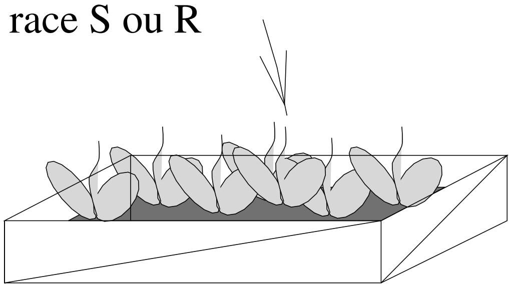
Ex. Tournesol : l'expérience

mildiou

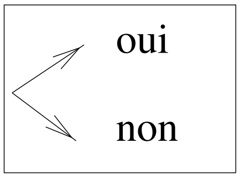
race S ou R



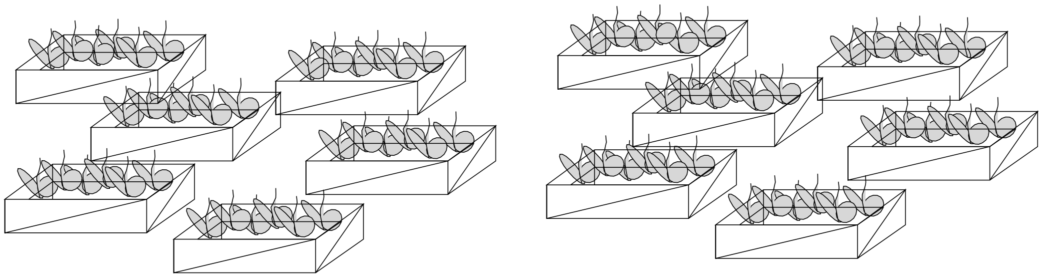
7 plantules

pour chaque plantule :

sporulation



RACE sensible RACE résistante



différentes doses

Ex. Tournesol : les données

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

Ex. Tournesol : schéma d'analyse

Données : graphiques

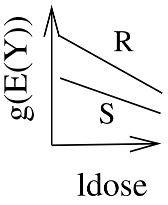
↓

Modèle GLM

g (E (Y)) = $\alpha_{race} + \beta_{race}$ .ldose

fct de lien :

loi :



↓

Estimation :

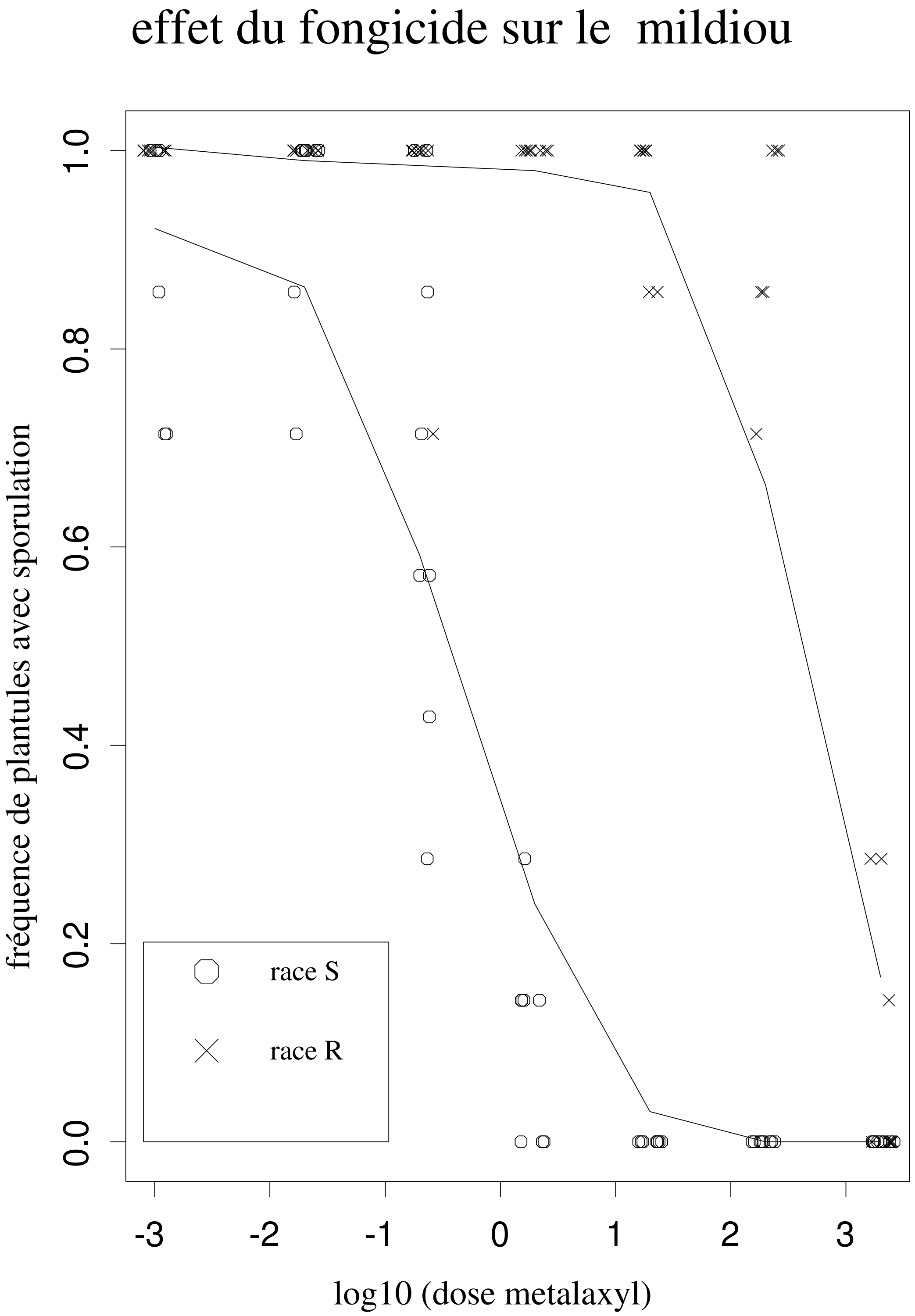
↓

Résultat → graphique
Résultat → numérique

↓

Validation

Ex. Tournesol : exploration des données

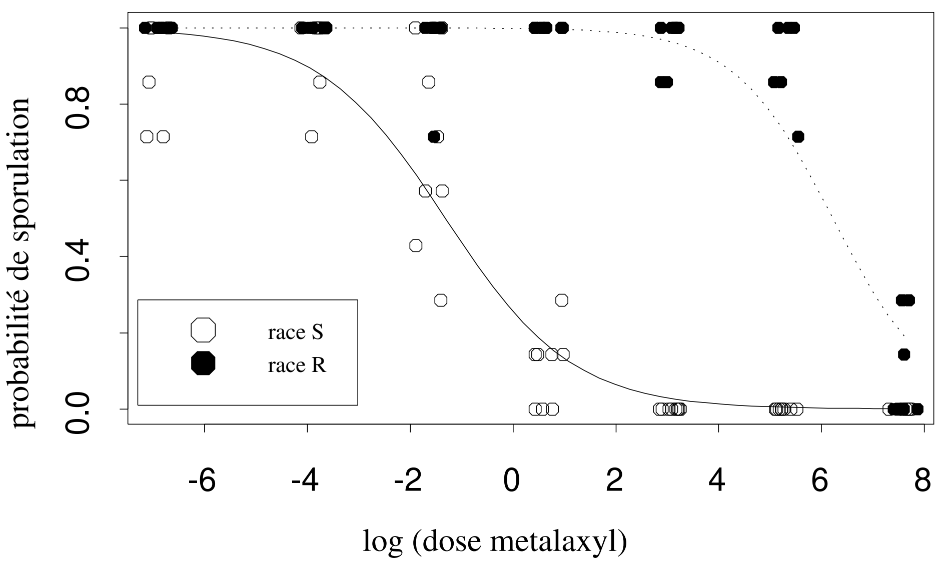


Ex. Tournesol : le modèle

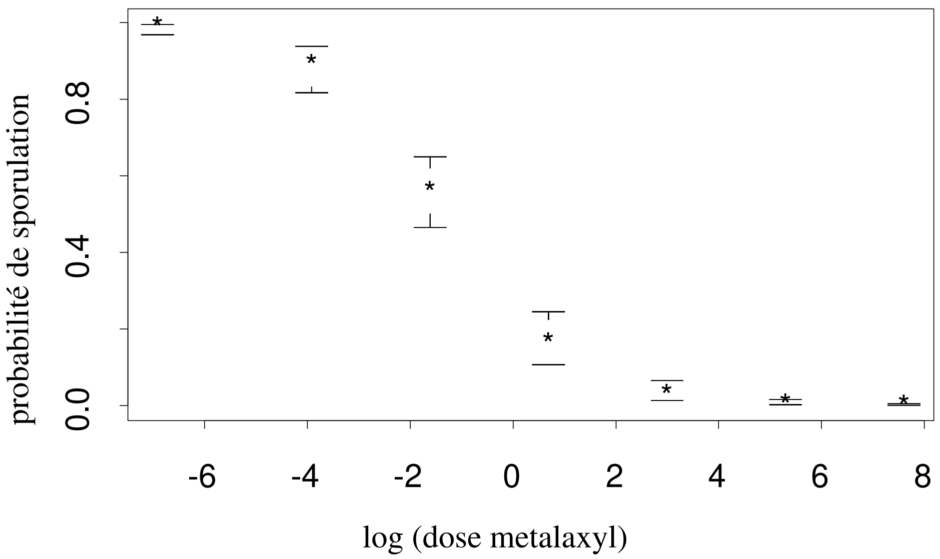
$$\text{GLM} : \log\left(\frac{p}{1-p}\right) = \alpha_{race} + \beta_{race}.ldose$$

---

effet du fongicide sur le mildiou



Race sensible



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
```

Ex. Tournesol : analyse

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

```
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
```

Ex. Tournesol : analyse

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
Scaled Pearson X2	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.

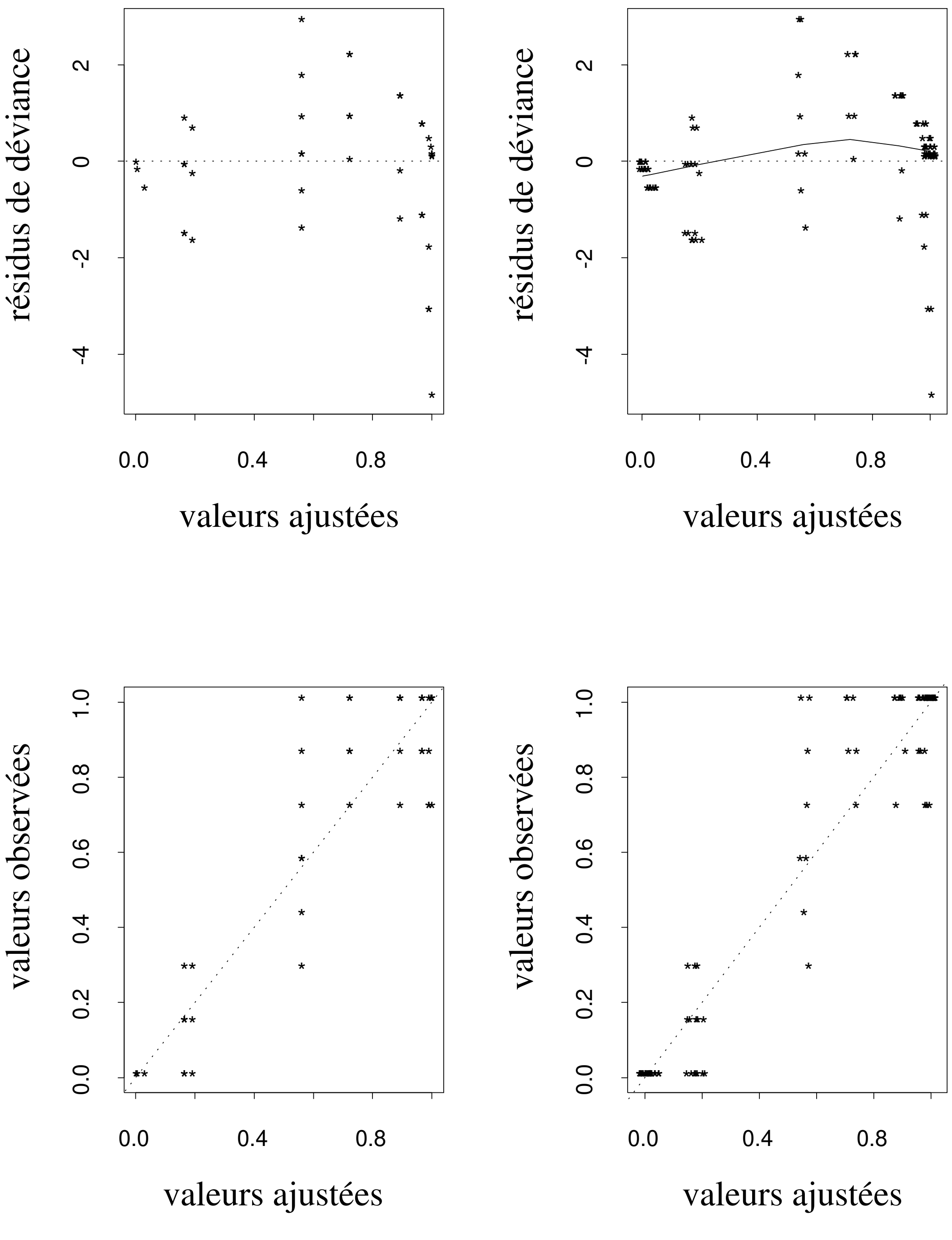
Ex. Tournesol : analyse

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

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

Ex. Tournesol : validation



Ex. Tournesol : comparaison de sous-modèles

M1 : modèle avec deux droites quelconques

M2 : modèle avec deux droites parallèles

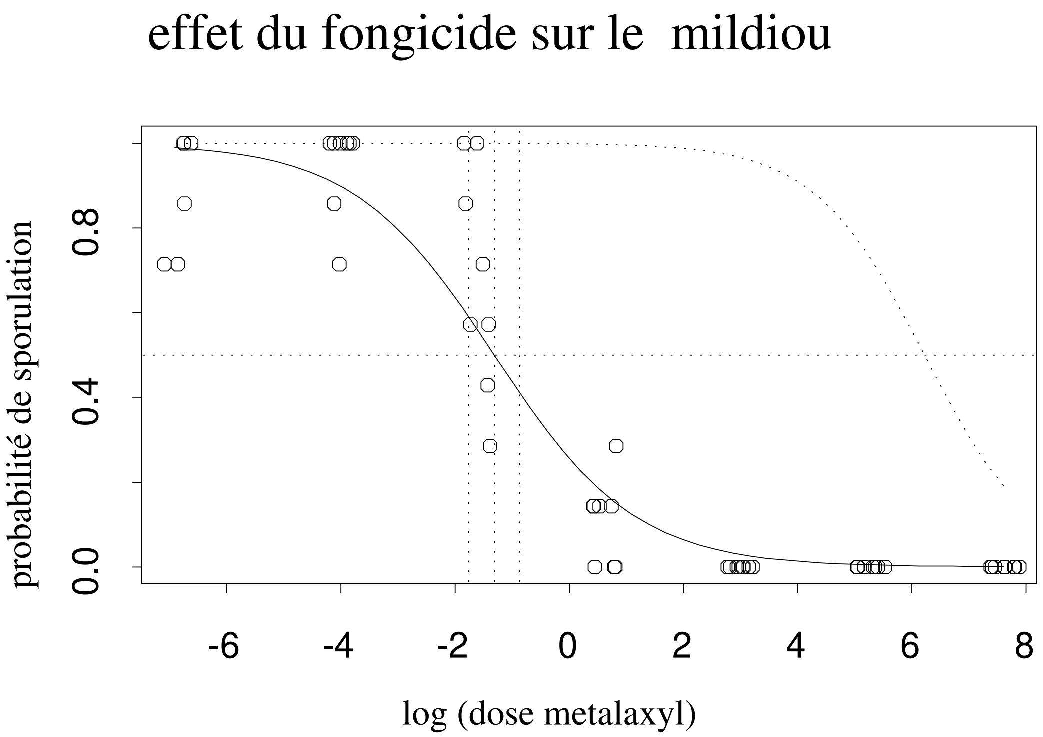
M3 : modèle avec une droite

```
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
```

Ex. Tournesol : estimation de DE_{50}



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

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

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 :

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

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

Loi binomiale : densité et densité conjointe

- **densité** de Y binomiale $B(n,p)$
  fonction $y \to P(Y = y)$

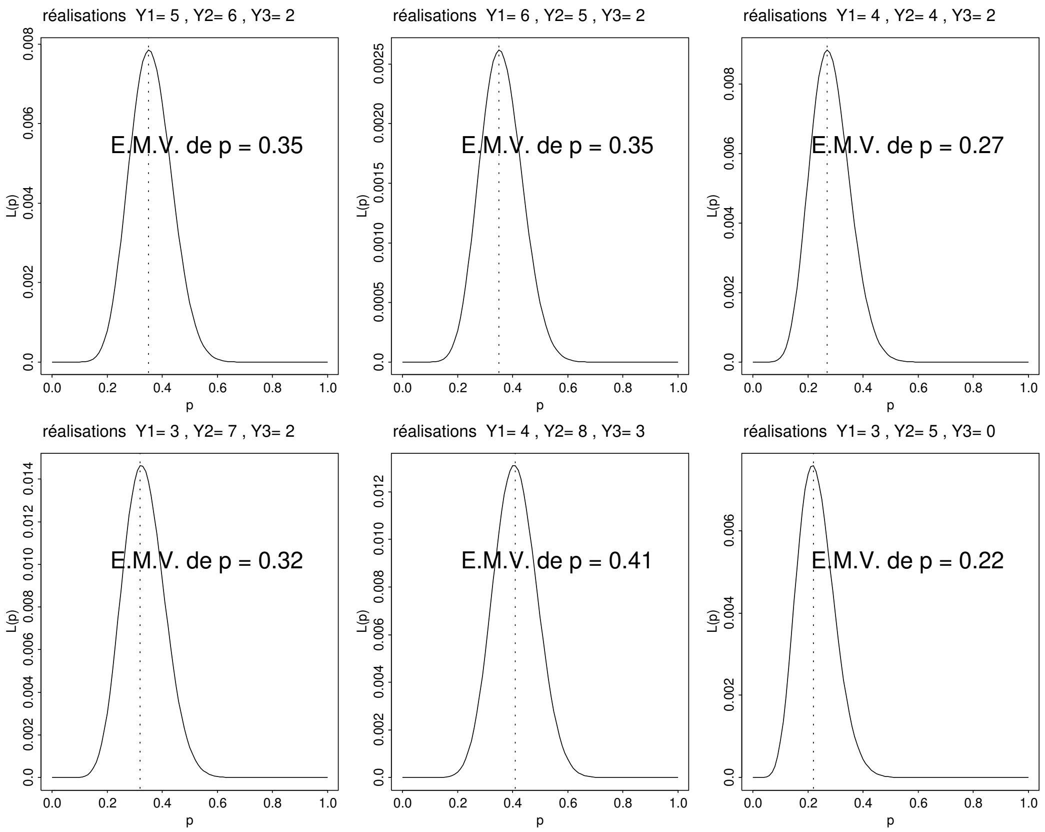
- **densité conjointe**
  variables indépendantes Y_i
  $Y_i \sim B(n_i, p_i) \qquad i = 1, ..., 3$
  c'est la densité du triplet, soit
  $(y_1, y_2, y_3) \to P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$

  mais encore...

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 :
  $$p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \to L(p; y_1, y_2, y_3)$$
  $$= P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3)$$
- plus généralement, si $p = p(\theta)$
  $L(\theta; y_1, y_2, y_3)$ vraisemblance de θ
  (ou $L(\theta)$ s'il n'y a pas d'ambiguïté)
  (L pour likelihood)
- **Maximum de vraisemblance** :
  Rechercher θ qui maximise L,
  soit à résoudre : $\frac{\partial}{\partial\theta} L = 0$.

Loi binomiale - vraisemblance (simulation)



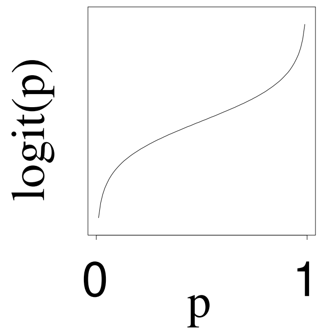
Modèle

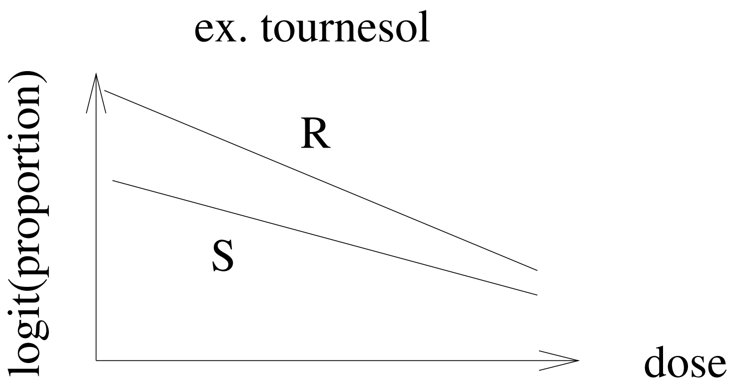
$$
\begin{aligned}
log\left(\frac{p}{1-p}\right) &= \alpha_{race} + \beta_{race} ldose \\
g(p) &= X\theta = \eta
\end{aligned}
$$

Modèle : fonction de lien

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

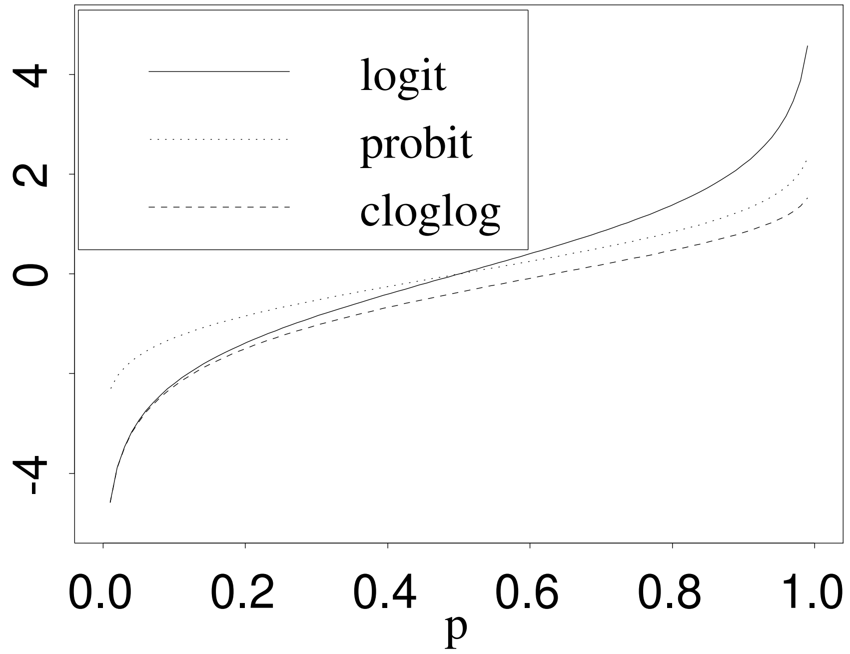
$$\text{logit} : [0,1] \longrightarrow]-\infty,+\infty[$$





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

Modèle : fonction de lien

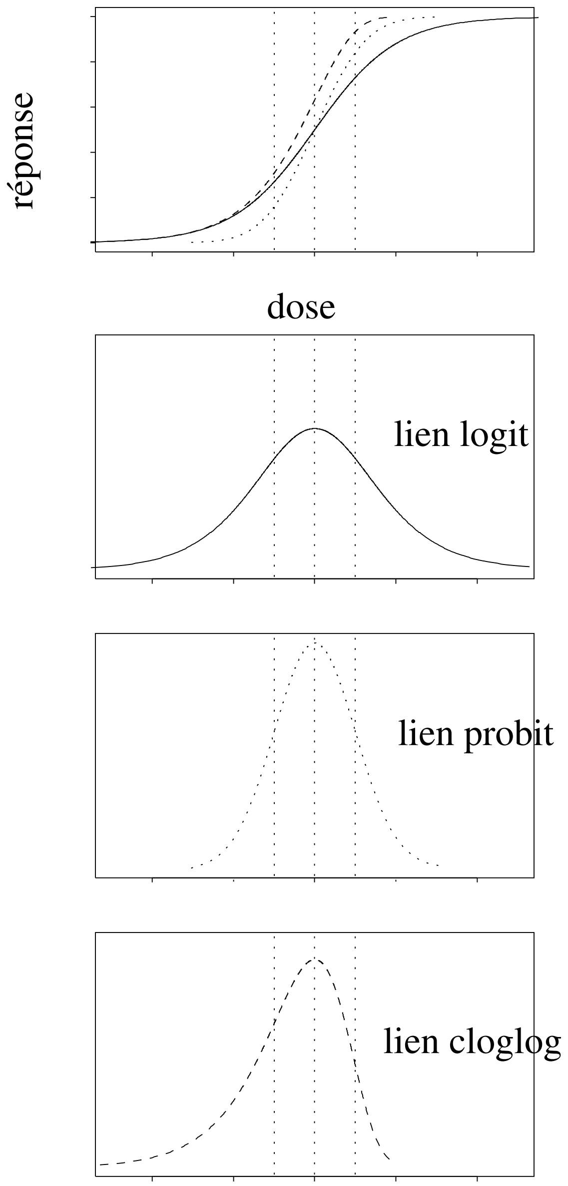


Loi binomiale

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

Fonction de lien et distribution de tolérance

Courbe dose-réponse

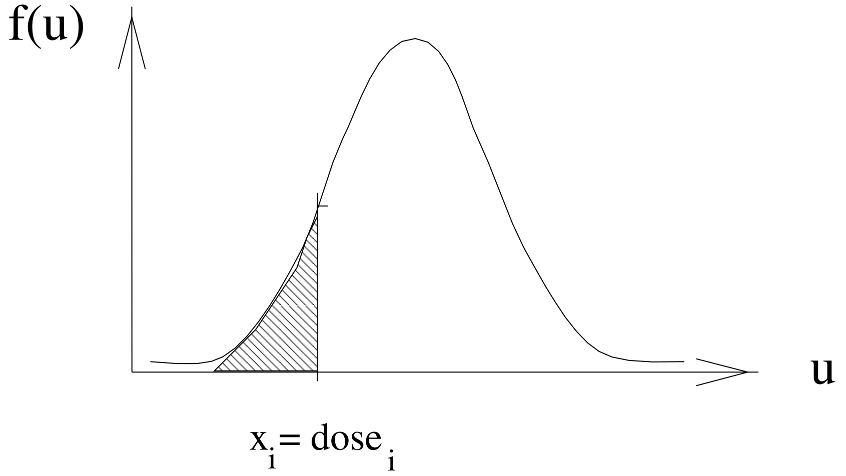


Fonction de lien et distribution de tolérance

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 :



$$p_i = P(Y_i = 1) \text{ pour 1 insecte si dose } x_i.$$

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

Fonction de lien et distribution de tolérance

Supposons les seuils distribués normalement

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

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

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

ϕ : 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

Modèle et vraisemblance

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

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

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

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

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

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

Estimation des paramètres

Maximiser

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

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

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

$\Rightarrow$ méthode itérative

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

Déviance

$$\begin{array}{ccc} & L & L_s \\ \vdash & \vert & \dashv \\ & M & M_s \end{array}$$

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

Déviance

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

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

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

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

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

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

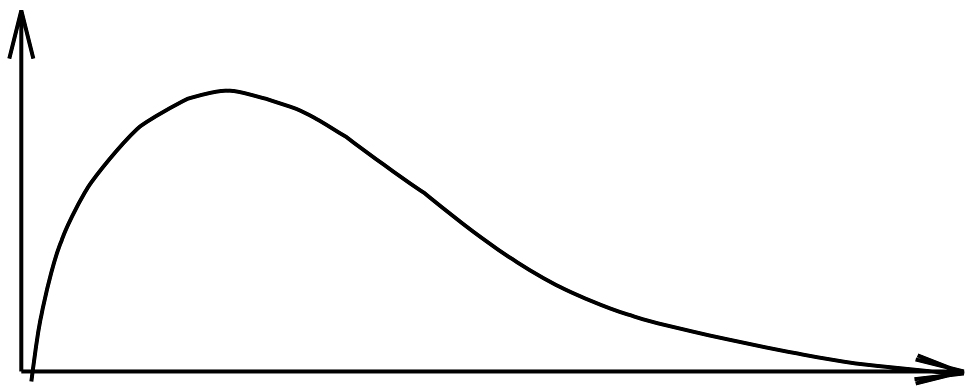
où $\quad \widehat{y_i} = n_i \widehat{p_i}$

Loi de la déviance

Données groupées :

Si n_i pas trop petits,

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



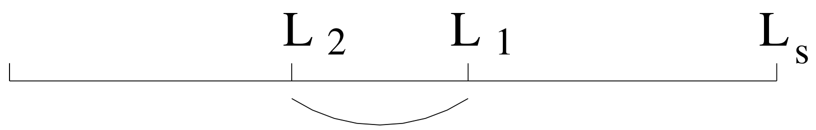
Test du rapport de vraisemblance

Comparer 2 modèles emboîtés :
test de sous-modèle

$M_1 \to L_1$

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

$M_2 \to L_2$



sous H_0 : “M_2 équivalent à M_1”

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

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

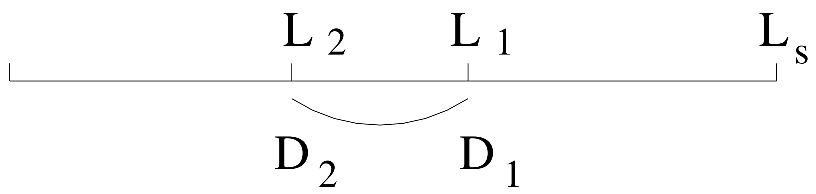
Test du rapport de vraisemblance

Test de sous-modèle

M_1
M_2 : sous-modèle

$$
\begin{aligned}
& (-2\log L_2) - (-2\log L_1) \\
= \; & (-2\log \tfrac{L_2}{L_s}) - (-2\log \tfrac{L_1}{L_s}) \\
= \; & D_2 - D_1
\end{aligned}
$$

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



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

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"

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

Données groupées
Données non groupées

Données groupées	$n_i > 1$
Données non groupées	$n_i = 1$

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

Propriétés asymptotiques et tests

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

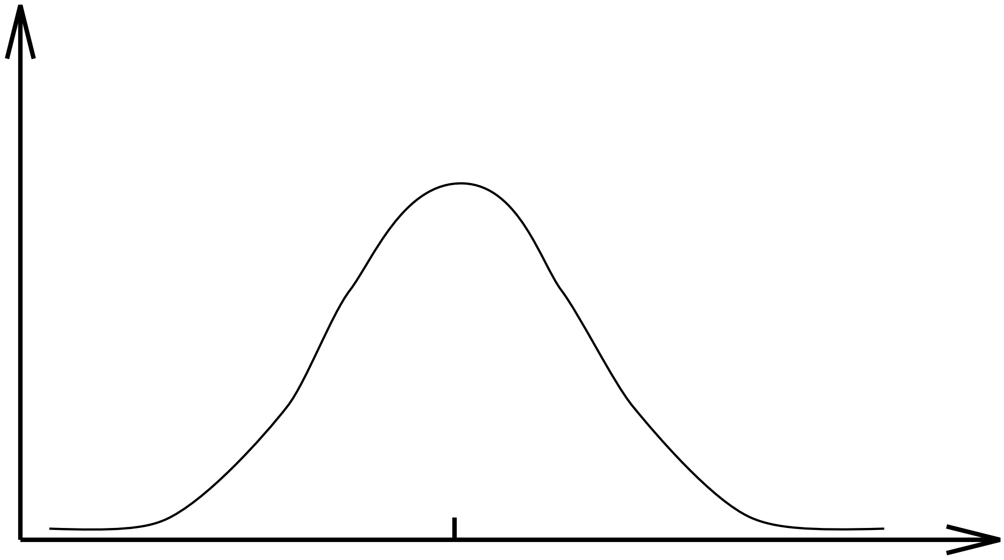
θ_j : 1 paramètre

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

Sous H_0 : “ $\theta_j = 0$”, $\quad \dfrac{\widehat{\theta}_j}{se(\widehat{\theta}_j)} \sim_{as} \mathcal{N}(0,1)$

$\longrightarrow$ Test de Wald

loi normale : densité



densité :

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

loi normale : densité conjointe et vraisemblance

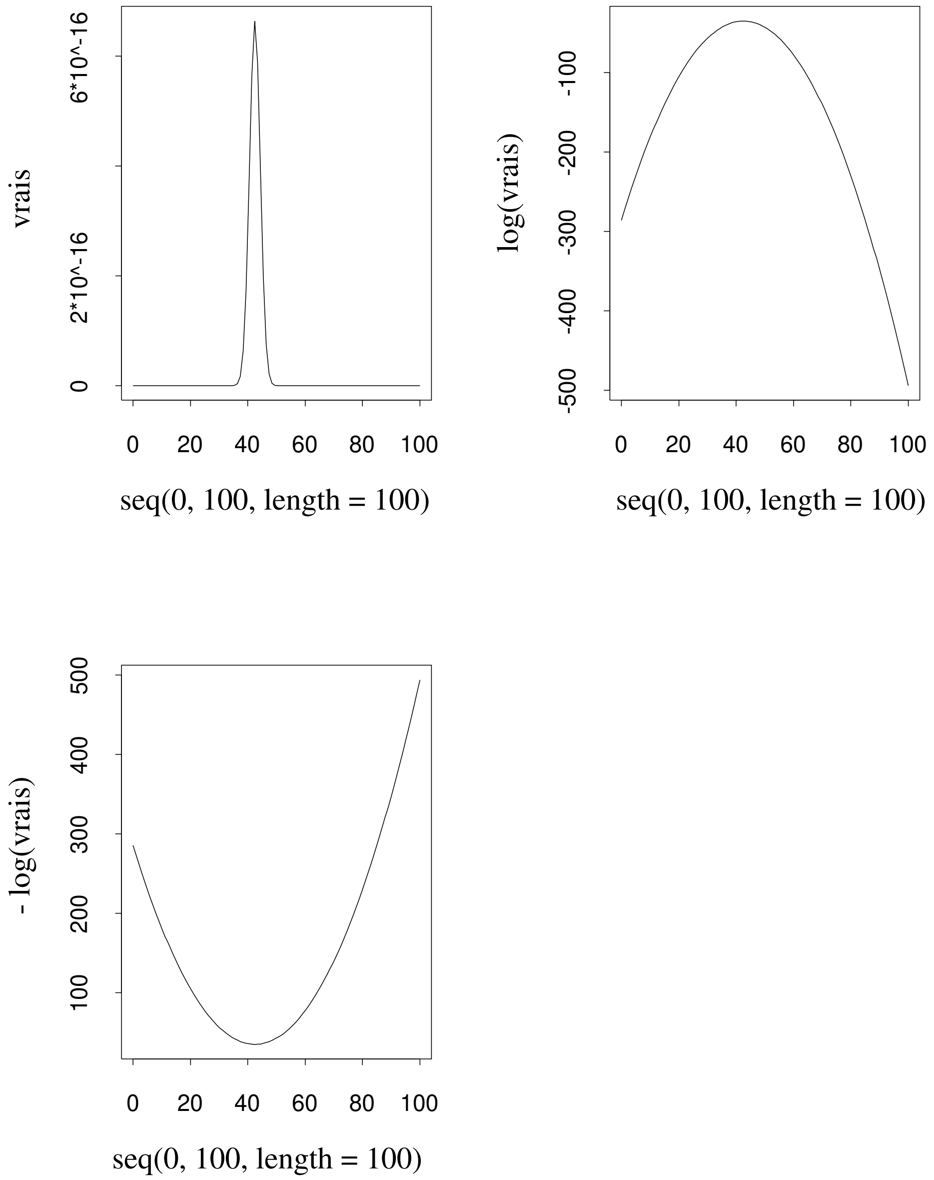
- densité conjointe de $Y_1, \ldots, Y_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$

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

- vraisemblance

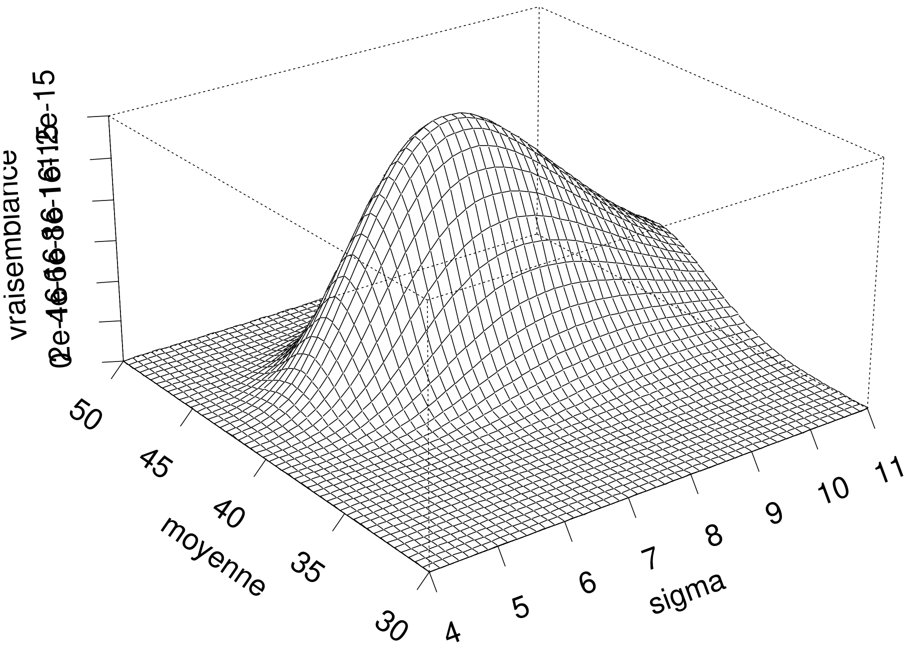
$$L(\mu, \sigma; y_1, \ldots, y_n) = \prod_{i=1}^{n} f(y_i)$$

loi normale : vraisemblance



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

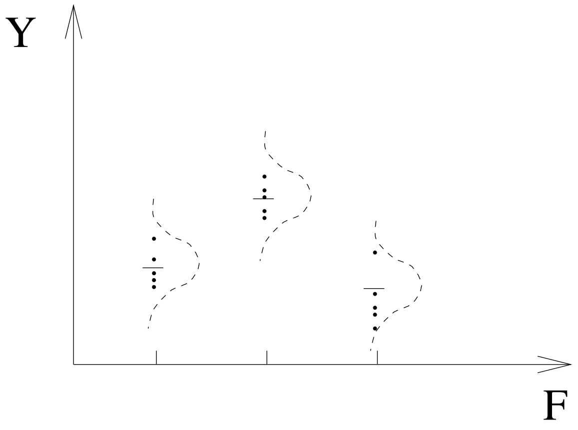
loi normale : vraisemblance



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

loi normale : cas de plusieurs populations

supposons σ^2 connu



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

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

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

Application de GLM à la loi normale

Loi normale, σ^2 connu

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

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

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

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

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

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

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

RÉSUMÉ

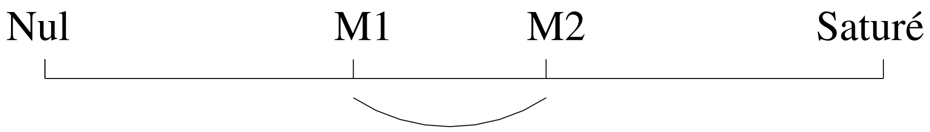
$$\boxed{\mu = E(Y) \qquad g(\mu) = X\theta \qquad \text{loi de Y}}$$

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

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

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

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



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