# A Novel Approach to Estimating Sub-threshold Aggregates in Official Statistics Using Truncated Empirical Distributions

Jacek Sadowski[†] Rafał Woźniak[‡]

[†] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw
[‡] Faculty of Economic Sciences, University of Warsaw

The 6th Congress of Polish Statistics

# British Royal Marine

- Wachter, K. W., & Trussell, J. (1982). Estimating Historical Heights. *Journal of the American Statistical Association*, 77(378), 279–293.
- https://doi.org/
- 67” ≈ 170 cm
- Restoring the full population distribution from the truncated distribution

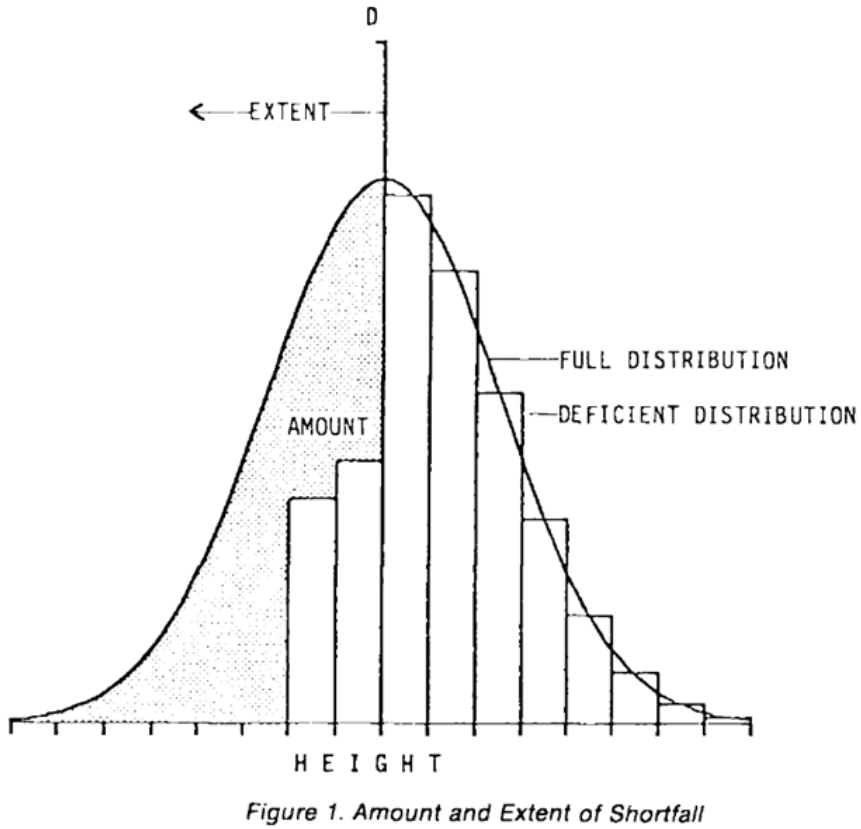



*Figure 1. Amount and Extent of Shortfall*

Source: Wachter & Trussell (1982)

# Aim

- The aim of this article is to present a novel method for estimating statistical aggregates for sub-threshold entities that are excluded from mandatory reporting due to their small scale of operation.
- The research problem addresses the high costs and respondent burden associated with traditional data-gap filling methods, such as censuses and surveys.
- The goal is to benefit from leveraging the mathematical properties of truncated empirical distributions allows for the precise reconstruction of parameters for the entire population without the need to collect additional data from small and medium-sized enterprises.

# Methods

- We propose a novel **cost-free** approach.
- The methodology is based on the assumption that the distributions of financial microdata (e.g., income, loan values) conform to normal, log-normal, or other known statistical distributions.
- A key element of the method is the estimation of truncated distribution parameters using the method of moments and the implicit function theorem.
- This enables the determination of parameters for the unobservable (sub-threshold) portion of the distribution.
- Statistical inference allows for the verification of the reliability of **this cost-free** approach compared to methods based on random sampling.

# Kod R: Generowanie danych

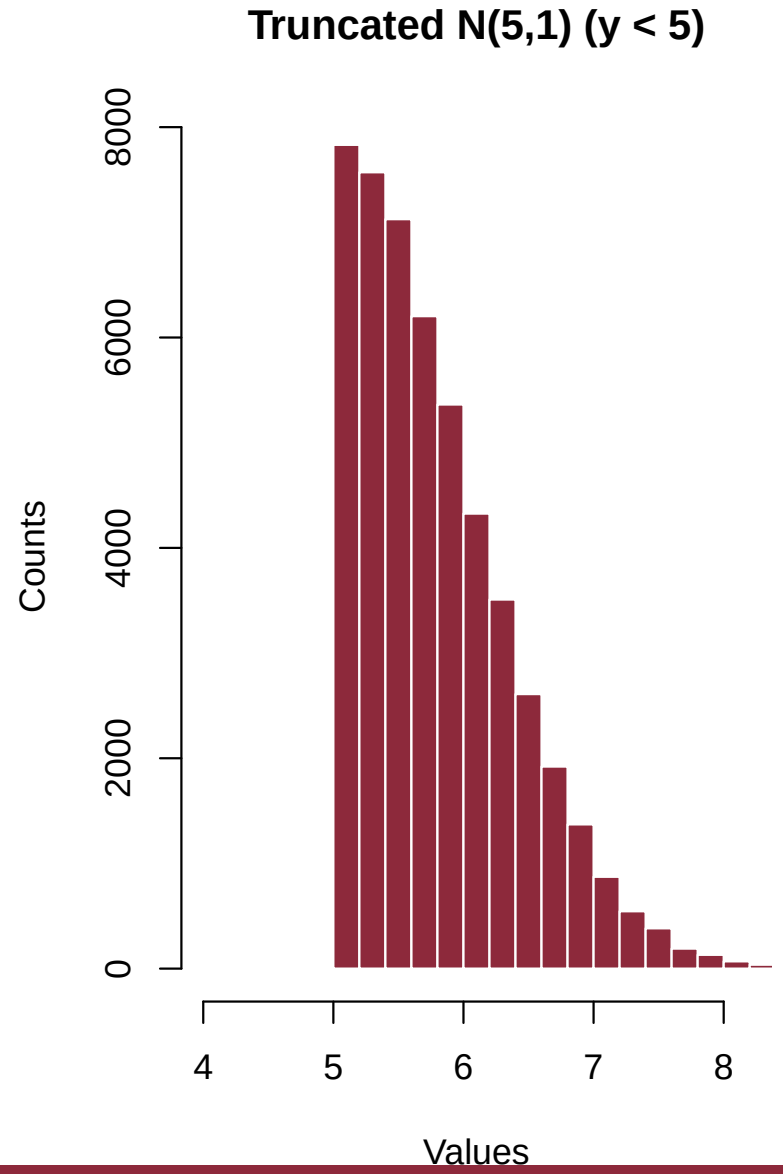
Poniższy fragment kodu odpowiada za ucięcie rozkładu:

```
set.seed("1234567")
N = 100000
mu = 5
sd = 1
a = 5

y = rnorm(n=N, mean=mu, sd=sd)
indeksy = which(y<5)
y = y[-indeksy]
```

# Illustration. Monte Carlo experiment

```r
set.seed("1234567")
N = 100000
mu = 5
sd = 1
a = 5

y = rnorm(n=N, mean=mu, sd=sd)
indeksy = which(y<5)
y = y[-indeksy]
podprogowe = y[y<a]
sum(podprogowe)
```

- Method of Moments for $\hat{\mu}$ and $\hat{\sigma}$
- Under-the-threshold sum: 210318.2
- Estimate: 279097.7



# Method of Moments

- Idea jest następująca: niech $X_1, ..., X_n$ będzie próbką z rozkładu $f(x|\theta_1, ..., \theta_k)$ z nieznanym wektorem $(\theta_1, ..., \theta_k)$.
- Momenty z populacji są funkcjami nieznanych parametrów $(\theta_1, ..., \theta_k)$, powiedzmy $\mu'_j(\theta_1, ...\theta_k)$. Metoda momentów polega na rozwiązaniu układu równań dla $(\theta_1, ..., \theta_k)$ względem $(m_1, ..., m_k)$:

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu'_1 = E(X) \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu'_2 = E(X^2) \\
&... \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu'_k = E(X^k)
\end{aligned}
$$

$$
\begin{cases}
m_1 = \mu'_1(\theta_1, ..., \theta_k), \\
m_2 = \mu'_2(\theta_1, ..., \theta_k), \\
\quad ... \\
m_k = \mu'_k(\theta_1, ..., \theta_k)
\end{cases}
$$

# For starters...

For $N(\mu, \sigma^2)$ the application of the Method of Moments is relatively simple:

$$\begin{cases} m_1 = \mu \\ m_2 = \mu^2 + \sigma^2 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = m_1 \\ \hat{\sigma^2} = m_2 - m_1^2. \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i / n, \\ \hat{\sigma^2} = \left( \sum_{i=1}^{n} X_i^2 - \left( \sum_{i=1}^{n} X_i / n \right)^2 \right) / n. \end{cases}$$

Complications emerge with truncation $X > a$. The moments for $X|_{X>a}$ depend on the inverse Mills ratio [e.g. Greene, Econometric analysis]:

$$K = \Phi'(\frac{\mu - a}{\sigma}) / \Phi(\frac{\mu - a}{\sigma}) \,, \tag{1}$$

where $\Phi$ is $N(0, 1)$ c.d.f. Then the set of equations involves the element with the implicit function $K = K(\mu, \sigma)$:

$$\begin{cases} m_1 = \mu + \sigma \cdot K(\mu, \sigma) \\ m_2 = \mu^2 + \sigma^2 + (\mu + a) \cdot \sigma \cdot K(\mu, \sigma) \end{cases} \tag{2}$$

There are some solutions in the literature: the method of successive approximations, the lookup table approach / using standardized statistical, etc.

# How to solve this set of equations?

- An analogous situation applies to random vectors truncated above a given threshold.
- Specifically, we obtain a nonlinear implicit system of equations, and the complexity of the system increases with the complexity of the threshold structure.
- In retrospect, the estimation with two distinct issues:
  1. On the one hand, estimators designed to be as accurate as possible will result in formulas that are relatively complex and computationally tedious.
  2. On the other hand, simplifying the estimator's formula may result in a less accurate estimation of the unknown parameters.
- Therefore, we can rely either on the optimization of computational complexity or on the optimization of estimation accuracy.

# An alternative for solving the set of equations

- But what if we introduced higher order moments?
- More equations, but simpler to solve!
- For the considered example of $Y = X \cdot 1_{X>a}$, where $X$ follows $N(\mu, \sigma^2)$

$$
\begin{aligned}
E(Y) &= \mu + \sigma K, && (3)\\
E(Y^2) &= \mu^2 + \sigma^2 + (\mu + a) \cdot \sigma K && (4)\\
E(Y^3) &= \mu^3 + 3\mu\sigma^2 + \left(a\mu + 2\sigma^2 + a^2 + \mu^2\right) \cdot \sigma K && (5)
\end{aligned}
$$

- and in consequence, this leads to relatively simple algebraic formulas for $\hat{\mu}$ and $\hat{\sigma^2}$ in terms of $m_k$ for $k = 1, 2, 3$.

# Truncated two dimensional distribution

- For a 2D case, the equations are longer. For a bivariate normal density function $(X, Y) \in \mathbb{R}^2$:

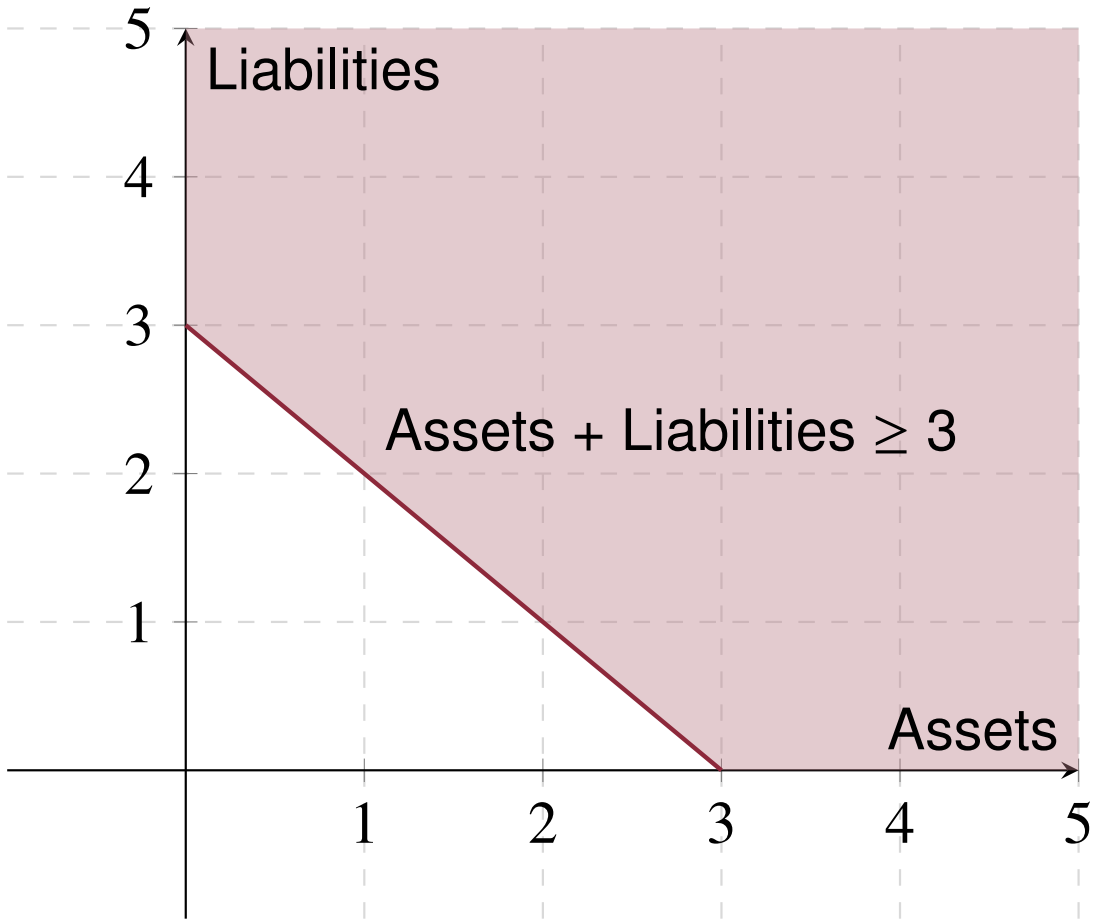
$$g(x, y) = \frac{(2\pi)^{-1}}{\sigma_x \cdot \sigma_y \cdot (1 - \varrho^2)^{1/2}} \cdot \exp\left(Formula\right) \tag{6}$$

$$Formula = \left( \frac{-1}{2(1 - \varrho^2)} \cdot \left[ \left( \frac{x - \mu_x}{\sigma_x} \right)^2 - 2\varrho \cdot \frac{x - \mu_x}{\sigma_x} \cdot \frac{y - \mu_y}{\sigma_y} + \left( \frac{y - \mu_y}{\sigma_y} \right)^2 \right]^2 \right) \tag{7}$$

- Truncation $X > a$, $Y > b$ leads to complicated moments.
- At first glance, it appears that the method of moments for a truncated normal vector ceases to be applicable due to computational complexity and the intricacy of the formulas.

# Balance of Payments – Trade Credits Statistics

- Balance of payments - assets and liabilities from trade credit
- Population above the reporting threshold (3 million PLN), which by definition creates a truncated set
- An alternative to costly census or survey methods



# Math for Trade Credits

- Consider a bivariate normal random vector $(X, Y)$ with a sample defined above a "triangle-shaped" threshold, described by the following relation: $\xi = \{(x, y) : x > 0, y > \max\{0, a - x\}\}$.
- Then, the moments of such a truncated vector will be related to the following integrals:

$$I_0 \equiv P(\xi) = \int_0^{\infty} \left[ \exp\left( -\frac{(x - \mu_x)^2}{2 \cdot (\sigma_x)^2} \right) \cdot \int_{\psi(x)}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-w^2/2} \, dw \right] dx$$

- the $k$-th moments of the variables $X$ and $Y$. However, these relationships are nonlinear, implicit, and do not possess explicit elementary forms.
- Another technical difficulty lies in evaluating the integral required to estimate the correlation coefficient $\varrho$.

# The Recursion Method as a Solution

- In the classical method of moments, it is necessary to compute the corresponding integrals and equate them to the successive $k$-th sample moments. However, in certain cases, evaluating these integrals becomes practically impossible.
- The solution we propose is based on finding an appropriate linear recurrence relation for successive moments, rather than computing the integrals and comparing them with the corresponding sample moments.
- Indeed, we note that the $k$-th moment of the variable $X$ truncated above the threshold $X > a$ takes the form:

$$m_k = \int_a^\infty x^k \cdot g(x)\,dx \Big/ \int_a^\infty g(x)\,dx \equiv I_k / I_0 \tag{8}$$

- and analogous relationships apply to variables defined above a triangle-shaped threshold.

# The Recursion Method as a Solution

- Next, we propose applying the Newton–Leibniz formula. The idea involves working with marginal densities:

$$g_X(x) = \frac{1_{x>0}}{P(\xi) \cdot \sqrt{2\pi} \cdot \sigma_x} \cdot \exp\left(-\frac{(x-\mu_x)^2}{2 \cdot (\sigma_x)^2}\right) \cdot \Phi\left(\frac{\mu_y + \min\{0, x-a\}}{\sigma_y \cdot \sqrt{1-\varrho^2}} + \frac{\varrho \cdot (x-\mu_x)}{\sigma_x \cdot \sqrt{1-\varrho^2}}\right)$$

- However, complication of the above force us to use the product rule for three functions, $(abc)' = a' \cdot bc + ab' \cdot c + ab \cdot c'$, and we obtain a **linear** recurrence:

$$I_{k+6} = c_5 \cdot I_{k+5} + ... + c_0 \cdot I_k$$

  where $c_0 = \mu_x$, $c_1 = \sigma_x^2$, and $c_2, \ldots, c_5$ are coefficients depending on the unknown parameters $(\mu_x, \sigma_x^2)$. Crucially, this yields a linear equation with respect to $(c_0, \ldots, c_5)$ i.e. the core idea of our method.

- The **benefit** of this approach is as follows: in the method of moments, we transform nonlinear systems of equations into higher-order linear recurrence systems. This allows for a direct comparison between theoretical moments and sample moments defined above a given threshold.

# Solution

Then, for example, for $k = 0$ we obtain:

$$\frac{I_6}{I_0} = c_5 \cdot \frac{I_5}{I_0} + \cdots + c_1 \cdot \frac{I_1}{I_0} + c_0$$

in other words, for the variable $X$ defined above the threshold, we arrive at a relationship between the moments of this variable:

$$m_6 = c_5 \cdot m_5 + \cdots + c_1 \cdot m_1 + c_0$$

and in this final equation, we can now substitute the sample moments computed above the threshold.
On the other hand, the estimates for the correlation coefficient $\varrho$ will be constructed by considering the recurrence for integrals of the form:

$$E(X^k \cdot Y) = \iint\limits_{\xi} x^k \cdot y \cdot \frac{g(x,y)}{P(\xi)}\, dxdy$$

where $\xi = \{(x,y) : x > 0, y > \max\{0, a - x\}\}$.

# Results and Conclusions

- The primary result of the work is a formalized algorithm for reconstructing a full statistical distribution based on its truncated fragment.
- The information gap created by reporting thresholds can be systematically closed with an acceptable statistical error.
- It was observed that the method of moments combined with the implicit function theorem effectively determines the cumulative distribution function values for sub-threshold entities while maintaining a low mean squared error.
- The quality of the results and the stability of the estimators can be assessed by comparing the results with data from periodic full censuses (if available) or through sensitivity tests regarding the truncation threshold.

# Thank you for your attention!

Questions?