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. 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 !

Première partie

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

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

› 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 $\hat{\sigma}^2$

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

› 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.

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

$\mathrm{B}(\hat{\theta}_n)$ 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** $\mathrm{B}(\hat{\theta}_n)$ 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 $\mathrm{B}(\hat{\theta}_n) = \mathrm{E}[\hat{\theta}_n|\theta] - \theta$, où $\mathrm{E}[\hat{\theta}_n|\theta]$ 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, $\mathrm{B}(\hat{\theta}_n) = 0$, 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, $\lim_{n\to\infty} \mathrm{B}(\hat{\theta}_n) = 0$, 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 $\text{Var}(\hat{\theta}_n) = \text{E}\left[\left(\hat{\theta}_n - \text{E}[\hat{\theta}_n]\right)^2\right]$.

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

$\textbf{MSE}_{\hat{\theta}_n}(\theta)$ 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)* ».

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

$\hat{\theta}_n$ 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 $\forall \varepsilon > 0$,

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

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

$\{Y_n\}$ 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 $\hat{\theta}_n \xrightarrow{P} \theta$.

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 $\boldsymbol{I}_n(\theta) = n\boldsymbol{I}(\theta) = -n\text{E}\left[\frac{\partial^2}{\partial\theta^2} \ln f(x;\theta)\right]$.

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 $\boldsymbol{\theta} = (\theta_1, \theta_2)$

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

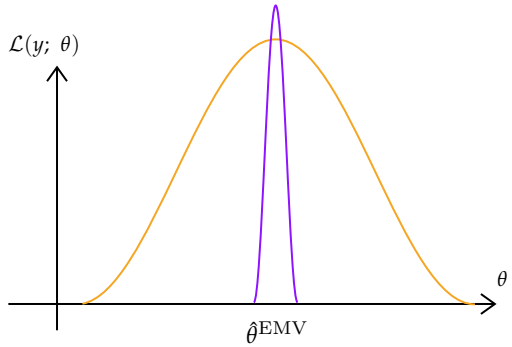
Borne inférieure Cramér-Rao

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



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*.

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 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 minimisé

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

Notation

$\hat{\theta}_L$ **et** $\hat{\theta}_U$ Fonctions de l'échantillon aléatoire $\{X_1, \ldots, X_n\}$ où $\hat{\theta}_L < \hat{\theta}_U$;

$(\hat{\theta}_L, \hat{\theta}_U)$ Intervalle de confiance de $100(1-\alpha)\%$ de θ si $\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) = 1-\alpha$;

› Avec les réalisations, on a un intervalle de nombres réels $(\hat{\theta}_l, \hat{\theta}_u)$.

$(1-\alpha)$ Niveau de confiance de l'intervalle où $\alpha \in (0,1)$.

Le type principal d'estimateur par intervalle est l'**intervalle de confiance** :

Intervalle de confiance

Nous sommes confiants à un niveau de $100(1-\alpha)\%$ que le paramètre inconnu θ est entre $(\hat{\theta}_L, \hat{\theta}_U)$.
De façon équivalente, nous sommes confiants à un seuil de $\alpha\%$ que θ est entre $(\hat{\theta}_L, \hat{\theta}_U)$.

Donc, $\theta \in (\hat{\theta}_L, \hat{\theta}_U)$ et nous pouvons dire que $\Pr(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \geq (1-\alpha)$ pour tout θ.

Ce qu'il faut bien saisir avec les intervalles de confiance, c'est que soit θ est contenu dans l'intervalle $(\hat{\theta}_l, \hat{\theta}_u)$ ou il ne l'est pas.
On peut conceptualiser les intervalles comme une distribution binomiale avec probabilité de succès de $(1-\alpha)$. Si l'on effectue M essais indépendants, on s'attend à ce que $(1-\alpha)M$ intervalles de confiance contiennent θ. Donc on se sent confiant à $(1-\alpha)\%$ que la vraie valeur de θ est contenue dans l'intervalle observé $(\hat{\theta}_l, \hat{\theta}_u)$.

Efficacité des intervalles de confiance Typiquement, la largeur de l'intervalle $(\hat{\theta}_L, \hat{\theta}_U)$ augmente si on augmente le niveau de confiance $(1-\alpha)$. 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 $(1-\alpha)$, l'intervalle avec la plus petite largeur est *plus efficace* que l'autre.

Statistiques

Rappel : Loi du khi-carré

Soit un échantillon aléatoire $(X_1, X_2, \ldots, X_n)$ de variables aléatoires normales de moyenne μ et variance σ^2.

Soit $Q = \sum_{i=1}^{n} (X_i - \mu)^2$.

Alors, $Q/\sigma^2 \sim \chi^2_{(n)}$.

Rappel : Loi de Student

Soit les variables aléatoires indépendantes :

› $Z \sim \mathcal{N}(0,1)$.

› $W \sim \chi^2_{(n)}$.

Alors, $T = \frac{Z}{\sqrt{W/n}} \sim t_{(n)}$.

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

Rappel : Loi de Fisher-Snedecor (F)

Soit les variables aléatoires indépendantes :

› $W_1 \sim \chi^2_{(\nu_1)}$.

› $W_2 \sim \chi^2_{(\nu_2)}$.

Alors, $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim \mathcal{F}_{(\nu_1,\nu_2)}$.

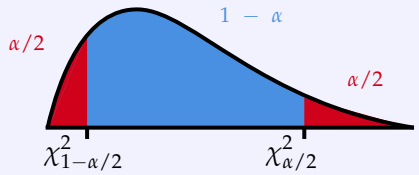
On peut relier la loi de Student et la loi F : $T^2 = \frac{Z^2}{W/n} \sim \mathcal{F}_{(1,n)}$ puisque $Z^2 \sim \chi^2_{(1)}$ où $Z \sim \mathcal{N}(0,1)$.

Intervalles de confiance

Intervalle de confiance sur la variance

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec σ^2 inconnue, $\Pr\left(\chi^2_{1-\alpha/2} \leq \frac{(n-1)S_n^2}{\sigma^2} \leq \chi^2_{\alpha/2}\right) = (1-\alpha)$.

Graphiquement :



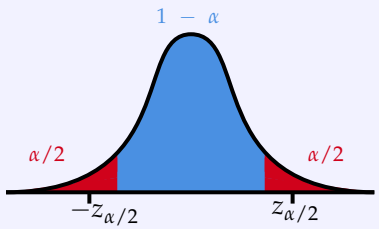
Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

$$\sigma^2 \in \left[\frac{(n-1)S_n^2}{\chi^2_{\alpha/2}}, \frac{(n-1)S_n^2}{\chi^2_{1-\alpha/2}}\right]$$

Intervalle de confiance sur la moyenne (σ^2 connue)

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec μ inconnu et σ^2 connue, $\Pr\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = (1-\alpha)$.

Graphiquement :



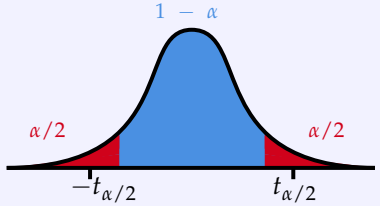
Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

Intervalle de confiance sur la moyenne (σ^2 inconnue)

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution normale avec σ^2 inconnue, $\Pr\left(-t_{\alpha/2,n-1} \leq \frac{\bar{X}-\mu}{S_n/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = (1-\alpha)$.

Graphiquement :



Nous sommes donc confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}, \bar{X} + t_{\alpha/2,n-1}\frac{S_n}{\sqrt{n}}\right].$$

Intervalle de confiance *approximatif* sur la moyenne

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution avec moyenne μ et une variance inconnue.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$\mu \in \left[\bar{X} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right].$$

Intervalle de confiance *approximatif* sur la proportion

Pour l'échantillon aléatoire $\{X_1, X_2, \ldots, X_n\}$ issu d'une distribution Bernoulli de paramètre p.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$p \in \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

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

Intervalle de confiance pour une différence de moyennes

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions normales de moyennes μ_1 et μ_2 et variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ inconnues.
Nous sommes confiants à un niveau de $100(1-\alpha)\%$ que :

$$(\mu_1 - \mu_2) \in \left[\bar{x}_n - \bar{y}_m \pm t_{\alpha/2, n+m-2} S_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right].$$

Intervalle de confiance *approximatif* pour une différence de moyennes

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions normales de moyennes μ_1 et μ_2 et variances σ_1^2 et σ_2^2 inconnues.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$(\mu_1 - \mu_2) \in \left[\bar{X}_n - \bar{Y}_m \pm z_{\alpha/2} \sqrt{\frac{S_n^2}{n} + \frac{S_m^2}{m}}\right].$$

Intervalle de confiance *approximatif* pour une différence de proportions

Pour les échantillons aléatoires $\{X_1, X_2, \ldots, X_n\}$ et $\{Y_1, Y_2, \ldots, Y_m\}$ issus de distributions Bernoulli de paramètres p_1 et p_2.

Pour n très grand, nous sommes *approximativement* confiants à un niveau de $100(1-\alpha)\%$ que :

$$(p_1 - p_2) \in \left[\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}\right].$$

Tests d'hypothèses

Introduction

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.

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

Notation

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

$\mathbf{H}_0$ Hypothèse nulle ;

› Représente généralement le statu quo jusqu'à preuve contraire.

$\mathbf{H}_1$ Hypothèse alternative.

› Représente généralement un changement du statu quo.

Test d'hypothèse

On spécifie une hypothèse nulle et, par conséquent, une hypothèse alternative :

$$\mathrm{H}_0 : \theta \in \Theta_0 \qquad \text{vs} \qquad \mathrm{H}_1 : \theta \in \Theta_1$$

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

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 para-

mètre q, il pourrait être n'importe quel chiffre sauf 0.50.

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 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 « extrême » pour être le statu quo et donc que l'on rejette.
On rejette H_0 si $\{X_1, \ldots, X_n\} \in \mathcal{C}$ puis on conserve 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 $\mathcal{C} = (0.44, 1]$.

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 ?? 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

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

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

Certitude du test

Lorsque nous voulons quantifier le degré auquel nous sommes confiants du test, nous utilisons la valeur p.
La valeur p 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.

- Ceci 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, dans notre cas, 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, représenté en rouge.

- Si la valeur p est petite, ceci indique que d'autres distributions pourraient potentiellement mieux s'ajuster aux données puisque l'événement est très rare ;
- Si la valeur p est grande, ceci indique que l'événement est très courant et que la distribution semble être bien ajustée.

Il y a plusieurs termes semblables qui peuvent devenir mélangeants.

Terminologie

p La **valeur** p du test.

- On peut la définir comme la probabilité d'un événement tout aussi (ou plus) rare sous l'hypothèse nulle ;
- On peut la définir comme 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 ;
- On peut la définir comme le **seuil de signifiance** ; c'est-à-dire, la probabilité de rejeter H_0 alors qu'elle est vraie ;
- Elle correspond donc également à la **probabilité d'une erreur de type I**.

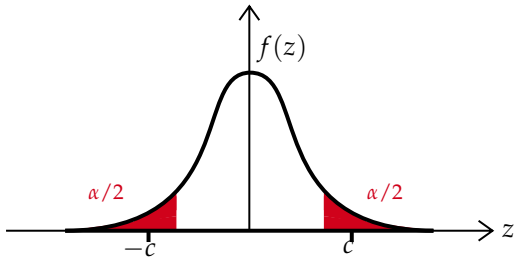
α Dénote habituellement le **seuil de signifiance** ou la **taille** du test.

- Même idée qu'avec les intervalles de confiance ;
- On peut parfois aussi utiliser α pour dénoter la valeur de p qui détermine si on rejette ou pas un test ;
- En anglais, « *threshold for significance* ».

Formellement, on définit $\alpha = \max_{\theta \in \Theta_0} \Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta\}$.
C'est-à-dire :

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

De façon générale, on représente la région critique pour un test bilatéral comme l'aire en rouge du graphique de la distribution normale ci-dessous :



Puissance d'un test

La puissance d'un test

La probabilité de *correctement* rejeter l'hypothèse nulle.

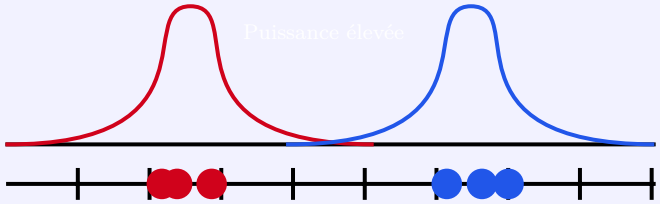
Une analyse de la puissance détermine le nombre d'observations qu'il faut afin d'avoir une probabilité élevée de correctement rejeter l'hypothèse nulle.

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

La forme de la distribution

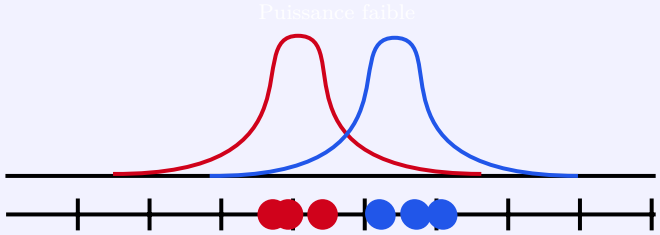
Si les deux distributions sont :

› 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** :

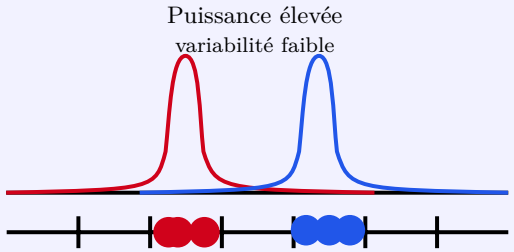


– La probabilité **d'incorrectement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;
– On peut aussi dire qu'il y a une forte probabilité **d'incorrectement** obtenir une faible valeur p ;
– Cependant, la puissance peut être augmentée avec plus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

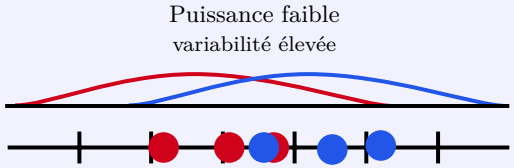
La variabilité des données

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 $d = \frac{\bar{x}-\bar{y}}{s_p^2}$.

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.

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

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

- C'est une fonction de θ ;
- Idéalement, si l'hypothèse nulle est :

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

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

Si, par exemple, on rejette l'hypothèse nulle, on pourrait tracer la fonction de puissance pour toutes les valeurs possibles de l'ensemble Θ_1.

Tests optimaux

Notation

δ (Procédure de) test ;

$\alpha(\delta)$ Probabilité d'une erreur de type I 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 pour un test δ ;
- $\beta(\delta) = \Pr\left\{(X_1, \ldots, X_n) \in \mathcal{C}^{\complement}; \theta \in \Theta_1\right\} = 1 - \gamma(\theta_1)$.

Pour mettre en contexte cette notation, revoici le tableau des types d'erreurs pour un test δ :

	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 $(X_1, \ldots, X_n) \in \mathcal{C}^{\complement}$	$1 - \alpha(\delta)$	$\beta(\delta)$

- *En théorie*, on minimise la probabilité d'une erreur de type I **et** de II ;
- *En réalité*, il y a un compromis et on ne pourra pas avoir des très petites probabilités pour les deux ;
- Le contexte va déterminer ce qu'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 qu'il n'a pas le cancer alors qu'il l'a (erreur de type II) que de dire qu'il a le cancer alors qu'il ne l'a pas (erreur de type I) ;
  - Dans ce contexte, on souhaiterait minimiser l'erreur de type II $\beta(\delta)$ plus que celle 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, k.
2. Trouve parmi tous les sous-ensembles de taille k celui qui minimise l'erreur de type II.

Tests optimaux

Soit un test δ^* avec les hypothèses simples :

$H_0 : \theta = \theta_0$

$H_1 : \theta = \theta_1$

› Par exemple, on pourrait avoir une distribution Bernoulli et poser $H_0 : p = 0.4$ v.s. $H_1 : p = 0.6$.

La procédure pour trouver la région critique $\mathcal{C}$ optimale de taille α du test δ^* est la suivante :

1. On trouve une région critique (alias, un sous-ensemble de $\mathcal{S}$) $\mathcal{C}$ tel que la probabilité $\alpha(\delta^*)$ d'une erreur de type I est de α.
   - › C'est-à-dire, $\alpha(\delta^*) = Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta = \theta_0\} = \alpha$.
   - › Cependant, ce critère n'identifie par un sous-ensemble unique ;
   - › Il y a une multitude de sous-ensembles $\mathcal{A}$ de $\mathcal{S}$ dont la probabilité que l'échantillon aléatoire y soit contenu (sous l'hypothèse nulle) est aussi α ;
   - › C'est-à-dire, $Pr\{(X_1, \ldots, X_n) \in \mathcal{A}; \theta = \theta_0\} = \alpha$.
2. On pose que la probabilité que l'échantillon aléatoire soit dans la région critique $\mathcal{C}$ (sous l'hypothèse alternative) est supérieure à la probabilité que l'échantillon aléatoire soit contenu dans tout autre sous-ensemble $\mathcal{A}$.
   - › C'est-à-dire, $Pr\{(X_1, \ldots, X_n) \in \mathcal{C}; \theta = \theta_1\} \geq Pr\{(X_1, \ldots, X_n) \in \mathcal{A}; \theta = \theta_1\}$.

Avec ces deux critères, on trouve ***la*** région critique $\mathcal{C}$ de taille α **optimale** pour tester les hypothèses simples.

En bref, on pose fixe à un seuil α la fonction de puissance posant que le vrai paramètre $\theta = \theta_0$ puis on trouve la région critique qui maximise la puissance posant que le vrai paramètre $\theta = \theta_1$.

Exemple avec une distribution binomiale

Soit :

› La variable aléatoire $X \sim Binom(n = 3, p = \theta)$.

– Alors, $\mathcal{S} = \{x : x = 0, 1, 2, 3\}$.

› Les hypothèses :

$H_0 : \theta = 0.50$

$H_1 : \theta = 0.75$

› Le seuil de signifiance $\alpha = 0.125$.

› Les sous-ensembles $\mathcal{A}_1 = \{x : x = 0\}$ et $\mathcal{A}_2 = \{x : x = 3\}$ de $\mathcal{S}$.

Alors, $\Pr(X \in \mathcal{A}_1; \theta = 0.50) = \Pr(X \in \mathcal{A}_2; \theta = 0.50) = 0.125$ et il n'y a pas d'autres sous-ensembles de $\mathcal{S}$ avec la même taille de 0.125.
Il s'ensuit que soit $\mathcal{A}_1$ ou $\mathcal{A}_2$ est la région critique $\mathcal{C}$ optimale de taille α pour tester H_0 contre H_1.

On trouve que $\Pr(X \in \mathcal{A}_1; \theta = 0.75) = 0.015625$ alors que $\Pr(X \in \mathcal{A}_2; \theta = 0.75) = 0.421875$.

› Dans le premier cas :

$$\underbrace{\Pr(X \in \mathcal{A}_1; \theta = 0.75) = 0.015625}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} < \underbrace{\Pr(X \in \mathcal{A}_1; \theta = 0.50) = 0.125}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)}$$

› Dans le deuxième cas :

$$\underbrace{\Pr(X \in \mathcal{A}_2; \theta = 0.75) = 0.421875}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est faux } (\theta = 0.75)} > \underbrace{\Pr(X \in \mathcal{A}_2; \theta = 0.50) = 0.125}_{\text{rejeter } H_0 \text{ alors que } H_0 \text{ est vraie } (\theta = 0.50)}$$

› Le premier sous-ensemble $\mathcal{A}_1$ n'est pas désirable, car on serait plus probable de incorrectement rejeter H_0 lorsqu'elle est vraie (erreur de type I) que de correctement la rejeter lorsqu'elle est fausse !

› Alors, on choisit $\mathcal{C} = \mathcal{A}_2 = \{x : x = 3\}$.

D'ailleurs, la région est choisie en incluant dans $\mathcal{C}$ les points x pour lesquels $f(x; \theta = 0.50)$ est petite par rapport à $f(x; \theta = 0.75)$.

› On peut d'ailleurs observer que le ratio $\frac{f(x;\theta=0.50)}{f(x;\theta=0.75)}$ évalué à $x = 5$ est un minimum.

On peut utiliser ce ratio comme outil pour identifier la région critique $\mathcal{C}$ optimale pour un seuil fixe de α.

Cas d'hypothèses simples

Théorème de Neymann-Pearson

Soit un test δ^* avec les hypothèses simple :

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1$$

Soit une constante $k > 0$ et le sous-ensemble $\mathcal{C} \in \mathcal{S}$ tel que :

1. $\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} \leq k$ pour tout $x \in \mathcal{C}$.
2. $\frac{\mathcal{L}(\theta_0;\boldsymbol{x})}{\mathcal{L}(\theta_1;\boldsymbol{x})} \geq k$ pour tout $x \in \mathcal{C}^{\complement}$.
   - En récrivant les équations comme $\mathcal{L}(\theta_1;\boldsymbol{x}) \leq (\geq)\ k\mathcal{L}(\theta_0;\boldsymbol{x})$ on peut l'interpréter comme qu'il doit être plus vraisemblable que $\theta = \theta_0(\theta_1)$ que $\theta_1(\theta_0)$ lorsque $x \in \mathcal{C}^{\complement}$ et que l'on rejette (accepte) H_0.
3. $\alpha = \Pr\{(X_1,\ldots,X_n) \in \mathcal{C};\theta_1\} = \alpha(\delta^*)$.

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

Test non biaisé

Soit les mêmes hypothèses que dans la définition du théorème de Neymann-Pearson.

Un test δ est non biaisé si sa puissance est toujours d'au moins α ; c'est-à-dire, $\Pr\{(X_1,\ldots,X_n) \in \mathcal{C};\theta\} \geq \alpha$.

Le meilleur test obtenu par le théorème de Neymann-Pearson est non biaisé.

Exemple avec une distribution normale

Soit :

- L'échantillon aléatoire $\boldsymbol{X} = (X_1,\ldots,X_n)$ d'une distribution normale $\mathcal{N}(\mu = \theta, \sigma^2 = 0)$.
  - Alors, $\mathcal{S} = \{x : x \in \mathbb{R}\}$.
- Les hypothèses :

$$H_0 : \theta = 0$$
$$H_1 : \theta = 1$$

On a :

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

Alors, la région critique $\mathcal{C}$ optimale est composée des points $(x_1, x_2, \ldots, x_n)$ tel que :

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

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, puisque $\bar{X} \overset{H_0}{\sim} \mathcal{N}(0, 1/n)$ on peut trouver c avec $\Pr\{\bar{X} \geq c; \theta = \theta_0\} = \alpha$.
Puis, on peut trouver la puissance du test quand H_1 est vraie avec $\Pr\{\bar{X} \geq c; \theta = \theta_1\}$.

Note sur les hypothèses Les hypothèses doivent entièrement spécifier la distribution. Si les hypothèses sont sur les paramètres, elles doivent être des hypothèses simples, mais elles peuvent être sur autre chose.
Par exemple, si on teste $H_0 : f_X(x) = g(x)$ v.s. $H_1 : f_X(x) = h(x)$ alors la vraisemblance sera un ratio de deux distributions différentes.

Cas d'hypothèses composées

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

Exemple avec une distribution normale

Soit :

- Un échantillon aléatoire $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ tiré d'une distribution normale $\mathcal{N}(0, \theta)$;
- Les hypothèses :
  $H_0 : \theta = 1$
  $H_1 : \theta > 1$

Alors, on trouve le ratio :

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

On voit que le ratio décroît alors que $\sum x_i^2$ augmente. Par conséquent, un test uniformément le plus puissance aura une région critique définie par $\sum x_i^2 > k$ avec un k choisi selon le seuil de signifiance.

L'idée est donc de poser un θ_1 fixe pour évaluer la forme du ratio de la vraisemblance. Selon la croissance ou décroissance de la fonction, ainsi que l'hypothèse, on peut établir une région pour laquelle une augmentation du θ_1 maintient la relation.

La région uniformément la plus puissante n'existe pas toujours, mais dans le cas qu'elle existe le théorème de Neymann-Pearson permet de la trouver.

Test du khi carré

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

Test d'adéquation (« *goodness-of-fit test* »)

Soit n répétitions (indépendantes) d'une expérience aléatoire.
On pose :

- L'espace d'échantillon des expériences $\mathcal{A}$ qui représente l'union de k différents ensembles (disjoints) $\mathcal{A} = \{A_1 \cup A_2 \cup \cdots \cup A_k\}$;
- On pose que pour $i = 1, 2, \ldots, k$, $\Pr(A_i) = p_i$ où $p_k = 1 - p_1 - \ldots - p_{k-1}$ et $O_k = n - O_1 - \ldots - O_{k-1}$;
  - p_i représente donc la *probabilité* que le résultat de l'expérience aléatoire fasse partie de l'ensemble A_i ;
  - O_i représente le *nombre d'observations* (la *fréquence*) pour lesquelles le résultat de l'expérience aléatoire fait partie de l'ensemble A_i.
- On pose que la distribution conjointe de $O_1, O_2, \ldots, O_{k-1} \sim MultiNom(n, p_1, \ldots, p_{k-1})$.

Soit le test d'hypothèse avec les nombres spécifiés $p_{1,0}, p_{2,0}, p_{k-1,0}$:
$H_0 : p_1 = p_{1,0}, p_2 = p_{2,0}, \ldots, p_{k-1} = p_{k-1,0}$
où $p_k = p_{k,0} = 1 - p_{1,0} - \ldots - p_{k-1,0}$.

Alors, sous l'hypothèse nulle : $Q = \sum_{i=1}^{k} \frac{(O_i - np_{i,0})^2}{np_{i,0}} \approx \chi^2_{(k-1)}$.

- Il y a seulement $k-1$ degrés de liberté, car on estime seulement $k-1$ paramètres.
  - Le nombre n total d'observations est fixe et on déduit n_k par la somme ;
  - Si on avait à

Tableau de contingence

Dans le cas de données à deux dimensions, alias un tableau de contingence, on définit :

E_{ij} L'espérance du nombre d'observations dans la cellule i, j ;

O_{ij} Le nombre observé d'observations dans la cellule i, j.

- On pose qu'il y a c colonnes au tableau pour r rangées (les rangées sont les différents ensembles) ;
- On peut donc tester si la distribution de la fréquence est identique pour les c colonnes avec $Q = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \approx \chi^2_{(r-1)\cdot(c-1)}$;

› Cette formule est beaucoup plus intuitive visuellement que par formule ; pour la comprendre, faites un exemple.

Test du rapport de vraisemblance

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

En fonction des observations, calculer :

1. Le maximum de vraisemblance sous l'hypothèse nulle ;
2. Le maximum de vraisemblance sous l'hypothèse alternative

La région critique correspond à la région pour laquelle le ratio des vraisemblances est en dessous d'une constante k.

› Si les deux hypothèses sont simples, ceci équivaut à utiliser le théorème de Neymann-Pearson.

Cependant, il peut s'avérer difficile d'isoler une distribution dans le ratio. Pour des grands échantillons, on peut plutôt utiliser la distribution asymptotique.
Soit une hypothèse nulle qui spécifie k paramètre et une hypothèse alternative qui en spécifie seulement l ($l < k$). Alors, la statistique du rapport de vraisemblance $Q = -2\left(\ln(\theta_0) - \ln(\theta_1)\right) \sim \chi^2_{k-l}$.

Cette statistique est vue dans les modèles linéaires généralisés aussi.

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-Neymann afin de prouver qu'une statistique est exhaustive.

Théorème de factorisation de Fisher-Neymann

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 faille 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 section *Famille exponentielle* du chapitre de *Modèles linéaires en actuariat* 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

Soit un échantillon aléatoire de taille n. Nous définissons la k^{e} **statistique d'ordre** $X_{(k)}$ comme étant la k^{e} plus petite valeur d'un échantillon.

› Les crochets sont utilisés pour distinguer la k^{e} statistique d'ordre $X_{(k)}$ de la k^{e} observation X_k.

› La k^{e} statistique d'ordre correspond au $\frac{k}{n+1}^{\text{e}}$ quantile.

Nous sommes habituellement intéressés au minimum $X_{(1)}$ et le maximum $X_{(n)}$:

Minimum

$$X_{(1)} = \min(X_1, \dots, X_n)$$
$$f_{X_{(1)}}(x) = n f_X(x) \big(S_X(x)\big)^{n-1}$$
$$S_{X_{(1)}}(x) = \prod_{i=1}^{n} \Pr(X_i > x)$$

Maximum

$$X_{(n)} = \max(X_1, \dots, X_n)$$
$$f_{X_{(n)}}(x) = n f_X(x) \big(F_X(x)\big)^{n-1}$$
$$F_{X_{(n)}}(x) = \prod_{i=1}^{n} \Pr(X_i \leq x)$$

De façon plus générale, on défini :

k^{e} statistique d'ordre

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

$$F_{X_{(k)}}(x) = \underbrace{\sum_{i=r}^{n} \binom{n}{i} [F_X(x)]^j [1 - F_X(x)]^{n-j}}_{\substack{\text{Probabilité qu'au moins } k \text{ des } n \\ \text{observations } X_k \text{ sont} \leq x}}$$

› On peut observer que $X_{(k)} \sim Beta(\alpha = k, \beta = n - k + 1)$

Nous pouvons également définir quelques autres statistiques d'intérêt :

L'étendue (« *range* »)

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

› 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 $16°C$.
  – 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 sur les statistiques d'ordre

Soit un échantillon de données météorologiques $\{-30°, -24°, -7°, -23°, +5°\}$ (Celsius).
Je suppose que ce sont des températures du 4 février observées lors des dernières années.

› La moyenne arithmétique ($-22.25°C$) m'intéresse, car je peux savoir, en moyenne, ce qu'est la température le 4 février.

› La mi-étendue ($-12.5°C$), tout comme l'étendue ($-35°C$), ne m'intéresse pas puisqu'elle ne prend pas en considération la vraisemblance des différentes températures.

Maintenant, je suppose que ces données sont des températures observées tout au long de l'hiver passé.

› La moyenne arithmétique ne m'intéresse pas puisqu'elle est beaucoup trop biaisée par les températures de cette même journée.

› Cependant, la mi-étendue et l'étendue me donnent maintenant une meilleure idée de la température de l'hiver.

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 d'une seule journée.

Nous pouvons définir 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.

Finalement, on définit la distribution conjointe du minimum et du maximum $\forall x < y$:

Distribution conjointe du maximum et du minimum

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

Graphiques

Le diagramme en boîte (« *boxplot* »)

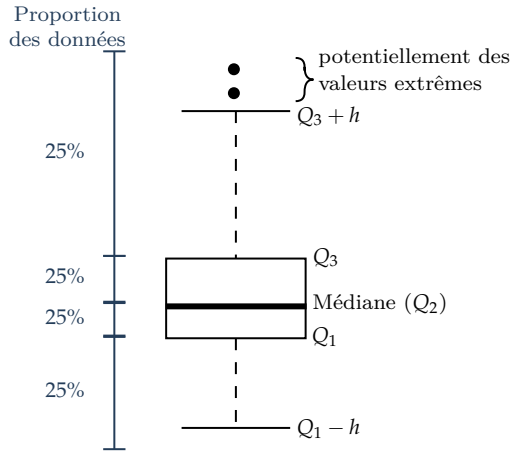
Le diagramme du « *sommaire à cinq chiffres* ».

Sommaire à cinq chiffres

Les cinq statistiques suivantes :

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 :



› La médiane est la ligne contenue dans la boîte.
  – La moitié des données sont au-dessus, et l'autre moitié en dessous, de la ligne.

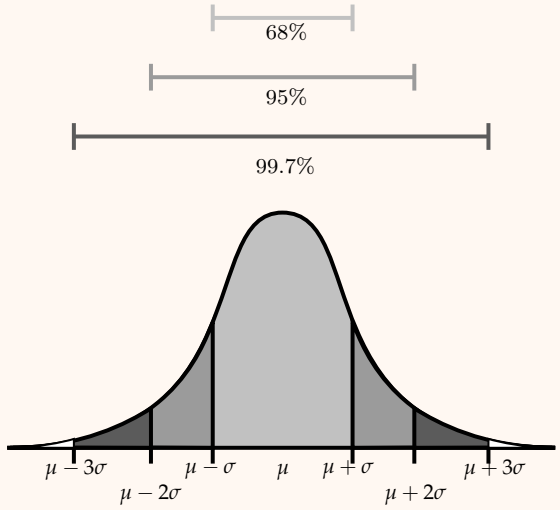
› La boîte est délimitée par le premier et le troisième quartile.
  – Il s'ensuit que la boîte contient la moitié des données.
  – De plus, 25% des données sont contenues entre la borne *supérieure* de la boîte et la médiane avec l'autre 25% qui est contenu entre la borne *inférieure* et la médiane.

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

– Les points qui sont à l'extérieur de ces bornes sont les données potentiellement aberrantes.
– $Q_3 - Q_1$ correspond à l'écart interquartile.
– Plus l'écart est élevé, plus la boîte sera large et, par conséquent, plus les moustaches seront situées loin de la médiane.
– Le 1.5 est basé sur la règle du 68-95-99.7 avec moins de 1% des données à l'extérieur de la borne supérieure.

Règle du 68-95-99.7

Pour une distribution normale, environ 68% des données sont en dedans d'un écart-type de la moyenne, 95% en dedans de 2 et 99.7% en dedans de 3.



En bref, le diagramme en boîte permet d'évaluer comment les données sont distribuées. Cette image de wikipedia résume bien :

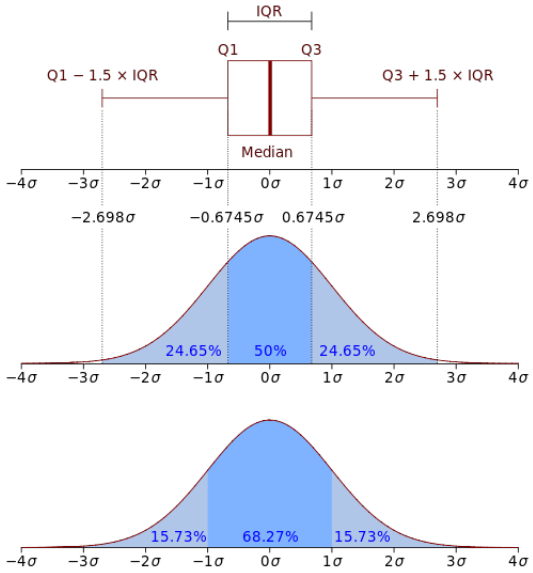
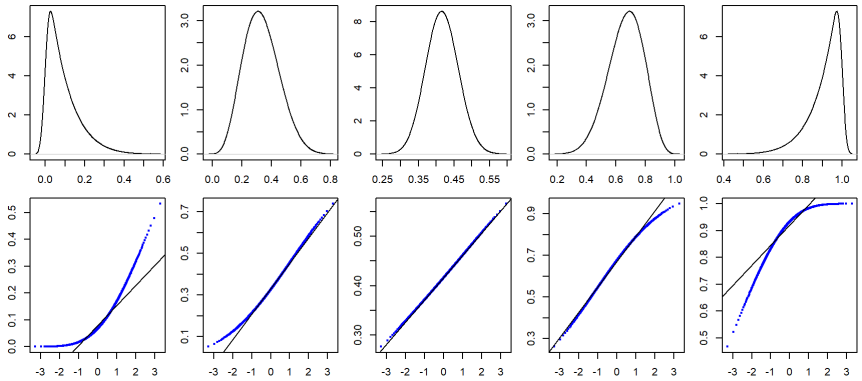


Diagramme quantile-quantile (« *Q-Q plot* »)

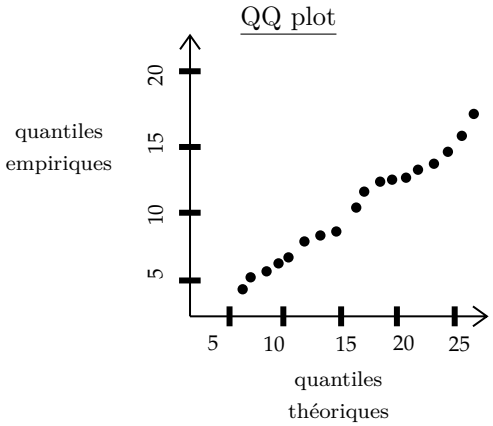
En pratique, on pose souvent que les données suivent une distribution. Un diagramme quantile-quantile permet de comparer les quantiles théoriques de la distribution aux quantiles empiriques des données.

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 :



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 « normaux » varient. Ci-dessous est une image de ce site qui montre quantiles selon la distribution avec une droite pour la normale :

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 Mathématiques actuarielles IARD I 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 :

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

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 :

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

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 *Mathématiques actuarielles IARD I* 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 $\mathrm{E}_{\theta_0}[M(X)] < \infty$ 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 : $\hat{\boldsymbol{\theta}}^{\text{EMV}} \approx \mathcal{N}\left(\boldsymbol{\theta}, \boldsymbol{I}(\boldsymbol{\theta})^{-1}\right)$.

Deuxième partie

Modèles linéaires en actuariat

Apprentissage statistique

Apprentissage statistique

L'apprentissage statistique est l'utilisation de statistiques pour estimer les relations entre des variables « ***explicatives*** » et un résultat (une variable « ***réponse*** »).

Variable réponse

Variable pour laquelle nous voulons effectuer des prévisions.

Variables explicatives

Variables utilisées pour les prévisions de la variable réponse.

Les modèles d'apprentissage statistique ont deux utilités principales :

1. Faire des **prévisions** de la valeur de la variable réponse pour des valeurs spécifiques des variables réponse.
2. Faire de l'**inférence** afin de comprendre quelles variables explicatives sont liées à la variable réponse, et à quel degré.

Il y a une multitude de modèles d'apprentissage statistique différents. Entre autres, ces modèles varient en **flexibilité** ; c'est-à-dire, certains modèles s'*ajustent* mieux aux données.

Par exemple, une régression linéaire correspond à une ligne droite (peu flexible). En réalité, il est peu probable que les données soient situées sur une droite. En contraste, une « *spline* » va passer à travers tous les points (très flexible).

Flexibilité du modèle

Il y a un compromis à faire entre la flexibilité d'un modèle et sa facilité d'interprétation :

- Les modèles *moins flexibles* sont plus facilement interprétables au coût de moins bonnes prévisions.
  - Ils sont généralement mieux pour faire des prévisions.
- Les modèles *plus flexibles* sont plus difficilement interprétables, mais ont l'avantage d'offrir de meilleures prévisions.
  - Ils sont généralement mieux pour faire des prévisions.

En revanche, si un modèle est ***sur ajusté*** aux données alors il pourrait être biaisé envers les données avec lesquelles il est entraîné et offrir des mauvaises prévisions pour des **nouvelles** données.

De façon générale, on sépare l'apprentissage statistique en deux types :

apprentissage supervisé comporte une variable réponse.

apprentissage non supervisé ne comporte pas de variable réponse.

Types de variables explicatives

Les variables explicatives prennent plusieurs formes :

Variable continue

Définie sur les nombres réels.

Par exemple :

- Les montants de perte d'accidents d'automobile.
- Le temps avant qu'une réclamation d'assurance soit réglée.

Variable catégorielle

Définie sur un petit nombre de valeurs catégorielles. On dit aussi *variable qualitative*.

Par exemple :

- Une **variable binaire** (seulement deux niveaux).
  - P. ex., une variable *"Maison a un système d'alarme"* prenant comme valeur *"oui"* ou *"non"*.
  - P. ex., le *sexe* d'un individu prenant comme valeur *"homme"* ou *"femme"*.
- Une variable avec plusieurs niveaux.
  - P. ex., la *marque d'une voiture* assurée prenant comme valeurs *"Toyota"*, *"Honda"*, etc.

Une variable catégorielle peut être :

1. **Nominale**

   Il n'y a pas d'ordre aux catégories.

   Par exemple :

   - Le *programme d'étude* d'un étudiant prenant comme valeur *"actuariat"*, *"comptabilité"*, etc.

2. **Ordinale**

   Il y a une d'ordre aux catégories.

   Par exemple :

   - La *sévérité d'un incendie* allant de *1* à *5*.

Variable de comptage

Définie sur les entiers positifs.

Par exemple :

- Le nombre de réclamations.

Régression

Famille exponentielle

Note La section *Statistique exhaustive minimale* du chapitre de *Famille exponentielle* détaille l'application de la famille exponentielle pour identifier le MVUE. Cette section couvre plus en détails la famille.

La famille exponentielle est de la forme : $f(y;\theta) = \mathrm{e}^{a(y)b(\theta)+c(\theta)+d(y)}$.

- Le GLM requiert que $a(y) = y$ que l'on nomme la **forme canonique**.
- Sous cette paramétrisation, $b(\theta)$ est le paramètre canonique (« *natural parameter* »).

Sous cette forme, on déduit que $\mathrm{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$ et $\mathrm{Var}(a(Y)) = \frac{b''(\theta)c'(\theta)-c''(\theta)b'(\theta)}{(b'(\theta))^3}$.

Classification

On fait de la **classification** lorsque nous voulons prédire une variable *catégorielle*. Il y a 3 trois types de variables : nominal, ordinal et binomial. Les deux premières ont été définies plus haut, une variable réponse binomiale est simplement une variable ayant 2 catégories.

Binomial

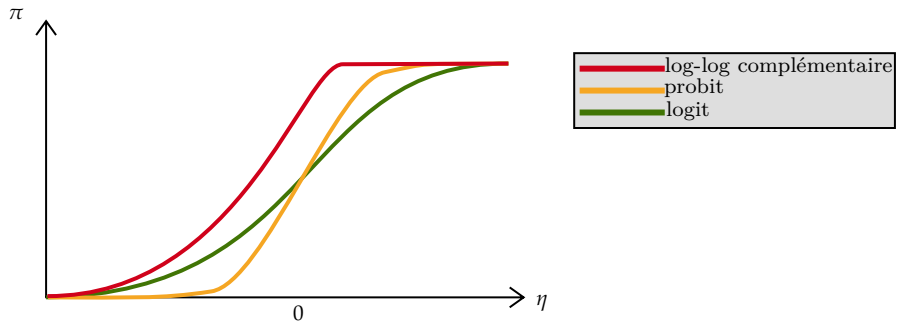
Soit $\eta = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$. Soit la probabilité π que $Y = 1$.

- Alors, on veut une fonction de lien $g(\pi) = \eta$ tel que $g(\pi) : [0,1] \mapsto (-\infty, \infty)$.
- Par exemple, la fonction quantile d'une distribution X.
  - On appelle cette distribution la « *tolerance distribution* ».
  - Ce nom provient de l'utilité du modèle pour évaluer si un médicament a un effet ou pas.
  - Une valeur élevée de η est plus probable de mener à une probabilité π élevée de oui ($Y = 1$).

Les 3 fonctions de lien les plus utilisées pour $\pi \in [0,1]$ sont les suivantes :

Nom	$\mu = \pi$	η
Logit	$\frac{\mathrm{e}^{\eta}}{1+\mathrm{e}^{\eta}}$	$\ln\left(\frac{\pi}{1-\pi}\right)$
Probit	$\Phi(\eta)$	$\Phi^{-1}(\mu)$
Log-log complémentaire	$1-\mathrm{e}^{-\mathrm{e}^{\eta}}$	$\ln(-\ln(1-\pi))$

Comme on peut observer ci-dessous, les fonctions de lien logit et probit sont **symétriques** à 0, mais pas la fonction de lien log-log complémentaire.



Note La cote, alias le « *odds ratio* », est $\frac{\pi}{1-\pi}$.

Nominal

On suppose qu'il y a J catégories possibles pour la variable réponse. Pour modéliser avec la régression logistique, on :

1. Choisit une catégorie comme catégorie de base 1.
2. Pour chacune des autres catégories, on trouve les cotes relatives (« *relative odds* »).

Le logarithme de la cote de la catégorie j relatif à la catégorie de base 1 est :

$$\ln \frac{\pi_j}{\pi_1} = \sum_{i=1}^{p} \beta_{ij} X_i = \eta_j, \quad j = 2, 3, \ldots, J$$

Alors, $\pi_j = \pi_1 e^{\eta_j}$ et puisque les probabilités doivent sommer jusqu'à 1 :

$$\pi_1 = \frac{1}{1 + \sum_{k=2}^{J} e^{\eta_k}}$$

$$\pi_j = \frac{e^{\eta_j}}{1 + \sum_{k=2}^{J} e^{\eta_k}}, \quad j = 2, 3, \ldots, J$$

Ordinal

Modèle logit cumulatif

$$\frac{\Pr(Y \leq j)}{1 - \Pr(Y \leq j)} = \frac{\sum_{k=1}^{j} \pi_k}{1 - \sum_{k=1}^{j} \pi_k} = \frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J}$$

Alors, avec les paramètres β qui varient par catégorie j, :

$$\ln \left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} \right) = \sum_{i=1}^{p} \beta_{ij} X_i$$

Modèle de cotes proportionnelles

Excepté l'intercepte, les paramètres β ne varient pas par catégorie j :

$$\ln \left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_J} \right) = \beta_{1j} + \sum_{i=2}^{p} \beta_i X_i$$

Modèle logit de catégories adjacentes

$$\ln \left(\frac{\pi_j}{\pi_{j+1}} \right) = \sum_{i=1}^{p} \beta_{ij} X_i$$

Modèle logit de ratio continu

$$\frac{\Pr(Y = j)}{\Pr(Y > j)} = \frac{\pi_j}{\pi_{j+1} + \ldots + \pi_J}$$

Autres

Régression	Type de variable réponse
Linéaire	Continue
Logistique	Binaire
Poisson	Données de comptage
Analyse de survie	Temps jusqu'à un événement

› Logistique prédit la probabilité qu'un événement ait lieu.

› Poisson prédit le « *rate* » auquel des événements aient lieu.

- C'est-à-dire, le nombre de fois, ou la fréquence, d'un événement sur une période de temps.
- Donc, le temps est fixé et on observe le nombre d'événements.
- On ne peut pas simplement appliquer un modèle linéaire, car les données suivent une distribution de Poisson, pas une distribution normale !
- Également, nous pouvons modéliser un « *offset* » pour considérer le temps d'exposition.

› Avec l'analyse de survie, on prédit le temps avant qu'un événement ait lieu.

- Donc, le nombre d'événements est fixé à un et on veut savoir le temps avant qu'il ait lieu.

Poisson

Il y a plusieurs façons de modéliser un même modèle de Poisson :

1. Modéliser le taux comme une fonction log-linéaire des x : $\lambda = \mathrm{e}^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p}$.
   › Ceci puisque le taux λ a une forme exponentielle.
2. Modéliser le log du taux comme une fonction linéaire des x : $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.
   › Ceci permet de traiter le taux λ avec un modèle linéaire.
   › L'avantage est la simplicité d'une ligne pour résumer le modèle.
   › Mathématiquement, les deux premières équations sont équivalentes.
3. Modéliser le log de la fréquence espérée avec un « *offset* » : $\ln(\mathrm{E}[Y]) = \ln(\mathrm{E}[\lambda t]) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \ln(t)$.
   › La deuxième équation représente ce que l'on fait en théorie alors que la troisième représente ce que l'on fait en pratique.

Hypothèses du modèle :

1. Les observations sont indépendantes.
   › Si, par exemple, avoir une récidive augmente la probabilité d'une deuxième récidive, alors le modèle n'est pas adéquat.
2. Le taux auquel les événements se produisent est une fonction log-linéaire de x.
   › C'est-à-dire, le log du taux est une fonction linéaire des x.
3. Les variations dans les x's ont des effets multiplicatifs sur le nombre d'événements.
   › Par exemple, si on modélise la fréquence d'accidents auto alors on s'attend à ce que le nombre d'accidents sur deux ans soit le double du nombre d'accidents sur un an.
4. La moyenne = variance = λ.
5. Le taux est constant.
   › Donc, on pose que le taux est fixe.
   › Par exemple, la probabilité d'une récidive pourrait diminuer dans le temps.

Le modèle à deux gros problèmes, premièrement la **sur-dispersion** où la variance est supérieure à la moyenne.

› C'est-à-dire que les données sont plus variables que ce qui est attendu.

› Contrairement à la régression linéaire où l'estimation de la moyenne et du SSE est séparée, l'estimation est la même pour le modèle de Poisson.

› Il y a des multiples raisons pour lesquelles ceci peut arriver :

1. Nous n'avons pas inclus toutes les variables explicatives significatives dans le modèle.
2. La forme fonctionnelle du modèle est inadéquate (p. ex., les données ne sont pas log linéaires).
3. Une variable est supposée d'être homogène alors qu'elle ne l'est pas.
   – P. ex., modéliser un groupe de fumeurs alors qu'il y a des sous-groupes (ceux qui font de l'exercice vs ceux qui n'en font pas, etc.)
4. etc.

On peut calculer la **dispersion** et on désire qu'elle soit environ de 1.
Pour résoudre la sur-dispersion, on peut :

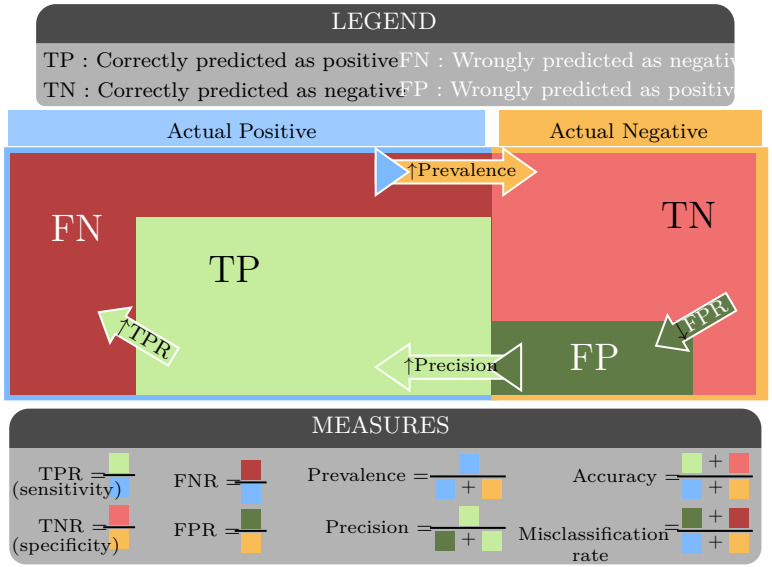
1. On peut pondérer l'erreur type de tous les coefficients par la racine du paramètre de dispersion.
   › Ceci ne change pas les prévisions, plutôt ça augmente l'erreur type pour tenir compte du fait que la variabilité des données observées est plus élevée que ce à quoi on s'attendait.
2. On peut ajuster un modèle avec une distribution binomiale négative.
   › Ceci permet d'estimer la fréquence et la variance séparément.
   › La variance sera plus grande, mais proportionnelle à la moyenne.

Deuxièmement, **Données gonflées à zéro**.

› L'idée est de modéliser la probabilité que l'événement ait lieu ou pas séparément de la fréquence.

› On peut modéliser la probabilité avec un modèle logistique est la fréquence avec un modèle de Poisson.

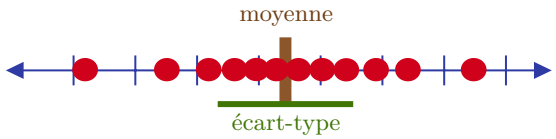
Matrice de confusion :



Erreur

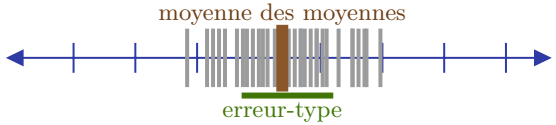
Écart-type Mesure la variation entre les observations d'**un** ensemble de données.

› « *standard deviation* ».



Erreur type Mesure la variation entre les moyennes de **plusieurs** ensembles de données.

› « *standard error* ».



Troisième partie

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 $\int_{-\infty}^{\infty} f(x)dx = 1$.

› 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

La fonction de hasard $h_X(x) = \frac{f(x)}{S(x)}$ tel que $h(x) \geq 0$.

› Par la définition, on déduit qu'elle est **seulement applicable pour les v.a. continues**.

› La fonction de hasard mesure la **vraisemblance** que la v.a. soit égale à x en gonflant la PDF moins il devient vraisemblable qu'elle soit supérieure à x.

› 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 *Sujets divers* 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 $M_X(t) = \mathrm{E}[e^{tX}]$.

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.

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* ».

› La CTE est le cas continu de la « *Tail-Value-at-Risk (TVaR)* ».

Mode

Contexte

Le mode est la réalisation qui a lieu le plus souvent.
Par exemple, en anglais la lettre E est la lettre la plus utilisée dans le dictionnaire. Elle représente donc le *mode* de la langue anglaise.

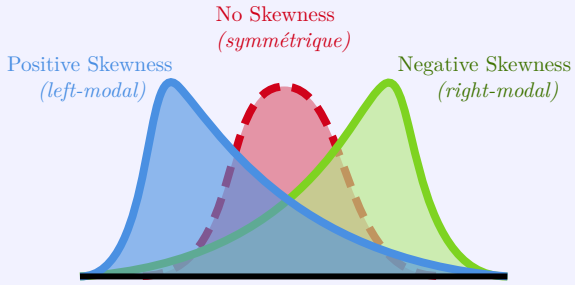
En termes mathématiques, le mode est le point qui maximise la PMF/PDF.

Dans le cas continu, si la distribution : on peut simplement dériver la PDF et trouver le point qui la rend égale à zéro.

- 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

$$\text{Skewness} = \frac{\mu_3}{\sigma^3} = \frac{\mu'_3 - 3\mu'_2\mu + 2\mu^3}{\sigma^3}.$$

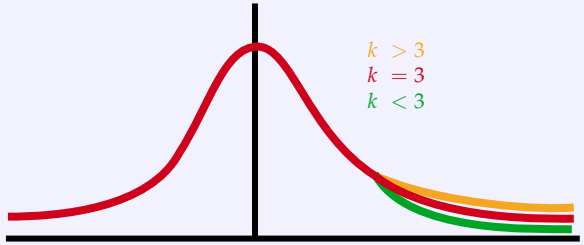


Kurtosis

$$\text{Kurtosis} = \frac{\mu_4}{\sigma^4} = \frac{\mu'_4 - 4\mu'_3\mu + 6\mu'_2\mu^2 - 3\mu^4}{\sigma^3}.$$

Le kurtosis mesure l'aplatissement d'une distribution et peut aider à juger la vraisemblance qu'une distribution produise des valeurs extrêmes (ou « *outliers* »).

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 *Sujets divers* 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 $n \in \mathbb{N}^+$	$x \geq 0$

$f(x)$	$= \dfrac{x^{n-1}\lambda^n \mathrm{e}^{-\lambda x}}{\Gamma(n)}$
$S(x)$	$= \displaystyle\sum_{k=0}^{n-1} \frac{(\lambda x)^{k-1}\mathrm{e}^{-\lambda x}}{k!}$

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, $f_W(w) = f_X\left(g^{-1}(w)\right)\left|\frac{\partial g^{-1}(w)}{\partial w}\right|$.

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.

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

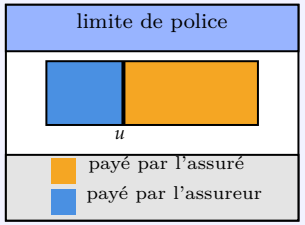
X Variable aléatoire du montant de perte.

Limite de police

Limite de police

Une **limite de police** u est le montant maximal qu'un assureur va payer pour une perte.

Visuellement :



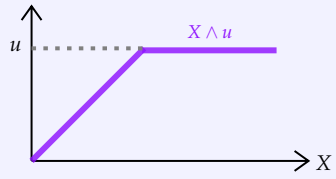
Montant de perte limité

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.

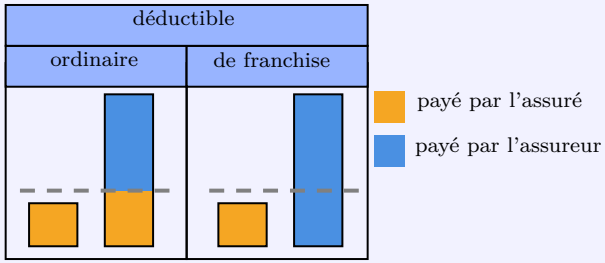
Il y a 2 types de déductibles :

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**

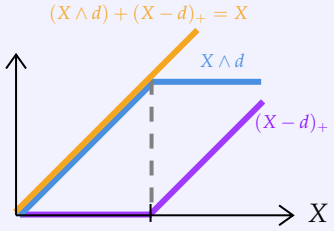
La variable aléatoire du montant de perte pour une police ayant un **déductible ordinaire** de d.

Assureur	Assuré
$(X-d)_+ = \begin{cases} 0, & X \leq d \\ X-d, & X > d \end{cases}$	$X \wedge d = \begin{cases} X, & X < d \\ d, & X \geq d \end{cases}$

› Il s'ensuit que $(X-d)_+ = \max(X-d;0)$.

› On observe que le montant de perte est la somme des contributions $X = X \wedge d + (X-d)_+$.

Visuellement :



L'espérance du montant de perte avec un déductible ordinaire

L'espérance du montant de perte, pour l'assureur, **avec un déductible *ordinaire*** $\mathrm{E}[(X-d)_+]$ 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, $LER = \frac{\mathrm{E}[X \wedge d]}{\mathrm{E}[X]}$.

« *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 $\mathrm{E}[Y^L] \geq \mathrm{E}[Y^L]$.

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	$(X-d\|X>d)$
$\mathrm{Exp}(\theta)$	$\mathrm{Exp}(\theta)$
$\mathrm{Unif}(a,b)$	$\mathrm{Unif}(0,b-d)$
$\mathrm{Pareto}(\alpha,\theta)$	$\mathrm{Pareto}(\alpha,\theta+d)$
$\mathrm{Beta}(1,b,\theta)$	$\mathrm{Beta}(1,b,\theta-d)$

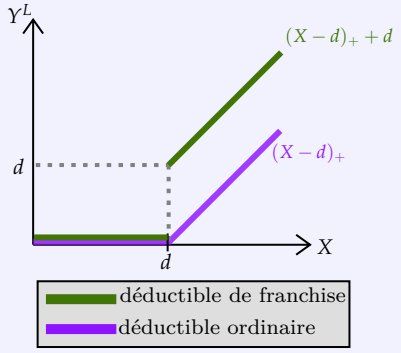
Déductible de franchise

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

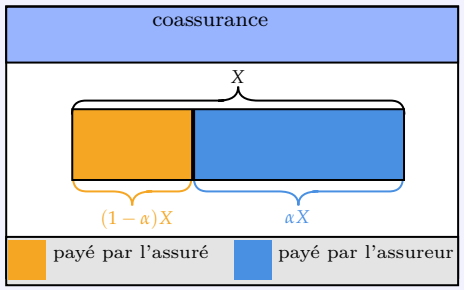
Pour la classe $(a,b,0)$ de distributions, on trouve les relations suivantes :

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

$$Y^L = \begin{cases} 0, & X \leq d \\ \alpha(X-d), & d < X < m \\ u, & X \geq m \end{cases}$$

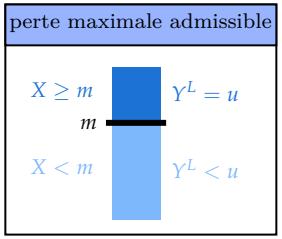
où m est la **perte maximale admissible**.

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

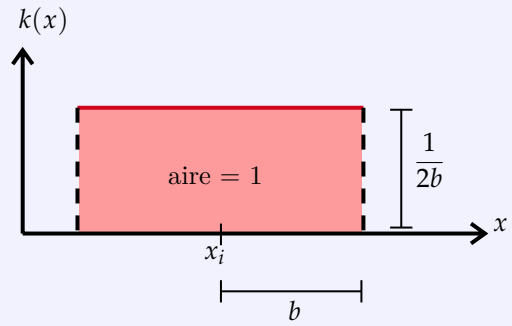
1. Choisir un type de fonction de densité pour $k()$.
2. Estimer la fonction de densité $f(x)$ comme la moyenne des fonctions noyau des n observations $k_1(x), \ldots, k_n(x)$:

$$\tilde{f}(x) = \frac{1}{n} \sum_{i=1}^{n} k_i(x)$$

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 $\frac{1}{2b}$.

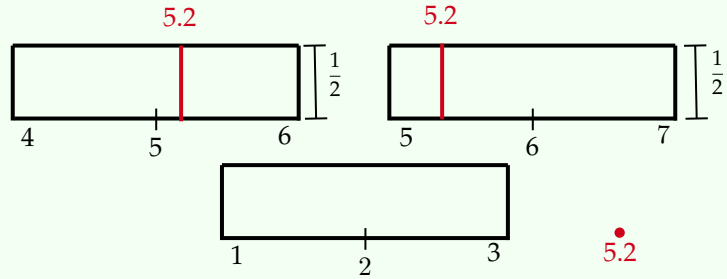
En termes mathématiques :

$$k_i(x) = \begin{cases} \frac{1}{2b}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{sinon} \end{cases}$$

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 :

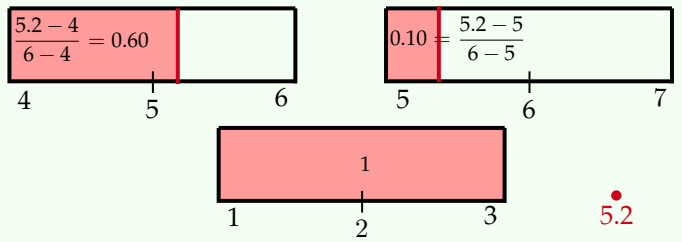


3. La fonction de densité estimée est donc :

$$\tilde{f}(5.2) = \frac{1}{3}\left(\frac{1}{2} + 0 + \frac{1}{2}\right) = \frac{1}{3}$$

Si on désire trouver la probabilité que la réclamation soit inférieure à 5.2 :

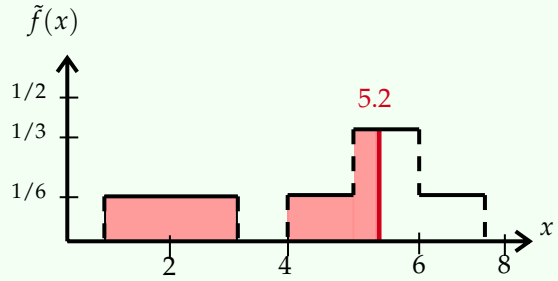
1. Visuellement, on voit comment l'équivalence géométrique du calcul des probabilités :



2. Donc :

$$\tilde{F}(5.2) = \frac{1}{3}(0.60 + 1 + 0.10) = 0.567$$

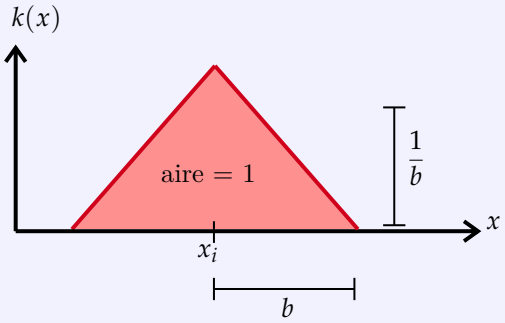
Visuellement, la densité par noyau est :



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 :

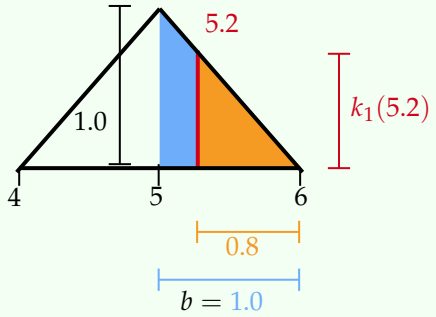
$$k_i(x) = \begin{cases} \frac{b-|x-x_i|}{b^2}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{sinon} \end{cases}$$

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 $k_1(5.2)$:

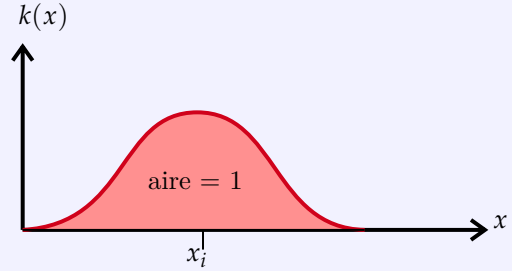


$$\frac{k_1(5.2)}{1.0} = \frac{0.8}{1.0}$$
$$k_1(5.2) = \frac{0.8}{1.0} \times 1.0$$

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 Analyse statistique des risques actuariels.

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 $j = 1,\dots,k$ et $\boldsymbol{n} = (n_1,\cdots,n_k)$, est :

$$\mathcal{L}(\theta;\boldsymbol{n}) = \prod_{j=1}^{k} \underbrace{\left[F(c_j;\theta) - F(c_{j-1};\theta)\right]^{n_j}}_{\substack{\text{probabilité qu'une observation} \\ \text{soit contenue dans l'intervalle}}}$$

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 :

$$\mathcal{L}(\theta;\boldsymbol{x}) = \underbrace{\left[\prod_{i=1}^{n_1} f(x_i;\theta)\right]}_{\substack{\text{probabilité de chaque} \\ \text{observation à la} \\ \text{valeur observée}}} \overbrace{\left[1 - F(u;\theta)\right]^{n_2}}^{\substack{\text{probabilité qu'une observation} \\ \text{soit supérieure, ou égale, à } u}}$$

tronquées vers la gauche

On pose un déductible de d :

$$\mathcal{L}(\theta;\boldsymbol{x}) = \underbrace{\frac{1}{\left[1 - F(d;\theta)\right]^{n}}}_{\substack{\text{pondère la vraisemblance par} \\ \text{la probabilité d'être supérieur} \\ \text{au déductible}}} \prod_{i=1}^{n} f(x_i;\theta)$$

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.

Le test de K.-S. compare la fonction de distribution (répartition) empirique à celle d'une distribution théorique. L'idée du test est donc de quantifier l'évaluation visuelle que l'on peut faire de l'ajustement.

Test de Kolmogorov-Smirnov

On teste si les données semblent suivre une distribution (« supportent l'hypothèse nulle ») avec la statistique de Kolmogorov-Smirnov :
$$D = \max_{x_{(1)} \leq x \leq x_{(n)}} \left| \hat{F}(x) - F^*(x) \right|.$$

- Ceci équivaut donc à calculer la différence maximale entre la fonction de répartition empirique et celle de la distribution.
- Puisque $\hat{F}()$ est une fonction à escalier, il faut seulement évaluer la fonction aux points observés $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$.
- De plus, le maximum peut seulement arriver soit au point de saut $x_{(i)}$ ou immédiatement avant $x_{(i-1)}$.

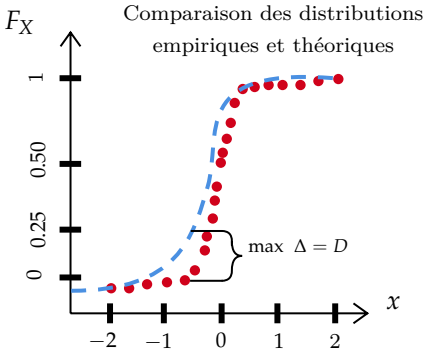
On peut donc récrire
$$D = \max_{i \in \{1,\ldots,n\}} \left\{ \max \left\{ \left| \hat{F}(x_{(i-1)}) - F^*(x_{(i)}) \right|, \left| \hat{F}(x_{(i)}) - F^*(x_{(i)}) \right| \right\} \right\}.$$

- Si les données sont bien ajustées, on s'attend à ce que D soit très petit.
- Lorsque la distribution est entièrement spécifiée (aucun paramètre n'est estimé), une table avec les valeurs critiques est donnée.
- S'il faut estimer des paramètres, la simulation Monte-Carlo est utilisée pour trouver des nouvelles valeurs critiques.

Test de K.-S. pour des données incomplètes

Pour des données tronquées à d et censurée (vers la droite) à u,
$$D = \max_{d \leq x \leq u} \left| \hat{F}(x) - F^*(x) \right|$$

Visuellement, le test de K.-S. ressemble à :



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 maximise 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.

Quatrième partie
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 des composantes.
La **théorie de la fiabilité** sert à quantifier la probabilité qu'un *système* fonctionne selon la fiabilité de ses composantes, et selon le rôle qu'elles ont dans le système.

Introduction aux systèmes

Notation

x_i **État** de la composante i.
$\phi(\boldsymbol{x})$ « ***Structure function*** » d'un système représentant son **état**.

L'état d'une composante

Chacune des composantes du sys tème a sa propre durée de vie (« *lifetime* »). Cette durée de vie est dénotée par la variable aléatoire binaire x_i représentant son état.

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. Il indique donc quelles composantes fonctionnent ou ne fonctionnent pas. Il est représentée sous la forme $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

Note Un système ayant n composantes et le vecteur des états peut être un de 2^n différentes combinaisons.

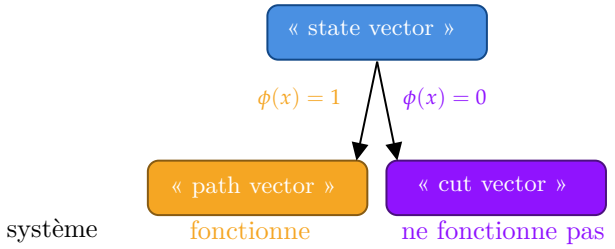
› Puisque les composantes du système sont binaires, chacune a deux valeurs possibles. Ceci résulte en $2 \times 2 \times \cdots 2 = 2^n$ différentes combinaisons possibles.

L'état d'un système

L'état d'un système dépend des états de ses composantes. L'état du système est représentée sous la forme d'une fonction $\phi(\boldsymbol{x})$ binaire :

$$\phi(\boldsymbol{x}) = \begin{cases} 1, & \text{si le système fonctionne} \\ 0, & \text{si le système ne fonctionne pas} \end{cases}$$

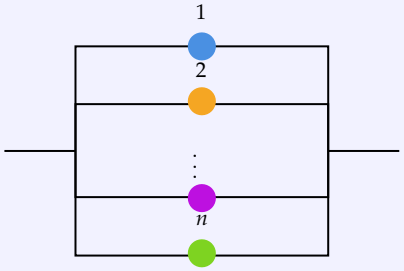
Nous verrons le type de vecteur d'état selon l'état de fonctionnement du système :



Systèmes communs

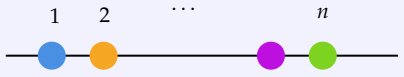
Système parallèle

Fonctionne tant qu'au moins une des composantes du système fonctionne.



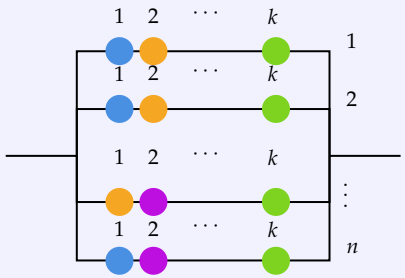
Système de série

Fonctionne seulement si toutes les composantes du système fonctionnent.



Système de k parmi n

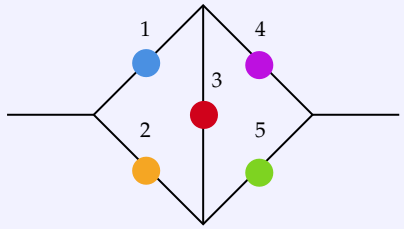
Fonction si au moins k des n composantes du système fonctionnent.



› Un système parallèle est donc un système de 1 parmi n et un système de série un système de n parmi n.

Système de pont

Il y a deux branches connectées par un pont dans le milieu.



Autres systèmes
En bref, il y a une infinité de systèmes qui peuvent être construits comme des combinaisons des systèmes précédents.

Minimal path and minimal cut sets

« *Path vector* »

Vecteur d'états pour lequel le système fonctionne ($\phi(\boldsymbol{x}) = 1$).

« *Minimal path vectors* »

« *Path vectors* » ayant le *minimum* de composantes pour **fonctionner**. Donc, le système cesse de fonctionner dès qu'une des composantes qui fonctionne échoue.

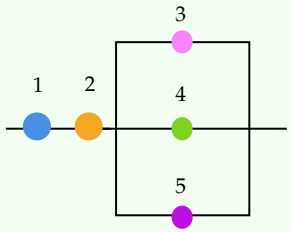
En termes mathématiques, $\boldsymbol{x}$ est un « *minimal path vector* » si $\phi(\boldsymbol{y}) = 0 \forall \boldsymbol{y} < \boldsymbol{x}$.

› $\boldsymbol{y} < \boldsymbol{x}$ 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* »

Ensembles minimaux des 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	
1	{1, 2, 3}
2	{1, 2, 4}
3	{1, 2, 5}

Minimal path vectors	
1	$\boldsymbol{x} = (1, 1, 1, 0, 0)$
2	$\boldsymbol{x} = (1, 1, 0, 1, 0)$
3	$\boldsymbol{x} = (1, 1, 0, 0, 1)$

Pour bien comprendre la condition pour qu'un « *mimimal path vector* », on observe les vecteurs $\boldsymbol{y}$ du premier « *minimal path vector* » $\boldsymbol{x}$:

Minimal path sets
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* »

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* »

« *Cut vectors* » ayant le *maximum* de composantes pour ne **pas fonctionner**. Donc, le système fonctionne dès qu'une des composantes qui ne fonctionne pas est réparée.

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).

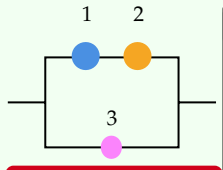
« *Minimal cut sets* »

Ensembles minimaux des 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 ($\psi(\boldsymbol{y}) = 1$).



Minimal cut vectors

Cut vectors	$\boldsymbol{y} > \boldsymbol{x}$
$(0, 0, 0)$	$(1, 1, 0), (0, 0, 1), (0, 1, 1),$ $(1, 0, 1), (1, 1, 1)$
$(1, 0, 0)$	$(1, 1, 0), (0, 0, 1), (0, 1, 1)$
$(0, 1, 0)$	$(1, 1, 0), (1, 0, 1), (0, 1, 1)$

$\psi(\boldsymbol{y}) = 1$ pour tous les vecteurs

	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 comment qu'il est organisé.

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 $\phi(\boldsymbol{x}) = 0$.

$$\begin{aligned}\phi(\boldsymbol{x}) &= \min\{x_1, \ldots, x_n\} \\ &= \prod_{i=1}^{n} x_i\end{aligned}$$

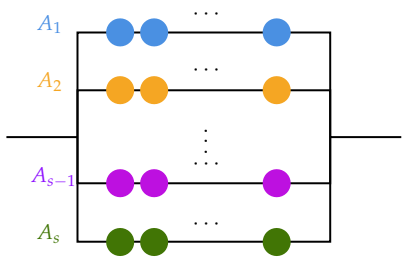
› La deuxième formulation découle du fait que les états sont des variables binaires.

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 le système comme :

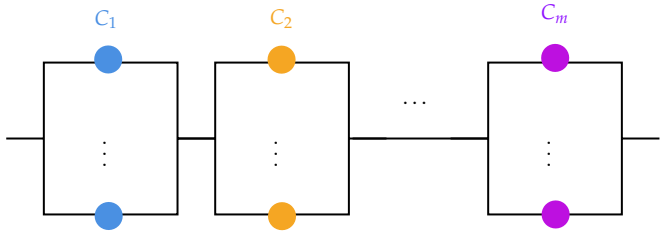
$$\phi(\boldsymbol{x}) = \max\left\{\min_{i\in A_1} x_i, \min_{i\in A_2} x_i, \ldots, \min_{i\in A_s} x_i\right\} = \min_{j} \prod_{i\in A_j} x_i$$

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 :

$$\phi(\boldsymbol{x}) = \min\left\{\max_{i\in C_1} x_i, \max_{i\in C_2} x_i, \ldots, \max_{i\in C_s} x_i\right\} = \prod_{j=1}^{m} \max_{i\in C_j} x_i$$

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

Fonction de la fiabilité des composantes $r(\boldsymbol{p})$ qui quantifie la probabilité que le système fonctionne.

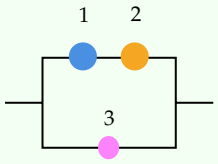
$$r(\boldsymbol{p}) = \underbrace{\Pr(\phi(\boldsymbol{X}) = 1)}_{\substack{\text{somme des probabilités} \\ \text{des « } \textit{path vectors} \text{ »}}} = 1 - \underbrace{\Pr(\phi(\boldsymbol{X}) = 0)}_{\substack{\text{somme des probabilités} \\ \text{des « } \textit{cut vectors} \text{ »}}}$$

$$= 0 \times \Pr(\phi(\boldsymbol{X}) = 0) + 1 \times \Pr(\phi(\boldsymbol{X}) = 1) = \mathrm{E}[\phi(\boldsymbol{X})]$$

› Puisque ϕ est fonction du vecteur de v.a. Bernoulli $\boldsymbol{X}$, 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 :

$$\begin{aligned} r(\boldsymbol{p}) &= \Pr(\phi(\boldsymbol{X}) = 1) \\ &= \Pr(\boldsymbol{X} = (0,\ 0,\ 1)) + \Pr(\boldsymbol{X} = (1,\ 1,\ 0)) + \Pr(\boldsymbol{X} = (1,\ 0,\ 1)) + \\ &\quad \Pr(\boldsymbol{X} = (0,\ 1,\ 1)) + \Pr(\boldsymbol{X} = (1,\ 1,\ 1)) \\ &= (1 - p_1)(1 - p_2)p_3 + p_1p_2(1 - p_3) + p_1(1 - p_2)p_3 + \\ &\quad (1 - p_1)p_2p_3 + p_1p_2p_3 \\ &= p_3 - p_2p_3 - p_1p_3 + p_1p_2p_3 + p_1p_2 - p_1p_2p_3 + p_1p_3 - p_1p_2p_3 + \\ &\quad p_2p_3 - p_1p_2p_3 + p_1p_2p_3 \\ &= p_3 + p_1p_2 - p_1p_2p_3 \end{aligned}$$

Bornes des fonctions de fiabilité

Contexte

Parfois, il n'est pas pratique ni nécessaire de trouver la fonction de fiabilité exacte. En lieu, on peut l'approximer en trouvant les bornes supérieures et inférieures de la fonction avec une des deux méthodes suivantes.

Méthode d'inclusion et d'exclusion

Rappel : Probabilités conjointes

$$\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$$

$$\Pr\left(\bigcup_{j=1}^{n} E_j\right) = \sum_{j=1}^{n} \Pr(E_j) - \sum_{j=1}^{n}\sum_{k>j} \Pr(E_j \cap E_k) + \sum_{j=1}\sum_{k>j}\sum_{l>k} \Pr(E_j \cap E_k \cap E_l) - \cdots + (-1)^{n+1} \Pr(E_1 \cap E_2 \cap \cdots \cap E_n)$$

Si on utilisait seulement la première somme de l'équation, on *sur*-estime la probabilité.
Si on utilise seulement les deux premières sommes, alors on *sous*-estime.
Ce qu'on en déduit est que la probabilité est **contenue entre ces deux estimations** et donc on peut établir des inégalités.

On peut établir les inégalités soit pour la probabilité que le système fonctionne ($r(\boldsymbol{p})$) ou pour la probabilité que le système ne fonctionne pas ($1 - r(\boldsymbol{p})$).

Minimal path sets On a que $\sum_{j=1}^{n} \Pr(E_j) = \sum_{j=1}^{s} \left(\prod_{i\in A_j} p_i\right)$.
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 $A_2 = (1,1,0)$:

$$\sum_{j=1}^{s} \left(\prod_{i\in A_j} p_i\right) = \prod_{i\in A_1} p_i + \prod_{i\in A_2} p_i = p + p^2$$

$$\sum_{j=1}^{s}\sum_{k>j} \left(\prod_{i\in A_j\cup A_k} p_i\right) = \prod_{i\in A_1\cup A_2} p_i = p^3$$

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.

Minimal cut sets Pour les « *minimal cut sets* » $C_1, \ldots, C_m$, 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,

$$\overbrace{\prod_{j=1}^{m}\left[1-\overbrace{\prod_{i\in C_j}(1-p_i)}^{\substack{\text{probabilité que toutes les composantes du } C_j \\ \text{échouent}}}\right]}^{\substack{\text{probabilité qu'au moins une composante} \\ \text{de chacun des « } \textit{minimal cut sets} \text{ »} \\ \text{fonctionne}}}^{\substack{\text{probabilité qu'au moins une} \\ \text{des composantes du } C_j \\ \text{fonctionne}}} \leq r(\boldsymbol{p}) \leq \overbrace{1-\prod_{j=1}^{s}\left[1-\overbrace{\prod_{i\in A_j}p_i}^{\substack{\text{probabilité que toutes les composantes du } A_j \\ \text{fonctionnent}}}\right]}^{\substack{\text{probabilité que toutes les composantes} \\ \text{d'au moins un des « } \textit{minimal path sets} \text{ »} \\ \text{fonctionnent}}}^{\substack{\text{probabilité qu'au moins une} \\ \text{des composantes du } A_j \\ \text{échoue}}}$$

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.

En bref :

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.

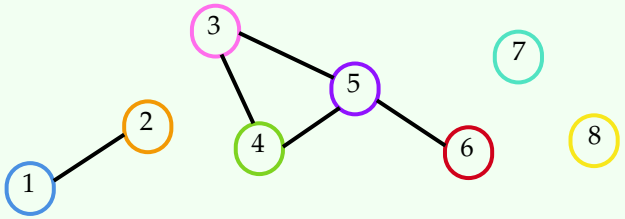
- Le nombre d'arcs est au plus $\binom{n}{2}$.
- C'est-à-dire, le nombre possibles de groupes de deux nœuds.

Également, un graphique peut être décomposé en sous-graphiques qu'on nomme les ***composantes***.

- Les composantes ne se chevauchent pas.
- Les composantes sont composées de nœuds connectés.
- Un graphique est ***connecté*** s'il a une seule composante.
- En autres mots, on peut aller d'un nœud à tout autre nœud du graphique via les arcs.

Exemple de graphique

Soit le graphique suivant :



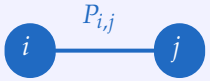
On trouve que :

- 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 aléatoiress

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.

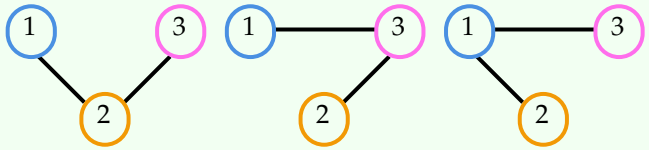
Un graphique aléatoire de n nœuds a :

- n^{n-2} « *minimal path sets* », et
- $2^{n-1} - 1$ « *minimal cut sets* »,
- $2^{\binom{n}{2}}$ graphiques possibles.

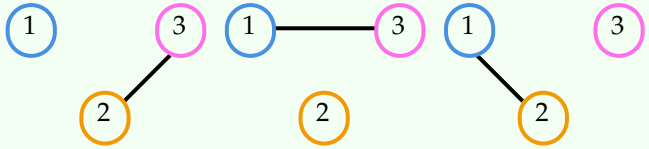
Exemple de connectivité

Soit un graphique aléatoire avec 3 nœuds.

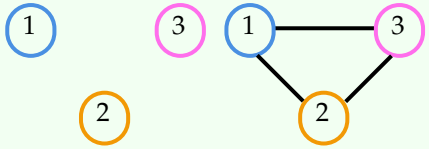
Les $3^{3-2} = 3$ « *minimal path sets* » sont les suivants :



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 chaque v.a. est iid avec $P_{i,j} = p$.

Alors, 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 pour 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 *Mathématiques actuarielles IARD I* 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 *Mathématiques actuarielles IARD* 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.

$$\begin{aligned} A_x &= \sum_{k=0}^{\omega-x-1} v^{k+1}{}_kp_x q_{x+k} \\ &= vq_x + v^2 p_x q_{x+1} + v^3{}_2p_x q_{x+2} + \dots \end{aligned}$$

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.

$${}_{m|}A_x = \sum_{k=m}^{\omega-x-1} v^{k+1}{}_kp_x q_{x+k}$$

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.

$$A^1_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1}{}_kp_x q_{x+k}$$

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.

$$A_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1}{}_kp_x q_{x+k} + {}_nE_x$$

› 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.

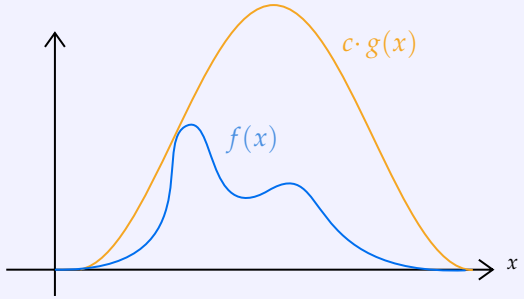
Supposons que nous pouvons simuler des réalisations d'une distribution ayant la fonction de densité g et que l'on veut simuler des réalisations d'une autre distribution ayant la fonction de densité f. Par exemple :

Simulation par la méthode d'acceptation-rejet

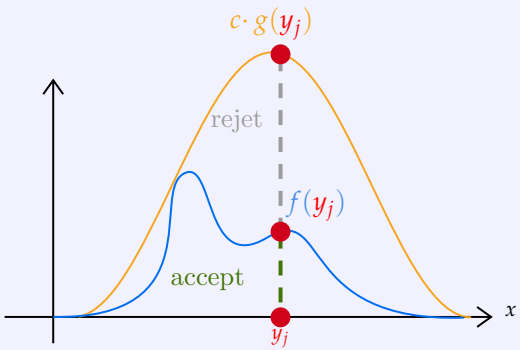
Pour une variable aléatoire X,

1. Trouver une constante c telle que $\frac{f(x)}{g(x)} \leq c$, $\forall x$.

   Par exemple,



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.

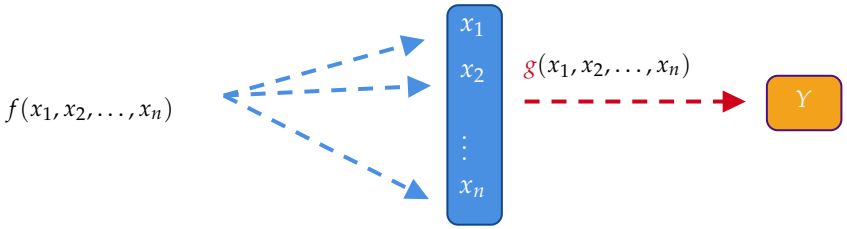
Simulation Monte-Carlo

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 :



Cinquième partie

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.

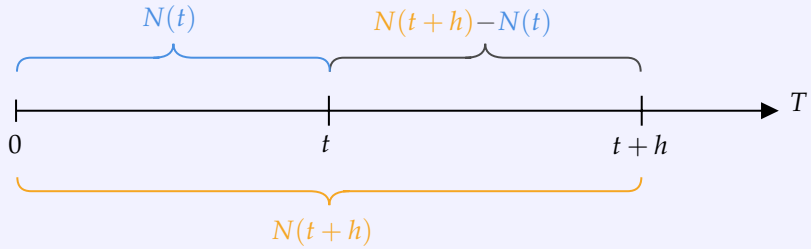
Processus de comptage

On dénote le processus de comptage 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 voir que l'**accroissement** $N(t+h) - N(t)$ représente le nombre d'événements produits sur l'intervalle $(t, t+h]$:



› Alias, processus de dénombrement.

Processus de Poisson

Processus de comptage dont :

1. chaque accroissement est une variable aléatoire 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 $N(t+h)-N(t) \sim \text{Poisson}(\lambda t)$.

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.

› $T_k = V_1 + V_2 + \cdots + V_k$.

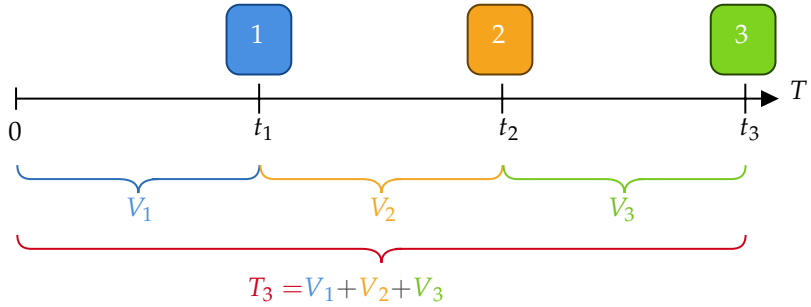
V_k Intervalle de temps entre la réalisation du $(k-1)^e$ et du k^e événement.

› 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 $T_k \sim \text{Gamma}\left(\alpha = k, \theta = \frac{1}{\lambda}\right) \sim \text{Erlang}\left(n = k, \lambda\right)$.

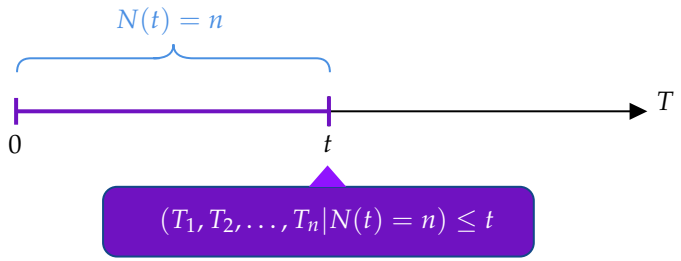
Note La loi Gamma avec un paramètre de forme α entier correspond à la loi Erlang. L'avantage de la loi Erlang est qu'elle a une fonction de répartition explicite qui découle de la relation entre les processus de Poisson et les temps d'occurrences. Voir la sous-section sur les **Distributions** du chapitre de Mathématiques actuarielles IARD I.

Temps d'occurrence conditionnels

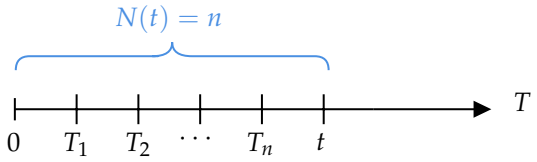
Note Voir la sous-section des **Statistiques d'ordre** du chapitre de Analyse statistique des risques actuariels.

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 $T_1, T_2, \ldots, T_n \leq t$:



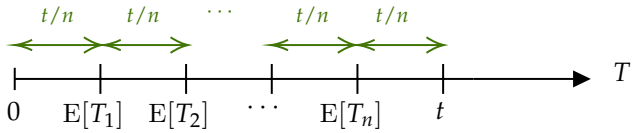
On en déduit que les temps d'occurrences sont en fait des **Statistiques d'ordre** avec $0 < T_1 \leq T_2 \leq \cdots \leq T_n \leq t$:



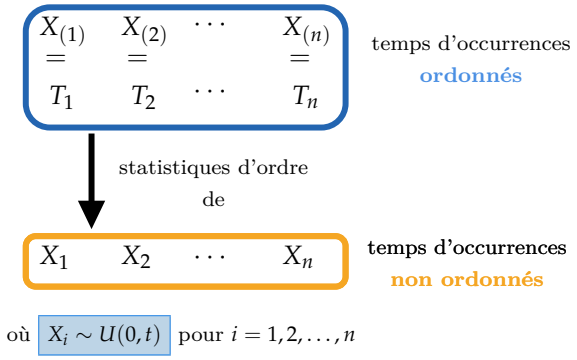
Pour déterminer la distribution de T_i, $i = 1, 2, \ldots, n$, on rappel ces deux propriétés des processus de Poisson homogènes :

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.

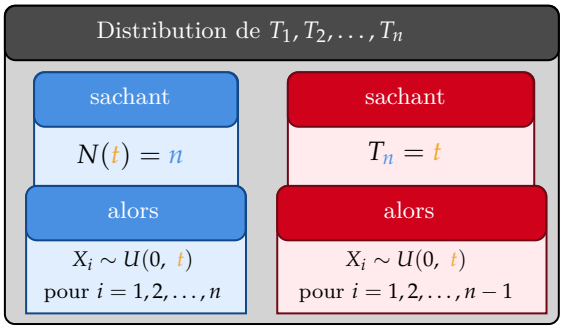
On en déduit que les temps d'occurrences des événements devraient être **uniformément distribués** en $n+1$ sous-intervalles :



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 $\mathrm{E}[X_{(k)}] = \mathrm{E}[T_k] = a + \frac{k(b-a)}{n+1}$.

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.

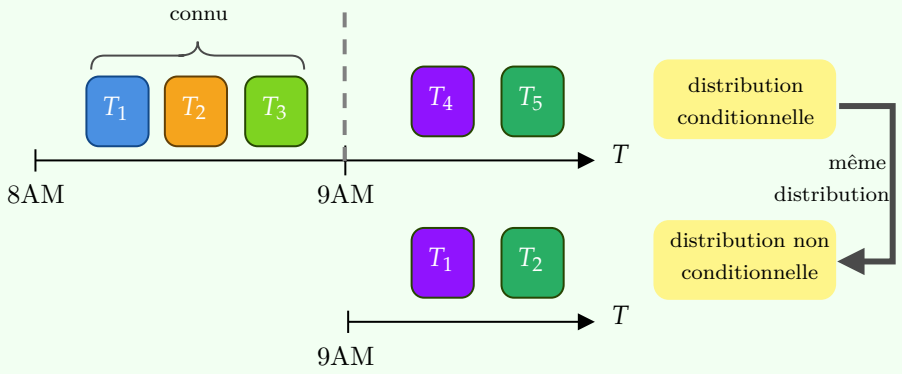
Calculer :

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 :



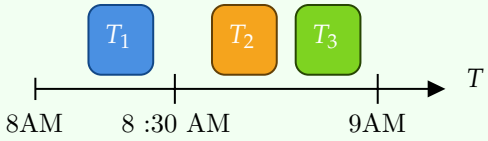
2. Puisque T_5 n'est pas borné, il suit une distribution Gamma$(2,1/4)$. Donc, $\mathrm{E}[T_2]=\frac{2}{4}=0.50$ ce qui équivaut à 9h30AM.

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, $\Pr(N(8,8.5] = 1|N(8,9] = 3) = \Pr\left(\substack{\text{un autobus arrive}\\\text{entre 8h00 et 8h30}} \bigcap \substack{2 \text{ autobus arrivent}\\\text{entre 8h30 et 9h00}}\right) = \Pr(1 \text{ succès}) = \binom{3}{1} 0.5^1 (1-0.5)^2 = 0.375$

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

$$\text{E}[S(t)] = \text{E}[N(t)]\text{E}[X] \qquad \text{Var}(S(t)) = \text{E}[N(t)]\text{E}[X^2]$$

Approximation de la distribution

Puisque la distribution de $S(t)$ est difficile à déterminer, elle peut être approximée avec le **théorème limite centrale** où $S(t) \approx \mathcal{N}\left(\text{E}[S(t)], \text{Var}(S(t))\right)$.

Il s'ensuit que :

$$\Pr(S(t) < s) = \Phi\left(\frac{s - \text{E}[S(t)]}{\sqrt{\text{Var}(S(t))}}\right)$$

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

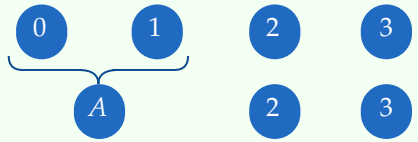
$\mathcal{A}$ L'ensemble des états à éviter.

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 :

$$\boldsymbol{A}^{(2)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \boldsymbol{A}^{(3)} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad \boldsymbol{A}^{(4)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

La chaîne de Markov est donc ***périodique***.

Chaîne de Markov ergodique

Une chaîne de Markov **irréductible**, **récurrente positive** et **apériodique** est ***ergodique***.

Time Spent in Transient States

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.

› $f_{i,j} = \frac{s_{i,j} - \delta_{i,j}}{s_{j,j}}$.

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

$\boldsymbol{A}^{-1}$ Inverse de la matrice $\boldsymbol{A}$ tel que $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^{-1} = I$.

Soit la matrice 2×2 $\boldsymbol{A}$ où :

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Alors son inverse $\boldsymbol{A}^{-1}$ est :

$$\boldsymbol{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

› 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 :

$$\boldsymbol{P} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ 0 & 0.8 & 0.2 \\ 0 & 0 & 1 \end{bmatrix}$$

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 :
   $\boldsymbol{P}_T = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0.8 \end{bmatrix}$
2. Trouver $\boldsymbol{I} - \boldsymbol{P}_T$:
   $\boldsymbol{I} - \boldsymbol{P}_T = \begin{bmatrix} 0.5 & -0.5 \\ 0 & 0.2 \end{bmatrix}$
3. Trouver l'inverse $(\boldsymbol{I} - \boldsymbol{P}_T)^{-1}$:
   $(\boldsymbol{I} - \boldsymbol{P}_T)^{-1} = \begin{bmatrix} 0.2 & 0.5 \\ 0 & 0.5 \end{bmatrix} \times \frac{1}{0.10 - 0} = \begin{bmatrix} 2 & 5 \\ 0 & 5 \end{bmatrix}$
4. Trouver l'élément $s_{1,2}$ de la matrice $\boldsymbol{S} = (\boldsymbol{I} - \boldsymbol{P}_T)^{-1}$ 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 $\boldsymbol{R}$.

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 of Markov Chains

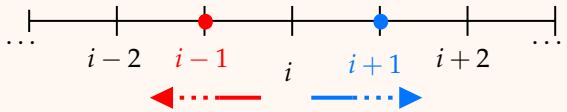
Random Walk

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 $p \in [0,1]$.

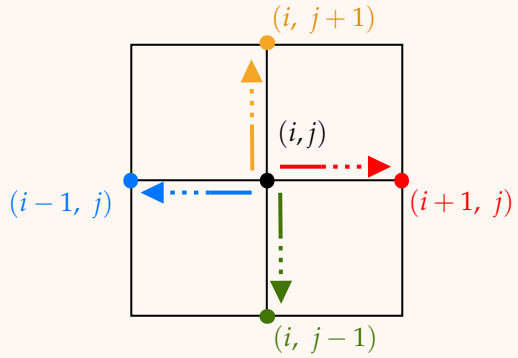
On peut donc visualiser une ligne :



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 $i \in \{1, 2, \dots, j-1\}$.
3. Il y a 3 classes : $\{0\}, \{1, 2, \dots, j-1\}, \{j\}$.
4. Les états $\{0\}$ et $\{j\}$ son récurrents puisqu'ils sont absorbants et les états $\{1, 2, \dots, j-1\}$ 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 :

$$P_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^j}, & p \neq \frac{1}{2} \\ \frac{i}{j}, & p = \frac{1}{2} \end{cases}$$

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, $\boxed{\text{Var}(X) = (b-a)^2 p(1-p)}$.

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 $\boxed{\mu = \sum_{j=0}^{\infty} j P_j}$.

La variance du nombre de nouveaux descendants qu'un individu produit est $\boxed{\sigma^2 = \sum_{j=0}^{\infty} (j-\mu)^2 P_j}$.

Notation

X_n Taille de la n^e génération.

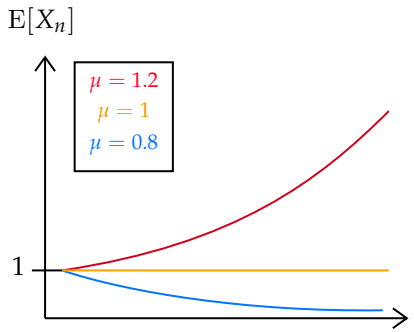
Si l'on pose une population initiale de 1 ($\boxed{X_0 = 1}$) :

$$\text{E}[X_n] = \mu^n$$

$$\text{Var}(X_n) = \begin{cases} \sigma^2 \mu^{n-1} \left(\frac{1-\mu^n}{1-\mu}\right), & \mu \neq 1 \\ n\sigma^2, & \mu = 1 \end{cases}$$

› Si la population initiale est de k ($\boxed{X_0 = k}$) alors la moyenne est de $k\text{E}[X_n]$ et la variance de $k\text{Var}(X_n)$.

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.

Sixième partie

Séries chronologiques