$

transformation logarithmique

Les écarts d'hypothèses de base sur le modèles statistiques sont-ils importants ?

non normalité?

Réponse de
John Sall, Bradley Jones (2005).
Leptokurtosiphobia = peur irrationnelle de la non-normalité
Six Sigma Forum magazine, vol 4, no 3, May 2005
Vol 15 for web.fm (jmp.com)

1. **Modèles à effets fixés : le manque de normalité n'est pas très important**

   tester la normalité des résidus est une étape non nécessaire car
   - pour de « grands échantillons » la non normalité est facile à détecter mais elle est sans conséquence
   - pour de « petits échantillons », la non normalité pourrait avoir des conséquences, mais elle quasiment impossible à détecter : aucun test est suffisamment puissant.

2. **Modèles à effets aléatoires : les conséquences sont plus importantes**

Les écarts d'hypothèses de base sur les modèles statistiques sont-ils importants ?

2. Le **test F est robuste** si les tailles n_i ne sont pas trop inégales.
3. **Indépendance** : conséquences importantes pour l'inférence.
   Une forte auto corrélation dans les valeurs de la réponse Y a comme conséquence pratique que les tailles échantillonnales sont plus faibles en réalité qu'elles le paraissent, rendant ainsi plus difficile la détection des différences significatives.
4. Les mesures répétées sur une même unité d'observation constituent un **cas fréquent de dépendance.**
   Important de savoir reconnaître cette situation lorsqu'elle est présente dans la structure des données et de faire une analyse appropriée.
   Cette méthode est vue plus loin.

ANOVA non paramétrique : test de Kruskall-Wallis

Les méthodes non paramétriques sont basées sur les rangs de la variable de réponse plutôt que les valeurs observées.

Assignation aux observations Y_{ij} le rang R_{ij} des valeurs ordonnées en ordre croissant de 1 à N.

On procède comme dans le test F usuel que l'on applique aux rangs R_{ij}.

1 ID	2 ville	3 intervalle	4 Y_durée entre pannes	5 Y_Rang	6 logY
1	A	1	4,41	2	0,644
2	A	2	100,65	13	2,003
3	A	3	14,45	6	1,160
4	A	4	47,13	9	1,673
5	A	5	85,21	12	1,930
6	B	1	8,24	4	0,916
7	B	2	81,16	11	1,909
8	B	3	7,35	3	0,866
9	B	4	12,29	5	1,090
10	B	5	1,61	1	0,207
11	C	1	106,19	14	2,026
12	C	2	33,83	7	1,529
13	C	3	78,88	10	1,897
14	C	4	342,81	15	2,535
15	C	5	44,33	8	1,647

1 ID	2 ville	3 interv alle	4 Y_durée entre pannes	5 Y_Rang	6 logY
10	B	5	1,61	1	0,207
1	A	1	4,41	2	0,644
8	B	3	7,35	3	0,866
6	B	1	8,24	4	0,916
9	B	4	12,29	5	1,090
3	A	3	14,45	6	1,160
12	C	2	33,83	7	1,529
15	C	5	44,33	8	1,647
4	A	4	47,13	9	1,673
13	C	3	78,88	10	1,897
7	B	2	81,16	11	1,909
5	A	5	85,21	12	1,930
2	A	2	100,65	13	2,003
11	C	1	106,19	14	2,026
14	C	4	342,81	15	2,535

Test de Kruskall-Wallis

$$F_{KW} = MSTR / MSE$$

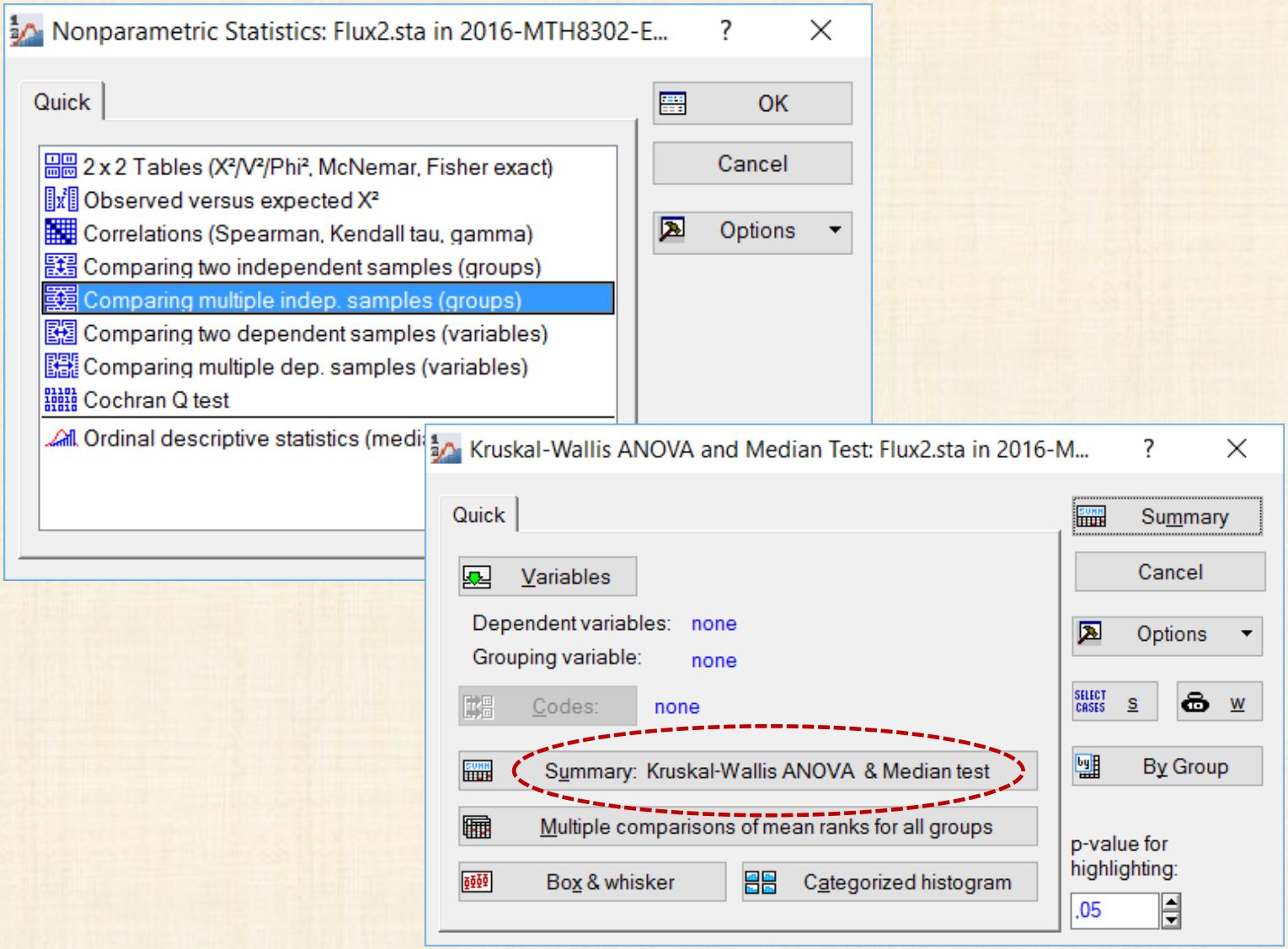
$$MSTR = \sum n_i (\overline{R}_{i.} - \overline{R}_{..})^2 / (g - 1)$$

$$MSE = \sum\sum (R_{ij} - \overline{R}_{i.})^2 / (N - g)$$

$$\overline{R}_{i.} = \sum R_{ij} / n_i$$

$$\overline{R}_{..} = \sum\sum R_{ij} / N = (N + 1) / 2$$

ANOVA non paramétrique : test de Kruskall-Wallis



ANOVA non paramétrique : test de Kruskall-Wallis + test Médiane
variable réponse : Y_durée entre pannes

Median Test, Overall Median = 44,3300;
Y_durée entre pannes
Independent (grouping) variable: ville
Chi-Square = 2,142857 df = 2 p = 0,3425

Dependent: Y_durée entre pannes	A	B	C	Total
<= Median: observed	2,000000	4,00000	2,000000	8,00000
expected	2,666667	2,66667	2,666667	
obs.-exp.	-0,666667	1,33333	-0,666667	
> Median: observed	3,000000	1,00000	3,000000	7,00000
expected	2,333333	2,33333	2,333333	
obs.-exp.	0,666667	-1,33333	0,666667	
Total: observed	5,000000	5,00000	5,000000	15,00000

Kruskal-Wallis ANOVA by Ranks;
Y_durée entre pannes
Independent (grouping) variable: ville
Kruskal-Wallis test: H (2, N= 15) =4,560000 p =,1023

Depend.: Y_durée entre pannes	Code	Valid N	Sum of Ranks	Mean Rank
A	1	5	42,00000	8,40000
B	2	5	24,00000	4,80000
C	3	5	54,00000	10,80000

pas de différence entre les villes

Analyse paramétrique de variance basée sur Y_rang : test ratio F

	SS	DF	MS	F	p
Intercept	960,00	1	960,00	61,02	0,0000
ville	91,20	2	45,60	2,90	0,0940
Error	188,80	12	15,73		

pas de différence entre les villes

introduction aux facteurs aléatoires : exemples
modèles mixtes : facteurs fixés + facteurs aléatoires

Régression : Y_log_price vs X_latitude town = facteur ignoré

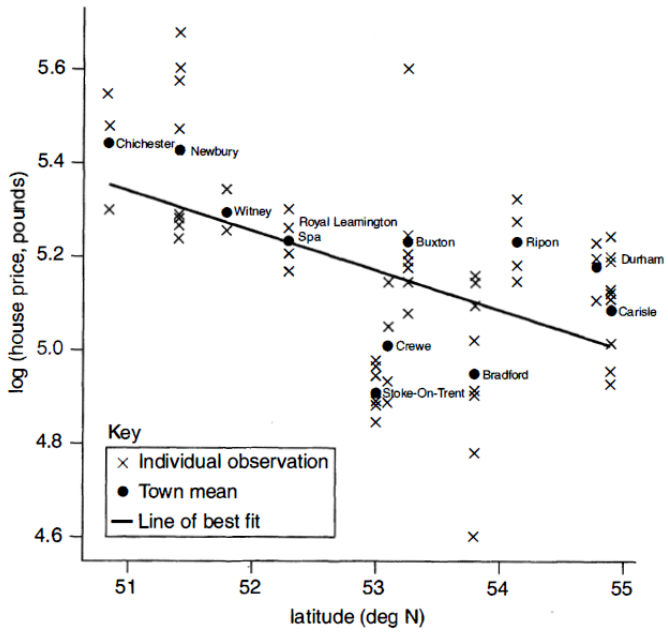
data

Source : N. W Galwey (2014) Introduction to Mixed Modeling, 2ed, Wiley
Y_Price (pounds) de 65 maisons en Angleterre, échantillon de 11 villes

1 ID	2 town	3 latitude	4 Y_price	5 Y_log_price	6 c5	7 town2	8 n_houses	9 mean-latitude	10 Y_log_mean_price
1	Bradford	53,7947	39950	4,6015		Bradford	9	53,7947	4,94745
2	Bradford	53,7947	59950	4,7778		Buxton	8	53,2591	5,23083
3	Bradford	53,7947	79950	4,9028		Carlisle	11	54,8923	5,08552
4	Bradford	53,7947	79995	4,9031		Chichester	3	50,8377	5,44034
5	Bradford	53,7947	79995	4,9031		Crewe	4	53,0998	5,00394
6	Bradford	53,7947	82500	4,9165		Durham	3	54,7762	5,17775
7	Bradford	53,7947	105000	5,0212		Newbury	8	51,4037	5,42456
8	Bradford	53,7947	125000	5,0969		Ripon	4	54,1356	5,23106
9	Bradford	53,7947	139950	5,1460		Leamington	4	52,2876	5,23337
10	Bradford	53,7947	145000	5,1614		Stoke	7	53,0041	4,90642
11	Buxton	53,2591	120000	5,0792		Witney	3	51,7871	5,29207
12	Buxton	53,2591	139950	5,1460					
13	Buxton	53,2591	149950	5,1759					
14	Buxton	53,2591	154950	5,1902					
15	Buxton	53,2591	159950	5,2040					

. . . . . .

51	Ripon	54,1356	210000	5,3222	
52	Leamington	52,2876	147000	5,1673	
53	Leamington	52,2876	159950	5,2040	
54	Leamington	52,2876	182500	5,2613	
55	Leamington	52,2876	199950	5,3009	
56	Stoke	53,0041	69950	4,8448	
57	Stoke	53,0041	69950	4,8448	
58	Stoke	53,0041	75950	4,8805	
59	Stoke	53,0041	77500	4,8893	
60	Stoke	53,0041	87950	4,9442	
61	Stoke	53,0041	92000	4,9638	
62	Stoke	53,0041	94950	4,9775	
63	Witney	51,7871	179950	5,2552	
64	Witney	51,7871	189950	5,2786	
65	Witney	51,7871	220000	5,3424	



Estimates of parameters

Parameter	Estimate	s.e.	t(62)	t pr.
Constant	9.68	1.00	9.68	<0.001
Latitude	−0.0852	0.0188	−4.53	<0.001

Regression analysis

Response variate: logprice
Fitted terms: Constant, latitude

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	0.710	0.70955	20.55	<0.001
Residual	62	2.141	0.03453		
Total	63	2.850	0.04524		

Percentage variance accounted for 23.7.
Standard error of observations is estimated to be 0.186.

introduction aux facteurs aléatoires : exemples
modèles mixtes : facteurs fixés + facteurs aléatoires

Régression : Y_log_price vs X_latitude + X_town town = facteur fixé

data

7 town2	8 n_houses	9 mean-latitude	10 Y_log_mean_price
Bradford	9	53,7947	4,94745
Buxton	8	53,2591	5,23083
Carlisle	11	54,8923	5,08552
Chichester	3	50,8377	5,44034
Crewe	4	53,0998	5,00394
Durham	3	54,7762	5,17775
Newbury	8	51,4037	5,42456
Ripon	4	54,1356	5,23106
Leamington	4	52,2876	5,23337
Stoke	7	53,0041	4,90642
Witney	3	51,7871	5,29207

Regression analysis

Response variate: logprice
Fitted terms: Constant + latitude + town

Estimates of parameters

Parameter	Estimate	s.e.	*t*(53)	*t* pr.
Constant	14.18	2.32	6.12	<0.001
latitude	−0.1717	0.0435	−3.95	<0.001
town Buxton	0.1914	0.0598	3.20	0.002
town Carlisle	0.3265	0.0885	3.69	<0.001
town Chichester	−0.015	0.136	−0.11	0.914
town Crewe	−0.0628	0.0761	−0.83	0.413
town Durham	0.399	0.106	3.75	<0.001
town Newbury	0.067	0.102	0.66	0.515
town Ripon	0.3421	0.0840	4.07	<0.001
town Royal Leamington Spa	0.0272	0.0874	0.31	0.757
town Stoke-On-Trent	−0.1767	0.0636	−2.78	0.007
town Witney	0	*	*	*

Parameters for factors are differences compared with the reference level:

Factor: Reference level
Town: Bradford

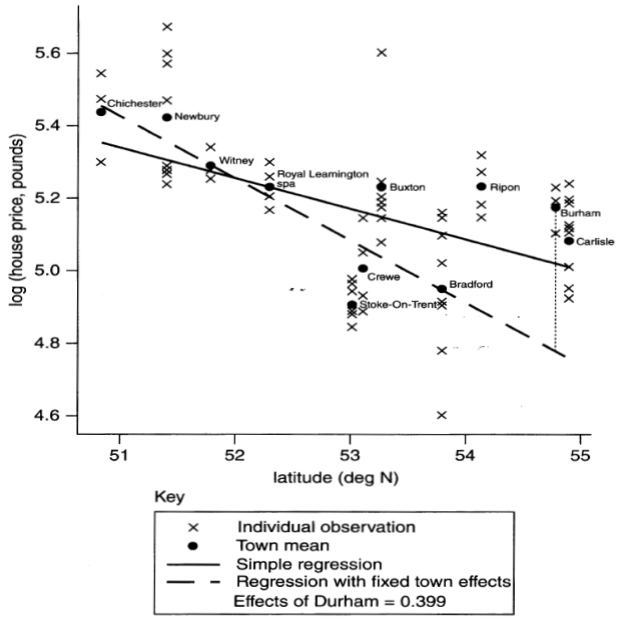


Figure 1.5 Relationship between latitude and house prices in a sample of English towns, comparing the lines of best fit from simple regression analysis and from analysis with town effects treated as fixed.

Regression analysis

Response variate: meanlogprice
Fitted terms: Constant, meanlatitude

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	*F* pr.
Regression	1	0.1160	0.11601	5.19	0.049
Residual	9	0.2010	0.02234		
Total	10	0.3170	0.03170		

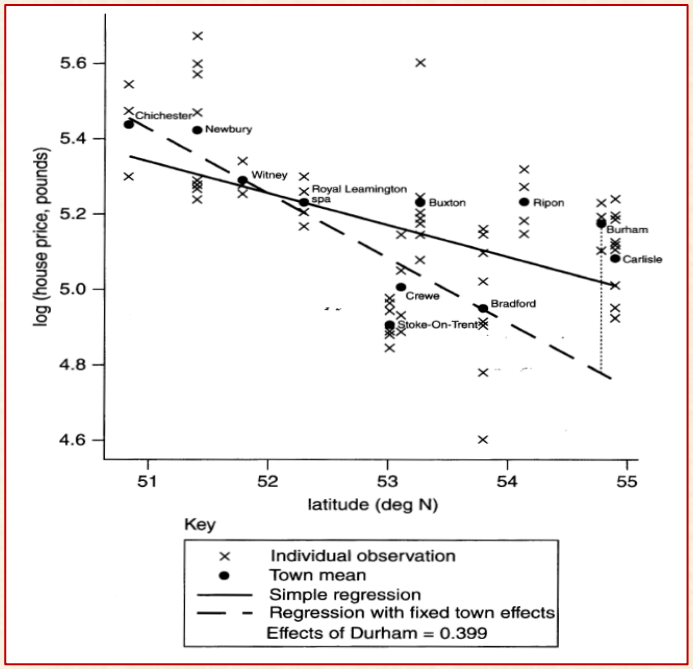
Percentage variance accounted for 29.5.
Standard error of observations is estimated to be 0.149.

Estimates of parameters

Parameter	Estimate	s.e.	*t*(9)	*t* pr.
Constant	9.44	1.87	5.04	<0.001
Meanlatitude	−0.0804	0.0353	−2.28	0.049

facteurs aléatoires
modèles mixtes : facteurs fixés + facteurs aléatoires

Régression : Y_log_price vs X_latitude + X_town **town = facteur aléatoire**



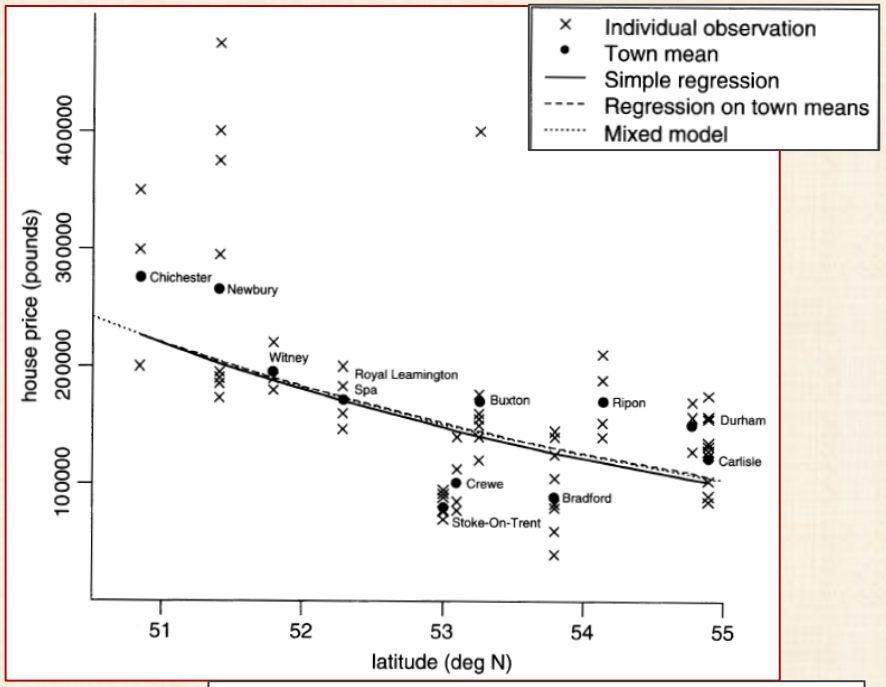


Figure 1.6 Relationship between latitude and house prices in a sample of English towns, comparing the lines of best fit from simple regression analysis, analysis on town means and mixed-model analysis.

Table 1.6 Comparison of the parameter estimates and their SEs, obtained from different methods of analysis of the effect of latitude on house prices in England.

Term	Method of analysis					
	Regression analysis ignoring towns (Model 1.1)		Regression analysis on town means		Mixed model analysis (Model 1.3)	
	Estimate	$SE_{Estimate}$	Estimate	$SE_{Estimate}$	Estimate	$SE_{Estimate}$
Constant	9.68	1.00	9.44	1.87	9.497	1.9248
latitude	−0.0852	0.0188	−0.0804	0.0353	−0.08147	0.036272