

Statistiques pour petits échantillons

Applications en recherche pré-clinique

Christophe Lalanne

ch.lalanne@mac.com

www.aliquote.org

Plan de l'exposé

rappels sur les principaux tests paramétriques pour la comparaison de moyennes et de proportions • présentation des approches non paramétriques • tests de permutation • plans d'expérience

TL;DR

Lorsque les conditions d'application sont vérifiées, le test de Student reste le plus puissant. On peut cependant envisager des méthodes alternatives pour la construction d'un test d'hypothèse nulle. Le modèle de régression linéaire est un outil versatile pour l'estimation ponctuelle et par intervalle. On peut néanmoins profiter des plans d'expérience pour gagner de la puissance statistique et de méthodes robuste d'estimation pour limiter l'influence de certaines observations.

Rappels sur la démarche inférentielle

Dicing with Death : Chance, Risk and Health

Statisticians are applied philosophers. Philosophers argue how many angels can dance on the head of a needle ; statisticians count them. Or rather, count how many can probably dance. (...) We can predict nothing with certainty but we can predict how uncertain our predictions will be, on average that is. Statistics is the science that tells us how.

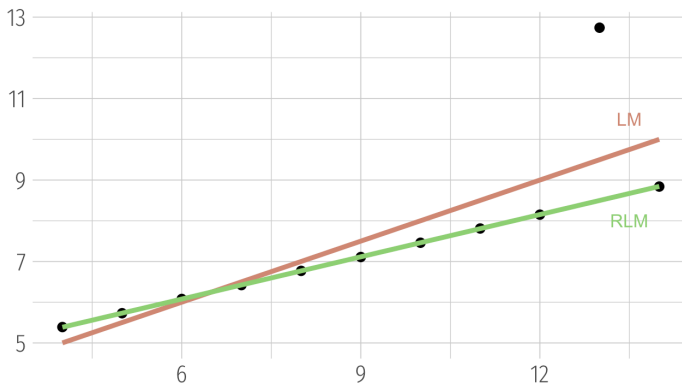
– Stephen Senn (2003)

Mots-clés :

- approche (fréquentiste) du test d'une hypothèse nulle, approche bayésienne, approche reposant sur la vraisemblance
- méthodes paramétriques et non paramétriques, bootstrap, procédures de re randomization, approches robustes, validation croisée
- dessin de l'étude, efficacité d'un estimateur et puissance statistique : structure d'appariement, variables d'ajustement et variables de confusion, réduction de dimension

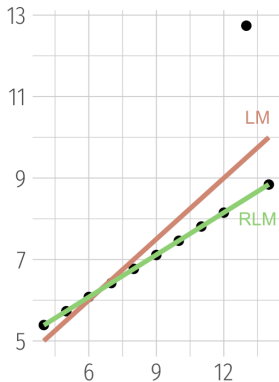
Les petits échantillons sont plus susceptibles de présenter des valeurs extrêmes (observations influentes), ils rendent difficile la mise en évidence de "petites différences", et il est plus difficile de vérifier les conditions de validité des tests statistiques usuels.

Moindres carrés ordinaires *versus* M-estimateur



Méthode des moindres carrés (LM) et approche robuste¹ (RLM)

Linéarité de la relation et points influents



Taille d'effet et nombre de sujets nécessaires

TABLE I—*Sample sizes required per group at the two sided 5% significance level for different values of d and power (d=expected mean difference/standard deviation)*

d	Power (1- β)				
	99	95	90	80	50
0.10	3676	2600	2103	1571	770
0.20	920	651	527	394	194
0.30	410	290	235	176	87
0.40	231	164	133	100	49
0.50	148	105	86	64	32
0.60	104	74	60	45	23
0.70	76	54	44	33	17
0.80	59	42	34	26	13
0.90	47	34	27	21	11
1.00	38	27	22	17	9
1.10	32	23	19	14	8
1.20	27	20	16	12	7
1.30	23	17	14	11	6
1.40	20	15	12	9	5
1.50	18	13	11	8	5

Tiré de Campbell, Julious, & Altman (1995)

Estimation et test d'hypothèse

Cas de la comparaison de deux moyennes

Test de Student

Pour deux échantillons indépendants, la statistique de test se définit ainsi :

$$t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \quad s_c = \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \right)^{1/2},$$

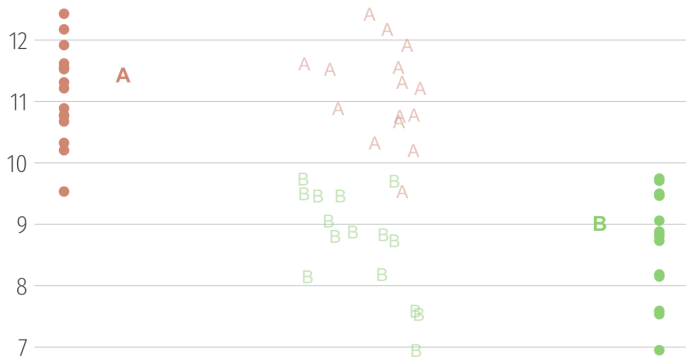
où les \bar{x}_i et n_i sont les moyennes et effectifs des deux échantillons, et s_c est la variance commune pour la différence de moyennes d'intérêt. Sous H_0 , cette statistique de test suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté.

Un intervalle de confiance à $100(1 - \alpha)\%$ pour la différence $\bar{x}_1 - \bar{x}_2$ peut être construit comme suit :

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha, n_1 + n_2 - 2} s_c \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

avec $P(t < t_{\alpha, n_1 + n_2 - 2}) = 1 - \alpha/2$.

Illustration



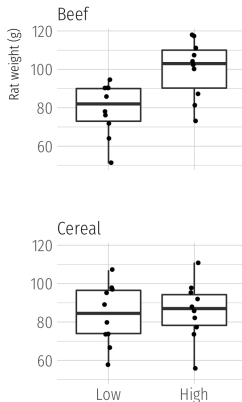
Deux échantillons ($n = 15$)

Application numérique

Experimental study of the gain in weight of rats fed on four different diets, defined by amount (low / high) and by source of protein (beef and cereal)

— Hand, Daly, McConway, & Ostrowski (1993)

Beef	Cereal
90 / 73	107 / 98
76 / 102	95 / 74
90 / 118	97 / 56
64 / 104	80 / 111
86 / 81	98 / 95
51 / 107	74 / 88
72 / 100	74 / 82
90 / 87	67 / 77
95 / 117	89 / 86
78 / 111	58 / 92



<R/>

```
d <- read.table("weight.dat")  
## comparaison Beef/Low vs. Beef/High  
t.test(d[,1], d[,2], var.equal = TRUE)
```

Résultat du test de Student

Two Sample t-test

```
data:  d[, 1] and d[, 2]  
t = -3.2021, df = 18, p-value = 0.00494  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
  -34.447192  -7.152808  
sample estimates:  
mean of x mean of y  
    79.2    100.0
```

Condition d'homoscédasticité

L'hypothèse de variances égales² peut être relaxée en utilisant une approximation des degrés de liberté sous la forme d'une simple combinaison linéaire des variances, $\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}$ (Satterthwaite, 1946). La statistique de test suit toujours une loi de Student, mais avec des degrés de liberté (ν) différents.

Formule de Satterthwaite et de Welch

L'approximation de Satterthwaite considère les degrés de liberté suivants :

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

La formule proposée par Welch (1947) (`var.equal = FALSE`, par défaut sous R) est à peu près comparable :

$$\nu = -2 + \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}}.$$

Cas des moyennes tronquées

Méthode proposée par Yuen (1974) : utilisation d'une statistique t basée sur des moyennes tronquées (typiquement, 20 %), qui se réduit à la statistique de Welch en l'absence de trimming.

Le test de Yuen

La statistique de test se construit comme dans le cas du test de Student, $t_{\text{obs}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{d_1 + d_2}}$, avec

$$d_j = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)}.$$

Ici, h_j représente le nombre d'observations dans chaque groupe après trimming, et s_{wj}^2 la variance winsorisée du groupe j . Les degrés de liberté associés à cette statistique t_{obs} sont

$$\frac{(d_1 + d_2)^2}{\frac{d_1^2}{h_1 - 1} + \frac{d_2^2}{h_2 - 1}}.$$

1. Approche non paramétrique
2. Test de nullité du paramètre par permutation
3. Estimation par intervalle par bootstrap
4. Test de nullité du paramètre par bootstrap
5. *Test de nullité basé sur un estimateur robuste* (Wilcox & Keselman, 2003)
6. *Estimation bayésienne d'un intervalle de crédibilité* (Kruschke, 2013)

Le test des rangs de (Mann-Whitney-)Wilcoxon

L'hypothèse nulle est que les deux échantillons comparés proviennent de deux distributions ayant les mêmes distributions (paramètre de position, i.e. médiane).

La statistique de test est construite comme la plus petite somme des rangs d'un des deux échantillons ; quand $n_1, n_2 > 15$ et qu'il n'y a pas d'ex-aequo, on a l'approximation suivante pour la statistique de test :

$$Z = \frac{S - n_1(n_1 + n_2 + 1)/2}{\sqrt{n_1 n_2 (n_1 + n_2 + 1)/12}} \sim \mathcal{N}(0; 1),$$

où S est la statistique de test pour l'échantillon avec n_1 observations.

Pour deux échantillons appariés, le test des rangs signés est utilisé : on calcule la somme T^+ des rangs des différences $z_i = x_{1i} - x_{2i}$ en valeurs absolues positifs ; dans le cadre asymptotique, la statistique de test correspondante est :

$$Z = \frac{T^+ - n(n + 1)/4}{\sqrt{n(n + 1)(2n + 1)/24}} \sim \mathcal{N}(0; 1).$$

Test de permutation bilatéral

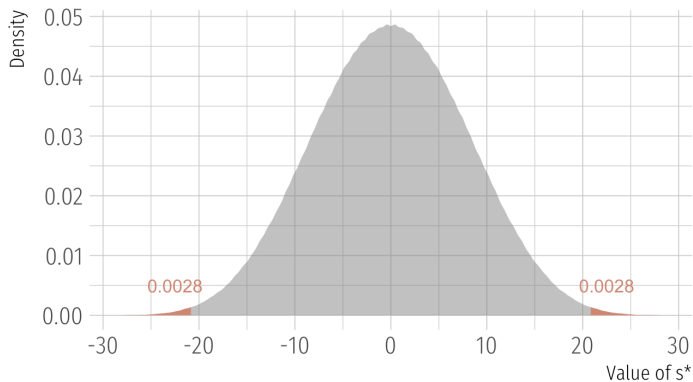
Dans la plupart des cas, la distribution de la statistique de test est symétrique (et $\mathbb{E}(T) = 0$ sous H_0). Soit $s_0 = \bar{x}_1 - \bar{x}_2$ la statistique empirique. On calculera le degré de signification en cumulant les résultats des permutations $s = \bar{x}_1^* - \bar{x}_2^*$ dans les deux queues de distribution.

```
1 s0 <- mean(d[,1]) - mean(d[,2])           ## statistique emp.
2 x <- c(d[,1], d[,2])
3 idx <- combn(seq(along = 1:20), 10)       ## {20 \choose 10}
4 f <- function(k) mean(x[k]) - mean(x[-k])
5 s <- apply(idx, 2, f)                     ## statistique s
6 pobs <- sum(abs(s) >= abs(s0)) / length(s)
```

Avec le package coin :

```
library(coin)
oneway_test(value ~ variable, data = reshape2::melt(d[1:2]),
             alternative = "two.sided", distribution = "exact")
```

Illustration



Distribution de la statistique sous permutation

Approche par bootstrap

Conditions de validité :

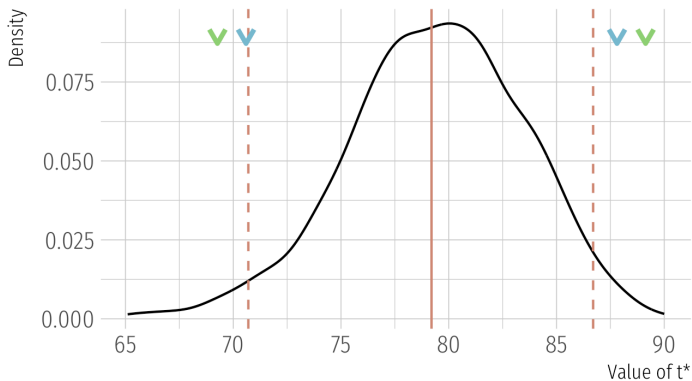
(a) le paramètre à estimer n'est pas dans un cas limite (borne de l'espace du paramètre, cas des statistiques d'ordre ou des proportions extrêmes) ; (b) l'échantillon peut être considéré représentatif de la population cible ; (c) la taille de l'échantillon est raisonnable.

1. Tirage avec remise de n échantillons de même taille n .³
2. Pour chaque échantillon, on calcule la statistique d'intérêt, $\hat{\theta}_i^*$.
3. Le biais par rapport à l'estimateur classique est $\hat{\theta} - \theta \approx \mathbb{E}(\hat{\theta}^*) - \hat{\theta}$.
4. Pour chaque échantillon bootstrap, on calcule $t_i^* = \frac{\hat{\theta}_i^* - \hat{\theta}}{se(\hat{\theta}_i^*)}$

($se(\hat{\theta}_i^*) = s_i^* \sqrt{n}^{-1}$), et la distribution bootstrap de t^* permet de construire un intervalle de confiance à 95 %.

Cette approche peut également être utilisée pour construire un test d'hypothèse nulle.

Illustration : intervalle de confiance pour une moyenne



IC à 95 % (bootstrap percentile) pour la moyenne Beef/Low (Student, Normal)

Illustration : test d'hypothèse pour une différence de moyenne

Dans le cas à deux échantillons, on rééchantillonne séparément les deux groupes (centrés sur leur moyennes respectives), on construit la statistique de test d'intérêt (ici, $s = \bar{x}_1^* - \bar{x}_2^*$) et on calcule le degré de signification $\#\{|s| > |\bar{x}_1 - \bar{x}_2|\}$ (sous H_0) :

```
1 x1 <- d[,1] - mean(d[,1]) + mean(x)
2 x2 <- d[,2] - mean(d[,2]) + mean(x)
3 B <- 10000          ## no. bootstrap samples
4 s <- numeric(B)    ## vector of test statistics
5 for (i in 1:B) {
6   x1s <- sample(x1, replace=TRUE)
7   x2s <- sample(x2, replace=TRUE)
8   s[i] <- mean(x1s) - mean(x2s)
9 }
10 pobs <- (1 + sum(abs(s) > abs(s0))) / (B+1)
```

Modèle linéaire et applications

Rappels sur la régression linéaire

Le modèle de régression simple

Soit y_i la réponse observée sur l'individu i , et x_i sa valeur observée pour le prédicteur x . Le modèle de régression linéaire s'écrit :

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

où β_0 représente l'ordonnée à l'origine et β_1 la pente de la droite de régression, et $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ est un terme d'erreur (résidus, supposés indépendants).

En minimisant les différences quadratiques entre les valeurs observées et les valeurs prédites (principe des MCO), on peut estimer les coefficients de régression, $\hat{\beta}_0$ et $\hat{\beta}_1$:

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \sum (y_i - \bar{y})(x_i - \bar{x}) / \sum (x_i - \bar{x})^2\end{aligned}$$

Sous H_0 , le rapport entre l'estimé de la pente ($\hat{\beta}_1$, de variance $\frac{SSR/(n-2)}{(n-1)s_x^2}$) et son erreur standard suit une loi de Student à $(n - 2)$ degrés de liberté.

Mesures d'influence

Les résidus “simples”, $e_i = y_i - \hat{y}_i$, sont de variance non constante, mais on montre que celle-ci est fonction de σ_ε^2 :

$$\text{Var}(e_i) = \sigma_\varepsilon^2(1 - h_i),$$

où h_i (“*hat value*”) est l'effet levier de la i ème observation sur l'ensemble des valeurs ajustées via le modèle de régression. Les points exerçant un fort effet levier tendent donc à avoir de plus faibles résidus.

- effet levier : large si $h_i > 2(k + 1)/n$
- distance de Cook : $D_i = \frac{h_i}{1-h_i} \times \frac{e_i'^2}{k+1}$, considérée large si $D_i > 4/(n - k - 1)$.
- DFFITS = $\frac{\hat{y}_i - \hat{y}_i^{(-i)}}{s_{e(-i)}} \sqrt{h_i^{(-i)}}$, considéré large si $> 2\sqrt{p/n}$.
- DFBETAS = $\hat{\beta}_j - \hat{\beta}_j^{(-i)} / s_{e(-i)} \sqrt{X'X_{jj}}$, pour le j ème coefficient du modèle (incluant l'intercept), considéré large si $> 2/\sqrt{n}$.

Approche robuste et résistante de la régression

1. Méthode robuste :

Estimer un modèle de régression classique, puis calculer les distances de Cook et exclure les observations pour lesquelles $D > 1$. Ensuite, de manière itérative on estime le modèle et on calcule le poids de chaque observation à partir des résidus. En utilisant la fonction de pondération proposée par Huber (1964), les observations avec des résidus élevés auront des poids de plus en plus petits. Il existe d'autres variantes de M-estimateurs.

2. Méthode résistante :

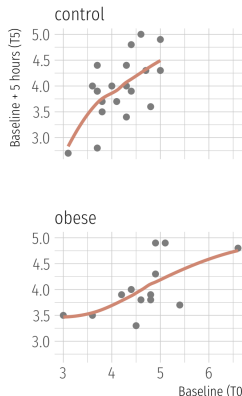
Utilisation de la médiane au lieu des MCO, $\min_b \text{median}_i \|y_i - x_i b\|^2$, ou d'une moyenne tronquée, $\min_b \sum_{i=1}^q \|y_i - x_i b\|_{(i)}^2$ (généralement, $q = \lfloor (n + p + 1)/2 \rfloor$) (Rousseeuw & Leroy, 1987).

Illustration numérique

Plasma inorganic phosphate levels from 33 subjects in two groups (20 control, 13 obese), taken every 30' or 60' from baseline (t0).

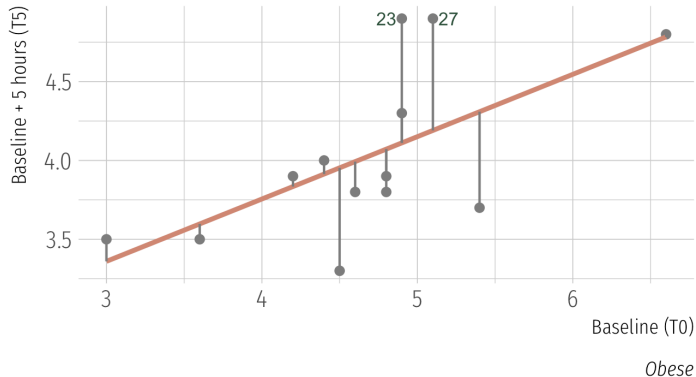
— Davis (2002)

ID	group	t0	t1	t4	t5
1	control	4.3	3.0	3.4	4.4
2	control	3.7	2.6	3.1	3.9
3	control	4.0	3.1	3.9	4.0
4	control	3.6	2.2	3.8	4.0
5	control	4.1	2.1	3.6	3.7
...					
29	obese	4.2	3.8	3.5	3.9
30	obese	6.6	5.2	4.2	4.8
31	obese	3.6	3.1	2.5	3.5
32	obese	4.5	3.7	3.1	3.3
33	obese	4.6	3.8	3.8	3.8



<R/>

```
tmp <- read.csv("phosphate.csv")  
d <- subset(tmp, group == "obese")
```



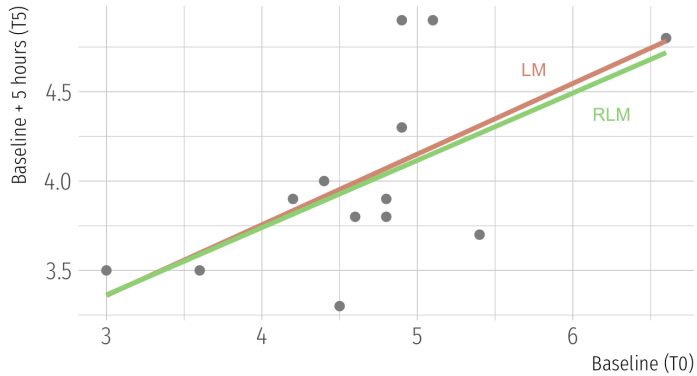
```

1 m <- lm(t5 ~ t0, data = d)
2 cooks.distance(m)
3 rob <- MASS::rlm(t5 ~ t0, data = d) ## Huber M-estimator
4 rob$w[c(3,7)]                      ## weights for obs. 23/27

```

Mesures d'influence 'influence.measure(m)'

	dfb.1_	dfb.t0	dffit	cov.r	cook.d	hat	inf
21	0.32600	-0.4144	-0.6286	0.895	0.17372	0.1360	
22	0.30425	-0.2843	0.3168	1.940	0.05437	0.3949	*
23	-0.05780	0.1686	0.6459	0.616	0.15678	0.0826	
24	-0.00301	-0.0168	-0.1138	1.274	0.00702	0.0786	
25	0.02863	-0.0191	0.0599	1.313	0.00197	0.0856	
26	-0.01154	0.0337	0.1289	1.271	0.00898	0.0826	
27	-0.18166	0.2811	0.6159	0.733	0.15426	0.0972	
...							



Analyse de variance et régression linéaire

Le modèle à effets pour un modèle à deux facteurs A et B s'écrit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk},$$

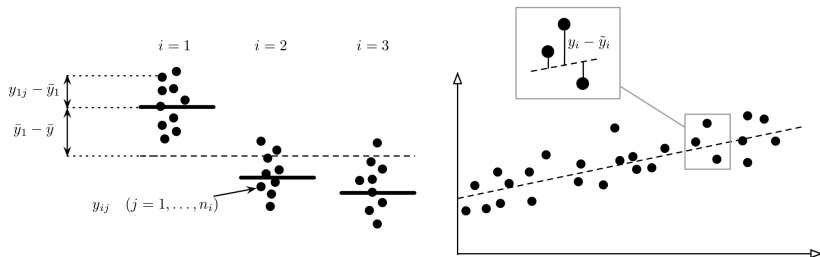
où y_{ijk} désigne la mesure sur le k^{e} sujet pour le niveau i de A et j de B , μ est la moyenne générale et les α_i et β_j désignent les effets principaux de A et B . Si le dessin expérimental est équilibré, la variance totale peut être partitionnée en sommes de carré indépendantes.

Un modèle de régression, en considérant des variables indicatrices codant les niveaux de A et B , se formule ainsi :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

où β_0 représente l'intercept et β_1 et β_2 les coefficients de régression associés à chaque facteur.

Données = modèle + erreur



- Part systématique (fonctionnelle, structurelle) du modèle, $\mathbb{E}(y \mid x) = f(x; \beta)$: choix de l'estimateur, linéarité de la relation (e.g., restricted cubic spline), régularisation ou shrinkage, effet bloc, ajustement
- Part aléatoire ou “bruit” : variance des erreurs, corrélation sérielle


```
1 tmp <- read.table("weight.dat")
2 d <- data.frame(weight = as.numeric(unlist(tmp)),
3               type = gl(2, 20, labels = c("Beef", "Cereal")),
4               level = gl(2, 10, labels = c("Low", "High")))
5 fm <- weight ~ type + level + type:level
6 summary(aov(fm, data = d))
```

Tableau d'analyse de variance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	221	220.9	0.988	0.3269
level	1	1300	1299.6	5.812	0.0211 *
type:level	1	884	883.6	3.952	0.0545 .
Residuals	36	8049	223.6		

```
1 m <- lm(fm, data = d)
2 summary(m)
```

Tableau des coefficients de régression

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	79.200	4.729	16.749	< 2e-16	***
typeCereal	4.700	6.687	0.703	0.48668	
levelHigh	20.800	6.687	3.110	0.00364	**
typeCereal:levelHigh	-18.800	9.457	-1.988	0.05447	.

L'analyse de covariance consiste à tester différents niveaux d'un facteur en présence d'un ou plusieurs co-facteurs continus. La variable réponse et ces co-facteurs sont supposées reliés, et l'objectif est d'obtenir une estimation des réponses corrigée pour les éventuelles différences entre groupes (au niveau des cofacteurs).

Ce type d'analyse est fréquemment utilisé comme méthode d'ajustement *a posteriori*, ou dans le cas des données pré/post avec des mesures continues et un facteur de groupe, et reste préférable à une simple analyse des scores de différences (Miller & Chapman, 2001 ; Senn, 2006).

Il existe également des alternatives non-paramétriques (Young & Bowman, 1995).

Équation de la droite de régression

Soit y_{ij} la j ème observation dans le groupe i . À l'image du modèle d'ANOVA à un facteur, le modèle d'ANCOVA s'écrit :

$$y_{ij} = \mu + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij},$$

où β est le coefficient de régression liant la réponse y et le cofacteur x (continu), avec \bar{x} la moyenne générale des x_{ij} , et toujours un terme d'erreur $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$.

Notons que l'on fait l'hypothèse que β est le même dans chaque groupe. Cette hypothèse de parallélisme peut se vérifier en testant la significativité du terme d'interaction $\alpha\beta$.

La réponse moyenne ajustée pour l'effet du co-facteur numérique s'obtient simplement comme $\bar{\alpha}_i + \hat{\beta}(\bar{x}_i - \bar{x})$, où \bar{x}_i est la moyenne des x dans le i ème groupe.

Cas des mesures corrélées

Les données corrélées ou hiérarchiques s'observent dans le cas du regroupement d'unités plus ou moins similaires (souris d'une même lignée, effet cage), de corrélation intra unité (mesures répétées répétées dans le temps ou un mélange des deux (Gelman & Hill, 2006 ; McCullagh & Searle, 2001).

Deux approches : modèles marginaux, où l'on modélise des effets moyens de population en faisant l'hypothèse d'une matrice de corrélation intra unité dite "de travail", *versus* modèles conditionnels, où l'on modélise des effets individuels (cas des modèles à effets mixtes). Cette dernière approche est généralement plus souple que l'approche par ANOVA à mesures répétées (ou la MANOVA) en cas de valeurs manquantes ou pour rendre compte de structure de corrélation intra unité particulière.

```
t.test(extra ~ group, data = sleep)
t.test(extra ~ group, data = sleep, paired = TRUE)
```

Ignorer la structure de corrélation résulte généralement en des tests moins puissants car on sait que :

$$\text{Var}(X_1 - X_2) = \text{Var}(X_1) + \text{Var}(X_2) - 2\text{Cov}(X_1, X_2),$$

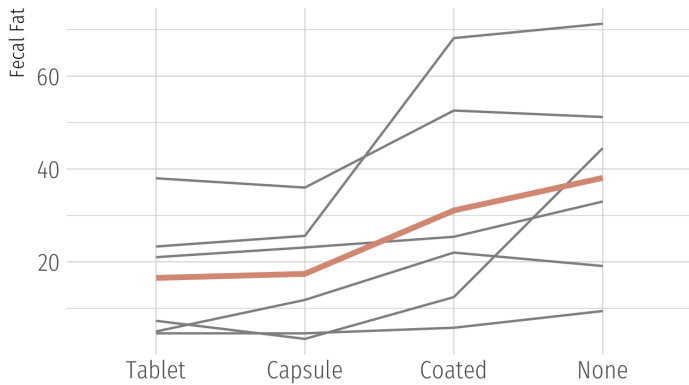
ou de manière équivalente, $\text{Cov}(X_1, X_2) = \rho\sqrt{\text{Var}(X_1)}\sqrt{\text{Var}(X_2)}$. Si on assume $\text{Cov}(X_1, X_2) = 0$, cela revient à sur-estimer la variance des différences de moyenne puisque $\text{Cov}(X_1, X_2)$ est généralement > 0 .

Application numérique

Lack of digestive enzymes in the intestine can cause bowel absorption problems, as indicated by excess fat in the feces. Pancreatic enzyme supplements can be given to ameliorate the problem

— Vittinghoff, Glidden, Shiboski, & McCulloch (2005)

ID	None	Tablet	Capsule	Coated	Avg.
1	44.5	7.3	3.4	12.4	16.9
2	33.0	21.0	23.1	25.4	25.6
3	19.1	5.0	11.8	22.0	14.5
4	9.4	4.6	4.6	5.8	6.1
5	71.3	23.3	25.6	68.2	47.1
6	51.2	38.0	36.0	52.6	44.5
Avg.	38.1	16.5	17.4	31.1	25.8



Modèles d'ANOVA

```
1 aov(fecfat ~ pilltype, data = fat) ## M1
2 aov(fecfat ~ pilltype + subject, data = fat) ## M2
3 aov(fecfat ~ pilltype + Error(subject), data = fat) ## M3
```

Source	DF	SS	MS	M1	M2*/M3
pilltype	3	2009	669.5	669.5/359.7 p=0.169	669.5/107.0 p=0.006
subject	5	5588	1117.7	—	1117.7/107.0 p<0.001*
Residuals	15	1605	107.0	—	—

Le modèle M1 (ANOVA à un facteur à effet fixe) suppose les observations indépendantes et ne tient pas compte des 78 % de variance entre sujets.

Les modèles M2 (ANOVA à deux facteurs à effet fixe) et M3 (ANOVA à deux facteurs dont un à effet aléatoire) incorporent un effet sujet :

$$y_{ij} = \mu + \text{subject}_i + \text{pilltype}_j + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2).$$

Dans M3, on suppose de plus que $\text{subject}_i \sim \mathcal{N}(0, \sigma_s^2)$, indépendants de ε_{ij} . L'inclusion d'un effet aléatoire permet de modéliser la corrélation intra sujet.

Composantes de variance

Corrélation intraclasse et symétrie composée

La corrélation entre deux mesures prises sur le même sujet vaut :

$$\text{Cor}(y_{ij}, y_{ik}) = \frac{\text{Cov}(y_{ij}, y_{ik})}{\text{Var}(y_{ij})^{1/2} \text{Var}(y_{ik})^{1/2}}.$$

Comme μ et pilltype sont fixés, et $\varepsilon_{ij} \perp \text{subject}_i$, on a alors :

$$\begin{aligned}\text{Cov}(y_{ij}, y_{ik}) &= \text{Cov}(\text{subject}_i, \text{subject}_i) \\ &= \text{Var}(\text{subject}_i) = \sigma_s^2\end{aligned}$$

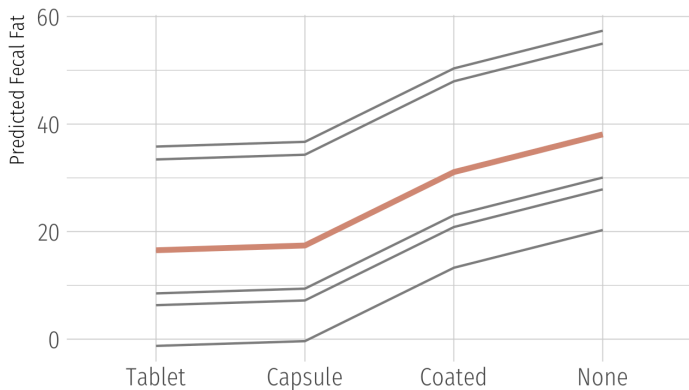
Les composantes de variance découlent de l'égalité $\text{Var}(y_{ij}) = \text{Var}(\text{subject}_i) + \text{Var}(\varepsilon_{ij}) = \sigma_s^2 + \sigma_\varepsilon^2$, supposée valide pour toutes les observations. On a donc : $\text{Cor}(y_{ij}, y_{ik}) = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\varepsilon^2}$, qui représente la proportion de variance attribuable aux sujets, et qu'on appelle également **corrélation intraclasse** (ρ).

Illustration : estimation de ρ

```
1 library(nlme)
2 m <- lme(fecfat ~ pilltype, data = fat,
3         random = ~ 1 | subject)           ## random-intercept
4 anova(lme.fit)                           ## ANOVA table
5 intervals(lme.fit)                       ## 95% CIs
6 sigma.s <- as.numeric(VarCorr(lme.fit)[1,2]) ## sigma_s
7 sigma.eps <- as.numeric(VarCorr(lme.fit)[2,2]) ## sigma_e
8 sigma.s^2 / (sigma.s^2 + sigma.eps^2)     ## rho = 0.7025
```

L'estimation des paramètres du modèle est réalisée par la méthode REML : les effets aléatoires sont estimés après avoir rendu compte des effets fixes (pilltype).

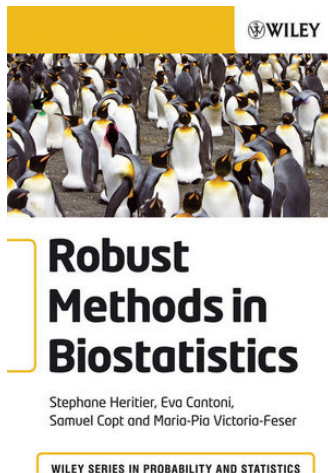
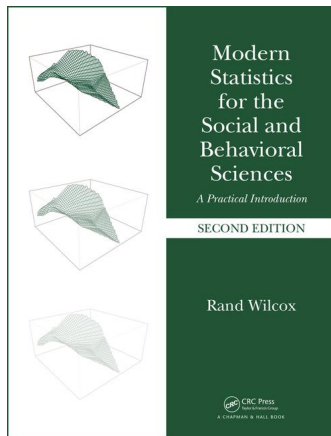
Modèle à intercept aléatoire et shrinkage



Prédictions du modèle à intercept aléatoire (symétrie composée)

Up

Pour aller plus loin



Références I

- Campbell, M., Julious, S., & Altman, D. (1995). Estimating Sample Sizes for Binary, Ordered Categorical, and Continuous Outcomes in Two Group Comparisons. *BMJ*, 311, 1145–1148. Repéré à <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2551061/>
- Davis, C. S. (2002). *Statistical Methods for the Analysis of Repeated Measurements*. New York : Springer.
- Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel Hierarchical Models*. (S.I.) : Cambridge University Press.
- Hand, D. J., Daly, F., McConway, K., & Ostrowski, E. (Éds). (1993). *A Handbook of Small Data Sets*. New York : Chapman & Hall/CRC Press. Repéré à <http://www.stat.ncsu.edu/research/sas/sicl/data/>
- Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics*, 35, 73–101.
- Kruschke, J. K. (2013). Bayesian Estimation Supersedes the T Test. *Journal of Experimental Psychology : General*, 142(2), 572–603.
- McCullagh, P., & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. (S.I.) : Wiley.
- Miller, G. A., & Chapman, J. P. (2001). Misunderstanding Analysis of Covariance. *Journal of Abnormal Psychology*, 110(1), 40–48.

Références II

- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. New York : Wiley.
- Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, 2, 110—114.
- Senn, S. (2006). Change From Baseline and Analysis of Covariance Revisited. *Statistics in Medicine*, 25(24), 4334—4344.
- Tufte, E. R. (1989). *The Visual Display of Quantitative Information*. Cheshire, CT, USA : Graphic Press.
- Vittinghoff, E., Glidden, D. V., Shiboski, S. C., & McCulloch, C. E. (2005). *Regression Methods in Biostatistics. Linear, Logistic, Survival, and Repeated Measures Models*. New-York : Springer.
- Welch, B. L. (1947). The Generalization of Student's Problem When Several Different Population Variances Are Involved. *Biometrika*, 34, 28—35.
- Wilcox, R. R., & Keselman, H. J. (2003). Modern Robust Data Analysis Methods : Measures of Central Tendency. *Psychological Methods*, 8(3), 254—274.
- Young, S. G., & Bowman, A. W. (1995). Nonparametric Analysis of Covariance. *Biometrics*, 51, 920—931.
- Yuen, K. K. (1974). The Two Sample Trimmed T for Unequal Population Variances. *Biometrika*, 61, 165—170.