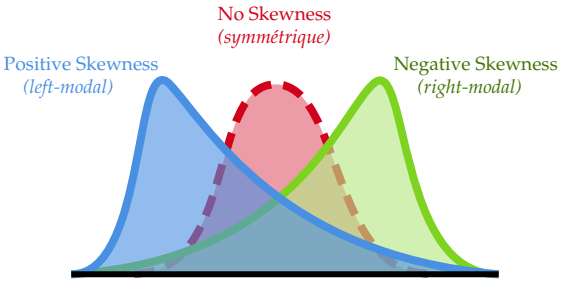
3 Estimation non-paramétrique

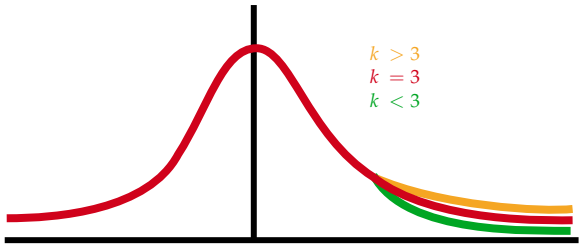
Moments à savoir

$$\mu'_k = \mathrm{E}\left[X^k\right] \qquad \mu_k = \mathrm{E}\left[(X-\mu)^k\right]$$

$$CV = \frac{\sigma}{\mu}$$

$$\text{Skewness} : \gamma = \frac{\mu_3}{\sigma^3} \qquad \text{Kurtosis} : \kappa = \frac{\mu_4}{\sigma^4}$$





3 critères pour évaluer les queues de distributions

1. La loi avec le moins de moments a la queue la plus lourde.
2. Première à diverger du quotient des distributions a la queue la plus lourde.
$$\lim_{x\to\infty} \frac{f_{X_1}(x)}{f_{X_2}(x)}$$
3. Si la fonction hasard $h(x) = \frac{f_X(x)}{S_X(x)}$ est croissante alors la queue est fine, sinon elle est lourde.
$h'_X(x) < 0$ queue lourde
$h'_X(x) > 0$ queue fine

Quantités des distributions à connaître

Y^P : **Excess loss**, alias **left truncated** and **shifted** variable.
On interprete comme le *montant de perte* en *excès d'un déductible d* sachant que la perte est au delà de ce montant.

Y^L : **Left censored** and **shifted** variable.
Elle est défini comme étant 0 pour toutes les pertes inférieures à d, alors que l'excès-moyen n'est simplement pas défini dans ces cas.
Donc, celle-ci a une masse à 0.

Y : **Limited loss**, alias **right censored** variable.

shifted : d est soustrait des valeurs restantes.
On peut visualiser le déplacement de la courbe de densité à la gauche.

left truncated : Toutes valeurs inférieures à d ne sont pas observées.

left censored : Toutes valeurs inférieures à d sont égale à 0.

right censored : Toutes valeurs supérieures à u sont égale à u.

Pour exemple, lorsqu'il y a une limite sur une police d'assurance les valeurs au-delà ne sont pas typiquement inscrites à leur vrai montant, mais plutôt comme la limite u.

Moments

$$\mathrm{E}\left[Y^P\right] = \mathrm{E}\left[X-d|X\geq d\right] = \frac{\int_d^\infty S_X(x)}{S_X(d)} = e_X(d)$$

$$\mathrm{E}\left[Y^L\right] = \mathrm{E}\left[(X-d)_+\right] = \int_d^\infty (x-d) f_X(x)dx$$

$$\mathrm{E}\left[Y\right] = E[X\wedge d] = \int_0^d f_X(x)dx \Leftrightarrow \int_0^d S_X(x)dx$$

8 Fréquence et sévérité avec modifications aux contrats

Déductible ordinaire

L'assureur paye tout montant en excédent du montant d.

$$Y_{(O)}^{(L|P)} = (X-d)_+ = \begin{cases} (0|\text{non-défini}) & , X\leq d \\ X-d & , X>d \end{cases}$$

Déductible franchise

L'assureur paye l'entièreté des coûts pour toute perte qui surpasse le montant d.
Pour éviter les petites réclamations

$$Y_{(F)}^{(L|P)} = (X-d)_+ = \begin{cases} (0|\text{non-défini}) & , X\leq d \\ X & , X>d \end{cases}$$

Moments

$$E[Y_{(O|F)}^{(L|P)}] = \frac{E[X] - E[X\wedge d] + dS_X(d)}{S_X(d)}$$

De plus, on note que :

$$E[Y_{(O)}^{(P)}] = e_X(d)$$

$$E[Y_{(O)}^{(L)}] = \pi_X(d)$$

Fonctions

$$f_{Y_{(O)}^{(L|P)}} = \frac{f_X(y+d)}{S_X(d)} \qquad h_{Y_{(O)}^{(L|P)}} = h_X(y+d)$$

$$S_{Y_{(O)}^{(L|P)}} = \frac{S_X(y+d)}{S_X(d)} \qquad F_{Y_{(O)}^{(L|P)}} = \frac{F_X(y+d) - F_X(d)}{S_X(d)}$$

LER et inflation du déductible ordinaire

Le LER nous donne le pourcentage de perte qu'on ne paie pas grâce au déductible

$$LER = \frac{\mathrm{E}\left[X\right] - \mathrm{E}\left[(X-u)_+\right]}{\mathrm{E}\left[X\right]} = \frac{\mathrm{E}\left[X\wedge u\right]}{\mathrm{E}\left[X\right]}$$

Soit $X^I = (1+r)X$

$$E[X^I\wedge u] = (1+r)E[X\wedge \frac{u}{1+r}]$$

$$f_{X^I}(x) = \frac{f_X\left(\frac{y}{1+r}\right)}{1+r}$$

$$F_{X^I}(x) = F_X\left(\frac{y}{1+r}\right)$$

Limite de police

L'assureur paye un maximum de u

$$Y = (X \wedge u) = \begin{cases} X, & X < u \\ u, & X \geq u \end{cases}$$

$$f_Y(y) = \begin{cases} f_X(y), & y < u \\ S_X(u), & y = u \end{cases}$$

$$F_Y(y) = \begin{cases} F_X(y), & y \leq u \\ 1, & y > u \end{cases}$$

Coassurance

L'assureur paye une fraction, α, de la perte.

Si la coassurance est la seule modification, alors nous obtenons $Y = \alpha X$.
L'impact sur les fonctions est le même qu'avec de l'inflation.

Formule récapitulative

Lorsque les 4 items sont présent (déductible *ordinaire*, limite, inflation et coassurance.

$$Y_{(O)}^{(L|P)} = \begin{cases} (0|\text{Non-défini}) & , x < \frac{d}{1+r} \\ \alpha\Big((1+r)x - d\Big) & , \frac{d}{1+r} \leq x < \frac{u}{1+r} \\ \alpha(u-d) & , x \geq \frac{u}{1+r} \end{cases}$$

$$\mathrm{E}\left[Y_{(O)}^{(L|P)}\right] = \frac{\alpha(1+r)\left(\mathrm{E}\left[X \wedge \frac{u}{1+r}\right] - \mathrm{E}\left[X \wedge \frac{d}{1+r}\right]\right)}{S_X\left(\frac{d}{1+r}\right)}$$

14 Estimation non-paramétrique des fonctions de répartition et de survie

Distribution empirique avec données complètes

I.C. au niveau $1-\alpha$ de $F(x) \in \left[F_n(x) \pm z_{1-\frac{\alpha}{2}}\sqrt{\widehat{V}(F_n(x))}\right]$

y_j : la j-ème des k valeurs unique de l'échantillon de n ($k \leq n$).
$y_1 < y_2 < \cdots < y_k$

s_j : Nombre de fois que l'observation y_j est observé dans l'échantillon.
$\sum_{j=1}^{k} s_j = n$

r_j : Nombre d'observations $\geq y_j$.
$\sum_{i=j}^{k} s_i = r_j$

$$F_n(x) = \frac{1}{n}\sum_{j=1}^{n} I_{\{x_j \leq x\}}$$

$$f_n(x) = \frac{1}{n}\sum_{j=1}^{n} I_{\{x_j = x\}}$$

$$nF_n(x) \sim \text{bin}(n, F(x))$$

$$E[F_n(x)] = F_n(x)$$

$$\widehat{Var}[F_n(x)] = \frac{F_n(x)(1-F_n(x))}{n}$$

$$F_n(x) = \begin{cases} 0, & x < y_1 \\ 1 - \frac{r_j}{n}, & y_{j-1} \leq x < y_j \\ 1, & x > y_k \end{cases}$$

$\forall j = 2, ..., k$

Distribution empirique avec données groupées

Fonction OGIVE

- Dans certains contextes, on a n données qui sont groupées en intervalles et la fonction OGIVE permet d'interpoler entre 2 points c_{j-1} et c_j.
- On défini n_j comme étant le nombre d'observations entre c_{j-1} et c_j.
- Soit x tel que

$$c_{j-1} \leq x \leq c_j$$

$$F_n(c_{j-1}) \leq F_n(x) \leq F_n(c_j)$$

Alors

$$F_n^{\text{OGIVE}}(x) = \alpha F_n(c_{j-1}) + (1-\alpha)F_n(c_j)$$

$$= \frac{c_j - x}{c_j - c_{j-1}}F_n(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}}F_n(c_j)$$

où $F_n(c_j) = \frac{\sum_{i=1}^{j} n_i}{n}$

$$f_n(x) = \frac{F_n(c_j) - F_n(c_{j-1})}{c_j - c_{j-1}} \Leftrightarrow \frac{n_j}{n(c_j - c_{j-1})}$$

Estimations empirique avec données censurées à droite

On représente les données censurées avec :

b_i : Nombre d'observations censurées à la droite dans l'intervalle $[y_i, y_{i+1})$ $\forall i = 1, 2, \ldots, k-1$

De plus, on interprète les valeurs définies plus haut.

s_i : Nombre de décès au temps i.

r_i : Le nombre ***à risque*** à l'observation y_i.

$$r_i = \begin{cases} n, & i = 1 \\ r_{i-1} - s_{i-1} - b_{i-1}, & i = 2, 3, \ldots, k+1 \end{cases}$$

On peut interpréter la fonction de survie comme une **probabilité conditionnelle**.

$$S(t) = \frac{S(t_1)}{S(t_0)} \times \frac{S(t_2)}{S(t_1)} \times \cdots \times \frac{S(t)}{S(t-1)}$$

$$\Leftrightarrow p_1 \times p_2 \times \cdots \times p_t = \prod_{j \leq t} p_j$$

où $p_t = P(T > t | T > t-1)$

On peut donc estimer p_j par :

$$\hat{p}_j = 1 - \frac{S_j}{r_j}$$

où $S_j \sim \text{Bin}(r_j, q_j)$
Ceci correspond donc à l'estimateur de **Kaplan-Meier** :

Estimateur Kaplan-Meier de la fonction de survie empirique

$$S_m(t) = \prod_{j \leq t}\left(1 - \frac{S_j}{r_j}\right)$$

La **Formule de Greenwood** estime la variance de la fonction de survie **Kaplan-Meier** :

Formule de Greenwood

$$\widehat{V}\big(S_m(t)\big) = \big(S_m(t)\big)^2 \sum_{j \leq t} \frac{S_j}{r_j(r_j - S_j)}$$

La ***cumulative hazard rate function*** est estimée par l'estimateur **Nelson-Åalen** :

Estimateur Nelson-Åalen du cumulative hazard rate function

$$\widehat{H}(t) = \sum_{j \le t} \frac{S_j}{r_j}$$

L'estimateur de la fonction de survie peut ensuite être déduit :

$$\widehat{S}(t) = e^{-\widehat{H}(t)}$$

La variance de l'estimateur Nelson-Åalen est estimée par la **formule de Klein** :

Formule de Klein

$$\widehat{V}\left(\widehat{H}(t)\right) = \sum_{j \le t} \frac{S_j(r_j - S_j)}{r_j^3}$$

Intervalles de confiance

$$H(t) \in \left[\widehat{H}(t) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}(\widehat{H}(t))}\right]$$

$$S(t) \in \left[S_m(t) \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{V}}(S_m(t))}\right]$$

Méthode d'efron : Poser que $S_m(y) = 0 \; \forall y \ge y_{\max}$.

Il est bon de noter que le résulat pour la variance de XXX à été obtenu avec la **méthode de Delta** qui consiste aux deux premiers termes de l'expansion Taylor :

méthode de Delta

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta)$$

6 Création de nouvelles lois

Multiplication par une constante

Ceci équivaut à appliquer l'inflation uniformément pour tous les niveaux de pertes et est donc un **changement d'échelle**.
Soit $Y = \theta X$, alors :

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(\theta X \le y) \\ &= \Pr\left(X \le \frac{y}{\theta}\right) = F_X\left(\frac{y}{\theta}\right) \\ \Rightarrow f_Y(y) &= \frac{1}{\theta} f_X\left(\frac{y}{\theta}\right) \end{aligned}$$

On déduit donc que le paramètre $\theta > 0$ est un *paramètre d'échelle* pour la variable aléatoire Y.

Élévation à une puissance

Soit la variable aléatoire de pertes X et $Y = X^{\frac{1}{\tau}}$, alors :

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(X^{\frac{1}{\tau}} \le y) \\ &= \Pr\left(X \le y^{\tau}\right) = F_X\left(y^{\tau}\right) \\ \Rightarrow f_Y(y) &= \text{-}\tau y^{\tau-1} f_X\left(y^{\tau}\right) \end{aligned}$$

Et on note la terminologie suivante :

$\tau > 0$	Y est une ***transformation*** de X
$\tau = -1$	Y est ***l'inverse*** de X
$\tau < 0$	Y est une ***transformation inverse*** de X

Exponentielle

Soit la variable aléatoire de pertes X et $Y = \exp\{X\}$, alors :

$$\begin{aligned} F_Y(y) &= \Pr(Y \le y) = \Pr(\mathrm{e}^X \le y) \\ &= \Pr\left(X \le \ln y\right) = F_X\left(\ln y\right) \\ \Rightarrow f_Y(y) &= \frac{1}{y} f_X\left(\ln y\right), y > 0 \end{aligned}$$

On dit donc que Y est la *log-loi* où $\ln Y = X$ est la loi.

Mélange

9 Fonction génératrice cumulante

Soit la fonction génératrice des moments $M_X(t)$, telle que

$$M_X(t) = \mathrm{E}\left[e^{tX}\right]$$

Alors, la fonction génératrice cumulante $K_X(t)$ est définie comme

$$K(t) = \ln M_X(t)$$

De plus, la fonction génératrice cumulante a les propriétés suivantes :

$$K'(t)\Big|_{t=0} = \mathrm{E}[X]$$

$$K''(t)\Big|_{t=0} = \mathrm{Var}(X)$$

10 Frequentist estimation

Méthode des moments

Soit un échantillon aléatoire de taille n (iid), on pose $\hat{\mu}'_k = \mu'_k$.

Estimateur par la méthode des moments

Un estimateur par la méthode des moments de θ est alors toute solution des p équations

$$\mu'_k(\theta) = \hat{\mu}'_k, \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que la distribution empirique aura les même p premiers moments centrés à 0 que la distribution paramétrique.
Lorsque les données sont **censurées**, on pose $\hat{\mu}'_k = \mathrm{E}[\min(X; u)^k]$.
Lorsque les données sont **tronquées**, on pose $\hat{\mu}'_k = \mathrm{E}[X^k | X > d]$.

Méthode des percentiles

Soit un échantillon aléatoire de taille n (iid), on pose $\hat{\pi}_g(\theta) = \pi_g(\theta)$.

Estimation de θ par la méthode du «Percentile Matching »

L'estimation de θ est alors toute solution des p équations :

$$F(\hat{\pi}_{g_k}|\theta) = g_k, \quad k = 1, 2, \ldots, p$$

La raison pour cet estimateur est que le modèle produit aura p pourcentiles qui vont «matcher »les données.

Lorsque les données sont **censurées**, on doit s'assurer que les pourcentiles sont en dedans de la portion des données non-censurées.

Lorsque les données sont **tronquées**, on doit «matcher »les pourcentiles aux pourcentiles de la distribution conditionnelle.

Il peut arriver que les pourcentiles de distributions ne soient pas unique, par exemple dans le cas de données discrètes lorsque le quantile recherché peut tomber entre 2 *marches* de la fonction empirique, ou mal-définis. Il est alors utile de définir une méthode d'interpolation des quantiles (bien qu'il n'en n'existe pas une officielle définitive).

Soit le «**smoothed empirical estimate** »d'un pourcentile :

«smoothed empirical estimate »

On utilise les statistiques d'ordre de l'échantillon $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ pour l'interpolation suivant :

$$\hat{\pi}_\kappa = (1-h)x_{(j)} + hx_{(j+1)}, \quad \text{où}$$
$$j = \lfloor (n+1)\kappa \rfloor \quad \text{et} \quad h = (n+1)\kappa - j$$

Méthode du maximum de vraisemblance (*MLE*)

Nous cherchons à maximiser la probabilité d'observer les données. Ceci est fait par la vraisemblance $\mathcal{L}(\theta; x)$ ou, puisque le logarithme ne change pas le maximum, la log-vraisemblance $\ell(\theta; x)$ où :

Maximum de vraisemblance

$$\mathcal{L}(\theta; x) = \prod_{i=1}^{n} f(x_i; \theta) \quad \text{et} \quad \ell(\theta; x) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

et l'**estimateur du maximum de vraisemblance** de $\boldsymbol{\theta}$ est celui qui maximise la fonction de vraisemblance.

Cependant, il peut être pratique de généraliser cette fonction de vraisemblance pour les cas de données censurées ou tronquées.

Soit un ensemble de données comportant n événements $A_1, \ldots, A_n$ avec A_j étant tout ce qui fut observé pour la j^e observation ; c'est-à-dire que A_j pourrait être une observation unique ou un intervalle (par exemple, dans le cas de données groupées).

De plus, un suppose que A_j est une observation de la variable aléatoire X_j et que les variables aléatoires $X_1, \ldots, X_n$ ne doivent pas obligatoirement avoir la même distribution paramétrique ; cependant, elles doivent tous dépendent du même vecteur paramétrique $\boldsymbol{\theta}$. Finalement, comme dans les deux autres cas, les variables aléatoires sont supposées indépendantes.

$$\mathcal{L}(\theta; x) = \prod_{j=1}^{n} \Pr(X_j \in A_j; \theta)$$

Pour faire le lien avec la définition précédente, dans le cas où A_j est un point unique et que la distribution est continue $\Pr(X_j \in A_j|\theta) = f(x_i; \theta)$.

Données modifiées

Pour le cas de données groupées, les observations $c_0 < c_1 < \cdots < c_k$ contiennent n_j observations par intervalle $(c_{j-1}, c_j]$.
Dans le cas des données censurées, on multiplie les fonctions de densité pour les données non-censurées et on remplace la densité pour la fonction de survie pour les données censurées ; soit n_{NC} le nombre de données non-censurées et n_C le nombre censurées tel que $n = n_{NC} + n_C$.
La fonction de vraisemblance est donc :

$$\mathcal{L}(\theta; x) = \prod_{j=1}^{n} \Big[F(c_j|\theta) - F(c_{j-1}|\theta)\Big]^{n_j}, \text{ groupé}$$
$$= \left(\prod_{j=1}^{n_{NC}} f(x_i; \theta)\right) \left(\prod_{j=n_{NC}+1}^{n} S(x_i; \theta)\right), \text{ censuré}$$
$$= \prod_{j=1}^{n} \frac{f(x_i; \theta)}{S(d; \theta)}, \text{ tronqué}$$

Variance des estimateurs et intervalle de confiance

Soit, sous certaines conditions de régularité, l'**information de Fisher** $I(\theta)$:

Information de Fisher

$$I(\theta) = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\ell(\theta)\right)^2\right]$$
$$\Leftrightarrow -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]$$

La simplification (2ème équation) est pour le cas où les observations sont (iid) et ont donc tous la même fonction de log-vraisemblance.

Si l'information n'est pas connue, on peut l'estimer avec **l'information observée** :

Information de Fisher observée

$$\hat{I}(\hat{\theta}) = \sum_{i=1}^{n} \left(\frac{\partial}{\partial\theta} \ln f(x_i;\theta)\Big|_{\theta=\hat{\theta}} \right)^2 = -\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \ln f(x_i;\theta)\Big|_{\theta=\hat{\theta}}$$

Estimation de la variance de $\hat{\theta}$

Ainsi, on peut calculer la variance de l'estimateur $\hat{\theta}_{MLE}$ tel que

$$\text{Var}(\hat{\theta}) = I(\theta)^{-1}$$

Intervalle de confiance pour $\hat{\theta}$

Lorsque $n \to \infty$, $\hat{\theta} \sim N(\theta, \text{Var}(\hat{\theta}))$ on peut trouver un IC pour l'estimateur au seuil de $1-\alpha$:

$$\theta \in \left[\hat{\theta} \pm z_{\alpha/2}\sqrt{\text{Var}(\hat{\theta})}\right]$$

Méthode delta pour estimer la variance d'une transformation de $\hat{\theta}$

Lorsqu'on veut calculer la variance d'une autre quantité que le paramètre $\hat{\theta}$ lui-même, on peut utiliser la méthode Delta :

$$\text{Var}\left(h(\hat{\theta})\right) = \left(\frac{\partial}{\partial\theta}h(\theta)\right)^2 \text{Var}\left(\hat{\theta}\right)$$

Dans un contexte multivarié, où $\hat{\boldsymbol{\theta}}$ est un vecteur d'estimateurs, alors on a

$$\text{Var}\big(h(\hat{\theta})\big) = \boldsymbol{h}^\top I(\theta)^{-1}\boldsymbol{h}$$

où $\boldsymbol{h}$ est le vecteur des dérivées partielle de $h(\theta)$:

$$\boldsymbol{h} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}h(\theta) \\ \frac{\partial}{\partial\theta_2}h(\theta) \\ \dots \\ \frac{\partial}{\partial\theta_k}h(\theta) \end{bmatrix}$$

15 Sélection de modèles

Il est essentiel de ne pas oublier que peu importe le modèle sélectionné, il sera une **approximation** de la réalité ; «all models are wrong, but some are more useful than others ».

Chi-Square *Goodness-of-fit*

On veut valider que les prévisions du modèle ne sont pas trop différentes des valeurs empiriques observées. Alias, that the predictions of the model has a **good fit** to the empirical data with similar percentiles.
On valide alors l'adéquation d'un modèle avec p variables explicatives et n observations en calculant la quantité X^2 :

Test d'adéquation du Khi-Carré

$$X^2 = \frac{\sum_{j=1}^{k} n(\hat{p}_j - p_{nj})^2}{\hat{p}_j} \sim \chi^2_{k-p-1}$$

où

- $\hat{p}_j$: Probabilité (selon le modèle) d'observer une valeur dans la j^e classe.
- p_{nj} : Proportion empirique des observations dans la i^e classe.

On peut récrire ce test sous la forme suivante :

Test d'adéquation du Khi-Carré (autre forme)

On pose

- n_j est le nombre de données dans le groupe $(c_{j-1}, c_j]$,
- $j \in \{1, \dots, k\}$,
- k est le nombre de groupes et
- $\sum_{j=1}^{k} n_j = n$.

La statistique de test est donc :

$$Q = \frac{\sum_{j=1}^{k}(E_j - n_j)^2}{E_j}$$

où

- Asymptotiquement, $Q \sim \chi^2_{k-p-1}$ et
- p est le nombre de paramètres estimés,
- $E_j = n[F(c_j;\hat{\boldsymbol{\theta}}) - F(c_{j-1};\hat{\boldsymbol{\theta}})]$.

On rejète H_0 si $Q > \chi^2_{k-p-1,\alpha}$.

On peut également effectuer le test LRT pour valider l'adéquation du modèle.

Test du rapport de vraisemblance (*LRT*)

On utilise la statistique du khi-carré pour comparer deux modèles ; la question est si le modèle sous H_1 est une représentation de la population mieux ajustée que le modèle sous H_0. On note que le modèle sous H_1 doit être une simplification de celui sous H_0.
En bref, on test si le modèle réduit avec θ_0 est une *bonne* simplification du modèle complet en testant si la différence des log-vraisemblances est significative :

Test du rapport de vraisemblance (LRT)

$$T = 2\left(\ell(\theta) - \ell(\theta_0)\right) \sim \chi^2_{dl_1 - dl_0, 1-\alpha}$$

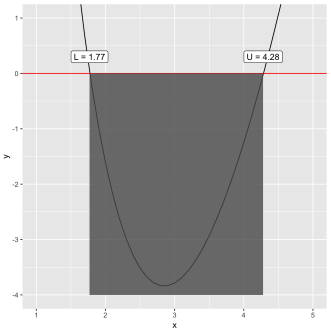
où dl_1 est le nombre de paramètres non-fixés du modèle complet et dl_0 le nombre de paramètres non-fixés du modèle réduit.
On va rejeter H_0 si $T > \chi^2_{dl_1-dl_0,1-\alpha}$ (test unilatéral) et conclure que le modèle réduit n'est pas une bonne simplification du modèle complet.

Construction d'un intervalle de confiance par inversion du *LRT*

Si θ_0 est un paramètre adéquat pour le modèle réduit, alors la statistique T du *LRT* ne dépassera pas le quantile théorique $\chi_{dl,1-\alpha^2}$. Alors, on veut trouver $\hat{\theta}_0$ tel que

$$2\left(\ell(\theta) - \ell(\theta_0)\right) \leq \chi^2_{dl_1-dl_0,1-\alpha}$$

On trouvera une équation du genre $g(\theta) \leq 0$, où g sera une fonction avec deux racines définies, qui correspondent aux bornes de l'intervalle de confiance pour les valeurs de $\hat{\theta}_0$:



Méthode de Delta multivariée

Méthode de Delta multivariée

Soit la variable aléatoire multivariée de **dimension** k et de **taille d'échantillon** n $\boldsymbol{X_n} = [X_{1n}, \ldots, X_{kn}]^\top$.
On pose :

1. $\boldsymbol{X_n}$ est asymptotiquement normale *(multivariée)* de
2. moyenne $\boldsymbol{\theta}$ et
3. matrice de covariance $\frac{\boldsymbol{\Sigma}}{n}$,

où les paramètres $\boldsymbol{\theta}$ et $\boldsymbol{\Sigma}$ ne dépendent pas de la taille d'échantillon n.
On note que puisque $\boldsymbol{X_n}$ est asymptotiquement normale, ils deviennent les paramètres de la normale multivariée.

Soit la fonction $\boldsymbol{G_n} = g(X_{1n}, \ldots, X_{kn})$.
On pose :

1. $\boldsymbol{G_n}$ est également asymptotiquement normale ***(pas** multivariée)* mais de
2. moyenne $g(\boldsymbol{\theta})$ et
3. ***variance*** $\frac{\boldsymbol{A}^\top \boldsymbol{\Sigma} \boldsymbol{A}}{n}$,

où $\boldsymbol{A}$ est le gradient (c.-à-d., le vecteur des k premières dérivées) de $g(\boldsymbol{\theta})$ évalué aux vrais paramètres $\boldsymbol{\theta}$ de $\boldsymbol{X_n}$; $\boldsymbol{A} = \left[\frac{\partial}{\partial\theta_1} g, \ldots, \frac{\partial}{\partial\theta_k} g\right]^\top \Big|_{\boldsymbol{\theta}}$.

Alors, on obtient :

Variance estimée méthode de Delta multivariée

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = \frac{\hat{\boldsymbol{A}}^\top \hat{\boldsymbol{\Sigma}} \hat{\boldsymbol{A}}}{n}$$

Exemple avec une lognormale

Puisque ce concept est très théorique et difficile à conceptualiser, nous jugeons pertinent d'inclure un exemple.

On veut l'estimateur de la moyenne d'une distribution log-normale où :

$$\mathrm{E}[X] = \mathrm{e}^{\mu+\sigma^2/2} = g(\mu, \sigma) = g(\boldsymbol{\theta})$$

On peut trouver (*pas démontré ici*) que les EMV sont :

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i \qquad \hat{\sigma} = \frac{1}{n}\sum_{i=1}^{n} (\ln x_i - \hat{\mu})^2$$

Donc, l'EMV de $g(\mu, \sigma)$ est :

$$g(\hat{\mu}, \hat{\sigma}) = \mathrm{e}^{\hat{\mu}+\hat{\sigma}^2/2} = g(\hat{\boldsymbol{\theta}})$$

De plus, on peut trouver *(encore, pas montré ici)* que la matrice de covariance est :

$$\frac{\boldsymbol{\Sigma}}{n} = \frac{1}{\boldsymbol{I}(\mu, \sigma)} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

où $\boldsymbol{I}$ est l'information de Fisher.
De plus :

$$\boldsymbol{A} = \begin{bmatrix} \frac{\partial}{\partial\mu} g(\mu, \sigma) \\ \frac{\partial}{\partial\sigma} g(\mu, \sigma) \end{bmatrix} \Leftrightarrow \begin{bmatrix} \frac{\partial}{\partial\theta_1} g(\boldsymbol{\theta}) \\ \frac{\partial}{\partial\theta_2} g(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \mathrm{e}^{\mu+\sigma^2/2} \\ \sigma\mathrm{e}^{\mu+\sigma^2/2} \end{bmatrix}$$

Finalement, on obtient :

$$\mathrm{E}[g(\hat{\mu}, \hat{\sigma})] = \mathrm{E}[g(\hat{\boldsymbol{\theta}})] \underset{n\to\infty}{=} \mathrm{e}^{\mu+\sigma^2/2}$$

$$\begin{aligned} \mathrm{Var}[g(\hat{\mu}, \hat{\sigma})] &= \mathrm{Var}[g(\hat{\boldsymbol{\theta}})] \\ &\underset{n\to\infty}{=} \hat{\boldsymbol{A}}^\top \frac{\hat{\boldsymbol{\Sigma}}}{n} \hat{\boldsymbol{A}} \\ &= \left[\begin{bmatrix} \mathrm{e}^{\hat{\mu}+\hat{\sigma}^2/2} \\ \sigma\mathrm{e}^{\hat{\mu}+\hat{\sigma}^2/2} \end{bmatrix}^\top \begin{bmatrix} \frac{\hat{\sigma}^2}{n} & 0 \\ 0 & \frac{2\hat{\sigma}^4}{n} \end{bmatrix} \begin{bmatrix} \mathrm{e}^{\hat{\mu}+\hat{\sigma}^2/2} \\ \sigma\mathrm{e}^{\hat{\mu}+\hat{\sigma}^2/2} \end{bmatrix} \right]_{\boldsymbol{\theta}} \\ &= \mathrm{e}^{2\mu+\sigma^2} \left(\frac{\sigma^2}{n} + \frac{\sigma^4}{2n} \right) \end{aligned}$$

Critères de sélection

Le livre introduit 2 concepts :

Parsimony : À moins d'évidence considérable au contraire, le modèle le plus simple est toujours préféré.

Si on essaye assez de modèles, éventuellement un va sembler être bien ajusté même s'il ne l'est pas. Il est donc essentiel de restreindre le nombre de modèles possibles.

De plus, le livre affirme qu'il y a deux méthodes de sélection de modèle ; par jugement et par valeur de test.
Pour choisir entre plusieurs modèles selon une valeur, on peut, entre autres, se baser sur les critères suivants :

1. la plus **faible valeur** pour le test **Kolmogorov-Smirnov** ;
2. la plus **faible valeur** pour le test **Anderson-Darling** ;
3. la plus **faible valeur** pour le test **Goodness-of-fit** ;
4. la plus **haute valeur** pour la ***p-value*** **du test Goodness-of-fit** ;
5. la plus **haute valeur** pour la **fonction de vraisemblance à son maximum**.

Le problème avec tous ces test (sauf la p-value du khi-carré) est qu'ils ne respectent pas le principe de *parsimony*. Le modèle le plus complexe sera toujours préférable puisqu'il aura un meilleur ajustement pour les données. La distinction avec la p-value du test du khi-carré est que le nombre de paramètres est pris en compte dans l'obtention du seuil.
Ceci est d'ailleurs d'où provient les tests d'AIC et de BIC. L'AIC pénalise les différents modèles selon le nombre de paramètres, le BIC les pénalisent proportionnellement au nombre d'observation.

Critère d'information d'Akaike (AIC) et critère Bayésien de Schwarz (BIC)

On obtient alors que pour n observation et un modèle avec p paramètres :

AIC : Ce critère est le plus utilisé dans la pratique et permet d'évaluer la qualité de l'ajustement d'un modèle.

$$AIC = -\ell + 2p$$

L'AIC prend en compte à la fois la qualité des prédictions du modèle et sa complexité.

BIC : BIC est similaire à l'AIC, mais la pénalité des paramètres dépend de la taille de l'échantillon.

$$BIC = \ell - \frac{p}{2}\ln n$$

› On cherche donc à minimiser ces 2 critères.

13 Estimation bayésienne

Avec l'estimation fréquentielle, on pose une distribution pour les données **fixée *mais* inconnue**. De plus, nos décisions sont axées plus avec les possibilités liés à d'autres échantillons pouvant être obtenus qu'avec notre échantillon de données.
L'approche Bayésienne quant à elle pose que seul les données réellement observées sont pertinentes et que c'est la distribution de la population qui est variable. L'approche est décrite par les définitions des distributions a priori et à postériori avec le théorème de Bayes pour trouver la solution.

Distribution *a priori*

Dénoté $\pi(\theta)$, représente nos opinions des chances que différentes valeurs de θ sont la vraie valeur du paramètre. Par conséquent, elle est définie sur l'espace des valeurs possible pour le paramètre θ.

La difficulté repose justement à déterminer une distribution a priori convaincante ; pour traiter la possibilité d'avoir aucune ce qu'est la distribution a priori, on peut élargir la définition :

Distribution *a priori* impropre

Une distribution pour laquelle les probabilités (ou fonction de densité) sont non-négatives et dont leur somme (ou intégral) est infini.

Distribution du modèle

Dénoté $f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$, elle est la distribution de probabilité des données tel que collecté étant donné une valeur précise de θ

Rappel distribution conjointe et marginale

Fonction conjointe :
$$f_{\boldsymbol{X},\Theta}(\boldsymbol{x},\theta) = f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)$$
Fonction marginale de $\boldsymbol{x}$:
$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \int f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)d\theta$$

Distribution *a posteriori*

Dénoté $\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$, elle représente la distribution de probabilité des paramètres conditionnelle aux données observées. La distribution *a posteriori* nous permet donc de savoir avec quelle probabilité (non-nulle) les paramètre $\boldsymbol{\theta}$ peuvent prendre certaine valeurs spécifiques sachant qu'on a observé certains x :

$$\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x}) = \frac{f_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})}{f_{\boldsymbol{X}}(x)} = \frac{f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)}{\int f_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)\pi(\theta)d\theta} \quad (1)$$

L'idée est de remplacer les différentes distributions dans l'Équation 1, et en déduire une distribution avec une paramétrisation différente[a].

a. Souvent, la distribution *a posteriori* aura la même distribution que celle *a priori*, mais avec des paramètres différents.

De plus, on défini la distribution pour les nouvelles observations :

Distribution prédictive du modèle

Dénoté $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$, elle est la distribution de probabilité d'une nouvelle observation y étant étant donné les données $\boldsymbol{x}$.
La fonction de densité d'une nouvelle observation étant donné la valeur du paramètre.
$$f_{Y|\boldsymbol{X}}(y|\boldsymbol{x}) = \int f_{Y|\Theta}(y|\theta)\pi_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})d\theta$$

L'estimateur Bayésien L'estimateur Bayésien est défini comme l'espérance du paramètre θ, sachant la distribution de X. En d'autres mots, on veut l'espérance de la distribution *a posteriori* :

Estimateur Bayésien

$$\hat{\theta}_{BAYES} = \mathrm{E}[\Theta|X] \quad (2)$$

0 Rappel de probabilité

Certaines lois à savoir

Loi	$\Pr(X=x)$ ou $f_X(x)$	$\mathrm{E}[X]$	$Var(X)$	
$Bin(n,p)$	$\binom{n}{x}p^x(1-p)^{n-x}$	np	$np(1-p)$	$((1$
$Pois(\lambda)$	$\frac{e^{-\lambda}\lambda^x}{x!}$	λ	λ	
$Gamma(\alpha,\lambda)$	$\frac{\lambda^\alpha x^{\alpha-1}e^{-\lambda x}}{\Gamma(\alpha)}$	$\frac{\alpha}{\lambda}$	$\frac{\alpha}{\lambda^2}$	
$Normale(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$	μ	σ^2	

Rappels d'algèbre linéaire

Matrice transposée

la matrice transposée est définie par $\boldsymbol{A}^\top$, telle que
$$\boldsymbol{A}^\top = \begin{bmatrix} a & -c \\ -b & d \end{bmatrix}$$

Déterminant d'une matrice

On peut calculer le déterminant $\det(\boldsymbol{A})$ de la matrice $\boldsymbol{A}$ tel que
$$\det(\boldsymbol{A}) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Inverse d'une matrice

L'équivalent de l'opération $\frac{1}{A}$ en algèbre linéaire est de calculer la matrice inverse de $\boldsymbol{A}^{-1}$, telle que
$$\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})}\begin{bmatrix} a & -c \\ -b & d \end{bmatrix}$$
où on multiple par la matrice adjointe de $\boldsymbol{A}$. Il faut normalement calculer les cofacteurs, mais le cas à 2 dimensions est un cas simplifié.