Description et inférence - DESINF (HAB711B)

Université Montpellier

Table des matières

Chapitre 1

Les tests classiques

1.1 Définition des risques et des valeurs critiques

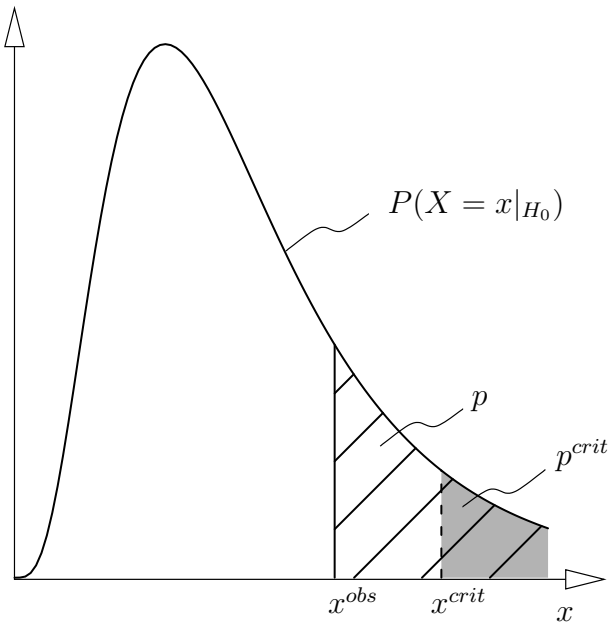
Par convention, on note en majuscule les variables aléatoires et en minuscule leurs réalisations qui ne sont pas aléatoires. Les quantités surmontées d'un « chapeau » sont des estimations de ces mêmes quantités caractérisant la population étudiée et obtenues à partir d'un échantillon. On note H_0 l'hypothèse nulle et H_1 l'hypothèse alternative. La notation $P(T \geqslant t_{obs}|_{H_0})$ veut donc dire « probabilité que la variable aléatoire T soit supérieure ou égale à la valeur numérique t_{obs} en supposant l'hypothèse nulle H_0 vraie ».

Pour toute variable aléatoire, il est possible de définir par une fonction, de telle sorte qu'on peut déterminer la probabilité qu'elle prenne une valeur donnée dans un intervalle ou un espace donné. Par exemple, une variable aléatoire peut être le gain obtenu à l'issue d'un jeu, et il est possible d'ajuster une fonction pour connaître la probabilité d'obtenir un gain donné en fonction de sa valeur. Ceci correspond à la fonction de densité de cette variable.

Soit X une statistique (ou variable aléatoire), et x^{obs} la valeur de cette statistique mesurée ou observée dans un échantillon. La figure ci-dessous représente la fonction de densité de probabilité f de la variable aléatoire X sous l'hypothèse nulle H_0. Pour chaque valeur x, on peut définir

$$P(X \geqslant x|_{H_0}) = \int_{z=x}^{\infty} f(z)dz,$$

ceci correspond la probabilité d'observer une valeur de X plus grande que x ; cette probabilité correspond à l'aire sous la courbe à droite du point x. On pourrait également représenter la fonction de répartition de X selon l'hypothèse nulle qui correspondrait à $F_X(x) = P(X \leq x)$,



La valeur p (ou « p-value » en anglais) est la probabilité d'observer une valeur de la statistique plus élevée que celle observée si jamais l'hypothèse nulle était vraie. On peut l'écrire de la façon suivante : $p = P(X \geqslant x^{obs}|_{H_0})$. Cette valeur corrspond à l'aire hachurée sous la courbe

Il est possible pour prendre une décision par rapport à un test de fixer une probabilité critique p^{crit} (par exemple 1%). A cette probabilité critique, on peut faire correspondre la valeur critique x^{crit} de X est telle que $P(X \geqslant x^{crit}|_{H_0}) = p^{crit}$. L'aire grisée sous la courbe correspond à p^{crit} et permet donc de positionner x^{crit}.

Le lien entre la valeur p et la valeur critique est établi par l'inégalité suivante $:x^{obs} > x^{crit} \Leftrightarrow p = P(X \geqslant x^{obs}|_{H_0}) < p^{crit}$

Par ailleurs, on peut, pour un test donné, estimer le risque de première espèce α. Ce risque est la probabilité de conclure, en observant une valeur de la statistique plus extrême que la valeur critique de rejeter l'hypothèse nulle alors que celle ci est vraie. Cette valeur est connue, et en général, le risque α correspond à la probabilité critique : c'est à dire que si on fixe une probabilité critique de 1%, on s'attend à ce que pour des valeurs observées x^{obs} plus extrêmes que x^{crit}, on rejette H_0 alors qu'elle est vraie dans 1% des cas.

Le risque de deuxième espèce β associé à un test est la probabilité de décider que l'hypothèse nulle est acceptable alors qu'elle est fausse. Pour l'estimer, il faudrait connaître et pouvoir modéliser la distribution de X sous l'hypothèse alternative H_1 ; ceci est très difficile en pratique. Comme la probabilité de se tromper dépend à la fois du risque de deuxième espèce et du risque de première espèce, il est rarement possible de pouvoir donner la probabilité de se tromper en faisant un choix ; tout au plus on peut estimer la probabilité de se tromper en rejettant H_O si celle ci était vraie

On rejette l'hypothèse nulle

- **si la valeur p est faible (typiquement en dessous d'une valeur critique ad-hoc à 5% ou 1%),**
- **si la statistique observée est plus extrême que la valeur critique.**

1.2 Comparaison d'une moyenne observée et à une moyenne théorique : test de Student

Problème : On mesure une variable X dans un échantillon de n individus. On cherche à comparer la moyenne de ces mesures à la valeur théorique m. Soient μ la moyenne de X et $\hat{\mu}$ son estimation dans l'échantillon.

Condition de validité : X suit une loi normale

Hypothèse nulle :

Test bilatéral : $\mu = m$

Test unilatéral : $\mu \leqslant m$ ou $\mu \geqslant m$ selon les données

Hypothèse alternative :

Test bilatéral : $\mu \neq m$

Test unilatéral : $\mu > m$ ou $\mu < m$ selon les données

Statistique : $t_{obs} = (\hat{\mu} - m)/\sqrt{\hat{\sigma}^2/n}$

Distribution de la statistique sous H_0 : T suit une distribution de Student à $(n-1)$ degrés de liberté

Risque de première espèce :

Test bilatéral : $p = P(|T| \geqslant |t_{obs}|\,|_{H_0})$

Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `t.test(x,mu=m)`

1.3 Comparaison de deux moyennes pour des échantillons indépendants

Problème : On mesure une même variable dans deux échantillons issus de deux populations et de tailles respectives n_1 et n_2. On cherche à comparer les moyennes entre les deux populations des mesures réalisées. Soient X_1 et X_2 les variables mesurées dans les populations 1 et 2 et de moyenne μ_1 et μ_2. Soient $\hat{\mu}_1$ et $\hat{\mu}_2$ les estimations de ces moyennes dans les deux échantillons.

1.3.1 Test paramétrique de Student

Conditions de validité : X_1 et X_2 sont indépendantes et suivent des lois normales de même variance (condition à vérifier par un test de comparaison de variance, c.f. page 9).

Hypothèse nulle :

Test bilatéral : $\mu_1 = \mu_2$

Test unilatéral : $\mu_1 \leqslant \mu_2$ ou $\mu_1 \geqslant \mu_2$ selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : $\mu_1 \neq \mu_2$

Test unilatéral : $\mu_1 > \mu_2$ ou $\mu_1 < \mu_2$ selon l'*a priori* sur les données

Statistique :

$$t_{obs} = (\hat{\mu}_1 - \hat{\mu}_2)/\sqrt{\hat{\sigma}^2(1/n_1 + 1/n_2)}, \quad \text{avec}$$
$$\hat{\sigma}^2 = \left((n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2\right)/(n_1 + n_2 - 2)$$

Distribution de la statistique sous H_0 : T suit une distribution de Student à $(n_1 + n_2 - 2)$ degrés de liberté.

Risque de première espèce :

Test bilatéral : $p = P(|T| \geqslant |t_{obs}|\,|_{H_0})$

Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `t.test(x,y, var.equal=T)`

1.3.2 Test paramétrique de Welch

Conditions de validité : X_1 et X_2 sont indépendantes et suivent des lois normales.

Hypothèse nulle :

Test bilatéral : $\mu_1 = \mu_2$

Test unilatéral : $\mu_1 \leqslant \mu_2$ ou $\mu_1 \geqslant \mu_2$ selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : $\mu_1 \neq \mu_2$

Test unilatéral : $\mu_1 > \mu_2$ ou $\mu_1 < \mu_2$ selon les données

Statistique : $t_{obs} = (\hat{\mu}_1 - \hat{\mu}_2)/\sqrt{(\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_2^2/n_2)}$

Distribution de la statistique sous H_0 : T suit une distribution de Student à $\left\{ (\hat{\sigma}_1^2/n_1) + (\hat{\sigma}_1^2/n_2)\right\}^2 / \left\{ (\hat{\sigma}_1^4/(n_1^2(n_1-1))) + (\hat{\sigma}_2^4/(n_2^2(n_2-1)))\right\}$ degrés de liberté.

Risque de première espèce :

Test bilatéral : $p = P(|T| \geqslant |t_{obs}, |_{H_0})$

Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `t.test(x,y)`

1.3.3 Test non-paramétrique de Wilcoxon-Mann-Whitney

Condition de validité : X_1 et X_2 sont indépendantes

Hypothèse nulle :

Test bilatéral : Les deux échantillons proviennent de populations qui suivent la même loi ($P(X_1 = x) = P(X_2 = x)$) si bien que $P(X_1 > X_2) = P(X_2 > X_1) = \frac{1}{2}$

Test unilatéral : Les deux échantillons provienne de populations de sorte que $P(X_1 > X_2) \geqslant P(X_2 > X_1)$ ou $P(X_1 > X_2) \leqslant P(X_2 > X_1)$ en fonction de l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : $P(X_1 > X_2) \neq P(X_2 > X_1) \neq \frac{1}{2}$

Test unilatéral : $P(X_1 > X_2) < P(X_2 > X_1)$ ou $P(X_1 > X_2) > P(X_2 > X_1)$

Statistique : $u_{obs} = \min(u_{1obs}, u_{2obs})$ ou $u_{obs} = \max(u_{1obs}, u_{2obs})$ selon la nature du test, avec $u_{1obs} = \sum_i R_{1i} - \frac{n_1(n_1+1)}{2}$, $\sum_i R_{1i}$ la somme des rangs pour les observations de l'échantillon issu de la population 1, et $u_{2obs} = n_1 n_2 - u_1$.

Pour calculer la statistique : on classe l'ensemble des $n_1 + n_2$ valeurs, et on affecte un rang (R_{1i} à chaque individu i de la population 1). En cas d'*ex aequo*, on affecte le même rang moyen à tous les rangs identiques.

Distribution de la statistique sous H_0 :

— On peut utiliser la véritable distribution de U à l'aide de tables conuues (cf. espace pédagogique).

— Si les échantillons sont grands ($n_1 > 20$ ou $n_2 > 20$), U suit approximativement une distribution normale de moyenne $\mu = \frac{1}{2}n_1n_2$ et de variance $\sigma^2 = \frac{1}{12}n_1n_2(n_1+n_2+1)$; ainsi $Z = (U-\mu)/\sigma$ suit une distribution normale centrée réduite.

Risque de première espèce :

— Si on utilise la distribution de U
Test bilatéral : $p = P(U \leqslant \min(u_{1obs}, u_{2obs})|_{H_0}) + P(U \geqslant \max(u_{1obs}, u_{2obs})|_{H_0})$
Test unilatéral : $p = P(U \leqslant \min(u_{1obs}, u_{2obs})|_{H_0})$ ou $P(U \geqslant \max(u_{1obs}, u_{2obs})|_{H_0})$

— Si on utilise l'approximation normale
Test bilatéral : $p = P(|Z| \geqslant |z_{obs}|\,|_{H_0})$
Test unilatéral : $p = P(Z \geqslant z_{obs}|_{H_0})$ ou $p = P(Z \leqslant z_{obs}|_{H_0})$

Code R : `wilcox.test(x,y)`

N.B. Le test de Wilcoxon-Mann-Whitney n'est pas, comme le montre la formulation de l'hypothèse nulle, un test de comparaison de moyennes mais un test d'identité de lois que suivent deux variables aléatoires. Toutefois dans de nombreux cas, particulièrement lorsque les variances des deux variables sont égales, l'hypothèse nulle n'est rejetée que si les espérances des deux distributions sont différentes. Probablement pour cette raison, ce test est utilisé en biologie comme une alternative non-paramétrique au test de Student. Il faut néanmoins être prudent car un risque de première espèce faible d'un test de Wilcoxon n'indique pas nécessairement une différence de moyennes ; il indique que l'un des deux échantillons comprend plus de valeurs élevées que l'autre.

1.4 Comparaison de deux moyennes pour des échantillons appariés

Problème : On mesure deux variables sur les mêmes individus i, pour i allant de 1 à n et on cherche à comparer les moyennes de ces deux mesures. Soient X_{1i} et X_{2i} les variables mesurées et $D_i = X_{1i} - X_{2i}$ leur différence. Soient μ_d la moyenne de D, $\hat{\mu}_d$ son estimation à partir des n individus, σ_d^2 la variance de D et $\hat{\sigma}_d^2$ son estimation à partir des n individus.

1.4.1 Test paramétrique de Student

Condition de validité : La différence $D_i = X_{1i} - X_{2i}$ suit une loi normale

Hypothèse nulle :
test bilatéral : $\mu_d = 0$
test unilatéral : $\mu_d \leqslant 0$ ou $\mu_d \geqslant 0$ selon les données

Hypothèse alternative :
test bilatéral : $\mu_d \neq 0$
test unilatéral : $\mu_d > 0$ ou $\mu_d < 0$ selon les données

Statistique : $t_{obs} = \hat{\mu}_d / \sqrt{\hat{\sigma}_d^2/n}$

Distribution de la statistique sous H_0 : T suit une distribution de Student à $(n-1)$ degré de liberté.

Risque de première espèce :
Test bilatéral : $p = P(|T| \geqslant |t_{obs}|\,|_{H_0})$
Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `t.test(X1,X2,paired=T)`

1.4.2 Test non-paramétrique de Wilcoxon

Hypothèse nulle :
Test bilatéral : les deux échantillons proviennent de populations qui suivent la même loi ($P(X_1 = k) = P(X_2 = k)$) si bien $P(X_1 > X_2) = P(X_2 > X_1) = \frac{1}{2}$
Test unilatéral : $P(X_1 > X_2) \geqslant P(X_2 > X_1)$ ou $P(X_1 > X_2) \leqslant P(X_2 > X_1)$ selon la nature du test

Hypothèse alternative :

Test bilatéral : $P(X_1 > X_2) \neq P(X_2 > X_1) \neq \frac{1}{2}$

Test unilatéral : $P(X_1 > X_2) < P(X_2 > X_1)$ ou $P(X_1 > X_2) > P(X_2 > X_1)$ selon la nature du test

Statistique : $u_{obs} = \min(u_{1obs}, u_{2obs})$ ou $u_{obs} = \max(u_{1obs}, u_{2obs})$ selon le test, avec u_1^{obs} la somme des rangs des différences appariées positives, u_2^{obs} la somme des rangs des différences appariées négatives.

Les différences $|d_i = X_{2i} - X_{1i}|$ sont alors classées et associées à un rang. Les différences nulles peuvent être éliminées mais alors la taille de l'échantillon diminue. Si les différences nulles sont conservées, le même rang moyen leur est attibué et dans ce cas $u_1^{obs} + u_2^{obs} = \frac{n(n+1)}{2}$.

Distribution de la statistique sous H_0 :

— On peut utiliser la véritable distribution de U

— Si l'échantillon est grand ($n > 54$), U suit approximativement une distribution normale de moyenne $\mu = \frac{1}{4}n(n+1)$ et de variance $\sigma^2 = \frac{1}{24}n(n+1)(2n+1)$; ainsi $z^{obs} = (u^{obs} - \mu)/\sigma$ suit une distribution normale centrée réduite.

Risque de première espèce :

— Si on utilise la vraie distribution de U

Test bilatéral : $p = P(U \leqslant \min(u_{1obs}, u_{2obs})|_{H_0}) + P(U \geqslant \max(u_{1obs}, u_{2obs})|_{H_0})$

Test unilatéral : $p = P(U \leqslant \min(u_{1obs}, u_{2obs})|_{H_0})$ ou $P(U \geqslant \max(u_{1obs}, u_{2obs})|_{H_0})$

— Si on utilise l'approximation normale

Test bilatéral : $p = P(|Z| \geqslant |z_{obs}|\,|_{H_0})$

Test unilatéral : Si H_1 est $P(X_1 > X_2) < P(X_2 > X_1)$, $p = P(Z \geqslant |z_{obs}||_{H_0})$, sinon $p = P(Z \leqslant -|z_{obs}||_{H_0})$

Code R : `wilcox.test(x,y,paired=TRUE)`

1.5 Comparaison de fréquences et de distributions

1.5.1 Comparaison de deux fréquences : test Z

Problème : On échantillonne n_1 individus dans une première population et n_2 individus dans une deuxième. Chaque individu appartient à une catégorie parmi deux possibles. On cherche à comparer la fréquence des individus d'une catégorie donnée entre populations. Soient f_1 et f_2 les fréquences des individus de la première catégorie étudiés dans les populations 1 et 2 et $\hat{f}_1$ et $\hat{f}_2$ leurs estimations dans les échantillons 1 et 2.

Conditions de validité : Les tailles des échantillons doivent être telles que $n_1 > 100$ et $n_2 > 100$ et les fréquences à comparer doivent satisfaire les inégalités suivantes : $0.1 \leqslant \hat{f}_1 \leqslant 0.9$ et $0.1 \leqslant \hat{f}_2 \leqslant 0.9$.

Hypothèse nulle :

Test bilatéral : $f_1 = f_2$

Test unilatéral : $f_1 \leqslant f_2$ ou $f_1 \geqslant f_2$ selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : $f_1 \neq f_2$

Test unilatéral : $f_1 > f_2$ ou $f_1 < f_2$ selon l'*a priori* sur les données

Statistique : $z_{obs} = (\hat{f}_1 - \hat{f}_2)/\sqrt{\hat{f}(1-\hat{f})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ avec $\hat{f} = \frac{n_1\hat{f}_1 + n_2\hat{f}_2}{n_1+n_2}$

Distribution de la statistique sous H_0 : Z suit une loi Normale centrée réduite

Risque de première espèce :

Test bilatéral : $p = P(|Z| \geqslant |z_{obs}|\,|_{H_0})$

Test unilatéral : $p = P(Z \geqslant z_{obs}|_{H_0})$ ou $p = P(Z \leqslant z_{obs}|_{H_0})$

Code R : `prop.test` ; l'utilisation de cette fonction dépasse le cadre de la comparaison de deux fréquences... Consultez l'aide de R pour utiliser la fonction.

1.5.2 Comparaison de a distributions : test du χ^2

Problème : On échantillonne un total de n individus appartenant à chacun à une populations i parmi a possibles. Chaque individu échantillonné appartient à une catégorie j parmi b possibles. On cherche à comparer la répartition des individus dans les b catégories entre les a populations.

Conditions de validité : les tailles d'échantillons doivent être telles que $n > 20$. Si $20 \leqslant n \leqslant 40$, l'échantillonnage doit suivre la règle de Cochran : au moins 5 observations doivent être faites dans chaque population et pour chaque catégorie ; si cette règle n'est pas respectée, il faut regrouper certaines catégories, et leur nombre de fait diminue.

Hypothèse nulle : La répartition des individus dans les b catégories est la même dans les a populations.

Hypothèse alternative : Il existe au moins une population dont les individus se répartissent d'une façon différente entre les b catégories.

Statistique :

$$\chi^2_{obs} = \sum_{i=1}^{a}\sum_{j=1}^{b} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}, \quad \text{avec}$$

n_{ij} l'effectif observé dans la population i et pour la catégorie j, $\hat{n}_{ij} = \dfrac{n_{i\cdot}.n_{\cdot j}}{n}$ l' effectif attendu sous l'hypothèse nulle dans la population i et pour la catégorie j. Celui ci est calculé de sorte que $n_{i\cdot}$ corresponde au nombre d'individus échantillonnés dans la population i toutes catégories confondues ($n_{i\cdot} = \sum_{j=1}^{b} n_{ij}$), $n_{\cdot j}$ au nombre d'individus échantillonnés dans la catégorie j toutes populations confondues ($n_{\cdot j} = \sum_{i=1}^{a} n_{ij}$), et n au nombre total d'individus échantillonnés toutes populations et toutes catégories confondues ($n = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij}$).

Distribution de la statistique sous H_0 : χ^2_{obs} suit une loi du χ^2 à ($\nu = (a-1)(b-1)$) degrés de liberté.

Risque de première espèce : $p = P(\chi^2 \geqslant \chi^2_{obs}|_{H_0})$

Code R : `chisq.test`

1.5.3 Comparaison d'une distribution observée à une distribution théorique : test du χ^2

Problème : On échantillonne n individus dans une population et chaque individu appartient à une catégorie parmi b possibles. On cherche à comparer la répartition observée des individus dans les b catégories à une répartition attendue.

La réalisation du test est en tout point identique au précèdent, les effectifs attendus étant calculés à partir de la distribution théorique. Il faut néanmoins retrancher au nombre de degrés de liberté $(b-1)$ le nombre ϕ de paramètres estimés pour calculer les effectifs attendus.

loi	ϕ	*paramètres*
Poisson	1	moyenne
Binomiale	1	probabilité de succès
Normale	2	moyenne et variance

1.6 Comparaison de variances

1.6.1 Test paramétrique de Fisher et Snedecor de comparaison de deux variances

Problème : On mesure les variables X_1 et X_2 dans deux populations 1 et 2. Les variances de X_1 et X_2 sont respectivement σ_1^2 et σ_2^2 et sont estimées par $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$ pour des échantillons de tailles n_1 et n_2. On cherche à comparer $\hat{\sigma}_1^2$ et $\hat{\sigma}_2^2$.

Conditions de validité : X_1 et X_2 sont indépendantes et suivent toutes les deux des lois normales.

Hypothèse nulle :

Test bilatéral : $\sigma_1^2 = \sigma_2^2$

Test unilatéral : $\sigma_1^2 \leqslant \sigma_2^2$ ou $\sigma_1^2 \geqslant \sigma_2^2$ selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : $\sigma_1^2 \neq \sigma_2^2$

Test unilatéral : $\sigma_1^2 > \sigma_2^2$ ou $\sigma_1^2 < \sigma_2^2$ selon l'*a priori* sur les données

Statistique : $f_{obs} = \hat{\sigma}_1^2/\hat{\sigma}_2^2$

Distribution de la statistique sous H_0 : F suit une loi de Fisher à $(n_1 - 1)$ et $(n_2 - 1)$ degrés de liberté.

Risque de première espèce : en supposant $f_{obs} > 1$

Test bilatéral : $p = P(F \geqslant f_{obs}|_{H_0}) + P(F \leqslant 1/f_{obs}|_{H_0})$

Test unilatéral : $p = P(F \geqslant f_{obs}|_{H_0})$ ou $p = P(F \leqslant 1/f_{obs}|_{H_0})$

Code R : `var.test(X1,X2)`

1.6.2 Test paramétrique de Bartlett de comparaison de k variances

Problème : On mesure les quantités X_i dans dans des populations i pour i allant de 1 à k. Les moyennes et variances de X_i sont μ_i et σ_i^2 et sont estimées, respectivement, par $\hat{\mu}_i$, $\hat{\sigma}_i^2$ pour des échantillons de taille n_i. On cherche à comparer simultanément les $\hat{\sigma}_i^2$.

Conditions de validité : Les X_i sont indépendantes et suivent toutes des lois normales

Hypothèse nulle : $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$

Hypothèse alternative : Il existe au moins une des k variances qui diffère des autres

Statistique : $b_{obs} = A/C$ avec

$$A = \left\{ \ln(\hat{\sigma}_p^2) \sum_{i=1}^{k} \{n_i - 1\} \right\} - \sum_{i=1}^{k} \left\{ (n_i - 1) \ln(\hat{\sigma}_j^2) \right\} \quad \text{avec}$$

$$\ln \text{ le logarithme népérien}, \quad \hat{\sigma}_p^2 = \frac{\sum_{i=1}^{k} (n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{k} (n_i - 1)}$$

,

$$\text{et} \quad C = 1 + \frac{1}{3(k-1)} \left\{ \sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{\sum_{i=1}^{k} \{n_i - 1\}} \right\}$$

Distribution de la statistique sous H_0 : B suit une loi du χ^2 à $(k - 1)$ degrés de liberté

Risque de première espèce : $p = P(B \geqslant b_{obs}|_{H_0})$

Un léger écart à la normalité a tendance à produire un résultat de test de Barlett significatif même si les k variances comparées sont en réalité identiques.

Code R : `bartlett.test(y~x)`

1.6.3 Autres tests non-paramétriques possibles

Test de Ansari-Bradley : Une version non-paramétrique du test de Fisher, recommandée par la documentation de R.

Code R : `ansari.test`

Test de Mood : Une version non-paramétrique du test de Fisher.

Code R : `mood.test`

Test de Fligner-Killeen : une version non paramétrique de comparaison de k variances, qui évite le problème de la sensibilité du test de Bartlett à de légers écarts à la normalité.

Code R : `fligner.test`

1.7 Estimation du lien et de la corrélation entre variables

1.7.1 Définition de la covariance et la corrélation

La covariance mesure le lien qui existe entre deux variables, elle est la même entre Y_1 et Y_2 qu'entre Y_2 et Y_1.

Soient deux variables aléatoires Y_1 et Y_2 de moyennes μ_{Y_1} et μ_{Y_2} et de variances $\sigma^2_{Y_1}$ et $\sigma^2_{Y_2}$. La covariance entre Y_1 et Y_2, $\sigma_{Y_1Y_2}$, est définie par [1] :

$$\begin{aligned}\sigma_{Y_1Y_2} =& \langle (Y_1 - \mu_{Y_1})(Y_2 - \mu_{Y_2}) \rangle \\ =& \langle Y_1Y_2 \rangle - \mu_{Y_1}\mu_{Y_2}\end{aligned}$$

Si Y_1 et Y_2 sont deux variables indépendantes, c'est-à-dire si la valeur de Y_1 ne renseigne en rien sur la valeur de Y_2, alors $P(Y_1 = y_1 \cap Y_2 = y_2) = P(Y_1 = y_1)P(Y_2 = y_2)$: la probabilité d'observer à la fois $Y_1 = y_1$ et $Y_2 = y_2$ est égale au produit des probabilités de chacun de ces deux évènements. Dans ce cas, $\langle Y_1Y_2 \rangle = \mu_{Y_1}\mu_{Y_2}$ et donc $\sigma_{Y_1Y_2} = 0$. Si la covariance est positive, les individus dont la mesure Y_1 est plus élevée que μ_{Y_1} ont en moyenne des valeurs de Y_2 plus élevées que μ_{Y_2}. Si la covariance est négative, les individus dont la mesure Y_1 est plus élevée que μ_{Y_1} ont en moyenne des valeurs de Y_2 moins élevées que μ_{Y_2}.

En pratique on utilise plus souvent la corrélation, qui correspond à la covariance standardisée par le produit des écarts types :

$$r_{Y_1Y_2} = \frac{\sigma_{Y_1Y_2}}{\sigma_{Y_1}\sigma_{Y_2}}$$

La corrélation varie donc entre -1 et $+1$ et la relation entre covariance et corrélation est résumée par les deux tableaux suivants :

matrice de covariance

	Y_1	Y_2
Y_1	$\sigma^2_{Y_1}$	σ_{Y_1,Y_2}
Y_2	σ_{Y_2,Y_1}	$\sigma^2_{Y_2}$

matrice de corrélation

	Y_1	Y_2
Y_1	1	r_{Y_1,Y_2}
Y_2	r_{Y_2,Y_1}	1

Estimation de la covariance et de la corrélation

$$\begin{aligned}\hat{\sigma}_{Y_1Y_2} =& \frac{\sum_{i=1}^{n}(y_{1i} - \hat{\mu}_{Y_1})(y_{2i} - \hat{\mu}_{Y_2})}{n-1} \\ \hat{r}_{Y_1,Y_2} =& \frac{\hat{\sigma}_{Y_1,Y_2}}{\hat{\sigma}_{Y_1}\hat{\sigma}_{Y_2}} \\ =& \frac{\sum_{i=1}^{n}(y_{1i} - \hat{\mu}_{Y_1})(y_{2i} - \hat{\mu}_{Y_2})}{\sqrt{\sum_{i=1}^{n}(y_{1i} - \hat{\mu}_{Y_1})^2}\sqrt{\sum_{i=1}^{n}(y_{2i} - \hat{\mu}_{Y_2})^2}}\end{aligned}$$

1.7.2 Tests de corrélation

Problème : On mesure deux variables Y_1 et Y_2 sur n individus issus d'une même population. On cherche à démontrer que Y_1 et Y_2 sont liées.

1.7.3 Test de corrélation de Pearson

Conditions de validité : Y_1 et Y_2 suivent une distribution binormale. En pratique, pour tester la validité de cette hypothèse, on vérifie que (1) chacune des deux variables est normalement distribuée, (2) la variance d'une variable ne change pas en fonction de l'autre variable (hypothèse d'homoscédasticité) et (3) la relation entre les deux variables est linéaire.

Hypothèse nulle :

Test bilatéral : $r_{Y_1Y_2} = 0$

Test unilatéral : $r_{Y_1Y_2} \leqslant 0$) ou ($r_{Y_1Y_2} \geqslant 0$) selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : Y_1 et Y_2 sont liées ($r_{Y_1Y_2} \neq 0$)

Test unilatéral : Y_1 et Y_2 sont liées positivement ($r_{Y_1Y_2} > 0$) ou négativement ($r_{Y_1Y_2} < 0$) selon l'*a priori* sur les données

Statistique : $t_{obs} = \hat{r}_{Y_1Y_2}/\sqrt{(1 - \hat{r}^2_{Y_1Y_2})/(n-2)}$

1. Par convention, on note entre crochets $\langle \ldots \rangle$ l'espérance mathématique d'une variable aléatoire. $\langle Y_1Y_2 \rangle$ désigne donc l'espérance du produit des variables aléatoires Y_1 et Y_2.

Distribution de la statistique sous H_0 : t_{obs} suit une distribution de Student à $(n-2)$ degrés de liberté

Risque de première espèce :
Test bilatéral : $p = P(|T| \geqslant |t_{obs}|\,|_{H_0})$
Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `cor.test(x,y)`

1.7.4 Test non-paramétrique de corrélation de Spearman

Condition de validité : Aucune ; le test ne dépend pas de la normalité des variables étudiées, ni de la forme de la relation entre Y_1 et Y_2 pourvue qu'elle soit monotone. Pour s'en convaincre, il suffit de réaliser le test sur Y_1 et Y_2, ou $\log(Y_1)$ et $\log(Y_2)$ ou bien encore $\exp(Y_1)$ et $\exp(Y_2)$, et de constater que le résultat est toujours le même.

Hypothèse nulle :
Test bilatéral : Y_1 et Y_2 sont indépendantes ($rs_{Y_1Y_2} = 0$)
Test unilatéral : $rs_{Y_1Y_2} \leqslant 0$ ou $rs_{Y_1Y_2} \geqslant 0$ selon les données

Hypothèse alternative :
Test bilatéral : Y_1 et Y_2 sont liées ($rs_{Y_1Y_2} \neq 0$)
Test unilatéral : $rs_{Y_1Y_2} > 0$ ou $rs_{Y_1Y_2} < 0$ selon les données

Statistique : $\hat{rs}_{Y_1Y_2} = r\hat{s}_{Y_1Y_2} = 1 - 2\frac{\sum_i d_i^2}{\frac{1}{3}n(n^2-1)}$, avec
d_i la différence de rangs entre Y_{1i} et Y_{2i} et n la taille de l'échantillon.
Si les classements de Y_1 et Y_2 sont identiques, $\sum_i d_i^2 = 0$ et $\hat{rs}_{Y_1Y_2} = 1$. Si les deux classements sont au contraire inverses, $\sum_i d_i^2 = \frac{1}{3}n(n^2-1)$ et $\hat{rs}_{Y_1Y_2} = -1$.

Distribution de la statistique sous H_0 : Si n est grand ($n > 30$), $\hat{rs}_{Y_1Y_2}$ suit une loi normale centrée de variance $\frac{1}{n-1}$. Sinon il faut utiliser la table de Spearman.

Risque de première espèce :
si $n > 30$, on définit Z une variable aléatoire qui suit une loi normale centrée réduite. On a alors :
Test bilatéral : $p = P(|Z| \geqslant |\sqrt{n-1}\hat{rs}_{Y_1Y_2}|)$
Test unilatéral : $p = P(Z \geqslant \sqrt{n-1}\hat{rs}_{Y_1Y_2})$ ou $p = P(Z \leqslant \sqrt{n-1}\hat{rs}_{Y_1Y_2})$

Code R : `cor.test(Y1,Y2,method="spearman")`

Exemple du calcul de la statistique de Spearman

Soit les valeurs observées suivantes :

i	y_1^i	rang sur y_{1i}	y_{2i}	rang sur y_{2i}	d_i	d_i^2
1	12	2	14	3	1	1
2	15	3	7	1	-2	4
3	18	4	20	5	1	1
4	22	5	18	4	-1	1
5	3	1	8	2	1	1

On classe les données selon Y_1, puis selon Y_2. On calcule ensuite pour chaque observation la différence d_i entre les deux classements et on élève cette différence au carré. Enfin, on additionne les valeurs obtenues. Avec les valeurs observées ci-dessus $\sum_i d_i^2 = 8$ et $\hat{rs}_{Y_1Y_2} = 1 - 2\frac{8}{\frac{1}{3}5(25-1)} = 0,6$

Autres tests non-paramétriques possibles

Test de corrélation de Kendall : Ce test utilise plus d'information et reste valide pour des faibles tailles d'échantillon ($n > 8$)

Code R : `cor.test` avec l'option `method="kendall"`

Chapitre 2

Le modèle linéaire

2.1 Définition d'un modèle statistique

Un modèle statistique est un outil mathématique qui permet de prédire des observations tout en évaluant les erreurs commises dans les prédictions. Il ne prétend pas décrire un mécanisme, mais de reproduire les observations dans les conditions dans lesquelles elles ont été réalisées. Un bon modèle est donc un modèle qui permet de bien décrire des données tout en restant le plus simple possible.

Cette notion de modèle n'est pas totalement absente des tests classiques présentés dans les chapitres précédents ; on pouvait cependant raisonner sans aborder la notion de modèle. La suite montre qu'intégrer cette notion permet de construire des outils statistiques très flexibles et très puissants.

Tous les modèles linéaires peuvent s'écrire sous la forme :

$$y_i = \hat{y}_i + \epsilon_i$$

avec y_i l'**observation** i, $\hat{y}_i$ la **prédiction** correspondante et ϵ_i l'écart entre l'observation et la prédiction, c'est-à-dire l'erreur commise dans la prédiction. Ce terme d'erreur est généralement appelé **résidu** et, dans le modèle linéaire classique, on suppose qu'il suit une **distribution normale**.

Les différentes catégories ou applications du modèle linéaire sont toutes construites sur le modèle ci-dessus ; elles se distinguent principalement par la façon dont la prédiction est faite, c'est-à-dire par la nature et le nombre de variables prédictives (facteurs explicatifs) utilisé pour calculer $\hat{y}_i$.

Ce chapitre commence par deux applications simples du modèle linéaire : la régression linéaire et l'ANOVA à un facteur de classification. La suite montre que l'on peut facilement généraliser les principes mis en œuvre dans ces deux outils, notamment grâce aux exemples de l'ANOVA à deux facteurs croisés et de l'ANCOVA.

Dans tous les cas, l'analyse de données se décompose en cinq étapes :

1. définition du modèle statistique,
2. ajustement des paramètres du modèle aux données,
3. vérification du bon ajustement du modèle et de la validité des conditions d'application (souvent *a posteriori*)
4. décomposition de la variance en variance expliquée et résiduelle
5. test des effets des variables prédictives incluses dans le modèle.

2.2 Deux modèles linéaires simples

2.2.1 La régression linéaire simple

Problème : On mesure dans une population des valeurs Y_i (variable **dépendante** pour certaines valeurs x_i contrôlées). On cherche à expliquer les variations de la variable aléatoire quantitative Y en fonction des variations de la variable quantitative x (variable **explicative** ou variable contrôlée). La variable x, notée en minuscule car non aléatoire, correspond à un **effet fixe** [a].

a. Contrairement à la corrélation où les deux variables sont aléatoires, ici une variable explique l'autre. L'objectif de la régression n'est donc pas de mesurer l'association entre variables, mais de savoir si une variable (la variable explicative) explique l'autre variable (la variable dépendante).

Première étape : décrire les données par un modèle linéaire

On cherche à prédire les valeurs possibles de Y étant données les valeurs de x. Pour y parvenir, on construit le modèle statistique suivant :

$$Y_i = \hat{y}_i + \epsilon_i$$

où $\hat{y}_i$ est la prédiction de y faite pour l'observation y_i et où ϵ_i est une variable aléatoire correspondant à la différence entre la valeur observée y_i et la valeur prédite $\hat{y}_i$. Cette variable aléatoire décrit donc l'erreur que l'on commet dans la prédiction. On va ensuite supposer qu'une relation linéaire unit x et Y pour pouvoir calculer la prédiction $\hat{y}_i$

$$\hat{y}_i = \hat{\mu}_Y + \beta(x_i - \hat{\mu}_x)$$

où $\hat{\mu}_Y$ est la moyenne des y_i, $\hat{\mu}_x$ la moyenne des x_i et où β est un paramètre[1]. L'erreur globalement commise dans les prédictions peut être quantifiée par la Somme des Carrés des Écarts Résiduels (SCER) :

$$\begin{aligned}\text{SCER} &= \sum_i (\hat{y}_i - y_i)^2 \\ &= \sum_i (\hat{\mu}_Y - y_i + \beta(x_i - \hat{\mu}_x))^2\end{aligned}$$

On cherche à estimer la valeur $\hat{\beta}$ pour β de sorte de minimiser les écarts entre les prédictions et les observations grace à la méthode dite des moindres carrés. Cette valeur est telle que $\partial SCER/\partial\beta|_{\hat{\beta}} = 0$. Or,

$$\begin{aligned}\text{SCER} &= \sum_i (y_i - \hat{\mu}_Y)^2 + \beta^2 \sum_i (x_i - \hat{\mu}_x)^2 - 2\beta \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y) \\ \frac{\partial \text{SCER}}{\partial \beta} &= 2\beta \sum_i (x_i - \hat{\mu}_x)^2 - 2 \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y)\end{aligned}$$

donc SCER est minimal pour $\beta = \hat{\beta}$ avec

$$\begin{aligned}\hat{\beta} &= \frac{\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y)}{\sum_i (x_i - \hat{\mu}_x)^2} \\ &= \frac{\hat{\sigma}_{X,Y}}{\hat{\sigma}_x^2}\end{aligned}$$

Deuxième étape : test de la pente de la droite de régression

On peut montrer que, si les erreurs $\epsilon_i = Y_i - \hat{y}_i$ sont normalement distribuées, $\hat{\beta}$ suit une distribution normale dont la variance peut être estimée par

$$\hat{\sigma}_{\hat{\beta}}^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{(n-2)\sum_i (x_i - \hat{\mu}_x)^2}$$

On peut se servir de ce résultat pour tester la significativité de la relation entre X et Y.

Conditions de validité : Les résidus $\epsilon_i = Y_i - \hat{y}_i$ sont normalement distribués

Hypothèse nulle :

Test bilatéral : La pente de la droite de régression est nulle ($\beta = 0$)
Test bilatéral : $\beta \leqslant 0$ ou $\beta \geqslant 0$ selon l'*a priori* sur les données

Hypothèse alternative :

Test bilatéral : la pente de la droite de régression n'est pas nulle ($\beta \neq 0$)
Test unilatéral : $\beta > 0$ ou $\beta < 0$

Statistique : $t_{obs} = \hat{\beta}/\hat{\sigma}_{\hat{\beta}}$.

Distribution de la statistique sous H_0 : T suit une distribution de Student à $n-2$ degrés de liberté.

1. On peut paramétrer différement l'équation en utilisant le classique Y : $\hat{y}_i = \alpha + \beta x_i$. Les deux formulations sont rigoureusement équivalentes et utilisent la même valeur $\hat{\beta}$. Ceci dit, notre formulation permet de montrer plus implicitement que la droite de régression passe par le point de coordonnées $(\hat{\mu}_x, \hat{\mu}_Y)$. Par ailleurs, notre formulation a le grand avantage d'épargner quelques étapes dans les calculs.

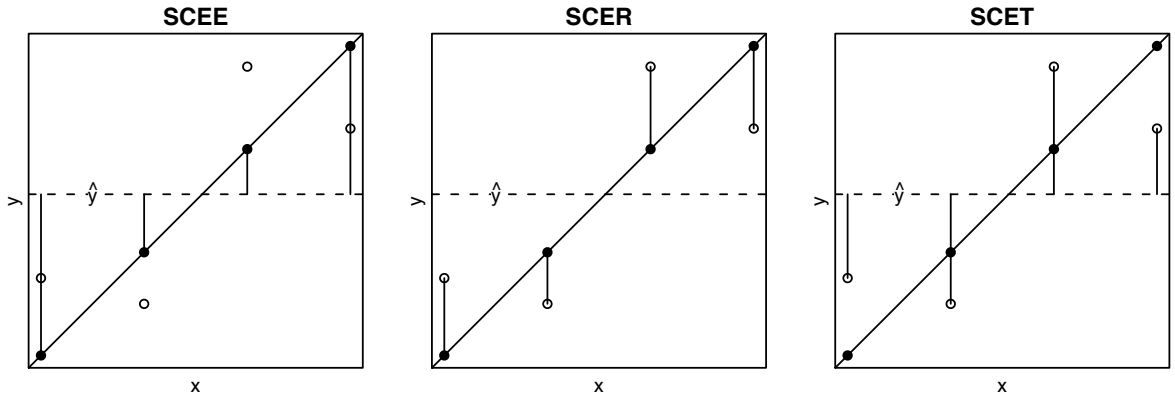


FIGURE 2.1 – Régression linéaire simple. Chaque cercle correspond à un couple (x_i, y_i) de valeurs, x étant une variable explicative et Y la variable à expliquer. La droite en trait plein correspond à la droite de régression. Chaque point placé sur cette droite correspond donc à une prédiction $\hat{y}_i$ réalisée pour une valeur de x_i. La droite en pointillé indique $\hat{\mu}_Y$, la moyenne des y_i. SCEE : somme des carrés des écarts expliqués ; chaque segment vertical représente l'écart entre la prédiction $\hat{y}_i$ et $\hat{\mu}_Y$. SCER : somme des carrés des écarts résiduels ; chaque segment vertical représente l'écart entre la prédiction $\hat{y}_i$ et l'observation y_i. SCET : somme des carrés des écarts totaux ; chaque segment vertical représente l'écart entre y_i et $\hat{\mu}_Y$ la moyenne des Y. Lorsque le modèle est ajusté, c'est-à-dire lorsque SCER est minimal et que $\hat{\beta} = \hat{\sigma}_{X,Y}/\hat{\sigma}_x^2$, on a $SCET = SCEE + SCER$.

Risque de première espèce :

Test bilatéral : $p = P(|T| \geqslant |t_{obs}|\,|_{H_0})$
Test unilatéral : $p = P(T \geqslant t_{obs}|_{H_0})$ ou $p = P(T \leqslant t_{obs}|_{H_0})$

Code R : `lm` pour ajuster le modèle aux données (c'est-à-dire pour estimer le paramètre β), puis `summary` pour réaliser le test de Student sur les paramètres estimés. Il faut vérifier *a posteriori* que les conditions de validité du modèle linéaire soient bien respectées (cf. la section suivante).

Deuxième étape (bis) : variance expliquée et variance résiduelle ; test de Fisher

Le carré des écarts entre prédiction et observation nous fournit une mesure de l'erreur de prédiction :

$$\text{SCER} = \sum_i (y_i - \hat{\mu}_Y)^2 + \hat{\beta}^2 \sum_i (x_i - \hat{\mu}_x)^2 - 2\hat{\beta} \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y)$$

or

$$\hat{\beta} = \frac{\sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y)}{\sum_i (x_i - \hat{\mu}_x)^2} \Rightarrow \sum_i (x_i - \hat{\mu}_x)(y_i - \hat{\mu}_Y) = \hat{\beta} \sum_i (x_i - \hat{\mu}_x)^2$$

donc

$$\text{SCER} = \sum_i (y_i - \hat{\mu}_Y)^2 + \hat{\beta}^2 \sum_i (x_i - \hat{\mu}_x)^2 - 2\hat{\beta}^2 \sum_i (x_i - \hat{\mu}_x)^2$$

et comme $\hat{\beta}(x_i - \hat{\mu}_x) = \hat{y}_i - \hat{\mu}_Y$ on obtient

$$\text{SCER} = \sum_i (y_i - \hat{\mu}_Y)^2 - \sum_i (\hat{y}_i - \hat{\mu}_Y)^2$$

La dispersion de Y peut donc se décomposer de la façon suivante :

$\sum_i (y_i - \hat{\mu}_Y)^2$	=	$\sum_i (\hat{y}_i - \hat{\mu}_Y)^2$	+	$\sum_i (y_i - \hat{y}_i)^2$
SCET		**SCEE**		**SCER**
dispersion totale	=	dispersion expliquée	+	dispersion résiduelle
(d.d.l. $= n - 1$)		(d.d.l. $= 1$)		(d.d.l. $= n - 2$)

De cette décomposition, illustrée par la figure 2.1 on peut déduire la quantité R^2 :

$$R^2 = \frac{SCEE}{SCET}$$

Le coefficient de détermination R^2 représente la part de la variation expliquée par le modèle. Plus R^2 est proche de 1, meilleure est la prédiction fournie par notre modèle. Dans le cas de la régression linéaire simple, on peut montrer que $R^2 = \hat{r}^2_{X,Y}$. La décomposition de la variation précédente est résumée dans la table qui suit, où σ^2 est la variance de ϵ et où σ^2_E est la variance expliquée par le modèle :

Source de variation		ddl	CM	CM attendu	F
Expliquée	$SCEE$	1	$CME = \frac{SCEE}{1}$	$\sigma^2 + \sigma^2_E$	$\frac{CME}{CMR}$
Résiduelle	$SCER$	$n-2$	$CMR = \frac{SCER}{n-2}$	σ^2	
Totale	$SCET$	$n-1$			

On peut exploiter cette décomposition pour tester la significativité de la relation entre X et Y. Le lien entre les deux variables ne peut en effet être significatif que si notre modèle est explicatif, autrement si $R^2 > 0$. Pour tester cette hypothèse on ne se sert pas de R^2 directement mais plutôt du rapport (F) entre le carré moyen expliqué (CME) et la carré moyen résiduel (CMR). D'après la table précédente, si le modèle n'est pas explicatif ($\sigma^2_E = 0$), ces deux carrés moyens sont égaux en espérance à σ^2 et leur rapport vaut donc 1.

Conditions de validité : Les erreurs $\epsilon_i = Y_i - \hat{y}_i$ sont distribuées normalement, leur variance est constante (elle ne dépend notamment pas de x_i).

Hypothèse nulle : Le modèle n'est pas explicatif ($\sigma^2_E = 0$)

Hypothèse alternative : Le modèle est explicatif ($\sigma^2_E > 0$)

Statistique : $f_{obs} = \dfrac{SCEE/1}{SCER/(n-2)}$

Distribution de la statistique sous H_0 : F suit une distribution de Fisher à 1 et $n-2$ degrés de liberté

Risque de première espèce : $p = P(F \geqslant f_{obs}|_{H_0})$

Code R : `lm` pour ajuster le modèle aux données (c'est-à-dire pour estimer le paramètre β), puis `anova` pour réaliser la décomposition de la variance et le test de Fisher. Il faut également vérifier *a posteriori* les conditions de validité du modèle construit. On peut utiliser les outils graphiques mis à disposition dans R (fonction générique `plot(modele)`) puis on doit réaliser les trois tests pour s'assurer des conditions de validité : (1) le test de Shapiro (`shapiro.test(residuals(modele))`) pour tester la normalité des résidus, (2) le test de Bartlett (`bartlett.test(y~x)`) ou plus généralement celui de Breusch-Pagan (`ncv.test` dans le package `car`, `bptest(y~x)` dans le package `lmtest`) pour tester l'homoscédasticité, (3) le test de Durbin-Watson (`dwtest(y~x)` dans le package `lmtest`) pour tester l'indépendance des résidus.

Corrélation ou régression ?

But de l'analyse	Y aléatoire, X fixé	Y_1 et Y_2 aléatoires
Décrire la relation fonctionnelle entre les deux variables	Régression de type I (vue ici)	Méthode des axes majeurs (cf. cours d'analyse multivariée)
Estimer l'association entre deux variables	Impossible	Corrélation

2.2.2 L'ANOVA à un facteur de classification

		Population					
		A_1	$\dots$	A_i	$\dots$	A_a	
Individu	1	y_{11}	$\dots$	y_{i1}	$\dots$	y_{a1}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	j	y_{1j}	$\dots$	y_{ij}	$\dots$	y_{aj}	
	$\vdots$	$\vdots$		$\vdots$		$\vdots$	
	n	y_{1n}	$\dots$	y_{in}	$\dots$	y_{an}	
			$\dots$	$\sum_{j=1}^{n} y_{ij}$	$\dots$		$\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}$

Problème : On mesure les valeurs Y_i sur n individus échantillonnés dans chacune des a populations ;les effectifs sont les mêmes dans chaque population et sont alors dits équilibrés ou balancés. On cherche à comparer les moyennes des mesures Y entre les a populations.

Définition du modèle statistique

Soient $\hat{\mu}$ la moyenne estimée sur l'ensemble des mesures et $\hat{\mu}_i$ la moyenne estimée pour une population i avec $\hat{\mu}_i = \hat{\mu} + \hat{\alpha}_i$. La quantité $\hat{\alpha}_i$ représente donc l'écart entre la moyenne de la population i et la moyenne générale. Le modèle statistique s'écrit :

$$Y_{ij} = \hat{y}_i + \epsilon_{ij}$$

où $\hat{y}_i$ correspond à une prédiction de la mesure dans la population i et où ϵ_{ij} mesure l'écart entre cette prédiction et la mesure réalisée sur l'individu j de la population i (c'est-à-dire l'erreur commise dans la prédiction). La valeur prédite des mesures dans la population i s'écrit simplement

$$\hat{y}_i = \mu + \alpha_i$$

l'estimation de μ étant $\hat{\mu}$ et celle de α_i étant $\hat{\alpha}_i$. La prédiction du modèle pour un individu sera donc la moyenne de la population à laquelle cet individu appartient.

Dans l'ANOVA à un facteur, l'ajustement du modèle aux données est immédiat : on peut en effet montrer que le modèle pour lequel la prédiction est la moyenne de la population à laquelle l'individu mesuré appartient minimise effectivement le carré des écarts résiduels.

Deuxième étape : test sur les paramètres estimés

Il est possible, à partir du modèle décrit précédemment, de réaliser un test de Student sur les paramètres estimés comme fait précédemment pour la régression linéaire, mais ici il y a autant de paramètres $\hat{\alpha}_i$ et donc de tests à réaliser qu'il y a de populations avec chacun l'hypothèse nulle que $\hat{\alpha}_i$ est égal à zéro. [2] Cette approche est intéressante seulement si on veut savoir s'il existe des différences entre les populations. Sinon, il faut réaliser un test unique qui est fondé sur la décomposition de la variance (cf. ci-dessous).

Deuxième étape (bis) : test de Fisher sur les variances expliquée et résiduelle

On pose

$$\hat{\mu} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}}{an}$$

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n} y_{ij}}{n}$$

La somme des carrés des écarts (SCE) totale peut s'écrire $SCET = SCEE + SCER$ avec

$$SCET = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \hat{\mu})^2 \qquad \text{SCE totale}$$

$$= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \left(\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right)^2 /an$$

$$SCEE = \sum_{i=1}^{a} n\,(\hat{\mu}_i - \hat{\mu})^2 \qquad \text{SCE expliquée}$$

$$= \sum_{i=1}^{a} \left(\sum_{j=1}^{n} y_{ij}\right)^2 /n - \left(\sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}\right)^2 /an$$

$$SCER = \sum_{i=1}^{a}\sum_{j=1}^{n} (y_{ij} - \hat{\mu}_i)^2 \qquad \text{SCE résiduelle}$$

2. Attention, les modèles d'ANOVA ne sont en fait généralement, et notamment dans R, pas écrits de cette façon : il prennent comme point de référence une population particulière (la première par défaut), pour laquelle le paramètre $\hat{\alpha}_i$ est fixé à zéro, et non pas la moyenne des populations. Dans ces modèles il y a donc $a - 1$ paramètres estimés et chacun de ces paramètres quantifie l'écart entre la moyenne de la population de référence et la moyenne de la population de référence.

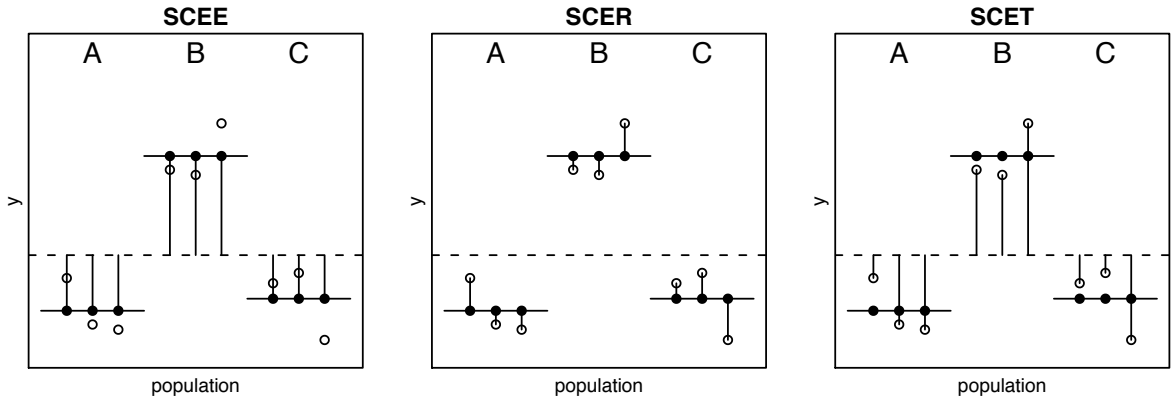


FIGURE 2.2 – Représentation des prédictions et des résidus d'une ANOVA à un facteur de classification. Chaque cercle correspond à une mesure y_i réalisée dans une des trois populations (A, B ou C) étudiées. Les segments horizontaux en trait plein correspondent à la prédiction du modèle pour chaque population. Chaque point placé sur ces segments correspond donc à une prédiction $\hat{y}_i$. La droite en pointillé indique $\hat{\mu}_Y$, la moyenne des y_i. SCEE : somme des carrés des écarts expliqués ; chaque segment vertical représente l'écart entre la prédiction $\hat{y}_i$ et $\hat{\mu}_Y$. SCER : somme des carrés des écarts résiduels ; chaque segment vertical représente l'écart entre la prédiction $\hat{y}_i$ et l'observation y_i. SCET : somme des carrés des écarts totaux ; chaque segment vertical représente l'écart entre y_i et $\hat{\mu}_Y$ la moyenne des Y. Lorsque le modèle est ajusté, c'est-à-dire lorsque SCER est minimal et que la prédiction correspond à la moyenne de la population à laquelle l'individu mesuré appartient, on a $SCET = SCEE + SCER$.

$$= \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}^2 - \sum_{i=1}^{a}\left(\sum_{j=1}^{n} y_{ij}\right)^2 /n$$

Cette décomposition est ensuite utilisée pour construire le tableau d'analyse de la variance :

Source de variation		ddl	CM	CM attendu	F
Expliquée	$SCEE$	$a-1$	$CME = \frac{SCEE}{a-1}$	$\sigma^2 + n\sigma_E^2$	$\frac{CME}{CMR}$
Résiduelle	$SCER$	$a(n-1)$	$CMR = \frac{SCER}{a(n-1)}$	σ^2	
Totale	$SCET$	$an-1$			

où σ^2 est la variance de la variable aléatoire ϵ_{ij} dans le modèle statistique et σ_E^2 est la part de la variance expliquée par les différences entre les a groupes.

La décomposition de la variance est alors utilisée pour tester l'effet du facteur population sur la variable mesurée :

Conditions de validité : $\epsilon_{ij} = Y_{ij} - \hat{y}_i$ suit une distribution normale centrée et sa variance est égale à σ^2 dans chacune des a populations

Hypothèse nulle : $\mu_1 = \mu_2 = \ldots = \mu_a$ si bien que la variation inter-population est du même ordre que la variation intra-population ($\sigma_E^2 = 0 \Rightarrow CME = CMR$)

Hypothèse alternative : Au moins une des a populations diffère des autres ($\sigma_E^2 > 0 \Rightarrow CME > CMR$)

Statistique : $f_{obs} = \frac{CME}{CMR}$

Distribution de la statistique sous H_0 : F suit une distribution de Fisher à $a-1$ et $a(n-1)$ degrés de liberté

Risque de première espèce : $p = P(F \geqslant f_{obs}|_{H_0})$

Code R : La procédure est identique à celle indiquée pour la régression linéaire : `lm` pour estimer les paramètres α_i, puis `summary` pour les obtenir, et enfin `anova` pour réaliser la décomposition de la variance et le test de Fisher. Il faut vérifier *a posteriori* que les conditions de validité du modèle construit sont remplies (cf. régression linéaire page 16).

Détails sur la signification de la décomposition de la variance

Nous allons ici expliquer de façon intuitive comment on peut tester l'existence de différences entre des moyennes en comparant des variances, un point délicat qu'il est important de comprendre.

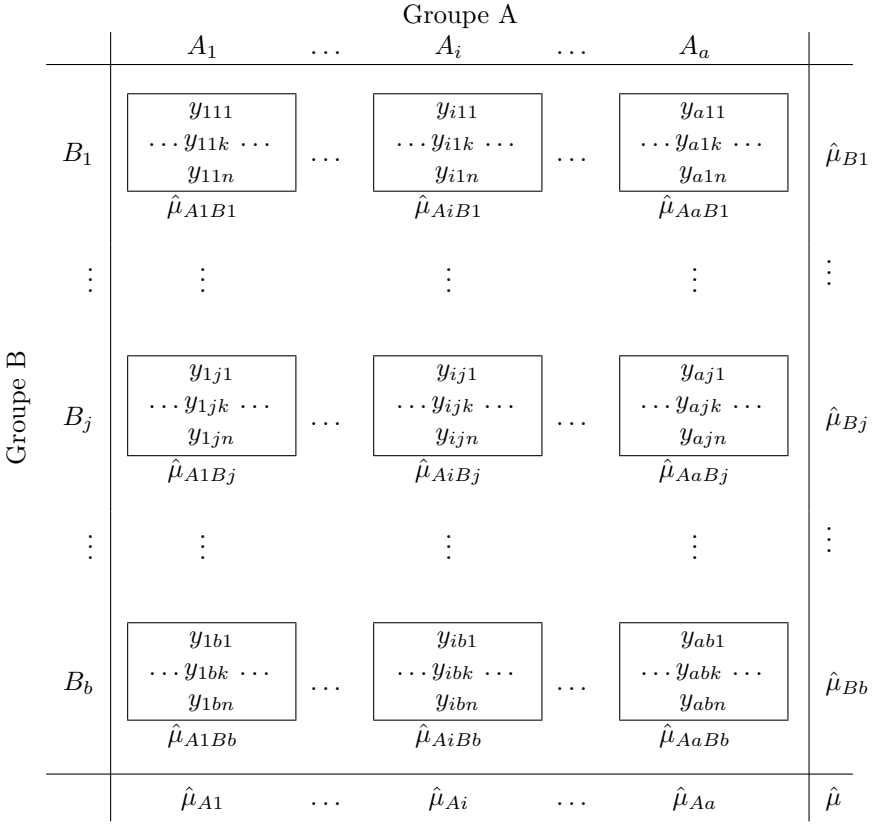
Imaginons que l'on réalise une mesure sur deux individus choisis dans deux populations différentes. Deux situations sont alors possibles : (1) si les populations sont identiques, on ne s'attend pas à observer plus de différences entre ces deux individus qu'entre deux individus choisis au hasard dans la même population, (2) si au contraire les populations sont différentes, on s'attend à observer plus de différences entre les deux individus qu'entre deux individus choisis au hasard dans la même population.

On peut estimer les variances intra-population (ou inter-individus) et inter-populations pour contraster ces deux situations. La variance intra-population - le carré moyen résiduel - estime les différences entre des individus issus d'une population peu importe laquelle. Les différences entre des individus issus de populations différentes - le carré moyen expliqué - dépendent à la fois de cette variance intra-population et de la variance entre populations (la variance inter-population).

Pour tester l'hypothèse nulle de l'égalité des moyennes de toutes les populations, il suffit alors de tester l'hypothèse selon laquelle le carré moyen expliqué est égal au carré moyen résiduel. L'hypothèse alternative serait que le carré moyen expliqué est supérieur au carré moyen résiduel, puisqu'il ne peut pas y avoir moins de différence entre deux individus issus de deux populations distinctes qu'entre deux individus issus de la même population. On peut utiliser le test de comparaison de variances de Fisher pour comparer le carré moyen expliqué au carré moyen résiduel.

2.3 Des modèles plus complexes

2.3.1 L'ANOVA à deux facteurs croisés



Problème : On mesure les valeurs Y_i sur des individus qui peuvent être regroupés selon deux critères (ou facteurs) A et B possédant respectivement les modalités i et j (i allant de 1 à a, et j allant de 1 à b). Pour chaque combinaison de ces deux critères (i.e. traitement), on a échantillonné n individus. On cherche à expliquer les variations de la variable Y en fonction des deux critères (ou facteurs) et de leur interaction. Par exemple on cherche à expliquer la production de graines de 3000 plantes de maïs provenant de dix localités différentes ($a = 10$) et ayant subit trois modalités d'apport en engrais ($b = 3$) ; le protocole étant équilibré, on dispose pour cela pour chaque traitement (i.e. chaque localité et chaque dose d'engrais) de 100 plantes.

Définition d'un modèle statistique

Soient μ la moyenne des mesures, $\mu_{Ai} = \mu + \alpha_i$ la moyenne pour la modalité i du facteur A et $\mu_{Bj} = \mu + \beta_j$ la moyenne pour la modalité j du facteur B. Soit $\mu_{AiBj} = \mu + \alpha_i + \beta_j + \gamma_{ij}$, la moyenne des individus combinant les modalités i du facteur A et j du facteur B. Le modèle statistique s'écrit :

$$Y_{ijk} = \hat{y}_{ij} + \epsilon_{ijk}$$

où $\hat{y}_{ij}$ correspond à une prédiction de la mesure pour les individus possédant à la fois les caractéristiques A_i et B_j et où ϵ_{ijk} mesure l'écart entre cette prédiction et la mesure réalisée sur l'individu k. La valeur prédite des mesures pour des individus issus de la modalité i du facteur A et j du facteur B s'écrit

$$\hat{y}_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

les paramètres α_i, β_j et γ_{ij} quantifiant respectivement l'effet du facteur A, celui du facteur B et l'interaction entre ces deux facteurs.

Ajustement du modèle aux données

Comme pour l'ANOVA à un facteur de classification, l'ajustement est réalisé en estimant les moyennes des mesures par groupe. Dans le cas d'une ANOVA à deux facteurs, on peut calculer quatre moyennes de nature différente, selon les différents regroupements possibles des données :

$$\begin{aligned}
\text{1 moyenne } \hat{\mu} &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{abn} & \quad a \text{ moyennes } \hat{\mu}_{Ai} &= \frac{\sum_{j=1}^{b}\sum_{k=1}^{n} y_{ijk}}{bn} \\
b \text{ moyennes } \hat{\mu}_{Bj} &= \frac{\sum_{i=1}^{a}\sum_{k=1}^{n} y_{ijk}}{an} & \quad ab \text{ moyennes } \hat{\mu}_{AiBj} &= \frac{\sum_{k=1}^{n} y_{ijk}}{n}
\end{aligned}$$

$\hat{\mu}$ est donc la moyenne globale des données, $\hat{\mu}_{Ai}$ la moyenne des indivius issus de la modalité i du facteur A, $\hat{\mu}_{Bj}$ la moyenne des indivius issus de la modalité j du facteur B, et $\hat{\mu}_{AiBj}$ la moyenne des individus issus conjointement de la modalité i du facteur A et de la modalité j du facteur B la case (à l'intersection de la ligne i et de la colone j dans le tableau ci-dessus). De ces moyennes estimées, on en déduit les estimations des paramètres du modèle :

$$\hat{\alpha}_i = \hat{\mu}_{Ai} - \hat{\mu}; \quad \hat{\beta}_j = \hat{\mu}_{Bi} - \hat{\mu}; \quad \hat{\gamma}_{ij} = \hat{\mu}_{AiBi} + \hat{\mu} - \hat{\mu}_{Ai} - \hat{\mu}_{Bi}$$

Décomposition de la variation

La somme des carrés des écarts (SCE) totale peut s'écrire $SCET = SCEE_A + SCEE_B + SCEE_{AB} + SCER$ avec

$$SCET = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \hat{\mu})^2 \qquad \text{SCE totale}$$

$$SCEE_A = nb\sum_{i=1}^{a}\hat{\alpha}_i^2 = nb\sum_{i=1}^{a}(\hat{\mu}_{Ai} - \hat{\mu})^2 \qquad \text{SCE expliquée par le facteur A}$$

$$SCEE_B = na\sum_{j=1}^{b}\hat{\beta}_j^2 = na\sum_{j=1}^{b}(\hat{\mu}_{Bj} - \hat{\mu})^2 \qquad \text{SCE expliquée par le facteur B}$$

$$SCEE_{AB} = n\sum_{i=1}^{a}\sum_{i=1}^{b}\hat{\gamma}_{ij}^2 \qquad \text{SCE expliquée par l'interaction AB}$$

$$SCER = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(y_{ijk} - \hat{\mu}_{AiBj})^2 \qquad \text{SCE résiduelle}$$

Pour simplifier, on peut dire que $SCEE_A$ est la dispersion entre les colonnes du tableau, $SCEE_B$ celle entre les lignes et $SCEE_{AB}$ celle entre les cases du tableau. Cette décomposition est ensuite utilisée pour construire le tableau d'analyse de la variance :

Source de variation		ddl	CM	CM attendu	F
Expliquée facteur A	$SCEE_A$	$a-1$	$CME_A = \frac{SCEE_A}{a-1}$	$\sigma^2 + nb\sigma_A^2$	$\frac{CME_A}{CMR}$
Expliquée facteur B	$SCEE_B$	$b-1$	$CME_B = \frac{SCEE_B}{b-1}$	$\sigma^2 + na\sigma_B^2$	$\frac{CME_B}{CMR}$
Expliquée interaction AB	$SCEE_{AB}$	$(a-1)(b-1)$	$CME_{AB} = \frac{SCEE_{AB}}{(a-1)(b-1)}$	$\sigma^2 + n\sigma_{AB}^2$	$\frac{CME_{AB}}{CMR}$
Résiduelle	$SCER$	$ab(n-1)$	$CMR = \frac{SCER}{ab(n-1)}$	σ^2	
Totale	$SCET$	$abn-1$			

Test des effets

Trois tests peuvent être réalisés pour une ANOVA à deux facteurs croisés :
— Les deux tests des effets simples (facteurs A et B)
— Le test de l'interaction entre les facteurs A et B
Ces trois tests sont dans leur principe identiques à celui détaillé pour l'ANOVA à un facteur : on réalise un test de Fisher en utilisant comme statistique le rapport du carré moyen expliqué par le facteur testé (CME_A, CME_B ou CME_{AB}) et du carré moyen résiduel (CMR). Avant de réaliser ces tests, il faut évidemment de vérifier que les résidus ϵ_{ijk} sont indépendants, normalement distribués et que leur variance est constante (c.f. page 16).

2.3.2 ANCOVA

		Population					
		A_1	...	A_i	...	A_a	
Individu	1	x_{11}, y_{11}	...	x_{i1}, y_{i1}	...	x_{a1}, y_{a1}	
	⋮	⋮		⋮		⋮	
	j	x_{1j}, y_{1j}	...	x_{ij}, y_{ij}	...	x_{aj}, y_{aj}	
	⋮	⋮		⋮		⋮	
	n	x_{1n}, y_{1n}	...	x_{in}, y_{in}	...	x_{an}, y_{an}	

Problème : On mesure des variable dépendante Y_{ij} et une variable contrôlée x_{ij} sur n individus provenant de a groupes. On cherche à expliquer les variations de Y par rapport au facteur qualitatif (groupes) A, le facteur quantitatif x, ainsi que par rapport à l'interaction entre A et x.

Définition du modèle statistique

Soient $\hat{\mu}_Y$ la moyenne générale estimée des mesures, $\hat{\mu}_{Yi} = \hat{\mu}_Y + \hat{\alpha}_i$ la moyenne des mesures pour les individus du groupe i. Soient $\hat{\mu}_x$ la moyenne générale estimée du facteur x et $\hat{\mu}_{x_i}$ la moyenne de ce facteur estimée pour les individus du groupe i. Soient $\hat{\beta}$ l'effet linéaire moyen de x sur Y et $\hat{\gamma}_i$ l'effet linéaire de x_i sur Y pour les individus du groupe i (interaction). Le modèle statistique décrivant les variations de Y_{ij} en fonction de x_{ij} s'écrit :

$$Y_{ij} = \hat{y}_i(x_{ij}) + \epsilon_{ij}$$

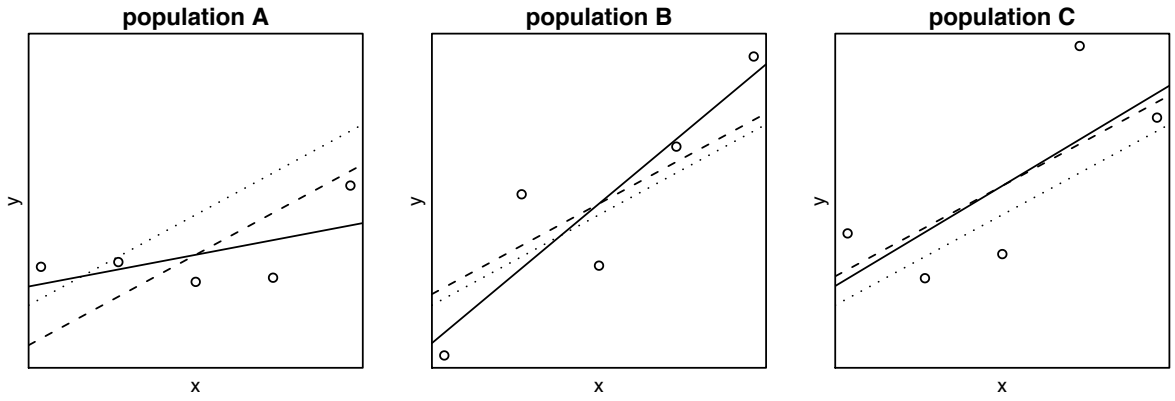


FIGURE 2.3 – Représentation des prédictions et des résidus d'une ANCOVA. Pour chaque population i, on peut calculer une droite de régression $\hat{y}_i = \mu_{Yi} + \hat{\beta}_i(x_i - \mu_{x_i})$, représentée ici en trait plein sur la figure. On peut aussi déterminer une droite de régression unique et identique pour toutes les populations $\hat{y} = \mu_Y + \hat{\beta}(x - \mu_x)$ (sans distinguer le facteur population), représentée ici en pointillé. On peut enfin calculer une droite de régression pour chaque population i $\hat{y}_i = \mu_{Yi} + \hat{\beta}(x - \mu_x)$ en supposant alors que la pente ne dépend pas de la population ; ces droites, représentées ici en tireté, ont toutes la même pente et sont donc parallèles.

où $\hat{y}_i(x_{ij})$ est une prédiction pour les individus du groupe i possédant une valeur x_{ij} de x. La variable aléatoire ϵ_{ij} decrit l'écart entre cette prédiction et l'observation. La prédiction s'écrit

$$\hat{y}_i(x_{ij}) = \mu_Y + \alpha_i + \beta(x - \mu_x) + \gamma_i(x_{ij} - \mu_{xi})$$

Décomposition de la variation

La variance totale peut être décomposer en quatre termes :

$SCEE_g$ La variance expliquée par la variance intergroupe (i.e. entre groupes) ; le calcul est identique à celui de l'anova à un facteur : $SCEE_g = \sum_i n(\hat{\mu}_{Yi} - \hat{\mu}_Y)$,

$SCEE_x$ La variance expliquée par l'effet linéaire de x sur Y ; le calcul est identique à celui de la régression linéaire simple : $SCEE_x = \sum_{i,j}(\hat{y}_{ij} - \hat{\mu}_Y)$,

$SCEE_{gx}$ La variance expliquée par l'interaction entre le facteur groupe et la variable explicative x,

$SCER$ La variance résiduelle.

Cette décomposition est ensuite utilisée pour construire, comme prédédemment, la table d'analyse de la variance :

Source de variation	**d.d.l.**	SCE	CM	F
Groupe	$a-1$	$SCEE_g$	$CM_g = SCEE_g/(a-1)$	CM_g/CM_e
x	1	$SCEE_x$	$CM_x = SCEE_x/1$	CM_x/CM_e
Groupe$\times x$	$a-1$	$SCEE_{gx}$	$CM_{gx} = SCEE_i/(a-1)$	CM_i/CM_e
Erreur	$an-2a$	$SCER$	$CM_e = SCER/(an-2a)$	
Total	$an-1$	$SCET$		

Il reste à estimer la variance (ou dispersion) résiduelle ($SCER$) et la variance (ou dispersion) due à l'interaction entre les groupes et la variable explicative x ($SCEE_i$).

Calcul de la somme des carrés résiduelle, $SCER$ Pour chaque groupe i, on estime une régression linéaire de la forme $\hat{y}_i(x_{ij}) = \hat{\mu} + \hat{\alpha}_i + \hat{\beta}_i(x_{ij} - \hat{\mu}_{x_i})$, et pour chaque régressions obtenue, on calcule l'erreur commise dans la prédiction :

$$SCER_i = \sum_{j=1}^{n}(y_{ij} - \hat{y}_i(x_{ij}))^2$$

On peut alors en déduire une erreur globale, qui est la somme des erreurs commises par chacune des a régressions :

$$SCER = \sum_{i=1}^{a} SCER_i$$

En procédant ainsi, on prend en compte à la fois l'effet de x sur Y et les différences entre les groupes, puisque les a régressions sont ajustées indépendamment pour les a groupes. La dispersion $SCER$ est donc bien la dispersion résiduelle, c'est-à -dire ce que l'on ne peut expliquer ni par l'effet de x, ni par les différences entre les groupes.

Calcul de la somme des carrés due à l'interaction, $SCEE_i$ S'il existe une interaction entre la variable x et le facteur groupe, alors l'effet de x sur Y varie d'un groupe à l'autre. Dans ce cas, les droites de régression calculées dans les a groupes n'ont pas toutes la même pente. Et si l'on contraint les a droites de régression à avoir la même pente (i.e. être parallèles) on commet une erreur dans les prédictions plus grandes que $SCER$; soit $SCER_{dp}$ cette erreur (dp pour droites parallèles). La dispersion due à l'interaction entre la variable x et le facteur groupe peut être calculée par rapport à $SCER_{dp}$:

$$SCEE_{gx} = SCER_{dp} - SCER$$

Pour calculer $SCER_{dp}$, il faut déterminer **LA** pente des a droites de régression parallèles; cette pente est en fait une sorte de moyenne des pentes $\hat{\beta}_i$. D'après les calculs réalisés pour une régression linéaire simple, on sait que $SCER = SCET(1 - R^2)$ avec $R^2 = r^2_{X,Y}$. On en déduit que $SCER = SCET - \sigma^2_{X,Y}/\sigma^2_X$. Par analogie (et on s'en contentera ici), on peut écrire :

$$SCER_{dp} = SCET - \frac{\left(\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \hat{\mu}_{x_i})(y_{ij} - \hat{\mu}_{Yi})\right)^2}{\sum_{i=1}^{a}\sum_{j=1}^{n}(x_{ij} - \hat{\mu}_{x_i})^2}$$

Test des effets

Pour les effets simple, se reporter au chapître sur l'ANOVA (effet groupe) et sur la régression linéaire (effet moyen de la variable prédictive x). Reste à tester l'interaction : ayant déterminé la part de dispersion due à l'interaction entre le facteur groupe et la variable x, on peut tester l'existence de cette interaction de la façon suivante :

Conditions de validité : les résidus sont normalement distribués et leur variance est constante.

Hypothèse nulle : il n'y a pas d'interaction, autrement dit les a droites de régression ont toutes la même pente.

Hypothèse alternative : il y a une interaction, autrement dit il y a au moins une droite de régression qui n'a pas la même pente que les autres droites.

Statistique : $f^{obs} = \frac{SCEE_{gx}/(a-1)}{SCER/(an-2a)}$.

Distribution de la statistique sous H_0 : F suit une distribution de Fisher à $a-1$ et $an-2a$ degrés de liberté.

Risque de première espèce : $p = P(F \geqslant f_{obs}|_{H_0})$

Code R : `lm` pour ajuster le modèle linéaire aux données (i.e. estimer les paramètres), suivi de `anova` pour tester les effets. Il faut aussi vérifier it a posteriori les conditions de validité du modèle construit (c.f. page 16).

Interaction et interprétation des effets simples

La figure 2.4 représente trois situations expérimentales (A, B et C) qui étudient l'effet de x sur Y dans deux groupes (en gris et noir). Le tableau ci-dessous résume les résultats de ces trois situations par $\star$ pour un effet significatif détecté ou par **n.s.** pour un effet non significatif détecté.

	cas A	Cas B	Cas C
Groupe	$\star$	n.s.	n.s.
x	$\star$	$\star$	n.s.
Groupe$\times x$	n.s.	$\star$	$\star$

L'**effet Groupe** se manifeste si, indépendamment de la relation entre x et Y, il existe une différence entre les moyennes de Y par groupe; c'est notoire pour le cas A mais pas pour les deux autres.

L'**effet** x se manisfeste si, indépendamment du groupe, x a un effet sur Y, c'est-à -dire si la pente estimée de la droite de régression entre Y et x est significativement différente de 0. Cet effet correspond

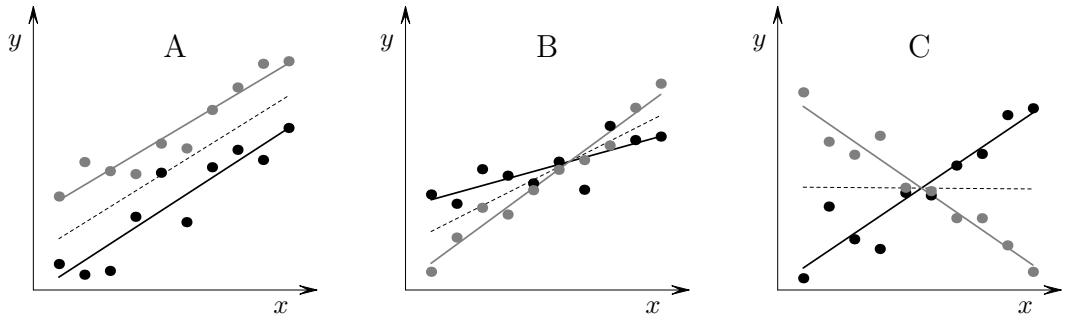


FIGURE 2.4 – En noir, les mesures prises dans la population A ; en gris, les mesures prises dans la population B. La ligne en pointillés indique la régression moyenne, l'analyse étant alors réalisée sur les données des deux populations mélangées. **A.** l'effet de x sur y est le même dans les deux populations. **B.** l'effet de x sur y est plus fort dans la population B que dans la population A. **C.** les effets de x sur y dans les deux populations sont inverses. Dans ce dernier cas, l'analyse de chaque facteur pris indépendamment n'aurait détecté aucun effet.

à la pente de la droite en pointillés dans la figure 2.4 et aux cas A et B ; dans le cas C, la pente de cette même droite est à peu près nulle et l'effet est alors non significatif.

L'**interaction Groupe**$\times x$ se manifeste si les pentes des droites de régression dans les deux groupes ne sont pas identiques. Cette interaction existe pour le cas B et surtout pour le cas C.

2.3.3 La méthode de décomposition de la variance dans le cas général

Les deux exemples précèdents montrent qu'il est possible de définir des modèles linéaires combinant plus d'un facteur explicatif et combinant des facteurs quantitatifs et qualitatifs. On peut montrer que, quel que soit le nombre et la nature des facteurs explicatifs, il est toujours possible :

1. de définir un modèle linéaire,
2. de trouver une estimation unique des paramètres de ce modèle qui minimise SCER,
3. de décomposer la dispersion totale en attribuant une part de cette dispersion à chacun des facteurs explicatifs.

Il est évidemment hors de question de calculer à la main les estimations de modèles compliqués. Il est cependant très utile de comprendre comment décomposer la dispersion totale dans n'importe quel modèle. Par exemple pour une ANOVA à deux facteurs croisés, le modèle correspondant s'écrit :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon$$

On peut définir une série de modèles emboités du plus simple au plus complexe, en ajoutant un par un les facteurs explicatifs jusqu'à obetnir le modèle complet décrit ci-dessus. A chacun de ces modèles correspond une somme des carrés des écarts résiduelle qui lui est propre.

$$\begin{array}{lll} Y_{ij} = & \mu + \epsilon_{ij} & \rightarrow SCER_0 \\ Y_{ij} = & \mu + \alpha_i + \epsilon_{ij} & \rightarrow SCER_1 \\ Y_{ij} = & \mu + \alpha_i + \beta_j + \epsilon_{ij} & \rightarrow SCER_2 \\ Y_{ij} = & \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij} & \rightarrow SCER_3 \end{array}$$

Plus un modèle comprend de facteurs explicatifs, mieux il s'ajustera aux données, et donc plus il minimisera les erreurs avec la réalité. On peut donc écrire :

$$SCER_3 < SCER_2 < SCER_1 < SCER_0$$

On peut dire que les erreurs qui ne sont plus comptabilisé en ajoutant au modèle l'effet du facteur A par rapport au modèle le plus simple font que $SCER_1 < SCER_0$. Ce "surplus" d'erreur dans le premier modèle est du au fait que le facteur A introduit explique une partie de la variance (dispersion) totale. On peut donc écrire $SCEE_A = SCER_0 - SCER_1$. De la même façon, on a $SCEE_B = SCER_1 - SCER_2$ et $SCEE_{AB} = SCER_2 - SCER_3$.

En résumé, pour calculer les dispersions expliquées par les facteurs d'un modèle linéaire, il suffit de définir une série de modèles emboîtés, du plus simple au plus complexe, et de calculer la dispersion résiduelle associée à chacun de ces modèles. La dispersion expliquée par un facteur se calcule alors comme la diminution de la dispersion résiduelle observée lorsque l'on ajoute ce facteur à un modèle plus simple. Cette méthode est applicable à n'importe quel modèle linéaire, quel que soit le nombre de facteurs explicatifs qui le composent.

Reste cependant un problème à régler : pour un même modèle linéaire, plusieurs décompositions sont possibles. Nous avons dans l'exemple précédent commencé la série de modèles en ajoutant au modèle le plus simple l'effet du facteur A ; on aurait tout aussi bien pu commencer par ajouter l'effet du facteur B et définir ainsi la série suivante

$$\begin{array}{lll} Y_{ij} = & \mu + \epsilon_{ij} & \\ Y_{ij} = & \mu + \beta_j + \epsilon_{ij} & \rightarrow SCER'_1 \\ Y_{ij} = & \mu + \beta_j + \alpha_i + \epsilon_{ij} & \\ Y_{ij} = & \mu + \beta_j + \alpha_i + \gamma_{ij} + \epsilon_{ij} & \end{array}$$

et on aurait estimé $SCEE'_A = SCER'_1 - SCER_2$ et $SCEE'_B = SCER_0 - SCER'_1$. Cette deuxième décomposition n'est équivalente à la première que si $SCER'_1 = SCER_1$.

On ne peut avoir $SCER'_1 = SCER_1$ que si les facteurs A et B sont indépendants. Si les facteurs A et B ne sont pas indépendants, une partie de la dispersion expliquée par un facteur devrait l'être par l'autre facteur. Le premier facteur introduit dans l'analyse explique alors une partie de la dispersion du second modèle. Par exemple si A et B sont corrélés, on a $SCEE_A > SCEE'_A$ et si B est introduit en premier dans le modèle, il expliquera une partie de la variance qui aurait sans cela été expliquée par A.

Si deux facteurs explicatifs sont corrélés, la forme proposé ici pour la décomposition de la variance, dite de **type I**, produit des résultats qui dépendent de l'ordre d'introduction des facteurs dans le modèle. Il faut remarquer que dans un jeu de données, deux facteurs peuvent être artificiellement corrélés si le plan d'expérience n'est pas « équilibré » (pas balancé), c'est-à-dire si les tailles d'échantillons ne sont pas les mêmes pour tous les traitements (i.e. combinaisons de facteurs), cas qui est courant en biologie. Par exemple, imaginons que les individus d'une expérience ont reçu les traitements A ou a pour un premier facteurs et B ou b pour un second facteur, et que 90 individus ont reçu les traitements A et B, 90 autres les traitements a et b alors que seulement 10 individus ont reçu les traitements A et b et 10 autres les traitements a et B. Le plan d'expérience est ici très déséquilibré et les variables décrivants les deux facteurs, deviennent, de façon artificielle non indépendantes : si l'on sait qu'un individu a reçu le traitement A, on sait aussi que la probabilité qu'il ait reçu le traitement B est forte et plus forte que la probabilité qu'il ait reçu le traitement b .

Pour contourner cette difficulté, on décomposer la variation autrement, selon le **type III**, et obtenir la série suivante de modèles à comparer :

$$\begin{array}{lll} Y_{ij} = & \mu + \beta_j + \alpha_i + \gamma_{ij} + \epsilon_{ij} & \rightarrow SCER_0 \\ Y_{ij} = & \mu + \beta_j + \alpha_i + \epsilon_{ij} & \rightarrow SCER_1 \\ Y_{ij} = & \mu + \beta_j + \epsilon_{ij} & \rightarrow SCER_2 \\ Y_{ij} = & \mu + \alpha_i + \epsilon_{ij} & \rightarrow SCER_3 \end{array}$$

Cette fois-ci la série est définie en partant du modèle complet et en enlevant les facteurs un à un et on peut calculer les dispersions expliquées en posant

$$\begin{array}{rl} SCEE_A = & SCER_2 - SCER_1 \\ SCEE_B = & SCER_3 - SCER_1 \\ SCEE_{AB} = & SCER_1 - SCER_0 \end{array}$$

La décomposition ne dépend plus, cette fois-ci, de l'ordre d'introduction des facteurs explicatifs.

Ces deux types I et II de décomposition sont possibles dans R, le type I étant appliqué par défaut.

Il faut retenir :

- Aucune des deux décompositions est plus fausse ou plus juste que l'autre ; les décompositions de type I et II testent des hypothèses nulles différentes. Il faut déterminer laquelle de ces hypothèses nulles correspond le mieux à la question biologique posée.
- Lorsque deux facteurs sont très corrélés, il peut devenir illusoire de chercher à tester leurs effets séparément. Puisque ces deux facteurs mesurent la même chose, autant n'en mettre qu'un seul dans le modèle.

Chapitre 3

Les écarts aux hypothèses du modèle linéaire

3.1 Les écarts aux hypothèses de normalité, d'indépendance et d'homoscédasticité

3.1.1 Inspection des résidus

Un modèle linéaire peut s'écrire sous la forme $Y_i = \hat{y}_i + \epsilon_i$. L'hypothèse principale faite porte sur la variable aléatoire ϵ_i. Cette variable doit être normalement distribuée et les ϵ_i doivent être indépendants et avoir la même variance pour que le modèle soit bien ajusté aux données.

Deux types d'outils permettent de détecter un écart à chacune de ces trois hypothèses : des outils graphiques et des tests statistiques. Tous ces outils permettent d'étudier, les comportements des résidus estimés après que le modèle ait été ajusté.

La réalisation des tests de normalité et d'homoscédasticté sous R a été présentée brièvement dans le chapitre précédent (cf. p.16). L'analyse graphique des résidus consiste, entre autres, à représenter les valeurs des résidus (les ϵ_i) en fonction des valeurs prédites par le modèle (les $\hat{y}_i$). La figure 3.1 présente quelques uns de ces graphiques pour différentes régressions linéaires. Le premier graphique indique que les résidus respectent les conditions de validité du modèle construit ; les deux autres correspondent à des cas problèmatiques qu'il faut savoir détecter et interpréter.

3.1.2 Les remèdes

Le modèle est-il bien construit ?

Si un écart à la normalité ou bien à l'homoscédasticité des résidus d'un modèle linéaire a été détecté, il faut avant toute chose, particulièrement avant de supposer que les données ont une distribution bizarre, se demander si cet écart peut être du à l'inadéquation du modèle construit. Par exemple, il faut se demander s'il ne manque pas dans le modèle un facteur explicatif ayant un effet potentiel sur la variable analysée.

Par exemple, considérons une mesure réalisée dans deux populations dont la distribution est normale dans chacune des deux populations. La variance de cette distribution est 1 dans les deux populations, mais la moyenne est 10 dans la première population et 20 dans la deuxième. On peut écrire un premier modèle qui prédit pour chaque observation leur moyenne générale qui vaut 15

$$y_{ij} = \mu_Y + \epsilon_{ij}$$

La distribution des résidus de ce modèle est représentée dans la figure 3.2A. Comme la moyenne de la première population vaut 10, la moyenne des résidus pour cette population est -5 ; similairement, la moyenne des résidus pour la deuxième population vaut 5. A cause de cette différence de moyennes, la distribution des résidus est bimodale et un test de Shapiro concluera logiquement à un écart significatif de la distribution des résidus à la normalité.

Pour améliorer les prédictions du modèle on réalise une ANOVA à un facteur de classification :

$$y_{ij} = \mu_Y + \alpha_i + \epsilon_{ij}$$

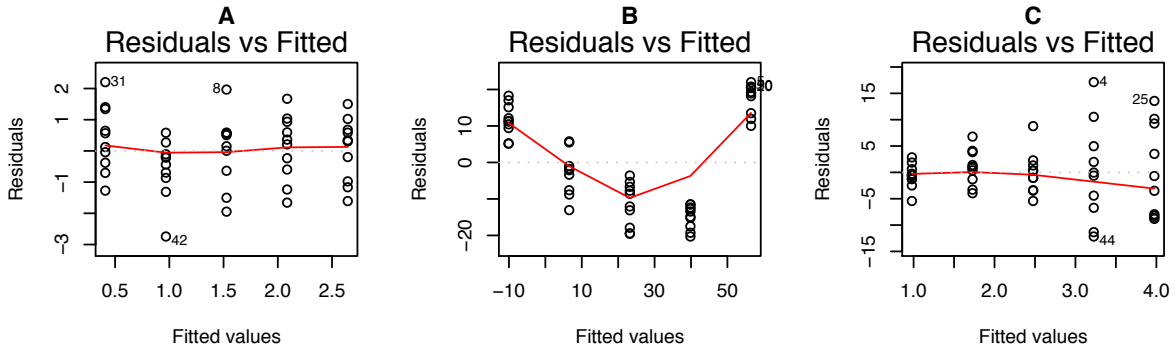


FIGURE 3.1 – Représentation des résidus d'une régression linéaire. A. Cas « normal » : les conditions de validité du modèle linéaire sont satisfaites. B. Le facteur explicatif a un effet non linéaire sur la mesure. Le modèle est donc mal ajusté ; ici les résidus ne sont ni normalement distribués ni indépendants (c.f. la figure 3.3). C. La variance résiduelle augmente nettement avec les valeurs prédites ; l'hypothèse d'homoscédasticité n'est donc pas satisfaite.

La prédiction pour chaque observation est la moyenne de la population dans laquelle l'observation a été réalisée. Avec ce nouveau modèle les moyennes des résidus dans les deux populations sont toutes les deux nulles, la distribution des résidus est unimodale et un test de Shapiro ne rejette pas l'hypothèse nulle de leur normalité. En conclusion, le premier modèle ne remplit pas les conditions de normalité des résidus parce qu'il n'inclut pas le facteur population.

Considérons maintenant le lien entre une variable Y et une variable explicative quantitative x. L'outil pour réaliser cette d'étude est la régression linéaire :

$$y_i = \mu_y + \beta(x - \mu_x) + \epsilon_i$$

Si le lien entre x et Y est en fait exponentiel, on peut alors écrire :

$$y_i = e^{\mu_y + \beta(x-\mu_x)+\epsilon_i}$$

Le premier modèle permet peut être de détecter le lien entre x et Y, mais il produit des résidus non normaux. La transformation de Y par la fonction logarithme, en rétablissant la linéarité de la relation avec x, permet d'obtenir un lien linéaire entre $\log(x)$ et $\log(Y)$ et des résidus normaux. Les résidus produits par chacun de ces deux modèles sont représentés dans la figure 3.3.

À partir des deux exemples ci-dessus, on peut formuler une série de questions qui permet de trouver l'origine possible d'un écart à la normalité, l'indépendance, à l'homoscédasticité des résidus d'un modèle linéaire :

1. Certains effets simples ont-ils été oubliés dans le modèle ?
2. Certaines interactions ont-elles été oubliées dans le modèle ?
3. Dans le cas d'une facteur explicatif quantitatif, le lien entre ce facteur et la variable analysée est-il correctement modélisé ? Peut-on améliorer la qualité du modèle en changeant le lien linéaire en un un lien log, en ajoutant un effet quadratique, etc. ?

Transformation des variables

Il est parfois impossible de corriger un écart à la normalité en modifiant la structure du modèle parce que (1) la variable qu'il faudrait ajouter au modèle pour corriger le problème demeure inconnue, ou (2) la distribution des résidus n'est réellement pas normale, même avec un modèle correctement construit. On peut tenter de résoudre le problème en transformant les données avant de les analyser ; on espère alors que la distribution de la variable transformée est plus proche d'une distribution normale que la distribution de la variable d'origine. Beaucoup de transformations sont possibles, la transformation de de Box-Cox est une des plus flexible :

$$\begin{cases} z = \dfrac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ z = \log(y) & \lambda = 0 \end{cases}$$

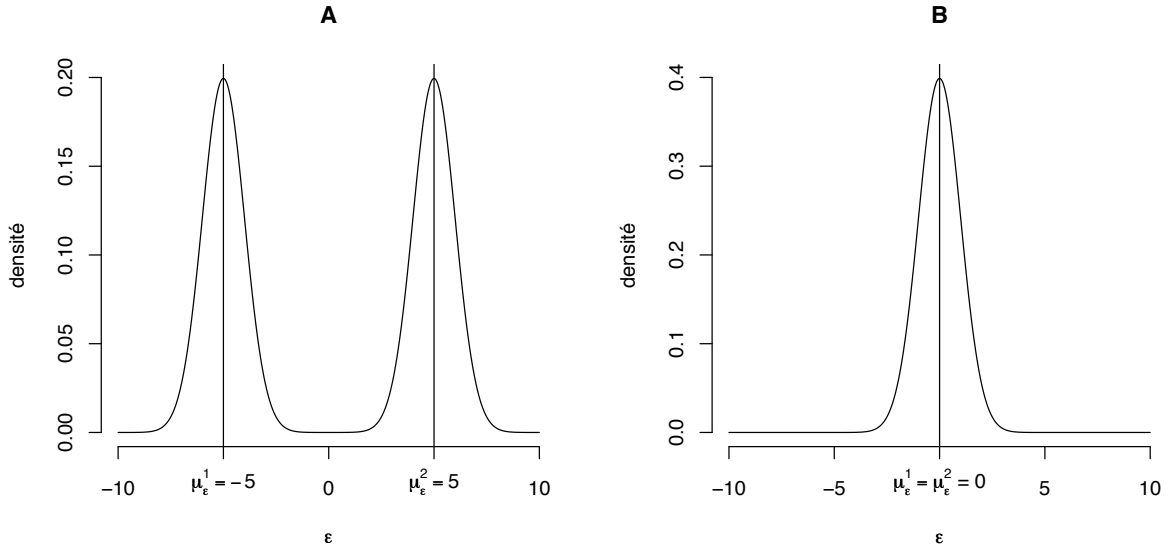


FIGURE 3.2 – A. Distribution des résidus du modèle $y_{ij} = \mu_Y + \epsilon_{ij}$; dans la population 1, la moyenne des ϵ_{ij} est -5, dans la population 2, elle vaut 5. Même si dans chaque population la distribution de ϵ est normale, la distribution globale de ϵ ne l'est pas. B. Distribution des résidus du modèle $y_{ij} = \mu_Y + \alpha_i + \epsilon_{ij}$ où α_i est l'écart entre la moyenne générale et la moyenne de la population i ($\mu_Y^i = \mu_Y + \alpha_i$). Les résidus à l'interieur au sein de chaque population sont centrés en zéro et leur distribution est unimodale et normale.

avec λ un paramètre à déterminer. La démarche à suivre pour construire et ajuster le modèle est la suivante :

1. On ajuste le modèle linéaire sur les données non transformées. On détecte un écart des résidus à la normalité, à l'indépendance ou à l'homoscédascitité.
2. On cherche la transformation Box-Cox la plus adéquate, c'est-à-dire la valeur du paramètre λ, permettant d'obtenir la normalitéd es résidus.
3. On transforme les données en utilisant le paramètre λ déterminé à l'étape précédente.
4. On ajuste le nouveau modèle linéaire sur les données ainsi transformées. On vérifie que les résidus du nouveau modèle satisfont les hypothèses de normalité, d'indépendance, et d'homoscédasticité des résidus car rien ne garantit que la transformation choisie est efficace.

Dans R un outil permettant d'estimer λ est proposé dans la bibliothèque MASS dont voici un exemple :

```
m  <- lm(y ~ x)

library(MASS)
bc <- boxcox(m,lambda=seq(-2,2,length=200))
lambda <- bc$x[which.max(bc$y)]

z <- (y^lambda-1)/lambda
m <- lm(z ~ x)
```

Le modèle linéaire généralisé : un bref aperçu

On sait parfois *a priori* que les données à analyser ne peuvent pas être normalement distribuées, quelle que soit la transformation choisie. Considérons par exemple que, avant une autorisation de mise sur le marché, on teste les effets toxiques d'un produit phytosanitaire en l'injectant à des rats. Le protocole est le suivant : des lots de vingts rats reçoivent une dose fixée de toxine, cette dose variant de zéro à une dose permettant de tuer tous les rats pendant la durée de l'expérience. On cherche à estimer la dose qui permet de tuer 50% des rats en un temps donné (i.e. la DL50). Les données correspondent alors au nombre de rats morts et au nombre de rats survivants en fonction d'une dose injectée.

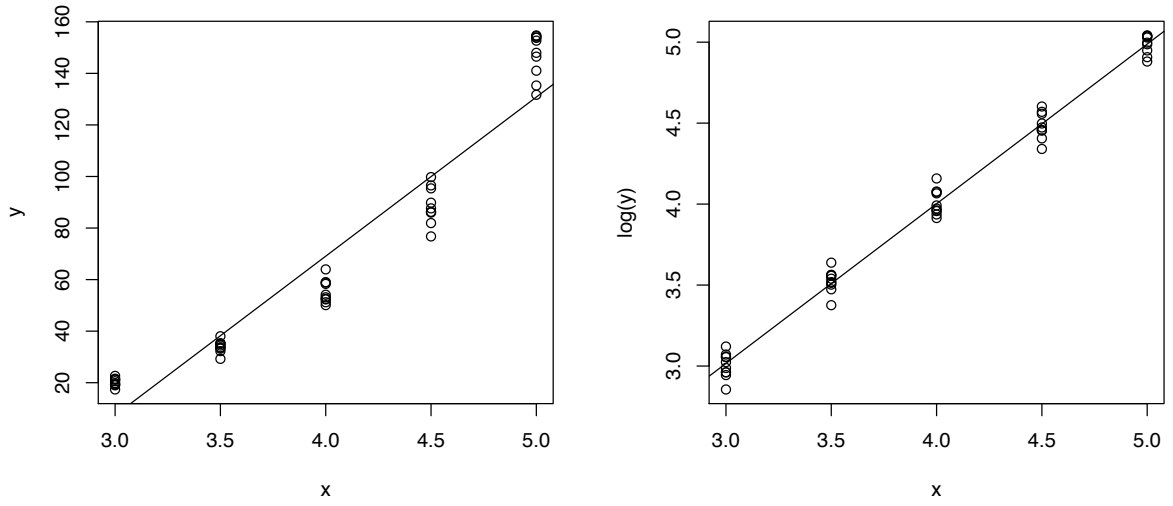


FIGURE 3.3 – A. Résidus d'une régression linéaire lorsque le lien entre les deux variables quantitatives est exponentiel. La valeur de Y prédite est trop faible pour les valeurs extrêmes de x et trop forte pour les valeurs intermédiaires. Les résidus sont alors systématiquement positifs pour les valeurs extrêmes de x et systématiquement négatifs pour les valeurs intermédiaires. B. Résidus d'une régression linéaire où les valeurs observées de Y ont été transformées par la fonction logarithme. Les valeurs prédites sont correctement ajustées et les résidus sont centrés en zéro.

Ce type de données (nombre de succès et d'échecs pour un nombre donné de tirages) est décrit par la distribution binomiale de paramètres N le nombre de tirages et p la probabilité de succès ($0 \leqslant p \leqslant 1$). La moyenne de la distribution est Np et sa variance $Np(1-p)$. Il est flagrant que l'hypothèse d'homoscédasticité n'est pas vérifiée pour ce type de distribution, puisque la moyenne et la variance de la distribution changent avec p. D'autre part, si p est faible, le nombre de succès est petit mais toujours positif. La distribution est donc fortement assymétrique et très différente d'une distribution normale.

Il est cependant possible d'écrire un modèle linéaire pour ce type de données ou de distribution en spécifiant explicitement la distribution des données est binomiale : on construit alors un modèle linéaire généralisé (GLM) ; c'est-à-dire un modèle linéaire étendu à d'autres distributions que la distribution normale.

Pour les données d'intoxication des rats, on aurait pu écrire un modèle sous la forme :

$$\hat{p} = \alpha + \beta x$$

avec x la dose de toxine injectée et $\hat{p}$ une prédiction de la probabilité de mourir pour cette dose. La quantité $\hat{p}$ ne peut cependant pas être négative ni dépasser 1. Or la relation linéaire $\alpha + \beta x$ ne garantit pas ces deux contraintes : si x est très élevé et β positif, on peut prédire des probabilités supérieur à 1 ; inversement, si α est négatif et x est petit, on peut prédire des probabilités négatives. Pour garantir que les prédictions ont un sens biologique, on construit plutôt le modèle suivant :

$$g(\hat{p}) = \alpha + \beta x$$

avec g une fonction de lien. Cette fonction est choisie de telle sorte que $g(p)$ prenne des valeurs entre $-\infty$ à $+\infty$ lorsque p varie de 0 à 1. Ainsi, on est assuré que $\hat{p} \in [0, 1]$. Le plus souvent dans le cas des données binomiales, cette fonction de lien est la fonction logit ($\text{logit}(p) = \log(p/(1-p))$). En résumé, pour ajuster un modèle linéaire généralisé, il faut connaître *a priori* la distribution des données et choisir une fonction de lien. Le traitement détaillé du modèle linéaire généralisé dépasse les objectifs de ce fascicule ; il est fourni par des ouvrages spécialisés (par exemple, The R Book de M. Crawley ou bien Generalized Linear Models McCullagh et Nelder, certains aspects seront vu plus en détail dans GMBE11C).

Pour les données de mortalité des rats en fonction d'une dose de toxine, le code R permettant d'ajsuter le modèle et de l'analyser est le suivant :

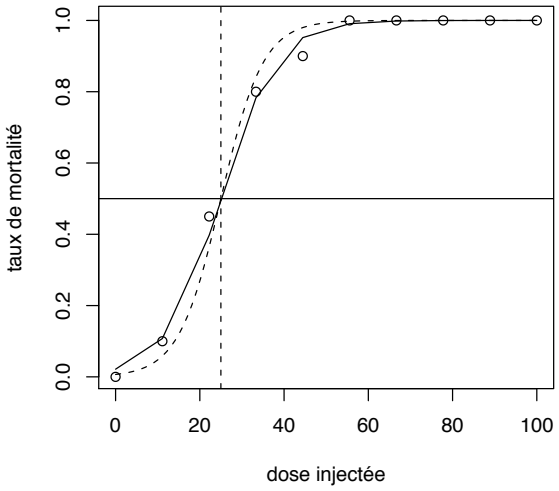


FIGURE 3.4 – Mortalité des rats en fonction de la dose de toxine injectée (c.f. texte pour plus de détails). Chaque dose a été injectée à vingt rats . Les points représentent la proportion de rats morts suite à l'injection. La courbe en pointillé représente la relation utilisée entre la dose injectée et la mortalité pour simuler ces données et la courbe en trait plein représente la relation estimée par un modèle linéaire généralisé (GLM). La droite verticale en pointillé correspond à la DL50.

```
m <- glm(cbind(mort,vivant) ~ dose,family=binomial(link=logit))
> summary(m)
Call:
glm(formula = cbind(mort, vivant) ~ dose, family = binomial(link = logit))
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95184  -0.07683   0.07927   0.24127   0.60669
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.81681    0.71429  -5.344 9.12e-08 ***
dose         0.15301    0.02579   5.932 2.99e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 161.7192  on 9  degrees of freedom
Residual deviance:   2.5020  on 8  degrees of freedom
AIC: 18.028

Number of Fisher Scoring iterations: 6
```

Il faut ici remarquer que la déviance résiduelle (l'équivalent de la SCER du modèle linéaire) ne dépasse pas le nombre de degrés de liberté résiduels. Si tel était le cas, la distribution choisie ne serait pas la bonne car la déviance résiduelle est plus grande que celle attendue sous l'hypothèse d'une distribution binomiale des données.

Pour les cas désespérés : les outils non-paramétriques

Lorsque les données ne peuvent pas être transformées et qu'aucune des distributions permises par le GLM ne semble les décrire correctement, il reste deux solutions :

1. Appliquer des méthodes non-paramétriques qui ne font pas l'hypothèse de la normalité ou d'homoscédasticité des résidus. La limite de cette approche est le nombre réduit d'outils disponibles.

2. Construire son propre outil statistique, par exemple en utilisant des méthodes de permutation ou bien en construisant un modèle dont les paramètres sont ajusté par maximum de vraisemblance. Beaucoup d'outils informatiques sont maintenant disponibles pour permettre cela démarche, mais dépasse le cadre de notre cours.

3.2 Les écarts à l'hypothèse d'indépendance

3.2.1 Un premier cas de pseudo-réplication

Considérons le lien entre le poids et la taille. La taille est mesurée une seule fois pour dix individus, mais le poids d'un individu étant susceptible de changer au cours du temps, est mesuré cinq fois pour chaque individu. On étudie la relation entre le poids et la taille en écrivant le modèle suivant :

$$Y_{ij} = \mu_Y + \beta(x_{ij} - \mu_x) + \epsilon_{ij}$$

avec x_{ij} la taille de l'individu i et Y_{ij} la mesure j du poids de l'individu i. L'analyse de ce modèle produit le tableau d'Anova présentée dans le tableau 3.1. Les données et les résidus du modèle sont présentés dans la figure 3.5.

Dans le modèle écrit ci-dessus chaque mesure est considérée indépendante des autres ; il y a donc 48 $(= 50 - 1 - 1)$ degrés de liberté résiduels. Or, le graphique des résidus (figure 3.5) montre clairement que l'hypothèse d'indépendance n'est pas valide car des paquets de points sont observables alors qu'aucune stucture particulière ne devrait être détectée. Les données sont dites pseudo-répliquées parce que deux mesures réalisées sur un même individu sont plus semblables entre elles que deux mesures réalisées sur deux individus différents, et puisque que le modèle construit ne prend pas en compte le facteur individu, ses résidus ne sont pas alors indépendants.

source	d.d.l.	SCE	CM	F
taille	1	$\text{SCEE}_{\text{taille}}$	$\text{CME}_{\text{taille}} = \frac{\text{SCEE}_{\text{taille}}}{1}$	$\frac{\text{CME}_{\text{taille}}}{\text{CMR}}$
résiduelle	48	SCER	$\text{CMR} = \frac{\text{SCER}}{48}$	

TABLE 3.1 – Tableau de l'ANOVA réalisée sur les données brutes.

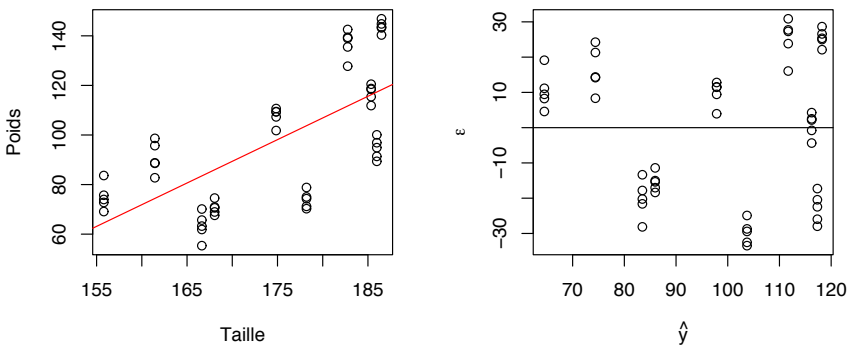


FIGURE 3.5 – Poids en fonction de la taille (graphique de gauche) et résidus du modèle (graphique de droite).

source	d.d.l.	SCE	CM	F
taille	1	$\text{SCEE}_{\text{taille}}$	$\text{CME}_{\text{taille}} = \frac{\text{SCEE}_{\text{taille}}}{1}$	$\frac{\text{CME}_{\text{taille}}}{\text{CMR}}$
résiduelle	8	SCER	$\text{CMR} = \frac{\text{SCER}}{8}$	

TABLE 3.2 – Tableau de l'ANOVA réalisée sur les poids moyens.

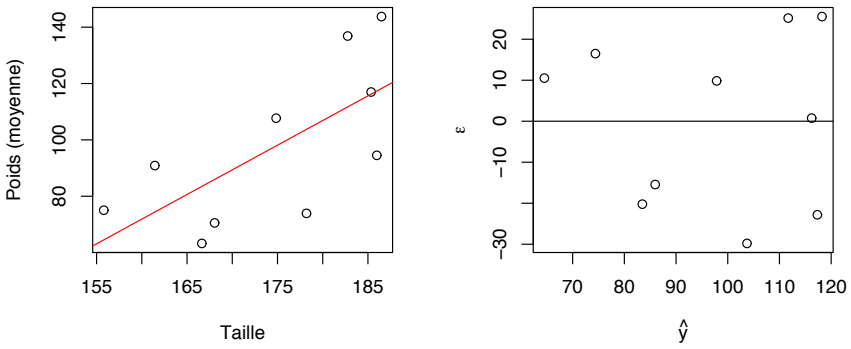


FIGURE 3.6 – Poids moyen d'un individu en fonction de sa taille (graphique de gauche) et résidus du modèle (graphique de droite).

Pour améliorer le modèle, on peut écrire une relation entre la taille et les moyennes individuelles du poids ;

$$Y_i = \mu_Y + \beta(x_i - \mu_x) + \epsilon_i$$

avec Y_i le poids moyen d'un individu et x_i la taille d'un individu. L'analyse de ce modèle produit le tableau d'Anova présentée dans le tableau 3.2. Les données et les résidus du modèle sont présentés dans la figure 3.6.

Dans ce nouveau modèle, seules les mesures réalisées sur des individus différents sont considérées comme indépendantes. Le nombre de degrés de liberté résiduels devient maintenant 8 $(= 10 - 1 - 1)$ et les résidus ne présentent plus de structure qui révélerait un écart à l'hypothèse de leur indépendance.

Un risque fréquent dans l'analyse de données pseudo-répliquées est de sur-estimer le nombre de degrés de liberté résiduels parce que le facteur de pseudoréplication n'est pas pris en compte dans le modèle. Cette sur-estimation engendre une sous-estimation du carré moyen résiduel et donc une sur-estimation de la statistique F permettant de tester les effets des facteurs du modèle. Par conséquent, on risque de déclarer comme significatif des effets qui ne le sont pas.

Le modèle proposé ci-dessus, puisqu'il établit un lien entre la moyenne individuelle du poids et la taille, ne prend pas en compte la variation intra-individuelle de poids. Cette variation « intra-individuelle » provient en partie des erreurs de mesures et des fluctuations environnementales du poids d'un individu en fonction du temps. Une autre partie de la variation du poids provient de différence existantes entre les individus, un individu pouvant s'écarter significativement de la relation établie entre le poids et la taille. Autrement dit, il existe des grands maigres et des petits gros. Cette variation contribue à la variation « inter-individuelle ». Le phénomène de pseudo-réplication apparait car deux mesures réalisées sur deux individus différents ont plus de chance d'être différentes entre elles que deux mesures réalisées sur un même individu.

On peut écrire un modèle qui prend en compte les deux sources de variation, intra et inter-individuelles :

$$Y_{ij} = \mu_Y + A_i + \beta(x_i - \mu_x) + \epsilon_{ij}$$

avec A_i l'écart entre le poids moyen observé d'un individu et la valeur prédite d'après sa taille. La quantité A_i est une variable aléatoire normalement distribuée centrée en 0 et de variance σ^2_{ind}. La variance σ^2_{ind}

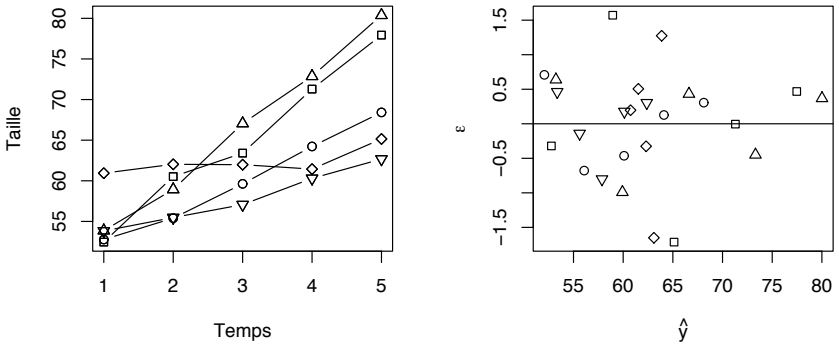


FIGURE 3.7 – Courbe de croissance de cinq individus (graphique de gauche). Résidus d'une analyse de ces données par une ANCOVA (graphique de droite). Les symboles différents dans le graphique de droite correspondent aux individus du graphique de gauche.

des A_i correspond à la variation « inter-individuelle », alors que la variance σ^2 des résidus ϵ_{ij} correspond à la variation « intra-individuelle ». Le modèle ainsi écrit combine un effet fixe, l'effet de la taille sur le poids quantifié par β, et un effet aléatoire, l'effet d'un individu décrit par la variable aléatoire A_i.

Les résidus du tout premier premier modèle construit comprenaient les deux sources de variance, intra et inter-individuelle ; c'est pourquoi ils n'étaient pas indépendants. En outre, la variance des résidus du deuxième modèle fondé sur les moyennes du poids correspond uniquement à la variance « inter-individuelle » ; cette variance est donc identique à la variance des A_i du troisième modèle mixte incluant un effet fixe et un effet aléatoire. Le modèle sur les moyennes et le modèle mixte permettent d'estimer correctement la variation « inter-individuelle ». Dans le modèle sur les moyennes, l'effet de la taille est testé en comparant la variance de poids due aux variations de la taille à la variance résiduelle (cf. tableau 3.2), correspondant à la variation « inter-individuelle ». Pour réaliser un test équivalent dans le modèle mixte, il faut comparer la variation expliquée par le facteur taille à la variation des A_i. Cet aspect de la décomposition de la variance est détaillé dans le paragraphe suivant.

3.2.2 Un deuxième cas de pseudo-réplication

Le problème

Etudions maintenant la courbe de croissance de cinq individus pour estimer s'ils ont grandi. Les données sont représentées par la figure 3.7.

On pourrait écrire un modèle qui ignore le facteur individu :

$$Y_{ij} = \mu_Y + \beta(x_j - \mu_x) + \epsilon_{ij}$$

avec Y_{ij} la mesure de taille de l'individu i à l'âge j. Comme précédemment les mesures obtenues sur un même individu ne sont pas indépendantes. Contrairement à l'exemple précédent, le modèle ne peut pas être fondé sur les valeurs moyennes individuelles, puisque le but de l'étude est d'expliquer les variations de taille en fonction de l'âge pour n'importe quel individu. Pour prendre en compte le problème de pseudoréplication, il faut écrire un modèle qui prend en compte les différences entre les individus :

$$Y_{ij} = \mu_Y + \alpha_i + \beta(x_{ij} - \mu_{x_i}) + \gamma_i(x_{ij} - \mu_{x_i}) + \epsilon_{ij}$$

où α_i quantifie l'effet du facteur individu, β celui du facteur temps et γ_i l'interaction entre les deux précédents facteurs. Le modèle construit correspond à une ANCOVA classique, avec tous les facteurs explicatifs considérés comme des effets fixes. Le tableau d'Anova produite par le modèle est présentée dans le tableau 3.3.

Comme le modèle prend en compte les différences entre les individus, il s'ajuste correctement aux données. Notamment, les résidus du modèle ne présentent pas d'écart significatif aux hypothèses habituelles du modèle linéaire, en particulier l'hypothèse de leur indépendance. Il reste néanmoins un problème lié à la pseudo-réplication et provient de la façon dont l'effet du temps est décrit et testé.

source	d.d.l.	SCE	CM	F
individu	4	SCEE_{ind}	$\text{CME}_{\text{ind}} = \frac{\text{SCEE}_{\text{ind}}}{4}$	$\frac{\text{CME}_{\text{ind}}}{\text{CMR}}$
temps	1	$\text{SCEE}_{\text{temps}}$	$\text{CME}_{\text{temps}} = \frac{\text{SCEE}_{\text{temps}}}{1}$	$\frac{\text{CME}_{\text{temps}}}{\text{CMR}}$
interaction	4	$\text{SCEE}_{\text{inter}}$	$\text{CME}_{\text{inter}} = \frac{\text{SCEE}_{\text{inter}}}{4}$	$\frac{\text{CME}_{\text{inter}}}{\text{CMR}}$
résiduelle	15	SCER	$\text{CMR} = \frac{\text{SCER}}{15}$	

TABLE 3.3 – Tableau de l'ANCOVA réalisée sur les données de croissance.

Dans une ANCOVA classique, l'effet temps est testé en calculant le rapport entre le carré moyen expliqué par le facteur temps et le carré moyen résiduel (c.f. tableau 3.3). Autrement dit, on considère que si les résidus étaient nuls (i.e. aucune erreur de prédiction n'est commise), on aurait une estimation parfaite de β, qui mesure l'effet du temps sur la taille ; seule la variance résiduelle contribue au bruit de fond des données. [1]

On veut tout de même évaluer la robustesse de l'estimation de β, en particulier si d'autres individus étaient échantillonés. Il est certain que ces nouveaux individus n'auraient pas tout à fait la même vitesse de croissance que ceux déjà mesurés et seraient donc caractérisés par des valeurs de γ_i différentes. Ces différences auraient logiquement un impact sur l'estimation de β. Donc, même si les résidus ϵ_{ij} étaient nuls, l'estimation de β serait associée à une erreur parce qu'elle est liée à un échantillon aléatoire d'individus. *Il faut alors considérer toute variation due aux différences de croissance entre les individus comme du bruit.* Le modèle décrivant cette situation s'écrit :

$$Y_{ij} = \mu_Y + A_i + \beta(x_{ij} - \mu_{x_i}) + G_i(x_{ij} - \mu_{x_i}) + \epsilon_{ij}$$

avec A_i et G_i deux variables aléatoires normalement distribuées correspondant respectivement aux écarts de la taille et de la vitesse de croissance dus aux différences entre les individus. Le modèle construit est mixte puisqu'il comprend un effet fixe (β) et deux effets aléatoires (A_i et G_i).

Comment tester l'effet du temps ?

Deux approches sont possibles pour tester l'effet du temps dans le modèle mixte. La plus simple, et la plus classique, est de comparer la variance expliquée par le facteur temps à la variance due aux différences de vitesse de croissance entre les individus, puisque le bruit auquel on veut comparer le facteur temps est produit par des différences de croissance entre les individus :

$$F = \frac{\text{CM}_{\text{temps}}}{\text{CM}_{\text{inter}}}$$

Il faut bien comprendre ici que si l'on compare la variance expliquée à la variance résiduelle on ne peut conclure que pour les individus étudiés. Si on considère au contraire que les différences entre les individus sont de l'ordre du bruit, en calculant la statistique F comme indiqué ci-dessus, on pourra aussi conclure quant aux autres individus de la population qui n'ont pas été échantillonnés. Par ailleurs, il est for probable que la variance inter-individuelle est plus forte que la variance résiduelle et on a :

$$\frac{\text{CM}_{\text{temps}}}{\text{CM}_{\text{inter}}} < \frac{\text{CM}_{\text{temps}}}{\text{CMR}}$$

Autrement dit, un effet significatif dans un modèle ne comprenant que des effets fixes (c'est-à-dire où l'on compare le carré moyen expliqué au carré moyen résiduel, c.f. le tableau 3.3) peut ne plus l'être dans un modèle mixte (et où la référence est le carré moyen de l'interaction). Le test de l'effet âge dans les données présentées dans la figure 3.7 produit par exemple un risque de première espèce de $0,024$ si l'on

1. On fait ici l'hypothèse que le modèle est correctement ajusté aux données, si bien qu'il ne reste que du bruit dans la variance résiduelle

compare le carré moyen expliqué par le facteur âge au carré moyen de l'interaction, alors que le risque est de $1,793 \times 10^{-14}$ si on compare le carré moyen expliqué par ce facteur au carré moyen résiduel.

Sous R, il est possible de changer la façon dont les F de la table d'Anova sont calculés en utilisant la fonction `aov` et l'option `error` dans la formule ; pour les données de croissance la syntaxe est :

```
> m <- aov(taille ~ temps + Error(individu/temps))
> summary(m)
```

Une autre méthode, plus récente, existe dans deux bibliothèques différentes de R : `nlme` et `lme4`. Les estimations des paramètres des modèles sont obtenues en maximisant la vraisemblance des données sous l'hypothèse du modèle. `lme4` propose également un outil puissant qui permet de spécifier des effets aléatoires lorsque les données ne suivent pas une distribution normale [2]. Les résultats ne peuvent pas se comparer directement à ceux obtenus par `aov` et les détails des analyses dépassent le cadre de ce fascicule (mais peuvent être trouvé sur le web ou dans l'ouvrage Mixed-Effects Models in S and S-PLUS de Pinheiro et Bates). Voici néanmoins les syntaxes pour l'une et l'autre de ces bilbliothèques :

```
> library(nlme)
> m <- lme(fixed=taille ~ temps,random=~1+temps|individu)
> anova(m)

> library(lme4)
> m <- lmer(taille ~ temps+(1+temps|individu))
> summary(m)
```

La notation `1+temps|individu` signifie que le facteur aléatoire `individu` a un effet sur l'ordonnée à l'origine du modèle (le `1` à gauche du `|`, qui correspond à la variable aléatoire A_i du modèle) et sur la pente (le `temps` à gauche du `|`, qui correspond à la variable aléatoire G_i du modèle). L'aide en ligne de R fournit beaucoup de détails sur l'utilisation de ces deux fonctions (vous pouvez également lire `http:www.r-project.org/doc/Rnews/Rnews_2005-1.pdf`, pages 27 à 30).

3.3 Résumé des écarts aux modèles linéaires

1. *Il existe deux problèmes distincts lorsque les données sont pseudo-répliquées*

   Si le facteur pseudo-répliqué n'est pas pris en compte dans le modèle

   le modèle s'ajuste très mal aux données, les résidus ne sont pas indépendants et le nombre de degrés de liberté associé aux résidus est très sur-estimé. On risque alors de déclarer comme significatif des effets qui ne le sont pas, mais le problème est facilement détecté par l'examen des résidus.

   Si le facteur pseudo-répliqué est pris en compte dans le modèle

   le modèle s'ajuste correctement aux données et l'examen des résidus ne révèle aucun problème flagrant. On risque cependant de déclarer comme significatif des effets qui ne le sont pas, parce que la variation entre les différentes modalités du facteur pseudo-répliqué n'est pas considérée comme du bruit. Pour résoudre ce problème, il faut utiliser les méthodes décrites ci-dessus.

2. *Quand doit-on s'attendre à la pseudo-réplication ?*

   Mesures répétées comme pour le premier exemple de ce chapitre, et c'est probablement le cas le plus simple à traiter,

   Séries temporelles comme pour le second exemple de ce chapitre. En plus des méthodes usuelles présentées ici, la fonction `lme` permet de prendre en compte explicitement l'autocorrélation temporelle qui peut subsister dans les résidus.

   Données spatialisées lorsque des populations plus ou moins proches géographiquement sont échantilonnées, on s'attend àce que deux populations proches se ressemblent plus que deux populations éloignées. Pour tirer des conclusions qui ne sont pas restreintes aux seules populations étudiées, il faut prendre en compte explicitement la distance entre les populations dans le modèle. La fonction `lme` permet de prendre en compte l'autocorrélation spatiale qui peut exister dans les résidus.

2. Dans la littérature, on fait référence à ce type d'analyse par l'acronyme GLMM, Modèle Linéaire Généralisé Mixte.

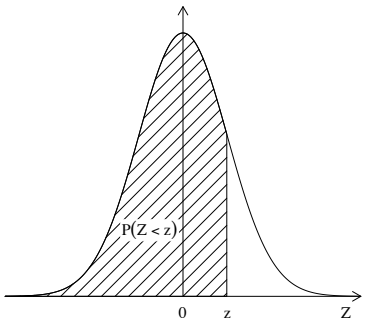
Données phylogénétiques lorsque les mesures réalisées peuvent, selon la vitesse d'évolution du trait mesuré, être plus semblables entre des espèces phylogénétiquement proches qu'entre des espèces plus distantes. La fonction `gls` de la bibliothèque `nlme` en utilisant les outils de la bibliothèque `ape` permet de considérer l'autocorrélation due à la phylogénie (c.f. l'aide de la fonction `corClasses` dans `ape`. D'autres méthodes comme les contrastes indépendants avec la fonction `pic` sont également disponibles dans cette bibliothèque).

La liste ci-dessus n'est pas exhaustive.

Chapitre 4

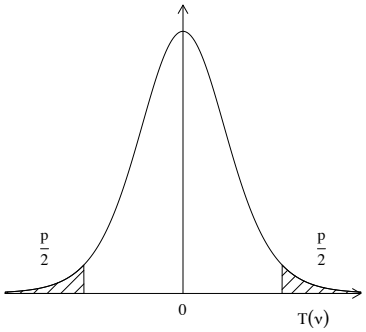
Tables statistiques

4.1 Fonction de répartition de la loi normale centrée réduite



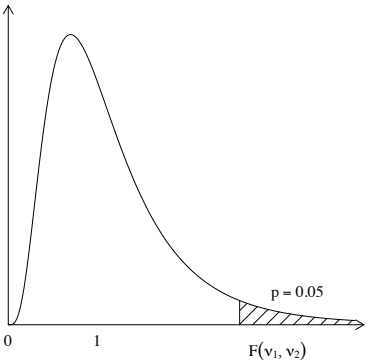
z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0,5	0,504	0,508	0,512	0,516	0,5199	0,5239	0,5279	0,5319	0,5359
0.1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0.2	0,5793	0,5832	0,5871	0,591	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0.3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,648	0,6517
0.4	0,6554	0,6591	0,6628	0,6664	0,67	0,6736	0,6772	0,6808	0,6844	0,6879
0.5	0,6915	0,695	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,719	0,7224
0.6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0.7	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0.8	0,7881	0,791	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0.9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,834	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1.1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,877	0,879	0,881	0,883
1.2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,898	0,8997	0,9015
1.3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1.4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1.5	0,9332	0,9345	0,9357	0,937	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1.6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1.7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1.8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1.9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,975	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2.1	0,9821	0,9826	0,983	0,9834	0,9838	0,9842	0,9846	0,985	0,9854	0,9857
2.2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,989
2.3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2.4	0,9918	0,992	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2.5	0,9938	0,994	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2.6	0,9953	0,9955	0,9956	0,9957	0,9959	0,996	0,9961	0,9962	0,9963	0,9964
2.7	0,9965	0,9966	0,9967	0,9968	0,9969	0,997	0,9971	0,9972	0,9973	0,9974
2.8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,998	0,9981
2.9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

4.2 Loi de Student



							p						
ν	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.02	0.01	0.001
1	0,1584	0,3249	0,5095	0,7265	1	1,3764	1,9626	3,0777	6,3138	12,7062	31,8205	63,6567	636,6192
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,92	4,3027	6,9646	9,9248	31,5991
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534	3,1824	4,5407	5,8409	12,924
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,941	1,1896	1,5332	2,1318	2,7764	3,7469	4,6041	8,6103
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,015	2,5706	3,3649	4,0321	6,8688
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432	2,4469	3,1427	3,7074	5,9588
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,896	1,1192	1,4149	1,8946	2,3646	2,998	3,4995	5,4079
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595	2,306	2,8965	3,3554	5,0413
9	0,1293	0,261	0,3979	0,5435	0,7027	0,8834	1,0997	1,383	1,8331	2,2622	2,8214	3,2498	4,7809
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125	2,2281	2,7638	3,1693	4,5869
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959	2,201	2,7181	3,1058	4,437
12	0,1283	0,259	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823	2,1788	2,681	3,0545	4,3178
13	0,1281	0,2586	0,394	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709	2,1604	2,6503	3,0123	4,2208
14	0,128	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,345	1,7613	2,1448	2,6245	2,9768	4,1405
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531	2,1314	2,6025	2,9467	4,0728
16	0,1277	0,2576	0,3923	0,535	0,6901	0,8647	1,0711	1,3368	1,7459	2,1199	2,5835	2,9208	4,015
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,069	1,3334	1,7396	2,1098	2,5669	2,8982	3,9651
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,862	1,0672	1,3304	1,7341	2,1009	2,5524	2,8784	3,9216
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,861	1,0655	1,3277	1,7291	2,093	2,5395	2,8609	3,8834
20	0,1273	0,2567	0,3909	0,5329	0,687	0,86	1,064	1,3253	1,7247	2,086	2,528	2,8453	3,8495
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207	2,0796	2,5176	2,8314	3,8193
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171	2,0739	2,5083	2,8188	3,7921
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139	2,0687	2,4999	2,8073	3,7676
24	0,127	0,2562	0,39	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109	2,0639	2,4922	2,7969	3,7454
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081	2,0595	2,4851	2,7874	3,7251
26	0,1269	0,256	0,3896	0,5309	0,684	0,8557	1,0575	1,315	1,7056	2,0555	2,4786	2,7787	3,7066
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033	2,0518	2,4727	2,7707	3,6896
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,056	1,3125	1,7011	2,0484	2,4671	2,7633	3,6739
29	0,1268	0,2557	0,3892	0,5302	0,683	0,8542	1,0553	1,3114	1,6991	2,0452	2,462	2,7564	3,6594
30	0,1267	0,2556	0,389	0,53	0,6828	0,8538	1,0547	1,3104	1,6973	2,0423	2,4573	2,75	3,646
40	0,1265	0,255	0,3881	0,5286	0,6807	0,8507	1,05	1,3031	1,6839	2,0211	2,4233	2,7045	3,551
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641	1,9901	2,3739	2,6387	3,4163
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6577	1,9799	2,3578	2,6174	3,3735
∞	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449	1,96	2,3264	2,5758	3,2905

4.3 Loi de Fisher



La statistique de Fisher est définie par

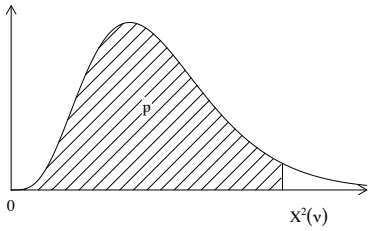
$$F = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

avec ici $\hat{\sigma}_1^2 > \hat{\sigma}_2^2$, ν_1 le nombre de d.d.l. correspondant à $\hat{\sigma}_1^2$ et ν_2 celui correspondant à $\hat{\sigma}_2^2$.

	ν_1								
ν_2	1	2	3	4	5	6	7	8	9
1	161,4	199,5	215,7	224,6	230,2	234	236,8	238,9	240,5
2	18,51	19	19,16	19,25	19,3	19,33	19,35	19,37	19,38
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999
5	6,608	5,786	5,409	5,192	5,05	4,95	4,876	4,818	4,772
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099
7	5,591	4,737	4,347	4,12	3,972	3,866	3,787	3,726	3,677
8	5,318	4,459	4,066	3,838	3,687	3,581	3,5	3,438	3,388
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,23	3,179
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,02
12	4,747	3,885	3,49	3,259	3,106	2,996	2,913	2,849	2,796
15	4,543	3,682	3,287	3,056	2,901	2,79	2,707	2,641	2,588
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393
24	4,26	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,3
30	4,171	3,316	2,922	2,69	2,534	2,421	2,334	2,266	2,211
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,18	2,124
60	4,001	3,15	2,758	2,525	2,368	2,254	2,167	2,097	2,04
120	3,92	3,072	2,68	2,447	2,29	2,175	2,087	2,016	1,959
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,01	1,938	1,88

	ν_1										
ν_2	9	10	12	15	20	24	30	40	60	120	∞
1	240,5	241,9	243,9	245,9	248	249,1	250,1	251,1	252,2	253,3	254,3
2	19,38	19,4	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,5
3	8,812	8,786	8,745	8,703	8,66	8,639	8,617	8,594	8,572	8,549	8,526
4	5,999	5,964	5,912	5,858	5,803	5,774	5,746	5,717	5,688	5,658	5,628
5	4,772	4,735	4,678	4,619	4,558	4,527	4,496	4,464	4,431	4,398	4,365
6	4,099	4,06	4	3,938	3,874	3,841	3,808	3,774	3,74	3,705	3,669
7	3,677	3,637	3,575	3,511	3,445	3,41	3,376	3,34	3,304	3,267	3,23
8	3,388	3,347	3,284	3,218	3,15	3,115	3,079	3,043	3,005	2,967	2,928
9	3,179	3,137	3,073	3,006	2,936	2,9	2,864	2,826	2,787	2,748	2,707
10	3,02	2,978	2,913	2,845	2,774	2,737	2,7	2,661	2,621	2,58	2,538
12	2,796	2,753	2,687	2,617	2,544	2,505	2,466	2,426	2,384	2,341	2,296
15	2,588	2,544	2,475	2,403	2,328	2,288	2,247	2,204	2,16	2,114	2,066
20	2,393	2,348	2,278	2,203	2,124	2,082	2,039	1,994	1,946	1,896	1,843
24	2,3	2,255	2,183	2,108	2,027	1,984	1,939	1,892	1,842	1,79	1,733
30	2,211	2,165	2,092	2,015	1,932	1,887	1,841	1,792	1,74	1,683	1,622
40	2,124	2,077	2,003	1,924	1,839	1,793	1,744	1,693	1,637	1,577	1,509
60	2,04	1,993	1,917	1,836	1,748	1,7	1,649	1,594	1,534	1,467	1,389
120	1,959	1,91	1,834	1,75	1,659	1,608	1,554	1,495	1,429	1,352	1,254
∞	1,88	1,831	1,752	1,666	1,571	1,517	1,459	1,394	1,318	1,221	1,002

4.4 Loi du χ^2



	p												
ν	0.995	0.99	0.975	0.95	0.9	0.75	0.5	0.25	0.1	0.05	0.025	0.01	0.005
1	7,879	6,635	5,024	3,841	2,706	1,323	0,4549	0,1015	0,01579	0,003932	0,0009821	0,0001571	3,927e-05
2	10,6	9,21	7,378	5,991	4,605	2,773	1,386	0,5754	0,2107	0,1026	0,05064	0,0201	0,01003
3	12,84	11,34	9,348	7,815	6,251	4,108	2,366	1,213	0,5844	0,3518	0,2158	0,1148	0,07172
4	14,86	13,28	11,14	9,488	7,779	5,385	3,357	1,923	1,064	0,7107	0,4844	0,2971	0,207
5	16,75	15,09	12,83	11,07	9,236	6,626	4,351	2,675	1,61	1,145	0,8312	0,5543	0,4117
6	18,55	16,81	14,45	12,59	10,64	7,841	5,348	3,455	2,204	1,635	1,237	0,8721	0,6757
7	20,28	18,48	16,01	14,07	12,02	9,037	6,346	4,255	2,833	2,167	1,69	1,239	0,9893
8	21,95	20,09	17,53	15,51	13,36	10,22	7,344	5,071	3,49	2,733	2,18	1,646	1,344
9	23,59	21,67	19,02	16,92	14,68	11,39	8,343	5,899	4,168	3,325	2,7	2,088	1,735
10	25,19	23,21	20,48	18,31	15,99	12,55	9,342	6,737	4,865	3,94	3,247	2,558	2,156
11	26,76	24,72	21,92	19,68	17,28	13,7	10,34	7,584	5,578	4,575	3,816	3,053	2,603
12	28,3	26,22	23,34	21,03	18,55	14,85	11,34	8,438	6,304	5,226	4,404	3,571	3,074
13	29,82	27,69	24,74	22,36	19,81	15,98	12,34	9,299	7,042	5,892	5,009	4,107	3,565
14	31,32	29,14	26,12	23,68	21,06	17,12	13,34	10,17	7,79	6,571	5,629	4,66	4,075
15	32,8	30,58	27,49	25	22,31	18,25	14,34	11,04	8,547	7,261	6,262	5,229	4,601
16	34,27	32	28,85	26,3	23,54	19,37	15,34	11,91	9,312	7,962	6,908	5,812	5,142
17	35,72	33,41	30,19	27,59	24,77	20,49	16,34	12,79	10,09	8,672	7,564	6,408	5,697
18	37,16	34,81	31,53	28,87	25,99	21,6	17,34	13,68	10,86	9,39	8,231	7,015	6,265
19	38,58	36,19	32,85	30,14	27,2	22,72	18,34	14,56	11,65	10,12	8,907	7,633	6,844
20	40	37,57	34,17	31,41	28,41	23,83	19,34	15,45	12,44	10,85	9,591	8,26	7,434
21	41,4	38,93	35,48	32,67	29,62	24,93	20,34	16,34	13,24	11,59	10,28	8,897	8,034
22	42,8	40,29	36,78	33,92	30,81	26,04	21,34	17,24	14,04	12,34	10,98	9,542	8,643
23	44,18	41,64	38,08	35,17	32,01	27,14	22,34	18,14	14,85	13,09	11,69	10,2	9,26
24	45,56	42,98	39,36	36,42	33,2	28,24	23,34	19,04	15,66	13,85	12,4	10,86	9,886
25	46,93	44,31	40,65	37,65	34,38	29,34	24,34	19,94	16,47	14,61	13,12	11,52	10,52
26	48,29	45,64	41,92	38,89	35,56	30,43	25,34	20,84	17,29	15,38	13,84	12,2	11,16
27	49,64	46,96	43,19	40,11	36,74	31,53	26,34	21,75	18,11	16,15	14,57	12,88	11,81
28	50,99	48,28	44,46	41,34	37,92	32,62	27,34	22,66	18,94	16,93	15,31	13,56	12,46
29	52,34	49,59	45,72	42,56	39,09	33,71	28,34	23,57	19,77	17,71	16,05	14,26	13,12
30	53,67	50,89	46,98	43,77	40,26	34,8	29,34	24,48	20,6	18,49	16,79	14,95	13,79
40	66,77	63,69	59,34	55,76	51,81	45,62	39,34	33,66	29,05	26,51	24,43	22,16	20,71
50	79,49	76,15	71,42	67,5	63,17	56,33	49,33	42,94	37,69	34,76	32,36	29,71	27,99
60	91,95	88,38	83,3	79,08	74,4	66,98	59,33	52,29	46,46	43,19	40,48	37,48	35,53
70	104,2	100,4	95,02	90,53	85,53	77,58	69,33	61,7	55,33	51,74	48,76	45,44	43,28
80	116,3	112,3	106,6	101,9	96,58	88,13	79,33	71,14	64,28	60,39	57,15	53,54	51,17
90	128,3	124,1	118,1	113,1	107,6	98,65	89,33	80,62	73,29	69,13	65,65	61,75	59,2
100	140,2	135,8	129,6	124,3	118,5	109,1	99,33	90,13	82,36	77,93	74,22	70,06	67,33