Analyse de variance (ANOVA)

- Modèle d'ANOVA à un facteur
  - Définition et lien avec la régression multiple
  - Les théorèmes du chapitre 1 revisités
    - Variance expliquée, variance résiduelle
    - Tests d'hypothèses linéaires
  - Reparamétrisation(s)
- Modèle d'ANOVA à 2 facteurs
  - Problématique
  - Modélisation
  - Reparamétrisation(s)
  - Tests d'hypothèses linéaires

Problématique de l'ANOVA

- Comment modéliser l'effet d'une variable «symbolique» (qualitative / facteur) *x* sur E[*Y*] ?
- Exemple :

A	26 27 35 36 38 38 41 42 45 50 65
B	26 26 30 30 33 36 38 38 39 46 47 51 51 56 75
C	29 42 44 44 45 48 48 50 56 56 58 58 60 61 63 63 69

Tab. 2 – Crise d'asthme

y : durée en jours entre 2 crises d'asthme
x : médicament $(\in\{A, B, C\})$

- Dans la régression linéaire simple, $\hat{B}_1 = \frac{\frac{1}{n}\sum_i (x_i - \bar{x}_n)(y_i - \bar{y}_n)}{\frac{1}{n}\sum_i (x_i - \bar{x}_n)^2}$

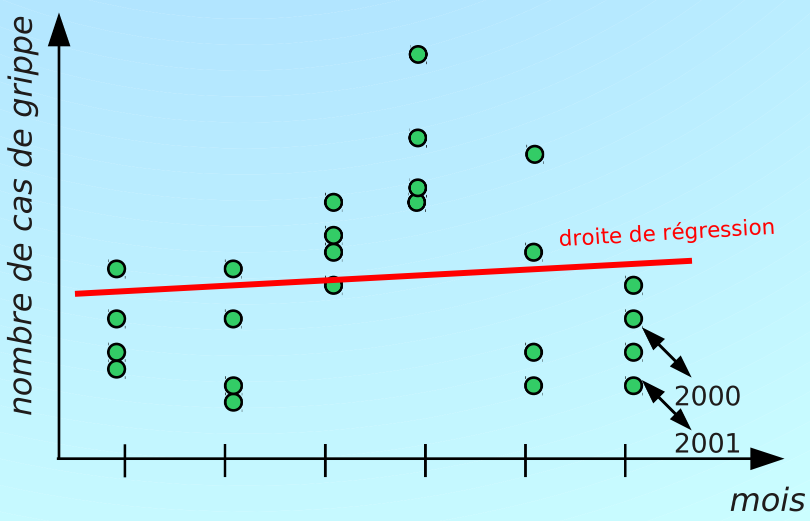
Nécessite que *x* soit quantitative $(x \in \mathbb{R})$

ANOVA ou régression linéaire simple ?

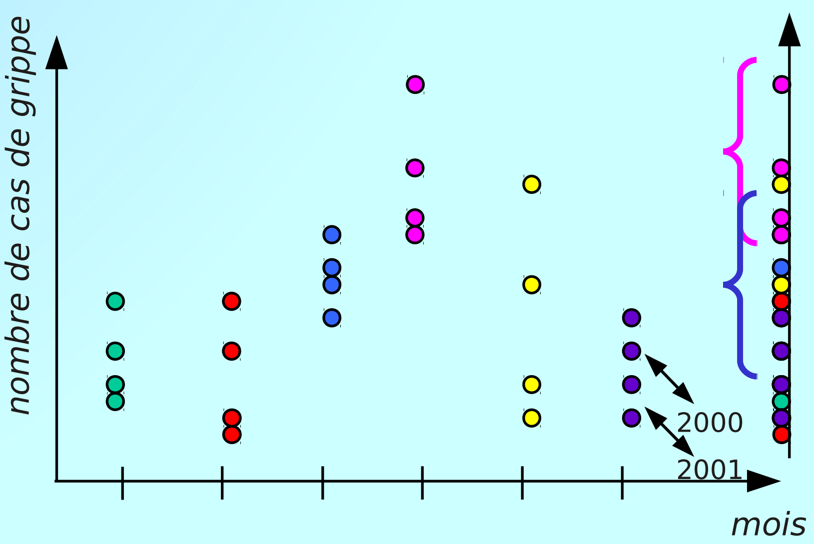
– Comment la fréquence de la gripppe dépend-elle du mois de l'année ?

– Représentations graphiques correspondant à :

Une régression linéaire simple



Une ANOVA



Modélisation

- $Y_l = m_{x_l} + \varepsilon_l$ où x_l est la valeur du facteur pour la $l^{\text{ème}}$ observation ($1 \leq x_l \leq p$)
- On réécrit le modèle en regroupant les observations avec même x_l; soit i une valeur du facteur x, n_i le nombre d'observations où x=i, et $(Y_{ij})_{j=1,\ldots,n_i}$ le vecteur de ces observations.

  Le modèle $Y_{ij} = m_i + \varepsilon_{ij}$ traduit le fait que E[Y] dépend de la valeur de x (=i)
- Ajout d'une hypothèse gaussienne (estimation, tests,...) :

$$\varepsilon_{ij} \text{ échantillon } \mathcal{N}(0, \sigma^2)$$

- Finalement c'est simplement un modèle avec p échantillons gaussiens mutuellement indépendants.

Analyse de variance et régression multiple

- On peut encore réécrire le modèle en considérant x_l comme un vecteur indicateur

$i^{\text{ème}}$ coordonnée

Ex. $\begin{pmatrix}1 & 0 & 0\end{pmatrix}$ si x=A $\quad \begin{pmatrix}0 & 1 & 0\end{pmatrix}$ si x=B $\quad \begin{pmatrix}0 & 0 & 1\end{pmatrix}$ si x=C $\quad \begin{pmatrix}0 & \cdots & 0 & 1 & 0 & \cdots & 0\end{pmatrix}$ si $x=i$

$\left(i \in \{A, B, C\}\right)$ cas général

$$\boxed{\Rightarrow Y = X\beta + \varepsilon}$$

$$Y = \begin{pmatrix} y_{11} \\ \vdots \\ y_{p n_p} \end{pmatrix} \qquad \beta = \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix} \qquad X = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \varepsilon = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{p n_p} \end{pmatrix}$$

- On se ramène alors au chapitre 1

Estimation

- Deux idées :
  - a) « intuitivement », estimer m_i par la moyenne empirique $\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij}$ des observations dont le facteur vaut *i;*
  - b) utiliser l'écriture du modèle sous la forme $Y = X\beta + \varepsilon$ et utiliser le $\hat{B} = ({}^t X X)^{-1} {}^t X Y$ du chapitre 1 (moindres carrés).

Les méthodes a) et b) coïncident :

pour toutes les valeurs de β et Y	pour certaines valeurs de β et Y	jamais	autres réponses

À faire par groupes de deux ou trois étudiants.
Durée : 10 min

Estimation par moindres carrés

Décomposition de la variance

$$S_Y^2 = S_{\hat{Y}}^2 + S_{Y|x}^2$$

variance empirique ou
« variance totale »

« variance expliquée »
ou « factorielle »

$$S_F^2 = \frac{1}{n} \sum_{i=1}^{p} n_i (\bar{Y}_{i.} - \bar{Y})^2$$

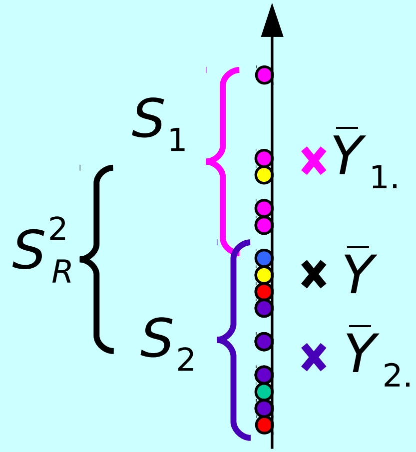
quantifie la séparation
entre les $\bar{Y}_{i.}$, i=1,...,p

Rappels : $\bar{Y}_{i.} = \hat{m}_i = \hat{B}_i$

$$\bar{Y} = \bar{Y}_{..}$$

« variance résiduelle »

$$S_R^2 = \frac{1}{n} \sum_{i=1}^{p} n_i S_i^2 = \frac{n-p}{n} \hat{\sigma}^2$$



Tests d'hypothèses linéaires

- Exemple du test de pertinence de l'ANOVA :

  y-a-t-il un effet de x sur (l'espérance de) Y ?

  y-a-t-il des mois où le virus de la grippe est plus virulent ?

On ne peut pas se contenter de comparer les moyennes empiriques $\overline{Y}_{i.}$

- Test de H0 : « $m_1 = \ldots = m_p$ » contre H1 : « $\exists i_1 \neq i_2,\ m_{i_1} \neq m_{i_2}$ »

  Sous H0, $\beta = m \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$

$$W = \left\{ \frac{n-p}{p-1} \frac{S_F^2}{S_R^2} > F^{-1}_{\mathcal{F}(p-1, n-p)}(1-\alpha) \right\}$$

Revient à comparer la séparation des $Y_{i.}$ $\rightarrow S_F^2$

avec la dispersion des Y_{ij} autour des $Y_{i.}$ $\rightarrow S_R^2$

L'ANOVA sous R

```
Call:
> Asthme.ANOVA <- aov(Duree ~ Medicament, Asthme)
> summary(Asthme.ANOVA)
            Df Sum Sq Mean Sq F value   Pr(>F)
Medicament   2 1398.6   699.3  5.3523 0.008715 **
Residuals   40 5226.0   130.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables(Asthme.ANOVA, type = "means")
Tables of means
Grand mean

45.55814

 Medicament
        A     B     C
    40.27 41.47 52.59
rep 11.00 15.00 17.00
```

commande / formule définissant le modèle et les données

L'ANOVA sous R

```
Call:
> Asthme.ANOVA <- aov(Duree ~ Medicament, Asthme)
> summary(Asthme.ANOVA)
            Df Sum Sq Mean Sq F value   Pr(>F)
Medicament   2 1398.6   699.3  5.3523 0.008715 **
Residuals   40 5226.0   130.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables(Asthme.ANOVA, type = "means")
Tables of means
Grand mean

45.55814

 Medicament
       A     B     C
   40.27 41.47 52.59
rep 11.00 15.00 17.00
```

nombre p de modalités du facteur-1

$n\,S_F^2$: somme de carrés factorielle

c_{MF} : carrés moyens factoriels $n\,S_F^2\,/\,p$ -1

$c_{MF}\,/\,c_{MR}$: statistique du test de pertinence

p-valeur du test de pertinence

c_{MR}: carrés moyens résiduels $n\,S_R^2\,/\,n\text{-}p$

$n\,S_R^2$: somme de carrés résiduelle

$n\text{-}p$

voir Section 2.3

Rappel

$$W=\left\{\frac{n-p}{p-1}\,\frac{S_F^2}{S_R^2}>F^{-1}_{\mathcal{F}(p-1,n-p)}(1-\alpha)\right\}$$

L'ANOVA sous R

```
Call:
> Asthme.ANOVA <- aov(Duree ~ Medicament, Asthme)
> summary(Asthme.ANOVA)
            Df Sum Sq Mean Sq F value   Pr(>F)
Medicament   2 1398.6   699.3  5.3523 0.008715 **
Residuals   40 5226.0   130.7
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> model.tables(Asthme.ANOVA, type = "means")
Tables of means
Grand mean

45.55814

 Medicament
       A     B     C
    40.27 41.47 52.59
rep 11.00 15.00 17.00
```

$\overline{Y}_{..}$

`summary` ne donne pas les estimations. Il faut utiliser une autre commande : **`model.tables`**

modalités du facteur

estimations $\overline{Y}_{i.}$

effectifs n_i

Exemple : 2.3 / effet « prof »

```
> alldata<-read.table("pms_anova.data",header=T)

#E.g. first mark in the list

> alldata[1,]

  marks prof specialization

1  14.5    1            MMIS
```

Construction du jeu de données

Analyse de variance

```
> pms.anova <- aov(lm(marks~prof,data=alldata))

> model.tables(pms.anova, type = "means")

Tables of means

11.91

 prof

          1     2     3

      11.88 11.32 12.61

rep 32.00 36.00 32.00

> summary(pms.anova)

            Df  Sum Sq Mean Sq F value Pr(>F)

prof         1    8.63  8.6289   0.6443 0.4241

Residuals   98 1312.56 13.3935

---

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

$$nS_F^2 = \sum_{i=1}^{p} n_i (\bar{Y}_{i.} - \bar{Y})^2$$

Exercice : retrouver les résultats de **R** ⇨

On donne :

```
> c(var(alldata[alldata[,2]==i,1]); i=1,2,3
[1]    20.48387        8.316468        11.83443
```

À faire par groupes de deux ou trois étudiants.
Durée : 5 min

Qui y arrive ?

```
> alldata<-read.table("pms_anova.data",header=T)

#E.g. first mark in the list

> alldata[1,]

  marks prof specialization

1  14.5    1           MMIS

> alldata$prof <- as.factor(alldata$prof)
```

```
> pms.anova <- aov(lm(marks~prof,data=alldata))

> model.tables(pms.anova, type = "means")

Tables of means

11.91

 prof

        1     2     3

    11.88 11.32 12.61

rep 32.00 36.00 32.00

> summary(pms.anova)

            Df  Sum Sq Mean Sq F value Pr(>F)

prof         2   28.25  14.123  1.0596 0.3506

Residuals   97 1292.94  13.329

---

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

Reparamétrisation du modèle

- Pour mieux interpréter les paramètres du modèle vis-à-vis de certaines applications, on peut le reparamétrer comme suit :

$$m_i = E[Y|X=i]$$

  - paramétrisation habituelle «absolue» :
  - autres paramétrisations «relatives» à une référence μ :

$$Y_{ij} = \mu + \beta_i + \varepsilon_{ij} \quad \text{où } \forall i, \ \beta_i = m_i - \mu$$

écart à la référence

$$\beta' = \begin{pmatrix} \mu \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \in \mathbb{R}^{p+1} \text{ non isomorphe à } \mathbb{R}^p \ni \beta = \begin{pmatrix} m_1 \\ \vdots \\ m_p \end{pmatrix}$$

$\Rightarrow$ non-identifiabilité du modèle

  - il faut contraindre β' à être dans un hyperplan de $\mathbb{R}^{p+1}$
    - la référence est m_1 (choix courant), $\beta_1 = 0 \Rightarrow \mu = m_1$
    - la référence est un m moyen, $\frac{1}{p}\sum_i \beta_i = 0 \Rightarrow \mu = \frac{1}{p}\sum_i m_i$

Courses automobiles

- Pour optimiser ses performances en rallye, un pilote teste 2 prototypes, X2V34 et XZAC, et pour chacun d'eux, 3 types de pneus, PMC119R, ACM7 et RM2000. Puis il effectue, pour chacune des 6 possibilités, 3 courses chronométrées dont on connaît les temps. Les ANOVA 1 sur les prototypes et les pneus donnent :

```
> summary(aov(Temps ~ Proto, T))
            Df Sum Sq Mean Sq F value    Pr(>F)
Proto        1 3307.6  3307.6  47.496 3.625e-06
Residuals   16 1114.2    69.6
> model.tables(..., type = "means")
Tables of means
Grand mean
191.8889
Proto
 X2V34   XZAC
178.33 205.44
```

```
> summary(aov(Temps ~ Pneu, T))
            Df Sum Sq Mean Sq F value Pr(>F)
Pneu         2  227.1   113.6  0.4061 0.6734
Residuals   15 4194.7   279.6
> model.tables(..., type = "means")
Tables of means
Grand mean
191.8889
Pneu
   ACM7 PMC119R  RM2000
 187.00  193.33  195.33
```

Lui conseillez-vous la combinaison suivante ?

	ACM7	PMC119R	RM2000
X2V34			
XZAC			

Autres réponses

À faire par groupes de deux ou trois étudiants.
Durée : 5 min

Modèle d'ANOVA à 2 facteurs

- $Y_{ijk} = m_{ij} + \varepsilon_{ijk}$ où $1 \leq i \leq p ; 1 \leq j \leq q ; 1 \leq k \leq n_{ij}$

  ε_{ijk} échantillon $\mathcal{N}(0, \sigma^2)$

- $m_{ij} = E[Y | F_1 = i ; F_2 = j]$ où i est la valeur du 1er facteur F_1, j celle du 2ème facteur F_2

- Le nombre de mesures répétées telles que $F_1=i$ et $F_2=j$ est n_{ij}

- En fin de compte, Y dépend du produit cartésien de 2 facteurs discrets à p et q modalités, ce qui n'est pas conceptuellement différend d'un seul facteur à pq modalités.

- Tout l'intérêt est dans la reparamétrisation, les tests d'hypothèses et leur interprétation !

Reparamétrisation(s) de l'ANOVA2

- $Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$ où $\forall (i, j),\ m_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$
- Contraintes algébriques pour l'identifiabilité

$$\frac{1}{p}\sum_i \alpha_i = 0 ; \frac{1}{q}\sum_j \beta_j = 0 ;\ \forall j,\ \frac{1}{p}\sum_i \gamma_{ij} = 0 ;\ \forall i,\ \frac{1}{q}\sum_j \gamma_{ij} = 0$$

- Du coup $\mu = \frac{1}{pq}\sum_{i,j} m_{ij} ;\ \alpha_i = \frac{1}{q}\sum_j m_{ij} - \mu ;\ \beta_j = \frac{1}{p}\sum_i m_{ij} - \mu$
- Interprétation :
  - α_i : ce que $F_1 = i$ induit comme changement par rapport à μ
  - β_j : ce que $F_2 = j$ induit comme changement par rapport à μ
  - γ_{ij} : ce que $F_1 = i$ et $F_2 = j$ induit comme changement par rapport à $\mu + \alpha_i + \beta_j$

valeur attendue en prenant en compte F1 et F2 séparément

Trois tests d'ANOVA 2

- a) H0 : « $\forall (i, j), \ \gamma_{ij}=0$ » (H1 : le contraire)

Sous H0 il n'y a pas d'interaction entre les deux facteurs.

- b) H0 : « m_{ij} » ne dépend pas de *i* (H1 : le contraire)

soit H0 : « $\forall j, \ \forall i, \ m_{ij}=m_{1j}$ » autrement dit $\forall i, \ \forall j, \ \alpha_i=\gamma_{ij}=0$

Sous H0 le facteur F_1 n'a pas d'effet sur (l'espérance de) *Y*

- c) H0 : « m_{ij} » ne dépend pas de *j* (H1 : le contraire)

soit H0 : « $\forall i, \ \forall j, \ m_{ij}=m_{i1}$ » autrement dit $\forall j, \ \forall i, \ \beta_j=\gamma_{ij}=0$

Sous H0 le facteur F_2 n'a pas d'effet sur *Y*

- Noter que s'il y a une interaction alors les facteurs **ont** un effet, même si $\forall i, \ \alpha_i=0$ ou $\forall j, \ \beta_j=0$.

- Ce sont des tests d'hypothèses linéaires dont les régions critiques sont du type

$$W=\left\{\frac{n-pq}{pq-l}\frac{RSS_0-RSS_1}{RSS_1}>F^{-1}_{\mathcal{F}(pq-l,\,n-pq)}(1-\alpha)\right\}$$

l : nombre de «paramètres indépendants»

Retour sur l'exemple des rallyes

```
> temps.anova <- aov(Temps ~ Proto * Pneu, T)
> model.tables(temps.anova, type = "means")
Tables of means / Grand mean / 191.8889
Proto
 X2V34   XZAC
178.33 205.44
Pneu
   ACM7 PMC119R  RM2000
 187.00  193.33  195.33
 Proto:Pneu
       Pneu
Proto   ACM7    PMC119R RM2000
  X2V34 180.00 172.67  182.33
  XZAC  194.00 214.00  208.33
```

$\bar{Y}$

les $\bar{Y}_{i..} = \frac{1}{n_{i.}} \sum_{j,k} Y_{ijk}$

les $\bar{Y}_{.j.} = \frac{1}{n_{.j}} \sum_{i,k} Y_{ijk}$

les $\hat{M}_{ij} = \frac{1}{n_{ij}} \sum_{k} Y_{ijk}$

```
> summary(temps.anova)
            Df Sum Sq Mean Sq  F value    Pr(>F)
Proto        1 3307.6  3307.6 122.5021 1.184e-07 ***
Pneu         2  227.1   113.6   4.2058  0.041289 *
Proto:Pneu   2  563.1   281.6  10.4280  0.002374 **
Residuals   12  324.0    27.0
```

test d'effet de F_1
test d'effet de F_2
test d'interaction

À comparer avec l'ANOVA 1 - `Pneu : Pr(>F) 0.6734`