Excess wealth transform-a useful tool for data analysis

Maochao Xu

Department of Mathematics
Illinois State University
mxu2@ilstu.edu

Common Measures of Variability

☛ Variance

$$\mathrm{Var}(X) = \frac{\sum (X_i - \mu)^2}{N}.$$

☛ Range

$$R = X_{n:n} - X_{1:n}.$$

☛ Interquartile range

$$\mathrm{IQR} = Q_3 - Q_1.$$

All of those measures are based on a single number and may not be very informative as observations in one particular part of the data may be more spread out than in other parts.

Excess wealth transform–Fernández-Ponce, Kochar and Muñoz-Perez (1998) and *Shaked and Shanthikumar* (1998)

For a random variable X with finite mean, distribution F, the transform is defined by

$$\begin{aligned} W(p\,;F) &= \mathrm{E}\left[(X - F^{-1}(p))^+\right] \\ &= \int_{F^{-1}(p)}^{\infty} \bar{F}(x)dx, \end{aligned}$$

where $(X)^+ = \max\{X, 0\}$,

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}, \quad \text{for } 0 \leq p \leq 1.$$

and

$$\bar{F} = 1 - F$$

is the survival function of X.

This transform is called the *right spread* transform in Fernández-Ponce, Kochar and Muñoz-Perez (1998). *Shaked and Shanthikumar* (1998) named it as *excess wealth* transform.

In the context of economics, $W(p\,;F)$ can be thought of as the additional wealth (on top of the pth percentile income) of the of the richest $100(1-p)\%$ individuals in the population.

Relation to variability

- ☛ Variance—Fernández-Ponce, Kochar and Muñoz-Perez (1998)

$$\mathrm{Var}(X) = \int_0^1 \left[\frac{W(p\,;F)}{1-p}\right]^2 dp.$$

- ☛ Spacing—Fernández-Ponce, Kochar and Muñoz-Perez (1998)

  If X_1 and X_2 are independent copies of X,

$$\mathrm{E} \mid X_1 - X_2 \mid = 2\int_0^1 W(p\,;F)dp.$$

- ☛ Truncated variance

$$\mathrm{Var}(X \mid X > F^{-1}(p_0)) = \frac{1}{1-p_0}\int_{p_0}^1 \left[\frac{W(p\,;F)}{1-p}\right]^2 dp.$$

- ☛ Functional variance—*Shaked and Shanthikumar* (1998)

$$W(p\,;F) \le W(p\,;G) \Longrightarrow \mathrm{Var}[h(X)] \le \mathrm{Var}[h(Y)],$$

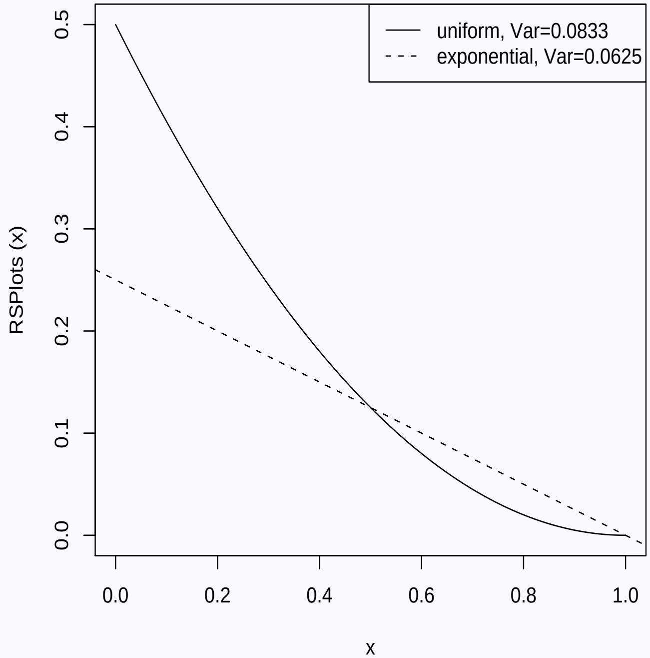
  for any increasing and convex function $h : [0, \infty) \to R$.

Example

Let X be a uniform $(0, 1)$ random variable, and Y be an exponential random variable with rate 4, then, for $0 < p < 1$,

$$W(p\,; F) = \int_{F^{-1}(p)}^{\infty} \bar{F}(x)dx = \frac{(1-p)^2}{2},$$

$$W(p\,; G) = \int_{G^{-1}(p)}^{\infty} \bar{G}(x)dx = \frac{1-p}{4}.$$



Detecting Ageing Class

The right spread transform may also be used to detect the aging property. For example, it could be used to detect the increasing failure rate (IFR) property of data sets.

☞ Proposition—Fernández-Ponce, Kochar and Muñoz-Perez (1998)

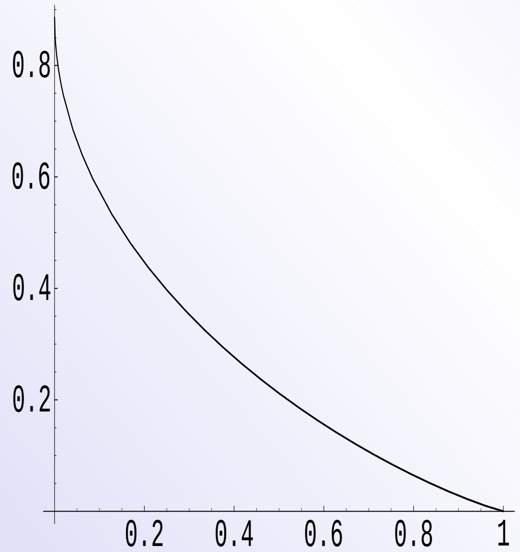
If X is IFR, then $W(p\,;F)$ is a convex shaped function of p.

Example: Weibull distribution with shape parameter 2 and scale parameter 1. The density function is

$$f(x) = 2xe^{-x^2}.$$

The hazard rate is

$$r(x) = 2x.$$



First Page
Title Page
Contents
Page 6 of 37
Go Back
Full Screen
Close
Quit

Excess wealth plot (EW-plot)

Assume $X_1, \cdots, X_n$ be a random sample from X. Let $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$ denote the order statistics corresponding to $X_1, \cdots, X_n$. The empirical distribution is defined as

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(X_i \le t).$$

Hence, the excess wealth transform is, for $0 \le i \le n-1$,

$$\begin{aligned} W_i = W\left(\frac{i}{n}; \hat{F}_n\right) &= \sum_{j=i}^{n-1} \int_{F_n^{-1}(j/n)}^{F_n^{-1}((j+1)/n)} \bar{F}_n(x)dx \\ &= \sum_{j=i}^{n-1} \frac{n-j}{n} \left(X_{j+1:n} - X_{j:n}\right). \end{aligned}$$

It may be written as

$$W\left(\frac{i}{n}; \hat{F}_n\right) = \sum_{j=i+1}^{n} \frac{n-j+1}{n} \left(X_{j:n} - X_{j-1:n}\right).$$

It is observed that

$$W_0 = \bar{X}, \qquad W_{i+1} = W_i - \frac{n-i}{n}(X_{i+1:n} - X_{i:n}), \quad 0 \le i \le n-2. \qquad (1)$$

A visual tool for detecting variability might be constructed as follows.

(a) Order the sample: $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$.

(b) Compute W_i as defined in equation (1) for $i = 0, \cdots, n-1$.

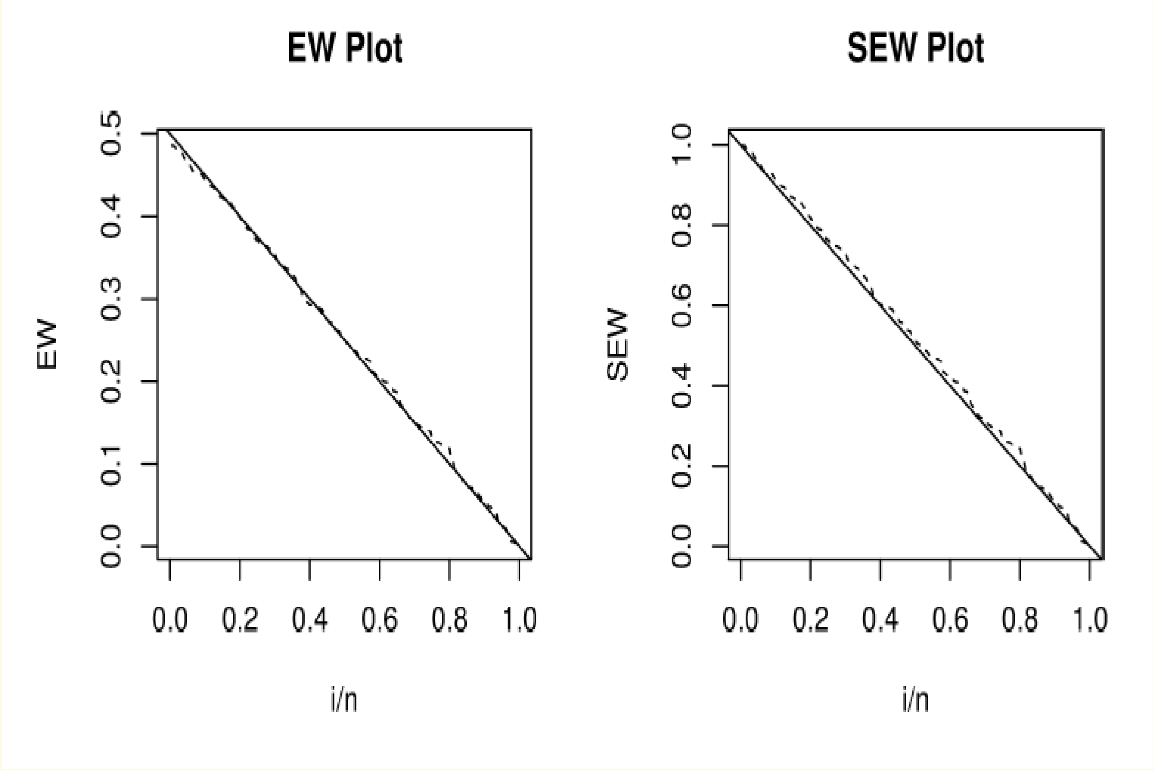
(c) Plot the pairs $(i/n, W_i)$, $i = 0, \cdots, n$, where $(1, W_n) = (1, 0)$, and connect the points by line segments.

First Page
Title Page
Contents
◀◀ ▶▶
◀ ▶

Go Back
Full Screen
Close
Quit

Example

Let X be an exponential random variable with rate 2, from which 200 samples are generated for EW-plot and scaled EW-plot. The EW-plot and scaled EW-plot are displayed in the following.



Theoretical support:

For any fixed p, if F^{-1} is continuous, then,

$$W(i/n; F_n) \xrightarrow{a.s.} W(p; F), \quad n \to \infty, \quad i/n \to p.$$

Excess wealth transform for censored data

Now, consider the random censoring mechanism,

$$X_i = \min\{T_i, C_i\},$$

where T_i is the real survival time with distribution F, and C_i is some independent censoring variable with distribution G. The observed quantity is the pair (X_i, D_i), where

$$D_i = \begin{cases} 1, & \text{if } X_i = T_i, \\ 0, & \text{if } X_i = C_i. \end{cases}$$

Assume the observed data set is $(x_1, d_1), \cdots, (x_n, d_n)$, where $d_i = 1(0)$ if the observation is uncensored (censored) for $i = 1, \cdots, n$. For the censored data, the Kaplan-Meier (K-M) estimator provides an estimate of the true distribution function of T_i, i.e.,

$$K_n(t) = 1 - \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1} \right)^{d_j}, \tag{2}$$

where k_t is the value of k such that $t \in [x_{k:n}, x_{k+1:n})$.

It is noted that if there is no censoring, the K-M estimator is the empirical distribution function.
Define

$$\bar{H}(x) = 1 - H(x) = \bar{F}(x)\bar{G}(x),$$

and $\tau = H^{-1}(1)$.

The censored excess wealth transform as

$$W^C(p\,; F) = \int_{F^{-1}(p)}^{\tau} \bar{F}(x)dx.$$

Hence,

$$W^C(p\,;K_n) = \int_{K_n^{-1}(p)}^{\tau} \bar{K}_n(x)dx,$$

where

$$\bar{K}_n(t) = 1 - K_n(t) = \prod_{j=1}^{k_t} \left(\frac{n-j}{n-j+1} \right)^{d_j}.$$

Now,

$$W_i^C = W(K_n(X_{i:n}); K_n) = \sum_{j=i}^{n-1} (X_{j+1:n} - X_{j:n}) \prod_{k=1}^{j} \left(\frac{n-k}{n-k+1} \right)^{d_k}. \qquad (3)$$

Similarly,

$$W_0^C = \tilde{X}, \qquad W_{i+1}^C = W_i^C - (X_{i+1:n} - X_{i:n}) \prod_{k=1}^{i} \left(\frac{n-k}{n-k+1} \right)^{d_k},$$

where $\tilde{X}$ is the censored mean as suggested in *Gill* (1983).

EW-Plot for censored data

A visual tool for detecting the variability for the censored data could be constructed as follows.

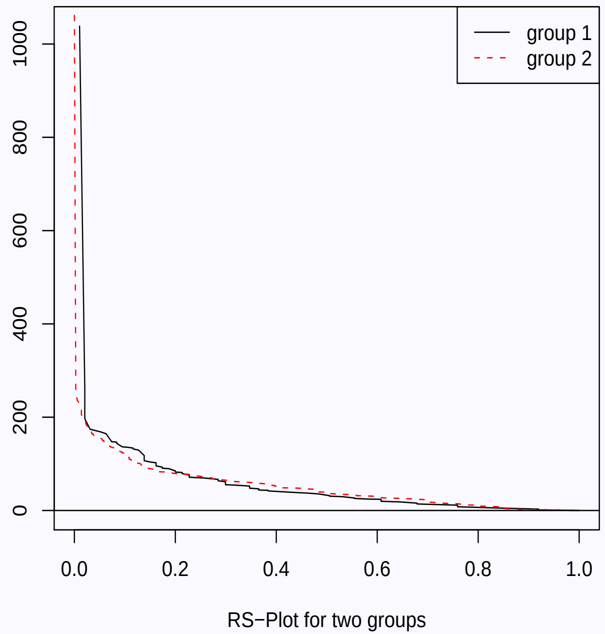
(a) Order the sample: $X_{1:n} \le X_{2:n} \le \cdots \le X_{n:n}$.

(b) Compute W_i^C as defined in equation (3) for $i = 0, \cdots, n-1$.

(c) Plot the pairs $\big(K_n(X_{i:n}), W_i^C\big)$, $i = 0, \cdots, n$, where $(1, W_n) = (1, 0)$.

EW-plot for Channing House data

Channing House is a retirement centre in Palo Alto, California. These data were collected between the opening of the house in 1964 until July 1, 1975. In that time 97 men and 365 women passed through the centre. For each of these, their age on entry and also on leaving or death was recorded. A large number of the observations were censored mainly due to the resident being alive on July 1, 1975 when the data was collected. Over the time of the study 130 women and 46 men died at Channing House.

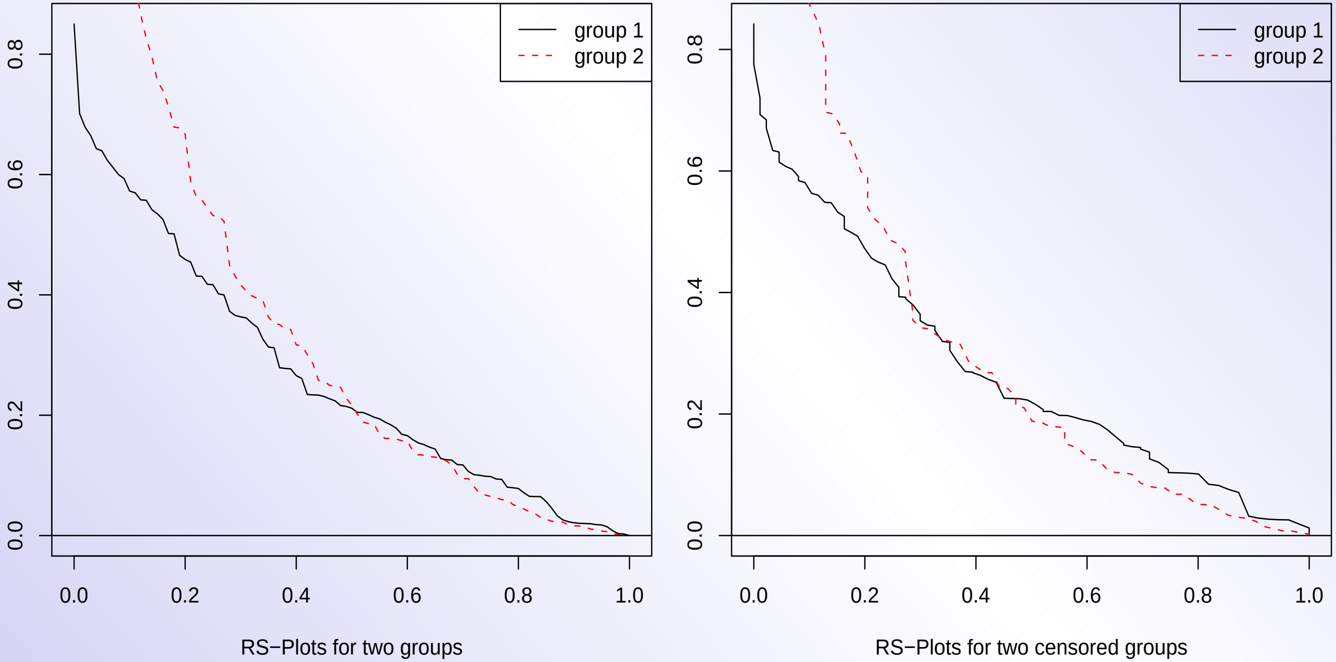


Standard deviation for male group (group 1): 73.71466
Standard deviation for female group (group 2): 73.81204

EW-plot for Weibull censored data

Group 1: $n = 100$, shape parameter 2 and scale parameter 1, variance is 0.215, 31 censored observations.

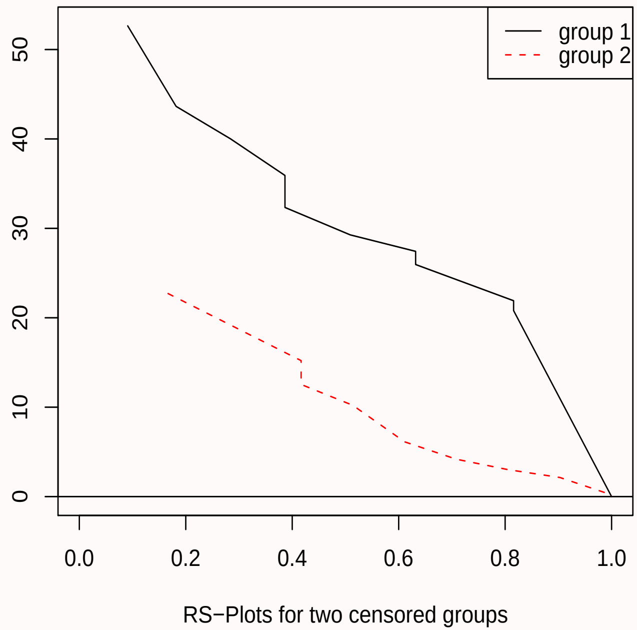
Group 2: $n = 100$, shape parameter 3 and scale parameter 2, variance is 0.359, 30 censored observations.



EW-plot for AML data

The data in the following Table are preliminary results from a clinical trial to evaluate the efficacy of maintenance chemotherapy for acute myelogenous leukemia (AML). The first group received maintenance chemotherapy; the second group did not. The purpose of the trial was to see if maintenance chemotherapy prolonged the time until relapse (cf. Embury, et al., 1977).

Group	weeks of complete remission
Group 1	9,13,13+,18,23,28+,31,34,45+,48,161+
Group 2	5, 5, 8, 8, 12, 16+, 23, 27, 30, 33, 43, 45



Theoretical support:

For any fixed p, if F is continuous, then,

$$W^C(K_n(x_{i:n}); K_n) \xrightarrow{a.s.} W^C(p; F), \quad n \to \infty, \quad K_n(x_{i:n}) \to p.$$

EW transform for heavy-tailed data

Distribution of the excess over a threshold μ:

$$F_\mu(t) = P\left(X - \mu \leq t \mid X \geq \mu\right).$$

Peaks over threshold (POT) modeling:

1. Hydrology: It is critical to model the level of water in a river or sea to avoid flooding.
2. Actuarial science: Insurance companies set premium levels based on models for large losses.
3. Survival analysis: The POT method is used for modeling lifetimes.
4. Environmental science: Public health agencies set standards for pollution levels.

Peaks over threshold modeling is based on the generalized Pareto class of distributions being appropriate for describing statistical properties of excesses. generalized Pareto distribution (GPD):

$$G_{\xi,\beta}(x) = \begin{cases} 1-(1+\xi x/\beta)^{-1/\xi} & \xi \neq 0, \\ 1-\exp(-x/\beta) & \xi = 0, \end{cases}$$

where $\beta > 0$, and $x \geq 0$ when $\xi \geq 0$, and $0 \leq x \leq -\beta/\xi$ if $\xi < 0$.

It is known that there exists a positive measurable function $\beta(\mu)$ such that

$$\lim_{\mu\to\infty} \sup_{x\geq\mu} \mid F_\mu(x) - G_{\xi,\beta(\mu)(x)} \mid = 0,$$

if

$$\lim_{x\to\infty} \frac{\bar{F}(tx)}{\bar{F}(x)} = t^{-1/\xi}.$$

The mean excess (ME) function is a tool popularly used to aid this choice of u and also to determine the adequacy of the GPD model in practice.

$$M(u) = \mathrm{E}(X - u \mid X > u),$$

if $\mathrm{E}(X) < \infty$. It is seen that if $\xi < 1$, then

$$M(u) = \frac{\beta}{1-\xi} + \frac{\xi}{1-\xi}u$$

Look at the EW transform:

$$W(p;F) = \frac{\beta}{1-\xi}(1-p)^{1-\xi}.$$

The scaled EW transform:

$$SEW(p;F) = \frac{W(p;F)}{\mathrm{E}(X)} = (1-p)^{1-\xi}.$$

The scaled EW transform could be used to detect heavy-tailed data:

$$\left(p, 1 - \frac{\ln(SEW(p;F))}{\ln(1-p)}\right).$$

It can be shown that

$$1 - \frac{\ln(SEW(i/n;F_n))}{\ln(1-i/n)} \xrightarrow{P} \xi.$$

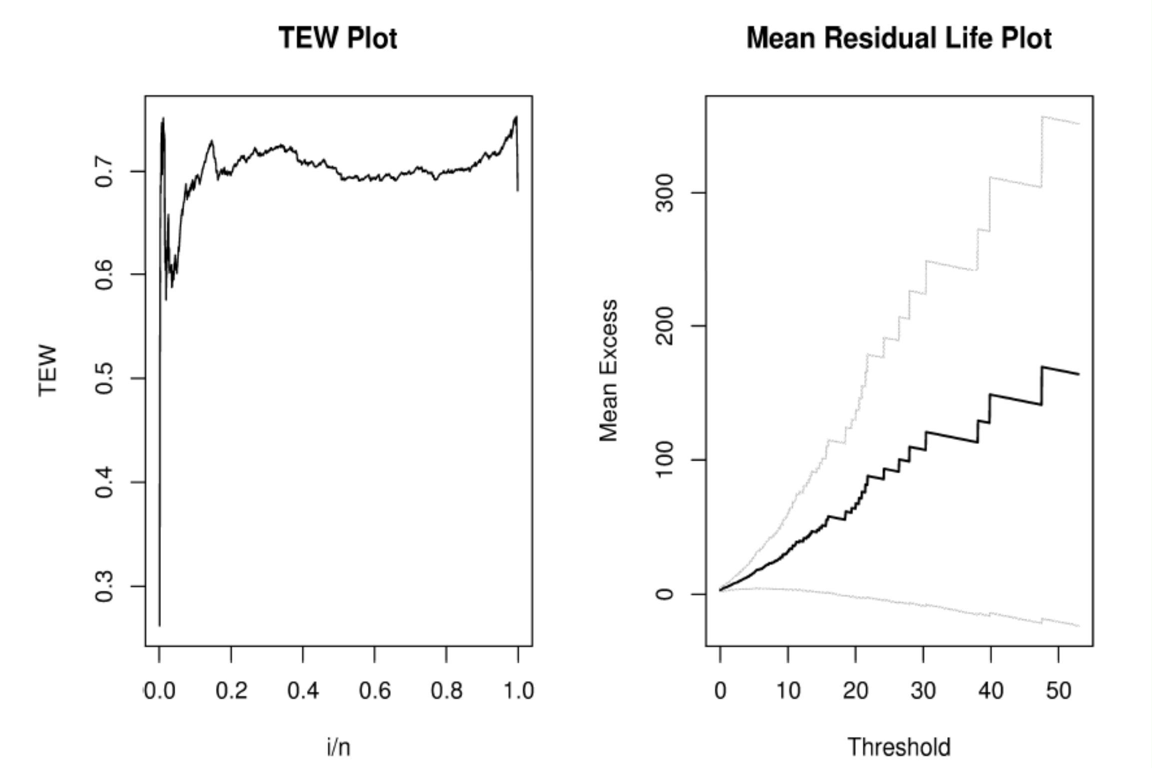
So, the following transformed plot (TEW) can be developed to detect the heavy-tailed data:

$$\left(\frac{i}{n}, 1 - \frac{\ln(SEW(i/n;F_n))}{\ln(1-i/n)}\right)$$

We generate 1000 data from the GPD distribution with $\beta = .9$ and $\xi = .7$, i.e.,

$$G_{\xi,\beta}(x) = 1 - (1 + .7x/.9)^{-1/.7}.$$

The TEW plot and ME plot (POT package in R) are displayed in the following. It is seen that the TEW plot is very informative, and also provides the rough estimate of ξ around .7.



Excess wealth order

A random variable X with distribution function F is said to be less than a random variable Y with distribution function G in the *excess wealth order*, denoted by $X \leq_{ew} Y$, if

$$W(p\,;F) \leq W(p\,;G), \quad \text{for all } p \in [0,1].$$

It is known that the right spread order is independent of location parameters (cf. *Kochar and Carriére*, 1997). That is, for any c,

$$X \leq_{ew} Y \Longleftrightarrow X \leq_{ew} Y + c.$$

Also it is known that,

$$X \leq_{ew} Y \Longrightarrow \mathrm{Var}X \leq \mathrm{Var}Y.$$

Actuarial application

The VaR is defined as

$$\mathrm{VaR}[X;\, p] = F^{-1}(p).$$

As the VaR at a fixed level only gives local information about the underlying distribution, actuaries proposed the so-called *expected shortfall* to overcome this shortcoming. Expected shortfall at probability level p is the stop-loss premium with retention $\mathrm{VaR}[X;\, p]$, that is, $\mathrm{E}\left(X - \mathrm{VaR}[X;\, p]\right)_+$, which is just the excess wealth transform of X. Hence, excess wealth order provides a natural way to compare the risks. *Sordo* (2010) proved the following interesting result.

X and Y bet two random variables with respective distribution functions F and G. Then

$$X \leq_{ew} Y \Longleftrightarrow H_{\phi,p}(X) \leq H_{\phi,p}(Y), \quad 0 < p < 1,$$

where

$$H_{\phi,p}(X) = \mathrm{E}(\phi(X - \mathrm{E}(X_p)) \mid X > F^{-1}(p)),$$

and ϕ is a convex function, and $X_p = (X \mid X > F^{-1}(p))$.

As a direct consequence,

$$X \leq_{ew} Y \Longrightarrow \mathrm{Var}(X \mid X > F^{-1}(p)) \leq \mathrm{Var}(Y \mid Y > G^{-1}(p)), \quad 0 < p < 1.$$

Hypothesis Testing

Belzunce, et al. (2001) established L-statistics to test the right spread order:

$$H_0 : X \stackrel{ew}{=} Y$$

vs. the alternative,

$$H_1 : X <_{ew} Y.$$

However, the test there may not be consistent since at some points the excess wealth transforms of X and Y may cross.

Denuit, et al. (2007) proposed a Kolmogorov-Smirnov type test for the shortfall dominance against parametric alternatives (one sample test), where the *shortfall order* is equivalent to the *excess wealth order* with replacing p by $1 - p$.

In the following, we will establish two sample Kolmogorov-Smirnov type test for the excess wealth order.

Two Sample Kolmogorov-Smirnov Test

Assume $X_1, \cdots, X_n$ be a random sample from continuous distribution F, and $Y_1, \cdots, Y_m$ be a random sample from continuous distribution G. The testing hypothesis we are interested in is

$$H_0 : W(p\,; F) \leq W(p\,; G) \quad \text{for all } p \in [0, 1],$$

versus,

$$H_a : W(p\,; F) > W(p\,; G) \quad \text{for some } p \in [0, 1].$$

Our test statistic based on empirical distributions for the null hypothesis is

$$S(\hat{F}_n; \hat{G}_m) = \left(\frac{nm}{n+m}\right)^{1/2} \sup_p \left(W(p\,; \hat{F}_n) - W(p\,; \hat{G}_m)\right).$$

We reject the null hypothesis H_0 when

$$S(\hat{F}_n; \hat{G}_m) > c_\alpha,$$

where c_α is the critical value based on the significant level α.

From the classical empirical process theory, it follows that

$$\sqrt{n}\left(\hat{F}_n - F\right) \Longrightarrow \mathbb{U}(F),$$

where $\mathbb{U}$ is a Brownian bridge process, and $\hat{F}_n$ is the empirical distribution based on n samples from X. Then, we have the following Proposition.

Assume F be continuous,

$$\sqrt{n}\left(W(p\,;\hat{F}_n) - W(p\,;F)\right) \Longrightarrow \mathbb{W}(p\,;F),$$

where

$$\mathbb{W}(p\,;F) = -\int_p^1 \frac{\mathbb{U}(s)}{f(F^{-1}(s))}ds + (1-p)\frac{\mathbb{U}(p)}{f(F^{-1}(p))}.$$

Theoretical Result

Assume F and G be continuous, and their inverses be also continuous. If $n/(n+m) \to \lambda$ as $n, m \to \infty$, $0 < \lambda < 1$,

(i)

$$\lim_{n,m\to\infty} P\left(\text{reject } H_0 \mid H_0 \text{ is true}\right) \le P(\sup_p \mathcal{S}(F;G)(p) > c_\alpha \mid H_0 \text{ is true}) = \alpha,$$

with equality holds when $X \overset{ew}{=} Y$, where

$$\mathcal{S}(F;G)(p) = (1-\lambda)^{1/2}\mathbb{W}(p\,;F) - \lambda^{1/2}\mathbb{W}(p\,;G).$$

(ii)

$$\lim_{n,m\to\infty} P\left(\text{reject } H_0 \mid H_0 \text{ is false}\right) = 1.$$

Bootstrapping for p-value

One method is to resample from the the combined sample with replacement (see *Van der Varrt and Wellner*, 1996). Assume $X_1^*, X_2^*, \cdots, X_n^*, Y_1^*, Y_2^*, \cdots, Y_m^*$ be a bootstrap sample from the combined sample $X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_n$. Denote

$$\hat{F}_n^*(t) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(X_i^* \leq t), \quad \hat{G}_m^*(t) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{I}(Y_i^* \leq t),$$

the empirical distributions based on the bootstrap samples. Then we compute the following test statistic.

$$S^{*,1}(\hat{F}_n; \hat{G}_m) = \left(\frac{nm}{n+m}\right)^{1/2} \sup_p \left(W(p\,; \hat{F}_n^*) - W(p\,; \hat{G}_m^*)\right).$$

The other method is to bootstrap independently from two samples. That is, let F_n^* be the empirical distribution based on bootstrap sample from $X_1, X_2, \cdots, X_n$, and G_m^* be the empirical distribution based on bootstrap sample from $Y_1, Y_2, \cdots, Y_m$. Then, compute the following statistic,

$$\begin{aligned} S^{*,2}(\hat{F}_n; \hat{G}_m) \;=\; & \left(\frac{nm}{n+m}\right)^{1/2} \sup_p \left(\left(W(p\,; \hat{F}_n^*) - W(p\,; \hat{F}_n) \right) \right. \\ & \left. - \left(W(p\,; \hat{G}_m^*) - W(p\,; \hat{G}_m) \right) \right). \end{aligned}$$

Then, the p-values are defined as, for $j = 1, 2$,

$$p_j^* = P\left(S^{*,j}(\hat{F}_n; \hat{G}_m) \geq S(\hat{F}_n; \hat{G}_m) \right).$$

Theoretical Support:
Under the same condition, assuming that $\alpha < 1/2$, a test for the right spread order based on any of the following rules:

$$\text{Reject } H_0 \text{ if } p_j^* < \alpha, \quad j = 1, 2,$$

satisfies the following,

$$\lim_{n\to\infty} P\left(\text{reject } H_0 \mid H_0 \text{ is true}\right) \le \alpha,$$

$$\lim_{n\to\infty} P\left(\text{reject } H_0 \mid H_0 \text{ is false}\right) = 1.$$

In practice, we will use Monte Carlo simulation to approximate the p-value. That is, the simulated p-value would be

$$p_j^* \approx \frac{1}{R}\sum_{i=1}^{R} \mathrm{I}\left(S_i^{*,j}(\hat{F}_n; \hat{G}_m) > S(\hat{F}_n; \hat{G}_m)\right), \quad j = 1, 2,$$

where R is the bootstrap replication.

Monte Carlo Results

The replication number of R is $1,000$. The rejection rates are computed for the bootstrap method with respect to the significant level of 0.05 and 0.01 and we consider equal sample size cases for the simulations.

Example 1

We generate X's samples from $N(1,1)$ and Y's samples from $N(2,4)$. Hence, the variance of X is smaller than the variance of Y.

N	25	50	100	500
p_1^*	0.9732	0.995	0.9982	1
p_2^*	0.997	1	1	1

From the table, it could be seen that both statistics suggest not rejecting the null hypothesis, i.e.,

$$H_0 : W(p\,;F) \le W(p\,;G) \quad \text{for all } p \in [0,1].$$

Example 2

We generate X's samples from $N(1,1)$ and Y's samples from $N(4,1)$. Hence, the variance of X is equal to the variance of Y. But, they are different in the location parameters.

N	25	50	100	500
p_1^*	0.761	0.968	0.807	0.995
p_2^*	0.558	0.937	0.679	0.974

From the table, in this case, there is no enough evidence to reject the null hypothesis. This coincides with the location free property of right spread order.

Example 3

Let X be a uniform $(0,1)$ random variable, and Y be an exponential random variable with rate 4, then, for $0 < p < 1$,

$$W(p\,;F) = \int_{F^{-1}(p)}^{\infty} \bar{F}(x)dx = \frac{(1-p)^2}{2},$$

$$W(p\,;G) = \int_{G^{-1}(p)}^{\infty} \bar{G}(x)dx = \frac{1-p}{4}.$$

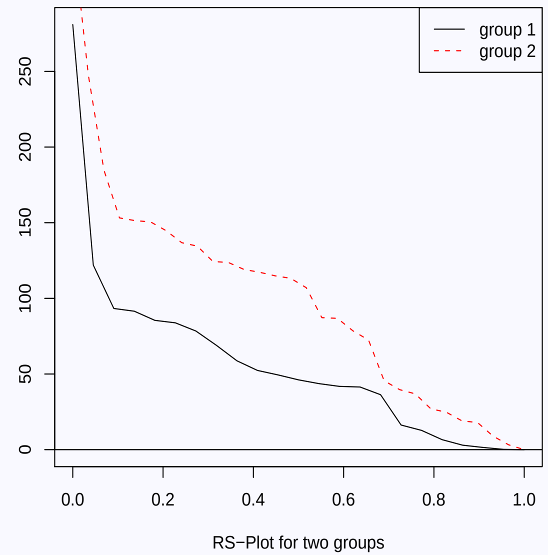
It could be seen that they have a cross at $p = 1/2$. Now, we generate data from both distributions. From Table 3, it is seen that the null hypothesis is rejected. It works well even in small samples!

N	25	50	100	500
p_1^*	0.004	0	0	0
p_2^*	0.001	0	0	0

Example: RFM data

The first data set was collected to study survival times in the presence of pollutants (cf. Hoel, 1972). The data set consists of two groups of survival times of RFM strain male mice, and the cause of death was thymic lymphoma. The first group with sample size 22 lived in a conventional laboratory environment, while the second group with sample size 29 was in a germ free environment.

The simulated p-values are $p_1^* = 0.978$ and $p_2^* = 0.665$ with 1000 bootstrap replications.



Thanks!