Contributeurs

MAS-I : Modern Actuarial Statistics I (ACT-2000, ACT-2003, ACT-2005)

aut., cre. Alec James van Rassel

Référence (manuels, YouTube, notes de cours) En ordre alphabétique :

src. Coaching Actuaries, Coaching Actuaries MAS-I Manual.
src. Cossette, H., ACT-1002 : Analyse probabiliste des risques actuariels, Université Laval, Québec (QC).
src. Côté, M.-P., ACT-2000 : Analyse statistique des risques actuariels, Université Laval, Québec (QC).
src. Hogg, R.V. ; McKean, J.W. ; and Craig, A.T., Introduction to Mathematical Statistics, 7th Edition, Prentice Hall, 2013.
src. Luong, A., ACT-2000 : Analyse statistique des risques actuariels, Université Laval, Québec (QC).
src. Luong, A., ACT-2005 : Mathématiques actuarielles IARD I, Université Laval, Québec (QC).
src. Marceau, É., ACT-2001 : Introduction à l'actuariat II, Université Laval, Québec (QC).
src. Starmer, J. (2015). StatQuest. Retrieved from https ://statquest.org/.
src. Tse, Y., Nonlife Actuarial Models, Theory Methods and Evaluation, Cambridge University Press, 2009.
src. Weishaus, A., CAS Exam MAS-I, Study Manual, 1st Edition, Actuarial Study Materials, 2018.

Contributeurs

pfr. Sharon van Rassel
pfr. Marianne Chouinard
pfr. Louis-Philippe Vignault
pfr. Philippe Morin

Cours reliés

ACT-2000 Analyse statistique des risques actuariels
ACT-2003 Modèles linéaires en actuariat
ACT-2005 Mathématiques actuarielles IARD I
ACT-2009 Processus stochastiques

En partie : mathématiques actuarielles vie I (**ACT-2004**), séries chronologiques (**ACT-2010**), introduction à l'actuariat II (**ACT-2001**) et méthodes numériques (**ACT-2002**).

Motivation

Inspiré par la chaîne de vidéos YouTube StatQuest et mon étude pour l'examen MAS-I, je crée ce document dans le but de simplifier tous les obstacles que j'ai encourus dans mon apprentissage des statistiques, et ainsi simplifier la vie des actuaires.

L'objectif est d'expliquer les concepts d'une façon claire, concise et visuelle ! Je vous prie de me faire part de tous commentaires et de me signaler toute erreur que

vous trouvez !

Table des matières

II Modèles linéaires en actuariat 53

I

Analyse statistique des risques actuariels

Échantillonnage et statistiques

Notation

X Variable aléatoire d'intérêt X avec fonction de densité $f(x;\theta)$;

Θ Ensemble des valeurs possible pour le paramètre θ tel que $\theta \in \Theta$;

Par exemple, pour une loi normale $\Theta = \{(\mu, \sigma^2) : \sigma^2 > 0, -\infty < \mu < \infty\}$.

$\{X_1, \ldots, X_n\}$ Échantillon de n observations (*variables aléatoires*).

On pose que toutes les observations ont la même distribution que X ;

On pose habituellement l'indépendance entre les observations ;

L'indépendance et la distribution identique rend l'échantillon un ***échantillon aléatoire*** ;

On dénote les *réalisations* de l'échantillon par $\{x_1, \ldots, x_n\}$.

Statistiques

Statistique T_n

Une statistique T_n est une fonction qui résume les n v.a. d'un échantillon aléatoire en une seule valeur.

Une statistique est donc également une *variable aléatoire* ;

Sa distribution est la **distribution d'échantillonnage** qui dépend de :

1. La statistique.
2. La taille de l'échantillon.
3. La distribution sous-jacente des données.

Moyenne échantillonnale $\bar{X}$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}.$$

Estime **sans biais** la moyenne μ ;

Si on pose que l'échantillon aléatoire est normalement distribué, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;

On centre et réduit pour trouver que $T_n = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$;

Si σ^2 est inconnue, on l'estime avec s_n^2 pour obtenir une distribution student— $T_n = \frac{\bar{X}-\mu}{S_n/\sqrt{n}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t_{(n-1)}$ où $W \sim \chi^2_{(n-1)}$.

Variance échantillonnale S_n^2

$$S_n^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}.$$

Estime **sans biais** la vraie variance σ^2 ;

S_n^2 n'est pas normalement distribuée, cependant la statistique $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Variance empirique $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}.$$

Estime **avec biais** la vraie variance σ^2.

Cependant, si la moyenne était connue et que nous n'avions pas à l'estimer avec $\bar{x}$, alors la variance empirique serait *sans biais*.

Statistique F

$$F = \frac{S_n^2/\sigma_1^2}{S_m^2/\sigma_2^2}.$$

Si on pose que les deux échantillons aléatoires indépendants $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ sont normalement distribués, $F \sim \mathcal{F}_{(n-1,m-1)}$.

Note sur majuscule vs minuscule On écrit les statistiques avec des majuscules lorsqu'elles sont aléatoires et avec des minuscules lorsque ce sont des réalisations. Par exemple, dans une probabilité on utilise une majuscule puisque la statistique est aléatoire. Pour un seuil α **fixé** d'un intervalle de confiance, le quantile n'est pas aléatoire et jusqu'à ce que l'on calcule l'intervalle avec l'échantillon observé, les statistiques sont également aléatoires.

Statistiques bivariées

Les statistiques bivariées sont définies pour un échantillon aléatoire bivarié $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

Covariance échantillonnale $cov_{X,Y}$

$$cov_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}.$$

Estime **sans biais** la vraie covariance $\sigma_{X,Y}$.

Corrélation échantillonnale $r_{X,Y}$

$$r_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}.$$ Également, $r_{X,Y} = \frac{cov_{X,Y}}{s_X s_Y}$.

Estime **sans biais** la vraie corrélation $\rho_{X,Y}$.

Vraisemblance

Notation

$\mathcal{L}(\theta; \boldsymbol{x})$ Fonction de vraisemblance de θ en fonction des observations $\boldsymbol{x}$;

$$\mathcal{L}(\theta; \boldsymbol{x}) = \prod_{i=1}^{n} f_X(x_i; \theta)$$

où $\boldsymbol{x}^\top = (x_1, \ldots, x_n)$.

$\{X_1, \ldots, X_n\}$ Échantillon de n observations.

Si les n observations sont indépendantes entres-elles et proviennent de la même distribution paramétrique (identiquement distribué) c'est un **échantillon aléatoire (iid)** ;

On peut le dénoter comme $\{X_n\}$.

Pour bien saisir ce que représente la fonction de vraisemblance $\mathcal{L}(\theta; \boldsymbol{x})$, il faut songer à ce que représente $f(x; \theta)$.
La fonction de vraisemblance $\mathcal{L}(\theta; \boldsymbol{x})$ se résume à une différente façon de voir la fonction de densité $f(x; \theta)$.

Au lieu de faire varier x pour un (ou des) paramètre θ fixe, on fait varier θ pour un échantillon d'observations $\boldsymbol{x}$ fixé.

Qualité de l'estimateur

La première section traite d'« **estimateurs ponctuels** ». C'est-à-dire, on produit une seule valeur comme notre meilleur essai pour déterminer la valeur de la population inconnue. Intrinsèquement, on ne s'attend pas à ce que cette valeur (même si c'en est une bonne) soit la vraie valeur exacte.

Une hypothèse plus utile à des fins d'interprétation est plutôt un **estimateur par intervalle** ; au lieu d'une seule valeur, il retourne un intervalle de valeurs plausibles qui peuvent toutes être la vraie valeur. Le type principal d'*estimateur par intervalle* est *l'intervalle de confiance* traité dans la deuxième sous-section.

En bref :

Estimateur ponctuel Règle (*fonction*) $\hat{\theta}_n$ qui décrit comment calculer une valeur précise estimée de θ en fonction de l'échantillon aléatoire.

Estimateur par intervalle *Intervalle aléatoire* qui produit un intervalle ayant une certaine probabilité de contenir la vraie valeur θ en fonction de l'échantillon aléatoire.

Estimation ponctuelle

Notation

θ Paramètre inconnu à estimer ;

$\hat{\theta}_n$ Estimateur de θ basé sur n observations ;

Souvent, on écrit $\hat{\theta}$ pour simplifier la notation.

Biais

Notation

$B(\hat{\theta}_n)$ Biais de l'estimateur $\hat{\theta}_n$.

Motivation

Lorsque nous avons un estimateur $\hat{\theta}_n$ pour un paramètre inconnu θ, on souhaite que, **en moyenne**, ses erreurs de prévision soient nulles. Le **biais** $B(\hat{\theta}_n)$ d'un estimateur quantifie les erreurs de l'estimateur dans ses prévisions de la vraie valeur du paramètre θ.

Biais d'un estimateur

Le biais est défini comme $B(\hat{\theta}_n) = E[\hat{\theta}_n|\theta] - \theta$, où $E[\hat{\theta}_n|\theta]$ est l'espérance de l'estimateur $\hat{\theta}_n$ sachant que la vraie valeur du paramètre est θ.

Estimateur sans biais

Lorsque le biais d'un estimateur est nul, $B(\hat{\theta}_n) = 0$, l'estimateur est *sans biais*.

Estimateur asymptotiquement sans biais

Lorsque le biais d'un estimateur tend vers 0 alors que le nombre d'observations de l'échantillon sur lequel il est basé tend vers l'infini, $\lim_{n\to\infty} B(\hat{\theta}_n) = 0$, l'estimateur est *asymptotiquement sans biais*.

Limitations

Bien que le biais quantifie les erreurs de prévisions de l'estimateur $\hat{\theta}_n$, il n'indique pas la variabilité de ses prévisions. Imagine une personne ayant ses pieds dans de l'eau bouillante et sa tête dans un congélateur. **En moyenne**, sa température corporelle est tiède. *En réalité*, sa température corporelle est à la fois extrêmement élevée et faible.

Variance

Notation

$\text{Var}(\hat{\theta}_n)$ Variance de l'estimateur $\hat{\theta}_n$.

Motivation

Les prévisions des estimateurs non biaisés seront toujours proches de la vraie valeur θ. Cependant, être bon ***en moyenne*** n'est pas suffisant et on souhaite évaluer la variabilité des prévisions d'un estimateur $\hat{\theta}_n$ avec sa variance $\text{Var}(\hat{\theta}_n)$.

Variance d'un estimateur

La variance est définie comme $\text{Var}(\hat{\theta}_n) = E\left[\left(\hat{\theta}_n - E[\hat{\theta}_n]\right)^2\right]$.

Limitations

Bien que la variance peut aider à dépister des estimateurs très variables, il a la limitation inhérente de ne pas prendre en considération le biais de l'estimateur. On cherche donc la juste balance entre le biais et la variance et utilisons l'erreur quadratique moyenne (EQM).

Erreur quadratique moyenne

Notation

$\mathbf{MSE}_{\hat{\theta}_n}(\theta)$ Erreur quadratique moyenne d'un estimateur $\hat{\theta}_n$

Motivation

L'erreur quadratique moyenne $\text{MSE}_{\hat{\theta}_n}(\theta)$ calcule la variance avec la vraie valeur du paramètre θ plutôt que l'espérance de l'estimateur $\text{E}[\hat{\theta}_n]$—il permet de quantifier l'écart entre un estimateur $\hat{\theta}_n$ et le vrai paramètre θ.

Erreur quadratique moyenne (EQM)

L'erreur quadratique moyenne est définie comme $\text{MSE}_{\hat{\theta}}(\theta) = \text{E}[(\hat{\theta}_n - \theta)^2]$.

Également, on peut réécrire l'expression comme $\text{MSE}_{\hat{\theta}}(\theta) = \text{Var}(\hat{\theta}_n) + \left[\text{B}(\hat{\theta}_n)\right]^2$.

Il s'ensuit que pour un estimateur non biaisé, $\text{MSE}_{\hat{\theta}}(\theta) = \text{Var}(\hat{\theta}_n)$.

En anglais, « *Mean Squared Error (MSE)* ».

Note Voir la section *Erreur quadratique moyenne* pour l'application de l'EQM dans le contexte *d'Apprentissage statistique*.

Convergence

Motivation

Nous voulons une mesure qui n'indique pas seulement qu'un estimateur arrive près de la bonne valeur souvent *(alias, une très petite variance)*, mais qu'il est mieux que d'autres estimateurs. Alors, un autre aspect à évaluer d'un estimateur est sa convergence pour de grands échantillons.

Par la loi des grands nombres, on s'attend à ce que la prévision d'un estimateur tend vers le vrai paramètre θ. On peut déduire avec intuition que le biais d'un estimateur « *consistent* » devrait tendre vers 0 et que sa variance devrait être très faible.

Il y a deux façons de définir la convergence d'un estimateur.

En fonction de la variance et du biais, $\hat{\theta}_n$ est un estimateur « *consistent* » de θ s'il est *asymptotiquement sans biais* et que sa *variance tend vers* 0 alors que la taille n de l'échantillon tend vers l'infini.

En termes mathématiques, $\hat{\theta}_n$ est un estimateur « *consistent* » de θ si la probabilité que sa prévision $\hat{\theta}$ du paramètre θ diffère de la vraie valeur par une erreur ε (presque nulle) tend vers 0 alors que la taille n de l'échantillon tend vers l'infini.

Cependant, la première façon est limitée car **l'inverse n'est pas vrai**—la variance et/ou biais d'un estimateur « *consistent* » ne tende(nt) pas nécessairement vers 0.

Convergence (« *consistency* ») d'un estimateur

$\hat{\theta}_n$ est un estimateur « *consistent* » de θ si :

1. $\lim_{n\to\infty} \text{B}(\hat{\theta}_n) = 0$.
2. $\lim_{n\to\infty} \text{Var}(\hat{\theta}_n) = 0$.

$\hat{\theta}_n$ est un estimateur « *consistent* » de θ si $\forall \varepsilon > 0$, $\lim_{n\to\infty} \Pr(|\hat{\theta}_n - \theta| > \varepsilon) = 0$.

Limitations

La convergence peut être manipulée. Dût à la sélection arbitraire de l'erreur ε, il est possible d'être sournois avec le choix de ε.

Note Les estimateurs par la méthode des moments sont « *consistent* » si ils sont uniques.

Détails mathématiques sur la convergence

On reprend les résultats de la section précédente en expliquant plus en détail la mathématique sous-jacente. **Vous pouvez sauter cette section**.

Convergence en probabilité

Notation

$\{Y_n\}$ Séquence de variables aléatoires ;
Y Variable aléatoire comprise dans $\{Y_n\}$.

On dit que Y_n converge en probabilité à Y si $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[|Y_n - Y| \geq \varepsilon\right] = 0$$

ou de façon équivalente,

$$\lim_{n\to\infty} \Pr\left[|Y_n - Y| < \varepsilon\right] = 1$$

On dénote la convergence en probabilité par : $Y_n \xrightarrow{P} Y$.

La convergence en probabilité est le théorème sous-jacent à la loi faible des grands nombres (vue en ACT-1002 : analyse probabiliste des risques actuariels).

Rappel : loi faible des grands nombres

Notation

$\{X_n\}$ Séquence de variables aléatoires iid avec moyenne μ et variance σ^2 où $\sigma^2 < \infty$;
$\overline{X}_n$ Moyenne empirique.

On pose que $\overline{X}_n \xrightarrow{P} \mu$.

Théorèmes résultant de la convergence en probabilité

Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n + Y_n \xrightarrow{P} X + Y$.
Soit $X_n \xrightarrow{P} X$ et une constante a. Alors $aX_n \xrightarrow{P} aX$.
Soit $X_n \xrightarrow{P} a$ et la fonction $g(\cdot)$ continue à a. Alors $g(X_n) \xrightarrow{P} g(a)$.
Soit $X_n \xrightarrow{P} X$ et la fonction continue $g(\cdot)$. Alors $g(X_n) \xrightarrow{P} g(X)$.
Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n Y_n \xrightarrow{P} XY$.

« *Consistency* »

Avec la notation définie ci-dessus, on simplifie la définition pour dire que $\hat{\theta}_n$ est un estimateur « *consistent* » de θ si $\hat{\theta}_n \xrightarrow{P} \theta$.

Note Voir la section *Tests sur la moyenne* de la section sur les *Tests d'hypothèses* pour la convergence en distribution.

Borne Cramér-Rao

Notation

$S(\theta)$ Fonction de Score, dérivée de la log-vraisemblance $S(\theta) = \frac{\partial \ln f(\theta;x)}{\partial \theta}$.

$\boldsymbol{I}_n(\theta)$ Matrice d'information de Fisher d'un échantillon aléatoire $\{X_n\}$;

La matrice d'information de Fisher pour une seule observation est dénotée $\boldsymbol{I}(\theta)$;

On obtient une "matrice" lorsque nous estimons plusieurs paramètres et donc $\boldsymbol{\theta}$ n'est pas juste un scalaire θ.

Information (de Fisher) de θ

Contexte

On peut penser à l'information de Fisher comme une mesure de la sensibilité de la dérivée de la log-vraisemblance $\ell'(\theta)$ aux données. Une information élevée, exprimée par une variabilité de $\ell'(\theta)$ élevée, suggère que la forme de $\ell(\theta)$ est sensible aux données.

L'information (de Fisher) de θ est $\boldsymbol{I}(\theta) = \text{Var}(\ell'(\theta))$.

Si les données sont (iid), on peut récrire $\boldsymbol{I}(\theta) = -\text{E}[\ell''(\theta)]$.

Pour des données (iid), on obtient que $\boldsymbol{I}_n(\theta) = n\boldsymbol{I}(\theta) = -n\text{E}\left[\frac{\partial^2}{\partial\theta^2}\ln f(x;\boldsymbol{\theta})\right]$.

Matrice d'information (de Fisher) de $\boldsymbol{\theta}$

Pour une distribution ayant plusieurs paramètres, l'information de Fisher devient une matrice des dérivées partielles de la log-vraisemblance $\ell(\boldsymbol{\theta})$.

Matrice d'information (de Fisher) pour $\boldsymbol{\theta} = (\theta_1, \theta_2)$

$$\boldsymbol{I}_n(\boldsymbol{\theta}) = \begin{bmatrix} -n\text{E}\left[\frac{\partial^2}{\partial\theta_1^2}\ln f(x;\boldsymbol{\theta})\right] & -n\text{E}\left[\frac{\partial^2}{\partial\theta_1\theta_2}\ln f(x;\boldsymbol{\theta})\right] \\ -n\text{E}\left[\frac{\partial^2}{\partial\theta_1\theta_2}\ln f(x;\boldsymbol{\theta})\right] & -n\text{E}\left[\frac{\partial^2}{\partial\theta_2^2}\ln f(x;\boldsymbol{\theta})\right] \end{bmatrix}$$

Borne inférieure Cramér-Rao

Motivation

Lorsque nous analysons la variance $\text{Var}(\hat{\theta}_n)$ d'un estimateur sans biais, la **borne inférieure de Cramér-Rao** sert de point de départ.

Sous certaines conditions de régularité, la borne inférieure Cramér-Rao est définie comme $\text{Var}(\hat{\theta}_n) \geq \frac{1}{\boldsymbol{I}_n(\theta)}$.

Dans le cas multivarié, $\text{Var}(\hat{\theta}_j) \geq \boldsymbol{I}_n^{-1}(\boldsymbol{\theta})_{j,j}$.

Détails mathématiques sur la borne Cramér-Rao

La borne de Cramér-Rao est un concept qui échappe souvent aux étudiants. Sur la base de ce vidéo et de ce vidéo, je vais tenter d'expliquer l'intuition sous-jacente au concept. Ce concept va réapparaître plus tard dans le bac et donc, s'il n'est pas clair d'ici la fin de la section, je vous conseille d'aller visionner les vidéos. Bien que je ne le recommande pas, **vous pouvez sauter cette section**.

Premièrement, on définit l'utilité des deux premières dérivées :

$\frac{\partial}{\partial\theta}\mathcal{L}(\theta)$: Représente le « *rate of change* » de la fonction ;

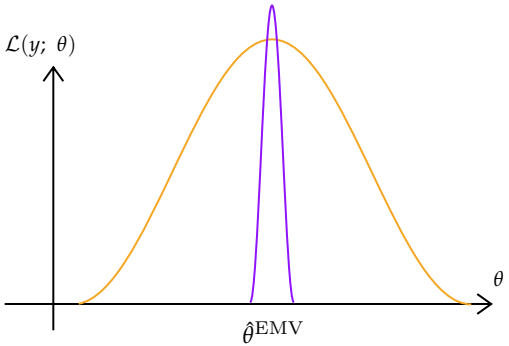
$\frac{\partial^2}{\partial\theta^2}\mathcal{L}(\theta)$: Représente la concavité de la fonction ; on peut y penser comme sa forme.

L'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}^{\text{EMV}}$ du paramètre θ d'une distribution maximise la fonction de vraisemblance en fonction d'un échantillon aléatoire. En posant la première dérivée de la fonction de vraisemblance comme étant égale à 0, on trouve le "point" auquel l'EMV est égale à θ—$\theta^{\hat{\text{EMV}}} = \theta$.

Note : L'EMV devient un "point" lorsqu'on le calcule pour un échantillon aléatoire d'observations.

La fonction de vraisemblance **est concave** et, puisque sa première dérivée est nulle à $\hat{\theta}_n^{\text{EMV}}$, elle va augmenter avant ce point puis diminuer par après. La première dérivée permet donc de trouver une fonction **qui est maximisée à** $\hat{\theta}_n^{\text{EMV}}$. Cependant, ceci ne permet pas d'identifier une fonction unique—plusieurs fonctions peuvent être maximisées au même **point** tout en ayant des formes différentes.

Par exemple, on trace ci-dessous la fonction de vraisemblance et une autre fonction également maximisée à $\hat{\theta}_n^{\text{EMV}}$:



On peut voir que la forme de la fonction de vraisemblance est plus comprimée. Alias, sa concavité est plus forte que l'autre fonction qui se maximise au même point. C'est-à-dire, la fonction de vraisemblance correspond à la fonction, dont le maximum est à $\hat{\theta}_n^{\text{EMV}}$, avec *la plus forte concavité*.

On peut observer que plus la concavité augmente, plus la variabilité de la fonction de vraisemblance se rapetisse. En effet, une faible concavité implique que la fonction de vraisemblance a un grand étendu de valeurs possibles et moins de points près de $\hat{\theta}^{\text{EMV}}$. En bref, la deuxième dérivée assure que, parmi les fonctions se maximisant à $\hat{\theta}_n^{\text{EMV}}$, la fonction de vraisemblance est la fonction dont la variabilité des prévisions est minimisée.

L'information de Fisher permet de quantifier cette fonction de la deuxième dérivée. Puis, la borne de Cramér-Rao se définit comme son réciproque $1/\boldsymbol{I}(\theta)$. L'intuition est que plus la concavité est faible, plus l'étendue est grand. Prendre le réciproque de l'information de Fisher permet donc de quantifier l'agrandissement de l'étendu.

Lorsque l'information de Fisher tend vers l'infini, alias la force de la concavité croît infiniment, on dit que la distribution de l'estimateur est "*asymptotiquement normale*" tel que $\hat{\theta}^{\text{EMV}} \xrightarrow{a.s.} \mathcal{N}\left(\mu = \theta, \sigma^2 = \frac{1}{\boldsymbol{I}(\theta)}\right)$ où a.s. veut dire asymptotiquement.

Efficacité

Notation

$\text{eff}(\hat{\theta}_n)$ Efficacité d'un estimateur $\hat{\theta}_n$;
$\text{eff}(\hat{\theta}_n, \tilde{\theta}_n)$ Efficacité de l'estimateur $\hat{\theta}_n$ relatif à l'estimateur $\tilde{\theta}_n$.

Motivation

Puisque la variance d'un estimateur ne peut être inférieure à la borne Cramér-Rao, il est désirable qu'un estimateur (sans biais) l'atteigne. On définit donc l'efficacité (« *efficiency* ») d'un estimateur (sans biais) comme le ratio la borne Cramér-Rao à sa variance.

Note Pour toute la section d'efficacité, on suppose que les estimateurs sont **sans biais**.

Efficacité (« *efficiency* ») d'un estimateur

L'« *efficiency* » d'un estimateur $\hat{\theta}_n$ est définie comme $\text{eff}(\hat{\theta}_n) = \frac{1/\boldsymbol{I}_n(\theta)}{\text{Var}(\hat{\theta})}$.

Estimateur « efficient »

Si $\text{eff}(\hat{\theta}_n) = 1$, alias la variance de l'estimateur est égale à la borne Cramér-Rao, l'estimateur est **« *efficient* »**.

Motivation

On peut utiliser le concept d'efficacité pour comparer des estimateurs entre-eux plutôt qu'à la borne Cramér-Rao. On obtient donc l'efficacité relative d'un estimateur relatif à un autre estimateur.

Efficacité (« *efficiency* ») relative

« *The relative efficiency* » de l'estimateur $\hat{\theta}_n$ à l'estimateur $\tilde{\theta}_n$ est définie comme $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\text{Var}(\hat{\theta}_n)}{\text{Var}(\tilde{\theta}_n)}$.

Si $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) < 1$, l'estimateur $\hat{\theta}_n$ est plus efficace que l'estimateur $\tilde{\theta}_n$ et vice-versa si $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) > 1$.

Estimateur non biaisé à variance minimale (MVUE)

Motivation

Si nous cherchons à minimiser la variance est désirons un estimateur sans biais, alors nous souhaitons un estimateur « *efficient* ». Cependant, cet estimateur n'existe pas toujours et donc nous voulons l'estimateur non biaisé ayant la plus petite variance possible.

Estimateur non biaisé à variance minimale (MVUE)

L'estimateur sans biais ayant la plus petite parmi tous les estimateurs *non biaisés*.

En anglais, « *minimum variance unbiased estimator (**MVUE**)* ».

Note On peut trouver cet estimateur comme l'estimateur non biaisé ayant la plus petite efficacité. Sinon, on peut l'identifier avec le *théorème de Lehmann-Scheffé* ou le théorème de Rao-Blackwell décrits dans la section *Statistiques exhaustives*.

Limitations

L'estimateur MVUE n'est pas nécessairement l'estimateur ayant la plus petite variance car un estimateur biaisé peut avoir une variance inférieure à celle du MVUE.

Estimation par intervalles

Contexte

Le type principal d'estimateur par intervalle est l'**intervalle de confiance**. Un intervalle de confiance suggère où est situé la valeur du paramètre d'intérêt à un certain niveau de confiance.

De façon générale, on requiert les éléments suivants pour obtenir un intervalle de confiance :

1. Une méthodologie
2. Une distribution adéquate
3. Un niveau de confiance
4. Des données

Intervalle de confiance

On décrit un **intervalle aléatoire** (L, U) d'un paramètre θ avec $\Pr(L \leq \theta \leq U) = k$ où k est le ***niveau de confiance***.

Lorsque L et U sont évalués pour les données, nous obtenons un **intervalle numérique** : l'**intervalle de confiance** (l, u) de $100k\%$ pour θ.

Note Les composantes L et U sont aléatoires alors que θ est fixe.

Pour construire des intervalles de confiance, nous utilisons la ***méthode du pivot*** basé sur *un pivot*.

Pivot (« *pivotal quantity* »)

Un pivot est une fonction **des observations** et **des paramètres inconnus** de la distribution de l'échantillon. Cependant, la **distribution du pivot** ne **dépend pas des paramètres inconnus**.

Le pivot est donc semblable, mais *distinct*, d'une statistique. Si la distribution d'une statistique dépend de paramètres inconnus, elle n'est pas un pivot. Si le pivot est composée de paramètres inconnus, alors il n'est pas une statistique.

Par exemple, le pivot $\frac{\bar{X}_n - \mu}{\sqrt{n}}$ est un pivot ; il est une fonction de l'échantillon $(\bar{X}_n)$ et du paramètre inconnu (μ), mais sa distribution est obtenue via le

théorème central limite et donc ne dépend pas de μ.

Méthode du pivot

1. Trouver un pivot $W = W(X_1, \ldots, X_n; \theta)$.
2. Trouver les quantiles de la distribution de W, tels que $\Pr(w_{(1-k)/2} \leq W \leq w_{(1+k)/2}) = k$ pour un intervalle bilatéral.
   Pour des intervalles unilatéraux, on ajuste le niveau des quantiles.
3. Isoler θ de l'équation $w_{(1-k)/2} \leq W(X_1, \ldots, X_n; \theta) \leq w_{(1+k)/2}$ pour obtenir l'intervalle de confiance.

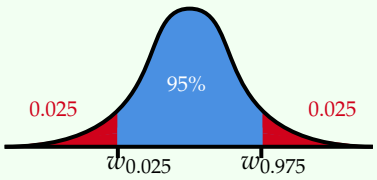
Exemple d'application de la méthode du pivot

Soit $X \sim \text{Pareto}(\alpha = 2, \theta)$ et le pivot $W = \frac{\theta}{X}$. Nous avons un échantillon d'une observation : $x = \{3.5\}$ et on désire trouver un intervalle de confiance bilatéral de 95% pour θ.

1. On doit vérifier que $W = \frac{\theta}{X}$ est un pivot valide.
   (a) W est une fonction de l'échantillon et du paramètre inconnu.
   (b) La distribution de W ne doit pas dépendre de θ, on confirme en trouvant sa fonction de répartition :

$$F_W(w) = \Pr(W \leq w) = \Pr\left(\frac{\theta}{X} \leq w\right) = \Pr\left(X < \frac{\theta}{w}\right)$$
$$= \left(\frac{\theta}{\left(\frac{\theta}{w}\right) + \theta}\right)^2$$
$$= \left(\frac{w}{w+1}\right)^2$$

2. On doit établir quels quantiles trouver pour un niveau de confiance de 95%. On déduit ces percentiles visuellement :



3. On isoles les valeurs des percentiles $w_{0.025}$ et $w_{0.975}$ avec la fonction de répartition :

$$0.025 = \left(\frac{w}{w+1}\right)^2 \Rightarrow w = 0.1878$$
$$0.975 = \left(\frac{w}{w+1}\right)^2 \Rightarrow w = 78.4968$$

4. On établit la probabilité désiré

$$\Pr(w_{0.025} \leq W \leq w_{0.975}) = 0.95$$
$$\Pr(0.1878 \leq W \leq 78.4968) = 0.95$$
$$\Pr\left(0.1878 \leq \frac{\theta}{X} \leq 78.4968\right) = 0.95$$
$$\Pr\left(0.1878 \times X \leq \theta \leq 78.4968 \times X\right) = 0.95$$

5. Finalement, pour obtenir « l'intervalle numérique », alias l'**intervalle de confiance** de niveau 95%, on insère l'échantillon de données et on trouve que nous sommes confiants à un niveau de 95% que $\theta \in [0.1878 \times 3.5, 78.4968 \times 3.5] = [0.6573, 274.7389]$.

Contexte

Ce qu'il faut bien saisir avec les intervalles de confiance, c'est que *soit* θ ***est*** **contenu** dans l'intervalle (L, U) *ou* **il ne l'est *pas***.

On peut conceptualiser les intervalles comme une distribution binomiale avec probabilité de succès de k. Si l'on effectue N essais indépendants, on s'attend à ce que $k \times N$ intervalles de confiance contiennent θ. Donc, nous sommes confiants à k% que la vraie valeur de θ est contenue dans l'intervalle **observé** (l, u).

Efficacité des intervalles de confiance Typiquement, la largeur de l'intervalle (L, U) augmente si on augmente le niveau de confiance k. Par exemple, pour être certain à 100% que l'intervalle va contenir la valeur, on a qu'à faire un intervalle $(-\infty, \infty)$.

Donc, un intervalle plus petit nous donne plus d'information si le niveau est adéquat. On dit que pour un même niveau k, l'intervalle avec la plus petite largeur est *plus* ***efficace*** que l'autre.

Intervalles de confiance

Note Ces sections sur les *Intervalles sur la moyenne*, les *Intervalles sur les proportions* et les *Intervalles sur la variance* sont semblables aux sections correspondantes des *Tests d'hypothèses*. La section sur les *Tests d'hypothèses* a été faite avant celle-ci sur les *Intervalles de confiance*, et donc je vous conseille de la lire en premier.

Intervalles sur la moyenne

Note Voir la section sur *Tests sur la moyenne* pour les rappels de la définition du théorème centrale limite et de la définition de la loi de Student.

1 échantillon

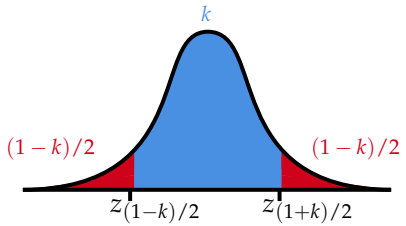
Pour un échantillon aléatoire de taille n avec moyenne μ et variance σ^2,

variance	distribution de l'échantillon	n grand ?	pivot	distribution du pivot
connue	n'importe quelle	oui	$\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}$	$\mathcal{N}(0,1)$
inconnue	normale	non	$\frac{\bar{X}_n-\mu}{S_n/\sqrt{n}}$	t_{n-1}

Donc, lorsque la variance est connue,

intervalle de confiance	intervalle aléatoire	intervalle numérique
bilatéral	(L,U)	$\mu \in \left[\bar{x}_n - z_{(1-k)/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{(1+k)/2}\frac{\sigma}{\sqrt{n}}\right]$
unilatéral à gauche	$(-\infty,U)$	$\mu \in \left[-\infty, \bar{x}_n + z_k\frac{\sigma}{\sqrt{n}}\right]$
unilatéral à droite	(L,∞)	$\mu \in \left[\bar{x}_n - z_k\frac{\sigma}{\sqrt{n}}, \infty\right]$

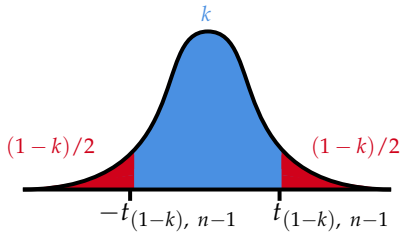
Visuellement, nous avons une aire de $(1-k)/2$ dans les deux queues et les points $z_{(1-k)/2}$ et $z_{(1+k)/2}$:



Puis, lorsque la variance n'est *pas* connue,

intervalle de confiance	intervalle aléatoire	intervalle numérique
bilatéral	(L,U)	$\mu \in \left[\bar{x}_n - t_{1-k,n-1}\frac{s}{\sqrt{n}}, \bar{x}_n + t_{1-k,n-1}\frac{s}{\sqrt{n}}\right]$
unilatéral à gauche	$(-\infty,U)$	$\mu \in \left[-\infty, \bar{x}_n + t_{2(1-k),n-1}\frac{s}{\sqrt{n}}\right]$
unilatéral à droite	(L,∞)	$\mu \in \left[\bar{x}_n - t_{2(1-k),n-1}\frac{s}{\sqrt{n}}, \infty\right]$

Visuellement, nous avons une aire de $(1-k)/2$ dans les deux queues et les points $-t_{(1-k),n-1}$ et $t_{(1-k),n-1}$:



Note Voir cette explication de la différence entre les quantiles de la loi normale et les quantiles de la loi de Student du chapitre sur les *Tests d'hypothèses*.

Note Puisque la loi normale est symétrique, $z_{(1-k)/2} = -z_{(1+k)/2}$. Il s'ensuit qu'on peut simplifier l'écriture de l'intervalle de confiance bilatéral comme $\mu \in \bar{x}_n \pm z_{(1-k)/2}\frac{\sigma}{\sqrt{n}}$.

Note Voir la sous-section *Intervalle de confiance et de prévision* de la section sur la *Régression linéaire simple* pour l'application de l'intervalle de confiance sur la moyenne à la régression linéaire simple.

2 échantillons

Pour 2 échantillons aléatoires **indépendants** de tailles n_1 et n_2,

variances	distribution des échantillons	autres conditions ?	n_k grands ?	pivot	distribution du pivot
connues	n'importe lesquelles	non	oui	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_1}}}$	$\mathcal{N}(0,1)$
inconnues	normales	$\sigma_1^2=\sigma_2^2$	non	$\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_1}}}$	$t_{n_1+n_2-2}$

où S_p est le « pooled estimator » S_p de l'écart-type.
Donc, lorsque la variance est connue,

intervalle de confiance	intervalle numérique
bilatéral	$(\mu_1-\mu_2)\in(\bar{x}_1-\bar{x}_2)\pm z_{(1-k)/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
unilatéral à gauche	$(\mu_1-\mu_2)\in\left[-\infty,(\bar{x}_1-\bar{x}_2)+z_k\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
unilatéral à droite	$(\mu_1-\mu_2)\in\left[(\bar{x}_1-\bar{x}_2)-z_k\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},\infty\right]$

Puis, lorsque la variance n'est *pas* connue,

intervalle de confiance	intervalle numérique
bilatéral	$(\mu_1-\mu_2)\in(\bar{x}_1-\bar{x}_2)\pm t_{(1-k),n_1+n_2-1}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$
unilatéral à gauche	$(\mu_1-\mu_2)\in\left[-\infty,(\bar{x}_1-\bar{x}_2)+t_{2(1-k),n_1+n_2-1}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$
unilatéral à droite	$(\mu_1-\mu_2)\in\left[(\bar{x}_1-\bar{x}_2)-t_{2(1-k),n_1+n_2-1}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},\infty\right]$

Si les données sont **appariées**, les intervalles de confiance sont les mêmes que ceux pour un échantillon avec les substitutions suivantes : $\bar{x}=\bar{d}$, $\sigma^2=\sigma_D^2$, $n=n_*$ et $s^2=s_D^2$. Voir les *Tests sur la moyenne* sur *2 échantillons* pour l'explication de données appariées.

Intervalles sur les proportions

Note Voir la boîte de contexte des *Tests sur les proportions* pour comprendre la notion de variance dans le cas d'une distribution Bernoulli.

1 échantillon

Pour un échantillon aléatoire de taille n tiré d'une distribution Bernoulli, on déduit du théorème centrale limite que, lorsque n est grand, le pivot $\frac{\hat{q}-q}{\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}} \xrightarrow{D} Z$.

Donc,

intervalle de confiance	intervalle numérique
bilatéral	$q \in \hat{q} \pm z_{(1-k)/2}\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}$
unilatéral à gauche	$q \in \left[-\infty, \hat{q} + z_k\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}\right]$
unilatéral à droite	$q \in \left[\hat{q} - z_k\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}, \infty\right]$

Note L'intervalle de confiance diffère de la statistique du test d'hypothèse correspondant car nous utilisons $\hat{q}$ pour l'erreur type et non q.

2 échantillons

Pour 2 échantillons aléatoires de tailles n_1 et n_2, on déduit du théorème centrale limite que, lorsque n_1 et n_2 sont grands, le pivot $\frac{(\hat{q}_1-\hat{q}_2)-(q_1-q_2)}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1}+\frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}} \xrightarrow{D} Z$.

Donc,

intervalle de confiance	intervalle numérique
bilatéral	$(q_1-q_2) \in (\hat{q}_1-\hat{q}_2) \pm z_{(1-k)/2}\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1}+\frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}$
unilatéral à gauche	$(q_1-q_2) \in \left[-\infty, (\hat{q}_1-\hat{q}_2) + z_k\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1}+\frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}\right]$
unilatéral à droite	$(q_1-q_2) \in \left[(\hat{q}_1-\hat{q}_2) - z_k\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1}+\frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}, \infty\right]$

Intervalles sur la variance

« *Pooled Estimator* »

Le « *pooled estimator* » est la moyenne pondérée des deux variances échantillonnales $S_p^2 = \frac{(n-1)S_n^2+(m-1)S_m^2}{n+m-2}$.

Note Voir la section sur *Tests sur la variance* pour les rappels de la définition de la loi du khi carré et de la définition de la loi de Fisher.

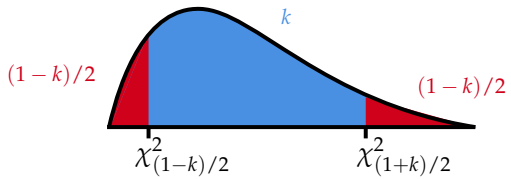
1 échantillon

Pour un échantillon aléatoire de taille n tiré d'une distribution normale, le pivot $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

Donc,

intervalle de confiance	intervalle numérique
bilatéral	$\sigma^2 \in \left[\frac{(n-1)s_n^2}{\chi^2_{(1+k)/2,n-1}}, \frac{(n-1)s_n^2}{\chi^2_{(1-k)/2,n-1}}\right]$
unilatéral à gauche	$\sigma^2 \in \left[0, \frac{(n-1)s_n^2}{\chi^2_{1-k,n-1}}\right]$
unilatéral à droite	$\sigma^2 \in \left[\frac{(n-1)s_n^2}{\chi^2_{k,n-1}}, \infty\right]$

Visuellement, nous avons une aire de $(1-k)/2$ dans les deux queues et les points $\chi^2_{(1+k)/2}$ et $\chi^2_{(1-k)/2}$:



Note Puisque la distribution du khi carré est asymétrique de droite, les intervalles de confiance n'ont pas de symétrie.

2 échantillons

Pour 2 échantillons aléatoires indépendants de tailles n_1 et n_2 qui sont normalement distribués avec variances σ_1^2 et σ_2^2, le pivot $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$.

Donc,

intervalle de confiance	intervalle numérique
bilatéral	$\sigma^2 \in \left[\frac{s_1^2}{s_2^2}\left(F_{(1-k)/2,n_1-1,n_2-1}\right)^{-1}, \frac{s_1^2}{s_2^2}\left(F_{(1-k)/2,n_2-1,n_1-1}\right)\right]$
unilatéral à gauche	$\sigma^2 \in \left[0, \frac{s_1^2}{s_2^2}F_{1-k,n_2-1,n_1-1}\right]$
unilatéral à droite	$\sigma^2 \in \left[\frac{s_1^2}{s_2^2}\left(F_{1-k,n_1-1,n_2-1}\right)^{-1}, \infty\right]$

Tests d'hypothèses

Hypothèses

Contexte

Les statistiques classiques posent que tout phénomène observable est régi par un *"**processus**" sous-jacent*. On ne peut jamais savoir exactement ce qu'est ce "processus", le mieux que l'on peut faire est d'émettre des ***hypothèses*** vraisemblables sur ce qu'il pourrait être.

Afin de déterminer la ***vraisemblance*** que les observations sont régies par le processus hypothétique, on analyse les observations en présumant qu'elles le sont. On accepte le processus hypothétique si la vraisemblance est suffisamment élevée.

Un test d'hypothèse prend deux possibilités de scénario opposantes et vérifie laquelle des deux est mieux supportée par les données. On pose un scénario de base (hypothèse nulle) représentant le statu quo. Puis, on résume les données en une seule statistique (de test) avec laquelle on évalue la vraisemblance que les données sont régies par un scénario alternatif que celui de base.

Notation

Θ_0 et Θ_1 Sous-ensembles disjoints de Θ tel que $\Theta_0 \cup \Theta_1 = \Theta$;

H_0 Hypothèse nulle.

H_1 Hypothèse alternative.

Test d'hypothèse

On spécifie une hypothèse nulle et une hypothèse alternative :

$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

Puis, on spécifie une expérience et un test pour décider si l'on accepte ou rejette l'hypothèse nulle.

Hypothèse nulle

Représente généralement le statu quo jusqu'à preuve contraire.

Hypothèse alternative

Représente généralement un changement du statu quo.

Décision du test d'hypothèse

Soit on :

1. Ne peut pas rejeter l'hypothèse nulle H_0.
2. On rejette l'hypothèse nulle H_0 pour l'hypothèse alternative H_a.

Contexte

Typiquement, on dit qu'« on ne peut pas rejeter l'hypothèse nulle » plutôt que dire qu'« on accepte l'hypothèse nulle ». Cette interprétation est plus précise, car un test d'hypothèse ne prouve pas quelle hypothèse est la bonne, elle décide quelle option est plus vraisemblable en fonction des données.

Il s'ensuit que nous sommes biaisés vers le statu quo—l'hypothèse nulle H_0. Donc, un test d'hypothèse n'est pas un choix entre deux scénarios, mais plutôt une évaluation pour voir s'il y a suffisamment d'évidence dans les données pour changer l'hypothèse du statu quo.

Terminologie

Hypothèse simple Spécifie **entièrement** une distribution de probabilité.

Par exemple, $\mathcal{H}_0 : q = 0.50$—on connaît la valeur exacte du paramètre q pour une distribution Bernoulli.

Hypothèse composite Spécifie **partiellement** une distribution de probabilité.

Par exemple, $\mathcal{H}_1 : q \neq 0.50$—on ne connaît pas la valeur exacte du paramètre q, il pourrait être n'importe quel chiffre sauf 0.50.

Par exemple, pour une distribution normale de moyenne μ et variance inconnue σ^2, $\mathcal{H}_0 : \mu = 0.50$—on ne connaît pas la variance et donc la distribution n'est pas *entièrement* spécifiée.

Exemple du laissez-passer universitaire (LPU)

Par exemple, on veut savoir si les étudiants utilisent l'autobus (oui ou non) avant et après l'implantation du LPU.
On pose que la proportion des gens qui utilisent l'autobus est $q = 0.44$.
Il y a deux types de tests qu'on peut faire,

Tester si l'utilisation est différente est un test "**bilatéral**", car on teste si elle a augmenté *ou* diminuée ;

$H_0 : q = 0.44$ $H_1 : q \neq 0.44$

Tester si l'utilisation a augmenté est un test "**unilatéral**", car on teste uniquement si elle a augmenté.

$H_0 : q = 0.44$ $H_1 : q > 0.44$

Un test unilatéral requiert que l'on sache déjà que la proportion de gens "*doit*" être supérieure. Un test bilatéral est plus conservatif et test les deux possibilités, il devrait donc être celui qu'on applique par défaut.

L'hypothèse :

nulle dans les deux cas est que, en moyenne, l'utilisation de l'autobus n'a pas *changée*.

alternative dans le cas d'un test :

unilatéral est que, en moyenne, l'utilisation a *augmentée*.

bilatéral est que, en moyenne, l'utilisation a *changée*.

Région et valeur critique

Région critique

Notation

$\mathcal{S}$ "Ensemble" de tous les résultats possible pour l'échantillon aléatoire ;

$\mathcal{C}$ **Région critique** du test qui est un sous-ensemble de $\mathcal{S}$.

La région critique $\mathcal{C}$ est l'ensemble des valeurs de la statistique, que l'on considère trop « extrêmes » pour être le statu quo, pour lesquelles on rejette l'hypothèse nulle H_0.

On rejette H_0 si $\{X_1, \ldots, X_n\} \in \mathcal{C}$ mais on ne peut pas rejeter H_0 si $\{X_1, \ldots, X_n\} \in \mathcal{C}^c$.

On peut aussi dire « **région de rejet** ».

Exemple du laissez-passer universitaire (LPU)

On reprend l'exemple du LPU.

L'ensemble des résultats possibles est $\mathcal{S} = [0, 1]$.

Un test "**bilatéral**" a comme région critique $\mathcal{C} = [0, 0.44) \cup (0.44, 1]$;

Un test "**unilatéral**" testant l'augmentation a comme région critique $\mathcal{C} = (0.44, 1]$.

En bref, voici un résumé des régions et valeurs critiques selon le type de test :

	unilatéral à gauche	bilatéral	unilatéral à droite
Région critique	$t \leq -c$	$\lvert t \rvert \geq c$	$t \geq c$
Valeur critique	$-z_{1-\alpha}$	$z_{1-\alpha/2}$	$z_{1-\alpha}$

Note Voir la section *Résumé graphique des régions critiques* pour un résumé graphique des régions critiques selon le type de test.

Erreurs de test

Contexte

Bien que nous tentons de prendre une décision informée sur quel test est le vrai, on ne peut jamais être certain que l'hypothèse sélectionnée est la bonne. Cependant, on peut évaluer l'impact d'une mauvaise décision selon que l'hypothèse nulle H_0 soit réellement la vraie hypothèse ou pas.

Avec cet approche, on trouve que l'on peut faire 2 types d'erreur, soit une erreur de type I (« *false positive* ») ou une erreur de type II (« *false negative* »). Le tableau ci-dessous montre ce qu'elles représentent, puis la section sur les *Tests optimaux (les plus puissants)* va plus en détails sur l'optimisation des erreurs.

	Vrai état	
Décision	H_0	H_1
Rejeter H_0	Erreur de type I	Bonne décision
Accepter H_0	Bonne décision	Erreur de type II

Certitude du test

Lorsque nous voulons quantifier le degré auquel nous sommes confiants du test, nous utilisons la ***valeur*** p qui a trois composantes :

1. La probabilité que l'événement se produise aléatoirement.
2. La probabilité qu'un événement tout aussi rare se produise.
3. La probabilité qu'un événement encore plus rare se produise.

Exemple de pile ou face

On souhaite tester si, en obtenant deux piles sur deux lancers, nous avons une pièce de monnaie truquée :

Hypothèse nulle Ma pièce de monnaie n'est pas truquée même si j'ai obtenu deux piles.

Étapes du calcul de la valeur p :

1. On calcule la probabilité d'obtenir 2 piles : $0.5 \times 0.5 = 0.25$.
2. Puis, on calcule la probabilité d'obtenir 2 faces (un événement tout aussi rare) : $0.5 \times 0.5 = 0.25$.
3. Finalement, il n'y a pas d'autres séquences plus rares.

Donc, la valeur p du test est de 0.50 ce qui est plutôt élevé. Souvent, on pose que la valeur p du test doit être d'au plus 0.05 ce qui veut dire que des événements tout aussi (ou plus) rares doivent arriver moins que 5% du temps pour que l'on considère la pièce de monnaie comme étant truquée.

Donc, avec un valeur p de 0.50, on ne peut pas rejeter l'hypothèse nulle que notre pièce de monnaie n'est pas spéciale.

Dans le cas continu, on somme les probabilités d'être plus rare ou d'être moins rare. C'est la même idée que les intervalles de confiance avec la valeur p, ou *seuil de signifiance* α, représentée en rouge. Si la valeur p est :

faible, ceci indique que d'autres distributions pourraient potentiellement mieux s'ajuster aux données puisque l'événement est très rare ;

élevée, ceci indique que l'événement est très courant et que la distribution semble être bien ajustée.

Terminologie

Il y a plusieurs termes semblables qui peuvent devenir mélangeants.

p La **valeur** p du test.

α Dénote habituellement le **seuil de signifiance** ou la **taille** du test.

La valeur p du test

On peut définir la valeur p de plusieurs façons :

1. la probabilité d'un événement tout aussi (ou plus) rare sous l'hypothèse nulle.
2. la **taille** de la région critique $\mathcal{C}$.

   C'est-à-dire, l'*aire* de la région de rejet de l'hypothèse nulle H_0 alors qu'elle est vraie.
3. la **probabilité d'une erreur de type I**.

   C'est-à-dire, la probabilité de rejeter H_0 alors qu'elle est vraie
4. le **seuil de signifiance**.

Le seuil de signifiance α du test

On dénote habituellement par α le **seuil de signifiance**, ou la **taille**, du test. Donc, c'est la valeur que la valeur p doit atteindre afin de pouvoir rejeter l'hypothèse nulle.

C'est la même idée qu'avec les intervalles de confiance.

En anglais, « *threshold for significance* ».

En termes mathématiques, on définit $\alpha = \max_{\theta \in \Theta_0} \Pr\{(X_1, \ldots, X_n) \in \mathcal{C} | \theta\}$.

En mots, on maximise la probabilité que l'échantillon aléatoire soit contenu dans la région critique (alias rejeter H_0) alors que la distribution est tracée en fonction du paramètre θ de l'hypothèse nulle.

Valeur p vs seuil α

On défini ces deux termes pour le test bilatéral en fonction de la valeur **observée** de la statistique t et de la valeur **critique** de la statistique c :

$p = \Pr(|T| \geq |t| | H_0 \text{ est vrai.})$

Comparaison	Décision
$p \leq \alpha$	Rejete H_0
$p > \alpha$	Ne rejete pas H_0

$\alpha = \Pr(|T| \geq |c| | H_0 \text{ est vrai.})$

Comparaison	Décision
$\lvert t\rvert \geq c$	Rejete H_0
$\lvert t\rvert < c$	Ne rejete pas H_0

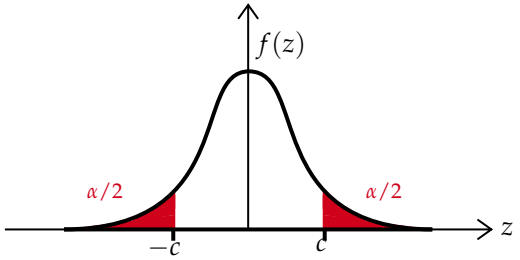
Résumé graphique des régions critiques

On résume les régions critiques pour les 3 types d'hypothèses :

Bilatéral

$p = \Pr(|T| \geq |t| \mid H_0 \text{ est vrai})$

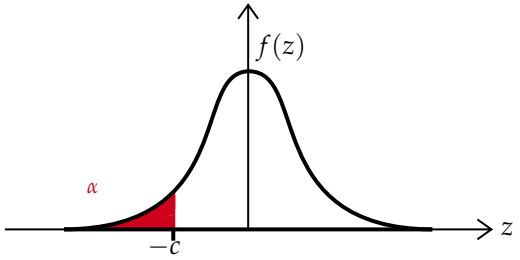
$\alpha = \Pr(|T| \geq c \mid H_0 \text{ est vrai})$



Unilatéral à la gauche

$p = \Pr(T \leq t \mid H_0 \text{ est vrai})$

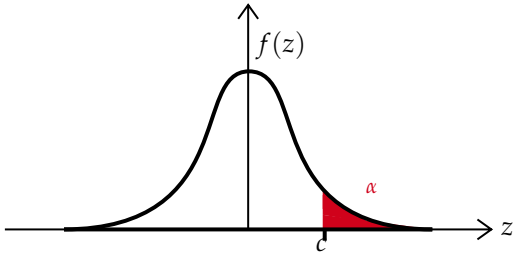
$\alpha = \Pr(T \leq -c \mid H_0 \text{ est vrai})$



Unilatéral à la droite

$p = \Pr(T \geq t \mid H_0 \text{ est vrai})$

$\alpha = \Pr(T \geq c \mid H_0 \text{ est vrai})$



Tests sur la moyenne

De façon générale, la statistique observée d'un test d'hypothèse sur la moyenne sera $t = \frac{\text{valeur estimée} - \text{valeur supposée}}{\text{erreur type}}$. Dans notre cas, l'erreur type correspond à l'écart-type de l'estimateur. Si la variance est connue, on utilise la ***vraie*** *erreur type* qui correspond à l'écart-type σ lui-même. Sinon, on utilise *l'erreur type* ***estimée*** qui correspond à l'estimation non biaisé de la variance s_n.

Note Voir la section *Erreur* du chapitre de *Propriétés* pour l'interprétation de l'erreur type en régression.

Pour tous les tests, on dénote la moyenne par μ et la variance par σ^2 même si les échantillons ne proviennent pas de distributions normales. Également, on dénote la valeur supposée par l'hypothèse nulle comme h.

Pour les tests sur la moyenne, on couvre 2 scénarios : *1 échantillon* et *2 échantillons*. Pour ce dernier, on distingue 2 cas : des échantillons indépendants ou des échantillons appariés (« *paired samples* »).

Rappels

Avant de détailler les tests, nous effectuons quelques rappels.

Premièrement, on généralise la convergence en probabilité de la section *Détails mathématiques sur la convergence* pour présenter la **convergence en distribution**.

Convergence en distribution

Notation

$\{Y_n\}$ Séquence de variables aléatoires indépendantes et identiquement distribuées.

Y Variable aléatoire comprise dans $\{Y_n\}$ avec moyenne μ et variance σ^2.

S_n La somme des n variables aléatoires, $S_n = Y_1 + Y_2 + \cdots + Y_n$.

On dit que S_n converge en distribution vers une distribution normale si

$$\lim_{n\to\infty} \Pr\left(\frac{S_n - \mathrm{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \leq z\right) = \Phi(z)$$

On dénote la convergence en distribution par : $S_n \xrightarrow{D} \mathcal{N}(\mathrm{E}[S_n], \mathrm{Var}(S_n))$.

La convergence en distribution est le théorème sous-jacent au théorème centrale limite (vue en ACT-1002 : analyse probabiliste des risques actuariels).

Rappel : théorème centrale limite

Notation

$\{X_n\}$ Séquence de variables aléatoires indépendantes et identiquement distribuées.

$\overline{X}_n$ La moyenne empirique des n variables aléatoires, $\overline{X}_n = \frac{S_n}{n}$.

On pose que $\overline{X}_n \xrightarrow{D} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$.

Le théorème s'applique directement pour *1 échantillon*. Pour *2 échantillons*, s'ils sont indépendants, on généralise pour trouver que la différence des moyennes empiriques $\overline{X}_1 - \overline{X}_2 \xrightarrow{D} \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$. S'ils sont appariés, on applique le théorème centrale limite à la différences des observations pour trouver que $\overline{D} \xrightarrow{D} \mathcal{N}\left(\mu_D, \sigma_D^2\right)$.

Lorsque la variance de l'échantillon aléatoire est connue, on applique le théorème centrale limite. Dans le cas où la variance est inconnue, nous avons recours à **la loi de Student** pour la distribution de la statistique de test.

Rappel : Loi de Student

La loi de Student se définit à partir d'une variable aléatoire normale et d'une variable aléatoire khi carré. Soit les 2 variabless aléatoires indépendantes $Z \sim \mathcal{N}(0,1)$ et $W \sim \chi^2_{(v)}$, alors $Y = \frac{Z}{\sqrt{W/v}} \sim t_{(v)}$.

Également, pour un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ tiré d'une distribution normale de moyenne μ et variance σ^2 inconnue, la statistique $T_n = \frac{\bar{X}_n - \mu}{\sqrt{S_n^2/n}} \sim t_{(n-1)}$.

La loi de Student tend vers la loi normale lorsque n est grand.

Le théorème centrale limite a l'avantage de s'appliquer ***peu importe*** la distribution de l'échantillon aléatoire au dépend de s'appliquer *juste lorsque n est grand*.
En revanche, la loi de Student s'applique ***juste*** **si** la distribution de l'échantillon aléatoire ***est normale*** mais ne *nécessite pas que n soit grand*.

1 échantillon

Pour un échantillon aléatoire de taille n,

variance	distribution de l'échantillon	n grand ?	valeur observée t de la statistique T	distribution de la statistique T
connue	n'importe quelle	oui	$\frac{\bar{x}_n - h}{\sigma/\sqrt{n}}$	$\mathcal{N}(0,1)$
inconnue	normale	non	$\frac{\bar{x}_n - h}{s_n/\sqrt{n}}$	t_{n-1}

Les valeur critiques sont :

Statistique T	unilatéral à gauche	bilatéral	unilatéral à droite
$\frac{\bar{X}_n - h}{\sigma/\sqrt{n}}$	$-z_{1-\alpha}$	$z_{1-\alpha/2}$	$z_{1-\alpha}$
$\frac{\bar{X}_n - h}{S_n/\sqrt{n}}$	$-t_{2\alpha,n-1}$	$t_{\alpha,n-1}$	$t_{2\alpha,n-1}$

Explication des percentile de loi normal vs de loi de Student

On dénote :

z_q $100q^{\text{e}}$ percentile de la loi normale standard.

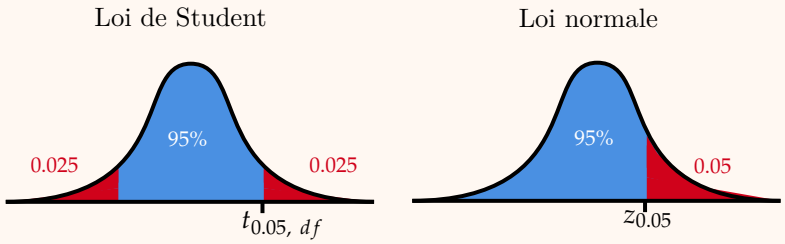
$t_{2(1-q),ddl}$ $100q^{\text{e}}$ percentile de la loi de student avec ddl degrés de liberté.

Ces deux percentiles correspondent à la même probabilité q. Ils sont notés différemment car les tables de l'examen les notent différemment :

La table de la loi *normale* standard comprend les **probabilités cumulatives** (e.g. $\Pr(Z \leq z)$).

La table de la loi de *Student* comprend les **probabilités des deux queues** (e.g. $\Pr(|T| \geq t)$).

Pour visualiser la différence, les valeurs des tables que l'on obtient pour $\alpha = 0.05$ sont :



2 échantillons

Contexte

Pour 1 échantillon, nous sommes habituellement intéressés à la *valeur* que prend la moyenne μ. Dans le cas où nous désirons comparer 2 échantillons **indépendants**, nous sommes plutôt intéressés aux *différences des valeurs* $\mu_1 - \mu_2$ et non les valeurs elles-mêmes.

Donc, au lieu de s'intéresser à tester si les moyennes sont égales, $\mu_1 = \mu_2$, on s'intéresse à tester si la différence entre les moyennes est nulle, $\mu_1 - \mu_2 = 0$. Bien que ça revient à la même chose mathématiquement, l'interprétation différente est importante. L'utilité du test devient donc de déterminer si deux échantillons proviennent de la même distribution.

Note La section sur la *Puissance d'un test* utilise ce scénario de deux échantillons. Les graphiques de la section aident à saisir l'idée de tester si deux échantillons proviennent de la même distribution.

Pour 2 échantillons aléatoires **indépendants** de tailles n_1 et n_2,

variances	distribution des échantillons	autres conditions ?	n_k grands ?	valeur observée t de la statistique T	distribution de la statistique T
connues	n'importe lesquelles	non	oui	$\frac{\bar{x}_1 - \bar{x}_2 - h}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	$\mathcal{N}(0,1)$
inconnues	normales	$\sigma_1^2 = \sigma_2^2$	non	$\frac{\bar{x}_1 - \bar{x}_2 - h}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	$t_{n_1+n_2-2}$

où s_p est la valeur observée du « pooled estimator » S_p de l'écart-type.

Contexte

Il peut y arriver que les 2 échantillons ne sont pas indépendants, mais plutôt **appariés**. Par exemple, les expériences pharmaceutiques ont un groupe de contrôle et un groupe expérimental et un groupe de contrôle qui sont appariés.

Pour 2 échantillons aléatoires **appariés** de taille n_*, où $n_1 = n_2 = n_*$, on définit les tests en fonction des différences D_i des paires de X.

Différence D_i de la i^e paire d'observations de X

La différence D_i de la i^e paire de X a une moyenne μ_D et une variance $\sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2$.

Si les variances sont inconnues, alors S_D^2 est la variance échantillonnale des différences.

Donc,

variance	distribution des différences	n_* grand ?	valeur observée t de la statistique T	distribution de la statistique T
connue	n'importe quelle	oui	$\frac{\bar{d}_n - h}{\sigma_D/\sqrt{n_*}}$	$\mathcal{N}(0,1)$
inconnue	normale*	non	$\frac{\bar{d}_n - h}{s_D/\sqrt{n_*}}$	t_{n_*-1}

Note Techniquement, le fait que les échantillons sont normalement distribués ne garanti pas que leurs différences D_i le sera. Il faudrait, par exemple, que les échantillons suivent une distribution normale bivariée. Cependant, l'examen n'a pas considéré cette distinction jusqu'à date et donc exiger que les échantillons sont normalement distribués est un critère que l'on considère adéquat.

Tests sur les proportions

Contexte

Le paramètre p d'une distribution Bernoulli est d'intérêt particulier car il représente la *proportion* d'une population qui est considérée un « succès ». Bien que p correspond à la moyenne d'une distribution Bernoulli, il y a quelques particularités qui distinguent le test sur une proportion.

La distinction principale revient à la variance, $p(1-p)$, d'une distribution Bernoulli. Puisqu'elle est fonction du paramètre p inconnu, il s'ensuit que la **variance est inconnue**. Auparavant, les tests sur la moyenne pour lesquels la variance est inconnue ont étés restreints à des échantillons tirés d'une distribution normale afin d'utiliser la loi de Student. Cependant, puisque l'échantillon est tiré de la distribution (discrète) de Bernoulli, on ne peut pas appliquer la loi de Student.

1 échantillon

Pour n qui est grand, le théorème centrale limite implique que $\bar{X}_n \xrightarrow{D} \mathcal{N}\left(q, \frac{q(1-q)}{n}\right)$.
Pour le cas d'une proportion, il est courant de dénoter la valeur observée de la moyenne empirique comme : $\bar{x}_n = \hat{q}$. Il s'ensuit que $t = \frac{\hat{q}-h}{\sqrt{\frac{h(1-h)}{n}}}$ où $T \sim \mathcal{N}(0,1)$.

2 échantillons

Pour n_1 et n_2 qui sont grands, le théorème centrale limite implique que $\bar{X}_1 - \bar{X}_2 \xrightarrow{D} \mathcal{N}\left(q_1 - q_2, \frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}\right)$. Il s'ensuit que $t = \frac{\hat{q}_1 - \hat{q}_2 - h}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}}$

où $T \sim \mathcal{N}(0,1)$.

Note Dans le cas où nous avons 1 échantillon, il faut juste connaître q pour connaître la variance. Dans le cas où nous avons 2 échantillons, il faut connaître q_1 et q_2 individuellement alors que nous connaissons uniquement la différence h. C'est pourquoi l'erreur type de la statistique diffère pour les scénarios.

Tests sur la variance

Rappels

Comme pour les test sur la moyenne, nous effectuons quelques rappels avant de détailler les tests.

Premièrement, on rappel la loi du khi carré que l'on utilise pour trouver une distribution à la statistique de test avec 1 échantillon.

Rappel : Loi du khi carré

Loi loi du khi carré peut être définie de plusieurs façons. En particulier, pour v variables aléatoires normales standards $Z_1, Z_2, \ldots, Z_v$, $\sum_{i=1}^{v} Z_i^2 \sim \chi^2_{(v)}$.

Également, pour un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ tirée d'une distribution normale de moyenne μ et variance σ^2, la statistique $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

La loi du khi carré a quelques propriétés qui la rendent intéressante. Contrairement à la loi normale, la loi du khi carré est :

1. non négative.
2. asymétrique vers la droite.

Puis, on rappel la loi de Fisher-Snedecor pour trouver une distribution à la statistique de test avec 2 échantillons.

Rappel : Loi de Fisher-Snedecor (F)

La loi de Fisher se définit à partir de variables aléatoire qui suivent la loi du khi carré.

Soit les 2 variables aléatoires indépendantes $W_1 \sim \chi^2_{(v_1)}$ et $W_2 \sim \chi^2_{(v_2)}$, alors $Y = \frac{W_1/v_1}{W_2/v_2} \sim \mathcal{F}_{(v_1,v_2)}$. De plus, la loi de Fisher a la propriété intéressante que $Y^{-1} \sim \mathcal{F}_{(v_2,v_1)}$.

De cette relation, on peut également relier la loi de Student à la loi de Fisher avec $Y = \frac{Z^2}{W/v} \sim \mathcal{F}_{(1,v)}$, car $W \sim \chi^2_{(v)}$ et $Z^2 \sim \chi^2_{(1)}$ où $Z \sim \mathcal{N}(0,1)$.

Également, pour deux échantillons aléatoires indépendantes de tailles n_1 et n_2 qui sont tirés de distributions normales avec variances σ_1^2 et σ_2^2, la statistique $T_n = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathcal{F}_{(n_1-1,n_2-1)}$.

La loi de Fisher a quelques propriétés qui la rendent intéressante. Contrairement à la loi normale mais comme la loi du khi carré, la loi de Fisher est :

1. non négative.
2. asymétrique vers la droite.

1 échantillon

Pour un échantillon aléatoire de taille n tiré d'une **distribution normale** de variance σ^2, la loi du khi carré implique que $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$. Il s'ensuit que $t = \frac{(n-1)s_n^2}{h}$.

Les régions critiques sont :

Région critique	unilatéral à gauche	bilatéral	unilatéral à droite
$\mathcal{C}$	$t \leq \chi^2_{\alpha,n-1}$	$\left\{t \leq \chi^2_{\alpha/2,n-1}\right\} \cup \left\{t \leq \chi^2_{1-\alpha/2,n-1}\right\}$	$t \geq \chi^2_{1-\alpha,n-1}$

Note Puisque la distribution khi carré est asymétrique, on ne peut pas simplifier d'avantage les régions critique avec des valeurs absolues comme avec la loi normale.

Percentiles de la loi du khi carré

On dénote :

$\chi^2_{q,ddl}$ Le $100q^{\text{e}}$ percentile de la loi du khi carré avec ddl degrés de liberté.

Comme la table de la loi normale, la table de la loi du khi carré comprend les **probabilités cumulatives** (e.g. $\Pr(W \leq w)$).

2 échantillons

Contexte

Pour 1 échantillon, nous sommes habituellement intéressés à la *valeur* que prend la variance σ^2. Dans le cas où nous désirons comparer 2 échantillons (indépendants), nous sommes plutôt intéressés au *ratio des variance* $\frac{\sigma_1^2}{\sigma_2^2}$ et non à la différence ni les valeurs.

Pour 2 échantillons aléatoires indépendants de tailles n_1 et n_2 qui sont normalement distribués avec variances σ_1^2 et σ_2^2, la loi de Fisher implique que $T_n = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1,n_2-1)}$. Il s'ensuit que $t = \frac{s_1^2}{s_2^2} \times \frac{1}{h}$.

Les régions critiques et l'hypothèse alternative pour $\text{H}_0 : \frac{\sigma_1^2}{\sigma_2^2} = h$ sont :

	unilatéral à gauche	unilatéral à droite
H_a	$\frac{\sigma_1^2}{\sigma_2^2} < h$	$\frac{\sigma_1^2}{\sigma_2^2} > h$
$\mathcal{C}$	$t \leq F_{1-\alpha,n_1-1,n_2-1}$	$t \geq F_{\alpha,n_1-1,n_2-1}$

	bilatéral
H_a	$\frac{\sigma_1^2}{\sigma_2^2} \neq h$
$\mathcal{C}$	$\left\{t \leq \left(F_{\alpha/2,n_2-1,n_1-1}\right)^{-1}\right\} \cup \left\{t \geq F_{\alpha/2,n_1-1,n_2-1}\right\}$

Note La valeur critique du test est $h \times F$. C'est-à-dire, on multiplie le quantile par la valeur h où souvent $h = 1$.

Percentiles de la loi de Fisher

On dénote :

$F_{q,ddl_{\textbf{num}},ddl_{\textbf{dén}}}$ Le $100q^{\text{e}}$ percentile de la loi de Fisher avec ddl_{num} degrés de liberté au numérateur et $ddl_{\text{dén}}$ degrés de liberté au dénominateur.

La table de la loi Fisher est la seule qui comprend des **probabilités de *survie*** (e.g. $\Pr(W > w)$).

Puissance d'un test

La puissance d'un test

La probabilité de *correctement* rejeter l'hypothèse nulle : $\Pr((X_1,\ldots,X_n) \in \mathcal{C} | \theta \in \Theta_1)$.

Une analyse de la puissance détermine le nombre d'observations qu'il faut afin d'avoir une probabilité élevée de correctement rejeter l'hypothèse nulle.

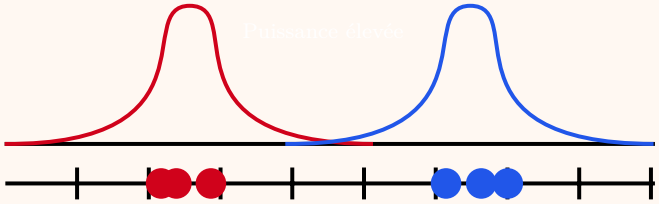
Facteurs influençant la puissance

Plusieurs facteurs influencent la puissance d'un test. Afin de les visualiser, on teste si deux échantillons d'observations proviennent de la même distribution.

1 La forme de la distribution

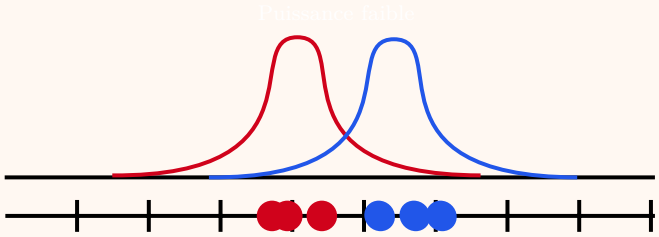
Si les deux distributions sont :

Très **distinctes**, la puissance sera très **élevée** :



- La probabilité de **correctement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;
- On peut aussi dire qu'il y a une forte probabilité de **correctement** obtenir une faible valeur p.

Se **chevauchent**, la puissance sera **faible** :

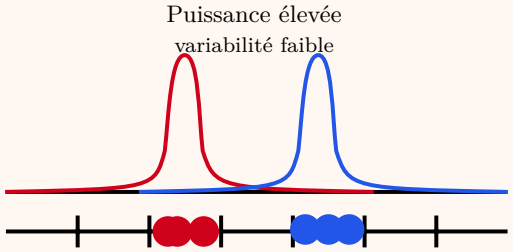


- La probabilité **d'incorrectement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;
- On peut aussi dire qu'il y a une forte probabilité **d'incorrectement** obtenir une faible valeur p ;
- Cependant, la puissance peut être augmentée avec plus d'observations.

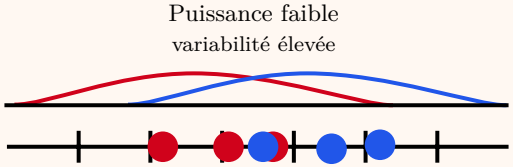
2 La variabilité des données

Si la variabilité de la distribution est

Faible, alors la variabilité de l'échantillon sera probablement faible aussi menant à une puissance très **élevée** :



Élevée, alors la variabilité de l'échantillon sera probablement élevée aussi menant à une puissance **faible** :



Il existe plusieurs mesures qui permettent de considérer la variabilité des données ainsi que la forme de la distribution. Entres autre, il y a le « ***effect size (d)*** » où $d = \frac{\bar{x}-\bar{y}}{s_p^2}$.

3 La taille de l'échantillon de données

Un grand échantillon de données peut compenser pour des distributions qui se chevauchent ou une variabilité élevée. Ça permet d'augmenter notre *confiance* qu'il y a bel et bien une différence entre les échantillons.

En contraste, nous n'avons pas besoin d'un grand échantillon de données pour des distributions très distinctes ou avec une faible variabilité ; nous sommes déjà confiants que les distributions sont différentes.

4 Le test statistique

Certains tests ont une puissance plus élevée que les autres. Cela dit, le test t habituel est très puissant.

La fonction de puissance

Contexte

La puissance est utile mais limitée à ce qu'il n'y ai qu'une seule hypothèse alternative. Il s'avère donc utile de définir la fonction de puissance qui permet de poser quelle hypothèse est la vraie.

La fonction de puissance permet donc d'analyser les valeurs possibles du paramètre. Par exemple, on pourrait tracer la fonction de puissance pour toutes les valeurs possibles de l'ensemble Θ_1.

Fonction de puissance

La fonction de puissance correspond à la probabilité de rejeter l'hypothèse nulle H_0 si la **vraie** valeur du paramètre est $\theta \in \Theta$: $\gamma(\theta) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C} | \theta\}$.

Ceci généralise la puissance en posant que la *vraie* hypothèse peut être la nulle ou tout autre hypothèse alternative s'il y en a une plusieurs. Bref, la fonction de puissance est une fonction de θ.

Idéalement, avec 2 hypothèses, si l'hypothèse nulle est

acceptée on souhaite que $\Pr\{(X_1, \ldots, X_n) \in \mathcal{C} | \theta \in \Theta_0\} = \gamma(\theta_0) = 0$.

rejetée on souhaite que $\Pr\{(X_1, \ldots, X_n) \in \mathcal{C} | \theta \in \Theta_1\} = \gamma(\theta_1) = 1$.

Note La puissance correspond à $\gamma(\theta_1)$.

Tests optimaux (les plus puissants)

Introduction

Notation

δ (Procédure de) test ;

$\alpha(\delta)$ Probabilité d'une erreur de type I, c'est-à-dire de incorrectement rejeter l'hypothèse nulle, pour un test δ.

$$\alpha(\delta) = \Pr\{(X_1, \ldots, X_n) \in \mathcal{C} | \theta \in \Theta_0\} = \gamma(\theta_0)$$

$\beta(\delta)$ Probabilité d'une erreur de type II, c'est-à-dire de incorrectement accepter l'hypothèse nulle, pour un test δ.

$$\beta(\delta) = \Pr\left\{(X_1, \ldots, X_n) \in \mathcal{C}^{\complement} | \theta \in \Theta_1\right\} = 1 - \gamma(\theta_1)$$

Λ **Ratio de vraisemblance**.

Pour mettre en contexte cette notation, voici le tableau des types d'erreurs pour un test δ repris de celui de la section des *Erreurs de test* :

	Vrai état	
Décision	$H_0 \Rightarrow \theta \in \Theta_0$	$H_1 \Rightarrow \theta \in \Theta_1$
Rejeter H_0 $(X_1, \ldots, X_n) \in \mathcal{C}$	$\alpha(\delta)$	$1 - \beta(\delta)$
Accepter H_0 $(X_1, \ldots, X_n) \in \mathcal{C}^{\complement}$	$1 - \alpha(\delta)$	$\beta(\delta)$

Bien qu'*en théorie* on minimise la probabilité d'une erreur de type I **et** de II, *en pratique* il y a un compromis entre les deux. On ne peut pas minimiser les deux erreurs. Selon le contexte, on détermine laquelle que l'on souhaite minimiser le plus.

Par exemple, soit l'hypothèse nulle que quelqu'un n'a pas le cancer. Il est plus *grave* de dire à quelqu'un réellement atteint du cancer qu'il n'a pas le cancer (erreur de type II) que de dire à quelqu'un qui n'est pas réellement atteint du cancer qu'il a le cancer (erreur de type I). Dans ce contexte, on souhaiterait minimiser l'erreur de type II $\beta(\delta)$ plus que l'erreur de type I $\alpha(\delta)$.

Puisqu'il est impossible de trouver un test δ pour lequel les probabilités d'erreurs de type I et II sont très petites, on :

1. Fixe l'erreur de type I à un seuil, alias une taille de région critique, α.
2. Trouve, parmi tous les régions (sous-ensembles) de taille α, la région de valeurs qui minimise l'erreur de type II.

Test le plus puissant

Tests optimaux, alias tests les plus puissants

Le test le plus puissant est le test, parmi tous les tests dont la taille de la région critique est de α et que les hypothèses sont simples, qui a la *meilleur région critique*. Le test qui a la *meilleur région critique* est le test qui a la **plus grande puissance**.

Pour trouver ce test optimal dénoté δ^*, on débute par poser deux conditions :

1. Les hypothèses doivent êtres simples :
   $H_0 : \theta = \theta_0$
   $H_1 : \theta = \theta_1$
2. La région, alias le sous-ensembles de $\mathcal{S}$, doit être de taille α. En autres mots, la probabilité de incorrectement rejeter l'hypothèse nulle (erreur de type I), $\alpha(\delta^*)$, est de α :
   $\alpha(\delta^*) = \Pr\left((X_1, \ldots, X_n) \in \mathcal{C} | \theta = \theta_0\right) = \alpha$

Avec ces deux conditions, on identifie toutes les régions, dénotées $\mathcal{A}$, de taille α qui pourraient être ***la*** région critique $\mathcal{C}$.

Puis, pour trouver la région critique *unique*, on pose que la probabilité que l'échantillon aléatoire soit contenu dans ***la*** région critique $\mathcal{C}$, sachant que l'hypothèse alternative est vraie, est supérieure à la probabilité que l'échantillon aléatoire soit contenu dans *tout autre sous-ensemble* $\mathcal{A}$:

$$\Pr\left((X_1, \ldots, X_n) \in \mathcal{C} | \theta = \theta_1\right) \geq \Pr\left((X_1, \ldots, X_n) \in \mathcal{A} | \theta = \theta_1\right)$$

Bref, on prend la région critique qui a la plus grande puissance.

Avec ces critères, on trouve ***la*** région critique $\mathcal{C}$ de taille α **optimale** pour tester les hypothèses simples. Le test qui y correspond est le **test le plus puissant**.

En bref, on pose la fonction de puissance en posant que l'hypothèse nulle est vraie fixe à un seuil α, puis on trouve la région critique qui maximise la fonction de puissance (qui pose que l'hypothèse alternative est vraie).

Exemple avec une distribution binomiale

Pour la variable aléatoire $X \sim Binom(n = 3, p = \theta)$, on fixe les hypothèses suivantes :

$H_0 : \theta = 0.50$
$H_1 : \theta = 0.75$

On souhaite identifier le test le plus puissant de taille $\alpha = 0.125$.

1. La première étape est d'identifier les régions de valeurs pour lesquels la variable aléatoire a une probabilité de 0.125 d'y être contenue.
   (a) On fait un tableau des valeurs sous les deux hypothèses :

FMP	$x = 0$	$x = 1$	$x = 2$	$x = 3$
$p(x\|\theta = 0.50)$	0.125	0.375	0.375	0.125
$p(x\|\theta = 0.75)$	0.01 5625	0.140 625	0.421 875	0.421 875

   (b) On trouve que les sous-ensembles de $\mathcal{S}$ correspondants sont $\mathcal{A}_1 = \{x = 0\}$ et $\mathcal{A}_2 = \{x = 3\}$.
   C'est-à-dire, $\Pr(X \in \mathcal{A}_1 | \theta = 0.50) = \Pr(X \in \mathcal{A}_2 | \theta = 0.50) = 0.125$ et il n'y a pas d'autres sous-ensembles de $\mathcal{S}$ avec la même "taille" de 0.125.
2. On doit identifier laquelle de $\mathcal{A}_1$ ou $\mathcal{A}_2$ est la région critique $\mathcal{C}$ optimale de taille α pour tester H_0 contre H_1.
3. On trouve la probabilité de correctement rejeter l'hypothèse nulle (faire partie de la région critique) pour les deux régions :
   $\Pr(X \in \mathcal{A}_1 | \theta = 0.75) = \Pr(X = 0 | \theta = 0.75) = 0.015\,625$
   $\Pr(X \in \mathcal{A}_2 | \theta = 0.75) = \Pr(X = 3 | \theta = 0.75) = 0.421\,875$
4. On compare les probabilités de correctement rejeter l'hypothèse nulle aux probabilités de incorrectement rejeter l'hypothèse nulle :
   Dans le premier cas :
   $\underbrace{\Pr(X \in \mathcal{A}_1 | \theta = 0.75)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} = 0.015625 \quad < \quad \underbrace{\Pr(X \in \mathcal{A}_1 | \theta = 0.50)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)} = 0.125$
   Dans le deuxième cas :
   $\underbrace{\Pr(X \in \mathcal{A}_2 | \theta = 0.75)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} = 0.421875 \quad > \quad \underbrace{\Pr(X \in \mathcal{A}_2 | \theta = 0.50)}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)} = 0.125$
5. Puisque le premier sous-ensemble a une probabilité plus élevée de incorrectement rejeter l'hypothèse nulle (erreur de type I) que de correctement la rejeter, on choisit le deuxième comme région critique : $\mathcal{C} = \mathcal{A}_2 = \{x = 3\}$.

Également, on peut observer que la région est choisie en incluant dans $\mathcal{C}$ les points x pour lesquels $p(x|\theta = 0.50)$ est petite par rapport à $p(x;\theta = 0.75)$. Il s'ensuit que le ratio $\frac{p(x|\theta=0.50)}{p(x;\theta=0.75)}$ évalué à $x = 5$ **est un minimum**.

Cet exemple mène au prochain théorème de Neyman-Pearson qui est une méthode plus efficace d'identifier le test le plus puissant. Grâce au théorème, on peut utiliser le ratio des fonctions de densité (vraisemblance) comme outil pour identifier la région critique $\mathcal{C}$ optimale à un seuil fixe de α.

Théorème de Neyman-Pearson

Le théorème de Neyman-Pearson permet de trouver le test le plus puissant. Soit un test δ^* avec les hypothèses simple :

$\mathrm{H}_0 : \theta = \theta_0$

$\mathrm{H}_1 : \theta = \theta_1$

Note Il est important que les hypothèses soient simples afin qu'elles **spécifient complètement la distribution**.

Soit une constante $k > 0$ et la région critique $\mathcal{C}$ de taille α tel que :

1. si l'échantillon aléatoire est contenu dans la région critique, $x \in \mathcal{C}$, alors $\frac{\mathcal{L}(\theta_0;x)}{\mathcal{L}(\theta_1;x)} \leq k$.
2. La taille $\alpha = \Pr\{(X_1,\ldots,X_n) \in \mathcal{C}|\theta_1\} = \alpha(\delta^*)$.

Alors $\mathcal{C}$ est ***la*** région critique **optimale** de taille α.

Note Réécrire la première égalité sous la forme de $\mathcal{L}(\theta_0;\boldsymbol{x}) \leq k\mathcal{L}(\theta_1;\boldsymbol{x})$ mène à l'interprétation qu'il doit être plus *vraisemblable* que l'échantillon soit distribué selon l'hypothèse alternative ($\theta = \theta_1$) que l'hypothèse nulle ($\theta = \theta_0$) lorsqu'il est contenu dans la région critique ($\boldsymbol{x} \in \mathcal{C}$) puisque que l'on rejette l'hypothèse nulle H_0. On peut dire que les données semblent *favoriser* l'hypothèse alternative.

Pour appliquer le *lemme de Neyman-Pearson* l'approche est typiquement d'écrire le ratio, puis de trouver une statistique permettant de calculer une probabilité avec sa distribution. Voici quelques exemples qui clarifient cette notion :

Exemple cas continu

Pour l'échantillon aléatoire $\boldsymbol{X} = (X_1,\ldots,X_n)$ tiré d'une distribution normale $\mathcal{N}(\mu = \theta, \sigma^2 = 1)$, on fixe les hypothèses suivantes :

$\mathrm{H}_0 : \theta = 0$

$\mathrm{H}_1 : \theta = 1$

On souhaite identifier le test le plus puissant pour un seuil fixé α.

1 La première étape est d'identifier le ratio des vraisemblances et d'essayer d'y trouver une statistique.

$$\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} = \frac{\exp\left\{-\sum_{i=1}^n x_i^2/2\right\}\frac{1}{(\sqrt{2\pi})^n}}{\exp\left\{-\sum_{i=1}^n (x_i-1)^2/2\right\}\frac{1}{(\sqrt{2\pi})^n}} = \exp\left\{-\sum_{i=1}^n x_i + \frac{n}{2}\right\}$$

2 Puis, avec le ratio des fonctions de vraisemblance, on trouve la région critique $\mathcal{C}$ avec la première égalité du théorème de Neyman-Pearson.

$$\mathrm{e}^{-\sum_{i=1}^n x_i + \frac{n}{2}} \leq k \quad \Rightarrow \quad -\sum_{i=1}^n x_i + \frac{n}{2} \leq \ln(k) \quad \Rightarrow \quad \sum_{i=1}^n x_i \geq \frac{n}{2} - \ln(k)$$

$$\therefore \frac{\sum_{i=1}^n x_i}{n} \geq \underbrace{\frac{1}{2} - \frac{\ln(k)}{n}}_{c}$$

Alors, la région critique optimale $\mathcal{C} = \left\{(x_1, x_2, \ldots, x_n) : \frac{1}{n}\sum_{i=1}^n x_i \geq c\right\}$ où c est une constante choisie telle que la taille de $\mathcal{C}$ est α.

Par exemple, sous l'hypothèse nulle $\bar{X} \overset{\mathrm{H}_0}{\sim} \mathcal{N}(0, 1/n)$,

on peut isoler c avec $\Pr(\bar{X} \geq c|\theta = \theta_0) = \alpha$.

on peut calculer la puissance du test avec $\Pr(\bar{X} \geq c|\theta = \theta_1)$.

Exemple cas discret

Pour la variable aléatoire discrète X, on fixe les hypothèses suivantes :

$\mathrm{H}_0 : \theta = 1$

$\mathrm{H}_1 : \theta = 2$

On souhaite identifier meilleur région critique pour les tests de taille $\alpha = 0.06$. Cependant, puisque la variable aléatoire est discrète, on ne peut pas manipuler le ratio des vraisemblances. Plutôt :

1 On produit le tableau des valeurs de la fonction de masse des probabilités

sous les deux hypothèses :

FMP	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$
$p(x\|\theta=1)$	0.01	0.05	0.50	0.43	0.01
$p(x\|\theta=2)$	0.02	0.24	0.25	0.25	0.24

2 On trouve les ratios dénotés Λ :

FMP	$x=1$	$x=2$	$x=3$	$x=4$	$x=5$
Λ	0.50	0.208	2	1.72	0.042

3 Par défaut, on suppose que les données sont distribuées selon l'hypothèse nulle. Dans le cas contraire, la vraisemblance sous l'hypothèse nulle sera grande et le ratio des vraisemblances faible. Donc, on cherche des petites valeurs du ratio. Puis, on trie les valeurs de x en ordre croissante du ratio des vraisemblances :

FMP	$x=5$	$x=2$	$x=1$	$x=4$	$x=3$
Λ	0.042	0.208	0.5	1.72	2

4 De ce tableau, on voit que la meilleur région critique de taille 0.06 est $C = \{x = 5, 2\}$ car $p(x=5|\theta=1) + p(x=2|\theta=1) = 0.01 + 0.05 = 0.06$.

Test uniformément le plus puissant

Motivation

Bien que le théorème de Neyman-Pearson est très utile, le fait qu'il est seulement applicable pour les hypothèses simples est une restriction importante puisque la majorité des tests d'hypothèses ont des hypothèses composées.

Bien qu'on ne peut pas l'utiliser directement pour des hypothèses composées, le théorème de Neyman-Pearson peut servir d'outil pour identifier le *test uniformément le plus puissant*. La situation typique est que l'hypothèse nulle est simple alors que l'hypothèse alternative est composée. L'astuce pour appliquer le théorème de Neyman-Pearson aux hypothèses composées est de voir une hypothèse composée comme un regroupement d'hypothèses simples (e.g. $\theta > 2$ implique $\theta = 3$, $\theta = 4$, $\theta = 5$, etc.)

On cherche donc la meilleur région critique pour tester l'hypothèse nulle à toutes les hypothèses alternatives simples contenues dans l'hypothèse alternative composée. Bref, le test uniformément le plus puissant a la meilleur région critique **pour toutes les combinaisons possibles d'hypothèses simples contenues dans l'hypothèse alternative composée**.

Test uniformément le plus puissant

Un test δ^* est le test de l'hypothèse simple H_0 contre l'hypothèse composée H_1 **uniformément le plus puissant** de taille α s'il a la plus grande puissance, $\gamma(\theta \in \Theta_1|\delta^*)$, parmi tous les tests δ de taille α, **pour toutes les valeurs possibles de l'hypothèse alternative** ($\theta \in \Theta_1$).

En termes mathématiques, c'est le test δ^* tel que $\gamma(\theta \in \Theta_1|\delta) \leq \gamma(\theta \in \Theta_1|\delta^*)$ où $\sup_{\theta\in\Theta_0} \gamma(\theta|\delta^*) \leq \alpha$ et δ est tout autre test tel que $\sup_{\theta\in\Theta_0} \gamma(\theta|\delta) \leq \alpha$.

En anglais, c'est le test « *uniformly most powerful (UMP)* ».

La procédure pour trouver le test *uniformément le plus puissant* est de poser un θ_1 fixe afin d'évaluer la forme du ratio de la vraisemblance. Selon la forme de l'hypothèse et la croissance, ou décroissance, du ratio Λ, on peut établir une région critique valide pour toutes les hypothèse alternatives simples contenues dans l'hypothèse alternative composée.

Exemple avec une distribution normale

Pour l'échantillon aléatoire $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ tiré d'une distribution normale $\mathcal{N}(0, \theta)$, on fixe les hypothèses suivantes :

$\mathrm{H}_0 : \theta = 1$

$\mathrm{H}_1 : \theta > 1$

1. On trouve le ratio Λ :

$$\Lambda = \frac{\mathcal{L}(\theta_0 = 1; \boldsymbol{x})}{\mathcal{L}(\theta_1; \boldsymbol{x})} = \frac{\frac{1}{(1)^n(\sqrt{2\pi})^n} \mathrm{e}^{-\frac{\sum_{i=1}^n x_i^2}{2(1)^2}}}{\frac{1}{\theta_1^n(\sqrt{2\pi})^n} \mathrm{e}^{-\frac{\sum_{i=1}^n x_i^2}{2\theta_1^2}}} = \theta_1^n \mathrm{e}^{-\frac{\sum_{i=1}^n x_i^2}{2}\left(1-\frac{1}{\theta_1^2}\right)}$$

2. On observe que le ratio Λ décroit alors que $\sum x_i^2$ augmente ce qui confirme que l'on peut isoler une région critique.
3. Puisque le ratio Λ est décroissant, un test *uniformément* le plus puissant aura une région critique définie par $\sum x_i^2 \geq k$, où k est choisi selon la taille α.

$$\theta_1^n \mathrm{e}^{-\frac{\sum_{i=1}^n x_i^2}{2}\left(1-\frac{1}{\theta_1^2}\right)} \leq c \qquad \Rightarrow \qquad \sum_{i=1}^n x_i^2 \geq -2\frac{\ln\left(\frac{k}{\theta_1^n}\right)}{1-\theta_1^{-2}}$$

Note La région qui correspond au test uniformément le plus puissant **n'existe pas toujours**. Si, par exemple, l'hypothèse alternative aurait été $\theta \neq 1$ alors la fonction de vraisemblance pourrait être décroissante pour $\theta < 1$ et la région critique ne serait pas unique.

L'approche prise dans l'exemple ci-dessus est de poser un $\theta_0 =$ un nombre, puis de poser une constante $\theta_1 > \theta_0$. Cependant, on peut généraliser l'approche au cas où les deux hypothèses sont composées (e.g. $\mathrm{H}_0 : \theta \leq 2$ et $\mathrm{H}_1 : \theta > 2$), on définit le **ratio de vraisemblance monotone**.

Ratio de vraisemblance monotone

Soit les constantes a et b tel que $a < b$ qui sont des valeurs possibles de θ (alias, $a, b \in \Theta$). Alors, on définit $\Lambda = \frac{\mathcal{L}(a)}{\mathcal{L}(b)}$.

Pour qu'un test ayant des hypothèses composées soit uniformément le plus puissant, le ratio de vraisemblance doit être monotone en fonction de la statistique y obtenue du ratio. La monotonicité assure que la relation de décroissance se maintient peut importe les valeurs prises par θ en dedans de son domain (e.g. pour tout $\theta_0 \leq 2$ et tout $\theta_1 > 2$).

On résume la région critique la taille du test selon le ratio de vraisemblance pour le test $\mathrm{H}_0 : \theta = h$:

Ratio de vraisemblance monotone	Région critique	α
décroissant	$y \geq c$	$\Pr(Y \geq c \mid \theta = h)$
croissant	$y \leq c$	$\Pr(Y \leq c \mid \theta = h)$

Note On voit donc qu'en posant $\theta = \theta_0 =$ un nombre, cette approche est équivalente à celle de l'exemple ci-dessus.

Note De façon générale, on peut insérer l'EMV dans le ratio de vraisemblance pour un test bilatéral.

Tests d'adéquation

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Notation

$F^*()$ Fonction de répartition d'une v.a. continue (hypothèse nulle).

$\hat{F}()$ Fonction de répartition empirique.

Contexte

L'objectif sous-jacent de cette section est d'évaluer la qualité de l'ajustement d'une distribution d'un échantillon de données. Jusqu'à présent, nous avons présenté les test d'hypothèses qui évaluent la valeur des paramètres d'une distribution. Cependant, le paramètre d'une distribution n'est qu'une partie de l'ajustement.

Cette section détaille plusieurs tests qui servent à évaluer la qualité, ou ***adéquation***, de l'ajustement au-delà des paramètres.

Test de Kolmogorov-Smirnov

Motivation

Le premier volet d'ajustement que l'on évalue est la fonction de répartition alias, la fonction de *distribution*. Le test de Kolmogorov-Smirnov compare la fonction de répartition *empirique* à la fonction de répartition *théorique* d'une distribution hypothétique.

Il est intuitif de visuellement évaluer l'ajustement des données, mais cette évaluation est très *subjective*. La « raison d'être » du test est de quantifier cette évaluation subjective afin d'obtenir une mesure **quantitative**.

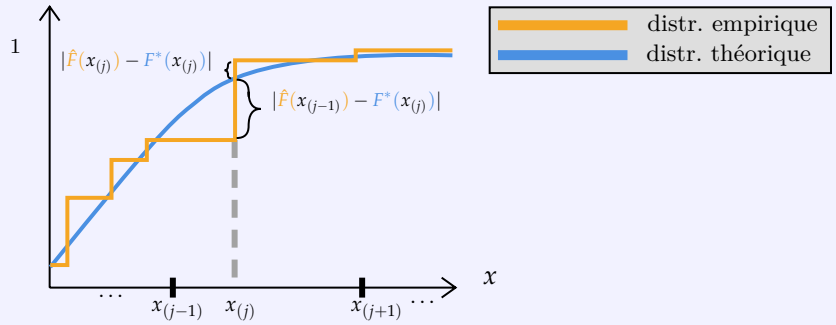
Test de Kolmogorov-Smirnov

Le test de Kolmogorov-Smirnov teste si les données semblent suivre une distribution avec **la statistique de Kolmogorov-Smirnov** $D = \max_j D_j$ où

$$D_j = \max_j \left\{ \left| \hat{F}(x_{(j-1)}) - F^*(x_{(j)}) \right|, \left| \hat{F}(x_{(j)}) - F^*(x_{(j)}) \right| \right\} \text{ avec } \hat{F}(x_{(0)}) = 0.$$

Ces équations reviennent donc à calculer **la différence maximale** entre la fonction de répartition empirique et théorique. Chacune des différences prend le maximum de l'écart entre la distribution théorique et la Il s'ensuit que si les données sont bien ajustées, on s'attend à ce que D soit très petit.

Visuellement :



Note Lorsque la distribution est entièrement spécifiée (aucun paramètre n'est estimé), une table avec les valeurs critiques est donnée. Cependant, il n'est pas nécessaire de l'apprendre car elle sera donnée en examen si elle est nécessaire.

Note La section de *Tests pour la qualité de l'ajustement* du chapitre de *Erreur* couvre l'application de ce test dans le contexte de données incomplètes.

Test d'adéquation du khi carré (« *Chi-Square Goodness-of-Fit Test* »)

Motivation

Le test de Kolmogorov-Smirnov a la limitation inhérente qu'il mesure seulement la plus grande divergence entre la distribution empirique et théorique. Il s'ensuit que nous désirons évaluer la similarité des distributions sur l'ensemble du domaine. Pour ce faire, on utilise le test d'adéquation du khi carré.

Contrairement au test de Kolmogorov-Smirnov, le test d'adéquation du khi carré **est un test d'hypothèse**.

Test d'adéquation du khi carré

Le test débute par définir k intervalles (distincts) des données dans lesquels les n données sont réparties.

Puis, pour $j \in \{1, \dots, k\}$, on défini :

Notation

q_j La probabilité de la distribution hypothétique d'être contenu dans l'intervalle j.

n_j Le nombre d'observations de l'échantillon de données contenues dans l'intervalle j.

L'hypothèse du test est que les espérances théoriques du nombre d'observations par intervalle (nq_j) vont être égaux aux nombres observés d'observations par intervalle (n_j). Donc, pour tout $j \in \{1, \dots, k\}$,

$$\mathrm{H}_0 : nq_j = n_j \qquad \mathrm{H}_1 : nq_j \neq n_j$$

La statistique du test calcule les divergences pour chacun des intervalles : $t = \sum_{j=1}^{k} \frac{(n_j - nq_j)^2}{nq_j}$ où la statistique est approximativement distribué selon la loi du khi carré : $T_n \sim \chi^2_{(k-1-r)}$.

Note Le $k-1$ des degrés de liberté est dut au fait que nous avons k données empiriques sous la contrainte que $\sum_{j=1}^{n} n_j = n$. Donc, il y a seulement $k-1$ données estimées. Le r correspond au nombre de paramètres estimés de la distribution théorique. Par exemple, pour une distribution normale avec $\sigma = 2$ et μ inconnu, $r = 1$. La région critique est $t \geq \chi^2_{(1-\alpha,k-1-r)}$.

Test de l'indépendance du khi carré (tableau de contingence)

Motivation

Le test de l'indépendance du khi carré est pour les données pouvant être représentées sous la forme d'un **tableau de contingence**. Le test examine la dépendance entre les deux variables avec une procédure semblable au test d'adéquation du khi carré. Il s'ensuit qu'il a une utilité assez différente des autres tests d'adéquation !

Tableau de contingence

Tableau de la fréquence d'observations décrites par 2 variables catégoriques. Donc, on peut visualiser la distribution empirique multivariée !

Note Voir plus bas pour un visuel de tableau de contingence.

Test d'indépendance du khi carré

Au lieu de tester si la distribution d'un échantillon suit celle d'une distribution théorique, on test si les 2 variables aléatoires sont indépendantes en fonction d'un tableau des fréquence de leurs observations. Les hypothèses sont donc :

H_0 : les 2 v.a. sont indépendantes $\quad$ H_1 : les 2 v.a. sont dépendantes

On pose qu'une variable aléatoire a a catégories et que l'autre a b catégories. Il s'ensuit que chacune des n observations appartient à une des combinaisons $a - b$.

On dénote, pour $i \in \{1, \dots, a\}$ et $j \in \{1, \dots, b\}$,

Notation

n_{ij} Le nombre d'observations dans la catégorie i de la première variable aléatoire et j de la deuxième.

$n_{i.}$ Le sous-total du nombre d'observations dans la catégorie i de la première variable aléatoire pour toutes les catégories de la deuxième.

$n_{.j}$ Le sous-total du nombre d'observations dans la catégorie j de la deuxième variable aléatoire pour toutes les catégories de la première.

Donc, le tableau de contingence est de la forme :

		Second variable				Total
		Cat. 1	Cat. 2	⋯	Cat. b	
First Variable	Cat. 1	n_{11}	n_{12}	⋯	n_{1b}	$n_{1\cdot}$
	Cat. 2	n_{21}	n_{22}	⋯	n_{2b}	$n_{2\cdot}$
	⋯	⋯	⋯	⋱	⋯	⋯
	Cat. a	n_{a1}	n_{a2}	⋯	n_{ab}	$n_{a\cdot}$
Total		$n_{\cdot 1}$	$n_{\cdot 2}$	⋯	$n_{\cdot b}$	n

La statistique du test a le même raisonnement que la statistique pour le test d'adéquation du khi carré : $t = \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{(n_{ij}n - n_{i\cdot}n_{\cdot j})^2}{n_{i\cdot}n_{\cdot j}}$ où la statistique est approximativement distribué selon la loi du khi carré : $T_n \sim \chi^2_{(a-1)(b-1)}$. La région critique est $t \geq \chi^2_{1-\alpha,(a-1)(b-1)}$.

Test du rapport de vraisemblance

Motivation

L'idée du test du rapport de vraisemblance (TRV) est de tester si les données peuvent être suffisamment bien expliquée par une simplification d'une distribution.

On l'appele le test du rapport de vraisemblance, « likelihood ratio test », car orginialement on calculait les 2 vraisemblances et on regardait le ratio. On utilisait des valeurs critiques calculées manuellement. Cependant, avec l'avancement de la statistique, on est venu à la réalisation que la statistique de test $T = 2(\ell_1 - \ell_0)$ suit approximativement une distribution du khi carré et on peut utiliser valeurs critiques d'une table du khi carré.

Test du rapport de vraisemblance

Les hypothèses sont

H_0 : les données proviennent d'une distribution A

H_1 : les données proviennent d'une distribution B

où la distribution A est un cas spécial de la distribution B.

Un des exemples le plus intuitif est le cas d'une distribution exponentielle. La distribution exponentielle est un cas spécial de la distribution gamma où le paramètre de forme $\alpha = 1$. Plus généralement, on pourrait avoir une distribution dont certains paramètres sont prédéterminés comme une simplification de la même distribution avec les paramètres qui sont inconnus. Par exemple, comparer une distribution « B » normale avec les 2 paramètres - d'emplacement μ et d'échelle σ^2 - inconnus avec la distribution « A » normale simplifiée où le paramètre d'emplacement $\mu = 2$ et le paramètre d'échelle σ^2 est inconnu.

Vu que la distribution B est plus, de soi, plus complexe que la distribution A, on déduit que que B a plus de "*paramètres libres*". L'idée du test est que à priori nous préférons un modèle plus simple (distribution A), mais que nous utiliserons un modèle plus complexe (distribution B) s'il y a une amélioration suffisamment importante pour le justifier.

On dénote donc :

Notation

r_0 **et** r_1 Nombre de paramètres libres de la distribution A et B.

$\mathcal{L}_0$ **et** $\mathcal{L}_1$ Maximum de vraisemblance sous l'hypothèse nulle et alternative.

On dénote les log-vraisemblances comme $\ell_0 = \ln(\mathcal{L}_0)$ et $\ell_1 = \ln(\mathcal{L}_1)$.

Puis, la statistique du test $t = 2(\ell_1 - \ell_0)$ et est approximativement distribué selon la loi du khi carré : $T_n \sim \chi^2_{r_1 - r_0}$.

Note La distribution asymptotique du khi carré dépend sur :

1. la condition qu'un modèle "emboîte" l'autre,
2. que n soit large,
3. les conditions de régularité typiques soit maintenues,
4. sous les deux hypothèses, les EMV sont des solutions « *consistent* » aux équations de Score.

Note Voir la section *Test du rapport de vraisemblance* de la section sur la *Régression linéaire généralisée* pour l'application du test du rapport de vraisemblance des les modèles linéaires généralisés.

Statistiques exhaustives

Soit l'échantillon aléatoire $(X_1, \ldots, X_n)$ d'une distribution avec paramètre θ inconnu.

Statistique exhaustive (« *sufficient* »)

La statistique T_n est une **statistique exhaustive** pour θ ssi la distribution de l'échantillon conditionnelle à la valeur de l'estimateur ne dépend pas de θ. C'est-à-dire, ssi $f(x_1, \ldots, x_n | t) = h(x_1, \ldots, x_n)$ où la fonction $h(\cdot)$ ne dépend pas de θ.

Donc, savoir la valeur t que prend la statistique T_n nous donne ***suffisamment*** d'information à propos de l'effet de θ sur l'échantillon sans avoir à connaître les n valeurs observées.

Exemple Bernoulli

Soit l'échantillon aléatoire d'une distribution Bernoulli de paramètre p.
Alors $T_n = \sum_{i=1}^n X_i$ est exhaustive pour p, car

$$\begin{aligned}\Pr(X_1 = x_1, \ldots, X_n = x_n | T_n = x_1 + \cdots + x_n) &= \frac{\prod_{i=1}^n p(x_i)}{p_{T_n}(t)} \\ &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \\ &= p^t (1-p)^{n-t}\end{aligned}$$

où $h_1(\cdot)$ dépend seulement de l'échantillon par t et $h_2(\cdot)$ ne dépend pas de p.

Note Pour une fonction injective (« *one-to-one* »), si T_n est une statistique exhaustive pour θ, alors $g(T_n)$ est une statistique exhaustive pour θ et T_n est une statistique exhaustive pour $g(\theta)$.

Limitations

La définition de l'exhaustivité nécessite de connaître la distribution de la statistique pour trouver $f_{T_n}(t)$ (ou $p_{T_n}(t)$ dans le cas discret). Cependant, ceci n'est pas toujours possible. Alors, nous pouvons utiliser l'approche du théorème de factorisation de Fisher-Neyman afin de prouver qu'une statistique est exhaustive.

Théorème de factorisation de Fisher-Neyman

La statistique T_n est une **statistique exhaustive** pour θ ssi on peut récrire la fonction de densité comme le produit d'une fonction $(h_1(\cdot))$ de la statistique $T_n$et du paramètre θ et d'une fonction $(h_2(\cdot))$ de l'échantillon. C'est-à-dire, ssi $f(x_1;\theta) \times \ldots \times f(x_n;\theta) = h_1(t;\theta) \times h_2(x_1, \ldots, x_n)$ où

1. $h_1(t;\theta)$ dépend de l'échantillon seulement par la statistique T_n.
2. $h_2(x_1, \ldots, x_n)$ ne dépend pas du paramètre θ.
3. $\forall i = 1, 2, \ldots, n$ $x_i \in \mathbb{R}$.

Cas multivarié

Pour $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)$, les statistiques $T_n^1, \ldots, T_n^r$ sont **conjointement exhaustives pour $\boldsymbol{\theta}$** si $f(x_1;\theta) \times \ldots \times f(x_n;\theta) = h_1(t_1, \ldots, t_r;\theta) \times h_2(x_1, \ldots, x_n)$ où

1. $h_1(t^1, \ldots, t^r;\theta)$ dépend de l'échantillon seulement par les statistiques $T_n^1, \ldots, T_n^r$.
2. $h_2(x_1, \ldots, x_n)$ ne dépend pas des paramètres θ.
3. $\forall i = 1, 2, \ldots, n$ $x_i \in \mathbb{R}$.

Exemple Bernoulli

Soit l'échantillon aléatoire d'une distribution Bernoulli de paramètre p.

Alors, par le théorème de factorisation, $T_n = \sum_{i=1}^n X_i$ est une statistique exhaustive pour p, car

$$\begin{aligned}p(x_1, \ldots, x_n) &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ &= p^{x_1+\cdots+x_n}(1-p)^{n-(x_1+\cdots+x_n)} \times 1 \\ &= h_1(x_1 + \cdots + x_n; p) h_2(x_1, \ldots, x_n)\end{aligned}$$

dépend seulement de l'échantillon par la valeur t de la statistique T_n.

Limitations

Le théorème de factorisation permet d'identifier des statistiques exhaustives. Cependant, il peut y avoir plusieurs statistiques exhaustives dont certaines qui offrent une plus grande réduction des données.

Par exemple, la moyenne empirique $\bar{X}_n$ réduit davantage les données que les statistiques d'ordre $(X_{(1)}, \ldots, X_{(n)})$. On cherche donc la statistique exhaustive offrant la **réduction maximale** qui retient cependant toute l'information sur le paramètre visé.

Statistique complète

Concept à clarifier, pas clair.

Statistique complète

La distribution de T_n provient d'une famille ***complète*** de distributions si le fait que $\mathrm{E}[g(T_n)] = 0$ **implique** que $\Pr(g(T_n) = 0) = 1$.

Il s'ensuit qu'il est possible que $\mathrm{E}[g(T_n)] = 0$ sans que la distribution de la statistique provienne d'une famille complète de distributions.

Le fait qu'une statistique soit complète veut dire que toute fonction $g(\cdot)$ qui entraîne la moyenne de $g(T_n)$ à être nulle doit être une fonction « *that maps to 0* ».

Il est hors du cadre de l'examen de devoir prouver que des statistiques sont complètes.

Le fait d'être complet implique qu'il existe une seule fonction T_n qui est un estimateur non biaisé de θ ; alias, $g(T_n)$ est **unique**.

Théorème de Lehmann-Scheffé

Si :

1. La statistique T_n est une statistique exhaustive pour θ.
2. La distribution de T_n provient d'une famille de distributions complète.
3. Il y existe une fonction unique $\varphi(\cdot)$ de T_n tel que $\varphi(T_n)$ est une estimateur non biaisé de θ.

alors la statistique $\varphi(T_n)$ est le MVUE de θ.

Contexte

Le théorème de Rao-Blackwell se base sur le théorème de Lehmann-Scheffé pour poser que la fonction unique $\varphi(\theta)$ doit être $\mathrm{E}_{\hat{\theta}_n}[\hat{\theta}_n | T_n]$.

Théorème de Rao-Blackwell

Si :

1. La statistique T_n est une statistique exhaustive pour θ.
2. La statistique $\hat{\theta}_n$ est un estimateur non biaisé de θ.

où T_n n'est pas fonction de $\hat{\theta}_n$, et vice-versa.

Le fait que T_n est exhaustif garanti que la distribution de $(\hat{\theta}_n | T_n)$ n'est pas fonction de θ. Alors, la fonction $\tilde{\theta}_n = \mathrm{E}_{\hat{\theta}_n}[\hat{\theta}_n | T_n]$ est une *statistique* non biaisé de θ avec $\mathrm{Var}(\tilde{\theta}_n) \leq \mathrm{Var}(\hat{\theta}_n)$.

Par le théorème de Lehmann-Scheffé, la distribution complète implique un MVUE unique. Donc, la statistique $\tilde{\theta}_n$ *doit être le* ***MVUE*** puisque sa variance est inférieure (ou égale) à tout autre estimateur non biaisé $\hat{\theta}_n$.

En bref, pour trouver le MVUE :

1. Trouver une statistique T_n complète exhaustive pour θ.
2. Trouver une fonction de T_n non biaisé pour θ.

Note Voir la section **Estimateur non biaisé à variance minimale (MVUE)** pour plus de détails sur le MVUE.

Statistique exhaustive minimale

Statistique exhaustive minimale

Une statistique exhaustive $T_n = T(X_1, \ldots, X_n)$ est "**minimale**" si pour toute autre statistique exhaustive $U_n = U(X_1, \ldots, X_n)$, il existe une fonction g telle que $T = g\{U(X_1, \ldots, X_n)\}$.

Critère de Lehmann-Scheffé

La statistique T_n est **exhaustive minimale** pour θ si $\frac{f(x_1;\theta)\times\ldots\times f(x_n;\theta)}{f(y_1;\theta)\times\ldots\times f(y_n;\theta)}$ ne dépend pas de θ ssi $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$
où, $\forall i = 1, 2, \ldots, n$, $x_i, y_i \in \mathbb{R}$.

Exemple Bernoulli

Soit l'échantillon aléatoire d'une distribution Bernouilli de paramètre p.

$$\frac{f(x_1;\theta)\times\ldots\times f(x_n;\theta)}{f(y_1;\theta)\times\ldots\times f(y_n;\theta)} = \left(\frac{p}{1-p}\right)^{(x_1+\cdots+x_n)-(y_1+\cdots+y_n)}$$

Le ratio est seulement indépendant de p si $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$ et donc $T_n = \sum_{i=1}^n X_i$ est **exhaustive minimale** pour p.

Famille exponentielle

Note La sous-section sur la *Famille exponentielle* de la sections sur la *Régression linéaire généralisée* couvre plus en détails la famille linéaire. Cette sous-section se limite à ses propriétés utiles pour identifier le MVUE.

Contexte

Dans le contexte du MVUE, la famille exponentielle est utile car, si une distribution provient de la famille exponentielle, il est beaucoup plus simple de le trouver.

La famille exponentielle

La variable aléatoire X fait partie de la famille exponentielle si l'on peut récrire sa fonction de probabilité comme : $f(x) = e^{a(x)\cdot b(\theta)+c(\theta)+d(x)}$ où

1. θ est le paramètre d'intérêt.
2. le domaine de X ne dépend pas du paramètre θ.

Exhaustivité et « *completeness* »

Pour un échantillon aléatoire tiré d'une distribution faisant partie de la famille exponentielle, on trouve que $f(x_1, \ldots, x_n) = h_1\left(\sum_{i=1}^n a(x_i); \theta\right) h_2(x_1, \ldots, x_n)$.

Par le théorème de factorisation, la statistique $\sum_{i=1}^n a(x_i)$ est une statistique exhaustive pour θ. De plus, la **distribution de la statistique** $\sum_{i=1}^n a(x_i)$ **provient d'une famille complète** de distributions (la preuve est hors du cadre de l'examen)

Plusieurs distributions font partie de la famille exponentielle, voici un tableau résumé :

Distribution	Paramètre d'intérêt	$\sum_{i=1}^n a(x_i)$	MVUE
Binomiale	q	$\sum_{i=1}^n X_i$	$\frac{1}{m}\bar{X}$
Normale	μ	$\sum_{i=1}^n X_i^2$	$\bar{X}$
Normale	σ^2	$\sum_{i=1}^n X_i$	$\frac{\sum_{i=1}^n (X_i^2)}{n} - \mu^2$
Poisson	λ	$\sum_{i=1}^n X_i$	$\bar{X}$
Gamma	θ	$\sum_{i=1}^n X_i$	$\frac{1}{\alpha}\bar{X}$
Inverse Gaussienne	μ	$\sum_{i=1}^n X_i$	$\bar{X}$
Binomiale Négative	β	$\sum_{i=1}^n X_i$	$\frac{1}{r}\bar{X}$

Statistiques d'ordre

Principes fondamentaux

k^{e} statistique d'ordre

La k^{e} statistique d'ordre est la k^{e} observation, en ordre croissant, dénotée $X_{(k)}$ $\forall k = 1, 2, \ldots, n$. Ceci correspond également au $\frac{k}{n+1}^{\text{e}}$ quantile.

Si l'échantillon est un **échantillon aléatoire**, on peut identifier la fonction de densité et de répartition :

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!1!(n-k)!} \underbrace{[F_X(x)]^{k-1}}_{\text{observations} < k} \overbrace{f_X(x)}^{\text{observation} = k} \underbrace{[S_X(x)]^{n-k}}_{\text{observations} > k}$$

$$F_{X_{(k)}}(x) = \underbrace{\sum_{i=k}^{n} \binom{n}{i} [F_X(x)]^i [S_X(x)]^{n-i}}_{\substack{\text{Probabilité qu'au moins } k \text{ des } n \\ \text{observations } X_k \text{ sont} \leq x}}$$

Note Les parenthèses sont utilisées pour distinguer la k^{e} statistique d'ordre $X_{(k)}$ de la k^{e} observation X_k.

Nous sommes habituellement intéressés au minimum $X_{(1)}$ et au maximum $X_{(n)}$. Nous les définissions ci-dessous avec leurs fonctions pour un **échantillon aléatoire** :

Minimum

$$X_{(1)} = \min(X_1, \ldots, X_n)$$

$$f_{X_{(1)}}(x) = n f_X(x) \big(S_X(x)\big)^{n-1}$$

$$S_{X_{(1)}}(x) = \prod_{i=1}^{n} \Pr(X_i > x) = [S_X(x)]^n$$

Maximum

$$X_{(n)} = \max(X_1, \ldots, X_n)$$

$$f_{X_{(n)}}(x) = n f_X(x) \big(F_X(x)\big)^{n-1}$$

$$F_{X_{(n)}}(x) = \prod_{i=1}^{n} \Pr(X_i \leq x) = [F_X(x)]^n$$

Cas spéciaux

Si l'échantillon aléatoire est tiré d'une **distribution uniforme** sur $[a, b]$, alors $\mathrm{E}[X_{(k)}] = a + \frac{k(b-a)}{n+1}$. De plus, si $X \sim U(0, \theta)$ alors

$$f_{X_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \left(\frac{x}{\theta}\right)^k \left(1 - \frac{x}{\theta}\right)^k \frac{1}{x}, \quad 0 \leq x \leq \theta$$

où il s'ensuit que $X_{(k)} \sim \text{Beta}(k, n-k+1, \theta)$.

Note Voir la section sur les *Temps d'occurrence conditionnels* pour l'application de cette espérance aux processus de Poisson.

Si l'échantillon aléatoire est tiré d'une **distribution exponentielle** de moyenne θ, alors $\mathrm{E}[X_{(k)}] = \theta \sum_{i=n-k+1}^{n} \frac{1}{i}$.

Pour un système de k parmi n, dont le nombre minimal de composantes est K, on dénote les durées de vie des n composantes par l'échantillon aléatoire $X_1, \ldots, X_n$. Pour lier les statistiques d'ordre à ce système, on pose que $k = n - K + 1$ et donc **la durée de vie du système** est $X_{(n-k+1)}$.

Autres statistiques

Nous définissons également quelques autres statistiques d'intérêt :

L'étendue (« *range* »)

L'étendue (range) est la différence entre le minimum et le maximum d'un échantillon : $R = X_{(n)} - X_{(1)}$.

Contexte

L'utilité de l'étendue est limitée puisqu'elle est très sensible aux données extrêmes.

Par exemple, supposons que l'on a des données historiques sur la température pour le 1er septembre. En moyenne, la température est de 16°C mais il y a un cas extrême de $-60°C$ en 1745. L'étendue sera de 86°C, ce qui n'est pas très représentatif des données. Donc, dans ce contexte, l'étendue n'est pas une mesure très utile.

La mi-étendue (« *midrange* »)

La moyenne entre du minimum et du maximum d'un échantillon : $M = \frac{X_{(n)}+X_{(1)}}{2}$.

Pour comprendre ce que représente la mi-étendue, on la compare à la moyenne arithmétique.

La moyenne arithmétique considère les données observées et calcule leur moyenne.

– Il s'ensuit qu'elle ne considère pas les chiffres qui ne sont pas observés.

La mi-étendue considère **tous** les chiffres—observés ou non—entre la plus grande et la plus petite valeur d'un échantillon, puis en prend la moyenne.

L'écart interquartile (« *interquartile range (IQR)* »)

Écart entre le troisième quartile et le premier quartile : $IQR = Q_3 - Q_1$.

L'IQR mesure la distribution du 50% des données qui sont situées au milieu de l'ensemble de données.

L'IQR est connu comme le « *midspread* ».

Exemple de contexte pour les statistiques

Nous cherchons à comprendre les contextes dans lesquels les différents mesures sont utiles. Pour ce faire, nous supposons un échantillon de données météorologiques $\{-30°, -24°, -7°, -23°, +5°\}$ (Celsius).

Pour un premier contexte, on suppose que les données représentent la température du 4 février des dernières années. Dans ce contexte, la moyenne arithmétique ($-22.25°C$) est une statistique intéressante, car elle nous informe que, en moyenne, la température ressentie le 4 février est de $-22.25°C$. En revanche, la mi-étendue ($-12.5°C$) et l'étendue ($-35°C$) ne m'intéressent pas, car elles ne considèrent pas la *vraisemblance* des différentes températures.

Pour un deuxième contexte, on suppose que ces données sont des températures observées tout au long de l'hiver passé. Dans ce contexte, la moyenne arithmétique n'est pas une statistique intéressante ; on ne peut pas supposer une température moyenne sur plusieurs mois à partir de 5 observations. De plus, il est illogique de vouloir connaître la température moyenne entre décembre et mars—ce chiffre ne veut rien dire.

En revanche, la mi-étendue et l'étendue me donnent maintenant une meilleure idée de la température au fil de l'hiver. Ils m'informent sur les valeurs extrêmes ce qui a du sens.

L'important à retenir est que l'utilité des mesures dépend de la situation. Également, ceci est un exemple **très** simpliste et, dans tous les cas, on ne peut pas tirer de conclusions sur les températures de l'hiver à partir de 5 observations.

Nous définissons la **médiane** en termes de statistiques d'ordre :

Médiane

$$\text{Med} = \begin{cases} X_{((n+1)/2)}, & \text{si n est impair} \\ \frac{X_{(n/2)}+X_{(n/2+1)}}{2}, & \text{si n est pair} \end{cases}$$

La moitié des données sont supérieures et inférieures à la médiane.

L'utilité de la médiane est qu'elle n'est pas aussi sensible aux données aberrantes que la moyenne.

Distribution conjointe

La distribution conjointe du minimum et du maximum pour un échantillon aléatoire a la fonction de densité, $\forall x < y$,

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)[F_X(y) - F_X(x)]^{n-2} f_X(x) f_X(y).$$

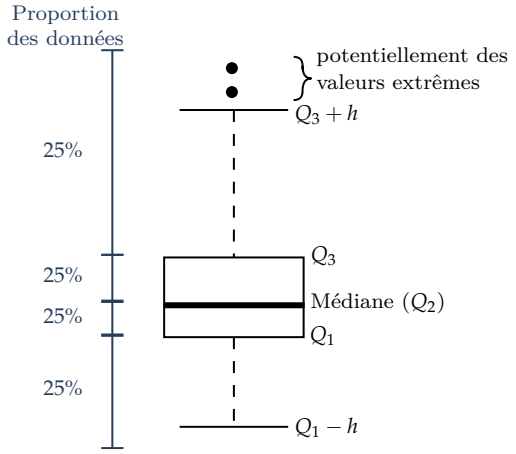
Graphiques

Diagramme en boîte (« *boxplot* »)

Le diagramme du « ***sommaire à cinq chiffres*** » composé de :

1. Le minimum.
2. Le premier quartile Q_1.
3. La médiane (deuxième quartile) Q_2.
4. Le troisième quartile Q_3.
5. Le maximum.

Visuellement :



Remarques :

La médiane (ligne dans la boîte) correspond au point où la moitié des données sont au-dessus et l'autre moitié en dessous.

La boîte est délimitée par le premier et le troisième quartile.

- Il s'ensuit que la boîte contient la moitié des données.
- De plus, 25% des données sont contenues entre la borne *supérieure* de la boîte et la médiane (l'autre 25% est contenu entre la borne *inférieure* et la médiane).

Les « moustaches » de la boîte sont tracées à un pas h des quartiles où $h = 1.5 \cdot (Q_3 - Q_1)$.

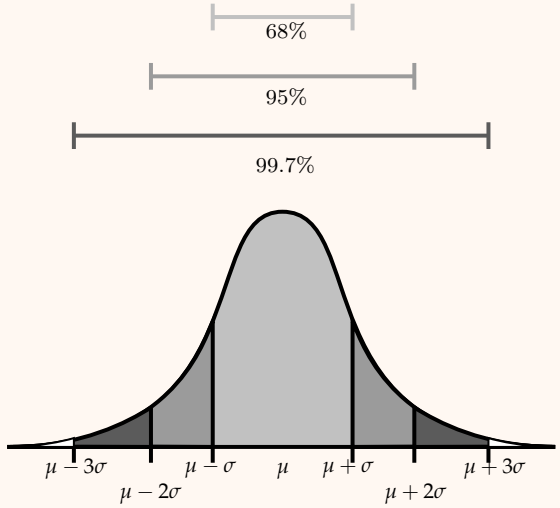
- Les points à l'extérieur de ces bornes sont *potentiellement* des données aberrantes.
- Plus *l'écart interquartile* $(Q_3 - Q_1)$ est élevé, plus la boîte sera large et, par conséquent, plus les moustaches seront loin de la médiane.

Note Le « 1.5 » du pas h découle de la ***règle du 68-95-99.7***. Selon la règle, moins de 1% des données seront situées à l'extérieur de la borne supérieure.

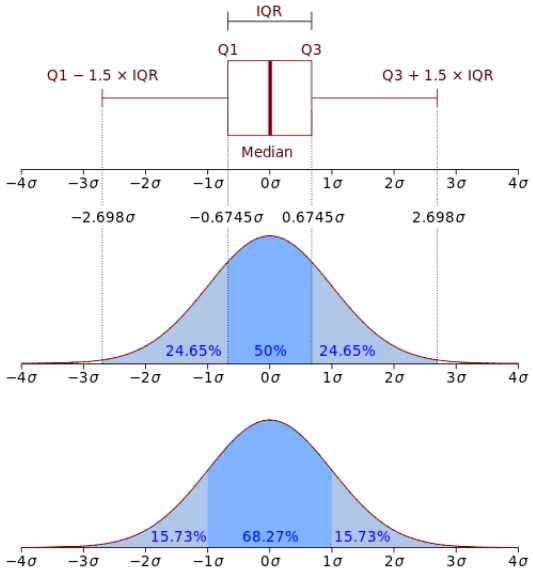
En premier, on défini la règle du 68-95-99.7.

Règle du 68-95-99.7

Pour un échantillon tiré d'une distribution normale, environ 68% des données se situent à moins de 1 écart-type de la moyenne, 95% à moins de 2 et 99.7% à moins de 3.



Puis, ce graphique tiré de *Wikipédia* illustre bien le lien entre le diagramme en boîte et la règle du 68-95-99.7 :



Donc, pourquoi utiliser 1.5 comme échelle de l'écart interquartile ? Puisque cela permet d'englober environ 99% des données.

Note Voir *Diagramme en boîte* pour l'interprétation des diagrammes en boîte.

Diagramme quantile-quantile (« *Q-Q plot* »)

En pratique, on pose souvent que les données suivent une distribution. Un diagramme quantile-quantile permet de comparer les quantiles théoriques de la distribution aux quantiles empiriques des données.

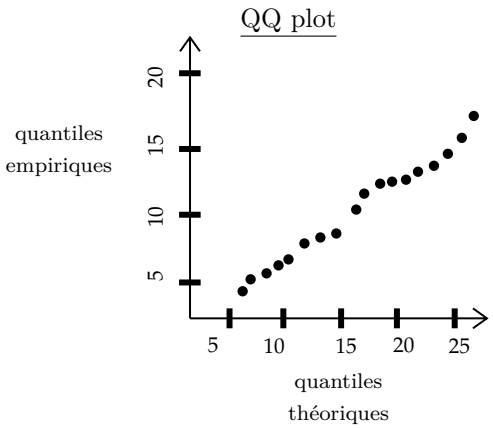
Dans un tel cas, on connaît la distribution, mais pas les paramètres.

Si les données sont normalement distribuées, on peut centrer et réduire pour obtenir la loi normale standard Z ;

– Ceci correspond à un diagramme quantile-quantile **normale** ;

Autrement, le diagramme quantile-quantile est tracé en estimant les paramètres de la distribution avec l'échantillon de données.

Par exemple :



Note Voir *Diagramme quantile-quantile* pour l'interprétation du diagramme quantile-quantile.

Construction d'estimateurs

Introduction

Contexte

Plus tôt, nous avons décrit les méthodes utilisées pour évaluer la **qualité** d'un estimateur. Cependant, nous n'avons pas décrit comment obtenir ces estimateurs. Non seulement il y a une panoplie de façons de construire un estimateur, mais aussi de façons d'estimer des paramètres.

La méthode vue dans le cadre du cours de statistique (et de l'examen MAS-I) est la méthode dite « **fréquentiste** ». Le cours de Erreur présente **l'estimation bayésienne**.

Dans le contexte de l'examen, nous voyons 3 méthodes. Les deux premières (méthode des moments et du « percentile matching ») sont les plus faciles à obtenir. Cependant, elles sont aussi les méthodes d'estimation les moins précises car elles utilisent seulement une *portion* des données. En revanche, la méthode du maximum de vraisemblance utilise *toutes* les données.

Cette distinction devient particulièrement importante dans le cas d'une distribution avec une queue de droite lourde (e.g. Pareto, Weibull, etc.). Pour ces distributions, il est essentiel de connaître précisément les valeurs extrêmes afin de bien estimer le(s) paramètre(s) de forme.

Les deux premières méthodes comporte également la limitation que les données doivent toutes provenir de la même distribution. Autrement, il ne serait pas clair ce que sont les moments et quantiles. Finalement, les deux premières méthodes peuvent être manipulées car car la décision de quels moments et centiles à utiliser est *arbitraire*.

Méthode des moments (MoM)

Terminologie

$\mu'_k(\hat{\theta})$ k^{e} moment centré à 0, $\mu'_k = \mathrm{E}[X^k]$.

$\widehat{=}$ Notation pour poser une égalité.

Méthode des moments (MoM)

Contexte

La méthode des moments applique l'idée, ou « hypothèse », qu'un échantillon de données devrait être semblable à sa distribution posée. Elle estime les paramètres avec les moments empiriques sous l'hypothèse que les moments empiriques devraient, en théorie, être égaux aux moments théoriques.

On pose les r premiers moments empiriques de l'échantillon égaux aux r premiers moments théoriques d'une distribution X ayant r paramètres $\boldsymbol{\theta}$.

L'estimation de $\boldsymbol{\theta}$ est donc la solution aux r équations suivantes :

$$\hat{\mu}'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k \widehat{=} \mathrm{E}\left[X^k\right] = \mu'_k(\theta), \quad k = 1, 2, \ldots, r$$

Note Pour des données incomplètes, on utilise le moment qui y correspond. Par exemple, si nous avons des données avec une limite de u alors on utilise $\mathrm{E}[X \wedge u]$.

Méthode du «Percentile Matching »

Notation

$\pi_q(\boldsymbol{\theta})$ $100q^{\text{e}}$ centile, $\pi_q(\boldsymbol{\theta}) = F_{\boldsymbol{\theta}}^{-1}(q)$, $q \in [0,1]$.

Méthode du « *percentile matching* »

Contexte

La méthode du « *percentile matching* » estime les paramètres avec les centiles empiriques sous l'hypothèse que les centiles empiriques devraient, en théorie, être égaux aux centiles théoriques.

Un désavantage de cette méthode est le choix des centiles à utiliser arbitraire. Ceci peut mener à des manipulations des données. Dans le contexte d'un examen cependant, les centiles à utiliser seront spécifiés.

On pose r centiles empiriques de l'échantillon égaux aux r centiles théoriques correspondants d'une distribution X ayant r paramètres $\boldsymbol{\theta}$.

L'estimation de $\boldsymbol{\theta}$ est donc la solution aux r équations suivantes :

$$\hat{\pi}_{q_k} \mathrel{\widehat{=}} \pi_{q_k}(\boldsymbol{\theta}), \quad k = 1,2,\ldots,r$$

Cependant, nous devons calculer ces centiles ! Il y existe une myriade de façons de le faire, mais pour l'examen on utilise le « *smoothed empirical estimate* » d'un centile. Entre autres, cette méthode permet d'interpoler des quantiles s'il y en a qui n'existent pas.

« *smoothed empirical estimate* »

Notation

$x_{(i)}$ La i^e statistique d'ordre de l'échantillon.

$b = \lfloor q(n+1) \rfloor$ Arrondi vers le bas du centile.

Étapes pour trouver les centiles :

1. Trier les observations en ordre croissante pour obtenir les statistiques d'ordre.
2. Calculer $q(n+1)$ et $b = \lfloor q(n+1) \rfloor$.
3. Si
   a) $q(n+1)$ est fractionnaire, calculer $\hat{\pi}_q$ comme l'interpolation linéaire de $x_{(b)}$ et $x_{(b+1)}$.
   b) $q(n+1)$ est entier, simplement poser $\hat{\pi}_q = x_{(b)}$.

En bref, pour $h = q(n+1) - b$, $\hat{\pi}_q = (1-h)x_{(b)} + hx_{(b+1)}$.

Note Pour des valeurs répétées (deux observations de l'échantillon ont la même valeur), on conserve uniquement le plus gros indice parmi les doublons. Si $x_{(2)} = x_{(3)}$ alors on conserve $x_{(3)}$ pour les interpolations.

Exemple « *smoothed percentile matching* » avec doublons

Soit l'échantillon de nombres $\{1,1,1,2,3,3,7,7,8,9,9,9\}$, quel est le 40^e centile ?

1. $0.40 \times (12+1) = 5.2$.
2. On récrit un tableau des indices et des valeurs :

Indice	Nombre
3	1
4	2
6	3
8	7
9	8
12	9

3. Puisque 5 est retiré comme doublon, on obtient que

$$\hat{\pi}_{0.4} = \left(1 - \frac{5.2-4}{6-4}\right) x_{(4)} + \left(\frac{5.2-4}{6-4}\right) x_{(6)} = 2.6$$

Méthode du maximum de vraisemblance

Contexte

La méthode du maximum de vraisemblance trouve le(s) paramètre(s) $\boldsymbol{x}$ qui maximise(nt) la probabilité d'avoir observé l'échantillon de données. On maximise la fonction de vraisemblance $\mathcal{L}(\theta;\boldsymbol{x})$ ou, puisque le logarithme ne change pas le maximum, la fonction de log-vraisemblance $\ell(\theta;x)$.

Voir la section *Vraisemblance* pour plus de détails sur la distinction de la fonction de vraisemblance à la fonction de densité. Également, la section *Estimation de modèles paramétriques* du chapitre *Erreur* explique la méthode du maximum de vraisemblance pour des données incomplètes.

Méthode du maximum de vraisemblance

On défini $\mathcal{L}(\theta;\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i;\theta)$ et $\ell(\theta;\boldsymbol{x}) = \sum_{i=1}^{n} \ln f(x_i;\theta)$, puis on calcule $\hat{\theta}^{\text{EMV}} = \max_{\theta}\{\mathcal{L}(\theta;\boldsymbol{x})\} = \max_{\theta}\{\ln \mathcal{L}(\theta;\boldsymbol{x})\}$.

Habituellement, on trouve la dérivée de la fonction de (log-)vraisemblance et trouve le paramètre θ tel que $\frac{d}{d\theta}\mathcal{L}(\theta;\boldsymbol{x}) = 0$.

Raccourcis

Si la fonction de vraisemblance est de la forme :

$\mathcal{L}(\gamma) = \gamma^{-a}\mathrm{e}^{-b/\gamma}$ alors $\hat{\gamma}^{\text{MLE}} = \frac{b}{a}$.

$\mathcal{L}(\lambda) = \lambda^{a}\mathrm{e}^{-\lambda b}$ alors $\hat{\lambda}^{\text{MLE}} = \frac{a}{b}$.

$\mathcal{L}(\theta) = \theta^{a}(1-\theta)^{b}$ then $\hat{\theta}^{\text{MLE}} = \frac{a}{a+b}$.

Propriétés

Propriété d'invariance

La propriété d'invariance implique que l'estimateur du maximum de vraisemblance d'une fonction $g(\cdot)$ du paramètre θ est la fonction évaluée à $\hat{\theta}^{\text{EMV}}$: $g(\hat{\theta}^{\text{EMV}})$ est l'EMV de $g(\theta)$.

Exemple de la propriété d'invariance

Afin de bien comprendre ce que veut dire la propriété d'invariance, on donne un exemple avec la loi de Poisson.

Pour une loi de Poisson, l'estimateur du maximum de vraisemblance est $\hat{\theta} = \bar{X}$. Par la propriété d'invariance, on peut déduire que l'estimateur du maximum de vraisemblance de la fonction $g(\lambda) = \mathrm{e}^{-\lambda}$ est $g(\hat{\lambda}) = \mathrm{e}^{-\hat{\lambda}}$.

Caractéristiques des estimateurs du maximum de vraisemblance

Les estimateurs du maximum de vraisemblance ont généralement ces 3 propriétés désirables :

1. $\hat{\theta}_n^{\text{EMV}}$ est un *estimateur « consistent »* pour θ.
2. $\hat{\theta}_n^{\text{EMV}}$ est asymptotiquement normalement distribué.
3. S'il y existe une statistique T_n *exhaustive* pour θ, alors $\hat{\theta}_n^{\text{EMV}}$ en est une fonction.

Les deux premières caractéristiques doivent cependant respecter ces 3 conditions :

1. Les conditions de régularité habituelles.
2. $\hat{\theta}_n^{\text{EMV}}$ est la solution unique de l'équation de score (des dérivées partielles).
3. $(X_1, X_2, \ldots, X_n)$ est un échantillon aléatoire.

Distribution asymptotique de l'estimateur du maximum de vraisemblance

Sous certaines conditions de régularité, la distribution de l'estimateur du maximum de vraisemblance $\hat{\theta}^{\text{EMV}}$ converge en distribution vers une distribution normale avec une moyenne θ et une variance égale à la borne de Cramér-Rao : $\hat{\theta}^{\text{EMV}} \approx \mathcal{N}\left(\theta, \frac{1}{\boldsymbol{I}_n(\theta)}\right)$.

En termes mathématiques,

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{D} \mathcal{N}\left(0, \frac{1}{\boldsymbol{I}_n(\theta)}\right)$$

La normalité de la distribution asymptotique implique :

1. $\hat{\theta}^{\text{EMV}}$ est *asymptotiquement* **sans biais**.
2. $\hat{\theta}^{\text{EMV}}$ est « ***consistent*** ».

3. $\hat{\theta}^{\text{EMV}}$ est, pour des grands échantillons, approximativement normalement distribué avec moyenne θ et variance $1/\boldsymbol{I}_n(\theta)$.
4. $\hat{\theta}^{\text{EMV}}$ est **asymptotiquement efficace**, car sa variance tend vers la borne Cramér-Rao.

Contexte

Souvent, nous voyons ces théorèmes et définitions sans vraiment voir ce que sont les mystérieuses conditions de régularité sous lesquelles les théorèmes sont valides. La raison et que ces conditions sont relativement compliquées pour leur utilité.
Je résume donc les conditions ci-dessous, mais ne vous en faites pas si vous ne les comprenez pas—vous pouvez sauter l'encadré.

Conditions de régularité

R0 Les variables aléatoires X_i sont iid ayant comme fonction de densité $f(x_i;\theta)$, pour $i = 1, 2, \ldots$.

R1 Le support des variables aléatoires X_i ne dépend pas des paramètres .

C'est-à-dire que, pour tout θ, le support des fonctions de densité reste le même.

Ceci permet entre autres de garantir que la vraisemblance sera maximisée à la vraie valeur θ_0 du paramètre.

C'est une condition restrictive que certains modèles ne respectent pas (e.g. la loi uniforme).

R2 La "vraie valeur" θ_0 de θ est contenue dans l'ensemble des valeurs possibles Θ.

R3 La fonction de densité $f(x;\theta)$ est différentiable deux fois comme fonction de θ.

Cette condition additionnelle assure que les deux premières dérivées existent pour calculer l'information de Fisher.

R4 L'intégrale $\int f(x;\theta)dx$ est différentiable deux fois sous l'intégrale comme fonction de θ.

Cette condition additionnelle assure que l'on peut utiliser la deuxième dérivée pour calculer l'information de Fisher.

R5 La fonction de densité $f(x;\theta)$ est différentiable trois fois comme fonction de θ. De plus, $\forall\theta \in \Theta$ il existe une constante c and une fonction $M(x)$ tel que $\left|\frac{\partial^3}{\partial\theta^3} \ln f(x;\theta)\right| \leq M(x)$ où $\mathrm{E}_{\theta_0}[M(\mathrm{X})] < \infty$ et $|\theta - \theta_0| < c$.

Celle-ci est la plus compliquée et assure la normalité asymptotique de l'EMV.

Contexte

La distribution normale asymptotique de l'estimateur du maximum de vraisemblance se généralise au cas multivarié avec une distribution normale multivariée.

Soit une distribution avec r paramètre tel que $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_r)$, on trouve que : $\hat{\boldsymbol{\theta}}^{\text{EMV}} \approx \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}\right)$.

II

Modèles linéaires en actuariat

Apprentissage statistique

Apprentissage statistique

L'apprentissage statistique est l'utilisation des statistiques pour pour essayer de trouver, puis quantifier, des relations entre des variables « ***explicatives*** » et une variable « ***réponse*** ».

Variables d'un modèle d'apprentissage statistique

Notation

Y variable réponse.

On dénote la i^{e} observation, d'un ensemble de n observations, par y_i où $i = 1, 2, \ldots, n$.

X_j j^{e} variable explicative où $j = 1, 2, \ldots, p$.

On dénote la i^{e} observation, de la j^{e} variable explicative, par $x_{i,j}$ où $i = 1, 2, \ldots, n$.

p Nombre de variables explicatives.

Variable réponse

Typiquement, nous voulons effectuer des prévisions sur la valeur de la *variable réponse*. Également, on veut essayer de mieux la comprendre avec les variables explicatives. Elle a plusieurs noms : « *output variable* », « *dependant variable* », « *outcome* », etc.

Variables explicatives

Les *variables explicatives* sont toutes les variables qui peuvent aider à comprendre la variable réponse. Ils ont plusieurs noms : « *independent variables* », « *features* », « *predictors* », etc.

Ces variables sont soit *quantitatives* ou *qualitatives*.

Variable quantitative

Les *variables quantitatives*, aussi appelées les « *covariates*, » prennent comme valeur une quantité et se séparent en 2 types :

1. **Variable continue**

   Les variables continues sont définies sur les **nombres réels**. Par exemple,

   les montants de perte d'un accident d'automobile,

   le temps avant qu'une réclamation d'assurance soit réglée,

   la probabilité de précipitations, etc.

2 Variable de comptage

Les variables de comptage sont définies sur les entiers positifs. Par exemple,

le nombre de d'accidents d'automobile,

le nombre d'étudiants dans une salle de cours, etc.

Variable qualitative

Les *variables quantitatives*, aussi appelées les ***variables catégorielles***, prennent comme valeur un petit nombre de résultats, ou catégories, possibles. On peut aussi considérer les catégories comme des **niveaux** ou des **classes**.

Les variables qualitatives se séparent en deux types selon la présence ou absence d'ordre à ses niveaux :

1 Variable nominale

Variable qualitative dont les **catégories n'ont pas d'ordre**. Par exemple,

le programme d'étude d'un étudiant (actuariat, informatique, etc.),

la couleur d'une voiture (rouge, bleu, vert, etc.),

le fabricant d'une voiture (Toyota, Subaru, etc.), etc.

2 Variable ordinale

Variable qualitative dont les **niveaux ont une ordre**. Par exemple,

la sévérité d'un accident de 1 à 5,

le degré de risque d'incendie de bas à élevé, etc.

Notes :

Si une variable qualitative a seulement 2 classes, on l'appelle une ***variable binaire***.

– Par exemple, la variable "sexe de l'assuré" prenant comme valeur *homme* ou *femme*.

Une variable explicative qualitative est appelée un ***facteur***.

Types de modèles d'apprentissage statistique

L'apprentissage statistique se distingue par 2 types ; soit il est supervisé, ou il ne l'est pas.

Apprentissage supervisé

L'apprentissage statistique *supervisé* évalue des données qui comportent une variable réponse. Toute l'analyse est concentrée sur l'évaluation de cette variable via les variables explicatives.

Par exemple,

prédire le nombre de buts marqués par un joueur de la LNH,

prédire la quantité de neige qui va tomber l'année prochaine,

prédire si un client est satisfait ou pas suite à un appel, etc.

Apprentissage non supervisé

L'apprentissage statistique *non supervisé* analyse les observations ou les variables d'un ensemble de données qui ne contient pas une variable réponse. L'idée de l'analyse est d'identifier des tendances qui pourraient exister, mais il n'y a pas d'objectif spécifique ni de façon de quantifier les résultats de l'analyse.

Par exemple,

analyser ce qu'achètent les consommateurs d'une certaine région,

évaluer les caractéristiques des étudiants en défaillance dans leurs prêts étudiants,

évaluer la composition du sol dans un environnement pollué, etc.

Note Dans le cadre de l'examen MAS-I, seule *l'analyse en composantes principales (ACP)* est couverte. Nous nous concentrons surtout sur les méthodes d'apprentissage supervisé.

Problèmes d'apprentissage supervisé

Les **problèmes** d'apprentissage supervisé se divisent typiquement en 2 types de problèmes :

1 Régression

Un problème de régression implique une variable réponse **quantitative**.

Note L'examen MAS-I se concentre davantage sur la régression que la classification.

2 Classification

Un problème de *Classification* implique une variable réponse **qualitative**.

Note Parfois, la différence entre la régression et la classification est minime. Par exemple, la classification d'une variable binaire *oui* ou *non* peut devenir un problème de régression en estimant la *probabilité* d'un oui.

Un problème de régression suppose qu'il y existe une relation entre la variable réponse et les variables explicatives. De façon générale, on représente la relation entre une observation des variables explicatives et de la variable réponse comme

$$(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = \underbrace{f(x_1, x_2, \ldots, x_p)}_{\substack{\text{une fonction des} \\ \text{variables explicatives}}} + \underbrace{\varepsilon}_{\substack{\text{une erreur} \\ \text{irreductible}}}$$

Note Ici, le conditionnement de Y sur les valeurs prises par les variables explicatives provient du fait que l'échantillon de données est observé. Cependant, ce niveau de détail n'est pas nécessaire pour comprendre les concepts. Donc, dans le but de *simplifier* la suite, on va simplement écrire Y.

On pose que $\mathrm{E}[\varepsilon] = 0$ et donc $\mathrm{E}[Y] = \mathrm{E}[f(x_1, x_2, \ldots, x_p)] = f(x_1, x_2, \ldots, x_p)$. Bref, une réalisation de la variable réponse est composée d'une composante **systématique** et d'une composante **aléatoire**. Souvent on dénote cette décomposition comme « *signal plus noise* ».

L'objectif de la régression et donc d'estimer f par $\hat{f}$ en présumant que *la relation ne change pas*. Le processus d'estimer f pour obtenir un $\hat{f}$ optimal est la ***phase d'entraînement***.

Objectifs de l'apprentissage supervisé

Les **objectifs** de l'apprentissage supervisé se résume à un des deux suivants :

1 Prévision

Effectuer des *prévisions* de la valeur prise par la variable réponse en fonction d'observations des variables explicatives.

2 Inférence

Chercher à comprendre l'effet des variables explicatives sur la variable réponse.

Les différents modèles d'apprentissage statistique varient, entres autre, en ***flexibilité***. La **flexibilité** décrit le degré auquel $\hat{f}$ peut s'*ajuster* aux données. Un *meilleur ajustement* implique que $\hat{f}$ est *plus flexible*, alors qu'un *moins bon ajustement* implique que $\hat{f}$ est *moins flexible*.

Par exemple, une régression linéaire correspond à une ligne droite et donc $\hat{f}$ est très peu flexible. L'ajustement aux données est moins bon, car il est peu probable que les données soient situées sur une droite. En revanche, une spline passe à travers tous les points (l'ajustement est alors « parfait ») et donc $\hat{f}$ est très flexible. Cependant, un modèle peut devenir ***surajusté*** aux données d'entraînement et effectuer de mauvaises prévisions. Voir la section <TBD> pour de plus amples détails sur le surajustement d'un modèle.

Le désavantage des modèles plus flexibles est leur **complexité**. Plus un modèle est flexible, plus il est complexe et plus il est difficilement interprétable. De façon générale, les modèles plus flexibles sont préférables pour la prévision alors que les modèles plus simples (moins flexibles) sont préférables pour l'inférence.

Précision des modèles d'apprentissage statistique

Notation

$\hat{Y}$ Prévision de Y où $\hat{Y} = \hat{f}(x_1, x_2, \ldots, x_p) = \hat{\mathrm{E}}[Y]$.

Erreur quadratique moyenne

Contexte

Il s'ensuit de sa définition que $\hat{Y}$ est un **estimateur** de $\mathrm{E}[Y]$. Pour quantifier l'erreur de prévision de la variable réponse, nous désirons calculer la variabilité de l'écart entre les prévisions $\hat{Y}$ des observations et leurs vraies valeurs Y. Cependant, ceci ne correspond **pas** à la variance de l'estimateur, car $\hat{Y}$ estime l'espérance $\mathrm{E}[Y]$ de l'observation et non sa valeur Y. Plutôt, ceci correspond à l'*Erreur quadratique moyenne*.

Il y a cependant une distinction à faire entre l'application de l'EQM pour évaluer la précision d'un modèle et son application pour évaluer la *Qualité de l'estimateur*. Lorsqu'on évalue la qualité d'un estimateur, on calcule la différence entre **une variable aléatoire** ($\hat{\theta}$) et **une constante** (θ). Lorsqu'on évalue l'erreur d'un modèle, on calcule la différence entre **deux variables aléatoires** ($\hat{Y}$ et Y).

L'erreur quadratique moyenne (EQM) mesure la **précision** du modèle où $\mathrm{MSE} = \mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$. Typiquement, on calcule l'EQM avec l'échantillon de données où $\mathrm{MSE} = \dfrac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n}$.

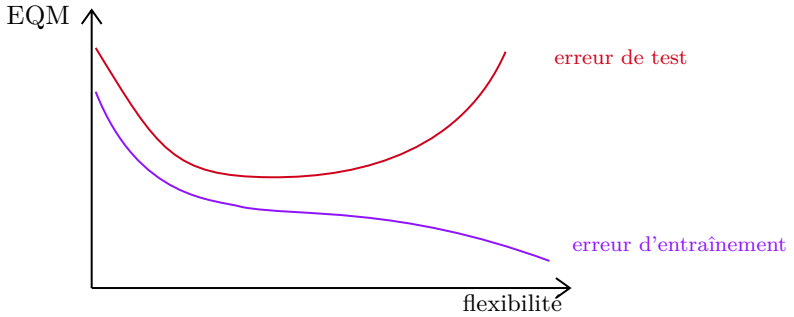
Si on calcule l'EQM avec l'échantillon de données de

test on obtient l'EQM de test.

d'entraînement on obtient l'EQM d'entraînement.

Si l'on ajuste le modèle avec l'EQM d'entraînement, on peut obtenir de très bonnes prévisions sur les données d'entraînement. Cependant, le modèle peut devenir ***surajusté*** et effectuer des mauvaises prévisions sur de nouvelles données ; il est difficilement généralisable. Alors, **on utilise l'EQM de test par défaut** évaluer la précision du modèle. C'est-à-dire, on ajuste le modèle avec les données d'entraînement puis on évalue la qualité de l'ajustement avec l'EQM calculé à partir des données de test.

De façon générale, l'EQM de test et d'entraînement ont le patron suivant :



Donc,

puisque l'EQM d'entraînement est optimisée pour les données d'entraînement, il sera toujours inférieur à l'EQM de test.

l'EQM d'entraînement décroît lorsque la flexibilité (et donc, la complexité aussi) du modèle augmente.

l'EQM de test est concave ce qui veut dire que le meilleur modèle fait un **compromis** entre la *flexibilité* et la *précision* du modèle.

Compromis biais-variance

Contexte

La variance de $\hat{f}$ mesure l'erreur due à la ***sensibilité*** *du modèle à l'ensemble de données*. C'est-à-dire, si la **forme** de $\hat{f}$ varie beaucoup selon l'ensemble de données utilisé pour ajuster le modèle. Si la variance de $\hat{f}$ est très élevée, ça peut signaler qu'il est surajusté aux données.

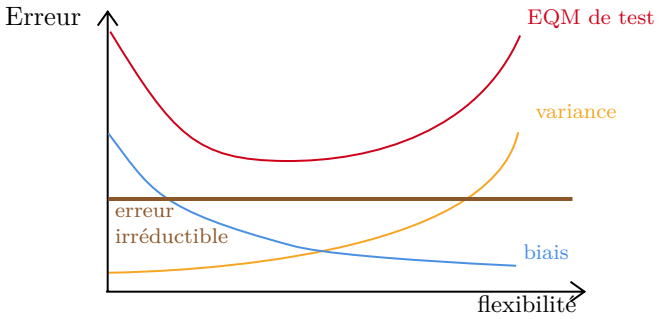
Le biais de $\hat{f}$ mesure le degré auquel $\hat{f}$ est proche de f. C'est une mesure de l'erreur **due aux hypothèses du modèle** d'apprentissage statistique. Un modèle très flexible s'ajuste bien aux données et comporte un faible biais. Par exemple, un spline a un faible biais et une variance élevée alors qu'une régression linéaire a un biais élevée et une faible variance.

Donc, augmenter la flexibilité diminue le biais et augmente la variance, puis vice-versa. Il y a donc un **compromis** à faire entre les deux.

On sait de la *définition de l'EQM* qu'on peut le récrire comme $\mathrm{MSE} = \mathrm{Var}(\hat{Y}) + \left[\mathrm{B}[\hat{Y}]\right]^2$. En décomposant le premier terme, on obtient que $\mathrm{MSE} = \mathrm{Var}(\hat{f}) + \left[\mathrm{B}[\hat{f}]\right]^2 + \mathrm{Var}(\varepsilon)$.

La troisième composante $\text{Var}(\varepsilon)$ correspond à la variance de **l'erreur irréductible**. Peu importe le modèle choisit, cette variance ne change pas et donc l'EQM de test ne peut pas y être inférieur. Les deux premières correspondent à la variance de ce qu'on appelle **l'erreur *réductible*** et peut être optimisée.

Pour choisir le meilleur modèle, nous allons donc vouloir trouver **le meilleur compromis** entre la variance et le biais du modèle :



Résumés numériques des modèles

Contexte

L'EQM, la variance et le biais d'un modèle permettent de quantifier sa précision. Cependant, avant d'ajuster un modèle, nous désirons avoir une image globale de l'ensemble de données sur lequel on l'ajuste.

Nous définissons alors quelques statistiques qui permettent d'évaluer les données.

Nous avons des statistiques

1 Univariées

Pour une variable x, nous utilisons typiquement la *moyenne échantillonnale* $\bar{x}$ et la *variance échantillonnale* s^2.

Également, on utilise les *statistiques d'ordre empiriques* et *l'écart interquartile*.

2 Bivariées

Pour deux variables x et y, on utilise typiquement la *covariance échantillonnale* $cov_{X,Y}$ et la *corrélation échantillonnale* $r_{X,Y}$.

Résumés graphiques des modèles

Contexte

Nous désirons avoir non seulement des résumés numériques, mais également des *résumés graphiques*.

Nuage de points (« *Scatterplots* »)

Les nuages de points permettent de visualiser les réalisations de 2 variables afin d'évaluer leur relation. Lorsqu'il y a plusieurs variables, on peut faire une **matrice de nuages de points** qui montre le diagramme de toutes les combinaisons des variables.

Diagramme en boîte

Contexte

Le *Diagramme en boîte (« boxplot »)* est défini dans la section *Graphiques* du *chapitre de statistiques*.

Dans cette section, on décrit ***l'interprétation*** du diagramme en boîte plutôt que son calcul.

Le diagramme quantile-quantile évalue si la distribution empirique est semblable à la distribution théorique. On peut, entre autres, évaluer la queue de la distribution. Selon la distribution, les quantiles considérés comme étant « normales » varient. Ci-dessous est un graphique montrant ce à quoi les quantiles devraient ressembler en fonction de la forme de la distribution.

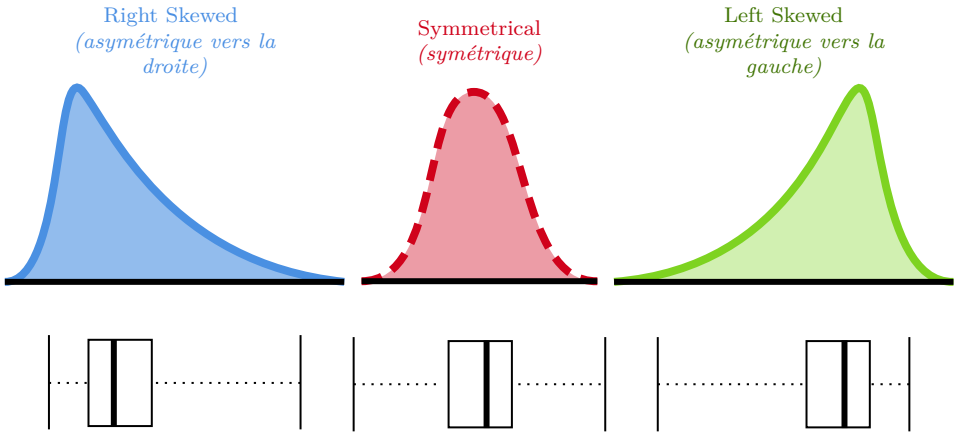
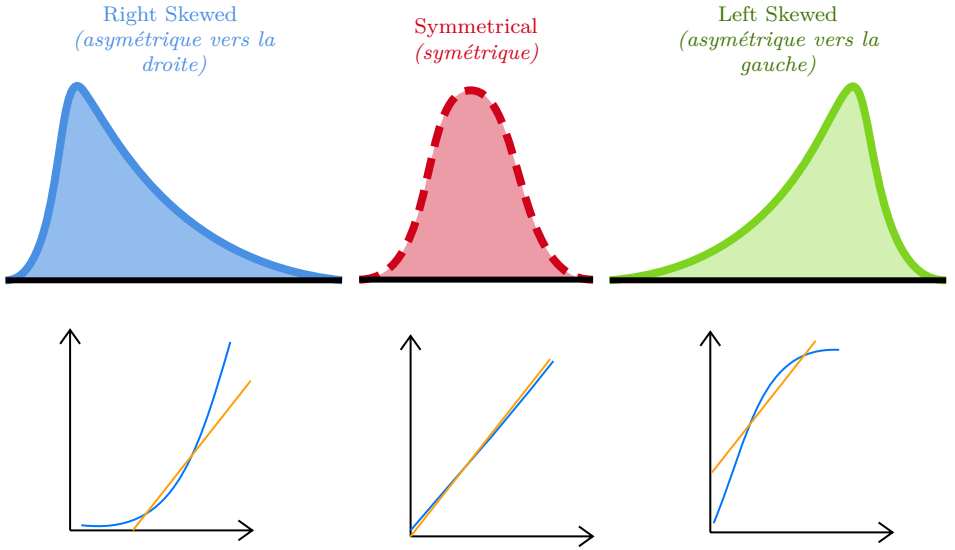


Diagramme quantile-quantile

Contexte

Le *Diagramme quantile-quantile (« Q-Q plot »)* est défini dans la section *Graphiques* du *chapitre de statistiques*.

Dans cette section, on décrit ***l'interprétation*** du diagramme en boîte plutôt que son calcul.



Régression linéaire simple

rewrite pour que ce soit moins une introduction et plus une intuition.

Contexte

On débute les méthodes de ***régression linéaire*** avec la **régression linéaire simple** qui utilise une variable explicative pour prédire une variable réponse quantitative.

Puis, on présente aussi les concepts fondamentaux et l'interprétation de la sortie d'un modèle de régression en le reliant aux tests d'hypothèse et les intervalles de confiance.

Contexte

La régression linéaire simple revient à prédire la variable réponse par une ligne droite. Puisque nous avons une fonction **linéaire**, on fait appel aux notions du secondaire pour obtenir une équation de la forme $f = \text{intercepte} + \text{pente} \times \text{variable explicative}$.

Définition du modèle

Notation

β_0 Paramètre d'intercepte.
β_1 Paramètre de pente.

Modèle de régression linéaire simple

On définit $Y = \beta_0 + \beta_1 x + \varepsilon$ sous *certains postulats*.

Postulats de la régression linéaire simple

On suppose que :

1. Y s'exprime en fonction de la variable explicative sous la forme $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $\forall i = 1, 2, \ldots, n$.
2. les réalisations x_i ne sont pas aléatoires $\forall i = 1, 2, \ldots, n$.

Puis, $\forall i = 1, 2, \ldots, n$, nous avons les postulats suivants sur ε_i :

$\mathbf{H}_1$ Linéarité : $\mathrm{E}[\varepsilon_i] = 0$.
$\mathbf{H}_2$ Homoscédasticité : $\mathrm{Var}(\varepsilon_i) = \sigma^2$.
$\mathbf{H}_3$ Indépendance : $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
$\mathbf{H}_4$ Normalité : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Des postulats, on déduit que valeurs observées de la variable réponse sont indépendantes et normalement distribuées avec $\mathrm{E}[Y_i] = \beta_0 + \beta_1 x_i$ et $\mathrm{Var}(Y_i) = \sigma^2$ telles que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.

Note L'**homoscédasticité** implique que la variance est constante pour toutes les observations.

Limitations

Par définition, le modèle pose qu'il y existe *systématiquement* une relation **linéaire** entre la variable réponse et la variable explicative. Donc, une variable explicative idéale pour une régression linéaire simple aura un important patron **linéaire** avec la variable réponse.

Il s'ensuit que le modèle de régression linéaire simple a des limitations inhérentes aux relations qu'il peut capturer. Si une variable réponse n'a pas de relation *linéaire* avec une variable explicative, on peut tenter de modéliser une fonction $g(\cdot)$ de la variable réponse Y. Par exemple, si Y a une forte relation *exponentielle* avec x, alors on peut modéliser $\log(Y)$ au lieu de Y pour obtenir une relation linéaire.

Estimation du modèle

Estimation des paramètres libres

Notation

b_0 Estimation du paramètre d'intercepte β_0.

b_1 Estimation du paramètre de pente β_1.

$\hat{y}$ Prévision de la variable réponse y.

Estimation du modèle de régression linéaire simple

Les prévisions $\hat{y}$ sont obtenues en fonction des estimations des paramètres : $\hat{y} = b_0 + b_1 x$.

Avec les données d'entraînement, on estime les paramètres libres par **la méthode des moindres carrés** qui minimise la somme des différences entre les valeurs observées et prédites de la variable réponse $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ afin d'obtenir les estimations $b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $b_0 = \bar{y} - b_1 \bar{x}$.

Note On peut récrire que la somme du produit des observations centrées de la variable explicative et de la variables réponse $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i$.

Note On peut récrire que la somme du carré des observations centrées de la variable explicative $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Estimation de la variance

Notation

MSE Erreur quadratique moyenne.

La racine du MSE, $\sqrt{MSE}$, est nommée l'**erreur type résiduelle** (« *residual standard error* »).

La variance de la valeur réponse, qui correspond à la variance de l'erreur de prévision, est estimée par l'erreur quadratique moyenne de prévision : $MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}$.

Note On divise par $n-2$ puisque 2 paramètres, β_0 et β_1, sont estimés. Le MSE de la section de *Précision des modèles d'apprentissage statistique* estime une mesure de la précision du modèle et donc diviser par n était suffisant. Cependant, ici on désire estimer un paramètre sans-biais et donc on doit diviser par $n-2$.

Représentation matricielle du modèle de régression linéaire simple

Contexte

Au lieu d'écrire que $\forall i = 1, 2, \ldots, n \; Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, on peut écrire l'expression sous forme matricielle avec $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$.

La représentation matricielle du modèle est superflue pour la régression linéaire simple puisqu'il y a seulement une variable explicative. Cependant, dès que nous en avons plusieurs et que nous obtenons une *Régression linéaire multiple* on doit utiliser la représentation matricielle.

On obtient que $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$. C'est-à-dire :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

où **la matrice d'incidence $\boldsymbol{X}$** est composée de 1 qui sont multipliés avec l'intercepte β_0 et de x_i qui sont multipliés avec la pente β_1.

Également, on peut récrire les estimations des paramètres en une matrice $\boldsymbol{b}$:

$$\boldsymbol{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix},$$

d'où $\boldsymbol{b} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}$.

Matrice de projection $\boldsymbol{H}$

La matrice $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top$ s'appelle la ***matrice de projection*** ou *matrice chapeau* (« *hat matrix* »).

Avec la matrice de projection, on obtient que $\hat{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{Hy}$.

Note Le nom « *hat matrix* » provient du fait que les prévisions de la variable réponse sont obtenues en multipliant les valeurs observées avec la matrice H—on met un chapeau sur $\boldsymbol{y}$.

Somme des carrés

Résidus e_i

Le i^{e} résidu est $e_i = y_i - \hat{y}_i$ ce qui correspond à la réalisation de ε_i.

Si le résidu est
positif la vraie valeur est **supérieure** à la valeur prédite.
négatif la vraie valeur est **inférieure** à la valeur prédite.

Nous minimisons l'EQM pour ajuster le meilleur modèle. Cependant, pour évaluer l'utilité du modèle, on partitionne la variabilité de la variable réponse en plusieurs sommes des carrés. On dénote la somme des carrés et la variance comme des fonctions, $SS(\cdot)$ et $\text{Var}(\cdot)$, de ce qui estime la variable réponse.

1 Total Sum of Squares (SST)

Contexte

La SST correspond à la somme des carrés si l'on prédit toujours la moyenne : $SS(\text{moyenne}) = (\text{données} - \text{moyenne})^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2$.
La variance $\text{Var}(\text{moyenne}) = \frac{(\text{données}-\text{moyenne})^2}{n-1} = s^2$ représente donc la variabilité des données autours de la moyenne.

On obtient que $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

2 Error Sum of Squares (SSE)

Contexte

La SSE correspond à la somme des carrés avec les valeurs prédites du modèle : $SS(\text{prévisions}) = (\text{données} - \text{prévisions})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
La variance $\text{Var}(\text{prévisions}) = \frac{(\text{données}-\text{prévisions})^2}{n-2} = MSE$ représente donc la variabilité des données autours de la droite de régression.

On obtient que $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. La SSE représente donc la variabilité qui n'est pas expliquée par la régression linéaire simple.

3 Regression Sum of Squares (SSR)

Contexte

Alors que la SST représente la variabilité totale et la SSE la variabilité qui n'est pas expliquée par le modèle, la SSR représente la variabilité qui ***est expliquée*** par le modèle.

On obtient que $SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$. Également, $SSR = b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$.

Note La SSE est parfois nommé la « *Sum of Squared Residuals* » et dénotée par RSS. Il est donc important de ne pas confondre SSR avec RSS.

En bref, on a que $SST = SSR + SSE$ où

Sum of Squares	Somme	Variabilité
SST	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	totale
SSR	$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	expliquée
SSE	$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	inexpliquée

Coefficient de détermination R^2

Contexte

Le R^2 est une mesure de corrélation qui correspond au carré du coefficient de corrélation : $R^2 = r_{Y,x}^2$. Il est plus facile d'interpréter le R^2 que le coefficient de corrélation de Pearson, particulièrement pour la régression linéaire.

Pour que la régression linéaire soit un bon modèle, on s'attend à ce qu'il y ait une forte dépendance entre la variable réponse Y et la variable explicative x. Si c'est le cas, $r_{Y,x}$ sera prêt de -1 ou de 1. Il s'ensuit que, pour obtenir un bon modèle, on désire **maximiser** le R^2.

On peut également définir le R^2 en fonction des sommes de carré : $R^2 = \frac{SS(\text{moyenne}) - SS(\text{prévisions})}{SS(\text{moyenne})}$. Donc, c'est le ratio de la variabilité expliquée par le modèle à la variabilité totale. Plus R^2 est élevé, plus le modèle explique la variabilité et le mieux qu'il est.

Le coefficient de détermination $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$. Également, $R^2 = b_1^2 \frac{s_X^2}{s_Y^2}$.

Note Voir *la définition du R_a^2* dans la section sur la *Régression linéaire multiple* pour les limitations du coefficient de détermination.

Note Voir *la définition du R_{ps}^2* dans la section sur la *Régression linéaire généralisée* pour l'application du concept du coefficient de détermination aux GLMs.

Estimateurs des paramètres

Estimateurs

En traitant les réalisations de la variable réponse comme une variable aléatoire et les réalisations de la variable explicative comme des constantes, b_0 et b_1 sont des combinaisons linéaires de Y et donc des estimateurs. De plus, nous nous intéressons à la statistique $\boxed{\mathrm{E}[Y] = \beta_0 + \beta_1 x}$ estimée par $\hat{Y}$.

Puisque les estimateurs sont des combinaisons linéaires de la v.a. normale Y, ils sont tous normalement distribués :

Estimateur	Distribution	Erreur type $se(\cdot)$
b_0	$\mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\right)$	$\sqrt{MSE\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$
b_1	$\mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$	$\sqrt{\frac{MSE}{\sum_{i=1}^n (x_i-\bar{x})^2}}$
$\hat{Y}$	$\mathcal{N}\left(\mathrm{E}[Y], \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\right)$	$\sqrt{MSE\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$

Note Il est important de distinguer le MSE qui estime $\mathrm{Var}(Y) = \sigma^2$ de $se(\hat{y})^2$ qui estime $\mathrm{Var}(\hat{Y})$.

Note Voir *la définition du VIF* de la section des *Hypothèses du modèle linéaire* pour sa relation avec l'erreur type des coefficients.

Matrice de variance-covariance

La matrice de variance-covariance de $\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ est $\boxed{\mathrm{Var}(\boldsymbol{b}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}}$ où

$$\mathrm{Var}(\boldsymbol{b}) = \begin{bmatrix} \mathrm{Var}(b_0) & \mathrm{Cov}(b_0, b_1) \\ \mathrm{Cov}(b_0, b_1) & \mathrm{Var}(b_1) \end{bmatrix}$$

Si σ^2 est inconnu, $\boxed{\widehat{\mathrm{Var}}(\boldsymbol{b}) = MSE\,(\boldsymbol{X}^\top \boldsymbol{X})^{-1}}$ où

$$\widehat{\mathrm{Var}}(\boldsymbol{b}) = \begin{bmatrix} \widehat{\mathrm{Var}}(b_0) & \widehat{\mathrm{Cov}}(b_0, b_1) \\ \widehat{\mathrm{Cov}}(b_0, b_1) & \widehat{\mathrm{Var}}(b_1) \end{bmatrix},$$

avec $\widehat{\mathrm{Var}}(b_0) = \mathrm{se}(b_0)^2$ et $\widehat{\mathrm{Var}}(b_1) = \mathrm{se}(b_1)^2$.

Bootstrapping

Contexte

Dans la section de *Simulation*, on a présenté la méthode de simulation Monte Carlo qui permet d'estimer l'erreur type comme l'écart-type d'un grand nombre d'estimations du paramètre d'intérêt. Cependant, cette méthode nécessite de connaître la vraie distribution de l'estimateur. Bien que l'on pose habituellement que la distribution est normale, ceci pourrait être une hypothèse erronée.

La méthode du « ***bootstrapping*** » permet d'éviter l'hypothèse de normalité ; elle extrait des échantillons aléatoires (avec remplacement) de l'ensemble de données originale. L'objectif est de créer plusieurs ensembles de données « *bootstrap* ».

Tests d'hypothèse

Test t bilatéral

S'il n'y a pas de relation linéaire entre la variable réponse Y et la variable explicative x, la pente β_1 sera nulle et $Y = \beta_0 + \varepsilon$. Donc, on effectue le test t bilatéral pour tester s'il y a une relation linéaire :

$$\mathrm{H}_0 : \beta_1 = 0 \qquad \mathrm{H}_1 : \beta_1 \neq 0$$

On fait un *test sur la moyenne* de l'estimateur b_1 avec $\boxed{t = \frac{b_1 - h}{se(b_1)}}$ où $\boxed{T \sim t_{(n-2)}}$. La zone critique est $\boxed{|t| \geq t_{\alpha, n-2}}$.

Note Puisque nous utilisons l'EQM et non la variance, la statistique est distribuée selon la loi de Student et non la loi normale. De plus, le calcul se base sur les n observations $e_1, \ldots, e_n$ qui sont soumis à 2 restrictions : $\boxed{\sum_{i=1}^n e_i = 0}$ et $\boxed{\sum_{i=1}^n x_i e_i = 0}$. Nous avons donc $(n-2)$ résidus **sans contraintes**.

Contexte

On peut également faire un test d'hypothèse unilatéral. Par exemple, on peut tester si la pente est **positive**, en supposant qu'elle ne l'est pas, avec le test unilatéral *vers la droite* $H_1 : \beta_1 > 0$ et $H_0 : \beta_1 \leq 0$. Également, on peut tester si la pente est **négative**, en supposant qu'elle ne l'est pas, avec le test unilatéral *vers la gauche* $H_1 : \beta_1 < 0$ et $H_0 : \beta_1 \geq 0$.

Test t unilatéral	Région critique
vers la gauche	$t \leq -t_{2\alpha,n-2}$
vers la droite	$t \geq t_{2\alpha,n-2}$

Intervalle de confiance et de prévision

Contexte

De façon générale, nous avons l'expression $\text{estimation} \pm (\text{percentile de la loi } t) \times (\text{erreur type})$.

Paramètre	Intervalle de confiance
β_0	$b_0 \pm t_{1-k,n-2} se(b_0)$
β_1	$b_1 \pm t_{1-k,n-2} se(b_1)$
$\mathrm{E}[Y]$	$\hat{y} \pm t_{1-k,n-2} se(\hat{y})$

Contexte

L'intervalle de confiance sur $\mathrm{E}[Y]$ prédit la **valeur moyenne** de $(Y|X=x)$. On peut cependant généraliser le concept d'intervalle de confiance pour *prédire* la vraie valeur de la variable réponse Y. Un ***intervalle de prévision*** trouve un intervalle de valeurs dans lequel la ***réalisation*** *d'une variable aléatoire* pourrait être contenue plutôt qu'un intervalle dans lequel un *paramètre* pourrait être contenu.

Intervalle de prévision

Contexte

Typiquement, l'intervalle de confiance est établit à partir de la distribution de la statistique. Cependant, pour un intervalle de prévision on pose que $Y_{n+1} - \hat{Y}_{n+1} \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]\right)$ avec $\hat{Y}_{n+1} = b_0 + b_1 x_{n+1}$.

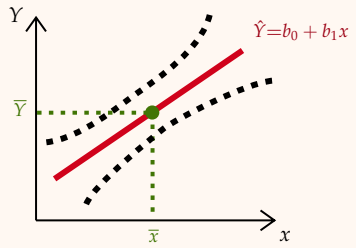
L'***intervalle de prévision*** est un intervalle de valeurs qui estime la valeur de la variable réponse pour la réalisation y_{n+1} d'une nouvelle observation Y_{n+1} : $y_{n+1} \in \hat{y}_{n+1} \pm t_{1-k,n-2} se\left(\hat{y}_{n+1}\right)$.

L'erreur type de l'intervalle de prévision $\sqrt{MSE\left[1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$ est presque identique à l'erreur type de l'intervalle de confiance $\sqrt{MSE\left[\frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right]}$. En fait, la distinction revient à distinguer 2 composantes :

1 « *Parameter risk* »

Le « *parameter risk* » est l'incertitude liée à l'estimation des paramètres. Il s'ensuit que cette composante fait partie des 2 erreurs type.

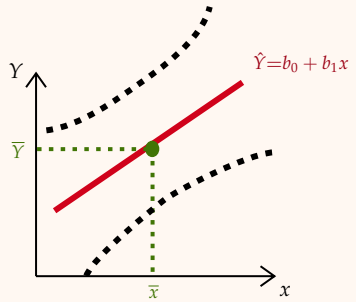
Visuellement, c'est l'intervalle des formes possibles pour la droite de régression estimée :



2 « *Process risk* »

Le « *process risk* » est l'incertitude liée à la fluctuation de la variable réponse auprès de sa moyenne. Il s'ensuit que cette composante fait seulement partie de l'erreur type sur la prévision.

Visuellement, l'impact du « *process risk* » sera d'élargir de $\sqrt{MSE}$ l'intervalle des deux bords :



Régression linéaire multiple

Contexte

La régression linéaire multiple généralise la régression linéaire simple en incluant p variables explicatives plutôt que juste une.

Définition du modèle

Notation

p Nombre de variables explicatives du modèle.
Pour inclure l'intercepte, on utilise souvent $p' = p + 1$.
β_j j^{e} coefficient de régression, $j \in \{0, 1, \ldots, p\}$.

Modèle de régression linéaire multiple

On définit $Y = \beta_0 + \beta_1 x_1 + \cdot + \beta_p x_p + \varepsilon$ sous certains postulats.

Ces postulats sont *les mêmes* que pour la régression linéaire simple, mais on ajoute la condition que le prédicteur x_j n'est pas une combinaison linéaire des p autres prédicteurs pour $j = 0, 1, \ldots, p$. Ceci évite qu'il y ait des variables ***redondantes*** dans le modèle.

Note Dans l'équation du modèle, on pose que $x_0 = 1$.

Estimation du modèle

Estimation des paramètres libres

Notation

b_j Estimation du j^{e} coefficient de régression, $j \in \{0, 1, \ldots, p\}$.

Estimation du modèle de régression linéaire simple

Les prévisions $\hat{y}$ sont obtenues en fonction des estimations des paramètres : $\hat{y} = b_0 + b_1 x_1 + \cdots + b_p x_p$.

Estimation de la variance

Pour la régression linéaire multiple, on récrit l'expression de l'EQM : $MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p'}$.

Note On divise par $n - p$ puisque p' paramètres, $\beta_0, \beta_1, \ldots, \beta_p$ sont estimés. La *Régression linéaire simple* comporte seulement deux paramètres β_0 et β_1 alors $p' = 2$.

Représentation matricielle du modèle de régression linéaire simple

Contexte

La représentation matricielle du modèle est essentielle pour la régression linéaire multiple puisqu'il y a p variables explicatives.

On généralise l'équation obtenue pour la *Régression linéaire simple* $\boldsymbol{Y = X\beta + \varepsilon}$ avec :

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Somme des carrés

Contexte

Les sommes des carrés ne changent pas en régression linéaire multiple. La seule différence est que l'on ne peut pas exprimer le coefficient de détermination en fonction de la corrélation puisqu'il y a plusieurs variables explicatives ($R^2 \neq r^2_{x,Y}$). Pour évaluer la qualité d'un modèle tout en tenant compte du nombre de paramètres, on définit une nouvelle mesure : le coefficient de détermination ajusté R^2_a.

Coefficient de détermination ajusté R^2_a

Contexte

Le désavantage du coefficient de détermination R^2 est qu'il va toujours augmenter lorsque nous ajoutons des variables explicatives. Plus nous avons des variables explicatives, mieux le modèle va prédire les observations (sur les données d'entraînement) et plus le R^2 sera élevé. La mesure ne considère donc pas l'utilité de ces variables explicatives, si elles valent la peine d'inclure dans le modèle.

Augmenter le nombre de prédicteurs équivaut à augmenter la complexité, ou *flexibilité*, du modèle. La section de *Précision des modèles d'apprentissage statistique* détaille comment que la variance augmente après un certain niveau de flexibilité. Le coefficient de détermination ajusté R^2_a considère donc ce compromis en utilisant les variances au lieu des sommes.

On définit le coefficient de détermination ajusté comme $R^2_a = 1 - \frac{SSE/(n-p')}{SST/(n-1)} = 1 - \frac{MSE}{s^2_y}$.

Également, on peut exprimer le R^2_a en fonction du R^2 : $R^2_a = 1 - (1-R^2)\left(\frac{n-1}{n-p'}\right)$. Puisque $\left(\frac{n-1}{n-p'}\right) > 1$, ce terme va gonfler la proportion de variabilité qui n'est pas expliquée par le modèle. Le résultat est que le $R^2_a < R^2$.

Note Contrairement au $R^2 \in [0,1]$, le $R^2_a \notin [0,1]$.

Variables explicatives spéciales

Termes d'ordre supérieur

Contexte

Le « linéaire » de régression linéaire ne provient pas de la linéarité des variables explicatives, mais plutôt de la linéarité des coefficients. Donc, on peut avoir un modèle avec des polynômes.

Régression linéaire avec polynôme d'ordre k

Soit une régression linéaire avec une seule variable explicative x_j, alors $Y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \cdots + \beta_k x_j^k + \varepsilon$.

Ce modèle suppose que la variable réponse est systématiquement reliée par un polynôme d'ordre k. On inclut aussi les polynômes de 1 à $k-1$ afin de mieux ajuster la forme de la droite de régression.

Note Voir la section de *Modèles additifs généralisés (GAM)* pour plus de détails sur la modélisation « non-linéaire ».

Variables « *dummy* »

Contexte

Si nous avons des variables explicatives catégoriques, nous devons les convertir en variables numériques pour appliquer le modèle de régression linéaire. Une variable « ***dummy*** » prend comme valeur 0 ou 1 et s'apparente à la fonction indicatrice $I(\cdot)$.

Variable « *dummy* »

Pour une variable catégorique, une variable « *dummy* » x_c est définie comme suit :

$$x_c = \begin{cases} 1, & \text{si } x_c = \text{catégorie } c \\ 0, & \text{si } x_c = \text{tout autre catégorie} \end{cases}$$

Pour une variable catégorique avec w catégories, nous devons utiliser $w-1$ variables « *dummy* ». Nous utilisons seulement $w-1$ et non w variables, car on définit une ***catégorie de base*** qui se réalise si toutes les variables « *dummy* » sont nulles.

Interaction de variables

Contexte

Pour modéliser la dépendance entre des variables explicatives dans le modèle on inclut une ***interaction*** qui équivaut au produit des variables explicatives.

Régression linéaire avec interaction

Soit une régression linéaire avec p variables explicatives et une interaction entre x_1 et x_2, alors $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \cdots + \beta_{p+1} x_p + \varepsilon$.

Ce modèle suppose que lorsqu'une des variable explicatives varie (p. ex. x_1), la variable réponse est impactée par son coefficient β_1 ***et*** le produit de x_2 avec un autre coefficient β_3. reliée par un polynôme d'ordre k. On inclut aussi les polynômes de 1 à $k-1$ afin de mieux ajuster la forme de la droite de régression.

Principe hiérarchique

Le ***principe hiérarchique*** stipule qu'une interaction significative implique que les termes individuels devraient être conservés dans le modèle, peu importe le résultat de leurs tests t. Lorsque l'interaction explique bien la variable réponse, la valeur des termes individuels n'importe peu. De plus, retirer les termes individuels pourrait changer l'interprétation et la signifiance de l'interaction.

Estimateurs des paramètres

Contexte

Pour la régression linéaire multiple, il n'y a pas de changements à apporter à la définition des estimateurs autre que le nombre de paramètres.

Test t

Contexte

Il y une différence entre le test t d'une régression linéaire multiple et le test t d'une *Régression linéaire simple* : la distribution de la statistique. Au lieu de suivre une loi de student de $n-2$ degrés de liberté, $T \sim t_{n-p'}$.

Cependant, l'interprétation est très différente et il est important de se rappeler qu'on effectue le test t sur **un coefficient à la fois**. Plutôt que tester s'il y existe une relation linéaire, on test si l'on peut ***simplifier le modèle*** d'un coefficient.

Note Si une variable catégorique est séparée en $w-1$ variables « *dummy* », on ne peut pas simplement retirer une des variables si elle n'est pas significative—le modèle n'aurait plus de sens. Plutôt, on doit examiner le modèle davantage pour comprendre les relations entre les variables.

Note Ne pas oublier le *principe hierarchique* pour un modèle ayant une interaction.

Test F

Contexte

La limitation inhérente au test t est qu'il peut seulement tester une variable à la fois. Pour tester l'importance de plusieurs variables explicatives, on doit utiliser un différent test. Cependant, avant de présenter ce test, on présente le ***tableau d'ANOVA*** et le test F.

tableau d'ANOVA

Le tableau de « *analysis of variance (ANOVA)* » élabore sur *le tableau résumé* des sommes des carrés présenté dans la section de *Régression linéaire simple*.

On défini la fonction $MS(\cdot) = \frac{SS(\cdot)}{df}$ pour trouver :

Source	Somme des carrés	Degrés de liberté	« *Mean square* »
régression	SSR	p	MSR
erreur	SSE	$n-p'$	MSE
totale	SST	$n-1$	s_y^2

La nouvelle mesure $MSR = \frac{SSR}{p}$ mesure la variance expliquée par la régression linéaire en moyenne par degré de liberté.

Note Bien que $SST = SSR + SSE$, $s_y^2 \neq MSR + MSE$.

Test F

Le test F test si l'on peut retirer **tous** les coefficients de régression sauf l'intercepte. Les hypothèses sont :

$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1 : au moins un $\beta_j \neq 0, j = 1, 2, \ldots, p$

Donc, on teste si le modèle nul est préférable au modèle de régression linéaire multiple. La statistique de test est $t = \frac{MSR}{MSE}$ où $T \sim F_{p,n-p'}$. On interprète p comme le nombre de degrés de liberté associé avec la régression et $n-p'$ comme le nombre de degrés de liberté associé avec l'erreur.

Puisque ce test est unilatéral vers la droite, on déduit de la *définition du test F* que l'on rejette l'hypothèse nulle si $t \geq F_{\alpha,p,n-p'}$.

Note Le test t est équivalent au test F pour la régression linéaire simple.

Contexte

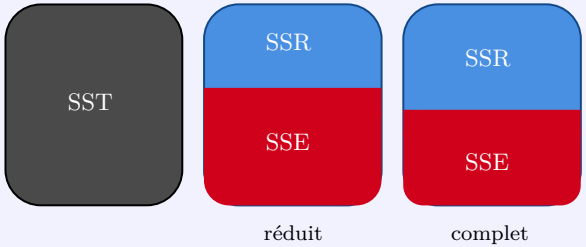
Une régression linéaire multiple avec au moins un prédicteur significatif devrait expliquer une grande proportion de la variabilité et donc avoir une valeur de MSR élevée. Le test F teste donc si le MSR est suffisamment grand relatif à le MSE. En fait, on peut interpréter le test F comme $F = \frac{\text{variabilité qui } \textbf{EST} \text{ expliquée par le modèle}}{\text{variabilité qui n'est } \textbf{PAS} \text{ expliquée par le modèle}}$.

Pour tester le retrait de **quelques** variables, on utilise le ***test F partiel***.

Test F partiel

Le test F partiel compare 2 modèles dont un qui a plus de coefficients (le modèle **complet**) que l'autre (le modèle **réduit**). L'objectif est de tester si un certain nombre de coefficients spécifiés expliquent suffisamment de variabilité pour être inclus dans le modèle.

Visuellement, on rejette l'hypothèse nulle que le modèle réduit est suffisant si l'aire en rouge est suffisamment réduite :



Il s'ensuit que la statistique $t = \frac{(SSE_r - SSE_f)/(p_f - p_r)}{SSE_f/(n-p'_f)}$ où SSE_r est la SSE du modèle **r**éduit et SSE_f du modèle complet (« ***f**ull* »). Également, $T \sim F_{p_f - p_r, n-p'_f}$.

Tout comme le test F, ce test est unilatéral vers la droite et on rejette l'hypothèse nulle si $t \geq F_{\alpha, p_f - p_r, n-p'_f}$.

Note Pour les **intervalles de confiance**, la seule différence en régressoin linéaire multiple est que la loi de Student a $n - p'$ degrés de liberté au lieu de $n - 2$.

ANOVA

Un facteur

Notation

w Nombre de niveaux, ou « traitements », du facteur.

n_j Nombre d'observations dans le j^{e} niveau.

$\sum_{i=1}^{j} n_j = n.$

$Y_{i,j}$ Variable réponse pour la i^{e} observation du j^{e} groupe où $i = 1, \dots, n_j$ et $j = 1, \dots, w$.

Contexte

L'ANOVA pour un GLM ayant une variable explicative (alias prédicteur, facteur) catégorique qui prédit la variable réponse s'appelle le modèle ANOVA « *one-factor* » ou « *one-way* ».

Modèle ANOVA pour un facteur

On pose que le niveau w est le niveau de base, puis on pose que $x_{i,j} = \boldsymbol{I}_{\left\{\begin{smallmatrix}\text{observation } i \text{ est}\\ \text{dans la catégorie } j\end{smallmatrix}\right\}}$ pour $j = 1, \dots, w-1$. Puis, on utilise la nouvelle notation $\boxed{Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}}$ pour $i = 1, \dots, n_j$ et $j = 1, \dots, w$. Ici, $\alpha_j = \beta_j$ et $\mu = \beta_0$.

Estimation

Contexte

Il est important de distinguer le i de $Y_{i,j}$ du i de Y_i auquel nous sommes habitués. Ici, on dénote une observation d'un échantillon des observations n_j et non de n en entier. Entre autre, cette nouvelle écriture implique que les $SS()$ doivent être modifiées.

Puisque nous cherchons à évaluer la variation par sous-groupe, on définie la moyenne par sous-groupe : $\boxed{\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{i,j}}{n_j}}$. Il s'ensuit que $\boxed{\hat{\mu} = \bar{y}_w}$ et que $\boxed{\hat{\alpha}_j = \bar{y}_j - \bar{y}_w}$ pour $j = 1, \dots, w-1$.

L'hypothèse nulle pour le modèle est que la réponse moyenne α_j pour chacun des w niveaux est égale et le tableau ANOVA est :

Source	Calcul	SS	ddl
Facteur	$\sum_{j=1}^{w} n_j(\bar{y}_j - \bar{y})^2$	SSR	$w-1$
Erreur	$\sum_{j=1}^{w}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y}_j)^2$	SSE	$n-w$
Total	$\sum_{j=1}^{w}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y})^2$	SST	$n-1$

Il s'ensuit que la statistique du *Test F* est la même avec $\boxed{t = \frac{MSR}{MSE}}$ où $\boxed{T \sim F_{p,n-p'}}$.

Deux facteurs

Notation

w, v Nombre de niveaux, ou « traitements », des facteurs A et B.
$Y_{i,j,k}$ Variable réponse pour la i^{e} observation du niveau $\{i, j\}$
α_j j^{e} **effet principal** pour le facteur A.
β_j k^{e} **effet principal** pour le facteur B.

Contexte

L'ANOVA pour un GLM ayant deux variables explicatives catégoriques s'appelle le modèle ANOVA « « *two-way* ».

Modèle ANOVA pour deux facteurs

Puis, on utilise la nouvelle notation $Y_{i,j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{i,j,k}$ pour $i = 1, \dots, n_*$, $j = 1, \dots, w$ et $k = 1, \dots, v$.

Modèle additif

Contexte

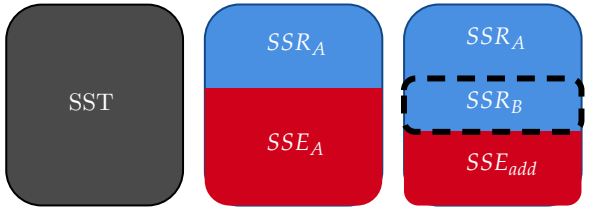
Dans le cadre de l'examen MAS-I, nous aurons seulement des ensembles de données balancés ayant n_* observations par niveau.

L'expression du modèle ayant 2 facteurs avec w et v niveaux s'écrit comme $Y_{i,j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{i,j,k}$ pour $i = 1, \dots, n_*$, $j = 1, \dots, w$ et $k = 1, \dots, v$.

Le tableau ANOVA du modèle additif s'écrit comme :

Source	SS	ddl
Facteur A	SSR_A	$w-1$
Facteur B	SSR_B	$v-1$
Erreur	SSE_{add}	$n-w-v+1$
Total	SST	$n-1$

Afin de bien comprendre l'impact d'un deuxième facteur sur l'analyse de la variance, on observe la visualisation de la relation $SSR_B = SSE_A - SSE_{add} = SSR_{add} - SSR_A$:



On peut tester deux hypothèses nulles. On peut tester l'hypothèse nulle que la réponse moyenne α_j pour chacun des w niveaux du facteur A est égale. Ceci équivaut à tester si le facteur A est significatif au modèle et la statistique de test est $t = \frac{SSR_A/(w-1)}{SSE_{add}/(n-(w+v-1))}$. On a que $T \sim F_{w-1,n-(w+v-1)}$

Sinon, on peut tester l'hypothèse nulle que la réponse moyenne β_k pour chacun des v niveaux du facteur B est égale. Ceci équivaut à tester si le facteur B est significatif au modèle et la statistique de test est $t = \frac{SSR_B/(v-1)}{SSE_{add}/(n-(w+v-1))}$. On a que $T \sim F_{v-1,n-(w+v-1)}$

Modèle avec interactions

Notation

$\gamma_{j,k}$ Interaction pour le niveau $\{j, k\}$.

L'expression du modèle ayant 2 facteurs et une interaction s'écrit comme $Y_{i,j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon_{i,j,k}$ pour $i = 1, \dots, n_*$, $j = 1, \dots, w$ et $k = 1, \dots, v$.

Le tableau ANOVA du modèle additif avec interactions s'écrit comme :

Source	SS	ddl
Facteur A	SSR_A	$w-1$
Facteur B	SSR_B	$v-1$
Interaction	SS_{diff}	$(w-1)(v-1)$
Erreur	SSE_{add}	$n-vw$
Total	SST	$n-1$

où $SS_{diff} = SSE_{add} - SSE_{int} = SSR_{int} - SSR_{add}$.

Puis, pour tester l'hypothèse nulle qu'il n'y a pas d'interactions (tous les $\gamma_{j,k}$ sont nuls) la statistique de test $t = \frac{SS_{diff}/((w-1)(v-1))}{SSE_{int}/(n-wv)}$ où $T \sim F_{(w-1)(v-1),n-wv}$.

Sinon, on peut tester chacun des facteurs de façon semblable à avant en substituant SSE_{add} par SSE_{int} au dénominateur.

Modèle additif sans réplication

Contexte

Le terme ***réplication*** implique qu'il y a plusieurs observations pour chacun des niveaux. Un modèle *sans* réplication implique qu'il y a une seule observation par niveau et que $n_* = 1$ et que $n = wv$.

Il s'ensuit que l'on peut simplifier le modèle sans l'indice i pour l'observation et que $Y_{j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k}$ pour $j = 1, \dots, w$ et $k = 1, \dots, v$.

De façon semblable au modèle avec un facteur, on définit la moyenne par sous-groupe avec $\bar{y}_{j\cdot} = \frac{\sum_{k=1}^{v} y_{j,k}}{v}$ et $\bar{y}_{\cdot k} = \frac{\sum_{j=1}^{w} y_{j,k}}{w}$.

Puis,

Source	Calcul	SS
Facteur A	$\sum_{j=1}^{w} v(\bar{y}_{j\cdot} - \bar{y})^2$	SSR_A
Facteur B	$\sum_{k=1}^{v} w(\bar{y}_{\cdot k} - \bar{y})^2$	SSR_B
Erreur	$\sum_{k=1}^{v}\sum_{j=1}^{w}(y_{j,k} - \bar{y}_{j\cdot} - \bar{y}_{\cdot k} + \bar{y})^2$	SSE_{add}
Total	$\sum_{k=1}^{v}\sum_{j=1}^{w}(y_{j,k} - \bar{y})^2$	SST

Autres

Modèle d'analyse de covariance (ANCOVA)

Contexte

Un modèle d'analyse de la covariance est un GLM ayant des prédicteurs quantitatifs et qualitatifs.

Total non-ajusté

Contexte

On appelle SST la « ***corercted*** *total sum of squares* » lorsque nous devons être plus spécifiques.

La $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ peut se décomposer en deux parties :

1. La SST ***non corrigée*** $\sum_{i=1}^{n} y_i^2$.
2. La ***correction*** de la SST $\sum_{i=1}^{n} \bar{y} y_i$.

Cette formulation permet d'écrire le tableau ANOVA sous la forme suivante :

Source	SS	ddl
Moyenne	$\sum_{i=1}^{n} \bar{y} y_i$	1
Régression	SSR	p'
Erreur	SSE	$n - p'$
Total	$\sum_{i=1}^{n} y_i^2$	n

Hypothèses du modèle linéaire

Contexte

On débute par *expliquer les conséquences* de ne pas respecter les hypothèses du modèle de régression linéaire. Également, les problèmes qui peuvent survenir avec les données. Puis, on détaille comment détecter ces problèmes et concluons en expliquant comment adresser ces problèmes.

Problèmes et enjeux

Il y a plusieurs problèmes qui peuvent survenir en régression linéaire :

1 Expression du modèle erronée

Cette problématique survient s'il est incorrect de poser qu'une fonction relie les variables explicatives à la variable réponse ou si on n'inclut pas les prédicteurs appropriés. Par exemple, ajuster un modèle linéaire alors qu'une relation polynomiale existe.

2 Résidus avec une moyenne non-nulle

Les résidus e, alias les réalisations des erreurs irréductibles ε, devraient être nuls en moyenne. Si la moyenne des résidus est loin d'être nulle, ça peut signaler qu'un aspect de la régression est erroné.

3 Hétéroscédasticité

L'hétéroscédasticité est l'inverse de l'homoscédasticité et implique que la variance des erreurs n'est pas constante pour toutes les observations. S'il y a hétéroscédasticité, cela implique que la MSE n'est pas fiable, car son utilité est fondée sur l'hypothèse qu'il y a une seule variance.

4 Erreurs corrélées

Les erreurs sont supposées aléatoires. Si ce n'est pas le cas, le comportement des erreurs serait prévisible d'une observation à l'autre et les covariances des observations Y ne seraient pas nulles.

S'il y a des erreurs dépendantes, les erreurs types seront sous-estimées et les valeurs p plus petites qu'elles devraient l'être. Ceci peut mener à un faux positif.

5 Erreurs non-normales

Si les erreurs ne suivent pas une distribution normale, on ne peut pas poser que les estimateurs suivent une loi t ou F. Il s'ensuit que de faire des tests d'hypothèse avec la mauvaise distribution mène à des conclusions mal fondées.

6 Multicolinéarité

Si un prédicteur est proche d'être une combinaison linéaire d'autres prédicteurs, il peut y avoir un problème de ***multicolinéarité***. L'impact est qu'il peut devenir difficile de déterminer quels coefficients sont importants. Ceci rend leur estimation instable, car la valeur estimée pourrait changer de façon importante d'un ensemble de données à un autre.

L'instabilité vient main en main avec des erreurs types plus élevés. L'impact est donc **sur l'interprétation** des coefficients et non **pas** sur la puissance prédictive du modèle, sur la fiabilité du MSE ni sur le résultats de tests F.

7 « *Influential points* »

Un ensemble données peut comporter des points « *influential* »—des observations ayant un impact important sur l'inférence du modèle. Il n'y a pas de méthode définitive pour mesurer l'influence, mais une observation peut être un point « *influential* » si elle est une donnée aberrante ou comporte un « *high leverage* ». Les données aberrantes et les points avec un « *high leverage* » sont donc des observations qui sont bizarres comparativement au reste des données.

Les données aberrantes (« *outliers* ») sont les observations avec un résidu extrême où la définition de « extrême » est relativement arbitraire. Ces points gonflent la SSE et doivent être évalués.

Les points ayant un « *high leverage* » sont les observations dont les prédicteurs prennent des valeurs inhabituelles. L'estimation des coefficients est très sensible aux valeurs bizarres d'observations et donc on évalue ces observations de près afin qu'elles ne causent pas de biais dans l'estimation des coefficients.

8 Dimensionalité élevée

La régression linéaire est conçue pour les ensembles de données où le nombre d'observations n est plus large que le nombre de prédicteurs p. Le sur-ajustement peut survenir pour des ensembles de données ayant beaucoup de dimensions (alias, p est trop grand).

Ce problème se résume par la ***malédiction de la dimensionalité***. Un ensemble de données avec beaucoup d'observations peut contenir beaucoup d'information. Cependant, une grande quantité de variables affaiblit la qualité des données.

Note Les 6 premier problèmes correspondent aux *postulats de la régression linéaire*.

Levier et résidus

Levier (« *leverage* »)

Le levier d'une observation mesure son impact dans la prévision de la variable réponse. Le levier de la i^{e} observation correspond à la i^{e} entrée diagonale de la *matrice de projection* $\boldsymbol{H}$. Pour la régression linéaire simple, on obtient que $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_j - \bar{x})^2}$.

Il s'ensuit que plus h_i est large, plus l'observation x_i est différente des autres. De plus, on obtient de l'expression que $h_i \in \left[\frac{1}{n}, 1\right]$ et que $\sum_{i=1}^{n} h_i = p'$. Une règle du pouce obtenue de cette condition est qu'une observation est **potentiellement aberrante** si son levier est plus que 3 fois le levier moyen : si $h_i > 3\left(\frac{p'}{n}\right)$.

Contexte

Le résidu e_i correspond à la différence entre la valeur de la variable réponse et sa prévision $y_i - \hat{y}_i$. Il s'ensuit que le résidu est *sensible à l'échelle* des valeurs de la variable réponse.

Nous devons donc standardiser les résidus. Cependant, diviser par la MSE serait erroné car il est possible que l'hypothèse d'homoscédasticité ne soit pas respectée. Plutôt, on doit utiliser une erreur type estimée. Bien qu'il y a plusieurs façons d'estimer cette erreur type, l'examen se concentre sur les deux façons les plus courantes.

Résidus standardisés

Les ***résidus standardisés*** pondèrent les résidus par une erreur type estimée qui gonfle la MSE par le réciproque du levier : $e_{sta,i} = \frac{e_i}{\sqrt{MSE(1-h_i)}}$.

Si le modèle est adéquat, les résidus standardisés sont *approximativement* distribués selon la **loi normale standard**.

Résidus studentisés

Les ***résidus studentisés*** pondère les résidus par une erreur type estimée qui gonfle une différente MSE par le réciproque du levier : $e_{stu,i} = \frac{e_i}{\sqrt{MSE^{(i)}(1-h_i)}}$.

La $MSE_{(i)}$ correspond à la MSE de la régression qui exclue la i^{e} observation.

Si le modèle est adéquat, les résidus studentisés sont distribués selon la **loi de Student** qui converge vers la loi normale standard.

Puisque les 2 résidus convergent vers une loi normale standard, on déduit de la règle du 68-95-99.7 que environ 95% des résidus seront entre -2 et 2, puis que environ 99.7% des résidus seront entre -3 et 3. La règle du pouce est donc que si la valeur absolue du résidu est supérieur à 3, l'observation est potentiellement une donnée aberrante : $e_{st,i} > 3$.

Pour évaluer le levier *et* les résidus en une seule mesure, on peut calculer soit « ***cook's distance*** » ou ***DFITS*** pour chaque observation.

Contexte

La mesure D**FF**IT (noter le double F et le manque de S) se définie comme la « ***diff**erence in **fit*** » : $DFFIT = \hat{y}_i - \hat{y}_i^{(i)}$. C'est donc la différence entre la valeur prédite du modèle $\hat{y}_i$ et la valeur prédite du modèle qui exclue la i^{e} observation $\hat{y}_i$.

Puis, la mesure D**FFITS** (noter le S) se définie comme la « ***S**tudentised DFFIT* » : $DFFITS = \frac{\hat{y}_i - \hat{y}_i^{(i)}}{\sqrt{MSE^{(i)}(1-h_i)}} = e_{stu,i}\sqrt{\frac{h_i}{1-h_i}}$.

Ces mesures servent à mesurer le degré d'influence qu'a une observation. Cependant, dans le cadre de l'examen MAS-I nous utilisons une mesure semblable, mais différente : DFITS. Les livres de référence ne donne pas de contexte sur cette mesure, c'est pourquoi j'ai présenté les deux mesures précédentes. La mesure DFITS, dont on note s'écrit avec un seul F, utilise le résidu ***standardisé*** $e_{sta,i}$ au lieu du résidu studentisé $e_{stu,i}$ du DFFITS.

DFITS

La mesure d'influence $DFITS = e_{sta,i}\sqrt{\frac{h_i}{1-h_i}}$.

Contexte

La mesure de « *Cook's distance* » est semblable aux autres. Afin de comprendre ce que la distance représente, on définie la mesure comme : $D_i = \frac{\left(\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_j^{(i)})^2\right)/p'}{\left(\sum_{j=1}^{n}(y_i - \hat{y}_i)^2\right)/(n-p')}$. On récrit ci-dessus d'autres écritures qui sont algébriquement équivalentes, mais plus faciles à calculer.

L'intuition est qu'on calcule le ratio des écarts moyens entre les prévisions du modèle complet et les prévisions du modèle qui exclut la i^{e} observation à la variance (MSE) du modèle. Pour comprendre l'intuition, on peut penser au « *cook's distance* » d'une façon semblable au coefficient de détermination ajusté.

« *Cooks's distance* »

La « ***Cook's distance*** » pour la i^{e} observation s'écrit de plusieurs façons :

1. En fonction du $DFITS$: $d_i = \frac{DFITS_i^2}{p'}$.
2. En fonction du résidu standardisé : $d_i = \frac{e_{sta,i}^2}{p'}\left(\frac{h_i}{1-h_i}\right)$.
3. En fonction du résidu : $d_i = \frac{e_i^2 h_i}{MSEp'(1-h_i)^2}$.

La règle du pouce est que la i^{e} observation est un « *influential point* » si d_i « *exceeds **unity*** ». C'est-à-dire, si $d_i > 1$.

Graphiques des résidus

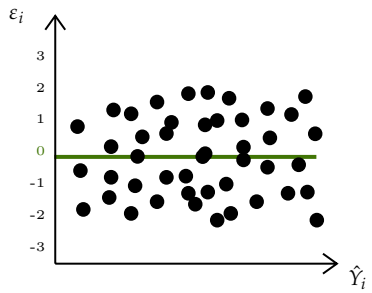
Contexte

Effectuer des tests diagnostiques sur les résidus nous permet d'identifier des problèmes potentiels et le non-respect des hypothèses du modèle. Puisque les résidus sont les réalisations des termes d'erreur, si leurs comportements diffèrent de ce qui est attendu on doit mettre en question la validité du modèle.

On présente 3 différents graphiques permettant d'analyser le patron général des résidus. Nous évaluons le comportement général et non individuel des résidus et donc les résidus que nous évaluons (non-ajustés, standardisés, studentisés) n'importe peu.

Graphique résidus contre prévisions

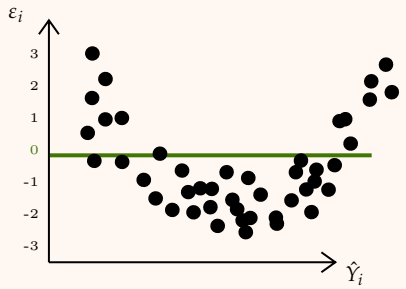
Voici un exemple d'un graphique ne présentant pas de problèmes :



Il y a 3 aspects que l'on cherche à vérifier visuellement :

1 Le patron des points

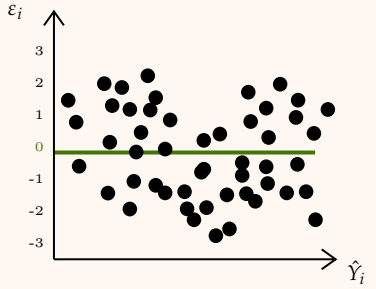
Comme ci-dessus, les **points** doivent sembler d'être **distribués de façon aléatoire** et ne **pas démontrer de patrons**. Par exemple, ce graphique démontre un patron indicatif d'une relation quadratique :



On observe que les résidus sont presque tous négatifs entre 5 et 15 et presque tous positifs avant 5 et après 15. Donc, si les prévisions sont bizarres de façon systématique, il est probable que l'équation du modèle ne soit pas bien spécifiée.

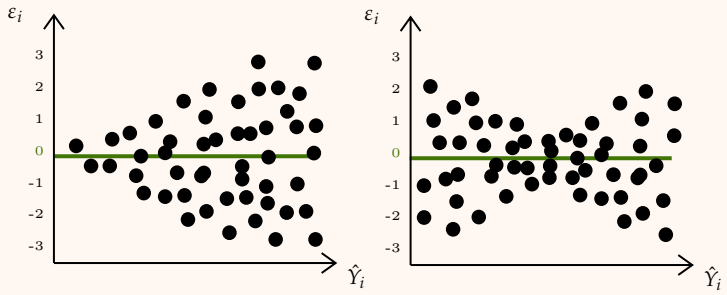
2 Moyenne des résidus non-nulle

L'hypothèse que la moyenne des résidus est nulle peut être mise en question **si on voit qu'une région du graphique a une moyenne non-nulle**. Par exemple, les observations au milieu du graphique suivant ont une moyenne différente de zéro :



3 Hétéroscédasticité

Le modèle de régression linéaire pose *l'homoscédasticité*. C'est-à-dire, que tous les termes d'erreur on la même variance. Cette hypothèse peut donc non-respectée si les résidus sont distribués de façon non-aléatoire et qu'on observe un patron dans les résidus. Typiquement, on utilise un entonnoir pour visualiser ; voici deux exemples :



Contexte

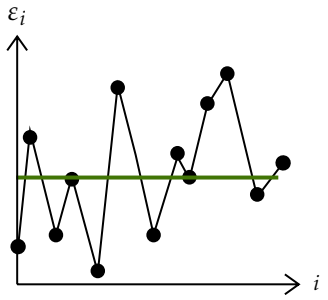
Il peut être difficile d'interpréter ces graphiques et il n'est pas évident de voir les patrons. Il est aussi important de voir comment le non-respect de certains postulats peut néanmoins respecter d'autres postulats. Par exemple, avoir de l'hétéroscédasticité n'empêche pas d'avoir une moyenne de zéro !

Il est important de lire l'interprétation des solutionnaires sur ces questions pour développer l'intuition nécessaire pour bien comprendre les graphiques. Également, de bien comprendre ce que représentent *les différents postulats* du modèle de régression linéaire.

Également, il est très important de ne pas oublier qu'on observe le comportement **général** des points. Alors, il ne faut pas se laisser influencer par une ou deux données aberrantes si le patron des observations en général semble adéquat.

Graphique résidus contre indice

Si les termes d'erreur sont indépendants, on **s'attend à ce que les résidus soient imprévisibles d'une observation à l'autre**. Voici un exemple d'un graphique ne présentant pas de problèmes :



Dans certains contextes, les termes d'erreur (particulièrement les termes adjacents) seront dépendants. Par exemple, ci-dessous les résidus d'une observation à l'autre ont tendance d'êtres semblables :

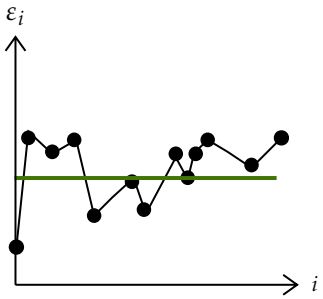
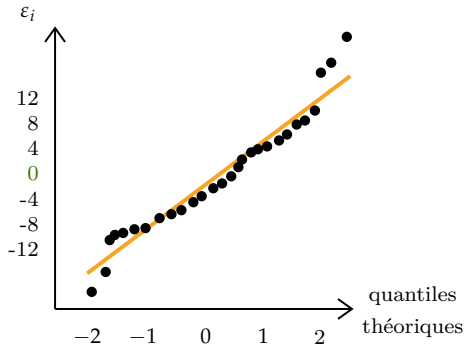


Diagramme quantile quantile des résidus

Le diagramme quantile quantile permet d'évaluer si la distribution des résidus semble être la même que celle d'une distribution normale standard. Voici un exemple d'un diagramme quantile quantile ne présentant pas de problèmes :



Bien que certains points sont loin de la droite, on observe le comportement **général** qui semble adéquat.

Facteur d'inflation de la variance (VIF)

Motivation

Pour évaluer la multicolinéarité, on cherche à détecter si un prédicteur est une combinaison linéaire des autres. Afin de quantifier ceci, on rappelle que le coefficient de détermination mesure le degré auquel un modèle explique une variable réponse. Pour évaluer la multicolinéarité, on applique le R^2 avec les prédicteurs comme variable réponse. C'est-à-dire, on calcule p coefficients de détermination, un à la fois, en posant que la variable réponse est x_j. On dénote ces coefficients $R^2_{(j)}$ pour $j = 1, 2, \ldots, n$.

Cependant, il est difficile de comparer les différents $R^2_{(j)}$ puisque l'échelle des valeurs du coefficient est si petite. Nous voulons donc **changer l'échelle de** $R^2_{(j)}$ afin d'augmenter les différences entre ses valeurs. On définit donc le VIF comme l'*inverse* du *réciproque* de $R^2_{(j)}$.

Facteur d'inflation de la variance (VIF)

Le facteur d'inflation de la variance (***VIF***) est donc une **mesure de la *multicolinéarité***. Le VIF pour le j^{e} prédicteur x_j est calculé en

1. effectuant une régression linéaire multiple avec x_j comme variable réponse prédite par les $p-1$ autres prédicteurs ;
2. calculant le coefficient de détermination de cette régression dénoté $R^2_{(j)}$;
3. calculant $VIF_j = \frac{1}{1-R^2_{(j)}}$.

Si le VIF est élevé, il est probable qu'il y a un problème de multicolinéarité. L'implication d'un VIF élevé peut être clarifié en notant que l'erreur type du j coefficient de régression $se(\hat{\beta}_j) = \sqrt{VIF_j}\sqrt{\frac{MSE}{(n-1)s^2_{x_j}}}$.

Donc, comme noté dans la section de *Problèmes et enjeux*, un VIF élevé implique des coefficients instables. On voit ici que c'est en raison de l'erreur type des coefficients qui est élevée.

La règle du pouce est qu'on devrait s'inquiéter d'un problème de multicolinéarité si le $VIF_j > 5$ ou (de façon équivalente) si $R^2_j = 0.80$. Cependant, le VIF comme toute autre mesure n'est qu'une *indication* de problèmes potentiels qu'il faut que nous devons analyser davantage.

Note En anglais, « *variance inflation factor (VIF)* ».

Résolutions potentielles

Dans le cas d'**hétéroscédasticité**, on veut stabiliser la variabilité. Par exemple, pour des résidus qui prennent la forme d'un entonnoir on peut transformer la variable réponse avec une **fonction concave** telle que le logarithme naturel ou la racine. La variable réponse est donc $\ln(Y)$ ou $\sqrt{Y}$ plutôt que Y.

Typiquement, si les **résidus sont dépendants** alors les données sont temporelles. Il faut donc utiliser un modèle de *« Branching Process »* pour la variable réponse.

Si les **termes d'erreur ne sont pas normalement distribués** alors la variable réponse est discrète. Il faut donc utiliser un modèle qui utilise une autre distribution que la distribution normale.

Si on identifie un **problème de multicolinéarité** avec des coefficients, on peut mitiger le problème avec une de ces deux approches :

1. exclure tous sauf un des prédicteurs.
2. combiner tous les prédicteurs en un.

Une méthode n'est pas supérieure à l'autre, ça dépend du contexte. Dans certains cas, on peut simplement signaler le problème sans retirer les prédicteurs. Par exemple, si on est seulement intéressés aux autres résultats du modèle qui ne sont pas touchés par ces prédicteurs. Sinon, on peut utiliser des **prédicteurs orthogonales**.

Deux prédicteurs sont ***orthogonales***, alias *non-corrélés* ou *perpendiculaires*, si leur produit scalaire est nul : $\sum_{i=1}^{n} x_{i,1}x_{i,2} = 0$. Un prédicteur est orthogonale à un ensemble de prédicteurs si son $VIF_j = 1$ la valeur minimale ($R^2_{(j)} = 0$).

Sélection du modèle

Sélection de sous-ensemble

Notation

g Nombre **total** de prédicteurs considéré.
p Nombre de prédicteurs du modèle où $p \leq g$.
M_p Meilleur modèle parmi les $\binom{g}{p}$ modèles possibles contenant p prédicteurs.

Algorithme de sélection « *best subset* »

1. Pour $p = 0, 1, \ldots, g$,
   (a) ajuster les $\binom{g}{p}$ modèles ayant p prédicteurs ;
   (b) poser que M_p est le modèle ayant le plus gros R^2.
2. Choisir le meilleur modèle parmi $M_0, M_1, \ldots, M_g$ selon un critère de sélection tel que le R_a^2.

Note Il ne serait pas adéquat d'utiliser le R^2 à la deuxième étape puisque les modèles ont tous un nombre différent de paramètres.

Limitation

La méthode de sélection du meilleur sous-ensemble considère tous les 2^g modèles. Puisque chaque prédicteur est dans le modèle ou il ne l'est pas, ou obtient qu'il y a $2 \times 2 \times \cdots \times 2 = 2^g$ différents modèles possible. Il s'ensuit que son intensité de calcul croît rapidement lorsque le nombre total de paramètres g augmente.

Méthodes de sélection « *stepwise* »

Algorithme de sélection « *Forward* »

1. Ajuster g modèles de régression linéaire simple, alias g modèles avec un seul prédicteur.
2. Poser que M_1 est le modèle ayant le plus gros R^2.
3. Pour $p = 2, 3, \ldots, g$,
   (a) ajuster les modèles qui ajoutent un prédicteur au modèle M_{p-1} ;
   (b) poser que M_p est le modèle ayant le plus gros R^2.
4. Choisir le meilleur modèle parmi $M_0, M_1, \ldots, M_g$ selon un critère de sélection tel que le R_a^2.

Algorithme de sélection « *Backward* »

1. Ajuster le modèle ayant g prédicteurs M_g.
2. Pour $p = g-1, g-2, \ldots, 1$,
   (a) ajuster les modèles qui retirent un prédicteur du modèle M_{p+1} ;
   (b) poser que M_p est le modèle ayant le plus gros R^2.
3. Choisir le meilleur modèle parmi $M_0, M_1, \ldots, M_g$ selon un critère de sélection tel que le R_a^2.

Limitations

La méthode de sélection ascendante est une méthode dite *gloutonne* (« ***greedy*** »), car elle ne fait qu'ajouter le meilleur prédicteur alors que le nombre de prédicteurs P augmente plutôt que trouver le meilleur sous-ensemble de prédicteurs.

L'application de la méthode de sélection descendante peut être limitée. Lorsque g est élevé il est possible que les modèles aient des problèmes de *dimensionalité*. Les modèles pour lesquels $n \leq p+1$ ne sont pas valides et même ceux pour lesquels $n \approx n-2$ peuvent être surajustés. Donc, puisque la « *backward selection* » commence avec M_g elle ne peut pas être utilisée.

Ces méthodes de sélection « *stepwise* » forcent les modèles $M_1, \ldots, M_g$ d'être des modèles *emboîtés*. De plus, les deux méthodes examinent $1 + \frac{g(g+1)}{2}$ modèles au lieu de 2^g. Il s'ensuit que le modèle choisit n'est pas nécessairement le meilleur et donc qu'on échange de la précision pour de l'efficacité.

Algorithme de sélection hybride

1. Débuter avec le modèle nul M_0.
2. Ajouter le prédicteur pour lequel l'ajout au modèle a la plus petite valeur p, pour un test t bilatéral, si la valeur p est au delà d'un seuil minimal.
   Si aucun modèle a une valeur p au-delà du seuil minimal, ne pas ajouter un prédicteur.
3. Retirer du modèle le prédicteur ayant la plus grande valeur p, pour un test t bilatéral, si la valeur p est au-delà d'un seuil maximal.
4. Répéter les étapes 1 à 3 jusqu'à ce que le modèle se stabilise.

Contexte

L'avantage de l'algorithme hybride est qu'il traite les limitations des algorithmes précédents. Cependant, il y a également des problèmes tel que choisir le seuil critique des tests t bilatérales.

Critère de sélection

Notation

MSE_g L'EQM du modèle utilisant tous les g prédicteurs.

Contexte

Les critères de sélection permettent de choisir le modèle. Dans les algorithmes le R_a^2 est mentionné, cependant ont en présente 4 autres dans les deux prochaines sections.

Cette section présente les 3 premières qui sont semblables à l'EQM d'entraînement, sauf qu'elles incluent une correction pour la flexibilité du modèle. La motivation est d'avoir une mesure qui peut mener à des conclusions semblables que celles obtenues de l'EQM de test sans les données de test. Cependant, la 4^{e} mesure qui utilise la validation croisée est une estimation plus directe de l'EQM de test.

Note Voir la section sur la *Précision des modèles d'apprentissage statistique* pour la définition de l'EQM de test et d'entraînement.

C_p de Mallow

Le C_p de Mallow est définit comme $C_p = \frac{SSE + 2p \times MSE_g}{n}$.

La formule, qui peut être décomposée comme $C_p = SSE/n + \left(2\frac{MSE_g}{n}\right) p$, équivaut au SSE moyen plus une pénalité par prédicteur. Puisque la SSE d'entraînement moyen décroît lorsque le nombre de prédicteurs p augmente, le C_p augmente si la réduction ne compense pas pour la pénalité.

Note Ce que l'indice p de C_p représente n'est pas clair. Donc, on écrit toujours C_p sans jamais substituer de valeur pour p.

Note Voir les *Critères d'information pour la sélection de modèles* du chapitre de *Erreur* qui contient la définition de l'AIC et du BIC avec la vraisemblance pour la sélection de modèles. Il s'ensuit que **la définition des mesures dépend du contexte**.

Critère d'information d'Akaike (AIC)

En **régression linéaire multiple**, $AIC = \frac{SSE+2p\times MSE_g}{n\times MSE_g}$.

L'AIC ne fait donc que changer l'échelle du C_p et choisira le même modèle.

Critère d'information bayésien (BIC)

En **régression linéaire multiple**, $BIC = \frac{SSE+\ln(n)\times p\times MSE_g}{n\times MSE_g}$.

Le BIC modifie donc la pénalité afin de tenir en compte le nombre d'observations. L'effet est que lorsque le nombre d'observations $n \geq 8$ le BIC va privilégier des modèles plus simples ayant moins de prédicteurs.

Note La règle du pouce de $n \geq 8$ provient du fait que 8 est le première nombre pour lequel $\ln(n) \geq 2$.

Motivation

L'avantage de ces mesure au R_a^2 est qu'elles ont des justifications théoriques. Cependant, tout comme le R_a^2, elles sont fonction de l'EQM d'entraînement et pas fiables pour des modèles surajustés.

Rééchantillonnage

Contexte

Les *méthodes de rééchantillonnage* consiste à ajuster un modèle plusieurs fois sur différents sous-échantillons de l'ensemble de données. Par exemple, on peut ajuster un modèle de régression linéaire sur plusieurs sous-échantillons d'un ensemble de données et comparer les modèles. L'idée est de séparer les données afin d'en utiliser une partie pour entraîner le modèle, puis une autre pour le tester.

Les deux méthodes couvertes dans le cadre de l'examen sont la ***validation croisée*** et le ***bootstrap***. La première méthode consiste à séparer l'ensemble de données en plusieurs sous-ensembles, puis la deuxième de tirer des échantillons avec remplacement de l'ensemble de données.

Le cas le plus simple de la validation croisée par k ensembles est d'avoir $k = 1$ ensemble—ceci correspond à l'ensemble de validation. Le cas le plus complexe tant qu'à lui est d'avoir $k = n$ sous-ensembles—ceci correspond au « *leave-one-out cross-validation* ».

Ensemble de validation (« *validation set* »)

Contexte

Avec l'ensemble de validation, on sépare les données en deux ensembles : les données d'entraînement et les données de test. Ceci s'apparente donc aux échantillons d'entraînement et de test mentionnés pour la première fois dans la section d'*Estimation du modèle* de la *Régression linéaire simple*.

La méthode ajuste le modèle avec **seulement** les n_1 données d'entraînement puis calcule l'EQM de test avec seulement les $n_2 = n - n_1$ données de test. Cette EQM de test estimée correspond à l'***erreur de l'ensemble de validation***.

Algorithme de validation croisée classique

1. Ajuster g modèles sur les n_1 données d'entraînement avec n'importe quelle méthode de sélection (p. ex. la sélection par *meilleur sous-ensemble*).
2. Déterminer le p correspondant au meilleur modèle parmi $M_0, M_1, \ldots, M_p$.

C'est-à-dire, le modèle ayant la plus faible erreur de l'ensemble de validation.

3. Utiliser la méthode de sélection utilisée précédemment pour ajuster tous les modèles comportant p prédicteurs sur **toutes** les observations, puis choisir le meilleur.

   Ceci permet de mieux ajuster les coefficients du modèle.

   Le modèle choisit n'aura pas nécessairement les 3 mêmes prédicteurs que celui choisit à l'étape 2.

Limitations

La validation croisée avec un ensemble de validation a deux enjeux dont il faut tenir en compte. Premièrement, puisque les données sont segmentées en deux de façon aléatoire, les **résultats sont volatiles** et peuvent varier en fonction de l'échantillon choisit. De plus, puisque nous utilisons seulement n_1 observations pour ajuster le modèle, le modèle est moins bien ajusté aux données et **l'erreur de l'ensemble de validation *surestime* l'EQM de test**.

Validation croisée par k sous-ensembles (« *k-fold validation* »)

Algorithme de validation croisée par k sous-ensembles

La validation croisée par k sous-ensembles divise de façon aléatoire toutes les observations en k sous-ensembles (« *folds* ») d'environ la même taille. Puis, on ajuste le modèle k fois selon l'algorithme suivant :

1. Pour $v = 1, \ldots, k$,
   (a) ajuster le v^{e} modèle avec toutes les données sauf celles dans le v^{e} sous-ensemble.
   (b) utiliser le v^{e} modèle pour calculer l'EQM de test avec les données du v^{e} sous-ensemble.
2. Calculer l'***erreur de validation croisée*** : la moyenne des k EQM de test calculées à l'étape 1.

   En anglais, « *CV error* ».

Algorithme de sélection par validation croisée par k sous-ensembles

1. Pour $p = 0, \ldots, g$ (ou $p = g, \ldots, 0$ pour la sélection « *backward* ») appliquer l'algorithme de validation croisée par k sous-ensembles.
2. Déterminer le p qui correspond au modèle ayant la plus faible erreur de validation croisée.
3. Utiliser la méthode de sélection utilisée précédemment pour ajuster tous les modèles comportant p prédicteurs sur **toutes** les observations, puis choisir le meilleur.

Contexte

La validation croisée par k sous-ensembles répond aux enjeux soulevés avec l'ensemble de validation. La sélection du nombre de prédicteurs se base sur la moyenne de plusieurs modèles ce qui permet de **diminuer la volatilité des résultats**. De plus, puisqu'on utilise la majorité des données pour ajuster chacun des modèles, les résultats sont moins biaisés.

Donc, on peut utiliser la validation croisée pour ***estimer l'erreur de test*** afin d'évaluer la performance d'un modèle ou pour ***choisir un niveau adéquat de flexibilité***. Il y a cependant un compromis à faire entre le biais et la volatilité, ou variance, des résultats. On développe sur cette idée avec la validation croisée par n sous-ensembles, alias la « *leave-one-out cross-validation (LOOCV)* ».

« *Leave-one-out cross-validation (LOOCV)* »

Contexte

Comme mentionné ci-dessus, la « *leave-one-out cross-validation (LOOCV)* » est un cas spécial de la validation croisée par k sous-ensembles lorsque $k = n$. Ceci implique que chacun des sous-ensembles comporte une seule observation. Il s'ensuit que l'intensité de calcul du LOOCV croît rapidement alors que le nombre d'observations n augmente puisque nous devons ajuster n modèles.

Erreur LOOCV

L'erreur LOOCV est la moyenne des résidus calculés à chaque étape comme l'EQM de test. Cependant, puisque l'erreur est calculée par observation on

peut simplifier son calcul comme $\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{1-h_i}\right)^2$.

Limitations

Plus le nombre de sous-ensembles augmente, plus les sous-ensembles de données se chevauchent, plus la variance dans l'estimation de l'EQM de test augmente. En revanche, le biais dans la prévision de l'EQM de test diminue. Par exemple, L'algorithme LOOCV aura le moins de biais, mais le plus de volatilité ! Le *Compromis biais-variance* est donc à considérer ici aussi.

La règle du pouce est d'utiliser $k = 5$ ou $k = 10$ sous-ensembles en validation croisée.

Le bootstrap

Contexte

Le bootstrap est applicable dans divers contexte. Notamment, on l'utilise pour évaluer la précision des estimations des paramètres d'un modèle.

Méthodes de régression alternatives

Contexte

Cette section discute les alternatifs à la régression linéaire multiple avec les « *shrinkage methods* » et les « *dimension reduction methods* ». Ces méthodes peuvent nécessiter de standardiser les variables.

Standardisation de variables

Une variable :

centrée	réduite	standardisée
$x - \bar{x}$	$\frac{\bar{x}}{\hat{\sigma}}$	$\frac{x - \bar{x}}{\hat{\sigma}}$

Note Par défaut, on utilise la variance empirique biaisée plutôt que la variance échantillonnale. Cependant, l'impact d'utiliser un ou l'autre est minime.

Contexte

Ces transformations de variables ont un impact sur l'estimation des coefficients d'une régression linéaire multiple. Si on centre les variables, l'intercepte devient équivalent à la moyenne empirique. Si on réduit les variables, l'intercepte reste pareille, mais les coefficients sont multipliés par l'écart-type de la variable explicative.

Par contre, en développant l'expression de $\hat{y}$, on trouve qu'elle est équivalente à l'expression originale. Donc, ces transformations ne changent pas les prévisions pour la régression linéaire puisque l'estimation est basée sur la **méthode des moindres carrés**. Cependant, si l'estimation est basée sur une ***autre méthode*** alors ces transformations **peuvent** avoir un impact !

Méthodes de réduction de la dimensionalité

Contexte

La méthode des moindres carrés (« *ordinary least squares (OLS)* ») estime les coefficients de régression en minimisant la $SSE = \sum_{i=1}^{n} \left(y_i - b_0 - b_1 x_{i,1} - \cdots - b_p x_{i,p}\right)^2$. Par le *Compromis biais-variance*, augmenter p (la flexibilité) diminue le biais de la méthode d'OLS, mais augmente sa variance.

Une méthode de diminuer la variance, et d'accorder un peu de biais, est de *restreindre* les valeurs prises par les estimations des coefficients.

Notation

a « *budget parameter* ».

λ « *tuning parameter* ».

Norme ℓ_1

La norme ℓ_1 du vecteur colonne $\boldsymbol{b}$ est $\|\boldsymbol{b}\|_1 = \sum_{j=1}^{p} |b_j|$.

Norme ℓ_2

La norme ℓ_2 du vecteur colonne $\boldsymbol{b}$ est $\|\boldsymbol{b}\|_2 = \sqrt{\sum_{j=1}^{p} b_j^2}$.

Régression Ridge

Contexte

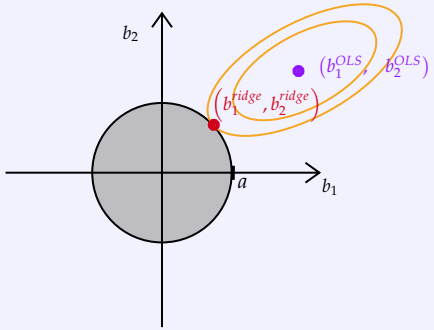
La régression Ridge va limiter tous les coefficients (sauf l'intercepte), sans toutefois éliminer de prédicteurs. Elle est donc utile lorsque nous avons une dimensionalité élevée et que seulement certains des prédicteurs sont importants.

Régression Ridge

La méthode de ***régression Ridge*** minimise également la SSE avec la contrainte additionnelle que $\sum_{j=1}^{p} b_j^2 \leq \alpha$.

Visualisation pour $p = 2$

Afin de comprendre l'idée sous-jacente à la régression Ridge, on trace le dessin suivant :



Le point mauve correspond au point qui minimise la SSE ; ses coordonnées correspondent donc aux estimations des paramètres b_1 et b_2 par la méthode d'OLS. Les ronds en orange représentent ce point plus un nombre (p. ex. $50, 100, \ldots$) correspondant à la pénalité additionnelle du modèle.

Le cercle gris correspond à la contrainte que $b_1^2 + b_2^2 \leq \alpha$. La pénalité choisie va donc correspondre au rond qui intersecte le cercle au point rouge dont les coordonnées sont les estimations des paramètres b_1 et b_2 par la méthode de régression Ridge.

On peut également voir pourquoi la régression Ridge réduit les coefficients, car le cercle les force de tendre vers zéro.

La régression Ridge va donc chercher à minimiser $\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - \cdots - b_p x_{i,p})^2 + \lambda \sum_{j=1}^{p} b_j^2$ où λ est le « *tuning parameter* » qui contrôle la réduction. Également, les variables explicatives doivent être **réduites** afin que les coefficients soient sur la ***même échelle*** pour la contrainte.

Le paramètre λ est inversement lié à la flexibilité ; ***augmenter*** λ ***diminue la flexibilité***. Si $\lambda = 0$, la régression Ridge équivaut à la méthode d'OLS alors que si $\lambda \to \infty$, la régression Ridge tend vers le modèle nul. Le processus de trouver la meilleur valeur de λ s'appelle le « ***tuning*** » et se fait via la *Rééchantillonnage*.

Note Augmenter λ implique de diminuer la pénalité, mais les coefficients individuels estimés b_j pourraient toutefois augmenter (en valeur absolue).

Régression Lasso

Contexte

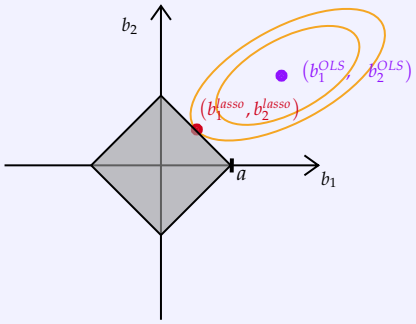
La distinction de la méthode de régression Lasso à la méthode de régression Ridge est que Lasso peut éliminer des prédicteurs plus aisément. Une utilité de la régression Lasso peut donc être d'éliminer des prédicteurs.

Régression Lasso

La méthode de ***régression Lasso*** minimise également la SSE avec la contrainte additionnelle que $\sum_{j=1}^{p} |b_j| \leq \alpha$.

Visualisation pour $p = 2$

Afin de comprendre l'idée sous-jacente à la régression Ridge, on trace le dessin suivant :



Le diamant gris correspond à la contrainte que $|b_1| + |b_2| \leq \alpha$. La pénalité choisie va donc correspondre au rond qui intersecte le diamant au point rouge dont les coordonnées sont les estimations des paramètres b_1 et b_2 par la méthode de régression Lasso.

On peut également voir pourquoi la régression Lasso peut éliminer des coefficients, car la forme du diamant peut impliquer que le point le plus près est sur une des axes sans nécessiter que $\lambda \to \infty$.

La régression Lasso va donc chercher à minimiser $\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - \cdots - b_p x_{i,p})^2 + \lambda \sum_{j=1}^{p} |b_j|$ où λ est le « *tuning parameter* » qui contrôle la réduction. Également, les variables explicatives doivent être **réduites** afin que les coefficients soient sur la ***même échelle*** pour la contrainte.

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Analyse et régression en composantes principales

Contexte

Dans la section *Statistiques*, on décrit une statistique comme une fonction qui résume n v.a. en une seule valeur. La même idée s'applique pour l'***analyse en composantes principales*** (PCA) où l'on résume p variables en un nombre inférieur de *nouvelles* variables.

Le PCA permet donc de représenter les données en moins de dimensions sans perdre une grande proportion des données. Après avoir expliquée comment fonctionne l'analyse en composantes principales, on explique comment utiliser ces nouvelles variables pour faire une régression linéaire multiple—une ***régression en composantes principales*** (PCR).

Pour ce faire, on pose que toutes les variables explicatives $x_1, \ldots, x_p$ sont ***centrées***.

Analyse en composantes principales (PCA)

Notation

z_m m^{e} composante principale.
$\phi_{j,m}$ j^{e} « *loading* » de la m^{e} composante principale.
$z_{i,m}$ Score de la m^{e} composante principale pour la i^{e} observation.

Composantes principales

Les ***composantes principales*** du PCA s'apparentent à une statistique T_n ; ce sont des fonctions, ou *combinaisons linéaires*, des p variables explicatives et des p « *loadings* » de la m^{e} composante : $z_m = \sum_{j=1}^{p} \phi_{j,m} x_j$.

Scores

On peut calculer le ***score*** pour la i^{e} observation de la m^{e} composante principale après avoir calculé les « *loadings* » comme $z_{i,m} = \sum_{j=1}^{p} \phi_{j,m} x_{i,j}$.

Les composantes principales sont obtenues de façon itérative. Pour obtenir les « *loadings* », on maximise la *variance empirique* de la composante z_m, $\frac{\sum_{i=1}^{n} z_{i,m}^2}{n}$ sous la contrainte que $\sum_{j=1}^{p} \phi_{j,m}^2 = 1$. De plus, on impose la contrainte que chaque composante n'est pas corrélée avec les précédentes : pour toutes les paires de m, u tel que $m \neq u$, $\sum_{j=1}^{p} \phi_{j,m}\phi_{j,u} = 0$. Cependant, calculer les loadings n'est pas dans le cadre de l'examen.

Contexte

L'idée du PCA que les *premières dimensions expliquent la majorité de la variabilité* afin de réduire le nombre de dimensions de p. **Standardiser** les variables aide à diminuer l'impact que l'échelle originale des variables pourrait avoir sur l'estimation des « *loadings* ».

Régression en composantes principales (PCR)

Notation

k Nombre de composantes principales utilisées où $k \in \{1, \dots, p\}$.
θ_j j^{e} coefficient de régression où $j \in \{0, \dots, k\}$.

L'équation du modèle est $Y = \theta_0 + \theta_1 z_1 + \cdots + \theta_k z_k + \varepsilon$. Poser $k < p$ nous permet donc de réduire le nombre de dimensions de p. La valeur de k idéale peut être obtenue par la validation croisée, c'est un hyperparamètre.

Note Si $k = p$, alors le modèle est équivalent à celui obtenu avec la régression linéaire multiple et aura les mêmes coefficients β.

Partial least squares (PLS)

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Notation

z_m m^{e} direction.

Contexte

De façon semblable au PCA, la méthode du PLS créé de nouvelles variables explicatives qui sont des combinaisons linéaires des p variables explicatives originales ; ces nouvelles variables sont des « ***directions*** » z_m.

La différence du PLS au PCA est que les **directions sont déduites de la variable réponse** y. Donc, contrairement au PCA, la méthode du PLS est de l'apprentissage supervisé. Il s'ensuit également que l'utilité des directions du « ***partial least squares regression*** *(PLR)* » est plus limitée que les « *loadings* » du PCA. Les « *loadings* » peuvent être utilisés pour analyser les données en plus de les modéliser.

Coefficients

Les coefficients $\phi_1, \dots, \phi_p$ des directions z_m se calculent en fonction de la variable réponse. Pour la première direction z_1, le coefficient ϕ_j équivaut à la pente la de régression linéaire simple de y en fonction de x_j et $z_1 = \sum_{j=1}^{p} \phi_j x_j$.

On prend la même procédure pour la deuxième direction, mais les prédicteurs utilisés x_j^* correspondent à la portion du prédicteur original pas prédit par z_1. Donc, $z_2 = \sum_{j=1}^{p} \phi_j^* x_j^*$ où ϕ_j^* est la pente estimée de la régression de y sur x_j^*.

La procédure est la même pour toutes les directions, puis on obtient que la i^{e} prévision est $\hat{y}_i = \hat{\theta}_1 z_{i,1} + \cdots + \hat{\theta}_k z_{i,k}$.

Note Il y a une faute dans le manuel dans la procédure qu'ils donnent pour calculer les directions du PLS. Il est donc peu probable que l'examen pose des questions

sur les spécifiques du calcul de la méthode du PLS. Plutôt, on devrait comprendre son utilité et son fonctionnement.

Régression linéaire généralisée

Contexte

La régression linéaire généralisée (GLM) *généralise* la régression linéaire qui est limitée par plusieurs hypothèses. Les GLMs regroupent plusieurs types de modèles sous une branche commune dont la similarité se base sur la famille exponentielle.

Famille exponentielle

Note La sous-section sur la *Famille exponentielle* de la section sur les *Statistiques exhaustives* détaille l'application de la famille exponentielle pour identifier le MVUE. Cette section couvre plus en détails la famille.

Famille exponentielle

Une distribution est membre de la famille exponentielle si sa fonction de densité peut être écrite sous la forme $f(y;\theta) = \mathrm{e}^{a(y)b(\theta)+c(\theta)+d(y)}$.

La fonction de densité implique que la distribution comporte un seul paramètre d'intérêt θ. On surnomme tous les autres paramètres des « ***nuisance parameters*** » que l'on pose fixes.

Forme canonique

Dans le contexte de régression linéaire généralisée, on pose que $a(y) = y$. La distribution est donc en ***forme canonique*** et $b(\theta)$ est le ***paramètre naturel*** de la distribution.

Moyenne et variance

La famille exponentielle permet d'établir un lien entre le paramètre θ et la moyenne par l'équation $\mathrm{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$. De plus, on peut lier la moyenne à la variance par l'équation $\mathrm{Var}(a(Y)) = \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{[b'(\theta)]^3}$.

Distribution Tweedie

Les distributions de la famille exponentielle dont $\mathrm{Var}(Y) = a\mathrm{E}[Y]^d$ sont des ***distributions Tweedie*** où a, d sont des constantes.
Selon la valeur de d, on y inclut :

$d = 0$ on obtient la distribution normale.

$d = 1$ on obtient la distribution de Poisson.

$1 < d < 2$ on obtient la distribution de Poisson gamma composée.

$d = 2$ on obtient la distribution gamma.

$d = 3$ on obtient la distribution inverse gaussienne.

L'important n'est pas de savoir ces liens, mais plutôt de savoir que si un lien existe entre la moyenne et la variance, alors on peut mieux choisir la distribution.

Définition du modèle

Modèle

Postulats de la régression linéaire généralise

1. La variable réponse suit une distribution qui est membre de la famille exponentielle.
2. La composante linéaire $\mathbf{x}_i^\top \boldsymbol{\beta}$ est une fonction de la moyenne $g(\mu_i)$

En bref, modéliser avec un GLM nécessite de choisir la composante aléatoire (alias, la distribution) et la composante systématique (alias, la fonction $g(\cdot)$. Choisir la distribution se résume habituellement à choisir une distribution dont le domaine est compatible avec les valeurs possibles de la variable réponse. Par exemple, une distribution continue serait inadéquate pour des données de comptage.

Fonctions de lien

Contexte

La ***fonction de lien*** lie la moyenne $\mathrm{E}[Y]$ à la composante linéaire $\mathbf{x}^\top \boldsymbol{\beta}$. Donc, elle établit le lien entre les prédicteurs et la moyenne de la variable réponse.

Les fonctions de lien doivent être **monotones** et **différentiables**, en voici quelques-unes où $g(\mu) = \mathbf{x}^\top \boldsymbol{\beta}$:

Fonction de lien	$g(\mu)$
identité	μ
canonique	$b(\theta)$
logit	$\ln\left(\frac{\mu}{1-\mu}\right)$
logarithmique	$\ln(\mu)$
inverse	$\frac{1}{\mu}$
puissance	μ^d

Contexte

Elle est typiquement choisie en fonction du domaine des valeurs possibles que prend la moyenne. Par exemple, $\mu \in \mathbb{R}$ pour une distribution normale et donc le lien d'identité est idéale. En revanche, la moyenne d'une distribution Bernoulli est entre 0 et 1 donc le lien logit $\boldsymbol{x}^\top \boldsymbol{\beta} = \ln\left(\frac{\mu}{1-\mu}\right)$ est plus approprié.

Note La fonction de lien canonique pose que le paramètre naturel $b(\theta) = \mathbf{x}^\top \boldsymbol{\beta}$. Selon la distribution, la fonction de lien canonique correspond à :

normale fonction de lien d'identité.

binomiale fonction de lien logit.

Poisson fonction de lien logarithmique.

gamma fonction de lien inverse.

Estimation des paramètres

Contexte

Les GLM estiment $\boldsymbol{\beta}$ par le MV. Sous l'hypothèse que les réalisations Y_i de la variable réponse sont indépendantes, on souhaite trouver les valeurs de $\boldsymbol{\beta}$ qui maximisent l'équation $\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i b(\theta_i) + c(\theta_i) + d(y_i)$.

On sait que les coefficients $\boldsymbol{\beta}$ sont reliés à la moyenne μ_i par la fonction de lien et que μ_i est relié au paramètre d'intérêt θ_i. Donc, les $\boldsymbol{\beta}$ sont imbriqués dans la fonction de vraisemblance et on peut obtenir les équations de Score en posant le vecteur de dérivées partielles par rapport à $\boldsymbol{\beta}$ égales à zéro $\left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} = 0, \ldots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p} = 0\right)$.

Sous la régression linéaire multiple, on peut obtenir une solution de façon analytique. Cependant, il est rate d'obtenir une expression analytique pour la régression linéaire généralisée. En lieu, on doit utiliser un *algorithme de résolution numérique* pour obtenir le vecteur $\boldsymbol{b}$ de façon itérative.

Une fois que l'EMV $\boldsymbol{b}$ est obtenue, on pose que $\hat{\mu} = g^{-1}\left(\mathbf{x}^\top \boldsymbol{\beta}\right)$. Ceci s'apparente donc à la notation de $\hat{y}$ de la SLR.

Note $\hat{\mu}$ n'est pas biaisé si l'on utilise le lien canonique, mais l'est sinon.

Méthode de « *Scoring* »

Contexte

La méthode de Scoring est basée sur l'approximation de Newton-Raphson qui approxime les racines d'une fonction. La méthode s'applique sur la dérivée d'une fonction et part d'une hypothèse initiale. Puis, on répète l'algorithme jusqu'à ce que l'on converge vers une solution.

Notation

u_j Fonction de score de β_j, $u_j = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}$ pour $j = 0, 1, \ldots, p$.

On dénote le vecteur des u_j par $\boldsymbol{u}$.

On dénote l'estimation de $\boldsymbol{u}$ à la fin de la m^{e} itération par $\boldsymbol{u}^{(m)}$.

$\boldsymbol{I}$ *Matrice d'information de Fisher* de $\boldsymbol{\beta}$.

On dénote l'estimation de $\boldsymbol{I}$ à la fin de la m^{e} itération par $\boldsymbol{I}^{(m)}$.

$\boldsymbol{b}^{(m)}$ Estimation de $\boldsymbol{\beta}$ à la fin de la m^{e} itération.

On trouve que $u_j = \sum_{i=1}^{n} \frac{(y_i - \mu_i) x_{i,j}}{\text{Var}(Y_i) g'(\mu_i)}$.

Note La façon d'arriver à l'expression de u_j n'est pas importante, il faut juste conceptuellement comprendre l'idée de passer de la log-vraisemblance à la fonction de Score.

Dans notre cas, *la matrice d'information de Fisher* est :

$$\boldsymbol{I} = \begin{bmatrix} -\text{E}\left[\frac{\partial^2}{\partial \beta_0^2} \ell(\boldsymbol{\beta})\right] & -\text{E}\left[\frac{\partial^2}{\partial \beta_0 \beta_1} \ell(\boldsymbol{\beta})\right] & \cdots & -\text{E}\left[\frac{\partial^2}{\partial \beta_0 \beta_p} \ell(\boldsymbol{\beta})\right] \\ -\text{E}\left[\frac{\partial^2}{\partial \beta_1 \beta_0} \ell(\boldsymbol{\beta})\right] & -\text{E}\left[\frac{\partial^2}{\partial \beta_1^2} \ell(\boldsymbol{\beta})\right] & \cdots & -\text{E}\left[\frac{\partial^2}{\partial \beta_1 \beta_p} \ell(\boldsymbol{\beta})\right] \\ \vdots & \vdots & \ddots & \vdots \\ -\text{E}\left[\frac{\partial^2}{\partial \beta_p \beta_0} \ell(\boldsymbol{\beta})\right] & -\text{E}\left[\frac{\partial^2}{\partial \beta_p \beta_1} \ell(\boldsymbol{\beta})\right] & \cdots & -\text{E}\left[\frac{\partial^2}{\partial \beta_p^2} \ell(\boldsymbol{\beta})\right] \end{bmatrix}$$

On trouve que l'élément $\boldsymbol{I}_{j+1,j^*+1} = \sum_{i=1}^{n} \frac{x_{i,j} x_{i,j^*}}{\text{Var}(Y_i) g'(\mu_i)^2}$.

Algorithme de Newton-Raphson

1. Poser des valeurs de départ pour le vecteur $\boldsymbol{b}^{(0)}$.
2. Pour $j = 1, 2, \ldots$:
   (a) Poser que $\boldsymbol{b}^{(m)} = \boldsymbol{b}^{(m-1)} + \left(\boldsymbol{I}^{(m-1)}\right)^{-1} \boldsymbol{u}^{(m-1)}$.
   (b) Si $\boldsymbol{b}^{(m)} \approx \boldsymbol{b}^{(m-1)}$ alors l'EMV a convergé et on cesse l'algorithme, sinon répéter.

On peut définir une approche alternative lorsqu'il y a hétéroscédasticité et que $\text{Var}(\varepsilon_i) = \frac{\sigma^2}{w_i}$. Puisque la variance est fonction d'un vecteur de poids $\boldsymbol{W}$ où

$$\boldsymbol{W} = \begin{bmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{bmatrix}$$

on définit la ***méthode des moindres carrés pondérée***. Au lieu d'avoir $\boldsymbol{b} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$, on a $\boldsymbol{b} = (\mathbf{X}^\top\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{W}\mathbf{y}$ où $w_i = \frac{1}{\text{Var}(Y_i)g'(\mu_i)^2}$. On peut donc définir que $\boldsymbol{I} = \sum_{i=1}^n \frac{\mathbf{x_i}\mathbf{x_i}^\top}{\text{Var}(Y_i)g'(\mu_i)^2}$ et $z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$.
Alors, on obtient par la « *iterative weighted least squares procedure* » que $\boldsymbol{b}^{(m)} = \left(\mathbf{X}^\top\mathbf{W}^{(\mathbf{m-1})}\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{W}^{(\mathbf{m-1})}\mathbf{z}^{(\mathbf{m-1})}$.

Note De façon générale, pour un paramètre d'intérêt θ la méthode de Score pose $\hat{\theta}^{(m)} = \hat{\theta}^{(m-1)} + I\left(\hat{\theta}^{(m-1)}\right)^{-1}\ell'\left(\hat{\theta}^{(m-1)}\right)$.

Résumés numériques

Contexte

Nous avons vu en régression linéaire *la relation* $SST = SSR + SSE$. Cependant, ceci ne se généralise pas directement pour les GLMs. Plutôt, on a que $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 + 2\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$.
Le terme additionnel est nul pour la MLR ce qui permet de définir le *coefficient de détermination*.

Puisque pour les GLMs $\frac{SSR}{SST} \neq 1 - \frac{SSE}{SST}$, on doit définir de nouvelles mesures pour évaluer le modèle.

Mesures basées sur la log-vraisemblance maximisée

Contexte

Nous avons utilisé la vraisemblance maximisée pour définir le *Test du rapport de vraisemblance* et la log-vraisemblance maximisée pour estimer les coefficients avec la *Méthode de « Scoring »*.

Les mesures se basent sur deux modèles :

1 Modèle nul

Le modèle ***nul***, ou *minimale*, comporte uniquement l'intercepte ($p' = 1$) : $Y = \beta_0$. Ceci implique que $\hat{\mu}_i = \bar{y}$ $\forall i$.

2 Modèle saturé

Le modèle ***saturé***, ou *complet*, comporte autant de prédicteurs que d'observations ($p' = n$) : $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{n-1}x_{n-1}$. Ceci implique que $\hat{\mu}_i = y_i$.

Log-vraisemblance maximisée

On dénote par :
ℓ_{null} La log-vraisemblance du **modèle nul**.
ℓ_{sat} La log-vraisemblance du **modèle saturé**.

La log-vraisemblance maximisée augmente avec le nombre de para-

mètre. Donc, pour un modèle ayant entre 1 et n coefficients on a que $\ell_{null} \leq \ell(\boldsymbol{b}) \leq \ell_{sat}$. De façon générale, une log-vraisemblance maximisée qui est élevée indique que le modèle est bien ajusté aux données. Cependant, si $\ell(\boldsymbol{b}) \approx \ell_{sat}$ alors le modèle peut être surajusté !

Déviance

Contexte

On peut voir la déviance comme l'analogue de l'EQM pour les GLMs. Pour comprendre l'intuition derrière la formule, on la récrit comme $D = \ln\left(\left(\frac{\mathcal{L}_{sat}}{\mathcal{L}(\boldsymbol{b})}\right)^2\right)$. Au lieu de prendre l'écart carré entre les prévisions et réalisations, on prend le carré du ratio du *modèle estimé* au *modèle* ***parfait***. Donc, on souhaite minimiser la déviance sans toutefois la rendre nulle (dans quel cas le modèle serait surajusté).

La déviance d'un GLM est $D = 2\left(\ell_{sat} - \ell(\boldsymbol{b})\right)$ où $D \sim \chi^2_{n-p'}$.

Si la distribution est normale, alias le cas de MLR, on a que $D = \frac{SSE}{\sigma^2}$ ou $\sigma^2 D = SSE$. On surnomme $\sigma^2 D$ la « *scaled deviance* ».

Note Puisque les log-vraisemblances sont des sommes, on peut récrire la déviance comme une somme de composantes avec $D = \sum_{i=1}^{n} D_i$.

Ratio de vraisemblance du khi carré

Contexte

Le test utilisant le ratio de vraisemblance du khi carré test l'hypothèse nulle que le modèle nul est supérieur au modèle ajusté. On rejette l'hypothèse nulle et acceptons le modèle ajusté si la valeur observée de la statistique est *supérieure* à la valeur critique. Ceci est donc le contraire d'un test utilisant la déviance qui rejette le modèle ajusté si la valeur observée dépasse la valeur critique. Ceci est cohérent, car on devrait s'attendre à ce que le modèle ajusté soit supérieur au modèle nul de la même façon que, en théorie, le modèle saturé soit « supérieur » au modèle ajusté.

Le ratio de vraisemblance du khi carré (« *likelihood ratio chi-square* ») est semblable à la déviance mais calcule l'écart entre le modèle ajusté et le modèle nul avec $C = 2\left(\ell(\boldsymbol{b}) - \ell_{null}\right)$.

On note que $C \sim \chi^2_{p'-1}$.

Pseudo R^2_{ps}

On peut également définir un analogue du R^2 pour les GLMs. Le R^2_{ps} capture l'apport du modèle proportionnellement au modèle nul avec $R^2_{ps} = \frac{\ell_{null} - \ell(\boldsymbol{b})}{\ell_{null}}$.

Contexte

La déviance et le pseudo-R^2 sont des mesures qui servent à évaluer la qualité d'ajustement d'un modèle. Pour comparer des modèles entre eux, on redéfinit l'AIC et le BIC. Dans le chapitre de *Erreur* on a définit l'AIC et le BIC *en fonction de la vraisemblance* alors que dans ce chapitre on les a définit *en fonction de l'EQM*.

L'utilité des mesures restent encore de comparer des modèles, mais nous les redéfinissons en fonction de la log-vraisemblance :

1. **Critère d'information d'Akaike (AIC)**

$$AIC = -2\ell(\boldsymbol{b}) + 2(\text{nombre de paramètres estimés})$$

2. **Critère d'information bayésien (BIC)**

$$BIC = -2\ell(\boldsymbol{b}) + (\text{nombre de paramètres estimés})\ln(n)$$

Note Le nombre de paramètres estimés est habituellement p', mais ça pourrait être plus que p' s'il faut estimer, par exemple en MLR, d'autre paramètres comme σ^2.

Note Ces définitions sont fournies dans l'examen.

Résidus

Contexte

Les *résidus*, ou « ***raw residuals*** », $e_i = y_i - \hat{\mu}_i$ ne sont pas directement utiles pour les GLMs. Au lieu de directement utiliser les résidus, on les modifient pour obtenir de différents résidus.

Résidus de Pearson

Les résidus de Pearson divisent le résidu par la variance évaluée à $\boldsymbol{b}$ au lieu de $\boldsymbol{\beta}$: $e_i^P = \frac{e_i}{\sqrt{\widehat{\text{Var}}(Y_i)}}$. On peut ensuite *standardiser les résidus comme on l'a avant* avec $e_{sta,i}^P = \frac{e_i^P}{\sqrt{1-h_i}}$.

Statistique du khi carré de Pearson

Avec les résidus de Pearson, on peut définir la ***statistique du khi carré de Pearson*** $\sum_{i=1}^n \left(e_i^p\right)^2$. Cette équation s'apparente au SSE, mais avec de différents résidus.

Résidus de déviance

Les résidus de déviance sont définis comme $e_i^D = \text{signe}(e_i)\sqrt{D_i}$ où $\text{signe}(e_i) \in \{-1, 1\}$ selon la valeur du résidu e_i. Également, on peut standardiser le résidu avec $e_{sta,i}^D = \frac{e_i^D}{\sqrt{1-h_i}}$.

Inférence statistique

Contexte

Les test t et F, ainsi que les intervalles de confiance, de la MLR ne sont pas directement applicables aux GLMs, mais des concepts semblables existent. On utilise la statistique de Score pour obtenir la distribution des coefficients. Puis, on montre comment on peut ajuster la variance s'il y a surdispersion.

Puis, on montre l'analogue du test t : la statistique de Wald. On l'utilise pour tester le retrait d'un coefficient. L'analogue du test F est le TRV : on l'utilise pour tester la simplification d'un modèle.

Statistique de Score

La ***statistique de Score*** U_j est la variable aléatoire qui correspond à u_j où $U_j = \sum_{i=1}^{n} \frac{(Y_i-\mu_i)x_{i,j}}{\mathrm{Var}(Y_i)g'(\mu_i)}$.

On dénote le vecteur colonne des statistique $\boldsymbol{U}$ où $\boldsymbol{U} \sim \mathcal{N}_{p'}(\boldsymbol{0}, \boldsymbol{I})$. De plus, $\mathbf{U}^\top \mathbf{I}^{-1}\mathbf{U} \approx \chi^2_{p'}$.

On généralise la distribution des coefficients indiquée pour la *Régression linéaire simple* aux GLMs avec la distribution **asymptotique** $\boldsymbol{b} \sim \mathcal{N}_{p'}(\boldsymbol{b}, \boldsymbol{I}^{-1})$ où

$$\boldsymbol{I}^{-1} = \begin{bmatrix} \widehat{\mathrm{Var}}(b_0) & \widehat{\mathrm{Cov}}(b_0,b_1) & \cdots & \widehat{\mathrm{Cov}}(b_0,b_p) \\ \widehat{\mathrm{Cov}}(b_0,b_1) & \widehat{\mathrm{Var}}(b_1) & \cdots & \widehat{\mathrm{Cov}}(b_1,b_p) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\mathrm{Cov}}(b_0,b_p) & \widehat{\mathrm{Cov}}(b_1,b_p) & \cdots & \widehat{\mathrm{Var}}(b_p) \end{bmatrix},$$

et $\widehat{\mathrm{Var}}(b_j) = se(b_j)^2$.

Théorie de Wald

Statistique de Wald

La ***statistique de Wald*** $(\boldsymbol{b}-\boldsymbol{\beta})^\top \boldsymbol{I}(\boldsymbol{b}-\boldsymbol{\beta}) \approx \chi^2_{p'}$.

On peut ensuite effectuer le ***test de Wald*** que $\beta_j = h$ avec $t = \left(\frac{b_j-h}{se(b_j)}\right)^2$ où $T \approx \chi^2_1$.

Surdispersion

Contexte

La ***surdispersion*** a lieu lorsque la variabilité est plus grande que celle estimée par le modèle. On peut détecter ceci avec $\frac{D}{\mathrm{E}[D]} = \frac{D}{n-p'}$. Plus le ratio est élevé, plus la surdispersion est grave !

La méthode de ***quasi-vraisemblance*** adresse ce problème en multipliant la variance par un paramètre $\phi > 1$: $\mathrm{Var}(Y_i) = \phi \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{(b'(\theta_i))^3}$.

Cette méthode établie une relation entre la moyenne et la variance qui est moins restrictive au dépens de rendre la distribution moins claire.

Test du rapport de vraisemblance

Le *test du rapport de vraisemblance* tel que définit dans la section *Test du rapport de vraisemblance* s'applique pour tester la simplification de modèles emboîtés. On pose que $t = 2\left[\ell(\boldsymbol{b}_f) - \ell(\boldsymbol{b}_r)\right]$ pour tester

- H_0 : modèle **réduit** est adéquat
- H_1 : modèle complet (« ***full*** ») est meilleur.

On teste donc si $p_f - p_r$ coefficients de régression sont nuls et $T \approx \chi^2_{p_r - p_f}$. Il s'ensuit que c'est un **test unilatéral vers la droite**.

Note On peut récrire la statistique comme $t = D_r - D_f$.

Classification

Préliminaire

Contexte

Puisque l'on peut récrire la f.m.p. d'une distribution binomiale sous la forme de la famille exponentielle, on sait qu'elle en fait partie. Cette section détaille donc l'application des GLMs pour des données binomiales.

Cote

La ***cote*** (« *odds* ») d'un événement est le ratio de la probabilité que l'événement se réalise par la probabilité que l'événement ne se réalise pas. Pour la distribution binomiale, $cote = \frac{q}{1-q}$.

log cote

La ***log cote*** (« *log odds* ») est le logarithme naturel de la cote : $\ln(cote) = \ln\left(\frac{q}{1-q}\right)$.

Fonctions de lien

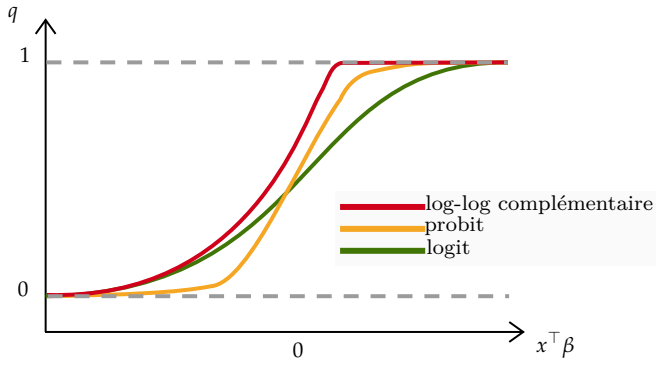
Contexte

Habituellement, on lie la moyenne μ aux coefficients $\boldsymbol{\beta}$ ce qui nous permet d'éviter d'avoir à spécifier la distribution. Dans le contexte de classification où on connaît la distribution, on peut simplement relier les coefficients **au paramètre d'intérêt** $\theta = q$ au lieu de la moyenne $\mu = mq$ avec $g(\theta) = \mathbf{x}^\top\boldsymbol{\beta}$.

Les 3 fonctions de lien les plus utilisées pour $q \in [0,1]$ sont les suivantes :

Nom	$q =$	$\mathbf{x}^\top\boldsymbol{\beta} =$
Logit	$\frac{e^{x^\top\beta}}{1+e^{x^\top\beta}}$	$\ln\left(\frac{q}{1-q}\right)$
Probit	$\Phi(\mathbf{x}^\top\boldsymbol{\beta})$	$\Phi^{-1}(q)$
Log-log complémentaire	$1-e^{-e^{x^\top\beta}}$	$\ln(-\ln(1-q))$

L'utilité de ces 3 fonctions est que leurs domaines sont contenus entre 0 et 1 ($g^{-1}(\mathbf{x}^\top\boldsymbol{\beta}) = q \in [0,1]$). Comme on peut observer ci-dessous, les fonctions de lien logit et probit sont **symétriques** à 0 alors que la fonction de lien log-log complémentaire ne l'est pas.



Modèle binomiale

Contexte

Lorsque la variable réponse prend une de deux catégories, alors elle suit une distribution binomiale.

Soit les v.a. indépendantes binomiales $Y_1,\dots,Y_n$ dont les m_is sont connus et $q_i = g^{-1}(\boldsymbol{x}_i^\top\boldsymbol{\beta})$. Alors, la fonction de log-vraisemblance $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \left[y_i \ln\left(\frac{q_i}{1-q_i}\right) + m_i \ln(1-q_i) + \ln\binom{m_i}{y_i}\right]$.

De cette expression, on obtient que la déviance $D = 2\sum_{i=1}^n \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i)\ln\binom{m_i - y_i}{m_i - \hat{\mu}_i}\right]$ et que le résidu de Pearson $e_i^P = \frac{y_i - m_i\hat{q}_i}{\sqrt{m_i\hat{q}_i(1-\hat{q}_i)}}$.

On obtient différents modèles selon la fonction de lien choisie. Par exemple, on obtient une ***régression logistique*** avec la fonction de lien logit. Puisque nous avons une fonction de q et non de la moyenne μ, on doit définir la fonction de Score directement avec $u_j = \sum_{i=1}^n (y_i - \mu_i)x_{i,j}$ et donc $\boldsymbol{I} = \sum_{i=1}^n m_i q_i(1-q_i)\mathbf{x_i}\mathbf{x_i}^\top$.

Note L'interprétation des paramètres se fait en fonction de ***ratios de cotes*** et non directement. On peut donc interpréter la (dé)croissance des prédicteurs en fonction d'un autre.

Réponse nominale

Notation

g Nombre de catégories.
c Catégorie pour $c = 1, \dots, g$.

Contexte

Lorsque la variable réponse est définie comme une de plusieurs catégories sans d'ordre, nous avons une *variable nominale*.

On modélise la variable réponse par une ***distribution multinomiale***. On pose que $Y_1, \dots, Y_g$ suit une distribution multinomiale avec m essais et avec probabilité de π_c d'être dans la **catégorie** c où $c = 1, \dots, g$.

On trouve que $p(y_1, \dots, y_g) = \frac{m!}{y_1! \cdots y_g!} \pi_1^{y_1} \cdot \cdots \cdot \pi_g^{y_g}$, puis que $\sum_{c=1}^{g} Y_c = m$ et $\sum_{c=1}^{g} \pi_c = 1$.

Puisque nous avons une contrainte, on fixe une catégorie arbitraire g comme la ***catégorie de référence***. Donc, $Y_1, \dots, Y_{g-1}$ suit une distribution multinomiale, puis que $y_g = m - \sum_{c=1}^{g-1} y_c$ et $\pi_g = 1 - \sum_{c=1}^{g-1} \pi_c$.

Régression logistique avec réponse nominale

Notation

k Catégorie de référence.
$\pi_{i,c}$ Probabilité que la i^{e} observation soit classifiée dans la catégorie c.

On pose que $\boldsymbol{x}_i^\top \boldsymbol{\beta}_t = \ln\left(\frac{\pi_{i,t}}{\pi_{i,k}}\right)$ où $t \neq k$. Il y a donc $g-1$ équations comme celle-ci desquelles on peut isoler

$$\pi_{i,c} = \begin{cases} \frac{1}{1+\sum_t \mathrm{e}^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_t}}, & c = k \\ \frac{\mathrm{e}^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_c}}{1+\sum_t \mathrm{e}^{\boldsymbol{x}_i^\top \boldsymbol{\beta}_t}}, & c \neq k \end{cases}$$

Note On estime $(p+1)(g-1)$ coefficients au total.

Réponse ordinale

Contexte

Lorsque la variable réponse est définie comme une de plusieurs catégories ordonnées, nous avons une *variable ordinale*.

La différence du modèle pour une réponse nominale est que nous voulons capturer l'ordre des catégories. Pour ce faire, on modéliser les probabilités ***cumulatives***.

Notation

Π_c Probabilité cumulative d'être classifié dans la catégorie c où $\Pi_c = \pi_1 + \cdots + \pi_c$.

Puisque $\Pi_g = \sum_{c=1}^{g} \pi_c = 1$ nous avons encore seulement $g-1$ probabilités à estimer.

Modèle de cotes proportionnelles

Notation

$\beta_{0,c}$ Intercepte pour la catégorie c.

Pour le modèle de cotes proportionnelles, on pose que tous les coefficients sauf l'intercepte sont les mêmes peu importe la catégorie. On modélise donc les probabilités cumulatives avec la relation $\ln\left(\frac{\Pi_{i,c}}{1-\Pi_{i,c}}\right) = \beta_{0,c} + \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

Note Le modèle estime $g - 1 + p$ coefficients au total

Modèles pour des données de comptage

Contexte

Puisque l'on peut récrire la f.m.p. d'une distribution de Poisson sous la forme de la famille exponentielle, on sait qu'elle en fait partie. Cette section détaille donc l'application des GLMs pour des données de comptage.

Régression de Poisson

Notation

a_i unité d'exposition pour la i^{e} observation.

λ_i moyenne par exposition pour la i^{e} observation.

Une distribution de Poisson modélise la fréquence d'un événement pour un contexte donné mesuré en ***unités d'exposition***.
Soit les v.a. indépendantes de Poisson $Y_1, \dots, Y_n$ de moyennes $\mu_i = a_i\lambda_i$. Avec la fonction de lien logarithmique, $\mu_i = a_i e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}}$.

Alors, la fonction de log-vraisemblance $\ell(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \ln \mu_i - \mu_i - \ln(y_i!)]$ et la fonction de Score $u_j = \sum_{i=1}^n (y_i - \mu_i) x_{i,j}$. De plus, la matrice d'information $\boldsymbol{I} = \sum_{i=1}^n \mu_i \boldsymbol{x}_i \boldsymbol{x}_i^\top$ et la déviance $D = 2\sum_{i=1}^n \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right]$.

De cette expression, on obtient que le résidu de Pearson $e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$.

Modèle log-linéaire

Pour modéliser les fréquences dans tes tableaux de contingence, on utilise des modèles log-linéaires.

Modèles additifs généralisés (GAM)

Contexte

Les méthodes alternatives de régression diminuent la complexité du modèle linéaire et diminuent la variance des prévisions. Cependant, ces modèles tiennent seulement sous l'hypothèse de normalité. Dans cette section, on cherche à assouplir l'hypothèse de normalité tout en gardant le modèle aussi interprétable que possible. On élargi le modèle linéaire avec des ajouts simples comme une fonction polynôme ou en escalier. Puis, on prend des approches plus avancées avec des splines, la régression locale et finalement un GAM.

« *Basis functions* »

Notation

$b_j(x)$ « *basis function* » où $j = 1, \dots, p$.

Contexte

On reprend la section de *Variables explicatives spéciales* en se limitant à une seule variable explicative x. On écrit l'équation du modèle en fonction de « *basis functions* » de la variable explicative : $Y = \beta_0 + \beta_1 b_1(x) + \cdots + \beta_p b_p(x) + \varepsilon$.

Cette section considère donc 3 choix de « *basis functions* ».

Régression d'une fonction polynôme

Contexte

La méthode la plus simple d'obtenir un ajustement non linéaire est d'ajouter des prédicteurs qui sont des puissances des prédicteurs préexistants. Par exemple, une régression cubique ajoute x^2 et x^3 comme prédicteurs.

L'équation a *déjà été présentée*, mais ont définit maintenant la fonction $b_j(x) = x^j$ pour $j = 1, \dots, d$ où l'équation est un polynôme d'ordre d.

Régression d'une fonction constante par morceaux

Notation

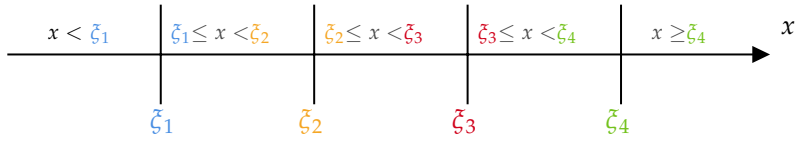
k Nombre de nœuds.

ξ_j « *cutpoints* » ou « ***knots*** » pour $j = 1, \dots, k$.

Contexte

Le désavantage de la régression d'un polynôme est que l'on impose une structure globale à la fonction non-linéaire de x. Avec une fonction constante par morceaux, on peut imposer des structures locales sur chaque intervalle. La méthode segmente l'étendu de la variable en k régions distinctes, on obtient une variable catégorielle sur laquelle on ajuste une fonction constante par morceaux.

On divise l'étendue de x en $k+1$ intervalles (« *bins* »). On obtient donc k nœuds $\xi_1, \dots, \xi_k$:



Il s'ensuit que les « *basis functions* » sont des fonctions indicatrices :

$$b_j(x) = \begin{cases} I_{\{\xi_j \leq x < \xi_{j+1}\}}, & j = 1, \dots, k-1 \\ I_{\{x \geq \xi_k\}}, & j = k \end{cases}$$

Ceci revient à écrire l'équation du modèle comme $Y = \beta_0 + \beta_1 I_{\{\xi_1 \leq x < \xi_2\}} + \cdots + \beta_k I_{\{\xi_k \leq x <}$

Note Il n'y a pas d'approche absolue, mais habituellement on utilise des centiles pour identifier les nœuds.

Régression d'une fonction polynôme par morceaux

Notation

d Ordre du polynôme.

Contexte

Cette approche combine la fonction polynôme avec la fonction en escalier pour obtenir une fonction polynôme par morceaux.

On combine les deux approches précédentes pour créer un polynôme d'ordre d pour chacun des k intervalles. Ceci revient à écrire l'équation du modèle comme :

$$Y = \begin{cases} \beta_{0,1} + \beta_{1,1}x + \cdots + \beta_{d,1}x^d + \varepsilon, & x < \xi_1 \\ \beta_{0,2} + \beta_{1,2}x + \cdots + \beta_{d,2}x^d + \varepsilon, & \xi_1 \leq x < \xi_2 \\ \vdots & \vdots \\ \beta_{0,k+1} + \beta_{1,k+1}x + \cdots + \beta_{d,k+1}x^d + \varepsilon, & x \geq \xi_k \end{cases}$$

Note Pratiquer le calcul de degrés de liberté pour ces régressions afin de comprendre combien de paramètres sont estimés.

Splines de régression

Contexte

Le désavantage de la régression d'une fonction polynôme par morceaux est qu'elle n'est pas continue. Afin d'obtenir une fonction continue, on impose une « contrainte » par nœud qui stipule que les extrémités doivent se connecter. Avec suffisamment de régions, on peut obtenir un modèle qui est très bien ajusté.

Pour assurer la continuité de la fonction, on requiert que les $d-1$ premières dérivées soient continues aux k nœuds. Donc, le nombre de prédicteurs d'une fonction polynôme d'ordre d est $p = (d-1) + k + \underbrace{k * (d-1)}_{\text{intéractions}} - \underbrace{k * (d-1)}_{\text{contraintes}} =$ $(d-1) + k$.

Les $d + k - 1$ « *basis functions* » sont donc :
x, x^2, x^3 ;
$b_4 = (x-\xi_1)^3_+$, $b_5 = (x-\xi_2)^3_+$, …, $b_{3+k} = (x-\xi_k)^3_+$.
surnommées les « ***truncated power*** *basis functions* ».

Splines de régression naturel

Contexte

Sans contraintes, il se peut que la fonction ait une courbe excessive aux extrémités le menant au-delà du domaine des réalisations. Afin d'éviter ce problème, on utilise un ***spline naturel*** qui ajoute des nœuds au maximum et au minimum de l'échantillon (les intervalles $\{x < \xi_1\}$ et $\{x > \xi_k\}$). On stipule que, passé ces « *boundary knots* », la fonction doit être linéaire.

En bref, pour un polynôme d'ordre d ayant k nœuds :

Modèle de régression	ddl
Polynôme	$(d+1)$
Polynôme par morceaux	$(d+1)(k+1)$
Polynôme par morceaux continue	$(d+1)(k+1) - (d-1)k$
Spline	$(d+1) + k$
Spline naturelle	$(d+1) + k - 4$

Splines de lissage

Notation

λ Paramètre de lissage (« *smoothness penalty* »).

Contexte

Les splines de lissage sont un alternatif aux splines de régression. Bien qu'ils sont semblables, ils sont utiles dans un différent contexte et obtenus en minimisant la SSE plus une pénalité de lissage.

Tout comme le spline de régression, le modèle est composé d'un polynôme par intervalle tel que la valeur de la fonction évaluée aux nœuds est la même que la valeur de deux premières dérivées de la fonction. La différence est qu'un spline de lissage a un nœud pour chaque point des données d'entraînement.

Un spline de lissage minimise l'expression $\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int_{-\infty}^{\infty} g''(t)^2 dt$. Donc, on minimise la SSE plus une pénalité qui est fonction du paramètre de lissage λ. Cette pénalité est un indicateur de la « *wiggliness* » ou « *roughness* » de la fonction. Bref, un spline de lissage équivaut à un spline de régression plus aplatit.

La fonction $g(x)$ est :

1. linéaire dans les intervalles $x < \min(x_1, \ldots, x_n)$ et $x > \max(x_1, \ldots, x_n)$;
2. cubique par partie entre les réalisations triées $x_1, \ldots, x_n$;
3. continue, a une première dérivée continue et une deuxième dérivée continue aux points $x_1, \ldots, x_n$.

Puisque la flexibilité de la courbe est contrôlée par λ, le nombre de degrés de liberté n'est pas une bonne mesure de sa flexibilité. Plutôt, on utilise le **nombre de degrés de liberté *effectif*** calculé avec une matrice $\boldsymbol{S}_\lambda$ qui s'apparente à la matrice chapeau $\boldsymbol{H}$ de la régression linéaire.

On dénote par $\hat{\boldsymbol{g}}_\lambda$ les prévisions de y_i pour un λ fixé, puis on trouve que $\hat{\boldsymbol{g}}_\lambda = \boldsymbol{S}_\lambda \boldsymbol{y}$. Le nombre de degrés de liberté effectif correspond à la somme des éléments de la diagonale $h_{\lambda i}$ de la matrice $\boldsymbol{S}_\lambda$ avec $ddl_\lambda = \sum_{i=1}^{n} h_{\lambda i}$.

Régression locale

Notation

s Étendu (« *span* »)

Contexte

La régression locale s'apparente aux splines de lissage, sauf que l'approche pour établir le modèle ressemble plus à un « *K-nearest neighbors* ».

On minimise $\sum_{i=1}^{n} w_i(y_i - \hat{y}_i)^2$. Puis, pour prédire la valeur de x_*, la procédure est :

1. Choisir un étendu $s \in [0,1]$ qui indique la proportion des observations utilisées pour la régression.
2. Identifier les $v = sn$ réalisations dont la valeur est la plus proche de x_*.
3. Assigner les n poids $K_{i,*} = K(x_i, x_*)$ selon leur proximité à x_* pour $i = 1, \ldots, n$.
4. Effectuer une régression pondérée, puis minimiser $\sum_{i=1}^{n} K_{i,*}(y_i - b_{0,*} - b_{1,*}x_i)^2$ afin d'obtenir les estimations $b_{0,*}$ et $b_{1,*}$.
5. Calculer la prévision comme $b_{0,*} + b_{1,*}x_*$.

Modèle additif généralisé (GAM)

Contexte

Les modèles additifs généralisés (« *generalized additive models* ») permettent de généraliser les méthodes ci-dessus ***pour plusieurs prédicteurs***.

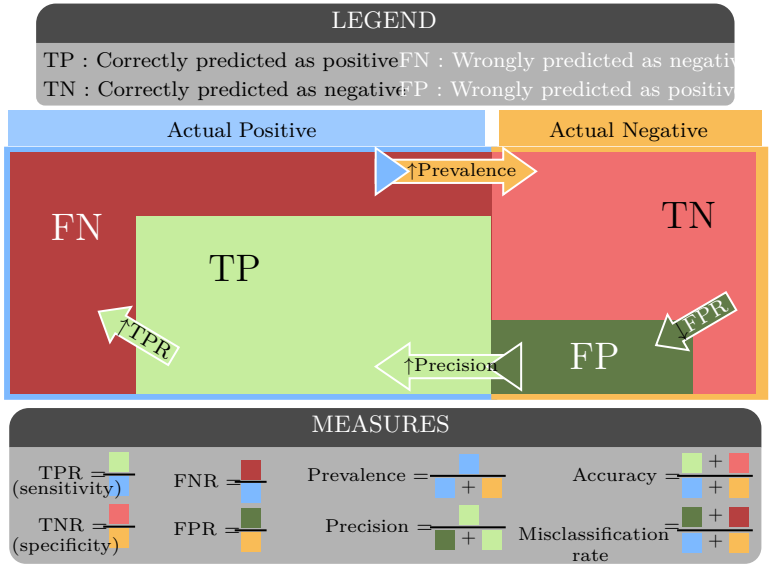
On pose que $Y = \beta_0 + f_1(x_1) + \cdots + f_g(x_g) + \varepsilon$ où chacune des fonctions peut être un modèle lui-même. Par exemple, on peut avoir que f_1 est une spline cubique, f_2 est une spline de lissage, etc.

Si toutes les fonctions $f_1, \ldots, f_g$ sont basées sur une régression des moindres carrés, alors le modèle GAM correspond à une régression linéaire multiple. Autrement, dès qu'au moins une des fonctions utilise une autre méthode comme une spline de lissage ou une régression locale, il faut utiliser une autre approche comme le « *backfitting* » qui ajuste séparément les modèles en fixant les $g-1$ autres variables de façon répétée jusqu'à ce que l'ajustement globale du modèle converge.

Un des désavantages du modèle GAM est qu'il ne tient pas compte des interactions entre les variables puisqu'elles sont toutes traitées séparément. Cependant, on peut manuellement en ajouter au modèle.

Autres

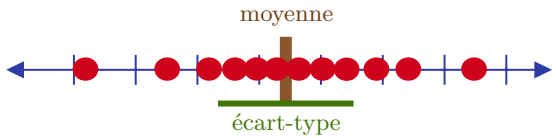
Matrice de confusion :



Erreur

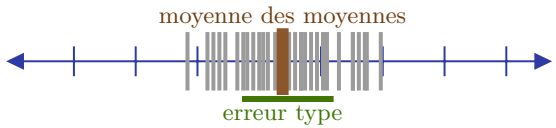
Écart-type Mesure la variation entre les observations d'**un** ensemble de données.

« *standard deviation* ».



Erreur type Mesure la variation entre les moyennes de **plusieurs** ensembles de données.

« *standard error* ».



III

Mathématiques actuarielles IARD I

Probabilité

Fonctions de variables aléatoires

Fonction de masse de probabilité (PMF)

Pour une variable aléatoire discrète X, on dénote sa fonction de masse de probabilité $p_X(x) = \Pr(X = x)$ tel que $0 \leq p(x) \leq 1$ et $\sum_x p(x) = 1$.

Fonction de densité (PDF)

Pour une variable aléatoire continue X, on dénote sa fonction de densité par $f_X(x)$ où $f_X(x) \neq \Pr(X = x)$.

La fonction de densité est évaluée sur des **intervalles de valeurs** pour obtenir la probabilité d'y être contenu, mais ne **représente pas une probabilité explicitement**.

De façon semblable à la PMF, $f(x) \geq 0$ et $\int_{-\infty}^{\infty} f(x)dx = 1$.

Contexte

La différence entre les conditions pour la PMF et la PDF est que la fonction de densité peut être supérieure à 1. Puisqu'elle ne représente **pas** une probabilité, elle ne doit pas être inférieure (ou égale) à 1 !

Fonction de répartition (CDF)

La fonction de répartition $F_X(x) = \Pr(X \leq x)$ tel que $F(-\infty) = 0$ et $F(\infty) = 1$.

En anglais, « *cumulative distribution function* ».

Fonction de survie

La fonction de survie $S_X(x) = \Pr(X > x)$ tel que $S(-\infty) = 1$ et $S(\infty) = 0$.

Fonction de hasard

Contexte

La fonction de hasard mesure la **vraisemblance** que la v.a. soit égale à x. La fonction de hasard « gonfle » la fonction de densité lorsqu'il devient de moins en moins vraisemblable qu'elle soit supérieure à x. De la définition, on déduit que c'est un concept seulement applicable pour les v.a. continues.

La fonction de hasard $h_X(x) = \frac{f(x)}{S(x)}$ tel que $h(x) \geq 0$.

En anglais, « *hazard function* », « *hazard rate* », « *failure rate function* » ou même « *force of mortality* ».

Fonction de hasard cumulative

La fonction de hasard cumulative $H_X(x) = \int_{-\infty}^{x} h(t)dt$. Également, $H(x) = -\ln S(x)$ ou $S(x) = \mathrm{e}^{-H(x)}$.

Note Voir la sous-section *Divers* de la section sur la *Théorie de la fiabilité* du chapitre *Critères d'information pour la sélection de modèles* pour l'interprétation de la distribution en fonction de la fonction de hasard et de la fonction de hasard cumulative.

Moments

Pour une v.a. X non-négative et une fonction $g(x)$ tel que $g(0) = 0$, $\mathrm{E}[g(X)] = \int_0^\infty g'(x)S(x)dx$.

Fonction génératrice des moments (MGF)

La fonction génératrice des moments (MGF) d'une v.a. X est dénoté comme $M_X(t) = \mathrm{E}[\mathrm{e}^{tX}]$.

Entre autres, la MGF sert à générer les moments d'une distribution avec $\mathrm{E}[X^n] = \frac{\partial^n M_X(t)}{\partial t^n}\big|_{t=0}$.

Fonction génératrice des probabilités (PGF)

La fonction génératrice des moments (PGF) d'une v.a. X est dénoté comme $P_X(t) = \mathrm{E}[t^X]$.

Entre autres, la PGF sert à :

1. Générer les masses de probabilité d'une distribution discrète avec $p(n) = \frac{1}{n!}\frac{\partial^n P_X(t)}{\partial t^n}\big|_{t=0}$.
2. Générer des espérances avec $\frac{\partial^n P_X(t)}{\partial t^n}\big|_{t=1} = \mathrm{E}\left[X(X-1)\dots(X-(n-1))\right]$.

Centiles, mode et statistiques

Centile

Contexte

Les centiles aident à quantifier la *vraisemblance* de pertes extrêmes. Bien que les actuaires se servent des centiles pour évaluer la ***fréquence*** des pertes extrêmes, ils ne sont **pas** utiles pour évaluer la *sévérité* de ces pertes.

Le $100q^{\mathrm{e}}$ ***centile*** d'une v.a. X est la valeur π_q tel que $\Pr(X < \pi_q) \leq q$ **et** $\Pr(X \leq \pi_q) \geq q$.

Dans le cas continu, $F_X(\pi_q) = q$ et $\pi_q = F_X^{-1}(q)$.

« *Conditionnal Tail Expectation (**CTE**)* »

Contexte

La CTE sert à évaluer la ***sévérité*** des pertes extrêmes. Ceci correspond au cas continu de la mesure « *Tail Value-at-Risk (TVaR)* » vue en introduction à l'actuariat II (ACT-2001).

Par exemple, si la $CTE_{0.95}(X) = 5000$ cela veut dire que la moyenne des pertes dans le top 5% est de 5 000$.

$$\begin{aligned} CTE_q(X) &= \mathrm{E}[X|X > \pi_q] \\ &= \pi_q + \mathrm{E}[X - \pi_q|X > \pi_q] \\ &= \pi_q + \frac{\mathrm{E}[X] - \mathrm{E}[X \wedge \pi_q]}{1-q} \end{aligned}$$

On surnomme $1-q$ la « *tolerance probability* ».

Mode

Contexte

Le mode est la réalisation qui survient le plus souvent.

Par exemple, la lettre E est la lettre la plus utilisée dans le dictionnaire anglais. Elle représente donc le *mode* de la langue anglaise.

En termes mathématiques, le mode est le point qui maximise la PMF/PDF.

Dans le cas continu, on peut simplement dériver la PDF et trouver le point qui la rend égale à zéro. Si la distribution :

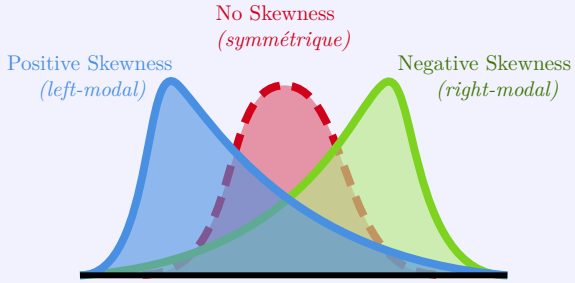
est unimodal, c'est-à-dire qu'elle a une « bosse », alors $\text{mode} = x \text{ tel que } f'(x) = 0$.

est strictement croissant ou décroissant, le mode sera une des deux extrémités.

– Par exemple, la loi exponentielle est strictement décroissante et a toujours un mode à 0 peu importe les paramètres.

Skewness

$$\text{Skewness} = \frac{\mu_3}{\sigma^3} = \frac{\mu'_3 - 3\mu'_2\mu + 2\mu^3}{\sigma^3}.$$



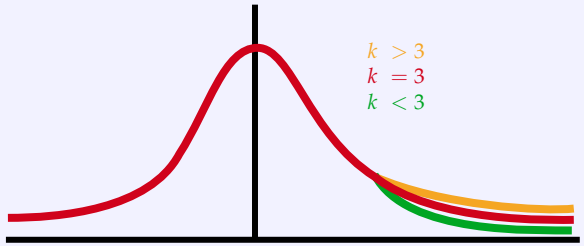
Kurtosis

Contexte

Le kurtosis mesure l'aplatissement d'une distribution et peut aider à juger la vraisemblance qu'une distribution produise des valeurs extrêmes (ou « *outliers* »).

$$\text{Kurtosis} = \frac{\mu_4}{\sigma^4} = \frac{\mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4}{\sigma^3}.$$

Le kurtosis de la distribution normale est de 3. On pose qu'il est plus vraisemblable pour une distribution dont le kurtosis supérieur à 3 de produire des valeurs extrêmes.



Distributions

Loi Pareto

Contexte

La distribution Pareto est un mélange de deux distributions exponentielles originalement conçue pour étudier des distributions de revenus.

Notation	Paramètres	Domaine
$X \sim \text{Pareto}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \dfrac{\alpha\theta^\alpha}{(x+\theta)^{\alpha+1}}$
$F(x)$	$= 1 - \left(\dfrac{\theta}{x+\theta}\right)^\alpha$

Si $X \sim \text{Pareto}(\alpha, \theta)$ alors $Y = (X - d | X > d) \sim \text{Pareto}(\alpha, \theta + d)$.

Loi Beta

Notation	Paramètres	Domaine
$X \sim \text{Beta}(a, b, \theta)$	$a, b > 0$ et $\theta \geq 0$	$x \in [0, \theta]$

$f(x)$	$= \dfrac{\theta}{\text{B}(a,b)} \left(\dfrac{x}{\theta}\right)^{a-1} \left(1 - \dfrac{x}{\theta}\right)^{b-1}$

$X \sim \text{Beta}(a = 1, b = 1, \theta) \sim \text{Unif}(0, \theta)$.

Si $X \sim \text{Unif}(a, b)$ alors $(X | X > d) \sim \text{Unif}(d, b)$ et $(X - d | X > d) \sim \text{Unif}(0, b - d)$.

Loi Gamma

Notation	Paramètres	Domaine
$X \sim \text{Gamma}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \dfrac{x^{\alpha-1} \text{e}^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}$

On appelle θ la moyenne et $\lambda = \frac{1}{\theta}$ le paramètre de fréquence (« *rate* »).

Soit n v.a. indépendantes $X_i \sim \text{Gamma}(\alpha_i, \theta)$ alors $\sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \theta)$.

Soit n v.a. indépendantes $X_i \sim \text{Exp}(\lambda_i)$ alors $Y = \min(X_1, \ldots, X_n) \sim \text{Exp}(\frac{1}{\sum_{i=1}^n \lambda_i})$.

Si $X \sim \text{Exp}(\theta)$ alors $(X - d | X > d) \sim \text{Exp}(\theta)$.

Loi de Weibull

Notation	Paramètres	Domaine
$X \sim \text{Weibull}(\tau, \beta)$	$\tau, \beta > 0$	$x \geq 0$

$f(x)$	$= \dfrac{\tau(x/\theta)^\tau \text{e}^{-(x/\theta)^\tau}}{x}$

La loi de Weibull est une transformation de la loi exponentielle ; pour $Y \sim \text{Exp}(\mu)$, alors $X = Y^{1/tau} \sim \text{Weibull}(\theta = \mu^{1/\tau}, \tau)$.

Note Voir la sous-section *Divers* de la section sur la *Théorie de la fiabilité* du chapitre *Critères d'information pour la sélection de modèles* pour l'interprétation de la fonction de hasard dans le contexte de la loi gamma, la loi exponentielle et la loi de Weibull.

Loi Erlang

Contexte

La loi Erlang est un cas spécial de la loi Gamma avec un paramètre de forme α entier. Elle est utile dans le contexte de **Processus de Poisson**, car nous pouvons trouver une forme explicite de la fonction de répartition (survie).

Notation	Paramètres	Domaine
$X \sim \text{Erlang}(n, \lambda)$	$\lambda > 0$ et $n \in \mathbb{N}^+$	$x \geq 0$

$f(x)$	$= \dfrac{x^{n-1}\lambda^n \mathrm{e}^{-\lambda x}}{\Gamma(n)}$
$S(x)$	$= \displaystyle\sum_{k=0}^{n-1} \dfrac{(\lambda x)^{k-1}\mathrm{e}^{-\lambda x}}{k!}$

Loi de Poisson

Notation	Paramètres	Domaine
$X \sim \text{Poisson}(\lambda)$	$\lambda > 0$	$x = 0, 1, 2, \ldots$

$\Pr(X = x)$	$= \dfrac{\mathrm{e}^{-\lambda}\lambda^x}{x!}$

Transformation

Changement d'échelle pour des v.a. continues

Toutes les distributions continues (sauf pour la lognormale, l'inverse gaussienne et la log-t) ont θ comme paramètre d'échelle. Alors, multiplier la v.a. par une constante c change uniquement le paramètre $\theta^* = c\theta$.

Trouver la PDF d'une v.a. transformée

Soit n v.a. $X_1, \ldots, X_n$ que l'on veut transformer en n autres variables aléatoires $W_1 = g_1(X_1, \ldots, X_n), \ldots, W_n = g_n(X_1, \ldots, X_n)$.

1. Trouver les inverses des équations de la transformation :

$$\begin{aligned} x_1 &= g_1^{-1}(w_1, \ldots, w_n) \\ &\vdots \\ x_n &= g_n^{-1}(w_1, \ldots, w_n) \end{aligned}$$

2. Calculer le déterminant de la matrice Jacobienne J :

$$J = \det \begin{bmatrix} \frac{\partial x_1}{\partial w_1} & \cdots & \frac{\partial x_1}{\partial w_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial w_1} & \cdots & \frac{\partial x_n}{\partial w_n} \end{bmatrix}$$

3. Trouver la fonction de densité conjointe avec

$$f_{W_1,\ldots,W_n}(w_1, \ldots, w_n) = f_{X_1,\ldots,X_n}\left(g_1^{-1}(w_1, \ldots, w_n), \ldots, g_n^{-1}(w_1, \ldots, w_n)\right) |J| .$$

Note Dans le cas univarié, $f_W(w) = f_X\left(g^{-1}(w)\right) \left|\frac{\partial g^{-1}(w)}{\partial w}\right|$.

Mélanges

Mélange discret de variables aléatoires

La variable aléatoire Y est un ***mélange discret*** des variables aléatoires $X_1, X_2, \ldots, X_n$ si sa fonction de densité (survie, répartition) peut être exprimée comme la moyenne pondérée des fonctions de densités des n v.a. : $f_Y(y) = \sum_{i=1}^n w_i f_{X_i}(y)$ où $\sum_{i=1}^n w_i = 1$ et $w_i \in [0, 1]$ pour $i = 1, 2, \ldots, n$.

Il s'ensuit également que $S_Y(y) = \sum_{i=1}^n w_i S_{X_i}(y)$, $F_Y(y) = \sum_{i=1}^n w_i F_{X_i}(y)$ et $\mathrm{E}[Y^k] = \sum_{i=1}^n w_i \mathrm{E}[X_i^k]$.

Note Voir *l'identité Poisson-Gamma* pour un exemple d'un mélange continue de variables aléatoires.

Queues de distributions

Contexte

Si une distribution a une queue de droite qui est lourde, « *thick* » ou « *fat* », alors elle a des probabilités élevées de pertes extrêmes. L'idée qu'une queue est « *épaisse* » ou « *lourde* » s'interprète comme l'écart entre la courbe de la fonction de densité et l'abscisse. Plus la courbe s'éloigne de l'abscisse, plus l'écart est plus « gras ». En revanche, une queue « *légère* » est proche de l'abscisse.

En situation d'examen nous ne pouvons pas visuellement évaluer la queue et donc nous utilisons un des 4 tests suivants :

1 Nombre de moments (positifs) qui existent

Plus la queue est **lourde**, *moins* il y a de moments qui existent.

Il devient de moins en moins probable que l'intégrale de $x^k f(x)$ va converger.

2 Ratio des fonctions de survie (ou PDF)

Plus la queue est **lourde**, *plus* la fonction de survie va tendre vers 0 ***lentement***.

Si $\lim_{x\to\infty} \frac{S_1(x)}{S_2(x)} = 0$ alors X_1 a une queue plus légère que X_2, et vice-versa si la limite tend vers ∞.

Par la règle de l'hôpital, ceci est équivalent pour le ratio des PDF.

3 Fonctions de hasard

Si la fonction de hasard est *décroissante*, il y a une probabilité plus élevée de pertes extrêmes et donc une queue ***lourde***.

4 CTEs (ou quantiles)

Plus le CTE (ou les quantiles) est large, plus les montants de pertes extrêmes sont larges et donc *plus* la queue est ***lourde***.

Estimations et types de données

Distributions empiriques

Notation

X Variable aléatoire de perte ;
θ Paramètre de la distribution de X ;
Le paramètre peut être un scalaire θ ou un vecteur $\boldsymbol{\theta}$;
Par exemple, pour une loi Gamma $\boldsymbol{\theta} = \{\alpha, \beta\}$;
Pour simplifier la notation, on le traite comme un scalaire θ.
$F_X(x;\theta)$ Fonction de répartition de X avec paramètre θ ;
Pour simplifier la notation, on écrit $F(x;\theta)$ sauf s'il faut être plus spécifique.
$f_X(x;\theta)$ Fonction de densité de X avec paramètre θ ;
Pour simplifier la notation, on écrit $f(x;\theta)$ sauf s'il faut être plus spécifique.
$\{X_1, \ldots, X_n\}$ Échantillon aléatoire de n observations de X ;
$\hat{\theta}$ Estimateur de θ établit avec l'échantillon aléatoire $\{X_1, \ldots, X_n\}$;
$F(x;\hat{\theta})$ Estimation *paramétrique* de la fonction de répartition de X ;
$f(x;\hat{\theta})$ Estimation *paramétrique* de la fonction de densité de X ;

Si θ est connu, la distribution de X est complètement spécifiée ;
En pratique, θ est inconnu et doit être estimé avec les données observées.

On peut estimer $F_X(x)$ et $f_X(x)$ directement pour toute valeur x sans présumer une forme paramétrique ;
Par exemple, un histogramme est une estimation *non paramétrique*.

Données complètes

Notation

X Variable d'intérêt (p. ex., la durée de vie ou la perte) ;
$\{X_1, \ldots, X_n\}$ Valeurs de X pour n individus ;
$\{x_1, \ldots, x_n\}$ n valeurs observées de l'échantillon ;
Il peut y avoir des valeurs dupliquées dans les valeurs observées.
$0 < y_1 < \ldots < y_m$ m valeurs distinctes où $m \leq n$;
w_j Nombre de fois que la valeur y_j apparaît dans l'échantillon pour $j = 1, \ldots, m$;

Il s'ensuit que $\sum_{j=1}^{m} w_j = n$;

Pour des données de mortalité, w_j individus décèdent à l'âge y_j ;
Si tous les individus sont observés de la naissance jusqu'à la mort c'est un « *complete individual data set* ».
r_j « *risk set* » au *temps* y_j ;
Le nombre d'individus exposés à la possibilité de mourir au temps y_j ;
Par exemple, $r_1 = n$, car tous les individus sont exposés au risque de décéder juste avant le temps y_1 ;

On déduit que $r_j = \sum_{i=j}^{m} w_i$, alias le nombre d'individus qui survivent juste avant le temps y_j.

Données incomplètes

Exemple

Soit une étude sur le nombre d'années nécessaire pour obtenir un diplôme universitaire. L'étude commence cette année et tient compte de tous les étudiants présentement inscrits, ainsi que ceux qui vont s'inscrire au courant de l'étude. Tous les étudiants sont observés jusqu'à la fin de l'étude et on note le nombre d'années nécessaire pour ceux qui complètent leurs diplômes.

Si un étudiant a commencé son cursus scolaire avant l'étude et suit présentement des cours, le chercheur a de l'information sur le nombre d'années qu'il a déjà investi. Cependant, d'autres étudiants qui se sont inscrits en même temps, mais ont cessé leurs études ne seront pas observés dans cet échantillon. Alors, l'individu est observé d'une population **tronquée à la gauche** puisque l'information sur les étudiants qui ont quitté l'université avant le début de l'étude n'est *pas disponible*.

Si un étudiant n'est pas encore diplômé lorsque l'étude prend fin, le chercheur ne peut pas savoir combien d'années supplémentaires seront nécessaires. Cet individu fait donc partie d'une population **censurée à la droite** puisque le chercheur a de l'information *partielle* (le nombre d'années minimal) sans savoir le nombre exact.

Notation

d_i État de troncature de l'individu i de l'échantillon ;

$d_i = 0$ s'il n'y a pas de troncature ;

Par exemple, un étudiant a commencé son programme universitaire d_i années avant le début de l'étude.

x_i Temps de "survie" de l'individu i ;

Par exemple, le nombre d'années avant d'obtenir son diplôme ;

Si l'étude prend fin avant que x_i soit observé, on dénote le temps de survie jusqu'à ce moment u_i ;

Donc chaque individu a *soit* une valeur x_i ou u_i, mais *pas les deux*.

Données groupées

Notation

$(c_0, c_1], (c_1, c_2], \ldots, (c_{k-1}, c_k]$ k intervalles regroupant les observations ;

$0 \leq c_0 < c_1 < \ldots < c_k$ Extrémités des k intervalles ;

n Nombre d'observations de x_i dans l'échantillon ;

n_j Nombre d'observations de x_i dans l'intervalle $(c_{j-1}, c_j]$;

Il s'ensuit que $\sum_{j=1}^{k} n_j = n$.

r_j « *risk set* » de l'intervalle $(c_{j-1}, c_j]$ lorsque les données sont complètes ;

Il s'ensuit que $r_j = \sum_{i=j}^{k} n_i$.

Applications en assurance

Notation

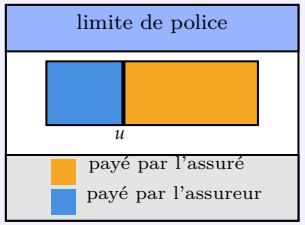
X Variable aléatoire du montant de perte.

Limite de police

Limite de police

Une **limite de police** u est le montant maximal qu'un assureur va payer pour une perte.

Visuellement :



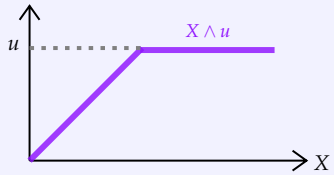
Montant de perte limité

La variable aléatoire du **montant de perte limité** $X \wedge u$ correspond au montant du paiement de l'assureur pour une police d'assurance ayant une limite de u :

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

Il s'ensuit que $X \wedge d = \min(X; d)$.

Visuellement :



L'espérance limitée du montant de perte

L'espérance limitée du montant de perte $\mathrm{E}[X \wedge u]$ correspond à l'espérance du paiement de l'assureur pour une police d'assurance ayant une limite de u :

$$\mathrm{E}[X \wedge u] = \int_0^u x f(x) dx + uS(u)$$

Déductibles

Déductible

Le **déductible d'une police** est le montant que l'assuré doit payer de sa poche avant que l'assureur débourse pour une perte.

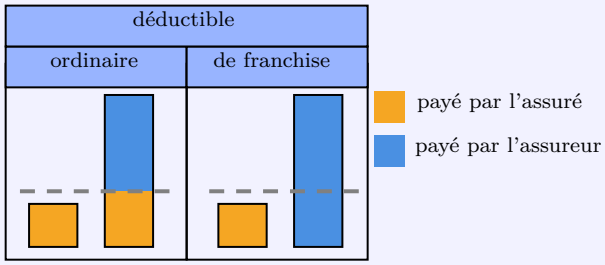
Il y a 2 types de déductibles :

déductible ordinaire Une fois que le montant de perte surpasse le déductible, l'assureur va payer le montant de la perte **en excès du déductible**.

déductible de franchise Une fois que le montant de perte surpasse le déductible, l'assureur va payer le montant **total** de la perte.

Par défaut, on suppose le déductible ordinaire.

Visuellement :



Déductible ordinaire

Montant de perte avec un déductible **ordinaire**

La variable aléatoire du montant de perte pour une police ayant un **déductible ordinaire** de d.

Assureur

$$(X-d)_+ = \begin{cases} 0, & X \leq d \\ X-d, & X > d \end{cases}$$

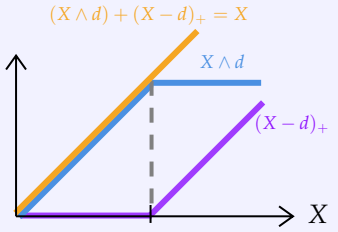
Assuré

$$X \wedge d = \begin{cases} X, & X < d \\ d, & X \geq d \end{cases}$$

Il s'ensuit que $\boxed{(X-d)_+ = \max(X-d;0)}$.

On observe que le montant de perte est la somme des contributions $\boxed{X = X \wedge d + (X-d)_+}$.

Visuellement :



L'espérance du montant de perte avec un déductible ordinaire

L'espérance du montant de perte, pour l'assureur, **avec un déductible *ordinaire*** $\boxed{\mathrm{E}[(X-d)_+]}$ correspond à :

$$\mathrm{E}[(X-d)_+] = \int_d^\infty (x-d) f(x) dx$$

« *Loss Elimination Ratio (LER)* »

Le « *Loss Elimination Ratio (LER)* » évalue combien qu'épargne l'assureur en imposant un déductible *ordinaire* de d, $\boxed{LER = \frac{\mathrm{E}[X \wedge d]}{\mathrm{E}[X]}}$.

« *payment per loss* » et « *payment per payment* »

Notation

Y^L Montant de perte.
« *payment per **loss*** »
Y^P Montant de paiement.
« *payment per **payment*** »

$\mathbf{E}[Y^L]$ Montant espéré de paiement **par perte *subie***.
$\mathbf{E}[Y^P]$ Montant espéré de paiement **par paiement *effectué***.

Par exemple, lorsqu'une police a un déductible, les pertes dont le coût est inférieur au déductible ne seront pas reportées à l'assureur.

Le montant de paiement est donc le montant que l'assureur va payer conditionnel à ce qu'il y ait un paiement.

Il s'ensuit que $\boxed{\mathrm{E}[Y^L] \geq \mathrm{E}[Y^P]}$.

Pour un déductible ordinaire de d,

$$\mathrm{E}[Y^L] = \mathrm{E}[(X-d)_+] \qquad \mathrm{E}[Y^P] = \mathrm{E}[X-d|X>d]$$

On trouve que $\mathrm{E}[Y^P] = \frac{\mathrm{E}[Y^L]}{S(d)}$.

Également, le montant espéré de paiement par paiement effectué *est* la fonction d'excès moyen $\mathrm{E}[Y^P] = e(d)$.

Si la police d'assurance comporte uniquement une limite, $Y^P = Y^L$.

Relations pour quelques distributions :

X	$(X-d\|X>d)$
$\mathrm{Exp}(\theta)$	$\mathrm{Exp}(\theta)$
$\mathrm{Unif}(a,b)$	$\mathrm{Unif}(0,b-d)$
$\mathrm{Pareto}(\alpha,\theta)$	$\mathrm{Pareto}(\alpha,\theta+d)$
$\mathrm{Beta}(1,b,\theta)$	$\mathrm{Beta}(1,b,\theta-d)$

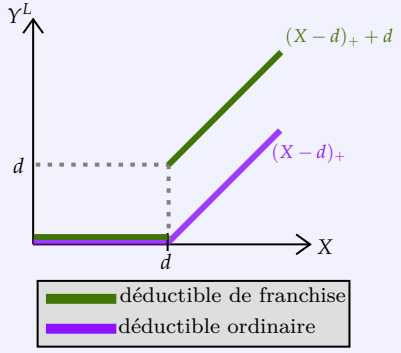
Déductible de franchise

Montant de perte avec un déductible de franchise

La variable aléatoire du montant de perte pour une police ayant un **déductible de franchise** de d.

$$(X|X>d) = \begin{cases} 0, & X \le d \\ X, & X > d \end{cases}$$

Visuellement :



L'espérance du montant de perte avec un déductible de franchise

L'espérance du montant de perte, pour l'assureur, **avec un déductible *de franchise*** $\mathrm{E}[X|X>d]$ correspond à :

$$\begin{aligned}\mathrm{E}[X|X>d] &= \int_d^\infty x f(x)dx = \int_d^\infty (x-d)f(x)dx + d\int_d^\infty f(x)dx \\ &= \mathrm{E}[(X-d)_+] + dS(d)\end{aligned}$$

Impacts du déductible sur la fréquence

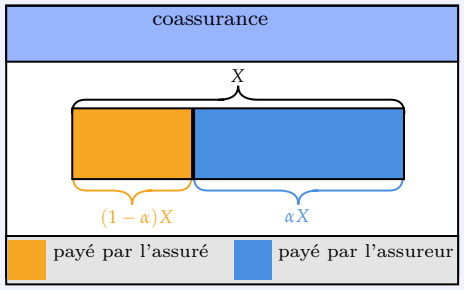
Pour la classe $(a,b,0)$ de distributions, on trouve les relations suivantes :

Nombre de pertes (N)	Nombre de paiements (N')
$\mathrm{Pois}(\lambda)$	$\mathrm{Pois}(S(d)\lambda)$
$\mathrm{Binom}(n,p)$	$\mathrm{Binom}(n,S(d)p)$
$\mathrm{BinNeg}(r,\beta)$	$\mathrm{BinNeg}(r,S(d)\beta)$

Coassurance

Coassurance α

Le pourcentage de coassurance α correspond à la portion de la perte payée par l'assureur. Pour une perte de X, l'assureur paye αX et l'assuré paye $(1-\alpha)X$.



L'espérance du montant de perte avec coassurance

L'espérance du montant de perte, pour l'assureur, **avec une coassurance de** α est $\mathrm{E}[\alpha X] = \alpha \mathrm{E}[X]$.

Combinaison des facteurs

Cas d'un déductible et de coassurance

Habituellement, la coassurance est appliquée ***après*** le déductible et la perte pour l'assureur est :

$$Y^L = \begin{cases} 0, & X \leq d \\ \alpha(X-d), & X > d \end{cases}$$

$$\mathrm{E}[Y^L] = \alpha\left(\mathrm{E}[X] - \mathrm{E}[X \wedge d]\right)$$

Si une question spécifie que la coassurance s'applique ***avant*** le déductible, il suffit de remplacer d par $\frac{d}{\alpha}$ et mettre le α en évidence comme avant :

$$Y^L = \begin{cases} 0, & \alpha X \leq d \\ \alpha X - d, & \alpha X > d \end{cases} = \begin{cases} 0, & X \leq \frac{d}{\alpha} \\ \alpha\left(X - \frac{d}{\alpha}\right), & X > \frac{d}{\alpha} \end{cases}$$

$$\mathrm{E}[Y^L] = \alpha\left(\mathrm{E}[X] - \mathrm{E}\left[X \wedge \frac{d}{\alpha}\right]\right)$$

Soit une police ayant :

1. une coassurance de α,
2. une limite de police de u,
3. un déductible ***ordinaire*** de d.

Alors, $\mathrm{E}[Y^L] = \alpha\left\{\mathrm{E}[X \wedge m] - \mathrm{E}[X \wedge d]\right\}$ et

$$Y^L = \begin{cases} 0, & X \leq d \\ \alpha(X-d), & d < X < m \\ u, & X \geq m \end{cases}$$

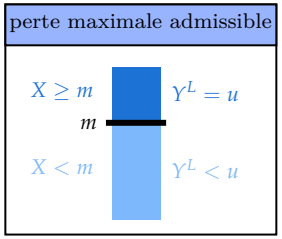
où m est la **perte maximale admissible**.

Perte maximale admissible m

Soit la perte maximale admissible $m = \frac{u}{\alpha} + d$ représentant la plus petite perte pour laquelle l'assureur paye la limite u.

En anglais, « *maximum covered loss* ».

Visuellement :



Inflation

Inflation r

L'inflation de r augmente les coûts, mais, de façon générale, ils sont couverts par la compagnie d'assurance et ne causent pas de changements à la police.

L'espérance du montant de perte avec inflation

L'espérance du montant de perte, pour l'assureur, **avec de l'inflation de r** est $\mathrm{E}[(1+r)X] = (1+r)\mathrm{E}[X]$.

Combiné avec les autres facteurs :

$$\mathrm{E}\left[Y^L\right] = \alpha(1+r)\left(\mathrm{E}\left[X \wedge \frac{m}{1+r}\right] - \mathrm{E}\left[X \wedge \frac{d}{1+r}\right]\right)$$

$$\mathrm{E}\left[Y^P\right] = \frac{\mathrm{E}[Y^L]}{S_X\left(\frac{d}{1+r}\right)}$$

Note Si la distribution de X comporte un paramètre d'échelle θ, on peut simplifier les équations en posant $\theta' = (1+r)\theta$.

Estimation de modèles non paramétriques

Contexte

Si on pose une distribution discrète, on utilise la fonction de répartition empirique pour l'estimer à partir d'un échantillon d'observations. Pour une observation x_i, la fonction de répartition empirique assigne une masse de probabilité de $1/n$ au point x_i.

Cependant, si l'on suppose une distribution continue, on désire *distribuer* cette masse *autour* de x_i. En lieu de supposer une distribution continue pour $f(x)$, puis d'estimer ses paramètres, on peut choisir de directement estimer la fonction de densité avec un **estimateur à noyau de la densité**.

On débute avec le cas continu en expliquant la *Distribution par noyau*, puis on explique le cas discret avec la *Distribution empirique*.

Distribution par noyau

Fonction noyau $k()$

La fonction noyau $k()$ est une *fonction de densité* à deux paramètres (x_i et b). Chaque observation a sa propre fonction noyau $k_i(x)$.

Contexte

Les fonctions noyau faisant partie de l'examen sont **symétriques** avec x_i comme point milieu.

i^{e} valeur observée x_i

La réalisation x_i est un paramètre pour la fonction noyau $k_i(x)$. Il est important de ne pas confondre **le paramètre x_i** avec le **point auquel on évalue la fonction de densité x**.

Contexte

Puisque la fonction noyau est symétrique et centrée sur l'observation x_i, la i^{e} valeur observée x_i **représente la moyenne de la distribution** liée à la fonction noyau.

Largeur de la bande b

L'interprétation de la largeur de la bande b varie selon la fonction noyau, mais de façon générale ça représente l'étendu de la densité.

En anglais, « *bandwith* ».

Estimer une fonction de densité par une fonction noyau

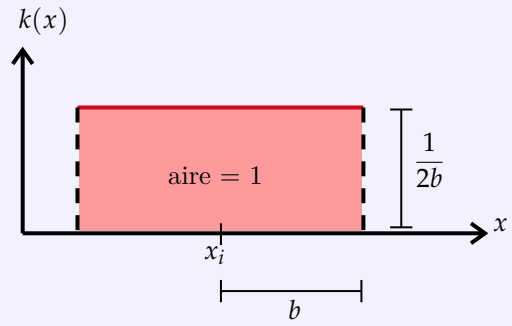
1. Choisir un type de fonction de densité pour $k()$.
2. Estimer la fonction de densité $f(x)$ comme la moyenne des fonctions noyau des n observations $k_1(x), \ldots, k_n(x)$:

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} k_i(x)$$

Noyau rectangulaire (uniforme)

Noyau rectangulaire ou uniforme

Le noyau rectangulaire, ou uniforme, suppose une densité distribuée uniformément :



La longueur de bande b représente donc la **distance du milieu x_i à la fin du domaine**.

Par géométrie, on obtient une largeur de $2b$ et, puisque l'aire doit être de 1, une hauteur de $\frac{1}{2b}$.

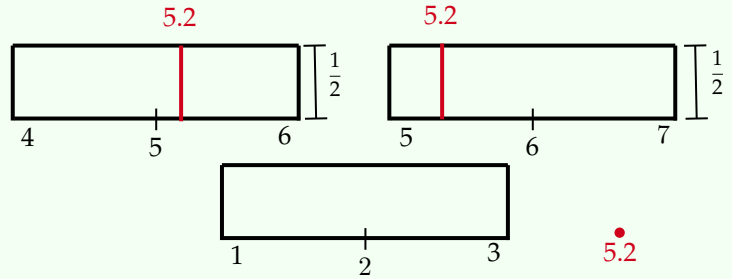
En termes mathématiques :

$$k_i(x) = \begin{cases} \frac{1}{2b}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{sinon} \end{cases}$$

Exemple de noyau rectangulaire

On observe les montants de réclamation $\{5, 2, 6\}$. Pour un noyau rectangulaire avec une longueur de bande $b = 1$, on désire estimer la fonction de densité évaluée à 5.2.

1. On interprète le problème comme $\tilde{f}(5.2) = \frac{1}{3}(k_1(x) + k_2(x) + k_3(x))$.
2. On visualise les fonctions de noyau :

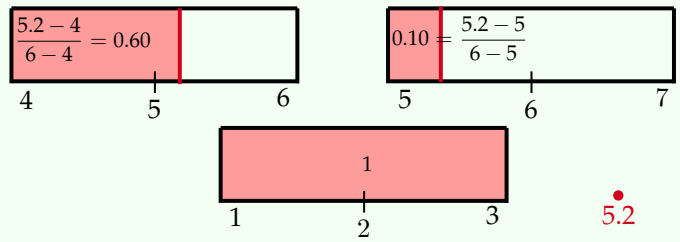


3. La fonction de densité estimée est donc :

$$\tilde{f}(5.2) = \frac{1}{3}\left(\frac{1}{2} + 0 + \frac{1}{2}\right) = \frac{1}{3}$$

Si on désire trouver la probabilité que la réclamation soit inférieure à 5.2 :

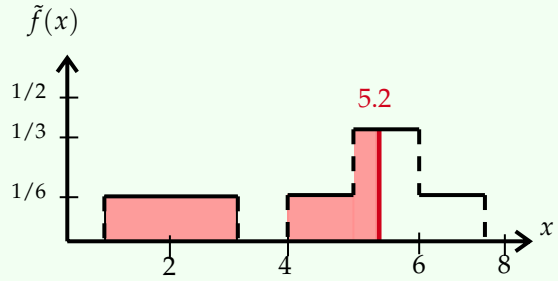
1. Visuellement, on voit comment l'équivalence géométrique du calcul des probabilités :



2. Donc :

$$\tilde{F}(5.2) = \frac{1}{3}(0.60 + 1 + 0.10) = 0.567$$

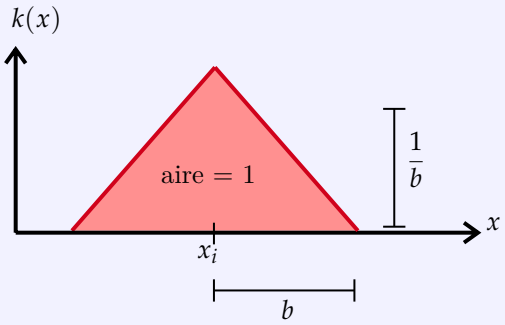
Visuellement, la densité par noyau est :



Noyau triangulaire

Noyau triangulaire

Le noyau triangulaire prend la forme d'un triangle isocèle :



La longueur de bande b représente donc la **distance du milieu x_i à la fin du domaine**.

Par géométrie, on obtient une largeur de $2b$ et, puisque l'aire doit être de 1, une hauteur de $\frac{1}{b}$.

En termes mathématiques :

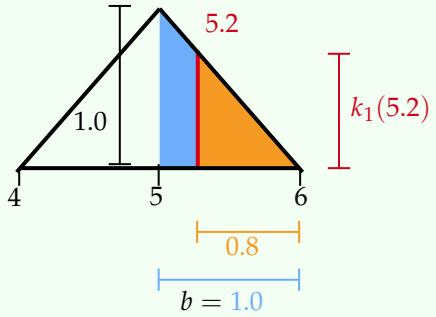
$$k_i(x) = \begin{cases} \frac{b-|x-x_i|}{b^2}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{sinon} \end{cases}$$

Note Pour calculer des probabilités, il est bien mieux de se **faire un dessin** et **utiliser la géométrie** que de mémoriser les formules.

Exemple noyau rectangulaire

Une propriété des triangles isocèles est que la ratio des hauteurs doit être égale au ratio des bases du triangle.

Pour le même exemple qu'avant, mais avec un noyau rectangulaire ce coups-ci, on trouve visuellement $k_1(5.2) = k_5(5.2)$:



$$\frac{k_1(5.2)}{1.0} = \frac{0.8}{1.0}$$
$$k_1(5.2) = \frac{0.8}{1.0} \times 1.0$$

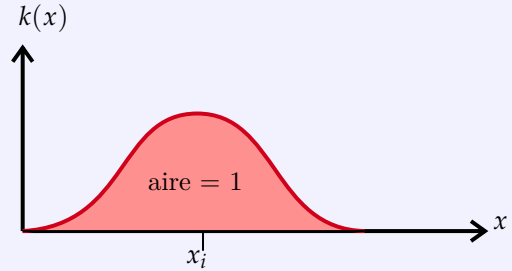
Puis, on désire estimer $F(5.5)$.

1. Trouver $K_5(5.5)$, $K_2(5.5)$ et $K_6(5.5)$.
   a) $2 \leq 5.5 - 1$ donc, $K_2(5.5) = 1$.
   b) Pour 6, on trouve l'aire du rectangle comme étant :
   $$\left(\frac{5.5-5}{6-5} \times \frac{1}{1} \times (5.5-5)\right)/2 = 0.125$$
   $$\Rightarrow K_6(5.5) = 0.125$$
   c) Pour 5, on trouve que $K_5(5.5) = (0.8 \times 0.8)/2 = 0.32$.
2. Calculer la moyenne pondérée des observations comme l'estimation : $\hat{F}(10) = \frac{1}{3}(0.125 + 1 + 0.32) = 0.482$.

Noyau gaussien

Noyau gaussien

Le noyau gaussien prend la forme d'une densité normale de moyenne x_i et variance b^2 :



La longueur de bande b représente donc **l'écart-type de la distribution**.

En termes mathématiques, pour $x \in (-\infty, \infty)$,

$$k_i(x) = \frac{1}{b\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-x_i}{b}\right)^2}$$

Note Le noyau gaussien est le seul dont les probabilités doivent être calculées algébriquement.

Distribution empirique

Section à compléter avec mes notes d'IARD et 11.2 de Nonlife Actuariel Models (tse).

Données complètes

Distribution empirique

Distribution discrète prenant comme valeurs $y_1, \ldots, y_m$ avec probabilités $\frac{w_1}{n}, \ldots, \frac{w_m}{n}$;

On peut également la définir comme la distribution discrète équiprobable des valeurs $x_1, \ldots, x_n$.

Notation

$\hat{f}()$ Fonction de densité empirique.

$\hat{F}()$ Fonction de répartition empirique.

$\tilde{F}()$ Fonction de répartition lissée ;

En anglais, « *smoothed empirical distribution function* ».

On appelle parfois la fonction de répartition *la fonction distribution* (« *distribution function* »).

$$\hat{f}(y) = \begin{cases} \frac{w_j}{n}, & \text{si } y = y_j \, \forall j \\ 0, & \text{sinon} \end{cases}$$

$$\hat{F}(y) = \begin{cases} 0, & y < y_1, \\ \frac{1}{n}\sum_{h=1}^{j} w_h, & y_j \leq y < y_{j+1},\ j = 1, \ldots, m-1 \\ 1, & y_m \leq y \end{cases}$$

On peut estimer la valeur de $\hat{F}()$ pour un une valeur de y pas dans l'ensemble $y_1, \ldots, y_m$ avec la fonction de répartition lissée $\tilde{F}()$. Pour $y_j \leq y < y_{j+1}$ et $j \in \{1, 2, \ldots, m-1\}$, $\tilde{F}(y)$ est une interpolation linéaire de $\hat{F}(y_{j+1})$ et $\hat{F}(y_j)$:

$$\tilde{F}(y) = \frac{y - y_j}{y_{j+1} - y_j} \hat{F}(y_{j+1}) + \frac{y_{j+1} - y_j}{y_{j+1} - y_j} \hat{F}(y_j)$$

Distribution binomiale de la fonction de répartition empirique

On peut écrire la fonction de répartition empirique comme $\hat{F}(y) = \frac{Y}{n}$ où Y est le nombre d'observations qui sont inférieures ou égales à y tel que $Y \sim \text{Bin}(n, p = F(y))$.

On trouve :

$$\text{E}[Y] = \frac{\text{E}[\hat{F}(y)]}{n} = F(y)$$
$$\text{Var}(Y) = \frac{\text{Var}(\hat{F}(y))}{n^2} = \frac{F(y)(1-F(y))}{n}$$

Données incomplètes

Section à compléter avec mes notes d IARD et 11.2 de Nonlife Actuariel Models (tse).

Estimateur de Kaplan-Meier Soit :

$$S(y_j) = \Pr(X > y_1)\Pr(X > y_2|X > y_1)\dots\Pr(X > y_j|X > y_{j-1}) = \Pr(X > y_1)\prod_{h=2}^{j}$$

Où on peut estimer $\widehat{\Pr}(X > y_1) = 1 - \frac{w_1}{r_1}$ et $\widehat{\Pr}(X > y_h|X > y_{h-1}) = 1 - \frac{w_h}{r_h}$ pour $h = 2, \dots, m$.

Il s'ensuit qu'on peut estimer $S(y_j)$ par :

$$\hat{S}(y_j) = \prod_{h=1}^{j}\left(1 - \frac{w_h}{r_h}\right)$$

Variance de l'estimateur Kaplan-Meier : $\text{Var}(\hat{S}_K(y_j)|\mathcal{C}) \approx \left(S(y_j)\right)^2\left(\sum_{h=1}^{j}\frac{1-s_h}{s_h r_h}\right)$

Approximation de Greenwood de la variance de l'estimateur Kaplan-Meier : $\widehat{\text{Var}}(\hat{S}_K(y_j)|\mathcal{C}) \approx \left(\hat{S}_K(y_j)\right)^2\left(\sum_{h=1}^{j}\frac{w_h}{r_h(r_h-w_h)}\right)$

Estimateur de Nelson-Aalen

Notation

$h(y)$ Fonction de hasard.
$H(y)$ Fonction de hasard cumulative.

$$H(y) = \int_0^y h(y)dy$$

Il s'ensuit que $S(y) = \text{e}^{-H(y)}$ et $H(y) = -\ln\left(S(y)\right)$.

Avec l'approximation $-\ln\left(1 - \frac{w_h}{r_h}\right) \approx \frac{w_h}{r_h}$ on trouve que $H(y) = \sum_{h=1}^{j}\frac{w_h}{r_h}$ qui correspond à l'**estimateur Nelson-Aalen** de la fonction de hasard cumulative.

Données groupées

Section à compléter avec mes notes d IARD et 11.3 de Nonlife Actuariel Models (tse).

Estimation de modèles paramétriques

Note Cette section ce veut une continuation de Méthode du maximum de vraisembla du chapitre Table des matières.

Estimation par maximum de vraisemblance pour des données incomplètes et groupées

Contexte

Lorsque les données sont groupées et/ou incomplètes, les observations ne sont plus iid. Cependant, on peut quand même formuler la fonction de vraisemblance et trouver l'estimateur du maximum de vraisemblance (EMV).

La première étape est d'écrire la fonction de (log) vraisemblance adéquate pour la méthode d'échantillonnage des données.

Fonction de vraisemblance

Données complètes

$$\mathcal{L}(\theta;\boldsymbol{x}) = \prod_{j=1}^{k} \underbrace{f(x_j;\theta)}_{\substack{\text{probabilité que chaque} \\ \text{observation soit égale à} \\ \text{la valeur observée}}}$$

Données groupées en k intervalles

La probabilité qu'une observation soit contenue dans l'intervalle $(c_{j-1}, c_j]$ est $F(c_j;\theta) - F(c_{j-1};\theta)$.

On pose que les observations *individuelles* sont iid afin d'obtenir que la vraisemblance d'avoir n_j observations dans l'intervalle $(c_{j-1}, c_j]$, pour $j = 1,\ldots,k$ et $\boldsymbol{n} = (n_1,\cdots,n_k)$, est :

$$\mathcal{L}(\theta;\boldsymbol{n}) = \prod_{j=1}^{k} \underbrace{[F(c_j;\theta) - F(c_{j-1};\theta)]^{n_j}}_{\substack{\text{probabilité qu'une observation} \\ \text{soit contenue dans l'intervalle}}}$$

Données censurées vers la droite

On pose que n_1 observations sont complètes et que n_2 observations sont censurées à la limite de u :

$$\mathcal{L}(\theta;\boldsymbol{x}) = \underbrace{\left[\prod_{i=1}^{n_1} f(x_i;\theta)\right]}_{\substack{\text{probabilité de chaque} \\ \text{observation à la} \\ \text{valeur observée}}} \overbrace{[1 - F(u;\theta)]^{n_2}}^{\substack{\text{probabilité qu'une observation} \\ \text{soit supérieure, ou égale, à } u}}$$

tronquées vers la gauche

On pose un déductible de d :

$$\mathcal{L}(\theta;\boldsymbol{x}) = \underbrace{\frac{1}{[1 - F(d;\theta)]^{n}}}_{\substack{\text{pondère la vraisemblance par} \\ \text{la probabilité d'être supérieur} \\ \text{au déductible}}} \prod_{i=1}^{n} f(x_i;\theta)$$

Évaluation et sélection de modèles

Cette section n'est pas suffisamment bien expliquée pour que je la considère complète.

Contexte

Évaluer les modèles avec des méthodes non paramétriques a l'avantage d'avoir très peu d'hypothèses. Cependant, il est plus difficile d'évaluer le modèle d'un point de vue théorique.
Évaluer les modèles avec des méthodes paramétriques a l'avantage de résumer le modèle à un petit nombre de paramètres. Cependant, ces méthodes sont une simplification et risquent d'imposer la mauvaise structure.

Graphiquement

Avec les méthodes d'évaluation visuelles, on peut détecter si les données diffèrent anormalement du modèle paramétrique.

On peut évaluer la fonction de répartition empirique et la fonction de répartition théorique sur un même graphique pour évaluer l'ajustement.

On peut évaluer le tracé des probabilités (« *P-P plot* ») qui trace la répartition empirique et la répartition théorique.

On peut tracer l'histogramme des données et superposer la densité théorique pour évaluer l'ajustement.

Le désavantage de ces méthodes est qu'elles ne fournissent pas des mesures quantitatives sur l'ajustement du modèle.

Tests pour la qualité de l'ajustement

Tests de spécification (« *misspecification tests* »)

Test de signifiance dont l'objectif est d'évaluer les hypothèses de distribution d'un modèle.

Notation

$F^*()$ Fonction de répartition d'une v.a. continue (hypothèse nulle).
$\hat{F}()$ Fonction de répartition empirique.

Les tests de Kolmogorov-Smirnov (K.-S.) et de Anderson-Darling sont idéaux lorsque l'on désire comparer les fonctions de répartition.

Voir la section *Tests d'adéquation* du chapitre de *Table des matières* pour la présentation du test de Kolmogorov-Smirnov. Nous présentons ci-dessous le test pour des données incomplètes.

Test de Kolmogorov-Smirnov pour des données incomplètes

Pour des données tronquées vers la gauche à d, il suffit d'ajuster la fonction de répartition théorique et de poser
$F^*(x) = \dfrac{F(x) - F(d)}{1 - F(d)}$
où $F(\cdot)$ est la fonction de répartition de la distribution théorique de X. Ceci nous donne donc que $F^*(\cdot)$ est la fonction de répartition de $(X|X > d)$.

Pour des données censurée vers la droite à u, la distribution théorique n'est pas affectée. Cependant, la valeur de $\hat{F}(m)$ n'est pas définie plutôt qu'être 1. De plus, n inclut les valeurs censurées pour compter le nombre total d'observations.

Lorsque les paramètres sont connus, le test de K.-S. n'est pas spécifique à aucune distribution avec des valeurs critiques générales. Le test de Anderson-Darling (A.-D.) considère toutes les différences $(\hat{F}(x) - F^*(x))$ et non seulement la différence maximale. Également, elle attribue plus de poids aux queues de la distribution en pondérant par la fonction de répartition et de survie :

$$A^2 = n \int \frac{(\hat{F}(x) - F^*(x))^2}{F^*(x)S^*(x)} f^*(x)dx$$

Donc, lorsque $F^*(x)$ ou $S^*(x)$ est petit, la différence est attribuée plus de poids.

Il s'ensuit que le test de A.-D. est « spécifique par distribution » dans le sens que les valeurs critiques sont différentes selon la distribution sous-jacente—il y a une table de valeurs critiques pour une distribution normale, Weibull, exponentielle, etc.

Test de Anderson-Darling

L'intégrale ci-dessus se simplifie à :

$$A^2 = -n - \frac{1}{n}\left[\sum_{j=1}^{n}(2j-1)\log\left(F^*(x_{(j)})\left[1 - F^*\left(x_{(n+1-j)}\right)\right]\right)\right]$$

Le test du khi carré sert à tester les hypothèses d'une distribution en comparant les fréquences observées aux fréquences théoriques.

Test d'adéquation du khi carré

Le test du rapport de vraisemblance teste la validité des restrictions d'un modèle et peut décider si un modèle peut être simplifié.

Test du rapport de vraisemblance

Critères d'information pour la sélection de modèles

Lorsque l'on compare deux modèles, on dit qu'un modèle est « emboîté » si l'autre comporte tous ses paramètres. Par exemple, un modèle basé sur une distribution exponentielle est emboîté par un modèle basé sur une distribution gamma ayant le même paramètre de fréquence β.

Il s'ensuit que le modèle comportant le plus de paramètres aura l'avantage de mieux s'ajuster aux données avec une fonction plus flexible et, possiblement, une log-vraisemblance plus élevée. Afin de comparer les modèles sur une même base, on utilise la **log-vraisemblance pénalisée**.

Critère d'information d'Akaike (AIC)

L'AIC pénalise les modèles ayant plus de paramètres en soustrayant le nombre de paramètres estimés p du modèle de la log-vraisemblance : $AIC = \log \mathcal{L}(\hat{\theta}_n^{\text{EMV}}; \boldsymbol{x}) - p$.

On choisit le modèle qui minimise l'AIC.

En anglais, « *Akaike Information Criterion (AIC)* ».

Le désavantage de l'AIC est que, pour deux modèles emboîtés, la probabilité de choisir le modèle plus simple (p. ex., un modèle basé sur la distribution exponentielle au lieu de la distribution gamma) *alors qu'il est vrai* (erreur de type I) ne tends pas vers 1 lorsque le nombre d'observations tend vers l'infini. On dit donc que c'est une mesure « *inconsistent* ».

Critère d'information bayésien (BIC)

Le BIC pénalise plus sévèrement les modèles ayant plus de paramètres : $BIC = \log \mathcal{L}(\hat{\theta}_n^{\text{EMV}}; \boldsymbol{x}) - \frac{p}{2} \log(n)$.

En anglais, « *Bayesian Information Criterion (BIC)* »

Le BIC est « *consistent* » et règle le désavantage de l'AIC avec une probabilité de 1 d'éviter une erreur de type I lorsque la taille de l'échantillon tend vers l'infini.

Dans les deux cas, la probabilité de rejeter le modèle plus simple lorsque le vrai modèle est entre les deux tend vers 1.

IV

Sujets divers

Optimisation numérique

Algorithmes « *Greedy* »

Méthode de résolution de problèmes qui prend la décision optimale **à chaque étape** d'obtenir la solution optimale d'un problème.

On dit que ces algorithmes sont « *greedy* », car, à chaque étape, ils prennent la meilleure décision sans tenir compte des choix futurs qui pourraient être plus optimaux. Donc, la solution trouvée n'est pas nécessairement la solution optimale.

Ces algorithmes ont l'avantage d'être **plus rapide**s au coût d'être **moins précis**.

Théorie de la fiabilité

Théorie de la fiabilité

Contexte : Un *système* ayant plusieurs *composantes*.
Idée : Le fonctionnement du système dépend du fonctionnement de ses composantes.
La **théorie de la fiabilité** évalue la probabilité qu'un *système* fonctionne selon la fiabilité de ses composantes et leurs rôles dans le système.

Introduction aux systèmes

Notation

x_i **État** de la composante i.
$\phi(\boldsymbol{x})$ « ***Structure function*** » désignant l'**état** d'un système.

L'état d'une composante

Chacune des composantes du système a sa propre ***durée de vie*** (« *lifetime* ») désignée par la variable aléatoire binaire de son état x_i.

Soit la composante **fonctionne**, ou elle **ne fonctionne pas** :

$$x_i = \begin{cases} 1, & \text{si la composante fonctionne} \\ 0, & \text{si la composante ne fonctionne pas} \end{cases}$$

Vecteur des états d'un système (« *path vector* »)

Le ***vecteur des états*** d'un système (« *state vector* ») regroupe les états de toutes les composantes d'un système et se dénote comme $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$. Il indique quelles composantes du système fonctionnent ou ne fonctionnent pas.

Note Un système ayant n composantes a 2^n différentes combinaisons possibles d'états de ses composantes. C'est-à-dire, 2^n différents *vecteur des états* possibles.

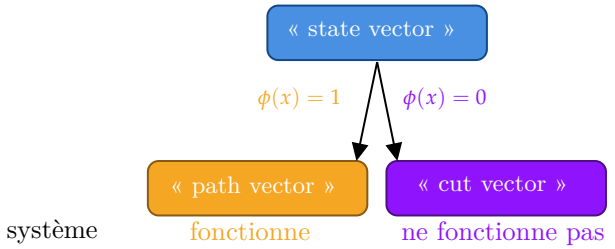
Puisque les composantes du système sont binaires, chacune prend une de deux valeurs ce qui résulte en $2 \times 2 \times \cdots 2 = 2^n$ différentes combinaisons possibles.

L'état d'un système

L'état d'un système dépend des états de ses composantes. L'état du système s'écrit sous la forme d'une fonction binaire $\phi(\boldsymbol{x})$:

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{si le système fonctionne} \\ 0, & \text{si le système ne fonctionne pas} \end{cases}$$

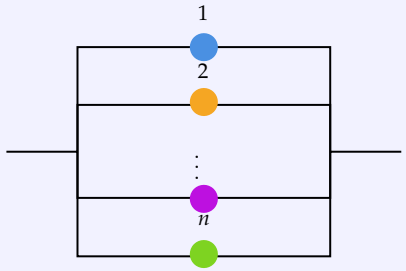
On distingue 2 types de vecteurs d'états selon le fonctionnement du système :



Types de systèmes les plus courants

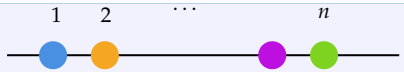
Système en parallèle

Un ***système en parallèle*** fonctionne tant qu'au moins une de ses composantes fonctionne.



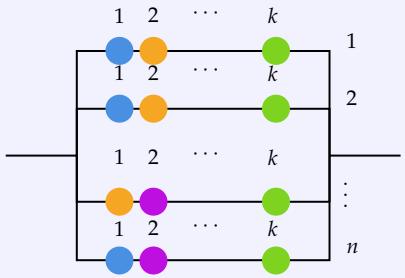
Système en série

Un ***système en série*** fonctionne seulement si toutes ses composantes fonctionnent.



Système de k parmi n

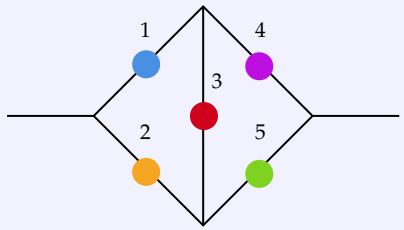
Un ***système de k parmi n*** n fonctionne si au moins k de ses n composantes fonctionnent.



Un système en parallèle est donc un système de 1 parmi n et un système en série un système de n parmi n.

Système de pont

Il y a deux branches connectées par un pont dans le milieu.



Contexte

On peut construire une infinité de systèmes comme des combinaisons des systèmes précédents. Entre autres, on peut combiner des systèmes ensembles. Par exemple : construire un système en série de systèmes en parallèle.

C'est lorsque nous créons des combinaisons que l'état de fonctionnement du système devient moins clair et c'est pourquoi les « *minimal path* » et « *minimal cut* » sets sont utiles.

Minimal path and minimal cut sets

« *Path vector* »

Le « *path vector* » est le vecteur d'états pour lequel le système fonctionne $(\phi(\boldsymbol{x}) = 1)$.

« *Minimal path vectors* »

Les « *minimal path vectors* » sont les « *path vectors* » ayant le *minimum* de composantes pour **fonctionner**. Il s'ensuit que dès qu'une des composantes d'un « *minimal path vector* » échoue, le système en entier cesse de fonctionner.

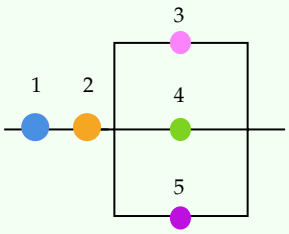
En termes mathématiques, $\boldsymbol{x}$ est un « *minimal path vector* » si $\phi(\boldsymbol{y}) = 0 \forall \boldsymbol{y} < \boldsymbol{x}$.

$\boldsymbol{y} < \boldsymbol{x}$ implique que tous les éléments y_i du vecteur $\boldsymbol{y}$ sont inférieurs ou égaux aux éléments x_i du vecteur $\boldsymbol{x}$ ($y_i \leq x_i \ \forall i$) avec au moins un élément qui est strictement inférieur ($y_i < x_i$ pour au moins un i).

« *Minimal path sets* »

Les « *minimal path sets* » sont les ensembles minimaux de composantes dont le fonctionnement garanti le fonctionnement du système. Donc, le système fonctionne uniquement si toutes les composantes d'au moins un des « *minimal path sets* » fonctionne.

Exemple de système



Minimal path sets	Minimal path vectors
1 $\{1, 2, 3\}$	1 $\boldsymbol{x} = (1, 1, 1, 0, 0)$
2 $\{1, 2, 4\}$	2 $\boldsymbol{x} = (1, 1, 0, 1, 0)$
3 $\{1, 2, 5\}$	3 $\boldsymbol{x} = (1, 1, 0, 0, 1)$

Afin de bien comprendre la condition selon laquelle un vecteur est classifié comme un « *mimimal path vector* », on observe les vecteurs $\boldsymbol{y}$ du premier « *minimal path vector* » $\boldsymbol{x}$:

Minimal path vector
1 $\boldsymbol{x} = (1, 1, 1, 0, 0)$

$\boldsymbol{y} < \boldsymbol{x}$
1 $(0, 0, 0, 0, 0), (0, 0, 1, 0, 0), (0, 1, 0, 0, 0),$ $(1, 0, 0, 0, 0), (0, 1, 1, 0, 0), (1, 1, 0, 0, 0),$ $(1, 0, 1, 0, 0)$

On note que $\phi(\boldsymbol{y}) = 0$ pour tous les vecteurs ce qui fait de x un « *minimal path vector* ».

« *Cut vector* »

Le « *cut vector* » est le vecteur d'états pour lequel le système ne fonctionne **pas** ($\phi(\boldsymbol{x}) = 0$). C'est donc l'inverse du « *path vector* ».

« *Minimal cut vectors* »

Les « *minimal cut vectors* » sont les « *cut vectors* » ayant le *maximum* de composantes pour ne **pas fonctionner**. Il s'ensuit que dès qu'une des composantes « brisée » est réparée, le système fonctionne.

En termes mathématiques, $\boldsymbol{x}$ est un « *minimal cut vector* » si $\phi(\boldsymbol{y}) = 1 \ \forall \boldsymbol{y} > \boldsymbol{x}$.

$\boldsymbol{y} > \boldsymbol{x}$ implique que tous les éléments y_i du vecteur $\boldsymbol{y}$ sont supérieurs ou égaux aux éléments x_i du vecteur $\boldsymbol{x}$ ($y_i \geq x_i \forall i$) avec au moins un élément qui est strictement supérieur ($y_i > x_i$ pour au moins un i).

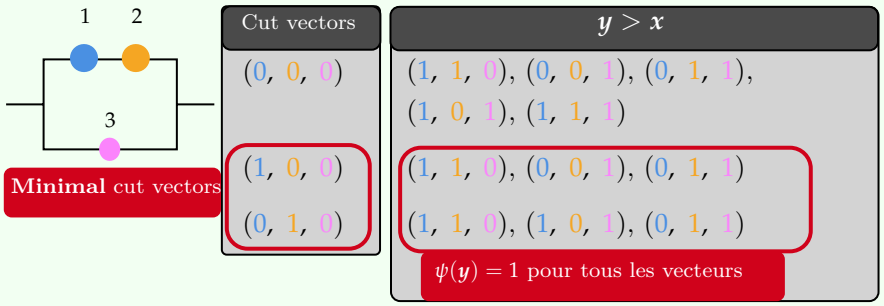
« *Minimal cut sets* »

Les « *minimal path sets* » sont les ensembles minimaux de composantes $\mathcal{C}$ dont l'échec garanti l'échec du système. Donc, le système cesse de fonctionner uniquement si toutes les composantes d'au moins un des « *minimal cut sets* » cessent de fonctionner.

En termes mathématiques, un « *minimal cut set* » C étant donné un « *minimal cut vector* $\boldsymbol{x}$ » est $\{i : x_i = 0\}$.

Exemple « *minimal cut sets* »

On peut visualiser ci-dessous que les « *minimal cut vectors* » sont les « *cut vectors* » pour lesquels tous les vecteurs $\boldsymbol{y}$ fonctionnent ($\psi(\boldsymbol{y}) = 1$).



	Nombre de	
Système	« miminal path sets »	« miminal cut sets »
Parallèle	n	1
Série	1	n
k parmi n	$\binom{n}{k}$	$\binom{n}{n-k+1}$
Pont	4	4

Pour un système composé de plusieurs systèmes, le nombre de vecteurs dépend de la façon dont il est construit :

Nombre de	Organisation du système	Action
« *minimal path sets* »	parallèle	somme
	série	produit
« *minimal cut sets* »	parallèle	produit
	série	somme

Structure Functions

Notation

$A_1, \ldots, A_s$ « *Minimal path sets* ».
$C_1, \ldots, C_m$ « *Minimal cut sets* ».

La « *structure function* » d'un système peut être déduite par deux approches :

1. Approche par les « *minimal path sets* ».
2. Approche par les « *minimal cut sets* ».

Cela dit, la fonction de base est fonction de la méthode d'organisation du système :

1 Système en parallèle

Un système en parallèle fonctionne tant qu'au moins une des composantes fonctionne. Alors, tant qu'au moins une des composantes i a un état de $x_i = 1$, l'état du système est de $\phi(\boldsymbol{x}) = 1$.

$$\begin{aligned}\phi(\boldsymbol{x}) &= \max\{x_1, \ldots, x_n\} \\ &= 1 - \prod_{i=1}^{n}(1 - x_i)\end{aligned}$$

La deuxième formulation découle du fait que les états sont des variables binaires.

2 Système en série

Un système en parallèle fonctionne ssi toutes les composantes fonctionnent. Alors, dès qu'une composante i a un état de $x_i = 0$, l'état du système est de $\phi(\boldsymbol{x}) = 0$.

$$\begin{aligned}\phi(\boldsymbol{x}) &= \min\{x_1, \ldots, x_n\} \\ &= \prod_{i=1}^{n} x_i\end{aligned}$$

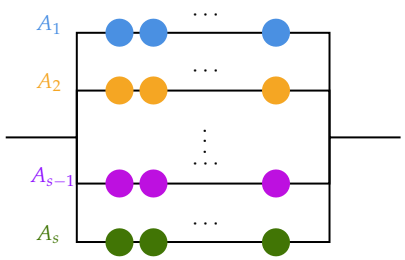
La deuxième formulation découle du fait que les états sont des variables binaires.

Approche par les « *minimal path sets* »

Soit ces deux constats :

1. Un système fonctionne ssi toutes les composantes d'au moins un des « *minimal path sets* » fonctionnent.
2. Un système en parallèle fonctionne ssi au moins une des composantes fonctionnent.

Alors, tout système peut être traité comme le système en parallèle de ses « *minimal path sets* » :



Il s'ensuit qu'on peut réécrire la fonction du système comme :

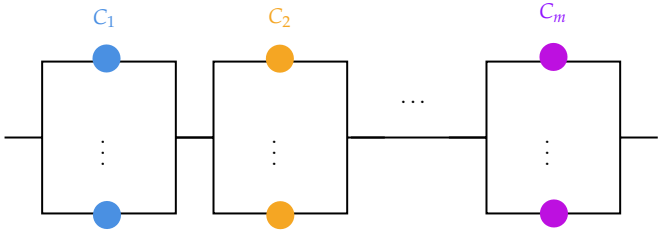
$$\begin{aligned}\phi(\boldsymbol{x}) &= \max\left\{\min_{i\in A_1} x_i, \min_{i\in A_2} x_i, \ldots, \min_{i\in A_s} x_i\right\} \\ &= \max_j \prod_{i\in A_j} x_i\end{aligned}$$

Approche par les « *minimal cut sets* »

Soit ces deux constats :

1. Un système cesse de fonctionner ssi toutes les composantes d'au moins un des « *minimal cut sets* » cessent de fonctionner.
2. Un système en série cesse de fonctionner ssi au moins une des composantes cesse de fonctionner.

Alors, tout système peut être traité comme le système en série de ses « *minimal cut sets* » :



Il s'ensuit qu'on peut réécrire le système comme :

$$\begin{aligned}\phi(\boldsymbol{x}) &= \min\left\{\max_{i\in C_1} x_i, \max_{i\in C_2} x_i, \ldots, \max_{i\in C_s} x_i\right\} \\ &= \prod_{j=1}^{m} \max_{i\in C_j} x_i\end{aligned}$$

Note Puisque l'état est une variable binaire, $x_i^k = x_i$.

Fiabilité des systèmes

Notation

X_i Variable aléatoire suivant une distribution Bernoulli : $X_i \sim \text{Bernoulli}(p_i)$.

$$X_i = \begin{cases} 1, & p_i \\ 0, & 1 - p_i \end{cases}$$

$\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ vecteur des v.a. Bernoulli.

p_i Fiabilité de la composante i.

$\boldsymbol{p} = (p_1, p_2, \ldots, p_n)$ vecteur des fiabilités.

$r(\boldsymbol{p})$ Fonction de fiabilité du système.

Fiabilité

La fiabilité d'une **composante** est la **probabilité que la composante fonctionne**.

La fiabilité d'un **système** est la **probabilité que le système fonctionne**.

Fonction de fiabilité

La ***fonction de fiabilité*** est une fonction de la fiabilité des composantes $r(\boldsymbol{p})$ qui quantifie la probabilité que le système fonctionne.

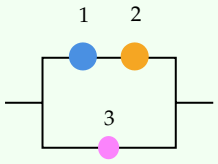
$$r(\boldsymbol{p}) = \underbrace{\Pr(\phi(\boldsymbol{X}) = 1)}_{\substack{\text{somme des probabilités} \\ \text{des « } path\ vectors \text{ »}}}$$

$$= 1 - \underbrace{\Pr(\phi(\boldsymbol{X}) = 0)}_{\substack{\text{somme des probabilités} \\ \text{des « } cut\ vectors \text{ »}}}$$

Puisque la fonction de structure ϕ est une fonction du vecteur de v.a. Bernoulli $\boldsymbol{X}$, alors ϕ est également une v.a. Bernoulli. Il s'ensuit que :

$$r(\boldsymbol{p}) = 0 \times \Pr(\phi(\boldsymbol{X}) = 0) + 1 \times \Pr(\phi(\boldsymbol{X}) = 1)$$
$$= \mathrm{E}[\phi(\boldsymbol{X})]$$

Exemple de calcul de la fonction de fiabilité



Cut vectors	Path vectors
(0, 0, 0)	(0, 0, 1)
(1, 0, 0)	(1, 1, 0)
(0, 1, 0)	(1, 0, 1)
	(0, 1, 1)
	(1, 1, 1)

On pose que les composante sont indépendantes, puis :

$$\begin{aligned} r(\boldsymbol{p}) &= \Pr(\phi(\boldsymbol{X}) = 1) \\ &= \Pr(\boldsymbol{X} = (0,\ 0,\ 1)) + \Pr(\boldsymbol{X} = (1,\ 1,\ 0)) + \Pr(\boldsymbol{X} = (1,\ 0,\ 1)) + \\ &\quad \Pr(\boldsymbol{X} = (0,\ 1,\ 1)) + \Pr(\boldsymbol{X} = (1,\ 1,\ 1)) \\ &= (1 - p_1)(1 - p_2)p_3 + p_1 p_2 (1 - p_3) + p_1 (1 - p_2) p_3 + \\ &\quad (1 - p_1) p_2 p_3 + p_1 p_2 p_3 \\ &= p_3 - p_2 p_3 - p_1 p_3 + p_1 p_2 p_3 + p_1 p_2 - p_1 p_2 p_3 + p_1 p_3 - p_1 p_2 p_3 + \\ &\quad p_2 p_3 - p_1 p_2 p_3 + p_1 p_2 p_3 \\ &= p_3 + p_1 p_2 - p_1 p_2 p_3 \end{aligned}$$

Bornes des fonctions de fiabilité

Contexte

Parfois, il n'est *pas pratique ni nécessaire* de trouver la fonction de fiabilité exacte. Plutôt, on peut l'approximer en trouvant les bornes supérieures et inférieures de la fonction avec une des deux méthodes qui suit.

Donc, l'utilité des bornes est d'approximer la probabilité sans nécessairement trouver la valeur exacte.

Méthode d'inclusion et d'exclusion

Pour comprendre d'où proviennent les formules pour les bornes, ont rappel le calcul de probabilité conjointes :

Rappel : Probabilités conjointes

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

$$\Pr\left(\bigcup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} \Pr(E_j) - \sum_{j=1}^{n}\sum_{k>j} \Pr(E_j \cap E_k) + \sum_{j=1}\sum_{k>j}\sum_{l>k} \Pr(E_j \cap E_k \cap E_l) - \cdots + (-1)^{n+1} \Pr(E_1 \cap E_2 \cap \cdots \cap E_n)$$

L'intersection de plusieurs événements correspond à la somme de sommes ci-dessus. Si l'on approxime la probabilité en tronquant l'équation à la première somme, alors on *sur*-estime la probabilité. Si on utilise seulement les deux premières sommes, alors on la *sous*-estime. Ce qu'on en déduit est que la probabilité est **contenue entre ces deux estimations** !

On peut donc établir des inégalités. Soit, pour la probabilité que le système fonctionne ($r(\boldsymbol{p})$) ou pour la probabilité qu'il ne fonctionne pas ($1 - r(\boldsymbol{p})$).

Minimal path sets On a que $\sum_{j=1}^{n} \Pr(A_j) = \sum_{j=1}^{s} \left(\prod_{i \in A_j} p_i\right)$.

Pour les « *minimal path sets* » $A_1, \ldots, A_s$, on établit :

$$r(\boldsymbol{p}) \leq \sum_{j=1}^{s} \left(\prod_{i \in A_j} p_i\right)$$

$$r(\boldsymbol{p}) \geq \sum_{j=1}^{s} \left(\prod_{i \in A_j} p_i\right) - \sum_{j=1}^{s}\sum_{k>j} \left(\prod_{i \in A_j \cup A_k} p_i\right)$$

$$\vdots$$

Exemple bornes avec minimal path sets

On reprend l'exemple de la sous-section sur les fonctions de fiabilité avec le système en parallèle ayant 3 composantes.

Ici, on pose que toutes les composantes ont une fiabilité de p, puis avec $A_1 = (0,0,1)$ et $A_2 = (1,1,0)$:

$$\sum_{j=1}^{2} \left(\prod_{i \in A_j} p_i\right) = \prod_{i \in A_1} p_i + \prod_{i \in A_2} p_i = p + p^2$$

$$\sum_{j=1}^{2}\sum_{k>j} \left(\prod_{i \in A_j \cup A_k} p_i\right) = \prod_{i \in A_1 \cup A_2} p_i = p^3$$

Donc $p + p^2 - p^3 \leq r(\boldsymbol{p}) \leq p + p^2$.

Si $p = 0.2$, $r(\boldsymbol{p}) \in [0.232, 0.24]$ mais si $p = 0.6$ alors $r(\boldsymbol{p}) \in [0.744, 0.96]$. On voit donc que plus p est petit, mieux l'intervalle approxime la fiabilité.

On peut aussi faire le calcul de façon plus intuitive en regardant les minimum path vectors dans les probabilités :

$$\begin{aligned} r(\boldsymbol{p}) &= \Pr(A_1 \cup A_2) \\ &= \Pr(A_1) + \Pr(A_2) - \Pr(A_1 \cap A_2) \\ &= \Pr(\{0,0,1\}) + \Pr(\{1,1,0\}) - \Pr(\{0,0,1\} \cap \{1,1,0\}) \\ &= \Pr(\{0,0,1\}) + \Pr(\{1,1,0\}) - \Pr(\{1,1,1\}) \\ &= p + p^2 - p^3 \end{aligned}$$

Minimal cut sets Pour les « *minimal cut sets* » $C_1, \dots, C_m$, on établit :

$$1 - r(\boldsymbol{p}) \leq \sum_{j=1}^{m} \left(\prod_{i \in C_j} (1 - p_i) \right)$$

$$1 - r(\boldsymbol{p}) \geq \sum_{j=1}^{m} \left(\prod_{i \in C_j} (1 - p_i) \right) - \sum_{j=1}^{m} \sum_{k>j} \left(\prod_{i \in C_j \cup C_k} (1 - p_i) \right)$$

$$\vdots$$

Exemple bornes avec minimal cut sets

On reprend l'exemple de la sous-section sur les fonctions de fiabilité avec le système en parallèle ayant 3 composantes.

Ici, on pose que toutes les composantes ont une fiabilité de p, puis avec $C_1 = (1,0,0)$ et $C_2 = (0,1,0)$:

$$\sum_{j=1}^{m} \left(\prod_{i \in C_j} (1 - p_i) \right) = \prod_{i \in C_1} (1 - p_i) + \prod_{i \in C_2} (1 - p_i) = (1-p)^2 + (1-p)^2 = 2(1-p)^2$$

$$\sum_{j=1}^{m} \sum_{k>j} \left(\prod_{i \in C_j \cup C_k} (1 - p_i) \right) = \prod_{i \in C_1 \cup C_2} (1 - p_i) = (1-p)^3$$

Donc $2(1-p)^2 - (1-p)^3 \leq r(\boldsymbol{p}) \leq 2(1-p)^2$.

Si $p = 0.2$, $1 - r(\boldsymbol{p}) \in [0.768, 1.28]$ mais si $p = 0.6$ alors $1 - r(\boldsymbol{p}) \in [0.256, 0.32]$. On voit donc que plus p est large, mieux l'intervalle approxime la fiabilité.

C'est donc l'inverse que l'approche par « *minimal path sets* ».

Méthode d'intersection

Contexte

Au lieu d'utiliser les probabilités d'union des événements, on utilise les probabilités d'intersection des événements.

Sous la **méthode d'intersection**,

$$\overbrace{\prod_{j=1}^{m} \left[1 - \overbrace{\prod_{i \in C_j} (1 - p_i)}^{\substack{\text{probabilité que toutes les composantes du } C_j \\ \text{échouent}}} \right]}^{\substack{\text{probabilité qu'au moins une des composantes du } C_j \\ \text{fonctionne} \\ \text{probabilité qu'au moins une composante de chacun des « } \textit{minimal cut sets} \text{ » fonctionne}}} \leq r(\boldsymbol{p}) \leq \overbrace{1 - \prod_{j=1}^{s} \left[1 - \overbrace{\prod_{i \in A_j} p_i}^{\substack{\text{probabilité que toutes les composantes du } A_j \\ \text{fonctionnent}}} \right]}^{\substack{\text{probabilité qu'au moins une des composantes du } A_j \\ \text{échoue} \\ \text{probabilité que toutes les composantes d'au moins un des « } \textit{minimal path sets} \text{ » fonctionnent}}}$$

Exemple bornes avec la méthode d'intersection

On reprend l'exemple de la sous-section sur les fonctions de fiabilité avec le système en parallèle ayant 3 composantes.

Ici, on pose que toutes les composantes ont une fiabilité de p, puis avec $C_1 = (1,0,0)$ et $C_2 = (0,1,0)$:

$$\prod_{j=1}^{m} \left[1 - \prod_{i \in C_j} (1 - p_i) \right] = \left(1 - (1-p)^2 \right) \left(1 - (1-p)^2 \right) = \left(1 - (1-p)^2 \right)^2$$

Avec $A_1 = (0,0,1)$ et $A_2 = (1,1,0)$,

$$1 - \prod_{j=1}^{s} \left[1 - \prod_{i \in A_j} p_i \right] = 1 - (1-p) \left(1 - p^2 \right)$$

Donc $\left(1 - (1-p)^2 \right)^2 \leq r(\boldsymbol{p}) \leq 1 - (1-p)\left(1-p^2\right)$.

Si $p = 0.2$, $r(\boldsymbol{p}) \in [0.1296, 0.232]$ et si $p = 0.6$ alors $1 - r(\boldsymbol{p}) \in [0.7056, 0.744]$. On voit donc que peut importe la valeur de p, l'intervalle approxime bien la fiabilité.

On résume l'efficacité des différents méthodes ci-dessous :

Approche	avec un petit p	avec un gros p
	Intervalle	
« *minimal path sets* »	large	**étroit**
« *minimal cut sets* »	**étroit**	large
intersection	**étroit**	**étroit**

Graphiques aléatoires

Graphique

Ensemble de nœuds connectés par des arcs.

Composantes des graphiques

N Ensemble des nœuds.

A Ensemble des arcs connectant les nœuds.

Le nombre d'arcs est d'au plus $\binom{n}{2}$.

C'est-à-dire, le nombre possibles de groupes de deux nœuds.

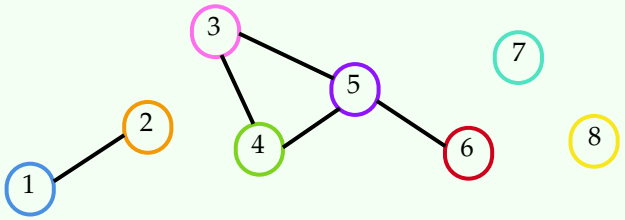
Contexte

Un graphique peut également être décomposé en sous-graphiques qu'on nomme ses ***composantes***. Les composantes ne se chevauchent pas et sont composées de nœuds connectés.

On dit qu'un graphique est ***connecté*** s'il a une seule *composante*. C'est-à-dire, si l'on peut passer d'un nœud à tout autre nœud du graphique via les arcs.

Exemple de graphique

Soit le graphique suivant :



On trouve que :

8 nœuds : $N = \{1,2,3,4,5,6,7,8\}$.

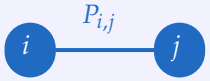
5 arcs : $A = \{\{1,2\},\{3,4\},\{3,5\},\{4,5\},\{5,6\}\}$.

4 composantes : $\{\{1,2\},\{3,4,5\},\{7\},\{8\}\}$.

Également, puisqu'il y a plusieurs composantes, le graphique n'est pas connecté.

Graphique aléatoire

Graphique avec n nœuds pour lequel deux composantes i et j ne sont pas reliées avec certitude, mais plutôt avec probabilité $P_{i,j}$:



Soit la v.a. $X_{i,j}$ représentant l'existence d'un arc entre les nœuds i et j avec probabilité $\Pr(X_{i,j} = 1) = P_{i,j}$ alors :

$$X_{i,j} = \begin{cases} 1, & \text{si } \{i,j\} \text{ est un arc} \\ 0, & \text{sinon} \end{cases}$$

Connectivité des graphiques aléatoires

Contexte

La connectivité des graphiques aléatoires est semblable à la fiabilité des systèmes.

Pour un système, il n'est pas nécessaire que toutes les composantes fonctionnent pour que le système fonctionne. De façon semblable, il n'est pas nécessaire que tous les nœuds d'un graphique aléatoire soient reliés pour qu'il soit connecté.

Alors, on peut appliquer les mêmes concepts de « *minimal path sets* » et de « *minimal cut sets* » des systèmes aux graphiques aléatoires.

Un graphique *aléatoire* est connecté tant que tous les arcs d'au moins un « *minimal path sets* » existent.

Un graphique aléatoire de n nœuds a :

n^{n-2} « *minimal path sets* », et

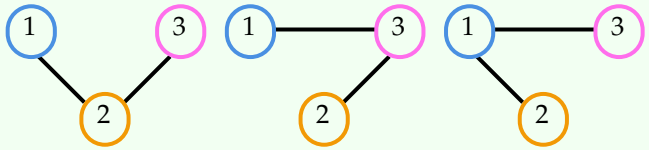
$2^{n-1}-1$ « *minimal cut sets* »,

$2^{\binom{n}{2}}$ graphiques possibles.

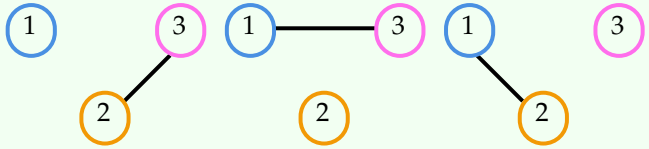
Exemple de connectivité

Soit un graphique aléatoire avec 3 nœuds.

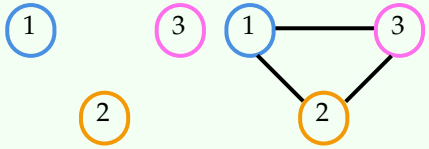
Les $3^{3-2} = 3$ « *minimal path sets* » sont les suivants :



Les $2^{3-1} - 1 = 3$ « *minimal cut sets* » sont les suivants :



Les deux autres graphiques possibles qui ne sont pas optimaux sont :



Probabilités de connectivité des graphiques

On pose que les v.a. $X_{i,j}$ sont iid avec $P_{i,j} = p$. Puis, on trouve la probabilité P_n qu'un graphique aléatoire de n nœuds soit connecté avec la formule récursive :

$$P_n = 1 - \sum_{k=1}^{n-1} \binom{n-1}{k-1} (1-p)^{k(n-k)} P_k, \quad n = 2, 3, \ldots$$

où $P_1 = 1, P_2 = p$.

On peut également trouver les **bornes** pour la probabilité afin de simplifier la tâche :

$$n(1-p)^{n-1} - \binom{n}{2}(1-p)^{2n-3} \leq 1 - P_n \leq (n+1)(1-p)^{n-1}$$

Finalement, on peut **approximer** la probabilité avec $P_n \approx 1 - n(1-p)^{n-1}$.

Durée de vie des systèmes

Contexte

Nous avons évalué la ***fiabilité*** d'un système et comment qu'elle est impactée par la fiabilité de ses composantes. Nous évaluons maintenant la **durée de vie** d'un système et comment qu'elle est impactée par la durée de vie de ses composantes.

Notation

T_i Durée de vie de la composante i.

$S_i(t)$ Fonction de survie de la durée de vie de la composante i.

$\boldsymbol{S}(t) = (S_1(t), \ldots, S_n(t))$ est le vecteur des fonctions de survie des n composantes.

T Durée de vie du système.

Calcul de probabilités de durée de vie

La probabilité que le système fonctionne passé t équivaut à la fonction de fiabilité évaluée au vecteur des fonctions de survie : $\Pr(T > t) = r[\boldsymbol{S}(t)]$.

Donc, on pose $p_i = S_i(t)$ pour $i = 1, 2, \ldots, n$.

Espérance de durée de vie

La durée de vie espérée équivaut à $\mathrm{E}[T] = \int_0^\infty r[\boldsymbol{S}(t)]dt$.

Exemple du calcul de la durée de vie espérée

On reprend l'exemple de la sous-section sur les fonctions de fiabilité avec le système en parallèle ayant 3 composantes.

On pose que les 3 composantes sont indépendantes et que la durée de vie est uniformément distribuée sur $(0, 2)$.

1 Trouver la fonction de survie de la composante i :
$$S_i(t) = \frac{2-t}{2-0} = \frac{2-t}{2}$$

2 Trouver la fonction de fiabilité.

Précédemment, nous avons trouvé que $r(\boldsymbol{p}) = p_3 + p_1 p_2 - p_1 p_2 p_3$.

3 Remplacer $\boldsymbol{p}$ par $\boldsymbol{S}(t)$:
$$\begin{aligned} r(\boldsymbol{p}) &= S_3(t) + S_1(t)S_2(t) - S_1(t)S_2(t)S_3(t) \\ &= \left(\frac{2-t}{2}\right) + \left(\frac{2-t}{2}\right)^2 - \left(\frac{2-t}{2}\right)^3 \\ &= \frac{t^3 - 4t^2 + 8}{8} \end{aligned}$$

4 Trouver $\mathrm{E}[T]$:
$$\begin{aligned} \mathrm{E}[T] &= \int_0^2 \frac{t^3 - 4t^2 + 8}{8} dt \\ &= 1.1667 \end{aligned}$$

Étapes du calcul de probabilités, ou de l'espérance, de la durée de vie

1. Déterminer la fonction de la structure du système $\phi(\boldsymbol{X})$.
   Soit avec les « *minimal path sets* » ou les « *minimal cut sets* ».
2. Déduire la fonction de fiabilité.
   Soit en trouvant $r(\boldsymbol{p}) = \mathrm{E}[\phi(\boldsymbol{X})]$, ou avec $r(\boldsymbol{p}) = Pr(\phi(\boldsymbol{X}) = 1)$.
3. Développer la fonction de survie $\Pr(T > t)$ de la fonction de fiabilité $r(\boldsymbol{S}(t))$.
4. Trouver la probabilité désirée ou l'espérance.

Raccourci Pour un système de k parmi n avec des durées de vie iid suivant une loi exponentielle de moyenne μ, $\mathrm{E}[T] = \mu \sum_{i=k}^{n} \frac{1}{i}$. Cette formule découle du coût espéré total pour les algorithmes « *greedy* » A et B.

Divers

Rappel : fonction de hasard

Dans le chapitre de *Erreur* à la sous-section *Fonctions de variables aléatoires* on a :

La fonction de hasard $h_X(x) = \frac{f(x)}{S(x)}$

La fonction de hasard cumulative $H_X(x) = \int_{-\infty}^{x} h(t)dt$

Système monotone

La fiabilité du système augmente lorsque la fiabilité de toute composante augmente.

Terminologie

IFR « *Increasing failure rate distribution* ».

DFR « *Decreasing failure rate distribution* ».

IFRA « *Increasing failure rate on the average distribution* ».

La distribution IFRA est une généralisation de la distribution IFR.

Il s'ensuit que si une distribution est IFR elle est également IFRA.

Distribution	$h(x)$ est une fonction _____ de x
IFR	croissante
DFR	décroissante
IFR et DFR	constante

Une distribution est IFRA si $\frac{H(x)}{x}$ est une fonction *croissante* de x, pour tout $x \geq 0$.

Note Si les distribution de durées de vies de toutes les composantes (*indépendantes*) d'un *système monotone* sont IFRA, alors la distribution de la durée de vie du système le sera aussi.

Distributions particulières

Puisque la fonction de hasard de la distribution exponentielle est fixe, elle est à la fois IFR et DFR. Cependant, lorsque la fonction de hasard varie, le type de distribution peut varier aussi. Par exemple, pour la loi gamma et la loi de Weibull :

Distribution	Weibull(τ, θ)	Gamma(α, β)
	Condition	
IFR	$\tau \geq 1$	$\alpha \geq 1$
DFR	$0 < \tau \leq 1$	$0 < \alpha \leq 1$
IFR et DFR	$\tau = 1$	$\alpha = 1$

Note Une loi gamma avec $\alpha = 1$, tout comme une loi de Weibull avec $\tau = 1$, revient à une distribution exponentielle.

Note Voir la sous-section *Distributions* du chapitre de *Erreur* pour une description de la loi gamma et de la loi de Weibull.

Assurance vie

Probabilités

Notation

ℓ_a Nombre d'individus initial dans une cohorte où $a = 0$ habituellement.

ℓ_{x+a} Nombre d'individus de la cohorte ayant survécu x années de a (donc âgés de $x + a$ années).

${}_td_x$ Nombre de décès entre les âges x et $x + t$.

$${}_td_x = l_x - l_{x+t}.$$

Probabilité de survie

La probabilité qu'un assuré de x ans survie au moins t années est $${}_tp_x = \frac{l_{x+t}}{l_x}.$$

Probabilité de décès

La probabilité qu'un assuré de x ans décède d'ici t années est $${}_tq_x = \frac{l_x - l_{x+t}}{l_x}.$$

Variable aléatoire du nombre de décédés entre les âges x et $x + t$ ${}_t\mathcal{D}_x$

On a que $${}_t\mathcal{D}_x \sim \text{Bin}(\ell_x, {}_tq_x).$$

Il s'ensuit que $\text{E}[{}_t\mathcal{D}_x] = {}_td_x$.

Également, ${}_t\mathcal{D}_x = \mathscr{L}_x - \mathscr{L}_{x+t}$.

Espérances de vie

Espérance de vie **abrégée** pour un individu d'âge x

$$e_x = \sum_{k=0}^{\omega-x-1} k \; {}_{k|}q_x$$

puis, si $\lim_{k\to\infty}(k+1)_{k+1}p_x = 0$,

$$= \sum_{k=1}^{\omega-x} {}_kp_x$$

En anglais, « ***curtate life expectancy*** ».

Espérance de vie **complète** pour un individu d'âge x

$$\mathring{e}_x = \int_0^{\omega-x} {}_tp_x\mu_{x+t}dt$$

$$= \int_0^{\omega-x} {}_tp_x dt \qquad \text{si } \lim_{t\to\infty} {}_tp_x = 0$$

Sous l'hypothèse d'une distribution uniforme des décès (DUD), $$\mathring{e}_x \overset{DUD}{=} e_x + \frac{1}{2}.$$

En anglais, « ***complete expectation of life*** ».

Contrats d'assurance vie

Notation

Z_x Variable aléatoire du contrat d'assurance pour un assuré d'âge x.

Y_x Variable aléatoire de la rente pour un rentier d'âge x.

Valeur présente actuarielle

On nomme l'actualisation de paiements conditionnels à la mortalité la ***valeur présente actuarielle (VPA)***.

Pour des contrats d'assurance, on la dénote par A_x et pour des contrats de rentes, par a_x.

En anglais, « *Actuarial Present Value (APV)* »

Assurance-vie entière Z_x

Est en vigueur tant que l'assuré est en vie et verse une prestation à la fin moment de l'année de son décès.

$$\begin{aligned} A_x &= \sum_{k=0}^{\omega-x-1} v^{k+1}{}_k p_x q_{x+k} \\ &= vq_x + v^2 p_x q_{x+1} + v^3 {}_2p_x q_{x+2} + \dots \end{aligned}$$

Capital différé de t années ${}_tE_x$

Si l'assuré ***ne décède pas*** dans les t années suivant l'émission du contrat, le capital différé ${}_tE_x$ paye une prestation de survie.

$${}_tE_x = v^t {}_tp_x$$

Alias, le **facteur d'actualisation actuariel**.

En anglais, « *mortality discount factor* ».

Assurance différée de m années ${}_{m|}Z_x$

Si l'assuré décède **après** les m années suivant l'émission du contrat, paye une prestation de décès.

$${}_{m|}A_x = \sum_{k=m}^{\omega-x-1} v^{k+1}{}_k p_x q_{x+k}$$

Assurance-vie temporaire $Z^1_{x:\overline{n}|}$

Si l'assuré décède dans les n années suivant l'émission du contrat, paye une prestation de décès.

$$A^1_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1}{}_k p_x q_{x+k}$$

Assurance mixte $Z_{x:\overline{n}|}$

Si l'assuré ***décède*** dans les n années suivant l'émission du contrat, paye une prestation de décès. S'il est toujours en vie, paye une prestation de survie.

$$A_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1}{}_k p_x q_{x+k} + {}_nE_x$$

En anglais, « *endowment insurance* ».

Note Si le contrat d'assurance est à double, ou j, force on remplace le facteur d'actualisation v par v^j.

Relations entre les contrats d'assurance

Assurance :

vie $A_x = vq_x + vp_x A_{x+1}$.

différée ${}_{m|}A_x = {}_mE_x A_{x+m}$.

temporaire $A^1_{x:\overline{n}|} = A_x - {}_{n|}A_x$.

mixte $A_{x:\overline{n}|} = A^1_{x:\overline{n}|} - {}_nE_x$.

Contrats de rentes

Rentes de base

Rente viagère de début de période $\ddot{Y}_x$

Pour $K = 0, 1, 2, \ldots$ on obtient que $\ddot{Y}_x = \ddot{a}_{\overline{K+1}|}$. Puis, $\mathrm{E}[\ddot{Y}_x] = \ddot{a}_x$.

$$\begin{aligned}\ddot{a}_x &= \sum_{k=0}^{\omega-x-1} v^k \, {}_kp_x \\ &= 1 + vp_x + v^2 {}_2p_x + \ldots \\ &= \frac{1 - A_x}{d}\end{aligned}$$

Relations

Rente
viagère $\ddot{a}_x = 1 + vp_x \ddot{a}_{x+1}$.

Vies conjointes

Rente vie entière du premier décès

La rente $\ddot{a}_{xy}$ effectue des paiements jusqu'au premier décès du couple (x, y).

En anglais, « *joint life annuity* ».

Rente vie entière du dernier survivant

La rente $\ddot{a}_{\overline{xy}}$ effectue des paiements jusqu'au dernier décès du couple (x, y).

En anglais, « *last survivor annuity* ».

si le premier décès est	alors
x	$\ddot{a}_{xy} = \ddot{a}_x$ et $\ddot{a}_{\overline{xy}} = \ddot{a}_y$
y	$\ddot{a}_{xy} = \ddot{a}_y$ et $\ddot{a}_{\overline{xy}} = \ddot{a}_x$

Il s'ensuit que $\ddot{a}_x + \ddot{a}_y = \ddot{a}_{xy} + \ddot{a}_{\overline{xy}}$.

Principe d'équivalence

Principe d'équivalence

Pose égale la VPA des primes aux prestations pour que les assurés reçoivent une couverture « équitable ». Du point de vue d'une compagnie d'assurance, on devrait aussi tenir en compte les dépenses et le profit pour la tarification.

Pour l'examen cependant, on les ignores et se restreint aux prestations et aux primes pour trouver que la prime nette est la prime telle que $VPA_{\text{primes}} = VPA_{\text{prestations}}$.

Assurance nivelée

Contexte

Typiquement, la mortalité n'est pas constante. En assurance vie, elle est moins élevée lorsqu'un assuré est jeune et augmente avec l'âge. En assurance dommages cependant, elle est plus élevée lorsqu'un assuré est jeune que lorsqu'il est âgé.

Charger une prime fixe dans le premier cas implique que l'assuré paye trop au début mais pas assez à la fin du contrat d'assurance. Dans le deuxième cas, il ne paye pas assez au début et trop à la fin. Si la prime est fixe, on peut équilibrer les paiements sur la durée de vie de l'assuré pour que ce soit équitable.

Cependant, si le détenteur de police « *lapses* » ou ne renouvelle pas sa police, alors les prestations reçues ne seront pas égales aux primes payées. Ceci est pourquoi **les assureurs chargent rarement des primes fixes lorsque la mortalité n'est pas constante**.

Simulation

On simule des réalisations de variables aléatoires à partir de nombres aléatoires distribués uniformément dans $[0,1)$.

Générer des nombres pseudo-aléatoires

On génère des nombres pseudo-aléatoires qui *simulent* des nombres réellement aléatoires.

1. Choisir l'ancrage : un nombre initial x_0.

   En anglais, « *seed* ».

2. Générer les nombres pseudo-aléatoires avec $x_{j+1} = (ax_j + c) \mod m$, $j \geq 0$.

   Les valeurs a, c, m sont spécifiées en avance pour ***imiter*** une simulation aléatoire.

   L'opérateur modulo revient à prendre le restant d'une division comme un nombre entier.

   Le nombre n'est pas fractionnaire, plutôt $x_{j+1} \in [0, m)$.

3. Calculer la réalisation $u_{j+1} = x_{j+1}/m$.

4. Répéter les étapes 1 à 3 le nombre de fois désiré.

Note Il est rare de devoir nous même simuler les nombres, habituellement ils sont donnés. Cependant, si c'est le cas, les nombres a, c, m seront donnés dans la question.

Méthode de l'inverse

Simulation par la méthode de l'inverse

Pour une variable aléatoire X avec fonction de répartition $F_X(x)$,

1. Simuler une réalisation u_j de la v.a. $U(0,1)$.
2. Poser $x_j = F_X^{-1}(u_j)$.
3. Répéter les étapes 1 et 2 le nombre de fois désiré.

Méthode d'acceptation-rejet

Contexte

Lorsqu'il est difficile, ou impossible, de trouver la fonction quantile on peut utiliser la méthode d'acceptation de rejet.

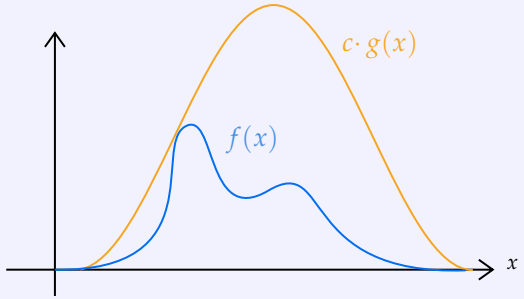
Supposons que nous pouvons simuler des réalisations d'une distribution ayant la fonction de densité g et que l'on veut simuler des réalisations d'une autre distribution ayant la fonction de densité f. Par exemple :

Simulation par la méthode d'acceptation-rejet

Pour une variable aléatoire X,

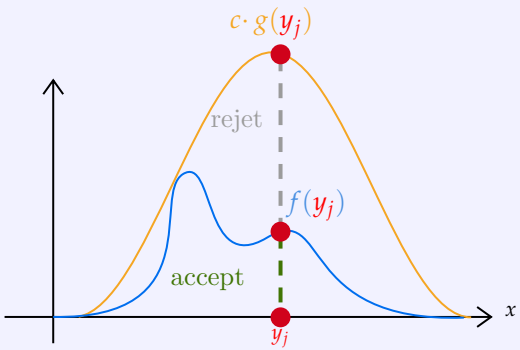
1. Trouver une constante c telle que $\frac{f(x)}{g(x)} \leq c$, $\forall x$.

   Par exemple,

   Pour ce faire : trouver la dérivée de $f(x)/g(x)$; la poser égale à zéro ; choisir la valeur critique x^* qui maximise la fonction $f(x)/g(x)$; poser que $c = f(x^*)/g(x^*)$.

2. Simuler une réalisation y_j de la variable aléatoire Y ayant la fonction de densité g et calculer $\frac{f(y_j)}{cg(y_j)}$.
3. Simuler une réalisation u_j de la variable aléatoire $U(0,1)$.
4. Comparer la réalisation u_j à $\frac{f(y_j)}{cg(y_j)}$, si $u_j \leq \frac{f(y_j)}{cg(y_j)}$ alors accepter la réalisation y_j, sinon la refuser et retourner à l'étape 2.



Les nombres simulés vont suivre la distribution associée à la fonction de densité f.

Note Le nombre d'itérations nécessaires pour obtenir un nombre aléatoire simulé suit une distribution géométrique de moyenne c.

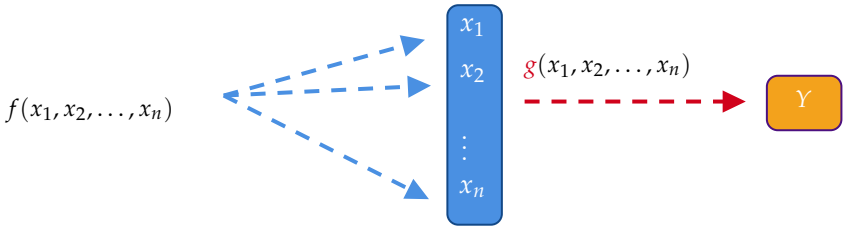
Simulation Monte-Carlo

Simulation Monte-Carlo

Pour une variable aléatoire X,

1. Simuler un vecteur de réalisation $(x_1, x_2, \ldots, x_n)$ d'une distribution dont la fonction de densité est $f(x_1, x_2, \ldots, x_n)$.
2. Appliquer une fonction g au vecteur des réalisations pour trouver $y_j = g(x_1, x_2, \ldots, x_n)$.
3. Répéter les étapes 1 et 2 r fois où r est grand.
4. Calculer la valeur désiré (espérance, variance, etc.) avec les réalisations $(y_1, y_2, \ldots, y_r)$.

Visuellement :



V

Processus stochastiques

Introduction

Notation

X_n État du processus au temps n.

Par exemple, si $X_n = i$ alors le processus est dit d'être dans l'état i au temps n.

Processus stochastique

Soit le processus stochastique $\{X_n, n = 0, 1, 2, \ldots\}$.

Processus de Poisson

Notation

$\lambda(t)$ Fonction d'intensité d'un processus de Poisson.

En anglais, « *rate function* ».

Processus stochastique

Une collection de variables aléatoires.

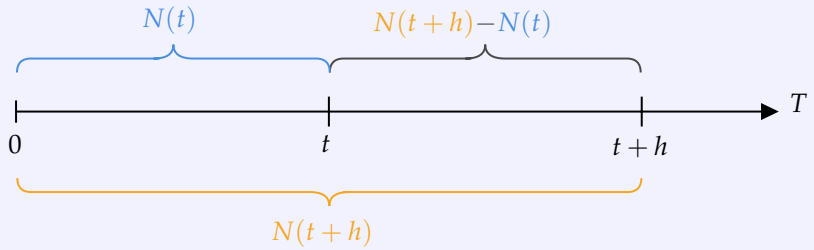
Processus de comptage

On dénote le processus de comptage, ou processus de dénombrement, par $\underline{N} = \{N(t), t \geq 0\}$. Le processus **compte le nombre d'événements** qui se produisent dans l'intervalle de temps $(0, t]$ où $t > 0$.

En termes mathématiques, c'est un processus stochastique dont les variables aléatoires prennent des valeurs non décroissantes et non négatives sous les conditions suivantes :

1. $N(0) = 0$;
2. $N(t) \geq 0$ *(valeurs non négatives)* ;
3. $N(t)$ est entier ;
4. $N(t+h) \geq N(h)$ pour $h > 0$ *(valeurs non décroissantes)*.

Visuellement, on voit que l'**accroissement** $N(t+h) - N(t)$ représente le nombre d'événements produits sur l'intervalle $(t, t+h]$:



Processus de Poisson

Processus de comptage dont :

1. chaque accroissement est distribué selon la loi de Poisson,
2. les accroissements qui ne se **chevauchent *pas*** sont indépendants.

Pour un processus de Poisson avec **fonction d'intensité** $\lambda(t)$, l'accroissement $N(t+h)-N(t) \sim \text{Poisson}\left(\lambda = \int_t^{t+h} \lambda(u)du\right)$.

On pose donc que le paramètre de la fréquence des accroissements λ est la *moyenne* de la fonction d'intensité des accroissements $\lambda(t)$ sur l'intervalle de temps $(t, t+h]$.

Processus de Poisson homogène

Si la fonction d'intensité est constante, $\lambda(t) = \lambda$, le processus $\underline{N}$ est un **processus de Poisson homogène** et $N(t+h)-N(t) \sim \text{Poisson}(\lambda h)$.

Processus de Poisson non homogène

Si la fonction d'intensité varie avec le temps t, le processus $\underline{N}$ est un **processus de Poisson non homogène**.

Temps d'occurrence

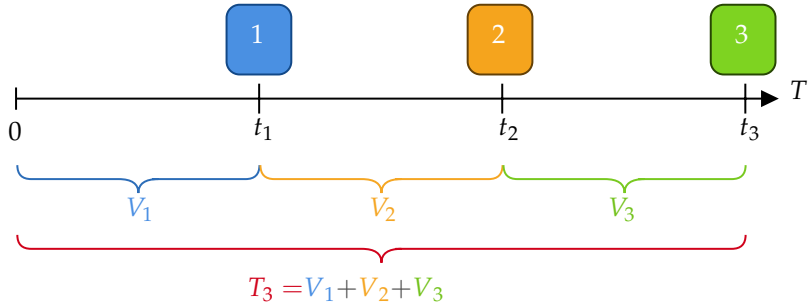
Notation

T_k Temps d'occurrence du k^e événement.
$T_k = V_1 + V_2 + \cdots + V_k$.

V_k Intervalle de temps entre la réalisation du $(k-1)^e$ et du k^e événement.
Alias, le temps inter arrivé.
$V_k = T_k - T_{k-1}$.

On pose que $T_0 = 0$, $V_0 = 0$ et que $V_1 = T_1$.

Visuellement :



Temps d'occurrence

On peut définir le processus de comptage en fonction du temps d'occurrence des événements au lieu nombre de sinistres : $N(t) = \sup\{k \geq 1 : T_k \leq t\}$, $\forall t \geq 0$.

On trouve que $\Pr(T_k > s) = \Pr(N(s) < k)$. C'est-à-dire, $\Pr\left(\begin{smallmatrix}\text{le } k^e \text{ événement se produise} \\ \text{après le temps } s\end{smallmatrix}\right) = \Pr\left(\begin{smallmatrix}\text{moins de } k \text{ événements se} \\ \text{produisent d'ici le temps } s\end{smallmatrix}\right)$.

Temps d'occurrence pour des processus de Poisson homogènes

Si $N(t) \sim \text{Poisson}(\lambda t)$ alors $V_k \sim \text{Exp}\left(\theta = \frac{1}{\lambda}\right)$ et $T_k \sim \text{Gamma}\left(\alpha = k, \theta = \frac{1}{\lambda}\right) \sim \text{Erlang}\left(n = k, \lambda\right)$.

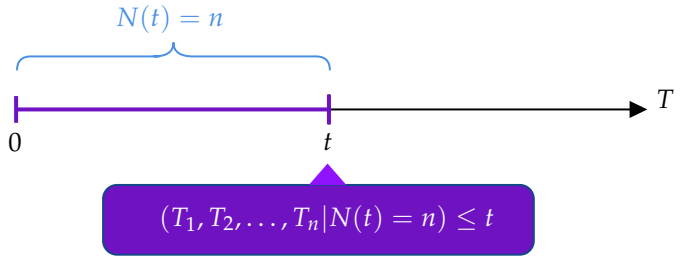
Note La loi Gamma avec un paramètre de forme α entier correspond à la loi Erlang. L'avantage de la loi Erlang est qu'elle a une fonction de répartition explicite qui découle de la relation entre les processus de Poisson et les temps d'occurrences. Voir la sous-section sur les **Distributions** du chapitre de Erreur.

Temps d'occurrence conditionnels

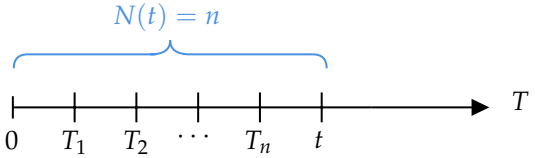
Note Voir la sous-section des **Statistiques d'ordre** du chapitre de Table des matières.

Lorsque nous savons qu'un certain nombre d'événements se produit d'ici un temps t, les temps d'occurrences $T_1, T_2, \ldots, T_n$ ne **suivent plus une distribution Gamma**. Ceci est puisque **leurs domaines sont *bornés* à** t au lieu d'être ***infinis***.

Par exemple, $N(t) = n$ implique que $T_1, T_2, \ldots, T_n \leq t$:



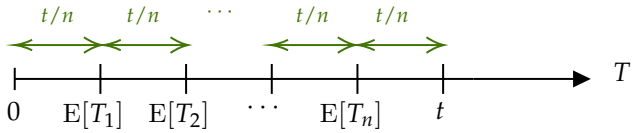
On en déduit que les temps d'occurrences sont en fait des **Statistiques d'ordre** avec $0 < T_1 \leq T_2 \leq \cdots \leq T_n \leq t$:



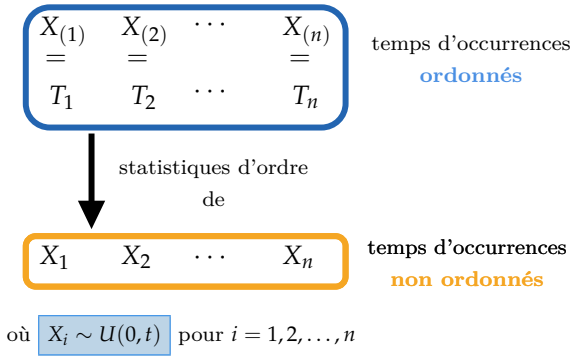
Pour déterminer la distribution de T_i, $i = 1, 2, \ldots, n$, on rappel ces deux propriétés des processus de Poisson homogènes :

1. Les intervalles qui ne se chevauchent pas sont indépendants.
2. Le paramètre de fréquence λ est proportionnel à la longueur d'un intervalle, ce qui implique qu'il est identique pour des intervalles de la même longueur.

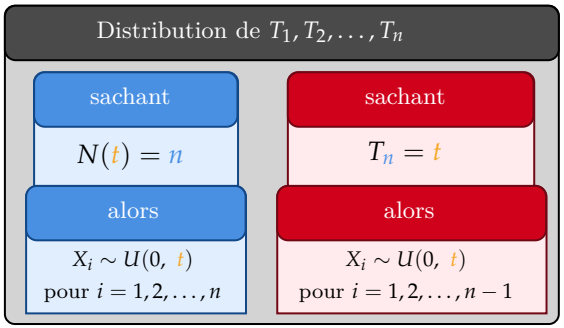
On en déduit que les temps d'occurrences des événements devraient être **uniformément distribués** en $n+1$ sous-intervalles :



Donc, $T_1, T_2, \ldots, T_n$ sont les statistiques d'ordre d'une distribution $U(0,t)$:



En bref :



Également, lorsque $X_k \sim U(a,b)$ pour $k = 1, 2, \ldots, n$, on trouve que $\mathrm{E}[X_{(k)}] = \mathrm{E}[T_k] = a + \frac{k(b-a)}{n+1}$.

Note Voir les *Cas spéciaux* des *Statistiques d'ordre* pour la définition de cette espérance.

Exemple

Des autobus arrivent à un arrêt d'autobus selon une distribution de Poisson avec un paramètre de fréquence de $\lambda = 4$ par heure. Les autobus commencent à arriver dès 8h du matin.

On sait qu'aujourd'hui, trois autobus sont passés entre 8h et 9h du matin.

Calculer :

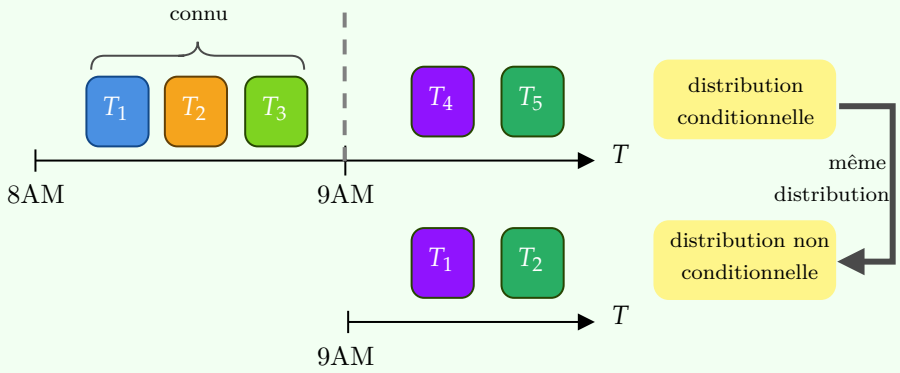
1. L'espérance du temps d'arrivé du 5^e bus,
2. L'espérance du temps d'arrivé du 2^e bus,
3. La probabilité que seulement un bus soit passé entre 8h et 8h30 du matin.

Premièrement, l'espérance du temps d'arrivé du 5^e bus :

1 On connaît l'intervalle de temps durant laquelle les 3 premiers autobus arrivent.
Ceci implique que le 5^e autobus peut arriver à tout moment passé 9AM—alias, T_5 est n'a **pas encore eu lieu** et **n'est *pas* borné**.

2 On peut donc récrire l'espérance conditionnelle :
$\mathrm{E}[T_5|N(8,9] = 3] = \mathrm{E}[T_2]$

Visuellement, on peut voir pourquoi ces deux écritures sont équivalentes :



2 Puisque T_5 n'est pas borné, il suit une distribution Gamma$(2, 1/4)$. Donc, $\mathrm{E}[T_2] = \frac{2}{4} = 0.50$ ce qui équivaut à 9h30AM.

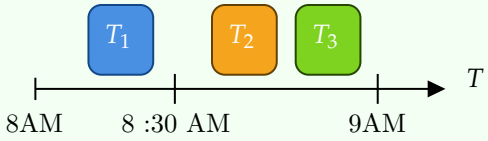
Deuxièmement, l'espérance du temps d'arrivé du 2^e bus :

1 On connaît l'intervalle de temps durant laquelle les 3 premiers autobus arrivent.
Ceci implique que le temps d'arrivé du 2^e doit être à, ou avant, 9AM—alias, T_2 a **eu lieu** et ***est* borné**.

2 Il s'ensuit que T_2 ne suit pas une distribution Gamma et que l'espérance conditionnelle de son temps d'arrivé, T_2, équivaut à l'espérance de la 2^e statistique d'ordre, $X_{(2)}$, des temps d'arrivés non ordonnés X_k distribués **uniformément** entre 8AM et 9AM $(U(8,9))$ pour $k = 1, 2, 3$:
$\mathrm{E}[T_2|N(8,9] = 3] = \mathrm{E}[X_{(2)}] = 8 + \dfrac{2 \times (9-8)}{3+1} = 8.5$
qui équivaut à 8h30AM.

Dernièrement, la probabilité que seulement un bus soit passé entre 8h et 8h30 du matin.

1 On observe la probabilité qu'on désire calculer :



2 Le « *twist* » pour calculer la probabilité est de la voir comme une binomiale.

3 D'abord, puisque $X_k \sim \mathrm{U}(8,9)$ alors la probabilité que n'importe lequel des autobus arrive dans la première demi-heure est $\Pr(X_k \leq 0.5) = \frac{1}{9-8+1} = 0.50$ pour $k = 1, 2, 3$.

4 Puis, on défini un « succès » comme « un autobus qui arrive dans la première demi-heure » ce qui implique que $\Pr(\text{succès}) = \Pr(X_k \leq 0.50) = 0.50$.

5 Finalement, $\Pr(N(8,8.5] = 1|N(8,9] = 3) = \Pr\left(\begin{smallmatrix}\text{un autobus arrive}\\ \text{entre 8h00 et 8h30}\end{smallmatrix} \cap \begin{smallmatrix}2 \text{ autobus arrivent}\\ \text{entre 8h30 et 9h00}\end{smallmatrix}\right) = \Pr(1 \text{ succès}) = \binom{3}{1} 0.5^1 (1-0.5)^2 = 0.375$

Propriétés des processus de Poisson

Décomposition de processus de Poisson

Décomposition de processus de Poisson (« *Thinning* »)

Si un processus de Poisson peut être décomposé en plusieurs sous-processus distincts, alors ces sous-processus distincts sont également des processus de Poisson avec une fonction d'intensité proportionnelle. Ce processus de décomposition s'appelle le « *thinning* ».

Soit :

le processus de Poisson N avec fonction d'intensité $\lambda(t)$,

les sous-processus distincts $N_1, N_2, \ldots, N_n$ de N dont les proportions sont $\pi_1, \pi_2, \ldots, \pi_n$.

Alors, $N_1, N_2, \ldots, N_n$ sont des processus de Poisson indépendants avec paramètre de fréquence $\pi_1\lambda(t), \pi_2\lambda(t), \ldots, \pi_n\lambda(t)$.

Si le processus N est homogène et que les **proportions** π_i sont **constantes**, pour $i = 1, 2, \ldots, n$, alors les sous-processus sont **homogènes**. Cependant, si les **proportions** ne sont ***pas*** **constantes** alors les sous-processus ne sont ***pas*** **homogènes**.

Superposition

Somme de processus de Poisson (« *Superposition* »)

La somme de plusieurs processus de Poisson s'appelle la « *superposition* ». Si les processus de Poisson sont indépendants, leur somme est également un processus de Poisson.

Soit :

les processus de Poisson indépendants $N_1, N_2, \ldots, N_n$ avec paramètres de fréquence $\lambda_1(t), \lambda_2(t), \ldots, \lambda_n(t)$.

Alors, $N_1 + N_2 + \cdots + N_n$ est un processus de Poisson avec paramètre de fréquence $\lambda = \lambda_1(t) + \lambda_2(t) + \cdots + \lambda_n(t)$.

Probabilités conjointes

Notation

N_1, N_2 Processus de Poisson indépendants avec paramètres de fréquence λ_1, λ_2.

$T_{1,n}$ Le temps jusqu'au n^e événement de N_1.

$T_{2,m}$ Le temps jusqu'au m^e événement de N_2.

$$\Pr\left(\begin{smallmatrix}\text{d'observer 1 événement de } N_1 \text{ avant}\\ \text{d'observer 1 événement de } N_2\end{smallmatrix}\right) = \Pr(T_{1,1} < T_{2,1}) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

On peut généraliser ceci pour trouver une distribution **binomiale négative** *ou* **binomiale** :

$$\begin{aligned}
\Pr\left(\begin{smallmatrix}\text{d'observer } n \text{ événements de } N_1 \text{ avant}\\ \text{d'observer } m \text{ événement de } N_2\end{smallmatrix}\right) &= \Pr(T_{1,n} < T_{2,m}) \\
&= \Pr\left(\begin{smallmatrix}\text{d'observer au plus } m-1 \text{ événements de } N_2 \text{ avant}\\ \text{d'observer le } n^e \text{ événement de } N_1\end{smallmatrix}\right) \\
&= \sum_{k=0}^{m-1} \binom{n+k-1}{n-1} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^n \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^k \\
&= \Pr\left(\begin{smallmatrix}\text{parmis les } n+m-1 \text{ premiers événements}\\ \text{au } \textbf{moins } n \text{ proviennent de } N_1 \text{ et}\\ \text{au } \textbf{plus } m-1 \text{ proviennent de } N_2\end{smallmatrix}\right) = \Pr\left(\begin{smallmatrix}n^e \text{ événement de } N_1 \text{ se produise avant}\\ \text{le } m^e \text{ événement de } N_2\end{smallmatrix}\right) \\
&= \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{(n+m-1)-k}
\end{aligned}$$

Notes sur la représentation sous la forme **binomiale négative** :

Dans l'équation, on traite une réalisation de N_1 comme un « succès » et une réalisation de N_2 comme un « échec ».

« *Au plus* $m-1$ », implique tout nombre d'événements du 2^e processus allant de 0 à $m-1$.

L'approche est donc de fixer n réalisations de N_1, puis de traiter tous les autres cas possibles en faisant varier le nombre de réalisations N_2 de 0 à $m-1$.

Au total, il y aura au moins n événements ($N_2 = 0$) et au plus $n+m-1$ événements ($N_2 = m-1$) qui vont se réaliser.

Ceci résulte en m différents scénarios possibles.

Notes sur la représentation sous la forme **binomiale** :

Dans l'équation, on traite une réalisation de N_1 comme un « succès ».

L'approche est donc de fixer le nombre de réalisations total à $n+m-1$ puis, d'attribuer le nombre d'événements aux deux processus en assurant *au moins* n réalisations de N_1.

Mélanges de processus de Poisson

Lorsque la fonction d'intensité *est* une variable aléatoire, nous obtenons un mélange de processus de Poisson. Ce **mélange** est un nouveau processus qui **n'est pas un processus de Poisson**.

Identité Poisson-Gamma

Si la v.a. conditionnelle $(N|\Lambda) \sim \text{Poisson}(\Lambda)$ et que $\Lambda \sim \text{Gamma}(n, \theta)$ alors la v.a. inconditionnelle $N \sim \text{Binomiale Négative}(r = n, \theta)$.

Processus de Poisson composés

Processus de Poisson composé

Contexte

Les distributions composées permettent aux compagnies d'assurance de conjointement modéliser la fréquence et la sévérité de sinistres.

Si la fréquence d'accidents est distribuée selon une loi de Poisson et que les montants sont iid, la somme des montants des sinistres est un **processus de Poisson composé**.

Soit :

le processus de Poisson N,

la suite de v.a. iid $X_1, X_2, \ldots, X_{N(t)}$.

Alors $S(t) = \sum_{i=1}^{N(t)} X_i$ est un **processus de Poisson composé** où $S(0) = 0$ et si $N(t) = 0$ alors $S(t) = 0$.

Fonctions du processus de Poisson composé

$\mathrm{E}[S(t)] = \mathrm{E}[N(t)]\mathrm{E}[X]$ $\quad$ $\mathrm{Var}(S(t)) = \mathrm{E}[N(t)]\mathrm{E}[X^2]$

Approximation de la distribution

Puisque la distribution de $S(t)$ est difficile à déterminer, elle peut être approximée avec le théorème centrale limite où $S(t) \approx \mathcal{N}\left(\mathrm{E}[S(t)], \mathrm{Var}(S(t))\right)$.

Il s'ensuit que :

$$\Pr(S(t) < s) = \Phi\left(\frac{s - \mathrm{E}[S(t)]}{\sqrt{\mathrm{Var}(S(t))}}\right)$$

Cependant, dans le cas où nous utilisons une distribution continue (normale) pour approximer une distribution **de sévérité** discrète, il faut appliquer une correction de continuité.

Correction de continuité

La correction de continuité s'applique lorsqu'une distribution continue approxime une distribution discrète.

Une distribution discrète est seulement définie sur les nombres entiers alors qu'une distribution continue est définie sur tous les nombres réels. La correction améliore donc l'estimation en remplaçant s par le point milieu entre s et la plus proche valeur de $S(t)$ qui est inférieure à s.

Sommer des processus de Poisson résulte en un processus de Poisson dont la v.a. de sévérité est la moyenne des v.a. de sévérités de chacun des processus. C'est-à-dire que $f_X(x) = \frac{\lambda_1}{\lambda_1+\lambda_2} f_{X_1}(x) + \frac{\lambda_2}{\lambda_1+\lambda_2} f_{X_2}(x)$.

Chaînes de Markov

Introduction

Contexte

Une chaîne de Markov est utilisée lorsqu'il y a un processus prenant une valeur précise dans chaque intervalle de temps.

Les ***états*** du processus sont les valeurs possibles qu'il peut prendre.

Typiquement, les états sont dénotes par des nombres entiers.

Le processus peut seulement être dans un seul état par intervalle de temps.

Par exemple, un pourrait avoir une chaîne de Markov dont les états correspondent au nombre de vélos qu'une boutique de sport a en stock à chaque jour au moment de la fermeture du magasin.

Souvent, nous sommes intéressés aux ***probabilités de transition*** d'un état à un autre.

Notation

X_m État du processus au temps m.

$P_{i,j}$ Probabilité de transition de l'état i à l'état j (en une période).

Chaîne de Markov

Une chaîne de Markov est un type de processus stochastique dénoté comme $\{X_m, m = 0, 1, 2, \ldots\}$. Le processus prend un ensemble (fini ou infini) de valeurs *dénombrable* représentant l'état du processus à différents moments dans le temps.

$X_m = i$ signifie que le processus est dans l'état i au temps m.

Homogénéité de la chaîne de Markov

Si les probabilités de transition sont :

fixes le processus est une chaîne de Markov ***homogène***, ou ***stationnaire***.

variables le processus est une chaîne de Markov ***non-homogène***.

Propriété sans-mémoire des chaînes de Markov

Une chaîne de Markov est un processus stochastique dont la distribution conditionnelle de l'état futur X_{m+1} dépend seulement du dernier état X_m et non de ceux avant.

En autres mots, le prochain état est indépendant des états passés et $P_{i,j} = \Pr(X_{m+1} = j | X_m = i)$.

On représente la **matrice des probabilités de transition** $\boldsymbol{P}$:

$$\boldsymbol{P} = \begin{bmatrix} P_{1,1} & P_{1,2} & \cdots & P_{1,j} & \cdots \\ P_{2,1} & P_{2,2} & \cdots & P_{2,j} & \cdots \\ \vdots & \vdots & \ddots & \vdots & \cdots \\ P_{i,1} & P_{i,2} & \cdots & P_{i,j} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

Chaque rangée somme à 1, mais pas nécessairement les colonnes.

Probabilités de transitions en plusieurs étapes

Contexte

Lorsque nous désirons savoir l'état plus qu'une étape dans le futur, nous devons généraliser les chaînes de Markov.

Par exemple, s'il pleut aujourd'hui, quel est la probabilité qu'il va pleuvoir dans 2 jours ?

Notation

$P_{i,j}^n$ Probabilité de transition de l'état i à l'état j en n périodes.

Équation de Chapman-Kolmogorov

L'équation de Chapman-Kolmogorov trouve la probabilité $P_{i,j}^{n+m}$ d'être dans l'état j au temps $n+m$ sachant qu'au temps 0 on était à l'état i.

Pour trouver cette probabilité, on considère tous les chemins possibles pour se rendre de i à j en $n+m$ étapes, puis on somme leurs probabilités : $P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^n P_{kj}^m$.

Cette équation équivaut à la **multiplication matricielle** de la matrice des transitions de probabilité.

En forme matricielle, $\boldsymbol{P}^{(n+m)} = \boldsymbol{P}^{(n)}\boldsymbol{P}^{(m)}$.

Rappel : Multiplication matricielle

Soit $\boldsymbol{A}_{m\times n}$ et $\boldsymbol{B}_{p\times q}$. Si $n = p$ alors $\boldsymbol{A}_{m\times n}\boldsymbol{B}_{p\times q} = \boldsymbol{AB}_{m\times q}$.

Par exemple, pour :

$$\boldsymbol{A} = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix}$$

Alors :

$$\boldsymbol{AB} = \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{bmatrix}$$

Raccourci On peut éviter deux multiplications de matrices en multipliant uniquement la rangée i et la colonne j : $P_{i,j}^n = \boldsymbol{P}_{i,} \cdot \boldsymbol{P}^{n-2} \cdot \boldsymbol{P}_{,j}$.

États absorbants

État absorbant

État dont on ne peut pas sortir un fois rentrée. Il s'ensuit que pour un état absorbant i, $P_{i,i} = 1$.

Par exemple, un état pour décédé sera absorbant.

Soit la probabilité qu'une chaîne de Markov débute à l'état i et se rend à l'état j au temps m sans avoir été dans les états d'un ensemble $\mathcal{A}$.

Pour calculer la probabilité, on défini une nouvelle chaîne de Markov qui contient tous les états ne faisant **pas** parti de l'ensemble $\mathcal{A}$ en plus d'un état absorbant représentant tous les états de $\mathcal{A}$.

Notation

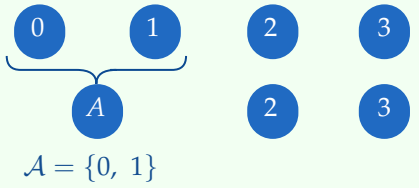
$\mathcal{A}$ L'ensemble des états à éviter.

A L'état absorbant qui combine tous les états de l'ensemble $\mathcal{A}$.

$Q_{i,j}$ Probabilité de transition de l'état i à l'état j (en une période) sans avoir accédé aux états de l'ensemble $\mathcal{A}$.

Exemple de regroupement

Par exemple, pour 4 états où on souhaite regrouper les états 0 et 1 :



$\mathcal{A} = \{0, 1\}$

Construction de la matrice Q

On construit $\boldsymbol{Q}$ de $\boldsymbol{P}$ selon les conditions suivantes :

1. Pour la transition entre des états qui ne font pas partie de l'ensemble

$\mathcal{A}$, la probabilité de transition demeure inchangée : $Q_{i,j} = P_{i,j}$ pour $i, j \notin \mathcal{A}$.

2. Pour la transition **de l'état non-absorbant i vers l'état absorbant** A, on somme les probabilités de transition de l'état i vers tous les états de l'ensemble $\mathcal{A}$: $Q_{i,A} = \sum_{k \in \mathcal{A}} P_{i,k}$ pour $i \notin \mathcal{A}$.
3. Par définition, $\Pr\left(\begin{smallmatrix}\text{transition d'un état absorbant} \\ \text{vers tout autre état}\end{smallmatrix}\right) = 0$: $Q_{A,i} = 0$ pour $i \notin \mathcal{A}$.
4. Par définition, $\Pr(\text{demeurer dans un état absorbant}) = 1$: $Q_{A,A} = 1$.

Finalement, on vérifie que change rangée de $\boldsymbol{Q}$ somme à 1.

Exemple de matrice de transition avec état absorbant

Soit la matrice des probabilités de transition suivante avec 4 états (1, 2, 3, 4) :

$$\boldsymbol{P} = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0 \\ 0 & 0.7 & 0.2 & 0.1 \\ 0.6 & 0.2 & 0 & 0.2 \\ 0.8 & 0.1 & 0.1 & 0 \end{bmatrix}$$

On sait qu'au temps 0, la chaîne de Markov est dans l'état 1. On souhaite trouver la probabilité d'atteindre l'état 2 au temps 4 sans jamais avoir été dans l'état 3 ni 4.

1. On défini l'ensemble $\mathcal{A} = \{3, 4\}$.
2. On défini la nouvelle chaîne de Markov $\boldsymbol{Q}$:

   De la première condition, le carré 2x2 en haut à gauche de la matrice des transitions demeure inchangée.

   La troisième colonne découle de la 2^e condition qui somme les probabilités de transitions vers les états faisant partie de $\mathcal{A}$.

   La troisième ligne découle des 4^e et 3^e conditions que l'état A est absorbant.

$$\boldsymbol{Q} = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 1 \end{bmatrix}$$

3. On trouve la matrice de transitions en 2 étapes :

$$\boldsymbol{Q} = \begin{bmatrix} 0.25 & 0.36 & 0.39 \\ 0 & 0.49 & 0.51 \\ 0 & 0 & 1 \end{bmatrix}$$

4. Finalement, on trouve $Q^4_{1,2} = \boldsymbol{Q}_1 \boldsymbol{Q}^{(2)} \boldsymbol{Q}_2$:

$$Q^4_{1,2} = \begin{bmatrix} 0.5 & 0.3 & 0.2 \end{bmatrix} \begin{bmatrix} 0.25 & 0.36 & 0.39 \\ 0 & 0.49 & 0.51 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.7 \\ 0 \end{bmatrix}$$
$$= 0.2664$$

Transitions de (ou vers) un état absorbant

Notation

$Q^m_{i,j}$ Probabilité de transition de l'état i à l'état j en m périodes sans avoir accédé aux états de l'ensemble $\mathcal{A}$.

Nous pouvons généraliser l'approche pour les cas où l'état de départ i ou l'état d'arrivé j peuvent faire partie de l'ensemble d'états $\mathcal{A}$.

Dans ces cas-ci la transition **de** (**vers**) l'état $\mathcal{A}$ doit être la **première** (**dernière**) transition.

On utilise donc la matrice des probabilités de transition P pour la **première** (**dernière**) transition où l'on **sort de** (**entre dans**) un état de l'ensemble $\mathcal{A}$, puis la matrice Q pour le restant des transitions.

État i	État j	Probabilité
$i \notin \mathcal{A}$	$j \notin \mathcal{A}$	$Q^m_{i,j}$
$i \notin \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} Q^{m-1}_{i,r} P_{r,j}$
$i \in \mathcal{A}$	$j \notin \mathcal{A}$	$\sum_{r \notin \mathcal{A}} P_{i,r} Q^{m-1}_{r,j}$
$i \in \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} \sum_{k \notin \mathcal{A}} P_{i,r} Q^{m-2}_{r,k} P_{k,j}$

Probabilités inconditionnelles

Notation

α_i Probabilité d'être à l'état i au temps 0.

$\alpha_i = \Pr(X_0 = i)$.

$\Pr(X_n = j)$ Probabilité "inconditionnelle" d'être dans l'état j au temps n. C'est-à-dire, la probabilité d'être dans l'état j au temps n peu importe l'état initial.

$\Pr(X_n = j) = \sum_{i=1}^{\infty} \alpha_i P^n_{i,j}$.

Rappel : Loi des probabilités totales

$\Pr(X = x) = \sum_y \Pr(X = x|Y = y) \Pr(Y = y)$.

Classification des états

Accessibilité d'états

Un état j est ***accessible*** de l'état i si $P_{i,j}^n > 0$ pour $n \geq 0$: $i \to j$.

C'est-à-dire qu'il est possible de faire la transition vers l'état j au moins une fois dans le futur ayant commencé dans l'état i.

Communication d'états

L'état i et l'état j se ***communiquent*** si l'état j est *accessible* de l'état i et que l'état i est *accessible* de l'état j : $i \leftrightarrow j$ si $i \to j$ et $j \to i$.

Note Un état absorbant communique seulement avec lui-même.

Propriétés des états qui se communiquent

1. $i \leftrightarrow i$
   L'état i communique avec lui-même.
2. $i \leftrightarrow j \Rightarrow j \leftrightarrow i$
   Si l'état i communique avec l'état j, alors l'état j communique avec l'état i.
3. $i \leftrightarrow j, j \leftrightarrow k \Rightarrow i \leftrightarrow k$
   Si l'état i communique avec l'état j et que l'état j communique avec l'état k, alors l'état i communique avec l'état k.

Classe d'états

Des états qui se communiquent entre-eux font partie de la même classe.

Propriétés de classe

Propriétés s'appliquant à tous les états de la classe.

Chaîne de Markov irréductible

Chaîne de Markov dont tous les états se communiquent entre-eux ayant donc **une seule classe**.

Nombre d'états d'une chaîne de Markov

Une chaîne de Markov ayant un nombre **fini** (**infini**) d'états est dite d'être **fini** (**infini**).

Notation

f_i Probabilité de retourner dans l'état i à tout point dans le future sachant que le processus débute dans l'état i.

Récurrence d'états

Un état est ***récurrent*** s'il est toujours possible d'y retourner un jour : $f_i = 1$.

Il s'ensuit que si un état i est récurrent, alors le nombre de fois que nous y retournons est **infini**. De cette interprétation, on déduit qu'un état est récurrent si $\sum_{n=1}^{\infty} P_{i,i}^n = \infty$.

Il s'ensuit qu'il est toujours possible de retourner dans l'état i à partir de tout autre état dans le futur.

Transitivité d'états

Un état est ***transitoire*** s'il est possible de ne pas y retourner un jour : $f_i < 1$.

Il s'ensuit que si un état i est transitoire, alors le nombre de fois que nouss y retournons est **fini**. De cette interprétation, on déduit qu'un état est transitoire si $\sum_{n=1}^{\infty} P_{i,i}^n < \infty$.

On déduit que si un état i est transitoire, alors il existe au moins un état duquel on ne peut pas retourner à l'état i.

Distribution géométrique

Si un processus débute dans un état transitoire i, il y a une probabilité de $1 - f_i$ de ne jamais y retourner. Il s'ensuit que la probabilité d'être dans l'état i n fois, sachant que nous y sommes initialement, est $f_i^{n-1}(1 - f_i)$ pour $n \geq 1$.

Donc, pour un processus qui débute dans l'état transitoire i, le nombre de fois que le processus est dans l'état i suit une **distribution géométrique** de paramètre $p = 1 - f_i$.

Il s'ensuit que l'espérance du nombre de visites est $\frac{1}{1-f_i}$.

On voit donc que pour $n \geq 1$, la probabilité désiré correspond à la fonction de masse des probabilités $p_n = p(1-p)^{n-1} = f_i^{n-1}(1 - f_i)$.

Exemple de transitivité et de récurrence

Soit la chaîne de Markov ayant la matrice des probabilité de transition suivante :

$$\boldsymbol{P} = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0 & 0.4 & 0.6 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

On trouve :

Aucun état est absorbant.

L'état 1 est *transitoire* et seulement l'état 2 est accessible de l'état 1 ($1 \to 2$).

L'état 2 et l'état 3 se *communiquent* ($2 \leftrightarrow 3$).

Les propriétés de *récurrence* et de *transitivité* sont des **propriétés de classes**.

Puisque tous les états d'une classe se communiquent, dès qu'un état est récurrent tous les états sont récurrents.

Pareillement, dès qu'un état est transitoire tous les états sont transitoires.

Donc, tous les états d'une classe sont soit transitoires ou récurrents.

Dans une chaîne de Markov finie, il doit y avoir au moins un état récurrent. Puis, puisqu'une chaîne de Markov irréductible n'a qu'une seule classe, **tous les états d'une chaîne de Markov finie irréductible sont récurrents**.

Probabilités stationnaires et limites

Notation

m_j Espérance du nombre de transitions pour qu'une chaîne de Markov ayant commencé dans l'état j y retourne.

π_j **Proportion de temps à long-terme** qu'une chaîne de Markov *irréductible* est dans l'état j.

En anglais, « *long-run proportion* ».

Alias, **probabilité stationnaire** d'être dans l'état j.

Types de récurrence

Soit l'état récurrent j,

1. si $m_j < \infty$, alors l'état j est **récurrent positif**.
2. si $m_j = \infty$, alors l'état j est **récurrent nul**.

La récurrence nulle peut seulement arriver dans une chaîne de Markov infinie ce qui implique que **les états d'une chaîne de Markov finie doivent être récurrent positifs**.

Puisque la récurrence est une propriété de classe, une classe est soit récurrente positive ou nulle.

Probabilités stationnaires

La probabilité stationnaire π_j de l'état j correspond au réciproque de l'espérance du nombre transitions pour qu'une chaîne de Markov ayant débuté dans l'état j y retourne : $\pi_j = \frac{1}{m_j}$.

Cependant, on isole habituellement les probabilités stationnaires à partir du système d'équations suivant :

$$\pi_j = \sum_{i=1}^{\infty} \pi_i P_{i,j}$$

$$\sum_{j=1}^{\infty} \pi_j = 1$$

Pour une chaîne de Markov composé de n états, il y aura $n+1$ équations.

Si aucune solution unique existe, la chaîne de Markov **n'est pas récurrente positive** (donc soit transitive ou récurrente nulle) et $\pi_i = 0$ pour tout i.

Note Tous les états d'une chaîne de Markov irréductible finie sont récurrent positifs.

Chaînes de Markov avec bénéfices

Notation

$r(j)$ Montant de bénéfice dans l'état j.

Contexte

On cherche à généraliser les chaînes de Markov pour le cas où un montant est transigé selon la classe dans laquelle le processus se situe.

Par exemple, pour une chaîne de Markov représentant le risque d'un assuré $r(j)$ pourrait représenter le montant de prime payable en fonction de classe dont l'assuré fait partie. Par exemple, il pourrait avoir une plus grosse prime payable pour une classe de risque risquée que standard.

En moyenne, le bénéfice sera $\sum_{j=1}^{\infty} r(j)\pi_j$.

Probabilités limites

Périodicité des chaînes de Markov

La matrice des probabilités de transition tend vers des ***probabilités limites*** lorsque le nombre de périodes tend vers l'infini. Ces probabilités limites correspondent aux probabilités stationnaires.

Si une chaîne de Markov **a des** (**n'a pas de**) probabilités limites, elle est **apériodique** (**périodique**).

Note Une chaîne de Markov peut avoir des probabilités stationnaires sans avoir de probabilités limites.

Exemple de chaîne de Markov périodique

Soit la chaîne de Markov suivante :

$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$

Cette chaîne de Markov ne converge pas vers des probabilités limites, à chaque période elle va inverser :

$\boldsymbol{A}^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\boldsymbol{A}^{(3)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ $\boldsymbol{A}^{(4)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

La chaîne de Markov est donc ***périodique***.

Chaîne de Markov ergodique

Une chaîne de Markov **irréductible**, **récurrente positive** et **apériodique** est ***ergodique***.

Temps passé dans les états transitoires

Notation

$\boldsymbol{P}_T$ Matrice des probabilités de transition contentant uniquement les états transitoires.

Les rangées ne somment donc pas nécessairement à 1.

$s_{i,j}$ Espérance du nombre de périodes que le processus est dans l'état transitoire j sachant que le processus a débuté dans l'état transitoire i.

$\boldsymbol{S}$ Matrice des valeurs de $s_{i,j}$.

$\boldsymbol{S} = (\boldsymbol{I} - \boldsymbol{P}_T)^{-1}$.

Note : Les indices de la matrice représentent les états et non la position dans la matrice.

Par exemple, si on retire la deuxième colonne alors les indices seront $s_{i,1}, s_{i,3}, s_{i,4}, \dots$.

$f_{i,j}$ Probabilité d'aller dans l'état j à tout point dans le futur sachant que le processus débute dans l'état i.

$f_{i,j} = \frac{s_{i,j} - \delta_{i,j}}{s_{j,j}}$.

Note En anglais, on dit « *Time Spent in Transient States* ».

Rappel : Matrice d'identité

La matrice d'identité $\boldsymbol{I}$ est la suivante :

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \cdots \\ 0 & 0 & 1 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

On exprime les valeurs de $\boldsymbol{I}$ avec la variable binaire $\delta_{i,j}$:

$$\delta_{i,j} = \begin{cases} 1, & \text{si } i = j \\ 0, & \text{si } i \neq j \end{cases}$$

Rappel : Inverse d'une matrice

Notation

A^{-1} Inverse de la matrice A tel que $A^{-1}A = AA^{-1} = I$.

Soit la matrice 2×2 A où :

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Alors son inverse A^{-1} est :

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

Pour plus de 3 dimensions, c'est long et peu probable d'être dans l'examen.

Exemple du calcul du temps espéré

Soit la chaîne de Markov à trois états (1, 2, 3) avec la matrice de transition suivante :

$$P = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$$

On souhaite trouver l'espérance du nombre de périodes passées dans l'état 2 sachant qu'on débute dans l'état 1.

1. Trouver la matrice de transitions pour les états transitoires :
   $$P_T = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0.8 \end{bmatrix}$$
2. Trouver $I - P_T$:
   $$I - P_T = \begin{bmatrix} 0.5 & -0.5 \\ 0 & 0.2 \end{bmatrix}$$
3. Trouver l'inverse $(I - P_T)^{-1}$:
   $$(I - P_T)^{-1} = \begin{bmatrix} 0.2 & 0.5 \\ 0 & 0.5 \end{bmatrix} \times \frac{1}{0.10 - 0} = \begin{bmatrix} 2 & 5 \\ 0 & 5 \end{bmatrix}$$
4. Trouver l'élément $s_{1,2}$ de la matrice $S = (I - P_T)^{-1}$ et donc $s_{1,2} = 5$.

« *Time Reversibility* »

Notation

$R_{i,j}$ Probabilité de transition de l'état i à l'état j (en une période) pour la chaîne de Markov inverse.

On dénote la matrice des probabilités de transition de la chaîne de Markov inverse par R.

Contexte

Lorsque l'on désire trouver la séquence des états à partir du dernier, on veut le processus inverse de la chaîne de Markov.

Chaîne de Markov inverse

Soit la chaîne de Markov **stationnaire** et **ergodique** $\{X_m, m \geq 0\}$. Alors, le processus inverse $(X_m, X_{m-1}, \dots)$ est lui-même une chaîne de Markov avec probabilités de transition $R_{i,j} = P_{j,i} \times \frac{\pi_j}{\pi_i}$.

Note On pose que la chaîne de Markov est *stationnaire* afin qu'elle soit "**homogène**" et que les probabilités de transition ne changent pas dans le temps.

Chaîne de Markov « *time reversible* »

Si $R_{i,j} = P_{i,j}$ pour tout i et j, la chaîne de Markov est « *time reversible* » et $\pi_i P_{i,j} = \pi_j P_{j,i}$.

Il s'ensuit que la probabilité que le processus fasse la transition d'un état i vers un état j est la même que pour la probabilité de la transition d'un état j vers un état i, et cela peu importe le chemin. C'est à dire, $P_{i,j}P_{j,k}P_{k,i} = P_{i,k}P_{k,j}P_{j,i}$

Note Un truc pour déterminer si une chaîne de Markov est réversible est de vérifier si pour un i et j que $P_{i,j} = 0$ alors $P_{j,i} = 0$.

Exemple de chaîne de Markov inverse

Soit la chaîne de Markov à 2 états (1, 2) avec la matrice des probabilités de transition suivante :

$$\boldsymbol{P} = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.1 & 0.6 & 0.3 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}$$

1 Trouver les probabilités limites :

$$\pi_1 = 0.3\pi_1 + 0.1\pi_2 + 0.2\pi_3 \Rightarrow \pi_1 = \frac{1}{7}\pi_2 + \frac{2}{7}\pi_3$$
$$\pi_2 = 0.2\pi_1 + 0.6\pi_2 + 0.1\pi_3 \Rightarrow \pi_2 = \frac{1}{2}\pi_1 + \frac{1}{4}\pi_3$$
$$\therefore \pi_2 = \frac{1}{2}\left(\frac{1}{7}\pi_2 + \frac{2}{7}\pi_3\right) + \frac{1}{4}\pi_3 = \frac{\frac{1}{7}\pi_3 + \frac{1}{4}\pi_3}{13/14}$$
$$= \frac{2}{13}\pi_3 + \frac{7}{26}\pi_3 = \frac{11}{26}\pi_3$$
$$\pi_1 + \pi_2 + \pi_3 = 1 \Rightarrow \frac{1}{7}\left(\frac{11}{26}\pi_3\right) + \frac{11}{26}\pi_3 + \pi_3 = 1 \Rightarrow \pi_3 = \frac{182}{270}$$
$$\pi_2 = \frac{11}{26} \times \frac{182}{270} = \frac{77}{270}$$
$$\pi_1 = \frac{1}{7}\frac{77}{270} + \frac{2}{7}\frac{182}{270} = \frac{7}{30}$$

2 Trouver probabilités de transition de la chaîne de Markov inverse $\boldsymbol{R}$:

(a) $\boldsymbol{R}_{11} = \boldsymbol{P}_{11}\frac{\pi_1}{\pi_1} = 0.3$

(b) $\boldsymbol{R}_{22} = \boldsymbol{P}_{22}\frac{\pi_2}{\pi_2} = 0.6$

(c) $\boldsymbol{R}_{33} = \boldsymbol{P}_{33}\frac{\pi_3}{\pi_3} = 0.7$

(d) $\boldsymbol{R}_{12} = \boldsymbol{P}_{21}\frac{\pi_2}{\pi_1} = 0.1 \times \frac{77/270}{7/30} = 0.12$

(e) $\boldsymbol{R}_{13} = \boldsymbol{P}_{31}\frac{\pi_3}{\pi_1} = 0.2 \times \frac{182/270}{7/30} = 0.58$

(f) $\boldsymbol{R}_{21} = \boldsymbol{P}_{12}\frac{\pi_1}{\pi_2} = 0.2 \times \frac{7/30}{77/270} = 0.16$

(g) $\boldsymbol{R}_{23} = \boldsymbol{P}_{32}\frac{\pi_3}{\pi_2} = 0.1 \times \frac{182/270}{77/270} = 0.24$

(h) $\boldsymbol{R}_{31} = \boldsymbol{P}_{13}\frac{\pi_1}{\pi_3} = 0.5 \times \frac{7/30}{182/270} = 0.17$

(i) $\boldsymbol{R}_{32} = \boldsymbol{P}_{23}\frac{\pi_2}{\pi_3} = 0.3 \times \frac{77/270}{182/270} = 0.13$

3 Construire la matrice des probabilités de transition inverse :

$$\boldsymbol{R} = \begin{bmatrix} 0.30 & 0.12 & 0.58 \\ 0.16 & 0.60 & 0.24 \\ 0.17 & 0.13 & 0.70 \end{bmatrix}$$

Applications des chaînes de Markov

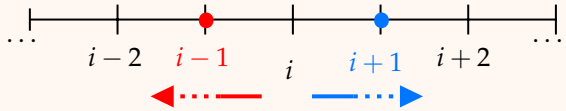
Marche aléatoire

Marche aléatoire

1 À une dimension

Une *marche aléatoire* à une dimension équivaut à une chaîne de Markov qui, de l'état i, peut seulement aller soit à l'état $i+1$ avec probabilité $P_{i,i+1} = p$ ou l'état $i-1$ avec probabilité $P_{i,i-1} = 1-p$ où $p \in [0,1]$.

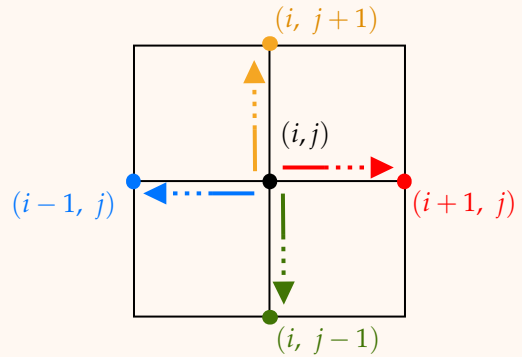
On peut donc visualiser une ligne :



2 À deux dimensions

Une marche aléatoire de deux dimensions représente chaque état comme une paire de chiffres (i,j) et donc le prochain état peut être un des quatre états suivants : $(i-1,j), (i+1,j), (i,j-1), (i,j+1)$.

On peut donc visualiser un carré en représentant l'état comme des coordonnées :



Marche aléatoire symétrique

S'il y a une probabilité égale d'aller dans toute direction, la marche aléatoire est *symétrique*. Par exemple, dans le cas d'une dimension $p = 0.5$ et dans 2 dimensions $p = 0.25$.

Les marches aléatoires sont seulement ***récurrentes*** si elles ont une ou deux dimensions et qu'elles sont symétriques. Autrement, elles sont ***transitoires***.

« *Gambler's ruin* »

Notation

P_i Probabilité de commencer i jetons et terminer avec j jetons.

Le complément $1 - P_i$ est la probabilité de commencer avec i jetons et de terminer avec aucun (0).

X Variable aléatoire du nombre de jetons que le « *gambler* » a à la fin.

« *Gambler's ruin problem* »

Soit un jeu où, à chaque ronde, un « *gambler* » *gagne* un jeton avec probabilité p ou *perd* un jeton avec probabilité $1 - p$. L'objectif est de se rendre à j jetons.

Le « *gambler's ruin problem* » est de calculer la probabilité qu'un « *gambler* » qui commence avec i jetons va terminer le jeu avec j jetons.

« *Gambling model* »

Le modèle qu'on utilise pour modéliser le « *gambler's ruin problem* » se nomme le « *gambling model* ». Il s'apparente à la marche aléatoire sauf qu'il comporte un nombre ***fini*** d'états. Les états correspondent au nombre de jetons.

Propriétés du « *gambling model* »

1. Puisque le « *gambler* » arrête lorsqu'il a soit 0 ou j jetons, $P_{0,0} = P_{j,j} = 1$.
   Il s'ensuit que les états 0 et j sont absorbants.
2. La probabilité de gagner $P_{i,i+1} = p$ et la probabilité de perdre $P_{i,i-1} = 1 - p$ où $i \in \{1, 2, \ldots, j-1\}$.
3. Il y a 3 classes : $\{0\}, \{1, 2, \ldots, j-1\}, \{j\}$.
4. Les états $\{0\}$ et $\{j\}$ son récurrents puisqu'ils sont absorbants et les états $\{1, 2, \ldots, j-1\}$ sont transitoires.

Distribution du nombre de jetons

La variable aléatoire X suit une distribution avec deux valeurs possibles : 0 ou j avec probabilités de P_i et $1 - P_i$ respectivement. Il s'ensuit que X suit une loi de Bernoulli :

$$\Pr(X = x) = \begin{cases} P_i, & x = j \\ 1 - P_i, & x = 0 \end{cases}$$

La probabilité d'un succès P_i est définie comme suit :

$$P_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^j}, & p \neq \frac{1}{2} \\ \frac{i}{j}, & p = \frac{1}{2} \end{cases}$$

Exemple de calcul de Gambler's Ruin

Soit un joueur ayant 10 jetons. À chaque ronde, il parie 1 jeton avec une stratégie qui lui garantit une probabilité de 0.6 de gagner. S'il gagne, il gagne un jeton de plus que celui qu'il a parié. Sinon, il en perd un additionnel. Quelle est la probabilité qu'il réussisse à amasser 25 jetons ?

1 Calculer la probabilité d'amasser 25 jetons en commençant avec 10 lorsqu'il y a une probabilité $p = 0.6$ de gagner :

$$\frac{1-\left(\frac{0.4}{0.6}\right)^{10}}{1-\left(\frac{0.4}{0.6}\right)^{25}} = 0.9827$$

Exemple de calcul de Gambler's Ruin

Tu as 6 jetons et ton ami en a 4. Vous pariez des jetons équiprobables jusqu'à ce que quelqu'un se rend à 10 jetons. Quelle est la probabilité que vous amassez 10 jetons ?

Avec $p = 0.5$, on obtient que $P_{10} = \frac{6}{10} = 0.60$.

Pour calculer la variance, on rappelle le raccourci de Bernoulli :

Rappel : Raccourci de Bernoulli

Soit la variable aléatoire X prenant une de deux valeurs :

$$X = \begin{cases} a, & p \\ b, & 1-p \end{cases}$$

Alors, $\mathrm{Var}(X) = (b-a)^2 p(1-p)$.

« *Branching Process* »

Contexte

On pose que nous avons une population d'individus dont chacun produit j descendants d'ici la fin de leur durée de vie avec probabilité P_j.

Le nombre moyen de nouveaux descendants qu'un individu produit est $\mu = \sum_{j=0}^{\infty} jP_j$.

La variance du nombre de nouveaux descendants qu'un individu produit est $\sigma^2 = \sum_{j=0}^{\infty}(j-\mu)^2 P_j$.

Notation

X_n Taille de la n^e génération.

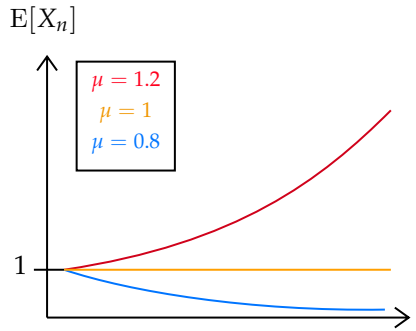
Si l'on pose une population initiale de 1 ($X_0 = 1$) :

$$\mathrm{E}[X_n] = \mu^n$$

$$\mathrm{Var}(X_n) = \begin{cases} \sigma^2\mu^{n-1}\left(\frac{1-\mu^n}{1-\mu}\right), & \mu \neq 1 \\ n\sigma^2, & \mu = 1 \end{cases}$$

Si la population initiale est de k ($X_0 = k$) alors la moyenne est de $k\mathrm{E}[X_n]$ et la variance de $k\mathrm{Var}(X_n)$.

On s'attend donc à ce que la population croît si $\mu > 1$ et décroît sinon :



On défini la probabilité que la population disparaisse π_0 si $X_0 = 1$ comme suit :

$$\pi_0 = \begin{cases} 1, & \mu \leq 1 \\ \sum_{j=0}^{\infty} \pi_0^j P_j, & \mu > 1 \end{cases}$$

Dans le cas où $\mu > 1$, il peut y avoir plusieurs solutions et donc on choisit la solution minimale.

Si la population initiale est de k ($X_0 = k$) alors la la probabilité que la population disparaisse est π_0^k.

VI

Séries chronologiques

Contexte

Les **données temporelles** (« *time series data* ») sont composées d'observations indexées par le temps. Typiquement, les données sont récoltées à des intervalles fixes nommées « ***sampling intervals*** ». Lorsque nous évaluons les séries chronologiques, on cherche à trouver des patrons dans les observations afin de prédire les prochaines valeurs de la série.

Notation

$\{x_t : t = 1, \ldots, n\}$ Série chronologique de longueur n, dénotée plus simplement par $\{x_t\}$.

Introduction

Notation

M_t « *Trends in time* ».
S_t « *Seasonal variations* ».
Z_t « *Random patterns* ».

De façon générale, on peut décomposer une série chronologique en 3 composantes :

1 Tendances avec le temps

La **tendance** (« *trend* ») M_t est la variation **systématique** d'une série chronologique. Par exemple, une croissance linéaire de la série chronologique avec le temps est une *tendance*.

2 Variations saisonnières

Une **variation saisonnière** S_t est un patron **cyclique** qui se répète dans une période de temps fixée. Par exemple, le volume de ventes d'une crèmerie sera plus élevé au cours de l'été que l'hiver.

Typiquement la période de temps est d'une année, mais on dénote par g la ***base saisonnière***. C'est-à-dire, le nombre d'unités de temps pour qu'un patron cyclique se répète.

3 Patrons aléatoires

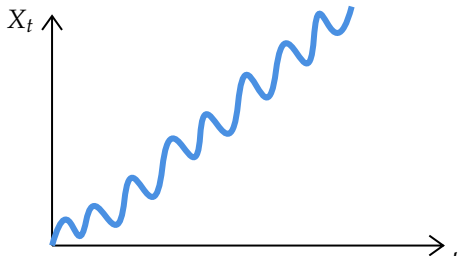
Tous **patrons aléatoires** Z_t sont des variations imprévisibles dans les données. Z_t est donc un terme d'erreur habituellement composé de v.a. corrélées de moyenne nulle.

Note Puisque X_t contient une composante aléatoire, ou ***stochastique***, Z_t, alors c'est un **processus stochastique**.

Il y existe une multitude de façons de combiner les 3 composantes, mais les modèles principaux sont :

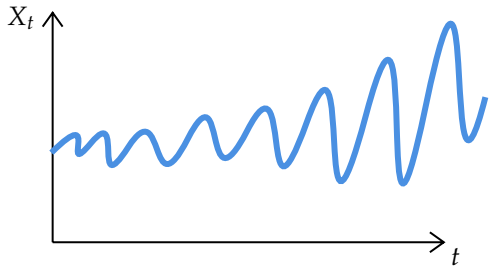
Modèle additif $X_t = M_t + S_t + Z_t$

Idéal si la variation saisonnière semble rester constante



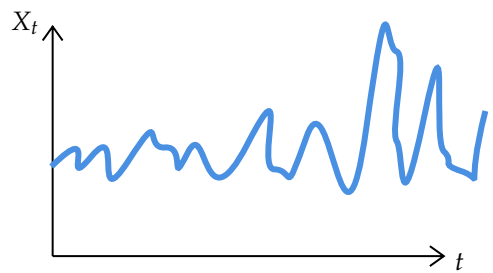
Modèle de saisonnalité multiplicatif $X_t = M_t \cdot S_t + Z_t$

Idéal si la variation saisonnière semble augmenter avec la tendance



Modèle multiplicatif $X_t = M_t \cdot S_t \cdot Z_t$

Idéal si la variation saisonnière et l'erreur aléatoire semblent augmenter avec la tendance



Note S'il n'y a pas de tendances saisonnières, on ignore la composante en fixant $S_t = 0$ pour le modèle additif ou $S_t = 1$ pour le modèle additif.

Stationnarité

Contexte

De façon générale, on désire qu'une série chronologique soit ***stationnaire***. La stationnarité implique que quelque-chose reste fixe dans le temps. Donc, un processus stochastique qui a une moyenne stationnaire aura la même espérance peu importe où on se situe dans le temps.

Les propriétés de second ordre d'un processus sont : sa moyenne, sa variance et sa corrélation. S'ils sont tous stationnaire, le processus est dit d'être ***stationnaire de second ordre***. Donc, $\mathrm{E}[X_t]$ ne varie pas selon t et $\mathrm{Cov}(X_t, X_s)$ dépend seulement de la différence $|t-s|$. Si **tous** les moments de X_t sont stationnaires, alors le processus stochastique est ***strictement stationnaire***. Dans ce cas-ci, les v.a. $X_1, X_2, \ldots, X_n$ sont **iid**.

Finalement, un modèle de séries chronologiques avec une moyenne stationnaire est dit d'avoir une moyenne ***ergodique*** si la moyenne, pour toute observation, tend vers la moyenne théorique alors que le temps tend vers l'infini.

Note La définition d'ergodique ici diffère donc de la définition d'une *chaîne de Markov ergodique*.

Puisqu'une série chronologique est beaucoup plus stable lorsqu'elle est stationnaire, on désire l'ajuster si elle ne l'est pas. Il y a plusieurs approches que l'on peut prendre pour le faire, mais les plus populaires sont :

1. Calculer les écarts de temps entre les observations consécutives $X_t - X_{t-1}$.
2. Appliquer une transformation sur les données temporelles. Par exemple, une transformation d'échelle logarithmique ($\ln(X_t)$).

Décomposition

Contexte

Nous avons mentionné que la tendance et la variation saisonnière sont déterministe alors que la variation aléatoire est stochastique. Nous voulons maintenant distinguer la tendance et la variation saisonnière afin des modéliser séparément.

Tendance

Notation

m_t Tendance au temps t.

g Base saisonnière.

Par exemple, pour des données mensuelles $g = 12$ et pour des données trimestrielles $g = 4$.

Moyenne mobile (« *moving average* »)

La tendance au temps t, m_t, est calculée comme étant la moyenne mobile centrée sur l'observation x_t. Afin de prendre en compte la variation saisonnière, la moyenne sera composée de g observations consécutives, soit :

1. Les $\frac{g-1}{2}$ observations avant x_t ;
2. L'observation x_t ;
3. Les $\frac{g-1}{2}$ observations après x_t.

Par exemple, pour des donnés mensuelles ($g = 12$) : $\hat{m}_t = \frac{0.5x_{t-6}+\sum_{i=t-5}^{t+5} x_i+0.5x_{t+6}}{12}$.

Note Si $\frac{g-1}{2}$ est fractionnaire, on pondère les extrémités par 0.5.

Contexte

La moyenne mobile centrée est une technique de ***lissage*** (« *smoothing* ») ou ***filtrage*** pour trouver une tendance sous-jacente aux données temporelles. Une autre technique populaire est une technique de régression avec des poids locaux : le ***loess***. Cependant, bien que cette technique est facilement applicable en `R` elle est trop compliquée pour faire à la main.

Finalement, nous avons seulement vu comment estimer la tendance. La section *Régression avec des séries chronologiques* présente des modèles pour ensuite effectuer des prévisions sur des séries chronologiques.

Variation saisonnière

Contexte

Une fois que les tendances sont estimées, on peut calculer les variations par rapport aux tendances pour trouver la variation saisonnière moyenne pour chaque saison. Puis, on peut soustraire ces variations des observations pour obtenir une série corrigée des variations saisonnières (« *seasonally adjusted series* »).

Variation saisonnière additive

Pour un modèle additif, chaque observation d'un ensemble de données temporel peut se décomposer comme $x_t = m_t + s_t + z_t$. Il s'ensuit que les *composantes de variation saisonnière* de la série se calculent comme $\hat{s}_t = x_t - \hat{m}_t$. Pour une variation saisonnière additive, on surnomme parfois ses composantes les ***effets additifs***.

Puis, on calcule la moyenne des composantes par saison pour obtenir g effets additifs moyens $\bar{s}_i$, pour $i = 1, 2, \ldots, g$, et on ajuste les effets moyens pour qu'ils aient une moyenne nulle avec $\bar{s}_i^* = \bar{s}_i - \frac{\sum_{i=1}^{g} \bar{s}_i}{g}$. Finalement, on calcule la série corrigée des variations saisonnières comme $x_t - \bar{s^*}_i$.

Variation saisonnière multiplicative

Pour un modèle multiplicatif, chaque observation d'un ensemble de données temporel peut se décomposer comme $x_t = m_t \cdot s_t + z_t$ ou $x_t = m_t \cdot s_t \cdot z_t$. Il s'ensuit que, peu importe le modèle multiplicatif, les *composantes de variation saisonnière* de la série se calculent comme $\hat{s}_t = \frac{x_t}{\hat{m}_t}$. On surnomme parfois les composantes les ***effets multiplicatifs***.

Puis, on calcule la moyenne des composantes par saison pour obtenir g effets multiplicatifs moyens $\bar{s}_i$, pour $i = 1, 2, \ldots, g$, et on ajuste les effets moyens pour qu'ils aient une moyenne de 1 avec $\bar{s}_i^* = \frac{\bar{s}_i}{\sum_{i=1}^{g} \bar{s}_i / g}$. Finalement, on calcule la série corrigé des variations saisonnières comme $\frac{x_t}{\bar{s}_i^*}$.

Autocorrélation

Contexte

Lorsque l'on retire la tendance et la variation saisonnière d'une série chronologique, il reste seulement la composante aléatoire qui est elle-même une série chronologique qu'on surnomme la ***série d'erreurs résiduelles*** (« *residual error series* »). Cependant, ces termes d'erreurs ne peuvent pas être supposés indépendants et il faut tenir en compte la corrélation d'observations consécutives.

Dans la section de *Apprentissage statistique*, les *Résumés numériques des modèles* utilisés pour des données bivariées sont *covariance* et la *corrélation*. Dans notre cas, nous voulons évaluer le degré de dépendance entre une même variable aléatoire à différents moments dans le temps. Donc, nous utilisons l'***autocovariance*** et l'***autocorrélation***.

Notation

lag k Décalage entre deux termes d'une série chronologique.
Par exemple, l'écart de temps entre X_t et X_{t+k}.

μ Moyenne de la série chronologique.

Les statistiques sont :

Autocovariance

L'autocovariance de la variable X avec un décalage de k est $\gamma_k = \text{Cov}(X_t, X_{t+k}) = \text{E}[(X_t - \mu)(X_{t+k} - \mu)]$.

Autocorrélation

L'autocorrélation de la variable X avec un décalage de k est $\rho_k = \text{Corr}(X_t, X_{t+k}) = \frac{\gamma_k}{\sigma^2}$.

Note On dénote la fonction d'autocorrélation comme l'***ACF***.

Note Les deux premiers moments de la série chronologique doivent être stationnaires pour calculer l'autocovariance et l'autocorrélation. C'est-à-dire que la série doit être stationnaire de second ordre.

Puis, pour un échantillon d'observation temporelles $x_1, \ldots, x_n$:

Autocovariance échantillonnale

L'autocovariance échantillonnale avec un décalage (« *lag* ») de k est $c_k = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{n}$.

Note On divise par n même s'il y a $n-k$ termes à la sommation afin que $c_k \in [-1, 1]$. Également, avec $k = 0$ on obtient la *variance empirique* (avec biais) : $c_0 = \hat{\sigma}^2$.

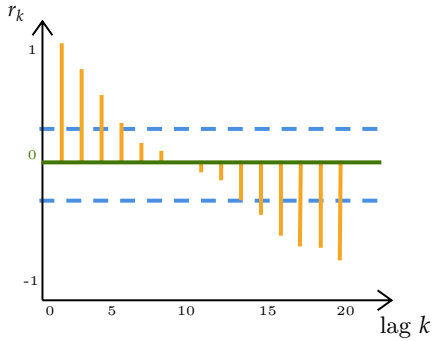
Autocorrélation échantillonnale

L'autocorrélation échantillonnale avec un décalage (« *lag* ») de k est $r_k = \frac{c_k}{c_0} = \frac{\sum_{t=1}^{n-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}$.

La distribution de r_k est approximativement normale avec $r_k \sim \mathcal{N}\left(-\frac{1}{n}, \frac{1}{n}\right)$ si $\rho_k = 0$.

Corrélogrammes

Une méthode d'évaluer s'il y existe d'importantes autocorrélations dans une série chronologique est le ***corrélogramme*** : un graphique de l'autocorrélation contre le décalage k :



Pour $k = 0$, on a toujours que $r_0 = 1$. Puis, on trace des lignes blueues aux points $-\frac{1}{n} \pm \frac{2}{\sqrt{n}}$ comme règle de pouce d'autocorrélations significatives ($z_{0.975 \approx 2}$). Si r_k se situe à l'extérieur de ces bornes, on rejette l'hypothèse nulle que $\rho_k = 0$ à un niveau de confiance de 95%.

De plus, on peut utiliser les corrélogrammes pour identifier s'il y a des patrons. Par exemple, si l'ACF décroît ça peut être indicatif d'une tendance qui demeure dans le données. Il s'ensuit qu'il est optimal d'utiliser le corrélogramme sur la série des erreurs résiduelles.

Cross-correlation

Contexte

Une série chronologique peut non seulement être corrélée avec un décalage d'elle-même, mais aussi avec un décalage d'une autre série chronologique. C'est le cas si une série devance (« *leads* ») une autre d'au moins une période. Par exemple, le nombre de meubles vendus au Ikea de Québec peut dépendre du nombre de nouvelles inscriptions à l'université Laval.

Pour quantifier cette relation, nous avons les statistiques suivantes :

Covariance croisée

La covariance croisée (« *cross-covariance* ») entre X et Y pour un décalage de k est $\gamma_k(X,Y) = \text{Cov}(X_{t+k}, Y_t) = \text{E}[(X_{t+k} - \mu_X)(Y_t - \mu_Y)]$.

Corrélation croisée

La fonction de corrélation croisée entre X et Y pour un décalage de k est $\rho_k(X,Y) = \text{Corr}(X_{t+k}, Y_t) = \frac{\gamma_k(X,Y)}{\sigma_X \sigma_Y}$.

Note On dénote la fonction de corrélation croisée comme la ***CCF***.

Puis, pour deux échantillons d'observations temporelles $x_1, \ldots, x_n$ et $y_1, \ldots, y_n$:

Covariance croisée échantillonnale

La covariance croisée échantillonnale avec un décalage de k est $c_k(x,y) = \frac{\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(y_t - \bar{y})}{n}$.

Note $\sigma_X^2 = c_0(x,x)$ et $\sigma_Y^2 = c_0(y,y)$.

Corrélation croisée échantillonnale

La corrélation croisée échantillonnale avec un décalage de k est $r_k(x,y) = \frac{c_k(x,y)}{\sqrt{c_0(x,x)c_0(y,y)}} = \frac{\sum_{t=1}^{n-k}(x_{t+k} - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2 \sum_{t=1}^{n}(y_t - \bar{y})^2}$.

Note Le décalage peut être négatif ce qui implique que $\gamma_k(X,Y) = \gamma_{-k}(Y,X)$, $\rho_k(X,Y) = \rho_{-k}(Y,X)$, etc.

Corrélogrammes croisés

De façon semblable au corrélogramme, on peut évaluer le corrélogrammes croisé pour détecter des corrélations croisées significatives entre les deux séries chronologiques. Ce coups-ci, on trace la CCF contre le décalage k.

Note On utilise les mêmes valeurs critiques pour tester si les termes sont nuls.

Modèles de séries chronologiques

Modèles de base

Contexte

Les deux premiers modèles de base que l'on voit sont le « *white noise* » et la marche aléatoire (avec et sans dérive). Également, on pose que l'espérance des séries chronologiques est nulle. Donc, pour le modèle $X_t = \alpha_1 \cdot X_{t-1} + Z_t$ si la moyenne est non-nulle il s'exprime comme $x_t - \mu = \alpha_1(x_{t-1} - \mu) + z_t$ ou $x_t = \alpha_0 + \alpha_1 x_{t-1} + z_t$ avec $\alpha_0 = (1-\alpha_1)\mu$, puis $\hat{x}_{t+1} = \mu + \alpha_1(x_t - \mu)$.

White noise

Contexte

Un processus de « *discrete white noise* » est une séquence de variable aléatoire iid idéale pour modéliser la série chronologique d'erreurs résiduelles. On s'attend à ce que cette série ***purement aléatoire*** soit horizontale sans tendance et qu'elle ait une variance constante sans de patrons saisonniers discernables.

Notation

$\{W_t\}$ Processus de « *white noise* » discret.

$\hat{w}_{n+l}$ prévision « *l-steps* » en avance basé sur la valeur observée w_n.

s_W^2 Erreur quadratique moyenne pour une série chronologique « *white noise* ».

Processus de « *white noise* »

Le processus $\{W_t\}$ est un processus de « *white noise* » discret si les variables $W_1, \ldots, W_n$ sont iid de moyenne 0. Il s'ensuit que les propriétés de deuxième ordre sont $\mathrm{E}[W_t] = 0$, $\mathrm{Var}(W_t) = \sigma_W^2$ et, $\forall k \neq 0$, $\gamma_k = 0$.

La prévision au temps n de la valeur au temps $n+l$ est $\hat{w}_{n+l} = 0$ puisque la valeur prédite sera toujours nulle !

Note Si les variables $W_1, \ldots, W_n$ suivent une distribution normale, W_t est un processus de « *white noise* » ***gaussien***.

Marche aléatoire

Contexte

La série d'erreurs résiduelles, telle que mentionnée dans la section sur l'*Autocorrélation*, est composée de termes qui ont tendance d'être dépendants. Une façon de traiter la série est avec une marche aléatoire définie comme des sommes partielles de processus de « *white noise* ».

Notation

$\{X_t\}$ Processus de marche aléatoire.

$\hat{x}_{n+l}$ prévision « *l-steps* » en avance basé sur la valeur observée x_n.

Marche aléatoire

Une marche aléatoire est un processus stochastique dont les termes $X_1, \ldots, X_n$ s'expriment comment la somme de termes d'une série « *white noise* » : $X_t = \sum_{i=1}^{t} W_i$.

La moyenne est également nulle pour le processus ($\mathrm{E}[W_t] = 0$). Il s'ensuit que la variance et l'autocovariance sont équivalentes. Cependant, ces derniers sont proportionnelles au temps écoulé : $\mathrm{Var}(X_t) = \gamma_k(t) = t\sigma_W^2$. Donc, la série chronologique comporte une moyenne stationnaire, mais pas une variance ni autocovariance stationnaire. On déduit que $\rho_k(t) = \frac{t\sigma_W^2}{\sqrt{t\sigma_W^2 + (t+k)\sigma_W^2}} = \frac{1}{\sqrt{1+k/t}}$.

Pour effectuer des prévisions, on récrit que $X_{n+l} = x_n + \sum_{i=1}^{l} W_{n+i}$. Puis, puisque $\mathrm{E}[W_t] = 0$ pour tout t, $\hat{x}_{n+l} = x_n$.

Erreur type prédite

L'erreur type prédite (« *forecast standard error* ») correspond à l'erreur type estimée de la différence entre l'observation au temps $n+l$ x_{n+l} et sa prévision $\hat{x}_{n+l}$. On trouve que $se(\hat{x}_{n+l}) = \sqrt{\widehat{\mathrm{Var}}\left(\sum_{i=1}^{l} W_{n+i}\right)} = s_W\sqrt{l}$.

Marche aléatoire avec dérive

Contexte

Si on trouve que la série d'erreurs résiduelles semble avoir un tendance dans le temps et qu'elle augmente ou décroît, on peut en tenir compte avec un paramètre de ***dérive***.

Notation

δ Paramètre de dérive (« *drift* »).

Marche aléatoire avec dérive

Une marche aléatoire avec dérive se définie comme $X_t = X_{t-1} + \delta + W_t$. Donc, on peut interpréter le paramètre de dérive δ comme la différence moyenne entre 2 termes consécutifs de la série chronologique. Il s'ensuit que la moyenne est non-nulle et que $\mathrm{E}[X_t] = t\delta$. Ni la variance ou l'autocorrélation est affectée.

Pour effectuer des prévisions, on récrit que $X_{n+l} = x_n + l\delta + \sum_{i=1}^{l} W_{n+i}$. Puis, en tenant en compte le paramètre de dérive δ : $\hat{x}_{n+l} = x_n + l\delta$. L'erreur type prédite quant à elle ne change pas non plus.

Propriétés

Opérateurs

Notation

∇ Symbole de l'opérateur de différenciation.

$\boldsymbol{B}$ Symbole de l'opérateur de décalage.

Opérateur de différenciation

Contexte

Une façon de traiter une série chronologique comme une marche aléatoire est de calculer les différences des observations pour obtenir une série « *white noise* ». C'est-à-dire, $W_t = X_t - X_{t-1}$. On définit l'opérateur $\nabla X_t = X_t - X_{t-1}$ pour généraliser à des différences d'ordre supérieure.

On a que $\nabla X_t = X_t - X_{t-1}$, puis $\nabla^2 X_t = \nabla(\nabla X_t) = \nabla X_t - \nabla X_{t-1} = X_t - 2X_{t-1} + X_{t-2}$, etc.

Opérateur de décalage (« *backward shift* »)

L'opérateur de décalage « *shift* » les observations d'un espace. Par exemple, on a que $\boldsymbol{B}X_t = X_{t-1}$, puis $\boldsymbol{B}^n X_t = X_{t-n}$. Également, $\nabla^n = (1 - \boldsymbol{B})^n$.

Équations caractéristiques

Équation caractéristique

Contexte

Polynôme écrit en fonction de $\boldsymbol{B}$ servant à déterminer si une série chronologique possède quelques propriétés.

Pour les paramètres α_i et β_j pour $i = 1, \ldots, p$ et $j = 1, \ldots, q$, on a l'équation caractéristique suivante :
$X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{1-p} + W_t + \beta_1 W_{t-1} + \cdots + \beta_q W_{t-q}$

Cette équation correspond aux *Modèles ARMA*, mais elle sert à déduire des propriétés pour les *Modèles autorégressifs* et les *Modèles de moyenne mobile*

qui sont des simplifications !

L'utilité devient plus claire lorsque l'on récrit l'équation comme :
$(1 - \alpha_1 \boldsymbol{B} - \cdots - \alpha_p \boldsymbol{B}^p) X_t = (1 + \beta_1 \boldsymbol{B} + \cdots + \beta_q \boldsymbol{B}^q) W_t$
$\Rightarrow \theta_p(\boldsymbol{B}) X_t = \phi_q(\boldsymbol{B}) W_t$

Propriétés des modèles autorégressifs et de moyenne mobile

Le modèle est :

stationnaire si aucune des racines de l'équation $\theta_p(\boldsymbol{B}) = 0$ est ≤ 1 en valeur absolue.

inversible si aucune des racines de l'équation $\phi_q(\boldsymbol{B}) = 0$ est ≤ 1 en valeur absolue.

composé de paramètres redondants si $\theta_p(\boldsymbol{B})$ et $\phi_q(\boldsymbol{B})$ ont un facteur en commun.

Modèles autorégressifs

Contexte

La modèle fondé sur la marche aléatoire a soulevé le concept qu'une observation soit dépendante d'une, ou plusieurs, observations précédentes. Le modèle autorégressif d'ordre p généralise ce concept avec un modèle qui est fonction des p observations précédentes.

Le nom « autorégressif » provient du fait qu'on effectue la régression d'une observation avec les autres termes de la série chronologique.

Modèle autorégressif d'ordre p $AR(p)$

Le modèle $AR(p)$ se définit comme $X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + W_t$ ou, de façon plus générale, $W_t = \theta_p(\boldsymbol{B}) X_t$.

Modèle d'ordre 1

Le modèle $AR(1)$ peut être vu comme une généralisation du processus de « *white noise* » et de la marche aléatoire où si :

$\alpha = 0$ le modèle est un processus de « *white noise* » avec $X_t = W_t$.

$\alpha = 1$ le modèle est une marche aléatoire avec $X_t = X_{t-1} + W_t$.

$\alpha \in (-1, 1)$ le modèle est stationnaire.

Puis, si le modèle $AR(1)$ est stationnaire, $\mathrm{E}[X_t] = 0$, $\mathrm{Var}(X_t) = \frac{\sigma^2_W}{1-\alpha^2}$, $\gamma_k = \frac{\alpha^k \sigma^2_W}{1-\alpha^2}$ et $\rho_k = \alpha^k$.

Contexte

Le cas où $p = 1$ est le cas plus simple du modèle autorégressif permettant de déduire les mesures et les critères de stationnarité. De façon plus générale, la stationnarité se détermine avec les racines de $\theta_p(\boldsymbol{B})$ et les mesures (espérance, variance, etc.) sont calculées de façon récursive.

Pour effectuer des prévisions, on utilise l'équation du modèle de façon récursive. De plus, on trouve que $se(\hat{x}_{n+l}) = \sqrt{\mathrm{Var}(X_{n+l} - \hat{X}_{n+l})}$.

Modèles de moyenne mobile

Contexte

Nous généralisons l'idée de centrer les observations avec le modèle de moyenne mobile applicable sur la série des erreurs résiduelles.

Modèle de moyenne mobile d'ordre q ($MA(q)$)

Le modèle $MA(q)$ se définit comme la combinaison linéaire du terme aléatoire (de « *white noise* ») au temps t en plus des q derniers termes les plus récents. L'équation est $X_t = W_t + \beta_1 W_{t-1} + \cdots + \beta_q W_{t-q}$ ou, de façon plus générale, $X_t = \phi_q(\boldsymbol{B}) X_t$.

Comme pour les autres modèles, $\mathrm{E}[X_t] = 0$. Cependant, puisque les termes d'erreurs sont indépendants, on peut déduire une équation explicite pour la variance $\mathrm{Var}(X_t) = \sigma_W^2 \sum_{i=0}^{q} \beta_i^2$ où $\beta_0 = 1$. Il s'ensuit que :

$$\gamma_k = \sigma_W^2 \sum_{i=0}^{q-k} \beta_i \beta_{i+k}, \quad 0 \leq k \leq q$$

$$\rho_k = \begin{cases} 1, & k = 0 \\ \frac{\sum_{i=0}^{q-k} \beta_i \beta_{i+k}}{\sum_{i=0}^{q} \beta_i^2}, & 1 \leq k \leq q \\ 0, & k > q \end{cases}$$

Note Bien que d'habitude les prévisions des termes de « *white noise* » sont nuls, nous pouvons utiliser les termes observés de la série d'erreurs résiduelles. Il s'ensuit que pour une série de longueur n, $\hat{x}_{n+q} = b_q w_n$.

Contexte

On déduit des équations caractéristiques que tous les modèles de moyenne mobile sont ***stationnaires*** et tous les modèles autorégressifs sont ***inversibles***.

Puis, avec les deux séries géométriques suivantes pour $x \in (-1, 1)$:

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \qquad \sum_{k=0}^{\infty} (-x)^k = \frac{1}{1+x}$$

On déduit que si un modèle $MA(q)$ est inversible, alors il peut s'exprimer comme un modèle $AR(\infty)$ stationnaire. Puis, si un modèle $AR(p)$ est stationnaire, alors il peut s'exprimer comme un modèle $MA(\infty)$ inversible.

Par exemple, soit un modèle $AR(2)$ stationnaire avec $X_t = \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + W_t$,

$$X_t(1 - \alpha_1 \boldsymbol{B} - \alpha_2 \boldsymbol{B}^2) = W_t$$

$$X_t = \frac{1}{1 - (\alpha_1 \boldsymbol{B} + \alpha_2 \boldsymbol{B}^2)} W_t$$

$$X_t = [1 + (\alpha_1 \boldsymbol{B} + \alpha_2 \boldsymbol{B}^2) + (\alpha_1 \boldsymbol{B} + \alpha_2 \boldsymbol{B}^2)^2 + \ldots] W_t$$

$$X_t = [1 + \alpha_1 \boldsymbol{B} + (\alpha_2 + \alpha_1^2) \boldsymbol{B}^2 + 2\alpha_1 \alpha_2 \boldsymbol{B}^3 + \alpha_2^2 \boldsymbol{B}^4 + \ldots] W_t$$

$$X_t = W_t + \alpha_1 W_{t-1} + (\alpha_2 + \alpha_1^2) W_{t-2} + \ldots$$

Modèles ARMA

Contexte

On peut combiner le modèle de moyenne mobile avec le modèle autorégressif pour obtenir un modèle autorégressif de moyenne mobile $ARMA(p,q)$.

Modèle autorégressif de moyenne mobile $ARMA(p,q)$

Le modèle $ARMA(p,q)$ se définit comme $X_t = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \beta_1 W_{t-1} + \cdots + \beta_q W_{t-q} + W_t$, ou de façon plus générale, $\theta_p(\boldsymbol{B})X_t = \phi_q(\boldsymbol{B})W_t$.

Il s'ensuit que un modèle $ARMA(p,0) = AR(p)$ et $ARMA(0,q) = MA(q)$.

Modèle d'ordre 1

Un modèle $ARMA(1,1)$ stationnaire avec $X_t = \alpha X_{t-1} + W_t + \beta W_{t-1}$ a une moyenne nulle et $\text{Var}(X_t) = \sigma_W^2 \left(\frac{1 + 2\alpha\beta + \beta^2}{1 - \alpha^2} \right)$. Également, pour $k > 0$, $\gamma_k = \sigma_W^2 (\alpha + \beta)\alpha^{k-1} \left(\frac{1+\alpha\beta}{1-\alpha^2} \right)$ et $\rho_k = \frac{\alpha^{k-1}(\alpha+\beta)(1+\alpha\beta)}{1+2\alpha\beta+\beta^2}$. Finalement, pour $k \geq 2$ alors $\rho_k = \alpha\rho_{k-1}$.

Modèles ARIMA

Contexte

Nous pouvons également tenir en compte la saisonnalité et les tendances (en différenciant) pour obtenir un modèle autorégressif de moyenne mobile intégré $ARIMA(p,d,q)$. Pour obtenir un modèle stationnaire, nous avons utilisé la différenciation. Le modèle ARIMA applique ceci avec l'opérateur de différenciation.

Une série chronologique est intégrée d'ordre d, $I(d)$, si la d^e différence de $\{X_t\}$ est une série « *white noise* » : si $\nabla^d X_t = W_t$.

Également, nous pouvons tenir en compte les effets saisonniers en incorporant la différenciation saisonnière.

Modèle autorégressif de moyenne mobile intégré $ARIMA(p,d,q)$

Le modèle $ARIMA(p,d,q)$ se définit comme $\theta_p(\boldsymbol{B})(1-\boldsymbol{B})^d X_t = \phi_q(\boldsymbol{B})W_t$.

Modèle saisonnier autorégressif de moyenne mobile intégré $SARIMA(p,d,q)(P,D,Q)_g$

Le modèle $SARIMA(p,d,q)(P,D,Q)_g$ se définit comme $\Theta_P(\boldsymbol{B}^g)\theta_p(\boldsymbol{B})(1-\boldsymbol{B}^g)^D(1-\boldsymbol{B})^d X_t = \Phi_Q(\boldsymbol{B}^g)\phi_q(\boldsymbol{B})W_t$.

Régression avec des séries chronologiques

Contexte

On peut modéliser les composantes déterministes (la tendance et la variation saisonnière) d'une série chronologique avec une régression temporelle. Habituellement, on peut trouver des explications vraisemblables pour les patrons d'une série chronologique et donc il s'ensuit qu'une régression peut estimer la tendance et la variation saisonnière en fonction du temps.

De façon générale, la différence entre une régression temporelle et une régression typique est que les erreurs résiduelles sont autocorrélées. S'il y a une ***autocorrélation positive*** des résidus, l'***erreur type*** des paramètres sera ***sous-estimée*** et donc l'importance de la tendance déterministe surestimée.

Modèles de régression avec tendance

Notation

$w_{t,j}$ Valeur de la j^e variable explicative au temps t.
M_t Tendance.

Modèle de régression linéaire

Le modèle linéaire pour la tendance de la série chronologique $\{X_t\}$ est $X_t = \beta_0 + \beta_1 w_{t,1} + \cdots + \beta_p w_{t,p} + Z_t$. Le modèle est linéaire puisque l'on somme les paramètres. Cependant, on peut modéliser des fonctions de variables explicatives comme $w_{t,j}$. Par exemple, on peut avoir $w_{t,j} = t^j$.

Puisque le modèle est fonction du temps, il s'ensuit qu'il n'est pas stationnaire. Également, la prévision sera fonction du temps avec $\hat{x}_{n+l} = b_0 + b_1(n+l) + \hat{\mathrm{E}}[Z_{n+l}]$.

Les termes d'erreur Z_t ont une moyenne nulle, mais ne sont pas nécessairement d'un processus de « *white noise* ». Si les variables aléatoires sont dépendantes, ont obtient que $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}\left[1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)\rho_k\right]$.

La méthode des moindres carrés ordinaire ne considère pas que les résidus peuvent être autocorrélés. Il faut donc utiliser une autre méthode : la ***méthode des moindres carrés généralisée*** (« *generalized least squares (GLS)* »). Cette méthode trouve les paramètres estimés qui maximisent la vraisemblance en considérant l'autocorrélation de l'ensemble de données.

Le besoin pour la méthode de GLS peut se déterminer à partir d'un corrélogramme. S'il y a de l'autocorrélation significative, alors un la GLS peut être appliquée pour ***améliorer les estimations*** des ***erreurs types***

Modèles de régression avec saisonnalité

Modèle indicateur de la saisonnalité

Le modèle indicateur de saisonnalité pour la série chronologique $\{X_t\}$ avec g saisons est $X_t = M_t + S_t + Z_t$ avec $S_t = \beta_j$ pour $j \in \{1, \dots, g\}$. Il s'ensuit qu'on peut réécrire le modèle comme :

$$X_t = \begin{cases} M_t + \beta_1 + Z_t, & t = 1, g+1, 2g+1, \dots \\ M_t + \beta_2 + Z_t, & t = 2, g+2, 2g+2, \dots \\ \vdots, & \vdots \\ M_t + \beta_g + Z_t, & t = g, 2g, 3g, \dots \end{cases}$$
$$= M_t + \beta_{1+(t-1) \mod g} + Z_t$$

Modèle de saisonnalité harmonique

Le modèle permet d'obtenir une courbe plus lisse et continue qu'avec des paramètres différents par saison. On utilise la fonction trigonométrique $h(t) = a\sin(2\pi f t + b)$ où :

a L'amplitude (la valeur maximale de la courbe).

f La fréquence (le nombre de cycles sur l'intervalle $[0, 1]$).

b Changement de phase (« *phase shift* »).

Avec les propriétés des fonctions trigonométriques, puis en posant que $\beta_1 = a\cos(b)$ et $\beta_2 = a\sin(b)$, on réécrit la fonction $h(t)$ comme une fonction linéaire : $h(t) = \beta_1 \sin(2\pi f t) + \beta_2 \cos(2\pi f t)$. Puis, la composante de variation saisonnière est $S_t = \sum_{j=1}^{\lfloor g/2 \rfloor} \beta_{1,j} \sin(2\pi j t/g) + \beta_{2,j} \cos(2\pi j t/g)$ pour le modèle $X_t = M_t + S_t + Z_t$.

Modèles de régression non-linéaires

Contexte

Une transformation qu'on peut appliquer est la transformation logarithmique. Ceci convertit un modèle multiplicatif en un modèle additif avec $\ln X_t = \ln M_t + \ln S_t + \ln Z_t$. La transformation est idéale lorsque le modèle prend de très grandes valeurs ou lorsque la variance n'est pas constante (hétéroscédasticité).

Cependant, si on effectue des *prévisions* sur une transformation de la v.a., on doit tenir en compte l'erreur de la distribution de la transformation. On voit deux facteurs de correction.

Facteur de correction lognormale

Contexte

Pour une série chronologique Z_t qui suit un processus gaussien de « *white noise* », on applique le ***facteur de correction lognormale*** qui correspond à ne pas oublier de prendre l'espérance de l'exponentiel du terme d'erreur. Cela puisque $\hat{X}_{n+l} = e^{\beta_0 + \beta_1 (n+l)} \mathrm{E}[e^{Z_{n+l}}]$ et donc la moyenne des résidus n'est pas nulle !

Le ***facteur de correction lognormale*** est $\mathrm{E}[e^{Z_{n+l}}] = e^{0+\sigma^2/2} = e^{\sigma^2/2}$.

Facteur de correction empirique

Contexte

Si la distribution des résidus Z_t n'est pas normale, on peut appliquer un ***facteur de correction empirique*** plus général.

Le ***facteur de correction empirique*** est $\mathrm{E}[e^{Z_{n+l}}] = \frac{1}{n} \sum_{t=1}^{n} e^{Z_t}$.