

# Pomiar nierówności w badaniach reprezentacyjnych z wykorzystaniem pakietu wINEQ

Dr Sebastian Wójcik  
Agnieszka Giemza  
Jarosław Napora

# Agenda

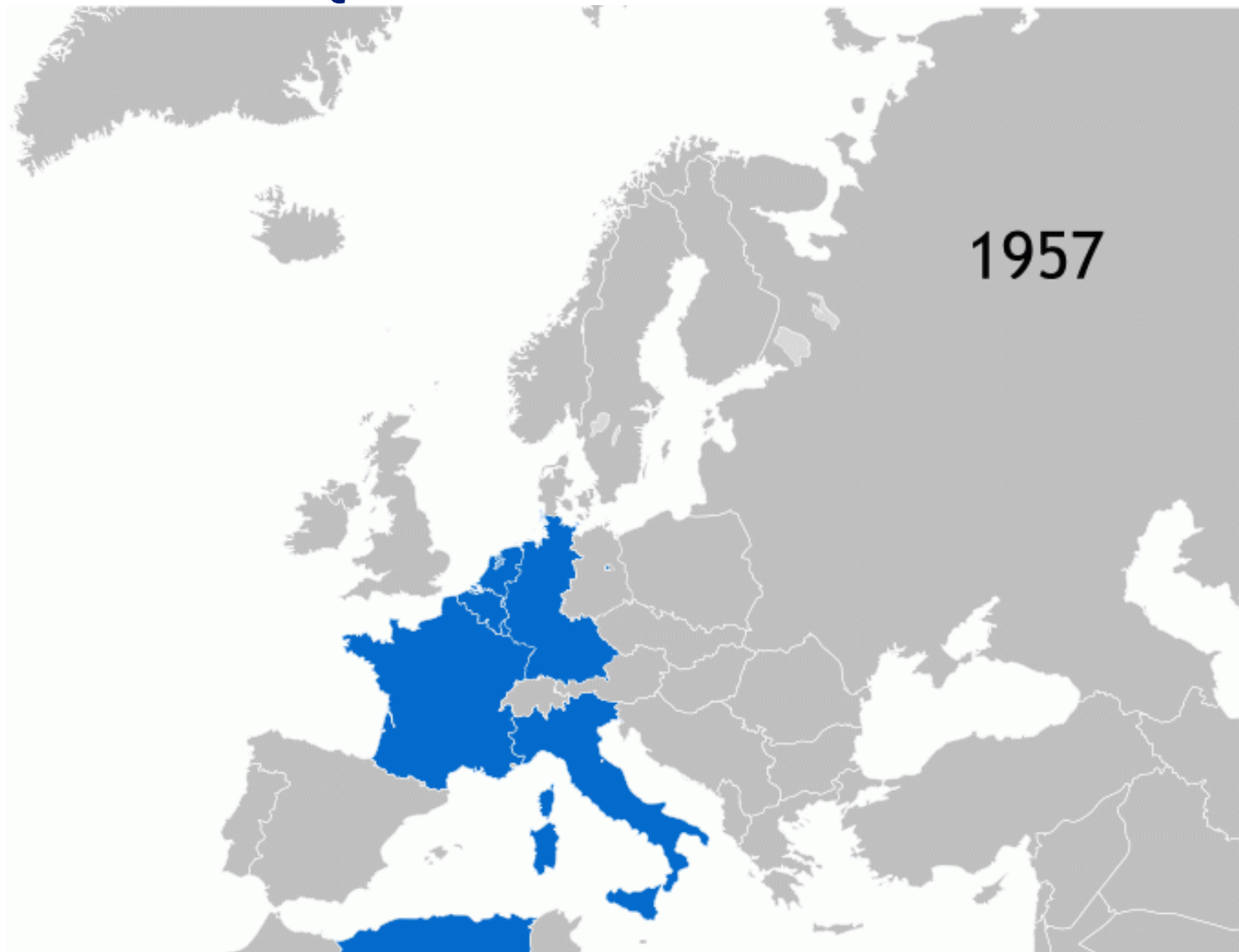


Pojęcie nierówności i jego aspekty

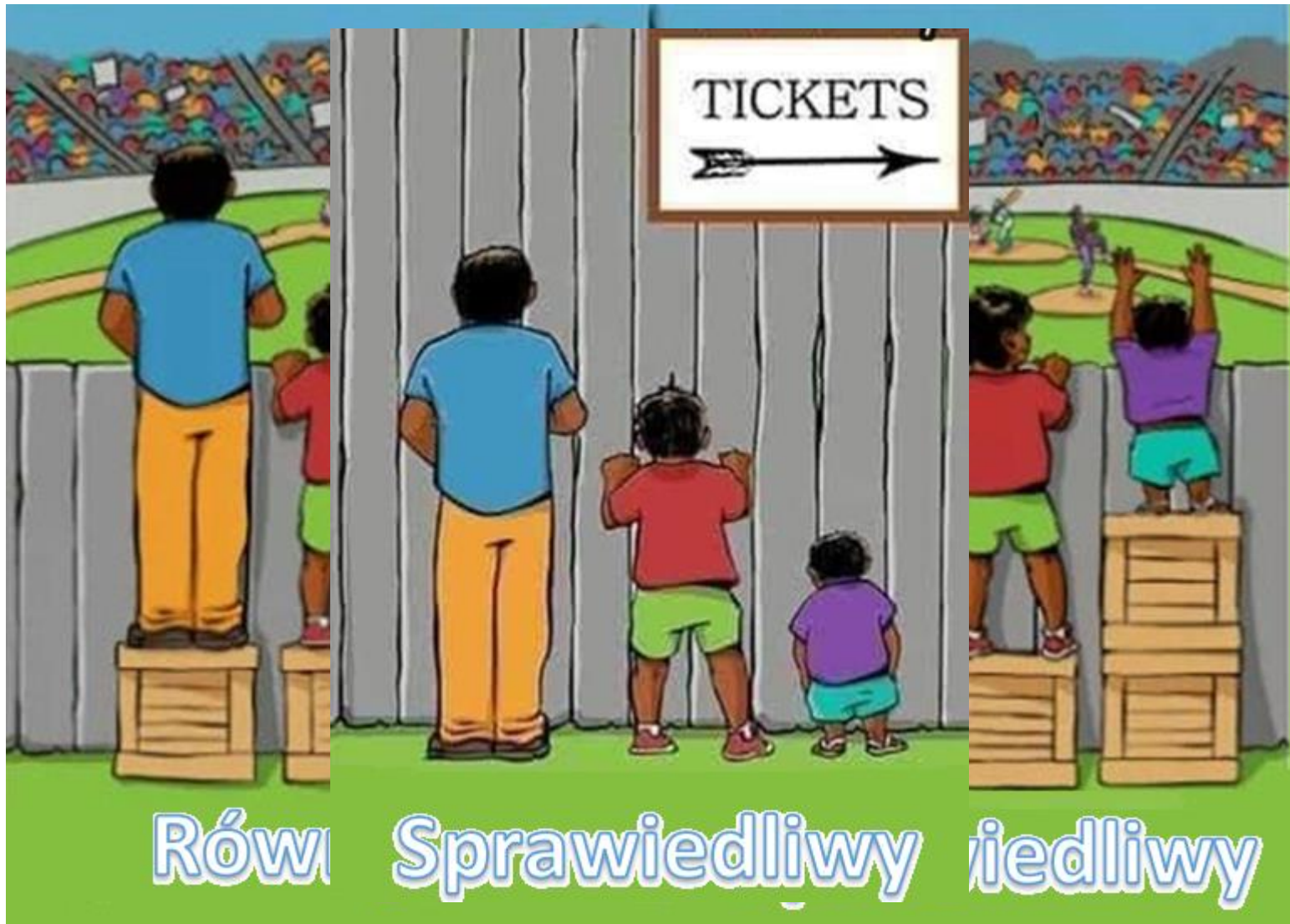
Pakiet wINEQ

Charakterystyka nierówności społecznych na podstawie  
Badania Budżetów Gospodarstw Domowych

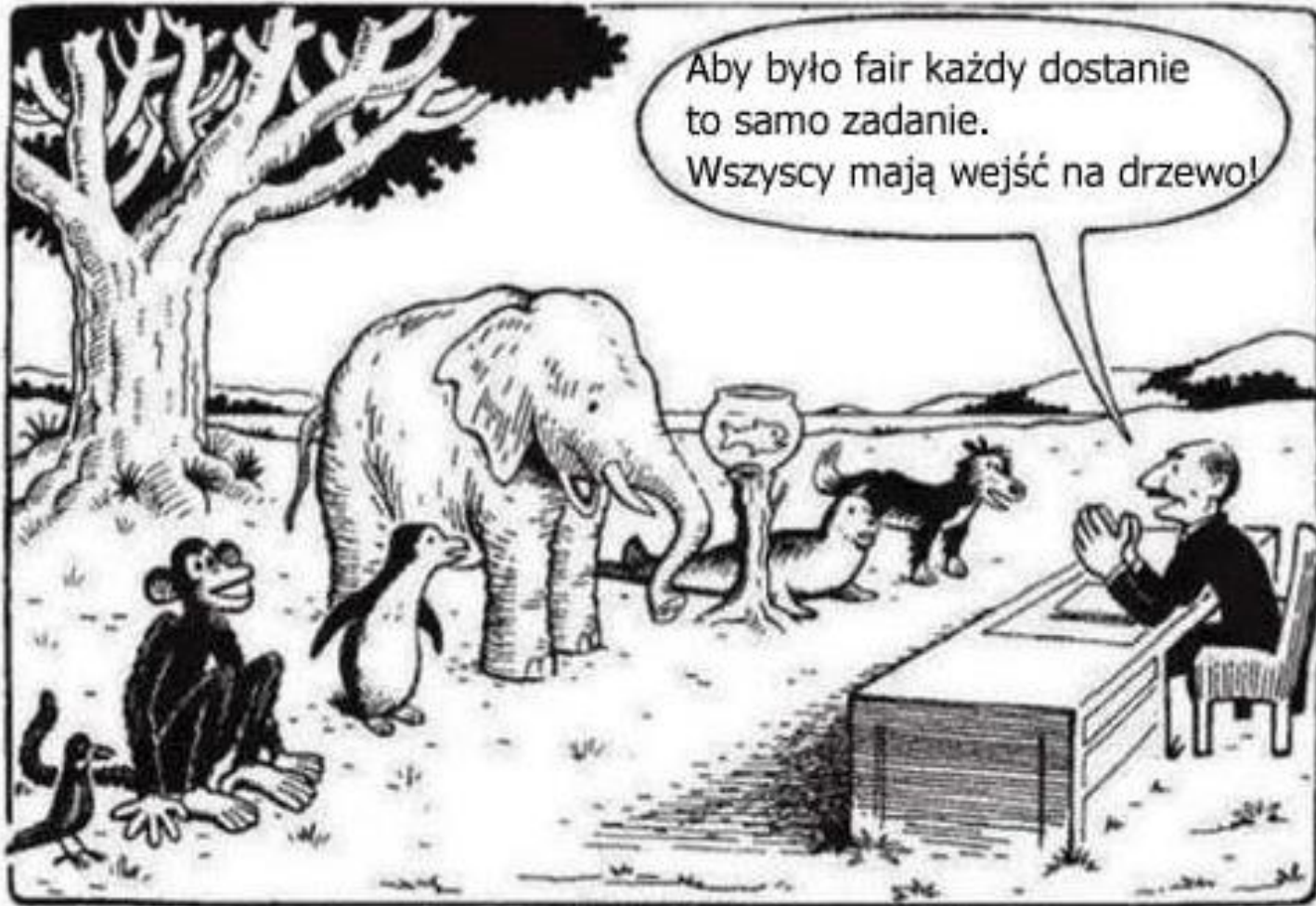
# Jednoczenie się a nierówności



# Nierówności a sprawiedliwość



# Równy $\neq$ sprawiedliwy?



# Nierówność w statystyce

Temat nierówności często pojawia się w kontekście:

- Poziomu rozwoju gospodarczego krajów i regionów
- Rozkładu dochodów obywateli krajów
- Ubóstwa
- Różnic w dochodach względem płci czy regionu

Podejście do badania nierówności

- Testy statystyczne (np. test dwóch średnich, dwóch wariancji czy test dwóch frakcji)
- Miary nierówności

# Badania reprezentacyjne – podstawowe działania statystyki

Już w pierwszych artykułach Ustawy o statystyce publicznej czytamy, że „Statystyka publiczna zapewnia **rzetelne, obiektywne** i systematyczne informowanie społeczeństwa, organów państwa i administracji publicznej oraz podmiotów gospodarki narodowej **o sytuacji ekonomicznej, demograficznej, społecznej oraz środowiska naturalnego**” (art. 3)

Pozyskiwanie informacji z każdej dziedziny życia społecznego i gospodarczego wymaga przeprowadzania **badania reprezentacyjnych**.



# Badania reprezentacyjne – podstawowe działania statystyki

- Integralną częścią badań reprezentacyjnych jest błąd próbkowania. Ma on silny wpływ na wiarygodność dostarczanych szacunków pod względem precyzji.
- Wyniki badań reprezentacyjnych są niekompletne bez ich precyzji lub przedziału ufności.
- Metody stosowane do przetwarzania danych z badań próbkowych muszą umożliwiać ważenie danych.
- Ważenie danych jest ważnym elementem w obliczaniu precyzji szacunków.



# Pakiet - historia

- Nierówności dochodowe, nierówność szans czy ubóstwo są ważnymi pojęciami w ekonomii.
- W praktyce stosowanych jest wiele różnych miar nierówności dla różnych skal pomiarowych.
- Wśród różnych pakietów R do analizy danych, nie wszystkie mają zaimplementowane metody odpowiednie dla danych ważonych.
- W przypadku analizy nierówności brakuje pakietów, które oferują szerokie spektrum miar nierówności, a także ważenie danych i pozwalają na wyznaczenie precyzji danej miary nierówności.
- Doprowadziło nas to do pomysłu stworzenia pakietu opartego na wykorzystaniu i rozszerzeniu funkcji w dostępnych w innych pakietach (**ineq**) oraz dodanych nowych funkcjonalności.

# Pakiet - historia

Naszą przygodę z zagadnieniem nierówności w odniesieniu do obliczania miar nierówności rozpoczęliśmy od przeszukania bazy pakietów CRAN-R (The Comprehensive R Archive Network).

Pakiety, które odpowiadały po części interesującym nas zagadnieniom to:

- **laeken**: Estimation of Indicators on Social Exclusion and Poverty - udostępnia m.in. funkcje do szacowania popularnych wskaźników zagrożenia ubóstwem i nierówności.
- **ineq**: Measuring Inequality, Concentration, and Poverty - oblicza różne miary nierówności (między innymi Gini, Theil, entropie), miary koncentracji (Herfindahl, Rosenbluth).

Pakiet **ineq** zawiera wiele interesujących nas miar nierówności, ale nie jest przeznaczony do bezpośredniego uwzględniania wag dla rekordów. Sugeruje się by je „uwzględnić” poprzez rozmnożenie rekordów.

# INEQ

Pakiet **ineq** zawiera wyniki badania reprezentatywnego budżetów domowych w Filipinach. Podane w pakiecie przykłady zastosowania funkcji nie uwzględniają ważenia danych pomimo, że są one dostępne we wspomnianym zbiorze.

Obliczony wskaźnik Theila bez ważenia

```
> start_time <- sys.time()
>   ineq::Theil(Ilocos$income)
[1] 0.3199159
> end_time <- sys.time();end_time - start_time
Time difference of 0.02000308 secs
```

oraz poprawny wynik przy rozmnożeniu rekordów

```
> start_time <- sys.time()
>   ineq::Theil(rep(Ilocos$income,Ilocos$AP.weight))
[1] 0.3163566
> end_time <- sys.time();end_time - start_time
Time difference of 0.1920581 secs
```

Czas działania takiego rozwiązania jest jednak stosunkowo bardzo długi

# Ineq vs wINEQ

```
> start_time <- Sys.time()
> ineq::Theil(rep(Ilocos$income,Ilocos$AP.weight))
[1] 0.3163566
> end_time <- Sys.time();end_time - start_time
Time difference of 0.1920581 secs
```

95,97 razy szybciej

```
> start_time <- Sys.time()
> wINEQ::Theil_T(Ilocos$income,Ilocos$AP.weight)
[1] 0.3163566
> end_time <- Sys.time();end_time - start_time
Time difference of 0.002001047 secs
```

```
> start_time <- Sys.time()
> for (i in 1:1000) {
+   ineq::Theil(rep(Ilocos$income,Ilocos$AP.weight))
+ }
> end_time <- Sys.time();end_time - start_time
Time difference of 3.699362 mins
```

1239,725 razy szybciej

```
> start_time <- Sys.time()
> for (i in 1:1000) {
+   wINEQ::Theil_T(Ilocos$income,Ilocos$AP.weight)
+ }
> end_time <- Sys.time();end_time - start_time
Time difference of 0.1790411 secs
```

# Pakiet wINEQ

Pakiet R został opracowany w następujących krokach:

- Wyprowadzenie formuł miar nierówności dla danych ważonych.
- Kodowanie nowych miar nierówności.
- Rozszerzenie miar nierówności z pakietu **ineq** w celu uwzględnienia danych ważonych.
- Dodanie funkcji **ineq.weighted** w celu dostarczenia wszystkich miar nierówności dla danego zestawu danych.
- Dodanie funkcji **ineq.weighted.boot** w celu rozszerzenia danych wynikowych **ineq.weighted** o bootstrap.

# Pakiet wINEQ - miary nierówności

Pakiet **wINEQ** zawiera miary nierówności:

- dostępne w pakiecie **ineq** (tylko dla zmiennych na skali ilorazowej),
- nowe miary dla zmiennych na skali: ilorazowej, interwałowej jak i porządkowej

Wszystkie metody zostały zaimplementowane z możliwością ważenia danych.

# Pakiet wINEQ - miary nierówności

	Skala ilorazowa	Skala interwałowa	Skala porządkowa
Indeks Allisona i Fostera		X	
Indeks Abul Naga and Yalcin			X
Indeks Apouey'a			X
Indeks Atkinsona	X		
Indeks Blair i Lacy			X
Współczynnik zmienności	X		
Entropia	X		
Współczynnik Giniego	X		
Indeks Hoovera	X		
Indeks Jenkinsa	X		
Indeks Cowella i Flachaire'a	X		
Indeks Kolma	X		
Indeks Letiego			X
Indeks Kolma	X		
Indeks Palma	X		
Proporcja 20:20	X		
Indeks Ricciego i Schutza	X		
Współczynnik Theila	X		

# Pakiet wINEQ - ważone wersje wskaźników nierówności

Ustalmy:

$x_i$  - i-ta wartość zmiennej

$w_i$  - waga i-tej wartości zmiennej

$\mu$  – średnia arytmetyczna

$\mu_w$  - ważona średnia arytmetyczna

$$\mu_w = \frac{\sum w_i x_i}{\sum w_i}$$



# Pakiet wINEQ - ważone wersje wskaźników nierówności

Współczynnik Giniego

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|}{2n^2 \mu}$$

$$G = \frac{\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j| w_i w_j}{2(\sum_{i=1}^n w_i)^2 \mu_w}$$

Wskaźnik Hoovera

$$H = \frac{1}{2} \frac{\sum_{i=1}^n |x_i - \mu|}{\sum_{i=1}^n x_i}$$

$$H = \frac{1}{2} \frac{\sum_{i=1}^n w_i |x_i - \mu_w|}{\sum_{i=1}^n w_i x_i}$$

# Pakiet wINEQ - ważone wersje wskaźników nierówności

## Współczynnik Theila

Theil T

$$T_1 = \frac{1}{n} \sum_{i=1}^n \frac{x_i}{\mu} \ln \frac{x_i}{\mu}$$

$$T_1 = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \frac{x_i}{\mu_w} \ln \frac{x_i}{\mu_w}$$

Theil L

$$T_0 = \frac{1}{n} \sum_{i=1}^n \ln \frac{\mu}{x_i}$$

$$T_0 = \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i \ln \frac{\mu_w}{x_i}$$

# Pakiet wINEQ - ważone wersje wskaźników nierówności

Współczynnik Atkinsona (z parametrem  $\epsilon$ )

$$A = \begin{cases} 1 - \frac{1}{\mu} \left( \frac{1}{n} \sum_{i=1}^n x_i^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}} & \text{dla } \epsilon \neq 1 \\ 1 - \frac{1}{\mu} \left( \prod_{i=1}^n x_i \right)^{\frac{1}{n}} & \text{dla } \epsilon = 1 \end{cases}$$
$$A = \begin{cases} 1 - \frac{1}{\mu_w} \left( \frac{1}{\sum_{i=1}^n w_i} \sum_{i=1}^n w_i x_i^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}} & \text{dla } \epsilon \neq 1 \\ 1 - \frac{1}{\mu_w} \left( \prod_{i=1}^n x_i^{w_i} \right)^{\frac{1}{\sum_{i=1}^n w_i}} & \text{dla } \epsilon = 1 \end{cases}$$

# Pakiet wINEQ – miary nierówności

- **AF**(X, W = rep(1, length(X)), norm = TRUE),
- **AN\_Y**(X, W = rep(1, length(X)), a = 1, b = 1),
- **Apouey**(X, W = rep(1, length(X)), a, b),
- **Atkinson**(X, W = rep(1, length(X)), e = 1),
- **BL**(X, W = rep(1, length(X)), withsqrt = FALSE),
- **CoefVar**(X, W = rep(1, length(X)), square = FALSE)
- **Entropy**(X, W = rep(1, length(X)), parameter = 0.5),
- **Gini**(X, W = rep(1, length(X))),
- **Hoover**(X, W = rep(1, length(X))),
- **Jenkins**(X, W = rep(1, length(X)), alfa = 0.8) obejmuje również indeks Cowella i Flachaire'a)
- **Kolm**(X, W = rep(1, length(X)), parameter = 1, scale = "None")
- **Leti**(X, W = rep(1, length(X)), norm = T)
- **medianf**(X, W = rep(1, length(X)))
- **Palma**(X, W = rep(1, length(X))),
- **Prop20\_20**(X, W = rep(1, length(X))),
- **RicciSchutz**(X, W = rep(1, length(X))),
- **Theil\_L**(X, W = rep(1, length(X))),
- **Theil\_T**(X, W = rep(1, length(X))),

gdzie: X jest wektorem danych, W jest wektorem wag.

# Pakiet wINEQ – funkcja agregująca

```
ineq.weighted(  
X,  
W = rep(1, length(X)),  
AF.norm = TRUE,  
Atkinson.e = 1,  
Jenkins.alfa = 0.8,  
Entropy.e = 0.5,  
Kolm.p = 1,  
Kolm.scale = "Standardization",  
Leti.norm = T,  
AN_Y.a = 1,  
AN_Y.b = 1,  
Apouey.a = 2/(1 - length(W[!is.na(W) & !is.na(X)])),  
Apouey.b = length(W[!is.na(W) & !is.na(X)])/(length(W[!is.na(W) & !is.na(X)]) - 1),  
BL.withsqrt = FALSE  
)
```

Wynik - funkcja generuje średnią ważoną i sumę X bądź medianę X oraz adekwatne miary nierówności.

# Pakiet wINEQ – funkcja agregująca

```
ineq.weighted.boot(  
X,  
W = rep(1, length(X)),  
B = 100,  
AF.norm = TRUE,  
Atkinson.e = 1, Jenkins.alfa = 0.8, Entropy.e = 0.5,  
Kolm.p = 1, Kolm.scale = "Standardization", Leti.norm = T,  
AN_Y.a = 1, AN_Y.b = 1,  
Apouey.a = 2/(1 - length(W[!is.na(W) & !is.na(X)])),  
Apouey.b = length(W[!is.na(W) & !is.na(X)])/(length(W[!is.na(W) & !is.na(X)]) - 1),  
BL.withsqrt = FALSE,  
keepSamples = FALSE, keepMeasures = FALSE,  
conf.alpha = 0.05, calib.boot = FALSE,  
Xs = rep(1, length(X)),  
total = sum(W), calib.method = "truncated",  
bounds = c(low = 0, upp = 10)  
)
```

# Pakiet wINEQ – zbiory danych

- Dzięki załączonym zbiorom danych pakiet daje możliwość sprawdzenia możliwości pakietu na danych pochodzących ze statystyki publicznej.
- Zamieszczone dane pochodzą z badania podróży oraz z badania jakości życia przeprowadzonego na pograniczu polsko-ukraińskim w latach 2015 i 2019.
- Oba zestawy danych zawierają wagi.
- Odpowiedzi zostały zmodyfikowane ze względu na potrzebę zachowania tajemnicy statystycznej. Stanowią tylko część pełnej bazy danych ale w zupełności wystarczają do testowania pakietu.



# Przykładowy opis funkcji

---

Entropy      *Generalized entropy index*

---

## Description

Computes generalized entropy index of a given variable taking into account weights.

## Usage

```
Entropy(X, W = rep(1, length(X)), parameter = 0.5)
```

## Arguments

X            is a data vector  
W            is a vector of weights  
parameter   is a entropy parameter

## Details

Entropy coefficient with respect to parameter  $\alpha$  is equal to Theil\_L(X,W) whenever  $\alpha = 0$ , is equal to Theil\_T(X,W) whenever  $\alpha = 1$ , and whenever  $\alpha \in (0, 1)$  we have

$$GE(\alpha) = \frac{1}{\alpha(\alpha-1)W} \sum_{i=1}^n w_i \left( \left( \frac{x_i}{\mu} \right)^\alpha - 1 \right)$$

where  $W$  is a sum of weights and  $\mu$  is the arithmetic mean of  $x_1, \dots, x_n$ .

## Value

The value of generalized entropy index

## References

Shorrocks A. F.: (1980) The Class of Additively Decomposable Inequality Measures. *Econometrica*  
Pielou E.C.: (1966) The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*

## Examples

```
# Compare weighted and unweighted result  
X=1:10  
W=1:10  
Entropy(X)  
Entropy(X,W)
```

```
data(Tourism)  
# Generalized entropy index for Total expenditure with sample weights  
X=Tourism$`Total expenditure`
```

• Nazwa funkcji

• Opis

• Konstrukcja funkcji

• Argumenty jakie funkcja posiada

• Szczegóły dot. funkcji (wzór, opis parametrów)

• Opis wyniku funkcji

• Literatura

• Przykłady (dla danych ważonych i nieważonych oraz dla danych zawartych w pakiecie)



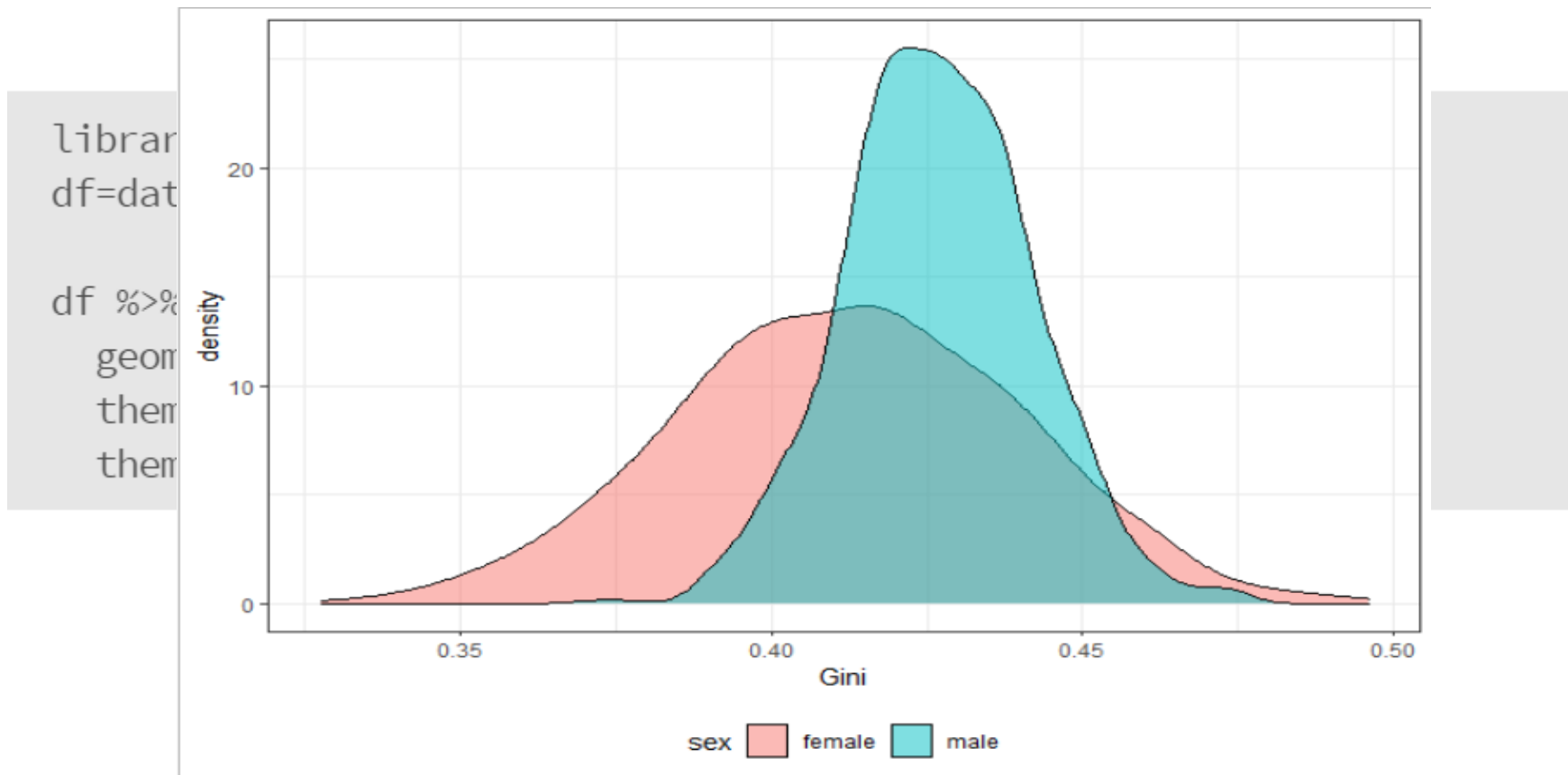
# Zastosowanie w praktyce

- Ilocos - Filipińskie badanie budżetów domowych – dane dostępne w pakiecie **ineq**.
- Uwzględniono wagi.
- Porównamy nierówność dochodów wśród mężczyzn i kobiet (głowy gospodarstwa domowego).

```
library(ineq)
data(Ilocos);attach(Ilocos)
Female <- ineq.weighted.boot(income[sex=='female'],
                             AP.weight[sex=='female'],
                             keepMeasures = T,B = 1000)
Male <- ineq.weighted.boot(income[sex=='male'],
                            AP.weight[sex=='male'],
                            keepMeasures = T,B = 1000)
```

# Zastosowanie w praktyce

Wizualizacja rozkładów współczynnika Giniego za pomocą ggplot2.



# Zastosowanie w praktyce

Testowanie normalności rozkładu za pomocą testu **Jarque-Bera** z pakietu **normtest**.

W obu przypadkach wartość p jest większa niż 0,05.

W związku z tym możemy stosować testy statystyczne, które opierają się na założeniu rozkładu normalnego.

```
library('normtest')
library('dplyr')
Female[[2]][, 'Gini'] %>% jb.norm.test()
```

```
##
## Jarque-Bera test for normality
##
## data: .
## JB = 2.5024, p-value = 0.257
```

```
Male[[2]][, 'Gini'] %>% jb.norm.test()
```

```
##
## Jarque-Bera test for normality
##
## data: .
## JB = 1.9303, p-value = 0.3775
```

# Zastosowanie w praktyce

- Testowanie różnicy w średnich za pomocą testu t. Wartość P jest niższa niż 0,05. Średnia wartość współczynnika Giniego dla kobiet wyniosła 0,4129 i jest istotnie niższa niż dla mężczyzn, która wyniosła 0,42678.

```
t.test(Female[[2]][, 'Gini'], Male[[2]][, 'Gini'])
```

```
##  
## Welch Two Sample t-test  
##  
## data: Female[[2]][, "Gini"] and Male[[2]][, "Gini"]  
## t = -14.106, df = 1559.8, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -0.01571447 -0.01187777  
## sample estimates:  
## mean of x mean of y  
## 0.4129748 0.4267709
```

# wINEQ na CRAN-R

- Pakiet jest dostępny na CRAN-R <https://cran.r-project.org/web/packages/wINEQ/index.html>
- Moderatorzy CRAN-R polecają nasz pakiet w zakładce *Task Views* w zakresie Econometrics ([CRAN Task View: Econometrics \(r-project.org\)](https://CRAN.R-project.org/view=OfficialStatistics)) oraz *OfficialStatistics* (<https://CRAN.R-project.org/view=OfficialStatistics>)
- Liczba pobrań naszego pakietu od publikacji w 2022 roku wyniosła **4500**
- Pakiet jest stale aktualizowany o nowe miary
- W ostatniej aktualizacji dodano indeksy Apouey, Abul Naga-Yelcin oraz Blair-Lacy

# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

- Badanie budżetów gospodarstw domowych stanowi podstawowe źródło informacji o poziomie i strukturze wydatków oraz dochodów badanych gospodarstw domowych
- Dostarcza danych wykorzystywanych do:
  - obliczania wag do indeksów cen towarów i usług konsumpcyjnych,
  - ustalenia wskaźnika kosztów utrzymania gospodarstw domowych,
  - ustalenia poziomu minimalnego wynagrodzenia za pracę
- BBGD to badanie reprezentacyjne ciągłe całoroczne
- Na przestrzeni blisko dwudziestu lat wielkość próby wahała się od 32 000 do 37 000 gospodarstw domowych

# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

- Analizie poddano następujące zmienne:
  - Przychody ogółem,
  - Rozchody ogółem,
  - Świadczenia otrzymane z ZUS i pomocy społecznej
  - Transfery w formie pieniężnej i w naturze między gospodarstwami
  - Przychody przed transferami
  - Płeć głowy rodziny
  - Czy głowa rodziny jest w związku małżeńskich/partnerskim
- Obliczono wskaźnik Theila, Hoovera, Atkinsona, Palmy i proporcję 20:20 wg lat i województw

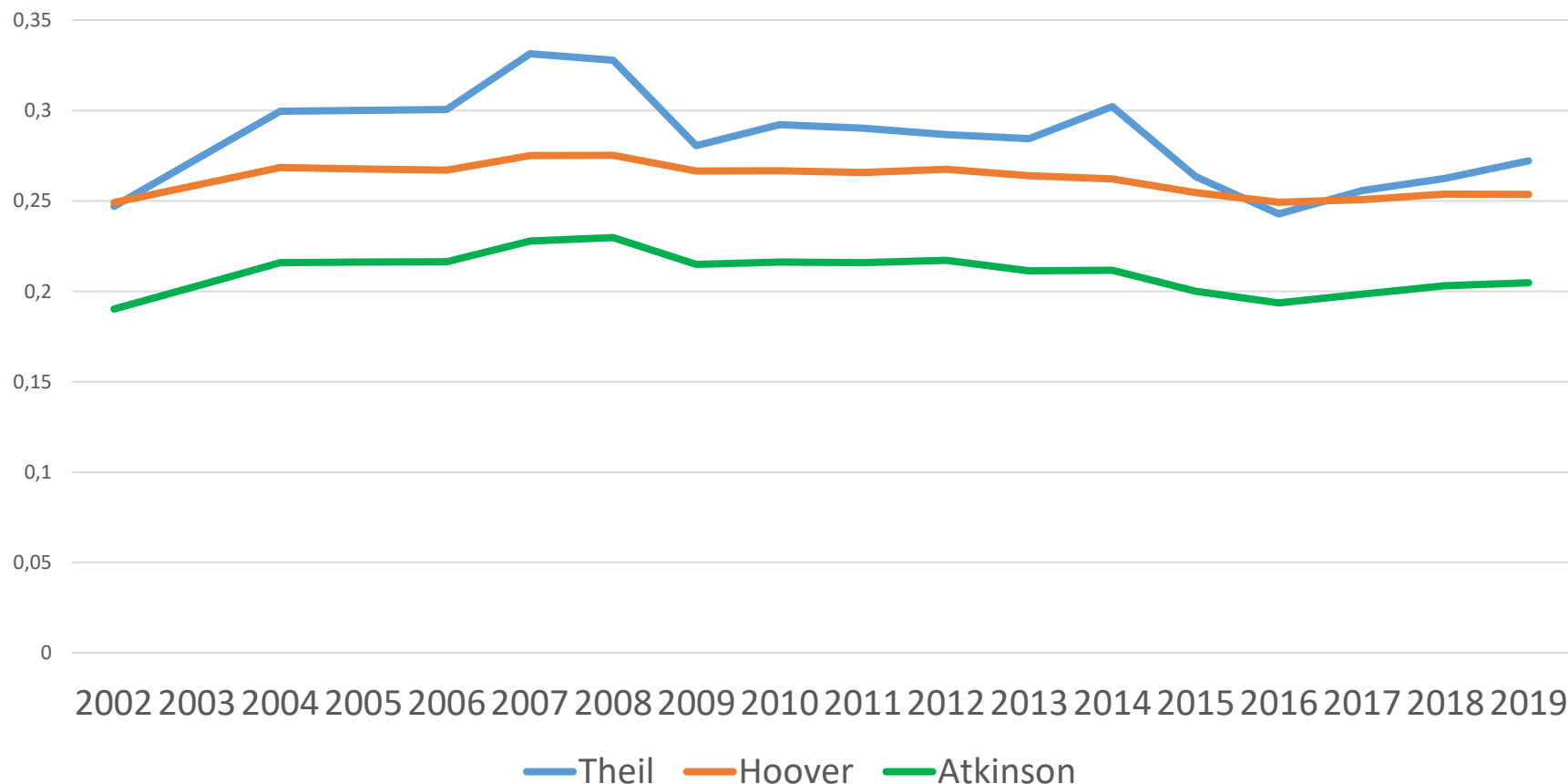
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

- Analizie poddano transfery w formie pieniężnej i w naturze między gospodarstwami
- Chcemy odpowiedzieć na pytanie czy transfery zmniejszają nierówności dochodowe
- Gospodarstwa sklasyfikowano do czterech grup:
  - Brak transferów
  - Darujący (tylko transfery z gospodarstwa)
  - Obdarowani (tylko transfery do gospodarstwa)
  - Transfery obustronne



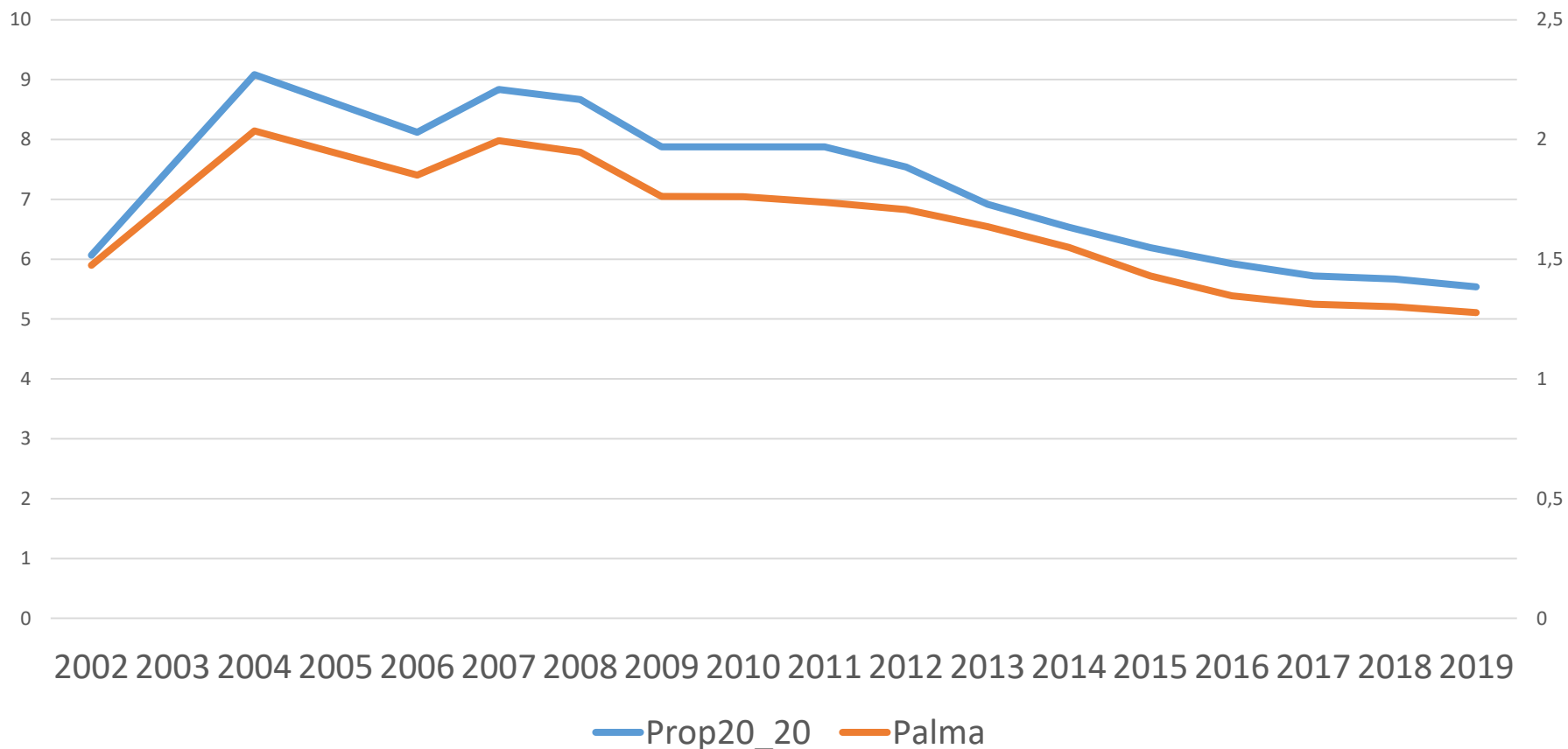
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach gospodarstw



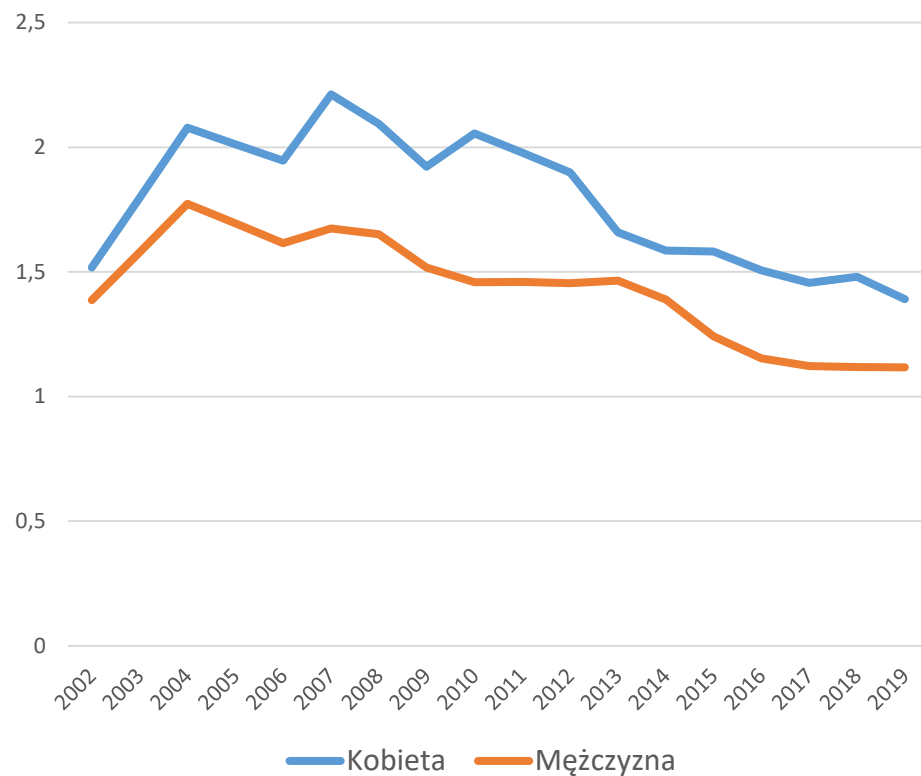
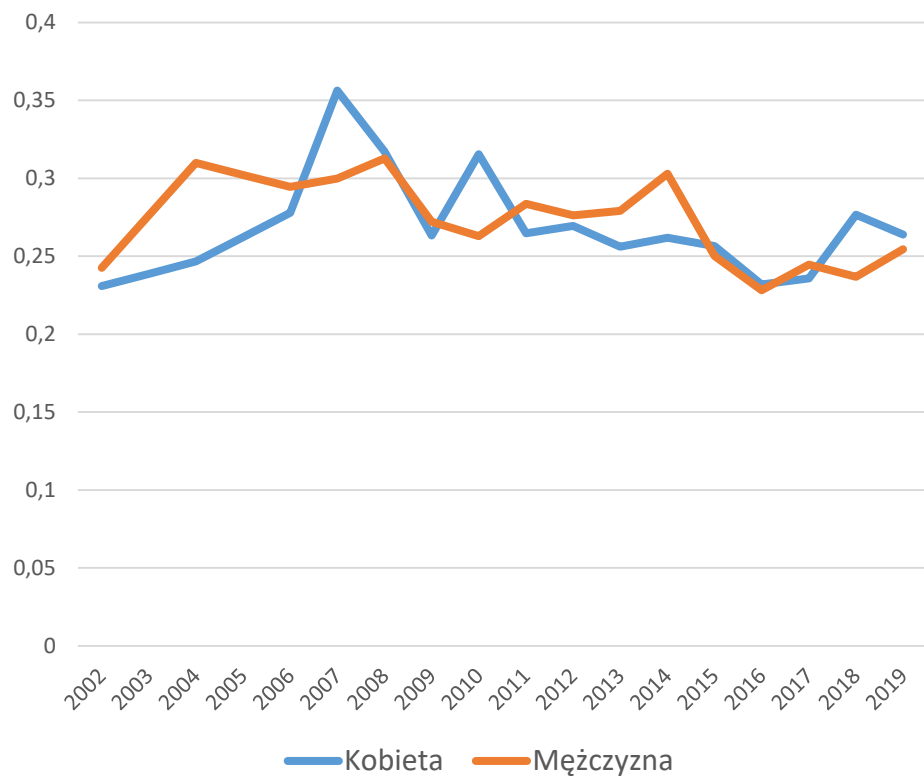
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach gospodarstw



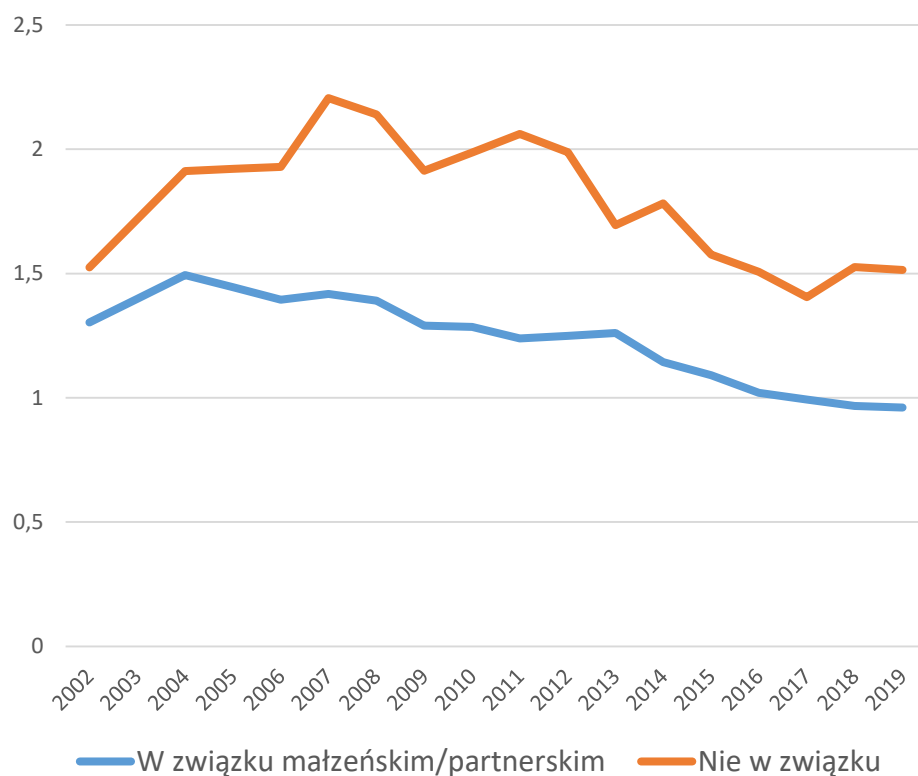
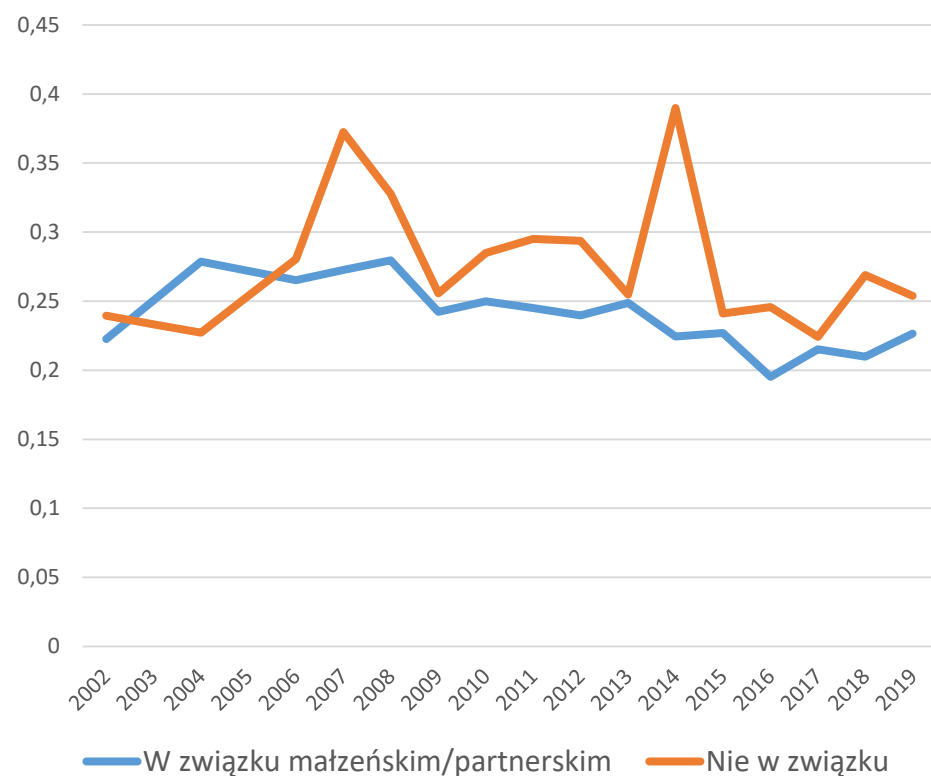
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach gospodarstw wg płci głowy rodziny



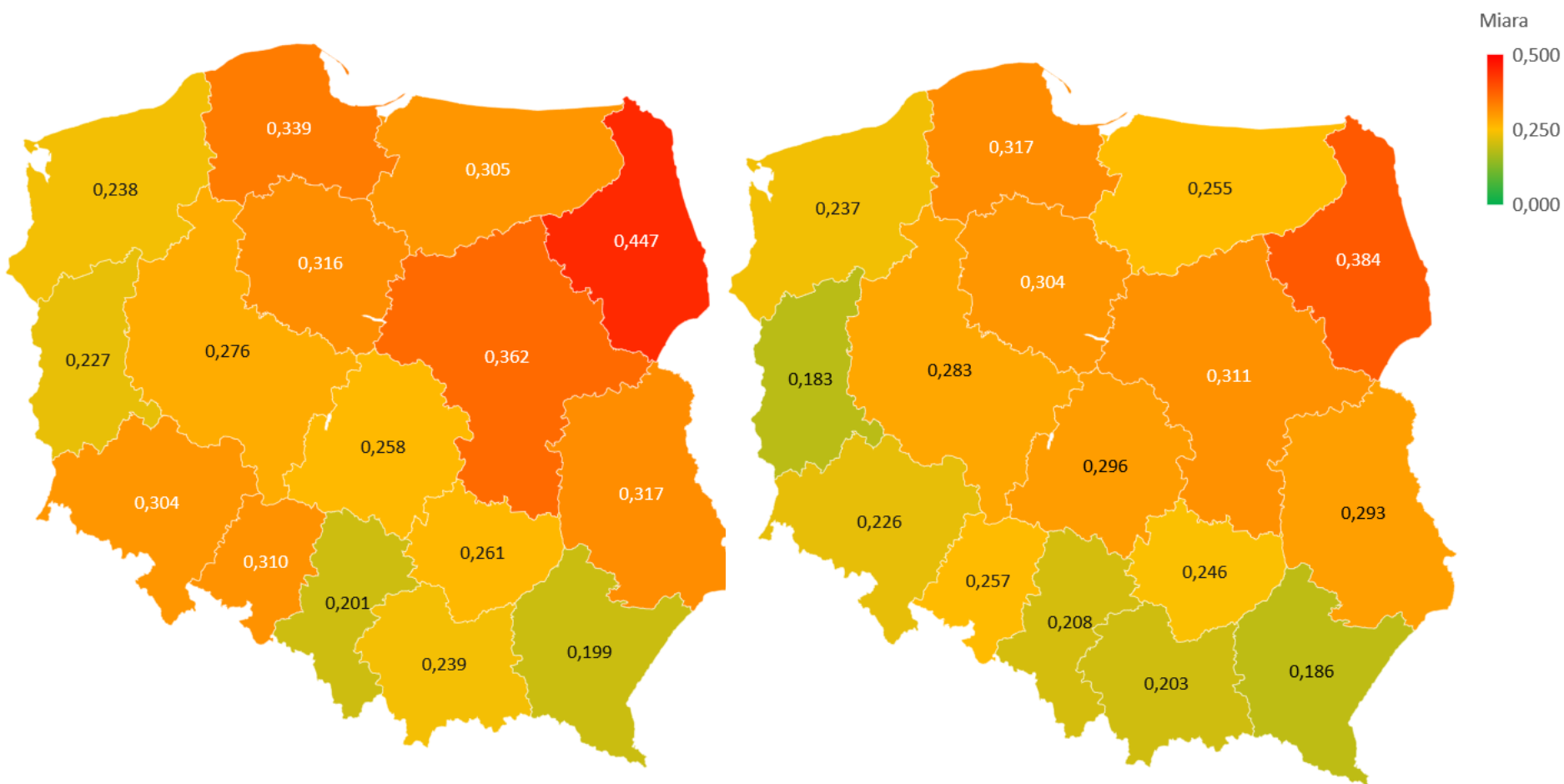
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach gospodarstw a status głowy rodziny



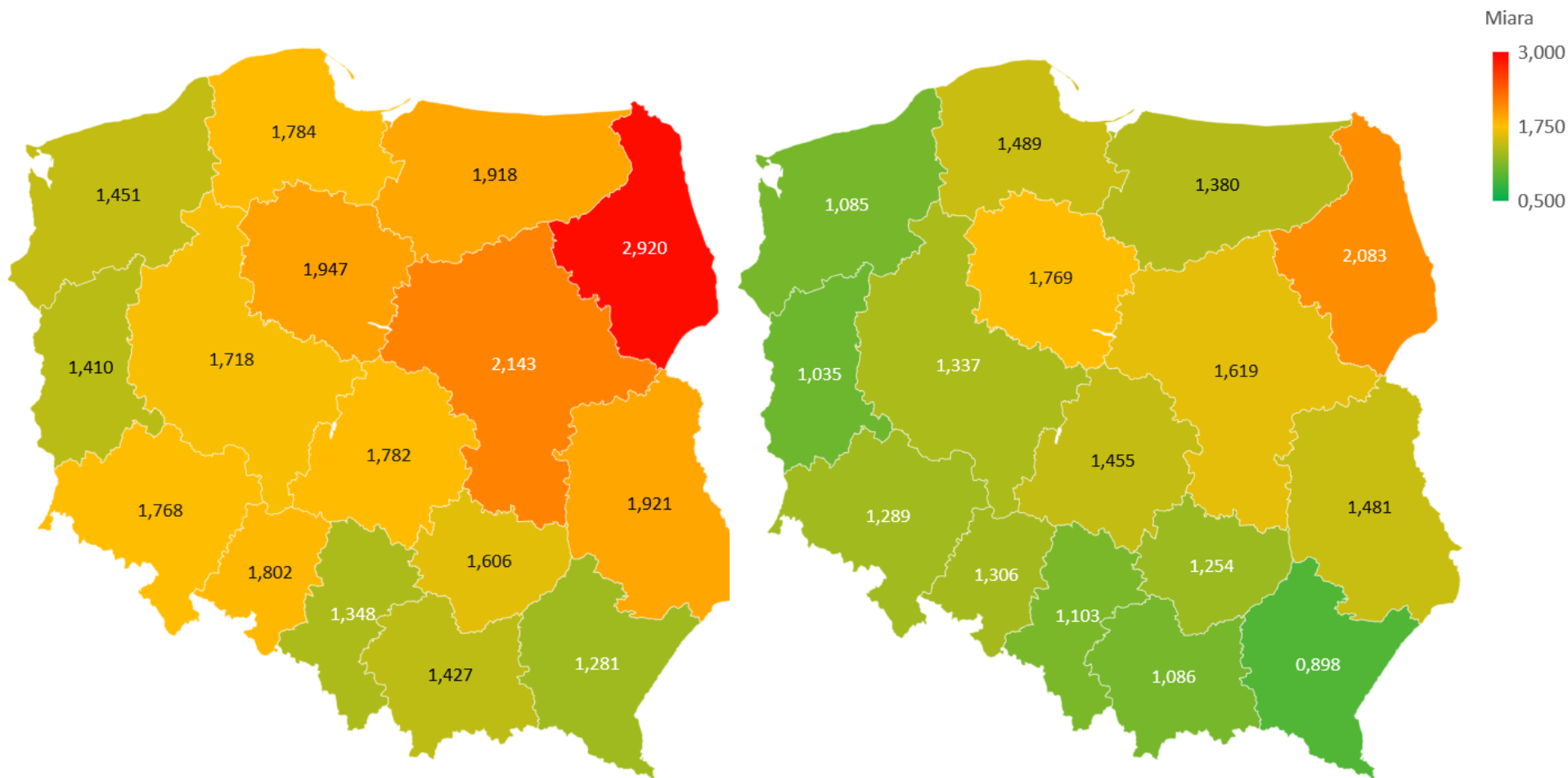
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach wg województw (2008-2013 i 2014-2019) – wskaźnik Theila



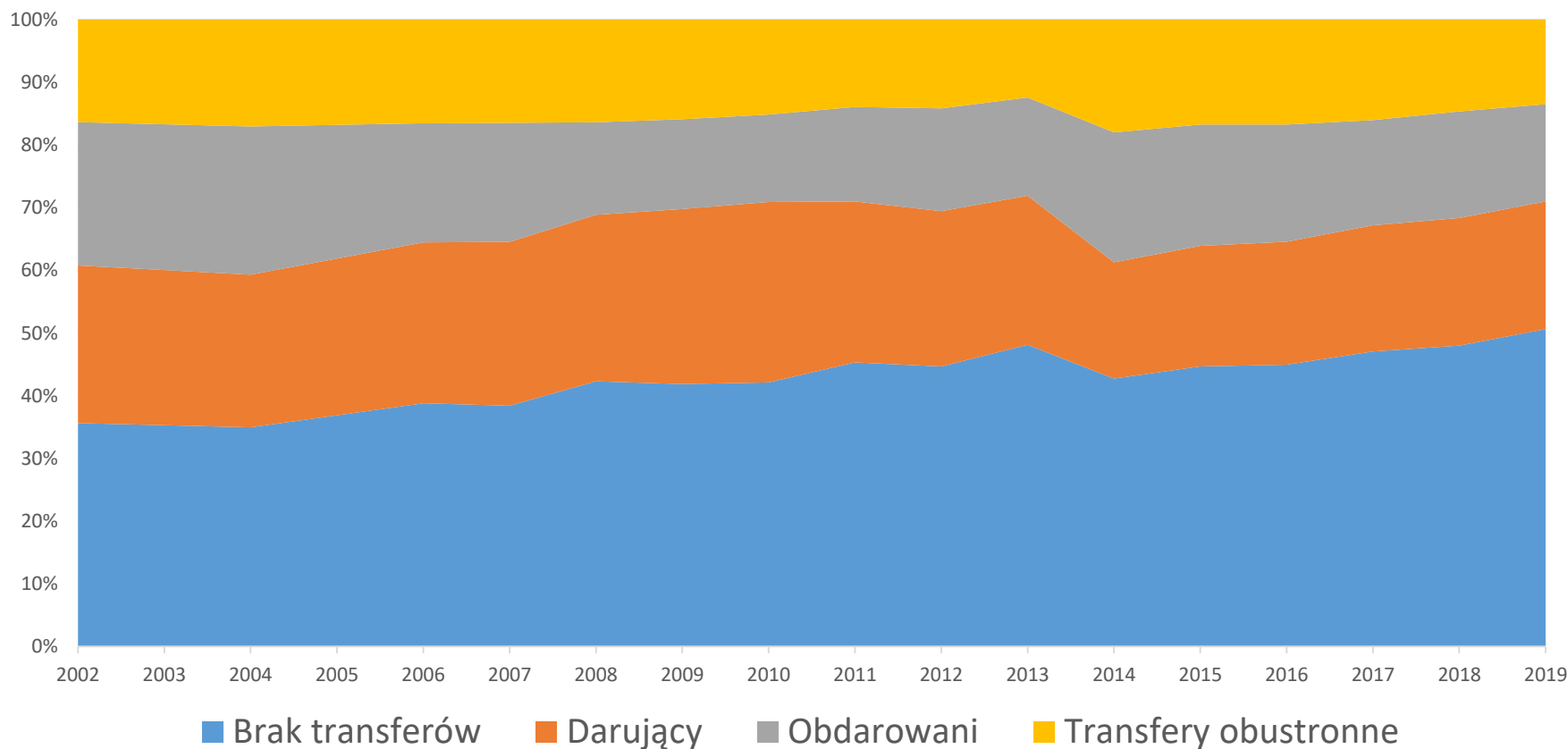
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Nierówności w przychodach wg województw (2008-2013 i 2014-2019) – wskaźnik Palmy



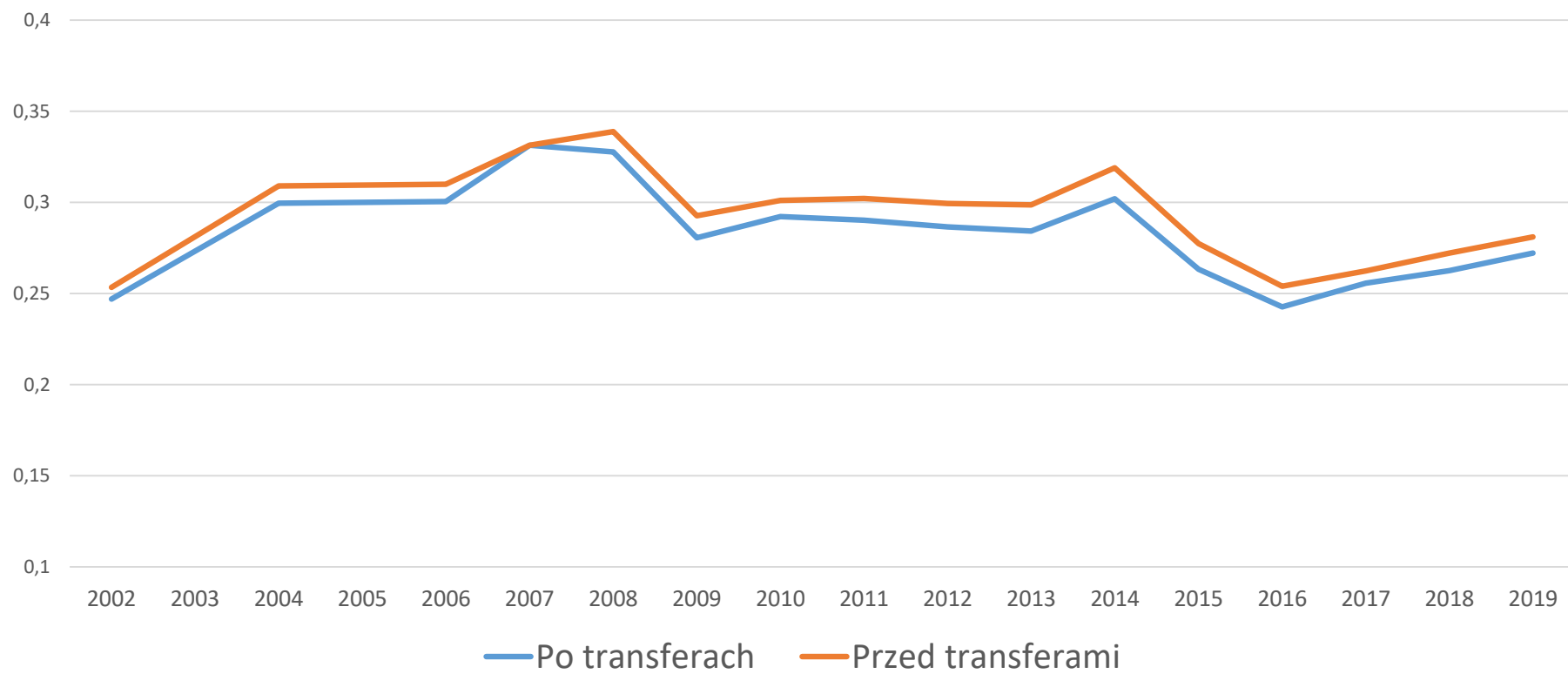
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Struktura gospodarstw domowych wg uczestnictwa w transferach



# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

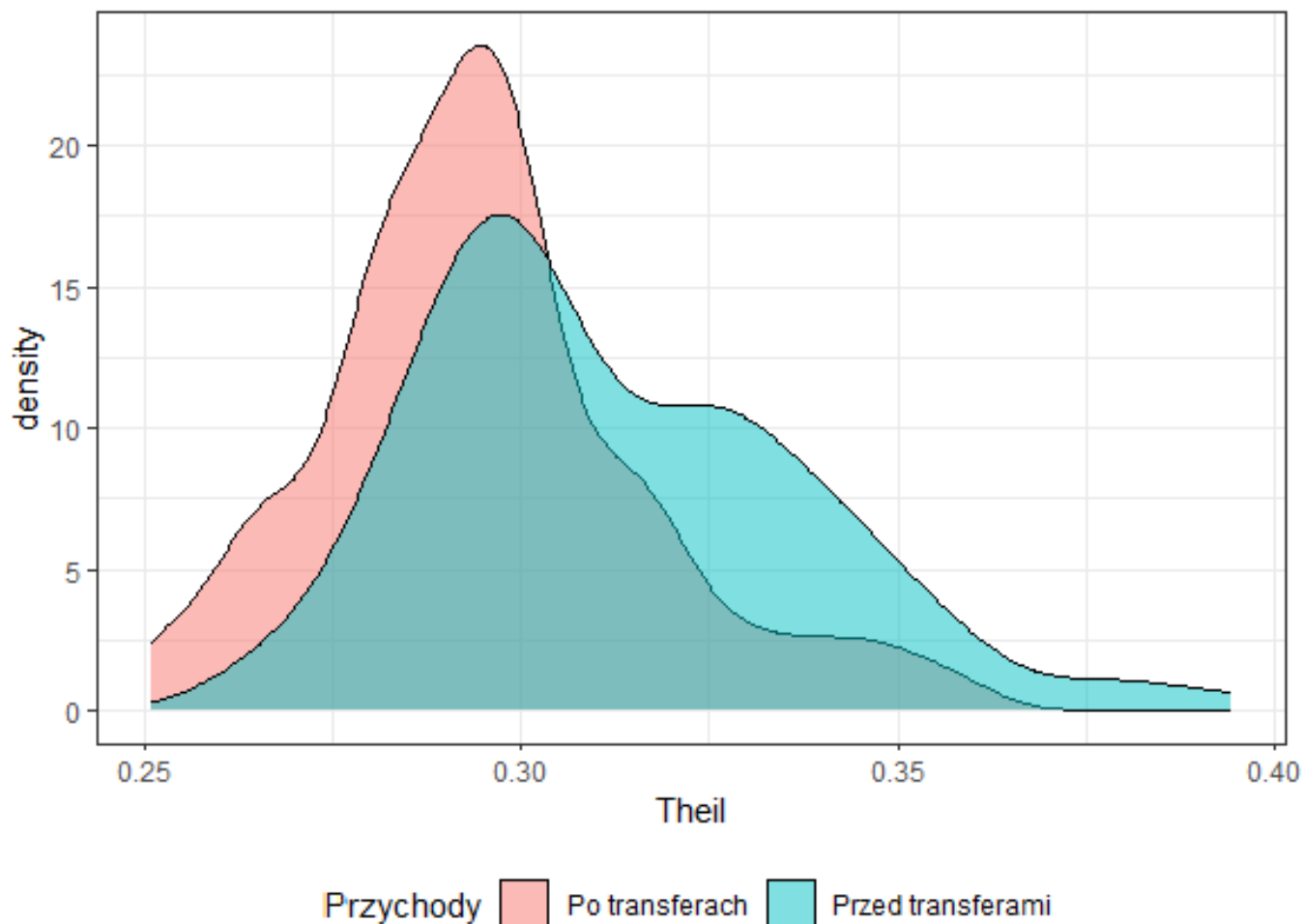
Wskaźnik Theila dla przychodów przed i po transferach





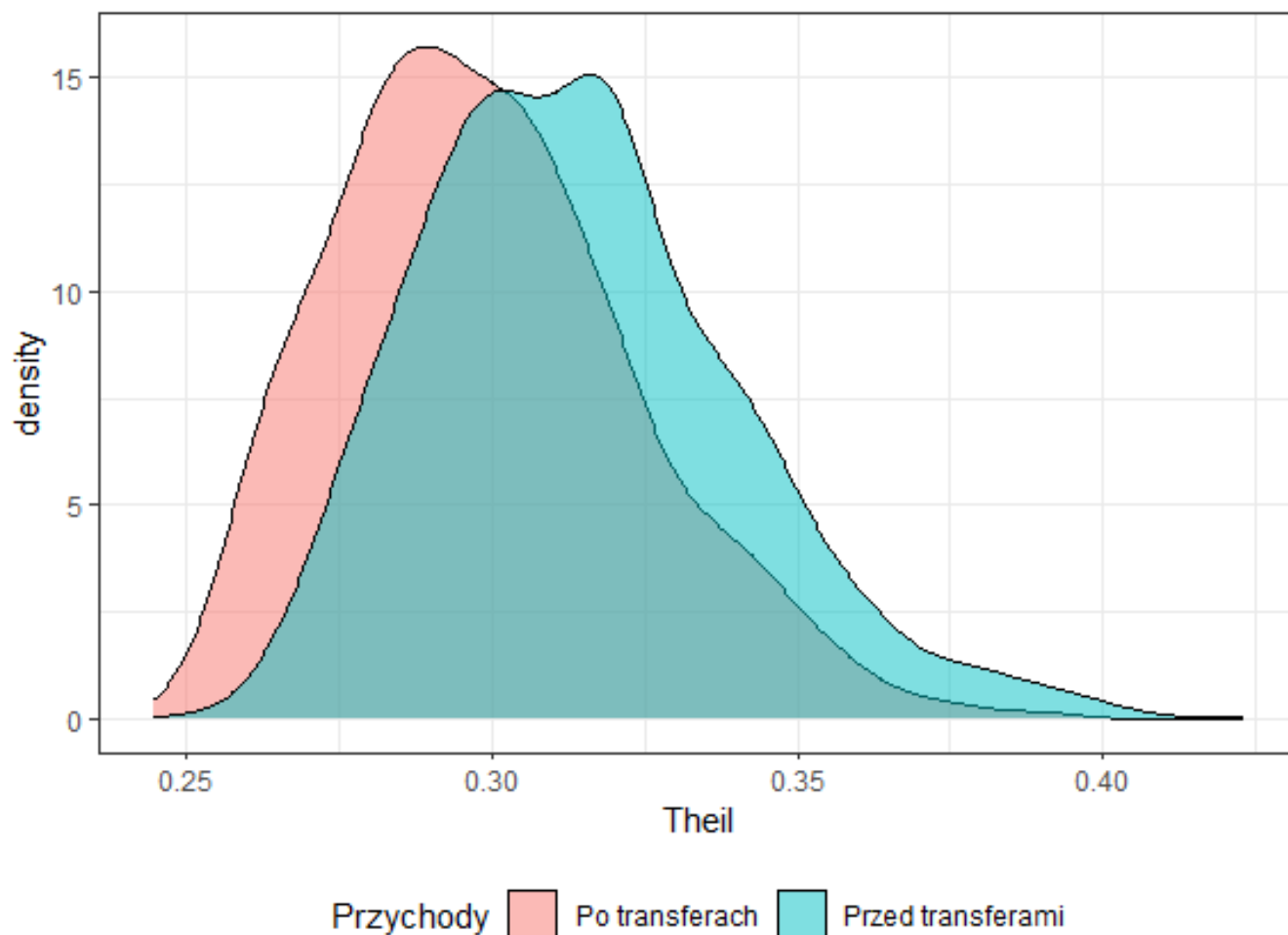
# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Wskaźnik Theila dla przychodów przed i po transferach w 2014 roku (B=100)



# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

Wskaźnik Theila dla przychodów przed i po transferach w 2014 roku (B=1000)



# Charakterystyka nierówności na podstawie Badania Budżetów Gospodarstw Domowych

- Rozkład wskaźnika Theila dla przychodów przed i po transferach w 2014 roku **różnił się** istotnie od rozkładu normalnego ( $p$ -wartość  $< 0,0001$ )
- Wariacje wskaźnika Theila **nie różniły się** istotnie ( $p$ -wartość =  $0,06743$ )
- Wykorzystaliśmy semiparametryczny test przesunięcia Manna-Whitneya do porównania wskaźnika Theila
- Istotne różnice zaobserwowano na poziomie  $p < 0,0001$

# Dziękujemy za uwagę!

Dr Sebastian Wójcik, Uniwersytet Rzeszowski, Urząd Statystyczny w Rzeszowie

Agnieszka Giemza, Urząd Statystyczny w Rzeszowie

Jarosław Napora, Politechnika Rzeszowska, Urząd Statystyczny w Rzeszowie