

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!

Table des matières

I Analyse statistique des risques actuariels

Échantillonnage et statistiques

Statistiques	5
Statistiques univariées	5
Statistiques bivariées	6

Vraisemblance

Qualité de l'estimateur

Estimation ponctuelle	7
Biais	8
Variance	8
Erreur quadratique moyenne	9
Convergence	9
Borne Cramér-Rao	11
Efficacité	12
Estimateur non biaisé à variance minimale (MVUE)	13
Estimation par intervalles	13

Intervalles de confiance

Intervalles sur la moyenne	15
1 échantillon	15
2 échantillons	15
Intervalles sur les proportions	16
1 échantillon	16
2 échantillons	16
Intervalles sur la variance	17
1 échantillon	17
2 échantillons	18

Tests d'hypothèses

Hypothèses	18
Région et valeur critique	20
Erreurs de test	20
Certitude du test	21
Valeur p vs seuil α	21
Résumé graphique des régions critiques	22
Tests sur la moyenne	22
Rappels	22
1 échantillon	23

2 échantillons	24
Tests sur les proportions	25
1 échantillon	25
2 échantillons	25
Tests sur la variance	25
Rappels	25
1 échantillon	26
2 échantillons	26
Puissance d'un test	27
Facteurs influençant la puissance	27
La fonction de puissance	28
Tests optimaux (les plus puissants)	29
Introduction	29
Test le plus puissant	29
Test uniformément le plus puissant	32
Tests d'adéquation	34
Test de Kolmogorov-Smirnov	34
Test d'adéquation du khi carré (« <i>Chi-Square Goodness-of-Fit Test</i> »)	35
Test de l'indépendance du khi carré (tableau de contingence)	35
Test du rapport de vraisemblance	36

Statistiques exhaustives

Statistique complète	38
Statistique exhaustive minimale	39
Famille exponentielle	40

Statistiques d'ordre

Principes fondamentaux	41
Cas spéciaux	41
Autres statistiques	42
Distribution conjointe	43
Graphiques	43
Diagramme en boîte (« <i>boxplot</i> »)	43
Diagramme quantile-quantile (« <i>Q-Q plot</i> »)	45

Construction d'estimateurs

Introduction	45
Méthode des moments (MoM)	46
Méthode du «Percentile Matching»	46
Méthode du maximum de vraisemblance	47
Raccourcis	47
Propriétés	47

3 de 127

Évaluation et sélection de modèles	90	Introduction	109
Graphiquement	90	Processus de Poisson	109
Tests pour la qualité de l'ajustement	90	Temps d'occurrence	110
Critères d'information pour la sélection de modèles	91	Temps d'occurrence conditionnels	111
IV Sujets divers	92	Propriétés des processus de Poisson	113
Optimisation numérique	92	Décomposition de processus de Poisson	113
Théorie de la fiabilité	93	Superposition	113
Introduction aux systèmes	93	Probabilités conjointes	113
Systèmes communs	93	Mélanges de processus de Poisson	114
Minimal path and minimal cut sets	94	Processus de Poisson composés	114
Structure Functions	96	Chaînes de Markov	115
Approche par les « <i>minimal path sets</i> »	96	Introduction	115
Approche par les « <i>minimal cut sets</i> »	96	Probabilités de transitions en plusieurs étapes	116
Fiabilité des systèmes	97	États absorbants	116
Bornes des fonctions de fiabilité	98	Transitions de (ou vers) un état absorbant	117
Méthode d'inclusion et d'exclusion	98	Probabilités inconditionnelles	117
Méthode d'intersection	99	Classification des états	118
Graphiques aléatoires	100	Probabilités stationnaires et limites	120
Durée de vie des systèmes	102	Chaînes de Markov avec bénéfices	121
Divers	103	Probabilités limites	121
Distributions particulières	103	Temps passé dans les états transitoires	122
Assurance vie	104	« <i>Time Reversibility</i> »	123
Probabilités	104	Applications des chaînes de Markov	124
Espérances de vie	104	Marche aléatoire	124
Contrats d'assurance vie	105	« <i>Gambler's ruin</i> »	124
Contrats de rentes	106	« <i>Branching Process</i> »	125
Rentes de base	106	VI Séries chronologiques	127
Vies conjointes	106		
Principe d'équivalence	106		
Assurance nivelée	106		
Simulation	107		
Méthode de l'inverse	107		
Méthode d'acceptation-rejet	107		
Simulation Monte-Carlo	108		

V Processus stochastiques 109

Première partie

Statistiques univariées

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 $\Theta = \{(\mu, \sigma^2) : \sigma^2 > 0, -\infty < \mu < \infty\}$.
- $\{X_1, \dots, X_n\}$ 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 $\{x_1, \dots, x_n\}$.

Statistiques

Statistique T_n

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

- › Une statistique est donc également une *variable aléatoire* ;
- › Sa distribution est la **distribution d'échantillonnage** qui dépend de :
 1. La statistique.
 2. La taille de l'échantillon.
 3. La distribution sous-jacente des données.

✓ Moyenne échantillonnale \bar{X}

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

- › Estime **sans biais** la moyenne μ ;
- › Si on pose que l'échantillon aléatoire est normalement distribué, $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma}{\sqrt{n}})$;
- › On centre et réduit pour trouver que $T_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$;
- › Si σ^2 est inconnue, on l'estime avec s_n^2 pour obtenir une distribution student — $T_n = \frac{\bar{X} - \mu}{s_n/\sqrt{n}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t_{(n-1)}$ où $W \sim \chi^2_{(n-1)}$.

✓ Variance échantillonnale S_n^2

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

- › Estime **sans biais** la vraie variance σ^2 ;
- › S_n^2 n'est pas normalement distribuée, cependant la statistique $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{(n-1)}$.

✓ Variance empirique $\hat{\sigma}^2$

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

- › Estime **avec biais** la vraie variance σ^2 .
- › Cependant, si la moyenne était connue et que nous n'avions pas à l'estimer avec \bar{x} , alors la variance empirique serait sans biais.

✓ Statistique F

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

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

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

Statistiques bivariées

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

✓ Covariance échantillonnale $cov_{X,Y}$

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

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

✓ Corrélation échantillonnale $r_{X,Y}$

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

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

échantillonnale
corrélation échantillonnale

Vraisemblance

Notation

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

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

où $\mathbf{x}^\top = (x_1, \dots, x_n)$.

$\{X_1, \dots, X_n\}$ 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 $\mathcal{L}(\theta; \mathbf{x})$, il faut songer à ce que représente $f(x; \theta)$.

La fonction de vraisemblance $\mathcal{L}(\theta; \mathbf{x})$ se résume à une différente façon de voir la fonction de densité $f(x; \theta)$.

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

Qualité de l'estimateur

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

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

En bref :

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

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

Estimation ponctuelle

Notation

θ Paramètre inconnu à estimer ;

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

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

Biais

Notation

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

Motivation

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

Biais d'un estimateur

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

Estimateur sans biais

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

Estimateur asymptotiquement sans biais

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

Limitations

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

Variance

Notation

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

Motivation

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

Variance d'un estimateur

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

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

$\text{MSE}_{\hat{\theta}_n}(\theta)$ 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 $E[\hat{\theta}_n]$ —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 $\text{MSE}_{\hat{\theta}}(\theta) = E[(\hat{\theta}_n - \theta)^2]$.

Également, on peut réécrire l'expression comme

$$\text{MSE}_{\hat{\theta}}(\theta) = \text{Var}(\hat{\theta}_n) + [B(\hat{\theta}_n)]^2.$$

- > Il s'ensuit que pour un estimateur non biaisé, $\text{MSE}_{\hat{\theta}}(\theta) = \text{Var}(\hat{\theta}_n)$.
- > En anglais, « *Mean Squared Error (MSE)* ».

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

Convergence

Motivation

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

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

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

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

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

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

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

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

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

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 $\forall \varepsilon > 0$,

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

ou de façon équivalente,

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

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

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

Rappel : loi faible des grands nombres

Notation

$\{X_n\}$ Séquence de variables aléatoires iid avec moyenne μ et variance σ^2 où $\sigma^2 < \infty$;

\bar{X}_n Moyenne empirique.

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

Théorèmes résultant de la convergence en probabilité

Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n + Y_n \xrightarrow{P} X + Y$.

Soit $X_n \xrightarrow{P} X$ et une constante a . Alors $aX_n \xrightarrow{P} aX$.

Soit $X_n \xrightarrow{P} a$ et la fonction $g(\cdot)$ continue à a . Alors $g(X_n) \xrightarrow{P} g(a)$.

Soit $X_n \xrightarrow{P} X$ et la fonction continue $g(\cdot)$. Alors $g(X_n) \xrightarrow{P} g(X)$.

Soit $X_n \xrightarrow{P} X$ et $Y_n \xrightarrow{P} Y$. Alors $X_n Y_n \xrightarrow{P} XY$.

« Consistency »

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

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

Borne Cramér-Rao

Notation

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

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

- La matrice d'information de Fisher pour une seule observation est dénotée $I(\theta)$;
- On obtient une "matrice" lorsque nous estimons plusieurs paramètres et donc θ 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 $I(\theta) = \text{Var}(\ell'(\theta))$.

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

- Pour des données (iid), on obtient que

$$I_n(\theta) = nI(\theta) = -nE\left[\frac{\partial^2}{\partial \theta^2} \ln f(x; \theta)\right].$$

Matrice d'information (de Fisher) de θ

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

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

$$I_n(\theta) = \begin{bmatrix} -nE\left[\frac{\partial^2}{\partial \theta_1^2} \ln f(x; \theta)\right] & -nE\left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln f(x; \theta)\right] \\ -nE\left[\frac{\partial^2}{\partial \theta_1 \partial \theta_2} \ln f(x; \theta)\right] & -nE\left[\frac{\partial^2}{\partial \theta_2^2} \ln f(x; \theta)\right] \end{bmatrix}$$

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 $\text{Var}(\hat{\theta}_n) \geq \frac{1}{I_n(\theta)}$.

Dans le cas multivarié, $\text{Var}(\hat{\theta}_j) \geq I_n^{-1}(\theta)_{jj}$.

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 ;

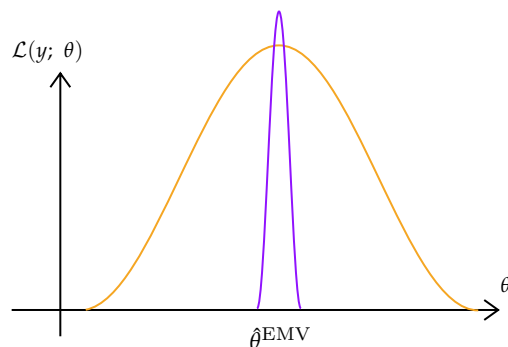
$\frac{\partial^2}{\partial \theta^2} \mathcal{L}(\theta)$: Représente la concavité de la fonction ; on peut y penser comme sa forme.

L'estimateur du maximum de vraisemblance (EMV) $\hat{\theta}_n^{\text{EMV}}$ 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 $\theta - \theta^{\text{EMV}} = \theta$.

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

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

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



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

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

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

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 $\hat{\theta}_n^{\text{EMV}} \xrightarrow{\text{a.s.}} \mathcal{N}\left(\mu = \theta, \sigma^2 = \frac{1}{I(\theta)}\right)$ 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 $\text{eff}(\hat{\theta}_n) = \frac{1/I_n(\theta)}{\text{Var}(\hat{\theta})}$.

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 $\text{eff}(\hat{\theta}_n, \tilde{\theta}_n) = \frac{\text{Var}(\tilde{\theta}_n)}{\text{Var}(\hat{\theta}_n)}$.

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

Estimateur non biaisé à variance minimale (MVUE)

Motivation

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

Estimateur non biaisé à variance minimale (MVUE)

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

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

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

Limitations

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

Estimation par intervalles

Contexte

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

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

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

Intervalle de confiance

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

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

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

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

Pivot (« *pivotal quantity* »)

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

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

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

théorème central limite et donc ne dépend pas de μ .

Méthode du pivot

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

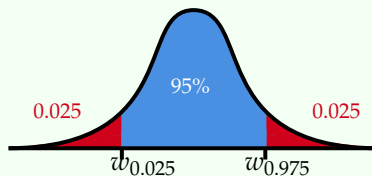
Exemple d'application de la méthode du pivot

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

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

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

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



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

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

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

- 4 On établit la probabilité désiré

$$\Pr(w_{0.025} \leq W \leq w_{0.975}) = 0.95$$

$$\Pr(0.1878 \leq W \leq 78.4968) = 0.95$$

$$\Pr\left(0.1878 \leq \frac{\theta}{X} \leq 78.4968\right) = 0.95$$

$$\Pr(0.1878 \times X \leq \theta \leq 78.4968 \times X) = 0.95$$

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

Contexte

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

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

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

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

Intervalles de confiance

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

Intervalles sur la moyenne

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

1 échantillon

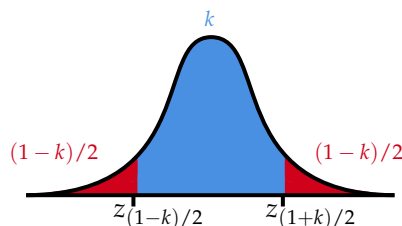
Pour un échantillon aléatoire de taille n avec moyenne μ et variance σ^2 ,

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

Donc, lorsque la variance est connue,

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

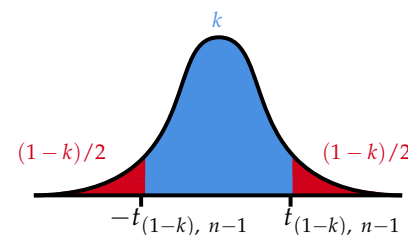
Visuellement, nous avons une aire de $(1-k)/2$ dans les deux queues et les points $z_{(1-k)/2}$ et $z_{(1+k)/2}$:



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

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

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



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

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

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

2 échantillons

Pour 2 échantillons aléatoires indépendants de tailles n_1 et n_2 ,

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

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

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

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

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

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

Intervalles sur les proportions

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

1 échantillon

Pour un échantillon aléatoire de taille n tiré d'une distribution Bernoulli, on déduit du théorème centrale limite que, lorsque n est grand, le pivot

$$\frac{\hat{q} - q}{\sqrt{\frac{\hat{q}(1-\hat{q})}{n}}} \xrightarrow{D} Z.$$

Donc,

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

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

2 échantillons

Pour 2 échantillons aléatoires de tailles n_1 et n_2 , on déduit du théorème centrale limite que, lorsque n_1 et n_2 sont grands, le pivot

$$\frac{(\hat{q}_1 - \hat{q}_2) - (q_1 - q_2)}{\sqrt{\frac{\hat{q}_1(1-\hat{q}_1)}{n_1} + \frac{\hat{q}_2(1-\hat{q}_2)}{n_2}}} \xrightarrow{D} Z.$$

Donc,

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

Intervalle sur la variance

« Pooled Estimator »

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

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

1 échantillon

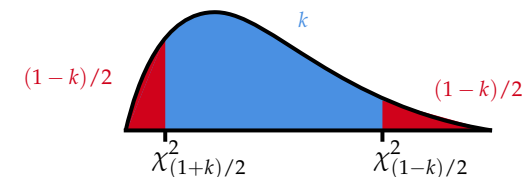
Pour un échantillon aléatoire de taille n tiré d'une distribution normale, le pivot

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{(n-1)}^2.$$

Donc,

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

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



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

2 échantillons

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

Donc,

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

Tests d’hypothèses

Hypothèses

Contexte

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

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 $\Theta_0 \cup \Theta_1 = \Theta$;

H_0 Hypothèse nulle.

H_1 Hypothèse alternative.

Test d’hypothèse

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

$H_0 : \theta \in \Theta_0$

vs

 $H_1 : \theta \in \Theta_1$

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

Hypothèse nulle

Représente généralement le statu quo jusqu'à preuve contraire.

Hypothèse alternative

Représente généralement un changement du statu quo.

✓ Décision du test d'hypothèse

Soit on :

- ① Ne peut rejeter l'hypothèse nulle H_0 .
- ② On rejete l'hypothèse nulle H_0 pour l'hypothèse alternative H_a .

Contexte

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

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

Terminologie

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

- Par exemple, $H_0 : q = 0.50$ —on connaît la valeur exacte du paramètre q pour une distribution Bernoulli.

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

- Par exemple, $H_1 : q \neq 0.50$ —on ne connaît pas la valeur exacte du paramètre q , il pourrait être n'importe quel chiffre sauf 0.50.
- Par exemple, pour une distribution normale de moyenne μ et variance inconnue σ^2 , $H_0 : \mu = 0.50$ —on ne connaît pas la variance et donc la distribution n'est pas entièrement spécifiée.

Exemple du laissez-passer universitaire (LPU)

Par exemple, on veut savoir si les étudiants utilisent l'autobus (oui ou non) avant et après l'implantation du LPU.

On **pose** que la proportion des gens qui utilisent l'autobus est $q = 0.44$.

Il y a deux types de **tests** qu'on peut faire,

- Tester si l'utilisation est différente est un test "**bilatéral**", car on teste si

elle a augmenté *ou* diminuée ;

$$H_0 : q = 0.44$$

$$H_1 : q \neq 0.44$$

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

$$H_0 : q = 0.44$$

$$H_1 : q > 0.44$$

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

L'hypothèse :

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

alternative dans le cas d'un test :

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

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

Région et valeur critique

Région critique

Notation

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

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

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

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

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

Exemple du laissez-passer universitaire (LPU)

On reprend l'exemple du LPU.

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

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

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

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

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

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

Erreurs de test

Contexte

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

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

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

Certitude du test

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

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 signifiante* 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 mélangés.

Terminologie

p La **valeur p** du test.

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

La valeur p du test

On peut définir la valeur p de plusieurs façons, soit comme :

1. la probabilité d'un événement tout aussi (ou plus) rare sous l'hypothèse nulle.
2. la **taille** de la région critique \mathcal{C} .
 - > C'est-à-dire, l'*aire* de la région de rejet de l'hypothèse nulle H_0 alors qu'elle est vraie.
3. le **seuil de signifiante**.
4. la **probabilité d'une erreur de type I**.
 - > C'est-à-dire, la probabilité de rejeter H_0 alors qu'elle est vraie

Le seuil de signifiante α du test

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

- > C'est la même idée qu'avec les intervalles de confiance.
- > En anglais, « *threshold for significance* ».

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

En mots,

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

Valeur p vs seuil α

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

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

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

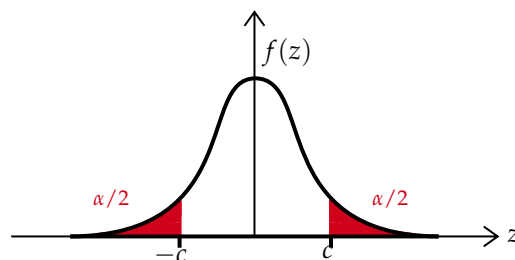
Comparaison	Décision
$p \leq \alpha$	Rejete H_0
$p > \alpha$	Ne rejete pas H_0

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

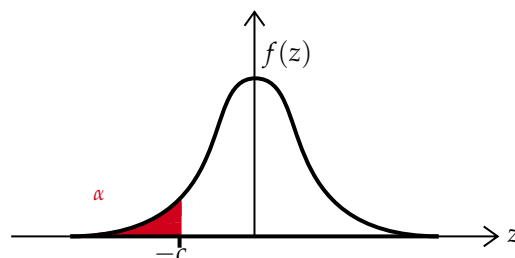
Résumé graphique des régions critiques

On résume les régions critiques pour les 3 type de test :

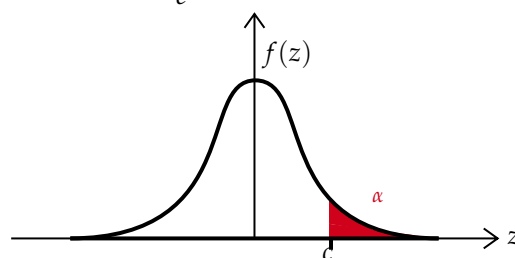
Bilatéral
$p = \Pr(T \geq t H_0 \text{ est vrai})$
$\alpha = \Pr(T \geq c H_0 \text{ est vrai})$



Unilatéral à la gauche
$p = \Pr(T \leq t H_0 \text{ est vrai})$
$\alpha = \Pr(T \leq -c H_0 \text{ est vrai})$



Unilatéral à la droite
$p = \Pr(T \geq t H_0 \text{ est vrai})$
$\alpha = \Pr(T \geq c H_0 \text{ est vrai})$



Tests sur la moyenne

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

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

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

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

Rappels

Avant de détailler les tests, nous effectuons quelques rappels.

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

Convergence en distribution

Notation

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

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

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

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

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

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

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

Rappel : théorème centrale limite

Notation

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

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

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

Le théorème s'applique directement pour 1 échantillon. Pour 2 échantillons, s'ils sont indépendants, on généralise pour trouver que la différence des moyennes empiriques

$\bar{X}_1 - \bar{X}_2 \xrightarrow{D} \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$. S'ils sont appariés, on applique le théorème

centrale limite à la différences des observations pour trouver que $\bar{D} \xrightarrow{D} \mathcal{N}(\mu_d, \sigma_D^2)$.

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

Rappel : Loi de Student

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

Également, pour un échantillon aléatoire (X_1, X_2, \dots, X_n) tiré d'une distribution normale de moyenne μ et variance σ^2 inconnue, la statistique

$$T_n = \frac{\bar{X}_n - \mu}{\sqrt{s_n^2/n}} \sim t_{(n-1)}.$$

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

Le théorème centrale limite a l'avantage de s'appliquer **peu importe la distribution de l'échantillon aléatoire** au dépend de s'appliquer **juste lorsque n est grand**.

En revanche, la loi de Student s'applique **juste si la distribution de l'échantillon aléatoire est normale** mais ne nécessite pas que n soit grand.

1 échantillon

Pour un échantillon aléatoire de taille n ,

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

Les valeurs critiques sont :

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



Explication des percentiles de loi normal vs de loi de Student

On dénote :

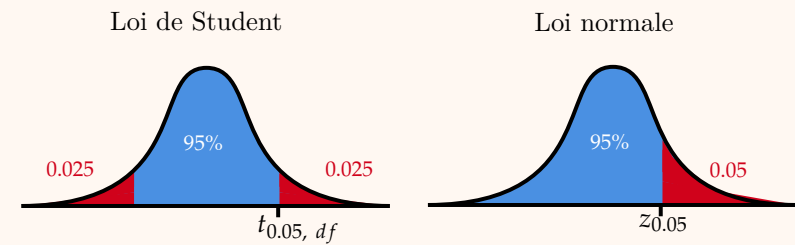
z_q 100^qe percentile de la loi normale standard.

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

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

- > La table de la loi *normale* standard comprend les **probabilités cumulatives** (e.g. $\Pr(Z \leq z)$).
- > La table de la loi de *Student* comprend les **probabilités des deux queues** (e.g. $\Pr(|T| \geq t)$).

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



2 échantillons

Contexte

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

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

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

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

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

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

Contexte

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

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

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

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

Donc,

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

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

Tests sur les proportions

Contexte

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

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

1 échantillon

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

2 échantillons

Pour n_1 et n_2 qui sont grands, le théorème centrale limite implique que $\bar{X}_1 - \bar{X}_2 \xrightarrow{D} \mathcal{N}\left(q_1 - q_2, \frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}\right)$. Il s'ensuit que $t = \frac{\hat{q}_1 - \hat{q}_2 - h}{\sqrt{\frac{q_1(1-q_1)}{n_1} + \frac{q_2(1-q_2)}{n_2}}}$ où $T \sim \mathcal{N}(0,1)$.

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

Tests sur la variance

Rappels

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

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

Rappel : Loi du khi carré

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

Également, pour un échantillon aléatoire (X_1, X_2, \dots, X_n) tirée d'une distribution normale de moyenne μ et variance σ^2 , la statistique

$$T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{(n-1)}^2.$$

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

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

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

Rappel : Loi de Fisher-Snedecor (F)

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

Soit les 2 variables aléatoires indépendantes $W_1 \sim \chi_{(v_1)}^2$ et $W_2 \sim \chi_{(v_2)}^2$, alors

$Y = \frac{W_1/v_1}{W_2/v_2} \sim \mathcal{F}_{(v_1, v_2)}$. De plus, la loi de Fisher a la propriété intéressante que $Y^{-1} \sim \mathcal{F}_{(v_2, v_1)}$.

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

Également, pour deux échantillons aléatoires indépendantes de tailles n_1 et n_2 qui sont tirés de distributions normales avec variances σ_1^2 et σ_2^2 , la statistique

$$T_n = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim \mathcal{F}_{(n_1-1, n_2-1)}.$$

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

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

1 échantillon

Pour un échantillon aléatoire de taille n tiré d'une **distribution normale** de variance σ^2 , la loi du khi carré implique que $T_n = \frac{(n-1)S_n^2}{\sigma^2} \sim \chi_{(n-1)}^2$. Il s'ensuit que

$$t = \frac{(n-1)s_n^2}{h}.$$

Les régions critiques sont :

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

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

✓ Percentiles de la loi du khi carré

On dénote :

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

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

2 échantillons

Contexte

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

Pour 2 échantillons aléatoires indépendants de tailles n_1 et n_2 qui sont normalement distribués avec variances σ_1^2 et σ_2^2 , la loi de Fisher implique que

$$T_n = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{(n_1-1, n_2-1)}. \text{ Il s'ensuit que } t = \frac{s_1^2}{s_2^2} \times \frac{1}{h}.$$

Les régions critiques et l'hypothèse alternative pour $H_0 : \frac{\sigma_1^2}{\sigma_2^2} = h$ sont :

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

✓ Percentiles de la loi de Fisher

On dénote :

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

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

Puissance d'un test

📖 La puissance d'un test

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

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

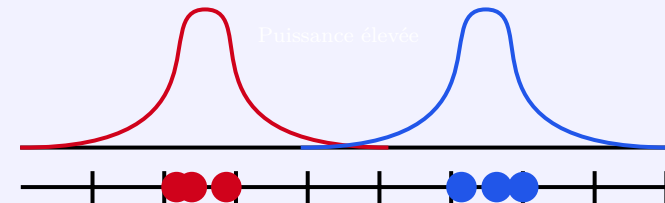
Facteurs influençant la puissance

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

☰ La forme de la distribution

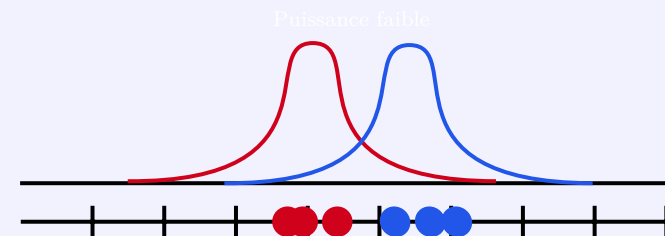
Si les deux distributions sont :

> 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** :



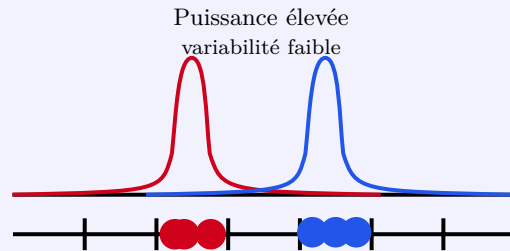
- La probabilité **d'incorrectement** rejeter l'hypothèse nulle (que les deux échantillons proviennent d'une même distribution) est élevée ;

- 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.

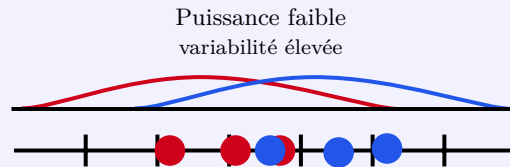
La variabilité des données

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. Entre autres, 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

Contexte

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

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

Fonction de puissance

La fonction de puissance correspond à la probabilité de rejeter l'hypothèse nulle H_0 si la **vraie** valeur du paramètre est $\theta \in \Theta$:

$$\gamma(\theta) = \Pr\{(X_1, \dots, X_n) \in \mathcal{C} | \theta\}.$$

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

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

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

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

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

Tests optimaux (les plus puissants)

Introduction

Notation

δ (Procédure de) test ;

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

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

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

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

Λ Ratio de vraisemblance.

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

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

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

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

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

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

Test le plus puissant

Tests optimaux, alias tests les plus puissants

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

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

- 1 Les hypothèses doivent être simples :

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

- 2 La région, alias le sous-ensembles de \mathcal{S} , doit être de taille α . En autres mots, la probabilité de incorrectement rejeter l'hypothèse nulle (erreur de type I), $\alpha(\delta^*)$, est de α :

$$\alpha(\delta^*) = \Pr((X_1, \dots, X_n) \in \mathcal{C} | \theta = \theta_0) = \alpha$$

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

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

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

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

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

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

Exemple avec une distribution binomiale

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

suivantes :

$$H_0 : \theta = 0.50$$

$$H_1 : \theta = 0.75$$

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

- 1 La première étape est d'identifier les régions de valeurs pour lesquels la variable aléatoire a une probabilité de 0.125 d'y être contenue.

(a) On fait un tableau des valeurs sous les deux hypothèses :

FMP	$x = 0$	$x = 1$	$x = 2$	$x = 3$
$p(x \theta = 0.50)$	0.125	0.375	0.375	0.125
$p(x \theta = 0.75)$	0.015625	0.140625	0.421875	0.421875

(b) On trouve que les sous-ensembles de \mathcal{S} correspondants sont $\mathcal{A}_1 = \{x = 0\}$ et $\mathcal{A}_2 = \{x = 3\}$.

➤ C'est-à-dire, $\Pr(X \in \mathcal{A}_1|\theta = 0.50) = \Pr(X \in \mathcal{A}_2|\theta = 0.50) = 0.125$ et il n'y a pas d'autres sous-ensembles de \mathcal{S} avec la même "taille" de 0.125.

- 2 On doit identifier laquelle de \mathcal{A}_1 ou \mathcal{A}_2 est la région critique \mathcal{C} optimale de taille α pour tester H_0 contre H_1 .

- 3 On trouve la probabilité de correctement rejeter l'hypothèse nulle (faire partie de la région critique) pour les deux régions :

$$\Pr(X \in \mathcal{A}_1|\theta = 0.75) = \Pr(X = 0|\theta = 0.75) = 0.015625$$

$$\Pr(X \in \mathcal{A}_2|\theta = 0.75) = \Pr(X = 3|\theta = 0.75) = 0.421875$$

- 4 On compare les probabilités de correctement rejeter l'hypothèse nulle aux probabilités de incorrectement rejeter l'hypothèse nulle :

➤ Dans le premier cas :

$$\Pr(X \in \mathcal{A}_1|\theta = 0.75) = 0.015625 < \Pr(X \in \mathcal{A}_1|\theta = 0.50) = 0.125$$

rejeter H_0 alors que H_0
est faux ($\theta = 0.75$)

rejeter H_0 alors que H_0
est vraie ($\theta = 0.50$)

➤ Dans le deuxième cas :

$$\Pr(X \in \mathcal{A}_2|\theta = 0.75) = 0.421875 > \Pr(X \in \mathcal{A}_2|\theta = 0.50) = 0.125$$

rejeter H_0 alors que H_0
est faux ($\theta = 0.75$)

rejeter H_0 alors que H_0
est vraie ($\theta = 0.50$)

- 5 Puisque le premier sous-ensemble a une probabilité plus élevée de incorrectement rejeter l'hypothèse nulle (erreur de type I) que de correctement la rejeter, on choisit le deuxième comme région critique : $\mathcal{C} = \mathcal{A}_2 = \{x = 3\}$.

Également, on peut observer que la région est choisie en incluant dans \mathcal{C} les points x pour lesquels $p(x|\theta = 0.50)$ est petite par rapport à $p(x|\theta = 0.75)$.

Il s'ensuit que le ratio $\frac{p(x|\theta=0.50)}{p(x|\theta=0.75)}$ évalué à $x = 5$ est un minimum.

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

Théorème de Neyman-Pearson

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

$$H_0 : \theta = \theta_0$$

$$H_1 : \theta = \theta_1$$

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

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

1. si l'échantillon aléatoire est contenu dans la région critique, $x \in \mathcal{C}$, alors

$$\frac{\mathcal{L}(\theta_0; \mathbf{x})}{\mathcal{L}(\theta_1; \mathbf{x})} \leq k.$$

2. si l'échantillon aléatoire n'est pas contenu dans la région critique, $x \in \mathcal{C}^c$,

$$\text{alors } \frac{\mathcal{L}(\theta_0; \mathbf{x})}{\mathcal{L}(\theta_1; \mathbf{x})} \geq k.$$

3. $\alpha = \Pr \{ (X_1, \dots, X_n) \in \mathcal{C} | \theta_1 \} = \alpha(\delta^*)$.

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

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

Note L'approche est typiquement d'écrire le ratio puis de tenter de trouver une statistique afin de calculer une probabilité avec sa distribution. Voir les exemples ci-dessous.

Exemple cas continu

Pour l'échantillon aléatoire $\mathbf{X} = (X_1, \dots, X_n)$ tiré d'une distribution normale $\mathcal{N}(\mu = \theta, \sigma^2 = 0)$, on fixe les hypothèses suivantes :

$$H_0 : \theta = 0$$

$$H_1 : \theta = 1$$

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

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

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

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

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

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

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

Par exemple, sous l'hypothèse nulle $\bar{X} \stackrel{H_0}{\sim} \mathcal{N}(0, 1/n)$,

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

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

Exemple cas discret

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

$$H_0 : \theta = 1$$

$$H_1 : \theta = 2$$

On souhaite identifier meilleure région critique pour les tests de taille 0.06. Cependant, puisque la variable aléatoire est discrète, on ne peut pas manipuler le ratio des vraisemblances. En lieu,

- 1 On produit le tableau des valeurs de la fonction de masse des probabilités sous les deux hypothèses :

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

- 2 On trouve les ratios dénotés Λ :

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

- 3 Puisqu'on cherche à rejeter l'hypothèse nulle, il s'ensuit qu'on cherche des petites valeurs du ratio. Donc, on trie en ordre croissant les valeurs de x :

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

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

Test uniformément le plus puissant

Motivation

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

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

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

Test uniformément le plus puissant

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

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

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

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

Exemple avec une distribution normale

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

$$H_0 : \theta = 1$$

$$H_1 : \theta > 1$$

1 On trouve le ratio Λ :

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

2 On observe que le ratio Λ décroît alors que $\sum x_i^2$ augmente.

3 Puisque le ratio est décroissant, un test uniformément le plus puissant aura une région critique définie par $\sum x_i^2 \geq k$, où k est choisi selon la taille α .

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

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

On résume la région critique la taille du test selon le ratio de vraisemblance pour le test $H_0 : \theta = h$:

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

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

L'approche prise dans l'exemple ci-dessus est de poser un $\theta_0 =$ un nombre puis de poser une constante $\theta_1 > \theta_0$. Cependant, on peut généraliser l'approche au cas où les deux hypothèses sont composées (e.g. $H_0 : \theta \leq 2$ et $H_1 : \theta > 2$), on définit le **ratio de vraisemblance monotone**.

≡ Ratio de vraisemblance monotone

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

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

Tests d'adéquation

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

Notation

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

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

Contexte

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

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

Test de Kolmogorov-Smirnov

Motivation

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

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

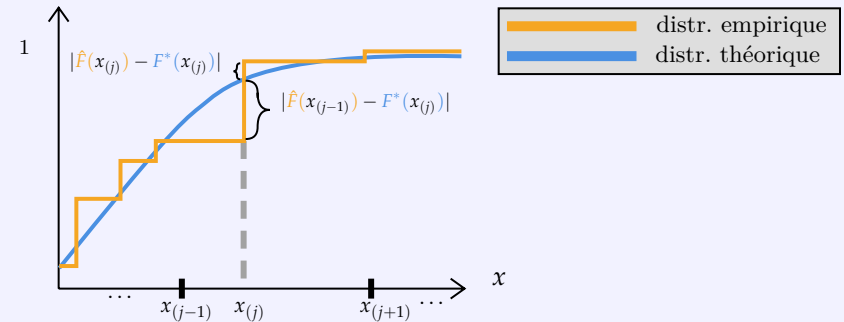
Test de Kolmogorov-Smirnov

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

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

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

Visuellement :



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

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

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

Motivation

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

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

Test d'adéquation du khi carré

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

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

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

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

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

$$H_0 : nq_j = n_j \quad H_1 : nq_j \neq n_j$$

La statistique du test calcule les divergences pour chacun des intervalles :

$$t = \sum_{j=1}^k \frac{(n_j - nq_j)^2}{nq_j} \quad \text{où la statistique est approximativement distribué selon}$$

la loi du khi carré : $T_n \sim \chi^2_{(k-1-r)}$.

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

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

Motivation

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

Tableau de contingence

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

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

Test d'indépendance du khi carré

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

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

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

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

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

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

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

Donc, le tableau de contingence est de la forme :

		Second variable				Total
		Cat. 1	Cat. 2	...	Cat. b	
First Variable	Cat. 1	n_{11}	n_{12}	...	n_{1b}	$n_{1.}$
	Cat. 2	n_{21}	n_{22}	...	n_{2b}	$n_{2.}$

	Cat. a	n_{a1}	n_{a2}	...	n_{ab}	$n_{a.}$
Total		$n_{.1}$	$n_{.2}$...	$n_{.b}$	n

La statistique du test a le même raisonnement que la statistique pour le test

d'adéquation du khi carré : $t = \frac{1}{n} \sum_{i=1}^a \sum_{j=1}^b \frac{(n_{ij}n - n_{i.}n_{.j})^2}{n_{i.}n_{.j}}$ où la statistique est approximativement distribué selon la loi du khi carré : $T_n \sim \chi_{(a-1)(b-1)}^2$.

Test du rapport de vraisemblance

Motivation

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

Test du rapport de vraisemblance

Les hypothèses sont

H_0 : les données proviennent d'une distribution A

H_1 : les données proviennent d'une distribution B

où la distribution A est un cas spécial de la distribution B . L'exemple intuitif est la distribution exponentielle qui est un cas spécial de la distribution gamma avec $\alpha = 1$. Cependant, plus généralement, on peut penser à une distribution avec des paramètres prédéterminés comme une simplification de la même distribution avec des paramètres inconnus (e.g. normale avec μ et σ^2 inconnus vs normale avec $\mu = 2$ et σ^2 inconnu).

Il s'ensuit que la distribution B est plus complexe que la distribution A et donc que B a plus de "paramètres libres". L'idée du test est que nous préférons un modèle plus simple (distribution A) mais utiliserons un modèle plus complexe (distribution B) s'il y a une amélioration suffisamment importante.

On dénote donc :

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

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

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

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

Note La distribution asymptotique du khi carré dépend sur :

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

Note Voir la section <section pas encore faite> du chapitre de *Modèles linéaires en actuariat* pour l'application du test du rapport de vraisemblance des les modèles linéaires généralisés.

Statistiques exhaustives

Soit l'échantillon aléatoire (X_1, \dots, 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 $f(x_1, \dots, x_n | t) = h(x_1, \dots, x_n)$ 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

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

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

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

Limitations

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

Théorème de factorisation de Fisher-Neyman

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

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

Cas multivarié

Pour $\theta = (\theta_1, \dots, \theta_r)$, les statistiques T_n^1, \dots, T_n^r sont **conjointement exhaustives** pour θ si

$$f(x_1; \theta) \times \dots \times f(x_n; \theta) = h_1(t_1, \dots, t_r; \theta) \times h_2(x_1, \dots, x_n) \text{ où}$$

- 1 $h_1(t_1, \dots, t_r; \theta)$ dépend de l'échantillon seulement par les statistiques T_n^1, \dots, T_n^r .
- 2 $h_2(x_1, \dots, x_n)$ ne dépend pas des paramètres θ .
- 3 $\forall i = 1, 2, \dots, n \ x_i \in \mathbb{R}$.

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

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

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 $(X_{(1)}, \dots, X_{(n)})$. 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 $E[g(T_n)] = 0$ **implique** que $\Pr(g(T_n) = 0) = 1$.

- > Il s'ensuit qu'il est possible que $E[g(T_n)] = 0$ sans que la distribution de la statistique provienne d'une famille complète de distributions.
- > Le fait qu'une statistique soit complète veut dire que toute fonction $g(\cdot)$ qui entraîne la moyenne de $g(T_n)$ à être nulle doit être une fonction « *that maps to 0* ».
- > Il est hors du cadre de l'examen de devoir prouver que des statistiques sont complètes.
- > Le fait d'être complet implique qu'il existe une seule fonction T_n qui est un estimateur non biaisé de θ ; alias, $g(T_n)$ est **unique**.

Théorème de Lehmann-Scheffé

Si :

- 1 La statistique T_n est une statistique exhaustive pour θ .
- 2 La distribution de T_n provient d'une famille de distributions complète.
- 3 Il y existe une fonction unique $\varphi(\cdot)$ de T_n tel que $\varphi(T_n)$ est un 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 $E_{\hat{\theta}_n}[\hat{\theta}_n | T_n]$.

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 $\tilde{\theta}_n = E_{\hat{\theta}_n}[\hat{\theta}_n | T_n]$ est une *statistique* non biaisé de θ avec $\text{Var}(\tilde{\theta}_n) \leq \text{Var}(\hat{\theta}_n)$.

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

Critère de Lehmann-Scheffé

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

Exemple Bernoulli

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

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

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 $f(x_1, \dots, x_n) = h_1(\sum_{i=1}^n a(x_i); \theta) h_2(x_1, \dots, x_n)$.

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

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

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

Statistiques d'ordre

Principes fondamentaux

k^e statistique d'ordre

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

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

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

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

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

$$f_{X_{(k)}}(x) = \frac{\Gamma(n+1)}{\Gamma(k)\Gamma(n-k+1)} \left(\frac{x}{\theta}\right)^k \left(10\frac{x}{\theta}\right)^k \frac{1}{x}, \quad 0 \leq x \leq \theta$$

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

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

Si l'échantillon aléatoire est tiré d'une **distribution exponentielle** de moyenne θ ,

alors $E[X_{(k)}] = \theta \sum_{i=n-k+1}^n \frac{1}{i}$.

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

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

Minimum

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

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

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

Maximum

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

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

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

Cas spéciaux

Si l'échantillon aléatoire est tiré d'une **distribution uniforme** sur $[a, b]$, alors

$$E[X_{(k)}] = a + \frac{k(b-a)}{n+1}. \text{ De plus, si } X \sim U(0, \theta) \text{ alors}$$

Autres statistiques

Nous définissons également quelques autres statistiques d'intérêt :

≡ L'étendue (« range »)

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

Contexte

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

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

≡ La mi-étendue (« midrange »)

La moyenne entre du minimum et du maximum d'un échantillon :

$$M = \frac{X_{(n)} + X_{(1)}}{2}.$$

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

- › La moyenne arithmétique considère les données observées et calcule leur moyenne.
 - Il s'ensuit qu'elle ne considère pas les chiffres qui ne sont pas observés.
- › La mi-étendue considère **tous** les chiffres—observés ou non—entre la plus grande et la plus petite valeur d'un échantillon, puis en prend la moyenne.

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

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

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

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

Exemple de contexte pour les statistiques

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

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

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

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

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

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

≡ Médiane

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

- › La moitié des données sont supérieures et inférieures à la médiane.
- › L'utilité de la médiane est qu'elle n'est pas aussi sensible aux données aberrantes que la moyenne.

Distribution conjointe

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

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

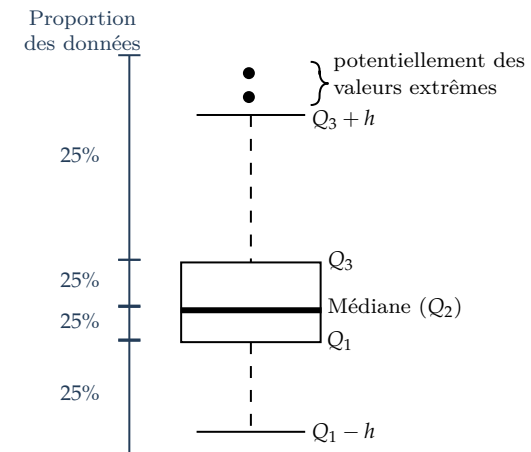
Graphiques

Diagramme en boîte (« *boxplot* »)

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

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

Visuellement :



Remarques :

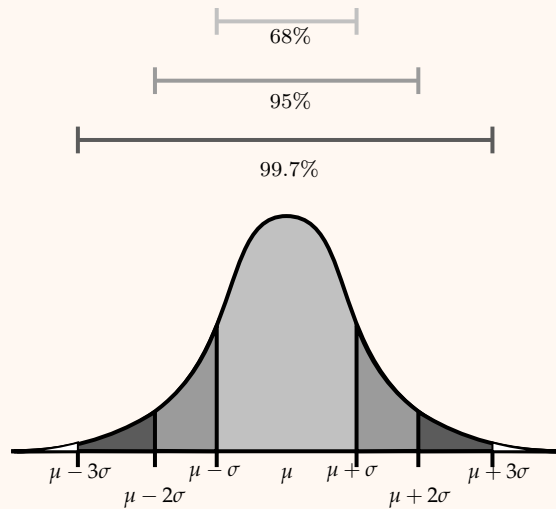
- > La médiane (ligne dans la boîte) correspond au point où la moitié des données sont au-dessus et l'autre moitié en dessous.
- > La boîte est délimitée par le premier et le troisième quartile.
 - Il s'ensuit que la boîte contient la moitié des données.
 - De plus, 25% des données sont contenues entre la borne *supérieure* de la boîte et la médiane (l'autre 25% est contenu entre la borne *inférieure* et la médiane).
- > Les « moustaches » de la boîte sont tracées à un pas h des quartiles où $h = 1.5 \cdot (Q_3 - Q_1)$.
 - Les points à l'extérieur de ces bornes sont potentiellement des données aberrantes.
 - Plus *l'écart interquartile* ($Q_3 - Q_1$) est élevé, plus la boîte sera large et, par conséquent, plus les moustaches seront loin de la médiane.

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

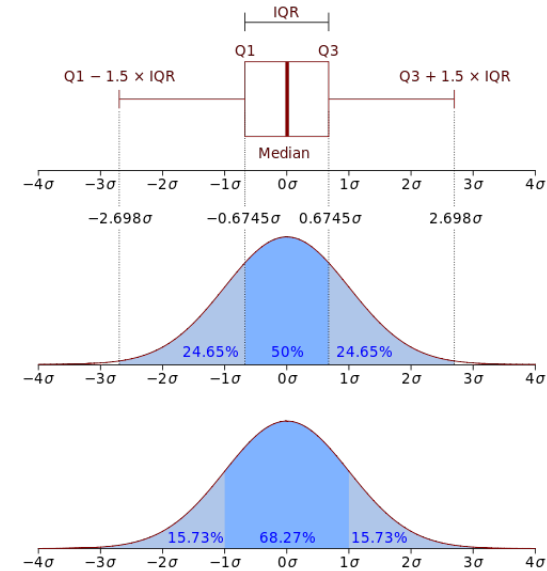
En premier, on définit la règle du 68-95-99.7.

✓ Règle du 68-95-99.7

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



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



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

Note Voir [Diagramme en boîte](#) pour l'interprétation des diagrammes en boîte.

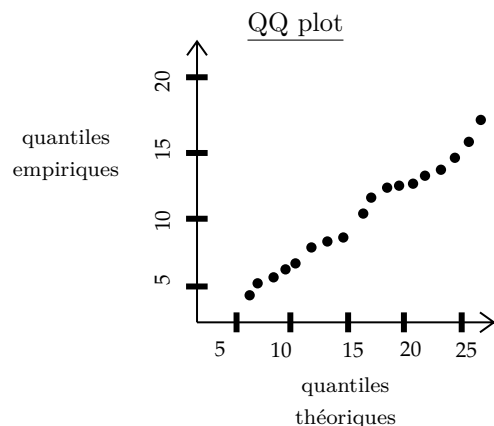
Diagramme quantile-quantile (« *Q-Q plot* »)

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

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

- › Si les données sont normalement distribuées, on peut centrer et réduire pour obtenir la loi normale standard Z ;
 - Ceci correspond à un diagramme quantile-quantile **normale** ;
- › Autrement, le diagramme quantile-quantile est tracé en estimant les paramètres de la distribution avec l'échantillon de données.

Par exemple :



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

Construction d'estimateurs

Introduction

Contexte

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

La méthode vue dans le cadre du cours de statistique (et de l'examen MAS-I) est la méthode dite « **fréquentiste** ». Le cours de 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 comportent 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 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, $\mu'_k = E[X^k]$.

$\hat{=}$ 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 θ .

L'estimation de θ est donc la solution aux r équations suivantes :

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

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 $E[X \wedge u]$.

Méthode du «Percentile Matching »

Notation

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

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 θ .

L'estimation de θ est donc la solution aux r équations suivantes :

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

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

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 définit $\mathcal{L}(\theta; \mathbf{x}) = \prod_{i=1}^n f(x_i; \theta)$ et $\ell(\theta; \mathbf{x}) = \sum_{i=1}^n \ln f(x_i; \theta)$, puis on calcule

$$\hat{\theta}^{\text{EMV}} = \max_{\theta} \{\mathcal{L}(\theta; \mathbf{x})\} = \max_{\theta} \{\ln \mathcal{L}(\theta; \mathbf{x})\}.$$

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

Raccourcis

Si la fonction de vraisemblance est de la forme :

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

> $\mathcal{L}(\lambda) = \lambda^a e^{-\lambda b}$ 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 $\hat{\theta}^{\text{EMV}}$:

$$g(\hat{\theta}^{\text{EMV}}) \text{ est l'EMV de } g(\theta).$$

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 $g(\lambda) = e^{-\lambda}$ est $g(\hat{\lambda}) = e^{-\hat{\lambda}}$.

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 « consistant »* 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, \dots, 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 $\hat{\theta}_n^{\text{EMV}}$ converge en distribution vers une distribution normale avec une moyenne θ et une variance égale à la borne de Cramér-Rao : $\hat{\theta}_n^{\text{EMV}} \approx \mathcal{N}\left(\theta, \frac{1}{I_n(\theta)}\right)$.

En termes mathématiques,

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

La normalité de la distribution asymptotique implique :

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

3. $\hat{\theta}_n^{\text{EMV}}$ est, pour des grands échantillons, approximativement normalement distribué avec moyenne θ et variance $1/I_n(\theta)$.
4. $\hat{\theta}_n^{\text{EMV}}$ 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 est 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 $i = 1, 2, \dots$.

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 $|\frac{\partial^3}{\partial \theta^3} \ln f(x; \theta)| \leq M(x)$ où $E_{\theta_0}[M(X)] < \infty$ 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 $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_r)$, on trouve que :

$$\hat{\theta}^{\text{EMV}} \approx \mathcal{N}(\theta, I(\theta)^{-1}).$$

Deuxième partie

Modèles linéaires en actuariat

Apprentissage statistique

Apprentissage statistique

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

Variables d'un modèle d'apprentissage statistique

Notation

Y variable réponse.

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

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

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

p Nombre de variables explicatives.

Variable réponse

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

Variables explicatives

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

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

Variable quantitative

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

1 Variable continue

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

- › les montants de perte d'un accident d'automobile,
- › le temps avant qu'une réclamation d'assurance soit réglée,
- › la probabilité de précipitations, etc.

2 Variable de comptage

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

- > le nombre de d'accidents d'automobile,
- > le nombre d'étudiants dans une salle de cours, etc.

≡ Variable qualitative

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

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

1 Variable nominale

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

- > le programme d'étude d'un étudiant (actuariat, informatique, etc.),
- > la couleur d'une voiture (rouge, bleu, vert, etc.),
- > le fabricant d'une voiture (Toyota, Subaru, etc.), etc.

2 Variable ordinale

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

- > la sévérité d'un accident de 1 à 5,
- > le degré de risque d'incendie de bas à élevé, etc.

Notes :

- > Si une variable qualitative a seulement 2 classes, on l'appelle une **variable binaire**.
 - Par exemple, la variable "sexe de l'assuré" prenant comme valeur *homme* ou *femme*.
- > Une variable explicative qualitative est appelée un **facteur**.

Types de modèles d'apprentissage statistique

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

≡ Apprentissage supervisé

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

Par exemple,

- > prédire le nombre de buts marqués par un joueur de la LNH,
- > prédire la quantité de neige qui va tomber l'année prochaine,
- > prédire si un client est satisfait ou pas suite à un appel, etc.

≡ Apprentissage non supervisé

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

Par exemple,

- > analyser ce qu'achètent les consommateurs d'une certaine région,
- > évaluer les caractéristiques des étudiants en défaillance dans leurs prêts étudiants,
- > évaluer la composition du sol dans un environnement pollué, etc.

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

Problèmes d'apprentissage supervisé

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

1 Régression

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

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

2 Classification

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

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

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

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

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

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

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

Objectifs de l'apprentissage supervisé

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

1 Prévion

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

2 Inférence

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

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

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

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

Précision des modèles d'apprentissage statistique

Notation

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

Erreur quadratique moyenne

Contexte

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

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

L'erreur quadratique moyenne (EQM) mesure la **précision** du modèle où $MSE = E[(Y - \hat{Y})^2]$. Typiquement, on calcule l'EQM avec l'échantillon de données où

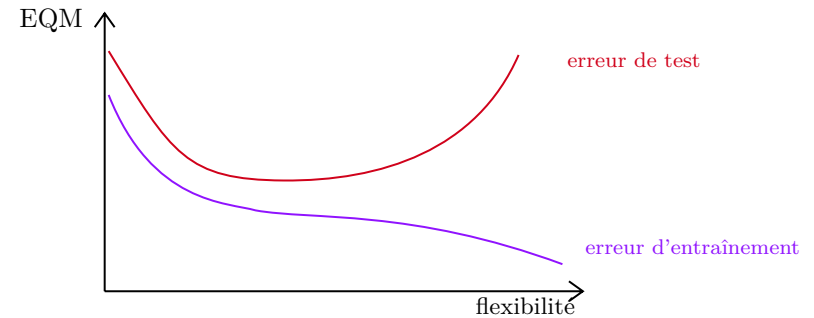
$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}.$$

Si on calcule l'EQM avec l'échantillon de données de **test** on obtient l'EQM de test.

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

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

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



Donc,

- > puisque est optimisé pour les données d'entraînement, il sera toujours inférieur à l'EQM de test.
- > l'EQM d'entraînement décroît lorsque la flexibilité (et donc, la complexité aussi) du modèle augmente.
- > l'EQM de test est concave ce qui veut dire que le meilleur modèle fait un **compromis** entre la *flexibilité* et la *précision* du modèle.

Compromis biais-variance

Contexte

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

Le biais de \hat{f} mesure le degré auquel \hat{f} est proche de f . Un modèle très flexible s'ajuste bien aux données et comporte un faible biais. Par exemple, un « *spline* » a un faible biais et une variance élevée alors qu'une régression linéaire a un biais élevée et une faible variance.

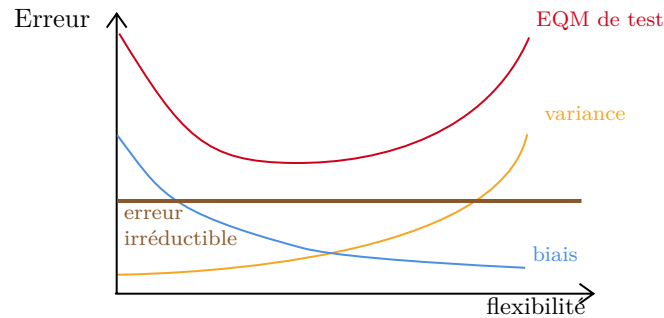
Donc, augmenter la flexibilité diminue le biais et augmente la variance, puis vice-versa. Il y a donc un **compromis** à faire entre les deux.

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

$$MSE = \text{Var}(\hat{f}) + [B[\hat{f}]]^2 + \text{Var}(\varepsilon).$$

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

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



Résumés numériques des modèles

Contexte

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

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

Nous avons des statistiques

1 Univariées

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

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

2 Bivariées

Pour deux variables x et y , on utilise typiquement la covariance échantillonnale $\text{cov}_{X,Y}$ et la corrélation échantillonnale $r_{X,Y}$.

Résumés graphiques des modèles

Contexte

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

Nuage de points (« Scatterplots »)

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

Diagramme en boîte

Contexte

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

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

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

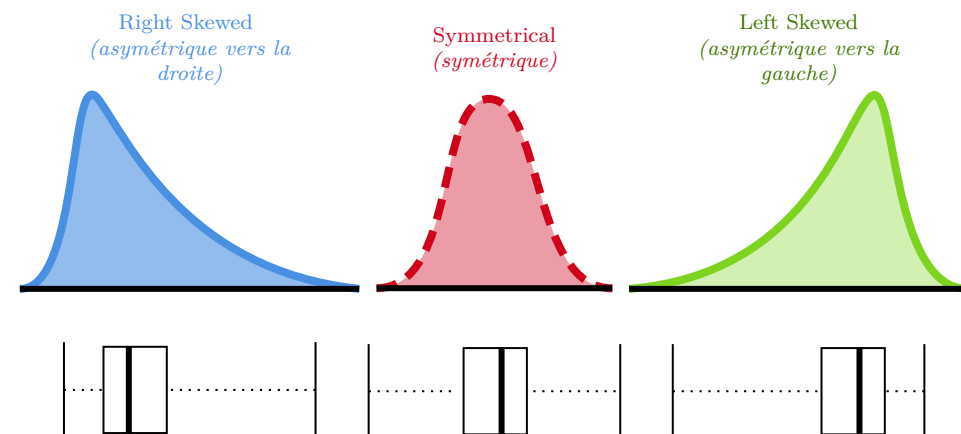
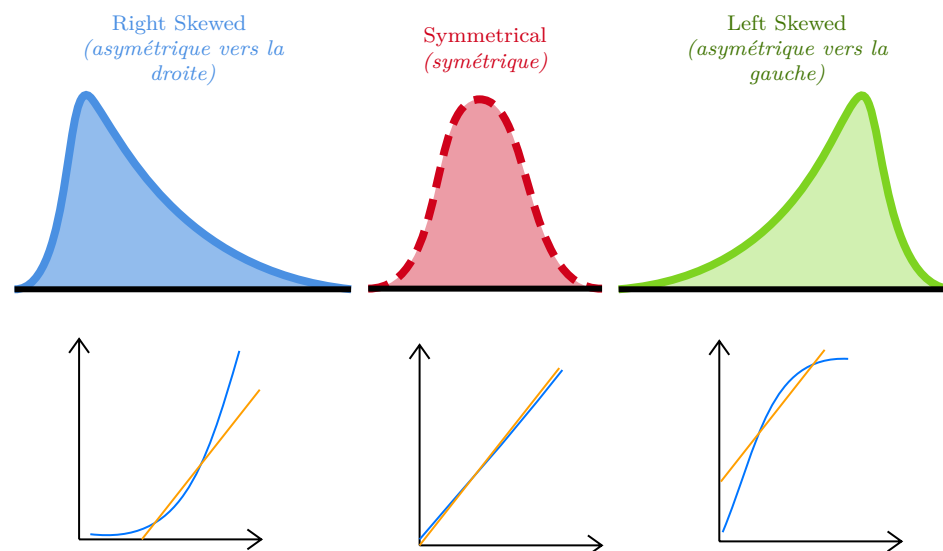


Diagramme quantile-quantile

Contexte

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

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



Régression linéaire simple

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

Contexte

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

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

Contexte

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

Définition du modèle

Notation

β_0 Paramètre d'intercepte.

β_1 Paramètre de pente.

Modèle de régression linéaire simple

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

Postulats de la régression linéaire simple

On suppose que :

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

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

H₁ Linéarité : $E[\varepsilon_i] = 0$.

H₂ Homoscédasticité : $\text{Var}(\varepsilon_i) = \sigma^2$.

H₃ Indépendance : $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$.

H₄ Normalité : $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Des postulats, on déduit que valeurs observées de la variable réponse sont indépendantes et normalement distribuées avec $E[Y_i] = \beta_0 + \beta_1 x_i$ et

$\text{Var}(Y_i) = \sigma^2$ telles que $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$.

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

Limitations

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

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

Estimation du modèle

Estimation des paramètres libres

Notation

b_0 Estimation du paramètre d'intercepte β_0 .

b_1 Estimation du paramètre de pente β_1 .

\hat{y} Préviation de la variable réponse y .

Estimation du modèle de régression linéaire simple

Les prévisions \hat{y} sont obtenues en fonction des estimations des paramètres :

$$\hat{y} = b_0 + b_1 x.$$

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

d'obtenir les estimations $b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$ et $b_0 = \bar{y} - b_1 \bar{x}$.

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

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

Estimation de la variance

Notation

MSE Erreur quadratique moyenne.

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

La variance de la valeur réponse, qui correspond à la variance de l'erreur de prévision, est estimée par l'erreur quadratique moyenne de prévision :

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}.$$

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

Représentation matricielle du modèle de régression linéaire simple

Contexte

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

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

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

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

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

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

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

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

Matrice de projection \mathbf{H}

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

Avec la matrice de projection, on obtient que $\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{H}\mathbf{y}$.

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

Somme des carrés

Résidus e_i

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

Si le résidu est

positif la vraie valeur est **supérieure** à la valeur prédite.

négatif la vraie valeur est **inférieure** à la valeur prédite.

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

1 Total Sum of Squares (SST)

Contexte

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

La variance $\text{Var}(\text{moyenne}) = \frac{(\text{données} - \text{moyenne})^2}{n-1} = s^2$ représente donc la variabilité des données autour de la moyenne.

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

2 Error Sum of Squares (SSE)

Contexte

La SSE correspond à la somme des carrés avec les valeurs prédites du modèle : $SS(\text{prévisions}) = (\text{données} - \text{prévisions})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$.

La variance $\text{Var}(\text{prévisions}) = \frac{(\text{données} - \text{prévisions})^2}{n-2} = \text{MSE}$ représente donc la variabilité des données autour de la droite de régression.

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

3 Regression Sum of Squares (SSR)

Contexte

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

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

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

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

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

📄 Coefficient de corrélation R^2

Contexte

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

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

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

La statistique $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$.

Estimateurs des paramètres

Estimateurs

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

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

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

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

Bootstrapping

Contexte

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

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

Matrice de variance-covariance

La matrice de variance-covariance de $\mathbf{b} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$ est

$\text{Var}(\mathbf{b}) = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ où

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

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

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

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

Tests d'hypothèse

Test t bilatéral

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

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

On fait un test sur la moyenne de l'estimateur b_1 avec $t = \frac{b_1 - h}{se(b_1)}$ où

$T \sim t_{(n-2)}$. La zone critique est $|t| \geq t_{\alpha, n-2}$.

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

Contexte

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

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

Intervalle de confiance et de prévision

Contexte

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

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

Contexte

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

Intervalle de prévision

Contexte

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

avec $\hat{Y}_{n+1} = b_0 + b_1 x_{n+1}$.

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

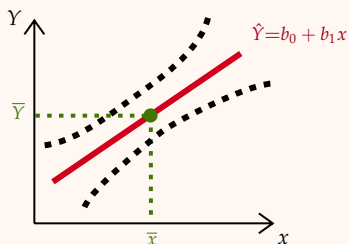
L'erreur type de l'intervalle de prévision $\sqrt{MSE \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$ est presque identique à l'erreur type de l'intervalle de confiance $\sqrt{MSE \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]}$.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

fait, la distinction revient à distinguer 2 composantes :

1 « *Parameter risk* »

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

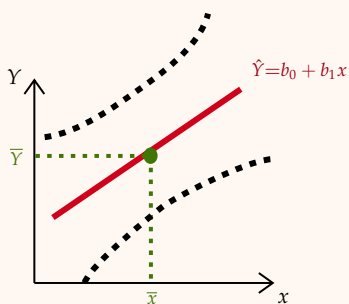
Visuellement, c'est l'intervalle des formes possibles pour la droite de régression estimée :



2 « *Process risk* »

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

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



Régression linéaire multiple

Régression linéaire généralisée

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 : $f(y; \theta) = e^{a(y)b(\theta) + c(\theta) + d(y)}$.

- › 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 $E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$ et $\text{Var}(a(Y)) = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{(b'(\theta))^3}$.

Classification

On fait de la **classification** lorsque nous voulons prédire une variable *catégorielle*. Il y a 3 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

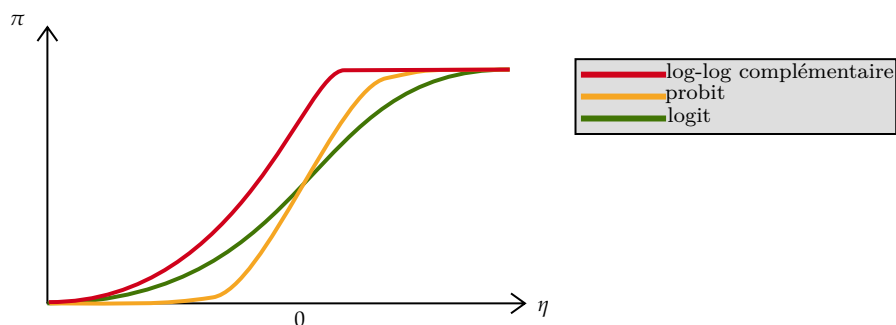
Soit $\eta = \beta_0 + \sum_{j=1}^p \beta_j x_j$. Soit la probabilité π que $Y = 1$.

- › 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	$\frac{e^\eta}{1+e^\eta}$	$\ln\left(\frac{\pi}{1-\pi}\right)$
Probit	$\Phi(\eta)$	$\Phi^{-1}(\mu)$
Log-log complémentaire	$1 - e^{-e^\eta}$	$\ln(-\ln(1 - \pi))$

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 :

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

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

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

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

Ordinal

Modèle logit cumulatif

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

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

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

Modèle de cotes proportionnelles

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

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

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

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

Analyse en composantes principales

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 : $\lambda = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}$.
 - › Ceci puisque le taux λ a une forme exponentielle.
2. Modéliser le log du taux comme une fonction linéaire des x : $\ln(\lambda) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.
 - › 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* » : $\ln(E[Y]) = \ln(E[\lambda t]) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \ln(t)$.
 - › 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 récurrence augmente la probabilité d'une deuxième récurrence, 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 récurrence 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.

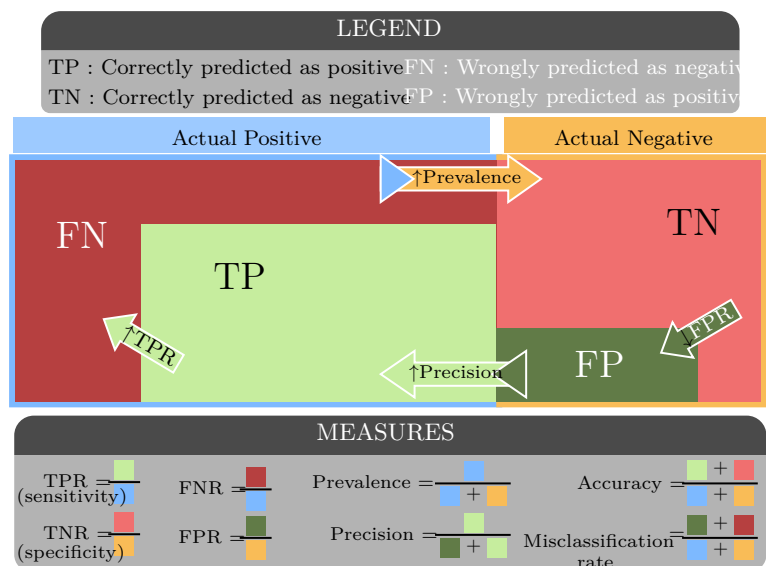
On peut calculer la **dispersion** et on désire qu'elle soit environ de 1.

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.

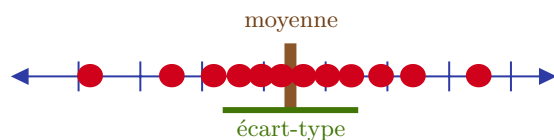
Deuxièmement, **Données gonflées à zéro**.

- › 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 et la fréquence avec un modèle de Poisson.

Matrice de confusion :**Erreur**

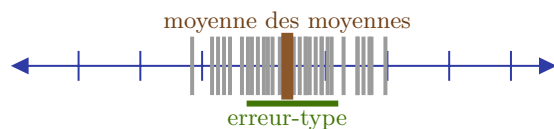
Écart-type Mesure la variation entre les observations d'un ensemble de données.

> « *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 $S(x) = e^{-H(x)}$.

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 **$g(0) = 0$** ,

$$E[g(X)] = \int_0^\infty g'(x)S(x)dx.$$

Fonction génératrice des moments (MGF)

La fonction génératrice des moments (MGF) d'une v.a. X est dénoté comme

$$M_X(t) = E[e^{tX}].$$

Entre autres, la MGF sert à générer les moments d'une distribution avec

$$E[X^n] = \frac{\partial^n M_X(t)}{\partial t^n} \Big|_{t=0}.$$

Fonction génératrice des probabilités (PGF)

La fonction génératrice des moments (PGF) d'une v.a. X est dénoté comme

$$P_X(t) = E[t^X].$$

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 $\frac{\partial^n P_X(t)}{\partial t^n} \Big|_{t=1} = E[X(X-1)\dots(X-(n-1))].$

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 $100q^e$ **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\$.

$$\begin{aligned} CTE_q(X) &= E[X|X > \pi_q] \\ &= \pi_q + E[X - \pi_q | X > \pi_q] \\ &= \pi_q + \frac{E[X] - E[X \wedge \pi_q]}{1 - q} \end{aligned}$$

- › 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.

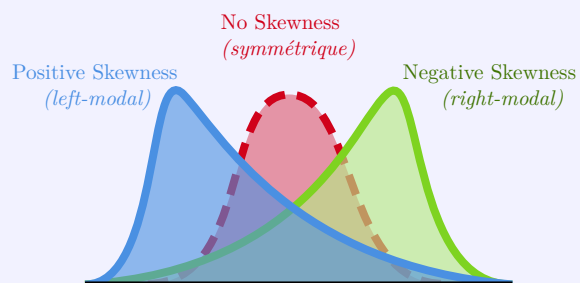
> est unimodal, c'est-à-dire qu'elle a une « bosse », alors
mode = x tel que $f'(x) = 0$.

> 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

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



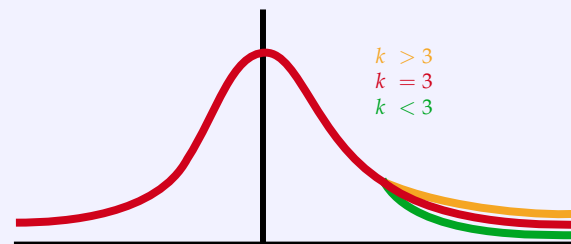
Kurtosis

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

Le kurtosis mesure l'aplatissement d'une distribution et peut aider à juger la vraisemblance qu'une distribution produise des valeurs extrêmes (ou « out-

liers »).

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
$X \sim \text{Pareto}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$f(x)$	$= \frac{\alpha \theta^\alpha}{(x + \theta)^{\alpha+1}}$
$F(x)$	$= 1 - \left(\frac{\theta}{x + \theta} \right)^\alpha$

> 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 $\theta \geq 0$	$x \in [0, \theta]$

$$f(x) = \frac{\theta}{B(a, b)} \left(\frac{x}{\theta} \right)^{a-1} \left(1 - \frac{x}{\theta} \right)^{b-1}$$

> $X \sim \text{Beta}(a = 1, b = 1, \theta) \sim \text{Unif}(0, \theta)$.

> 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
$X \sim \text{Gamma}(\alpha, \theta)$	$\alpha, \theta > 0$	$x \geq 0$

$$f(x) = \frac{x^{\alpha-1} e^{-x/\theta}}{\Gamma(\alpha) \theta^\alpha}$$

> 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 $Y = \min(X_1, \dots, X_n) \sim \text{Exp}(\frac{1}{\sum_{i=1}^n \lambda_i})$.

> Si $X \sim \text{Exp}(\theta)$ alors $(X - d | X > d) \sim \text{Exp}(\theta)$.

Loi de Weibull

Notation	Paramètres	Domaine
$X \sim \text{Weibull}(\tau, \beta)$	$\tau, \beta > 0$	$x \geq 0$

$$f(x) = \frac{\tau (x/\theta)^\tau e^{-(x/\theta)^\tau}}{x}$$

> 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 $n \in \mathbb{N}^+$	$x \geq 0$

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

Loi de Poisson

Notation	Paramètres	Domaine
$X \sim \text{Poisson}(\lambda)$	$\lambda > 0$	$x = 0, 1, 2, \dots$

$\Pr(X = x)$	$= \frac{e^{-\lambda} \lambda^x}{x!}$
--------------	---------------------------------------

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, \dots, X_n que l'on veut transformer en n autres variables aléatoires $W_1 = g_1(X_1, \dots, X_n), \dots, W_n = g_n(X_1, \dots, X_n)$.

- 1 Trouver les inverses des équations de la transformation :

$$x_1 = g_1^{-1}(w_1, \dots, w_n)$$

$$\vdots$$

$$x_n = g_n^{-1}(w_1, \dots, w_n)$$

- 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

$$f_{W_1, \dots, W_n}(w_1, \dots, w_n) = f_{X_1, \dots, X_n}(g_1^{-1}(w_1, \dots, w_n), \dots, g_n^{-1}(w_1, \dots, w_n)) |J|.$$

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

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 **lente-ment**.

- › Si $\lim_{x \rightarrow \infty} \frac{S_1(x)}{S_2(x)} = 0$ 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.

$\{X_1, \dots, X_n\}$ Échantillon aléatoire de n observations de X ;

$\hat{\theta}$ Estimateur de θ établi avec l'échantillon aléatoire $\{X_1, \dots, X_n\}$;

$F(x; \hat{\theta})$ 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);

$\{X_1, \dots, X_n\}$ Valeurs de X pour n individus;

$\{x_1, \dots, x_n\}$ n valeurs observées de l'échantillon;

› Il peut y avoir des valeurs dupliquées dans les valeurs observées.

$0 < y_1 < \dots < y_m$ m valeurs distinctes où $m \leq n$;

w_j Nombre de fois que la valeur y_j apparaît dans l'échantillon pour $j = 1, \dots, m$;

› 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

$(c_0, c_1], (c_1, c_2], \dots, (c_{k-1}, c_k]$ k intervalles regroupant les observations ;

$0 \leq c_0 < c_1 < \dots < c_k$ 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

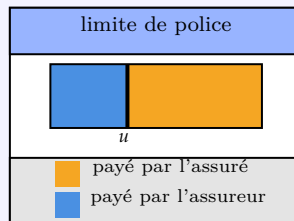
X Variable aléatoire du montant de perte.

Limite de police

Limite de police

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

Visuellement :



L'espérance limitée du montant de perte

L'espérance limitée du montant de perte $E[X \wedge u]$ correspond à l'espérance du paiement de l'assureur pour une police d'assurance ayant une limite de u :

$$E[X \wedge u] = \int_0^u x f(x) dx + u S(u)$$

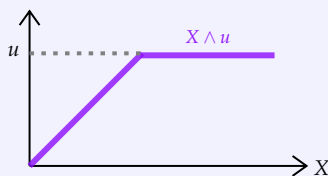
Montant de perte limité

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 :



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.

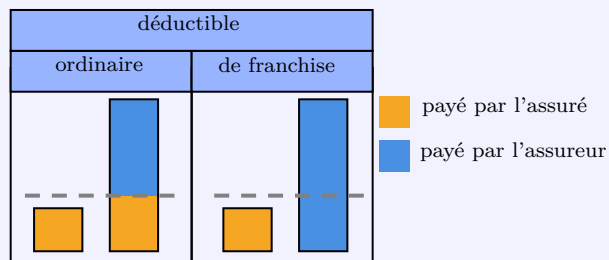
Il y a 2 types de déductibles :

déductible ordinaire Une fois que le montant de perte dépasse 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 dépasse 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

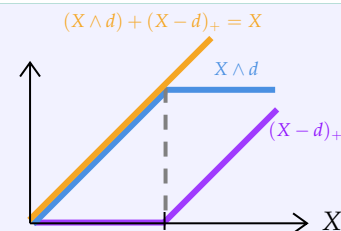
La variable aléatoire du montant de perte pour une police ayant un **déductible ordinaire** de d .

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

> 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 $E[(X-d)_+]$ correspond à :

$$E[(X-d)_+] = \int_d^\infty (x-d)f(x)dx$$

« Loss Elimination Ratio (LER) »

Le « Loss Elimination Ratio (LER) » évalue combien qu'épargne l'assureur en imposant un déductible ordinaire de d , $LER = \frac{E[X \wedge d]}{E[X]}$.

« payment per loss » et « payment per payment »

Notation

Y^L Montant de perte.

> « payment per loss »

Y^P Montant de paiement.

> « payment per payment »

$E[Y^L]$ Montant espéré de paiement **par perte subie**.

$E[Y^P]$ 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 $E[Y^L] \geq E[Y^P]$.

Pour un déductible ordinaire de d ,

$$E[Y^L] = E[(X - d)_+] \quad E[Y^P] = E[X - d | X > d]$$

› On trouve que $E[Y^P] = \frac{E[Y^L]}{S(d)}$.

› Également, le montant espéré de paiement par paiement effectué est la fonction d'excès moyen $E[Y^P] = e(d)$.

› Si la police d'assurance comporte uniquement une limite, $Y^P = Y^L$.

Relations pour quelques distributions :

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

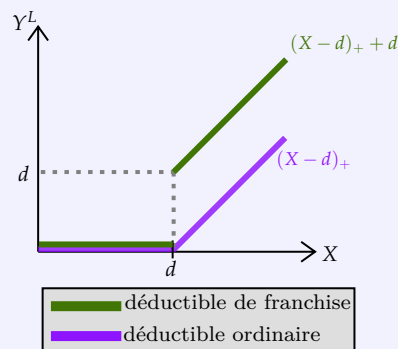
Déductible de franchise

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 .

$$(X | X > d) = \begin{cases} 0, & X \leq d \\ X, & X > d \end{cases}$$

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** $E[X | X > d]$ correspond à :

$$\begin{aligned} E[X | X > d] &= \int_d^\infty x f(x) dx = \int_d^\infty (x - d) f(x) dx + d \int_d^\infty f(x) dx \\ &= E[(X - d)_+] + d S(d) \end{aligned}$$

Impacts du déductible sur la fréquence

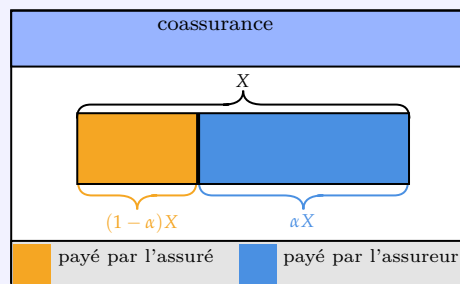
Pour la classe $(a, b, 0)$ de distributions, on trouve les relations suivantes :

Nombre de pertes (N)	Nombre de paiements (N')
$\text{Pois}(\lambda)$	$\text{Pois}(S(d)\lambda)$
$\text{Binom}(n, p)$	$\text{Binom}(n, S(d)p)$
$\text{BinNeg}(r, \beta)$	$\text{BinNeg}(r, S(d)\beta)$

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 $E[\alpha X] = \alpha E[X]$.

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 :

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

$$E[Y^L] = \alpha (E[X] - E[X \wedge d])$$

› 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 :

$$Y^L = \begin{cases} 0, & \alpha X \leq d \\ \alpha X - d, & \alpha X > d \end{cases} = \begin{cases} 0, & X \leq \frac{d}{\alpha} \\ \alpha \left(X - \frac{d}{\alpha} \right), & X > \frac{d}{\alpha} \end{cases}$$

$$E[Y^L] = \alpha \left(E[X] - E \left[X \wedge \frac{d}{\alpha} \right] \right)$$

Soit une police ayant :

1. une coassurance de α ,

2. une limite de police de u ,

3. un déductible **ordinaire** de d .

Alors, $E[Y^L] = \alpha \{ E[X \wedge m] - E[X \wedge d] \}$ et

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

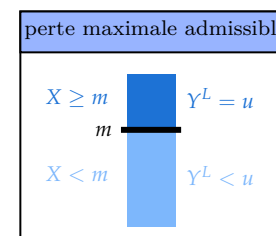
où m est la **perte maximale admissible**.

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 $E[(1+r)X] = (1+r)E[X]$.

Combiné avec les autres facteurs :

$$E[Y^L] = \alpha(1+r) \left(E\left[X \wedge \frac{m}{1+r}\right] - E\left[X \wedge \frac{d}{1+r}\right] \right)$$

$$E[Y^P] = \frac{E[Y^L]}{S_X\left(\frac{d}{1+r}\right)}$$

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, « *bandwidth* ».

Estimer une fonction de densité par une fonction noyau

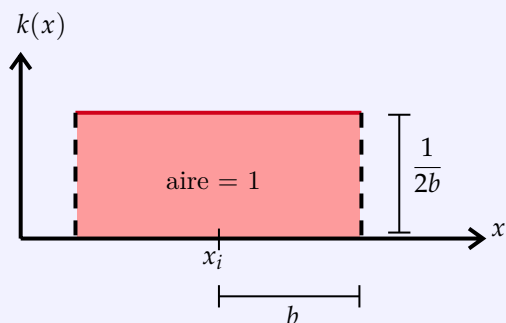
- 1 Choisir un type de fonction de densité pour $k()$.
- 2 Estimer la fonction de densité $f(x)$ comme la moyenne des fonctions noyau des n observations $k_1(x), \dots, k_n(x)$:

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

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

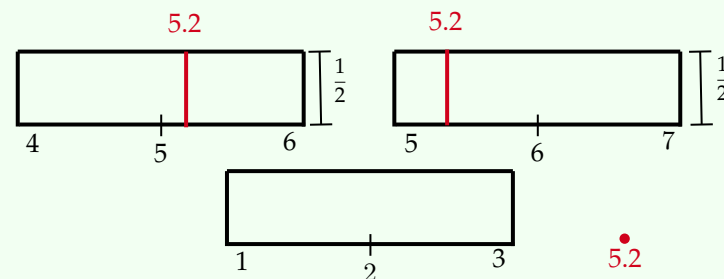
En termes mathématiques :

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

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 :

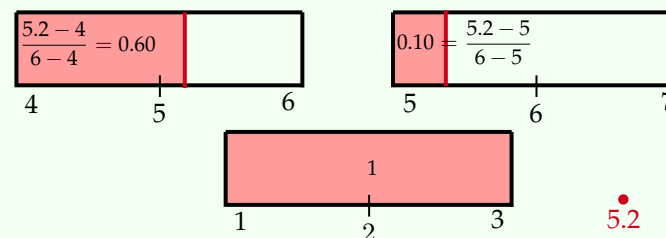


- 3 La fonction de densité estimée est donc :

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

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

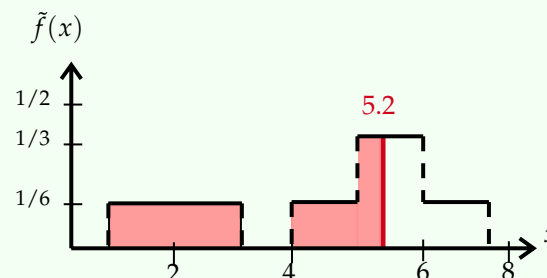
- 1 Visuellement, on voit comment l'équivalence géométrique du calcul des probabilités :



- 2 Donc :

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

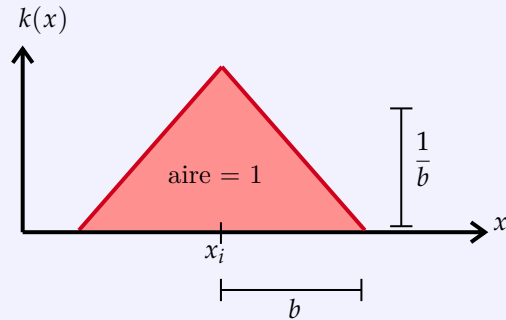
Visuellement, la densité par noyau est :



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 :

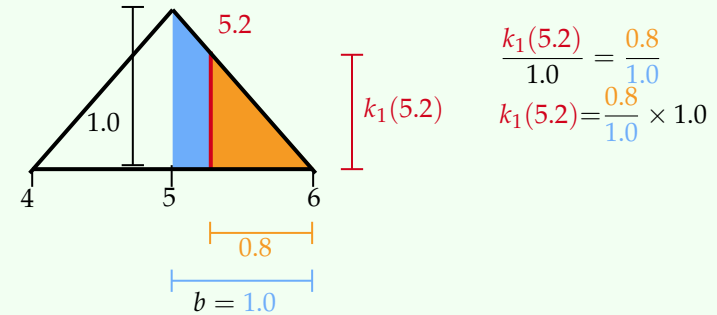
$$k_i(x) = \begin{cases} \frac{b - |x - x_i|}{b^2}, & x_i - b \leq x \leq x_i + b \\ 0, & \text{sinon} \end{cases}$$

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.

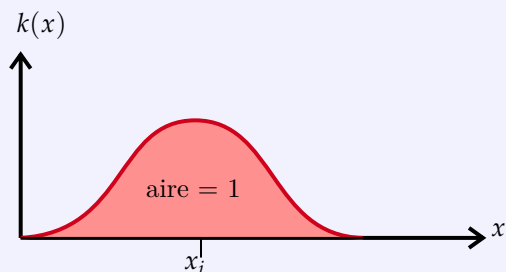
Pour le même exemple qu'avant, mais avec un noyau rectangulaire ce coup-ci, on trouve visuellement $k_1(5.2)$:



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 $x \in (-\infty, \infty)$,

$$k_i(x) = \frac{1}{b\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_i}{b}\right)^2}$$

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, \dots, y_m avec probabilités $\frac{w_1}{n}, \dots, \frac{w_m}{n}$;

> On peut également la définir comme la distribution discrète équiprobable des valeurs x_1, \dots, 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* »).

$$\hat{f}(y) = \begin{cases} \frac{w_j}{n}, & \text{si } y = y_j \forall j \\ 0, & \text{sinon} \end{cases}$$

$$\hat{F}(y) = \begin{cases} 0, & y < y_1, \\ \frac{1}{n} \sum_{h=1}^j w_h, & y_j \leq y < y_{j+1}, j = 1, \dots, m-1 \\ 1, & y_m \leq y \end{cases}$$

On peut estimer la valeur de $\hat{F}()$ pour une valeur de y pas dans l'ensemble y_1, \dots, y_m avec la fonction de répartition lissée $\tilde{F}()$. Pour $y_j \leq y < y_{j+1}$

et $j \in \{1, 2, \dots, m-1\}$, $\tilde{F}(y)$ est une interpolation linéaire de $\hat{F}(y_{j+1})$ et $\hat{F}(y_j)$:

$$\tilde{F}(y) = \frac{y - y_j}{y_{j+1} - y_j} \hat{F}(y_{j+1}) + \frac{y_{j+1} - y}{y_{j+1} - y_j} \hat{F}(y_j)$$

✓ 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 :

$$E[Y] = \frac{E[\hat{F}(y)]}{n} = F(y)$$

$$\text{Var}(Y) = \frac{\text{Var}(\hat{F}(y))}{n^2} = \frac{F(y)(1 - F(y))}{n}$$

Données incomplètes

Section à compléter avec mes notes d'IARD et 11.2 de Nonlife Actuariel Models (tse).

Estimateur de Kaplan-Meier Soit :

$$S(y_j) = \Pr(X > y_1) \Pr(X > y_2 | X > y_1) \dots \Pr(X > y_j | X > y_{j-1}) = \Pr(X > y_1) \prod_{h=2}^j \Pr(X > y_h | X > y_{h-1})$$

Où on peut estimer $\hat{\Pr}(X > y_1) = 1 - \frac{w_1}{r_1}$ et $\hat{\Pr}(X > y_h | X > y_{h-1}) = 1 - \frac{w_h}{r_h}$ 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 : $\text{Var}(\hat{S}_K(y_j) | \mathcal{C}) \approx (S(y_j))^2 \left(\sum_{h=1}^j \frac{1 - S_h}{S_h r_h} \right)$

Approximation de Greenwood de la variance de l'estimateur Kaplan-Meier :

$$\widehat{\text{Var}}(\hat{S}_K(y_j) | \mathcal{C}) \approx (\hat{S}_K(y_j))^2 \left(\sum_{h=1}^j \frac{w_h}{r_h(r_h - w_h)} \right)$$

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 $S(y) = e^{-H(y)}$ et $H(y) = -\ln(S(y))$.

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 est une continuation de Méthode du maximum de vraisemblance 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

$$\mathcal{L}(\theta; \mathbf{x}) = \prod_{j=1}^k \underbrace{f(x_j; \theta)}_{\text{probabilité que chaque observation soit égale à la valeur observée}}$$

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 $\mathbf{n} = (n_1, \dots, n_k)$, est :

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

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 :

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

tronquées vers la gauche

On pose un déductible de d :

$$\mathcal{L}(\theta; \mathbf{x}) = \frac{1}{[1 - F(d; \theta)]^n} \prod_{i=1}^n f(x_i; \theta)$$

pondère la vraisemblance par la probabilité d'être supérieur au déductible

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 signifiante dont l'objectif est d'évaluer les hypothèses de distribution d'un modèle.

Notation

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

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

Les tests de Kolmogorov-Smirnov (K.-S.) et de Anderson-Darling sont idéaux lorsque l'on désire comparer les fonctions de répartition.

Voir la section *Tests d'adéquation* du chapitre de *Analyse statistique des risques actuariels* pour la présentation du test de Kolmogorov-Smirnov. Nous présentons ci-dessous le test pour des données incomplètes.

Test de Kolmogorov-Smirnov pour des données incomplètes

Pour des données tronquées vers la gauche à d , il suffit d'ajuster la fonction de répartition théorique et de poser

$$F^*(x) = \frac{F(x) - F(d)}{1 - F(d)}$$

où $F(\cdot)$ est la fonction de répartition de la distribution théorique de X . Ceci nous donne donc que $F^*(\cdot)$ est la fonction de répartition de $(X|X > d)$.

Pour des données censurées vers la droite à u , la distribution théorique n'est pas affectée. Cependant, la valeur de $\hat{F}(m)$ n'est pas définie plutôt qu'être 1. De plus, n inclut les valeurs censurées pour compter le nombre total d'observations.

Lorsque les paramètres sont connus, le test de K.-S. n'est pas spécifique à aucune distribution avec des valeurs critiques générales. Le test de Anderson-Darling (A.-D.) considère toutes les différences ($\hat{F}(x) - F^*(x)$) et non seulement la différence maximale. Également, elle attribue plus de poids aux queues de la distribution en pondérant par la fonction de répartition et de survie :

$$A^2 = n \int \frac{(\hat{F}(x) - F^*(x))^2}{F^*(x)S^*(x)} f^*(x) dx$$

Donc, lorsque $F^*(x)$ ou $S^*(x)$ est petit, la différence est attribuée plus de poids.

Il s'ensuit que le test de A.-D. est « spécifique par distribution » dans le sens que les valeurs critiques sont différentes selon la distribution sous-jacente—il y a une table de valeurs critiques pour une distribution normale, Weibull, exponentielle, etc.

Test de Anderson-Darling

L'intégrale ci-dessus se simplifie à :

$$A^2 = -n - \frac{1}{n} \left[\sum_{j=1}^n (2j-1) \log \left(F^*(x_{(j)}) [1 - F^*(x_{(n+1-j)})] \right) \right]$$

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

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.

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 :

$$AIC = \log \mathcal{L}(\hat{\theta}_n^{\text{EMV}}; \mathbf{x}) - p.$$

- > 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 :

$$BIC = \log \mathcal{L}(\hat{\theta}_n^{\text{EMV}}; \mathbf{x}) - \frac{p}{2} \log(n).$$

- > En anglais, « *Bayesian Information Criterion (BIC)* »

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 rapides** 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 .

$\phi(x)$ « **Structure function** » d'un système représentant son **état**.

L'état d'une composante

Chacune des composantes du système 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

$$\mathbf{x} = (x_1, x_2, \dots, x_n).$$

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

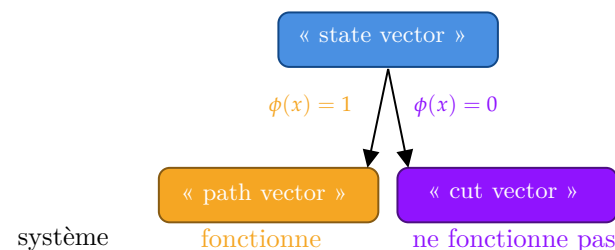
possibles. Ceci résulte en $2 \times 2 \times \dots \times 2 = 2^n$ différentes combinaisons possibles.

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 $\phi(x)$ binaire :

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

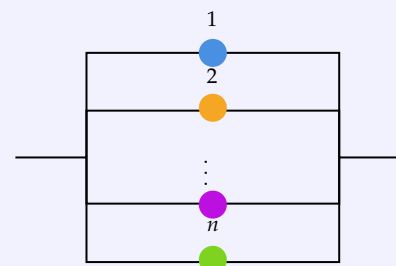
Nous verrons le type de vecteur d'état selon l'état de fonctionnement du système :



Systèmes communs

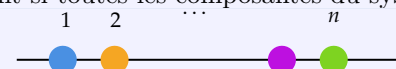
Système parallèle

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



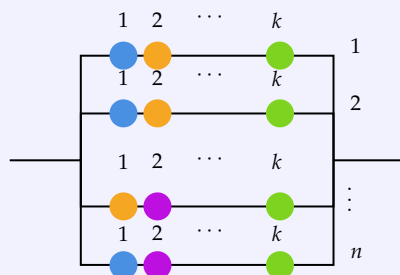
Système de série

Fonctionne seulement si toutes les composantes du système fonctionnent.



Système de k parmi n

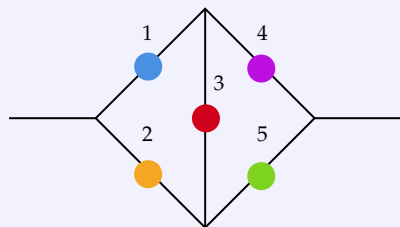
Fonctionne si au moins k des n composantes du système fonctionnent.



- > 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 ($\phi(\mathbf{x}) = 1$).

« 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.

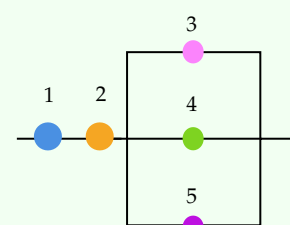
En termes mathématiques, \mathbf{x} est un « minimal path vector » si $\phi(\mathbf{y}) = 0 \forall \mathbf{y} < \mathbf{x}$.

- > $\mathbf{y} < \mathbf{x}$ implique que tous les éléments y_i du vecteur \mathbf{y} sont inférieurs ou égaux aux éléments x_i du vecteur \mathbf{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 » fonctionnent.

Exemple de système



Minimal path sets

1	{1, 2, 3}
2	{1, 2, 4}
3	{1, 2, 5}

Minimal path vectors

1	$\mathbf{x} = (1, 1, 1, 0, 0)$
2	$\mathbf{x} = (1, 1, 0, 1, 0)$
3	$\mathbf{x} = (1, 1, 0, 0, 1)$

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

Minimal path sets

$$1 \quad \mathbf{x} = (1, 1, 1, 0, 0)$$

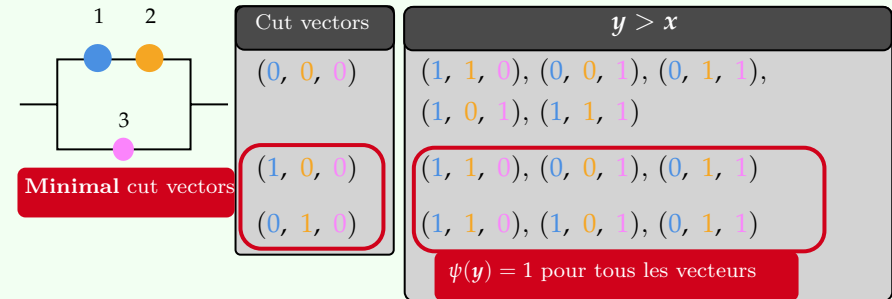
 $\mathbf{y} < \mathbf{x}$

$$1 \quad (0, 0, 0, 0, 0), (0, 0, 1, 0, 0), (0, 1, 0, 0, 0), \\ (1, 0, 0, 0, 0), (0, 1, 1, 0, 0), (1, 1, 0, 0, 0), \\ (1, 0, 1, 0, 0)$$

On note que $\phi(\mathbf{y}) = 0$ pour tous les vecteurs ce qui fait de \mathbf{x} un « *minimal path vector* ».

Exemple « *minimal cut sets* »

On peut visualiser ci-dessous que les « *minimal cut vectors* » sont les « *cut vectors* » pour lesquels tous les vecteurs \mathbf{y} fonctionnent ($\psi(\mathbf{y}) = 1$).

« *Cut vector* »

Vecteur d'états pour lequel le système ne fonctionne **pas** ($\phi(\mathbf{x}) = 0$).

> 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, \mathbf{x} est un « *minimal cut vector* » si $\phi(\mathbf{y}) = 1 \forall \mathbf{y} > \mathbf{x}$.

> $\mathbf{y} > \mathbf{x}$ implique que tous les éléments y_i du vecteur \mathbf{y} sont supérieurs ou égaux aux éléments x_i du vecteur \mathbf{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* » \mathcal{C} étant donné un « *minimal cut vector* » \mathbf{x} est $\{i : x_i = 0\}$.

Système	Nombre de	
	« <i>miminal path sets</i> »	« <i>miminal cut sets</i> »
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
« <i>minimal path sets</i> »	parallèle	somme
	série	produit
« <i>minimal cut sets</i> »	parallèle	produit
	série	somme

Structure Functions

Notation

A_1, \dots, A_s « Minimal path sets ».

C_1, \dots, 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 $\phi(\mathbf{x}) = 1$.

$$\begin{aligned}\phi(\mathbf{x}) &= \max\{x_1, \dots, x_n\} \\ &= 1 - \prod_{i=1}^n (1 - x_i)\end{aligned}$$

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

② Système en série

Un système en série 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 $\phi(\mathbf{x}) = 0$.

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

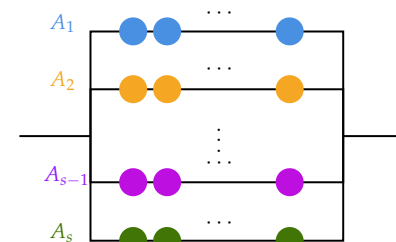
› La deuxième formulation découle du fait que les états sont des variables binaires.

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 :

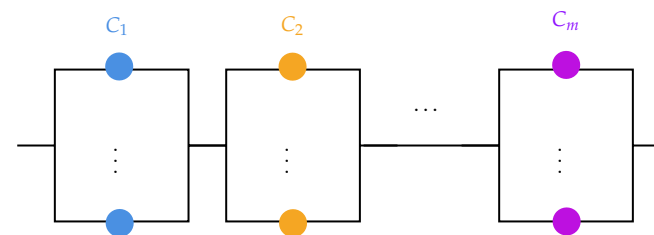
$$\phi(\mathbf{x}) = \max \left\{ \min_{i \in A_1} x_i, \min_{i \in A_2} x_i, \dots, \min_{i \in A_s} x_i \right\} = \min_j \prod_{i \in A_j} x_i$$

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 :

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

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 :

$$X_i \sim \text{Bernoulli}(p_i).$$

$$X_i = \begin{cases} 1, & p_i \\ 0, & 1 - p_i \end{cases}$$

> $\mathbf{X} = (X_1, X_2, \dots, X_n)$ vecteur des v.a. Bernoulli.

p_i Fiabilité de la composante i .

> $\mathbf{p} = (p_1, p_2, \dots, p_n)$ vecteur des fiabilités.

$r(\mathbf{p})$ 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 $r(\mathbf{p})$ qui quantifie la probabilité que le système fonctionne.

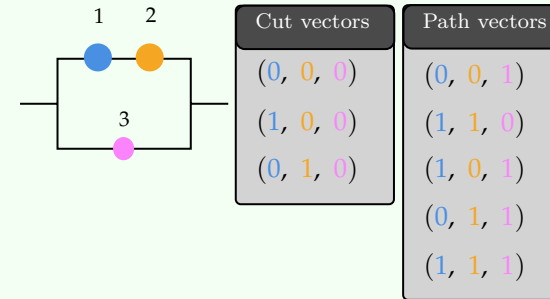
$$r(\mathbf{p}) = \underbrace{\Pr(\phi(\mathbf{X}) = 1)}_{\text{somme des probabilités des « path vectors »}} = 1 - \underbrace{\Pr(\phi(\mathbf{X}) = 0)}_{\text{somme des probabilités des « cut vectors »}}$$

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

> Puisque ϕ est fonction du vecteur de v.a. Bernoulli \mathbf{X} , ϕ est également une v.a. Bernoulli.

> Il s'ensuit que $r(\mathbf{p}) = 0 \times \Pr(\phi(\mathbf{X}) = 0) + 1 \times \Pr(\phi(\mathbf{X}) = 1) = E[\phi(\mathbf{X})]$.

Exemple de calcul de la fonction de fiabilité



On pose que les composantes sont indépendantes, puis :

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

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

$$\begin{aligned}\Pr(E_1 \cup E_2) &= \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2) \\ \Pr\left(\bigcup_{j=1}^n E_j\right) &= \sum_{j=1}^n \Pr(E_j) - \sum_{j=1}^n \sum_{k>j} \Pr(E_j \cap E_k) + \sum_{j=1}^n \sum_{k>j} \sum_{l>k} \Pr(E_j \cap E_k \cap E_l) - \\ &\quad \dots + (-1)^{n+1} \Pr(E_1 \cap E_2 \cap \dots \cap E_n)\end{aligned}$$

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 ($r(\mathbf{p})$) ou pour la probabilité que le système ne fonctionne pas ($1 - r(\mathbf{p})$).

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, \dots, A_s , on établit :

$$\begin{aligned}r(\mathbf{p}) &\leq \sum_{j=1}^s \left(\prod_{i \in A_j} p_i \right) \\ r(\mathbf{p}) &\geq \sum_{j=1}^s \left(\prod_{i \in A_j} p_i \right) - \sum_{j=1}^s \sum_{k>j} \left(\prod_{i \in A_j \cup A_k} p_i \right) \\ &\vdots\end{aligned}$$

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 $A_2 = (1, 1, 0)$:

$$\begin{aligned}\sum_{j=1}^s \left(\prod_{i \in A_j} p_i \right) &= \prod_{i \in A_1} p_i + \prod_{i \in A_2} p_i = p + p^2 \\ \sum_{j=1}^s \sum_{k>j} \left(\prod_{i \in A_j \cup A_k} p_i \right) &= \prod_{i \in A_1 \cup A_2} p_i = p^3\end{aligned}$$

$$\text{Donc } p + p^2 - p^3 \leq r(\mathbf{p}) \leq p + p^2.$$

Si $p = 0.2$, $r(\mathbf{p}) \in [0.232, 0.24]$ mais si $p = 0.6$ alors $r(\mathbf{p}) \in [0.744, 0.96]$. On voit donc que plus p est petit, mieux l'intervalle approxime la fiabilité.

Minimal cut sets Pour les « *minimal cut sets* » C_1, \dots, C_m , on établit :

$$\begin{aligned}1 - r(\mathbf{p}) &\leq \sum_{j=1}^m \left(\prod_{i \in C_j} (1 - p_i) \right) \\ 1 - r(\mathbf{p}) &\geq \sum_{j=1}^m \left(\prod_{i \in C_j} (1 - p_i) \right) - \sum_{j=1}^m \sum_{k>j} \left(\prod_{i \in C_j \cup C_k} (1 - p_i) \right) \\ &\vdots\end{aligned}$$

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 $C_2 = (0, 1, 0)$:

$$\sum_{j=1}^m \left(\prod_{i \in C_j} (1 - p_i) \right) = \prod_{i \in C_1} (1 - p_i) + \prod_{i \in C_2} (1 - p_i) = (1 - p)^2 + (1 - p)^2 = 2(1 - p)^2$$

$$\sum_{j=1}^m \sum_{k>j} \left(\prod_{i \in C_j \cup C_k} (1 - p_i) \right) = \prod_{i \in C_1 \cup C_2} (1 - p_i) = (1 - p)^3$$

$$\text{Donc } 2(1 - p)^2 - (1 - p)^3 \leq r(\mathbf{p}) \leq 2(1 - p)^2.$$

Si $p = 0.2$, $1 - r(\mathbf{p}) \in [0.768, 1.28]$ mais si $p = 0.6$ alors $1 - r(\mathbf{p}) \in [0.256, 0.32]$. 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* ».

Ici, on pose que toutes les composantes ont une fiabilité de p , puis avec $C_1 = (1, 0, 0)$ et $C_2 = (0, 1, 0)$:

$$\prod_{j=1}^m \left[1 - \prod_{i \in C_j} (1 - p_i) \right] = (1 - (1 - p)^2) (1 - (1 - p)^2) = (1 - (1 - p)^2)^2$$

Avec $A_1 = (0, 0, 1)$ et $A_2 = (1, 1, 0)$,

$$1 - \prod_{j=1}^s \left[1 - \prod_{i \in A_j} p_i \right] = 1 - (1 - p) (1 - p^2)$$

$$\text{Donc } (1 - (1 - p)^2)^2 \leq r(\mathbf{p}) \leq 1 - (1 - p) (1 - p^2).$$

Si $p = 0.2$, $r(\mathbf{p}) \in [0.1296, 0.232]$ et si $p = 0.6$ alors $1 - r(\mathbf{p}) \in [0.7056, 0.744]$. On voit donc que peu importe la valeur de p , l'intervalle approxime bien la fiabilité.

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,

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

Exemple bornes avec la méthode d'intersection

On reprend l'exemple de la sous-section sur les fonctions de fiabilité avec le système en parallèle ayant 3 composantes.

En bref :

Approche	avec un petit p	avec un gros p
	Intervalle	
« <i>minimal path sets</i> »	large	étroit
« <i>minimal cut sets</i> »	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.

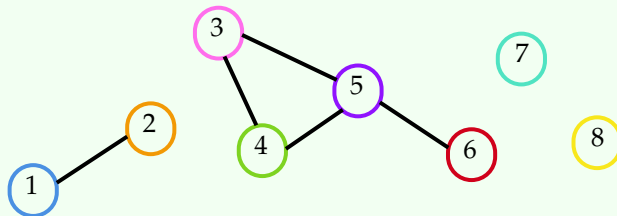
- > Le nombre d'arcs est au plus $\binom{n}{2}$.
- > C'est-à-dire, le nombre possibles de groupes de deux nœuds.

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 :



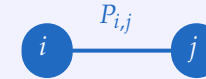
On trouve que :

- > 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 aléatoires

Contexte

La connectivité des graphiques aléatoires est semblable à la fiabilité des systèmes.

Pour un système, il n'est pas nécessaire que toutes les composantes fonctionnent pour que le système fonctionne. De façon semblable, il n'est pas nécessaire que tous les nœuds d'un graphique aléatoire soient reliés pour qu'il soit connecté.

Alors, on peut appliquer les mêmes concepts de « *minimal path sets* » et de « *minimal cut sets* » des systèmes aux graphiques aléatoires.

Un graphique *aléatoire* est connecté tant que tous les arcs d'au moins un « *minimal path sets* » existent.

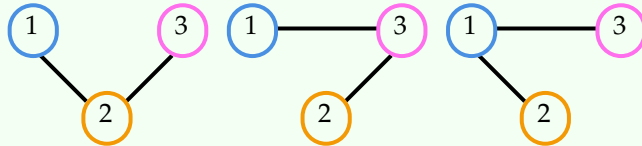
Un graphique aléatoire de n nœuds a :

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

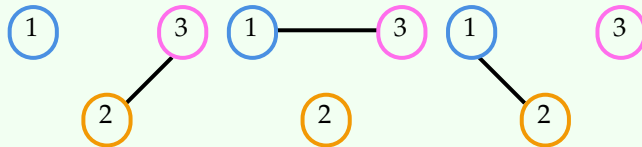
Exemple de connectivité

Soit un graphique aléatoire avec 3 nœuds.

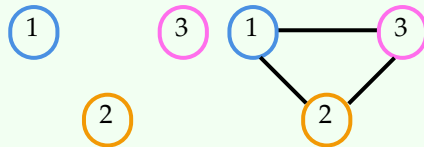
Les $3^{3-2} = 3$ « *minimal path sets* » sont les suivants :



Les $2^{3-1} - 1 = 3$ « *minimal cut sets* » sont les suivants :



Les deux autres graphiques possibles qui ne sont pas optimaux sont :



Finalement, on peut **approximer** la probabilité avec $P_n \approx 1 - n(1 - p)^{n-1}$.

✓ 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 réursive :

$$P_n = 1 - \sum_{k=1}^{n-1} \binom{n-1}{k-1} (1-p)^{k(n-k)} P_k, \quad n = 2, 3, \dots$$

où $P_1 = 1, P_2 = p$.

On peut également trouver les **bornes** pour la probabilité pour simplifier la tâche :

$$n(1-p)^{n-1} - \binom{n}{2} (1-p)^{2n-3} \leq 1 - P_n \leq (n+1)(1-p)^{n-1}$$

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 .

› $\mathbf{S}(t) = (S_1(t), \dots, S_n(t))$ 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 : $\Pr(T > t) = r[\mathbf{S}(t)]$.

Donc, on pose $p_i = S_i(t)$ pour $i = 1, 2, \dots, n$.

✓ Espérance de durée de vie

La durée de vie espérée équivaut à $E[T] = \int_0^\infty r[\mathbf{S}(t)] dt$.

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 $r(\mathbf{p}) = p_3 + p_1 p_2 - p_1 p_2 p_3$.

3 Remplacer \mathbf{p} par $\mathbf{S}(t)$:

$$\begin{aligned} r(\mathbf{p}) &= S_3(t) + S_1(t)S_2(t) - S_1(t)S_2(t)S_3(t) \\ &= \left(\frac{2-t}{2}\right) + \left(\frac{2-t}{2}\right)^2 - \left(\frac{2-t}{2}\right)^3 \\ &= \frac{t^3 - 4t^2 + 8}{8} \end{aligned}$$

4 Trouver $E[T]$:

$$\begin{aligned} E[T] &= \int_0^2 \frac{t^3 - 4t^2 + 8}{8} dt \\ &= 1.1667 \end{aligned}$$

□ Étapes du calcul de probabilités, ou de l'espérance, de la durée de vie

- 1 Déterminer la fonction de la structure du système $\phi(\mathbf{X})$.
› Soit avec les « *minimal path sets* » ou les « *minimal cut sets* ».
- 2 Dédire la fonction de fiabilité.
› Soit en trouvant $r(\mathbf{p}) = E[\phi(\mathbf{X})]$, ou avec $r(\mathbf{p}) = \Pr(\phi(\mathbf{X}) = 1)$.
- 3 Développer la fonction de survie $\Pr(T > t)$ de la fonction de fiabilité $r(\mathbf{S}(t))$.
- 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 , $E[T] = \mu \sum_{i=k}^n \frac{1}{i}$. 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).

${}_t d_x$ Nombre de décès entre les âges x et $x + t$.

$$> {}_t d_x = \ell_x - \ell_{x+t}.$$

Probabilité de survie

La probabilité qu'un assuré de x ans survie au moins t années est ${}_t p_x = \frac{\ell_{x+t}}{\ell_x}$.

Probabilité de décès

La probabilité qu'un assuré de x ans décède d'ici t années est ${}_t q_x = \frac{\ell_x - \ell_{x+t}}{\ell_x}$.

Variable aléatoire du nombre de décédés entre les âges x et $x + t$ ${}_t \mathcal{D}_x$

On a que ${}_t \mathcal{D}_x \sim \text{Bin}(\ell_x, {}_t q_x)$.

> Il s'ensuit que $E[{}_t \mathcal{D}_x] = {}_t d_x$.

> Également, ${}_t \mathcal{D}_x = \mathcal{L}_x - \mathcal{L}_{x+t}$.

Espérances de vie

Espérance de vie abrégée pour un individu d'âge x

$$e_x = \sum_{k=0}^{\omega-x-1} {}_k | q_x$$

puis, si $\lim_{k \rightarrow \infty} (k+1) {}_{k+1} p_x = 0$,

$$= \sum_{k=1}^{\omega-x} {}_k p_x$$

> En anglais, « *curtate life expectancy* ».

Espérance de vie complète pour un individu d'âge x

$$e_x = \int_0^{\omega-x} {}_t p_x \mu_{x+t} dt$$

$$= \int_0^{\omega-x} {}_t p_x dt \quad \text{si } \lim_{t \rightarrow \infty} {}_t p_x = 0$$

Sous l'hypothèse d'une distribution uniforme des décès (DUD),

$$e_x \stackrel{DUD}{=} e_x + \frac{1}{2}.$$

> 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.

$$A_x = \sum_{k=0}^{\omega-x-1} v^{k+1} {}_k p_x q_{x+k} \\ = v q_x + v^2 {}_1 p_x q_{x+1} + v^3 {}_2 p_x q_{x+2} + \dots$$

Capital différé de t années ${}_t E_x$

Si l'assuré **ne décède pas** dans les t années suivant l'émission du contrat, le capital différé ${}_t E_x$ paye une prestation de survie.

$${}_t E_x = v^t {}_t p_x$$

> Alias, le **facteur d'actualisation actuariel**.

> En anglais, « *mortality discount factor* ».

Assurance différée de m années ${}_m | Z_x$

Si l'assuré décède **après** les m années suivant l'émission du contrat, paye une prestation de décès.

$${}_m | A_x = \sum_{k=m}^{\omega-x-1} v^{k+1} {}_k p_x q_{x+k}$$

Assurance-vie temporaire $Z_{x:\overline{n}|}$

Si l'assuré décède dans les n années suivant l'émission du contrat, paye une prestation de décès.

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

Assurance mixte $Z_{x:\overline{n}|}$

Si l'assuré **décède** dans les n années suivant l'émission du contrat, paye une prestation de décès. S'il est toujours en vie, paye une prestation de survie.

$$A_{x:\overline{n}|} = \sum_{k=0}^{n-1} v^{k+1} {}_k p_x q_{x+k} + {}_n E_x$$

> 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 $A_x = v q_x + v p_x A_{x+1}$.

différée ${}_m | A_x = {}_m E_x A_{x+m}$.

temporaire $A_{x:\overline{n}|}^1 = A_x - {}_n | A_x$.

mixte $A_{x:\overline{n}|} = A_{x:\overline{n}|}^1 + {}_n E_x$.

Contrats de rentes

Rentes de base

☰ Rente viagère de début de période \ddot{Y}_x

Pour $K = 0, 1, 2, \dots$ on obtient que $\ddot{Y}_x = \ddot{a}_{\overline{K+1}|}$. Puis, $E[\ddot{Y}_x] = \ddot{a}_x$.

$$\begin{aligned}\ddot{a}_x &= \sum_{k=0}^{\omega-x-1} v^k {}_k p_x \\ &= 1 + v p_x + v^2 {}_2 p_x + \dots \\ &= \frac{1 - A_x}{d}\end{aligned}$$

☑ Relations

Rente

viagère $\ddot{a}_x = 1 + v p_x \ddot{a}_{x+1}$.

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 ignore 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 $x_{j+1} = (ax_j + c) \bmod m$,
 $j \geq 0$.
 > 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

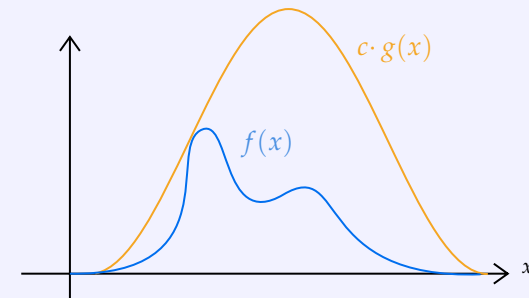
Lorsqu'il est difficile, ou impossible, de trouver la fonction quantile on peut utiliser la méthode d'acceptation de rejet.

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 :

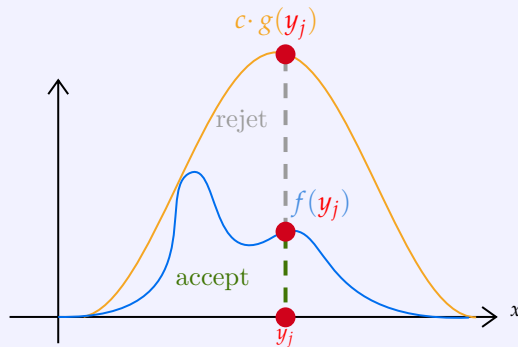
Simulation par la méthode d'acceptation-rejet

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 .

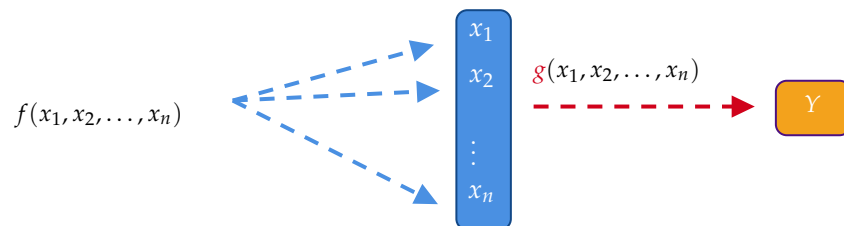
Simulation Monte-Carlo

Simulation Monte-Carlo

Pour une variable aléatoire X ,

- 1 Simuler un vecteur de réalisation (x_1, x_2, \dots, x_n) d'une distribution dont la fonction de densité est $f(x_1, x_2, \dots, x_n)$.
- 2 Appliquer une fonction g au vecteur des réalisations pour trouver $y_j = g(x_1, x_2, \dots, x_n)$.
- 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, \dots, 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 $\{X_n, n = 0, 1, 2, \dots\}$.

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.

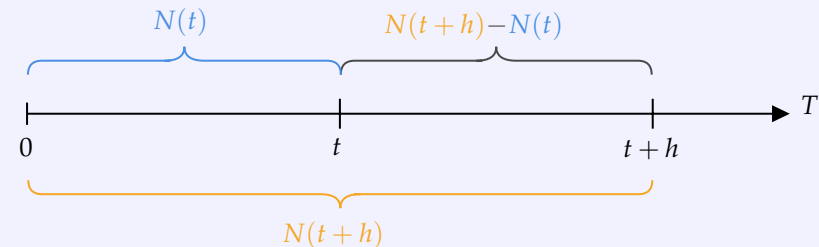
Processus de comptage

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. $N(t+h) \geq N(t)$ pour $h > 0$ (*valeurs non décroissantes*).

Visuellement, on voit que l'**accroissement** $N(t+h) - N(t)$ représente le nombre d'événements produits sur l'intervalle $(t, t+h]$:



> 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.

> $T_k = V_1 + V_2 + \dots + V_k$.

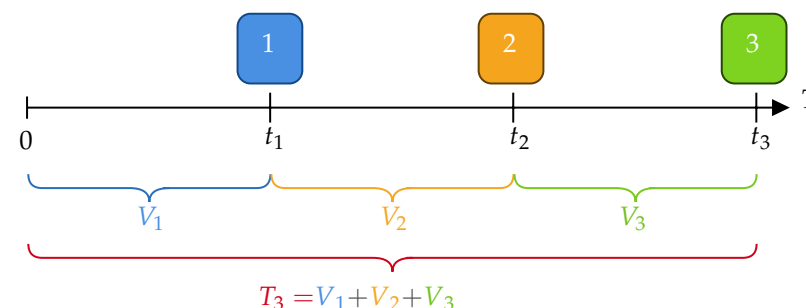
V_k Intervalle de temps entre la réalisation du $(k-1)^e$ et du k^e événement.

> 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

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

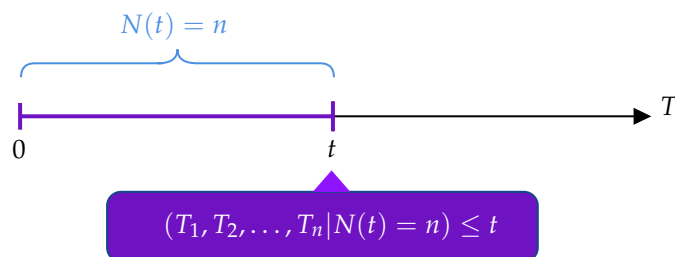
Note La loi Gamma avec un paramètre de forme α entier correspond à la loi Erlang. L'avantage de la loi Erlang est qu'elle a une fonction de répartition explicite qui découle de la relation entre les processus de Poisson et les temps d'occurrences. Voir la sous-section sur les **Distributions** du chapitre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

Temps d'occurrence conditionnels

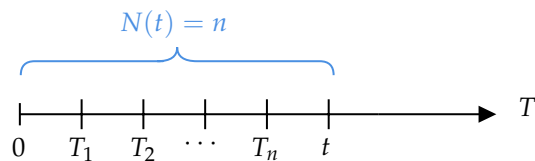
Note Voir la sous-section des **Statistiques d'ordre** du chapitre de Analyse statistique des risques actuariels.

Lorsque nous savons qu'un certain nombre d'événements se produit d'ici un temps t , les temps d'occurrences T_1, T_2, \dots, 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 $T_1, T_2, \dots, T_n \leq t$:



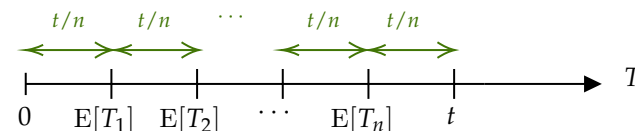
On en déduit que les temps d'occurrences sont en fait des **Statistiques d'ordre** avec $0 < T_1 \leq T_2 \leq \dots \leq T_n \leq t$:



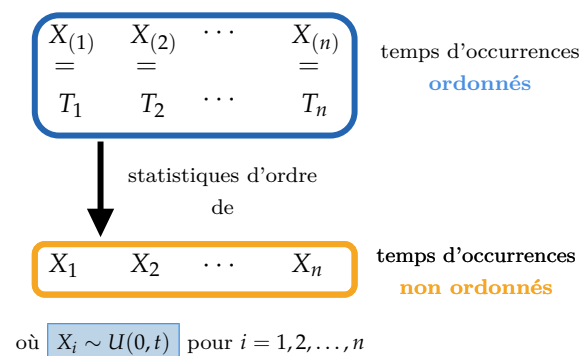
Pour déterminer la distribution de T_i , $i = 1, 2, \dots, n$, on rappelle ces deux propriétés des processus de Poisson homogènes :

- 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.

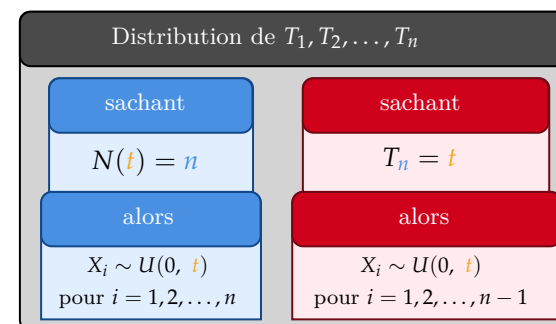
On en déduit que les temps d'occurrences des événements devraient être **uniformément distribués** en $n+1$ sous-intervalles :



Donc, T_1, T_2, \dots, T_n sont les statistiques d'ordre d'une distribution $U(0, t)$:



En bref :



Également, lorsque $X_k \sim U(a, b)$ pour $k = 1, 2, \dots, n$, on trouve que $E[X_{(k)}] = E[T_k] = a + \frac{k(b-a)}{n+1}$.

Note Voir la Cas spéciaux des Statistiques d'ordre pour la définition de cette espérance.

Exemple

Des autobus arrivent à un arrêt d'autobus selon une distribution de Poisson avec un paramètre de fréquence de $\lambda = 4$ par heure. Les autobus commencent à arriver dès 8h du matin.

On sait qu'aujourd'hui, trois autobus sont passés entre 8h et 9h du matin.

Calculer :

1. L'espérance du temps d'arrivée du 5^e bus,
2. L'espérance du temps d'arrivée 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'arrivée du 5^e bus :

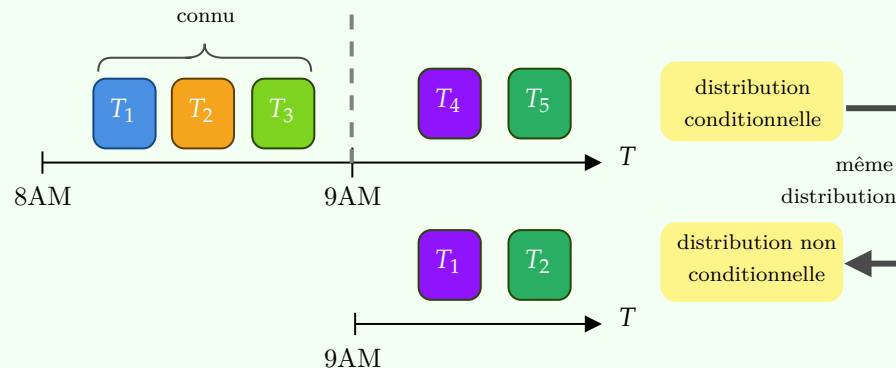
- 1 On connaît l'intervalle de temps durant laquelle les 3 premiers autobus arrivent.

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 :

$$E[T_5 | N(8,9) = 3] = E[T_2]$$

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, $E[T_2] = \frac{2}{4} = 0.50$ ce qui équivaut à 9h30AM.

Deuxièmement, l'espérance du temps d'arrivée du 2^e bus :

- 1 On connaît l'intervalle de temps durant laquelle les 3 premiers autobus arrivent.

Ceci implique que le temps d'arrivée du 2^e doit être à, ou avant, 9AM—alias, T_2 a **eu lieu** et **est borné**.

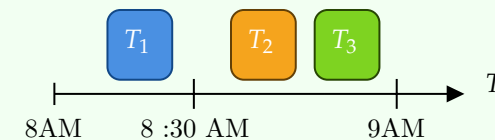
- 2 Il s'ensuit que T_2 ne suit pas une distribution Gamma et que l'espérance conditionnelle de son temps d'arrivée, T_2 , équivaut à l'espérance de la 2^e statistique d'ordre, $X_{(2)}$, des temps d'arrivées non ordonnés X_k distribués **uniformément** entre 8AM et 9AM ($U(8,9)$) pour $k = 1, 2, 3$:

$$E[T_2 | N(8,9) = 3] = E[X_{(2)}] = 8 + \frac{2 \times (9 - 8)}{3 + 1} = 8.5$$

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 $X_k \sim U(8,9)$ 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 définit 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, $\Pr(N(8,8.5]) = 1 | N(8,9) = 3) =$
 $\Pr(\text{un autobus arrive entre 8h00 et 8h30} \cap \text{2 autobus arrivent entre 8h30 et 9h00}) = \Pr(1 \text{ succès}) =$
 $\binom{3}{1} 0.5^1 (1 - 0.5)^2 = 0.375$

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, \dots, N_n de N dont les proportions sont $\pi_1, \pi_2, \dots, \pi_n$.

Alors, N_1, N_2, \dots, N_n sont des processus de Poisson indépendants avec paramètre de fréquence $\pi_1\lambda(t), \pi_2\lambda(t), \dots, \pi_n\lambda(t)$.

Si le processus N est homogène et que les **proportions** π_i sont **constantes**, pour $i = 1, 2, \dots, n$, 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, \dots, N_n avec paramètres de fréquence $\lambda_1(t), \lambda_2(t), \dots, \lambda_n(t)$.

Alors, $N_1 + N_2 + \dots + N_n$ est un processus de Poisson avec paramètre de fréquence $\lambda = \lambda_1(t) + \lambda_2(t) + \dots + \lambda_n(t)$.

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 .

$$\Pr \left(\begin{array}{c} \text{d'observer 1 événement de } N_1 \text{ avant} \\ \text{d'observer 1 événement de } N_2 \end{array} \right) = \Pr(T_{1,1} < T_{2,1}) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

On peut généraliser ceci pour trouver une distribution **binomiale négative** ou **binomiale** :

$$\begin{aligned} \Pr \left(\begin{array}{c} \text{d'observer } n \text{ événements de } N_1 \text{ avant} \\ \text{d'observer } m \text{ événement de } N_2 \end{array} \right) &= \Pr(T_{1,n} < T_{2,m}) \\ &= \Pr \left(\begin{array}{c} \text{d'observer au plus } m-1 \text{ événements de } N_2 \text{ avant} \\ \text{d'observer le } n^e \text{ événement de } N_1 \end{array} \right) \\ &= \sum_{k=0}^{m-1} \binom{n+k-1}{n-1} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^n \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^k \\ &= \Pr \left(\begin{array}{c} \text{parmi les } n+m-1 \text{ premiers événements} \\ \text{au moins } n \text{ proviennent de } N_1 \text{ et} \\ \text{au plus } m-1 \text{ proviennent de } N_2 \end{array} \right) = \Pr \left(\begin{array}{c} n^e \text{ événement de } N_1 \text{ se produise avant} \\ \text{le } m^e \text{ événement de } N_2 \end{array} \right) \\ &= \sum_{k=n}^{n+m-1} \binom{n+m-1}{k} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^k \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{(n+m-1)-k} \end{aligned}$$

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 $X_1, X_2, \dots, X_{N(t)}$.

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

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

✓ Approximation de la distribution

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

Il s'ensuit que :

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

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 dénotés 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 $\{X_m, m = 0, 1, 2, \dots\}$. 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** P :

$$P = \begin{bmatrix} P_{1,1} & P_{1,2} & \dots & P_{1,j} & \dots \\ P_{2,1} & P_{2,2} & \dots & P_{2,j} & \dots \\ \vdots & \vdots & \ddots & \vdots & \dots \\ P_{i,1} & P_{i,2} & \dots & P_{i,j} & \dots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

- › 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, $P^{(n+m)} = P^{(n)} P^{(m)}$.

Rappel : Multiplication matricielle

Soit $A_{m \times n}$ et $B_{p \times q}$. Si $n = p$ alors $A_{m \times n} B_{p \times q} = AB_{m \times q}$.

Par exemple, pour :

$$A = \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \quad B = \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix}$$

Alors :

$$AB = \begin{bmatrix} a_{1,1}b_{1,1} + a_{1,2}b_{2,1} & a_{1,1}b_{1,2} + a_{1,2}b_{2,2} \\ a_{2,1}b_{1,1} + a_{2,2}b_{2,1} & a_{2,1}b_{1,2} + a_{2,2}b_{2,2} \end{bmatrix}$$

Raccourci On peut éviter deux multiplications de matrices en multipliant uniquement la rangée i et la colonne j : $P_{i,j}^n = P_{i,\cdot} \cdot P^{\cdot,n-2} \cdot P_{\cdot,j}$.

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 définit 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

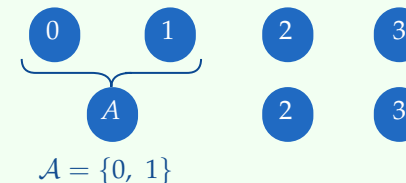
\mathcal{A} L'ensemble des états à éviter.

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 :



Construction de la matrice Q

On construit Q de 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, $\Pr(\text{transition d'un état absorbant vers tout autre état}) = 0$: $Q_{A,i} = 0$ pour $i \notin \mathcal{A}$.

④ Par définition, $\Pr(\text{demeurer dans un état absorbant}) = 1$: $Q_{A,A} = 1$.

Finalement, on vérifie que chaque rangée de 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) :

$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 & 0 \\ 0 & 0.7 & 0.2 & 0.1 \\ 0.6 & 0.2 & 0 & 0.2 \\ 0.8 & 0.1 & 0.1 & 0 \end{bmatrix}$$

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 définit l'ensemble $\mathcal{A} = \{3, 4\}$.

② On définit la nouvelle chaîne de Markov 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.

$$Q = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0 & 0.7 & 0.3 \\ 0 & 0 & 1 \end{bmatrix}$$

③ On trouve la matrice de transitions en 2 étapes :

$$Q = \begin{bmatrix} 0.25 & 0.36 & 0.39 \\ 0 & 0.49 & 0.51 \\ 0 & 0 & 1 \end{bmatrix}$$

④ Finalement, on trouve $Q_{1,2}^4 = Q_{1,2} Q_{1,2}^2 Q_{1,2}$:

$$Q_{1,2}^4 = \begin{bmatrix} 0.5 & 0.3 & 0.2 \end{bmatrix} \begin{bmatrix} 0.25 & 0.36 & 0.39 \\ 0 & 0.49 & 0.51 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.3 \\ 0.7 \\ 0 \end{bmatrix} = 0.2664$$

Transitions de (ou vers) un état absorbant

Notation

$Q_{i,j}^m$ 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'arrivée 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
$i \notin \mathcal{A}$	$j \notin \mathcal{A}$	$Q_{i,j}^m$
$i \notin \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} Q_{i,r}^{m-1} P_{r,j}$
$i \in \mathcal{A}$	$j \notin \mathcal{A}$	$\sum_{r \notin \mathcal{A}} P_{i,r} Q_{r,j}^{m-1}$
$i \in \mathcal{A}$	$j \in \mathcal{A}$	$\sum_{r \notin \mathcal{A}} \sum_{k \notin \mathcal{A}} P_{i,r} Q_{r,k}^{m-2} P_{k,j}$

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.

> $\Pr(X_n = j) = \sum_{i=1}^{\infty} \alpha_i P_{i,j}^n$.

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 $n \geq 0$: $i \rightarrow j$.

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 $i \rightarrow j$ et $j \rightarrow i$.

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 futur 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 nous 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 désirée 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 :

$$P = \begin{bmatrix} 0.7 & 0.3 & 0 \\ 0 & 0.4 & 0.6 \\ 0 & 0.5 & 0.5 \end{bmatrix}$$

On trouve :

- > Aucun état est absorbant.
- > L'état 1 est *transitoire* et seulement l'état 2 est accessible de l'état 1 ($1 \rightarrow 2$).
- > 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,

puisque 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 :

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Cette chaîne de Markov ne converge pas vers des probabilités limites, à chaque période elle va inverser :

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

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

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

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

☰ Chaîne de Markov ergodique

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

Temps passé dans les états transitoires

Notation

P_T Matrice des probabilités de transition contenant 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 .

S Matrice des valeurs de $s_{i,j}$.

> $S = (I - P_T)^{-1}$.

> **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 .

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

Note En anglais, on dit « *Time Spent in Transient States* ».

Rappel : Matrice d'identité

La matrice d'identité I est la suivante :

$$I = \begin{bmatrix} 1 & 0 & 0 & \dots \\ 0 & 1 & 0 & \dots \\ 0 & 0 & 1 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

On exprime les valeurs de I avec la variable binaire $\delta_{i,j}$:

$$\delta_{i,j} = \begin{cases} 1, & \text{si } i = j \\ 0, & \text{si } i \neq j \end{cases}$$

Rappel : Inverse d'une matrice

Notation

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

Soit la matrice 2×2 A où :

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

Alors son inverse A^{-1} est :

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

> Pour plus de 3 dimensions, c'est long et peu probable d'être dans l'examen.

Exemple du calcul du temps espéré

Soit la chaîne de Markov à trois états (1, 2, 3) avec la matrice de transition suivante :

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

On souhaite trouver l'espérance du nombre de périodes passées dans l'état 2 sachant qu'on débute dans l'état 1.

1 Trouver la matrice de transitions pour les états transitoires :

$$P_T = \begin{bmatrix} 0.5 & 0.5 \\ 0 & 0.8 \end{bmatrix}$$

2 Trouver $I - P_T$:

$$I - P_T = \begin{bmatrix} 0.5 & -0.5 \\ 0 & 0.2 \end{bmatrix}$$

3 Trouver l'inverse $(I - P_T)^{-1}$:

$$(I - P_T)^{-1} = \begin{bmatrix} 0.2 & 0.5 \\ 0 & 0.5 \end{bmatrix} \times \frac{1}{0.10 - 0} = \begin{bmatrix} 2 & 5 \\ 0 & 5 \end{bmatrix}$$

4 Trouver l'élément $s_{1,2}$ de la matrice $S = (I - P_T)^{-1}$ et donc $s_{1,2} = 5$.

« Time Reversibility »

Notation

$R_{i,j}$ Probabilité de transition de l'état i à l'état j (en une période) pour la chaîne de Markov inverse.

➤ On dénote la matrice des probabilités de transition de la chaîne de Markov inverse par \mathbf{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 :

$$\mathbf{P} = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.1 & 0.6 & 0.3 \\ 0.2 & 0.1 & 0.7 \end{bmatrix}$$

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 \mathbf{R} :

$$(a) \mathbf{R}_{11} = \mathbf{P}_{11} \frac{\pi_1}{\pi_1} = 0.3$$

$$(b) \mathbf{R}_{22} = \mathbf{P}_{22} \frac{\pi_2}{\pi_2} = 0.6$$

$$(c) \mathbf{R}_{33} = \mathbf{P}_{33} \frac{\pi_3}{\pi_3} = 0.7$$

$$(d) \mathbf{R}_{12} = \mathbf{P}_{21} \frac{\pi_2}{\pi_1} = 0.1 \times \frac{77/270}{7/30} = 0.12$$

$$(e) \mathbf{R}_{13} = \mathbf{P}_{31} \frac{\pi_3}{\pi_1} = 0.2 \times \frac{182/270}{7/30} = 0.58$$

$$(f) \mathbf{R}_{21} = \mathbf{P}_{12} \frac{\pi_1}{\pi_2} = 0.2 \times \frac{7/30}{77/270} = 0.16$$

$$(g) \mathbf{R}_{23} = \mathbf{P}_{32} \frac{\pi_3}{\pi_2} = 0.1 \times \frac{182/270}{77/270} = 0.24$$

$$(h) \mathbf{R}_{31} = \mathbf{P}_{13} \frac{\pi_1}{\pi_3} = 0.5 \times \frac{7/30}{182/270} = 0.17$$

$$(i) \mathbf{R}_{32} = \mathbf{P}_{23} \frac{\pi_2}{\pi_3} = 0.3 \times \frac{77/270}{182/270} = 0.13$$

3 Construire la matrice des probabilités de transition inverse :

$$\mathbf{R} = \begin{bmatrix} 0.30 & 0.12 & 0.58 \\ 0.16 & 0.60 & 0.24 \\ 0.17 & 0.13 & 0.70 \end{bmatrix}$$

Applications des chaînes de Markov

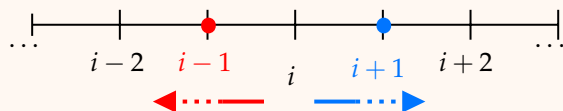
Marche aléatoire

Marche aléatoire

1 À une dimension

Une *marche aléatoire* à une dimension équivaut à une chaîne de Markov qui, de l'état i , peut seulement aller soit à l'état $i+1$ avec probabilité $P_{i,i+1} = p$ ou l'état $i-1$ avec probabilité $P_{i,i-1} = 1 - p$ où $p \in [0, 1]$.

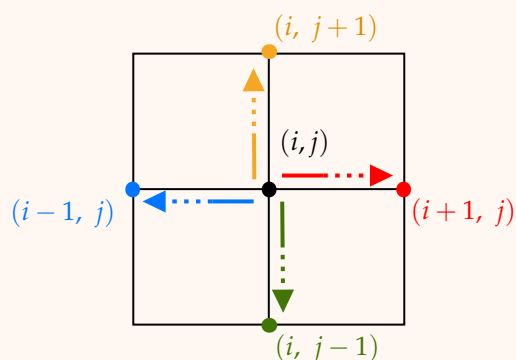
On peut donc visualiser une ligne :



2 À deux dimensions

Une marche aléatoire de deux dimensions représente chaque état comme une paire de chiffres (i, j) et donc le prochain état peut être un des quatre états suivants : $(i-1, j)$, $(i+1, j)$, $(i, j-1)$, $(i, j+1)$.

On peut donc visualiser un carré en représentant l'état comme des coordonnées :



Marche aléatoire symétrique

S'il y a une probabilité égale d'aller dans toute direction, la marche aléatoire est *symétrique*. Par exemple, dans le cas d'une dimension $p = 0.5$ et dans 2 dimensions $p = 0.25$.

Les marches aléatoires sont seulement **récurrentes** si elles ont une ou deux dimensions et qu'elles sont symétriques. Autrement, elles sont **transitoires**.

« Gambler's ruin »

Notation

P_i Probabilité de commencer i jetons et terminer avec j jetons.

> Le complément $1 - P_i$ est la probabilité de commencer avec i jetons et de terminer avec aucun (0).

X Variable aléatoire du nombre de jetons que le « gambler » a à la fin.

« Gambler's ruin problem »

Soit un jeu où, à chaque ronde, un « gambler » *gagne* un jeton avec probabilité p ou *perd* un jeton avec probabilité $1 - p$. L'objectif est de se rendre à j jetons.

Le « gambler's ruin problem » est de calculer la probabilité qu'un « gambler » qui commence avec i jetons va terminer le jeu avec j jetons.

« Gambling model »

Le modèle qu'on utilise pour modéliser le « gambler's ruin problem » se nomme le « gambling model ». Il s'apparente à la marche aléatoire sauf qu'il comporte un nombre **fini** d'états. Les états correspondent au nombre de jetons.

Propriétés du « gambling model »

- 1 Puisque le « gambler » arrête lorsqu'il a soit 0 ou j jetons, $P_{0,0} = P_{j,j} = 1$.
 > Il s'ensuit que les états 0 et j sont **absorbants**.
- 2 La probabilité de gagner $P_{i,i+1} = p$ et la probabilité de perdre $P_{i,i-1} = 1 - p$ où $i \in \{1, 2, \dots, j-1\}$.
- 3 Il y a 3 classes : $\{0\}, \{1, 2, \dots, j-1\}, \{j\}$.
- 4 Les états $\{0\}$ et $\{j\}$ sont récurrents puisqu'ils sont **absorbants** et les états $\{1, 2, \dots, j-1\}$ sont transitoires.

« Branching Process »

Contexte

On pose que nous avons une population d'individus dont chacun produit j descendants d'ici la fin de leur durée de vie avec probabilité P_j .

Le nombre moyen de nouveaux descendants qu'un individu produit est

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

La variance du nombre de nouveaux descendants qu'un individu produit est

$$\sigma^2 = \sum_{j=0}^{\infty} (j - \mu)^2 P_j.$$

Notation

X_n Taille de la n^e génération.

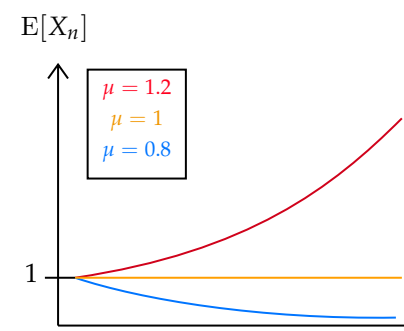
Si l'on pose une population initiale de 1 ($X_0 = 1$) :

$$E[X_n] = \mu^n$$

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

> Si la population initiale est de k ($X_0 = k$) alors la moyenne est de $kE[X_n]$ 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 définit la probabilité que la population disparaisse π_0 si $X_0 = 1$ comme suit :

$$\pi_0 = \begin{cases} 1, & \mu \leq 1 \\ \sum_{j=0}^{\infty} \pi_0^j P_j, & \mu > 1 \end{cases}$$

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

- › 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 probabilité que la population disparaisse est π_0^k .

Sixième partie

Séries chronologiques