

HLMA408: Traitement des données

ANOVA: cas d'un unique facteur

Joseph Salmon

<http://josephsalmon.eu>

Université de Montpellier



Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Introduction

Pour ce cours on se servira des données récoltées sur le syndrome de Down⁽¹⁾

Syndrome de Down = chromosome 21 supplémentaire

<https://www.stat.berkeley.edu/~statlabs/labs.html>

≈ 0.25 million de personnes aux US.

Seuls les gènes “en bas” du chromosome 21 sont à l’origine de ce syndrome.

But de l’étude : trouver les gènes responsables de la maladie.

Méthode : ajouter à des souris de laboratoire des portions du chromosome 21 humain, et observer l’apparition de symptômes.

⁽¹⁾voir aussi le livre par Nolan and Speed(2001), Chapitre 11

Mesurer l'apparition de symptômes chez la souris

- ▶ Tests pour mesurer leurs facultés d'apprentissage, leur intelligence. Ces tests sont basés sur des signaux visuels. PB: 500 souris de l'étude sont nées aveugles.
- ▶ Pour les souris aveugles : on ne dispose que d'une mesure de masse

Mesurer l'apparition de symptômes chez la souris

- ▶ Tests pour mesurer leurs facultés d'apprentissage, leur intelligence. Ces tests sont basés sur des signaux visuels. PB: 500 souris de l'étude sont nées aveugles.
- ▶ Pour les souris aveugles : on ne dispose que d'une mesure de masse

On espère que des comparaisons des masses de ces souris aveugles fourniront des preuves supplémentaires pour détecter la région du chromosome 21 à l'origine du syndrome.

Mesurer l'apparition de symptômes chez la souris

- ▶ Tests pour mesurer leurs facultés d'apprentissage, leur intelligence. Ces tests sont basés sur des signaux visuels. PB: 500 souris de l'étude sont nées aveugles.
- ▶ Pour les souris aveugles :
on ne dispose que d'une mesure de masse

On espère que des comparaisons des masses de ces souris aveugles fourniront des preuves supplémentaires pour détecter la région du chromosome 21 à l'origine du syndrome.

Données: *Human Genome Center à Lawrence Berkeley Laboratory*

Panel de souris transgéniques, chacune contenant un des quatre fragments d'ADN



Élevage et reproduction avec d'autres souris non transgéniques



On attend plusieurs générations. . .



souris de notre échantillon

Les quatre fragments d'ADN sont

230E8 (670 Kb)

141G6 (475 Kb)

152F7 (570 Kb)

285E6 (430 Kb)


Variables

Sur les souris aveugles⁽²⁾ :

- ▶ **DNA** : le fragment d'ADN inséré dans la souris ancêtre avec les codes '1'=141G6, '2'=152F7, '3'=230E8 et '4'=285E6 ('0' : groupe témoin, sans trisomie)
- ▶ **Line** : la lignée
- ▶ **Transgenic** : (binaire 0 ou 1)
- ▶ **Sex** : (1= mâle, 0=femelle)
- ▶ **Age** : âge au moment de la pesée (en jours)
- ▶ **Weight** : en grammes, à 0.1 g près.
- ▶ **Cage** : numéro de la cage où la souris est élevée

⁽²⁾données: <http://josephsalmon.eu/enseignement/datasets/mouse.data>

ANOVA : Motivation

ANOVA ( : *Analysis of variance*) : méthode d'analyse permettant d'étudier la dépendance d'une variable quantitative par rapport à 1 ou 2 variables qualitatives (=facteurs).

Exemple : influence du sexe sur la masse des souris

- ▶ 1 facteur explicatif: le sexe
- ▶ 2 modalités : mâle (M) / femelle (F)

Exemple : influence du fragment d'ADN sur la masse des souris

- ▶ 1 facteur explicatif: le fragment d'ADN
- ▶ 5 modalités : '141G6', '152F7', '230E8', '285E6', et 'sans trisomie'

Méthodologie : aperçu

Idée sous-jacente:

1. calculer les moyennes (et les variances) de la variable quantitative sur les sous-populations par modalités.
2. utiliser des tests classiques pour identifier des différences significatives dans ces sous-populations.

Pour cela :

- ▶ tester si l'espérance de la variable quantitative est homogène sur l'ensemble des modalités de la variable qualitative
- ▶ tester (test de Fischer) l'hypothèse nulle \mathcal{H}_0 d'égalité des espérances par l'analyse des termes de variance
- ▶ si l'on rejette cette hypothèse (on suppose donc qu'il y a un effet du facteur) on cherche alors à tester pour chaque modalité si elle a un impact ou non (test de Student)

Rappel sur la moyenne

Théorème

Si y_1, \dots, y_n sont n observations d'une variable y , la moyenne empirique \bar{y}_n minimise

$$f(c) := \sum_{i=1}^n (y_i - c)^2 .$$

Ce que l'on note:

$$\bar{y}_n = \arg \min_{c \in \mathbb{R}} f(c) .$$

Preuve : prendre la condition de premier ordre, c'est-à-dire que $f'(c^*) = 0$ pour la solution c^* de ce problème d'optimisation, *i.e.*, $2 \sum_{i=1}^n (y_i - c^*) = 0$ et donc finalement $c^* = \bar{y}_n$

Interprétation : Au sens des moindres carrés la moyenne est la meilleure approximation d'un échantillon par une constante!

Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Moyennes et moindres carrés

Pour estimer la masse moyenne des souris mâles et femelles

Notation :

- ▶ y_i : masse de la i^{e} souris,
- ▶ \bar{y}_M : masse moyenne des mâles
- ▶ \bar{y}_F : masse moyenne des femelles
- ▶ $\mathbb{1}_{M,i}$: variable **binaire** (ou **indicatrice**) codant le sexe

$$\mathbb{1}_{M,i} = \begin{cases} 1, & \text{si la } i^{\text{e}} \text{ souris est un mâle} \\ 0, & \text{si la } i^{\text{e}} \text{ souris est une femelle} \end{cases}$$

Théorème

$$(\bar{y}_F, \bar{y}_M - \bar{y}_F) = (\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{(\beta_0, \beta_1) \in \mathbb{R}^2} \sum_{i=1}^n \left(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \right)^2$$

Preuve

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 \mathbf{1}_{M,i}))^2 \\ &= \sum_{i:\text{m\^a}le} (y_i - (\beta_0 + \beta_1))^2 + \sum_{i:\text{femelle}} (y_i - \beta_0)^2 \end{aligned}$$

Pour minimiser f , les conditions n\u00e9cessaires du 1^{er} ordre donnent:

Preuve

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 \mathbf{1}_{M,i}))^2 \\ &= \sum_{i:\text{m\^ale}} (y_i - (\beta_0 + \beta_1))^2 + \sum_{i:\text{femelle}} (y_i - \beta_0)^2 \end{aligned}$$

Pour minimiser f , les conditions n\u00e9cessaires du 1^{er} ordre donnent:

$$\begin{cases} \frac{\partial f}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) + 2 \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \frac{\partial f}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases}$$

Preuve

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}))^2 \\ &= \sum_{i:\text{m\^ale}} (y_i - (\beta_0 + \beta_1))^2 + \sum_{i:\text{femelle}} (y_i - \beta_0)^2 \end{aligned}$$

Pour minimiser f , les conditions n\u00e9cessaires du 1^{er} ordre donnent:

$$\begin{aligned} &\begin{cases} \frac{\partial f}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) + 2 \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \frac{\partial f}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases} \\ \Leftrightarrow &\begin{cases} \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases} \end{aligned}$$

Preuve

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 \mathbf{1}_{M,i}))^2 \\ &= \sum_{i:\text{m\^ale}} (y_i - (\beta_0 + \beta_1))^2 + \sum_{i:\text{femelle}} (y_i - \beta_0)^2 \end{aligned}$$

Pour minimiser f , les conditions n\u00e9cessaires du 1^{er} ordre donnent:

$$\begin{aligned} &\begin{cases} \frac{\partial f}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) + 2 \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \frac{\partial f}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases} \\ \Leftrightarrow &\begin{cases} \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases} \quad \Leftrightarrow \quad \begin{cases} \hat{\beta}_0 = \bar{y}_F \\ \hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_M \end{cases} \end{aligned}$$

Preuve

$$\begin{aligned} f(\beta_0, \beta_1) &= \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 \mathbf{1}_{M,i}))^2 \\ &= \sum_{i:\text{m\^ale}} (y_i - (\beta_0 + \beta_1))^2 + \sum_{i:\text{femelle}} (y_i - \beta_0)^2 \end{aligned}$$

Pour minimiser f , les conditions n\u00e9cessaires du 1^{er} ordre donnent:

$$\begin{cases} \frac{\partial f}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) + 2 \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \frac{\partial f}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 2 \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \sum_{i:\text{femelle}} (\hat{\beta}_0 - y_i) = 0 \\ \sum_{i:\text{m\^ale}} (\hat{\beta}_0 + \hat{\beta}_1 - y_i) = 0 \end{cases} \Leftrightarrow \begin{cases} \hat{\beta}_0 = \bar{y}_F \\ \hat{\beta}_0 + \hat{\beta}_1 = \bar{y}_M \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\beta}_0 = \bar{y}_F \\ \hat{\beta}_1 = \bar{y}_M - \bar{y}_F \end{cases}$$

Variance

Théorème

Pour un modèle gaussien $y_i = \beta_0^* + \beta_1^* \mathbf{1}_{M,i} + \varepsilon_i$, $\varepsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

$$\text{Var}(\hat{\beta}_0) = \text{Var}(\bar{y}_F) = \frac{\sigma^2}{n_F}$$

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\bar{y}_M - \bar{y}_F) = \frac{\sigma^2}{n_F} + \frac{\sigma^2}{n_M},$$

où n_F est le nombre de femelles et n_M le nombre de mâles. Enfin un estimateur sans biais de la variance est alors:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n-2} \sum_{i=1}^n \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{M,i}) \right]^2$$

Preuve : utiliser la formule de la variance d'une moyenne et que les variances de deux variables aléatoires indépendantes s'ajoutent

Généralisation: cas des souris

5 classes / modalités possibles pour le chromosome 21:

0. souris non trisomique
1. souris trisomique + fragment de type 1 (141G6)
2. souris trisomique + fragment de type 2 (152F7)
3. souris trisomique + fragment de type 3 (230E8)
4. souris trisomique + fragment de type 4 (285E6)

Généralisation: cas des souris

5 classes / modalités possibles pour le chromosome 21:

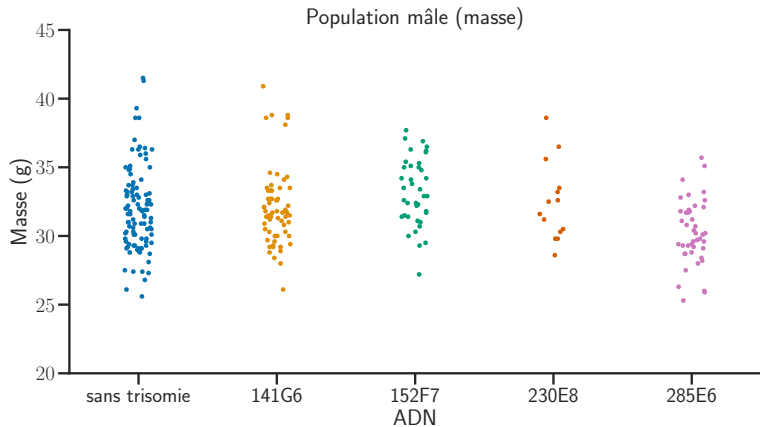
0. souris non trisomique
1. souris trisomique + fragment de type 1 (141G6)
2. souris trisomique + fragment de type 2 (152F7)
3. souris trisomique + fragment de type 3 (230E8)
4. souris trisomique + fragment de type 4 (285E6)

On introduit 4 variables binaires $\mathbb{1}_1$, $\mathbb{1}_2$, $\mathbb{1}_3$ et $\mathbb{1}_4$ qui indiquent quel fragment est présent :

$$\mathbb{1}_{1,i} = \begin{cases} 1 & \text{si 141G6 est présent,} \\ 0 & \text{sinon.} \end{cases} \quad \mathbb{1}_{2,i} = \begin{cases} 1 & \text{si 152F7 est présent,} \\ 0 & \text{sinon.} \end{cases}$$

$$\mathbb{1}_{3,i} = \begin{cases} 1 & \text{si 230E8 est présent,} \\ 0 & \text{sinon.} \end{cases} \quad \mathbb{1}_{4,i} = \begin{cases} 1 & \text{si 285E6 est présent,} \\ 0 & \text{sinon.} \end{cases}$$

Données



Étude sur les souris: statistiques descriptives (II)

n_0 = nombre d'observations de modalité 0 = 105

n_1 = nombre d'observations de modalité 1 = 64

n_2 = nombre d'observations de modalité 2 = 14

n_3 = nombre d'observations de modalité 3 = 43

n_4 = nombre d'observations de modalité 4 = 105

Ainsi $n = n_0 + n_1 + n_2 + n_3 + n_4 = 265$

Étude sur les souris: statistiques descriptives (II)

\bar{y}_0 = moyennes des observations de modalité 0 = 31.93

\bar{y}_1 = moyennes des observations de modalité 1 = 32.10

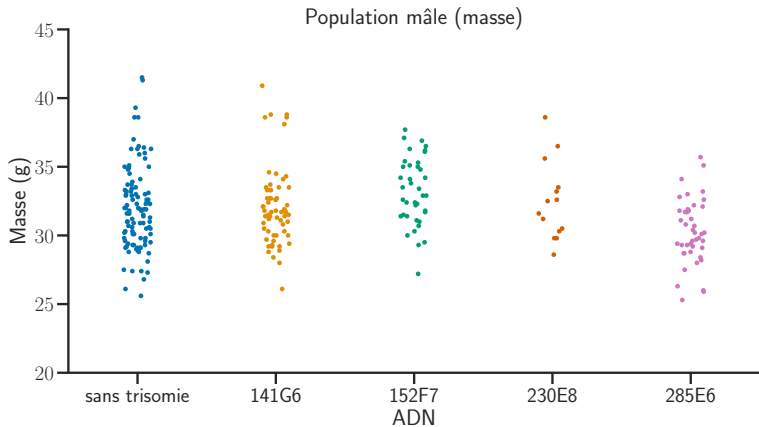
\bar{y}_2 = moyennes des observations de modalité 2 = 33.21

\bar{y}_3 = moyennes des observations de modalité 3 = 32.45

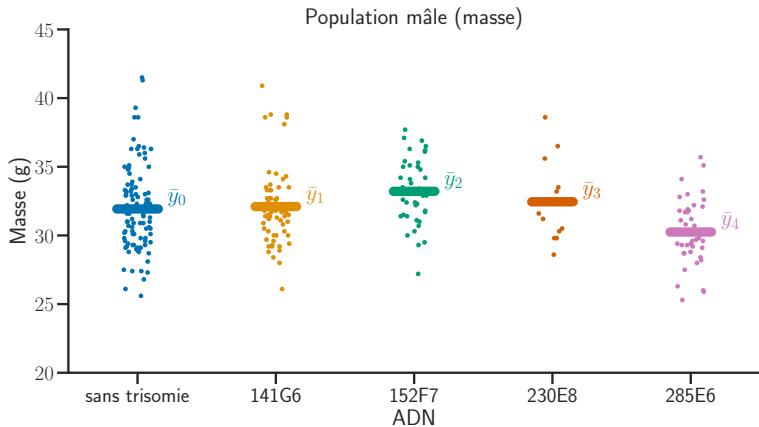
\bar{y}_4 = moyennes des observations de modalité 4 = 30.25

Ainsi $\bar{y} = (n_0\bar{y}_0 + n_1\bar{y}_1 + n_2\bar{y}_2 + n_3\bar{y}_3 + n_4\bar{y}_4)/n = 31.91$

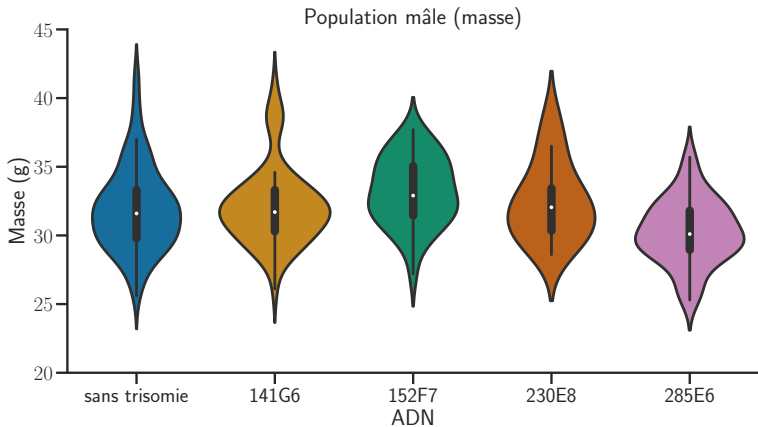
Données, moyennes, violons



Données, moyennes, violons



Données, moyennes, violons



Formulation multivariée

On minimise alors

$$f(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 \mathbf{1}_{1,i} + \beta_2 \mathbf{1}_{2,i} + \beta_3 \mathbf{1}_{3,i} + \beta_4 \mathbf{1}_{4,i}])^2$$

Formulation multivariée

On minimise alors

$$\begin{aligned} f(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) &= \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 \mathbf{1}_{1,i} + \beta_2 \mathbf{1}_{2,i} + \beta_3 \mathbf{1}_{3,i} + \beta_4 \mathbf{1}_{4,i}])^2 \\ &= \sum_{i=1}^n \left(y_i - \left[\beta_0 + \sum_{g=1}^4 \beta_g \mathbf{1}_{g,i} \right] \right)^2 \end{aligned}$$

Formulation multivariée

On minimise alors

$$\begin{aligned} f(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) &= \sum_{i=1}^n (y_i - [\beta_0 + \beta_1 \mathbf{1}_{1,i} + \beta_2 \mathbf{1}_{2,i} + \beta_3 \mathbf{1}_{3,i} + \beta_4 \mathbf{1}_{4,i}])^2 \\ &= \sum_{i=1}^n \left(y_i - \left[\beta_0 + \sum_{g=1}^4 \beta_g \mathbf{1}_{g,i} \right] \right)^2 \\ &= \sum_{i \text{ sans trisomie}} (y_i - \beta_0)^2 \\ &\quad + \sum_{i \text{ de type (1)}} (y_i - \beta_0 - \beta_1)^2 \\ &\quad + \sum_{i \text{ de type (2)}} (y_i - \beta_0 - \beta_2)^2 \\ &\quad + \sum_{i \text{ de type (3)}} (y_i - \beta_0 - \beta_3)^2 \\ &\quad + \sum_{i \text{ de type (4)}} (y_i - \beta_0 - \beta_4)^2 \end{aligned}$$

Solution du problème

En adaptant la preuve pour le cas de deux modalités, on obtient que le minimum de la fonction f précédente est atteint en

$$\begin{cases} \hat{\beta}_0 &= \bar{y}_0 \\ \hat{\beta}_1 &= \bar{y}_1 - \bar{y}_0 \\ \hat{\beta}_2 &= \bar{y}_2 - \bar{y}_0 \\ \hat{\beta}_3 &= \bar{y}_3 - \bar{y}_0 \\ \hat{\beta}_4 &= \bar{y}_4 - \bar{y}_0 \end{cases}$$

Conclusion : les moyennes par **modalité** (les valeurs possibles de la variable) sont primordiales

Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Retour sur les moyennes

Fait : les moyennes de chaque groupe peuvent être calculées avec une méthode de type moindres carrés

Estimation : les quantités calculées sont des estimateurs de

$$\mathbb{E}(y_i) = \begin{cases} \beta_0^*, & \text{si la } i^{\text{e}} \text{ souris est non trisomique,} \\ \beta_0^* + \beta_1^*, & \text{si la } i^{\text{e}} \text{ souris a 141G6 présent,} \\ \beta_0^* + \beta_2^*, & \text{si la } i^{\text{e}} \text{ souris a 152F7 présent,} \\ \beta_0^* + \beta_3^*, & \text{si la } i^{\text{e}} \text{ souris a 230E8 présent,} \\ \beta_0^* + \beta_4^*, & \text{si la } i^{\text{e}} \text{ souris a 285E6 présent.} \end{cases}$$

Exemple : β_1 représente la **différence** entre la moyenne des masses des souris transgéniques de type (1) et des souris non-trisomiques de type (0)

Rem. : la modalité "0" est la **modalité de référence**

Estimateur / Prédiction

Pour tout $i \in \llbracket 1, n \rrbracket$, la prédiction associée à la i^{e} observation est:

$$\hat{y}_i := \hat{\beta}_0 + \hat{\beta}_1 \mathbf{1}_{1,i} + \hat{\beta}_2 \mathbf{1}_{2,i} + \hat{\beta}_3 \mathbf{1}_{3,i} + \hat{\beta}_4 \mathbf{1}_{4,i}$$

Interprétation : la prédiction du modèle pour l'observation i est la (masse) moyenne du groupe correspondant à i :

$$\hat{y}_i = \begin{cases} \bar{y}_0, & \text{si } i \text{ est dans le groupe 0} \\ \bar{y}_1, & \text{si } i \text{ est dans le groupe 1} \\ \bar{y}_2, & \text{si } i \text{ est dans le groupe 2} \\ \bar{y}_3, & \text{si } i \text{ est dans le groupe 3} \\ \bar{y}_4, & \text{si } i \text{ est dans le groupe 4} \end{cases}$$

Retour sur les variances

Sous l'hypothèse d'un modèle gaussien

$$y_i = \beta_0^* + \beta_1^* \mathbf{1}_{1,i} + \cdots + \beta_4^* \mathbf{1}_{4,i} + \varepsilon_i, \text{ avec } \varepsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\left\{ \begin{array}{l} \text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n_0} \\ \text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_0} \\ \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_0} \\ \text{Var}(\hat{\beta}_3) = \frac{\sigma^2}{n_3} + \frac{\sigma^2}{n_0} \\ \text{Var}(\hat{\beta}_4) = \frac{\sigma^2}{n_4} + \frac{\sigma^2}{n_0} \end{array} \right.$$

et

n_0 = nombre d'observations de modalité 0

n_1 = nombre d'observations de modalité 1

\vdots

n_4 = nombre d'observations de modalité 4

Rem. : 0 est modalité de référence (ici modalité “non trisomique”)

Estimateur de la variance totale

Théorème

Sous l'hypothèse d'un modèle gaussien

$y_i = \beta_0^* + \beta_1^* \mathbf{1}_{1,i} + \dots + \beta_4^* \mathbf{1}_{4,i} + \varepsilon_i$ avec $\varepsilon_i \stackrel{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ alors un estimateur sans biais de σ^2 est

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-5} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-5} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 \mathbf{1}_{1,i} - \hat{\beta}_2 \mathbf{1}_{2,i} - \hat{\beta}_3 \mathbf{1}_{3,i} - \hat{\beta}_4 \mathbf{1}_{4,i})^2\end{aligned}$$

i.e.,

$$\mathbb{E}(\hat{\sigma}^2) = \sigma^2$$

On peut alors prendre pour estimateurs non-biaisés des variances des coefficients: $\hat{\sigma}^2(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{n_0}$, et $\hat{\sigma}^2(\hat{\beta}_g) = \frac{\hat{\sigma}^2}{n_g} + \frac{\hat{\sigma}^2}{n_0}$ pour $g = 1, \dots, 4$

Forme alternative

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-5} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-5} \sum_{g=0}^4 \sum_{i \text{ de classe } g} (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n-5} \sum_{g=0}^4 \sum_{i \text{ de classe } g} (y_i - \bar{y}_g)^2 \\ &= \frac{1}{n-5} \sum_{g=0}^4 (n_g - 1) \left(\frac{1}{n_g - 1} \sum_{i \text{ de classe } g} (y_i - \bar{y}_g)^2 \right) \\ &= \frac{1}{n-5} \sum_{g=0}^4 (n_g - 1) \hat{\sigma}^2(y_g)\end{aligned}$$

où pour tout $g \in \llbracket 0, 5 \rrbracket$, $\hat{\sigma}^2(y_g) = \frac{1}{n_g - 1} \sum_{i \text{ de classe } g} (y_i - \bar{y}_g)^2$

(estimateur sans biais de la variance)

Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Décomposition de la somme des carrés

Dès que l'on fait une estimation avec un critère des moindres carrés, on peut écrire

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Rappel : dans notre étude \hat{y}_i est la moyenne sur le sous-groupe contenant la i^{e} souris.

Rem. : ce n'est rien d'autre que le théorème de Pythagore!

Preuve : il faut développer le carré et réaliser que les doubles produits s'annulent

Simplification

G : nombre de groupes

n_g : nombre d'observations dans le g^e groupe

\bar{y}_g : moyenne empirique des observations du g^e groupe

\bar{y} : moyenne empirique sur l'échantillon complet

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{g=0}^{G-1} n_g (\bar{y}_g - \bar{y})^2$$

et de plus

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{g=0}^{G-1} \sum_{i \in g} (y_i - \bar{y}_g)^2$$

Simplification

G : nombre de groupes

n_g : nombre d'observations dans le g^e groupe

\bar{y}_g : moyenne empirique des observations du g^e groupe

\bar{y} : moyenne empirique sur l'échantillon complet

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \sum_{g=0}^{G-1} n_g (\bar{y}_g - \bar{y})^2$$

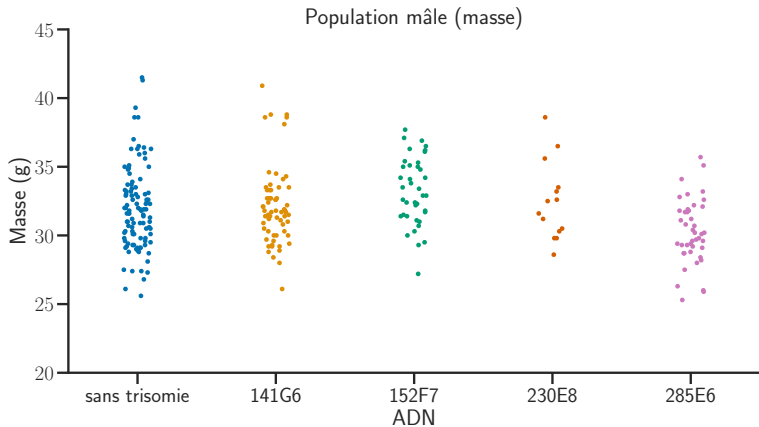
et de plus

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{g=0}^{G-1} \sum_{i \in g} (y_i - \bar{y}_g)^2$$

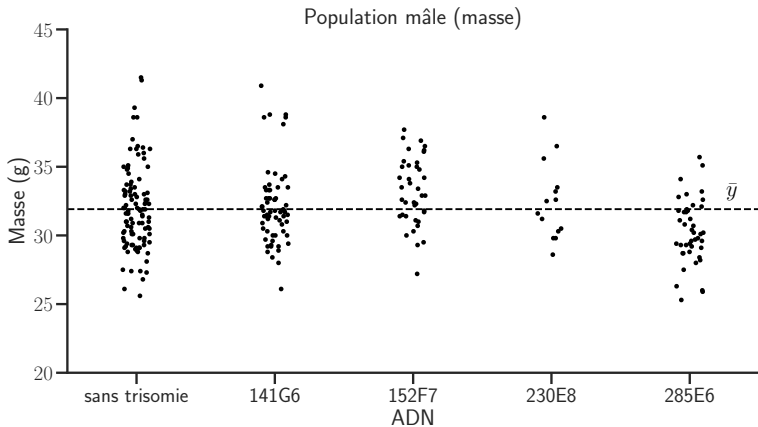
Ainsi,

$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{variabilité totale}} = \underbrace{\sum_{g=0}^{G-1} \sum_{i \in g} (y_i - \bar{y}_g)^2}_{\text{variabilité intra-groupe}} + \underbrace{\sum_{g=0}^{G-1} n_g (\bar{y}_g - \bar{y})^2}_{\text{variabilité inter-groupe}}$
--

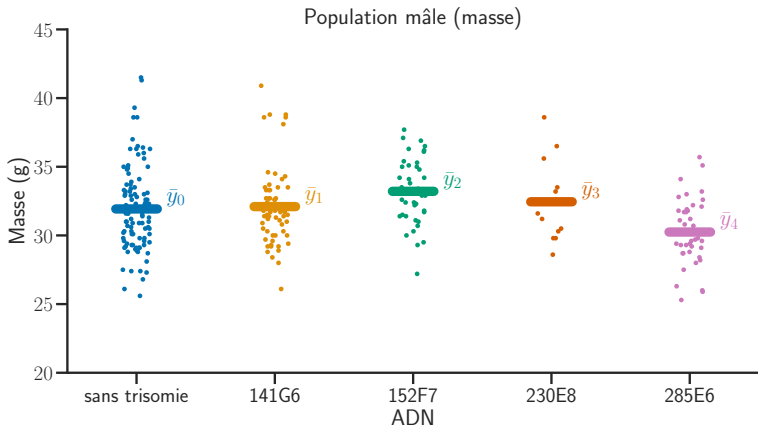
Données



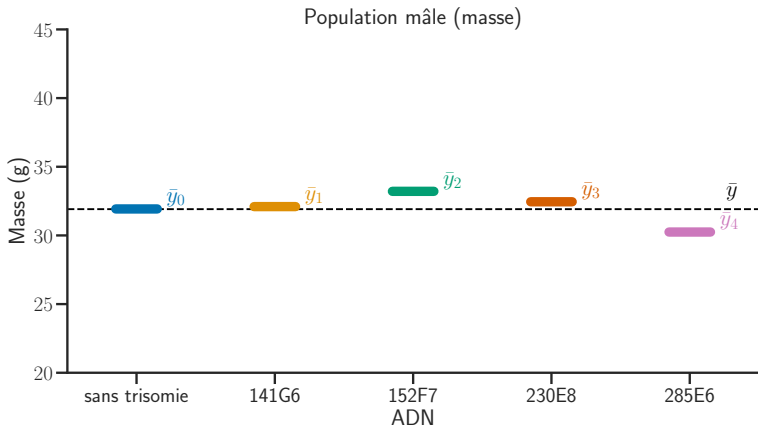
Variabilité totale



Variabilité intra-groupe



Variabilité inter-groupe



Sommaire

Introduction

Rappel sur l'espérance et les moindres carrés

Modèle statistique pour la moyenne

Somme des carrés et effet de la variable

Tests

Test (de Fisher) de l'effet du facteur

Hypothèse nulle, \mathcal{H}_0 : “ $\beta_1^* = \dots = \beta_{G-1}^* = 0$ ”
 \iff “les G groupes ont même espérance”
 \iff “ $y_i = \beta_0^* + \varepsilon_i, \quad \forall i \in \llbracket 1, n \rrbracket$ ”

Hypothèse alternative, \mathcal{H}_1 : “ $\exists g \in \{0, \dots, G-1\}$ tel que $\beta_g^* \neq 0$ ”

Statistique de test (de Fisher):

$$F = \frac{\frac{1}{G-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-G} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \frac{\frac{1}{G-1} \sum_{g=0}^{G-1} n_g (\bar{y}_g - \bar{y})^2}{\hat{\sigma}^2}$$

Sous \mathcal{H}_0 : $F \sim \mathcal{F}(G-1, n-G)$ loi de Fisher à $(G-1, n-G)$ degrés de liberté

Preuve: résultat admis⁽³⁾

Rem. : La difficulté principale est de prouver que

$\frac{1}{G-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ et $\frac{1}{n-G} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ sont indépendants (au sens des variables aléatoires).

⁽³⁾B. Delyon. "Régression". 2015. URL: <https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf>.

Suite du test de Fisher

Interprétation : F quantifie la variabilité observée des \bar{y}_i compte tenu de la variabilité mesurée par $\hat{\sigma}$

- ▶ \bar{y} est l'estimateur obtenu sous l'hypothèse nulle
- ▶ le numérateur $\frac{1}{G-1} \sum_{g=0}^{G-1} n_g (\bar{y}_g - \bar{y})^2$ mesure la variabilité inter-groupe (pondérée par la taille des groupes)
- ▶ le dénominateur $\frac{1}{n-G} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ mesure la totale des prédictions.

Conclusion : rejeter \mathcal{H}_0 si F est grande.⁽⁴⁾ Valeurs critiques à α fixé / p -valeurs obtenues par la loi de Fisher à $(n-1, n-G)$ degrés de liberté (G =nombre de groupes)

Objectif : tester si le facteur a un impact (linéaire) sur la réponse

⁽⁴⁾ici le test est unilatéral: si la statistique est petite cela signifie que les moyennes par classes sont anormalement proches (!). Noter la différence avec le test de Fisher pour deux populations.

Application: exemple des souris

Y a-t-il une influence des fragments sur la masse des souris mâles ?

- ▶ Variable quantitative : masse des souris mâles
- ▶ Variable qualitative : type de fragment (141G6, 152F7, 230E8, 285E6, et “non trisomique”)

Exemple : Pour l'étude sur les souris mâles on trouve ⁽⁵⁾
 $F = 6.14$, ce qui correspond à une p-valeur associée de $9.72 \cdot 10^{-5}$.

Interprétation : la statistique F est très grande, sa p-valeur est extrêmement faible \implies rejet de \mathcal{H}_0 avec grande confiance.

Conclusion : on rejette \mathcal{H}_0 ; on admet que le patrimoine génétique a un impact sur la masse des souris mâles vues les données

Mais quelle fragment a un impact sur la masse des souris?

⁽⁵⁾ cf. notebook associé

Test d'impact d'une modalité du facteur g

Hypothèse nulle, \mathcal{H}_0 : " $\beta_g^* = 0$ "

Interprétation : pour $g \in \{1, \dots, G - 1\}$ il n'y a pas de différence entre les masses des souris non-trisomiques et celles du groupe g .

Rappel :

- ▶ on estime β_g^* par $\hat{\beta}_g = \bar{y}_g - \bar{y}_0$
- ▶ on estime la variance de l'erreur sur β_g^* par $\hat{\sigma}^2(\hat{\beta}_g) = \frac{\hat{\sigma}^2}{n_g} + \frac{\hat{\sigma}^2}{n_0}$

Rem. : pour le $g = 0$, ce test est peu pertinent car il consiste à tester l'hypothèse saugrenue "la masse de souris non-trisomique est nulle"

Test de signification du coefficient β_g^*

Test de nullité de β_g^* (hypothèse nulle: $\mathcal{H}_0 : \beta_g^* = 0$)

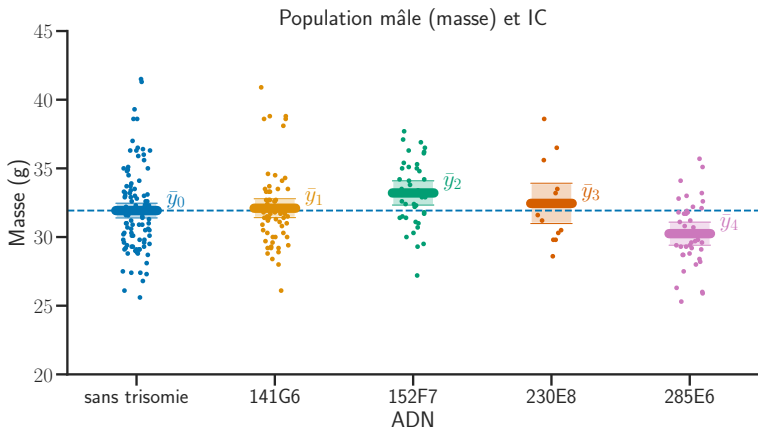
- ▶ Statistique de test : $\hat{T}_g = \frac{\hat{\beta}_g}{\hat{\sigma}(\hat{\beta}_g)}$
- ▶ Comportement sous: $\mathcal{H}_0 : T_g \sim t(n - G)$
- ▶ $\begin{cases} \text{Si } |T_g| \text{ est grande,} & \text{rejeter } \mathcal{H}_0 \\ \text{Sinon,} & \text{conserver } \mathcal{H}_0 \end{cases}$
- ▶ Valeur critique à α fixé (et p -value) obtenue par la loi de Student à $n - G$ degrés de liberté (G =nombre de groupes)

Exemple : $G = 5$ dans l'étude des souris, et on accepte l'hypothèse " $\beta_g = 0$ " pour $g = 1, 3$; on rejette cette hypothèse pour $g = 2, 4$ (cf. notebook pour les détails): le groupe 2 ('152F7') et le groupe 4 (285E6) ont des masses significativement différentes du groupe témoin.

Intervalle de confiance

IC au niveau $1 - \alpha$ pour β_g^* :

$[\hat{\beta}_g - \hat{\sigma}(\hat{\beta}_g)t_{1-\frac{\alpha}{2}, n-G}, \hat{\beta}_g + \hat{\sigma}(\hat{\beta}_g)t_{1-\frac{\alpha}{2}, n-G}]$ avec $t_{1-\frac{\alpha}{2}, n-G}$
 $1 - \frac{\alpha}{2}$ -quantile d'une loi de Student à $n - G$ degré de liberté



Biographie du jour : Susan P. Holmes⁽⁶⁾



- ▶ Statisticienne
- ▶ Doctorat à l'université de Montpellier (II)
- ▶ Travaille dans l'application de statistiques multivariées non paramétriques, de méthodes *bootstrap* et de visualisation de données à la biologie
- ▶ Auteure du livre “[Modern Statistics for Modern Biology](#)” avec Wolfgang Huber.

⁽⁶⁾https://en.wikipedia.org/wiki/Susan_P._Holmes

Bibliographie I

- ▶ Delyon, B. “Régression”. 2015. URL: <https://perso.univ-rennes1.fr/bernard.delyon/regression.pdf>.
- ▶ Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.