



ÉCOLE
D'INGÉNIEURS
PARIS-LA DÉFENSE

Risques extrêmes et dépendance entre risque

Anir Ben Chabane

Le but de ce rapport est de présenter les résultats obtenus sur des exercices abordant les différentes techniques qui interviennent dans le cadre de la quantification des risques, où nous allons plus précisément essayer de comprendre l'impact des événements rares sur les mesures de ces risques.

La résolution de cet exercice a été implémentée sur python, deux script « `exo_9.py` » et « `exo_16.py` », et je fais référence dans mon rapport afin de faire un rapprochement avec le raisonnement décrit et ce qui est recherché.

NB : Lorsque je mentionne (Cf l.1 :10 `exo_9.py`), il faut se référer à la pièce jointe qui correspond au code python de la ligne 1 à la ligne 10 de fichier source « `exo_9.py` »

Table des matières

I – Exercice N°1 - TD 09/04/17	4
Question 1	4
a) Disposition des histogrammes des échantillons :	4
b) Description statistique des échantillons :	5
c) Construction de la courbe de concentration.....	5
d) Estimation du paramètre ξ de la loi des extrêmes généralisée (GEV)	7
e) Estimation du paramètre ξ à l'aide de l'estimateur Pickand :.....	8
Question 2	10
a) Identification des lois	10
b) Qqplots par échantillon :	11
II – Exercice N°1 - TD 16/04/17	13
Question 1	13
a) Détermination de la loi de cout individuelle – (l.22 exo_16.py)	13
b) Estimation de paramètres & QQ-plots – (cf l.33 :124 exo_16.py)	13
Question 2 – (cf l.127 : 168 exo_16.py).....	15
II – Exercice N°2 - TD 16/04/17	16
Question 1 – (cf l.191 :221 exo_16.py).....	16
Question 2 – (cf l.222 :242 exo_16.py).....	16
Question 3 – (cf l.253 :269 exo_16.py).....	17
Question 4 – (cf l.277 :362 exo_16.py).....	17

I – Exercice N°1 - TD 09/04/17

Nous disposons de deux échantillons de taille $n=20\,000$, chacun suivant une loi de probabilité différente.

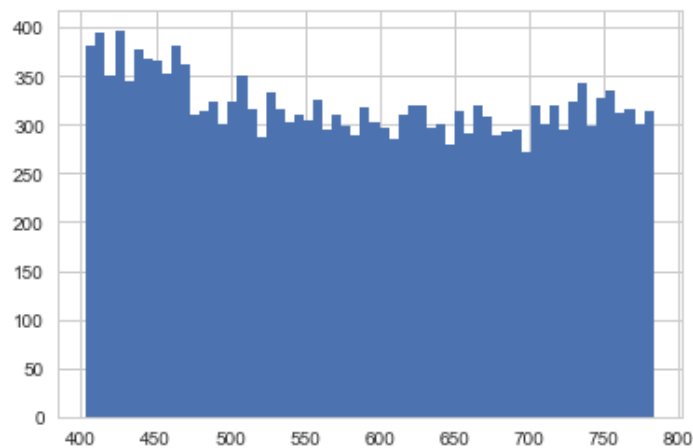
- Echantillon C
- Echantillon D

Question 1

a) Disposition des histogrammes des échantillons :

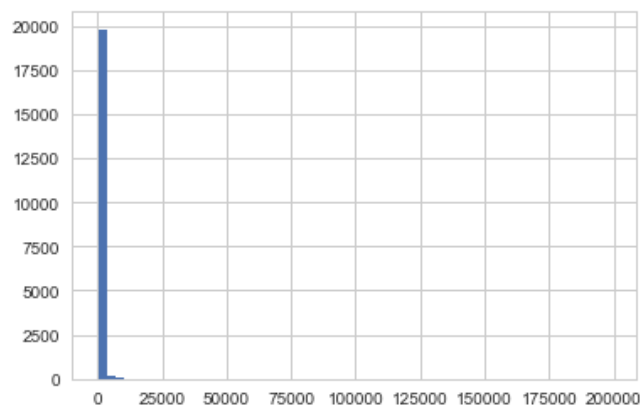
- Echantillon C (*cf l.28 exo_9.py*)

Histogramme avec $K=60$ quantiles, en considérant toutes les valeurs de l'échantillon. Notons la distribution à tendance équitable.



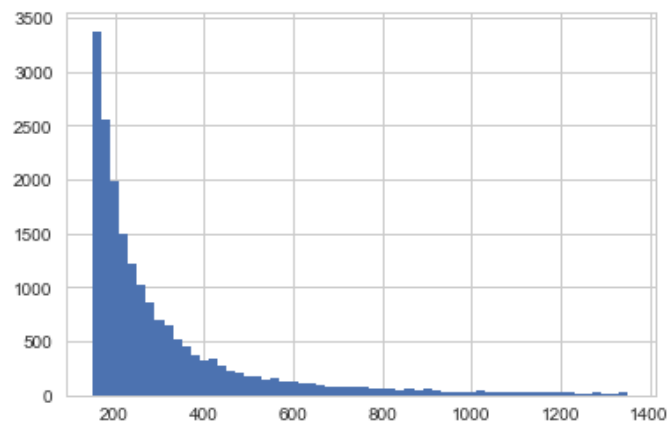
- Echantillon D

$K=60$, toutes les lignes sont considérées pour graphe de l'historgramme : On obtient une très mauvaise visualisation, expliqué par la présence de valeurs extrêmes dans l'échantillon



- Echantillon D, **n = 19200**

Reprenons l'échantillon D, mais cette fois, nous n'allons pas visualiser $n = 20\,000$ lignes, mais les $n'' = 19200$ premières valeurs de l'échantillon ordonné. On obtient une visualisation un peu plus remarquable.



b) Description statistique des échantillons :

Mean = moyenne, std = écart-type

	count	mean	std	min	25%	50%	75%	max	echantillon
value	20000	461.179	2321.94	150.002	182.012	238.978	378.752	198918	d
value	20000	595.532	116.599	403.498	491.633	593.123	698.328	798.424	c

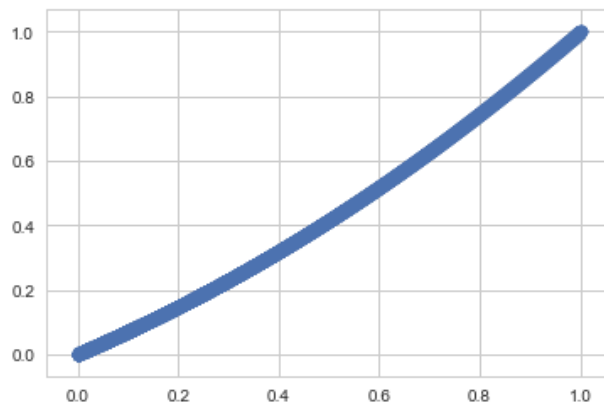
La nature de la variance et de la répartition des valeurs au sein des 3 premiers quartiles caractérisent ces échantillons.

c) Construction de la courbe de concentration

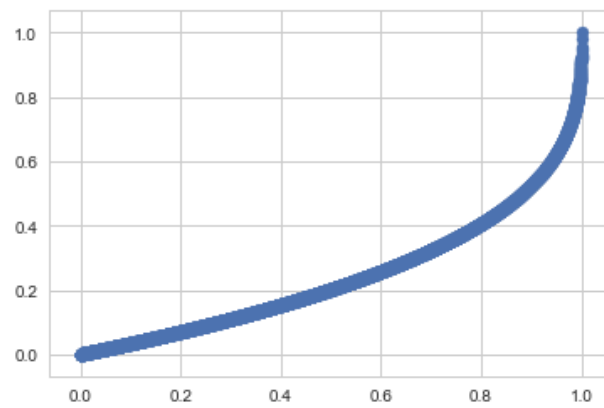
La courbe de concentration permet de donner une autre forme d'observation de la distribution de population dans un échantillon. On en déduit l'indice de Gini

- 0 – *égalité totale de distribution*
- 1 – *inégalité totale de la distribution*

- Courbe de concentration échantillon C - (cf l.43 exo_9.py)
Indice de Gini = 0.113



- Courbe de concentration échantillon D – (cf l.70 exo_9.py)
Indice de Gini = 0.51



Plus la courbe est proche de $y=x$, plus la distribution est égale

d) Estimation du paramètre ξ de la loi des extrêmes généralisée (GEV)

$$G\xi(x) = \exp(-(1 + \xi x)^{-1/\xi}), x \in \mathbb{R}$$

La démarche à suivre est d'estimer ce paramètre pour chacun de nos échantillons, pour cela, il faudra découper l'échantillon en k classes de taille m , puis on note cet échantillon par $M(k)$.

On prendra $k = 200$ et $m = 100$. Le but est de construire le diagramme de probabilité de retour. On retiendra alors dans $M(k)$ les valeurs maximales par classe k

Nous observerons ces diagrammes pour déduire le domaine d'attraction des maximums des deux échantillons. Si le diagramme est :

- Concave, $X \sim \text{Weibull}$
- Convexe, $X \sim \text{Frechet}$
- Sinon, $X \sim \text{Gumbel}$

Ces lois sont des cas particuliers de la loi d'extremum généralisée. Cela nous simplifie en pratique l'estimation des paramètres de la loi pour l'ajuster au mieux selon l'échantillon.

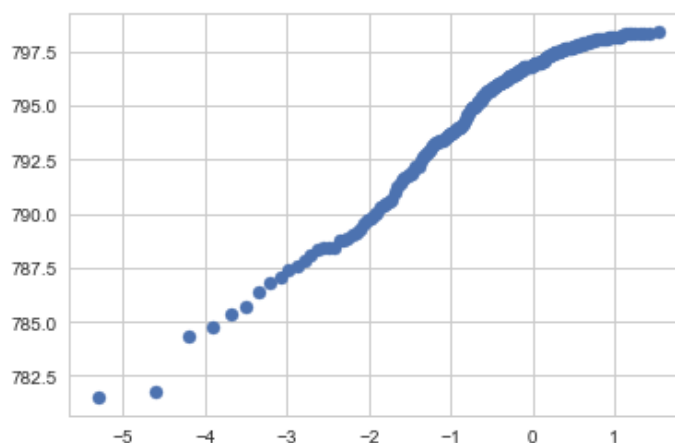
Expression mathématique du diagramme :

$$(\ln(y_u), \hat{Q}_k(u))$$

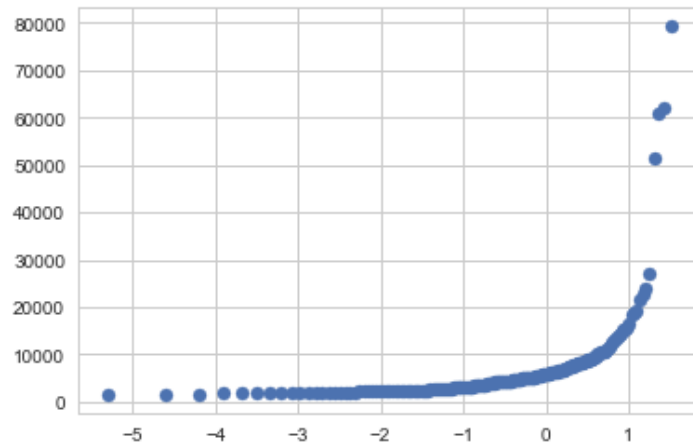
où $y_u = -\ln(1 - u)$ avec $u = i/(k + 1)$ pour $i = 1, 2, \dots, k$.

Où $\hat{Q}_k(u)$ est en fait l'échantillon et $u = X \in \text{échantillon C ou D}$.

- Diagramme de probabilité de retour échantillon C - (cf l.76 exo_9.py)
Echantillon C concave, $C \sim \text{Weibull}$



- Diagramme de probabilité de retour échantillon D - (cf l.87 exo_9.py)
Echantillon D convexe, $D \sim \text{Fréchet}$



e) Estimation du paramètre ξ à l'aide de l'estimateur Pickand :

On a notre échantillon $M(k)$ = maximum de chaque classe k ,

- $R(i,k) = 1/\ln(2) * \ln (M(k-i+1:k) - M(k-2i+1:k) / M(k-2i+1:k) - M(k-4i+1:k))$, pour $i = 1 \dots n/4$,
 $k=200$ – (cf l.101 exo_9.py)

Le vecteur $R(i,k)$ va donner une représentation graphique qui va nous permettre de fixer arbitrairement un ξ

- $(i, R(i,k))$ Echantillon C



- $(i, R(i, k))$ Echantillon D



On peut estimer qu'on atteint le paramètre ξ lorsque les valeurs se stabilisent.

L'échantillon **C** est autour de $\xi = -1$, tandis que l'échantillon **D** est autour de $\xi=0.7$

Question 2

a) Identification des lois

Pour aboutir à l'Identification de la loi pour chaque échantillon, nous allons essayer d'ajuster certaines loi en donnant des paramètres estimés à l'aide du maximum de vraisemblance, ainsi que la méthode des moments pour estimer les paramètres de la loi Pareto.

La suite est de valider la concordance d'une loi avec un échantillon, pour cela, on va se baser sur l'observation des diagrammes **Quantile(échantillon)-Quantile(loi[paramètres])**, ainsi que la distance **Kolmogorov-Smirnov**

« En [statistiques](#), le test de Kolmogorov-Smirnov est un [test d'hypothèse](#) utilisé pour déterminer si un [échantillon](#) suit bien une loi donnée connue par sa [fonction de répartition](#) continue, ou bien si deux échantillons suivent la même loi ». [wikipedia](#)

- Echantillon C – (cf l.135 :178 exo_9.py)

Loi	Paramètres	Distance KS
Exponentielle	$0.016 = (1/E(X))$	0.49
Normale	(595.5, 161.6) (moy. ,ecart.)	0.0681
Pareto	(6.20, 499.5) (k,Xo)	0.27
Log Normale	(6.37, 0.018)	0.077

- ➔ Pareto : on résout l'équation $E(X)^2 + 2V(X)^k - V(X)^{2k} = 0$, sachant qu'on a l'espérance et la variance empirique. On trouve $k = 6.20$ environs.
Nous avons maintenant le k qui nous permet de résoudre cette seconde équation : $E(X) - k * X_o / (k-1) = 0$, ou l'inconnue est X_o . On trouve $X_o = 499.5$. (On considère $E(X) = \text{moyenne}(\text{échantillon})$, et $V(X) = \text{écart-type}(\text{échantillon})$).
- ➔ Log Normal :

$$\mu = \ln(E(X)) - \frac{1}{2} \ln\left(1 + \frac{\text{Var}(X)}{(E(X))^2}\right),$$
$$\sigma^2 = \ln\left(1 + \frac{\text{Var}(X)}{(E(X))^2}\right).$$

D'où, $\mu = 6.37$ et $\sigma = 0.018$

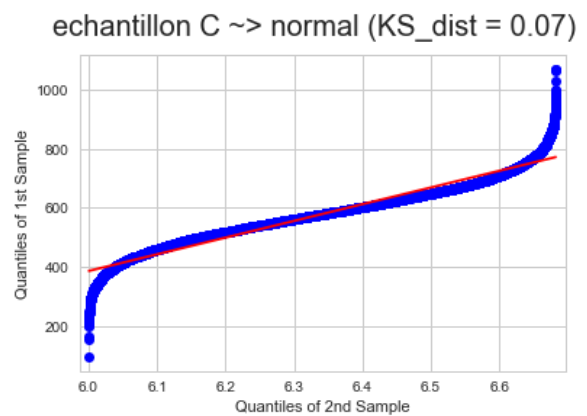
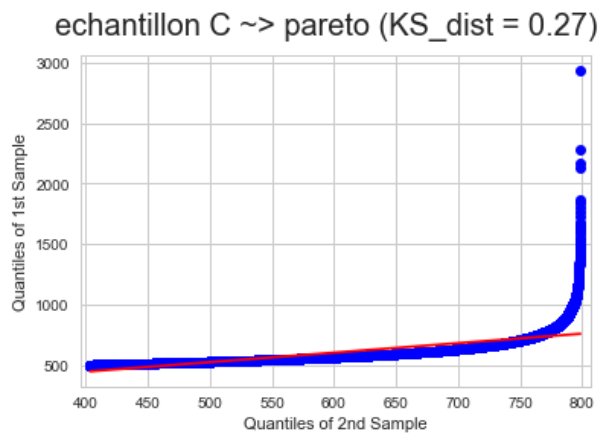
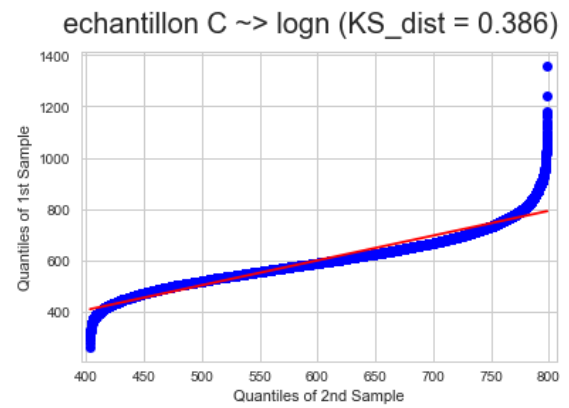
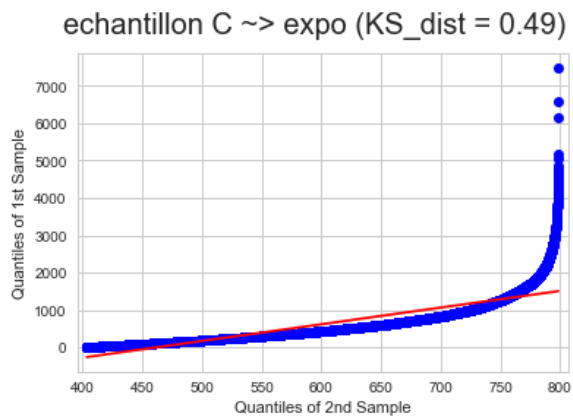
- Echantillon D – (cf l.189 :231 exo_9.py)

Loi	Paramètres	Distance KS
Exponentielle	$0.276 = (1/E(X))$	0.276
Normale	(461.2, 2321.9) (moy. ,ecart.)	0.44
Pareto	(2.02, 232.82) (k,Xo)	0.48
Log Normale	(4.5, 1.64)	0.56

On peut admettre que $C \sim N(595.5, 161.6)$ et que $D \sim \text{exp}(0.276)$.

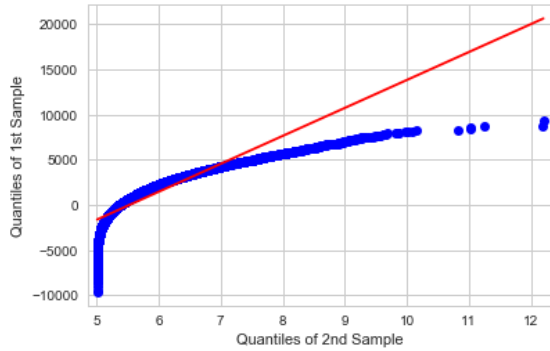
b) Qqplots par échantillon :

- Échantillon C

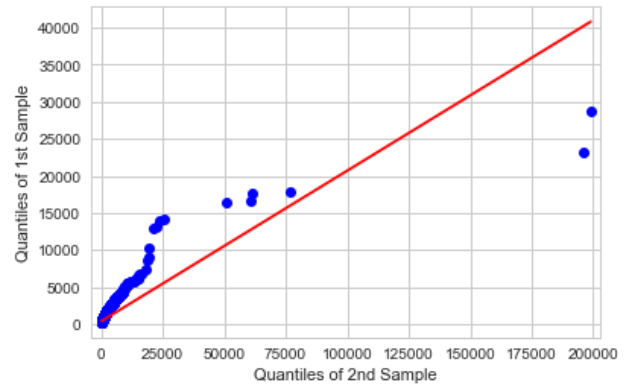


- Echantillon D

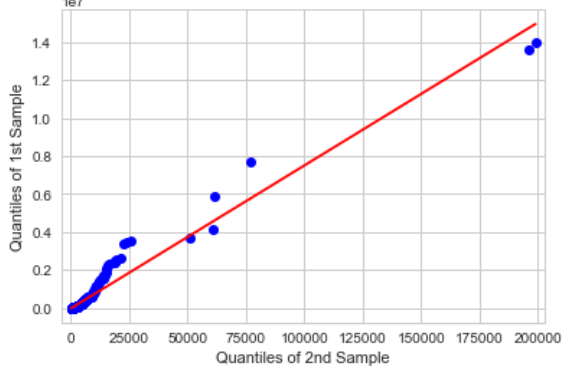
echantillon D ~> normal (KS_dist = 0.44)



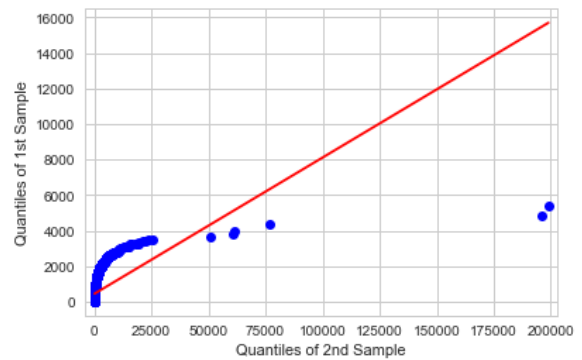
echantillon D ~> pareto (KS_dist = 0.48)



echantillon D ~> lognormal (KS_dist = 0.56)



echantillon D ~> expo (KS_dist = 0.276)



II – Exercice N°1 - TD 16/04/17

Cet exercice traite de la distribution observée du **nombre** et du **coût individuel** de sinistre pour le risque cyclonique.

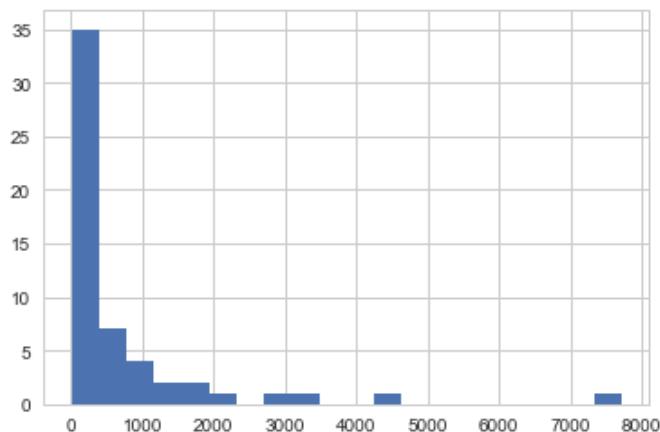
Question 1

a) Détermination de la loi de cout individuelle – (*l.22 exo_16.py*)

Nous allons nous reposer sur la méthode utilisée dans l'exercice 1) question 2) a) du TD - 09/04.

Pour commencer, on dispose l'histogramme des couts de sinistres dans l'échantillon soumis à notre étude.

- Histogramme couts de sinistres échantillon « Coûts cyclones USA »



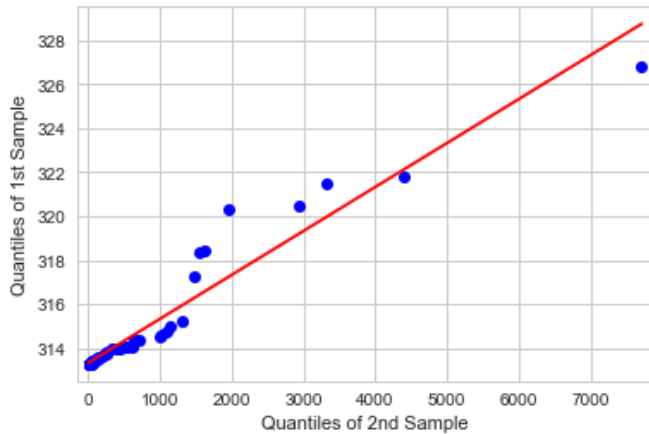
b) Estimation de paramètres & QQ-plots – (*cf l.33 :124 exo_16.py*)

- Loi de référence, paramètres estimés, distance relative à l'échantillon de cout

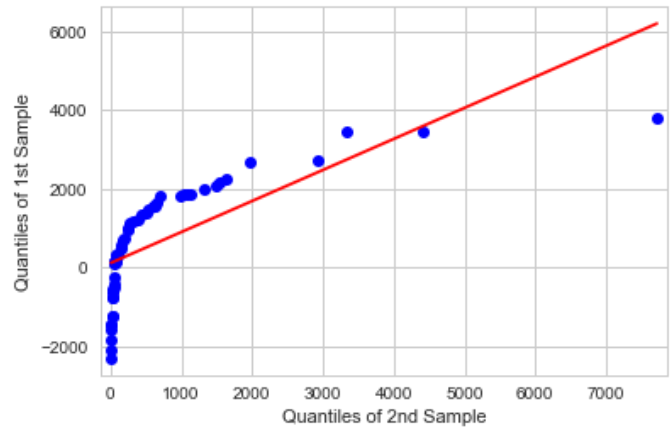
Loi	Paramètres	Distance KS
Exponentielle	0.0145 (1/esp)	0.35
Normale	(689.5, 1311.8) (esp, ec-type)	0.27
Pareto	(2.13, 366.11) (k, x0)	0.54
Log Normale	(5.78 ,1.81) (mu, sigma)	0.14 (p-value = 0.008)
Weibull	(0.48, 313.3) (k, lambda)	0.6
Gamma	(2450.34,0.28) (theta, k)	0.16 (p-value = 0.42)

- QQ-plots relatifs aux lois décrites antérieurement

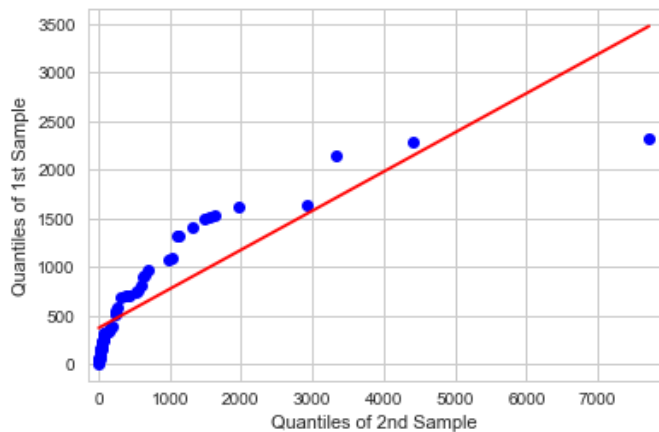
echantillon -> Weibull (KS_dist = 0.6)



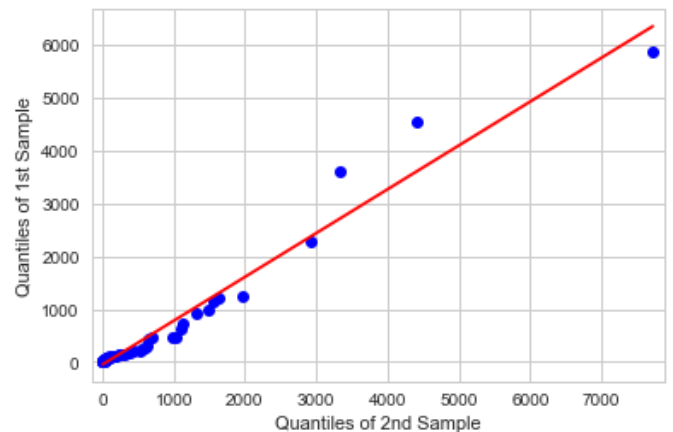
echantillon -> Normal (KS_dist = 0.27)



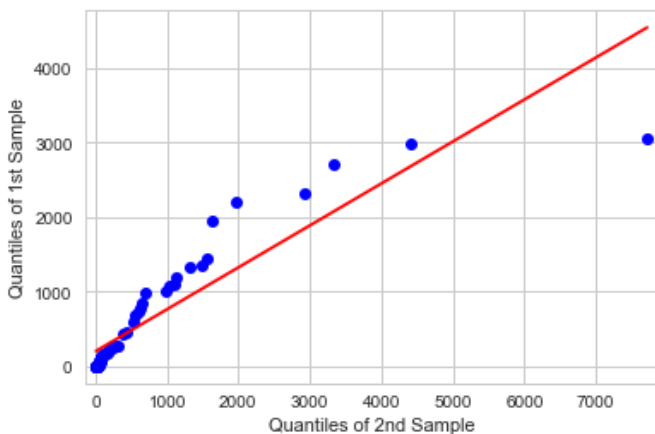
echantillon -> expo (KS_dist = 0.35)



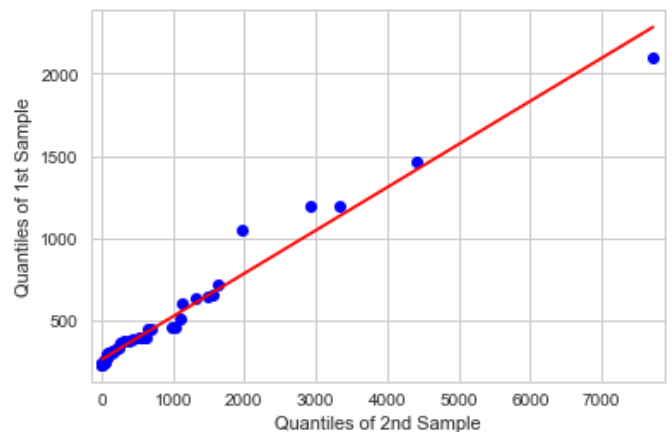
echantillon -> LogNorm (KS_dist = 0.14)



echantillon -> Gamma (KS_dist = 0.16)



echantillon -> pareto (KS_dist = 0.54)



Finalement, les deux lois qui approchent le mieux notre échantillon est la loi Gamma & la loi LogNorm, pour arbitrer celle qu'on retient on s'appuie sur la distance KS, et la p-value de la distance.

La distance de la loi LogNorm est plus basse que celle que présente la loi Gamma, par contre la p-value du test KS de la loi Gamma est beaucoup plus élevée que celle de LogNorm ($0.42 > 0.008$).

Malgré la meilleure disposition du QQ-plot de la loi LogNormal, je retiendrais la loi Gamma qui présente de bons résultats au niveau du test KS.

Question 2 – (cf l.127 : 168 exo_16.py)

- Simulation de la loi de sinistralité sur 10 ans :
- Nombre annuel de cyclones est ajusté par la loi de poisson $\lambda = 1.7$
- Cout sinistre individuel ajusté par la loi Gamma, $\theta = 2450.30$ et $k = 0.28$

La démarche à suivre est de simuler un nombre de sinistre par année, et puis pour chaque sinistre X on va simuler le cout de chacun de ces sinistre en utilisant la loi retenue à la question précédente, et dresser les coûts individuels de chaque sinistre, et les couts cumulés par période.

	annee	cout_unit	cout cumule
0	1950	167.887	182.575
1	1950	14.688	182.575
2	1951	167.522	4073.23
3	1951	13.306	4073.23
4	1951	1162.17	4073.23
5	1951	656.978	4073.23
6	1951	2073.25	4073.23
7	1953	9.17621	9.17621
8	1954	336.791	336.791
9	1955	0.0131547	0.0131547
10	1956	15.8105	153.784
11	1956	137.974	153.784
12	1957	503.93	510.208
13	1957	6.27823	510.208
14	1958	1.86449	502.015
15	1958	484.951	502.015
16	1958	15.1998	502.015
17	1959	40.1258	40.1258

II – Exercice N°2 - TD 16/04/17

- Des études antérieures ont montré que N sinistres sur une période donnée $\sim \text{geom}(p=0.1)$
- Cout individuel du sinistre $C \sim \text{expo}(\alpha = 0.01)$
- Montants cumulés des sinistres est noté X

Question 1 – (cf l.191 :221 exo_16.py)

Le but est de calculer l'espérance et l'écart type de X en utilisant les formules du modèle collectif, et retrouver ces valeurs par simulation de la loi exponentielle ($\lambda = 0.01$)

- $E(X) = 1/(\alpha \cdot p)$, $\sigma(X) = E(X) = 1000$
- Moyenne(X) = 995.04, Ecart-Type = 1008.16

Loi de X	N	err abs esperance	err rel esperance	erreur abs sigma	err rel sigma	confiance
X ~> expo	10000	4.95796	-0.00495796	8.16071	0.00816071	['1014.80500809', '975.279073599']

Question 2 – (cf l.222 :242 exo_16.py)

L'assureur souhaite souscrire un traité de réassurance « stop-loss », dont le principe est de se débarrasser des sinistres à un cout supérieures à C , où C est la valeur du quantile à 95%, ou autrement dit, $P(X > C) = 0.05$

Nous allons nous appuyer sur l'échantillon simulé dans la question précédente :

- $X \rightarrow \text{expo}(\lambda=0.01)$, X valeur du cout de sinistre cumulé par période.
- L'échantillon contient 10 000 valeurs simulées.
- $C = 2976.6$

Nous reprenons notre échantillon, et on retient pour tout X_i , la valeur $\max(0, X_i - C)$ (on se débarrasse des valeurs $> C$)

- La prime pure de l'échantillon net des valeurs $X_i > C$ est : 52.71

esperance	écart type	confiance
52.7079	333.152	['59.2387192036', '46.1771576351']

On donne ici une estimation de la prime pure, et un intervalle de confiance à 95% de celle-ci.

Question 3 – (cf l.253 :269 exo_16.py)

L'assureur s'intéresse au montant du plus gros sinistre sur une période, on va alors construire un échantillon qui va simuler :

- ➔ Nombre de sinistre par période
- ➔ Cout individuel de sinistre sur chaque période
- ➔ On retient dans notre échantillon le cout le plus élevé de sinistre sur nos périodes

L'objectif est d'approcher son espérance mathématique et de construire l'intervalle de confiance lié à l'échantillon de l'extremum des sinistres.

esperance	écart type	confiance
253.336	148.199	['256.241291222', '250.430993262']

Question 4 – (cf l.277 :362 exo_16.py)

L'objectif de cette question est de changer la loi de cout de sinistres, on passe de la loi exponentielle ($\lambda = 0.01$) à Pareto ($k=5$, $\min=80$)

La démarche à suivre est de reprendre ce qui a été dans la question 1),2) et 3).

- Etape 1 : Simulation des couts de sinistres individuels à l'aide de pareto
 - $E(X)$ théorique = 1000
 - Moyenne(X) = 990.01

Loi de X	N	err abs esperance	err rel esperance	erreur abs sigma	err rel sigma	confiance
X ~> pareto(a=5,x0=80)	10000	9.99194	-0.00999194	0.403144	0.000423385	['1008.68174119', '971.334370177']

- Etape 2 : Prime pure avec « stop-loss », et intervalle de confiance
 - Valeur du « stop-loss » $C = 2885.60$

esperance	écart type	confiance
49.1756	311.456	['55.2810919299', '43.0701511697']

- Etape 3 : Etude des valeurs extrêmes de l'échantillon simulé dans l'étape 1

esperance	écart type	confiance
254.553	150.422	['257.501337978', '251.60387794']