Premières notions de statistique
Analyse de la variance

Franck Picard

UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

La variabilité

- Lorsque l'on raisonne sur un groupe d'individus (ou une population), on cherche souvent à établir des tendances moyennes
- A ce titre, on aura toujours un compromis à trouver entre "tendance" et variabilité
- La tendance moyenne permet de synthétiser un ensemble de valeurs en une seule valeur indicative
- Une tendance forte accompagnée d'une variabilité élevée sera interprétée différemment d'une tendance forte accompagnée d'un bruit faible.

Causes de la variabilité

- Elle peuvent être connues et plus ou moins contrôlées
  - le type de sol est une source de variabilité pour le rendement d'une culture
  - Les conditions de croissance d'une colonie bactérienne
- ou non contrôlées
  - L'erreur de mesure d'un appareil de spectroscopie
  - Le génotype des individus (contrôlé ? contrôlable ? à contrôler ?)

L'objectif des modèles linéaires est d'analyser et de mettre au point des expériences basées sur du matériel expérimental variable

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

Rappels du le test de Student

- Le test de Student permet de tester l'effet de deux modalités sur les moyennes de deux échantillons
- Vocabulaire : on s'intéresse en fait à un **facteur** comportant deux **modalités** : $i = 1, 2$
- Le modèle sous-jacent au test est le modèle gaussien sur deux populations **indépendantes** :

$$Y_j^1 \underset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \;\; Y_j^2 \underset{iid}{\sim} \mathcal{N}(\mu_2, \sigma^2)$$

- Dans ce modèle i correspond au niveau du facteur, et j est l'indice de répétition, $j = 1, ...n_i$
- On souhaite tester l'hypothèse **sur le paramètre d'espérance**

$$H_0 : \{\mu_1 = \mu_2\}$$

Comparer I moyennes ?

- On considère maintenant un facteur avec I modalités
- Le modèle sous-jacent au test est toujours le modèle gaussien sur I populations **indépendantes** :

$$Y_j^1 \underset{iid}{\sim} \mathcal{N}(\mu_1, \sigma^2), \ \ldots Y_j^I \underset{iid}{\sim} \mathcal{N}(\mu_I, \sigma^2)$$

- Par analogie, on souhaite tester l'absence globale d'influence du **facteur** sur la variable observée **y**

$$H_0 : \{\mu_1 = \ldots = \mu_I\}$$

Notations générales

- L'échantillon observé $\mathbf{y}$ est un vecteur de taille $n = \sum_i n_i$ composé de I vecteurs $\mathbf{y}_i$:

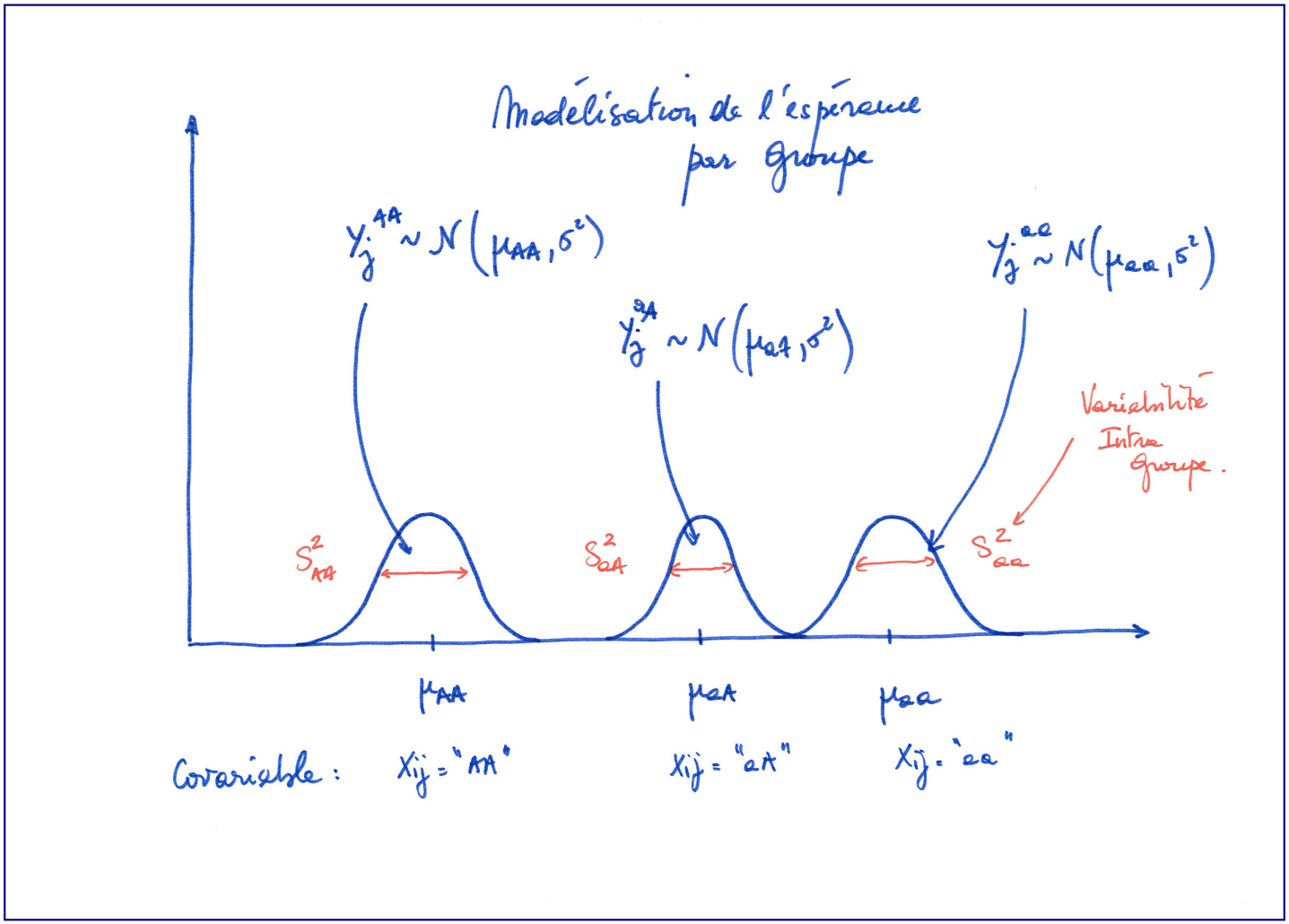
$$\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_I]$$

- Chaque $\mathbf{y}_i = [y_{i1}, \ldots, y_{i,n_i}]$, ou $\mathbf{y}_i = [y_{ij}]$ $j = 1, \ldots, n_i$.
- Le modèle porte sur les $\mathbf{y}_i$ qui sont modélisées par :

$$\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I}_n), \ \boldsymbol{\mu}_i = \mu_i 1_{n_i}$$

- On fait toujours les mêmes hypothèses : les Y_{ij} sont indépendants (intra et inter-traitement)
- σ^2 est constante : le modèle est **homoscédastique**

Illustration



Exemple : Anxiété des sportifs

- On s'intéresse au niveau d'anxiété chez les sportifs : diffère-t-il en fonction du niveau de la compétition ?
- On récolte des données chez des sportifs de 4 niveaux différents

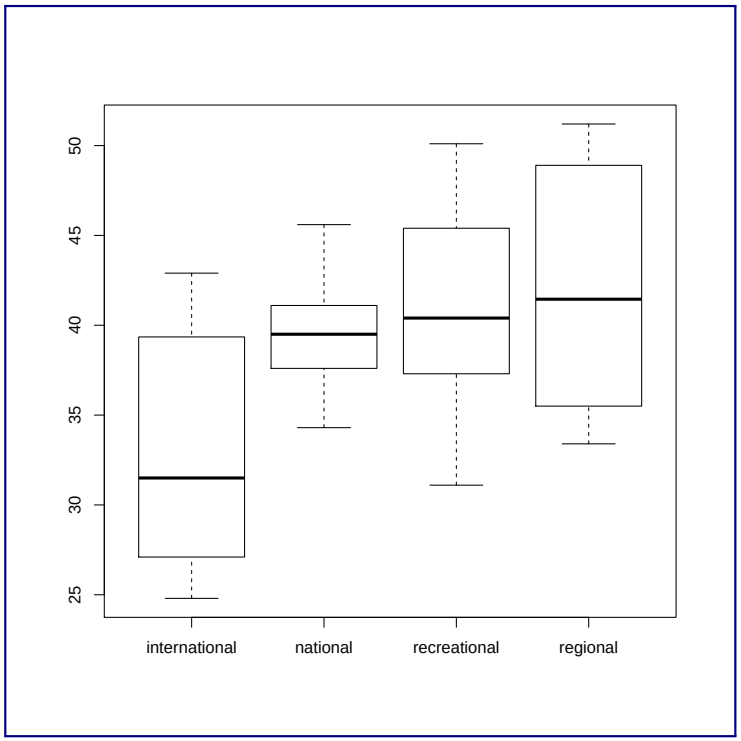
international	national	régional	recreational
24.80	45.60	33.40	31.10
26.70	41.10	34.60	35.70
27.50	34.30	36.40	37.30
30.60	37.60	39.10	39.40
32.40	39.50	43.80	40.40
38.20		47.90	44.50
40.50		49.90	45.40
42.90		51.20	49.80
			50.10
$n_1 = 8$	$n_2 = 5$	$n_3 = 8$	$n_4 = 9$

Exemple : Anxiété chez les sportifs

- On note y_{ij} l'anxiété mesurée pour un sportif j au niveau de compétition i
- On suppose que les niveaux d'anxiétés sont indépendants d'un sportif à un autre, et d'un niveau à un autre
- On pose le modèle suivant :

$$\mathbf{Y}_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I}_n), \ \boldsymbol{\mu}_i = \mu_i 1_{n_i}$$

- μ_i est le l'anxiété moyenne du niveau de compétition i



Une espérance conditionnelle

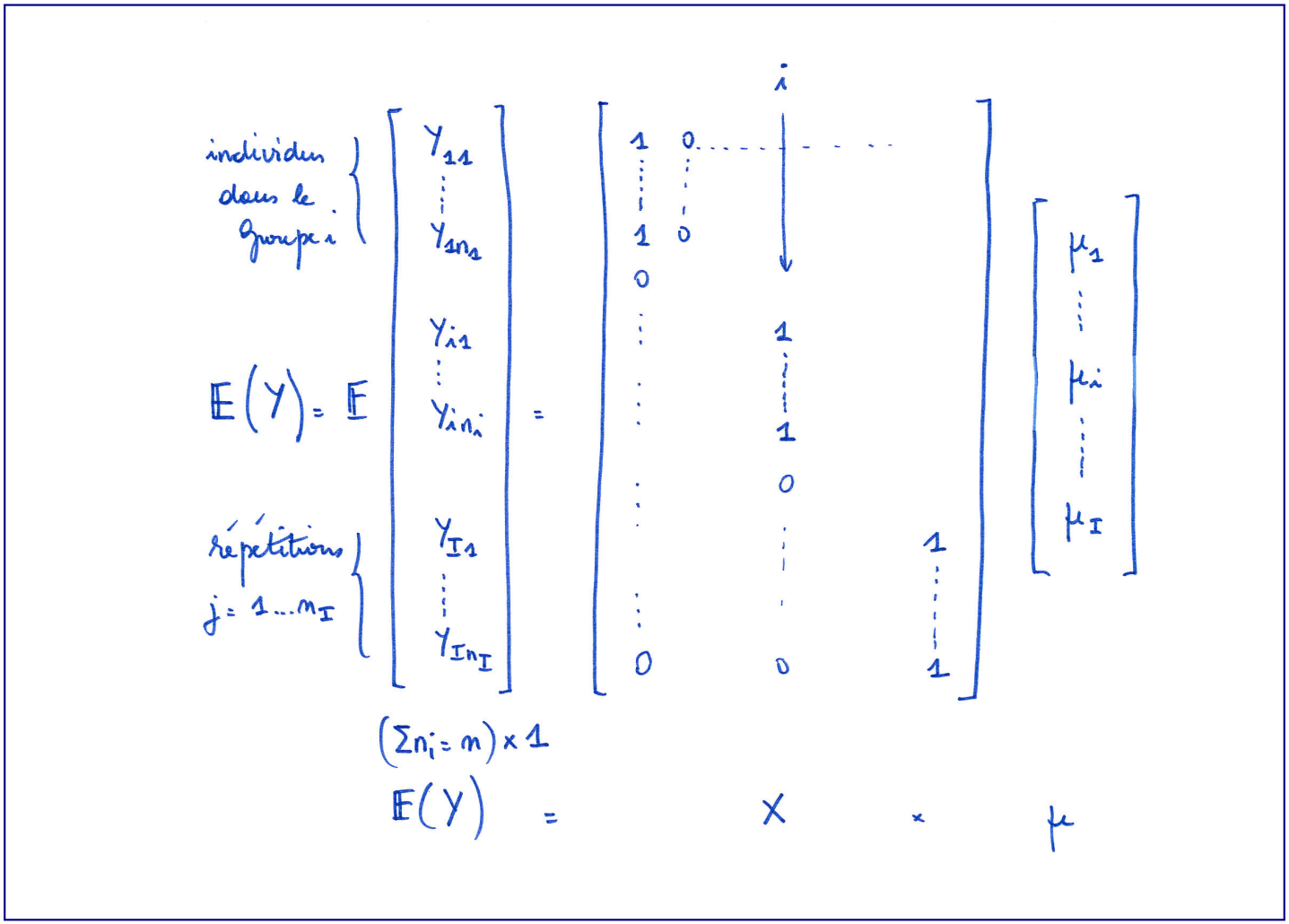
- On introduit une covariable X_{ij} qui définit le niveau du facteur : $X_{ij} = 1$ si l'individu j est dans le groupe i, 0 sinon
- Si on devait modéliser la loi de X_{ij} on pourrait choisir une loi multinomiale
- Or on ne s'intéresse pas aux variations de $\mathbf{X}$, mais à celles de $\mathbf{Y}$ avec un $\mathbf{X}$ fixé
- La modélisation consiste à donner une forme à l'espérance conditionnelle des observations ayant observé les X_{ij} :

$$\mathbb{E}(Y_{ij}|X_{ij} = 1) = \mu_i, \ \mathbb{V}(Y_{ij}|X_{ij} = 1) = \sigma^2$$

- On peut aussi écrire :

$$\mathbb{E}(Y_{ij}|X_{ij} = x_{ij}) = \sum_{i=1}^{I} \mu_i x_{ij}$$

Notation matricielle



Définition des résidus

- Le modèle standard est $\mathbf{Y} \sim \mathcal{N}(\mathbf{X}\boldsymbol{\mu}, \sigma^2 \mathbf{I}_n)$.
- On cherche à décomposer l'espérance du signal en fonction de covariables $\mathbf{X}$
- On peut s'interroger sur ce qu'il reste une fois cette modélisation effectuée
- On introduit une nouvelle variable, que l'on appelle **résidu** :

$$E_{ij} = Y_{ij} - \mu_i, \;\; E_{ij} \sim \mathcal{N}(0, \sigma^2) \;\; (iid)$$

- C'est un terme d'erreur aléatoire, l'écart entre les observations et le modèle

Une nouvelle écriture du modèle

- Les modèles de régression s'écrivent souvent d'une manière canonique (on suppose implicitement que les covariables sont fixées)
- En utilisant la notion de résidus, le modèle devient :

$$Y_{ij} = \mu_i + E_{ij},\ E_{ij} \sim \mathcal{N}(0, \sigma^2)\ (iid)$$

- C'est une décomposition très courante :

observations aléatoires= signal + bruit aléatoire

- On suppose que le signal contenu dans l'espérance du modèle
- Le terme d'erreur devient un terme "mystère" qui contient tout ce qu'on n'a pas pris en compte dans le modèle
- On s'interrogera sur le caractère résiduel des résidus !

Paramètres et estimateurs du modèle

- Paramètres $(\mu_i)_i$, I paramètres de moyenne + 1 paramètre de variance
- Le critère des moindre-carrés :

$$d^2(\mathbf{Y}, \boldsymbol{\mu}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \mu_i)^2 = \sum_{i=1}^{I} \sum_{j=1}^{n_i} E_{ij}^2$$

- L'estimateur des moindre-carrés pour le paramètre μ_i est donc la moyenne empirique du niveau i :

$$\widehat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \frac{1}{n_i} Y_{i+} = Y_{i\bullet}$$

- Un estimateur de la variance est la somme des carrés résiduelle :

$$\widehat{\sigma}^2 = \frac{1}{n - I} \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - \widehat{\mu}_i)^2$$

Prédictions à partir du modèle estimé

- Une fois les paramètres estimés, on peut utiliser les estimateurs pour prédire des valeurs de la réponse
- Sur les données déjà observées, on peut résumer les Y_{ij} par la moyenne de chaque groupe $Y_{i\bullet}$:

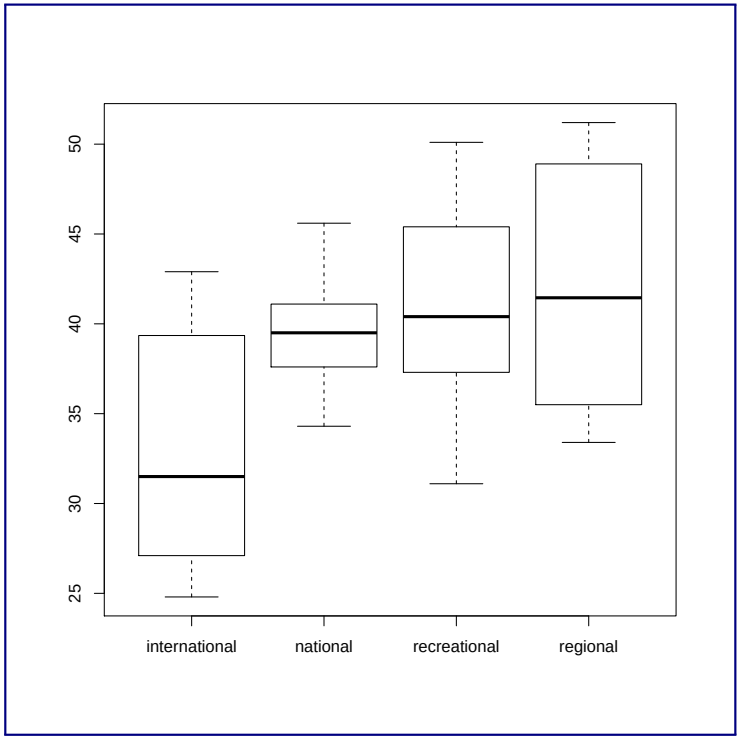
$$\widehat{Y}_{ij} = \widehat{\mu}_i$$

- L'idée est de prédire la valeur de Y pour un individu pour lequel on ne dispose que d'une mesure de x_{ij}. Pour un individu tel que $x_{ij} = 1$ alors on prédira sa valeur de Y à l'aide de $\widehat{\mu}_i$
- Plus généralement, on notera $\widehat{\mathbf{Y}}$ la prédiction de $\mathbf{Y}$ à partir des estimateurs du modèle

Exemple : Anxiété chez les sportifs

- On estime l'anxiété moyenne par la moyenne empirique pour chaque modalité
- Si un nouveau sportif de niveau national se présentait, on prédirait sa valeur d'anxiété par 39.62

facteur	$Y_{i\bullet}$	variance intra
international	32.95	46.37
national	39.62	17.59
recreational	41.52	41.51
regional	42.04	50.51



Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

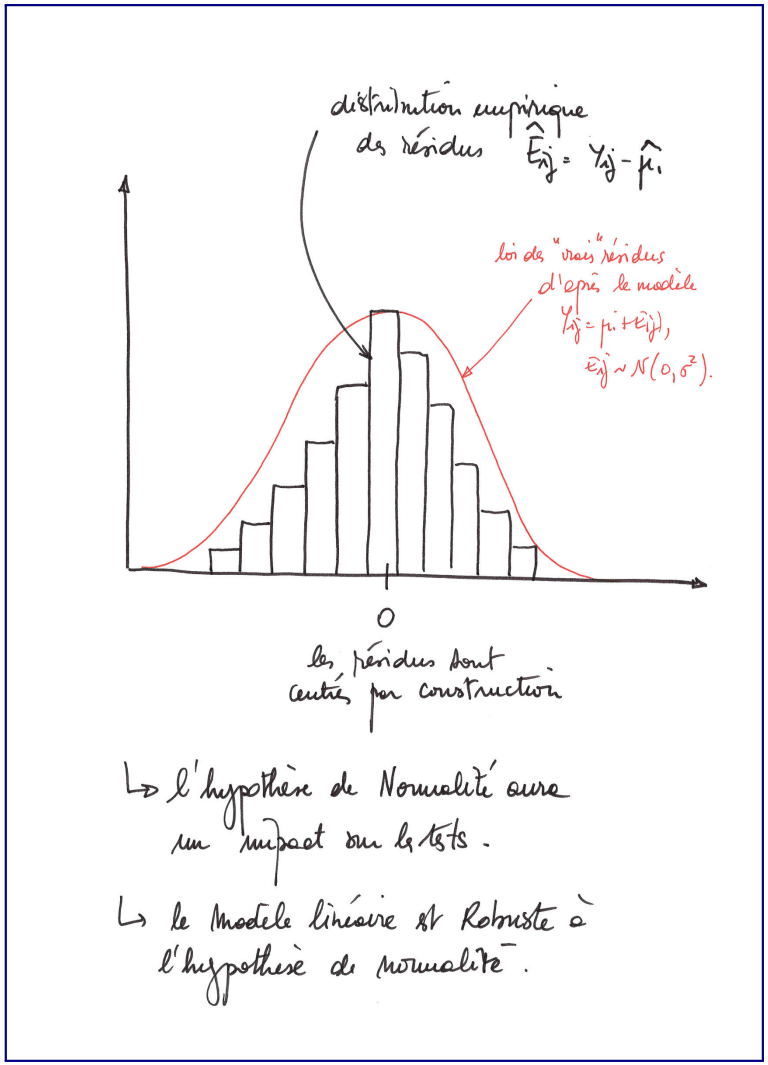
7 Comparaisons de traitements

Validation du modèle

- Après avoir estimé les paramètres du modèle, une première étape consiste à vérifier que les hypothèses du modèle sont vérifiées
- Quelles hypothèses ?
  - l'indépendance des erreurs
  - la normalité des erreurs
  - l'homoscédasticité du modèle (σ^2 constante)
- Le premier point peut être contrôlé en amont (plan d'expérience)
- Les deux points suivants concernent les résidus : $\widehat{E}_{ij} = Y_{ij} - \widehat{\mu}_i$

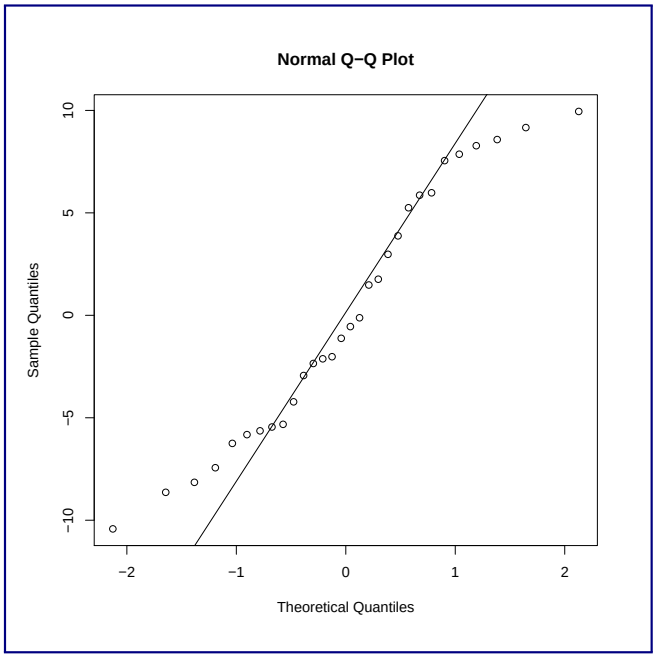
Validation de l'hypothèse de normalité - 1

- Après avoir calculé les prédictions des résidus, on peut appliquer un test d'adéquation à la loi normale
- Remarque : les résidus sont centrés par construction
- Exemple de test : Shapiro-Wilks, Kolmogorov-Smirnov
- Le modèle linéaire est robuste à l'hypothèse de normalité



Validation de l'hypothèse de normalité - 2

- On utilise souvent un outil graphique : Q-Q plot
- L'idée est de représenter les quantiles empiriques en fonction de quantiles théoriques d'une loi (ici la loi normale)
- C'est un outil qui permet souvent d'identifier les points qui s'écartent particulièrement de l'hypothèse de normalité (ici les queues de distribution, les valeurs extrêmes d'anxiété)



Q-Q plot des résidus pour l'anxiété chez les sportifs

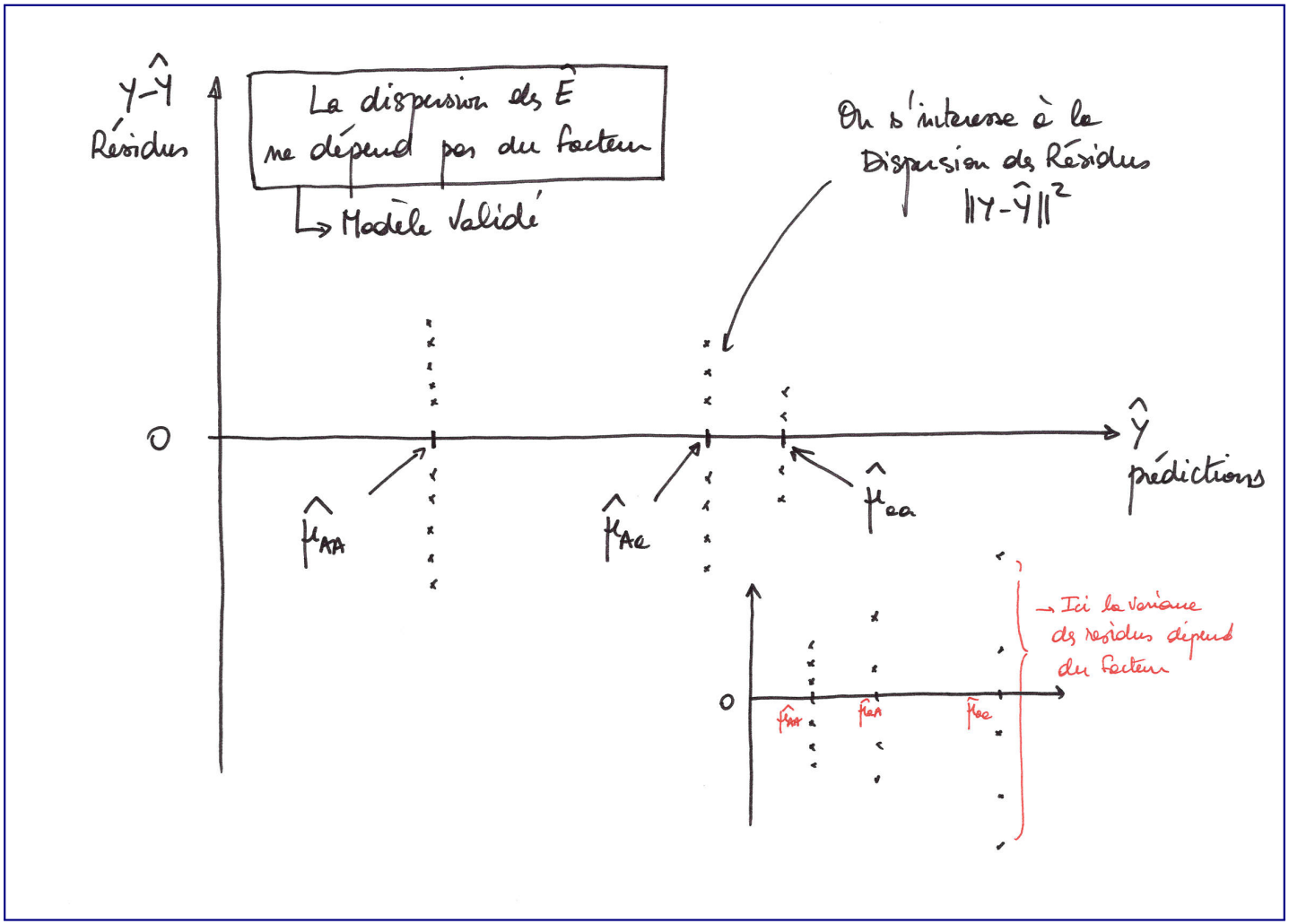
Le théorème de Cochran et la variance des résidus

- L'hypothèse fondamentale qui doit absolument être respectée est l'hypothèse d'homoscedasticité
- On suppose que la variance des erreurs σ^2 ne dépend pas de la covariable, elle est supposée constante
- Pour vérifier cette hypothèse on utilise un résultat provenant du théorème de Cochran (théorème à la base de tous les développements du modèle linéaire) :

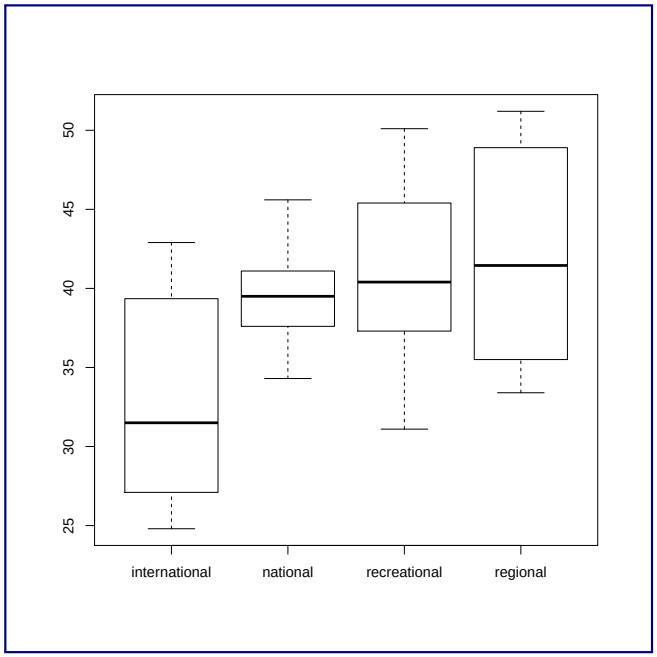
$$\|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 \perp \widehat{\mathbf{Y}}$$

- La dispersion des résidus doit être indépendante de l'intensité de la prédiction.
- A partir de ce résultat, on construit ce qui s'appelle le graphe des résidus qui sert de diagnostic visuel

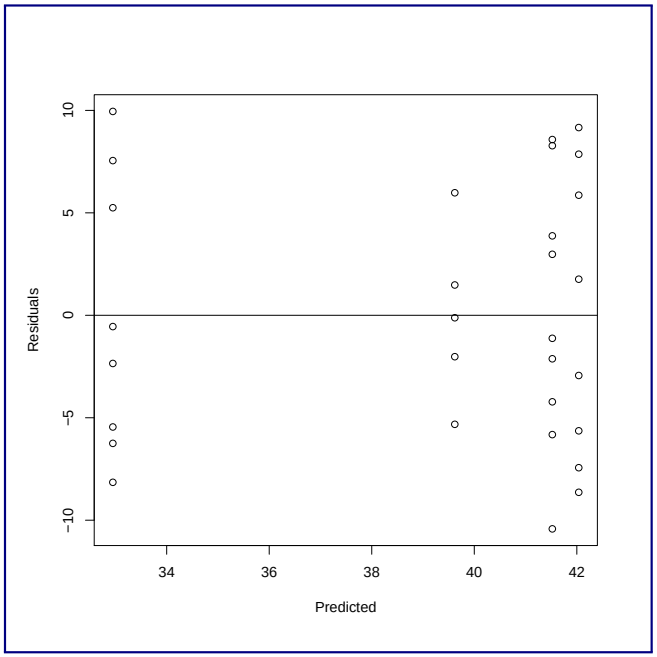
Le graphe des résidus



Exemple : anxiété des sportifs



Boxplot du signal en fonction des modalités du facteur



Résidus du modèle en fonction des prédictions données par les $Y_{i\bullet}$ (32.95,39.62,41.52,42.04)

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

La décomposition de la somme des carrés

- Le modèle porte sur l'espérance de la variable étudiée, mais son étude repose sur la décomposition de la variabilité (**AN**alysis **O**f **VA**riance)
- La variance initiale est celle des observations c'est la somme des carrés totale

$$\mathrm{SCT}(\mathbf{Y}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{\bullet\bullet})^2$$

- Elle décrit la variabilité du jeu de données autour de sa moyenne générale $Y_{\bullet\bullet}$
- Elle est **fixée** pour un jeu de données $\mathbf{Y}$ particulier

La somme des carrés résiduelle (Within Sum of Squares)

- Elle décrit la variabilité des observations autour de la moyenne de chaque groupe

$$\text{SCR}(\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - Y_{i\bullet})^2$$

- Elle **dépend du modèle** (ici des niveaux de X)
- Elle est faible quand les observations sont proches de la moyenne de chaque groupe (quand $Y_{i\bullet}$ décrit bien chaque groupe)

La SCR est une **variance résiduelle** qui décrit la variabilité restant après avoir retranché la part du signal expliquée par le modèle. C'est la variance des résidus.

La variance inter groupes (Between Sum of Squares)

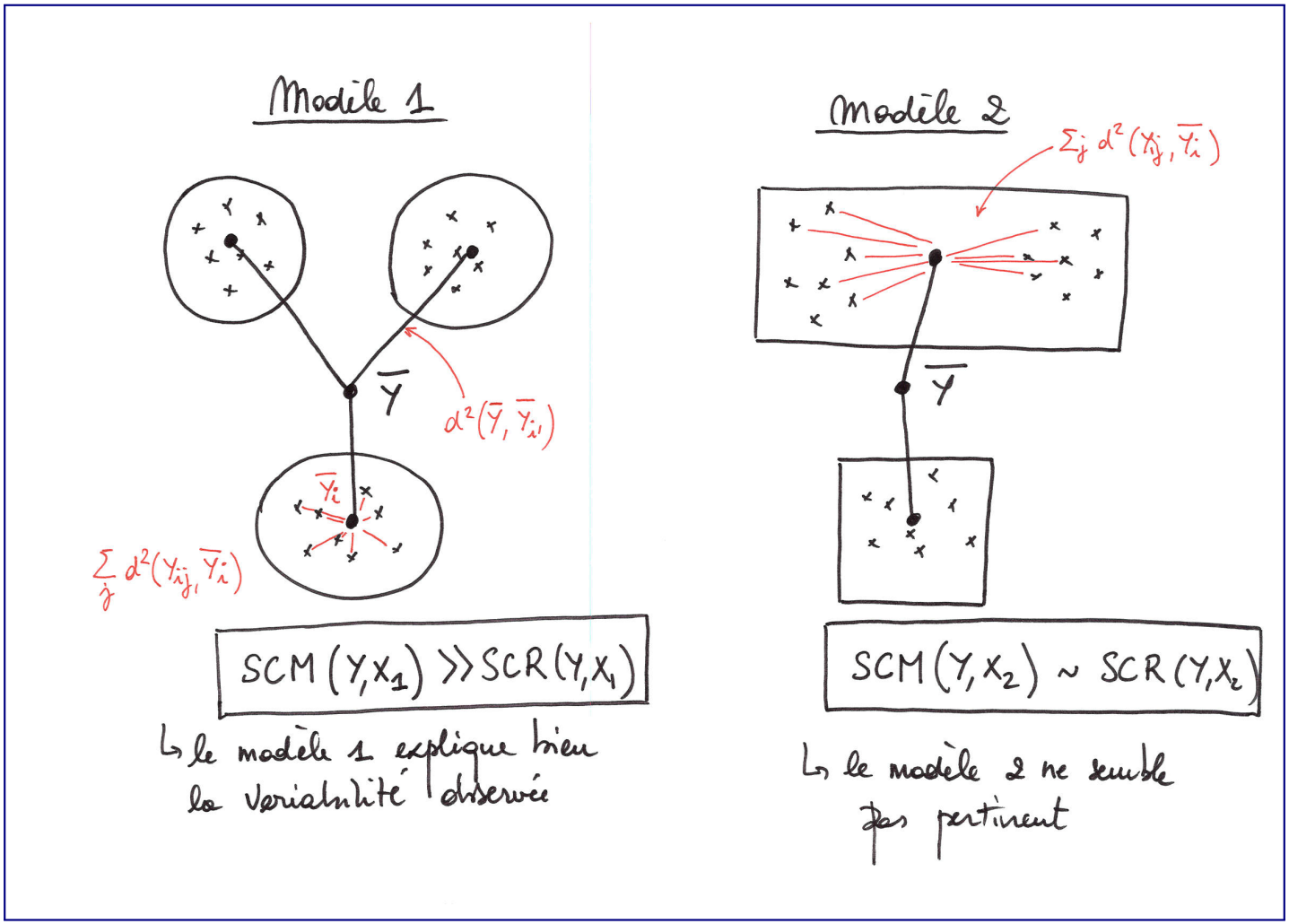
- Elle décrit la variabilité de chaque groupe autour de la moyenne générale

$$\mathrm{SCM}(\mathbf{Y}, \mathbf{X}) = \sum_{i=1}^{I} n_i (Y_{i\bullet} - Y_{\bullet\bullet})^2$$

- Si on résumait les observations de chaque groupe par leur moyenne quelle variabilité observerait-on ?
- Plus elle est grande, plus chaque centre de groupe se distingue des autres

La SCM décrit la séparabilité des groupes et donne une indication sur la pertinence du modèle considéré

Illustration



La formule magique

- C'est LA formule

$$\mathrm{SCT}(\mathbf{Y}) = \mathrm{SCM}(\mathbf{Y}, \mathbf{X}) + \mathrm{SCR}(\mathbf{Y}, \mathbf{X})$$

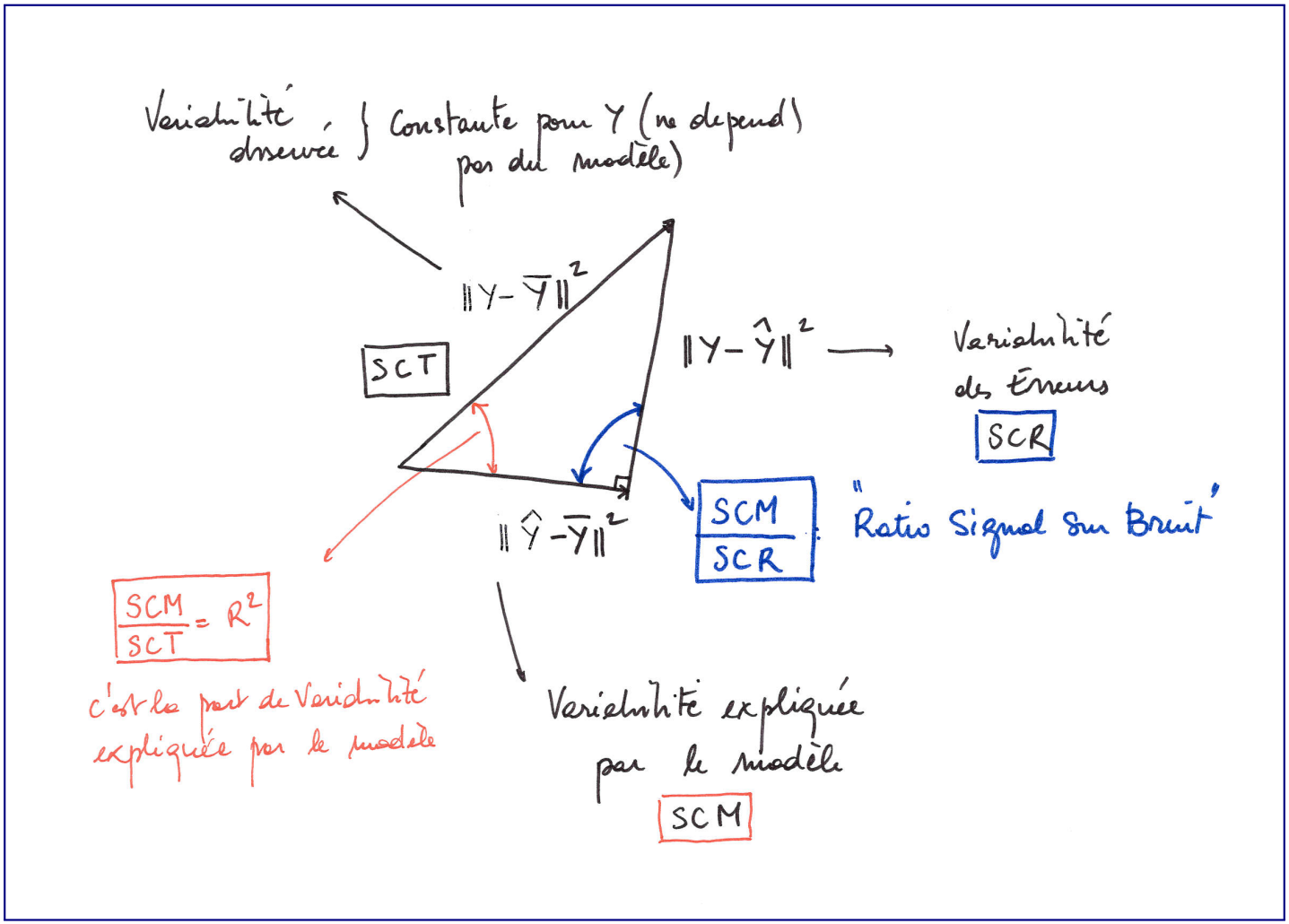
- Pour la retrouver :

$$(Y_{ij} - Y_{\bullet\bullet})^2 = (Y_{ij} - Y_{i\bullet} + Y_{i\bullet} - Y_{\bullet\bullet})^2$$

- La formule plus générale est issue du théorème de Pythagore

$$\|\mathbf{Y} - \overline{\mathbf{Y}}\|^2 = \|\widehat{\mathbf{Y}} - \overline{\mathbf{Y}}\|^2 + \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2$$

Un peu de géométrie



Comment quantifier l'apport du modèle ?

1/ Le facteur a-t-il globalement un effet ? Pour répondre à cette question, on s'intéresse au rapport

$$\frac{\text{SCM}(\mathbf{Y},\mathbf{X})}{\text{SCR}(\mathbf{Y},\mathbf{X})} = \frac{\text{Variance expliquée par le modèle}}{\text{Variance résiduelle}}$$

2/ Si oui en 1/ on peut s'interroger sur le pouvoir explicatif du modèle. On introduit le coefficient de **détermination** :

$$R^2 = \frac{\text{SCM}(\mathbf{Y},\mathbf{X})}{\text{SCT}(\mathbf{Y})} = \frac{\text{Variance expliquée par le modèle}}{\text{Variance Totale}}$$

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

Comment décider ? Le modèle est-il pertinent ?

- La construction des sommes de carrés peut se faire sans hypothèse sur le modèle, c'est une décomposition algébrique
- On souhaiterait faire des tests pour prendre en compte la variabilité des données dans la prise de décision
- Idée : grâce aux hypothèses de normalité et d'indépendance, on peut déterminer la loi des sommes de carrés (loi du χ^2)
- On utilisera la loi de Fisher qui décrit les variations de rapports de deux χ^2

Espérance des sommes de carrés

- La somme des carrés résiduelle

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}}(\mathrm{SCR}(\mathbf{Y},\mathbf{X})) &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}\mathbb{E}\left[(Y_{ij}-Y_{i\bullet})^2\right] \\ &= (n-I)\sigma^2\end{aligned}$$

- La somme des carrés du modèle

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}}(\mathrm{SCM}(\mathbf{Y},\mathbf{X})) &= \sum_{i=1}^{I} n_i\mathbb{E}\left[(Y_{i\bullet}-Y_{\bullet\bullet})^2\right] \\ &= (I-1)\sigma^2+\sum_{i=1}^{I} n_i\left(\mu_i-\mu\right)^2\end{aligned}$$

Les sommes de carrés moyennes (Mean Squares)

- Les carrés moyens résiduels $\mathrm{SCR}/n-I$, avec

$$\mathrm{SCR} \sim \sigma^2 \chi^2(n-I)$$

- Les carrés moyens du modèle

$$\frac{\mathrm{SCM}}{I-1} = \sigma^2 + \frac{1}{I-1}\sum_{i=1}^{I} n_i \left(\mu_i - \mu\right)^2$$

- Dans la suite, on utilisera la loi de Fisher qui décrit les variations d'un rapport de variables indépendantes

$$\frac{\chi^2(\mathrm{df}_1)/\mathrm{df}_1}{\chi^2(\mathrm{df}_2)/\mathrm{df}_2} \sim \mathcal{F}(\mathrm{df}_1, \mathrm{df}_2)$$

Test d'hypothèse d'absence d'effet du facteur

- On adopte une approche globale : le modèle est-il informatif ?
- Dans le cas de l'ANOVA-1 cela revient à poser l'hypothèse de l'absence de différence de moyennes entre les modalités du facteur

$$H_0 : \{\mu_1 = \ldots = \mu_I = \mu\}$$

- Sous cette hypothèse : $\text{SCM} \underset{H_0}{\sim} \sigma^2 \chi^2(I-1)$
- L'hypothèse faite sur les **espérances** se répercute globalement sur les **variances** et on teste H_0 avec la statistique de Fisher :

$$F = \frac{\text{SCM}/(I-1)}{\text{SCR}/(n-I)} \underset{H_0}{\sim} \mathcal{F}(I-1, n-I)$$

La table d'ANOVA

- Elle permet en une table de résumer la décomposition globale des sommes de carré
- Elle permet de faire le premier test global de Fisher (modèle nul *vs*. modèle complet)
- Elle est présente après chaque utilisation du modèle linéaire dans les logiciels usuel

Source	Sum of Squares	dF	Mean Squares	F-stat	Pv
Factor	SCM	I-1	SCM / (I-1)	MSM/MSR	.
Residual	SCR	n-I	SCR / (n-I)		
Total	SCT				

Exemple : l'anxiété des sportifs

- On considère le modèle $Y_{ij} = \mu_i + E_{ij}$ avec Y_{ij} le niveau d'anxiété des sportifs dans une compétition de type i. On a vérifié le graph des résidus pour ce modèle (homoscédasticité ok)
- En étudiant la table d'analyse de la variance, on constate que la statistique de Fisher du modèle nul contre le modèle complet présente une p-value de 0.0322 plus petite que $\alpha = 5\%$
- On rejette l'hypothèse du modèle nul : le modèle est informatif globalement, et explique $R^2 = 425.7/(425.7 + 1080.6) \simeq 28\%$ de la variabilité observée

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
compet	3	425.7	141.90	3.414	0.0322 *
Residuals	26	1080.6	41.56		

Signif. codes : 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Le test de Kruskal–Wallis et l'approche non-paramétrique

- Si l'analyse des résidus invalide fortement l'hypothèse de normalité, ou si le nombre d'observations est trop faible pour une approche paramétrique
- On peut construire une approche fondée sur les rangs : on note R_{ij} le rang de Y_{ij} dans l'échantillon et $R_{i\bullet}$ le rang moyen des observations du groupe i et $R_{\bullet\bullet} = (n+1)/2$
- Pour déterminer si le rang des observations dépend du groupe, on considère la statistique

$$K = (n-1)\frac{\sum_{i=1}^{I} n_i (R_{i\bullet} - R_{\bullet\bullet})^2}{\sum_{i=1}^{I}\sum_{j=1}^{n_i} (R_{ij} - R_{i\bullet})^2}$$

- On connaît la loi de K sous l'hypothèse nulle (dépend de la présence d'ex-aequos)

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

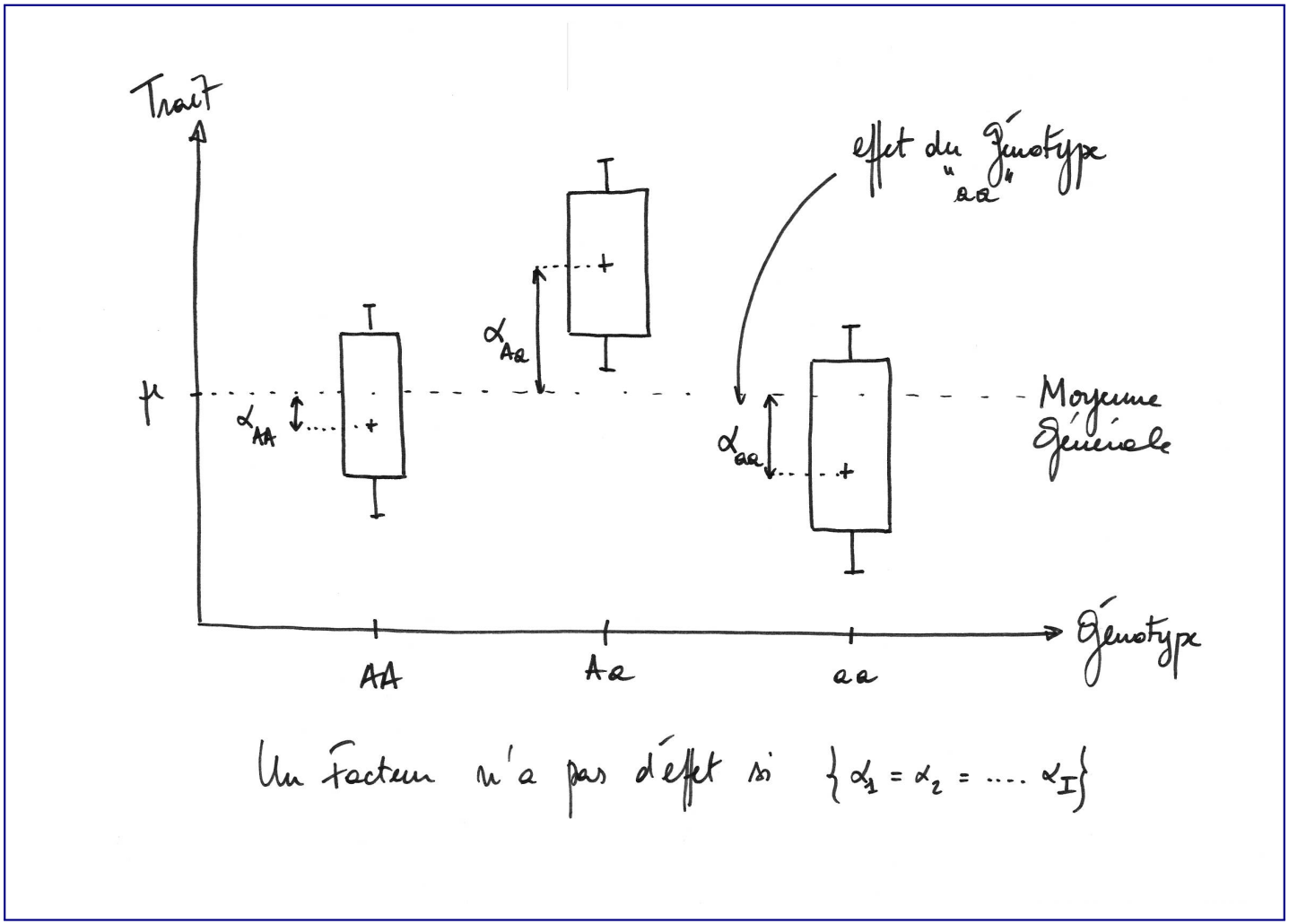
Une nouvelle formulation du modèle

- On introduit une nouvelle notation

$$\mathbb{E}(Y_{ij}) = \mu_i = \mu + \alpha_i$$

- μ s'interprète comme la moyenne générale
- α_i s'interprète comme l'effet du niveau i
- On gagne en interprétation (surtout pour les modèles à plusieurs facteurs)
- Attention ! On augmente le nombre de paramètres (avec un nombre d'observations constant)

Illustration du modèle additif



Estimation des paramètres du nouveau modèle

- On utilise la technique des moindre-carrés

$$d^2(\mathbf{Y}, \mu, \boldsymbol{\alpha}) = \sum_{i=1}^{I} \sum_{j=1}^{n_i} (Y_{ij} - [\mu + \alpha_i])^2$$

- On dérive cette expression par rapport à μ et à tous les α_i (système à $I + 1$ inconnues)

$$\begin{cases} Y_{++} - n\widehat{\mu} - \sum_i n_i \alpha_i = 0 \\ Y_{i+} - n_i \widehat{\mu} - n_i \alpha_i = 0 \end{cases}$$

- Le système admet une infinité de solution si on ne rajoute pas de contrainte

Le modèle additif n'est pas **identifiable** : les estimateurs dépendront de certaines contraintes.

Contraintes usuelles

- $\alpha_I = 0$ c'est une contrainte utile d'un point de vue numérique (utilisée dans R). Dans ce cas, le dernier niveau devient un niveau de référence :

$$\widehat{\mu} = Y_{I\bullet},\ \widehat{\alpha}_i = Y_{i\bullet} - Y_{I\bullet}$$

- $\sum_i n_i \alpha_i = 0$: les effets se compensent. Cette contrainte redonne les estimateurs naturels

$$\widehat{\mu} = Y_{\bullet\bullet},\ \widehat{\alpha}_i = Y_{i\bullet} - Y_{\bullet\bullet}$$

On ne peut pas interpréter les valeurs des $\widehat{\alpha}_i$ dans l'absolu.

Exemple : anxiété chez les sportifs

- Les moyennes empiriques des 4 groupes sont :

facteur	international	national	recreational	regional
$Y_{i\bullet}$	32.95	39.62	41.52	42.04

- Or la sortie R issue du modèle linéaire est :

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           32.950      2.279  14.456  6.1e-14 ***
competnational         6.670      3.675   1.815  0.08111 .
competrecreational     8.572      3.133   2.736  0.01105 *
competregional         9.088      3.223   2.819  0.00909 **
```

- Le niveau $Y_{\text{international}\bullet}$ a été pris comme référence

Combinaisons Linéaires estimables

- La nécessité des contraintes implique qu'il faut être très prudent si on souhaite tester la valeur d'un paramètre isolément ($\{\alpha_i = 0\}$).
- Certaines combinaisons linéaires de paramètres ne dépendent pas des contraintes (invariantes)
- On appelle **combinaison linéaire estimable** une combinaison linéaire des paramètres qui ne dépend pas des contraintes.
- Les **prédictions** ne dépendent pas des contraintes $\widehat{\mu}_i = \widehat{\mu} + \widehat{\alpha}_i$
- En général on cherche à comparer les traitements entre eux. On souhaite tester $\alpha_i - \alpha_{i'} = 0$. C'est un **contraste** qui ne dépend pas des contraintes

Outline

1 Introduction à l'ANOVA

2 Introduction du modèle de régression

3 Introduction à l'analyse des résidus

4 Les sommes de carrés

5 Construction des tests

6 Retour sur les paramètres et les estimateurs

7 Comparaisons de traitements

Une démarche en deux temps

- La démarche globale du modèle linéaire consiste d'abord à déterminer si le facteur étudié a un effet ou non
- On répond à cette question à l'aide de la table d'analyse de la variance (sommes de carrés)
- Si on a détecté une influence d'un facteur sur la réponse on peut se demander quels sont les niveaux de ce facteur qui sont responsables de l'effet global
- On teste un **ensemble** d'hypothèses nulles $H_0^{i,i'} : \{\alpha_i - \alpha_{i'} = 0\}$

Retour sur la statistique de Student, et intérêt de l'ANOVA

- Dans le cas de deux modalités, la statistique de Student s'écrit :

$$T(\mathbf{Y}) = \frac{\widehat{\mu}_i - \widehat{\mu}_{i'}}{S(\mathbf{Y})\sqrt{1/n_i + 1/n_{i'}}} \sim \mathcal{T}(n_i + n_{i'} - 2)$$

- L'intérêt du modèle linéaire est qu'il propose un estimateur global de la variance résiduelle :

$$T(\mathbf{Y}) = \frac{\widehat{\alpha}_i - \widehat{\alpha}_{i'}}{\widehat{\sigma}\sqrt{1/n_i + 1/n_{i'}}} \sim \mathcal{T}(n - I)$$

- On gagne en puissance grâce à une meilleure estimation de σ qui prend en compte toutes les données (et pas seulement les deux traitements que l'on teste)

Comparaisons multiples, tests multiples, risques multiples ?

- Lorsque l'on teste I traitements deux à deux, on effectue $I(I-1)/2$ tests avec chacun un risque α (notation différente du α_i)
- A l'issue des $I(I-1)/2$ tests on peut s'interroger sur le risque global de la procédure α_G : quelle erreur globale obtient-on ? (faux positifs)
- Dans le pire des cas, si tous les tests étaient indépendants, l'espérance du nombre de faux positifs serait $I(I-1)\alpha/2$ (on cumule les risques)

Plusieurs réponses possibles

- La définition du risque dans le cas d'un ensemble d'hypothèses n'est pas unique
- Une première méthode consiste à contrôler le risque global d'obtenir au moins un faux positif (Family-Wise Error Rate) :

  $\alpha_{\text{global}} = \mathbb{P}(\text{Rejet d'au moins une hypothèse nulle} | \text{Toutes sont vraies})$

- La **correction de Bonferroni** consiste à diviser le risque avec lequel est testée chaque hypothèse individuelle par le nombre de tests
- C'est une procédure **conservative** qui manque de puissance en général. Il existe d'autres procédures de contrôle comme le False Discovery Rate (FDR)

Probabilité critique ajustée

- Si on souhaite contrôler le risque global de la procédure α_{global} alors chaque hypothèse individuelle sera testée avec le nouveau risque corrigé $\widetilde{\alpha}$ tel que :

$$\widetilde{\alpha} = \alpha_{\text{global}}/\{\#\text{tests}\}$$

- Plutôt que de modifier le risque avec lequel est testé chaque hypothèse, on peut modifier les p-values
- La règle de décision standard est si $P_v(t) \leq \alpha$ alors on rejette H_0
- On définit la p-value ajustée telle que :

$$\widetilde{P}_v(t) = \min\{\#\text{tests} \times P_v(t), 1\}$$

- On compare ensuite la probabilité critique ajustée au risque global qui est contrôlé.

Exemple : Anxiété chez les sportifs, méthode de Tukey

	diff	lwr	upr	p adj
national-international	6.67	-3.41	16.75	0.29
recreational-international	8.57	-0.02	17.17	0.05
regional-international	9.09	0.24	17.93	0.04
recreational-national	1.90	-7.96	11.77	0.95
regional-national	2.42	-7.67	12.50	0.91
regional-recreational	0.52	-8.08	9.11	1.00

Exemple : Anxiété chez les sportifs, méthode de Tukey