180 Anova I aléatoire

180.1 Principe

On cherche à déterminer si le taux de graisse dans le foie de rat est influencé par l'absorption d'alcool. L'expérimentateur a la possibilité de prendre plusieurs rats (réplicat[1] ou unités du premier degré d'échantillonnage, variance σ^2A) et d'effectuer plusieurs mesures par foie (duplicat[2] ou unité du second degré, variance σ^2).

Rat 1	Rat i	Rat 3	Rat 4
-	-	-	-
-	-	-	-
-	$x_{(i)j}$	-	-
-	-	-	-
-	-	-	-
-	-	-	-
	Mx_i		Mx

Tableau 180-1 Représentation schématique d'un tableau de résultat.

Le plan d'expérience correspond à une analyse à un critère de classification aléatoire
(ANOVA I Aléatoire)

La variance de la mesure dépend de deux sources de variabilité qui s'additionnent :

$$\sigma_x^2 = \sigma_A^2 + \sigma^2$$

Par exemple, on sélectionne dans une population de rats, dont la teneur en graisse du foie est une v.a. N de moyenne 120 et d'écart -type 20, un individu dont le taux de graisse moyen dans le foie est 130
Puis, sur le foie homogénéisé, on effectue une mesure dans une population de moyenne 130 et d'écart -type 10 (cette population représente toutes les mesures -presque une infinité - que l'on pourrait réaliser sur le foie homogénéisé). On obtient pour une mesure particulière la valeur 143.

---

[1] *Un réplicat est une expérience entièrement recommencée (répliquée).*

[2] *Un duplicat est une mesure réalisée deux fois dans les mêmes circonstances (on dit aussi triplicat (3X), quadruplicat (4X)…).*

Cette valeur correspond au modèle : $X_{(i)j} = \mu + A_i + E_{(i)j}$ avec

$A_{(i)}$ v.a. N (0 ;400) et $E_{(i)j}$ v.a. N (0 ;100)

dans ce cas particulier, μ = 120, Ai = 10 et $E_{(i)j}$ = 13 : $X_{(i)j}$ = 120 +10 +13 = 143.

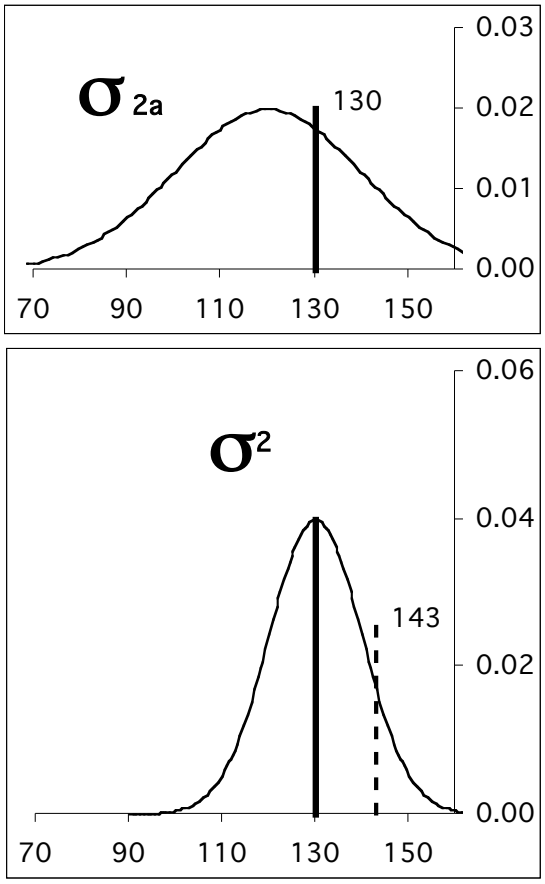

Figure 180-1 Echantillonnage à deux niveaux.

180.2 Echantillonnage optimal

En notant *p* le nombre de rats et *n* le nombre de mesures sur le foie de rat, l'application du théorème central limite donne les variances suivantes :

$$\sigma^2_{M_{xi}} = \sigma^2_A + \sigma^2 / n$$

$$\sigma^2_{Mx} = \frac{\sigma^2_A + \sigma^2 / n}{nA}$$

$$\sigma^2_{Mx} = \frac{\sigma^2_A}{n_A} + \frac{\sigma^2}{N}$$

Ceci permet de mettre en évidence que pour N mesures (N = n.na) l'échantillonnage idéal (variance de Mx minimale) correspond à n = 1 et na = N, soit un seul réplicat par rat et N rats. En effet, cela donne :

$$\sigma^2_{Mx} = \frac{\sigma^2_A + \sigma^2/1}{N}$$

$$\sigma^2_{Mx} = \frac{\sigma^2_A}{N} + \frac{\sigma^2}{N}$$

Toutes les autres valeurs de na < N produisent une valeur de variance supérieure.

> *Toutefois, il n'est guère réaliste expérimentalement de prétendre ne faire qu'une seule mesure par animal. Tout protocole requiert au minimum des duplicats, ne fut -ce que pour détecter des erreurs grossières... Les triplicats permettent non seulement de détecter une erreur, mais ils identifient aussi la mesure inexacte.*
>
> *Par ailleurs, il peut être plus intéressant de consentir un nombre total d'observations globalement plus grand, mais réalisées sur un plus petit nombre d'animaux, pour des raisons éthiques, et parfois simplement économiques.*

En considérant le rapport des coûts (moral, temps, matériel, réactifs, animalerie) de chaque type d'observations
C_A/C
et le rapport des variances
σ^2/σ^2_A

on peut montrer que le nombre de réplicats optimal n≠1 est déterminé par l'équation :

$$n \approx \sqrt{\frac{CA \;\; \sigma^2}{C \;\; \sigma^2_A}}$$

expression qui suppose que σ^2_A ne soit pas nul.

180.3 Equation

Soit trois individus et quatre mesures.

	Individu 1	Individu 2	Individu 3
	68	55	84
	54	65	70
	39	54	75
	59	86	71
Mx_i	55	65	75

SCE	442.0	662.0	122.0	Total
SCER				1226
d.l.	3	3	3	9
CMR	SCER/9	136.2		Moy
S^2i	147.3	220.7	40.7	136.2

Tableau 180-2 Calcul de SCER et CMR dans un tableau reprenant un échantillonnage à 2 niveaux : n_a = 3 et n = 4. On remarquera que CMR = SCER/dl est aussi la moyenne des variances des 3 échantillons de même taille n.

	Individu 1	Individu 2	Individu 3	
	55	65	75	
	55	65	75	
	55	65	75	
	55	65	75	
Mx		65		
d.l.		2		
SCE	400	0	400	
SCEF				800
CMF		800/2		400

Tableau 180-3 Calcul de SCEF et CMF dans un tableau reprenant un échantillonnage à 2 niveaux : a = 3 et n = 4. Pour Visualiser le principe du calcul, les valeurs individuelles ont été remplacées par la moyenne de l'échantillon correspondant.

Individu 1	Individu 2	Individu 3
68	55	84
54	65	70
39	54	75
59	86	71

Mx	65
SCET	2026

Tableau 180-4 Calcul de SCET dans un tableau reprenant un échantillonnage à 2 niveaux : a = 3 et n = 4.. On remarquera que SCET = SCEF +SCER (2026 = 1226 +800).

On peut définir les sommes de carrés d'écarts qui respectent les équations suivantes :
$SCE_T = SCE_F + SCE_R$
d.l.T = d.l.F +d.l.R

d.l. = degrés de liberté

Dans notre exemple :

2026 = 1226 +800
11 = 2 + 9

A partir des SCE, on définit des carrés moyens qui sont une généralisation de la notion de variance.

$$CM = \frac{SCE}{d.l.}$$

$$CMR = \frac{SCER}{n_A(n-1)}$$ **Équation 180-1**

n étant constant dans les différents groupes, on constate que CMR = 136.2 représente bien la variance moyenne :

(147.3 +220.7 +40.7)/4 = 136.2

Le carré moyen résiduel mesure la variance entre les duplicats.

sous réserve d'homogénéité des variances $\sigma^2 i = \sigma^2$

E(CMR) = σ^2

$$CMF = \frac{SCEF}{n_A - 1.}$$ **Équation 180-2**

Le carré moyen factoriel mesure la variance entre les moyennes Mx_i.

sous réserve d'homogénéité des variances $\sigma^2_i = \sigma^2$

$E(CMF) = \sigma^2 + n\,\sigma^2_A$

180.4 Test d'hypothèse sur σ^2

Exemple : Un biochimiste étudie le taux de catalase dans le foie de rat et obtient les résultats suivants

	Rat 1	Rat 2	Rat 3	Rat 4
	100	55	64	76
	96	61	70	44
	80	63	72	42
	88	81	62	72
	92	43	76	74
	86	47	82	60
Moyennes	90.33	58.33	71.00	61.33
Moyenne générale		70.25		
Variances	51.87	183.47	55.60	233.07

	S.CE.	D.L.	C.M.	F
totale	6372.50			
factoriel	3752.50	3.00	1250.83	
résiduel	2620.00	20.00	131.00	9.55

Tableau 180-5 Etude du taux de catalase (6 réplicats sur 4 rats). Moyennes et table d'ANOVA I.

Ho : homogénéité des variances σ^2_i

$$H_{5;4} = \frac{233.07}{51.87} = 4.49$$

$$H_{5;4;0.95} = 13.7$$

Il y a acceptation de Ho, au risque d'erreur β inconnu.

180.5 Test d'hypothèse sur σ^2_a

Sous Ho = $\sigma^2 A = 0$, $n\,\sigma^2 A = 0$
on constate donc que

$E(CMF) = \sigma^2 + n\,\sigma^2_A = \sigma^2 = E(CMR)$

Sous Ho, le rapport CMF/CMR est une v.a. F avec n_A -1 d.l. au numérateur et n_A (n -1) d.l. au dénominateur.

On peut donc tester si $\sigma^2 A$ est > 0 par un test de comparaison de deux variances.
Ho : $\sigma^2_A = o$

$$F_{3;20} = \frac{1250.83}{131} = 9.55$$

$$F_{3;20;0.95} = 3.1$$

Il y a rejet de Ho. On choisit donc H_1 : $\sigma^2_A > 0$ au risque d'erreur α 5% (soit avec une confiance de 95%).

Puisque $\sigma^2 A$ n'est pas nulle, on peut donc l'estimer :
$E(CMF) = \sigma^2 + n\,\sigma^2_A$
$E(CMF) = E(CMR) + n\,\sigma^2_A$
$E((CMF - CMR)/n) = \sigma^2_A = 186.64$

180.6 Nombre d'unités du 2° niveau

On peut dès lors estimer le nombre de mesures optimal, en utilisant l'équation

$$n \approx \sqrt{\frac{C_A\ \sigma^2}{C\ \sigma^2_A}}$$ **Équation 180-3**

En supposant un coût de 100 euros pour le rat et 10 euros pour le test

$$n \approx \sqrt{\frac{100 \times 131}{10 \times\ 186.6}} = 2.64$$

soit 2.64 < 3 ce qui signifie des triplicats pour chaque rat.

180.7 Nombre d'unités du 1° niveau

En fixant la plus petite différence sensée Δ_{min}, il est possible d'utiliser l'équation 6.6 pour déterminer le nombre de rats nécessaires à l'expérience.

$$n_a \approx \frac{16\sigma^2_{Mxi}}{\Delta^2}$$

Par exemple, pour mettre en évidence une différence d'au moins 10 unités de catalase, on estime σ^2_{Mx} :

$$\sigma^2_{M_{xi}} = \sigma^2_A + {\sigma^2}/{n}$$

$$S^2_{M_{xi}} = 186.64 + {131}/{3} = 230.31$$

et ensuite on peut estimer n_a :

$$n_a \approx \frac{16 \times 230.31}{10^2} = 36.85$$

soit 37 rats.

Si l'on prétend utiliser moins de rats, il faut réévaluer leur coût.

Soit le coût du rat = 1000 euros :

$$n \approx \sqrt{\frac{1000 \times 131}{10 \times \ 186.6}} = 8.38 \approx 9$$

$$S^2_{M_{xi}} = 186.64 + {131}/{9} = 201.2$$

$$n_a \approx \frac{16 \times 201.2}{10^2} = 32.19 \approx 32$$

On remarquera que dans le cas présent, la variance d'un animal à l'autre étant grande, on ne peut pas réduire substantiellement le nombre d'animaux.

180.8 Intervalle de confiance

$$\sigma^2_{Mx} = \frac{\sigma^2_A + {\sigma^2}/{n}}{n_A}$$

$$= \frac{n\sigma^2_A + \sigma^2}{na}$$

$E(CMF) = \sigma^2 + n\,\sigma^2_A$

La variance de la moyenne générale M_x
elle est donc estimée par $CMF/n_A n$ ou encore CMF/N
ce qui représente, approximativement, pour une confiance de 95%

$$\varepsilon \approx 2\sqrt{\frac{CMF}{N}}$$

Dans notre exemple, la moyenne générale est 70.25.

$$\varepsilon \approx 2\sqrt{\frac{1250.83}{24}} = 14.44$$

soit une moyenne générale du taux de catalase du foie de rat qui a 95% de chances environ d'être comprise entre 55.81 et 84.69

Une estimation plus précise de l'intervalle de confiance est obtenue par référence à la variable t de Student :

$$\varepsilon \approx t_{n_a - 1;1-\alpha/2}\sqrt{\frac{CMF}{N}}$$ **Équation 180-4**

avec, dans notre exemple, $n_a - 1 = 3$ et $t_{3;0.975} = 3.18$

$$\varepsilon \approx 3.18\sqrt{\frac{1250.83}{24}} = 22.96$$

soit une moyenne générale du taux de catalase du foie de rat qui a 95% de chances environ d'être comprise entre 47.29 et 93.21.

> *On notera que lorsque a est petit la référence à la variable t est nécessaire. Le nombre de degrés de liberté de t étant uniquement lié au nombre d'unités du premier degré, l'intervalle de confiance ne sera guère plus précis en multipliant le nombre de réplicats.*

Attention, si je commets l'erreur de ne pas considérer le critère de classification aléatoire "Rat", et j'applique la formule de l'intervalle de confiance pour 24 observations indépendantes

$$\varepsilon \approx t_{n-1;1-\alpha/2}\sqrt{\frac{S_x^2}{n}} \quad = \quad \varepsilon \approx t_{23;0.975}\sqrt{\frac{277}{24}}$$