Mathematical Statistics

Anna Janicka

Lecture II, 26.02.2018

DESCRIPTIVE STATISTICS, PART II

Plan for today

1. Descriptive Statistics, part II:
   - mode
   - quantiles
   - measures of variability
   - measures of asymmetry
   - the boxplot

Measures of central tendency – reminder

- *Classic:*
  - *arithmetic mean*
- *Position (order, rank):*
  - *median*
  - mode
  - quartile

Example 1 – cont.

Grade	*Number*	*Frequency*
2	59	31.89%
3	63	34.05%
3.5	33	17.84%
4	14	7.57%
4.5	10	5.41%
5	6	3.24%
Total	185	100%

Mode – examples
Quartiles – example
Variance – example

Example 3 – cont.

Interval	*Class mark*	*Number*	*Frequency*	*Cumulative number* cn_i	*Cumulative frequency* cf_i
(30,40]	35	11	0,11	11	0,11
(40,50]	45	23	0,23	34	0,34
(50,60]	55	33	0,33	67	0,67
(60,70]	65	12	0,12	79	0,79
(70,80]	75	6	0,06	85	0,85
(80,90]	85	8	0,08	93	0,93
(90,100]	95	3	0,03	96	0,96
(100,110]	105	2	0,02	98	0,98
(110,120]	115	2	0,02	100	1
Total		100	1		

[Mode – example]
[Quartiles – example]
[Variance – example]

Mode

the value that appears most often

- for grouped data:

  Mo = most frequent value

- for grouped class interval data:

$$Mo \cong c_L + \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})} \cdot b$$

where

n_{Mo} – number of elements in mode's class,

c_L, b – analogous to the median

Mode – examples

Example 1 – cont.

Example 1:

$Mo = 3$

Example 3 – cont.

Example 3:

the mode's interval is (50,60], with 33 elements

$n_{Mo} = 33$, $c_L = 50$, $b = 10$, $n_{Mo\text{-}1} = 23$, $n_{Mo+1} = 12$

$$Mo \cong 50 + \frac{33-23}{(33-23)+(33-12)} \cdot 10 \approx 53.23$$

Which measure should we choose?

- Arithmetic mean: for typical data series (single max, monotonous frequencies)
- Mode: for typical data series, grouped data (the lengths of the mode's class and neighboring classes should be equal)
- Median: no restrictions. The most robust (in case of outlier observations, fluctuations etc.)

Quantiles, quartiles

- p-th quantile (quantile of rank p): number such that the fraction of observations less than or equal to it is at least p, and values greater than or equal to it at least $1-p$

- Q_1 : first quartile = quantile of rank ¼
- Second quartile = median = quantile of rank ½
- Q_3: Third quartile = quantile of rank ¾

Quantiles – cont.

Empirical quantile of rank *p*:

$$Q_p = \begin{cases} \dfrac{X_{np:n} + X_{np+1:n}}{2} & np \in Z \\ X_{[np]+1:n} & np \notin Z \end{cases}$$

Quartiles – cont.

- Quantiles for p = ¼ and p = ¾.
- For grouped class interval data – analogous to the median

$$Q_k \cong c_L + \frac{b}{n_{M_k}}\left(\frac{k \cdot n}{4} - \sum_{i=1}^{M_k - 1} n_i\right)$$

for k=1 or 3

where M_1, M_3 – number of the quartile's class

b – length of quartile class interval

c_L – lower end of the quartile class interval

Quartiles – examples

Example1 – cont.

Example 1:

$$185 \cdot \tfrac{1}{4} = 46.25 \qquad 185 \cdot \tfrac{3}{4} = 138.75$$

so

$$Q_1 = X_{47:185} = 2, \qquad Q_3 = X_{139:185} = 3.5$$

Example 3 – cont.

Example 3:

$$100 \cdot \tfrac{1}{4} = 25 \qquad 100 \cdot \tfrac{3}{4} = 75$$

$$M_1 = 2, \quad \mathrm{M}_3 = 4 \quad \text{so}$$

$$Q_1 \cong 40 + \frac{10}{23}(25 - 11) \approx 46{,}09 \qquad Q_3 \cong 60 + \frac{10}{12}(75 - 67) \approx 66{,}67$$

Variability measures

- Classical measures
  - variance, standard deviation
  - average (absolute) deviation
  - coefficient of variation
- Measures based on order statistics
  - range
  - interquartile range
  - *quartile deviation*
  - *coefficients of variation (based on order stats)*
  - *median absolute deviation*

Measures based on order statistics

- Range

the most simple measure, does not take into account anything but the extreme values

$$r = X_{n:n} - X_{1:n}$$

- Inter Quartile Range (midspread, middle fifty)

more robust than the range

$$IQR = Q_3 - Q_1$$

length of the interval that covers the middle 50% observations

may be further used to calculate **quartile deviation *Q***= *IQR*/2, and coefficients of variation $V_Q = Q/Med$ or $V_{Q1Q3} = IQR/(Q_3+Q_1)$ (quartile variation coefficient) or the typical range: [*Med* – *Q*, *Med* + *Q*]

Range, interquartile range – examples

Example 1:

$$r = 5 - 2 = 3,$$

$$IQR = 3.5 - 2 = 1.5$$

Example 3:

$$r \cong 120 - 30 = 90$$

(in reality 118,9-32,45 $=$ 86,45)

$$IQR \cong 66,67 - 46,09 = 20,58$$

Classical measures of dispersion

Variance

- raw data

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^{\,2} - (\overline{X})^2$$

- grouped data

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{k} n_i (X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i X_i^{\,2} - (\overline{X})^2$$

- grouped class interval data

$$\hat{S}^2 \cong \frac{1}{n}\sum_{i=1}^{k} n_i (\overline{c}_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{k} n_i \overline{c}_i^{\,2} - (\overline{X})^2$$

c=length of class interval (for equal intervals)

+ Sheppard's correction

$$\overline{S}^2 \cong \hat{S}^2 - \frac{c^2}{12}$$

in general

$$\overline{S}^2 \cong \hat{S}^2 - \frac{1}{12n}\sum_{i=1}^{k} n_i (c_i - c_{i-1})^2$$

Variance – examples

Example 1 – cont.

Example 1:

$$\hat{S}^2 \approx \tfrac{1}{185}\left((2-2.99)^2\cdot 59+(3-2.99)^2\cdot 63+(3.5-2.99)^2\cdot 33+(4-2.99)^2\cdot 14+(4.5-2.99)^2\cdot 10+(5-2.99)^2\cdot 6\right)$$

$$\approx 0.69$$

Example 3:

Example 3 – cont.

$$\hat{S}^2 \approx {}^{1}\!/_{100}\cdot \quad ((35-58.7)^2\cdot 11+(45-58.7)^2\cdot 23+(55-58.7)^2\cdot 33+(65-58.7)^2\cdot 12$$
$$+(75-58.7)^2\cdot 6+(85-58.7)^2\cdot 8+(95-58.7)^2\cdot 3+(105-58.7)^2\cdot 2+(115-58.7)^2\cdot 2)$$

$$=331.31$$

$$\overline{S}^2 = 331.31 - {}^{10^2}\!/_{12} \approx 322.98$$

in reality $\hat{S}^2 = 333.85$

distrubution not normal or sample too small for Sheppard's correction – larger errors from small sample size than from class grouping.

Standard deviation

In the same units as the initial variable

$$\hat{S} = \sqrt{\hat{S}^2}, \qquad \overline{S} = \sqrt{\overline{S}^2}$$

Example 1:

$$\hat{S} \approx 0.83 \; [grade]$$

Example 3:

$$\hat{S} \approx 18.2 \; [m^2]$$

Average (absolute) deviation, mean deviation

Nowadays seldom used. Simple calculations.

for raw data

$$d = \frac{1}{n}\sum_{i=1}^{n} | X_i - \overline{X} |$$

etc...

We have: $d<S$

Coefficient of variation (classical)

For comparisons of the same varaible accross populations or different variables for the same population

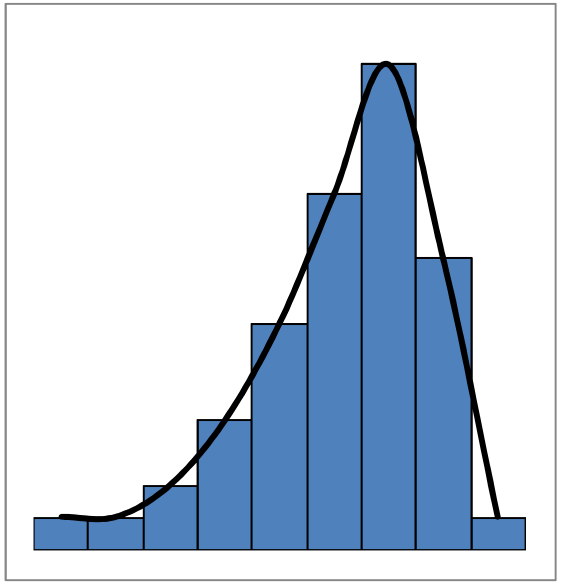
$$V_S = \frac{\hat{S}}{\overline{\overline{X}}}(\cdot 100\%),$$

$$\text{or } V_d = \frac{d}{\overline{\overline{X}}}(\cdot 100\%)$$

Skewness (asymmetry)

left

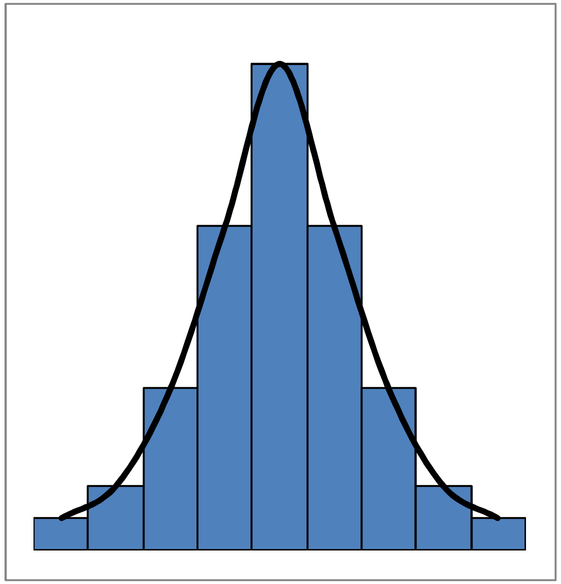
(negative)



$\overline{X} < Med < Mo$

symmetry

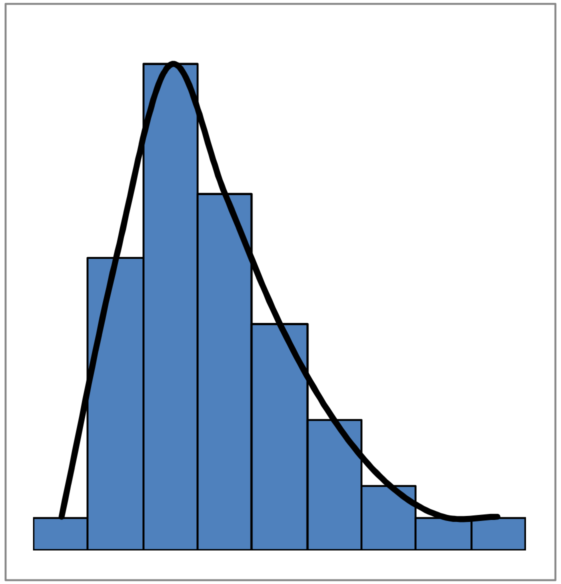
(zero)



$\overline{X} = Med = Mo$

right

(positive)



$\overline{X} > Med > Mo$

(typical order)

Measures of asymmetry

- Skewness

$$A = \frac{M_3}{\hat{S}^3}$$

where M_3 is the third central moment

- Skewness coefficient

$$A_1 = \frac{\overline{X} - Mo}{\hat{S}} \quad \text{or} \quad A_1 = \frac{\overline{X} - Med}{\hat{S}}$$

- Quartile skewness coefficient

$$A_2 = \frac{Q_3 - 2Med + Q_1}{Q_3 - Q_1}$$

measures skewness for the middle 50% observations only

Interpretation

- positive values= positive asymmetry (right skewed distribution)
- negative values = negative asymmetry (left skewed distribution)
- For the skewness coefficient (with the median) and the quartile skewness coefficient the strength of asymmetry (absolute value):
  - 0 – 0.33: weak
  - 0.34 – 0.66: medium
  - 0.67 – 1: strong

Asymmetry – examples

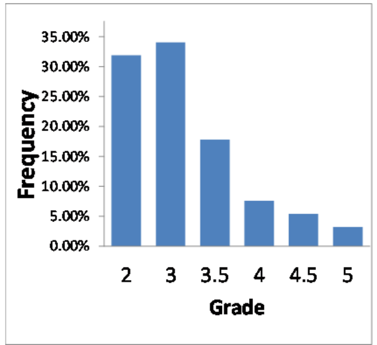
Example 1:

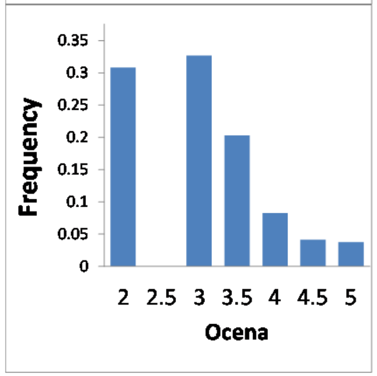
$$A \approx 0.41$$

$$A_1 = \frac{2.99-3}{0.83} \approx -0.01\,(Med)$$

$$A_1 = \frac{2.99-3}{0.83} \approx -0.01\,(Mo)$$

$$A_2 = \frac{3.5-2\cdot 3+2}{3.5-2} = -\frac{1}{3}$$



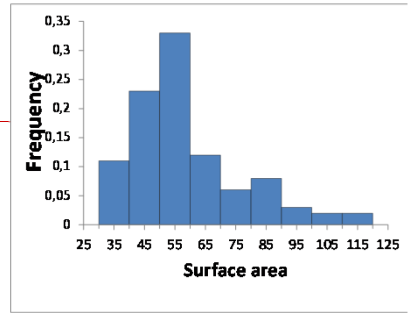


Example 3:

$$A \cong 1.15,$$

$$A_1 \cong \frac{58.7-53.23}{18.2} \approx 0.3\,(Mo) \text{ or } A_1 = \frac{58.7-54.85}{18.2} \approx 0.24\,(Med)$$

$$A_2 \cong \frac{66.67-2\cdot 54.85+46.09}{66.67-46.09} \approx 0.15$$



Boxplot
(Box and whisker plot)

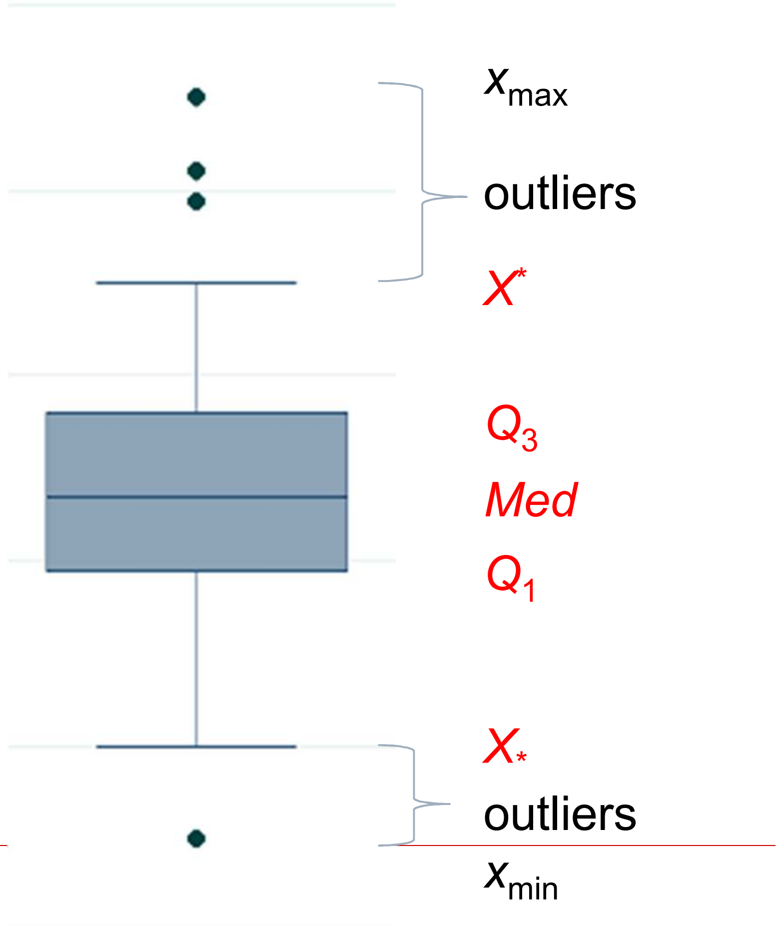
Allows to compare two (or more) populations

$$X_* = \min\{X_i : X_i \in [Q_1 - 3/2\, IQR, Q_1]\}$$

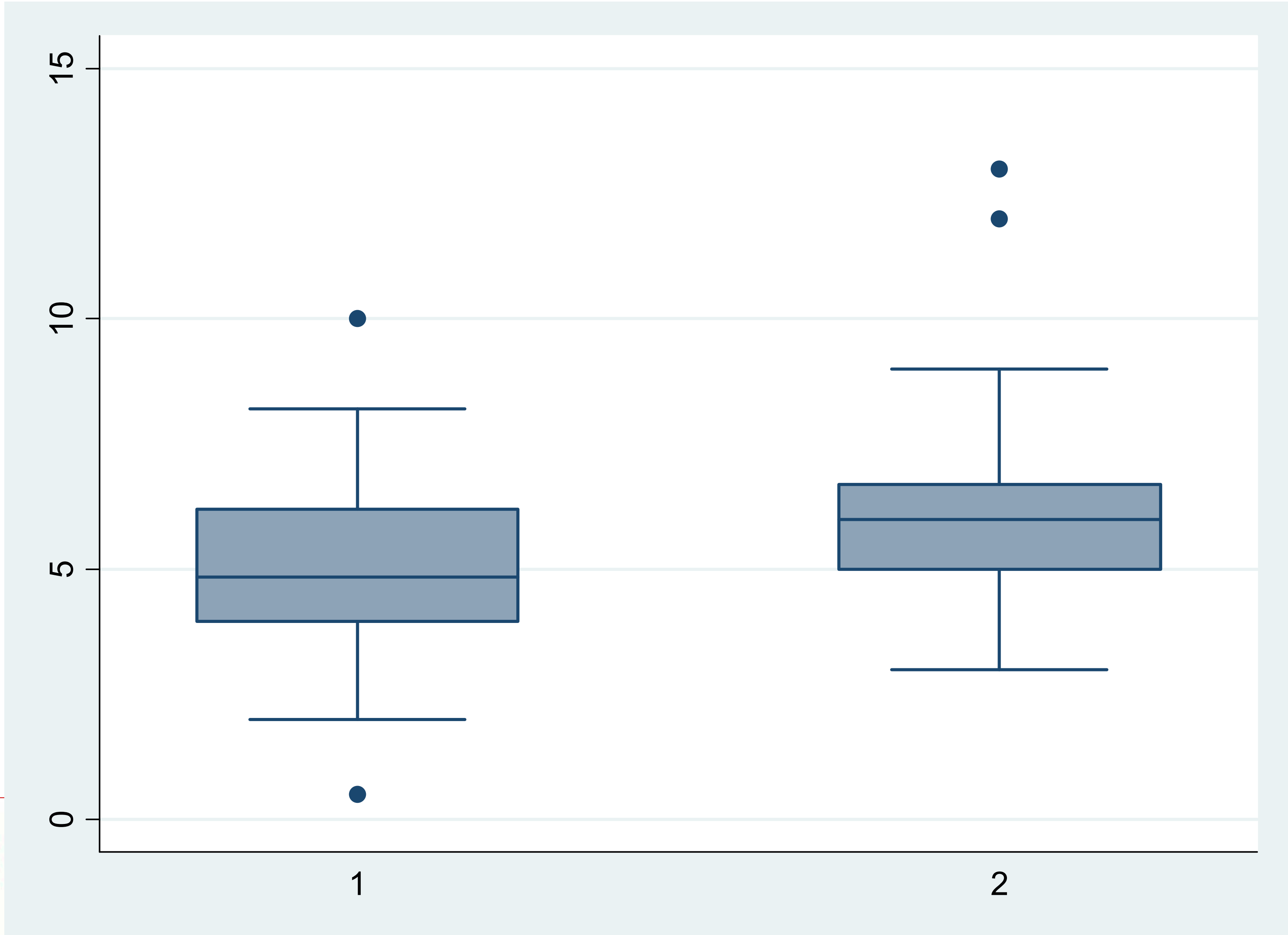
$$X^* = \max\{X_i : X_i \in [Q_3, Q_3 + 3/2\, IQR]\}$$

outliers:

$$x < X_* \text{ or } x > X^*$$



Boxpolot – example of comparison



Examples (1)

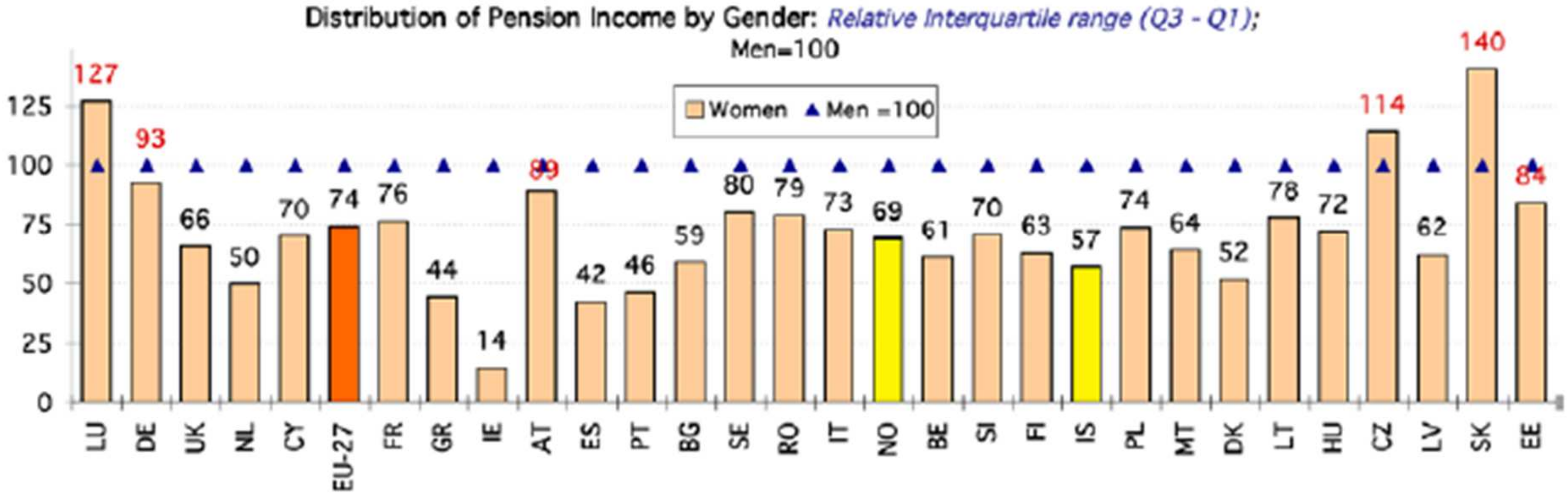
TABL. 16. CHARAKTERYSTYKI OPISOWE CEN DETALICZNYCH NIEKTÓRYCH TOWARÓW ŻYWNOŚCIOWYCH W CZERWCU 2009 R.

DESCRIPTIVE CHARACTERISTICS OF RETAIL PRICES OF SELECTED FOODSTUFFS IN JUNE 2009

Towary *Goods*	Średnia cena w złotych *Average price in zł*				Współczynnik zmienności w % *Variability coefficient in %*	
	najniższa *lowest*	najwyższa *highest*	krajowa *national*			
	województwo *voivodship*		arytme-tyczna *arithmetic*	mediana *median*	ogółem *total*	między-woje-wódzkiej *intervoi-vodship*
Chleb pszenno-żytni - za 0,5 kg *Wheat-rye bread - per 0.5 kg*	1,59 świętokrzyskie	2,19 śląskie	1,88	1,86	19,6	7,6
Mąka pszenna - za 1 kg *Wheat flour - per kg*	1,74 lubuskie	2,19 świętokrzyskie	1,93	1,99	17,5	4,9
Kasza jęczmienna - za 0,5 kg *Pearl-barley - per 0.5 kg*	1,54 podkarpackie	2,13 mazowieckie	1,82	1,80	32,3	10,0
Mięso wołowe z kością (rostbef) - za 1 kg *Beef meat, bone-in (roast beef) - per kg*	17,93 łódzkie	23,51 kujawsko-pomorskie	19,84	19,50	15,9	6,8
Mięso wołowe bez kości (z udźca) - za 1 kg *Beef meat, boneless (gammon) - per kg*	21,89 małopolskie	29,63 zachodniopomorskie	26,15	25,80	15,4	9,5

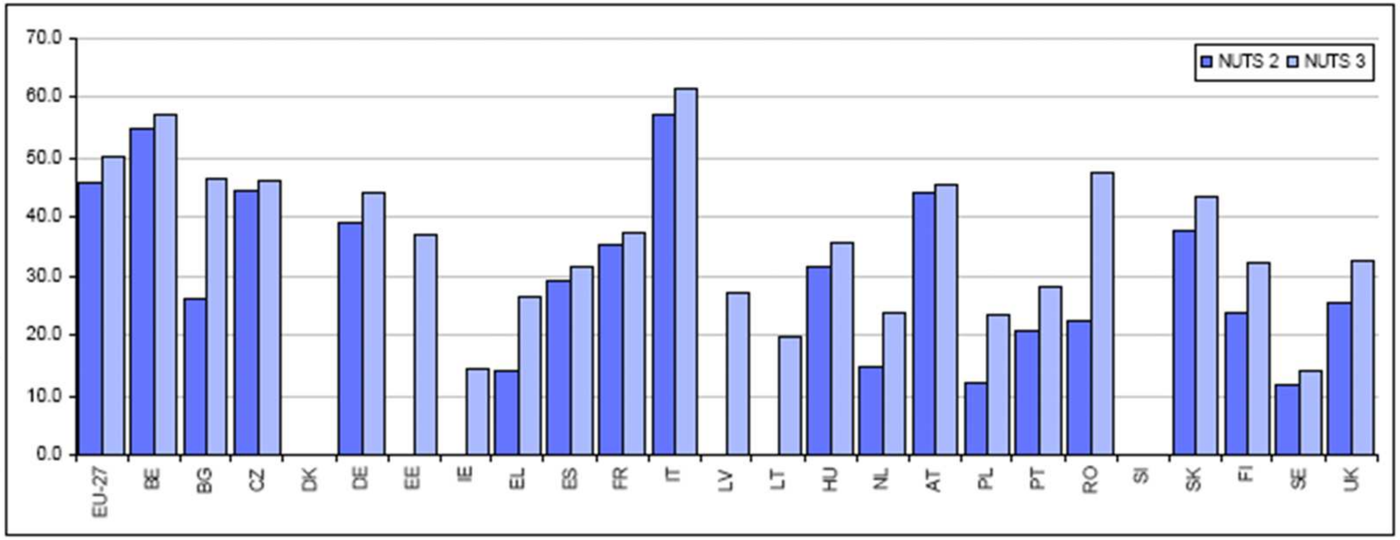
Source: CSO Poland, 2009

Examples (2)



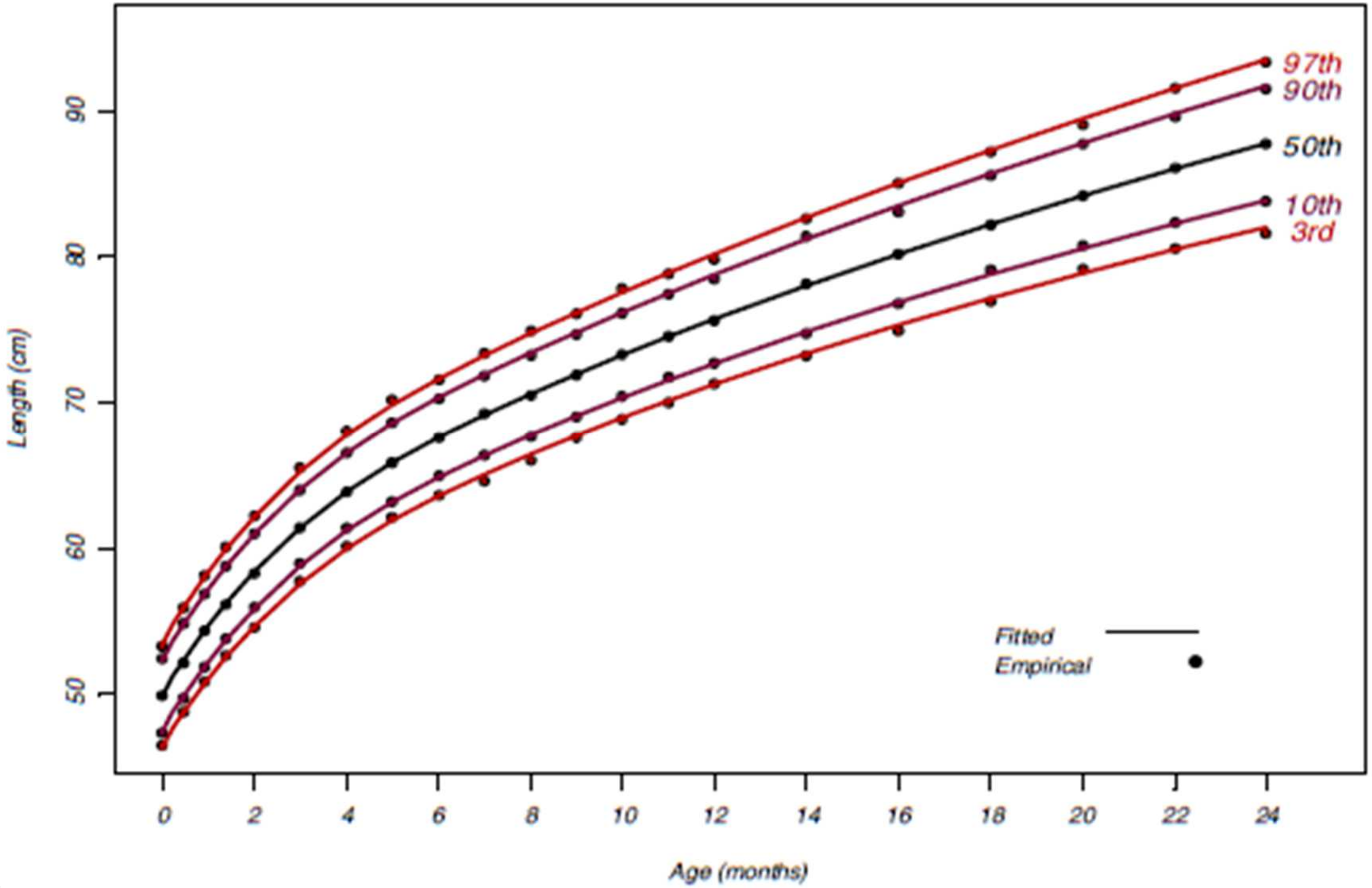
Dispersion (coeffcient of variability) of unemployment rates 2006

Source: European Commission



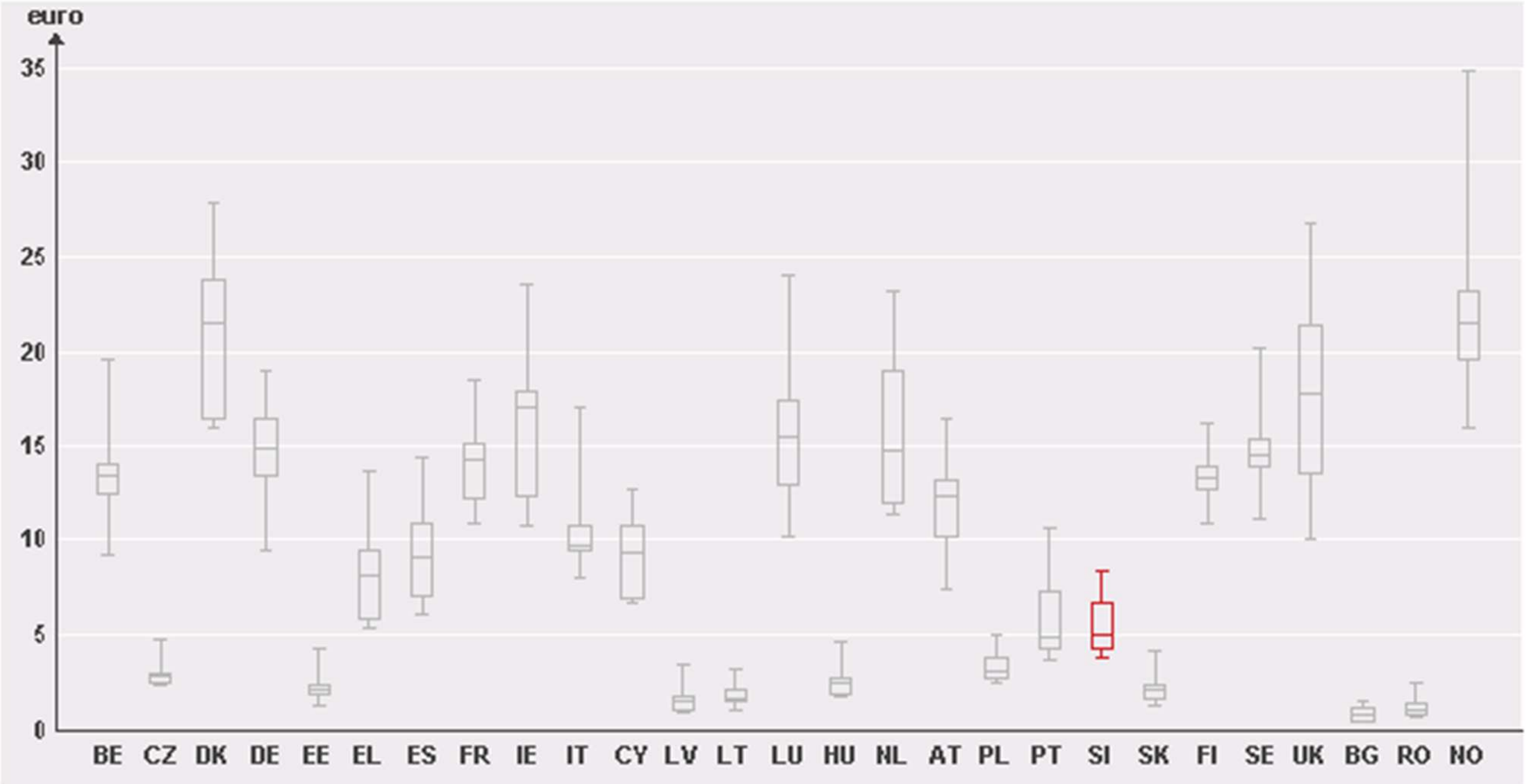
Warsaw University
Faculty of Economic

Examples (3)
Growth charts



Source: WHO

Examples (4)
Gross hourly earinings, 2002



Source: European Commission 2005