OS13 : Test d'estimation et loi extrême des lois de Cauchy et géométrique

Adrien WARTELLE
TRAN Quoc Nhat Han

3 octobre 2018

Sommaire

Résumé

Ce rapport est à montrer l'application de certaines tests statistiques pour la loi géométrique et la loi de Cauchy.

1 Loi de Cauchy

1.1 Rappel

Soit f_X la fonction de densité de Cauchy de deux paramètres x_0 et a ($a > 0$), définie par :

$$f_X(x) = \frac{1}{\pi a \left(1 + \left(\frac{x - x_0}{a}\right)^2\right)} = \frac{1}{\pi} \frac{a}{(x - x_0)^2 + a} \tag{1.1}$$

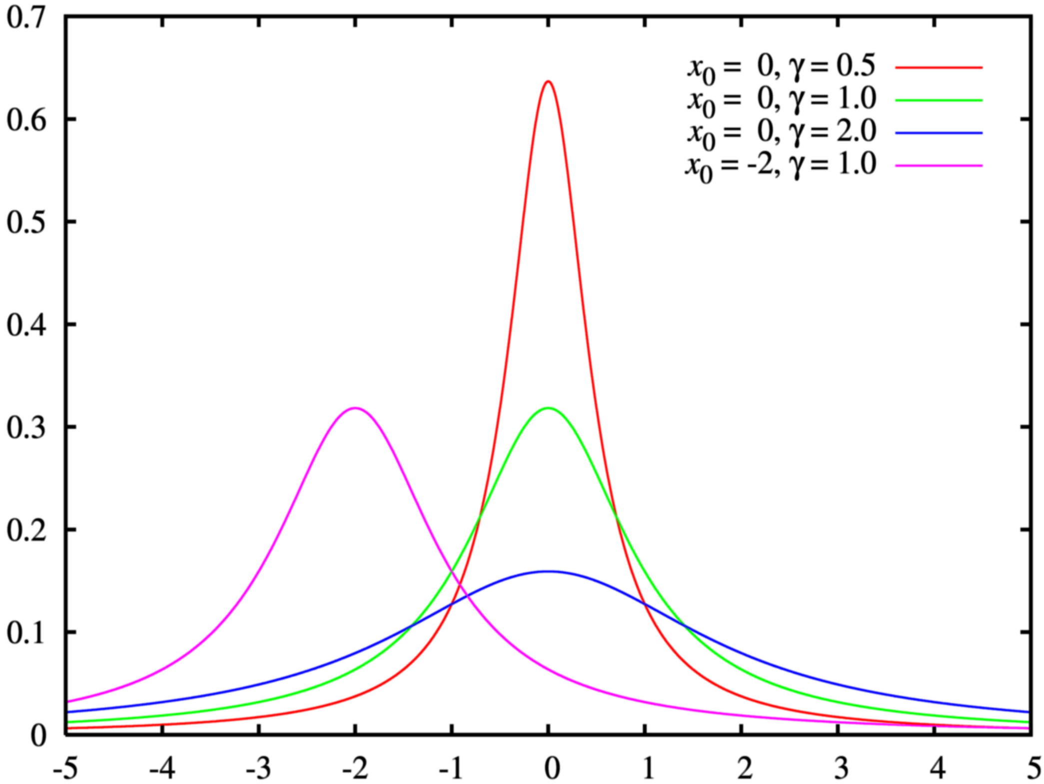

FIGURE 1 – Distribution théorique de la loi Cauchy. Source : [1]

a est dite l'échelle de la fonction, et x_0 est son médian.
La loi de Cauchy n'admet ni espérance ni écart type.
La fonction de répartition :

$$F_X(x) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{a}\right) + \frac{1}{2} \tag{1.2}$$

1.2 Estimer les paramètres inconnus

Calcul théorique

Soient n réalisations $x_1, x_2, ..., x_n$. Assumons que ces données suivent la loi de Cauchy (1.1).

Nous allons utiliser la méthode de maximum de rapport de vraisemblance.

En l'absence d'information de la distribution, nous assumons que ces mesures sont indépendants. Posons une variable aléatoire X_i correspondante à chaque réalisation x_i pour $i = \overline{1, n}$.

On écrit la loi conjointe de ces n variables :

$$
\begin{aligned}
L(X) = L\left(X_1, X_2, ..., X_n\right) &= \prod_{i=1}^{n} f_X\left(X_i\right) \\
&= \prod_{i=1}^{n} \frac{1}{\pi} \frac{a}{\left(x_i - x_0\right)^2 + a} \\
&= \pi^{-n} \prod_{i=1}^{n} \frac{a}{\left(x_i - x_0\right)^2 + a}
\end{aligned}
$$

On cherche l'optimum maximale.

$$
\begin{aligned}
\frac{\partial L}{\partial x_0}(X) &= \pi^{-n} \sum_{j=1}^{n} \left(- \frac{a\left(2x_j - 2x_0\right)}{\left[\left(x_0 - x_j\right)^2 + a\right]^2} \prod_{i=1; i \neq j}^{n} \frac{a}{\left(x_i - x_0\right)^2 + a} \right) \\
&= \pi^{-n} \left(\prod_{i=1}^{n} \frac{a}{\left(x_i - x_0\right)^2 + a} \right) \left(- \sum_{j=1}^{n} \frac{2x_0 - 2x_j}{\left(x_0 - x_j\right)^2 + a} \right) \\
\frac{\partial L}{\partial x_0}(X) = 0 &\Leftrightarrow \sum_{j=1}^{n} \frac{x_0 - x_j}{\left(x_0 - x_j\right)^2 + a} = 0 : \text{insolvable par la main}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial L}{\partial a}(X) &= \pi^{-n} \sum_{j=1}^{n} \left(\frac{\left(x_j - x_0\right)^2 + a - a}{\left(x_j - x_0\right)^2 + a} \prod_{i=1; i \neq j}^{n} \frac{a}{\left(x_i - x_0\right)^2 + a} \right) \\
&= \pi^{-n} a^{n-1} \frac{\sum\limits_{j=1}^{n} \left(x_0 - x_j\right)^2}{\prod\limits_{i=1}^{n} \left(\left(x_i - x_0\right)^2 + a\right)} > 0 \forall a > 0
\end{aligned}
$$

Résolution par R

On génère une échantillon avec a et x_0 au choix. Puis, utiliser la fonction *mle* (Maximum Likelihood Estimator) de librairie *stats4* avec 2 valeurs initiales $\hat{a}$ et $\hat{x_0}$ pour estimer a et x_0.

Nous essayons d'estimer x_0 et a directment. L'algorithme d'approximation implémenté dans R a besoin un bon point de départ, sinon le résultat obtenu variera grossièrement.

— Etant donné que la médiane est théoriquement aussi x_0, choissisons x_0 comme la médiane de l'échantillon.

— Nous avons $f_X(x_0) = \frac{1}{\pi a}$. En plus, $f_X(x_0)$ est la valeur maximale d de densité. Prenons alors $\hat{a} = \frac{1}{\pi d}$.

Testons avec $x_0 = 13$ et $a = 0.5$.

```
# fixer le gemme et paramètres réels
set.seed(2018)
real_x0 = 13
real_a = 0.5
nbreaks = 40

# générér échantillons
N = 1000
x = rcauchy(n = N,
            location = real_x0,
            scale = real_a)

# définir fonction log négative pour MLE
ll = function(location, scale) {
    dist = suppressWarnings(dcauchy(x,
                                    location,
                                    scale,
                                    log = TRUE))
    -sum(dist)
}

# MLE (Maximum Likelihood Estimator)
library(stats4)
hat_x0 = median(x)
d = hist(x,
         breaks = nbreaks,
         plot = FALSE) # histogram
hat_a = 1 / (pi * max(d[["density"]]))
result = mle(ll,
             start = list(location = hat_x0, scale = hat_a))

# illustrer
hist(x,
     freq = F,
     breaks = nbreaks,
     col = "green",
     xlab = "x",
     ylab = "Densité",
     ylim = c(0, 0.2),
     main = "Loi de Cauchy - réalité vs théorique") # histogram
curve(dcauchy(x, real_x0, real_a),
      add = TRUE,
```

```
      col = "black") # theoretic
curve(dcauchy(x, result@coef[1], result@coef[2]),
      add = TRUE,
      col = "red") # calculated
legend("topright",
       legend = c("histogramme", "théorique", "réalité"),
       fill = c("green", "black", "red"))
```

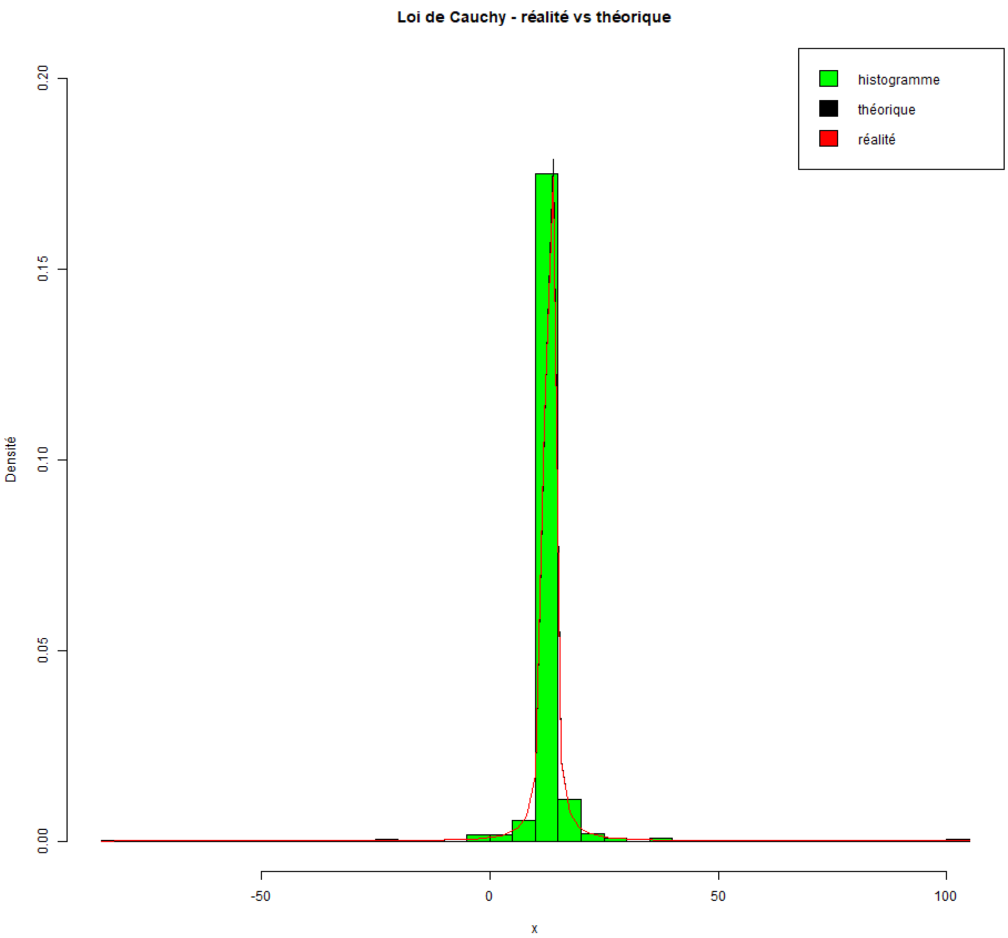

FIGURE 2 – Représentation de l'histogramme et de la courbe de densité

Pour $x_0 = 13$ et $a = 0.5$, les valeurs trouvées par R sont $\hat{x_0} = 12.9877139$ et $\hat{a} = 0.4951073$.

1.3 Test de paramètres

Nous testons ici si les valeurs trouvées par R sont vraiment approchées à celles théoriques. Pour l'instant, nous n'avons pas d'outils pour vérifier 2 variables en couple.

1.4 Test d'adéquation

R nous donne la fonction *ks.test* pour valider la compabilité d'une échantillon avec une loi selon le test de Kolmogorov-Smirnov.

```
ks.test(x, "pcauchy", real_x0, real_a)
```

On a trouvé $D = 0.02696$, $p-value = 0.4614$, signifiant l'écart maximale est 0.02696 et le niveau d'acceptance est 0.4614. Vu que nous fixons $\alpha = 5\% < p - value$, l'échantillon dépasse largement le test. Autrement dit, il se distribue selon la loi de Cauchy.

1.5 Etude de biais d'estimateur

Maintenant on s'intéresse au biais d'estimateur. On relance l'algorithme en-dessus avec de différents observations ($N = 100, 125, 150, ..., 1100$). $x_0 = 13$ et $a = 0.5$.

```
library(stats4)
# définir fonction log négative pour MLE
ll = function(location, scale) {
    dist = suppressWarnings(dcauchy(x,
                                    location,
                                    scale,
                                    log = TRUE))
    - sum(dist)
}

# fixer les paramètres réels
set.seed(2018)
real_x0 = 13
real_a = 0.5
nbreaks = 40
minN = 100
numN = 40
delta = 25
maxN = minN + (numN - 1) * delta

Ns = seq.int(minN, maxN, delta) # tous 40 valeurs de n de 100 à 1000
results = array(NA, c(2, numN)) # à contenir tous les résultats

for (i in 1:numN) {
    N = minN + (i - 1) * delta

    # générér échantillon
    x = rcauchy(n = N,
                location = real_x0,
                scale = real_a)

    # MLE (Maximum Likelihood Estimator)
    hat_x0 = median(x)
    d = hist(x,
             breaks = nbreaks,
```

```r
              plot = FALSE) # histogram
    hat_a = 1 / (pi * max(d[["density"]]))
    result = mle(ll,
              start = list(location = hat_x0, scale = hat_a))

    # récupérer le résultat
    results[1, i] = result@coef[1] # x_0
    results[2, i] = result@coef[2] # a
}

# illustrer
plot(x = results[1, ],
     y = results[2, ],
     col = rgb(0, 100, 0, 70, maxColorValue = 255),
     pch = 16,
     cex = 1.5,
     main = "Convergence d'estimateurs",
     xlab = "Location x_0",
     ylab = "Echelle a"
     )
points(x = real_x0,
     y = real_a,
     col = "red",
     pch = 16,
     cex = 2
     )

legend("topleft",
       legend = c("valeur théorique", "valeur estimée"),
       fill = c("red", rgb(0, 100, 0, 70, maxColorValue = 255)))
```

Nous observons que les valeurs estimées se rassemblent assez proches de la valeur théorique.

1.6 Etude de loi d'extremum

Générons donc des échantillons de n valeurs ($n = 10, 100, 1000, etc.$). Parmi chacun, $\frac{n}{10}$ valeurs les plus grandes seront gardées pour étudier la loi d'extremum.

Comme les réalisations sont indépendantes et identiquement distribuées, théoriquement :

$$\begin{aligned} F_{x_{\max}}(x) &= P\left(X_{\max} \leq x\right) \\ &= P\left(X_1 \leq x, ..., X_n \leq x\right) \\ &= \prod_{i=1}^{n} F_X(x) \\ &= F_X(x)^n \end{aligned}$$

Avec $F_X(x) = \frac{1}{\pi} \arctan \frac{x-x_0}{a} + \frac{1}{2}$, la fonction de répartition de la loi de Cauchy. Alors,

$$F_{x_{\max}}(x) = \left(\frac{1}{\pi} \arctan \frac{x - x_0}{a} + \frac{1}{2}\right)^n$$

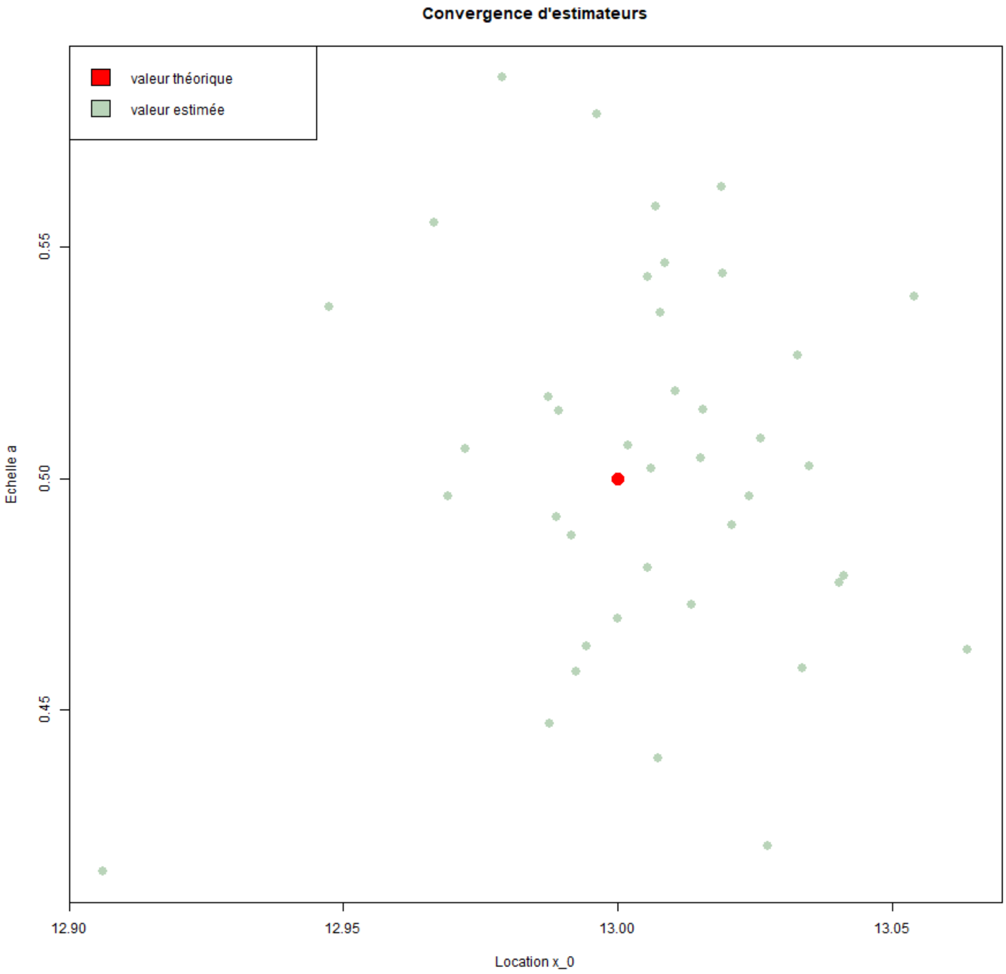

FIGURE 3 – Convergences de $\hat{x_0}$ et $\hat{a}$ vers valeurs théoriques

Or,

$$\left|\arctan\frac{x-x_0}{a}\right| < \frac{\pi}{2} \Rightarrow \frac{-\pi}{2} < \arctan\frac{x-x_0}{a} < \frac{\pi}{2} \Rightarrow 0 < \arctan\frac{x-x_0}{a} + \frac{\pi}{2} < \pi$$

Par conséquence,

$$\begin{aligned}
&\tan\left(\arctan\tfrac{x-x_0}{a} + \tfrac{\pi}{2}\right) = -\cot\left(\arctan\tfrac{x-x_0}{a}\right) = -\tfrac{a}{x-x_0} \\
&\Rightarrow \arctan\tfrac{x-x_0}{a} + \tfrac{\pi}{2} = \arctan\left(-\tfrac{a}{x-x_0}\right) = \arctan\tfrac{a}{x_0-x} \\
&\Rightarrow F_{x_{\max}}(x) = \pi^{-n}\left(\arctan\tfrac{x-x_0}{a} + \tfrac{\pi}{2}\right)^n = \pi^{-n}\arctan^n\tfrac{a}{x_0-x}
\end{aligned}$$

Nous prenons 3 valeurs les plus grandes pour chaque itération, et établissons leur histogramme.

```
# fixer les paramètres réels
set.seed(2018)
real_x0 = 13
real_a = 0.5
minN = 1000
numN = 100
delta = 500
maxN = minN + (numN - 1) * delta
pickupNum = 3
results = rep(NA, 3000) # à contenir tous les résultats
countResults = 0

for (i in 1:numN) {
    N = minN + (i - 1) * delta

    # générér échantillon
    x = rcauchy(n = N,
                location = real_x0,
                scale = real_a)

    # trier
    sort(x, decreasing = TRUE) # dans l'ordre décroissant

    # tirer
    extremes = head(x, pickupNum)
    for (j in 1:length(extremes))
        results[countResults + j] = extremes[j]
    countResults = countResults + length(extremes)
}

# histogramme
hist(results,
     freq = F,
     breaks = 100,
     col = "green",
     xlab = "Extremum",
     ylab = "Densité",
     main = "Loi d'extremum au cas de Cauchy"
     )
```

En fin d'approximer la loi d'extremum, nous ne regardons que les réalisations positives. En utilisant le package *poweRlaw*, nous calculons la compabilité de la loi d'extremum.

```
results = na.omit(results) # filtrer les NA
library(poweRlaw)
positiveResults = results[which(results > 0)] # filtrer les négatifs
cauchy_pl = conpl$new(positiveResults) # objet de calculer la loi d'extremum
est = estimate_xmin(cauchy_pl) # estimer la borne inférieure
cauchy_pl$setXmin(est) # mettre à jour l'objet de distribution
plot(cauchy_pl, col = "green")
lines(cauchy_pl, col = "red")
bs_p = bootstrap_p(cauchy_pl) # tester l'estimation
bs_p$p
```

Calcul de $p - value$ nous retourne 0, qui signifie l'incompabilité de ces deux lois.

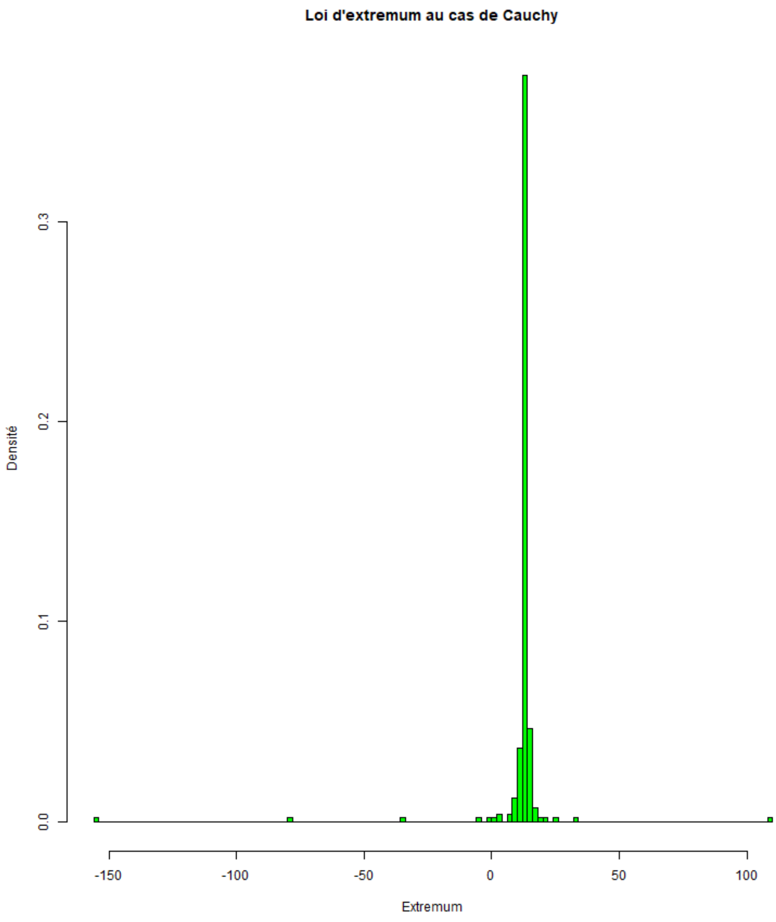

FIGURE 4 – Histogramme de valeurs extremum de la distribution de Cauchy

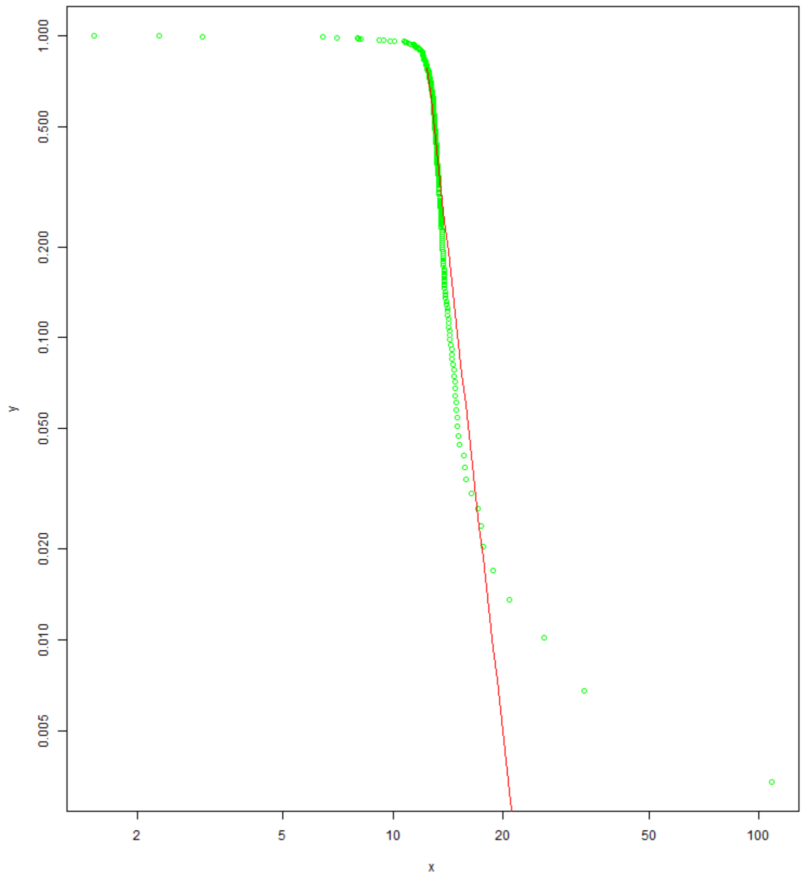

FIGURE 5 – Approcher la loi d'extremum avec celle des maximales de Cauchy. Les données (verte) ne s'approchent pas totalement l'approximation (rouge).

2 Loi géométrique

La loi géomètrique est une loi discrète de distribution utilisé dans le cadre d'un enchainement (non fini) d'épreuves de Bernouilli. Soit X une varaible aléatoire suivant cette loi, on a :

$$P(X = k) = (1 - p)^(k - 1)p \ k \in \mathbb{N}_+$$

La variable X est le numéro de la première épreuve où l'on obtient un succès qui a une probabilité p. Ainsi, p est le seul paramètre de loi (à estimer). La fonction de répartition $F(k) = P(X \leq k) \ k \in \mathbb{N}_+$ est :

$$F(k) = 1 - (1 - p)^k$$

Démonstration. On peut remarquer que :

$$P(X = k) = (1 - p)^{k-1}(1 - (1 - p)) = (1 - p)^{k-1} - (1 - p)^k$$

Donc :

$$F(k) = \sum_{l=1}^{k} P(X = l) = \sum_{l=1}^{k}((1 - p)^{l-1} - (1 - p)^l)$$

$$F(k) = \sum_{l=0}^{k-1}(1 - p)^l - \sum_{l=1}^{k}(1 - p)^l$$

Soit $q = 1 - p$, on a :

$$F(k) = \sum_{l=0}^{k-1} q^l - \sum_{l=1}^{k} q^l$$

$$F(k) = \frac{1 - p^k}{1 - q} - (\frac{1 - q^{k+1}}{1 - q} - 1)$$

$$F(k) = \frac{1 - q^k - 1 + q^{k+1} + 1 - q}{1 - q}$$

$$F(k) = \frac{1 - q + q^{k+1} - q^k}{1 - q}$$

$$F(k) = \frac{1 - q - q^k(1 - q)}{1 - q}$$

$$F(k) = \frac{1 - q - q^k(1 - q)}{1 - q}$$

$$F(k) = 1 - q^k$$

On retrouve bien :

$$F(k) = 1 - (1 - p)^k$$

□

2.1 Génération de n variables aléatoires

Afin de voir à quoi ressemble la distribution et de pouvoir tester l'estimateur que nous allons calculer par la suite, on génère des échantillons de 50, 500 et 5000 variables. On utilise ainsi le code Matlab (Octave) ce dessous :

```
p=0.1; %probability to win
pop1 = geomGen(50,p);
pop2 = geomGen(500,p);
pop3 = geomGen(5000,p);
k=1:max(pop3)+1;
distrib = (1-p).^k.*p;
subplot(221);
hist(pop1,k,1);
xlabel('k');
ylabel('Freq(k)');
title('Histogramme de 50 variables');
subplot(222);
hist(pop2,k,1);
xlabel('k');
ylabel('Freq(k)');
title('Histogramme de 500 variables');
subplot(223);
hist(pop3,k,1);
xlabel('k');
ylabel('Freq(k)');
title('Histogramme de 5000 variables');
subplot(224);
bar(k,distrib);
xlabel('k');
ylabel('P(X=k)');
title('Distribution théorique géomètrique');
```

On obtient ainsi la figure 6.

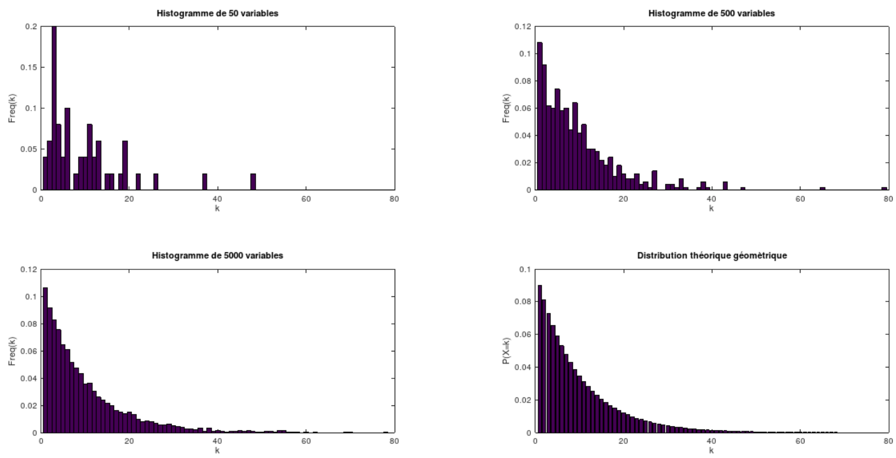

Figure 6 – Histogrammes et distribution de loi géométrique

Pour générer les populations (voir figure 6 , on utilise la fonction "geomGen" qui simule n expériences où, pour chacune d'entre elle, on effectue un enchainement d'épreuve de Bernouilli en générant une variable de loi uniforme sur $[0;1]$. Si la valeur obtenu est supérieure à p, on arrête la boucle et la valeur générée est égale au nombre d'épreuves, sinon on continue jusqu'à obtenir cette condition. Le code Matlab (Octave) utilisé est :

```
function[X]=geomGen(n,p)
X=ones(1,n);
for (i=1:n)
  trial=rand(1,1);
  k=1;
  while (trial > p)
    trial=rand(1,1);
    k++;
  endwhile
  X(i)=k;
endfor
endfunction
```

2.2 Estimateurs du maximum de vraisemblance

L'estimateur $\hat{p}$ du maximum de vraisemblance est la valeur qui maximise la loi de vraisemblance $L(p)$: $\hat{p} = argmax(L(p))$. On calcule $L(p)$ dans un premier temps :

$$L(p) = L(X_1(p), X_2(p), ..., X_n(p)) = \prod_{i=1}^{n} (1-p)^{X_i - 1} p$$

avec $X_i, \ i \in \{1, .., n\}$ les variables d'échantillon.

$$L(p) = p^n (1-p)^{(\sum\limits_{i=1}^{n} X_i) - n}$$

On peut effectuer un passage au logarithme pour trouver un maximum car il s'agit d'une fonction strictement croissante (et défini sur $]0;1]$). L'argument du maximum du logarithme de $L(p)$ et du maximum de $L(p)$ sont les mêmes.

$$\log L(p) = n \log p + (\sum_{i=1}^{n} X_i) - n$$

Références

[1] Wikipedia,
`https://en.wikipedia.org/wiki/Cauchy_distribution`