Ex. seuil, intérêts, particularités

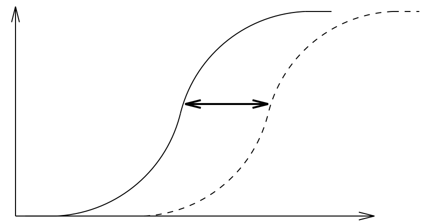
1. $\boxed{\text{fct de lien dérivée}}$ du lien logit
2. $$\left.\begin{array}{r}\text{I.C. autour de la DE}_{50}\\ \text{des proba.}\end{array}\right\} \rightarrow \begin{array}{c}\text{approximation}\\ \text{ou}\\ \boxed{\text{Max. de vrais.}}\end{array}$$
3. $\boxed{\text{Surdispersion}}$

   $\rightarrow$ tenir compte des effets séance et individu
4. $\boxed{\text{“efficacité relative”}}$

L'exemple “seuil” est une application directe du GLM, qui permet de présenter une extension du lien logit, le calcul d'un intervalle de confiance autour d'une fonction des paramètres (qui n'est pas forcément une combinaison linéaire des paramètres), la notion de surdispersion, ainsi que la notion d'efficacité relative.

La fonction de lien dérivée du lien logit est introduite par Collett (chap. V p.4) pour tenir compte d'une mortalité naturelle. L'estimation de la DE_{50} est traitée par Collett (chap. V p.2). La surdispersion est traitée par Collett (chap.6), et l'efficacité relative aussi (chap.4.3).

La fonction de lien dérivée et la surdispersion seront traités en détail dans ces transparents. L'efficacité relative ne sera pas développée. Pour plus de détail, on pourra se reporter à Collett. Il s'agit simplement de la comparaison des perceptions dans la base fromagère et la solution tampon. On suppose les pentes égales, et on estime le décalage horizontal.

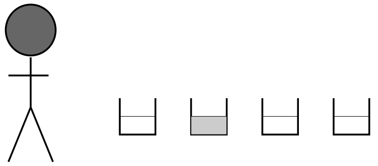


Cet exemple peut aussi être le support d'une réflexion sur la distribution supposée des seuils individuels, qui oriente le choix de la fonction de lien. (voir Collett, chap 4.1 : distribution de tolérance, chap5 : choix de la transformation log pour la concentration).
Lorsque la distribution des seuils est fortement dissymétrique, la transformation log appliquée aux concentrations permet d'obtenir une distribution des seuils plus symétrique, et donc d'envisager l'utilisation du lien logit ou du lien probit.

Exemple seuil

⇒ Déterminer un seuil de détection

Test à choix forcé, 1 parmi 4



Q : Lequel diffère?

$$\left\langle \begin{array}{ll} Y = 0 & \text{PERDU} \quad ! \\ Y = 1 & \text{BRAVO} \quad ! \end{array} \right.$$

L'objectif de l'étude est de déterminer la concentration nécessaire, pour que 50 % des sujets détectent une molécule (en goûtant ou en sentant, la molécule pouvant être intégrée dans différents substrats : eau, solution tampon, base fromagère, ...)

On réalise des tests à choix forcé 1 parmi 4, c'est à dire que pour chaque test, un seul des gobelets contient la molécule étudiée.

On obtient des observations binaires.

- 6 à 8 concentrations sont testées à chaque séance.
- plusieurs séances sont nécessaires, car le nombre de tests par sujet et par séance est limité.
- environ 20 sujets par séance.

  Ce sont les mêmes sujets, d'une séance à l'autre, à part les absents.

Ex. seuil, modèle, fct de lien

$$p = P(Y = 1) \qquad \begin{cases} \text{détection réelle} & \Rightarrow\ p_c \text{ (corrigée)} \\ \text{hasard} & \Rightarrow\ \pi_0 = 1/4 \end{cases}$$

$$\boxed{p = p_c + (1 - p_c)\pi_0}$$

$$\boxed{\text{logit}(p_c) = \log(\frac{p_c}{1 - p_c}) = \alpha + \beta \log(conc)}$$

or, $p_c = \frac{p-\pi_0}{1-\pi_0}$, $1 - p_c = \frac{1-p}{1-\pi_0}$

$$\boxed{\log(\frac{p - \pi_0}{1 - p}) = \alpha + \beta \log(conc)}$$

- Le sujet doit donner une réponse. S'il a donné la bonne réponse, c'est soit par hasard, soit parce qu'il a réellement détecté la substance.

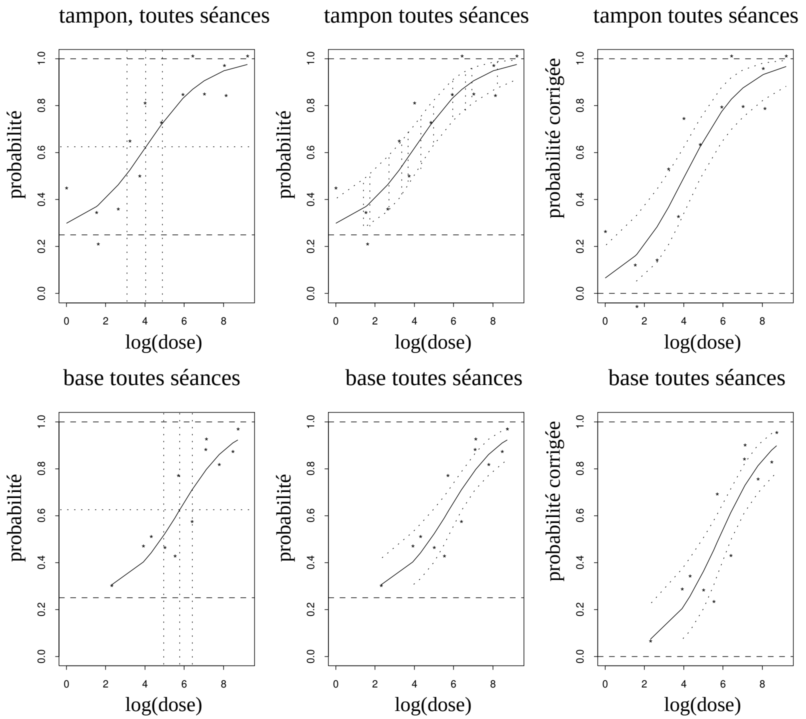
  On note p la probabilité de bonne réponse, p_c la probabilité corrigée qui est en fait la probabilité de détection réelle.

- L'ensemble des bonnes réponses est composé des bonnes réponses obtenues par hasard, avec la probabilité $(1 - p_c)\ \pi_0$ (il n'y a pas détection, et il y a bonne réponse au hasard) et des bonnes réponses obtenues par détection réelle avec la probabilité p_c.

- Une nouvelle fonction de lien s'introduit naturellement. Elle est dérivée du lien logit.

  On écrit le modèle linéaire généralisé sur la probabilité de détection réelle p_c, qui est liée au log(conc), par l'intermédiaire de la fonction de lien logit. On remplace p_c par son expression en fonction de p et de π_0 . On obtient un modèle pour p. On a introduit naturellement une nouvelle fonction de lien, dérivée du lien logit, et qui tient compte du paramètre connu π_0. $\pi_0 = 1/4$: est une probabilité de base, analogue à une "réponse naturelle", ou "mortalité naturelle", selon le contexte.

Ex. seuil, résultats



Le but de l'étude est la détermination du seuil de perception d'une molécule responsable du goût chèvre.

La molécule est testée dans une solution tampon, ou dans une base fromagère. Trois séances de tests, pour une vingtaine de sujets à chaque séance, ont été réalisées.

Les proportions observées sont représentées, en fonction du log de la concentration.
La courbe estimée est superposée.

On définit le seuil comme la concentration détectée dans 50 % des cas. On cherche donc à estimer la concentration correspondant à une probabilité corrigée égale à 0.5, et un intervalle de confiance. Les résultats sont représentés sur le premier et le quatrième graphiques.

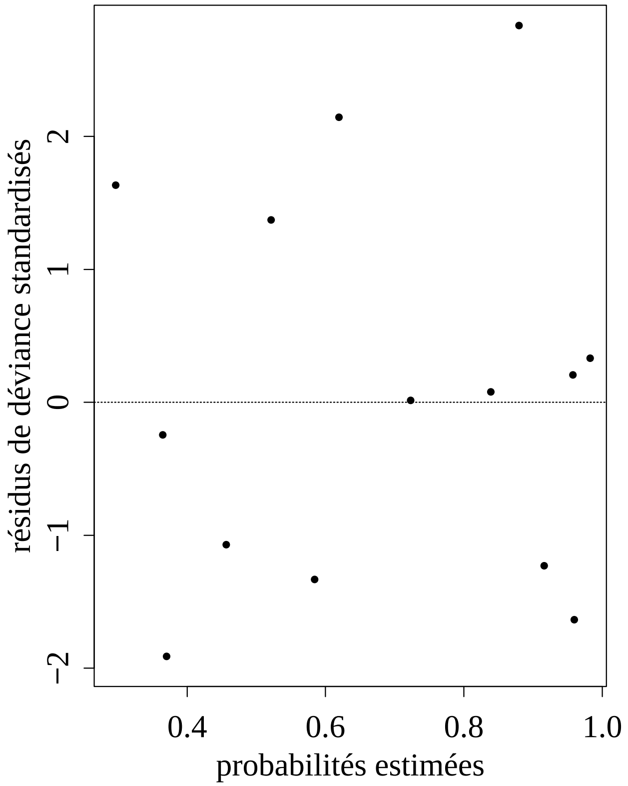
On estime ensuite, pour chaque concentration observée, un intervalle de confiance autour de la probabilité (il ne s'agit pas d'un intervalle de confiance pour un individu, mais d'un intervalle de confiance pour la probabilité $\widehat{p}$). On joint les bornes des intervalles de confiance ponctuels ainsi obtenus. Les résultats sont représentés dans les autres graphiques, dans l'échelle des probabilités d'abord, dans l'échelle des probabilités corrigées ensuite.

Surdispersion, introduction

```
Deviance Residuals:
  Min   1Q Median 3Q Max
 -1.8 -1.1 0.0085  1 2.6

Coefficients:
                   Std.
            Value  Error tvalue
(Intercept) -2.80   0.71   -3.9
  log(dose)  0.69   0.15    4.7

(Dispersion Parameter
  taken to be 2 )

   Null Deviance: 104 on 13 df

Residual Deviance: 24 on 12 df

Number of
Fisher Scoring Iterations: 4
```


Le modèle $\log(\frac{p-\pi_0}{1-p}) = \alpha + \beta \log(conc)$ est estimé pour les données obtenues avec la solution tampon. La déviance (Residual Deviance) est environ le double du nombre de degrés de liberté résiduels (12). Le paramètre de surdispersion est supérieur à 1, il vaut 2 (24/12).

Surdispersion, introduction



Pourtant, on n'observe pas de structure systématique des résidus. On en conclut que la fonction de lien, et le régresseur (logarithme de la concentration) conviennent.

Si l'on fait l'hypothèse que le modèle a été correctement écrit, on interprète comme un phénomène de surdispersion le fait que la déviance est plus grande que la déviance attendue, alors que les résidus ne présentent pas de structure particulière.

Les résidus sont surdispersés, leur variance est plus grande qu'attendue sous l'hypothèse de lois binomiales indépendantes.

Surdispersion, introduction

Loi binomiale

$Y \sim B(n,p)$ $E(Y)=np$ $\underbrace{V(Y)=np(1-p)}_{\text{si les n obs. indépendantes}}$

Si dépendance, variance $\neq n\,p\,(1-p)$

~~binomiale~~

obs. binaires corrélées positivement → variance $> n\,p\,(1-p)$ surdispersion

obs. binaires corrélées négativement → variance $< n\,p\,(1-p)$ sousdispersion

On n'a de loi binomiale que dans le cas où les observations binaires sont indépendantes. Ce qui suppose, dans notre cas, que toutes les réponses sont indépendantes les unes des autres, qu'elles sont indépendantes d'une concentration à l'autre, au cours de la même séance, d'une séance à l'autre ...

Cela suppose donc que la probabilité de bonne réponse donnée par un individu ne dépend ni de l'individu, ni de la séance. Cela revient à considérer que l'unité expérimentale est le test (fait par un individu à un instant donné), et à supposer que le fait de faire une série de tests sur un même individu, revient à faire les mêmes tests sur des individus différents.

Des observations binaires corrélées positivement produisent une surdispersion, c'est à dire une variance plus grande que celle attendue sous l'hypothèse d'indépendance. Au contraire, des observations binaires corrélées négativement produisent une sousdispersion. Cela est moins fréquent que la surdispersion. Cela reflète des phénomènes de compétition. Imaginons une compétition entre des arbres pour leur croissance. Si un arbre est particulièrement grand, ses voisins ont une plus forte probabilité d'être petits.

La surdispersion est parfois nommée "extra-binomial variation", ou hétérogénéité.

Rq : le formateur peut ajouter un indice i à chacun des n et des y.

Surdispersion, diagnostic

approx. χ^2 de la déviance pas valide

– si données binaires (non groupées)

– si certains des n_i petits.

déviance grande si :

- modèle inapproprié
  - $\rightarrow$ var. expl. oubliée
  - $\rightarrow$ var. expl. à transformer
  - $\rightarrow$ outlier
  - $\rightarrow$ fct de lien inadéquate
- $\boxed{\text{surdispersion}}$

$\boxed{\text{Surdispersion} \Rightarrow \text{déviance plus grande}}$

Il faut se souvenir que la déviance n'est pas une mesure de la qualité d'ajustement, si les données sont binaires, non groupées. La loi de la déviance est alors inconnue. Une déviance grande ne donne aucune indication. Ceci est vrai pour les données binaires non groupées (tous les $n_i = 1$), mais c'est aussi vrai dans des cas proches, où certains des n_i sont petits.

Dans les autres cas, la surdispersion se traduit par une déviance supérieure à celle que l'on attend, sous l'hypothèse d'indépendance entre les observations binaires.

$D \sim_{as} \chi^2(ddl)$, si indépendance

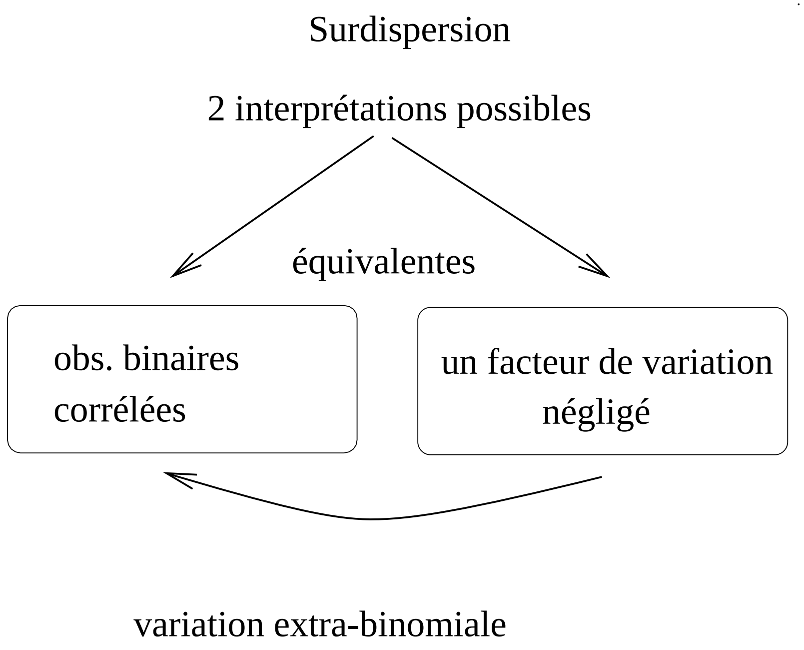
$$E(\chi^2(ddl)) = ddl$$

Toute déviance supérieure au $(1 - \alpha)$ quantile d'un $\chi^2(ddl)$ est suspecte.

Mais il peut y avoir plusieurs explications à une déviance élevée. Avant de conclure à une surdispersion, il faut éliminer les autres causes possibles.

- la déviance peut être trop grande, si le modèle est inadéquat. Il faut utiliser les méthodes de validation, pour s'assurer de l'adéquation du modèle.
- Enfin, la dernière cause possible est la surdispersion.

Surdispersion, interprétation intuitive



Les 2 interprétations sont strictement équivalentes.

<u>intuitivement :</u> les observations binaires sont corrélées positivement, parce qu'elles ont quelque chose en commun, elles appartiennent au même individu, ou à la même séance.

Si un individu a une probabilité élevée, c'est qu'il perçoit bien la molécule, son seuil de perception est élevé. Un nouveau test chez le même individu a toutes les chances de donner un bon résultat. Les 2 probabilités sont corrélées. Même chose pour une séance : si, à une séance, 50 % des sujets sont enrhumés, la probabilité d'un individu a de fortes chances d'être faible, celle de son voisin aussi. Les probabilités sont corrélées.

Les observations sont corrélées parce qu'elles ont quelque chose en commun, elles ont en commun "l'appartenance" à un même individu, ou à une même séance.

<u>Autres exemples :</u> effet portée, famille, fermenteur, boîte de pétri, région, . . . qui peuvent s'interpréter en termes de corrélation entre individus d'une même portée, famille, . . .

L'existence d'une corrélation entre les réponses binaires est indissociable de l'existence d'un facteur de variation négligé. En effet, la corrélation entre les données binaires conduit à des variations entre les probabilités de réponse et vice-versa.

<u>ex portée :</u> si un fœtus est anormal, les autres de la même portée ont une probabilité d'anomalie plus grande.

<u>ex. germination :</u> si une graine d'un lot germe, une substance chimique est émise, qui favorise les germination des autres graines.

De façon générale, la surdispersion est liée à une variable explicative qui n'est pas prise en compte, pas mesurée, ou pas contrôlée de façon adéquate.

Surdispersion, interprétation, hétérogénéité

$\boxed{\text{Variabilité des proba. de réponse}}$

ex. facteur portée

obs : y_i/n_i, $i = 1, \ldots, n$, k var. expl. $X_1 \ldots X_k$

ϑ_i : vraie proba. inconnue de la portée,

varie autour de p_i.

$\hookrightarrow$ V.A.

$$E(\vartheta_i) = p_i$$

$$V(\vartheta_i) = \gamma.p_i.(1 - p_i)$$

γ : param. d'échelle inconnu

<u>facultatif :</u> mais intéressant :

on montre, par le calcul, l'équivalence entre “corrélation des données binaires” et “var. expl. pas prise en compte”, qui se conçoit bien, intuitivement.

Imaginons le cas où les individus sont issus de différentes portées. On observe une variable binaire sur chacun des individus, certains des individus ont en commun l'appartenance à une même portée.

S'il y a variabilité des probabilités de réponse d'une portée à l'autre, ϑ_i est une variable aléatoire, et sa variance est non nulle. Une expression raisonnable de cette variance est $V(\vartheta_i) = \gamma p_i(1 - p_i)$, $V(\vartheta_i)$ est nulle si p_i vaut 0 ou 1, non nulle sinon.

facultatif

Surdispersion, interprétation, hétérogénéité

ϑ_i : V.A non observable : variable latente.

$$y_i|\vartheta_i \sim \mathcal{B}(n_i\vartheta_i, n_i\vartheta_i(1-\vartheta_i))$$

$$E(y_i|\vartheta_i) = n_i\vartheta_i$$

$$V(y_i|\vartheta_i) = n_i\vartheta_i(1-\vartheta_i)$$

Surdispersion, interprétation, hétérogénéité

$$\begin{aligned} E(Y) &= E\left[E(Y|X)\right] \\ V(Y) &= E[V(Y|X)] + V[E(Y|X)] \end{aligned}$$

$$\begin{aligned} E(y_i) &= E[E(y_i|\vartheta_i)] = E(n_i\vartheta_i) = n_iE(\vartheta_i) = n_ip_i \\ V(y_i) &= E[V(y_i|\vartheta_i)] + V[E(y_i|\vartheta_i)] \end{aligned}$$

$$\begin{aligned} E[V(y_i|\vartheta_i)] &= E[n_i\vartheta_i(1-\vartheta_i)] \\ &= n_i(E(\vartheta_i) - E(\vartheta_i^2)) \end{aligned}$$

$$\text{or, } V(\vartheta_i) = E\left[(\vartheta_i)^2\right] - \left[E(\vartheta_i)\right]^2$$

$$\begin{aligned} E[V(y_i|\vartheta_i)] &= n_i(E(\vartheta_i) - V(\vartheta_i) - [E(\vartheta_i)]^2) \\ &= n_i(p_i - \gamma\ p_i\ (1-p_i) - p_i^2) \\ &= n_ip_i(1-p_i)(1-\gamma) \end{aligned}$$

$$V[E(y_i|\vartheta_i)] = V(n_i\vartheta_i) = n_i^2V(\vartheta_i) = n_i^2\gamma\ p_i\ (1-p_i)$$

$$\Rightarrow \boxed{V(y_i) = n_i\ p_i\ (1-p_i)\ \underbrace{(1+(n_i-1)\ \gamma)}_{\phi}}$$

<u>facultatif</u>

On a besoin de la moyenne et de la variance inconditionnelle de y_i. On applique les 2 premières formules, et on arrive au résultat encadré.

- S'il n'y avait pas de variation aléatoire des proba. de réponse, y_i suivrait une loi $\mathcal{B}(n_i, p_i)$,et on aurait $V(y_i) = n_i\ p_i\ (1-p_i)$.

$\phi = 0 \iff$ pas de variation aléatoire des proba. de réponse.

$\phi > 0 \implies$ variation des proba.
$var(y_i) > n_i\ p_i\ (1-p_i)$

$$\boxed{\text{facteur de surdispersion} : 1 + (n_i - 1)\gamma}$$

- Si données non groupées, $n_i = 1$,

  $Var(y_i) = p_i\ (1-p_i) =$ variance d'une réponse binaire.

  Donc,

$$\boxed{\text{les données binaires ne peuvent pas donner d'info sur } \phi}$$

<u>Rq</u> : il faut supposer les n_i à peu près constants pour corriger par un ϕ constant.

Surdispersion, interprétation, corrélation entre réponses binaires

$y_i/n_i \qquad R_{i1},\ R_{i2},\ \ldots,R_{in_i} \quad : V.A.$binaires $\sim$ Bernouilli

$$P(R_{ij}=1)=p_i$$

$$E(R_{ij})=p_i, \quad V(R_{ij})=p_i\,(1-p_i)$$

y_i : valeur observée de la variable $\sum_j R_{ij}$

$$E(y_i)=\sum_j E(R_{ij})=n_i\,p_i$$

$$V(y_i)=\sum_j V(R_{ij})+\sum_j\sum_{k\neq j} Cov(R_{ij},R_{ik})$$

On pose $\quad \delta = \ Corr(R_{ij},R_{ik})$

On a $\quad \delta = \dfrac{Cov(R_{ij},R_{ik})}{\sqrt{V(R_{ij}).V(R_{ik})}}$

$$V(R_{ij})=V(R_{ik})=p_i\,(1-p_i)$$

$$Cov(R_{ij},R_{ik})=\delta\,p_i\,(1-p_i)$$

<u>facultatif</u>

On modélise la corrélation entre réponses binaires.

<u>Rq</u> : δ est supposé constant. Cela peut raisonnablement s'appliquer à des exemples de type “facteur bloc” ou “facteur portée” et moins à des corrélations de type spatiale ou temporelle, où la corrélation dépend de l' “écart” (en temps ou dans l'espace) entre les observations.

Surdispersion, interprétation, corrélation entre réponses binaires

$$
\begin{aligned}
V(y_i) &= \sum_j p_i\,(1-p_i) + \sum_j \sum_{k\neq j} \delta\; p_i\,(1-p_i) \\
&= n_i\; p_i\,(1-p_i) + n_i\,(n_i-1)\;\delta\; p_i\,(1-p_i)
\end{aligned}
$$

$$
\boxed{V(y_i) = n_i\; p_i\,(1-p_i)\;\boxed{(1+(n_i-1)\;\delta)}}
$$

facultatif

$$
\begin{aligned}
&\text{Si } \delta = 0 \quad , \quad V(y_i) = n_i\; p_i\,(1-p_i) \\
&\text{Si } \delta > 0 \quad , \quad V(y_i) > n_i\; p_i\,(1-p_i)
\end{aligned}
$$

Une corrélation positive conduit à une plus grande variation dans le nombre de succès, par rapport à la variation attendue sous l'hypothèse d'indépendance.

Surdispersion, interprétation, hétérogénéité et corrélation

$$\boxed{V(y_i) = n_i\, p_i\, (1 - p_i)\, (1 + (n_i - 1)\gamma)} \qquad \mathbf{1}$$

$$\boxed{V(y_i) = n_i\, p_i\, (1 - p_i)\, (1 + (n_i - 1)\, \delta)} \qquad \mathbf{2}$$

1 hétérogénéité : $\gamma \geq 0 \quad \Rightarrow$ surdispersion

2 corrélation : $-1 \leq \delta \leq 1 \Rightarrow \delta \geq 0$ surdispersion

$\delta \leq 0$ sousdispersion

$$(1 + (n_i - 1)\delta) \geq 0 \quad \Rightarrow \quad -\frac{1}{n_i - 1} \leq \delta \leq 1$$

limite : $-\frac{1}{n_i-1} \leq \delta$

Si n_i gd, **2** ne permet pas de modéliser une sousdispersion.

<u>facultatif :</u> ce transparent peut être présenté, même si les précédents n'ont pas été présentés.

Les 2 équations sont semblables, mais il y a une différence importante.

1 ne modélise que la surdispersion, car $\phi \geq 0$.

2 modélise la surdispersion, et la sousdispersion dans certaines conditions ...

limite du modèle : ... **2** modélise la sousdispersion sauf si n_i est grand, car alors δ est borné par 0 (0 est une borne inférieure). En conséquence, on peut utiliser **2** pour modéliser la sousdispersion, seulement si les n_i sont petits.

Si on modélise une surdispersion (c'est le cas le plus fréquent), les équations **1** et **2** sont identiques. Cela confirme que **l'effet de la corrélation entre réponses binaires, et l'effet de la variation entre probabilités de réponse, ne peuvent pas être distingués**.

La corrélation, si elle est négative, doit être supérieure à $-\frac{1}{n_i-1}$. Conclusion : si les n_i sont grands, ce modèle ne peut pas prendre en compte des corrélations négatives.

Surdispersion, modélisation, facteur d'hétérogénéité

Modéliser la surdispersion

$$V(y_i) = n_i\, p_i\, (1-p_i)\phi$$

$$\phi = \sigma^2$$

Si les $n_i \simeq Ct$,

$$\sigma_i^2 = \sigma^2 = \text{facteur d'hétérogénéité}$$

$$\widehat{\sigma}^2 = \frac{\text{déviance (modèle complet)}}{ddl}$$

inférence :

- $e.t.(\widehat{\beta}) \rightarrow$ multiplié par $\sqrt{\widehat{\sigma}^2}$
- $\frac{\widehat{\beta}}{\widehat{e.t.(\widehat{\beta})}} \sim t(ddl)$
- Modèles emboîtés comparés par test F

$$\frac{D_1 - D_0}{(dl_1 - dl_0)\widehat{\phi}} \sim F(dl_1 - dl_0, m-k)$$

Si possible, ϕ doit être estimé avant la modélisation; à défaut, il doit estimé à l'aide du modèle le plus complet possible. ϕ est souvent noté σ^2.

Si les n_i sont constants ou à peu près constants, σ^2 est une constante appelée facteur d'hétérogénéité. On peut l'estimer par la déviance par unité de degré de liberté, calculée à partir du modèle le plus complet.

Pour faire de l'inférence, il faut :

- corriger les écarts-types des paramètres, en les multipliant par $\sqrt{\widehat{\sigma}^2}$.
- les $\frac{\widehat{\beta}}{\widehat{e.t.(\widehat{\beta})}}$ sont des t-value, et non plus des z-value, car on a estimé le paramètre $\widehat{\sigma}^2$ (analogie avec $Y \sim \mathcal{N}(\mu, \sigma^2)$, $\frac{Y-\widehat{\mu}}{\sigma} \sim \mathcal{N}(0,1)$, $\frac{Y-\widehat{\mu}}{\widehat{\sigma}} \sim t$).
- on utilise un test de Fisher pour comparer 2 modèles emboîtés. La statistique de test a une distribution indépendante de σ^2.

La procédure n'est pas trop sensible à des différences sur les n_i.

<u>test de Fisher : pourquoi cette statistique de test ?</u>
(on trouve l'explication dans la documentation SAS, proc genmod)

$D_1 - D_0 \sim \phi\chi^2(dl_1 - dl_0)$ ou $(D_1 - D_0)/\phi \sim \chi^2(dl_1 - dl_0)$
On estime ϕ par $\widehat{\phi} = Dev/dl$ avec le modèle le plus complet.
On a : $\left((m-k)\widehat{\phi}\right)/\phi \sim \chi^2(m-k)$.

Si les statistiques $\frac{D_1-D_0}{\phi}$ et $\frac{(m-k)\widehat{\phi}}{\phi}$ sont indépendantes,

$$F = \frac{(\frac{D_1-D_0}{\phi})/(dl_1 - dl_0)}{\frac{(m-k)\widehat{\phi}}{\phi}/(m-k)} = \frac{D_1 - D_0}{(dl_1 - dl_0)\widehat{\phi}} \sim F(dl_1 - dl_0, m-k)$$

En cas de n_i assez différents, voir la méthode de William.

Surdispersion, modélisation, méthode de William pour estimer ϕ

principe :

- Modèle complet :

$$\begin{cases} \text{on égalise le } \chi^2 \text{ de Pearson à son espérance,} \\ \text{mais le } \chi^2 \text{ dépend de } \gamma \end{cases}$$

$\hookrightarrow$ méthode itérative $\Rightarrow \widehat{\gamma}$

- Modèles ajustés, avec pondération par $w_i = \frac{1}{\widehat{\sigma_i}^2}$, avec $\widehat{\sigma_i}^2 = 1 + (n_i - 1)\widehat{\gamma}$

Inférence :

- $\frac{\widehat{\beta}}{e.t.(\widehat{\beta})} \sim t(ddl)$
- Modèles emboîtés comparés par test du χ^2

$$(D_1 - D_2) \sim \chi^2(dl_1 - dl_2)$$

facultatif

- La méthode de William est une méthode itérative, plus générale que la méthode de correction par le facteur d'hétérogénéité, elle s'applique dans tous les cas, que les n_i soient constants ou pas.
- On considère le modèle complet. On égalise le χ^2 de Pearson à son espérance, mais le χ^2 dépend de γ. Cela fournit un algorithme itératif d'estimation de γ, à partir du modèle complet. On obtient une estimation de γ.

  Alors la déviance est égale au degré de liberté, et toute la surdispersion a été expliquée par le modèle.

  Puis les modèles considérés sont tous ajustés avec la pondération $w_i = \frac{1}{\widehat{\sigma_i}^2} = \frac{1}{1+(n_i-1)\widehat{\gamma}}$.

  L'effet de la pondération par w_i, est d'augmenter les écarts-types des $\widehat{\beta}$ (par rapport à la non prise en compte de la surdispersion).

  Les méthodes d'inférence sont les mêmes que dans le cas où il n'y a pas de surdisperion. Les paramètres estimés divisés par leur écart-type suivent des lois de student. On compare des modèles emboîtés en calculant la différence des déviances, et en la comparant au quantile d'un χ^2.

On ne présente que deux méthodes de prise en compte de la surdispersion, la méthode de William, et la méthode de correction par le facteur d'hétérogénéité. Il existe d'autres méthodes, comme le modèle béta-binomial, ou les modèles à effets aléatoires, qui dépassent le cadre de ce cours.

facultatif

Surdispersion, des méthodes pour la prendre en compte

- p_i varie, pas de loi précisée

  $\rightarrow$ facteur de surdispersion ($n_i \simeq$ Ct)

  $\rightarrow$ méthode de William

- $p_i \sim beta$

  $\rightarrow$ modèle beta-binomial

- effets $\sim \mathcal{N}$

  $\rightarrow$ modèles à effets aléatoires

I.C. pour une fonction des paramètres

- <u>ex. 1 :</u> **I.C. pour la DL_{50}**

$$\log(\frac{p-\pi_0}{1-p}) = \alpha + \beta\log(\text{conc})$$

$$\log(\text{conc}) = \boxed{\frac{\log(\frac{p-\pi_0}{1-p}) - \alpha}{\beta}}$$

avec $p = 0.5$ et $\pi_0 = 1/4$

- <u>ex. 2 :</u> **I.C. pour une proba. $\widehat{p_0}$**

I.C. pour $\widehat{\eta_0} = \widehat{\alpha} + \widehat{\beta}\log(\text{conc}_0)$,
puis lien réciproque.

- On revient à l'ex. seuil : I.C. autour de la DL_{50} ($= DE_{50}$, dose efficace 50), c'est à dire autour de la concentration nécessaire, pour que 50 % des sujets la perçoivent.

  Il faut calculer un I.C. autour de $\frac{\log(\frac{p-\pi_0}{1-p})-\alpha}{\beta}$.

  Ce n'est pas une combinaison linéaire des paramètres, mais une fonction plus complexe de deux paramètres. La détermination de l'I.C. n'est donc pas immédiate, mais elle est possible moyennant quelques approximations (voir Collett chap. 4). Collett propose deux méthodes, l'une basée sur une approximation de la variance d'une fonction de deux paramètres, l'autre basée sur le théorème de Fieller qui donne la variance du rapport de deux variables aléatoires, lorsque les deux variables aléatoires sont supposées suivre des lois normales.

On peut utiliser la fonction Splus dose.p, présentée dans Venables and Ripley[3] pour calculer la DL_{50} et l'écart-type associé. La fonction dose.p appelle la fonction vcov.

```
dose.p_function(glm, cf = 1:2, p = 0.5, surdisp = F)
{
        eta <- family(glm)$link(p)
        b <- coef(glm)[cf]
        dose <- (eta - b[1])/b[2]
        names(dose) <- paste("p = ", format(p), ":", sep = "")
        pd <- cbind(1, dose)/b[2]
        vr <- ((pd %*% vcov(glm, surdisp = surdisp)[cf, cf]) * pd) %*% c(1, 1)
        attr(dose, "SE") <- sqrt(vr)
        attr(dose, "p") <- p
        class(dose) <- "glm.dose"
        dose
}

vcov_function(obj)
{so <- summary(obj)
 so$dispersion * so$cov.unscaled
}
```

- I.C. autour d'une probabilité $\widehat{p_0}$, calculée pour une dose donnée. On calcule l'I.C. autour de $\widehat{\eta_0}$, puis on applique aux bornes de l'I.C. obtenu, la fonction réciproque de la fonction de lien.

3. Modern Applied Statistics with Splus, Venables and Ripley, Springer, 1996

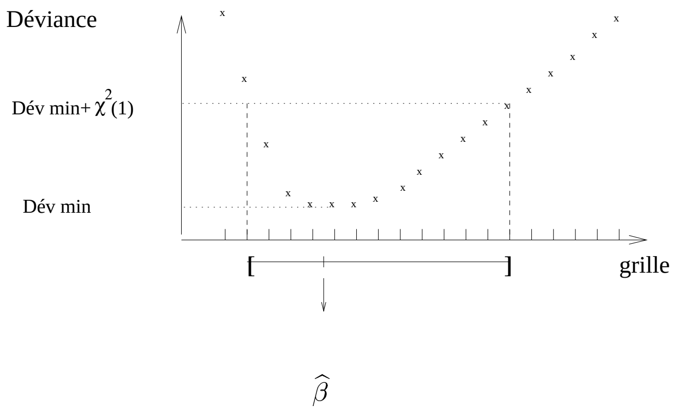
I.C. pour une fonction des paramètres

I.C. pour une fct des param. → approx. normale → I.C. symétrique

I.C. pour une fct des param. → $\boxed{\text{Max. de vrais.}}$ → I.C. dissymétrique

cas particulier :
I.C. pour le paramètre β, par max. de vrais.

$$\text{logit}(p) = \alpha + \underbrace{\beta}_{grille} \log(conc)$$



Nous présenterons 2 méthodes, l'une basée sur une approximation normale, l'autre sur le maximum de vraisemblance. Dans certains cas, le résultat sera peu différent selon la méthode employée.

De façon générale, **la méthode basée sur le maximum de vraisemblance est préférable**, elle fournit des I.C. qui ne sont pas nécessairement symétriques. Le principe de la méthode est basé sur le test de sous-modèle.

On présente la méthode dans le cas plus simple du calcul de l'intervalle de confiance pour un paramètre (β), avant de la présenter dans le cas plus général du calcul de l'intervalle de confiance pour une fonction des paramètres.

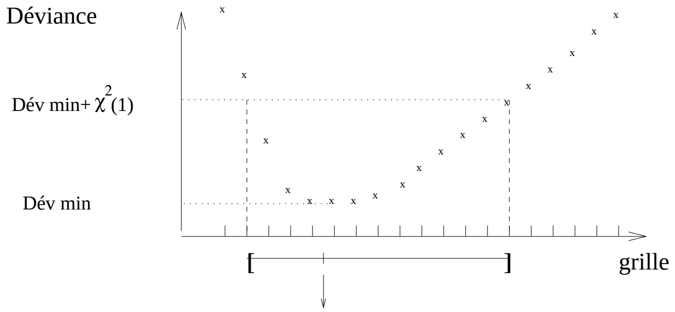
Le principe de la méthode est le suivant :

- on génère une grille de valeurs autour de $\widehat{\beta}$,
- on estime le modèle, c'est à dire le paramètre α, en imposant à β de prendre tour à tour chacune des valeurs de la grille. Le modèle est estimé pour chaque valeur de la grille. On calcule la déviance associée à chaque estimation. On représente graphiquement les déviances en fonction des valeurs de la grille. On trace la droite horizontale à la hauteur : déviance $+\chi^2(1)$ où $\chi^2(1)$ est le $1-\alpha$ quantile d'une loi de χ^2 à 1 ddl , (car on a fait varier un paramètre, en imposant une contrainte pour chaque valeur de la grille), et on lit graphiquement l'I.C pour β. On obtient un I.C. à 95 %, en positionnant la ligne horizontale à la hauteur Déviance $+$ 3.84, un I.C. à 90 % en positionnant la ligne horizontale à la hauteur Déviance $+$ 2.71 ...

I.C. pour une fonction des paramètres

ex. 1 : I. C. pour $\left.\frac{\log(\frac{p-\pi_0}{1-p})-\alpha}{\beta}\right|_{p=0.5}$

$$\beta\left[\log(\text{conc})-\underbrace{\frac{\log(\frac{p-\pi_0}{1-p})-\alpha}{\beta}}_{grille}\right]+\underbrace{\log(\frac{p-\pi_0}{1-p})}_{offset}=\alpha+\beta\ \log(\text{conc})$$



$$\left.\frac{\log(\frac{p-\pi_0}{1-p})-\widehat{\alpha}}{\widehat{\beta}}\right|_{p=0.5}$$

Le prédicteur $\alpha+\beta\log(\text{conc})$ s'écrit de façon équivalente :

$$\beta\left[\log(\text{conc})-\frac{\log(\frac{p-\pi_0}{1-p})-\alpha}{\beta}\right]+\log(\frac{p-\pi_0}{1-p})$$

- on génère une grille de valeurs autour de $\left.\frac{\log(\frac{p-\pi_0}{1-p})-\widehat{\alpha}}{\widehat{\beta}}\right|_{p=0.5}$
- on considère $\log(\frac{p-\pi_0}{1-p})$ comme offset, c'est à dire comme variable à mettre dans le modèle, avec un coeff. égal à 1 (pas de paramètre à estimer).
- on estime le modèle pour chaque valeur de la grille, on calcule la déviance. On représente graphiquement les déviances en fonction des valeurs de la grille. On trace la droite horizontale à la hauteur : déviance $+\ \chi^2(1)$, (car on a fait varier un paramètre), et on lit graphiquement l'I.C.

Ces quelques lignes donneront une idée de la façon dont on peut calculer l'intervalle de confiance par maximum de vraisemblance, avec Splus. Elles doivent être adaptées, notamment s'il faut tenir compte d'une surdispersion. Supposons que IC contienne l'intervalle de confiance obtenu par approximation normale (borne inf, estimation, borne sup), que glm1 soit le résultat d'une estimation.

```
grille_seq(IC[1]-diff(range(IC))/4,IC[3]+diff(range(IC))/4,length=20)
dev_NULL
offset0_rep(log((p-p0)/(1-p)),nrow(don))
for (i in 1:20)
{ldose_log(don$dose)-grille[i]
 dev_c(dev,glm(cbind(k,n-k)~-1+offset(offset0)+ldose,
              family=glm1$call$family,data=don)$deviance) }
plot(grille,dev)
abline(h=glm1$deviance+qchisq(.95,1))
ldose.inf_approx(dev[1:10],grille[1:10],glm1$deviance+sigma2*qchisq(.95,1))$y
ldose.sup_approx(dev[11:20],grille[11:20],glm1$deviance+sigma2*qchisq(.95,1))$y
```

I.C. pour une fonction des paramètres

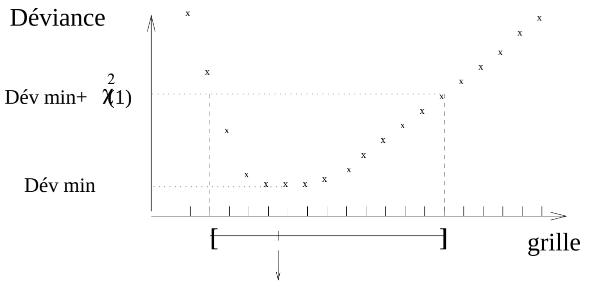
ex. 2 : I. C. pour $\widehat{p_0}$, pour la concentration conc_0

$$\log(\frac{p-\pi_0}{1-p}) = \alpha + \beta \log \text{conc}$$

I.C. pour $\alpha + \beta \log \text{conc}_0$?

$$\alpha + \beta \log \text{conc}_0 + \beta(\log \text{conc} - \log \text{conc}_0) = \alpha + \beta \log \text{conc}$$

$$\underbrace{\alpha + \beta \log \text{conc}_0}_{offset_{\hookrightarrow grille}} + \beta \ \log \frac{\text{conc}}{\text{conc}_0} = \alpha + \beta \log \text{conc}$$



On écrit le modèle équivalent :

- la variable explicative est $\log(\frac{\text{conc}}{\text{conc}_0})$.
- on fait varier $\alpha + \beta \log \text{conc}_0$ sur une grille.

Rq facultative : principe de la méthode

H_0 : " $p = p_0$ ", où p_0 est un nombre réel.

H_1 : "$p \neq p_0$"

$H_0 \rightarrow M_0$ $\qquad$ M_0 : sous-modèle de M_1
$H_1 \rightarrow M_1$

Test de H_0 $\Leftrightarrow$ test de sous-modèle
$\Rightarrow$ statistique du rapport de vraisemblance

La région d'acceptation du test est l'ensemble des valeurs de p telles que la statistique du rapport de vraisemblance est inférieure à un $\chi^2(1)$.

Superposer au graphique le résultat qu'on pourrait obtenir avec un test de Wald (courbe à l'allure parabolique), pour montrer que le test de Wald, basé sur la normalité asymptotique du paramètre conduit à un I.C. symétrique, et insister sur le fait que l'I.C. obtenu par la méthode du maximum de vraisemblance exposée ci-contre est préférable, lorsqu'il est dissymétrique.

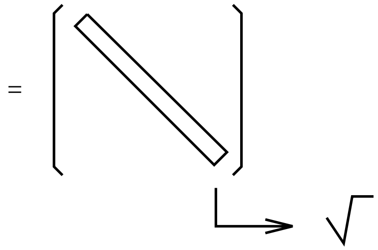
Test de Wald : région d'acceptation : $\{t / \frac{(t-\widehat{\beta}_j)^2}{(se(\widehat{\beta}_j))^2} \leq \chi^2(1)\}$

Remarque : la même méthode permet de calculer des intervalles de confiance simultanés pour deux paramètres : la grille est une grille dans IR^2, on trace l'isocontour correspondant à la déviance $+ \chi^2(2)$, et on en déduit les intervalles de confiance pour les deux paramètres. La méthode peut être intéressante lorsque les deux paramètres sont corrélés.

I.C. pour une comb. lin. des params (par approximation)

$$\widehat{\eta}_0 = X_0\widehat{\theta} = \begin{pmatrix} \cdots \\ \cdots \\ \boxed{\cdots\cdots\cdots} \\ \cdots \end{pmatrix} \begin{pmatrix} \widehat{\theta}_0 \\ \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_p \end{pmatrix}$$

$$Var(X_0\ \widehat{\theta}) = X_0\ Var(\widehat{\theta})\ X_0'$$



$$[\ \widehat{\eta}_0 \pm z_{\alpha/2}\sqrt{Var(\widehat{\eta}_0)}\ \ ]$$
$$logit^{-1}\ [\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots]$$

Supposons qu'on veuille estimer plusieurs combinaisons linéaires des paramètres, et leurs I.C. respectifs.

Une combinaison linéaire est caractérisée par un vecteur de coefficients. On écrit (en ligne) les vecteurs ligne des coefficients dans une matrice X_0.

Le vecteur colonne des combinaisons linéaires estimées est obtenu par le produit matriciel $X_0\ \widehat{\theta}$.

Pour calculer les I.C. autour de ces valeurs, il nous faut une estimation de leur variance. On calcule donc la matrice de variance-covariance de $X_0\ \widehat{\theta}$. On l'obtient par le produit matriciel $X_0\ Var(\widehat{\theta})\ X_0'$. Les variances recherchées sont dans la diagonale de cette matrice. On extrait la diagonale, on en prend la racine carrée. On obtient les écarts-type estimés des combinaisons linéaires.

Les paramètres $\widehat{\theta}$ suivent asymptotiquement une loi normale. Les combinaisons linéaires des paramètres aussi. Ce qui nous permet de calculer un I.C. de la façon suivante :

$$[\widehat{\eta}_0 \pm z_{\alpha/2}\sqrt{Var(\widehat{\eta}_0)}]$$

Il s'agit ici d'un I.C. pour η_0, où η_0 est une combinaison linéaire particulière des paramètres, c'est le prédicteur associé à la dose x_0. Cet I.C. est calculé dans l'échelle du prédicteur linéaire.

Pour obtenir un I.C. dans l'échelle d'origine (dans l'échelle des probabilités), il suffit d'appliquer la fonction réciproque de la fonction de lien aux bornes de l'intervalle, c'est la fonction $logit^{-1}[\ldots] = \frac{exp[\ldots]}{1+exp[\ldots]}$.

Cette méthode de calcul d'un I.C. pour une combinaison linéaire des paramètres est basée sur l'approximation normale des paramètres (les paramètres suivent asymptotiquement une loi normale). De manière générale, la méthode basée sur le maximum de vraisemblance est préférable.

RÉSUMÉ

Surdispersion (Sousdispersion)
obs. bin. corrélées positivement (négativement)

Si $n_i \simeq Ct$,

$$\sigma^2 = \text{facteur d'hétérogénéité}$$
$$\widehat{\sigma}^2 = \frac{\text{Dev(modèle complet)}}{\text{ddl}}$$

$e.t.(\widehat{\beta}) \longrightarrow$ multipliés par $\sqrt{\widehat{\sigma}^2}$

$$\frac{\widehat{\beta}}{\widehat{e.t.(\widehat{\beta})}} \sim t(ddl)$$

modèles emboîtés comparés par tests F.

I.C. pour une fct des param. $\longrightarrow$ approx. normale
I.C. pour une fct des param. $\longrightarrow$ Max. de vrais.

La surdispersion est la conséquence d'une corrélation positive entre les observations binaires. La corrélation entre les observations binaires peut être positive lorsque certaines des observations partagent en commun l'appartenance à un même groupe (famille, région, fermenteur, boîte de pétri ...). Ce qui explique qu'une surdispersion puisse aussi bien être interprétée en terme de corrélation entre les réponses binaires, qu'en terme de facteur de variation négligé. Une sousdispersion (moins fréquente que la surdispersion) provient d'une corrélation négative entre les réponses binaires. Elle peut être interprétée en terme de compétition entre les individus (ex : on peut expliquer une sousdispersion dans la modélisation de la hauteur des arbres d'une forêt par la compétition entre les arbres voisins).

Si les n_i sont à peu près constants, on peut estimer un facteur d'hétérogénéité, à partir du modèle le plus complet, et utiliser ce facteur d'hétérogénéité pour corriger les écarts-type des paramètres estimés. Les statistiques de test (de Wald) sont alors des t-value (et non plus des z-value). Les modèles emboîtés sont comparés par des tests de Fisher.

Lorsqu'on s'intéresse à une fonction des paramètres, on peut en calculer un intervalle de confiance, soit en utilisant une approximation normale, soit par la méthode du maximum de vraisemblance. De façon générale, c'est la méthode du maximum de vraisemblance qui est préférable. L'approximation normale donnera des résultats convenables, et proches de ceux du maximum de vraisemblance, si l'intervalle est symétrique. Sinon, il faut utiliser la méthode du maximum de vraisemblance, un peu moins facile à mettre en œuvre, un peu plus longue en temps calcul, mais plus générale, car on peut l'utiliser quelle que soit la fonction des paramètres, qu'il s'agisse d'une combinaison linéaire ou pas.