2001

Mémoire de Stage
de
troisième année

Belguise Olivier

sur

Tempêtes : Etude des dépendances entre les branches Auto et Incendie avec la théorie des copulas

Stage effectué chez

Guy Carpenter

47/53,rue Raspail
92594 Levallois-Perret

du 28 mai au 31 octobre 2001

Maître de Stage :

M. Charles Lévi
Tel :01-56-76-48-06
e-mail :Charles.Levi@guycarp.com

Remerciements

Je tiens, en tout premier lieu, à remercier Monsieur Charles Lévi pour toute l'attention qu'il m'a portée et pour avoir su me faire profiter de son expérience tout au long de ce stage.

Je remercie également Monsieur Denis Frebourg qui m'a fait partager sa connaissance du monde de la réassurance et avec qui j'ai pris plaisir à travailler pendant ces cinq mois.

Mes remerciements vont également à Messieurs Christian Robert et Stuart Klugman qui ont pris le temps de répondre à mes nombreuses questions sur le sujet.

Je remercie enfin tout le personnel de Guy Carpenter pour son accueil chaleureux.

Sommaire

Introduction 1

1 Première Partie: Contexte de l'étude et actualisation des données 5

1.1 Présentation de la base de données 6

1.2 Présentation du mécanisme de l'excédent de sinistres. 9

1.3 Actualisation du montant des tempêtes 12

1.3.1 Prise en compte de l'inflation 12

1.3.2 Prise en compte de l'évolution du nombre de contrats en portefeuille 13

1.3.3 Prise en compte de l'effet franchise 14

1.3.4 Prise en compte de l'élargissement du champ de la garantie 16

1.4 Reconstitutions des tempêtes et statistiques descriptives sommaires 16

2 Seconde Partie: Théorie des copulas 19

2.1 Quelques rappels concernant les distributions multivariées et la loi uniforme 21

2.1.1 Définition de la fonction de répartition conjointe 21

2.1.2 Définition des lois marginales 21

2.1.3 Lois de probabilités absolument continues 21

2.1.4 Notion d'indépendance 22

2.1.5 Les bornes de Fréchet-Hoeffding 22

2.1.6 Définition et propriétés de la loi uniforme 22

2.2 Définition et propriétés des copulas 23

2.2.1 Définition 23

2.2.2 Le théorème de Sklar 24

2.2.3 Corollaire du théorème de Sklar 24

2.2.4 Propriété d'invariance 25

2.2.5 "Survival Copula" ou "Flipped Copula" 25

2.2.6 Densité des copulas 26

2.2.7 Distribution conditionnelle et copulas 27

2.3 Exemples de copulas 27

2.3.1 La copula d'indépendance 27

2.3.2 La copula comonotone ou de dépendance totale positive 28

2.3.3 La copula antimonotone ou de dépendance totale négative 29

2.3.4 La copula normale 29

2.3.5 Les copulas Archimédiennes 31

2.3.5.1 Définition d'une copula archimédienne 31

2.3.5.1.1 Définition préliminaire 31

2.3.5.1.2 Théorème 31

2.3.5.2 La copula de Frank 32

2.3.5.3 La copula de clayton 34

2.3.5.4 La copula HRT 35

2.3.5.5 La copula de Gumbel 36

2.4 Dépendance 39

2.4.1 Le coefficient de correlation linéaire 39

2.4.2 Notion de dépendance parfaite 40

2.4.3 Les coefficient de corrélation de Kendall et de Spearmann 40

2.4.3.1 Le τ de Kendall 41

2.4.3.2 Le ρ de Spearman 42

2.4.4 La notion de "tail dépendence" ou "dépendance de queue" 43

2.4.4.1 Upper tail dependence 44

2.4.4.2 Lower tail dependence 45

2.5 Approche empirique des copulas ____ 46

2.5.1 Quelques rappels sur les distributions empiriques ____ 46

2.5.2 Expression empirique des copulas et des mesures de dépendance associées ____ 46

2.6 Le choix de la bonne copula ____ 49

2.6.1 La fonction K(z) ____ 49

2.6.2 La fonction J(z) ou de tau cumulatif ____ 52

2.6.3 La fonction M(z) ____ 55

2.6.4 Les fonctions L(z) et R(z) ou "tail concentration functions" ____ 57

3 Troisième partie: Adéquation à une loi bivariée grâce aux copulas ____ 60

3.1 Modélisation du phénomène tempête pour le premier niveau de sélection ____ 61

3.1.1 Détermination des copulas à utiliser pour modéliser (X,Y) ____ 63

3.1.2 Détermination des lois marginales ____ 67

3.1.3 Vérification de la qualité des ajustements ____ 70

3.1.3.1 Le test de Kolmogorov ____ 70

3.1.3.2 Le test d'Andersson-darling ____ 72

3.1.3.3 Le test du khi-deux ____ 72

3.1.4 Détermination des paramètres de la distribution bivariée ____ 73

3.1.5 Test d'adéquation d'une distribution bivariée ____ 76

3.1.5.1 La ligne de régression médiane ____ 76

3.1.5.2 Test du Khi-deux ____ 78

3.1.5.2.1 Extension du test à deux dimensions ____ 78

3.1.5.2.2 Utilisation de deux test unidimensionnels ____ 79

3.1.6 Etude des lois marginales annexes ____ 80

3.2 Modélisation du phénomène tempête pour le second niveau de sélection 82

3.2.1 Détermination de la copula à utiliser 82

3.2.2 Détermination des lois marginales à utiliser et test d'adéquations 84

3.2.2.1 Mean excess function 84

3.2.2.2 Résultats obtenus 85

3.2.2.3 test du khi-deux 87

3.3.3 Détermination des paramètres de la distribution bivariée et tests d'adéquation 88

3.3.4 Etude des lois marginales annexes 89

4 Quatrième partie: Simulations 90

4.1 Présentation de la méthode employée pour la simulation 91

4.1.1 Synthèse des résultats obtenus 91

4.1.2 Utilisation de la méthode de Monte-Carlo 91

4.2 Présentation des résultats obtenus 94

4.2.1 Comparaison entre deux excédents de sinistres par branche et un excédent sur la somme des deux branches 94

4.2.2 Etude d'un excédent de sinistre automobile dont le paiement est conditionné par la branche incendie. 97

4.2.3 Un contrat "particulier" 99

Conclusion 101

Bibliographie 105

Annexes 107

Introduction

Cette fin de siècle a été marquée par de nombreux événements climatiques de toute première importance.
Les tempêtes de 1990, dont tout le monde s'accordait à l'époque à dire qu'il s'agissait des tempêtes du siècle, en sont un premier exemple. Que dire alors de «Lothar » et «Martin » survenus fin 1999 et qui touchèrent la France de plein fouet. «Tempête du millénaire » ou «événement exceptionnel » furent les qualificatifs les plus souvent employés.
Toujours est-il que ces événements ont amené assureurs et surtout réassureurs à se pencher de manière plus approfondie sur ce type de phénomène.
Même s'il est vrai que lorsque surviennent des tempêtes, les bâtiments sont les principaux biens assurés endommagés, l'assurance automobile constitue également une part non négligeable du montant total imputable aux tempêtes.

Afin de coter certains contrats de réassurance prenant en compte l'impact des tempêtes sur les branches «Automobile » et «Incendie-Tempête-Grêle-Neige », une estimation de la distribution conjointe de ces deux types de dommages s'avère nécessaire voire impérative. Il est évident que nous ne pouvons supposer l'indépendance entre ces deux branches sous peine d'aboutir à des résultats erronés.

Jusqu'à présent la plupart des techniques utilisées présentaient certains inconvénients. Nous pouvons, à titre d'exemple, citer l'utilisation de distributions bidimensionnelles comme la Pareto bivariée. Mais dans ce précis, les lois marginales seront nécessairement des lois de Pareto. L'idéal serait donc de pouvoir créer des distributions multivariées et à fortiori bivariées en choisissant nous-mêmes nos lois marginales.

La théorie des copulas va nous permettre de répondre à toutes ces attentes.

Le mot copula, dont la signification en latin est littéralement *lien*, a été employé pour la première fois en 1959 par Abe Sklar dont le théorème constitue la clef de voûte de toute la théorie.

Ces copulas sont tout simplement, et sans rentrer dans les détails, des fonctions qui vont constituer un véritable *lien* entre la distribution multivariée et les lois marginales.

Mais bien que cette théorie date de la fin des années 50, les exemples d'application ne sont pourtant pas légion, et les mémoires d'actuariat traitant du sujet sont quasiment inexistants. A vrai dire seuls deux articles parus relativement récemment (« Fitting bivariate loss distributions with copulas » de Klugman-Parsa et « Understanding relationship using copulas » de Frees-Valdez) ont une approche actuarielle des copulas. En ce qui concerne les bases théoriques et mathématiques à proprement parler, l'ouvrage de Nelsen « An introduction to copulas » fait figure de référence.

Les données qui serviront de base à notre étude nous ont été fournies par une grande compagnie d'assurance française.
La base de données utilisée comprend tous les sinistres dus aux tempêtes sur une période de 11 ans, à compter du 1er janvier 1990 jusqu'au 31 décembre 2000.

Cette base de données sera présentée de manière plus explicite dans notre première partie.
En outre, et compte tenu du fait que la période sur laquelle notre historique est basé est relativement étendue, l'actualisation du montant des sinistres fera également l'objet d'une présentation au cours de ce chapitre.
Nous y donnerons également quelques notions très élémentaires de réassurance et nous y présenterons notamment le traité en excédent de sinistre.

La deuxième partie présentera la théorie des copulas dans ses grandes lignes, notamment le théorème de Sklar, clarifiera la notion de dépendance. Pour des informations plus précises au sujet de ce concept, le lecteur pourra se reporter à l'article *Correlation and dependence in risk management : properties and pitfall* écrit par Embrechts-McNeil-Strauman.
Cette partie mettra également en avant les copulas utilisées au cours de notre étude, et donnera toutes les formules nécessaires à l'estimation de paramètres par la méthode du maximum de vraisemblance.
D'autre part, des techniques nous permettant de choisir la copula susceptible de correspondre au mieux à une certaine série de données y seront également exposées.

Dans notre troisième partie, nous utiliserons les copulas dans le but d'évaluer une distribution bivariée des montants « automobile » et « incendie ». Nous déterminerons dans un premier temps la ou les copulas que nous allons juger susceptibles de convenir grâce aux techniques exposées dans la seconde partie. Nous déterminerons également la ou les lois marginales permettant la meilleure adéquation possible dans les branches automobiles et multi-risques.
Enfin, nous estimerons les paramètres de notre distribution bivariée avec les lois marginales et les copulas retenues. Bien évidemment, nous effectuerons également les tests usuels pour voir si notre adéquation est acceptable.

L'objet de la quatrième partie sera l'évaluation de divers contrats de réassurance grâce à notre distribution bivariée et à la méthode de simulation de Monte-Carlo.
La plupart de ces contrats ne pourrait en aucun cas être évalués si nous n'avions au préalable effectué une adéquation à une loi bivariée.

Les outils utilisés pour mener à bien cette étude seront les logiciels SAS (modules Base, Stat) et Excel.

Première Partie

Contexte de l'étude et Actualisation des données

1.1Présentation de la base de données

Comme nous l'avons mentionné dans l'introduction, la base de données utilisée provient d'une grande compagnie d'assurance française et comprend leur historique tempête établi sur la période débutant au 1er janvier 1990 et s'arrêtant au 31 décembre 2000.
Il s'agit donc uniquement d'un fichier « sinistres ».

Par sinistre nous entendons toute atteinte portée à l'intégrité d'un bâtiment ou d'un bien, en un endroit et un lieu précis, et dont les conséquences sont éventuellement garanties par un contrat d'assurance souscrit auprès de notre compagnie.
Notre base de données ne comprendra donc que des sinistres aux conséquences matérielles directes (pas de perte d'exploitation) et en aucun cas des éléments corporels.
En fait la garantie tempête, pour ce qui est des dommages causés aux bâtiments n'est qu'une **des garanties incluse dans la garantie incendie**. Néanmoins, les assureurs éprouvent quelques difficultés à harmoniser leur définition de la tempête. La définition la plus couramment admise est la suivante :
«*Evénement météorologique au cours duquel le vent a une violence telle qu'il détruit, brise ou endommage un certain nombre de bâtiments de bonne construction, d'arbres et autres objets dans un rayon de 5 km autour du risque assuré »*.

Cette définition n'est donc pas très tranchée. Nous risquons donc ne trouver dans cette base des sinistres ne résultant que de simples coups de vents très localisés.

Il est également important de noter que notre base se limite aux sinistres survenus en France Métropolitaine.

Elle se présente sous la forme de fichiers Excel et comporte un peu plus de 155000 enregistrements. Néanmoins, ces données seront converties au format SAS afin d'en faciliter le traitement.

Ces sinistres sont «vus » dans le courant de l'année 2001, mais dans la mesure où les sinistres tempêtes sont réglés en général très rapidement, nous considérerons ces montants comme la charge finale du sinistre.
Un exemple d'enregistrements :

Département	Exercice	Catégorie de risque	Coût	Date Survenance
11	1995	MR Agricole	67 038,00	11/01/95
75	1995	MR Agricole	2 678,00	26/01/95
66	1995	MR Agricole	1 199,69	23/08/95
30	1995	Auto 4 roues	415,1	01/01/95
48	1995	Auto 4 roues	5 349,11	11/01/95
69	1995	Auto Lourds	10 025,00	18/01/95
29	1995	Auto 4 roues	2 023,13	19/01/95
80	1995	Auto 4 roues	480,33	08/01/95
49	1995	Auto 4 roues	3 366,45	17/01/95
37	1995	Auto 4 roues	14 799,07	22/01/95
12	1995	Auto 4 roues	260,92	20/01/95
66	1995	Auto 2 roues	850,2	12/03/95
79	1995	MR Particulier	39 460,51	22/01/95
64	1995	MR Particulier	5 445,78	18/01/95
56	1995	MR Particulier	2 623,72	19/01/95
83	1995	MR Particulier N Occupant	1 067,40	13/01/95
41	1995	MR Particulier	28 692,00	26/01/95
56	1995	MR Particulier N Occupant	17 576,20	19/01/95
83	1995	MR Particulier N Occupant	5 430,03	13/01/95

Cette base de données comprend donc cinq champs :

- **Département :** Celui-ci ne correspond pas, comme certains auraient pu s'y attendre, au numéro du département où se trouve le bien assuré, mais à celui ou se trouve le souscripteur. Ainsi, dans de nombreux cas, le département mentionné pourra se trouver dans la région parisienne, alors que le sinistre se sera produit dans une habitation secondaire, en Normandie par exemple. Néanmoins, comme nous ne prendrons pas en compte le critère géographique dans la suite de notre étude, cela ne portera pas à conséquence.

- **Exercice :** Ce champ correspond tout simplement à l'exercice auquel le sinistre est rattaché.

- **Catégorie de risque :** Cette variable comprend la nature du sinistre occasionné par la tempête. Elle nous permettra de rattacher un sinistre à la branche auto ou à la branche incendie. Cependant, cette variable ne se décline pas uniquement en fonction de ses deux branches. En fait elle se décompose en des «niveaux plus précis » que nous allons ensuite classer dans l'une des deux branches. La branche automobile est par exemple subdivisée en «auto 4 roues », «auto 2 roues », «auto remorques légères »…
  La branche incendie est constituée quant à elle des dommages pouvant affecter aussi bien les particuliers, que les entreprises ou encore les exploitations agricoles. Nous pourrons trouver dans la base des enregistrements de type « MR Particulier », « MR commerces » ou encore « MR Agricole ». **C'est pourquoi, nous utiliserons parfois le terme de multirisque à la place d'incendie**.

- **Coût :** Ce champ contient l'information la plus importante, à savoir le préjudice subi par la compagnie lors de la survenance du sinistre. Il correspond donc à la charge finale du sinistre réglé par cette compagnie à ses assurés exprimée en francs. Il conviendra d'ajuster ce coût à des données comme l'inflation par exemple.

- **Date survenance :** Cette variable correspond comme son nom l'indique à la date de survenance de l'événement ayant mis en cause la garantie tempête. C'est donc à partir de ce champ que nous allons regrouper les sinistres individuels pour estimer le coût total de l'événement.

Le nombre de champs disponibles est donc somme toute plutôt restreint. La présence d'un numéro d'événement, ou encore de l'origine exacte des dommages (tempête, grêle, neige) auraient également constitué un apport appréciable.

L'intérêt de cette remarque réside dans le fait que la plupart des contrats de réassurance portant sur les tempêtes sont intimement liés à la notion d'événement.

Par abus de langage, dans le reste de l'exposé, nous parlerons de tempête quel que soit le type d'événement atmosphérique étudié.
Nous allons présenter maintenant le contrat de réassurance en excédent de sinistres et les clauses horaires s'y rapportant.

1.2 Présentation du mécanisme de l'excédent de sinistres

Avant de faire la présentation d'un excédent de sinistres, il convient tout d'abord de situer ce type de réassurance par rapport aux autres modes de réassurance envisageables en en effectuant une présentation succincte.

La réassurance peut tout d'abord être à caractère facultatif ou obligatoire.
Lorsque l'on parle de réassurance facultative, les polices réassurées sont prises une à une et examinées au cas par cas. Celle-ci peut donc être utilisée par exemple pour des grands risques industriels.

Dans le cadre de la réassurance obligatoire, il s'agit de réassurer un ensemble de polices qui appartiennent à un cadre déterminé au préalable. Ainsi, toute police entrant dans ce cadre prédéfini doit être réassurée. Il s'agit en fait de la forme de réassurance la plus courante.

Les traités de réassurance obligatoire se décomposent en fait en deux types de réassurance

- Les traités proportionnels pour lesquels le principe consiste à répartir dans le même rapport la prime et les capitaux de la police réassurée ainsi que tous les sinistres touchant cette police.

- Les traités non-proportionnels, dans lesquels, comme leur nom l'indique, le réassureur ne couvre pas proportionnellement les sinistres en fonction des primes.

La réassurance en Excédent de sinistres entre dans le cadre de ces traités non proportionnels.
Ces traités en excédent de sinistres se décomposent eux-mêmes en deux catégories.

- Les excédents de sinistres par risque dans lesquels le réassureur couvre un sinistre touchant un risque individuel.

- Les excédents de sinistres par événements pour lesquels le réassureur couvre un ensemble de sinistres individuels liés par un même événement.

Pour les tempêtes, l'excédent de sinistre par événement est, est-il besoin de le préciser, la forme utilisée.
Un événement débute au plus tôt à la survenance du premier sinistre à la charge de la cédante.
Un événement tempête contient une clause horaire : celle-ci est de 72 heures dans le cas d'une tempête, cyclone, ouragan ou tornade et de 24h dans le cas
de la grêle. Un de ces phénomènes naturels qui se poursuivrait éventuellement au-delà des durées maximales spécifiées constituerait ainsi plusieurs événements.
Nos sinistres individuels seront donc regroupés par période de trois jours afin de reconstituer le montant total de l'événement.

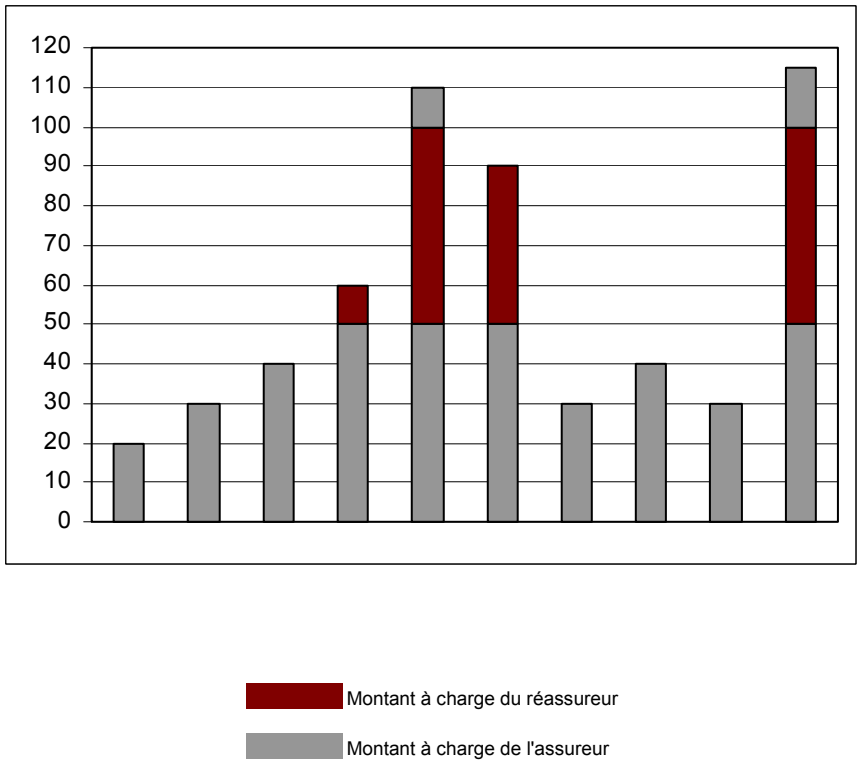
Le principe de l'excédent de sinistre est comparable à celui d'une police avec franchise : le réassureur paie l'assureur dès que le sinistre dépasse un certain montant, que l'on appelle priorité, et ce dans la limite d'un montant défini de manière contractuelle que l'on appelle la portée.
On utilise très couramment la notation :
Portée Xs Priorité.
Nous appellerons plafond la somme portée+priorité.

Donc si X est le montant total occasionné par une tempête, la répartition entre l'assureur et le réassureur s'effectue donc comme suit :

$$\text{Cédante} \begin{cases} X & \text{si } X<\text{Priorité} \\ \text{Priorité} & \text{si Priorité} \le X < \text{Portée+Priorité} \\ X-\text{Portée} & \text{si } X \ge \text{Portée+Priorité} \end{cases}$$

$$\text{Réassureur} \begin{cases} 0 & \text{si } X<\text{Priorité} \\ X-\text{Priorité} & \text{si Priorité} \le X < \text{Portée+Priorité} \\ \text{Portée} & \text{si } X \ge \text{Portée+Priorité} \end{cases}$$

Le traité présenté ci-dessus correspond à un 50 Xs 50.



Ainsi, les sinistres dont le montant n'a pas atteint la priorité ainsi que ceux qui dépassent le plafond restent donc à la charge de la cédante.(après déduction de la portée)

Il est également très fréquent que le besoin de couverture soit trop important pour faire l'objet d'un seul et unique traité. Les engagements sont alors couverts par le ou les réassureurs dans le cadre d'un programme de réassurance non-proportionnel en «superposant plusieurs traités en excédents de sinistre » et on parle alors de tranches.

Il est évident que portée et priorité sont relatives à la taille de la compagnie. La comparaison des primes en montant absolu entre les différentes compagnies devenant dès lors non significative, ceux-ci sont souvent exprimés sous forme de rate-on-line. Le rate-on-line est tout simplement le rapport entre prime et portée de la tranche.

1.3 Actualisation des montants des tempêtes.

La période sur laquelle repose notre historique nous contraint à prendre en compte tous les facteurs ayant pu influencer de par leur évolution les règlements des dommages résultants des tempêtes. Cette étape préliminaire nous a semblé indispensable à la modélisation statistique du phénomène tempête.
Imaginez que Lothar par exemple soit survenu dans le courant de l'année 1992. Combien aurait-il alors coûté ?
Le but vous l'aurez compris, est en quelques sorte d'homogénéiser les conditions de survenance de ces tempêtes.
Pour ce faire, nous avons décider de retenir quatre critères à savoir : l'inflation, l'évolution du nombre de contrats en portefeuille, la prise en compte d'un effet de franchise, et l'élargissement du champ de la garantie sur certains risques.

1.3.1 Prise en compte de l'inflation

En fait, plus que l'indice d'inflation a proprement parler, ce qui nous intéresse tout particulièrement, c'est un indice approprié à la branche à étudier.
Pour la branche multirisque l'indice qui nous a semblé le plus révélateur est l'indice FFB, qui est en fait l'indice du coût de la construction, alors que pour la branche

automobile notre préférence s'est portée sur celui du coût des réparations automobiles. Nous avons jugé l'utilisation de ce dernier indice plus opportune que celle de l'indice du prix des véhicules. L'évolution de cet indice a été un peu plus difficile à obtenir que pour l'indice FFB puisque le seul moyen de se la procurer aura été de consulter tous les argus sur une période de 11 ans.
Ces deux indices ont été globalement croissants au cours du temps, traduisant évidemment une hausse du coût de la construction et des réparations dans le secteur automobile.
Bien que nous disposions de l'évolution de ces indices tous les trois mois, nous nous sommes limités à ne retenir qu'une seule valeur pour chaque année.

Nous avons choisi comme année de référence 1999, et ceci bien évidemment pour tous nos autres facteurs d'actualisation.
En notant FFB_j l'indice retenu pour l'année j (j=1990…2000), tous les montants entrant dans la branche multirisque de l'année ont donc été **multipliés** par le coefficient (FFB_{1999}/FFB_j). Nous avons effectué la même règle de trois pour les sinistres entrant dans la catégorie automobile.
Donc, pour synthétiser, les règlements antérieurs à 1999 ont donc été légèrement augmentés. A titre indicatif, les montants de l'année 1992 par exemple ont été multipliés par 1,2 pour la branche multirisque et par 1,33 pour l'automobile.

1.3.2 Prise en compte de l'évolution du nombre de contrats en portefeuille.

Il est impératif de prendre en compte ce facteur.
Malheureusement, nous ne disposons pas dans notre base de données des éléments nécessaires à établir l'évolution du nombre de polices ou des sommes assurées sur une période de 10 ans. Nous nous sommes renseignés auprès de la compagnie nous ayant fourni les données pour savoir si il aurait été possible d'obtenir ces renseignements complémentaires. Notre interlocuteur nous a signifié que cela aurait nécessité une mobilisation trop importante de leurs moyens informatiques.

Nous avons donc eu recours à des données de «substitution » nous permettant d' «approcher » le mieux possible cette évolution.
Les données utilisées sont issues de deux sources distinctes, à savoir certains documents FFSA et les numéros annuels spéciaux de l'Argus exposant les résultats IARD de toutes les compagnies françaises.
Nous nous sommes tout particulièrement intéressés aux primes émises par «notre » compagnie dans les catégories

- Dommages automobiles
- Dommages aux biens particuliers
- Dommages aux biens professionnels
- Dommages aux biens agricoles

A partir des documents FFSA, nous avons reconstitué, lorsque cela a été possible une prime moyenne de marché chaque année. Nous avons donc pu retracer, certes de manière très approximative, l'évolution du portefeuille de notre compagnie dans chacune de ces quatre branches.
Nous avons à nouveau effectué une simple règle de trois afin de revaloriser les montants des tempêtes et ainsi obtenir des montants pour un portefeuille «homogène » en termes de nombre de polices et de sommes assurées.

$$\text{cout revalorisé} = \text{cout observé l'année j} \times \frac{\text{nombre de polices année de référence}}{\text{nombre de polices année j}}$$

1.3.3 Prise en compte de l'effet franchise

Les sinistres présents dans notre base de données ont bien entendu fait l'objet de règlements auprès des assurés.
Or tous ces montants sont amputés de la franchise qui est resté à la charge de l'assuré.
Le seul problème, c'est que cette franchise a également évolué pendant notre période d'observation. Cette évolution a forcément des répercutions, tant en termes de coût moyen que de nombre de règlements effectués.
Supposons par exemple que la franchise par police ait baissé.
Dans ce cas le nombre de sinistres dépassant le seuil va forcément s'en retrouver augmenté.

Dans une étude interne menée par M. Charles Lévi pour la Compagnie Transcontinentale de Réassurance, l'impact de l'évolution de la franchise sur le coût moyen et le nombre de polices sinistrées a été évalué. Nous avons donc utilisé ces résultats sans faire d'investigations plus poussées. Ceux-ci sont présentés ci-après.

Risques Particuliers		
Franchise	Coût moyen	Nombre
1 FNB	21,10%	4,20%
2 FNB	14,90%	4,20%
3 FNB	9,10%	3,60%
4 FNB	4,10%	2,30%
5 FNB	0%	0%
6 FNB	-3,20%	-3,10%
7 FNB	-5,70%	-7%
8 FNB	-7,70%	-11,30%
9 FNB	-9,10%	-16%
10 FNB	-10,10%	-20,80%

Risques agricoles		
Franchise	Coût moyen	Nombre
5 FNB	7,20%	20,80%
10 FNB	1,50%	12,40%
15 FNB	0,00%	0,00%
20 FNB	0,60%	-12,90%
25 FNB	2,3%	-24,8%
30 FNB	4,70%	-35,30%
35 FNB	7,40%	-44,4%
40 FNB	10,40%	-52,00%

Le lecteur remarquera que les franchises sont exprimées en termes de FFB. Le passage d'une franchise de 5 FFB à 3 FFB a donc un effet correspondant à une augmentation du nombre de sinistres de l'ordre de 3,6% et du coût moyen aux environ de 9,1%. L'impact estimé sur un événement tempête reviendrait donc à multiplier le coût global par un facteur de 1,13.
Le passage de 10 à 5 FFB quant à lui occasionnerait une augmentation du coût de l'événement de 40,4%. ((1/(1-0.101))*(1/(1-0.208))=1.404).

D'après nos sources, la franchise de notre compagnie se situait à 5 FFB en 1990, passant ensuite à 3 FFB jusqu'en 1998, pour enfin finir à 1,6 FFB (soit à peu près 900F) en 1999 et 2000.
Donc une simple extrapolation linéaire nous a permis, par exemple, d'estimer que l'impact du passage de 5 FFB à 1,6 nécessitait l'ajustement des montants par un coefficient de 1,22.
Néanmoins, nous n'avons pu prendre en compte l'effet franchise sur les sinistres automobiles, dans la mesure où le montant de la franchise est fonction du contrat de chaque assuré.

1.3.4 Prise en compte de l'élargissement du champ de la garantie

Avant les tempêtes de 1990, la garantie tempête prenait en compte pour le remboursement des sinistres un coefficient de vétusté. Le principe de ce coefficient de vétusté consiste à rembourser l'assuré en fonction de l' «ancienneté » du bien endommagé. Cependant, suite aux importantes tempêtes survenues dans le courant de l'année 1990 et au mécontentement somme toute légitime d'une bonne partie des assurés, les compagnies d'assurance ont donc décidé d'abandonner l'utilisation de ce coefficient de vétusté et par conséquent de passer au remboursement en valeur à neuf. D'autres part des éléments supplémentaires se sont retrouvés inclus dans le champ de cette garantie comme les gouttières, antennes et autres clôtures.
L'impact de toutes ces transformations a été estimé à 12% d'augmentation sur les risques des particuliers. Il convient donc de multiplier les montants de l'année 1990 relatifs à ce type de risques par ce coefficient.

Il est clair que certaines des modifications envisagées doivent être apportées plutôt au niveau de l'événement tempête (comme par exemple la prise en compte de l'effet franchise) alors que d'autres (comme l'inflation) ont un sens au niveau du sinistre pris de manière individuelle.
A titre indicatif, nous avons opéré en multipliant tous les sinistres individuels par les coefficients appropriés, et en agrégeant ensuite les montants des sinistres individuels revalorisés.

1.4 Reconstitution des tempêtes et statistiques descriptives sommaires

Nous avons vu dans les sections précédentes que le traité en excédent de sinistre comprenait une clause horaire, qui était de 72h pour une tempête et de 24h pour la grêle.

Néanmoins, nous ne disposons malheureusement ni de l'origine du phénomène (grêle, tempête), ni de l'heure à laquelle un sinistre s'est produit, ce qui est plus facilement compréhensible. Nous n'avons pas non plus accès aux données de Météo-France car celles-ci sont beaucoup trop coûteuses.
Certes, les tempêtes les plus «célèbres » sont facilement répertoriées, mais notre connaissance se limite tout au plus à une dizaine de phénomènes.

Nous avons donc décidé de regrouper les sinistres individuels, revalorisés par branche automobile ou incendie, par période de 3 jours consécutifs sauf pour les orages de grêle clairement identifiés ou nous nous sommes limités à un jour.
Nous avions dans un premier temps pensé à regrouper tous les sinistres survenus en été uniquement en fonction de leur date de survenance, soit sur une période de un jour. Mais les résultats obtenus avec cette méthode ne s'étant pas révélés probants, nous avons décidé de l'abandonner.

Le principal problème auquel nous nous sommes retrouvés confronté et qui nous a quand même surpris il faut bien l'admettre, c'est le nombre impressionnant d'événements «tempête » ainsi reconstitué.
En fait, nous avons dénombré plus de 1100 tempêtes, soit à peine moins que le nombre de périodes de 3 jours compris dans une période de 11 ans.
La garantie tempête est donc très souvent touchée. Même s'il est clair que nous ne pouvons considérer tous ces événements comme des tempêtes au sens météorologique du terme, cela n'est pas non plus complètement dénué de sens dans la mesure où le réassureur s'intéresse uniquement au montant global des sinistres sur une période de 3 jours.
Les principaux événements reconstitués en terme de montant total, et non pas par branches, sont les suivants :

Date début Evénement	Date fin Evénement	Auto en KF	MR en KF	Total en KF
26/12/99	26/12/99	18 736	1 041 264	1 060 000
27/12/99	29/12/99	9 315	580 306	589 621
03/02/90	05/02/90	1 451	290 782	292 233
25/01/90	27/01/90	963	188 939	189 902
27/02/90	01/03/90	1 056	91 366	92 422
12/12/90	14/12/90	231	48 483	48 714

Comme on pouvait facilement le présager, les tempêtes de fin 1999 et celles de 1990 constituent les événements les plus conséquents.

Afin de mieux observer la répartition des montants des événements tempêtes dans chaque branche, le calcul des quantiles constitue une solution facile à mettre en œuvre et relativement parlante :

Automobile	
Quantile	Montant en KF
100% Max	18 736
99%	958
95%	206
90%	87
75% Q3	30
50% Median	7
25% Q1	0
10%	0
5%	0
1%	0
0% Min	0

Multirisque	
Quantile	Montant en KF
100% Max	1 041 264
99%	20 497
95%	2 146
90%	1 102
75% Q3	270
50% Median	55
25% Q1	16
10%	5
5%	2
1%	0
0% Min	0

Plus d'un quart des tempêtes ont donc un coût nul dans la branche automobile, et moins de 5% dans la branche incendie.
Force est de constater que nous ne pourrons vraisemblablement pas retenir la totalité de ces événements pour la suite de notre étude, d'autant que les montants nuls gênent considérablement.
Nous allons donc être amenés à faire une sélection parmi ces événements, et à étudier les dépendances entre ces deux branches au travers de plusieurs niveaux de sélections. Nous reviendrons bien évidemment plus en détail sur ce point lors de notre troisième partie, au cours de laquelle nous tenterons d'établir une distribution bivariée.

Auparavant, il est impératif d'évoquer le concept de copulas et toute la théorie s'y rattachant.

Seconde Partie

Théorie des copulas

La distribution normale multivariée a souvent servi de modèle. Ce type de distribution est attrayant dans la mesure où les lois marginales sont elles aussi normales et que la dépendance est entièrement déterminée par le coefficient de corrélation.
Mais la loi normale n'est malheureusement pas toujours la plus appropriée à l'étude de certains phénomènes.
Pour pallier à ceci de nombreuses distributions bivariées ont été développées, nous pensons notamment à la Pareto bivariée ou bien à la la loi Gamma comme mentionnées en introduction.
Mais, encore une fois, ces lois n'ont pas que des avantages. D'une part, le "paramètre d'association" se retrouve souvent dans les lois marginales, d'autre part ces lois marginales sont de la même famille.

Les copulas vont nous permettre la construction de distribution multivariées qui n'ont pas tous ces défauts. Les copulas, qui ne sont ni plus ni moins que des fonctions, vont nous permettre d'établir un lien entre les lois marginales et la distribution multivariée.

L'objet de cette partie est de donner des bases théoriques sommaires mais néanmoins nécessaires à la bonne compréhension des copulas.
Nous effectuerons tout d'abord quelques rappels sur les distributions multivariées et sur une propriété essentielle de la loi uniforme.Puis nous présenterons le concept de copula et nous donnerons quelques exemples qui nous serviront par la suite. Nous aborderons enfin la notion de dépendance, et nous présenterons des mesures de dépendance qui peuvent être exprimées à partir des copulas. Nous finirons par donner quelques exemples de fonctions nous permettant d'affiner le choix d'une copula pour une certaine série données.

2.1 Quelques rappels concernant les distributions multivariées et la loi uniforme

Avant d'aborder les copulas à proprement parler, il nous a semblé nécessaire de devoir effectuer quelques rappels.
L'objet de cette partie n'est pas d'entrer dans des problèmes théoriques, c'est pourquoi les définitions présentées seront souvent brèves et donc réductrices.

2.1.1 Définition de la fonction de répartition conjointe

Considérons un vecteur aléatoire en n-dimensions $\mathbf{X}=(X_1,X_2,\ldots,X_n)$.
On appelle alors fonction de répartition conjointe de $X=(X_1,X_2,\ldots,X_n)$, la fonction de n variables réelles définie par :
$F(x_1,x_2,\ldots,x_n)=P(X_1\leq x_1,\ldots,X_n\leq x_n)$.

2.1.2 Définition des lois marginales

Cette fonction de répartition conjointe nous permet d'établir les fonctions de répartition $F_1,\ldots,F_n$ des variables aléatoires marginales $X_1,..X_n$ (donc les fonctions de répartition marginales) comme suit :
$F_i(x_i)=P(X_i\leq x_i)=F(\infty,..,\infty,x_i,..,\infty,..,\infty)$.
La dernière expression doit se comprendre comme la limite lorsque chaque composante x_j $(j\neq i)$ tend vers l'infini.

2.1.3 Lois de probabilité absolument continues

La fonction de répartition de **X** est dite absolument continue s'il existe une fonction f telle que

$$F(x_1,\ldots,x_n)=\int_{-\infty}^{x_1}\cdots\int_{-\infty}^{x_n} f(t_1,\ldots,t_n)\,dt_1\ldots dt_n$$

La fonction f est alors appelée densité de probabilité conjointe de la loi de **X.**

Si la loi de **X** est absolument continue, alors ses variables aléatoires marginales sont absolument continues et sa densité conjointe détermine les densités marginales $f_i(x_i)$ ainsi

$$f_i(x_i) = \int_{\infty}^{\infty} \dots \int_{\infty}^{\infty} f(x_1, \dots, x_i, \dots, x_n)\, dx_1 \dots dx_{i-1}\, dx_{i+1} \dots dx_n$$

2.1.4 Notion d'independance

Les composantes du vecteur X sont mutuellement indépendantes si pour tout $(x_1,x_2,\dots x_n)$ dans $\mathbf{R}^n$

$$F(x_1,\dots,x_n) = \prod_{i=1}^{n} F_i(x_i)$$

ou si X a pour densité

$$f(x_1,\dots,x_n) = \prod_{i=1}^{n} f_i(x_i)$$

2.1.5 Les bornes de Frechet-Hoeffding

Le théorème suivant nous permet d'établir les bornes d'une distribution multivariée en fonction des lois marginales de cette distribution et celui-ci s'énonce comme suit: Pour toute fonction de répartition multivariée $F(x_1,..x_n)$ ayant des distributions marginales $F_1,\dots,F_n$ alors

$$\max\left\{\sum_{i=1}^{n} F_i(x_i) + 1 - n, 0\right\} \le F(x_1,\dots,x_n) \le \min\{F_1(x_1),\dots,F_n(x_n)\}$$

2.1.6 Définition et propriétés de la loi uniforme

Une variable alétoire U à valeurs dans [0,1] est dite uniformément répartie sur [0,1] si elle est absolument continue et admet pour densité: $f(x) = 1_{[0,1]}(x)$

Il est très facile de montrer que: E[U]=1/2 et Var(U)=1/12.

Théorème

Soit X une variable aléatoire ayant pour fonction de répartition F. Notons F^{-1} la fonction-quantile de F, et $\alpha \in [0,1]$

$$F^{-1}(\alpha)=inf\{x / F(x)\geq\alpha\}$$

Alors

- *Pour toute variable aléatoire U uniformément répartie sur [0,1] ($U\sim U(0,1)$) alors $F^{-1}(U) \sim F$.*
- *Si F est continue alors la variable aléatoire F(X) est uniformément répartie sur [0,1], $F(X) \sim U(0,1)$.*

La notation $X\sim F$ signifie bien entendu que la variable aléatoire X a F pour fonction de répartition. Ce théorème est des plus importants.

2.2 Définition et propriétés des copulas

Nous allons, au cours de cette partie, enfin donner la définition d'une copula et énoncer le théorème de Sklar qui permet de faire le lien entre une distribution multivariée et ses fonctions de répartition marginales.

2.2.1 Définition

Une copula est définie comme étant une fonction de répartition multivariée ayant des lois marginales uniformes sur [0,1].

Donc $C(u_1,\ldots,u_n)=P(U_1\leq u_1,\ldots,U_n\leq u_n)$ est donc une fonction definie de $[0,1]^n$ vers [0,1] qui vérifie les trois propriétés suivantes:

1 $C(u_1,\ldots,u_n)$ est une fonction croissante dans chaque composante u_i

2 $C(1,..,1,u_j,1..,1)=u_j$ pour tout $j \in \{1,\ldots,n\}$

3 Quels que soient $(a_1,\ldots,a_n)$ et $(b_1,\ldots,b_n) \in [0,1]^n$ et $a_i\leq b_i$, alors

$$\sum_{i_1}^{2}\ldots\sum_{i_n}^{2}(-1)^{i_1+\ldots+i_n}C(u_{1i_1},\ldots,u_{ni_n})\geq 0$$

avec $u_{j1}=a_j$ et $u_{j2}=b_j$ pout tout $j \in \{1,\ldots,n\}$

La seconde propriété découle du fait que les lois marginales sont uniformes.
La troisième nous assure que si un vecteur **U** a pour fonction de distribution C, alors la probabilité $P(a_1 \le U_1 \le b_1,\ldots,a_n \le U_n \le b_n)$ ne peut pas être négative.

Dans le cas à deux dimensions, la propriété 3 se résume à
Pour tout $a_1 \le a_2$ et $b_1 \le b_2$:
$C(a_2,b_2)-C(a_1,b_2)-C(a_2,b_1)+C(a_1,b_1) \ge 0$

2.2.2 Le théorème de Sklar

Soit H une fonction de distribution en n dimensions ayant des marginales $F_1,F_2,\ldots,F_n$. Alors il existe une n-copula C telle que pour tout $\mathbf{x} \in [-\infty,\infty]^n$

$$H(x_1,x_2,\ldots,x_n) = C(F_1(x_1),\ldots,F_n(x_n))$$

Si les marginales sont continues alors C est unique; sinon C est uniquement déterminée en fonction de $RanF_1 * RanF_2 * \ldots RanF_n$. Inversement si C est une copula à n dimensions (n-copula) et que $F_1,\ldots,F_n$ sont des fonctions de distribution, alors la fonction H définie comme ci-dessus est une fonction de distribution à n-dimensions ayant pour marginales $F_1,F_2,\ldots,F_n$.
Nb:Par Ran nous entendons l'abbréviation du range de F qui est "l'ensemble d'arrivée de la fonction".

2.2.3 Corollaire du théorème de Sklar

Avant d'aborder ce corollaire qui utilise des *fonctions inverses généralisées,* il convient tout d'abord de donner une définition de celles-ci.
Si F est une fonction de distribution, alors une fonction *inverse généralisée* de F pourra être toute fonction $F^{(-1)}$ définie sur I=[0,1] telle que
1 si $t \in RanF$, alors $F^{(-1)}(t)$ a pour valeur $\mathbf{x} \in [-\infty,\infty]$ et x est tel que F(x)=t.
Donc , pour tout $t \in RanF$ $F(F^{(-1)}(t))=t$
2 si $t \notin RanF$ alors $F(-1)(t)=\inf\{x \mid F(x) \ge t\}=\sup\{x \mid F(x) \le t\}$

Si F est strictement croissante, alors elle a une seule fonction inverse généralisée, qui est la fonction inverse, que nous noterons bien evidemment F^{-1}. Ce sera le cas lors de notre étude.

Ce petit point étant eclairci, nous pouvons dès lors, énoncer le corollaire du théorème de Sklar.

Définissons H,C, $F_1,\ldots,F_n$ comme dans le théorème précedent et notons $F_1^{(-1)},\ldots,F_n^{(-1)}$ les fonctions inverses généralisées de $F_1,\ldots,F_n$.
Alors, pour tout vecteur $\mathbf{u} \in [0,1]^n$

$$C(u_1,...,u_n) = H(F_1^{(-1)}(u_1),...,F_n^{(-1)}(u_n))$$

2.2.4 Propriété d'invariance.

C est invariante par transformation croissante.
Si $T_1,\ldots,T_n$ sont des fonctions strictement croissantes, alors $(T_1(X_1),\ldots,T_n(X_n))$ a la même copula que $(X_1,\ldots,X_n)$.

Preuve (*McNeil[10]*): Notons F_i la fonction de répartition de X_i et $T_i^{(-1)}$ la fonction inverse généralisée de T_i.
La variable $T_i(X_i)$ a pour fonction de répartition $G_i = F_i o\, T_i^{(-1)}$ et donc

$$\begin{aligned} C(u_1,\ldots,u_n) &= P(X_1 \le F_1^{(-1)}(u_1),\ldots, X_n \le F_n^{(-1)}(u_n)) \\ &= P(T_1(X_1) \le T_1 o F_1^{(-1)}(u_1),\ldots, X_n \le T_n o F_n^{(-1)}(u_n)) \\ &= P(T_1(X_1) \le G_1^{(-1)}(u_1),\ldots,T_n(x_n) \le G_n^{(-1)}(u_n)). \end{aligned}$$

2.2.5 Survival Copula ou Flipped Copula

Une traduction francaise étant impossible ou tout du moins ridicule, nous avons préféré opter pour l'appelation anglophone.
Par soucis de clarté, nous allons présenter les résultats pour un modèle bivarié.

La notation S(x) est très souvent employée pour décrire la fonction de survie P(X>x).
La fonction de survie conjointe S(x,y)=P(X>x,Y>y) n'est pas 1-F(x,y) comme certains auraient pu le penser (en fait il s'agit de la probabilité P(X>x ou Y>y)) mais

$S(x,y)=1-F_x(x)-F_y(y)+F(x,y)$

Par analogie, pour une copula nous savons que $C(u,v)=P(U<u,V<v)$, la fonction de survie est $C_s(u,v)=P(U>u,V>v)=1-u-v+C(u,v)$.

Comme $C(F_x(x),F_y(y))=F(x,y)$, nous obtenons $C_s(F_x(x),F_y(y))=S(x,y)$.

C_s n'est pas une copula dans la mesure où elle n'est pas nulle au point (0,0), mais que par contre elle est égale à zéro en (1,1).

Nous définissons alors $C_f(u,v)=C_s(1-u,1-v)=1-(1-u)-(1-v)+C(1-u,1-v)$

$$=u+v-1+C(1-u,1-v)$$

Ainsi $C_f(S_x(x),S_y(y))= C_s(F_x(x),F_y(y))=S(x,y)$.

La fonction C_f est, quant à elle, une Copula. Nous l'appelerons donc "flipped" copula (on peut également la retrouver sous le nom de "survival copula" dans la littérature), et, en appliquant celle-ci aux fonctions de survie marginales, nous obtenons alors la distribution de survie conjointe.

Néanmoins, les "flipped copula" peuvent être appliquées aux fonctions de distributions marginales, et avoir ainsi des propriétés inverses à la copula originale. Un exemple sera exposé dans la suite de ce rapport.

2.2.6 Densité des Copulas

Nous savons que si elle existe, la densité f d'une fonction de distribution F est définie comme suit:

$$f(x_1,\dots x_n)=\frac{\partial\ F(x_1,\dots x_n)}{\partial x_1\dots\partial x_n}$$

L'expression de le densité de la copula que nous noterons indifferement c ou C_{12} (dans le cas ou n=2) s'exprime donc comme suit

$$c(u_1,\dots,u_n)=\frac{\partial\, C(u_1,\dots,u_n)}{\partial u_1\dots\partial u_n}$$

Si nous appelons f_i la densité de la i-ème marginale, alors la densité f de F s'exprime alors

$$f(x_1,\dots,x_n)=c(F_1(x_1),\dots,F_n(x_n))\times\prod_{i=1}^{n}f_i(x_i)$$

Cette formule prendra toute son importance lorsque nous effectuerons des estimations paramétriques par la méthode du maximum de vraisemblance.

2.2.7 Distribution conditionnelle et Copulas

Les distributions conditionnelles peuvent également s'exprimer à l'aide des copulas.
Notons $C_1(u,v)$ la dérivée de C(u,v) par rapport à u.

$$C_1(u,v) = \frac{\partial C(u,v)}{\partial u}$$

Si la distribution jointe de X et Y est $F(x,y)=C(F_x(x),F_y(y))$ alors la distribution conditionelle de $Y|X=x$ est donc

$F_{Y|X}(y)=C_1(F_x(x),F_Y(y))$.

Nous verrons ultérieurement que si C_1 est inversible, alors la simulation de probabilité jointe pourra être faite par ce biais.

Cette formule revêt donc également une importance capitale.

2.3 Exemples de Copulas

Les principales propriétés des Copulas ayant été définies, nous pouvons dès à présent donner quelques exemples de celles-ci.

L' échantillon proposé ne sera bien entendu pas exhaustif, mais nous tacherons de présenter les Copulas les plus célèbres et celles qui présentent un interêt dans l'étude des distributions de montant de sinistres

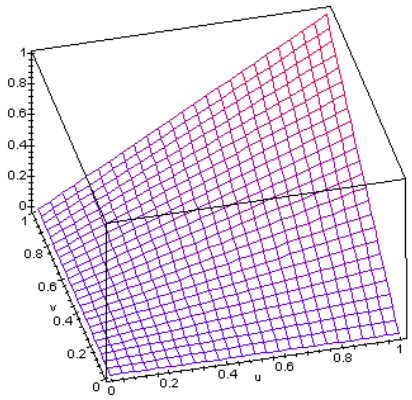
Nous commencerons par des exemples triviaux, puis nous aborderons les Copulas elliptiques et enfin nous nous attacherons à exposer les Copulas Archimédiennes.

2.3.1 La Copula d'indépendance

La copula d'indépendance a la forme suivante:

$$C^{ind}(u_1,\ldots,u_n) = \prod_{i=1}^{n} u_i$$

Il est évident que des variables aléatoires ayant cette copula sont indépendantes.



Représentation 3-D de la copula d'indépendance

2.3.2 La copula comonotone ou de dépendance totale positive

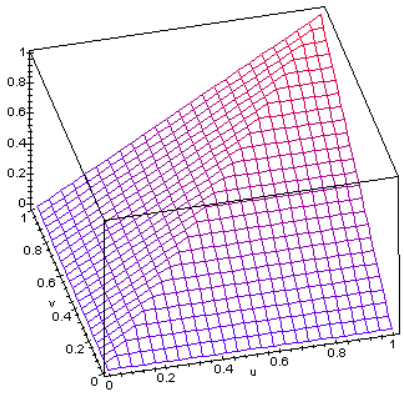
Nous avons vu, dans la partie "rappels concernant les distributions multivariées" en quoi consistaient les bornes de Frechet-Hoeffding.
Pour une copula, cette formule se réecrit de la manière suivante:

$$\max\left\{\sum_{i=1}^{n} u_i + 1 - n, 0\right\} \le C(u_1,\ldots,u_n) \le \min(u_1,\ldots,u_n)$$

La copula comonotone est donc définie comme suit:

$$C^u(u_1,\ldots u_n) = \min(u_1,\ldots,u_n)$$

Dans le cas n=2, cette copula se représente de la manière suivante:

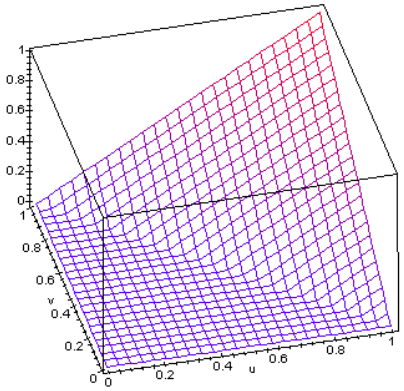


Représentation 3-d de la copula comonotone

2.3.3 La copula antimonotone ou de dépendance totale négative

A partir de la borne de Fréchet inférieure nous pouvons définir une autre copula. Attention, cette borne inférieure est une copula uniquement dans le cas ou n=2. Cette copula se définit donc ainsi:

$$C^1(u,v) = \max(u+v-1, 0)$$



Représentation 3-d de la Copula antimonotone

2.3.4 La Copula Normale

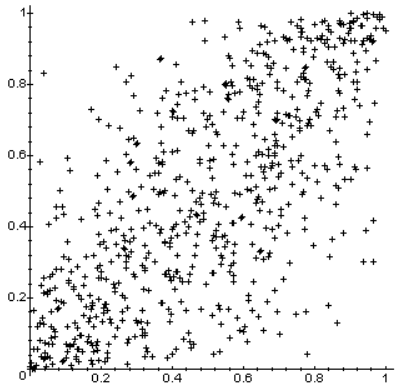
Cette copula fait partie de la famille des copulas elliptiques (de même que la Copula de Student que nous ne présenterons pas). La copula d'une distribution normale bivariée (avec ρ comme coefficient de correlation) est définie comme suit:

$$C_\rho^{Ga}(u,v) = \int_{\infty}^{\Phi^{-1}(u)} \int_{\infty}^{\Phi^{-1}(v)} \frac{1}{2\pi\sqrt{(1-\rho^2)}} \exp\left\{\frac{t_1^2 + t_2^2 - 2\rho t_1 t_2}{2(\rho^2-1)}\right\} dt_1 dt_2$$

avec, bien entendu, -1<ρ<1 et ϕ la fonction de distribution de la loi normale.
La densité s'écrit alors

$$c_\rho^{Ga}(u,v) = \frac{1}{\sqrt{1-\rho^2}} \exp\left(\frac{1}{2}\left(\frac{\rho^2\phi^{-1}(u)^2 + \rho^2\phi^{-1}(v)^2 - 2\rho\phi^{-1}(u)\phi^{-1}(v)}{1-\rho^2}\right)\right)$$

Il est très facile de simuler une paire de variable aléatoire (U,V) ayant pour distribution jointe la Copula Normale.
La première étape consiste à simuler une paire de variable aléatoire (X,Y) ayant pour distribution la loi normale bivariée N(0,R). R étant la matrice de corrélation ayant ρ comme élément non diagonal, et bien sûr 1 sur la diagonale.
En prenant ensuite $U=\Phi(X)$ et $V=\Phi(Y)$, alors (U,V) aura pour distribution la copula Normale.
Le coefficient de correlation de Kendall, que nous introduirons dans la partie suivant a pour expression
$\tau=(2/\pi)\arcsin(\rho)$.



Simulation de 800 paires ayant pour distribution la copula $C^{ga}_{0.6}$ et des lois marginales uniformes

Dire qu'un vecteur (X,Y) est normal revient donc à affirmer

- Que les lois marginales F_X et F_Y sont gaussiennes
- Que ces marginales sont reliées entre elles par la Copula décrite ci-dessus.

Néanmoins, l'hypothèse de normalité des lois marginales peut être levée et F_X , F_Y pourront être des lois ayant une queue plus épaisse comme par exemple une Pareto ou une Student, et toujours avoir une structure de dépendance normale.

2.3.5 Les Copulas Archimédiennes

Dans les applications liées à l'assurance, il existe très certainement une plus forte dépendance entre les gros sinistres qu'entre les petits. De telles asymétries ne peuvent être mises en valeur à l'aide des copulas elliptiques (les copulas elliptiques vérifient $C=C_F$).

Les Copulas Archimédiennes sont définies de manière beaucoup plus simple que les copulas elliptiques, en tout cas lorsque l'on étudie des phénomènes en deux dimensions. Nous nous limiterons à ce cas dans nos applications ultérieures.

2.3.5.1 Définition d'une Copula Archimédienne

2.3.5.1.1 Définition préliminaire

Soit φ une fonction continue, strictement décroissante définie de [0,1] vers [0,∞] telle que $\varphi(1)=0$. La fonction pseudo-inverse de φ , notée $\varphi^{[-1]}$, avec Dom $\varphi^{[-1]}=[0,\infty]$ et Ran $\varphi^{[-1]}=[0,1]$ est définie comme suit:

$$\varphi^{[-1]}(t)=\begin{cases}\varphi^{-1}(t) \text{ si } 0\le t\le\varphi(0)\\ \\ 0 \text{ si } \varphi(0)\le t\le\infty\end{cases}$$

Remarquez que $\varphi^{[-1]}$ est continue et non-croissante sur [0,∞], et strictement décroissante sur [0, $\varphi(0)$].

En outre $\varphi(\varphi^{[-1]}(u))=u$ sur [0,1] et

$$\varphi(\varphi^{[-1]}(t))=\begin{cases}t \text{ si } 0\le t\le\varphi(0)\\ \\ \varphi(0) \text{ si } \varphi(0)\le t\le\infty\end{cases}$$

Ainsi si $\varphi(0)=\infty$, alors $\varphi^{-1}=\varphi^{[-1]}$

2.3.5.1.2 Théorème

Soient φ une fonction continue, strictement décroissante de [0,1] vers [0,∞] telle que $\varphi(1)=0$, et $\varphi^{[-1]}$ la fonction pseudo-inverse de φ.

Soit C la fonction définie de $[0,1]^2$ vers [0,1] telle que

$C(u,v)=\varphi^{[-1]}(\varphi(u)+\varphi(v))$.

Cette fonction est une Copula si et seulement si φ est convexe.

Ce type de copula est appelé Copula Archimédienne, et la fonction φ est plus connue sous le nom de générateur de la copula.
De nombreuses copulas ont été inventés en suivant ce procédé et elles forment la famille des Copulas Archimédiennes qui possède des propriétés particulières et intéressantes.

Nous allons, dans la suite de cette partie citer quelques copulas parmis les plus célèbres de cette famille.

2.3.5.2 La copula de Frank

Prenons

$$\varphi(t) = -\ln\left(\frac{e^{-at}-1}{e^{-a}-1}\right)$$

avec a≠0.
Nous obtenons ainsi la copula de Frank qui est donc définie comme suit:

$$C_a(u,v) = -\frac{1}{a}\ln\left(1+\frac{\left(e^{-au}-1\right)\left(e^{-av}-1\right)}{e^{-a}-1}\right)$$

Les cas limites sont donc les suivants:

$$\lim_{a\to-\infty} C_a = C^l, \lim_{a\to 0} C_a = C^{ind}, \lim_{a\to\infty} C_a = C^u$$

En définissant $g_z=e^{-az}-1$ nous obtenons:

$$C_1(u,v) = \frac{\partial C(u,v)}{\partial u} = \frac{g_u g_v + g_v}{g_u g_v + g_1}$$

et finallement

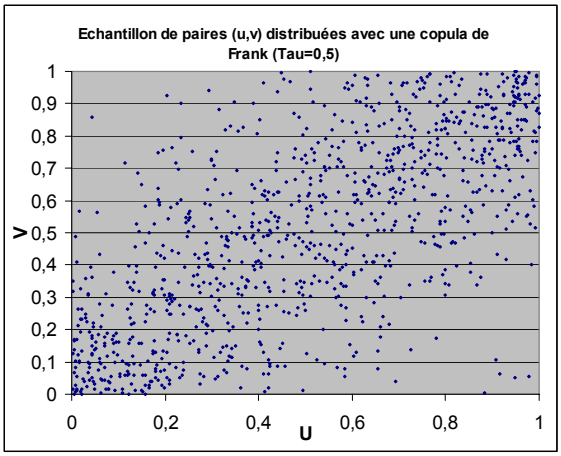
$$c(u,v) = -ag_1\left(\frac{1+g_{u+v}}{(g_u g_v + g_1)^2}\right)$$

Nous rappelons à nouveau que le calcul de la densité est très important dans la mesure où celle-ci va être utilisée ultérieurement pour des estimations de paramètres par la méthode du maximum de vraisemblance.

La copula de Frank étant la seule copula Archimédienne à respecter l'équation $C(u,v)=C_F(u,v)$, celle-ci risque de ne pas être tout à fait appropriée à des applications dans le domaine de l'assurance.

Pour simuler des paires (u,v) ayant comme distribution la Copula de Frank, nous pourrons utiliser la distribution conditionnelle.
La démarche à suivre consiste donc à simuler indépendemment u et p toutes deux issues de la loi uniforme sur [0,1] . Mais ici p sera compris comme une réalisation de la distribution conditionnelle de $V \mid u$. Comme cette distribution n'est n'est plus ni moins que C_1 , v sera déterminé en résolvant l' équation $v=C_1^{-1}(p \mid u)$.
C_1 étant inversible, v se détermine alors ainsi

$$v = -\frac{1}{a}\ln\left(1+\frac{pg_1}{1+g_u(1-p)}\right)$$



Une fois u et v simulés, nous pourrons simuler des réalisations des variables X et Y en inversant tout simplement les distributions marginales, c'est à dire en prenant $x=F_X^{-1}(u)$ et $y=F_Y^{-1}(v)$.
Le coefficient de corrélation de Kendall que nous définirons dans la partie suivante, se définit, pour la copula de Frank de la manière suivante:

$$\tau(a) = 1 - \frac{4}{a} + \frac{4}{a^2}\int_0^a \frac{t}{e^t - 1}dt$$

La dernière partie de l'expression est également connue sous le nom de fonction de Debye.

2.3.5.3 La Copula de Clayton

Prenons

$$\varphi(t) = a\left(t^{-1/a} - 1\right)$$

avec a>0.

Nous obtenons ainsi la copula de Clayton

$$C_a(u,v) = \left(u^{-1/a} + v^{-1/a} - 1\right)^{-a}$$

Les cas limites suivant la valeur du paramètre sont donc:

$$\lim_{a\to\infty} = C^{ind}, \lim_{a\to 0} = C^{u}$$

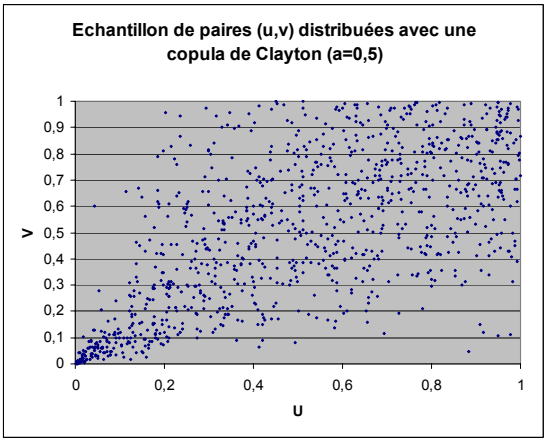
La dérivée première par rapport à u s'écrit:

$$C_1(u,v) = u^{-1-1/a}\left[u^{-1/a} + v^{-1/a} - 1\right]^{a-1}$$

et ainsi la densité s'exprime:

$$c(u,v) = (1+\frac{1}{a})\,(uv)^{-1-1/a}\,(u^{-1/a} + v^{-1/a} - 1)^{-a-2}$$

et $\tau(a)=1/(2a+1)$.



La technique de simulation utilisée est la même que pour la copula de Frank, sauf que dans ce cas là $v=[(pu^{1+1/a})^{1/(-a-1)}-u^{-1/a}+1]^{-a}$.

Il est visible sur cet échantillon qu'il y a une importante concentration de points près de (0,0). Cette copula aurait donc tendance a "correler" entre eux les petits sinistres et pas du tout les gros. Cela va à l'encontre de ce que nous recherchons lorsque nous étudions des distributions de montants de sinistres.

L'idée serait donc de définir une copula présentant des propriétés opposées.
Nous avons défini dans les pages précédentes la notion de "survival Copula".
Nous allons donc appliquer cette formule pour obtenir la copula HRT "heavy right tail".

2.3.5.4 La copula HRT

Cette copula est donc définie comme étant la "survival copula" de la copula de Clayton.
Elle ne fait pas partie de la famille des copulas archimédienne, mais comme sa formule est dérivée de la copula de clayton, elle a sa place ici.
Elle s'écrit donc:

$$C_a(u,v) = u + v - 1 + \left((1-u)^{-1/a} + (1-v)^{-1/a} - 1\right)^{-a}$$

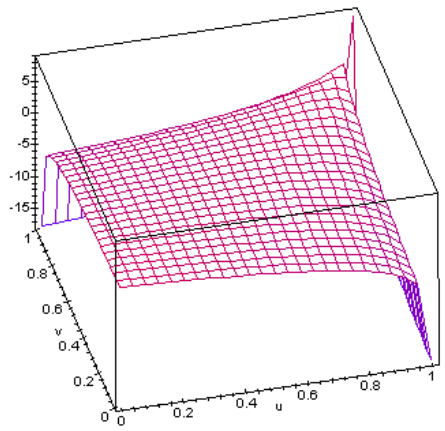
et par suite

$$C_1(u,v) = 1 - \left[(1-u)^{-1/a} + (1-v)^{-1/a} - 1\right]^{-a-1}(1-u)^{-1-1/a}$$

la densité étant donc

$$c(u,v) = (1+\frac{1}{a})\left[(1-u)^{-1/a} + (1-v)^{-1/a} - 1\right]^{-a-2}\left[(1-u)(1-v)\right]^{-1-1/a}$$

et $\tau(a)=1/(2a+1)$, comme pour la copula de Clayton.

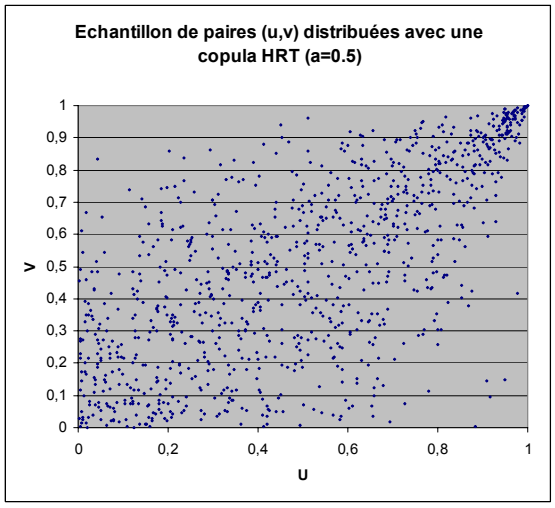


La densité de la copula HRT sur une échelle logarithmique(a=0.5)

Comme pour les autres copulas présentées jusqu'à présent, la distribution conditionnelle C_1 est inversible.

Dans ce cas

$v=1-[(1-(1-u)^{-1/a}+[(1-p)(1-u)^{1+1/a}]^{(-1/(1+a))}]^{-a}$



La copula HRT ayant été concue dans cette optique, la concentration de points près de (1,1) est la plus importante. Avec cette structure de dépendance, les gros sinistres auront tendance à survenir ensembles.

2.3.5.5 La copula de Gumbel

Prenons comme générateur

$$\varphi(t) = (-\ln(t))^a$$

avec a>1. Cette fonction satisfaisant à toutes les conditions du théorème sur les copulas Archimédiennes, nous pouvons générer la copula de Gumbel en prenant:

$$C_a(u,v) = \exp\left(-\left[(-\ln(u))^a + (-\ln(v))^a\right]^{1/a}\right)$$

Les cas limites suivant la valeur du paramètre sont donc:

$$\lim_{a\to\infty} C_a = C^u,\ C_{a=1} = C^{ind}$$

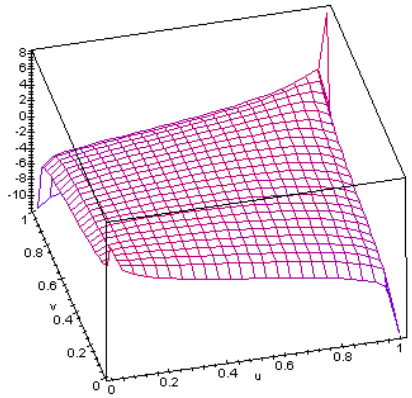
La dérivée par rapport à la composante u s'écrit:

$C_1(u,v)=C(u,v)[(-\ln u)^a +(-\ln v)^a]^{-1+1/a} (-\ln u)^{a-1}/u$

Et la densité

$c(u,v)=C(u,v)u^{-1}v^{-1}[(-\ln u)^a+(-\ln v)^a]^{-2+2/a}[(\ln u)(\ln v)]^{a-1}\{1+(a-1)[(-\ln u)^a +(-\ln v)^a]^{-1/a}\}$

et $\tau(a)=1-1/a$



La densité de la copula de Gumbel (a=2) sur une échelle logarithmique

Par contre, la distribution conditionnelle C_1 n'est pas inversible. Il existe néanmoins une technique de simulation propre aux copulas archimédiennes.

Définissons tout d'abord pour tout $t \in [0,1]$ la fonction:

$$K(t) = t - \frac{\varphi(t)}{\varphi'(t^+)}$$

Supposons qu'un vecteur aléatoire (U,V) ait pour fonction de distribution une Copula Archimédienne, alors cette fonction K sera tout simplement la fonction de distribution de la variable aléatoire C(U,V).

Dans le cas de la Copula de Gumbel, cette fonction sera donc:

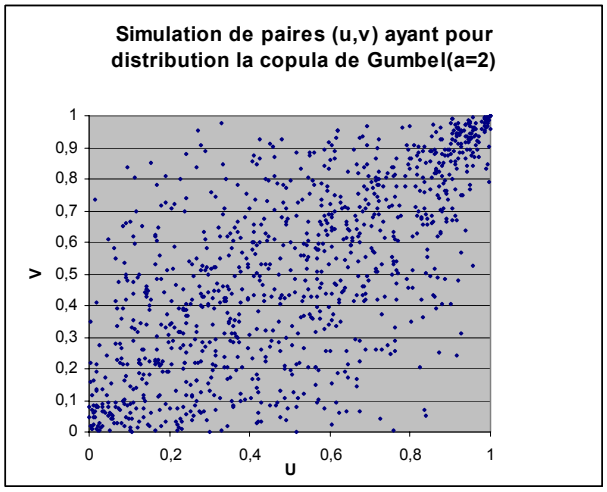
$$K(t) = t(1 - \frac{1}{a}\ln(t))$$

La procédure à suivre pour simuler des paires (u,v) est donc

- Simuler indépendamment s et q suivant la loi uniforme U(0,1)
- Déterminer $t=K_c^{-1}(q)$

- Prendre $u=\varphi^{[-1]}(s\ \varphi(t))$ et $v=\varphi^{[-1]}((1-s)\ \varphi(t))$

La détermination de t ne peut se faire que par résolution numérique de l'équation, et donc par une méthode itérative.
Pour ce faire, nous avons utilisé la méthode de Newton,



Comme le montrent la simulation effectuée ci-dessus et le graphique représentant la densité de cette copula, celle-ci présente une densité de probabilité plus élevée aux abords des points (0,0) et surtout (1,1).
Cette copula pourrait donc eventuellement convenir dans le cadre d' études de distributions de montants de sinistres.
La présentation des principales copulas nous a permis d'examiner la forme des différentes structures de dépendance possibles, et les méthodes pour simuler de telles distributions.
L' échantillon proposé est très loin d'être exhaustif, néanmoins il a le mérite de présenter les copulas les plus souvent utilisées, et surtout les densités de celles-ci.
Nous allons dès à présent nous pencher de manière plus approfondie sur les différentes mesures de dépendance possibles.

2.4 Dépendance

Lorsque l'on évoque ce concept, le premier mot qui vient généralement à l'esprit est celui de coefficient de corrélation linéaire.

Ce coefficient de corrélation linéaire, également connu sous le nom de coefficient de correlation de Pearson, est en effet le plus couramment utilisé.

Ce coefficient est tout à fait approprié lorsque nous étudions des distributions normales ou de student multivariées, mais celui-ci perd de son interet si le modèle est différent.

Malheureusement, les distributions dans le domaine de l'assurance suivent très rarement de telles lois.

Nous présenterons deux autres coefficient de corrélation, à savoir le coefficient de Kendall et celui de Spearman, qui présentent l'avantage de demeurer inchangés sous l'hypothèse d'une transformation strictement croissante des variables aléatoires.

Une mesure de dépendance doit en fait nous permettre de nous faire une idée de la structure de dépendance entre deux variables aléatoires, et ceci exprimé à l'aide d'un seul nombre.

Nous examinerons également les notions de "tail dependence" ou de dépendance de queue, très importantes dans l'étude des dépendances entre les valeurs extrêmes et directement associés aux copulas.

2.4.1 Le coefficient de corrélation linéaire.

Définition

Soient X et Y deux variables aléatoires ayant une variance finie.

Alors, le coefficient de corrélation linéaire des variables X et Y est donné par:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}$$

Où Cov(X,Y)=E[XY]-E[X]E[Y] est la covariance entre X et Y, et Var(X) Var(Y) correspondent aux variances respectives des variables X et Y.

Dans le cadre d'une dépendance linéaire parfaite Y=aX+b ($a\neq 0$, $b\in R$), alors le coefficient de corrélation est égal à 1 ou –1 selon le signe de a.

D'autre part ce coefficient de corrélation reste inchangé par des transformations linéaires croissantes des variables aléatoires.

En effet $\rho(aX+b,cY+d)=sign(ab)\ \rho(X,Y)$.
Néanmoins, ce coefficient ne demeure pas constant sous l'hypothèse d'une transformation croissante non-linéaire.
Il est également important de signaler que l'indépendance de deux variables conduit à un coefficient de correlation linéaire nul, mais que la réciproque n'est pas exacte.
Il suffit pour s'en convaincre de prendre X distribuée selon la loi normale centrée réduite et $Y=X^2$. Il est clair que ces variables sont dépendantes, et pourtant leur covariance est nulle, et par conséquent le coefficient de corrélation également.

2.4.2 Notion de dépendance parfaite

Définition
Soit un couple de variables aléatoires (X,Y) ayant pour copula C^l ou C^u.(cf 2.3.2 ou 2.3.3)
En d'autres termes $F(x,y)=\max\{F_x(x)+F_Y(y)-1;0\}$ ou $F(x,y)=\min\{F_x(x),F_Y(y)\}$ respectivement.
Alors, il existe deux fonctions monotones α, β définies de R vers R et une variable aléatoire Z telle que

$$(X,Y)=_d(\alpha(Z),\ \beta(Z))$$

avec α croissante et β décroissante dans le premier cas et les deux fonctions croissantes dans le second. La réciproque est également vraie.

Si (X,Y) a pour copula C^u alors X et Y sont définies comme comonotones, et comme étant antimonotones si elles ont C_l pour copula.
Ceci explique le nom donné aux copulas aux sections (2.3.2) et (2.3.3)
Dans le cas de distributions continues, et en notant F_x et F_y les distributions respectives de X et Y, alors
$C=C^l \Leftrightarrow Y=T(X)$ p.s, $T=F_Y^{-1}o(1-F_x)$ décroissante.
$C=C^u \Leftrightarrow Y=T(X)$ p.s, $T=F_Y^{-1}o(F_x)$ croissante.

2.4.3 Les coefficients de corrélation de Kendall et de Spearman

La définition de ces deux coefficients est intimement liée à la notion de concordance. Ils constituent une alternative au coefficient de correlation linéaire, qui n'est pas comme nous l'avons montré auparavant la mesure de dépendance la plus appropriée et souffre de certaines lacunes.

Définition de la concordance

Notons (x,y) et (x',y') deux observations d'un vecteur aléatoire continu (X,Y).
Alors (x,y) et (x',y') sont dites concordantes si (x-x')(y-y')>0 et discordantes dans le cas contraire.

2.4.3.1 Le τ de Kendall

Soit (X,Y) un vecteur aléatoire et notons (X',Y') un vecteur en tout point identique.
Le τ de Kendall est tout simplement la probabilité de concordance moins celle de discordance, à savoir

$\tau(X,Y)=P((X-X')(Y-Y')>0)-P((X-X')(Y-Y')<0)$

ainsi

- $-1\leq\tau\leq1$
- si X et Y sont comonotones alors $\tau=1$
- si X et Y sont antimonotones alors $\tau=-1$
- si X et Y sont independantes alors $\tau=0$
- si α et β sont des fonctions strictement croissantes, alors $\tau(\alpha(X), \beta(Y))= \tau(X,Y)$
- τ ne dépend que de la copula de (X,Y)

Malheureusement, lorsque $\tau=0$ les variables X et Y ne sont pas forcement independantes.
Si (X,Y) a pour copula C, alors

$$\tau(X,Y) = 4\int_0^1\int_0^1 C(u,v)c(u,v)dudv - 1 = 4E[C(U,V)] - 1$$

Si X et Y ont pour copula une copula Archimédienne C, alors

$$\tau(X,Y) = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}$$

A titre d'exemple, pour la copula de Clayton, le rapport $\varphi(t)/\varphi'(t)$ s'écrit

$$\frac{\varphi(t)}{\varphi'(t)} = at(t^{1/a} - 1)$$

et on retrouve donc

$$\tau = 1 + 4a\int_0^1 t^{1+1/a} - t\, dt = 1 + 4a\left(\frac{a}{1+2a} - \frac{1}{2}\right) = \frac{1}{2a+1}$$

Nous avons vu que le tau de Kendall était la probabilité de concordance moins la probabilité de discordance.

Donc pour calculer ce coefficient de manière empirique à partir d'un échantillon $\{(x_1,y_1)\ldots(x_n,y_n)\}$ d'observations d'un vecteur aléatoire (X,Y), il est nécessaire d'établir toutes les combinaisons possibles de paires entre elles. Des notions de dénombrement nous permettent d' établir leur nombre à C_n^2 , soit (n*(n-1))/2.

Il suffit ensuite de compter le nombre de paires concordantes, d' y retrancher le nombre de paires discordantes, puis de diviser le tout par le nombre total de paires possibles, soit C_n^2.

En comptant 1 pour une paire concordante et –1 pour une paire discordante, alors la formule pour τ est

$$\tau = \binom{n}{2}^{-1} \sum_{i<j} \text{sign}(x_i - x_j)(y_i - y_j)$$

Comme on peut le constater en visualisant cette formule, le coefficient de Kendall ne dépend en fait que du rang de chaque observation. C'est pourquoi celui ci demeure inchangé lors de transformation croissante, que celle-ci soit linéaire ou pas.

2.4.3.2 Le ρ de Spearman

Pour donner une définition du coefficient de correlation de Spearman , il est non seulement nécessaire de conserver les paires (X,Y) et (X',Y') définies dans la partie précédente, mais en plus d'y rajouter une autre paire identique (X*,Y*).

Ce coefficient se définit:

ρ_s =3 (P{(X-X')(Y-Y*)>0} – P{(X-X')(Y-Y*)>0}).

Les propriétés du ρ de Spearman sont les mêmes que celles du τ de Kendall, à savoir

- $1 \le \rho_s \le 1$

- si X et Y sont comonotones alors $\rho_s = 1$
- si X et Y sont antimonotones alors $\rho_s = -1$
- si X et Y sont independantes alors $\rho_s = 0$
- si α et β sont des fonctions strictement croissantes, alors

  $\rho_s(\alpha(X), \beta(Y)) = \rho_s(X,Y)$
- ρ ne dépend que de la copula de (X,Y)

Si X et Y ont des distributions marginales F_x et F_y continues et une copula C unique (ie Fx(X),Fy(Y)~C) , alors le coefficient de Spearman est donné par:

$$\rho_s(X,Y) = 12\iint_{I^2} C(u,v) - uv \, dudv = 12\iint_{I^2} C(u,v)dudv - 3$$

ce qui peut se réécrire

$$\rho_s(X,Y) = 12E(UV) - 3 = \frac{E(UV) - 1/4}{1/12} = \frac{E(UV) - E(U)E(V)}{\sqrt{Var(U)Var(V)}}$$

et l'on peut en déduire que

$\rho_s(X,Y) = \rho(F_x(X),F_y(Y))$

Effectuons le calcul dans le cas de la copula

$C^u = \min(u,v)$

$$\int_0^1\int_0^1 \min(u,v)dudv = \int_0^1\int_0^u v \, dvdu + \int_0^1\int_0^v u \, dvdu = \int_0^1 \frac{u^2}{2}du + \int_0^1 \frac{v^2}{2}dv = \frac{1}{6} + \frac{1}{6} = \frac{4}{12}$$

ainsi

$\rho_s = 12*(4/12) - 3 = 1$

ce qui était bien entendu le résultat attendu.

2.4.4 La notion de "tail dependence" ou "dépendance de queue"

Cette notion est très importante dans l'étude de la dépendance asymptotique entre deux variables aléatoires. Cela va nous permettre de voir le niveau de dépendance dans les valeurs extremes (upper tail dependence) et dans les valeurs petites (lower tail dependence).

Ce concept sera totalement basé sur celui des copulas.

2.4.4.1 Upper tail dependence.

L'objet est donc l'étude de la dépendance dans la queue commune de la distribution bivariée.

Prenons donc deux variables aléatoires continues X et Y ayant pour fonction de distribution respectives F_x et F_y.

Le coefficient d' "upper tail dependence" de X et Y est défini

$$\lambda_u = \lim_{u \to 1^-} P\left(Y > F_Y^{-1}(u) \middle| X > F_X^{-1}(u)\right) = \lim_{u \to 1^-} \frac{C(u,u) - 2u + 1}{1-u}$$

si toutefois cette limite $\lambda_u \in [0,1]$ existe.

La quantité λ est une fonction de la copula et est donc invariante par transformation croissante.

Si $\lambda \in (0,1]$, il existe alors une dépendance asymptotique.

Si $\lambda=0$, on dit qu'il y a indépendance asymptotique.

La copula la plus frequemment donnée en exemple dans la litterature pour illustrer ce concept est celui de la copula de Gumbel

Exemple pour la copula de Gumbel

Nous savons que

$$C_a^{Gu}(u,u) = u^{2^{1/a}}$$

donc

$$\lambda_u = \lim_{u \to 1} \frac{1 - 2u + u^{2^{1/a}}}{1-u} = 2 - \lim_{u \to 1} \frac{u^{2^{1/a}} - 1}{1-u}$$

et nous appliquons ensuite la règle de l'Hospital pour obtenir

$$\lambda_u = 2 - \lim_{u \to 1} \frac{\partial\left(u^{2^{1/a}}\right)}{\partial u} = 2 - 2^{1/a}$$

La dépendance augmente vers 1 au fur et à mesure que a tend vers l'infini.

A l'inverse, la copula Normale par exemple ne présente aucune dépendance asymptotique.

2.4.6.2 Lower tail dependence.

En gardant, les mêmes notations qu'auparavant, le coefficient de "lower tail dependence" de X et Y est défini comme étant

$$\lambda_l = \lim_{u\to 0+} P\left(Y \le F_Y^{-1}(u) \middle| X \le F_X^{-1}(u)\right) = \lim_{u\to 0+} \frac{C(u,u)}{u}$$

si cette limite λ_l existe.

Exemple pour la copula de Clayton

Nous savons que

$$C_a^{cl}(u,u) = \left(2u^{-1/a} - 1\right)^{-a}$$

donc

$$\lambda_l = \lim_{u\to 0} \frac{1}{u\left(2u^{-1/a} - 1\right)^a} = \lim_{u\to 0} \frac{1}{\left(2 - u^{1/a}\right)^a} = \frac{1}{2^a}$$

La copula de Clayton possède donc un coefficient de "lower tail dependence".

Il est facile de démontrer que le coefficient de "lower tail dependence" de la copula de Clayton est le même que le coefficient d "upper tail dependence" de la copula HRT.

Nous rappelons à toutes fins utiles que ces deux copulas sont liées par la relation

$$C_a^{hrt}(u,v) = u + v - 1 + C_a^{cl}(1-u, 1-v)$$

donc, en posant z=(1-u).

$$\lim_{u\to 1-} \frac{1 - 2u + C_a^{hrt}(u,u)}{1-u} = \lim_{u\to 1-} \frac{C_a^{cl}(1-u,1-u)}{1-u} = \lim_{z\to 0+} \frac{C_a^{cl}(z,z)}{z}$$

En fait, cette assertion se généralise à toutes les copulas et leurs versions "flipped".

2.5 Approche empirique des Copulas

2.5.1 Quelques rappels sur les distributions empiriques

La définition d'une fonction de répartition empirique est la suivante

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}} = \frac{\text{card}\{x_i \le x\}}{n}$$

c'est à dire le nombre d'observations d'un echantillon comprenant n élement notés $(x_1,\ldots,x_n)$ inférieures à x divisé par n.
Dans le cas d'une distribution bivariée, en supposant l'échantillon noté $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, celle ci s'écrit:

$$F_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \le x\}} 1_{\{y_{i \le y}\}}$$

2.5.2 Expression empirique des Copulas et des mesures de dépendance associées

Notons $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ un echantillon d'observations d'un vecteur aléatoire à deux dimensions noté (X,Y). Nous notons $x_{(i)}$ et $y_{(j)}$ respectivement les i-^{eme} et j-^{eme} élements des statistiques ordonnées.

Alors, la copula empirique associée à l'échantillon est définie comme

$$C_n\left(\frac{i}{n},\frac{j}{n}\right) = \frac{\text{nb de paires de l'échantillon vérifiant } x \le x_{(i)} \text{ et } y \le y_{(j)}}{n}$$

Deux exemples triviaux

Supposons que nous disposions des 6 observations suivantes:

{xi,yi}	Rang x	Rang y
{5.2,15.23}	1	1
{6.3,18}	2	2
{9,19.2}	3	3
{12,28.4}	4	4
{14,30.3}	5	5
{17.2,35}	6	6

Alors, dans ce cas, la copula empirique est

(i/n,j/n)	**1/6**	**2/6**	**3/6**	**4/6**	**5/6**	**1**
1/6	1/6	1/6	1/6	1/6	1/6	1/6
2/6	1/6	2/6	2/6	2/6	2/6	2/6
3/6	1/6	2/6	3/6	3/6	3/6	3/6
4/6	1/6	2/6	3/6	4/6	4/6	4/6
5/6	1/6	2/6	3/6	4/6	5/6	5/6
1	1/6	2/6	3/6	4/6	5/6	1

On retrouve la copula de dépendance totale positive puisque, C(i/n,j/n)=min(i/n,j/n)
Supposons maintenant que l'échantillon soit

{xi,yi}	Rang x	Rang y
{5.2,35}	1	6
{6.3,30.3}	2	5
{9,28.4}	3	4
{12,19.2}	4	3
{14,18}	5	2
{17.2,15.23}	6	1

Alors, dans ce cas la copula empirique est

(i/n,j/n)	1/6	2/6	3/6	4/6	5/6	1
1/6	0	0	0	0	0	1/6
2/6	0	0	0	0	1/6	2/6
3/6	0	0	0	1/6	2/6	3/6
4/6	0	0	1/6	2/6	3/6	4/6
5/6	0	1/6	2/6	3/6	4/6	5/6
1	1/6	2/6	3/6	4/6	5/6	1

On retrouve dans ce cas là, la copula de dépendance totale négative puisque C(i/n,j/n)=max(i/n+j/n-1,0)

Les formulations non-paramétriques des coefficients de Kendall et Spearmann peuvent être également définies à partir des copulas empiriques et s'établissent comme suit:

$$\tau = \frac{2n}{n-1}\sum_{i=2}^{n}\sum_{j=2}^{n} C_n\left(\frac{i}{n},\frac{j}{n}\right)C_n\left(\frac{i-1}{n},\frac{j-1}{n}\right) - C_n\left(\frac{i}{n},\frac{j-1}{n}\right)C_n\left(\frac{i-1}{n},\frac{j}{n}\right)$$

$$\rho = \left(\frac{12}{n^2-1}\right)\left(\sum_{i=1}^{n} C_n\left(\frac{i}{n},\frac{j}{n}\right) - \frac{ij}{n^2}\right)$$

Dans le cadre des deux exemples présentés ci-dessus, les coefficients non-paramétriques de Kendall et Spearmann sont tous les deux respectivement égaux à 1 et –1.

2.6 Le choix de la bonne copula.

Au vu de toute la théorie exposée précédemment, nous sommes en droit de nous demander quelle copula pourrait le mieux correspondre à une certaine série de données.

Deux copulas ayant le même coefficient de correlation de Kendall peuvent en effet avoir des comportements tout à fait différents. Il suffit pour s'en convaincre de jeter un bref coup d'oeuil aux graphiques des simulations présentés dans la partie précédente.

Nous allons donc dans cette partie donner quelques méthodes nous permettant de faire le tri parmi les copulas afin de ne choisir que celles qui pourraient présenter des caractéristiques semblables à celles de notre série de données.

En fait, nous allons exposer plusieurs fonctions, dont certaines ont été exposée par Mr Gary Venter (Guy Carpenter New-York) dans son article 'Tails of Copulas' et qui auront des caractéristiques tout à fait différentes selon la copula choisie.

Ces fonctions pouvant également être établies de façon empirique, uniquement à partir du rang de chaque observation , une simple comparaison graphique nous permettra de ne retenir qu'une ou deux familles de copulas pour la poursuite de notre étude.

2.6.1 La fonction K(z)

Cette fonction, qui a déjà été évoquée dans notre présentation de la copula de Gumbel, n'est ni plus ni moins que la fonction de répartition de la variable aléatoire C(U,V).

Il a été démontré que pour une copula de type Archimédienne, cette fonction se définissait comme suit :

$$K(z) = z - \frac{\phi(z)}{\phi'(z)}$$

Dans le cadre des copulas Archimédiennes présentées auparavant cette fonction K(z) est donc la suivante

Gumbel

$$K_a(z) = z(1 - \frac{1}{a}\ln(z))$$

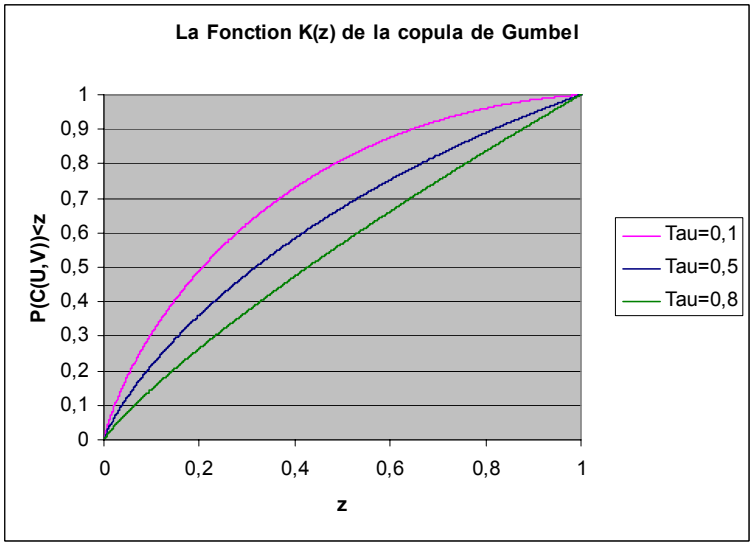
Frank

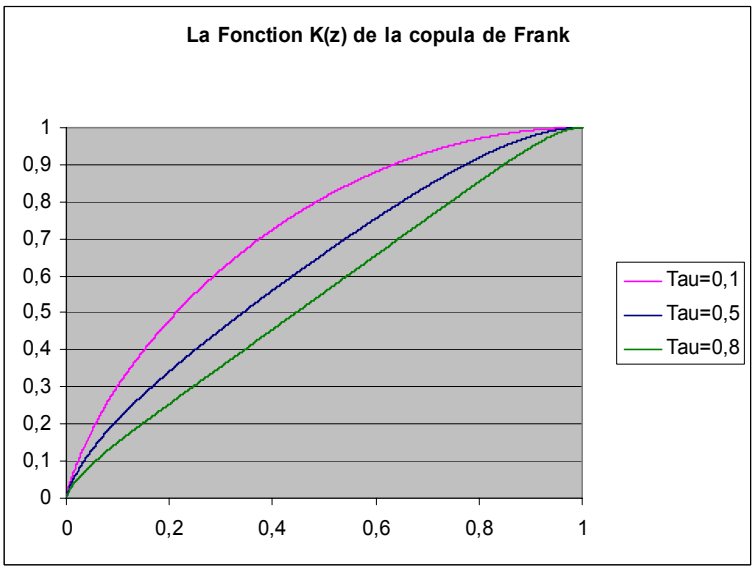
$$K_a(z) = z + \frac{1}{a}\ln\left[\frac{1 - e^{-az}}{1 - e^{-a}}\right]$$

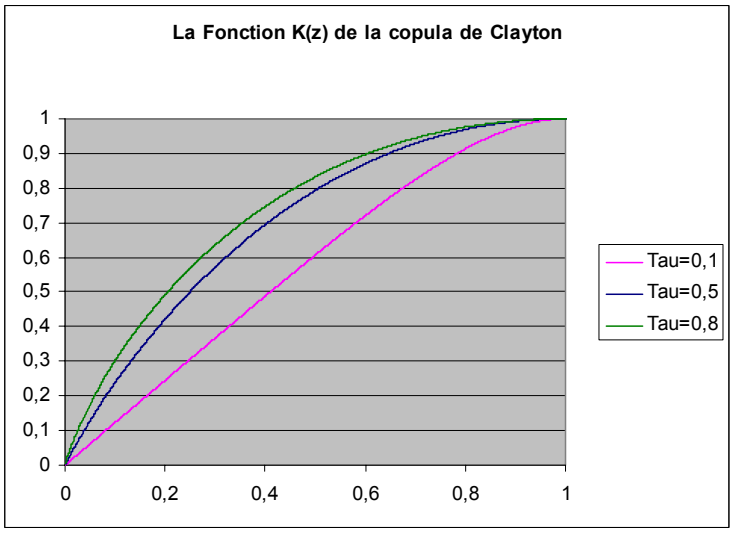
Clayton

$$K_a(z) = z + az\left(1 - z^{1/a}\right)$$

Les graphiques de ces différentes fonctions et pour différentes valeurs du Tau de Kendall sont présentées ci-après.







Supposons maintenant que nous disposions d'un échantillon d'observations $(x_1,y_1)\ldots(x_n,y_n)$ issu d'un vecteur aléatoire (X,Y).

Pour établir un estimateur non-paramétrique de la fonction K à partir de cet échantillon, la procédure à suivre consiste à :

1 Définir la pseudo-observation z_i pour chaque i=1…n

$$z_i = \frac{\text{nombre de paires } \{x_j, y_j\} \text{ telles que } x_j < x_i \text{ et } y_j < y_i}{n-1}$$

2 Définir l'estimateur non-paramétrique de K comme suit

$$K_n(z) = \frac{\{\text{nombre de } z_i \le z\}}{n}$$

Cette estimation non-paramétrique de K pourra ensuite être comparée graphiquement aux versions paramétriques de K pour les différentes copulas archimédiennes.

Le paramètre « *a* » de la copula pourra être établi, par exemple à partir du coefficient de Kendall empirique de l'échantillon.

En effet nous avons vu qu'il existe une *relation directe* entre le coefficient de Kendall et le paramètre de la copula. Il suffit donc simplement de résoudre une équation pour déterminer le paramètre de la copula, même si cette opération peut se révéler un peu plus délicate dans le cas de la copula de Frank par exemple.

Une autre méthode pour déterminer ce paramètre pourra être celle du maximum de vraisemblance que nous exposerons par après.

Nous pourrons ensuite être en mesure d'effectuer une comparaison graphique entre K_a et l'estimateur non paramétrique de K_n calculé à partir de l'échantillon.

En règle générale cette simple comparaison graphique mettra assez nettement en évidence la copula archimédienne à utiliser.

Dans le cas ou le graphique ne soit pas assez expressif nous pouvons également utiliser comme critère la minimisation de la distance $\int(K_a(z)-K_n(z))^2\, dK_n(z)$, ou bien tirer profit d'une représentation graphique sous forme de q-qplots.

2.6.2 La fonction J(z) ou de Tau cumulatif

Cette fonction, inventée par Gary Venter, est basée sur le coefficient du τ de Kendall dont nous rappelons ici la formule :

$$\tau = -1 + 4\int_0^1 \int_0^1 C(u,v)c(u,v)dudv$$

L'idée de Gary Venter a ensuite été de définir une fonction de τ cumulatif en prenant :

$$J(z) = -1 + 4\frac{\int_0^z \int_0^z C(u,v)c(u,v)dudv}{C(z,z)^2}$$

Il est évident que $J(1)= \tau$.

Bien que pour certaines copulas des formules aient été établies par les actuaires de Guy Carpenter à New-York (dont les formules seront données en annexe), la meilleure solution pour établir les graphiques de ces fonctions reste encore l'intégration numérique. En effet, au vu de complexité de la fonction à intégrer, cela semble la solution la plus indiquée.

Des macros en VBA ont donc été crées dans le but de calculer cette fonction J(z). Le carré $[0,1]^2$ a été divisé en un nombre N^2 très grand (N est bien entendu paramétrable) de très petits carrés de taille [1/N,1/N], dans le but de calculer la double intégrale au numérateur.

Nous pouvons noter $I_{i/N,j/N}$ par exemple le carré [i-1/N,i/N][j-1/N,j/N] pout tout i=1…N et j=1…N.

Nous pouvons appeler $X_{i/N,j/N}$ la valeur de fonction C(u,v)c(u,v) prise au **centre** du carré $I_{i/N,j/N.}$

La valeur Int_z de la double intégrale a ensuite été déterminée en chaque point z =1/N,2/N,…,1 en prenant

$$Int_z = \frac{1}{N^2}\sum_{i=1}^{zN}\sum_{j=1}^{zN} X_{\frac{i}{N},\frac{j}{N}}$$

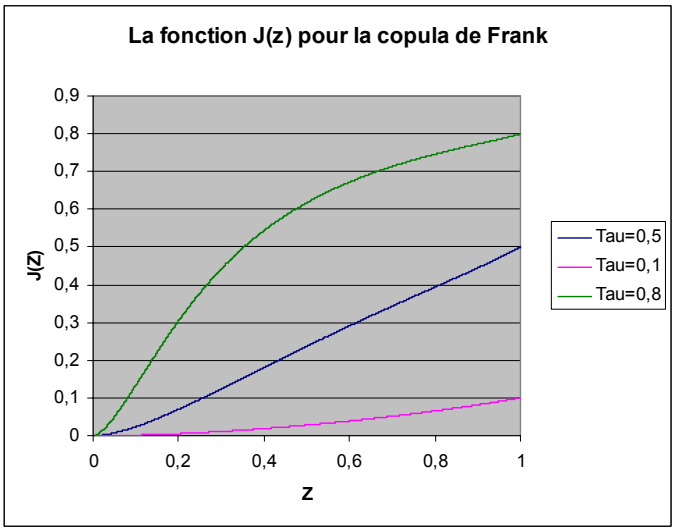
Et donc

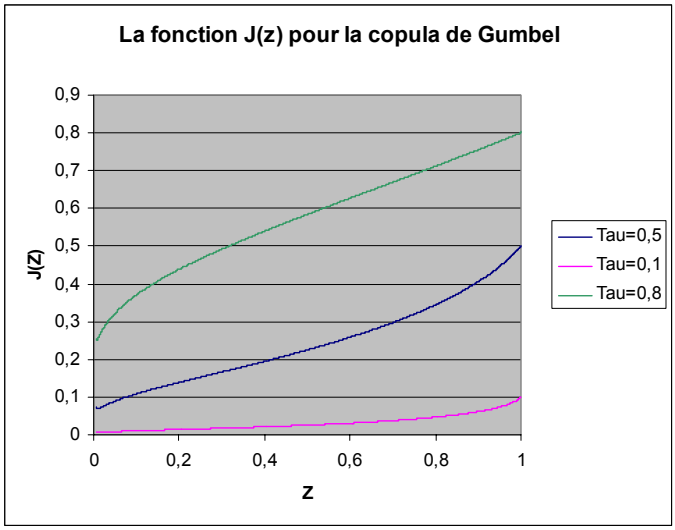
J(z) a été établie en chaque point z=1/N,2/N,…,1 d'après

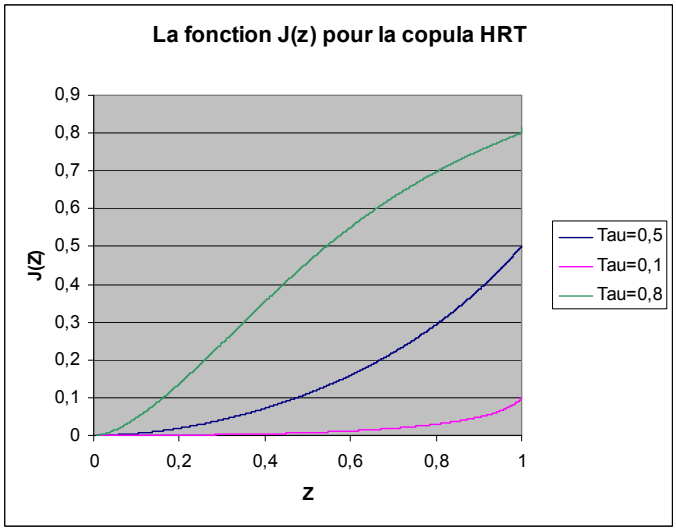
$$J(z) = -1 + 4\frac{Int_z}{C(z,z)^2}$$

Il est évident que l'approximation est très mauvaise pour des valeurs de z très faibles. Mais en découpant $[0,1]^2$ en 1 million de petits carrés (N=1000) et en représentant graphiquement J(z) à partir de 1/200, nous pouvons considérer le résultat acceptable.

Les graphiques de la fonction J(z) pour différentes copulas et différentes valeurs de τ sont représentés ci-après :







On constate des différences notables dans le comportement de la fonction J(z) pour les différentes copulas, la fonction J(z) peut démarrer avec une faible corrélation (Frank) ou une correlation plus importante (Gumbel). Nous n'avons pas représenté la copula de Clayton dans la mesure où J(z) est représentée très rapidement par une ligne horizontale dont la valeur est bien entendu égale à τ. Ceci s 'explique aisément dans la mesure ou la copula de Clayton possède un coefficient de « right tail dependence » et donc la « dépendance est surtout concentrée près du point (0,0) ».

L'intérêt de cette fonction réside là encore dans le fait qu'une version empirique de celle-ci puisse être établie.
Nous allons, comme pour la fonction K(z), donner la procédure pour établir la fonction J(z) de manière empirique. Considérons le même échantillon d'observations $(x_1,y_1)\ldots(x_n,y_n)$. Notons rang x_i le rang de la i-ème observation dans l'ensemble des valeurs $x_1,\ldots,x_n$. La procédure à suivre consiste à
1 Définir C(z,z) en prenant

$$C(z,z)=\frac{\text{nombre de paires } \{x_j,y_j\} \text{ telles que rang } x_j < z(n+1) \text{ et rang } y_j < z(n+1)}{n}$$

2 Définir la pseudo-observation z_i pour chaque i=1…n comme précédemment

$$z_i=\frac{\text{nombre de paires } \{x_j,y_j\} \text{ telles que } x_j < x_i \text{ et } y_j < y_i}{n-1}$$

3 Définir ensuite la quantité

$$I(z)=\frac{1}{n}\sum_{i=1}^{n}\left(z_i \times 1\{\text{rang } x_i < z(n+1) \text{ et rang } y_i < z(n+1)\}\right)$$

4 Nous définissons ensuite J(z) par

$$J(z)=-1+\frac{4I(z)}{C(z,z)^2}$$

Il ne reste plus qu'à comparer la version « empirique » ainsi obtenue à la version « paramétrique » de J(z) pour différentes copulas.

2.6.3 La fonction M(z)

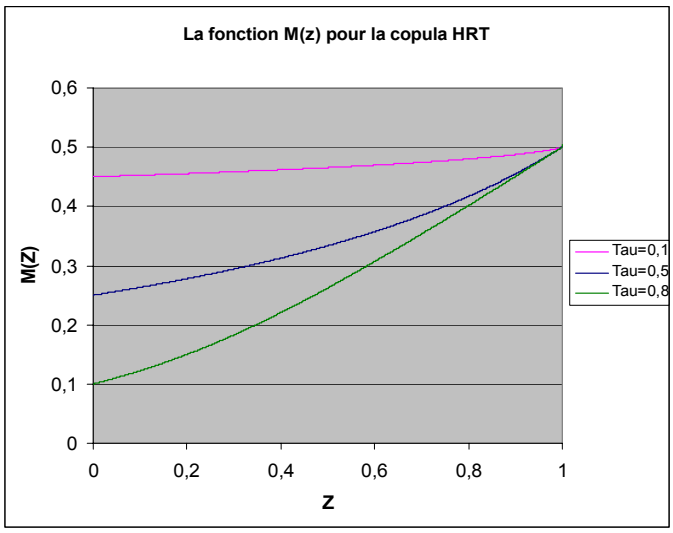
Posons

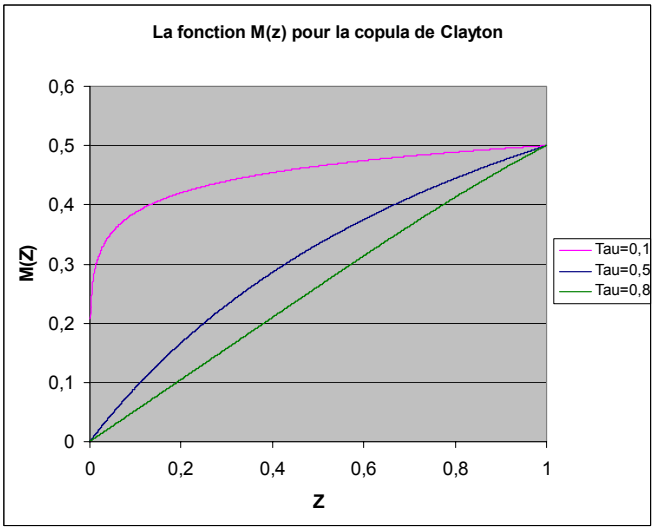
$$M(z)=E\left(V\middle|U<z\right)=\frac{\int_{u=0}^{z}\int_{v=0}^{1} vc(u,v)dudv}{z}$$

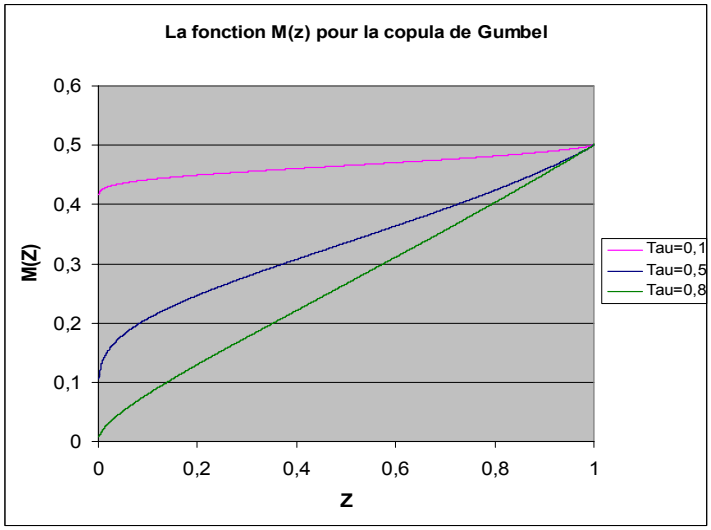
Comme E[V]=1/2, quelle que soit la copula choisie, la fonction M vérifiera M(1)=1/2. La différence entre les copulas se fera donc au niveau des petites valeurs de z et à l'allure de la courbe au fur et à mesure que l'on approche du point z=1.
Afin d'établir le graphique de cette fonction, nous avons à nouveau eu recours à des méthodes d'intégration numérique.

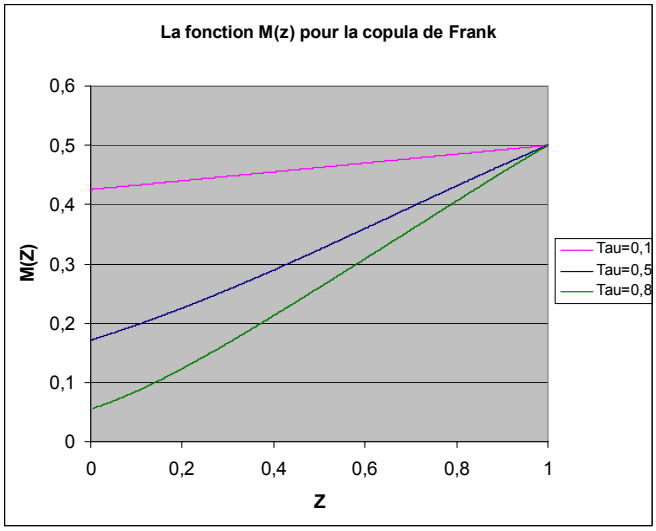
Le principe de cette intégration numérique reste à peu près le même que pour la fonction J(z), à savoir que nous avons divisé le carré $[0,1]^2$ en un nombre très grand de petits carrés. Nous prenons la valeur de la fonction vc(u,v) au centre de chaque carré, et nous modifions les bornes pour la sommation.

Les graphiques ainsi obtenus pour plusieurs copulas sont présentés ci-après.









La procédure a suivre pour calculer cette fonction à partir des données $(x_1,y_1)\ldots(x_n,y_n)$ consiste à :

1 Définir D(z) en prenant

$$D(z)=\sum_{i=1}^{n}1\{\text{rang}\,x_i < z(n+1)\}$$

2 et définir N(z) suivant la formule

$$N(z)=\sum_{i=1}^{n}1\{\text{rang}\,x_i < z(n+1)\}\times\frac{\text{rang}\,y_i}{n+1}$$

3 et enfin M(z)=N(z)/D(z).

Il est évident M(1)=1/2 puisque

$$N(1)=\frac{1}{n+1}\sum_{i=1}^{n}i=\frac{1}{n+1}\times\frac{n(n+1)}{2}=\frac{n}{2}$$

Et comme D(1)=n, on retrouve bien M(1)=1/2 .

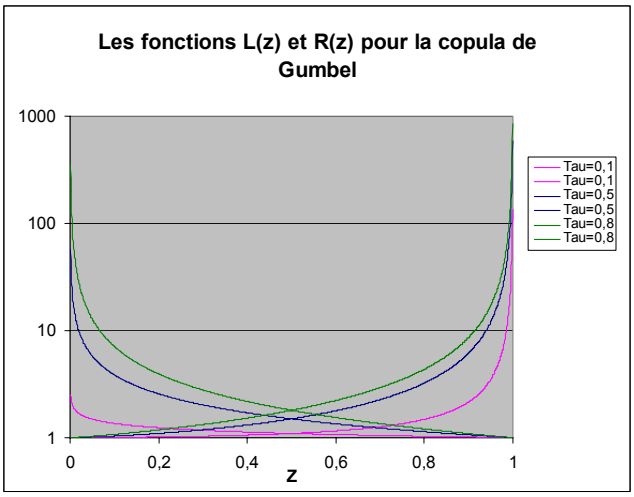
2.6.4 Les fonctions L(z) et R(z) ou “Tail concentration functions”

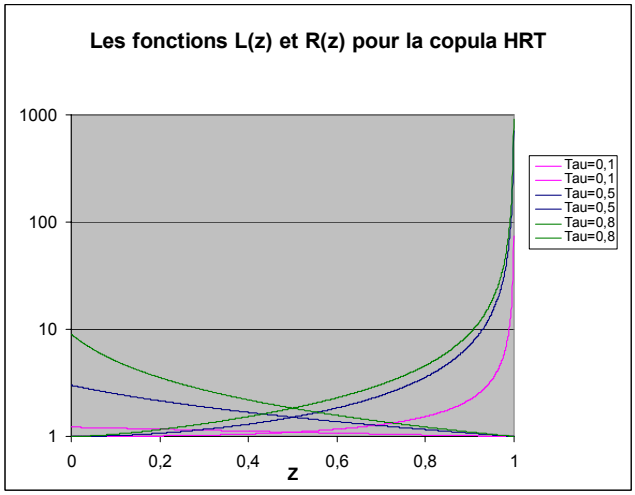
Nous nous somme déjà penchés précedemment sur les concepts de “right tail” et “left tail dependence”.

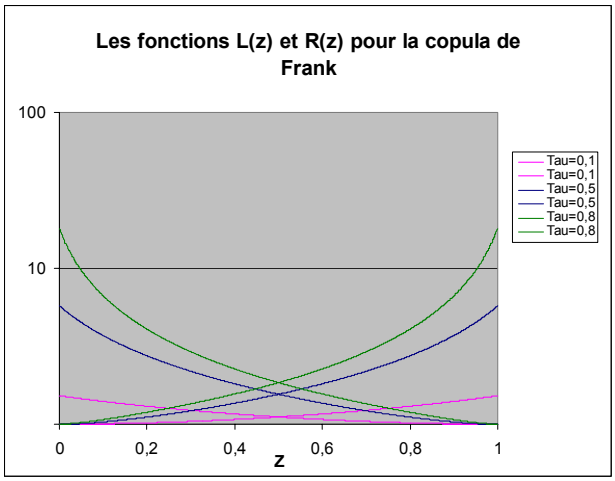
Les fonctions que nous allons présenter maintenant sont totalement basées sur ces concepts, et vont nous permettre encore une fois de comparer le comportement de ces fonctions aux versions non-paramétriques de celles-ci établies à partir des données.

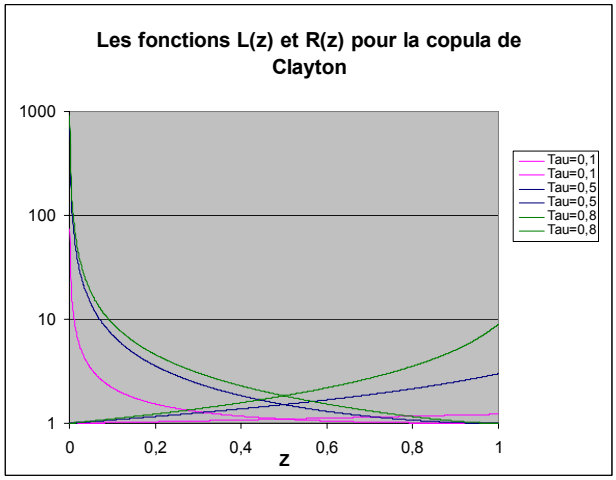
Pour tout z appartenant à (0,1), nous définissons
$L(z)=P(U<z,V<z)/z^2$ et $R(z)=P(U>z,V>z)/(1-z)^2$

Ces fonctions mettent donc en valeur la "concentration de probabilité" aux abords des points (0,0) et (1,1). Celles ci-peuvent donc se traduire très facilement en termes de copulas, puisque:
$L(z)=C(z,z)/z^2$ et $R(z)=(1-2z+C(z,z))/(1-z)^2$.
Nous avons expliqué précedemment (lors de la présentation de la fonction de tau cumulatif) comment calculer C(z,z) à partir des données. Etablir la version non-paramétrique de ces fonctions ne pose donc à priori aucun problème.
Les graphiques de ces fonctions sont présentés ci-dessous pour différentes valeurs du tau de Kendall.
Nous savons que R(0)=L(1)=1, donc ces fonctions pourront très facilement être différenciées l'une de l'autre.









Une constatation s'impose d'emblée au vu de ces graphiques: plus le τ de Kendall est important, plus les valeurs de ces fonctions sont élevées aux abords des points dits intéressants (à savoir 0 pour L(z) et 1 pour R(z))

Comme nous pouvions nous y attendre, les graphiques des copulas HRT et Clayton sont parfaitement symétriques.

La copula de Gumbel présente une forte concentration dans les deux queues, même si la droite est de loin la plus épaisse. Au fur et à mesure que la corrélation diminue, l'importance relative de la queue à gauche décroit par rapport à celle de droite.

La copula HRT présente une concentration à droite aussi importante que la copula de Gumbel, mais à gauche celle-ci est quasiment nulle.

La copula de Frank est, quant à elle, totalement symétrique.

Nous venons donc de passer en revue les différentes fonctions qui serviront de préliminaire à notre étude. Nous sommes maintenant en mesure d'étudier le phénomène des tempêtes en utilisant ces copulas.

Troisième Partie

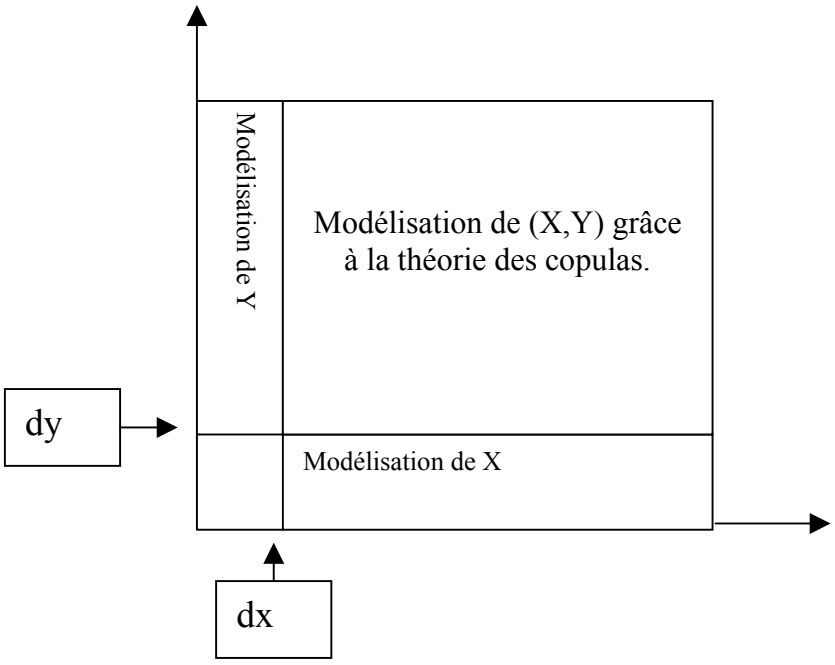
Adéquation à une loi bivariée grâce aux Copulas

Dans cette partie, nous allons estimer les paramètres d'une distribution bivariée grâce à la théorie des copulas, et ainsi modéliser les dépendances entre la branche Auto d'une part et la branche incendie d'autre part.
Nous allons travailler à la hauteur de deux niveaux de sélection, car nous avons vu en première partie que nous ne pourrions pas garder toutes les tempêtes afin de mener à bien notre étude.
Nous avons en fait choisi de retenir deux niveaux de sélection, qui sont les suivants:

- Tempêtes dépassant le montant de 1 000 Francs en Auto **et** en incendie.
- Tempêtes dépassant 17 500 Francs en Auto et 570 000 Francs en incendie.

Nous pouvons donc considérer les montants d'une tempête en auto et en incendie comme la réalisation d'un vecteur aléatoire (X,Y). Nous noterons X la branche incendie, et Y la branche automobile.
Nous noterons donc respectivement dx et dy les seuils retenus dans chaque branche.

Nous avons choisi de modéliser le phénomène suivant le schema présenté ci-dessous:



Nous ne pouvions décemment pas prendre en considération les montants inférieurs à 1000 Francs dans l'une des deux branches, car nous avons considéré qu'il s'agissait de valeurs abérrantes, et nous les avons dès lors considérées comme nulles.
Néanmoins, nous ne pouvions ignorer les valeurs prises par l'autre branche.

Nous exposerons de manière approfondie la méthodologie suivie pour le premier niveau de sélection, alors que nous n'exposerons que brièvement les résultats obtenus pour le second.

Nous allons donc suivre les étapes suivantes pour modéliser (X,Y)

- Identification de la ou des copulas à utiliser.
- Détermination de la ou des lois marginales correspondant au phénomène grâce à la méthode du maximum de vraisemblance.
- Vérification de la qualité de l'ajustement de ces lois grâce aux tests de Kolmogorov, Andersson-Darling et du Khi-Deux.
- Utilisation conjointe des copulas et lois marginales retenues pour estimer les paramètres de notre distribution bivariée par la méthode du maximum de vraisemblance.
- Vérification de la qualité de l'ajustement.

Et donc, à coté de tout ceci, nous estimerons également la loi marginale de X lorsque Y est inférieur à dy, et vice-versa.

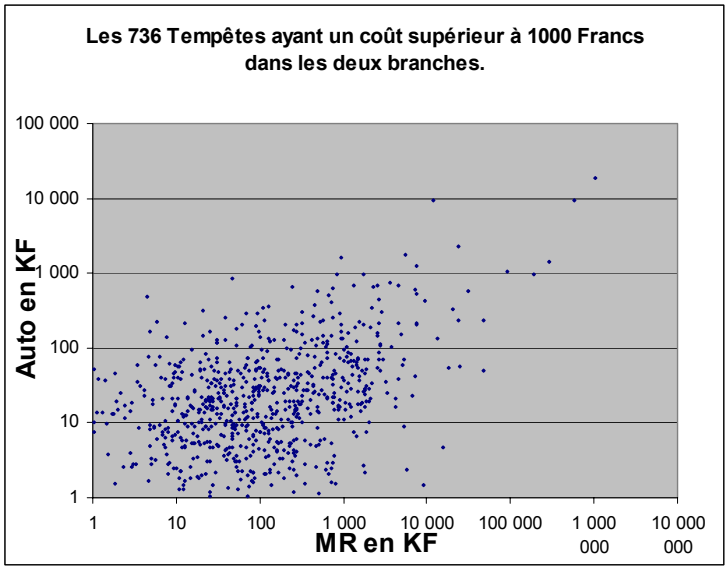
Nous utiliserons le solveur d'Excel pour tous les calculs de maximisation de la vraisemblance, et un outil appelé Instrat Fit développé par le service actuariel de Guy Carpenter afin de mieux cerner les lois marginales susceptibles de nous intéresser.

3.1 Modélisation du phénomène tempête pour le premier niveau de sélection.

3.1.1 Détermination des copulas à utiliser pour modéliser (X,Y)

Nous ne retenons dans un premier temps, que les tempêtes dont le montant a dépassé 1 000 Francs à la fois dans la branche Auto et dans la branche Incendie.
Nous travaillerons cependant en KF. Le nombre de tempêtes répondant à ces conditions est de 736.

A toutes fins utiles, le graphique représentant les montants dans les deux branches est le suivant:



Ce nuage de points montre de manière évidente qu'il existe bien une dépendance entre ces deux branches.

Le **τ de Kendall** empirique est de 0,243.

Le **ρ de Spearmann**, quant à lui, a pour valeur 0,357.

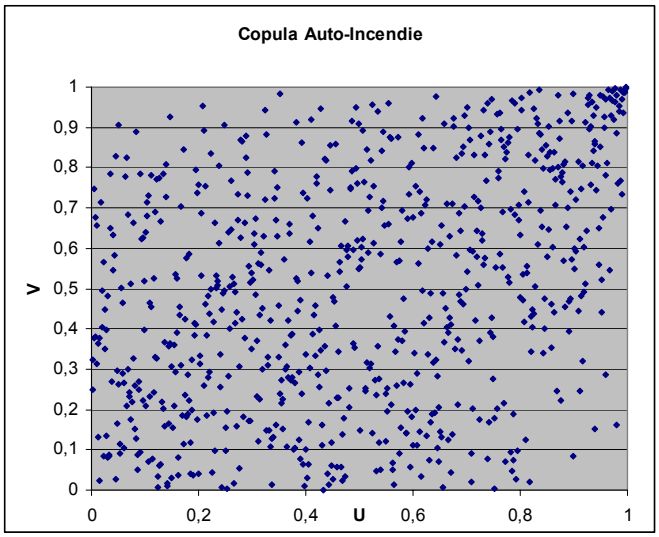
Le **ρ de Pearson** est de 0,877.

Nous allons maintenant estimer le paramètre d'association pour différentes copulas sans faire d'hypothèses sur la forme paramétrique des lois marginales.

Notons donc n le nombre de nos observations (soit 736).

La méthode consiste donc à transformer notre série de données (x_i,y_i) $\{i=1\ldots n\}$ en (u_i,v_i) en utilisant pour cela la fonction de répartition empirique.

Il suffit, pour ce faire, de repérer le rang de chaque x_i et y_i dans leurs branches respectives et de le diviser par (n+1).



Ensuite, nous pouvons utiliser la méthode du maximum de vraisemblance pour estimer le paramètre de notre copula. Conformément aux notations adoptées dans notre seconde partie, nous noterons celui-ci « a ».

Donc,

$$\hat{a} = \arg\max \sum_{i=1}^{n} \ln c(u_i, v_i; a)$$

où c représente la fonction de densité de la copula.

Nous allons donc estimer le paramètre pour les copulas de Frank, Gumbel, Clayton, HRT, et enfin la copula normale.

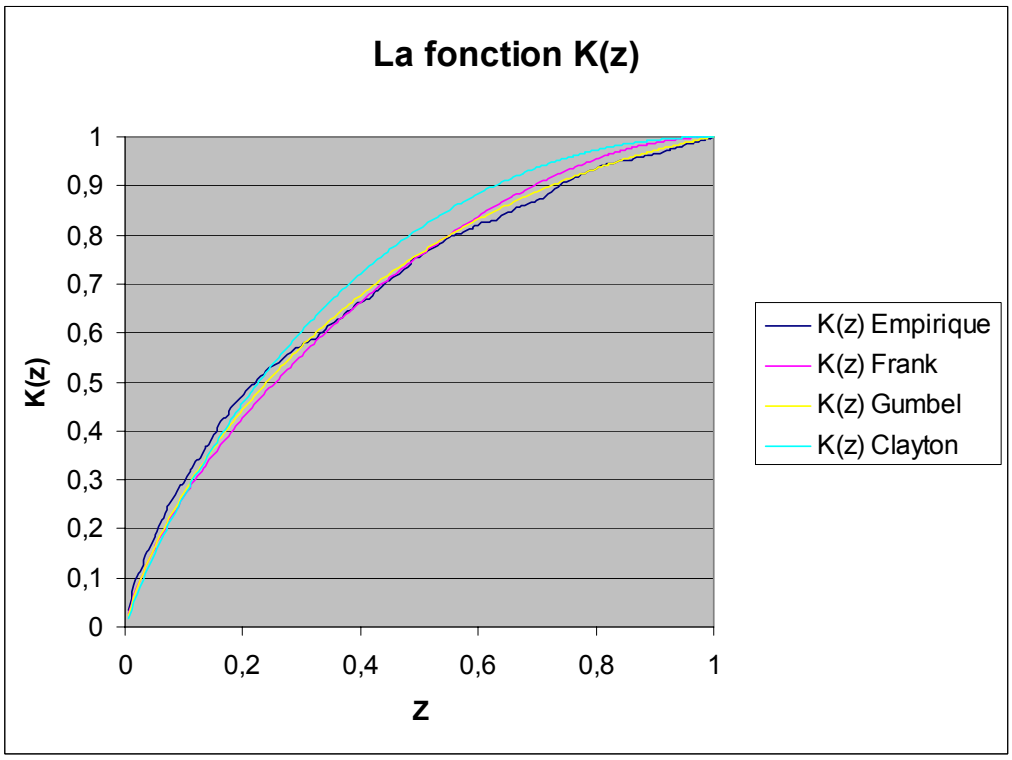
Les fonctions utilisées pour la densité sont celles qui ont été précedemment exposées.

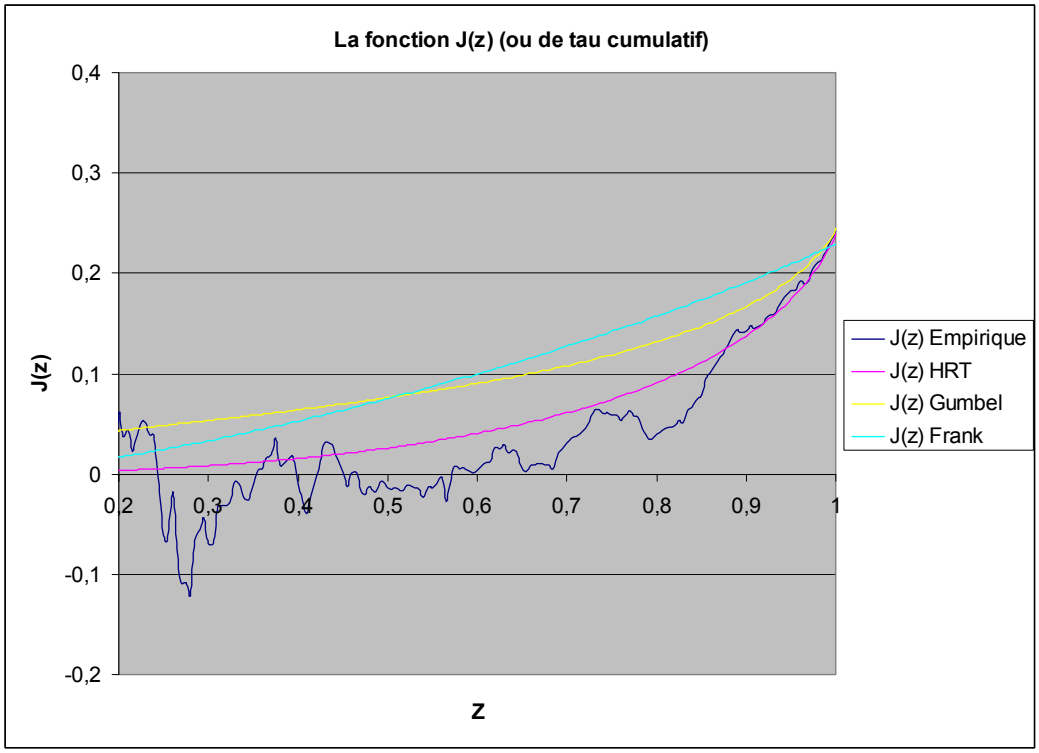
Les résultats obtenus par la méthode du maximum de vraisemblance sont donc les suivants :

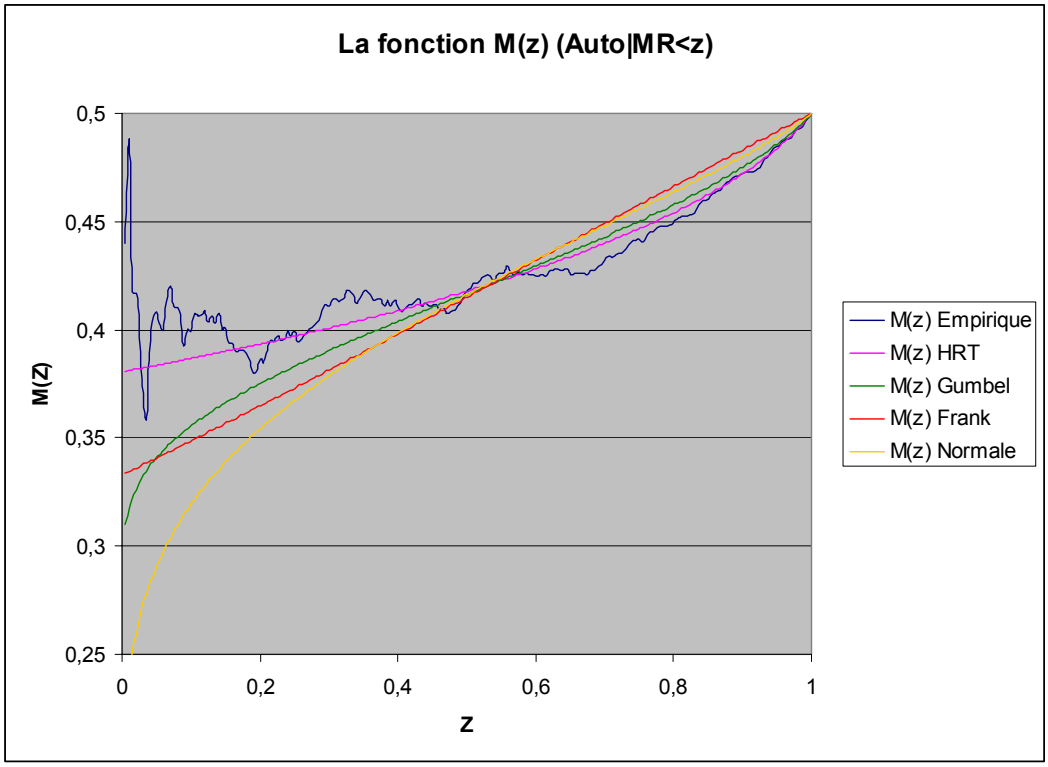
	Gumbel	Normale	HRT	Frank	Clayton
Paramètre	1,323	0,378	1,445	2,318	3,378
Log Vraisemblance	77,223	55,428	84,070	50,330	16,447
τ de Kendall	0,244	0,247	0,257	0,245	0,129

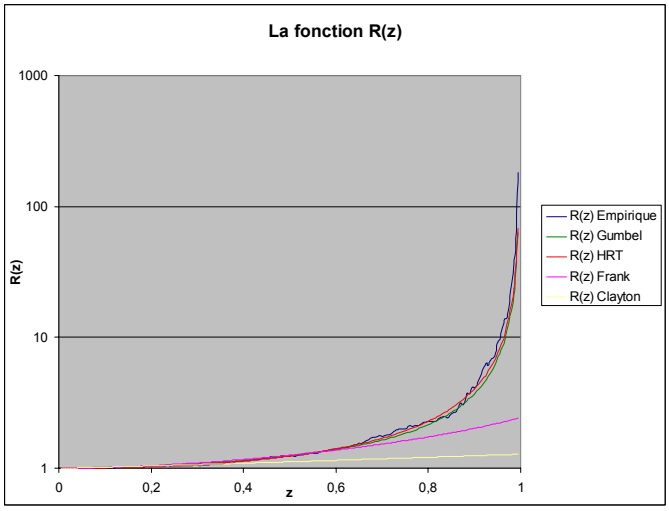
Comme la relation entre le paramètre et le τ de Kendall est « bijective », nous avons également mentionné dans le tableau la valeur de τ résultant de cette maximisation.

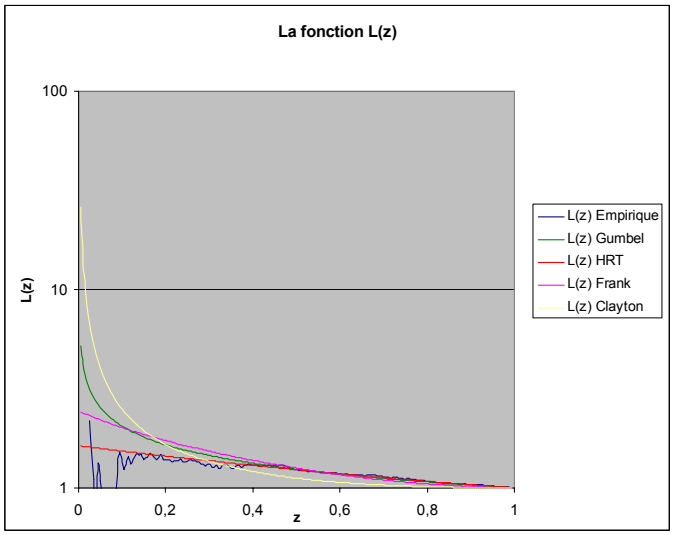
Par ailleurs, nous avons établi les versions non-paramétriques des fonctions J(z), M(z), K(z), R(z) et L(z) exposées à la fin de la seconde partie de ce mémoire.
Nous pouvons les comparer aux versions paramétriques obtenues en prenant les paramètres trouvés ci-dessus.
Les graphiques obtenus sont les suivants :






La fonction M(z) (Auto|MR<z)
0,5
0,45
0,4
0,35
0,3
0,25
M(Z)
0
0,2
0,4
0,6
0,8
1
z
M(z) Empirique
M(z) HRT
M(z) Gumbel
M(z) Frank
M(z) Normale


La fonction R(z)
1000
100
10
1
R(z)
0
0,2
0,4
0,6
0,8
1
z
R(z) Empirique
R(z) Gumbel
R(z) HRT
R(z) Frank
R(z) Clayton


La fonction L(z)
100
10
1
L(z)
0
0,2
0,4
0,6
0,8
1
z
L(z) Empirique
L(z) Gumbel
L(z) HRT
L(z) Frank
L(z) Clayton

Au vu des graphes présentés et de la valeur de la fonction du logarithme de la vraisemblance à son maximum, il apparait que deux copulas pourraient effectivement correspondre à nos données.
La meilleure candidate est sans conteste la copula HRT puisque c'est elle qui maximise « le mieux »la fonction de vraisemblance, et se rapproche au plus près des différentes fonctions indicatrices calculées. Il semble d'ailleurs, au vu du graphique nommé « Copula Auto-Incendie » que la seule concentration visible de points se trouve aux abords de (1,1), ce qui corrobore le fait que cette copula soit la plus appropriée.
Nous allons néanmoins garder la copula de Gumbel, car au vu du graphe de K(z) il s'agit de la meilleure copula Archimédienne, et en terme de vraisemblance elle se place en seconde position. La fonction R(z) nous montre bien que son comportement « à droite » est similaire à celui de la copula HRT, même si « à gauche » elle ne retranscrit pas tout à fait le comportement exact du phénomène observé.
Ceci est du au fait que la copula de Gumbel induit également une dépendance entre les petites valeurs de u et de v, et explique sûrement pourquoi la courbe de J(z) passe légèrement au dessus de la courbe non-paramétrique.

3.1.2 Détermination des lois marginales

Comme nous n'avons pris en compte que les données à partir d'un certain seuil, il est nécessaire d'utiliser des lois tronquées.
Une loi tronquée à gauche au seuil dx est définie de la manière suivante :
Supposons que X soit une variable aléatoire continue ayant F_x comme fonction de répartition. Nous définissons la variable aléatoire tronquée T de fonction de répartition F_{Tx} en prenant T_x=X si X>dx, et T_x est non définie dans le cas contraire.
La fonction de répartition de T est donc la suivante :

$$F_{T_X}(x) = P(X \le x | X > dx) = \frac{F_X(x) - F_X(dx)}{1 - F_X(dx)} \quad \text{si} \quad x > dx$$

et 0 sinon.Pour une loi donnée, et en notant $\theta=(\theta_1,\theta_2,\ldots,\theta_n)$ le vecteur des paramètres de cette loi, la fonction de log-vraisemblance à maximiser est la suivante :

$$l(\theta) = \sum_{i=1}^{n} \ln\left(\frac{f_X(x_i;\theta)}{1 - F_X(dx;\theta)}\right)$$

Nous avons grâce au logiciel Instrat-Fit effectué cette opération pour une quinzaine de lois différentes.
Les lois marginales testées furent les suivantes :

- Lois à un paramètre : Exponentielle, Pareto simple, Generalized Extreme value limit.
- Lois à deux paramètres : Gamma, Weibull, Gamma Inverse, Log Gamma, Inverse Weibull, Loglogistique, Paralogistique, Inverse Paralogistique, Log-Normale.
- Lois à trois paramètres : Burr, Inverse Burr.

L'expression de toutes ces lois se trouve en annexe, sous le titre « Instrat continuous distribution ».

Le critère de selection du logiciel est appelé HQ, qui est tout simplement la valeur absolue de la fonction de log-vraisemblance à son maximum à laquelle on rajoute le produit entre le nombre de paramètres et le logarithme du nombre de données divisé par 2π.

Les deux lois que nous avons retenues aussi bien pour la branche Automobile que pour la branche Incendie sont les lois Loglogistique et Paralogistique.
Les fonctions de répartition et les densités de ces lois sont les suivantes :

La loi Loglogistique

$$F_X(x) = 1 - \frac{1}{1 + \left(\frac{x}{\theta}\right)^{\alpha}} \qquad f_X(x) = \frac{\alpha}{\theta} \frac{\left(x/\theta\right)^{\alpha-1}}{\left(1 + \left(x/\theta\right)^{\alpha}\right)^2}$$

La loi Paralogistique

$$F_X(x) = 1 - \frac{1}{\left(1 + \left(\frac{x}{\theta}\right)^{\sqrt{\alpha}}\right)^{\sqrt{\alpha}}} \qquad f_X(x) = \frac{\alpha}{\theta} \frac{\left(\frac{x}{\theta}\right)^{\sqrt{\alpha}-1}}{\left(1 + \left(\frac{x}{\theta}\right)^{\sqrt{\alpha}}\right)^{\sqrt{\alpha}+1}}$$

En fait ces deux lois, ne sont que des cas spéciaux de la fonction de distribution Burr qui est elle plus connue et dépend de 3 paramètres.

La loi Burr

$$F_X(x) = 1 - \frac{1}{\left(1 + \left(\frac{x}{\theta}\right)^{\beta}\right)^{\frac{\alpha}{\beta}}} \qquad f_X(x) = \frac{\alpha}{\beta} \frac{\left(\frac{x}{\theta}\right)^{\beta-1}}{\left(1 + \left(\frac{x}{\theta}\right)^{\beta}\right)^{\frac{\alpha}{\beta}+1}}$$

Cette loi est de type « heavy tailed distribution » ou loi à queue épaisse. La loi Paralogistique est donc une loi Burr avec la contrainte $\beta=\alpha^{1/2}$, et la Loglogistique satisfait la contrainte $\beta=\alpha$. Néanmoins, nous n'avons pas retenu la loi Burr, car la maximisation de la log-vraisemblance par cette loi nécessitait une amélioration d'au moins 4,76 (=ln(736/2π)), ce qui n'a pas été le cas.

L'estimation des paramètres par la méthode du maximum de vraisemblance a donc conduit aux résultats suivants :

	Incendie		Automobile	
	Loglogistique	Paralogistique	Loglogistique	Paralogistique
estimateur de theta	97,690	71,574	18,245	19,245
estimateur de alpha	0,771	0,703	1,063	1,087
Max Log vraisemblance	-5 088,752	-5 089,080	-3 587,450	-3 587,442
Critère HQ	5 098,279	5 098,606	3 596,976	3 596,969

3.1.3 Vérification de la qualité des ajustements

Nous avons choisi de vérifier la qualité des ajustements par trois tests.
Nous utiliserons tout d'abord les tests de Kolmogorov-Smirnov et Andersson-Darling qui sont de type EDF (empirical distribution fonction) et qui indiquent l'écart entre la fonction de répartition empirique et la fonction de répartition de la loi testée.
Nous utiliserons également le test du Khi-deux pour vérifier ces ajustements.

3.1.3.1 Le test de Kolmogorov

Nous voulons donc tester le fait que des observation suivent une loi théorique donnée de fonction de répartition F_0.
Or nous savons que F_n, qui est la fonction de répartition empirique doit converger vers F_o.
La statistique de ce test que nous noterons D_n est définie par $D_n=\max(D_n^+,D_n^-)$ en prenant

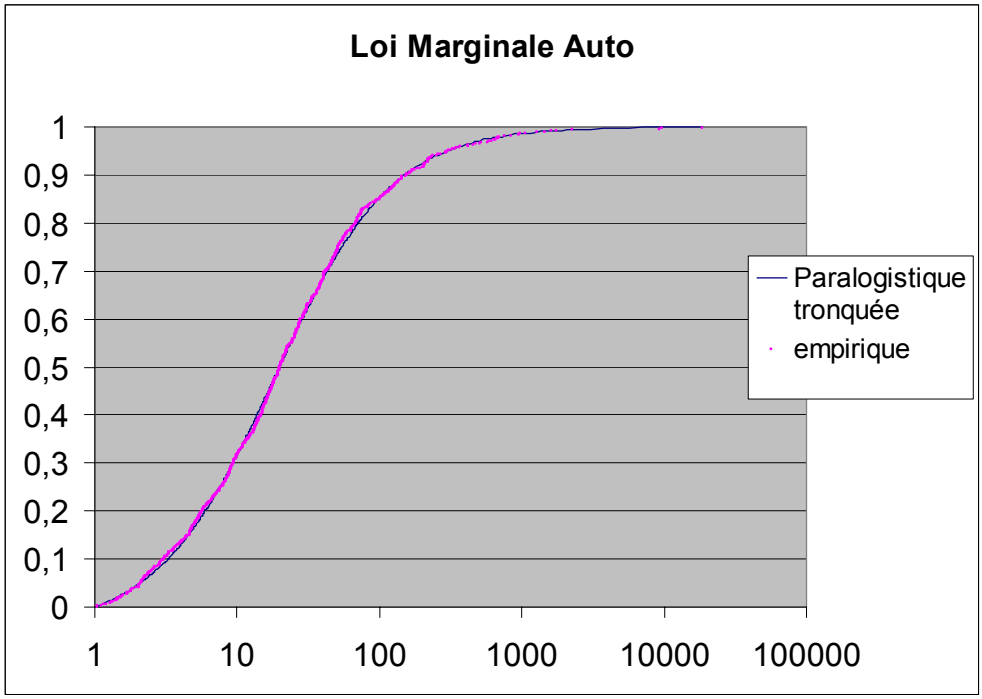
- $D_n^+ = \underset{1\le i\le n}{\text{Max}}\left|\frac{i}{n} - F_o\left(x_{(i)}\right)\right|$
- $D_n^- = \underset{1\le i\le n}{\text{Max}}\left|\frac{i-1}{n} - F_o\left(x_{(i)}\right)\right|$

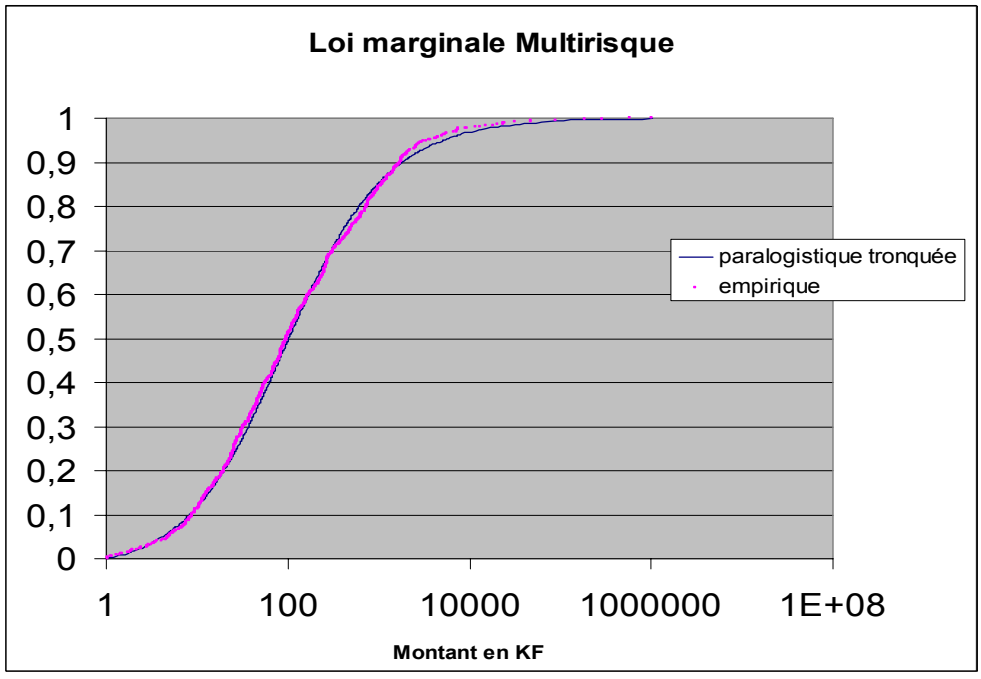
La région critique de ce test : $W=\{x_1,\ldots,x_n / D_n>c\}$ et $P_{Fo}(W)=\alpha$.
Pour n>100, au seuil $\alpha=0.05$, la valeur critique c est donnée par la table des quantiles pour le test de Kolmogorov-Smirnov qui est $1,36/\sqrt{n}$, soit dans notre cas, 0,0501.
Les résultats obtenus pour nos adéquations par les différentes lois sont :

	Automobile		Incendie	
	Paralogistique	Loglogistique	Paralogistique	Loglogistique
Stat de Kolmogorov	0,0216	0,0218	0,0324	0,0311
Valeur critique	0,0501	0,0501	0,0501	0,0501

Il semblerait que malgré le nombre très important de données, notre adéquation ne soit pas foncièrement mauvaise puisque les 4 adéquations satisfont le test de Kolmogorov. Le seul problème du test de Kolmogorov est son champ d'application puisque si le nombre de données est faible, il arrive très rarement qu'un modèle soit rejeté, alors que si il est important, le test du khi-deux est plus approprié.

A titre indicatif, les graphiques suivants permettront de mieux s'apercevoir de l'ajustement réalisé pour les deux lois de type paralogistique :





L'adéquation est globalement meilleure pour la branche automobile même si le résultat en incendie semble convenable. Il semblerait de prime abord que la loi paralogistique constitue un meilleur ajustement pour l'automobile et la loglogistique pour l' incendie.

3.1.3.2 Le test d'Andersson-Darling

La statistique de ce test est basée sur l'importance de la queue de distribution.
Son expression est définie par :

$$A_n^2 = n\int_{-\infty}^{+\infty}\left(F_n(x)-F_0(x)\right)^2\frac{1}{F_0(x)(1-F_0(x))}dF_0(x)$$

Ce qui peut se réecrire facilement sous la forme :

$$A_n^2 = -n-\frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln F_0\left(x_{(i)},\hat{\theta}\right)+\ln\left(1-F_0\left(x_{(n-i+1)},\hat{\theta}\right)\right)\right]$$

La règle de décision est la suivante : Nous sommes amenés à rejeter l'hypothèse H_o (F_o=F_n) si notre statistique est supérieure à une valeur critique $c_{(1-\alpha)}$.

Pour α=0.05 et notre nombre de données la valeur critique de ce test est de 2.492.

	Automobile		Incendie	
	Paralogistique	Loglogistique	Paralogistique	Loglogistique
Statistique AD	0,262	0,270	0,835	0,819
Valeur critique	2,492	2,492	2,492	2,492

Toutes nos lois semblent satisfaire le test.

3.1.3.3 Le test du Khi-deux

Ce test est le test d'adéquation le plus communément employé.

Comme d'habitude nous noterons

$H_o \quad F_X(x) = F(x;\theta)$

$H_1 \quad F_X(x) \neq F(x;\theta)$

Le principe de ce test consiste à effectuer une partition de notre échantillon en J ensembles.

Pour le j^{eme} groupe, le test est basé sur n_j qui est tout simplement le nombre d'observations dans chaque groupe et sur

$$E_j = n \times P\left(X \in j^{ème} \text{ groupe}; \hat{\theta}\right)$$

où n représente la taille de l'échantillon, et la probabilité est celle qu'une observation se trouve dans le j^{eme} groupe avec θ prenant ses valeurs préalablement estimées.
En d'autres termes, E_j est tout simplement le nombre d'observations attendu dans le j^{eme} groupe avec le modèle et les estimations paramètriques utilisés.
La statistique du test est alors :

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

La valeur critique $c_{1-\alpha}$ est le nombre tel que $P(\chi^2 > c_{1-\alpha}) = \alpha$, où χ^2 suit une distribution du khi-deux à (J-1-nombre de paramètres estimés).
Le choix des classes de partition est relativement important dans la mesure où celui-ci peut dénaturer les résultats de manière significative.
Certaines personnes insistent sur le fait que le nombre théorique E_j soit supérieur à 5 dans chacune des classes, et que le nombre de classes choisi pour le découpage soit compris entre $n^{1/3}$ et $n^{2/3}$.

Nous avons décidé de faire une partition de notre échantillon en 16 intervalles aussi bien dans la branche incendie que pour l'automobile.
Le nombre de degrès de liberté de notre loi du khi-deux est par conséquent de 13 puisque deux paramètres ont été estimés pour chaque loi et pour chaque branche.
La valeur critique au seuil 0,05 est donc de 22,36 dans tous les cas.

	Automobile		Incendie	
	Paralogistique	Loglogistique	Paralogistique	Loglogistique
Stat du Khi-deux	16,484	16,639	17,750	17,272
valeur critique	22,362	22,362	22,362	22,362

Au vu des résultats, l'hypothèse H_o n'est pas rejetée et ces résultats confirment ceux produits par les deux tests précedents.

3.1.4 Détermination des paramètres de la distribution bivariée.

Nous allons donc conserver pour estimer au mieux les paramètres de notre distribution bivariée la copula HRT bien sûr, la copula de Gumbel, et aussi les lois marginales paralogistique et loglogistique.
Nous avons vu, que, dans un cadre général, la densité de (X,Y) se présentait sous la forme:

$$f(x,y) = f_X(x)\, f_Y(y)\, c\big(F_X(x), F_Y(y)\big)$$

où c représente la densité de la copula, F_X et F_Y les distributions marginales de X et Y respectivement, et f_X, f_Y leurs densités marginales.

Nous notons toujours θ le vecteur des paramètres. Celui-ci a dans le cas présent la dimension de ce vecteur est de 5×1 puisque nous avons en fait deux paramètres à estimer pour chaque marginale plus le paramètre de dépendance.
En notant θ_X et θ_Y les composantes du vecteur θ relatives aux lois marginales de X et Y, la fonction de log-vraisemblance devient par suite:

$$l(\theta) = \sum_{i=1}^{n} \ln\big(\, c\big(F_X(x_i;\theta_X), F_Y(y_i;\theta_Y); a\big)\,\big) + \sum_{i=1}^{n} \ln f_X(x_i;\theta_X) + \sum_{i=1}^{n} \ln f_Y(y_i;\theta_Y)$$

En notant $\hat{\theta}_{ML}$ l'estimateur obtenu par la méthode du maximum de vraisemblance, celui-ci possède une propriété de normalité asymptotique sous certaines hypothèses que nous ne détaillerons pas ici, ce qui d'un point de vue mathématique se réecrit:

$$\sqrt{n}\big(\hat{\theta}_{ML} - \theta_0\big) \to N\big(0, \tau^{-1}(\theta_0)\big)$$

où $\tau(\theta_o)$ représente la matrice d'information de Fischer.
Cette assertion est exacte pour toutes les estimations par la méthode du maximum de vraisemblance et permet ainsi de donner des intervalles de confiance pour les paramètres. Malheureusement, au vu des calculs nécessaires et particulièrement compliqués pour obtenir cette matrice de Fischer, nous nous dispenserons de cette estimation.

Comme dans notre cas nous avons utilisé des lois marginales tronquées, notre fonction de log-vraisemblance à maximiser prendra la forme:

$$l(\theta) = \sum_{i=1}^{n} \ln\left(c\left(\frac{F_X(x_i) - F_X(dx)}{1 - F_X(dx)}, \frac{F_Y(y_i) - F_Y(dy)}{1 - F_Y(dy)} \right) \right) + \sum_{i=1}^{n} \ln \frac{f_X(x_i)}{1 - F_X(dx)} + \sum_{i=1}^{n} \ln \frac{f_Y(y_i)}{1 - F_Y(dy)}$$

$$l(\theta) = \sum_{i=1}^{n} \ln\left(c\left(F_{T_X}(x_i), F_{T_Y}(y_i)\right) \right) + \sum_{i=1}^{n} \ln f_{T_X}(x_i) + \sum_{i=1}^{n} \ln f_{T_Y}(y_i)$$

Par soucis de clarté, nous n'avons pas fait apparaitre dans la formule de log-vraisemblance les composantes du vecteur θ.

Il parait peu probable que l'on puisse trouver une solution analytique à ce problème de maximisation. Nous ne pouvons utiliser que des techniques numériques itératives. Nous nous sommes donc renseignés auprès de Stuart Klugman ("Fitting bivariate loss distributions with copulas") et Gary Venter ("Tails of copulas") pour savoir quels logiciels ils avaient mis à contribution pour effectuer ce genre de calcul. Dans le premier cas un add-on Excel utilisant le simplex a été utilisé, dans le second c'est le solveur d'Excel.

Comme le solveur procède par itérations, nous avons dans un premier temps utilisé comme valeurs de départ celles obtenues lors des maximisations précedemment effectuées, tant pour le paramètre de la copula que ceux des lois marginales. Puis, nous avons changé de valeurs de départ (au moins une dizaine de répétition) pour s'assurer que nous retrouvions le même résultat. Ceci fut systématiquement le cas. Les calculs ont néanmoins été effectué avec un ordinateur possédant 1 G_o de Ram et deux processeur Pentium 3 agissant de concert.

Les résultats obtenus pour différentes combinaisons furent les suivants:

	Copula HRT	
	Auto Paralogistique	MR Paralogistique
estimateur de theta	19,127	73,532
estimateur de alpha	1,075	0,711
estimateur de a	1,448	
Max log-vraisemblance	-8592	

	Copula HRT	
	Auto Paralogistique	MR Loglogistique
estimateur de theta	19,277	98,916
estimateur de alpha	1,085	0,774
estimateur de a	1,472	
Max log-vraisemblance	-8591	

	Copula de Gumbel	
	Auto Paralogistique	MR Paralogistique
estimateur de theta	19,861	85,221
estimateur de alpha	1,115	0,766
estimateur de a	1,323	
Max log-vraisemblance	-8613	

	Copula de Gumbel	
	Auto Paralogistique	MR Loglogistique
estimateur de theta	20,053	108,620
estimateur de alpha	1,129	0,833
estimateur de a	1,317	
Max log-vraisemblance	-8615	

Les meilleurs résultats en termes de maximisation de la vraisemblance sont obtenus avec la **combinaison copula HRT, loi paralogistique pour l'automobile et loi loglogistique pour l' incendie**. La maximisation obtenue en supposant l'indépendance entre les deux lois marginales est tout simplement la somme des deux résultats obtenus pour les lois marginales soit à peu près –8676. L'utilisation des copulas pour modéliser le phénomène constitue donc une amélioration certaine.

3.1.5 Tests d'adéquation d'une distribution bivariée.

Nous allons exposer ici les différents tests d'adéquation possibles pour une distribution bivariée. Nous concentrerons nos efforts sur la présentation des résultats pour **la meilleure combinaison** possible au regard de la maximisation de la fonction de log-vraisemblance.

Avant d'effectuer nos tests d'adequation nous allons tracer la ligne de regression médiane

3.1.5.1 La ligne de regression médiane

Pour une valeur donnée d'une tempête en incendie x, la valeur y sur cette ligne est la solution de l'équation $0,5=F_{Ty|Tx}(y \mid x)$, ce qui en terme avec la copula HRT peut se réecrire

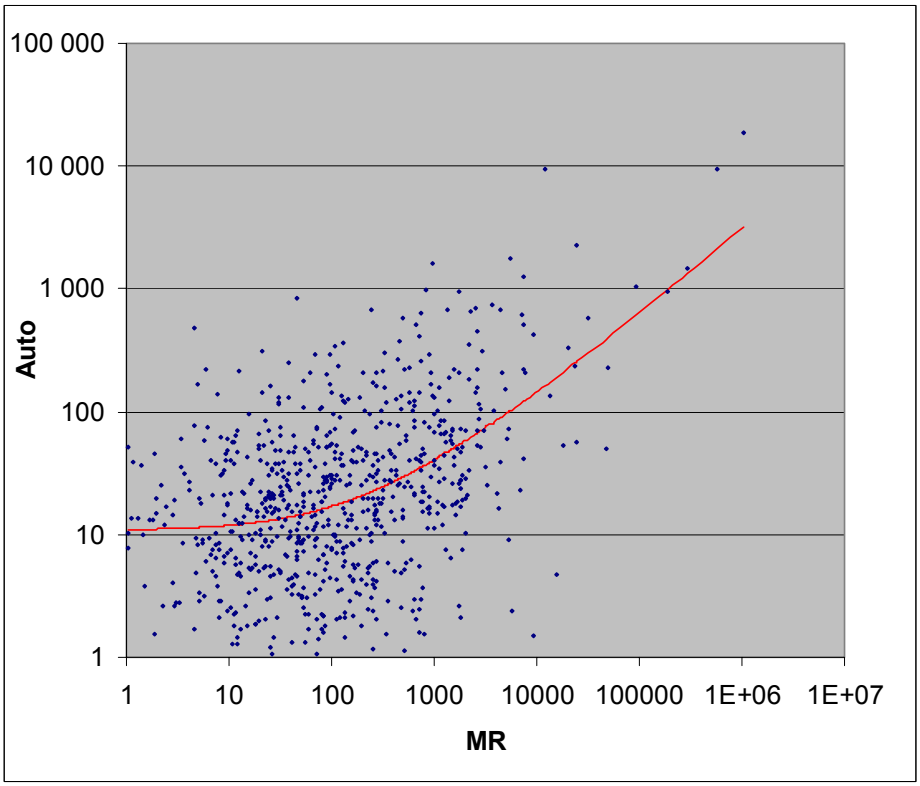
$$0{,}5 = C_1\left(F_{T_X}(x), F_{T_Y}(y)\right) = 1 - \left[\left(1 - F_{T_X}(x)\right)^{-1/a} + \left(1 - F_{T_Y}(y)\right)^{-1/a} - 1\right]^{-a-1}\left(1 - F_{T_X}(x)\right)^{-1-1/a}$$

En résolvant par rapport à F_Y , nous obtenons

$$v = F_{T_Y}(y) = 1 - \left[1 - \left(1 - F_{T_X}(x)\right)^{-1/a} + \left(\frac{1}{2}\left(1 - F_{T_X}(x)\right)^{1+1/a}\right)^{\frac{-1}{1+a}}\right]^{-a}$$

Puis nous obtenons y en inversant v

$$y = F_{T_Y}^{-1}(v) = \theta\left[\left(1 - \left(v\left(1 - F_y(d_y)\right) + F_y(d_y)\right)\right)^{\frac{-1}{\sqrt{\alpha}}} - 1\right]^{\frac{1}{\sqrt{\alpha}}}$$



Le résultat au premier coup d'oeuil semble plutôt bon, dans la mesure où notre ligne semble bien passer au milieu de notre nuage de points.

3.1.5.2 Test du Khi-deux

Nous allons donc vérifier la qualité de notre distribution bivariée par l'intermédiaire du test du khi-deux dont le principe a été precédemment exposé.
Nous exposerons deux techniques différentes pour vérifier la justesse de l'adéquation

- Le première consiste à étendre le test à deux dimensions
- La seconde permet de se ramener à deux tests à une dimension.

3.1.5.2.1 Extension du test à deux dimensions.

Nous avons en fait découpé l'espace en un certain nombre de pavés, dont certains sont bien entendu de taille infinie. La taille des pavés diminue au fur et à mesure que l'on s'approche du mode.
Nous avons ensuite déterminé la probabilité théorique qu'une observation se retrouve dans ce pavé.
Nous savons que $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = F(x_2,y_2) - F(x_2,y_1) - F(y_2,x_1) + F(x_1,y_1)$.
Ce qui se réecrit donc,

$$C(F_X(x_2), F_Y(y_2)) - C(F_X(x_2), F_Y(y_1)) - C(F_X(x_1), F_Y(y_2)) + C(F_X(x_1), F_Y(y_1))$$

Là encore cette formule est écrite dans le cas général. Nous avons à nouveau pris en compte le caractère tronqué de nos lois dans nos calculs.

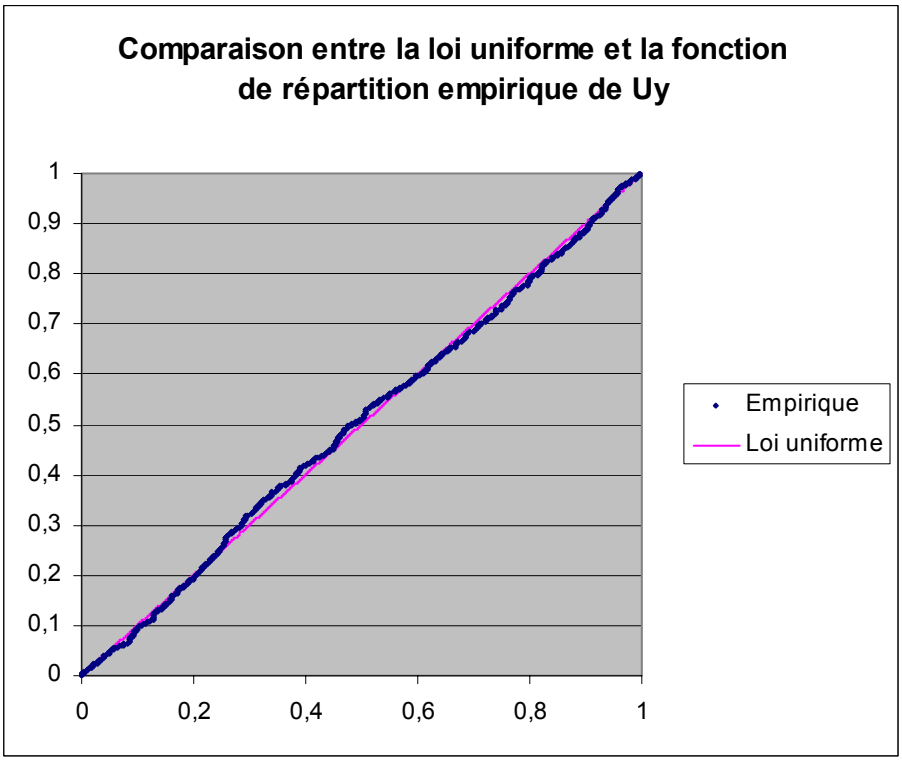
$$C(F_{T_X}(x_2), F_{T_Y}(y_2)) - C(F_{T_X}(x_2), F_{T_Y}(y_1)) - C(F_{T_X}(x_1), F_{T_Y}(y_2)) + C(F_{T_X}(x_1), F_{T_Y}(y_1))$$

Nous avons donc découpé notre espace en 22 rectangles plus ou moins réguliers.
Le nombre de degrès de liberté de notre loi du khi-deux est de 22-5-1, soit 16.
Notre statistique du khi-deux a pour valeur 23,86 alors que notre valeur critique se situe à 26,3. Le test se révèle donc concluant.

A titre indicatif, cette même statistique a également été calculée avec la copula de Gumbel et les deux lois paralogistiques et a été évaluée à 26,3. Elle constitue donc un ajustement moins bon.
Nous avons également mené ce test en supposant l'indépendance entre les deux branches: la statistique ainsi déterminée a pris pour valeur 186.

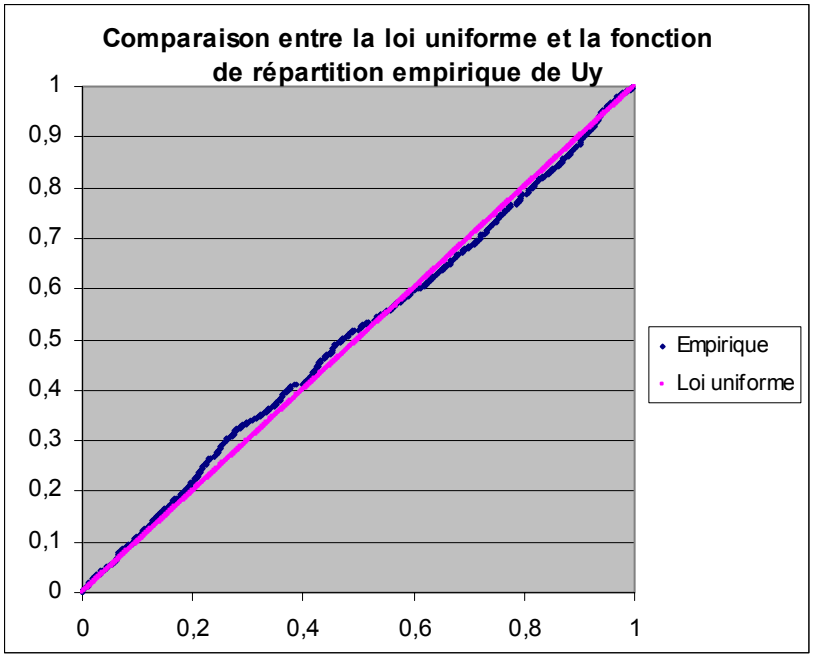
3.1.5.2.2 Utilisation de deux tests uni-dimensionnels

Cette méthode utilise le fait que les variables aléatoires $U_{Tx}=F_{Tx}(T_x)$ et $U_{Ty}=F_{Ty|Tx}(T_y|T_x)$ sont indépendantes et ont une distribution uniforme sur (0,1).
Nous allons donc tester dans un premier temps le fait que U_{Tx} suive une loi uniforme, puis si cela s'avère être le cas, nous ferons de même pour U_{Ty}. Nous avons là encore besoin de la dérivée partielle de notre copula par rapport à u. (ie:$C_1(u,v)$)
Nous avons, dans un premier temps, divisé l'intervalle (0,1) en 20 segments d'une longueur de 0,05 chacun, et utilisé le test du khi-deux avec un nombre de degrés de liberté de 14.(=20-5-1)
Notre statistique est de 14,65 alors que la valeur critique est de 23,68.
La première condition étant remplie, nous avons ensuite effectué le second test avec toujours 20 intervalles.
La statistique du Khi-deux pour ce second test est de 21.86, soit toujours en deça de notre valeur critique.



Ce graphique permet de mieux se rendre compte de la qualité de l'adéquation, puisque la loi empirique est très semblable à la loi uniforme.

A titre de comparaison, ce même graphique, établi à partir de l'adéquation à la copula de Gumbel, et deux paralogistiques pour les lois marginales a l'allure suivante:



L'adéquation est dans ce cas nettement moins bonne. Le premier test s'est révélé concluant, mais ceci ne fut pas le cas pour le second.

Malgré le nombre important de données nous avons donc réussi à trouver une distribution bivariée qui satisfasse à tous nos tests d'adéquations.

3.1.6 Etude des lois marginales annexes.

Nous allons maintenant étudier, pour pouvoir être tout à fait complet dans notre démarche, le comportement de la variable X lorsque Y est inférieur à dy et vice-versa.

Nous nous contenterons ici de donner les résultats obtenus, en sachant toutefois que les adéquations retenues ont satisfait tous les tests.

La loi marginale retenue pour l' incendie lorsque l'automobile était inférieure à 1000 Francs, a été évaluée par rapport à 406 données.
La meilleure adéquation possible s'est révélée être une loi lognormale tronquée.
Les paramètres μ et σ furent respectivement estimés 3,39 et 1,764.

En ce qui concerne l'automobile, nous avons ajusté la fonction de répartition à une loi exponentielle de paramètre λ égal à 24,01.

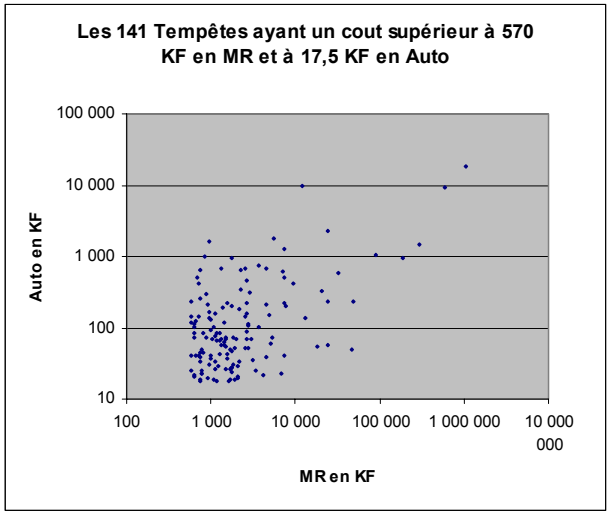
En définitive, nous venons de montrer comment établir les paramètres d'une distribution bivariée par l'intermédiaire des copulas. Nous avons trouvé un modèle satisfaisant tous nos tests d'adéquations. Néanmoins, il est très difficile d'estimer la fréquence annuelle de survenance de telles tempêtes si tant est que l'on puisse les nommer ainsi, et on peut supposer qu'aux niveaux de seuils trés élevés qui intéressent les réassureurs nos adéquations ne sont guère significatives.
Les articles qui nous ont servis de référence pour cette étude sont *"fitting bivariate loss distribution with copulas[3]*" et "*understanding relationship using copulas[2]*" mentionnés dans notre introduction. L'ordre de grandeur des données utilisées dans ces articles est pourtant à peu près le même que le nôtre en termes de montants.

Nous allons néanmoins restreindre le nombre de "tempêtes" prises en comptes, et recommencer toutes les opérations d'estimation des lois marginales et des copulas adéquates . Bien évidemment, nous ne rentrerons pas de manière aussi approfondie dans les explications théoriques, notamment au niveau des tests.
Nous sommes conscient que le fait de scinder notre problème en trois parties *(modélisation de (X,Y) lorsque X et Y sont supérieurs à dx et dy, modélisation de X uniquement lorsque Y est inférieur à dy et vice-versa)* n'est pas sans poser quelques problèmes. Nous avons néanmoins retenu cette solution car c'est la seule qui nous permettait de mettre en place des indicateurs du choix de la copula qui soient rigoureux au niveau théorique, et nous permettaient ensuite de vérifier la qualité de nos adéquations.

3.2 Modélisation du phénomène tempête pour le second niveau de sélection.

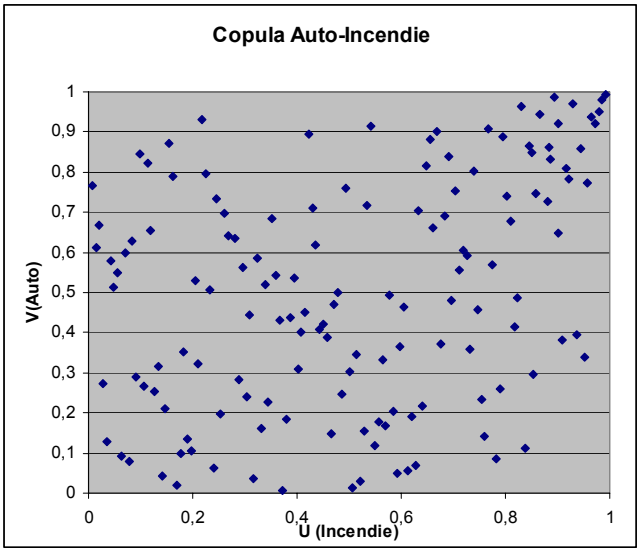
Nous allons donc procéder exactement de la même manière que précedemment, sauf que les seuils retenus ne seront plus de 1000 Francs dans chaque branche, mais de 17 500 (dy) en automobile et de 570 000 Francs (dx) en incendie.
L'ensemble des sinistres Automobile dépassant 17500 représente 97% de la "sinistralité tempêtes" sur 11 ans dans cette branche, et il en est de même pour la branche incendie. C'est ce qui a justifié notre choix.
Les montants retenus sont cependant nettement inférieurs aux priorités usuelles dans les contrats de réassurance en matière d'évènements "tempêtes". C'est pourquoi nous avons jugé qu'il n'était en fait pas très génant de n'étudier que les montants "incendie" supérieurs à 570 KF lorsque l'automobile était inférieure à 17.5 KF.

3.2.1 Détermination de la copula à utiliser pour modéliser (X,Y)



Le graphique ci-dessus représente donc un "zoom" des précedents sur l'ensemble des sinistres dépassant les seuils en incendie **ET** en Automobile. C'est donc la structure de dépendance entre ces sinistres que nous nous attacherons à étudier.
Le coefficient de Kendall empirique est de 0,23.

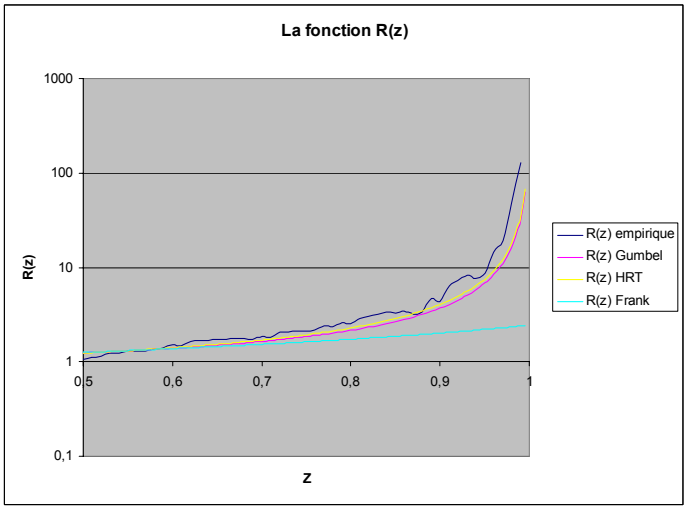
Nous allons donc, comme auparavant déterminer le paramètre de dépendance sans faire d'hypothèse sur les lois marginales.



La maximisation de la fonction de log-vraisemblance a conduit aux résultats suivants:

	Gumbel	Normale	HRT	Frank	Clayton
Paramètre	1,356	0,365	1,302	2,056	4,387
Log Vraisemblance	15,049	9,123	17,439	7,423	1,603
Tau de Kendall	0,263	0,238	0,277	0,219	0,102

Nous avons a nouveau calculé les diverses fonctions nous permettant d'affiner notre choix. Nous ne représenterons ici que le comportement de R(z).



Contrairement à la partie précedente, nous ne conserverons que la copula HRT.

3.2.2 Détermination des lois marginales à utiliser et tests d'adéquation

Comme nous avons augmenté le seuil à partir duquel nous prenons en compte les données, il peut être utile d'utiliser une loi de type "Generalized Pareto" pour modéliser les lois marginales. Nous allons dans un premier temps tracer la "Mean Excess Function" pour voir si l'utilisation de cette loi est nécessaire.

3.2.2.1 Mean excess function

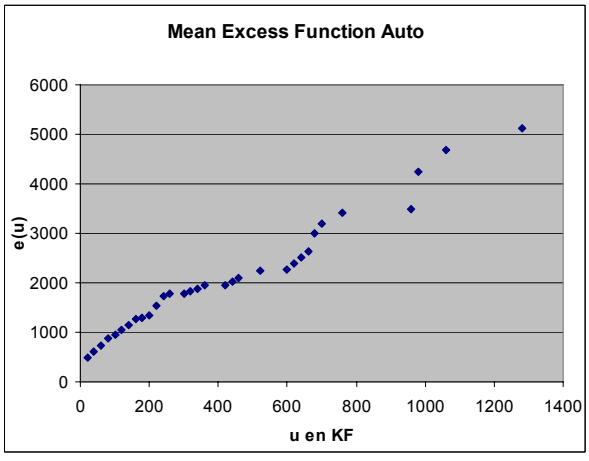
Si X est une variable aléatoire, alors la mean excess function est définie par

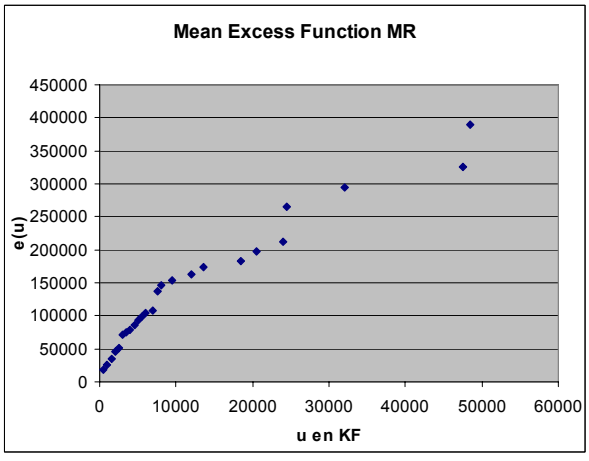
$$e(u) = E[X - u | X > u]$$

Son expression empirique, en considérant que l'on a un échantillon $x_1,\ldots,x_n$ de la v.a X est donc la suivante:

$$e(u) = \frac{\sum_{i=1}^{n} (x_i - u)^+}{\sum_{i=1}^{n} 1_{\{x_i > u\}}}$$

Si cette fonction est croissante, alors nos lois marginales seront de type "heavy tailed", et nous pourrons éventuellement essayer d'ajuster nos données sur une loi de Pareto généralisée.





Nous avons arrêté les graphiques dès lors que nous disposions de moins de six données au dessus du seuil.

Ces "deux mean excess function" sont strictement croissantes, et par conséquent nous pouvons ajouter la loi Pareto généralisée au panel déjà testé par le logiciel Instrat Fit.

Cependant, nous sommes tout à fait conscient que cette loi de Pareto généralisée devrait en fait être appliquée à partir de seuils beaucoup plus élevés que les seuils retenus, et que rien ne nous permet d'affirmer que l'adéquation risque d'être meilleure.

3.2.2.2 Résultats obtenus

Nous ajoutons donc au panel de lois d'Instrat Fit la loi de Pareto généralisée dont l'expression est:

$$G_{dx,\xi,\tau}(x) = 1 - \left(1 + \frac{\xi(x - dx)}{\tau}\right)^{-1/\xi} \quad \text{si} \quad x \geq dx$$

et 0 sinon.

La fonction de log-vraisemblance à maximiser est donc

$$l(\xi,\tau) = -n\ln(\tau) - \left(1 + \frac{1}{\xi}\right)\sum_{i=1}^{n} \ln\left(1 + \xi\frac{x_i - dx}{\tau}\right)$$

qui dépend ainsi de deux paramètres.

Nous utiliserons, une fois de plus, le solveur d'Excel pour résoudre ce problème de maximisation.

Les deux meilleurs ajustements possibles pour la branche automobile et la branche incendie ont été respectivement une loi log-normale tronquée et une loi de Pareto simple. Nous allons comparer la qualité de ces ajustements avec celle obtenue pour la loi de Pareto généralisée.

La loi de Pareto

$$F_X(x) = 1 - \left(\frac{c}{x}\right)^{\alpha} \quad x \geq c \qquad\qquad f_X(x) = \frac{\alpha}{x}\left(\frac{c}{x}\right)^{\alpha} \quad x \geq c$$

La loi log-normale tronquée

En notant ϕ la loi normale centrée-réduite, l'expression de la loi log-Normale tronquée est donc:

$$F_T(x) = \frac{\phi\left(\frac{\ln(x) - \mu}{\sigma}\right) - \phi\left(\frac{\ln(dx) - \mu}{\sigma}\right)}{1 - \phi\left(\frac{\ln(dx) - \mu}{\sigma}\right)} \qquad f_T(x) = \frac{\exp\left(-\frac{1}{2}\left(\frac{\ln(x) - \mu}{\sigma}\right)^2\right)}{\sigma\sqrt{2\pi}x\left(1 - \phi\left(\frac{\ln(dx) - \mu}{\sigma}\right)\right)}$$

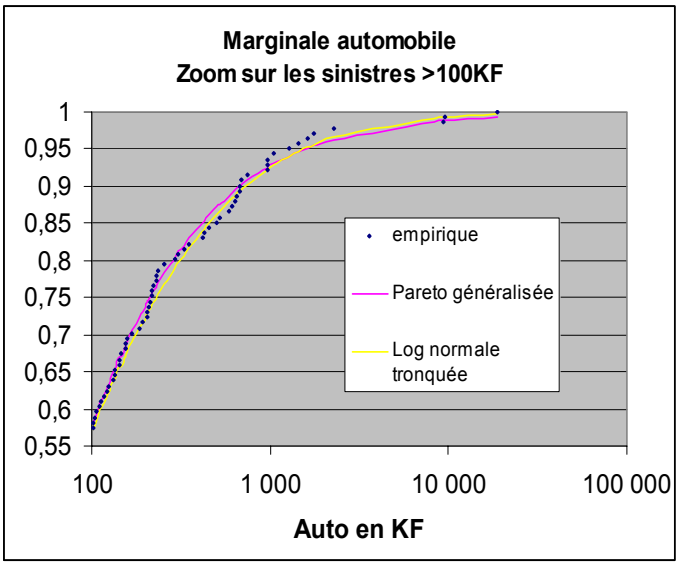
si x≥dx et 0 sinon.

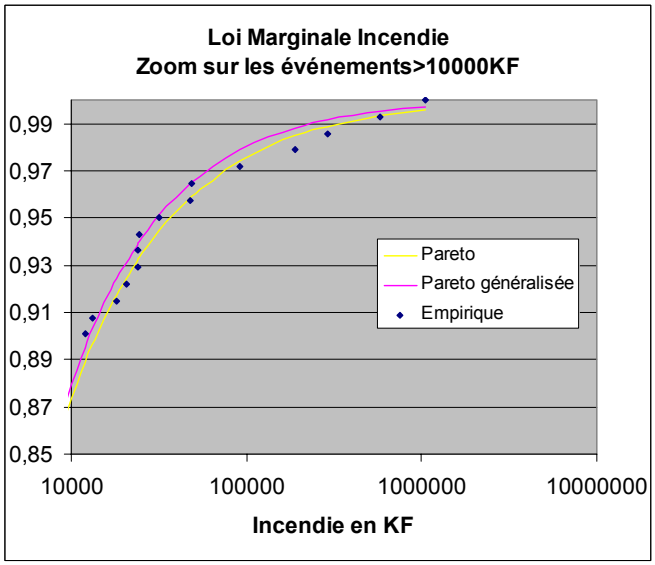
Les estimations paramétriques ont conduit aux résulats suivants:

	Marginale Auto
	Lognormale tronquée
estimateur de mu	2,399
estimateur de sigma	2,435
log vraisemblance	-872,742
	Pareto généralisée
estimateur de Ksi	1,176
estimateur de tau	55,571
log vraisemblance	-873,360

	Marginale Incendie
	Pareto simple
estimateur de alpha	0,721
log vraisemblance	-1277,283
	Pareto généralisée
estimateur de Ksi	1,256
estimateur de tau	898,03
log vraisemblance	-1276,930

Nota bene: Afin de vérifier l'exactitude des maximisations du solveur d'Excel, nous avons repris avec celui-ci des travaux effectués en interne chez Guy Carpenter et qui utilisaient des techniques appropriées aux cas des lois log-normale tronquée et Pareto généralisée.
Nous avons abouti strictement aux mêmes estimations paramétriques.





3.2.2.3 Test du Khi-deux

Les résultats du test du khi-deux pour l'automobile et un découpage en 14 tranches sont les suivants:

	Auto	
	Lognormale tronquée	Pareto généralisée
stat du khi-deux	8,639	9,702
Valeur critique	19,675	19,675

Et en Incendie

	Incendie	
	Pareto simple	Pareto généralisée
stat du khi-deux	14,725	12,496
Valeur critique	21,026	19,675

Dans la branche automobile, le choix est très facile puisque la loi lognormale tronquée minimise la statistique du khi-deux et maximise la fonction de log-vraisemblance et ceci pour un nombre de paramètres estimés identiques.

La réponse est un peu plus ambigue pour la seconde branche. Apparemment la Pareto généralisée présente une statistique du khi-deux plus faible, mais nous estimons un paramètre supplémentaire.

Nous avons retenu, à tort ou a raison, les deux lois, car en tout état de cause elles satisfont toutes deux le test et elles ont un comportement similaire.

3.3.3 Détermination des paramètres de la distribution bivariée et tests d'adéquation

Les estimations paramétriques ont donc conduit aux résultats présentés ci-dessous:

	Copula HRT		
	Lognormale Auto		Pareto MR
estimateur de mu	2,585	estimateur de alpha	0,722
estimateur de sigma	2,367		
	Copula HRT		
estimateur de a	1,365	Max log-vraisemblance	-2132,051

	Copula HRT		
	Lognormale Auto		Pareto Generalisée MR
estimateur de mu	2,874	estimateur de Ksi	1,209
estimateur de sigma	2,231	estimateur de Tau	941,061
	Copula HRT		
estimateur de a	1,473	Max log-vraisemblance	-2131,320

Comme nous n'utilisons que 141 données, l'estimation des paramètres de notre distribution bivariée conduit à des changements beaucoup plus significatifs des paramètres des lois marginales. En outre, l'incertitude autour des valeurs estimées est beaucoup plus importante que dans la partie précedente.
L'amélioration de la valeur de la fonction de log-vraisemblance à son maximum n'est pas assez significative pour préferer le second modèle.

Nous allons maintenant effectuer le test d'adéquation du khi-deux pour le **premier modèle**. Vu le nombre de données assez retreint dont nous disposons, nous ne pourrons effectuer le test en deux dimensions qui semble plus adapté. Nous nous ramènerons donc à **deux tests unidimensionnels**.
Le nombre de paramètres estimés est de 4, et par conséquent la valeur critique est de 23,68. Le nombre de tranches choisies est de 19.
Le premier test a conduit a une statistique du khi-deux de 15,38.
Le second a une statistique égale à 16,60.
Pour plus de sureté ce test a également été conduit en sens inverse.($F_y(Y)$ puis $F_x(X| Y=y)$.)

Le premier test a pour statistique 17,65 le second 16,88.

3.3.4 Etude des lois marginales "annexes".

L'étude des 54 événements en incendie ayant un montant supérieur à 570 KF alors que le montant en automobile était inférieur à 17,5 conduit à un ajustement sur une loi de Pareto simple de paramètre α=1,307.

L'étude des sinistres automobile supérieurs à 17,5 KF alors que le montant en incendie était inférieur à 570 KF a donc amené à l'ajustement sur une loi de Weibull ($F_x(x)=1-\exp(-(x/\theta)^{\beta})$) tronquée au seuil 17,5 de paramètres θ=1,47 et β=0,33.
Ces estimations de paramètres satisfont les divers tests d'adéquations.

Nous sommes maintenant en mesure de simuler des tempêtes dont le montant est supérieur à 570 KF pour la branche MR et 17,5 KF pour la branche automobile.

Quatrième Partie

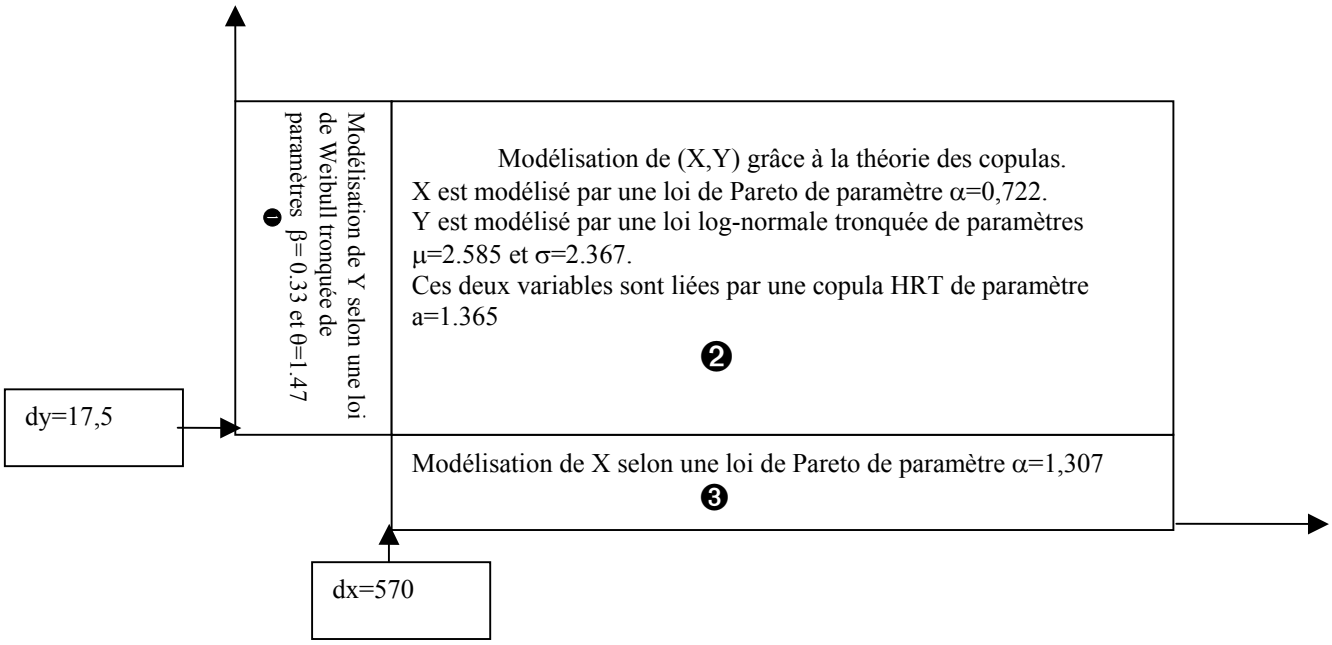
Simulations

4.1 Présentation de la méthode employée pour la simulation

4.1.1 Rappels des résultats obtenus

Nous avons étudié dans notre précédente partie, le moyen d'estimer au mieux les paramètres d'une fonction de répartition bivariée en utilisant les copulas. Nous allons à présent nous servir de ces estimations afin de coter divers contrats de réassurance de type excédent de sinistre.

Pour ce faire, nous allons nous servir des résultats obtenus pour le second niveau de sélection.



4.1.2 Utilisation de la méthode de Monte-Carlo

Nous avons donc observé sur une période de 11ans

- 255 événements entrant dans la catégorie ❶
- 141 dans la ❷
- 55 dans le ❸.

Mais l'essentiel de la sinistralité est concentrée dans les sinistres de type ❷ (plus de 95%)

Nous allons par conséquent faire l'hypothèse que les variables aléatoires N_1,N_2,N_3 qui représentent les fréquences annuelles des différentes catégories d'évènements suivent une loi de Poisson de paramètres respectifs $\lambda_1=255/11$, $\lambda_2=141/11$, $\lambda_3=55/11$.
Cette hypothèse est très souvent utilisée en matière de réassurance.
Nous aurions pu également utiliser une loi de fréquence de type binomiale négative .

Nous disposions des logiciels Excel (avec l'add-on @risk) et SAS pour réaliser ces simulations: notre choix s'est porté sur ce second logiciel.
En effet, le générateur de nombre aléatoire d'Excel est réputé pour être d'une qualité douteuse. Par ailleurs, le temps d'éxécution de telles simulations avec le logiciel SAS est tout à fait acceptable.

Nous allons nous baser sur un nombre n (=50000) d'années de simulations.

Nous avons donc simulé 3 n variables aléatoires $N_j^1,\ldots,N_j^n \quad \text{pour } j=1,2,3$
suivant des lois de Poisson de paramètre λ_j simulées grâce à la fonction ranpoi de SAS.
Supposons maintenant que pour la $i^{ème}$ année de simulation nous ayons 3 réalisations N_1^i, N_2^i, N_3^i .
Nous allons donc simuler ensuite des couples de variables aléatoires
$\left\{ {}_1X_j^i,{}_1Y_j^i \right\},\ldots,\left\{ {}_{N_j^i}X_j^i,{}_{N_j^i}Y_j^i \right\} \quad j=1,2,3$

- <u>Simulation pour les sinistres de type ❶</u>

Nous allons considérer que $X_1=0$.
Y_1 va suivre une loi de Weibull tronquée au seuil dy (=17,5) de paramètres θ et β.
Pour simuler une telle loi, nous notons tout d'abord

$$F_Y(dy)=1-e^{\left(-\left(\frac{dy}{\theta}\right)^{\beta}\right)}$$

Ensuite nous simulons une variable aléatoire v uniformément répartie sur [0,1].
Nous notons ensuite:

$$v'=\left(\left(1-F_Y(dy)\right)v\right)+F_Y(dy)$$

La simulation y_1 de Y_1 est alors

$$y_1 = \theta\left[\ln\left(\frac{1}{1-v'}\right)\right]^{\frac{1}{\beta}}$$

- Simulation pour les sinistres de type ❷

Pour le second type d'événement, nous allons donc simuler un couple de variables aléatoires (X_2,Y_2) de lois marginales respectives Pareto(α) et log-normale tronquée (μ,σ) "reliées" entre elles par une structure de dépendance de type copula HRT.
Nous allons pour ce faire simuler deux variables aléatoires aléatoires **indépendantes** et uniformément réparties sur [0,1] dont les réalisations seront notées **u et p**.

Nous allons ensuite déterminer **v** en prenant
$v=1-[(1-(1-u)^{-1/a}+[(1-p)(1-u)^{1+1/a}]^{(-1/(1+a))}]^{-a}$.

Nous déterminons ensuite, en appelant ϕ la fonction de répartition de la loi normale centrée réduite:

$$F_Y(d_y) = \phi\left(\frac{\ln(dy)-\mu}{\sigma}\right)$$

Nous définissons comme précedemment

$$v' = \left(\left(1-F_Y(dy)\right)v\right)+F_Y(dy)$$

et enfin nous obtenons y_2 en prenant

$$y_2 = \exp\left(\mu+\sigma\phi^{-1}\left(v'\right)\right)$$

Pour simuler la variable aléatoire X_2, il suffit de prendre

$$x_2 = \frac{dx}{(1-u)^{1/\alpha}}$$

Ainsi, nous avons simulés des réalisations du vecteur aléatoire (X_2,Y_2) ayant les propriétés désirées.

- Simulation pour les sinistres de type ❸

Nous allons considérer que $Y_3=0$.

Comme précedemment, nous allons prendre X_3

$$x_3 = \frac{dx}{(1-u)^{1/\alpha}}$$

ou u représente une variable aléatoire uniformément répartie.

Le paramètre α de la loi de Pareto n'est pas le même que pour les sinistres de type ❷.

Dès lors, toutes les cotations de contrats de type excédent de sinistre sont réalisables, quelles que soient les conditions sur chacune des deux branches.

4.2 Présentation des résultats obtenus

4.2.1 Comparaison entre deux excédents de sinistres par branche et un excédent de sinistre sur la somme des deux branches.

Nous allons, pour chacune des deux branches, fixer une priorité et une portée.

Nous les noterons par exemple: $Prio_{auto}$ et $Port_{auto}$ pour la branche automobile et $Prio_{MR}$ et $Port_{MR}$ pour la branche incendie.

Pour le cumul des deux branches, nous les nomerons $Prio_{Total}$ et $Port_{Total}$.

Les variables aléatoires représentant la charge du réassureur pour chaque événement sont définies par exemple pour l'automobile par:

$${}_kW_j^i = \min\left(\left(\max\left({}_kY_j^i - Prio_{auto}, 0\right), Port_{auto}\right) \right) \quad j = 1,2.$$

La charge du réassureur, pour tous les sinistres automobiles survenus sur une période de un an est donc

$$S_{auto}{}^i = S_1^i + S_2^i = \sum_{k=1}^{N_1^i} {}_kW_1^i + \sum_{k=1}^{N_2^i} {}_kW_2^i$$

Le lecteur n'aura aucun mal à établir l'équivalent de cette formule pour la branche incendie.

Pour le cumul des deux branches, la formule de la charge du réassureur pour chaque événement est donc

$${}_kW_j^i = \min\left(\left(\max\left({}_kY_j^i + {}_kX_j^i - Prio_{Total}\right),0\right), Port_{Total}\right) \quad j = 1,2,3.$$

La prime pour chacun de ces contrats sera donc établie en prenant la moyenne sur les années ainsi simulées, et ainsi pour l'automobile par exemple:

$$P_{auto} = \frac{1}{n}\sum_{i=1}^{n} S_{auto}^i$$

En fait, nous verrons que, vu les priorités que nous allons retenir, la quasi-totalité de la prime sera constituée par les sinistres de type ❷. Nous aurions peut-être pu nous passer de simuler ❶ et ❸, ce qui aurait considérablement allégé les formules et n'aurait sans doute pas changé l'interprétation des résultats.

Nous ne fixerons pas les montants de priorité aléatoirement. En fait nous allons choisir ceux-ci de manière à ce que, avec les paramètres retenus, la charge du réassureur soit non nulle en moyenne une année sur quatre.
Pour les plafonds, nous allons déterminer les montants par rapport au sinistre de référence, à savoir 90% de Lothar.

	Prime	ecart-type	Priorité en KF	Plafond en KF	
Total	93 501	1 093	113 000	946 800	dont 0,99% du à ❸
Incendie	92 110	1 077	111 000	930 000	dont 1,02% du à ❸
Auto	1 274	15	3 220	16 800	dont 0% du à ❶

Nous constatons que la différence entre les deux types de couvertures (séparation des deux branches, et au total) n'est pas très importante et en tout cas pas significative au

vu des erreurs commises tant au niveau de l'estimation des paramètres que de celles inhérentes à la simulation. Par contre, la quasi-totalité de la prime provient des événements simulés du type ❷, ce qui nous conforte dans notre opinion que nous aurions pu en théorie nous limiter à ce cas.
En fait, le véritable problème c'est le poids écrasant de la branche incendie par rapport à l'automobile, ce qui conduit en fait à ce que la prime XL automobile soit inférieure à 2% du montant de la prime XL en incendie.
Nous avons effectué ces mêmes simulations en conservant des priorités et des plafonds semblables, mais en supposant l'indépendance entre les deux branches.
La prime de l'XS sur le total était de 94 098 soit une légère augmentation de 600. Cela va à l'encontre de ce que nous aurions pu penser. Mais parallèlement, la prime de l'XS sur la seule branche incendie est passée à 93200, soit une augmentation d'à peu près 1100 qui est donc due aux aléas de la simulation. Nous ne pouvons donc tirer **aucune conclusion**, puisque les variations dues à la simulation sont plus importantes que l'impact de la dépendance elle-même. Pour observer cet impact, il aurait été nécessaire de multiplier les montants simulés dans la branche automobile par un facteur 20 minimum.

Cela reste néanmoins riche d'enseignements, dans la mesure où, si l'on veut s'intéresser à l'impact de la dépendance sur une prime de réassurance, une première étude de la prime pure empirique s'avère nécessaire.

En outre, nous aurions pu étudier directement le comportement de la loi marginale du total, sans passer au préalable par la théorie des copulas.

En fait, la structure de dépendance joue pleinement son rôle si l'on veut étudier des contrats "spéciaux", comme ceci fut le cas pour *Frees-Valdez*[2], où le paiement d'un excédent sur une branche est conditionné par le montant dans l'autre branche.

4.2.2 Etude d'un excédent de sinistre automobile dont le paiement est conditionné par la branche incendie.

Le principe de ce contrat est des plus simples. Nous allons simplement évaluer la prime annuelle d'un contrat qui paye l'excédent de sinistres automobiles uniquement si, parallelement, l'événement tempête a généré un montant supérieur à la priorité dans la branche incendie.
Nous allons donc fixer un plafond et priorité dans chaque branche et évaluer la prime pour ce type de contrat. Nous n'utiliserons pour ce faire que les sinistres de type ❷, avec une fréquence suivant toujours une loi de Poisson de paramètre λ_2.
A titre indicatif, et pour mieux mettre en valeur la nécessité d'utiliser une distribution bivariée, nous allons donner les résultats obtenus si nous avions supposé l'indépendance entre les deux branches.

	Copula HRT a=1,365			
	Priorité Auto	Plafond Auto	Prime annuelle	Estimation de la fréquence de dépassement de la priorité.
Automobile	3220	32000	1010	0,11
Incendie	111000			0,2502

Par le vocable "estimation de la fréquence de dépassement de la priorité" nous entendons le nombre moyen d'années au cours de laquelle le réassureur va devoir effectuer un paiement sur ce type de contrat. Sur 25% des années simulées, il y a eu un événement dépassant la priorité de 111 000 KF en incendie, et seuleument 11% des années ont vu un paiement effectué par le réassureur pour le contrat automobile.
La valeur théorique du τ de Kendall entre les deux branches est de 0,268. Nous noterons cet exemple : cas1.

Les résultats obtenus en supposant l'indépendance entre les deux branches (cas 2):

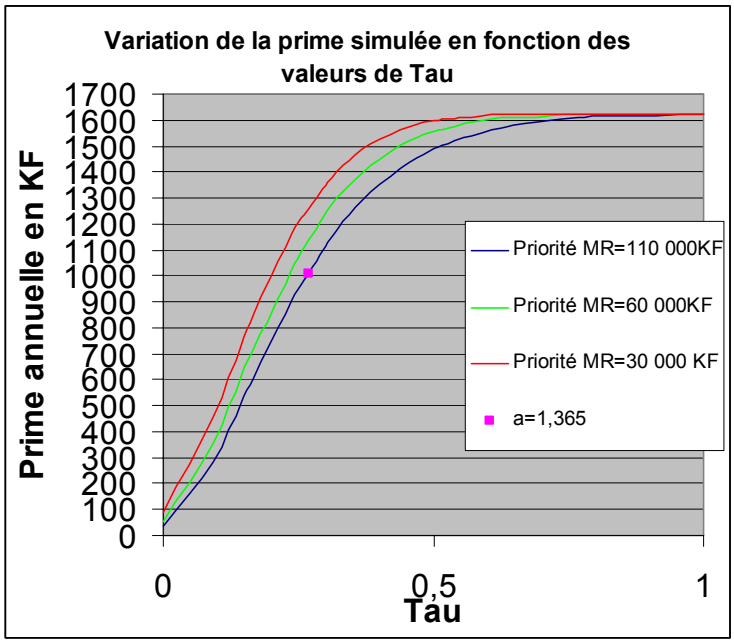
	Indépendance			
	Priorité Auto	Plafond Auto	Prime annuelle	Estimation de la fréquence de dépassement de la priorité.
Automobile	3220	32000	33	0,0064
Incendie	111000			0,2506

Ceci a conduit, comme cela était prévisible, à une réduction considérable et erronée de la prime, et pourtant les lois marginales simulées sont strictement les mêmes.

Nous avons, par curiosité, calculé cette même prime pour des lois marginales reliées par la copula comonotone(τ=1, situation de parfaite dépendance, cas 3): l'application numérique a conduit au résultat de 1621.
La différence absolue de prime est donc plus élevée entre la situation de totale indépendance (cas2) et le cas 1 d'une part, qu'entre le cas1 et le cas3 d'autre part. Cependant, ceci n'est pas conforme à ce que nous aurions pu escompter en nous intéressant à la différence en termes de τ de Kendall. Passer de 0 à 0,26 conduit à une augmentation de la prime équivalente à 1000KF , alors que passer de 0,26 à 1 fait seuleument augmenter la prime de 600KF.
Ceci est du à deux facteurs agissant conjointement:

- Le premier est la structure de dépendance utilisée (copula HRT)
- Le second concerne les priorités définies, relativement élevées.

Or nous avons vu que la copula HRT possède un coefficient de "Right Tail Dependence" qui vaut $1/2^a$. Or, à partir d'une certaine valeur de a et vu la structure de la copula HRT, tous les événements qui excèdent la priorité en incendie vont également conduire à un dépassement de la priorité en automobile. En outre, nous avons choisi nos montants de priorité de sorte à ce qu'ils soient homogènes dans les deux branches.(ie une proba équivalente à 25% qu'au moins un événement auto ou incendie atteigne la priorité chaque année).



Le graphique ci-dessus a été établi en gardant les mêmes paramètres de lois marginales, la même priorité sur la branche automobile (soit 3220KF), la même structure de dépendance, mais nous avons fait simplement varier le paramètre de dépendance de la copula HRT et le montant de la priorité en incendie qui conditionne le paiement de la branche automobile.

Plus nous faisons tendre la priorité en Incendie vers 0, moins la courbure de la courbe sera accentuée. Si nous fixons la priorité en Incendie à 570,il n'y aura plus de conditionnement et la prime en Automobile sera de 1620 quelle que soit la valeur de τ.

4.2.3 Un contrat "particulier"

Nous avons vu que la simulation d'une distribution bivariée pouvait conduire à la tarification de tous les contrats possibles quelles que soient les conditions sur les différentes branches.

Le contrat dont nous allons examiner la prime ici pourra être qualifié d'exotique.

Il s'agit d'un contrat qui rembourse le cumul annuel des excédents de sinistres auto dont la priorité et le plafond sont respectivement fixés à 1000 et 31 600 KF

uniquement si la somme annuelle des sinistres “incendie” supérieurs à 5000KF dépasse le montant de 125 000KF.

La prime annuelle obtenue par simulation pour ce type de contrat à été évaluée à 1937KF.
En supposant l’indépendance entre les deux branches, nous aurions aboutis au résultat de 966KF.

Là encore, la différence varie du simple au double. Ceci s’explique aisément dans la mesure où la survenance de sinistres incendie supérieurs à 5000KF s’accompagne d’une augmentation de la fréquence des sinistres auto supérieurs à 1000KF. Mais le “facteur dépendance” joue un rôle moins important que précedemment (partie 4.2.2), car il est en quelque sorte couplé avec la fréquence annuelle des sinistres.

Conclusion

Nous avons utilisé les copulas pour étudier l'impact des tempêtes sur les branches automobile et incendie d'une grande compagnie d'assurances française.

Nous avons mis en évidence la copula représentant le mieux le phénomène, à savoir la copula HRT, qui n'est rien d'autre que la "Survival copula" de la copula de Clayton. Ceci traduit le fait que les grosses tempêtes conduisent à de montants élevés à la fois en automobile et en incendie, alors que pour les tempêtes de plus faible intensité, nous n'observons pas vraiment de dépendance.

Nous avons également donné au cours de ce mémoire, des techniques nouvelles permettant de sélectionner une ou plusieurs familles de copulas, par l'intermédiaire de comparaison entre des fonctions empiriques et des fonctions paramétriques.
Nous nous sommes limités à cinq copulas, mais cet échantillon est très loin d'être exhaustif.
Les formules des densités des diverses copulas pourront également être utiles au lecteur qui voudrait mener des travaux similaires.

Le véritable problème auquel nous nous sommes retrouvés confrontés, c'est la modélisation en elle-même. Nous avons scindé notre problème en trois parties et utilisé les copulas sur une seule de ces parties. Etait-ce la meilleure solution?
Une autre possibilité aurait consisté à sélectionner les tempêtes de la même manière et à considérer une copula reliant des marginales définies non pas sur l'intersection ($X>dx$ et $Y>dy$) mais sur l'ensemble $[0,\infty[^2$ en considérant les données ($X<dx$ et $Y<dy$) comme censurées. Certes, dans ce cas là, nous nous ramènerions à une seule estimation de paramètres dont la fonction de log-vraisemblance est relativement simple à écrire, mais nous ne pourrions pas passer par les étapes préliminaires de sélection des lois marginales et de choix de la copula. En outre, le fait de se priver des points près de (0,0) est tout à fait préjudiciable puisque un des intérêts des copulas est justement d'étudier la dépendance entre ces points là. Comment faire ainsi, la différence entre une copula de Gumbel et une copula HRT par exemple.

En outre, dans notre modèle,nous avons décidé, de manière somme toute arbitraire, de ne pas tenir compte des tempêtes ayant conduit à des montants inférieurs à dx ou dy Francs.
Nous avons de fait du utiliser des lois tronquées, ce qui a singulièrement compliqué les formules. Nous aurions pu essayer, par exemple, d'effectuer une adéquation de la loi marginale de (X-1000).

Ce problème de choix aurait pu être en partie contourné si nous avions disposé de données de Météo France. En effet, Météo France a une définition propre de la tempête, qui prend en compte l'intensité et la durée du vent. Nous aurions ainsi pu sélectionner uniquement les montants relatifs aux journées de tempêtes telles que Météo France les définit, et nous épargner ainsi d'avoir à choisir des seuils de sélection.

Une autre question que l'on est en droit de se poser concerne l'estimation des divers paramètres. Nous avons vu, que dans un premier temps, nous avons évalué le paramètre de nos copulas sans faire d'hypothèses sur les lois marginales et que parallèlement à cela nous avons estimé les meilleures lois marginales possibles. Puis, par la méthode du maximum de vraisemblance, nous avons estimé conjointement les cinq paramètres. Or, cette méthode conduit à de petites variations dans l'estimation des paramètres des lois marginales.
Nous aurions pu utiliser une méthode intermédiaire, qui aurait consisté à garder les estimations paramétriques obtenues pour les lois marginales. Puis, estimer ensuite le paramètre de la copula par la méthode du maximum de vraisemblance où la contribution de chaque donnée sera donnée par $\ln(c(F_x(x_i),F_y(y_i);a))$. De fait, les paramètres des lois marginales resteront inchangés, et l'estimation du paramètre d'association sera un peu plus exacte.
Comme vous pouvez le constater, les alternatives dans la modélisation et l'estimation des paramètres sont nombreuses.

Par ailleurs, même si nos adéquations à une distribution bivariée sont relativement bonnes, celles-ci sont-elles toujours aussi significatives aux niveaux qui intéressent les réassureurs? *Frees-Valdez (2)* ont bien estimé des primes de réassurance mais les priorités retenues étaient moins élévées que les nôtres .

Mais nous sommes néanmoins tenus de garder un nombre conséquent de données afin que l'utilisation des copulas garde un sens, et que nous simulions ainsi des sinistres qui sont tantôt au dessus, tantôt en dessous de la priorité.Simuler uniquement des événements au dessus des priorités ne présenterait pas très grand intérêt puisque nous ne pourrions plus conditionner le paiement de l'XL d'une branche par l'autre.

Cependant, il reste quelques points dans le travail effectué pour lesquels aucune amélioration ne semble réellement envisageable.
En effet, au vu du nombre de lois testées, il nous semble difficile d'obtenir de meilleures adéquations que celles que nous avons déterminées pour les lois marginales, et le choix de la copula HRT semble être le plus pertinent.
En fait, dans leurs articles, *Klugman-Parsa* sélectionnent les lois marginales mais choisissent la copula de Frank arbitrairement, alors que *Frees-Valdez* ajustent leurs lois marginales sur des Pareto de manière arbitraire, mais utilisent une procédure pour sélectionner la copula de Gumbel. Dans ce rapport, nous avons à la fois cherché à trouver les meilleures lois marginales possibles et la copula la plus adaptée.

Néanmoins, les copulas, même si elles demeurent un outil très puissant n'ont pour l'instant pas souvent été utilisées. Les exemples exposés dans ce mémoire traitent de problèmes bivariés. En théorie, la méthode doit pouvoir être étendue à plus de deux dimensions. Mais, la plupart des articles parus à ce jour se contentent d'utiliser les copulas pour des travaux en deux dimensions. Lorsque le nombre de dimensions est supérieur à deux, les auteurs se ramènent toujours à une copula elliptique (Normale ou de Student) pour la simple et bonne raison que les densités de ces copulas sont "faciles" à calculer et que ces copulas sont aisément simulables. Or, comme nous l'avons maintes fois répété, ce type de copulas n'est pas forcément le plus adapté à l'étude de la sinistralité en assurance. Même si il existe des extensions ,et encore sous certaines conditions, des copulas Archimédiennes à plus de deux dimensions, (cf [9]), établir les dérivées nécessiterait des calculs tout à fait fastidieux.
En outre, l'utilisation des copulas deviendrait quasi-impossible avec des logiciels de bureautique standard comme Excel. Ceci en limite donc le champ d'application possible.

Bibliographie

1 **Gary Venter**: *Tails of copulas.* 2000

2 **Edward Frees-Emiliano A Valdez**: *Understanding relationship using copulas.* 1999

3 **Stuart A Klugman-Rahul Parsa:** *Fitting bivariate loss distributions with copulas. Insurance: Mathematics and economics.* 1998

4 **Roger Nelsen**: *An introduction to copulas.Springer Lecture notes in statistics* 1999

5 **Groupe de recherche opérationnelle Crédit Lyonnais:** *Copulas for finance-A reading guide and some applications.* 2000

6 **Groupe de recherche opérationnelle Crédit Lyonnais:** *Which copula is the right one?* 2000

7 **Alexis Bailly:** *Mémoire magistère d'actuariat Strasbourg – Modelling dependent random variable* 2000

8 **Embrechts-McNeil-Straumann:** *Correlation and dependence in risk management: properties and pitfalls* 1999

9 **Embrechts-McNeil-Lindskog:** *Modelling dependence with copulas and applications to risk management.* 2001

10 **McNeil:** *Multivariate models:theory- Cours magistère d'actuariat Strasbourg* 2001.

11 **P. Nobelis:** *Notes de cours- Magistère d'actuariat 1ère et 2e année.*

12 **A. McNeil:** *Estimating the tails of loss severity distributions using extreme value theory.* 1997

13 **U. Schmock:** *Estimating the value of the wincat coupons of the winterthur insurance convertible bond.* 2001.

14 **D. Foata- A.Fuchs:** *Calcul des probabilités. Masson* 1996

15 **Klugman-Panjer-Willmot:** *Loss models. Wiley-Interscience* 1998

16 **L'Argus:** *Numéros de 1990 à 2000 et dossiers annuels IARD.*

17 **FFSA:** *Diverses études statistiques.*

18 **Descheemaekere-Perron:** *Les événements naturels en France: étude du risque tempête pour une compagnie d'assurances. Mémoire IAF.* 1996

19 **C. Robert:** *Gestion de risques multiples. Présentation.*

20 **Ghoudi:** *Propriétés statistiques des copules de valeurs extrêmes bidimensionnelles.* 1997

21 **S.Wang:** *Aggregation of correlated risk portfolios. Proceedings of the Casualty Actuarial society* 1999

Annexes

Instrat continuous distribution

L'ensemble des lois testées pour les lois marginales fut le suivant:

Lois à un paramètre

Exponentielle

$$F(x) = 1 - e^{-\frac{x}{\lambda}} \qquad f(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$$

Pareto simple

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

Avec $x \geq c$

Lois à deux paramètres

Gamma

$$F(x) = \frac{1}{\Gamma(\beta)} \int_0^{\frac{x}{\theta}} y^{\beta-1} e^{-y} dy \qquad f(x) = \frac{1}{\theta \Gamma(\beta)} \left(\frac{x}{\theta}\right)^{\beta-1} e^{-\frac{x}{\theta}}$$

Weibull

$$F(x) = 1 - e^{-\left(x/\theta\right)^{\beta}} \qquad f(x) = \left(\frac{x}{\theta}\right)^{\beta} \frac{\beta}{x} e^{-\left(x/\theta\right)^{\beta}}$$

Gamma inverse

$$F(x) = 1 - \frac{1}{\Gamma(\beta)} \int_0^{\frac{\theta}{x}} y^{\beta-1} e^{-y} dy \qquad f(x) = \frac{\theta^{\beta} e^{-\frac{\theta}{x}}}{x^{\beta+1} \Gamma(\beta)}$$

Log Gamma

$$F(x) = \frac{1}{\Gamma(\beta)} \int_0^{\alpha \ln\left(\frac{x}{c}\right)} y^{\beta-1} e^{-y} dy \qquad f(x) = \frac{\alpha^{\beta} [\ln(x/c)]^{\beta-1}}{\Gamma(\beta)} \frac{c^{\alpha}}{x^{\alpha+1}}$$

Avec $x \geq c$.

En fait si X suit une loi log-gamma, cela siginifie tout simplement que Y=ln(X/c) suit une loi Gamma.

Weibull Inverse

$$F(x) = e^{-\left(\theta/x\right)^{\beta}} \qquad f(x) = \left(\frac{\theta}{x}\right)^{\beta} \frac{\beta}{x} e^{-\left(\theta/x\right)^{\beta}}$$

Loglogistique

$$F_X(x) = 1 - \frac{1}{1+\left(\frac{x}{\theta}\right)^{\alpha}} \qquad f_X(x) = \frac{\alpha}{\theta} \frac{\left(x/\theta\right)^{\alpha-1}}{\left(1+\left(x/\theta\right)^{\alpha}\right)^2}$$

Paralogistique

$$F_X(x) = 1 - \frac{1}{\left(1+\left(\frac{x}{\theta}\right)^{\sqrt{\alpha}}\right)^{\sqrt{\alpha}}} \qquad f_X(x) = \frac{\alpha}{\theta} \frac{\left(\frac{x}{\theta}\right)^{\sqrt{\alpha}-1}}{\left(1+\left(\frac{x}{\theta}\right)^{\sqrt{\alpha}}\right)^{\sqrt{\alpha}+1}}$$

Paralogistique inverse

$$F_X(x) = \frac{1}{\left(1+\left(\frac{\theta}{x}\right)^{\alpha}\right)^{\alpha}} \qquad f_X(x) = \frac{\alpha^2}{\theta} \frac{\left(x/\theta\right)^{\alpha^2-1}}{\left(1+\left(\theta/x\right)^{\alpha}\right)^{\alpha+1}}$$

Normale

En notant

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{x^2}{2}\right) dx$$

alors

$$F(x) = \phi\left(\frac{x-\mu}{\sigma}\right) \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

Log-Normale

$$F(x) = \phi\left(\frac{\ln(x)-\mu}{\sigma}\right) \qquad f(x) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{x} \exp\left(-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\sigma}\right)^2\right)$$

Lois à trois paramètres

Burr

$$F_X(x) = 1 - \frac{1}{\left(1+\left(\frac{x}{\theta}\right)^{\beta}\right)^{\frac{\alpha}{\beta}}} \qquad f_X(x) = \frac{\alpha}{\beta}\frac{\left(\frac{x}{\theta}\right)^{\beta-1}}{\left(1+\left(\frac{x}{\theta}\right)^{\beta}\right)^{\frac{\alpha}{\beta}+1}}$$

Burr Inverse

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

La fonction J(z)

Pour une copula ayant pour fonction de distribution C(u,v), nous définissons:

$$I(z) = \int_0^z \int_0^z C(u,v)c(u,v)dvdu$$

Alors J(z) peut-être définie par

J(z)=4I(z)/C(z,z)²-1

Nous allons donner les formes analytiques de 4I(z) pour les copulas de Gumbel et HRT.

Gumbel

$$\left(2-\frac{1}{a}\right)\exp\left[2^{1+\frac{1}{a}}\ln(z)\right]-4(-\ln(z))^{a}\left(1-\frac{1}{a}\right)\int_{-2^{\frac{1}{a}}\ln(z)}^{\infty} e^{-2w}w^{-a}dw$$

HRT (Heavy right Tail)

$$8z-8+4(2y-1)^{-a}+\frac{\left[4a(1-z)^2+2\left(1+(2y-1)^{-2a}\right)(a+1)\right]}{2a+1}+8a\int_1^y (w+y-1)^{-a-1}w^{-a}dw$$

avec $y=(1-z)^{-1/a}$.