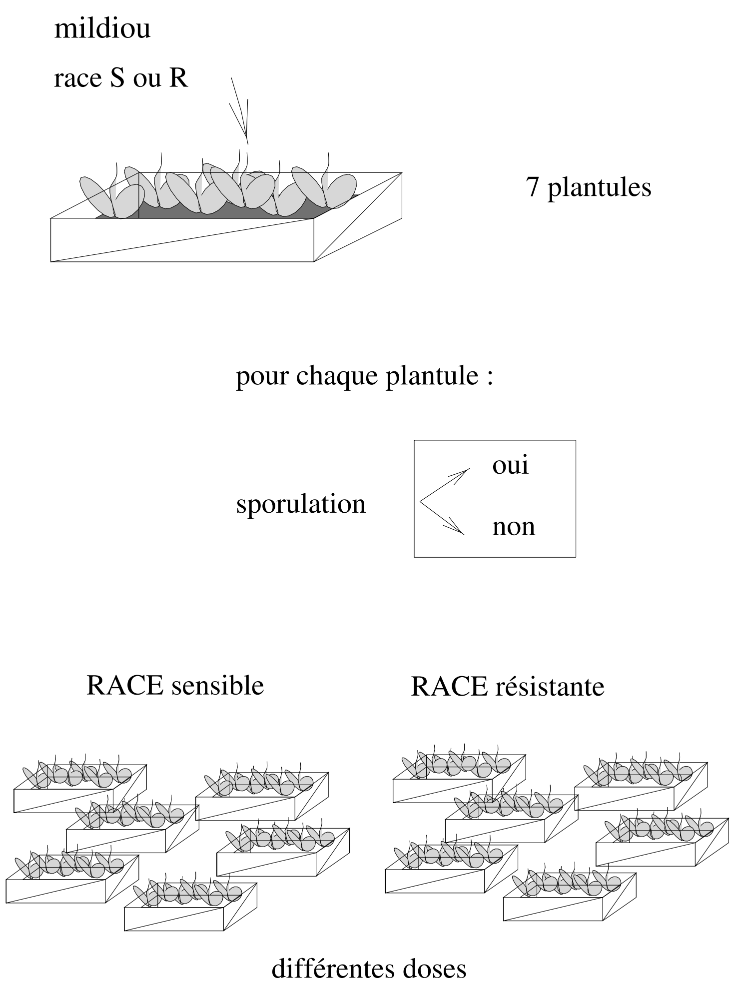
## Ex. Tournesol : l'expérience






Ce chapitre présente la modélisation d'un jeu de données, à l'aide d'un GLM, et l'interprétation des résultats, sans expliquer le pourquoi de la méthode employée. Des termes seront employés (déviance par exemple), et les notions associées seront utilisées. Ces termes ne seront définis qu'au chapitre suivant. Cette stratégie est volontaire. Le but est de montrer jusqu'où on peut aller, la méthode à employer, les nouveautés par rapport à des méthodes connues, comme celle du modèle linéaire. La théorie sous-jacente sera décortiquée au chapitre suivant.

- Les données de l'exemple "Tournesol" présenté ci-dessous ont été recueillies par l'équipe pathologie du tournesol, laboratoire INRA de pathologie et de mycologie de la station d'amélioration des plantes de Clermont-Ferrand. (Albourie JM, Tourvieille J, Tourvieille de Labrouhe D 1998: Resistance to metalaxyl in isolates of the sunflower pathogen Plasmopara halstedii. Europ. J. Plant Pathol. 104:235-242)

- on étudie l'**effet d'un traitement fongicide sur deux races de mildiou**, champignon parasite du tournesol. La race R est nouvelle, mutante, elle présente une résistance au traitement fongicide.

  Le fongicide est déposé sur les graines de tournesol, à différentes doses. Des groupes de 7 plantules sont inoculées avec l'une des races de mildiou. On compte le nombre de plantules présentant des symptômes de mildiou : c'est le **nombre de plantules pour lesquelles le champignon a sporulé**. Il sporule ou ne sporule pas, selon la plantule, sa race, et la dose de fongicide (metalaxyl).

## Ex. Tournesol : les données

| | race | dose | sp | n | 37 | S | 2e+01 | 0 | 7 | 74 | R | 2e-01 | 5 | 7 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | S | 1e-03 | 5 | 7 | 38 | S | 2e+01 | 0 | 7 | 75 | R | 2e-01 | 7 | 7 |
| 2 | S | 1e-03 | 6 | 7 | 39 | S | 2e+02 | 0 | 7 | 76 | R | 2e-01 | 7 | 7 |
| 3 | S | 1e-03 | 7 | 7 | 40 | S | 2e+02 | 0 | 7 | 77 | R | 2e-01 | 7 | 7 |
| 4 | S | 1e-03 | 7 | 7 | 41 | S | 2e+02 | 0 | 7 | 78 | R | 2e-01 | 7 | 7 |
| 5 | S | 1e-03 | 7 | 7 | 42 | S | 2e+02 | 0 | 7 | 79 | R | 2e+00 | 7 | 7 |
| 6 | S | 1e-03 | 5 | 7 | 43 | S | 2e+02 | 0 | 7 | 80 | R | 2e+00 | 7 | 7 |
| 7 | S | 2e-02 | 7 | 7 | 44 | S | 2e+02 | 0 | 7 | 81 | R | 2e+00 | 7 | 7 |
| 8 | S | 2e-02 | 7 | 7 | 45 | S | 2e+02 | 0 | 7 | 82 | R | 2e+00 | 7 | 7 |
| 9 | S | 2e-02 | 7 | 7 | 46 | S | 2e+02 | 0 | 7 | 83 | R | 2e+00 | 7 | 7 |
| 10 | S | 2e-02 | 6 | 7 | 47 | S | 2e+03 | 0 | 7 | 84 | R | 2e+00 | 7 | 7 |
| 11 | S | 2e-02 | 7 | 7 | 48 | S | 2e+03 | 0 | 7 | 85 | R | 2e+00 | 7 | 7 |
| 12 | S | 2e-02 | 5 | 7 | 49 | S | 2e+03 | 0 | 7 | 86 | R | 2e+00 | 7 | 7 |
| 13 | S | 2e-02 | 7 | 7 | 50 | S | 2e+03 | 0 | 7 | 87 | R | 2e+01 | 7 | 7 |
| 14 | S | 2e-02 | 7 | 7 | 51 | S | 2e+03 | 0 | 7 | 88 | R | 2e+01 | 7 | 7 |
| 15 | S | 2e-01 | 5 | 7 | 52 | S | 2e+03 | 0 | 7 | 89 | R | 2e+01 | 7 | 7 |
| 16 | S | 2e-01 | 4 | 7 | 53 | S | 2e+03 | 0 | 7 | 90 | R | 2e+01 | 6 | 7 |
| 17 | S | 2e-01 | 6 | 7 | 54 | S | 2e+03 | 0 | 7 | 91 | R | 2e+01 | 7 | 7 |
| 18 | S | 2e-01 | 2 | 7 | 55 | R | 1e-03 | 7 | 7 | 92 | R | 2e+01 | 6 | 7 |
| 19 | S | 2e-01 | 7 | 7 | 56 | R | 1e-03 | 7 | 7 | 93 | R | 2e+01 | 7 | 7 |
| 20 | S | 2e-01 | 7 | 7 | 57 | R | 1e-03 | 7 | 7 | 94 | R | 2e+01 | 7 | 7 |
| 21 | S | 2e-01 | 4 | 7 | 58 | R | 1e-03 | 7 | 7 | 95 | R | 2e+02 | 7 | 7 |
| 22 | S | 2e-01 | 3 | 7 | 59 | R | 1e-03 | 7 | 7 | 96 | R | 2e+02 | 7 | 7 |
| 23 | S | 2e+00 | 1 | 7 | 60 | R | 1e-03 | 7 | 7 | 97 | R | 2e+02 | 6 | 7 |
| 24 | S | 2e+00 | 2 | 7 | 61 | R | 1e-03 | 7 | 7 | 98 | R | 2e+02 | 5 | 7 |
| 25 | S | 2e+00 | 0 | 7 | 62 | R | 1e-03 | 7 | 7 | 99 | R | 2e+02 | 6 | 7 |
| 26 | S | 2e+00 | 1 | 7 | 63 | R | 2e-02 | 7 | 7 | 100 | R | 2e+02 | 7 | 7 |
| 27 | S | 2e+00 | 1 | 7 | 64 | R | 2e-02 | 7 | 7 | 101 | R | 2e+03 | 1 | 7 |
| 28 | S | 2e+00 | 0 | 7 | 65 | R | 2e-02 | 7 | 7 | 102 | R | 2e+03 | 0 | 7 |
| 29 | S | 2e+00 | 0 | 7 | 66 | R | 2e-02 | 7 | 7 | 103 | R | 2e+03 | 0 | 7 |
| 30 | S | 2e+00 | 1 | 7 | 67 | R | 2e-02 | 7 | 7 | 104 | R | 2e+03 | 0 | 7 |
| 31 | S | 2e+01 | 0 | 7 | 68 | R | 2e-02 | 7 | 7 | 105 | R | 2e+03 | 0 | 7 |
| 32 | S | 2e+01 | 0 | 7 | 69 | R | 2e-02 | 7 | 7 | 106 | R | 2e+03 | 0 | 7 |
| 33 | S | 2e+01 | 0 | 7 | 70 | R | 2e-02 | 7 | 7 | 107 | R | 2e+03 | 2 | 7 |
| 34 | S | 2e+01 | 0 | 7 | 71 | R | 2e-01 | 7 | 7 | 108 | R | 2e+03 | 2 | 7 |
| 35 | S | 2e+01 | 0 | 7 | 72 | R | 2e-01 | 7 | 7 | | | | | |
| 36 | S | 2e+01 | 0 | 7 | 73 | R | 2e-01 | 7 | 7 | | | | | |

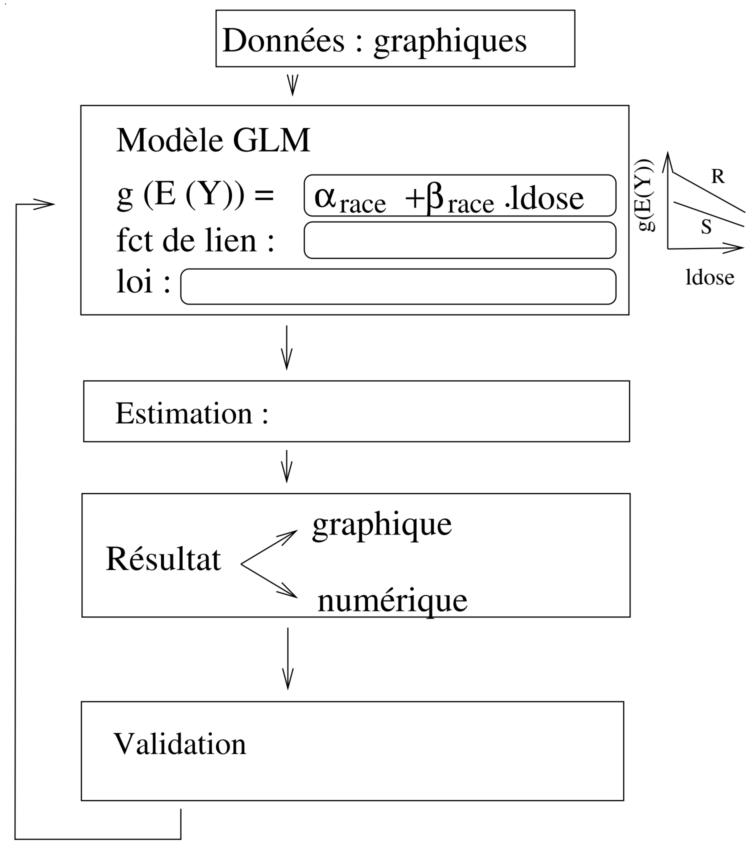


Avant même de visualiser les données, remarquons que les observations brutes sont des **observations binaires**. Le mildiou sporule ou ne sporule pas. Si l'on regroupe les observations brutes par lot de 7 plantules, on obtient des **données groupées**, des **proportions**. Les **observations binaires** sont supposées toutes **indépendantes** les unes des autres (le fait que le champignon sporule sur une plantule n'a pas d'influence sur le fait qu'il sporule sur la plantule voisine). Elles sont supposées suivre la **même loi de Bernoulli** (même probabilité, dans les mêmes conditions de race et de dose). Les données groupées suivent donc des **lois binomiales**.

Toutes les observations sont reportées. On note $n_i$ le nombre de plantules, il est toujours égal à 7. Dans d'autres expériences, on peut avoir des $n_i$ qui ne sont pas tous égaux. On trouve dans la colonne “sp” le nombre de plantules pour lesquelles il y a eu sporulation, parmi les 7 plantules testées.

Remarque : en regroupant les observations on a défini une nouvelle unité statistique qui est le lot. Le taux de sporulation d'un lot est jugé plus intéressant que la réponse individuelle d'une plantule. La réponse individuelle d'une plantule n'apporte aucune information par rapport à l'information apportée par le lot auquel elle appartient : une plantule d'un lot ne peut être distinguée d'une plantule du même lot. On a alors intérêt à étudier le taux de sporulation du lot plutôt que le taux individuel. On a 108 observations binomiales. (on notera $m$ le nombre d'observations binomiales ici $m$=108).

## Ex. Tournesol : schéma d'analyse






Ce schéma d'analyse est très général. Il peut être appliqué à toute modélisation. Seule la boîte "modèle" est caractéristique du modèle linéaire généralisé.

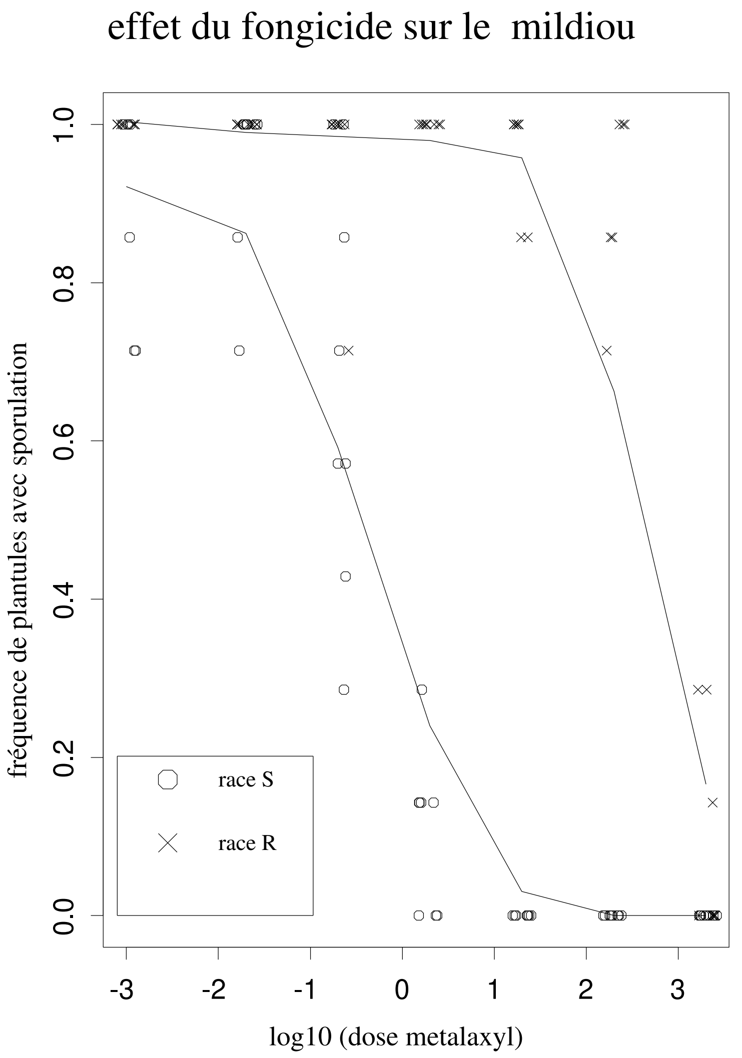
Ce schéma d'analyse peut être écrit au tableau, et rester affiché tout au long des chapitres 1 et 2. On s'y réfèrera lors des différentes phases d'analyse. On le complètera au fur et à mesure, en précisant les fonctions de lien (logit, probit, cloglog, ... ), lois (binomiale, poisson, gamma, normale, multinomiale, ... ), et la méthode d'estimation (méthode du maximum de vraisemblance). Il pourra aussi être utilisé lorsqu'on reviendra au modèle linéaire, pour préciser la fonction de lien identité, la loi normale, la méthode d'estimation des moindres carrés.

Rq : la **fonction de lien** n'est absolument **pas une transformation** des données. Le modèle GLM n'est pas un modèle linéaire sur une transformée des $Y$ ; il décrit la liaison linéaire entre une transformée de l'espérance de $Y$ et des variables explicatives.

Rq : le terme "modèle linéaire" prête à confusion, car il peut être interprété au sens statistique de "modèle linéaire EN LES PARAMETRES", et au sens commun de "liaison d'allure linéaire". Lorsqu'il sera employé, il faudra préciser son sens.

- La première étape "Données" est TOUJOURS une représentation graphique des données : diagramme de corrélation, lissage empirique...
- Le modèle postulé suppose que, **pour chaque race**, la liaison entre ldose et Y peut être représentée par **une droite**, au travers d'une fonction de lien (transformation) appliquée à l'espérance de Y. Si on écrit, dans la partie systématique du modèle, un effet race, et une interaction entre le facteur race et le régresseur ldose (race+race:ldose), on suppose que le modèle se résume à une droite par race, et que les deux droites ont une ordonnée à l'origine et une pente qui leur sont propres.
- On estime les quatre paramètres du modèle (deux pentes, et deux ordonnées à l'origine). Il faut une méthode d'estimation.
- Des sorties graphiques et des résultats numériques sont indispensables, et complémentaires pour une interprétation correcte.
- La dernière étape est une étape de validation statistique pour juger la pertinence des choix de la fonction de lien, de la loi de Y et de la partie systématique du modèle. On utilisera notamment des représentations graphiques.

## Ex. Tournesol : exploration des données



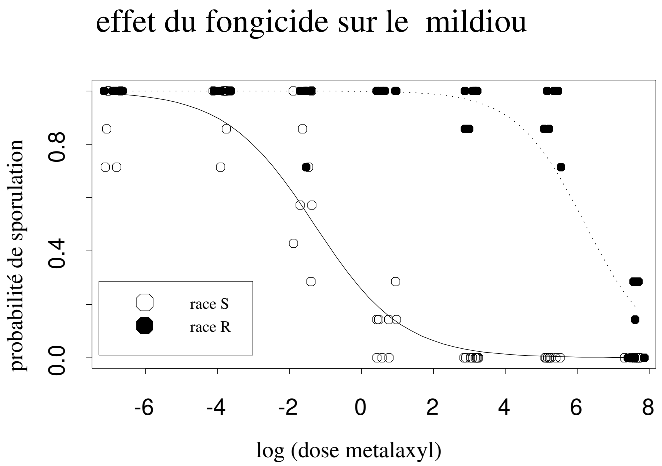


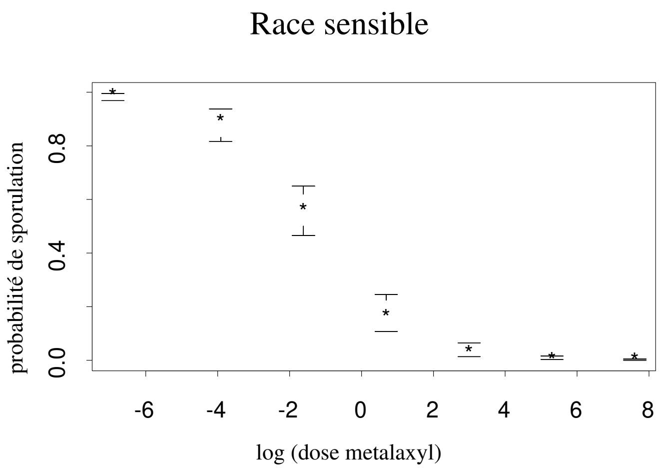

Les observations reportées (données groupées) sont des proportions, elles sont **bornées par 0 et 1**. La **liaison** entre la variable explicative log(dose) et les proportions semble **sigmoïdale**. La variance est hétérogène.

- La variable "nombre de plantules avec sporulation" suit une loi binomiale. Un modèle binomial est donc plus appropriée qu'un modèle gaussien. Ceci est un premier point, mais ce n'est pas du tout le point essentiel. La variance d'une loi $\mathcal{B}(n,p)$ vaut $np(1-p)$, elle dépend de $p$, elle n'est donc pas constante (on l'observe nettement sur le graphique). Les hypothèses d'homogénéité des variances et dans une moindre mesure de normalité, faites sur la partie aléatoire d'un modèle linéaire ne conviennent pas.
- D'autre part, un **modèle linéaire** sur l'espérance **n'est pas satisfaisant**, car certaines des probabilités estimées seraient extérieures à l'intervalle $[0,1]$. Un modèle satisfaisant cette dernière contrainte est recherché.
- On a choisi de porter en abscisse les **logarithmes des doses** pour deux raisons :
  - si on porte la dose en abscisse, on visualise très mal la liaison, parce que les observations sont tassées. D'une suite en progression géométrique on passe par la transformation log à une suite en progression arithmétique. $(\log_{10}(2 \times 10^k) = \log_{10} 2 + k)$
  - pour symétriser la distribution des seuils de tolérance c'est souvent le logarithme de la dose, et non pas la dose, qui est choisi comme variable explicative (ce point est approfondi plus loin).
- Les abscisses ont été légèrement perturbées, de façon à visualiser les nombreux points superposés.
- Dans la partie systématique du modèle, il faut prendre en compte un facteur, le **facteur race**, et un régresseur, la **variable log(dose)**.

# Ex. Tournesol : le modèle

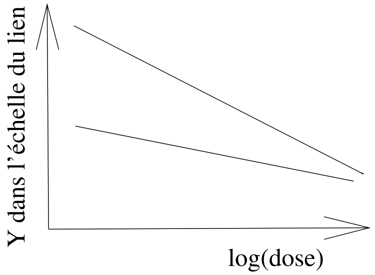
## GLM : $\log(\frac{p}{1-p}) = \alpha_{race} + \beta_{race}.ldose$






Le modèle estimé "race + race : ldose" peut être représenté par deux droites quelconques, dans l'échelle de la fonction de lien. La fonction de lien est définie plus tard. La fonction de lien utilisée ici est la fonction $\log(\frac{p}{1-p})$. C'est une fonction qui envoit $[0,1]$ dans $I\!R$. Ce n'est pas $E(Y)$ que l'on suppose lié linéairement à la dose, mais $g(E(Y))$.




On représente les observations et les courbes estimées.

- On a bien une **courbe sigmoïdale**.
- L'**ajustement** semble **bon dans l'ensemble**. On pourrait, dans le but d'améliorer l'estimation, remettre en cause le lien, ou la partie systématique du modèle.
- On représente, pour la race sensible S, les probabilités estimées, et les intervalles de confiance associés, à 95%. Les intervalles de confiance sont plus larges au centre qu'aux extrémités. **On a modélisé l'hétérogénéité des variances**. Ils sont dissymétriques, et d'autant plus dissymétriques que l'on s'approche des bornes 0 et 1.

## Ex. Tournesol : analyse

$$\text{GLM}: \log(\frac{p}{1-p}) = \alpha_{race} + \beta_{race}.ldose$$

```
Coefficients :
             Value Std. Error t value
     raceS -1.0651    0.21290  -5.003
     raceR  6.4821    0.89403   7.250
raceSldose -1.8688    0.20507  -9.113
raceRldose -2.4004    0.32796  -7.319

(Dispersion Parameter for Binomial family
          taken to be 1 )

  Null Deviance : 836.7 on 107 deg. of freedom

Residual Deviance : 131 on 104 deg. of freedom

Number of Fisher Scoring Iterations : 6
```



**Version S**

- $H_0$ : "$\log \frac{p}{1-p} = \alpha_{race} + \beta_{race}.ldose$" : "modèle adéquat".
  $H_1$ : "$\log \frac{p}{1-p} = \widetilde{p}$ : proportion observée par combinaison (race, dose)"
  La **déviance résiduelle** vaut 131. On peut interpréter cette déviance comme une mesure de l'écart du modèle aux données. Elle est bien plus faible que la déviance résiduelle du modèle réduit à un terme constant (modèle nul). On utilise la déviance résiduelle pour valider le modèle (**test d'adéquation du modèle**), parce que les nombres $\boldsymbol{n_i}$ **ne sont pas trop petits** (ici $n_i = 7$; il n'y a pas de règle absolue pour dire quand un effectif est suffisant!). On la compare à un $\chi^2$ à 104 degrés de liberté ($m-k$ : le nombre d'observations binomiales moins le nombre de paramètres = 108-4). Le $(1-\alpha)$ quantile d'un $\chi^2(104)$, avec $\alpha = 0.05$, vaut 128.8. Si l'on choisit un risque $\alpha = 0.05$, on rejette l'hypothèse nulle d'adéquation du modèle (p=4%). Le modèle n'est pas tout à fait adéquat, mais il n'est "pas trop mauvais", et on se permet d'aller plus loin dans l'interprétation.

  On résoudra plus tard ce problème d'inadéquation.

- $H_0$ : "$\log \frac{p}{1-p} = \text{Cte}$"
  $H_1$ : "$\log \frac{p}{1-p} = \alpha_{race} + \beta_{race}.ldose$"
  Le modèle apporte une **explication significative**, car la différence de déviance entre le modèle nul et ce modèle vaut $836.7 - 131$, elle est à comparer à un $\chi^2$ à $107 - 104 = 3$ degrés de liberté.

- On a estimé quatre paramètres, une ordonnée à l'origine par race, une pente par race. On lit directement les pentes et ordonnées à l'origine, parce que le modèle a été décrit sans intercept (-1). On obtient **l'estimation de chacun des paramètres**, une estimation de son **écart-type**, et une "**t-value**" qui permet de tester la nullité du paramètre. Ces "t-value" sont, dans notre cas, des "z-value" (on verra pourquoi chapitre 8), et sont à comparer au $(1-\alpha)$ quantile d'une loi normale $\mathcal{N}(0,1)$ c'est à dire à 1.96 ($\alpha = 0.05$) (test de Wald).

  Ici, les deux pentes sont négatives, et significativement non nulles.

- La méthode d'estimation des paramètres est une **méthode itérative**. La procédure a convergé en 6 itérations.

- Bug Splus (version 3.4) : la sortie Splus réellement obtenue n'est pas exactement celle-ci : on obtient 866.7 au lieu de 836.7 comme déviance du modèle nul, et 108 au lieu de 107 comme degré de liberté. C'est parce que Splus "ne sait plus" ce qu'est le modèle nul, lorsqu'on met un terme "-1" dans le modèle (ici, le modèle a été écrit "-1 + race + race:ldose" de façon à obtenir directement les ordonnées à l'origine et les pentes de chaque race, dans la sortie numérique). Pour obtenir une déviance et un degré de liberté corrects pour le modèle nul, il suffit de refaire tourner le même modèle, mais avec une autre paramétrisation "race+race:ldose".

## Ex. Tournesol : analyse

GLM : $\log(\frac{p}{1-p}) = \alpha_{race} + \beta_{race}.ldose$

```
Analysis of Deviance Table

Terms added sequentially (first to last)
           Df   Dev Res.Df Res.Dev Pr(Chi)
      NULL               107   836.7
      race  1 199.1      106   637.6  0.0000
     ldose  1 504.6      105   133.0  0.0000
race:ldose  1   2.0      104   131.0  0.1585
```



**Version S**

On affiche la **table d'analyse de déviance** du modèle. Les termes race, ldose et race:ldose sont introduits successivement dans le modèle. Cela revient à considérer dans l'ordre, la **succession de modèles** : "modèle nul","race","race+ldose", "race+ldose+race:ldose".

On trouve dans la colonne "Res.Dev" les **déviances (résiduelles) des modèles successifs**, dans la colonne Res.Df, **les degrés de liberté associés** aux modèles. Dans la colonne "Dev", sont affichées les **déviances associées à chaque terme**, et dans la colonne Df, les **degrés de liberté associés** à chaque terme. Par exemple, la déviance 199.1 associée au terme race s'obtient par différence entre les déviances 836.7 et 637.6, du modèle nul et du modèle "race". Enfin, on trouve dans la colonne "Pr(Chi)" la **pvalue** correspondant au test de l'hypothèse nulle d'absence d'effet du terme considéré : c'est la probabilité, sous hypothèse nulle, de dépasser la déviance observée.

La déviance expliquée par le terme race vaut 199, elle est à comparer à un $\chi^2(1)$, le terme race est donc hautement significatif ($p < 0.0001$), les ordonnées à l'origine sont donc différentes. De même, le terme ldose est hautement significatif, ce qui signifie que la pente commune est non nulle.

Par contre, le terme d'interaction race:ldose n'est pas significatif (p=0.16), ce qui signifie que l'hypothèse d'égalité des pentes est acceptée.

# Ex. Tournesol : analyse

$$\text{GLM : } \log(\frac{p}{1-p}) = \alpha_{race} + \beta_{race}.ldose$$

```
   Criteria For Assessing Goodness Of Fit
Criterion              DF   Value  Value/DF

Deviance              104   131.0    1.3
Scaled Deviance       104   131.0    1.3
Pearson Chi-Square    104  2166.7   20.8
Log Likelihood          .  -156.2      .

       Analysis Of Parameter Estimates
Parameter   DF Estimate Std.Err Chisq  Pr>Chi

INTERCEPT    0   0.0000  0.00       .       .
RACE       S 1  -1.0651  0.21   25.03  0.0001
RACE       R 1   6.4824  0.90   51.83  0.0001
LDOSE*RACE S 1  -1.8688  0.21   83.04  0.0001
LDOSE*RACE R 1  -2.4004  0.33   52.89  0.0001
SCALE        0   1.0000  0.00       .       .

NOTE :  The scale parameter was held fixed.
```



**Version SAS**

- La **déviance résiduelle** vaut 131. On peut interpréter cette déviance comme une déviance du modèle par rapport aux données. On l'utilise pour valider le modèle (**test d'adéquation du modèle**), parce que les nombres $\boldsymbol{n_i}$ **ne sont pas trop petits** ($n_i = 7$). On la compare à un $\chi^2$ à 104 degrés de liberté ($m - k$ : le nombre d'observations binomiales moins le nombre de paramètres = 108-4). Le $(1-\alpha)$ quantile d'un $\chi^2(104)$, avec $\alpha = 0.05$, vaut 128.8. Si l'on choisit un risque $\alpha = 0.05$, on rejette l'hypothèse nulle d'adéquation du modèle (p=4%). Le modèle n'est pas tout à fait adéquat, mais il n'est "pas trop mauvais", et on se permet d'aller plus loin dans l'interprétation.

  On résoudra plus tard ce problème d'inadéquation.

  En testant ainsi l'hypothèse d'adéquation du modèle, on a testé l'hypothèse nulle $H_0$ contre l'alternative $H_1$ :
  $H_0$ : "$\log \frac{p}{1-p} = \alpha_{race} + \beta_{race}.ldose$" : "modèle adéquat".
  $H_1$ : "$\log \frac{p}{1-p} = \widetilde{p}$ : proportion observée par combinaison (race, dose)"

- On a estimé quatre paramètres, une ordonnée à l'origine par race, une pente par race. On obtient **l'estimation de chacun des paramètres**, une estimation de son **écart-type**, et une "**statistique du** $\chi^2$" qui permet de tester la nullité du paramètre. Ces "statistiques du $\chi^2$" sont les carrés des "estimations des paramètres divisées par leur écart-type" (ce sont les carrés des "z-value" affichées par Splus). La pvalue associée au test de nullité de chaque paramètre est affichée.

  Ici, les deux pentes sont négatives, et significativement non nulles (test de Wald).

  l'option "noint" (pas d'intercept) a été choisie pour décrire le modèle. C'est grâce à cela que les pentes et ordonnées à l'origine sont obtenues directement.

## Ex. Tournesol : analyse

$$\text{GLM} : \log(\frac{p}{1-p}) = \alpha_{race} + \beta_{race}.ldose$$

```
LR Statistics For Type 1 Analysis

Source      Deviance   DF ChiSquare  Pr>Chi

INTERCEPT  836.7388   0          .       .
RACE       637.6401   1   199.0987  0.0001
LDOSE      133.0299   1   504.6102  0.0001
LDOSE*RACE 131.0416   1     1.9883  0.1585
```



**Version SAS**

On affiche la **table d'analyse de déviance** du modèle.

Le modèle apporte une **explication significative**, car la différence des déviances entre le modèle nul et ce modèle vaut $836.7 - 131$, elle est à comparer à un $\chi^2$ à $107 - 104 = 3$ degrés de liberté. On teste ainsi l'hypothèse nulle $H_0$ contre l'alternative $H_1$ :
$H_0$ : "$\log \frac{p}{1-p} = \text{Cte}$"
$H_1$ : "$\log \frac{p}{1-p} = \alpha_{race} + \beta_{race}.ldose$"

Les termes race, ldose et race*ldose sont introduits successivement dans le modèle. Cela revient à considérer, dans l'ordre, la **succession de modèles** : "modèle nul","race","race+ldose", "race+ldose+race*ldose".

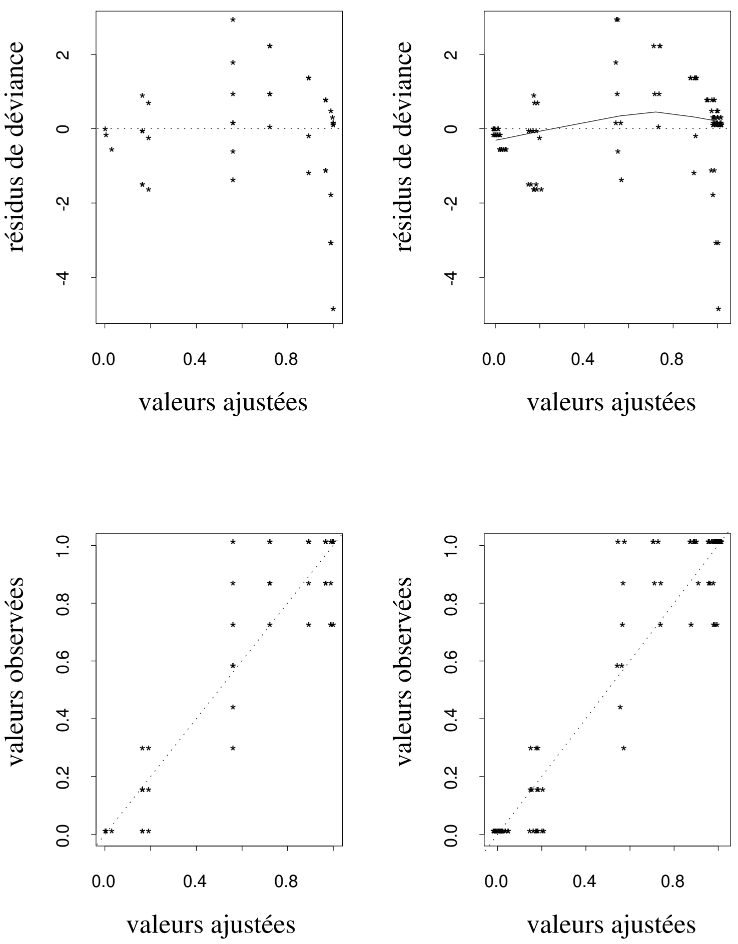
On trouve dans la colonne "Deviance" les **déviances des modèles successifs**, dans la colonne "Chisquare", sont affichées les **déviances associées à chaque terme**, et dans la colonne Df, les **degrés de liberté associés** à chaque terme. Par exemple, la déviance 199.1 associée au terme race s'obtient par différence entre les déviance 836.7 et 637.6, du modèle nul et du modèle "race". Enfin, on trouve dans la colonne "Pr>Chi" la **pvalue** correspondant au test de l'hypothèse nulle d'absence d'effet du terme considéré.

La déviance expliquée par le terme race vaut 199, elle est à comparer à un $\chi^2(1)$, le terme race est donc hautement significatif ($p < 0.0001$), les ordonnées à l'origine sont donc différentes. De même, le terme ldose est hautement significatif, ce qui signifie que la pente commune est non nulle.

Par contre, le terme d'interaction race*ldose n'est pas significatif (p=0.16), ce qui signifie que l'hypothèse d'égalité des pentes est acceptée.

Rq : le titre "LR statistics" signifie "likelihood ratio statistics". En effet, les statistiques de test sont des différences de déviance, qui sont aussi des différences entre des -2log(vraisemblance). Les tests réalisés sont donc bien des tests du rapport de vraisemblance.

## Ex. Tournesol : validation






La déviance nous a permis de tester la validité du modèle. Il reste qu'il est toujours indispensable de confronter les données au modèle à l'aide de représentations graphiques.

Les résidus, (ici les **résidus de déviance**), sont représentés, **en fonction des valeurs ajustées**. Ce graphe présente de nombreux points superposés, parce qu'il n'y a que 8 réponses possibles, pour un ensemble de 7 plantules : 0/7, 1/7, ..., 7/7. Pour visualiser les points superposés, le même graphique est à nouveau représenté, après une légère **perturbation aléatoire** des valeurs ajustées. On note 1 (ou 3) résidus négatifs, particulièrement grands, mais pas de structuration systématique des résidus, sauf peut-être une légère tendance à présenter des résidus négatifs, puis positifs.

On représente les **proportions observées en fonction des proportions ajustées**, et la première bissectrice. Ce nouveau graphique n'apporte pas d'information nouvelle par rapport au précédent, mais il peut être plus parlant. Comme précédemment, on perturbe aléatoirement les valeurs ajustées, pour visualiser les points superposés.

Ce coup d'œil rapide à quelques représentations graphiques ne révèle **pas d'inadéquation flagrante**. L'ajustement sera étudié plus en détail dans un exercice.

## Ex. Tournesol : comparaison de sous-modèles

M1 : modèle avec deux droites quelconques

M2 : modèle avec deux droites parallèles

M3 : modèle avec une droite

```
Analysis of Deviance Table

   Res.Df  Res.Dev   Df    Dev   Pr(Chi)
M1    104    131.0
M2    105    133.0  -1    -2.1    0.1585
M3    106    502.2  -1  -369.2    0.0000
```

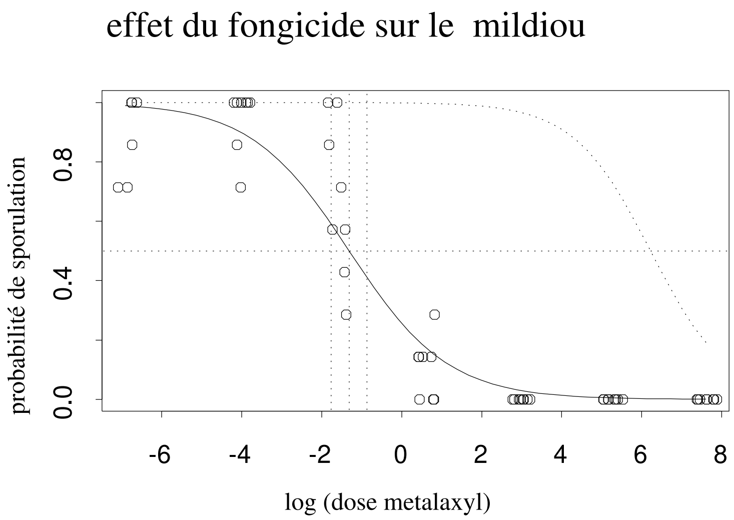


On a choisi un modèle "loi binomiale, lien logit, $\log \frac{p}{1-p} = \alpha_{race} + \beta_{race}.ldose$". L'analyse des résultats ne conduit pas à rejeter ce modèle. On peut rechercher un modèle plus simple, sous-modèle du précédent, dont la qualité d'explication restera acceptable, l'objectif étant d'obtenir un modèle parcimonieux, c'est à dire aussi explicatif que possible, avec le plus petit nombre de paramètres possible.

Voyons un exemple de test de sous-modèles, puis de calcul de la dose efficace 50 $DE_{50}$ et de son intervalle de confiance.

**Test pour comparer deux sous-modèles :** On considère le modèle $M_1$ comme adéquat ($p = 4\%$). On teste successivement l'égalité des pentes, et l'égalité des droites, par des tests de sous-modèles : le plus grand modèle, en nombre de paramètres est celui, déjà estimé, où la pente et l'ordonnée à l'origine diffèrent selon la race (4 paramètres). On considère ensuite le sous-modèle à 3 paramètres, où la pente est commune aux deux races, puis le sous-modèle à 2 paramètres, où pente et ordonnée à l'origine sont communes aux deux races. Le grand modèle correspond à deux droites quelconques, le deuxième modèle correspond à deux droites parallèles, et le dernier à une seule droite.
Plus on contraint le modèle, plus la déviance résiduelle augmente (la déviance résiduelle est une mesure de la variabilité résiduelle, non expliquée par le modèle).
Lorsqu'on compare les modèles M1 et M2, la déviance résiduelle augmente de 2.1. Cette augmentation est à comparer à un $\chi^2$ à 1 degré de liberté, car la différence entre les nombres de paramètres des deux modèles vaut 1. Cette augmentation de la déviance n'est pas significative ($p = 0.16$). On conserve l'hypothèse d'égalité des pentes.
Le passage du modèle M2 au modèle M3 produit une augmentation de déviance égale à 369, pour un paramètre consommé. 369 est à comparer à un $\chi^2(1)$. Cette différence entre les déviances est hautement significative. On rejette l'hypothèse d'égalité des ordonnées à l'origine (p<0.001).

# Ex. Tournesol : estimation de $DE_{50}$






***$DE_{50}$*** : On s'intéresse à la race sensible. On cherche à estimer la **dose de fongicide nécessaire**, pour obtenir une **probabilité de sporulation égale à 0.5** . On note $DE_{50}$ la dose recherchée (DE comme dose efficace, ou DL comme dose léthale, dans d'autres contextes).
Ici, la $DE_{50}$ est estimée à -1.3 (dans l'échelle log), avec un intervalle de confiance à 95% qui vaut $[-1.76, -0.87]$ dans l'échelle log, donc $[0.17, 0.42]$ dans l'échelle initiale.

Une fois le modèle estimé et validé, des tests utiles peuvent être réalisés, et des estimations intéressantes calculées. Par exemple, on peut estimer la dose efficace à différents niveaux. Ainsi, dans une étude sur les substances perçues par les consommateurs, on pourra estimer la $DE_5$ si on souhaite une substance peu détectable ou bien la $DE_{90}$ dans un objectif contraire.

## RÉSUMÉ

GLM $\longrightarrow \widehat{p} \in ]0,1[$

$\longrightarrow$ hétérogénéité des variances modélisée

$\longrightarrow$ régresseur : $log(dose)$

$\longrightarrow$ effet linéaire de log(dose) sur logit(p)

1 modèle $\longrightarrow$ table d'ANODEV

modèles emboîtés $\longrightarrow$ table d'ANODEV

$\widehat{param}$, $\widehat{e.t.}$, t-value $\longrightarrow$ test



Les deux avantages du GLM sont
1) de pouvoir contraindre les probabilités estimées à appartenir à l'intervalle ]0,1[ au moyen de la fonction de lien,
2) de permettre la modélisation de l'hétérogénéité des variances par l'intermédiaire de la loi choisie (loi binomiale : $Var(Y) = np(1-p)$).

Ici, on a choisi la transformation logarithmique pour le régresseur dose. Le modèle posé suppose la linéarité de l'effet de log(dose) sur logit(p).

Deux types de table d'analyse de déviance peuvent être obtenus.

- Une table par modèle. Cette table présente les déviances associées à chacun des termes du modèle. Ces déviances dépendent de l'ordre d'introduction des termes dans le modèle. Elles sont ajustées par rapport aux termes précédemment introduits dans le modèle. Elles permettent donc de tester la significativité de l'explication apportée par un terme, une fois pris en compte les termes précédents.
- D'autres tables présentent une succession de modèles emboîtés. La comparaison de 2 modèles emboîtés se fait par un test de sous-modèle, il s'agit d'un test de rapport de vraisemblance. La statistique de test est égale à la différence des déviances (résiduelles) des modèles.

Le principe de ces deux types de tables d'analyse de déviance est similaire. Dans les deux cas, ce sont des tests de sous-modèles (test du rapport de vraisemblance) qui sont présentés.

Enfin, les paramètres estimés sont présentés, leur écart-type estimé, la t-value correspondante (paramètre estimé divisé par l'écart-type). Ces statistiques (t-values) permettent de tester la nullité du paramètre correspondant.

## Chapitre 3 : plan

1. GLM : cas de la loi binomiale
   - Loi de Bernoulli et loi binomiale
   - Loi binomiale : densité et densité conjointe
   - Loi binomiale : vraisemblance
   - Modèle
   - Fonction de lien
   - Estimation des paramètres
   - Déviance, loi asymptotique
   - Test du rapport de vraisemblance
   - Données groupées - non groupées
   - Propriétés asymptotiques et tests
2. GLM : cas général
   - Famille exponentielle
   - Application : loi binomiale
   - Application : loi normale
   - Résultats généraux
   - Relation entre LM et GLM



Chaque fois qu'une notion, ou un résultat est introduit, faire le lien avec l'exemple tournesol, les sorties numériques et graphiques.

## Loi de Bernoulli et loi binomiale

- **loi de Bernoulli** $B(1,p)$

  $X$ prenant la valeur $r=0$ ou $r=1$ selon le schéma :

$$\boxed{P\,(\;X=r\;)=p^r(1-p)^{1-r}}$$

$$E(X)=p\;,\;Var(X)=p(1-p)$$

- **loi binomiale** $B(n,p)$

  $X_i, i = 1,..,n$, binaires, **indépendantes** de loi $B(1,p)$, alors

  $Y=\sum_{i=1}^{n} X_i$ prend la valeur $y \in [0,n]$ selon :

$$\boxed{P\,(\;Y=y\;)=\mathcal{C}_n^y\;p^y(1-p)^{n-y}}$$

$$E(Y)=np\;,\;Var(Y)=np(1-p)$$



- exemples de variables binaires : jeu pile ou face, germination d'une plante, mortalité suite à un traitement,...
- Quand le nombre d'occurences $y$ d'un phénomène observé sur $n$ individus est la somme de variables binaires indépendantes et de même probabilité $p$, ce nombre suit une loi **binomiale**.
- $n$ est **fixé a priori (non aléatoire).** $\mathcal{C}_n^y$ se lit "nombre de combinaisons de $y$ parmi $n$".
- la distribution d'une binomiale est symétrique si $p=.5$. Graphiques des distributions pour $n=10$ et $p=.1,\ .25,\ .5,\ .9$ (cf. livre de Collett p. 21 et p. 22).
- la loi binomiale est introduite quand l'unité statistique est un groupe préalablement défini dans l'expérimentation ou bien par agrégation d'observations binaires supposées se réaliser dans des conditions identiques. On parlera dans ce cas de **données groupées**.
- il est plus facile, dans un cadre binomial, de comparer les estimations issues d'un modèle aux observations (analyse des résidus) que dans le cas d'une variable binaire.
- $E$ désigne l'espérance et $Var$ la variance. Remarquer que la variance d'une binomiale n'est pas constante, elle est maximale pour p=0.5, nulle pour p=0 et p=1. Dans le modèle linéaire classique la variance de la variable expliquée est supposée constante.
- dans le cas de proportions, on a $E(\frac{Y}{n})=p$ et $Var(\frac{Y}{n})=\frac{p(1-p)}{n}$
- si $np(1-p)\geq 2$ alors $\dfrac{Y-np}{\sqrt{np(1-p)}}$ suit approximativement une loi normale centrée réduite. Si $p=.5$, l'approximation est bonne pour $n\geq 10$. D'où les méthodes approchées pour les tests et calculs d'intervalles de confiance sur les proportions.

## Loi binomiale : densité et densité conjointe

- **densité** de $Y$ binomiale $B(n,p)$

  fonction $y \to P(Y=y)$

- **densité conjointe**

  variables indépendantes $Y_i$

  $Y_i \sim B(n_i, p_i) \qquad i = 1, ..., 3$

  c'est la densité du triplet, soit

  $(y_1, y_2, y_3) \to P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

  mais encore...



- la densité pour une loi discrète est la fonction qui donne la probabilité de chaque valeur de la variable. La somme des probabilités est égale à 1.
- le triplet correspond à 3 réalisations de binomiales indépendantes; la densité est liée à la définition des unités statistiques et de l'échantillon (ici de taille 3).
- si l'on considère des unités statistiques où la réponse est binaire, la densité conjointe est le produit de densités de lois de bernoulli.
- faire compléter le transparent par les stagiaires;

  produit des probabilités car indépendance, puis

  $$P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \prod_{i=1}^{3} \mathcal{C}_{n_i}^{y_i}\, p_i^{y_i} (1-p_i)^{n_i - y_i}$$

  la somme de ces probabilités est aussi égale à 1.
- généralisation immédiate pour plus de 3 réalisations de binomiales.

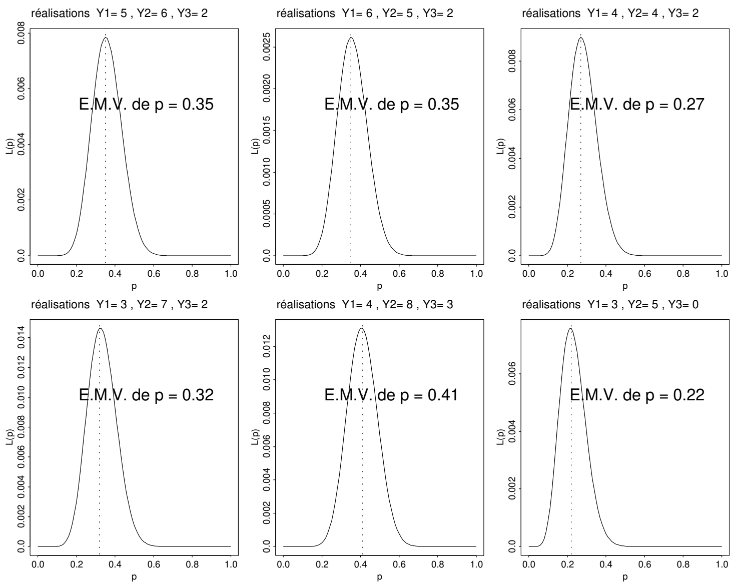
## Loi binomiale - vraisemblance

- $y_i$ réalisation de $Y_i \sim B(n_i, p_i)$, $i = 1, 3$
- $Y_i$ indépendantes et $p_i$ **inconnues**
- **vraisemblance** des paramètres $p_i$ :
  $p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \rightarrow L(p; y_1, y_2, y_3)$
  $= P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$
- plus généralement, si $p = p(\theta)$
  $L(\theta; y_1, y_2, y_3)$ vraisemblance de $\theta$
  (ou $L(\theta)$ s'il n'y a pas d'ambiguïté)
  ($L$ pour likelihood)
- **Maximum de vraisemblance** :
  Rechercher $\theta$ qui maximise $L$,
  soit à résoudre : $\frac{\partial}{\partial\theta} L = 0$.



- pour une valeur du triplet des paramètres $p_i$, la vraisemblance donne la probabilité de la réalisation du triplet $(y_1, y_2, y_3)$, compte tenu du modèle. Chaque valeur de la vraisemblance est calculée à partir des réalisations.
- dans l'exemple tournesol, on a 14 $p_i$ (car 7 doses et 2 races), et $\theta$ est un vecteur de 4 paramètres (2 pentes et 2 ordonnées à l'origine). Les $p_i$ dépendent de $\theta$, on les notera parfois $p_i(\theta)$ pour insister sur le fait que ces probabilités sont fonction des paramètres du modèle considéré. $\theta$ est le vecteur des paramètres du modèle.
- l'idée du principe du maximum de vraisemblance consiste à rechercher $\theta$ qui rende la réalisation la plus probable possible.
- dans le cas de 3 paramètres à estimer, on doit résoudre 3 équations à 3 inconnues : $\frac{\partial}{\partial\theta_i} L = 0$. Dans la pratique, résolution numérique d'un système d'équations non-linéaires.
  $\theta = (\alpha_R, \alpha_S, \beta_R, \beta_S)$ dans l'exemple tournesol.
- on maximise en fait $Log(L)$.

# Loi binomiale - vraisemblance (simulation)






simulation d'un exemple pour le max de vraisemblance.

On tire 1 réalisation pour chacune des binomiales $B(10, .3)$, $B(20, .3)$ et $B(7, .3)$. Soient $y_1$, $y_2$ et $y_3$ les valeurs obtenues. On calcule la vraisemblance $L$ et l'EMV[1] de la probabilité $p$ supposée commune aux 3 réalisations $y_i$. La répétition de ce processus 6 fois donne une idée de la variabilité de l'EMV.

```
function(p = 0.3)
{
# simulation
# (y1,y2,y3) realisations de 3 binomiales
# B(n1,p),B(n2,p), B(n3,p)
        n1 <- 10
        n2 <- 20
        n3 <- 7
        y1 <- rbinom(1, n1, p)
        y2 <- rbinom(1, n2, p)
        y3 <- rbinom(1, n3, p)
        p0 <- seq(0, 1, 0.01)
        nn <- length(p0)
        L <- (1:nn) * 0
        for(i in 1:nn) {
 L[i] <- dbinom(y1, n1, p0[i]) * dbinom(y2, n2, p0[i])
                               * dbinom(y3, n3, p0[i])
                       }
        plot(p0, L, type = "l", xlab = "p", ylab = "L(p)")
        title(paste("realisations  Y1=", y1, ", Y2=", y2, ",
                    Y3=", y3))
        mp <- p0[L == max(L)]
        abline(v = mp, lty = 2)
        text(0.6, 0.7 * max(L), paste("E.M.V. de p =", mp),
             cex = 1)
        invisible()
}
```

1. Estimateur du Maximum de Vraisemblance

## Modèle

$$log\left(\frac{p}{1-p}\right) = \alpha_{race} + \beta_{race} ldose$$

$$g(p) = X\theta = \eta$$



On écrit d'abord le modèle dans le cas particulier de l'exemple tournesol, puis dans le cas général.

$\boldsymbol{p}$ : vecteur des **probabilités théoriques inconnues**. $\boldsymbol{p_i} = \boldsymbol{E}(\frac{Y_i}{n_i})$ .

Les probabilités sont liées aux variables explicatives par l'intermédiaire d'une **fonction de lien**, qui peut être la fonction logit ($log(\frac{p}{1-p})$) par exemple, et qui sera plus généralement notée $g$ $\boxed{g(p) = X\theta = \eta}$

$g$ : fct de lien.
$X$ : matrice d'incidence, qui définit le plan d'expérience.
$\theta$ : vecteur des paramètres du modèle.
$\eta$ : **prédicteur linéaire**.

La fonction de lien décrit la relation entre le prédicteur linéaire $\eta$ et l'espérance $E(\frac{Y}{n}) = p$.

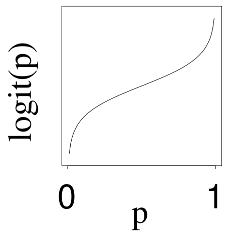
Exercice : ex. tournesol : $g(p) = \log(\frac{p}{1-p})$ : lien logit
Ecrire $\theta$ et $X$, dans le cas du modèle vu au chapitre précédent, avec deux droites quelconques.

$$X = \begin{pmatrix} 1 & 0 & ldose & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & ldose & 0 \\ 0 & 1 & 0 & ldose \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & ldose \end{pmatrix} \theta = \begin{pmatrix} \alpha_R \\ \alpha_S \\ \beta_R \\ \beta_S \end{pmatrix}$$

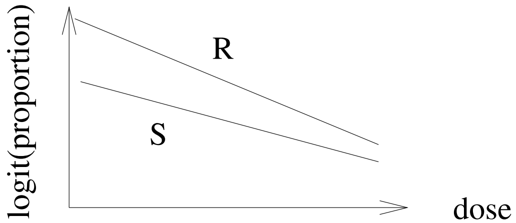
## Modèle : fonction de lien

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

$$\text{logit} : [0,1] \longrightarrow ]-\infty, +\infty[$$




ex. tournesol



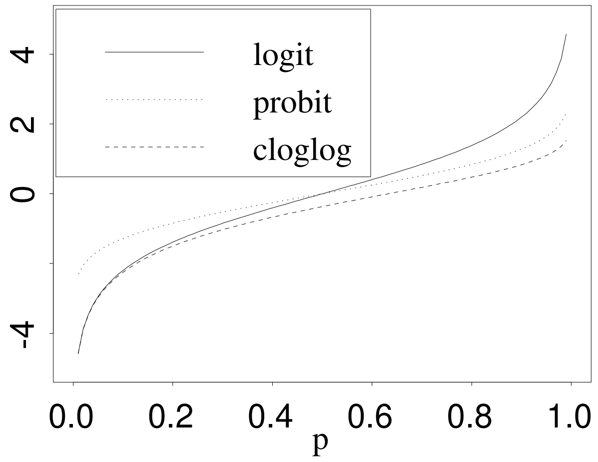

$$log\left(\frac{p}{1-p}\right) = \alpha_{race} + \beta_{race} ldose$$



Le lien naturel (dit canonique) est le lien logit, en référence à certaines propriétés des lois de la famille exponentielle à laquelle appartient la loi binomiale. C'est le lien par défaut dans la plupart des logiciels.

- La fonction **logit envoie $[0,1]$ sur $]-\infty,+\infty[$** . Lorsqu'on travaille dans l'échelle logit, on travaille donc dans l'intervalle $]-\infty,+\infty[$. Les $logit(p) \in ]-\infty,+\infty[$, donc les $p$ obtenus par la transformation réciproque du lien appartiennent à $]0,1[$. Il n'y a aucun risque d'obtenir des valeurs ajustées ou des bornes d'intervalle de confiance à l'extérieur de l'intervalle ]0,1[.
- Si le modèle est $log\left(\frac{p}{1-p}\right) = \alpha_{race} + \beta_{race} ldose$, cela signifie que c'est **dans l'échelle logit que l'effet ldose est supposé avoir une forme linéaire**. Dans l'échelle d'origine, les observations sont bornées par 0 et 1 et l'effet est sigmoïdal.

## Modèle : fonction de lien




Loi binomiale

fonctions de lien $\begin{cases} \text{logit} \\ \text{probit} \\ \text{cloglog} \end{cases}$



Pour la loi binomiale, la fonction de lien la plus fréquemment utilisée est le **lien logit**. Les **liens probit** et **cloglog** sont couramment utilisés aussi. On peut inventer une infinité d'autres liens. Toute fonction continue monotone, de $[0,1]$ dans $]-\infty,+\infty[$ convient, donc en particulier, toute fonction de répartition.

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

$$probit(p) = \phi^{-1}(p)$$

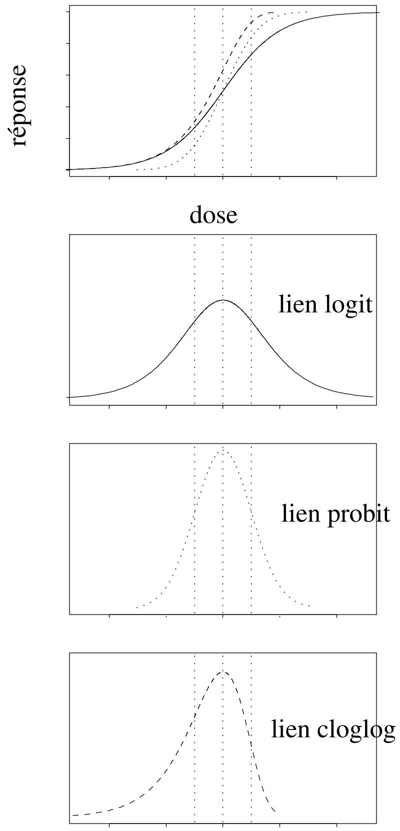
où $\phi$ est la fonction de répartition de la loi $\mathcal{N}(0,1)$.

$$cloglog(p) = log(-log(1-p))$$

Lorsque p tend vers 0, sa transformée tend vers $-\infty$, lorsque p tend vers 1, sa transformée tend vers $+\infty$. Les 3 fonctions de lien sont légèrement différentes les unes des autres. Les liens logit et cloglog sont très proches, si les probabilités sont proches de 0. Si les probabilités sont proches de 1, les liens cloglog et probit sont plus proches l'un de l'autre, que ne le sont les liens logit et cloglog.

# Fonction de lien et distribution de tolérance

## Courbe dose-réponse






Dans le cas des courbes dose-réponse le **choix du lien** est **lié à la forme de la distribution des seuils individuels**. (appelée distribution des seuils, ou distribution de tolérance, selon le contexte).

La courbe dose-réponse estimée pour une population est le résultat des réponses individuelles. Chaque individu a son propre seuil de tolérance au produit.

Le **lien probit** est associé à une **distribution normale**.
Le **lien logit** est associé à une **distribution logistique**.
Le **lien cloglog** est associé à une **distribution de Gumbel** des seuils.
Les distributions normale et logistique sont symétriques, la distribution normale est plus pointue que la distribution logistique.
La distribution de Gumbel est dissymétrique.

Le premier graphique représente la proportion cumulée d'individus qui répondent pour une dose d inférieure ou égale à la dose reportée en abscisse. Les 3 courbes suivantes s'obtiennent en dérivant la première, et représentent les distributions des seuils implicitement choisies lorsqu'on utilise un modèle avec un lien logit, probit, ou cloglog.

Le lien cloglog s'introduit naturellement dans certains problèmes de dilution quand on estime le nombre $N$ d'organismes dans une solution, et où on observe, à chaque dilution, le nombre de boîtes de Pétri sur lesquelles l'organisme se développe.

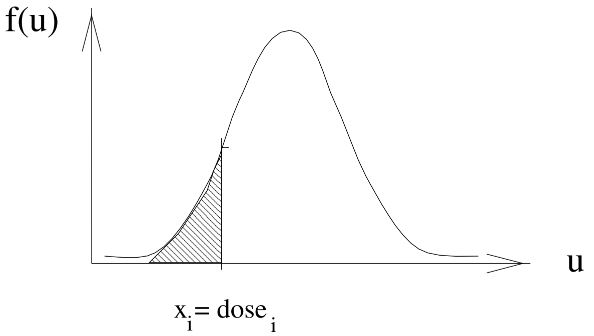
Ce lien est introduit aussi dans les cas de survie censurés (par intervalle), dans des tests sérologiques où le nombre d'individus séropositifs par tranche d'âge permet d'estimer un taux moyen d'infection, et d'analyser sa variation selon la tranche d'âge, un taux constant signifiant une maladie endémique (cf. Collett chap.4,6).

## Fonction de lien et distribution de tolérance

Lien probit : loi normale des seuils de tolérance

1 insecte $\longrightarrow$ 1 seuil de tolérance

$x_i$ : dose administrée
U : V.A. "seuil de tolérance"
u : seuil de tolérance d'un insecte. densité :




$p_i = P(Y_i = 1)$ pour 1 insecte si dose $x_i$.

$$p_i = P(U \leq x_i) = \int_{-\infty}^{x_i} f(u)du$$



On applique différentes doses d'insecticide à des lots d'insectes. La réponse observée pour chaque insecte est binaire. La réponse observée sur chaque lot est binomiale.

Chaque insecte a son propre seuil de tolérance au composé. Un insecte à faible seuil de tolérance a plus de chances de mourir qu'un insecte à seuil élevé, pour une dose donnée.

Soit $x_i$ la dose administrée. Les insectes dont le seuil est inférieur à $x_i$ meurent. Soit U la variable aléatoire "seuil de tolérance", u le seuil de tolérance d'un individu. On représente la densité de la variable aléatoire U. La zone hachurée représente la probabilité $p_i$ de réponse d'un insecte exposé à la dose $x_i$ (et tiré au sort dans la population étudiée).

## Fonction de lien et distribution de tolérance

Supposons les seuils distribués normalement

$$U \sim \mathcal{N}(\mu, \sigma^2)$$

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right)$$

$$p_i = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_i} exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du = \phi\left(\frac{x_i-\mu}{\sigma}\right)$$

$\phi$ : fct de répartition d'une loi $\mathcal{N}(0,1)$

On pose $\alpha = -\frac{\mu}{\sigma}$, $\beta = \frac{1}{\sigma}$

$$p_i = \phi(\alpha + \beta\, x_i)$$

$$p_i = F(\alpha + \beta x_i)$$

$$\boxed{F^{-1}(p_i) = \alpha + \beta\, x_i \qquad F = \phi}$$

régression probit



Si on suppose que la distribution des seuils individuels est une loi $\mathcal{N}(\mu, \sigma^2)$, on arrive à un modèle probit. La fonction de lien est la réciproque de la fonction de répartition d'une loi $\mathcal{N}(0,1)$.

La pente $\beta$ est l'inverse de l'écart-type de la distribution des seuils individuels. Plus les seuils individuels sont variables, plus la pente est faible. L'ordonnée à l'origine est proportionnelle à la moyenne $\mu$ des seuils individuels, et vaut $-\frac{\mu}{\sigma}$.

De manière générale, on note F la fonction de répartition choisie. Ici, $F = \phi$.

## Modèle et vraisemblance

$$\theta = (\alpha_R, \alpha_S, \beta_R, \beta_S)$$

$$logit(p) = log\left(\frac{p}{1-p}\right) = \eta = \alpha_{race} + \beta_{race} ldose$$

$$\Rightarrow \boxed{p = \frac{exp(\eta)}{1+exp(\eta)}}$$

---

$$L(\theta) = \prod_{i=1}^{n} \mathcal{C}_{n_i}^{y_i} \boxed{p_i}^{y_i} (1-p_i)^{n_i - y_i}$$

$$p_i = \frac{exp(\eta_i)}{1+exp(\boxed{\eta_i})} \quad \hookleftarrow$$

$$\hookrightarrow \eta_i = \alpha_{race(i)} + \beta_{race(i)} ldose_i$$



Ce transparent est destiné à bien comprendre où interviennent les paramètres dans l'écriture de la vraisemblance.

On travaille dans l'échelle logit, c'est à dire dans l'échelle du prédicteur linéaire. Dans cette échelle, on obtient des estimations qui appartiennent à $]-\infty, +\infty[$. Mais ce sont les probabilités qui nous intéressent. On revient aux probabilités, c'est à dire à l'échelle d'origine, en appliquant la **transformation réciproque de la fonction de lien**, la transformation $\frac{exp}{1+exp}$.

On remplace, dans l'expression de la vraisemblance, les probabilités $p_i$ par les $\frac{exp(\eta_i)}{1+exp(\eta_i)}$. C'est dans $\eta_i$ que les paramètres du modèle interviennent.

# Estimation des paramètres

Maximiser

$$L(\theta) = \prod_{i=1}^{n} \mathcal{C}_{n_i}^{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}$$

$$\log(L) = \sum_{i=1}^{n} \log\left(\mathcal{C}_{n_i}^{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}\right)$$

$$\frac{\partial \log L}{\partial \theta_j} = 0 \;\Rightarrow \text{k eq. non lin. en les } \theta_j, \quad j = 1, \ldots, k$$

$$\Rightarrow \text{méthode itérative}$$

$$\widehat{\theta} \text{ E.M.V.}$$



On a vu **comment les paramètres $\boldsymbol{\theta}$** du modèle **interviennent dans la vraisemblance**. On estime $\theta$ par **maximisation de la vraisemblance**.

(cf. le schéma d'analyse chapitre 2 p.3).

On maximise en fait le log(vrais) qui est plus facile à manipuler (maximiser la vrais ou son log revient au même, puisque la fonction log est monotone croissante).

Au maximum, les dérivées premières par rapport à chaque paramètre sont nulles. On obtient un système de k équations non linéaires en les paramètres (k=nombre de paramètres à estimer).

E.M.V.: estimateur (estimation) du maximum de vraisemblance

On ne peut pas donner d'expression explicite de $\theta$ (contrairement au modèle linéaire, où $\widehat{\theta} = (X'X)^{-1}X'Y$). On résoud le système d'équations par une **méthode numérique itérative**.

L'**algorithme** consiste à faire une régression linéaire pondérée d'une variable de travail, à chaque itération. Les poids sont remis à jour à chaque itération.

Algo. IRLS : iteratively reweighted least-square.
Dans le cas particulier de la loi binomiale et du lien logit :
variable de travail : $z_i = \eta_i + \frac{y_i - n_i p_i}{n_i p_i (1-p_i)}$
poids : $w_i = n_i p_i (1-p_i)$
A chaque itération, régression pondérée de $z_i$ sur les variables explicatives.

On sait maintenant estimer les paramètres d'un modèle. Le modèle est-il bon ? comment valider le modèle ? comment comparer plusieurs modèles ?

**Exercice 1**

# Déviance


| modèle nul | modèle candidat | modèle saturé |
|---|---|---|
| nb param = 1 | nb param = k | nb param = nb obs = m |

$M_s \quad (\widetilde{p})$

comparer $M$ à $M_s$?

$M \quad (\widehat{p})$

$$(-2\log L) - (-2\log L_s) = -2\log\left(\frac{L}{L_s}\right)$$

$$\boxed{D = -2\log\left(\frac{L}{L_s}\right) \sim_{as} \chi^2(m-k)}$$

D = mesure de la qualité de l'ajustement
si données groupées



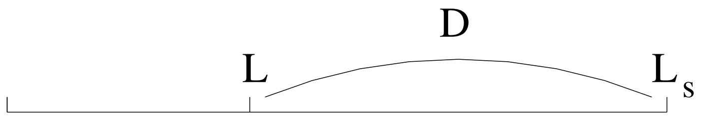
• **Le modèle est-il bon ?**
Le **modèle saturé** est le modèle de référence, il correspond à un **ajustement parfait** (tous les résidus sont nuls). Avec le modèle saturé, les probabilités estimées sont notées $\widetilde{p}$, elles sont égales aux proportions observées.
$\widehat{p}_i = \widetilde{p}_i = \frac{y_i}{n_i}$.

Le **modèle nul** est le modèle qui n'explique rien, celui dans lequel aucune variable explicative n'est introduite, la partie systématique est réduite au terme constant, les probabilités sont supposées toutes égales.
$p_i = \text{Cte}, \forall i$

Le **modèle candidat** est intermédiaire entre le modèle nul et le modèle saturé. Avec le modèle candidat, les probabilités estimées sont notées $\widehat{p}$.
$p_i = p_i(\theta) \qquad \widehat{p}_i = p_i(\widehat{\theta})$.

Soient $M$ le modèle candidat, et $M_s$ le modèle saturé. On veut comparer $M$ à $M_s$. $M$ est-il aussi bon que $M_s$ ? $M$ est-il valide ?

On fait un test du rapport de vraisemblance (sans le savoir ...). La **statistique de test** s'appelle **déviance**. Elle est égale à la **différence des $-2\log(vrais)$ entre le modèle candidat, et le modèle saturé**. La déviance mesure l'écart entre le modèle candidat et le modèle saturé. Sous l'hypothèse nulle d'équivalence entre les modèles $M$ et $M_s$, elle suit **asymptotiquement** une **loi de $\chi^2$ à $m-k$ degrés de liberté**.




Rq formateur : attention, $D$ n'est pas une mesure de la qualité de l'ajustement si les données sont non groupées. Dans le **cas des données non groupées**, les valeurs ajustées obtenues avec le modèle saturé sont exactement égales aux valeurs observées. Pour chaque observation, soit $\widehat{p}_i$ soit $1-\widehat{p}_i$ est nul. La vraisemblance du modèle saturé est donc nulle. Et la **déviance** ne dépend que des probabilités estimées $\widehat{p}_i$, et ne peut **pas** constituer une **mesure d'adéquation**.

Rq : $m$ est le nombre d'observations binomiales (ex : 108 dans l'exemple tournesol), et non pas le nombre d'observations binaires. $k$ est le nombre de paramètres du modèle candidat.

Rq : le -2 du $-2\log(\text{vrais})$ permet à la statistique calculée de suivre une loi du $\chi^2$.

## Déviance

$$L = \prod_{i=1}^{n} \mathcal{C}_{n_i}^{y_i}\, p_i^{y_i}(1-p_i)^{n_i-y_i}$$

$$\log(\widehat{L}) = \sum_{i=1}^{n} \left(\log \mathcal{C}_{n_i}^{y_i} + y_i \log \widehat{p_i} + (n_i - y_i)\log(1-\widehat{p_i})\right)$$

$$\log(\widetilde{L_s}) = \sum_{i=1}^{n} \left(\log \mathcal{C}_{n_i}^{y_i} + y_i \log \widetilde{p_i} + (n_i - y_i)\log(1-\widetilde{p_i})\right)$$

$$D = -2\log(\widehat{L}) + 2\log(\widetilde{L_s})$$

$$= 2\sum_{i=1}^{n}\left(y_i \log\frac{\widetilde{p_i}}{\widehat{p_i}} + (n_i-y_i)\log(\frac{1-\widetilde{p_i}}{1-\widehat{p_i}})\right)$$

$$\boxed{D = 2\sum_{i=1}^{n}\left(y_i \log\frac{y_i}{\widehat{y_i}} + (n_i-y_i)\log(\frac{n_i-y_i}{n_i-\widehat{y_i}})\right)}$$

$$\text{où} \quad \widehat{y_i} = n_i\widehat{p_i}$$



La déviance est la différence entre le -2log(vrais) du modèle candidat, et le -2log(vrais) du modèle saturé.

Voici une autre expression de la déviance (dans le cas binomial), qui montre clairement que la déviance est une mesure du défaut d'ajustement ("lack of fit").
$\widetilde{p_i} = \frac{y_i}{n_i}$ : proba estimée sous l'hypothèse du modèle saturé.

$\widehat{p_i}$ : proba estimée sous l'hypothèse du modèle candidat.

La **déviance compare les observations** $y_i$ **aux valeurs ajustées** $\widehat{y_i}$. Elle **mesure** donc bien l'**inadéquation** du modèle. C'est une mesure globale de l'inadéquation qui peut être due à un mauvais choix de la fonction de lien, de la loi (modélisation de la variance) ou de la partie systématique du modèle. Elle peut encore être due à une ou plusieurs observations extérieures.

On note qu'un développement limité à l'ordre 2 de $D(x) = y_i \log(y_i/x) + (n_i - y_i)\log(\frac{n_i - y_i}{n_i - x})$ au voisinage de $y_i$ approxime la déviance par $D \sim \sum_{i=1}^{n} n_i \frac{(\widehat{p_i} - \widetilde{p_i})^2}{\widetilde{p_i}(1-\widetilde{p_i})}$ (suffisante si $\widehat{p_i} - p_i \leq 0.1 p_i$). C'est une somme de carrés d'écarts réduits.

Rq : dans Splus, la déviance s'appelle déviance résiduelle.

**Exemple tournesol :** la déviance vaut 131, elle est à comparer à un $\chi^2(104 = 108 - 4, 0.95)$ qui vaut 128.8. On rejette l'hypothèse d'adéquation du modèle (p=4 %).

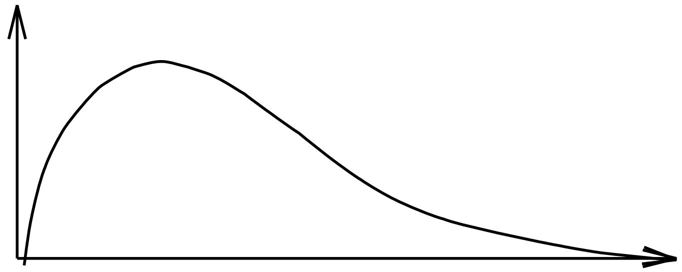
Rq : dans le cas où $y_i = 0$ ou $y_i = n_i$ pour tout $i$ (exemple des données binaires non groupées), $\log(\widetilde{L_s}) = 0$. En effet, dans le cas d'un modèle saturé, si $y_i = 0$ alors $\widetilde{p_i} = \widehat{p_i} = 0$, si $y_i = n_i$ alors $\widetilde{p_i} = \widehat{p_i} = 1$.

# Loi de la déviance

Données groupées :

Si $n_i$ pas trop petits,

$\boxed{D \sim_{as} \chi^2(m-k)}$ $\boxed{\text{Test d'adéquation}}$

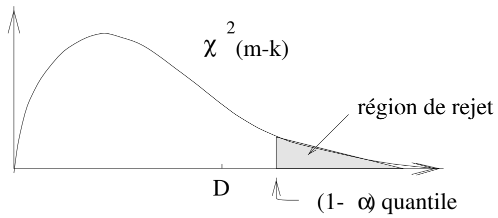




Dans le cas de données groupées, à la condition que les $n_i$ ne soient pas trop petits, on connaît la loi asymptotique de la déviance. C'est une loi du $\chi^2$. D est donc une statistique de test qui permet de tester l'adéquation du modèle. Ce test est valide si les $n_i$ ne sont pas trop petits, car c'est un test asymptotique. Les $n_i$ doivent être d'autant plus grands que les $p_i$ sont proches de 0 ou de 1.

**Test d'adéquation :** calculer la **statistique D**, et la **comparer au $(1-\alpha)$ quantile d'une loi $\chi^2$** dont le nombre de degrés de liberté est $m-k$ (le degré de liberté résiduel).

Attention : $m$ est le nombre d'observations binomiales (et pas le nombre d'observations binaires), et $k$ est le nombre de paramètres.





Loi asymptotique de la déviance

On accepte l'hypothèse d'adéquation du modèle, lorsque la déviance est assez petite. Le test est réalisé comme schématisé sur le graphique.
Une façon rapide (et approximative) de faire le test consiste à comparer la déviance au degré de liberté, parce que $E(\chi^2(m-k)) = m-k$. On accepte l'hypothèse d'adéquation lorsque la déviance est inférieure ou égale au degré de liberté (dl).

Attention : si la déviance est très inférieure au degré de liberté (pvalue très proche de 1), le modèle est probablement surparamétré (risque d'interprétation abusive).

**exercice 2**

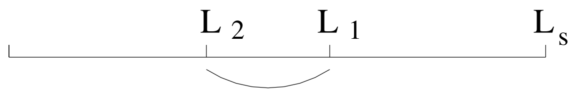
## Test du rapport de vraisemblance

Comparer 2 modèles emboîtés :
test de sous-modèle

$M_1 \rightarrow L_1$

$$(-2\log L_2) - (-2\log L_1)$$

$M_2 \rightarrow L_2$




sous $H_0$ : “$M_2$ équivalent à $M_1$”

$$\boxed{(-2\log L_2) - (-2\log L_1) \sim_{as} \chi^2(dl_2 - dl_1)}$$

$$-2\log\frac{L_2}{L_1} \sim_{as} \chi^2(dl_2 - dl_1)$$



On cherche un modèle satisfaisant le principe de parcimonie : on cherche à diminuer autant que possible le nombre de paramètres du modèle tout en conservant un modèle “convenable”. Pour cela, on est amené à comparer des modèles emboîtés. On peut aussi être amené à comparer deux modèles emboîtés pour tester une hypothèse.

Soient $M_1$ le grand modèle, et $M_2$ un sous-modèle, un cas particulier du modèle $M_1$.
On passe d'un modèle à un sous-modèle, par égalité entre plusieurs paramètres, ou par annulation d'un ou de plusieurs paramètres, ou en imposant une contrainte d'égalité d'un paramètre à une valeur fixée, ou plus généralement, en imposant une contrainte linéaire sur les paramètres.

$H_0$ : $M_2$ équivalent à $M_1$
$M_2$ aussi bon que $M_1$
$M_2$ valide (on a supposé $M_1$ valide)

La statistique du test : c'est la **différence des $-2\log$**(vrais) des deux modèles.
Cette statistique **suit**, sous $H_0$, **un $\chi^2$** dont le **nombre de degrés de liberté** est égal à la **différence des nombres de paramètres** des 2 modèles (résultat asymptotique).

Si la statistique de test est supérieure à la valeur critique, on rejette $H_0$.

Rq : $dl_1$ est le degré de liberté résiduel de $M_1$. La différence $dl_2 - dl_1$ est la différence des nombres de degrés de liberté résiduels des 2 modèles. C'est aussi la différence des nombres de paramètres des 2 modèles.
Rq : le $-2$ du $-2\log(vrais)$ est là pour que la statistique suive une loi de $\chi^2$.
Rappel : en modèle linéaire gaussien, pour comparer 2 modèles emboîtés, on utilise le test de Fisher :

$$\frac{(SCR_2 - SCR_1)/(dl_2 - dl_1)}{CMR_1} \sim F(dl_2 - dl_1, dl_1)$$

## Test du rapport de vraisemblance

Test de sous-modèle

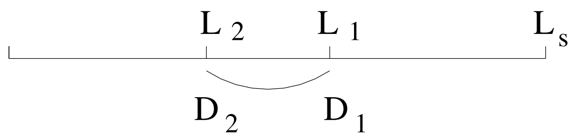
$M_1$
$M_2$ : sous-modèle

$$(-2\log L_2) - (-2\log L_1)$$

$$= (-2\log \tfrac{L_2}{L_s}) - (-2\log \tfrac{L_1}{L_s})$$

$$= D_2 - D_1$$

$$\boxed{D_2 - D_1 \sim_{as} \chi^2(dl_2 - dl_1)}$$




valide sauf si
$M_1 \simeq M_s$ et données non groupées



On a vu qu'on peut comparer 2 modèles emboîtés, en utilisant le test du rapport de vraisemblance :

$$(-2\log L_2) - (-2\log L_1) \sim_{as} \chi^2(dl_2 - dl_1)$$

Bien souvent, les logiciels ne fournissent pas la vraisemblance, ni le $-2\log(vrais)$, mais la déviance.

On peut faire ce test à l'aide des déviances, puisque la **différence des $-2\log(vrais)$ n'est rien d'autre que la différence des déviances**. Donc, $D_2 - D_1 \sim_{as} \chi^2(dl_2 - dl_1)$

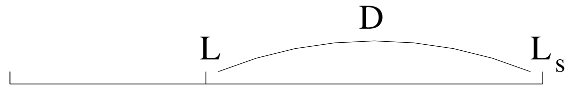
Ce résultat est un résultat asymptotique.
Il reste **valide, en cas de données non groupées** ($n_i = 1$), **si le plus grand modèle n'est pas le modèle saturé** ou trop proche du modèle saturé. En effet, si $M_1$ est très proche du modèle saturé, le test se ramène à un test d'adéquation, qui n'est pas valide, lorsque $n_i = 1$.

Résumé :
en cas de **données groupées**, le test d'adéquation et les tests de sous-modèles sont valides.
En cas de **données non groupées**, le test d'adéquation n'est pas valide, mais les tests de sous-modèles sont valides, dès lors que le plus grand modèle n'est pas saturé.

**Exemple tournesol :**
On a fait deux tests de sous-modèle (chap. II p. 7). On a comparé les modèles "race" et "race+ldose" d'abord, puis les modèles "race+ldose" et "race+ldose+ldose:race". L'ajout de ldose modifie significativement le modèle. L'ajout de race:ldose n'apporte pas de modification significative. On en conclu que les ordonnées à l'origine diffèrent, mais que les pentes sont égales.

# Test du rapport de vraisemblance

Test d'adéquation




$H_0$ : "le modèle est adéquat"

$$D \sim_{as} \chi^2(dl)$$

Test d'explication


$H_0$ : "le modèle n'apporte pas d'explication"

$$D_{nul} - D \sim_{as} \chi^2(dl_{nul} - dl)$$



Le test du rapport de vraisemblance permet de comparer deux modèles emboîtés. Les tests d'adéquation (déjà vus) et d'explication en sont des cas particuliers.

Un **test d'adéquation** revient à comparer le modèle candidat au modèle saturé.

$dl - dl_s = dl = m - k$, car le degré de liberté du modèle saturé est nul.

Rq : un test d'adéquation concluant à l'adéquation du modèle n'est pas suffisant pour conclure définitivement. Un examen graphique des résidus et des points influents est indispensable. Dans certains cas, le test peut conduire à accepter l'adéquation du modèle, alors qu'une observation est très influente pour l'un des termes, si influente qu'elle justifie à elle seule la significativité du terme en question. Ce test d'adéquation n'est basé que sur une valeur numérique, qui n'est qu'un résumé incomplet de la qualité du modèle.

Rq : le choix du modèle est un compromis entre adéquation et nombre de paramètres.

Un **test d'explication** revient à comparer le modèle candidat au modèle nul.

Obtenir un modèle adéquat est important, encore plus important si l'objectif est une prédiction, que si l'objectif se limite à la compréhension du phénomène et la connaissance des facteurs influents. Il peut arriver qu'on obtienne un modèle inadéquat, que l'on ne reussisse pas à améliorer, mais qui apporte une explication significative. Ce modèle, même s'il est inadéquat, permet de mettre en évidence l'influence de certains facteurs de variation. En ce sens, il est utile, et sera interprété.

Exemple tournesol :
Test d'explication (cf. chap. II p. 6, p.7) : $836.7 - 131 \sim \chi^2(107 - 104)$.

705.7 est à comparer à un $\chi^2(3)$. Le modèle apporte une explication significative.

**exercice 3**

## Données groupées
## Données non groupées

Données groupées $n_i > 1$

Données non groupées $n_i = 1$



La distinction entre données groupées et données non groupées est importante, car l'utilisation de la déviance, pour réaliser un test d'adéquation, n'est pas valide en cas de données non groupées.

Les données groupées sont caractérisées par $n_i > 1$, alors que les données non groupées sont caractérisées par des $n_i$ tous égaux à 1.

**Le caractère groupé ou non groupé des données dépend donc à la fois des données et du modèle utilisé !**

Exemple tournesol :
Dans cet exemple, quel que soit le modèle, les données sont groupées, car $n_i = 7$.

## Données groupées
## Données non groupées

Données non groupées :

$$n_i = 1$$

$$D \sim \quad ?$$

$$\boxed{D \neq \text{mesure de la qualité d'ajustement}}$$

$\Rightarrow$ autres méthodes de validation

| | Test adéquation | Test explication | Test sous-modèle |
|---|---|---|---|
| Données groupées | oui | oui | oui |
| Données non groupées | **NON** | oui | **OUI, SI ...** |



Dans le cas de **données non groupées**, on ne connaît pas la loi de D. **D n'est pas une mesure de la qualité d'ajustement**. On utilise donc d'autres méthodes pour valider un modèle.

Données non groupées : on peut faire un test de sous-modèle. OUI SI ... le grand modèle n'est pas trop proche d'un modèle saturé.

**Exercice 4**

## Propriétés asymptotiques et tests

$$\widehat{\theta} \sim_{as} \mathcal{N}(\theta, (X'WX)^{-1})$$

$\theta_j$ : 1 paramètre

$$\frac{\widehat{\theta}_j - \theta_j}{se(\widehat{\theta}_j)} \sim_{as} \mathcal{N}(0,1)$$

Sous $H_0$ : " $\theta_j = 0$",
$$\frac{\widehat{\theta}_j}{se(\widehat{\theta}_j)} \sim_{as} \mathcal{N}(0,1)$$

$\longrightarrow$ Test de Wald



Notations :
$X$ : matrice d'incidence.
$W$ : matrice de pondération (l'algo. d'estimation des paramètres consiste à faire, à chaque itération, une régression linéaire pondérée).

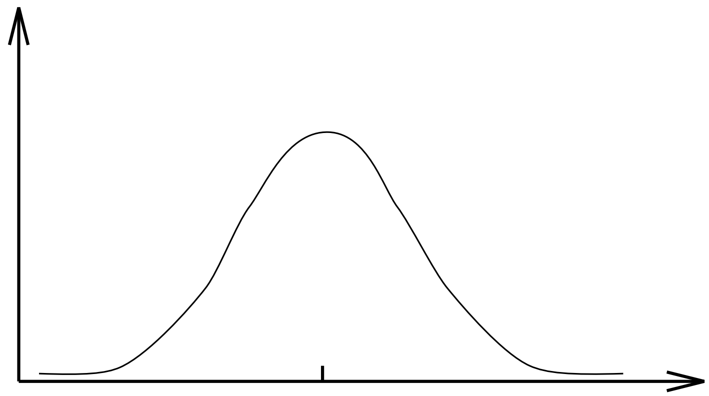
Lorsqu'on a estimé un modèle, on souhaite souvent réaliser des tests sur certains paramètres, tester la nullité de chacune des pentes de l'exemple tournesol, par exemple.

La **loi asymptotique de $\frac{\widehat{\theta}_j}{se(\widehat{\theta}_j)}$ est la loi $\mathcal{N}(0,1)$** sous l'hypothèse de nullité du paramètre.

Remarquons que le terme asymptotique n'a pas le même sens, selon que les données sont groupées ou pas. Si les données sont non groupées (c'est à dire si tous les $n_i$ valent 1), le terme asymptotique signifie que $\sum_i n_i \to \infty$. Si les données sont groupées, il signifie que tous les $n_i$ tendent vers $\infty$ "à la même vitesse", autrement dit que $\frac{n_i}{N} \to C_i$, où $N = \sum_i n_i$ et $C_i$ est une constante non nulle.

**exercice 5**

## loi normale : densité



densité :

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$



On peut passer directement au transparent p. 28, à moins que l'on souhaite insister sur la notion de vraisemblance. L'objectif de ces 4 transparents est de revoir la notion de vraisemblance dans le cadre de la loi normale. La densité de la loi normale est en général plus familière que la densité d'une loi binomiale, et la présentation de la vraisemblance peut être facilitée.

distribution d'une V.A. Y qui suit une **loi normale** $\mathcal{N}(\mu, \sigma^2)$
**densité** de probabilité de la V.A. Y.

Compléter ce graphe avec : $\mu$ et

$$P(Y \in [a,b]) = \int_a^b f(y)dy$$

## loi normale : densité conjointe et vraisemblance

- densité conjointe de $Y_1, \ldots, Y_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$

$$f(y_1, ..., y_n) = \prod_{i=1}^{n} f(y_i)$$

- vraisemblance

$$L(\mu, \sigma; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$$



- Supposons qu'on ait une population, dans laquelle on tire un échantillon composé de $n$ individus. Pour chaque individu, on observe la V.A. Y. On observe donc $n$ valeurs indépendantes $y_1, \ldots, y_n$.

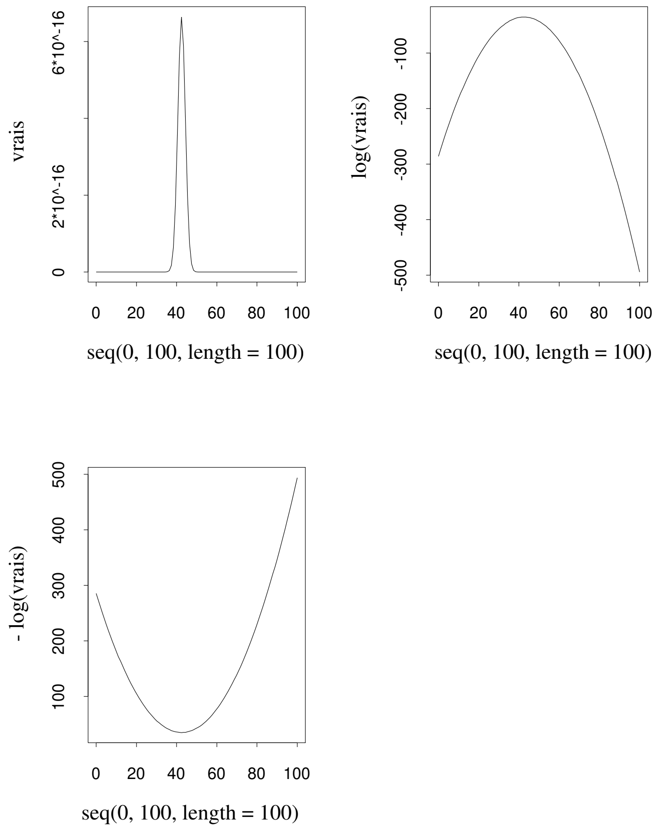
  Q : que vaut $\mu$, la vraie moyenne inconnue ?

  Classiquement, on estime $\mu$ par $\widehat{\mu} = \bar{y}$.

  **$\widehat{\mu}$ est la valeur de $\mu$ la plus vraisemblable**, étant données les observations.

  $\widehat{\mu}$ maximise la vraisemblance, **$\widehat{\mu}$ est l'estimateur du maximum de vraisemblance.**

## loi normale : vraisemblance




$$\widehat{\mu} = 42.5$$



Le but de ce transparent et du suivant est de donner un exemple simple de maximisation de la vraisemblance.

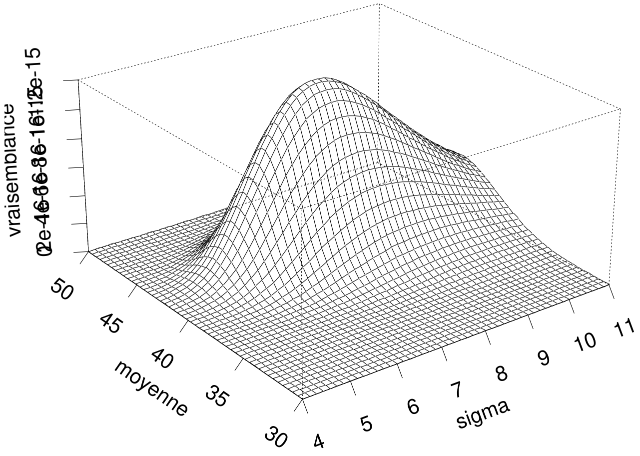
On se place dans le cas d'une population, donc d'un échantillon, et de l'estimation du paramètre $\mu$. On suppose le paramètre $\sigma$ connu. On estime $\mu$ par maximisation de la vraisemblance.

Un échantillon de taille 10 est simulé à partir d'une loi $\mathcal{N}(41, 36)$. Les données obtenues sont : 55 41 41 46 34 28 48 50 38 44.

La vraisemblance est calculée, et représentée graphiquement, ce qui permet d'estimer $\mu$.

```
y_rnorm(10,41,6)

sigma_6

vrais.calc_function(sigma,y,mu)
{
densite_NULL
for (i in 1:length(y))
densite_c(densite,1/(sigma*sqrt(2*pi)) *
                  exp(-((y[i]-mu)^2)/(2*sigma^2)))
prod(densite)
}

vrais_NULL
for (mu in seq(0,100,length=100))
vrais_c(vrais,vrais.calc(sigma,y,mu))

par(mfrow=c(2,2))
plot(seq(0,100,length=100),vrais,type="l")
plot(seq(0,100,length=100),log(vrais),type="l")
plot(seq(0,100,length=100),-log(vrais),type="l")
```

## loi normale : vraisemblance




$\widehat{\mu} = 42.5 \qquad \widehat{\sigma} = 7.92$

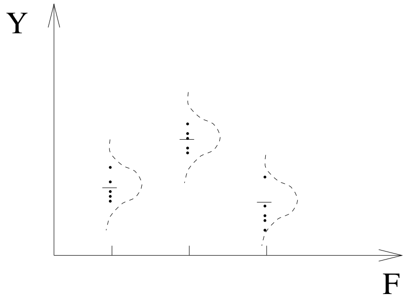


Avec le même jeu de données simulées, on suppose maintenant $\sigma$ inconnu. On estime donc simultanément $\mu$ et $\sigma$.

```
vrais.calc_function(sigma,y,mu)
{
densite_NULL
for (i in 1:length(y))
densite_c(densite,1/(sigma*sqrt(2*pi)) *
                  exp(-((y[i]-mu)^2)/(2*sigma^2)))
prod(densite)
}

vrais_NULL
mu.grille_seq(30,50,length=50)
sigma.grille_seq(4,11, length=50)
for (mu in mu.grille)
{for (sigma in sigma.grille)
vrais_c(vrais,vrais.calc(sigma,y,mu))}

vrais_matrix(vrais,50,50)

persp(sigma.grille,mu.grille,vrais,
   ylab="moyenne",xlab="sigma",zlab="vraisemblance")
```

## loi normale : cas de plusieurs populations

supposons $\sigma^2$ connu




$$\theta = (\mu_1, \mu_2, \mu_3)$$

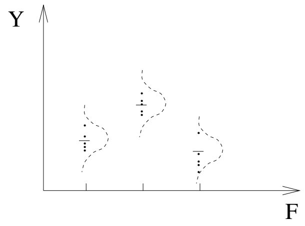
$$L(\theta; y_{11}, \dots) = \prod_{i,j} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_{ij}-\mu_i)^2}{2\sigma^2}\right)$$

$$\begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ y_{31} \\ \vdots \\ y_{3n_3} \end{pmatrix} \xrightarrow[notation]{chgt} \begin{pmatrix} y_1 \\ \vdots \\ y_{n_1} \\ y_{n_1+1} \\ \vdots \\ y_{n_1+n_2} \\ y_{n_1+n_2+1} \\ \vdots \\ y_n \end{pmatrix}$$



L'objectif de ces transparents est de montrer que le modèle linéaire gaussien, avec $\sigma^2$ constant est un cas particulier du modèle linéaire généralisé, que dans ce cas, la déviance est proportionnelle à la somme de carrés résiduelle. On ne traite pas le cas plus compliqué du modèle linéaire gaussien avec $\sigma^2$ non constant, qui lui, n'est pas un cas particulier du modèle linéaire généralisé.

- Supposons qu'on ait maintenant plusieurs populations. exemple : ANOVA à 1 facteur à 3 niveaux. On suppose $\sigma^2$ connu.




$i = 1, \dots, 3$

$j = 1, \dots, n_i$

$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$

$\widehat{\mu}, \widehat{\mu}_1, \widehat{\mu}_2, \widehat{\mu}_3$ à positionner sur le graphe.

- **densité conjointe**

  Les paramètres du modèle sont les $\mu_i$ ($\mu_i = \mu + \alpha_i$).

  Atention, changement de notation, pour des questions de simplicité.

  Après changement de notation, $i$ varie de 1 à $n$. $y_i$ prend $n$ valeurs éventuellement toutes distinctes. $\mu_i$ ne prend que 3 valeurs possibles, $\mu_1$, $\mu_2$ ou $\mu_3$ selon le cas, mais $i$ varie bien de 1 à $n$.

## Application de GLM à la loi normale

Loi normale, $\sigma^2$ connu

$$L = \prod_i \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$\log \widehat{L} = \sum_i \left(\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(y_i-\widehat{\mu}_i)^2}{2\sigma^2}\right)$$

$$\log \widehat{L_s} = \sum_i \left(\log\left(\frac{1}{\sigma\sqrt{2\pi}}\right)\right)$$

$$D = 2\sum_i \frac{(y_i-\widehat{\mu}_i)^2}{2\sigma^2}$$

$$\boxed{D = \frac{1}{\sigma^2}\sum_i (y_i-\widehat{\mu}_i)^2} = \frac{1}{\sigma^2} \qquad \ldots$$



Exercice : cette page pourrait éventuellement être laissée en exercice.

Q : on se place dans le cas d'une distribution normale, et on suppose $\sigma^2$ connu. Que vaut la déviance?

Dans le cas du modèle saturé, les observations sont égales aux valeurs ajustées. Donc, $y_i = \widehat{\mu}_i = \bar{y}_i$.

$\sum_i (y_i - \widehat{\mu}_i)^2$ est la somme des carrés résiduelle.

$$D = \frac{1}{\sigma^2}\mathrm{SC}_{\mathrm{res}}$$

$\mathrm{SC}_{\mathrm{res}} \sim_{ex} \chi^2(\mathrm{dl}_{\mathrm{res}})$ (*ex* : exactement)

$D \sim_{ex} \sigma^2\chi^2(\mathrm{dl}_{\mathrm{res}})$ (notation : cela signifie que $\sigma^2 D \sim \chi^2$)

Dans le cas de la **loi normale**, avec $\sigma^2$ connu, on connaît la **distribution exacte de la déviance**, c'est une **loi de $\boldsymbol{\chi^2}$** . La déviance peut être considérée comme l'analogue, en GLM, de la somme des carrés résiduels, bien connue en modèle linéaire.

## Relation entre LM et GLM

$$\begin{cases} \text{LM} \\ \sigma^2 \text{connu} \end{cases} \Longleftrightarrow \text{GLM} \begin{cases} \text{lien identité} \\ \text{loi normale} \end{cases}$$

1. lien identité

$$\begin{aligned} g(\mu) = X\theta \;&\Rightarrow\; \mu = X\theta \\ &\Rightarrow\; E(Y) = X\theta \\ &\Rightarrow\; \boxed{Y = X\theta + \varepsilon} \end{aligned}$$

2. loi normale

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right)$$

$$-\log(L) = \sum_{i=1}^{n} \left[\log(\sigma\sqrt{2\pi}) + \boxed{\frac{(y_i-\mu_i)^2}{2\sigma^2}}\right]$$

Max de vrais $\Longleftrightarrow$ moindres carrés

Même critère, même estimateur



Comparons le modèle linéaire (où l'on suppose $\sigma^2$ connu), et un cas particulier du GLM, où le lien est le lien identité, et où la distribution est la loi normale.

1. le fait de choisir un lien identité revient à écrire $Y = X\theta + \varepsilon$ comme en modèle linéaire.
2. Si on postule une loi normale, maximiser la vraisemblance revient à minimiser $\sum(y_i - \mu_i)^2$, donc à utiliser un critère des moindres carrés.

Le critère à minimiser est le même. Comme la partie fixe du modèle est la même ($X\theta$), on conclut que l'estimateur est le même, et que les méthodes sont équivalentes.

On montre donc l'**équivalence entre** le **modèle linéaire avec $\sigma^2$ connu**, **et** le **GLM, avec lien identité, et distribution normale**.

## GLM et LM

GLM :

$$\mu = E(Y)$$

$$g(\mu) = X\theta$$

hyp. loi de $Y$,
$Var(Y) = fct(\mu) = V(\mu)$,
observations indépendantes.

Maximum de vraisemblance

LM :

$$Y = X\theta + \varepsilon$$

hyp. loi normale
$Var(Y) = \sigma^2 = Cte$
observations indépendantes.

Moindres carrés



Tout modèle linéaire avec $\sigma^2$ connu est un cas particulier du modèle linéaire généralisé, c'est un modèle linéaire généralisé avec loi normale et lien identité. Dans les deux cas, on suppose l'indépendance entre les observations. En **modèle linéaire**, la loi supposée des observations est la **loi normale** (compte tenu des facteurs explicatifs), et la **variabilité** est **homogène** ($\sigma^2$ est constant, il n'est pas indicé par $i$), en **modèle linéaire généralisé**, on peut supposer un **autre type de loi**. C'est le **choix de la loi** qui **détermine implicitement la modélisation de la variance**, c'est à dire la manière dont la variance varie en fonction de l'espérance. Par exemple, dans le cas de la loi binomiale, $Var(\frac{Y_i}{n_i}) = \frac{p_i(1-p_i)}{n_i}$, la variance varie avec $p_i$.

Remarque : sur ce transparent, dans la notation $V(\mu)$, V ne signifie pas variance, mais cela signifie que la variance de Y est une fonction de $\mu$.

En **modèle linéaire**, c'est **dans l'échelle initiale** des données que l'on a **additivité des effets**, et des effets du type de ceux que l'on décrit dans la partie systématique du modèle (effets linéaires, quadratiques, ... ). En **modèle linéaire généralisé**, c'est **dans l'échelle de la fonction de lien**, que l'on a **additivité des effets**, et des effets du type de ceux qu'on décrit dans la partie systématique du modèle. Par exemple, avec un lien log, on a additivité des effets dans l'échelle du lien log, mais des effets multiplicatifs, dans l'échelle initiale. Si on introduit dans le modèle un effet linéaire de la variable X, c'est dans l'échelle log (si on a choisi un lien log) que l'on a un effet linéaire.

Dans le cas du **modèle linéaire généralisé**, l'estimateur est l'estimateur du **maximum de vraisemblance**. Dans le cas du **modèle linéaire**, c'est l'estimateur des **moindres carrés**, qui est exactement l'estimateur du maximum de vraisemblance si $\sigma^2$ est connu.

## RÉSUMÉ

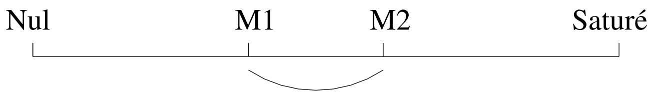
$$\boxed{\mu = E(Y) \qquad g(\mu) = X\theta \qquad \text{loi de Y}}$$

$$L(\theta) = \prod_{i=1}^{n} \text{densités}$$

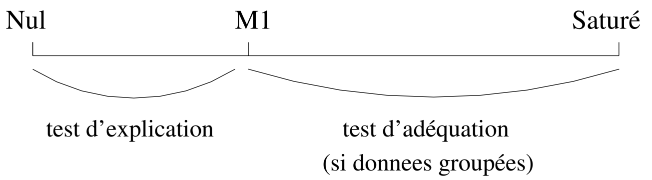
Max. de vraisemblance $\longrightarrow \widehat{\theta}$

$$\frac{\widehat{\theta}_j - \theta_j}{se(\widehat{\theta}_j)} \sim_{as} \mathcal{N}(0,1) \qquad \text{Test de Wald}$$

Test de sous-modèle : test du rapport de vrais.




$$D = -2 \log \frac{L_{M_1}}{L_{M_2}} \sim_{as} \chi^2(dl_1 - dl_2)$$






On choisit la fonction de lien g, la loi de Y, et les termes intervenant dans la partie systématique du modèle. C'est le choix de la loi de Y qui détermine implicitement le choix de la fonction de variance.

On estime les paramètres par la méthode du maximum de vraisemblance, la vraisemblance étant définie comme une fonction des paramètres, égale au produit des densités.

On dispose de résultats asymptotiques concernant les paramètres estimés. Ils suivent asymptotiquement une loi normale. On peut ainsi réaliser des tests, dits tests de Wald, sur les paramètres. Ces tests sont “valables” dès que le nombre d'observations est suffisant. Le cadre asymptotique suppose $\sum_i n_i \to \infty$ si les données sont non groupées, et tous les $n_i \to \infty$ (à la même vitesse) si les données sont groupées.

Tous les tests, qu'il s'agisse de tests réalisés dans la phase de recherche d'un modèle, ou de tests réalisés dans la phase d'exploitation d'un modèle, peuvent être réalisés au moyen de tests de sous-modèles. Ce sont des tests du rapport de vraisemblance, qui permettent de comparer deux modèles emboîtés. La statistique de test est une déviance, elle est égale à la différence des -2 log(vrais) des deux modèles, ou (c'est équivalent) à -2 fois le log du rapport des vraisemblances. Elle suit, sous l'hypothèse nulle d'égalité entre les deux modèles, une loi de $\chi^2$ dont le nombre de degrés de liberté est égal à la différence des degrés de liberté entre les deux modèles. Ce test est un test asymptotique également.

Ce principe, appliqué au cas où le sous-modèle est le modèle nul, permet de réaliser un test d'explication. Ce test permet de s'assurer que les régresseurs et facteurs apportent une explication significative.

Le test d'adéquation est un test asymptotique qui permet de s'assurer que toute la variabilité a bien été prise en compte par les facteurs et régresseurs. Ce résultat asymptotique est valide si tous les $n_i$ tendent vers $\infty$ (à la même vitesse que $\sum_i n_i$). Ce test d'adéquation n'est donc valide que si les données sont groupées “avec des $n_i$ pas trop petits”.