Institutul de Statistică Matematică şi Matematică Aplicată
”Gheorghe Mihoc – Caius Iacob” al Academiei Române
Calea 13 Septembrie nr. 13, sector 5, 050711 Bucureşti
Tel. 021 318 2433 Fax 021 318 2439
E-mail: office@ismma.ro

ISMMA Preprint Series

No. 1/2022

On a conjecture concerning the optimality of the uniform distribution in the broken stick model

Gheorghiţă Zbăganu

“Gheorghe Mihoc – Caius Iacob” Institute of Mathematical Statistics and Applied Mathematics of the Romanian Academy

Recommended by

Stelian Ion

On a conjecture concerning the optimality of the uniform distribution in the broken stick model

Gheorghe Zbăganu, ISMMA Bucuresti
gheorghitazbaganu@yahoo.com

1. Abstract

A stick of length 1 is broken in n segments $(Y_{j,n})_{1\le j\le n}$ by a sequence of i.i.d. random variables $\boldsymbol{X}=(X_j)_{j\ge 1}$ as follows: sort $(X_1,\dots,X_{n-1})$ ascending and obtain $O(\boldsymbol{X})=(X_{j:(n-1)})_{1\le j\le n-1}$. Next add $X_{0:(n-1)}=0$ and $X_{n:(n-1)}=1$. Finally define $Y_{j,n}=X_{j:(n-1)}-X_{(j-1):(n-1)}$ for $j=1,\dots,n$ and denote the vector of these segments by $\boldsymbol{Y}_n=(Y_{j,n})_{1\le j\le n}$. Let F be the distribution function of X_j. We prove the following result, conjectured in [5]:

Proposition *Suppose that F is absolutely continuous,its density f is positive and continuous and F has a finite barycenter $b(F)=\int_0^1 xf(x)dx$. Then the sequence of empirical Lorenz curves L_n of the vectors $\boldsymbol{Y}_n$ converges to the Lorenz curve L of some absolutely continuous probability distribution G. Moreover, $L(p)\le p+(1-p)\ln p$, the Lorenz curve of the standard exponential distribution $Exp(1)$.*

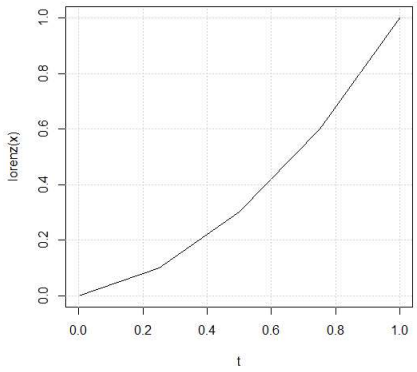
The meaning is that the the uniform distribution is the most egalitarian in the brocken stick model at least among all the absolutely continuous ones .

2. Definitions and preliminary results

Let $\mathbf{x}=(x_1,\dots,x_n)$ be a non-negative vector of reals. Sort it as $O(\mathbf{x})=(x_{1:n},\dots,x_{n:n})$. Let $s=\sum_{j=1}^n x_j=\sum_{j=1}^n x_{j:n}$. The polygonal line which joins the points $(0,0),(\frac{1}{n},\frac{x_{1:n}}{s}),(\frac{2}{n},\frac{x_{1:n}+x_{2:n}}{n}),\dots,(\frac{n-1}{n},\frac{x_{1:n}+x_{2:n}+\dots+x_{n-1:n}}{s}),(1,1)$ is called the **Lorenz curve** of the vector $\mathbf{x}$. We denote it by $L_{\mathbf{x}}$.

Important remark: the **Lorenz curve** of a vector is invariant to homotheties. Meaning that $\mathbf{x}$ and $\lambda\mathbf{x}$ have the same Lorenz curve.

Example. If $\mathbf{x}=(4,1,2,3)$ then $L_{\mathbf{x}}$ is the polygonal line which joins the points $(0,0),\left(\frac{1}{4},\frac{1}{10}\right),\left(\frac{2}{4},\frac{3}{10}\right),\dots,\left(\frac{3}{4},\frac{6}{10}\right),(1,1)$ as in the figure below



If F is a probability distribution on $[0,\infty)$ and it has a finite **positive** barycenter $\mu = \int x \mathrm{d}F(x) = \int_0^1 F^{-1}(u)du$, then its **Lorenz curve** is defined by $L_F(p) = \frac{\int_0^p F^{-1}(u)du}{\mu}$.

Here $F^{-1}(p) = \sup\{x : F(x) \leq p\}$ stands for the superior quantile of F.

For example if F is the standard exponential distribution, then $F(x) = 1 - e^{-x}$ hence $F^{-1}(p) = -\ln(1-p), \mu = 1$ therefore $L_F(p) = p + (1-p)\ln(1-p), p \in [0,1]$

The relation between these two definitions is that

$$\text{If } \mathbf{x} = (x_1, \ldots, x_n), \quad \text{then } L_{\mathbf{x}} = L_F \text{ where } F = \frac{1}{n}\sum_{j=1} \delta_{x_j}$$

Here $\delta_a(A) = 1_A(a)$ stands for the Dirac distribution concentrated at a.

If F and G are two distributions on $[0,\infty)$ one says that F is **less egalitarian** than G (or that F exhibits more inequality than G) if $L_F \leq L_G$. [1]

Or, in terms of random variables: if F is the distribution of X, G is the distribution of Y then we say that X **is less egalitarian than** Y if $L_F \leq L_G$. It is known $[1,3]$ that

$$L_F \leq L_G \text{ if and only if } Eu\left(\frac{X}{\mathrm{E}X}\right) \leq Eu\left(\frac{Y}{\mathrm{E}Y}\right) \text{ for all convex u:} \mathbb{R}_+ \to \mathbb{R}$$

The following fact was proved in $[5]$

Proposition *2.1. Suppose that F_n ,F are distributions on $\mathbb{R}_+$,F_n converges weakly to F and $b(F_n) \to b(F) > 0$. Then $L_{F_n} \to L_F$ pointwise. Conversely, if $L_{F_n} \to L_F$ pointwise and $b(F_n) \to b(F) > 0$ it follows that F_n converges weakly to F .*

Corollary *2.2. Let $\boldsymbol{Y}_n = (Y_{j,n})_{1\leq j\leq n}$ be defined as in abstract. Suppose that F is the uniform distribution on $[0,1]$. Let $F_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{nY_{j,n}}$. Then $F_n \Rightarrow Exp(1)$*

Proof Let $(\xi_j)_{j\geq 1}$ be a sequence of i.i.d. exponentially distributed random variables. It is well known (see for instance [4]) that the vector $\boldsymbol{Y}_n$ has the same distribution as $\boldsymbol{Z}_n = \left(\frac{\xi_j}{\xi_1+\ldots+\xi_n}\right)_{1\leq j\leq n}$. It follows that the random variable F_n has the same distribution as $G_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{\frac{n\xi_j}{\xi_1+\ldots+\xi_n}}$. But the Lorenz curve of G_n is the same as the Lorenz curve of the

vector $(\xi_1,\ldots,\xi_n)$ or, which is the same thing with the Lorenz curve of empirical distribution $H_n = \frac{1}{n}\sum_{j=1}^{n}\delta_{\xi_j}$. By Glivenko theorem we know that $H_n \Rightarrow Exp(1)$ uniformly s.s.

As all the barycenters $b(F_n) = b(G_n) = b(H_n) = 1$ we may apply Proposition 2.1 to infer that $L_{H_n} = L_{G_n} \to L_{Exp(1)}$. According to the second part from Proposition 2.1 it follows that $G_n \Rightarrow Exp(1)$, hence $F_n \Rightarrow Exp(1)$. QED

Corollary *2.3. Let $\boldsymbol{Y}_n = (Y_{j,n})_{1\le j\le n}$ be defined as in abstract. Suppose that F is the uniform distribution on $[0,1]$. Then $L_{Y_n}(p) \to p + (1-p)\ln(1-p), p \in [0,1]$ as $n \to \infty$.*

Proof $L_{Y_n}(p) = L_{F_n}(p) \to L_{Exp(1)}(p) = p + (1-p)\ln(1-p)$. QED

3. The main result.

The following result was proved in [5],Corollary 2, p 21:

Proposition *3.1. Let $0 = T_0 < T_1 <\ldots< T_k = 1$,$I_j = [T_{j-1}, T_j], \pi_j = T_j - T_{j-1}$ and $F = \sum_{j=1}^{k} p_j U(I_j)$.*

Suppose that all p_j are positive and that they sum up to 1.

Let $\boldsymbol{Y}_n = (Y_{j,n})_{1\le j\le n}$ be defined as in abstract and $G_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{nY_{i,n}}$ be the normalized empirical distribution of the lags.

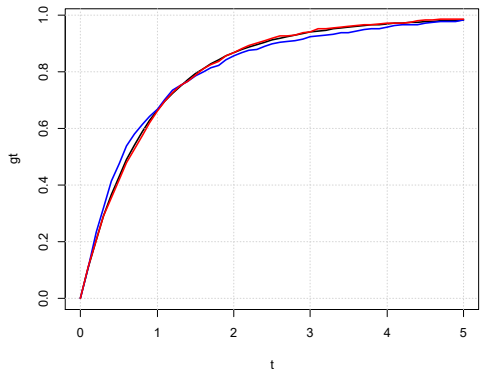
Then $G_n \Rightarrow G := \frac{1}{k}\sum_{j=1}^{k} p_j Exp\left(\frac{p_j}{\pi_j}\right)$ therefore the Lorenz curves of the lags have a limit: $L_{G_n} \to L_G$ as $n \to \infty$.

If $T_j = \frac{j}{k}$ then $G = \frac{1}{k}\sum_{j=1}^{k} p_j Exp(kp_j)$. Moreover, this limit is invariant with respect to permutations of the weights $(p_j)_j$,it has the barycenter equal to 1 and it is less egalitarian than $Exp(1)$.

As a byproduct $L_G(p) \le p + (1-p)\ln(1-p)$.

Remark. In the terminlogy from [5] if the mother distribution F is a mixture of uniform distributions, then the born distribution G is again a mixture, but of exponential distributions. All these distributions are DFR ones: their hazard rate is decreasing. See, for instance [2].

Example *The mother distribution $F = \left(U\left(0,\frac{1}{3}\right) + 2 * U\left(\frac{1}{3},\frac{2}{3}\right) + 3 * U\left(\frac{2}{3},1\right)\right)/6$ bears the distribution $G = \left(Exp\left(\frac{1}{2}\right) + 2 * Exp(1) + 3 * Exp\left(\frac{3}{2}\right)\right)/6$. We have checked if by simulations. In the figure below G is the black curve, G_{500} the blue one and G_{1000} is red*



We generalize this result.

Proposition *3.2. Suppose that the mother distribution F admits a continuous density $f > 0$. Then it bears the distribution G with the distribution function*

$$G(x) = 1 - \int_0^1 f(t)e^{-f(t)x}dt$$

Notice that the barycenter of G is equal to 1.

Proof Let $k \geq 2$ be fixed and $T_j = \frac{j}{k}, 0 \leq j \leq k.$ Consider the densities $f_k = \sum_{j=1}^{k} kI_j 1_{\left[\frac{j-1}{k}, \frac{j}{k}\right]}$ with $I_j = \int_{\frac{j-1}{k}}^{\frac{j}{k}} f(t)dt.$ As f is continuous we can write it as $f_k = \sum_{j=1}^{k} f(\theta_j) 1_{\left[\frac{j-1}{k}, \frac{j}{k}\right]}$ with $\theta_j \in \left[\frac{j-1}{k}, \frac{j}{k}\right]$ such that $kI_j = f(\theta_j).$ Notice that $f_k \to f$ uniformly and consequently, the distributions F_k converge in total variation to $F.$ Here $dF_k = f_k, dF = f.$

But the mother distribution F_k bears the distribution G_k with $1 - G_k(x) = \sum_{j=1}^{k} \frac{f(\theta_j)}{k} e^{-xf(\theta_j)}.$

If $k \to \infty$ then $\sum_{j=1}^{k} \frac{f(\theta_j)}{k} e^{-f(\theta_j)x} \to \int_0^1 f(t)e^{-xf(t)}dt.$ Let G be the distribution with the tail $1 - G(x) = \int_0^1 f(t)e^{-xf(t)}dt.$ This is the distribution born by $F.$

As its barycenter is $b(G) = \int_0^\infty (1 - G(x))dx = \int_0^\infty \int_0^1 f(t)e^{-xf(t)}dtdx = \int_0^1 \int_0^\infty f(t)e^{-xf(t)}dxdt = \int_0^1 1dx = 1,$ the Lorenz curves of G_k converge to the Lorenz curve of $G.$ QED

Corollary *3.3. If the mother distribution F has a positive continuous density then the distribution G born by it is less egalitarian than the distribution born by $U(0,1)$. Or, otherwise written $L_G(p) \leq p + (1-p)\ln(1-p)$*

Proof. It is obvious: $L_{G_k}(p) \leq p + (1-p)\ln(1-p)$ and $L_{G_k} \to L_G$ as $G \to \infty.$ QED.

Definition *The Gini index of a distribution G is Gini*$(G) = 1 - 2\int_0^1 L_G(p)dp.$

Remark *One can prove that Gini*$(G) = 1 - \frac{E\min(X,Y)}{EX}$*, where* X, Y *are two independent variables having the distribution* G*. Otherwise written, .*

$$Gini(G) = 1 - \frac{\int_0^\infty (1-G(x))^2 dx}{\int_0^\infty (1-G(x))dx}$$

In our case $1 - G(x) = \int_0^1 f(t)e^{-f(t)x}dt$ has the property that $\int_0^\infty (1-G(x))dx = 1$ hence we can write Gini$(G) = 1 - \int_0^\infty (1-G(x))^2 dx = 1 - \int_0^\infty \left(\int_0^1 f(t)e^{-f(t)}dt\right)^2 dt.$

Or, if we apply Fubini theorem, $\left(\int_0^1 f(t)e^{-f(t)}dt\right)^2 = \int\int f(s)f(t)dtds$, we arrive at

Corollary *3.4. The Gini index of the born distribution* G *has the formula* $Gini(G) = 1 - \int_0^1\int_0^1 \frac{f(s)f(t)}{f(s)+f(t)}dtds$*. This quantiy is always greater than the Gini index of the exponential distribution. Therefore* $Gini(G) \geq \frac{1}{2}$.

Proof By Fubini, we can write Gini$(G) = 1 - \int_0^\infty \left(\int_0^1 f(t)e^{-f(t)}dt\right)^2 dt$

$= 1 - \int_0^1\int_0^1\int_0^\infty f(s)f(t)e^{-(f(s)+f(t))x}dxdtds = 1 - \int_0^1\int_0^1 \frac{f(s)f(t)}{f(s)+f(t)}dtds$

The claim that Gini$(G) \geq \frac{1}{2}$ is a consequence of the fact that G exhibits more inequality than the standard exponential distribution $Exp(1)$. QED

Example *Suppose that the density of the mother distribution is* $f(t) = (a+bt)1_{(0,1)}(t)$*. Here* $a, b, a+b$ *are non negative and* $a + \frac{b}{2} = 1$.

Then $G_{a,b}(x) = \begin{cases} 1 - \frac{1}{bx^2}((1+ax)e^{-ax} - (1+(a+b)x)e^{-(a+b)x}) & \text{if } b \neq 0 \\ 1 - e^{-x} & \text{if } b = 0 \end{cases}$

Its Gini coefficient is $1 - \int_0^1\int_0^1 \frac{(a+bs)(a+bt)}{2a+bs+bt}dtds$

For instance

- if $a = 0, b = 2,$ then Gini$(G) = 1 - 2\int_0^1\int_0^1 \frac{st}{s+t}dtds = \frac{4}{3}\ln 2 - \frac{1}{3} = 0.59086$

- if $a = 1.5, b = -1$ then Gini$(G) = 1 - \frac{1}{2}\int_0^1\int_0^1 \frac{(3-2s)(3-2t)}{3-s-t}dtds \approx 0.5212$

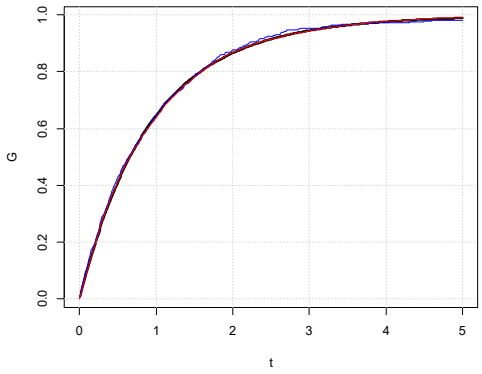
In the figure below there are three curves, plotted for $a = 1.5, b = -1$: the black one is the graph of the predicted $G,$ the blue one is the graph of G_{500} and the red one the graph of G_{5000}.

The plot was made in R with the script

```
gab<-function(a,b,x) {A=(1+a*x)*exp(-a*x);B=(1+(a+b)*x)*exp(-(a+b)*x)
rez=(A-B)/b/x/x;rez=1-rez;rez}          ## the function G
```

$G(x) = 1 - \frac{1}{x^2}\left(e^{-\frac{x}{2}}\left(\frac{x}{2}+1\right) - e^{-\frac{3x}{2}}\left(\frac{3x}{2}+1\right)\right)$

```
t=seq(0,5,by=.01);nt=length(t);b=-1;a=1-b/2;ft=t; G=t ## initializations
for (k in 1:nt) {G[k]=gab(a,b,t[k])};plot(t,G,type="l",lwd=3);grid()
do<-function(x)
{n=length(x);y=sort(x);n1=n-1;z=1:n1;for ( k in 1:n1) {z[k]=y[k+1]-y[k]};z}
n=5000;x=(sqrt(a^2+2*b*runif(n))-a)/b;y=do(x);z=n*y
for (k in 1:nt) {ft[k]=femp(z,t[k])};lines(t,ft,col=2,lwd=1.5)
```



Open problems.

1.What happens if f is equal to 0 on some interval? Or if it is not continuous?

2. What if in the broken stick evolution only the longest stick is broken?

References

1. Arnold B.C. : *The Lorenz Order in the Space of Distribution Functions*. In *Majorization and the Lorenz Order: A Brief Introduction. Lecture Notes in Statistics*, vol 43. Springer, NY,1987.

2. Barlow, R. E. & Prochan, F.: *Mathematical theory of reliability*. John Wiley &Sons, Inc., New York, 1965.

3. Shaked, M., Shanthikumar, G.:*Stochastic Orders*,Springer Series in Statistics, Springer New York, 2007

4 . S. S. Wilks, *Mathematical Statistics*, Wiley . Chapter 7.7,177-182,1962.

5. Zbăganu, G. Asymptotic Results in Broken Stick Models: The Approach via Lorenz Curves. Mathematics 2020, 8, 625. (https://www.mdpi.com/2227-7390/8/4/625)