Czesław Domański
Uniwersytet Łódzki

ROZKŁAD LAMBDA-TUKEY'A I PRÓBA JEGO ZASTOSOWANIA*

Wprowadzenie

W literaturze przedmiotu prezentowane są różnorodne rozkłady empiryczne, z których do najważniejszych należą: system krzywych K. Pearsona (1894), zawarty w pracy Pearsona (1948, por. także Domański, Pruska, 2000), system Johnsona przedstawiony w pracy Hahna i Shapiro (1967), rozkład Burra (1973) czy rozkład Tukey'a (1960).

W artykule przedstawiony będzie rozkład Lambda-Tukey'a z czterema parametrami, pozwalający na prezentację wielu różnorodnych kształtów krzywych. Zamieszczono także fragmenty tablic wartości parametrów opracowane dla tego rozkładu, które ułatwiają szacowanie jego parametrów.

Do innych ważnych zastosowań prezentowanego rozkładu należy generowanie liczb losowych dla badań symulacyjnych oraz analiz Monte Carlo sprawdzających odporność procedur statystycznych.

1. Uogólnienie rozkładu λ Tukey'a

Z reguły ciągły rozkład prawdopodobieństwa definiuje się za pomocą dystrybuanty lub funkcji gęstości. Alternatywnie można go określić przez funkcję kwantylową (percentylową). Ujmując to najprościej, funkcja kwantylowa jest funkcją odwrotną do dystrybuanty.

Badania nad uogólnionym rozkładem λ Tukey'a prowadzili m.in. Ramberg, Tadikamalla, Dudkiewicz, Mykytka (1979). Prezentowany przez tych autorów rozkład jest czteroparametrowy, uwzględniający parametry: położenia, skali, skośności i kurtozy.

* Praca napisana w ramach projektu sfinansowanego ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/01/B/HS4/02746.

Szczególnym przypadkiem funkcji kwantylowej jest funkcja λ Tukey'a (1960):

$$R(p) = \frac{p^{\lambda} - (1-p)^{\lambda}}{\lambda}, \quad 0 \le p \le 1 \tag{1}$$

określona dla wartości $\lambda \neq 0$.

Jeżeli $\lambda \to 0$, to otrzymamy rozkład logistyczny.

Ramberg i Schmeiser (1974) przedstawili rozkład (1) z czterema parametrami danego funkcją kwantylową postaci:

$$R(p) = \lambda_1 + \frac{p^{\lambda_3} - (1-p)^{\lambda_4}}{\lambda_2}, \quad 0 \le p \le 1 \tag{2}$$

gdzie:

λ_1 – parametr położenia,

λ_2 – parametr skali,

λ_3 – parametr skośności,

λ_4 – parametr kurtozy.

Funkcja gęstości odpowiadająca (2) dana jest wzorem:

$$f(x) = f[R(p)] = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}}, \quad 0 \le p \le 1 \tag{3}$$

Wyznaczenie funkcji gęstości dla ustalonych parametrów $\lambda_1, \lambda_2, \lambda_3 \, i \, \lambda_4$ wymaga znalezienia wartości (2) i (3) dla argumentu p z przedziału [0,1]. Następnie nanosi się wartości $f[R(p)]$ na osi Y względem wartości *R(p)* odłożonych na osi X.

Rozkład ten, którego szczególnym przypadkiem jest oryginalny rozkład λ pozwala uzyskać również skośne krzywe. Zauważmy, że dystrybuanta tego rozkładu nie występuje w postaci jawnej.

Wzory na wartość oczekiwaną, wariancję oraz współczynnik skośności i kurtozy uogólnionego rozkładu λ dane są wzorami:

$$\mu \equiv E(X) = aE(Y) + b = \lambda_1 + \frac{1}{\lambda_2}\left(\frac{1}{\lambda_3 + 1} - \frac{1}{\lambda_4 + 1}\right)$$

$$\sigma^2 \equiv E(X - E(X))^2 = a^2 E(Y - E(Y))^2 = \frac{1}{\lambda_2^2}(A_2 - A_1^2)$$

$$\beta_3 \equiv \frac{1}{\sigma^3} E(X - E(X))^3 = \frac{\lambda_2^3}{(A_2 - A_1^2)^{\frac{3}{2}}} \cdot \frac{1}{\lambda_2^3}(A_3 - 3A_1A_2 + 2A_1^3) \tag{4}$$

$$\beta_4 \equiv \frac{1}{\sigma^4}(E(X - E(X))^4) = \frac{1}{(A_2 - A_1^2)^2}(A_4 - 4A_1A_3 + 6A_1^2A_2 + 3A_1^4)$$

gdzie:

$$A_1 = \sum_{j=0}^{1}(-1)^j\binom{1}{j}\beta(\lambda_3(1-j)+1,\lambda_4 j+1) = \beta(\lambda_3+1,1)-\beta(1,\lambda_4+1) = \frac{1}{\lambda_3+1}-\frac{1}{\lambda_4+1}$$

$$A_2 = \sum_{j=0}^{2}(-1)^j\binom{2}{j}\beta(\lambda_3(2-j)+1,\lambda_4 j+1) = \beta(2\lambda_3+1,1)-2\beta(\lambda_3+1,\lambda_4+1)+\beta(1,2\lambda_4+1) = \tag{5}$$

$$= \frac{1}{2\lambda_3+1}+\frac{1}{2\lambda_4+1}-2\beta(\lambda_3+1,\lambda_4+1)$$

$$A_3 = \sum_{j=0}^{3}(-1)^j\binom{3}{j}\beta(\lambda_3(3-j)+1,\lambda_4 j+1) = \beta(3\lambda_3+1,1)-3\beta(2\lambda_3+1,\lambda_4+1)+3\beta(\lambda_3+1,2\lambda_4+1)-$$

$$-\beta(1,3\lambda_4+1) = \frac{1}{3\lambda_3+1}-\frac{1}{3\lambda_4+1}-3\beta(2\lambda_3+1,\lambda_4+1)+3\beta(\lambda_3+1,2\lambda_4+1)$$

$$A_4 = \sum_{j=0}^{4}(-1)^j\binom{4}{j}\beta(\lambda_3(4-j)+1,\lambda_4 j+1) = \beta(4\lambda_3+1,1)-4\beta(3\lambda_3+1,\lambda_4+1)+6\beta(2\lambda_3+1,2\lambda_4+1)-$$

$$-4\beta(\lambda_3+1,3\lambda_4+1)+\beta(1,4\lambda_4+1) = \frac{1}{4\lambda_3+1}+\frac{1}{4\lambda_4+1}-4\beta(3\lambda_3+1,\lambda_4+1)+6\beta(2\lambda_3+1,2\lambda_4+1)-$$

$$-4\beta(\lambda_3+1,3\lambda_4+1)$$

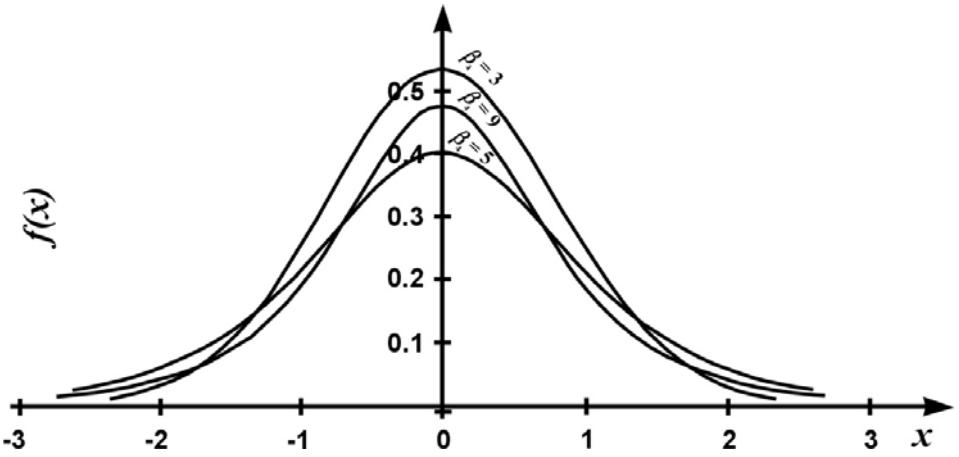
przy czym β oznacza funkcję beta.

Stąd *k*-ty moment można otrzymać, gdy $\min(\lambda_3,\lambda_4)\rangle\frac{1}{k}$.

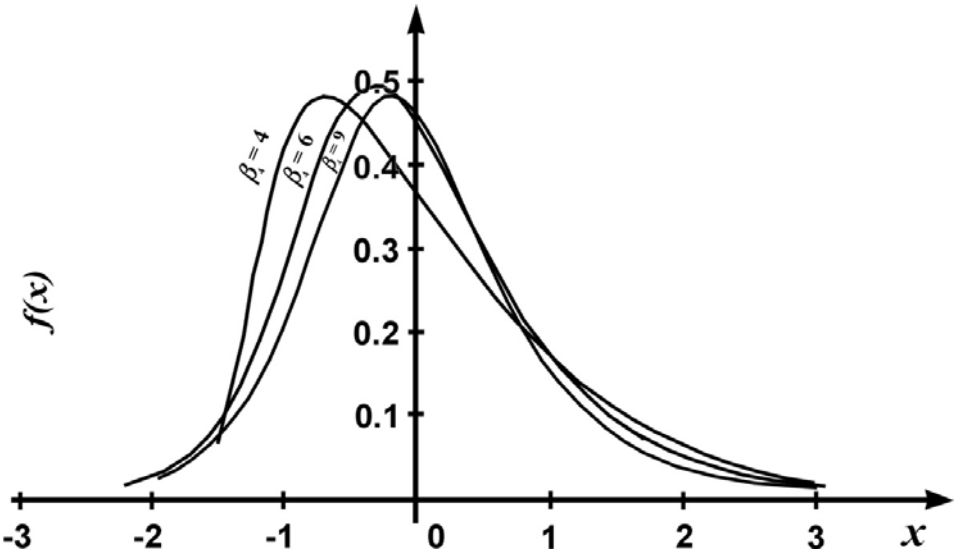
Wartość ta zależy tylko od parametrów $\lambda_3 i \lambda_4$, w konsekwencji współczynniki skośności i kurtozy również zależą tylko od tych parametrów.

Prezentowany rozkład z czteroma parametrami pozwala uzyskać wiele różnorodnych kształtów krzywych, co zostało pokazane na rys. 1-3. Na rys. 1 przedstawiona została funkcja gęstości dla parametrów β_3 = 0 oraz β_4 = 3,5,9, natomiast na rys. 2 – dla β_3 = 1 oraz β_4 = 1,6,9, a na rys. 3 dla parametrów β_3 = 0,0.5,1 oraz β_4 = 4.

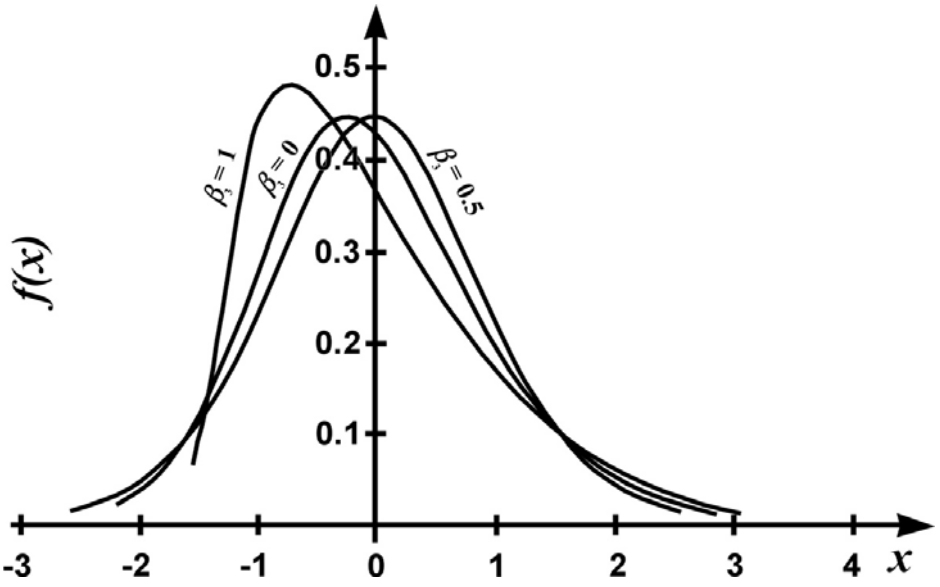
Ramberg, Dudewicz, Tadikamalla i Mykytka (1979) przedstawili tablice wartości $\lambda_1, \lambda_2, \lambda_3 i \lambda_4$ dla wybranych parametrów β_3 i β_4 oraz dla μ = 0 i σ = 1. Wielkości zamieszczone w tych tablicach zostały uwzględnione przy konstrukcji rozkładów prezentowanych na rys. 1-3.



Rys. 1. Funkcja gęstości dla $\beta_3 = 0$ oraz $\beta_4 = 3,5,9$



Rys. 2. Funkcja gęstości dla $\beta_3 = 1$ oraz $\beta_4 = 4,6,9$



Rys. 3. Funkcja gęstości dla $\beta_3 = 0,0.5,1$, $\beta_4 = 4$

Wartości A_k dla k = 1,2,3,4 (por. wzór 5) zależą tylko od parametrów $\lambda_3 \, i \, \lambda_4$, stąd współczynniki skośności i kuriozy zależą tylko od tych parametrów.

Parametry λ_i uogólnionego rozkładu λ Tukey'a obliczamy z równań:

$$\begin{aligned} &\mu = 0 \\ &\sigma^2 = 1 \\ &\beta_3 = \beta_3^* \\ &\beta_4 = \beta_4^* \end{aligned} \tag{6}$$

gdzie $\beta_3^* \, i \, \beta_4^*$ są obliczone na podstawie wyników z próby.

Uwzględniając wzory (4), otrzymujemy:

$$\begin{cases} \lambda_1 + \dfrac{1}{\lambda_2}\left(\dfrac{1}{\lambda_3 + 1} - \dfrac{1}{\lambda_4 + 1}\right) = 0 \\ \dfrac{1}{\lambda_2^2}(A_2 - A_1^2) = 1 \\ \dfrac{1}{(A_2 - A_1^2)^{\frac{3}{2}}}(A_3 - 3A_1A_2 + 2A_1^3) = \beta_3^* \\ \dfrac{1}{(A_2 - A_1^2)^2}(A_4 - 4A_1A_3 + 6A_1^2A_2 - 3A_1^4) = \beta_4^* \end{cases} \tag{7}$$

Do równań (7) podstawimy parametry λ_i (i = 1,2,3,4) z tablicy 4 artykułu Ramberga i in. (1979).

Na podstawie programu „Mathematica"* lewe strony równań (6) oznaczone są literami f, g, h i w.

Po podstawieniu parametrów λ powinno się otrzymać:

$$f = 0$$

$$g = 1$$

$$h = \beta_3$$

$$w = \beta_4$$

Obliczenia wykonane są dla 4 przypadków:

1. Podstawiamy $\lambda_1 = -1{,}245, \lambda_2 = 0{,}2445, \lambda_3 = 0{,}0178, \lambda_4 = 0{,}4748$

W wyniku otrzymujemy:

$$\begin{cases} f = 0{,}0002007596 \\ g = 0{,}999776 \\ h = 0{,}500145 \\ w = 2{,}40007 \end{cases}$$

W cytowanych tablicach dla wybranych parametrów λ $\beta_3 = 0{,}5$ $\beta_4 = 2{,}4$.

2. $\lambda_1 = -0{,}045, \lambda_2 = -0{,}1198, \lambda_3 = -0{,}0569, \lambda_4 = -0{,}0617$

Wyniki:

$$\begin{cases} f = 0{,}000277763 \\ g = 0{,}999848 \\ h = -0{,}148876 \\ w = 5{,}39963 \end{cases}$$

W tablicach $\beta_3 = 0{,}15, \beta_4 = 5{,}4$.

3. $\lambda_1 = -0{,}134, \lambda_2 = -0{,}2501, \lambda_3 = -0{,}0977, \lambda_4 = -0{,}1242$

Wyniki:

$$\begin{cases} f = 0{,}0000837921 \\ g = 0{,}99988 \\ h = -0{,}651818 \\ w = 8{,}20468 \end{cases}$$

* Obliczenia zostały wykonane przez dr Katarzynę Bolonek-Lasoń.

W tablicach $\beta_3 = 0{,}65, \beta_4 = 8{,}2.$

4. $\lambda_1 = -0{,}499, \lambda_2 = 0{,}1497, \lambda_3 = 0{,}0538, \lambda_4 = 0{,}1438$

Wyniki:

$$\begin{cases} f = -0{,}000216105 \\ g = 1{,}00005 \\ h = 0{,}550225 \\ w = 3{,}40036 \end{cases}$$

W tablicach $\beta_3 = 0{,}55, \beta_4 = 3{,}4.$

W przedstawionych przypadkach otrzymujemy wyniki zgodne z wartościami z tablicy 4 artykułu Ramberga i in. (1979).

2. Przykłady zastosowań dla indeksów giełdowych

Dane empiryczne dotyczą tygodniowych notowań indeksu DAX z okresu 03.01.1997-27.07.2012 (813 obserwacji, por. rys. 4). Na podstawie tych danych wyznaczamy parametry rozkładu:

$\mu = 5391{,}972$

$\sigma = 1338{,}23$

$$\beta_3 = \frac{\mu_3}{\sigma^3} = -0{,}05$$

$$\beta_4 = \frac{\mu_4}{\sigma^4} = 2{,}14$$

Z tab. 1 dla $\beta_3 = 0{,}05$ i $\beta_4 = 2{,}2$ odczytujemy $\lambda_1 = -0{,}802$, $\lambda_2 = 0{,}3314$, $\lambda_3 = 0{,}1128$, $\lambda_4 = 0{,}5802$. Przekształcamy wielkości parametrów λ_1 i λ_2 według wzorów (uwzględniamy wartość bezwzględną ze względu na to, że wartości λ_1 i λ_2 w tab. 1 podane są dla zmiennej o wartości oczekiwanej zero i wariancji jeden):

$$\lambda_1(\mu,\sigma) = \lambda_1(0{,}1)\sigma + \mu = -0{,}802 \cdot 1338{,}23 + 5391{,}972 = 4318{,}7$$

$$\lambda_2(\mu,\sigma) = \lambda_2(0{,}1)/\sigma = 0{,}3314/1338{,}23 = 0{,}00025$$

gdzie μ i σ to średnia i odchylenie standardowe obliczone na podstawie danych empirycznych.

Zmienna X oraz odpowiadająca jej funkcja gęstości przyjmuje postać:

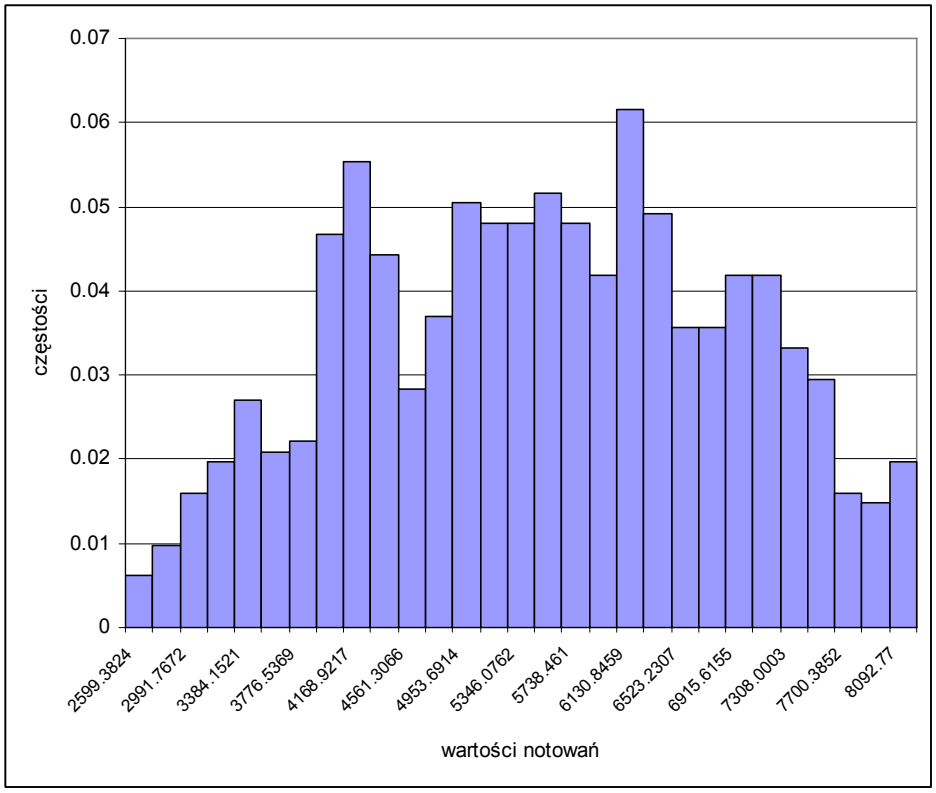
$$X = \frac{\lambda_1 + p^{\lambda_3} + (1-p)^{\lambda_4}}{\lambda_2}$$
$$f(x) = \frac{\lambda_2}{\lambda_3 p^{\lambda_3 - 1} + \lambda_4 (1-p)^{\lambda_4 - 1}} \tag{8}$$

Tabela 1

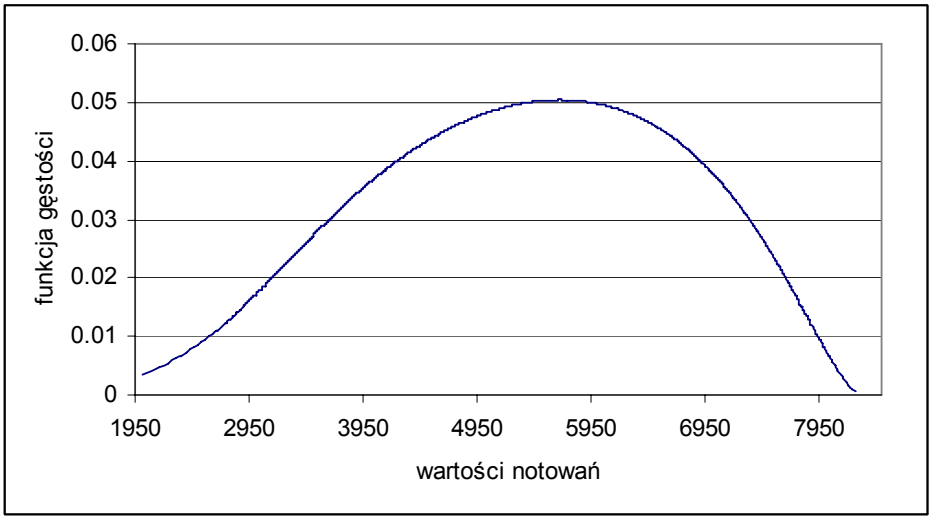
Wybrane wartości parametrów $\lambda_1, \lambda_2, \lambda_3 \, i \, \lambda_4$ dla współczynników skośności β_3 = 0.0,0.05,1 kurtozy β_4 = 1.0, …,9.0 *gdy* μ = 0 i σ = 1

$\beta_3 = 0.0$				
β_4	λ_1	λ_2	λ_3	λ_4
1.8	0.0	.5774	1.0000	1.0000
2.0	0.0	.4952	.5843	.5843
2.2	0.0	.4197	.4092	.4092
2.4	0.0	.3533	.3032	.3032
2.6	0.0	.2949	.2303	.2303
2.8	0.0	.2433	.1765	.1765
3.0	0.0	.1974	.1349	.1349
4.0	0.0	.0262	.0148	.0148
5.0	0.0	-.0676	-.0443	-.0443
6.0	0.0	-.1686	-.0802	-.0802
7.0	0.0	-.2306	-.1045	-.1045
8.0	0.0	-.2800	-.1223	-.1223
9.0	0.0	-.3203	-.1359	-.1359
$\beta_3 = 0.05$				
β_4	λ_1	λ_2	λ_3	λ_4
1.8	-1.703	.2861	.0000	.9502*
2.0	-1.229	.3122	.0505	.7603
2.2	-.802	.3314	.1128	.5802
2.4	-.375	.3328	.1876	.3941
2.6	-.143	.2924	.1973	.2605
2.8	-.083	.2429	.1625	.1903
3.0	-.059	.1975	.1276	.1425
4.0	-.026	.0264	.0146	.0153
5.0	-.016	-.0867	-.0435	-.0448
6.0	-.013	-.1682	-.0791	-.0810
7.0	-.011	-.1034	-.1034	-.1054
8.0	-.928+	-.2797	-.1212	-.1232
9.0	-.837+	-.3201	-.1348	-.1368
$\beta_3 = 1.00$				
β_4	λ_1	λ_2	λ_3	λ_4
3.4	-1.253	.1772	.0000*	.2854*
4.0	-.886	.1333	.0193	.1588
5.0	-.533	.0340	.9695+	.0285
6.0	-.379	-.0562	-.0187	-.0388
7.0	-.297	-.1291	-.0453	-.0790
8.0	-.248	-.1878	-.0670	-.1058
9.0	-.215	-.2356	-.0844	-.1249

Źródło: Na podstawie (Ramberg i in., 1979).



Rys. 4. Histogram wartości notowań indeksu DAX w latach 1997-2012



Rys. 5. Funkcja gęstości wyznaczona na podstawie równań (6) odpowiadająca notowaniom indeksu DAX

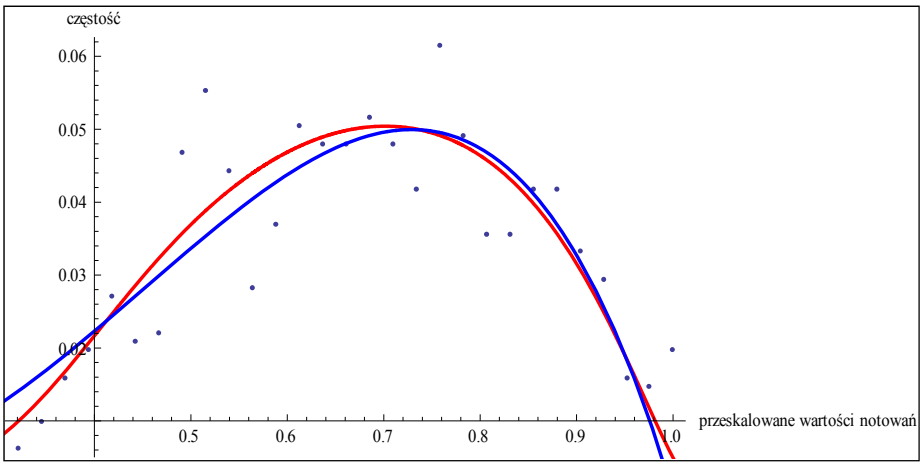
Wartości trzeciego i czwartego momentu danych empirycznych znajdują się w obszarze rozkładu beta, zatem w programie „Mathematica" dopasowujemy ten rozkład metodą najmniejszych kwadratów do danych empirycznych (rys. 6), otrzymując parametry rozkładu beta:

```
{BestFitParameters→ {a→3.64419, b→1.98305, c→2.50873},

                        |  Estimate  Asymptotic SE  CI
ParameterCITable→       a | 3.64419   0.452338      {2.7144, 4.57399}
                        b | 1.98305   0.190936      {1.59058, 2.37552}
                        c | 2.50873   0.132889      {2.23557, 2.78189}

EstimatedVariance→ 0.745372,

                                  DF     SumOfSq     MeanSq
                    Model         3      388.158     129.386
ANOVATable→         Error         26     19.3797     0.745372,
                    Uncorrected Total    29     407.538
                    Corrected Total      28     62.7104
```

$$f(x)=\frac{2{,}50873}{\beta(3{,}64419;\ 1{,}98305)}x^{2{,}64419}(1-x)^{0{,}98305}$$

Punkty na rys. 6 oznaczają dane empiryczne, czyli częstości występowania zmiennej w każdym przedziale, niebieska krzywa prezentuje funkcję gęstości rozkładu beta, natomiast czerwona krzywa przedstawia funkcję gęstości wyznaczoną na podstawie funkcji kwantylowej.

Stopień dopasowania rozkładu beta do danych empirycznych mierzony współczynnikiem determinacji wynosi około 74%, dla rozkładu wyznaczonego na podstawie funkcji kwantylowej współczynnik ten wynosi 76%.



Rys. 6. Funkcja gęstości rozkładu beta oraz rozkładu wyznaczonego na podstawie funkcji kwantylowej

Nie ma podstaw do odrzucenia hipotezy zerowej o zgodności rozkładu empirycznego z rozkładem wyznaczonym na podstawie funkcji kwantylowej (wartość statystyki testowej $\chi^2 = 30{,}53$, wartość krytyczna dla poziomu istotności α = 0,05 wynosi $\chi^2_{24} = 36{,}415$). Dla rozkładu beta test zgodności χ^2 odrzuca hipotezę zerową o zgodności tego rozkładu z rozkładem empirycznym (wartość statystyki testowej $\chi^2 = 38{,}85$).

Podsumowanie

Omawiany rozkład pozwala uzyskać szeroką gamę kształtów krzywych, które jako najprostsze przykłady pokazane są na rys. 1-3. Ze względu na wysoką elastyczność tego rozkładu znajduje on wiele różnorodnych zastosowań w przypadku, gdy rzeczywisty rozkład nie jest znany.

Wielu autorów zajmowało się badaniami własności rozważanego rozkładu (por. np. Chalabi, Scott, Wuertz, 2012). Literatura z tego zakresu jest stosunkowo bogata, co świadczy o dużych możliwościach zastosowań uogólnionego rozkładu λ. Tristano (2010) prezentuje np. uogólniony rozkład λ z pięcioma parametrami, który w dalszych badaniach autora będzie rozważany.

Bibliografia

Burr I.W. (1973): *Parameters for a General System of Distributions to Match a Grid of* α_3 *and* α_4. Comm. Statist., 2,1-21.

Chalabi Y., Scott D.J., Wuertz D. (2012): *An Asymmetry-Steepness Parameterization of the Generalized Lambda Distribution*, http://mpra.ub.uni-muenchen.de/37814.

D'Addaro R. (1949): *Ricerche sulla* curva dei redditi. „Giornale degli Economisti e Annali di Economia", 8, s. 91-114.

Domański Cz., Pruska K. (2000): *Nieklasyczne metody statystyczne*. Polskie Wydawnictwo Ekonomiczne, Warszawa.

Edgeworth F.Y. (1898): *On the Representation of Statistics by Mathematical Formule*. „Journal of the Royal Statistical Society", 1, s. 670-700.

Hahn G.J., Shapiro S.S. (1967): *Statistical Models in Engineering*. John Wiley & Sons, New York.

Johnson N.L. (1949): *Systems of Frequency Curves Generated by Methods of Translation*, „Biometrika" 44, s. 147-176.

Pearson K. (1894): *Contributions* to the *Mathematical Theory* of Evolution. Transactions of the Royal Society, 184. W: K. Pearson (1948), s. 1-40.

Pearson K. (1895): *Contributions* to the *Mathematical Theory* of Evolution. II Skew Variation in Homogeneous Material Philosophical, 186. W: K. Pearson (1948): s. 41-112.

Pearson K. (1948): *Karl Pearson's Early Statistical Papers*. Cambridge University Press.

Ramberg J.S., Schmeister B.W. (1974): *An Approximate Method for Generating Asymmetric Random Variables*. Canon. ACM, 17, s. 78-82.

Ramberg J.S., Tadikamalla P.R., Dudewicz E.J., Mykytka E.F., (1979): *A Probability Distribution and its Uses in Fitting Data*. „Technometries 21", No. 2, s. 201-214.

Tarsitano A. (2010): *Comparing Estimation Methods for the FPLD*. „Journal Probability and Statistics", Vol. 1, No. 1, s. 1-16.

Tukey J.W. (1960): *The Practical Relationship Between the Common Transformations of Percentages of Counts and of Amounts*. Technical Report 36, Statistical Techniques Research Group, Princeton University.

LAMBDA-TUKEY DISTRIBUTION AND APPLICATION ATTEMPT

Summary

In the article the generalized Lambda-Tukey distribution was presented with the following four parameters of: location, scale, skewness and kurtosis.

The distribution presented, due to its high flexibility is widely applied, especially when empirical distributions are sophisticated and do not show desired accordance with known classical theoretical distributions.

The examples presented on the fitting of the DAX index distribution to the four parameter Tukey distribution turn out to be better than the ones for the beta distribution.