HLMA408: Traitement des données

cas d'un seul facteur

Joseph Salmon

http://josephsalmon.eu

Université de Montpellier

Sommaire

Sommaire

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 millions de personnes aux US.

Seuls les gènes "en bas" du chromosome 21 sont à l'origine de ce syndrome.

But de l'étude : comprendre quel(s) gène(s) sont responsables de cette 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 poids.

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 poids.

On espère que des comparaisons de poids 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 poids.

On espère que des comparaisons de poids 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

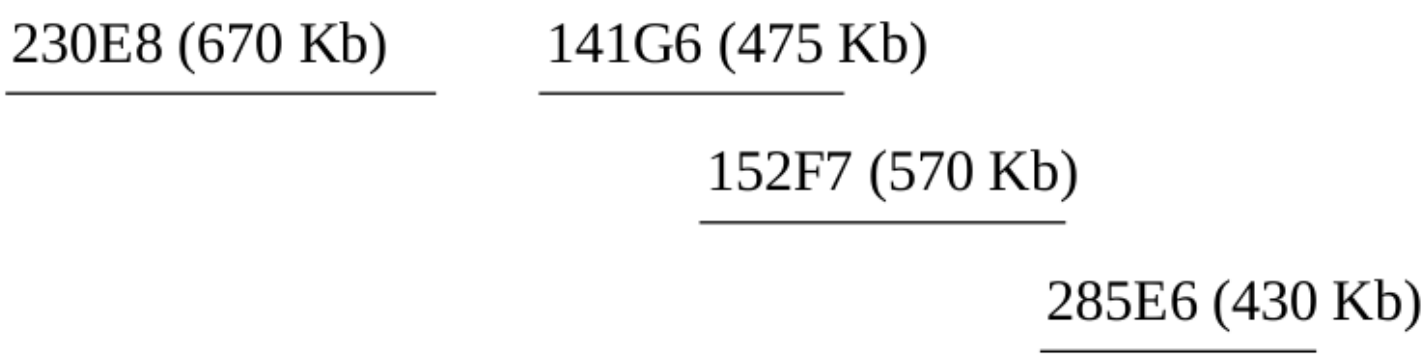
↓

On attend plusieurs générations...

↓

souris de notre échantillon

Les quatre fragments d'ADN sont



Variables

Sur les souris aveugles:

- **DNA**: le fragment d'ADN inséré dans la souris ancêtre avec les codes '1'=141G6, '2'=152F7, '3'=230E8 et '4'=285E6
- **Line** : la lignée
- **Transgenic** : (binaire 0 ou 1)
- **Sex** : (1= mâle, 0=femelle)
- **Age** : 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

Sommaire

Rappel sur la moyenne

Théorème

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

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

Ce que l'on note:

$$\overline{y}_n = \arg\min_{c \in \mathbb{R}} \varphi(c)$$

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

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

Moyennes sur deux populations et moindres carrés

But: estimer le poids moyen des souris mâles et femelles
Notation:

- y_i : poids de la i^{e} souris,
- $\bar{y}_M$: poids moyen des mâles
- $\bar{y}_F$: poids moyen 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

$$(\bar{y}_F, \bar{y}_M - \bar{y}_F) = \underset{(\beta_0,\beta_1)\in\mathbb{R}^2}{\arg\min}\ \varphi(\beta_0,\beta_1) := \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1\, \mathbb{1}_{M,i}) \Big)^2$$

Preuve

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \Big)^2$$

$$= \sum_{i:\text{mâle}} \Big(y_i - (\beta_0 + \beta_1) \Big)^2 + \sum_{i:\text{femelle}} \Big(y_i - \beta_0 \Big)^2$$

Les conditions nécessaires du premier ordre s'écrivent alors:

Preuve

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \Big)^2$$

$$= \sum_{i:\text{mâle}} \Big(y_i - (\beta_0 + \beta_1) \Big)^2 + \sum_{i:\text{femelle}} \Big(y_i - \beta_0 \Big)^2$$

Les conditions nécessaires du premier ordre s'écrivent alors:

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

Preuve

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \Big)^2$$

$$= \sum_{i:\text{mâle}} \Big(y_i - (\beta_0 + \beta_1) \Big)^2 + \sum_{i:\text{femelle}} \Big(y_i - \beta_0 \Big)^2$$

Les conditions nécessaires du premier ordre s'écrivent alors:

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

$$\iff \begin{cases} \sum_{i:\text{femelle}} (y_i - \hat{\beta}_0) = 0 \\ \sum_{i:\text{mâle}} (y_i - \hat{\beta}_0 - \hat{\beta}_1) = 0 \end{cases}$$

Preuve

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \Big)^2$$

$$= \sum_{i:\text{mâle}} \Big(y_i - (\beta_0 + \beta_1) \Big)^2 + \sum_{i:\text{femelle}} \Big(y_i - \beta_0 \Big)^2$$

Les conditions nécessaires du premier ordre s'écrivent alors:

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

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

Preuve

$$\varphi(\beta_0, \beta_1) = \sum_{i=1}^{n} \Big(y_i - (\beta_0 + \beta_1 \mathbb{1}_{M,i}) \Big)^2$$

$$= \sum_{i:\text{mâle}} \Big(y_i - (\beta_0 + \beta_1) \Big)^2 + \sum_{i:\text{femelle}} \Big(y_i - \beta_0 \Big)^2$$

Les conditions nécessaires du premier ordre s'écrivent alors:

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

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

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

Variance

Théorème

Sous l'hypothèse d'un modèle gaussien $y_i = \beta_0^\star + \beta_1^\star \mathbb{1}_{M,i} + \varepsilon_i$ avec $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$

$$\mathbb{V}\mathrm{ar}(\hat{\beta}_0) = \mathbb{V}\mathrm{ar}(\bar{y}_F) = \frac{\sigma^2}{n_F}$$

$$\mathbb{V}\mathrm{ar}(\hat{\beta}_1) = \mathbb{V}\mathrm{ar}(\bar{y}_M - \bar{y}_F) = \frac{\sigma^2}{n_F} + \frac{\sigma^2}{n_M}$$

Enfin un estimateur sans biais de la variance est alors:

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

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

Généralisation

5 possibilités 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

5 possibilités 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 :

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

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

Formulation multivariée

On minimise alors

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

Formulation multivariée

On minimise alors

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

$$= \sum_{i=1}^{n} \left(y_i - [\beta_0 + \sum_{g=1}^{4} \beta_g \mathbb{1}_{g,i}]\right)^2$$

Formulation multivariée

On minimise alors

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

Solution du problème

En adaptant la preuve pour le cas de deux modalités, on obtient que le minimum de cette fonction φ 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 **modalités** (les valeurs possibles de la variable) sont primordiales

Diagrammes en violons

Comparaison de diagrammes de violons pour les poids (en grammes) des souris mâles, regroupées en sous-populations suivant la partie du chromosome 21 dans leur ADN.

Sommaire

Retour sur les moyennes

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

Estimation: les quantités calculées estiment les paramètres ci-dessous

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

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

Rem: la modalité "0" (non trisomique) est ici la **modalité de référence**

Estimateur / Prédiction

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

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

Interprétation: estimateur donné par le modèle pour l'observation i, correspond à la moyenne de la classe associée à i, *i.e.*,

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

Retour sur les variances

Sous l'hypothèse d'un modèle gaussien

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

$$\begin{cases} \mathbb{V}\text{ar}(\hat{\beta}_0) &= \frac{\sigma^2}{n_0} \\ \mathbb{V}\text{ar}(\hat{\beta}_1) &= \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_0} \\ \mathbb{V}\text{ar}(\hat{\beta}_2) &= \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_0} \\ \mathbb{V}\text{ar}(\hat{\beta}_3) &= \frac{\sigma^2}{n_3} + \frac{\sigma^2}{n_0} \\ \mathbb{V}\text{ar}(\hat{\beta}_4) &= \frac{\sigma^2}{n_4} + \frac{\sigma^2}{n_0} \end{cases}$$

avec

$n_0 =$ 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 $y_i = \beta_0^\star + \beta_1^\star \mathbb{1}_{1,i} + \cdots + \beta_4^\star \mathbb{1}_{4,i} + \varepsilon_i$ avec $\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$ alors un estimateur sans biais de σ^2 est

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

i.e.,

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

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

Interprétation sous forme de modèle linéaire

"$\beta_1^\star = 0$" : signifie que dans le modèle, il n'y a pas de différence (en moyenne) entre les poids des souris non-trisomiques celles avec le fragment 141G6 (type 1).

Rappel :

- on estime $\beta_1^\star$ avec $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$
- on estime l'écart-type de l'erreur sur $\beta_1^\star$ avec $\hat{\sigma}_1^2 = \frac{\hat{\sigma}^2}{n_1} + \frac{\hat{\sigma}^2}{n_0}$

Tester "$\beta_1^\star$ est **significativement** différent de 0"
$\Longleftrightarrow$
vérifier que $\bar{y}_1 \neq \bar{y}_0$

Test de signification d'un coefficient

Procédure de test de nullité de $\beta_1^\star$:

- Statistique de test : $\hat{T}_1 = \dfrac{\hat{\beta}_1}{\hat{\sigma}_1}$
- Hypothèse nulle: $\mathcal{H}_0$: "$\beta_1^\star = 0$"
- Comportement sous: $\mathcal{H}_0$: $T_1 \sim t(n-G)$
- $\begin{cases} \text{Si } |T_1| \text{ est grande,} & \text{rejeter } \mathcal{H}_0 \\ \text{Sinon,} & \text{conserver } \mathcal{H}_0 \end{cases}$
- 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

Sommaire

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

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

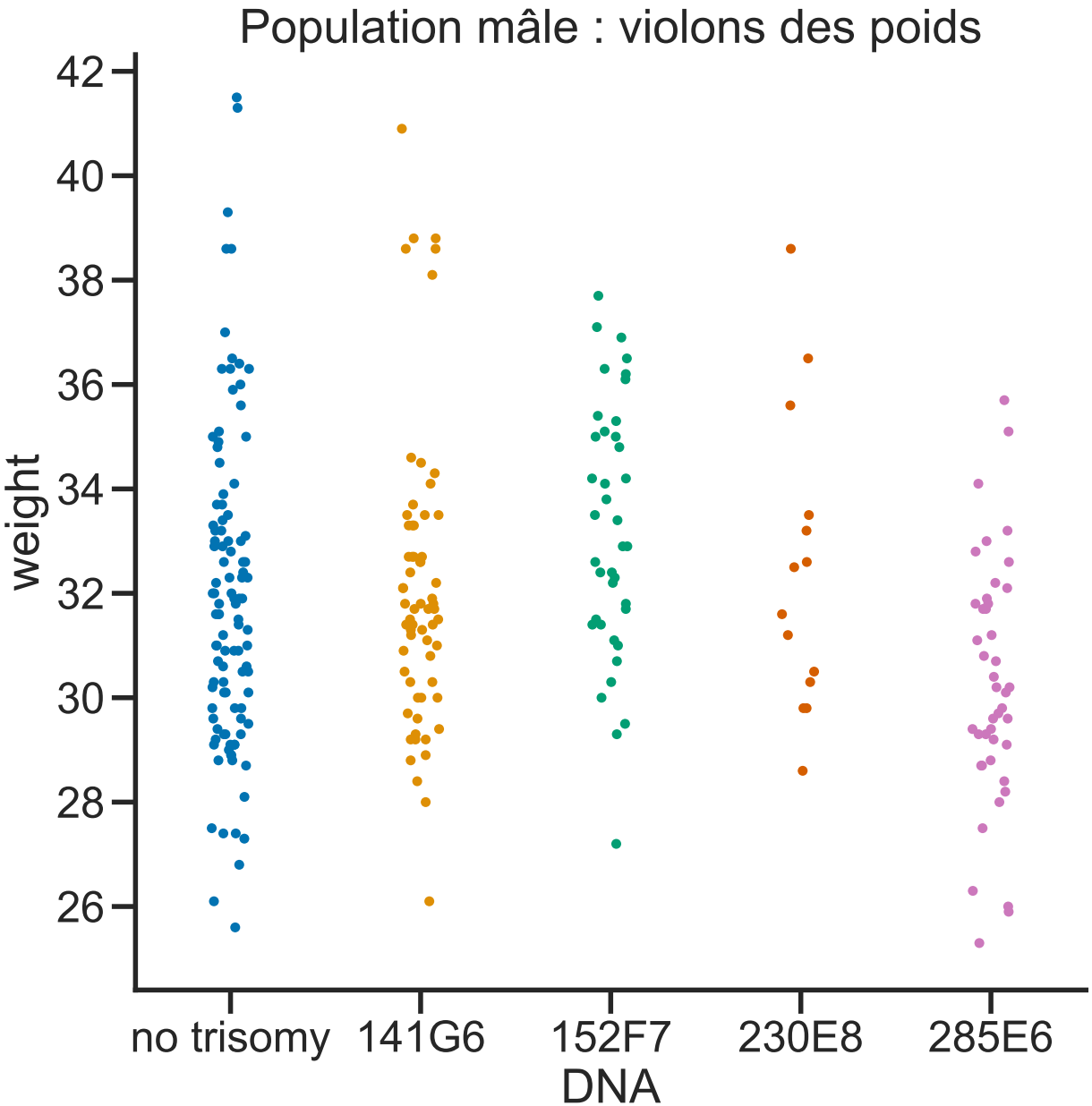
et de plus

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

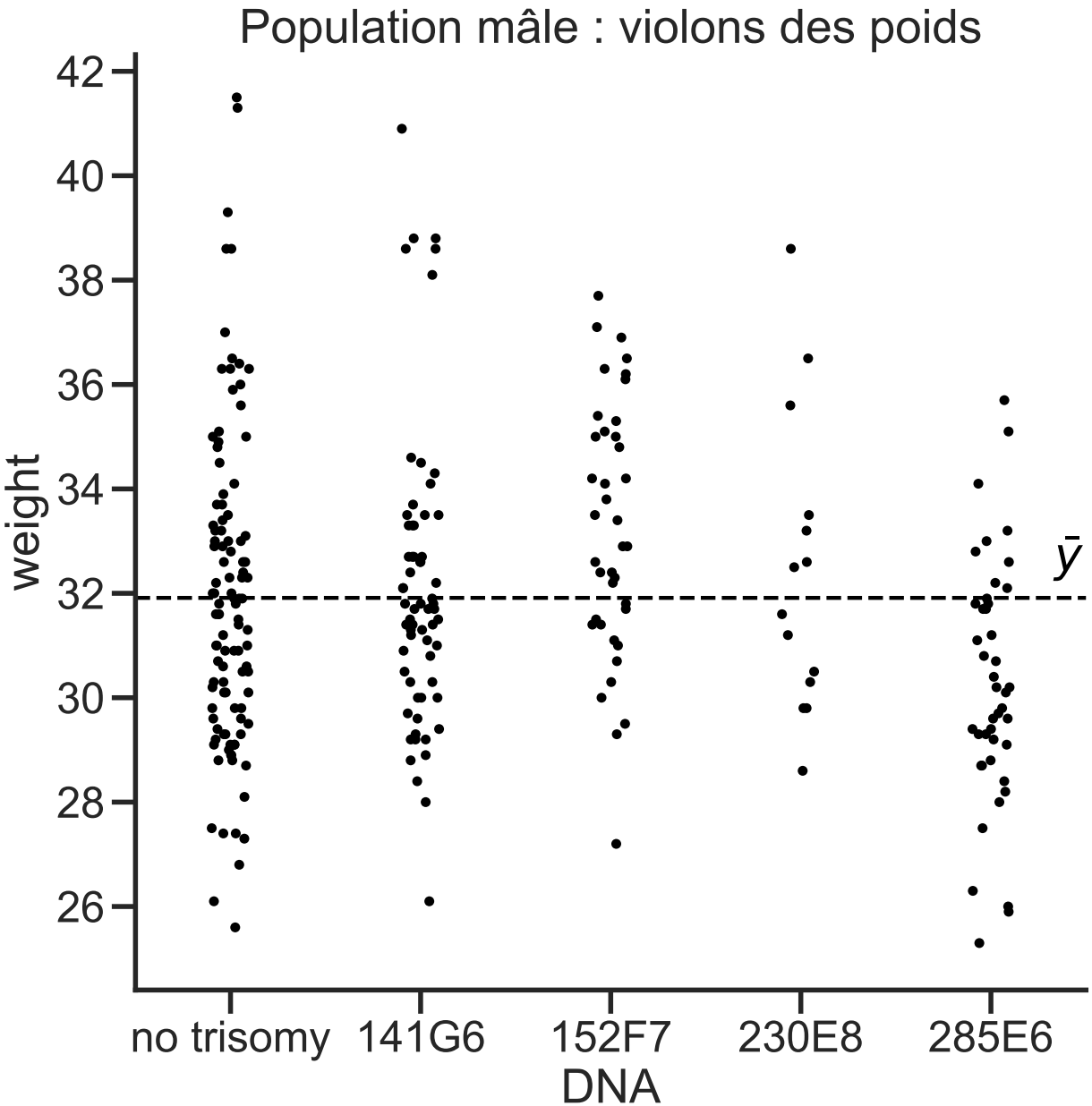
Ainsi,

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

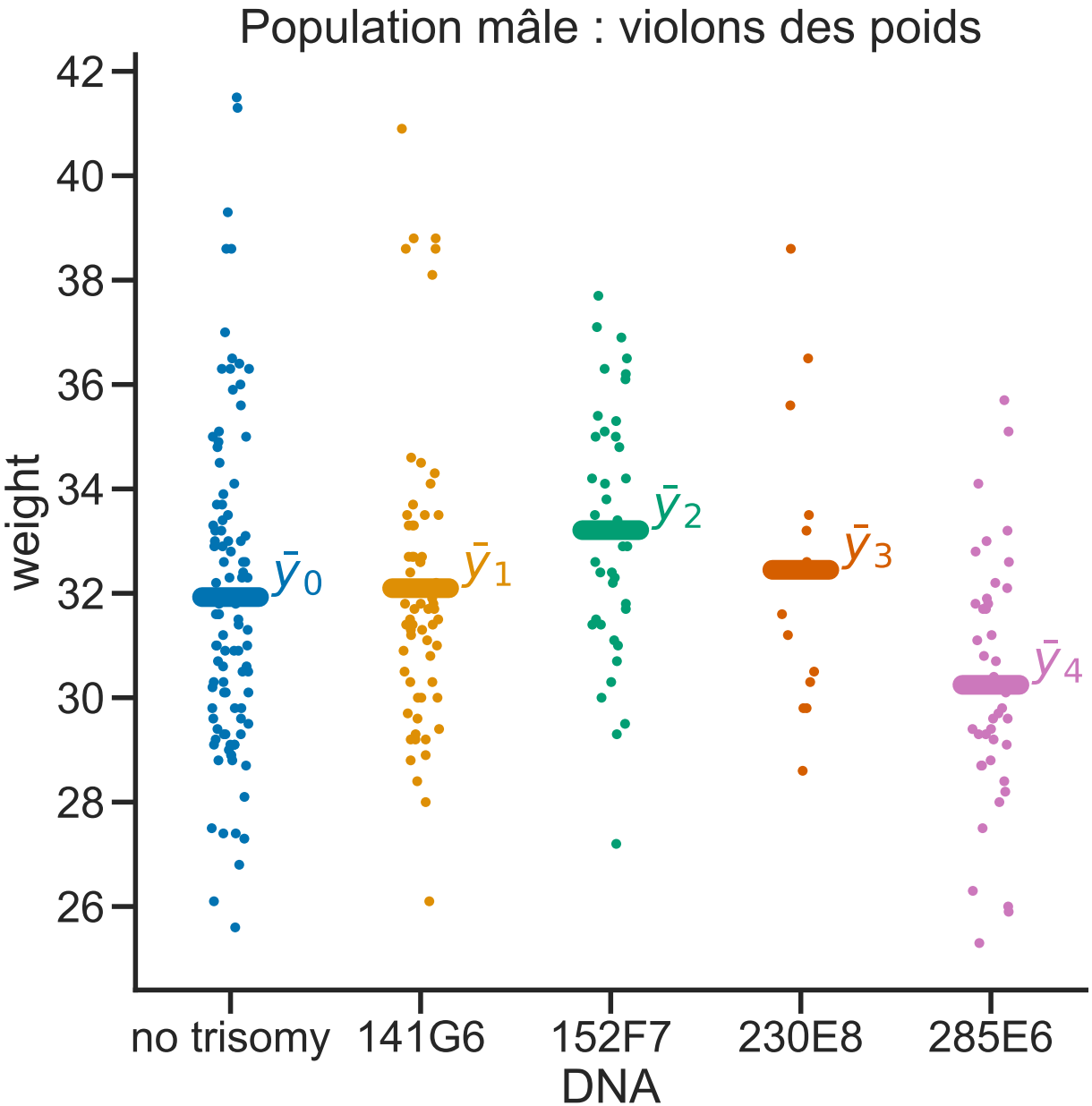
Données


Population mâle : violons des poids
weight
42
40
38
36
34
32
30
28
26
no trisomy
141G6
152F7
230E8
285E6
DNA

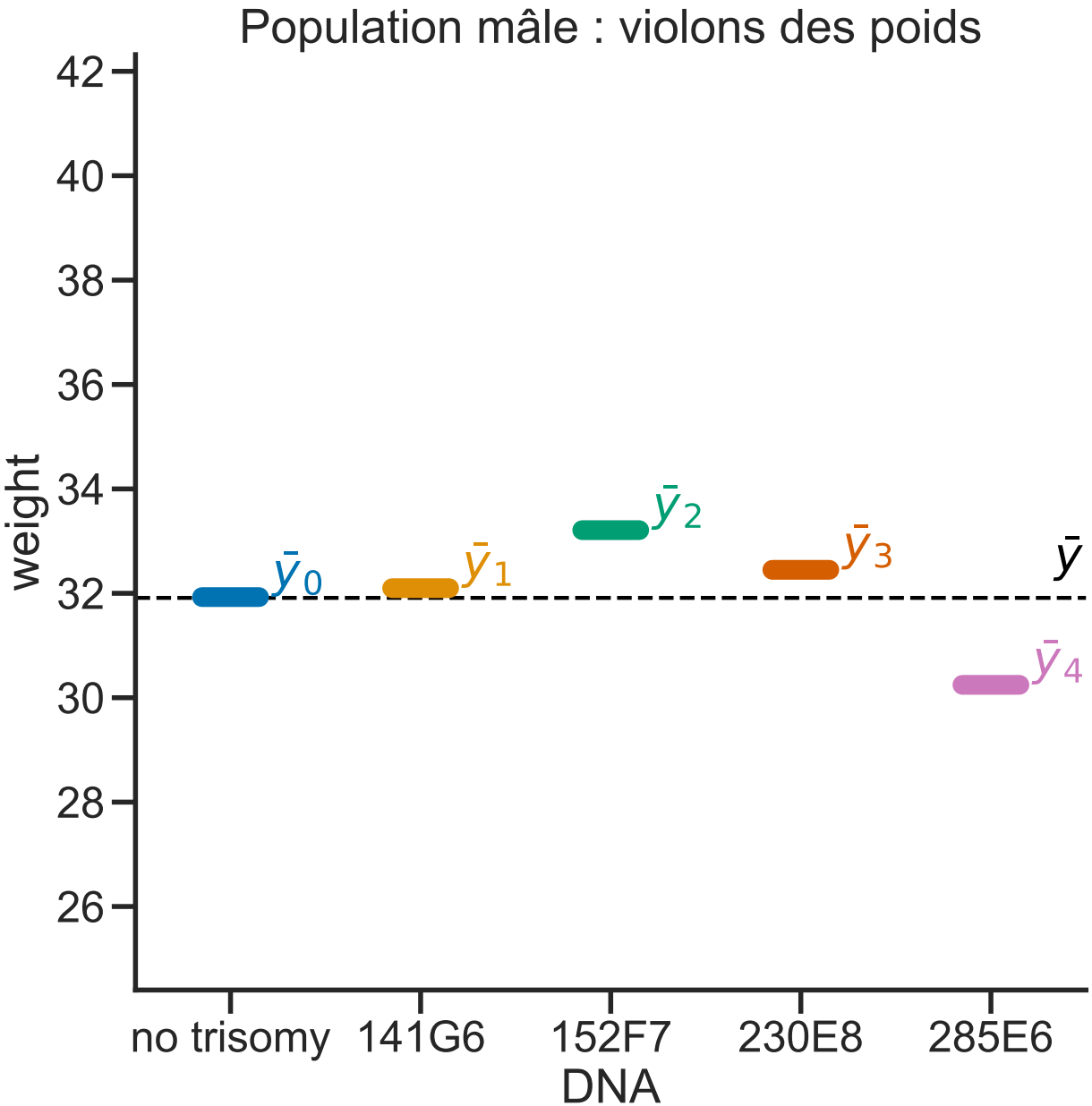
Variabilité totale



Variabilité intra-groupe



Variabilité inter-groupe



Test global de l'effet du facteur

$\mathcal{H}_0$: "$\beta_1^\star = \cdots = \beta_{G-1}^\star = 0$"$\Longleftrightarrow$"les groupes ont la même espérance"

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

Alors sous l'hypothèse nulle:

$$F \sim \mathcal{F}(G-1, n-G)$$

loi de Fisher à $(G-1, n-G)$ degrés de liberté [(2)]

[(2)] *cf.* leçon sur les gaussiennes

Suite

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. Valeur critique à α fixé (et p-valeur) obtenue par la loi de Fisher à $(n-1, n-G)$ degrés de liberté (G=nombre de groupes)

Bibliographie I

- Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.