II. Analiza struktury – opis statystyczny -wykład

2.1. Filozofia statystyki – STATYSTYKA- jako dyscyplina naukowa. Historia, tradycja, nowoczesność, rzeczywistość i fikcja.

2.2. Podstawowe pojęcia: zbiorowość statystyczna (populacja generalna, próba losowa, próba celowa), cecha statystyczna, liczebność zbiorowości i próby.

2.3. Pojęcie dystrybuanty empirycznej, jej własności.

2.4. Źródła informacji statystycznej (bieżąca sprawozdawczość, badanie statystyczne, materiały wtórne).

2.5. Podstawowe schematy szeregów empirycznych opartych na: liczebnościach, częstościach względnych, skumulowanych liczebnościach i skumulowanych częstościach względnych.

2.6. Podstawowe formy prezentacji materiału (tabelaryczna: tablice proste i złożone; graficzna: różne typy wykresów, schematy rozkładów).

2.7. Zakres analiz statystycznych

2.8. Analiza struktury

- ***2.8.1. Miary przeciętnego poziomu***
- ***2.8.2. Miary zróżnicowania***
- ***2.8.3. Miary asymetrii (skośności) i spiczastości***
- ***2.8.4. Koncentracja zjawiska i jej pomiar***
- ***2.8.5. Położenie miar przeciętnego poziomu, podstawowe typy rozkładów***
- ***2.8.6. Przykład empiryczny***

Zajęcia lab: 1,2

Literatura:

E. Frątczak , A.Korczyński, Statystyka od podstaw z systemem SAS, SGH, Warszawa 2013

J.B. Davis ., Statistics Using SAS. Enterprise Guide, SAS Publishing, SAS Institute Inc., Cary, NC 2007.

G. Keller, Managerial Statistics, 9th international ed., South-Western CENGAGE Learning, UK, USA 2012,

J. Jóźwiak, J.Podgórski: Statystyka od podstaw. Wyd. VI zmienione, PWE, Warszawa 2012

H. Kassyk-Rokicka, Mierniki statystyczne, PWE, Warszawa 2011

Statystyka. Zbiór zadań pod red. H. Kassyk-Rokickiej. Wyd. IV, PWE, Warszawa 2012

2.1. Filozofia statystyki – STATYSTYKA- jako dyscyplina naukowa. Historia, tradycja, nowoczesność, rzeczywistość i fikcja.

Statystyka to dyscyplina naukowa zajmująca się badaniem prawidłowości w procesach masowych, to jest takich, które realizują się na dużą skalę (np. procesy produkcji, procesy dystrybucji, procesy ludnościowe, procesy klimatyczne, itd.). W potocznym znaczeniu – statystyka to zbiór danych liczbowych, to określone bazy danych statystycznych i systemy ewidencji (nazywane rejestrami, np. system PESEL).

Warto pamiętać, że termin „statystyka” pochodzi od łacińskiego słowa „status” (państwo) – począwszy od spisów ludności i bogactw naturalnych w Chinach i Egipcie przeprowadzonych ponad 2000 l p.n.e. aż po czasy współczesne opis słowny przekształcany był coraz częściej na liczbowy opis statystyczny.

Statystyka, jako dyscyplina naukowa obejmuje dwa podstawowe działy: statystykę opisową (opis statystyczny) i statystykę matematyczną (wnioskowanie statystyczne). W ramach dyscypliny wyodrębniło się szereg subdyscyplin: jak statystyka społeczna, statystyka ekonomiczna, statystyka matematyczna, itp.

Jeden z wielkich statystyków XX wieku, prof. C.R.Rao w pracy „Statystyka i prawda”, W-wa, 1994 napisał: **„ statystyka jest nauką, techniką i sztuką”. Sztuką dla statystyka jest skłanianie liczb by same mówił , co zależy od wprawy i dużego doświadczenia statystyka.**

2.2. Podstawowe pojęcia: zbiorowość statystyczna (populacja generalna, próba losowa, próba celowa), cecha statystyczna, liczebność zbiorowości i próby.

Zbiorowość statystyczna – to zbiór jednostek wyodrębnionych pod względem czasu, przestrzeni i więzi logicznej pomiędzy nimi. Zatem każda jednostka wchodząca w skład zbiorowości statystycznej powinna być jasno określona pod względem trzech, uprzednio wymienionych kryteriów. Jeśli zbiorowość statystyczna obejmuje wszystkie jednostki – to nosi nazwę populacji (np. populacja - liczba ludności Polski).

Ogólna liczba jednostek wchodzących w skład zbiorowości statystycznej (populacji) nazywa się liczebnością zbiorowości. W przypadku, kiedy zbiorowość statystyczna jest skończona i policzalna ogólną jej liczebność zapisuje się:

$N = n_1 + n_2 + n_3 + \ldots\ldots\ldots + n_{N-1} + n_N = \sum n_i$

N – liczebność ogólna zbiorowości;

i = 1,2,3,…, N

Część jednostek populacji pobranych w sposób losowy lub celowy nosi nazwę **próby**.

Liczebność próby oznaczać się będzie jako *n*.

$n = n_1 + n_2 + n_3 + \ldots\ldots\ldots + n_{N-1} + n_k = \sum n_i$

i = 1,2,3,…, k

Cecha statystyczna – to właściwość jednostek zbiorowości statystycznej powodująca, że jednostki różnią się między sobą – daje podstawę do agregacji i dezagregacji jednostek.

Przyjmuje się umownie, że dla oznaczenia cech statystycznych stosuje się duże litery:

X, Y, Z , zaś do oznaczenia ich realizacji – stosuje się małe litery: x, y, z. Duża różnorodność występujących cech umożliwia dokonanie ich odpowiedniej klasyfikacji. Podstawowy podział cech statystycznych wyodrębnia dwie grupy: cechy mierzalne

(ilościowe) cechy niemierzalne (jakościowe).

Cechy niemierzalne to takie właściwości jednostek zbiorowości statystycznej, które nie dadzą się przedstawić liczbowo, a zamienność ich jest możliwa do opisania jedynie za pomocą określeń słownych. Do tej grupy zaliczane są takie cechy jak: płeć, wykształcenie, miejsce pracy, stan cywilny, pochodzenie społeczne.

Cechy mierzalne to takie właściwości jednostek zbiorowości statystycznej, których zmienność daje się opisać liczbowo przy pomocy odpowiednich jednostek miary. Do tej grupy należą takie cechy jak: wiek (w latach), waga (w kg), wzrost (w cm), wynagrodzenie (w zł), liczba osób w rodzinie.

Wśród cech mierzalnych (ilościowych) istnieje możliwość dokonania ich dalszego podziału podstawą, którego jest zakres zmienności cechy w dowolnie obranym przedziale przyjmowanych wartości.

Jeżeli wartości przyjmowane przez cechę różnią się między sobą o pełną jednostkę – to takie cechy przyjęto nazywać **cechami dyskretnymi (skokowymi).** Przykładem takich cech są: liczba osób wchodząca w skład gospodarstwa domowego, liczba pracowników zatrudnionych w danej firmie, liczba wypadków drogowych, które miały miejsce w określonym czasie, itd. Wartości przyjmowane przez ww. cechy są zawsze liczbami całkowitymi.

W przypadku, kiedy wartości przyjmowane przez cechy różnią się między sobą o część jednostki, w jakiej wyrażana jest zmienność cech, to cechy takie nazywane są **cechami ciągłymi**. Oznacza to, że pomiędzy dowolnie wybranymi wartościami cechy mogą występować inne wartości pośrednie (zachowana jest ciągłość w dowolnie wybranym przedziale zmienności). Przykładem takich cech są: staż pracy (w latach), waga (w kg.), długość (w cm), itd.

Dodatkową podgrupą są **cechy quasi ilościowe**, **zwane porządkowymi**. Cechy te kwantyfikują zwykle natężenie badanej właściwości przestrzennej w sposób opisowy, porządkując w ten sposób zbiorowość: na przykład ocena z przedmiotu: bardzo dobra, dobra, dostateczna, lub krócej w zapisie: 5; 4; 3.

Generalnie dla cech skokowych (dyskretnych) charakterystyczne jest ujęcie punktowe:

$X = x_i$ lub $Y = y_j$.

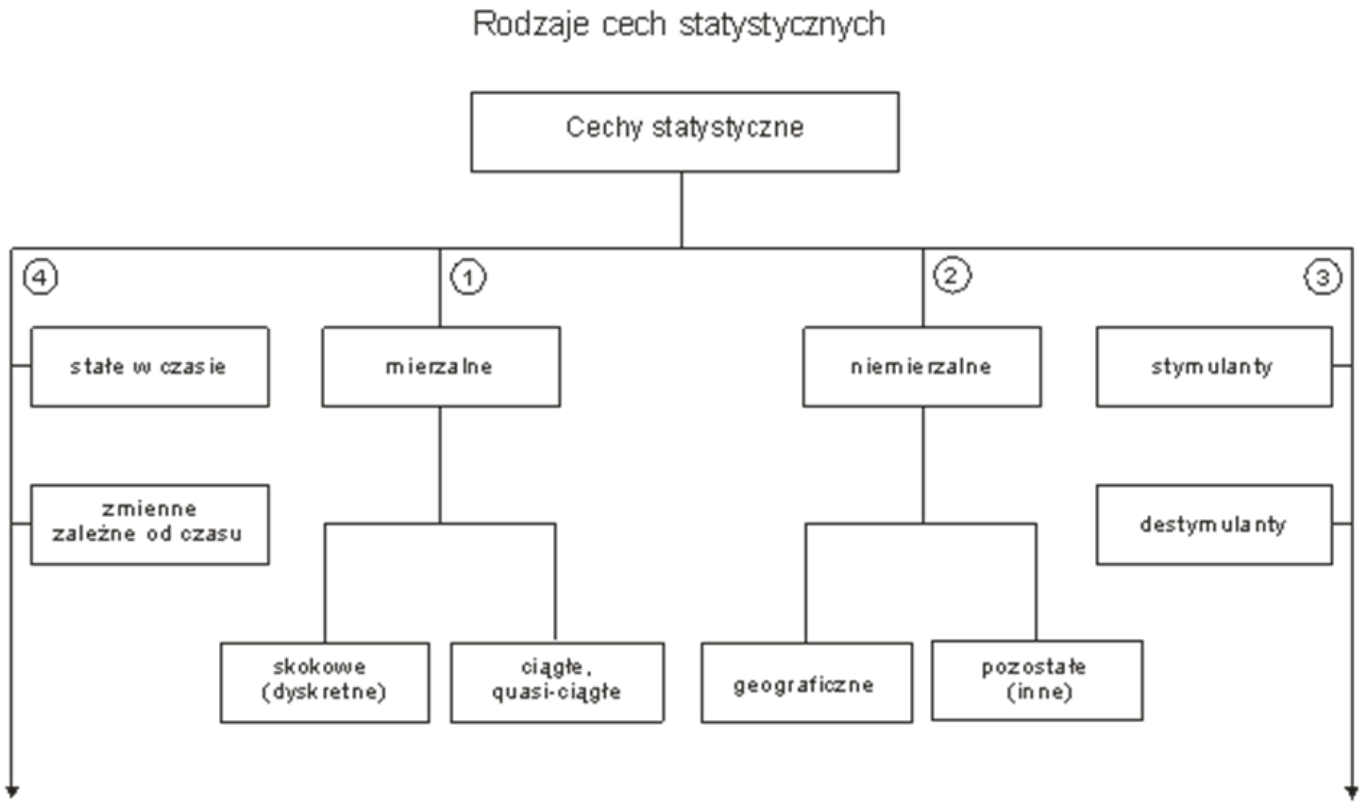
Dla cech ciągłych charakterystyczne jest ujęcie przedziałowe, co można zapisać:

$X \in [x_{0i} - x_{1i}]$, $Y \in [y_{0i} - y_{1i}]$

Oprócz ww. podziału cech statystycznych można spotkać inne jeszcze podziały.

Reasumując rozważania na temat cech statystycznych można je pogrupować jak przedstawiono na Schemacie nr 1.

Schemat 1. Rodzaje cech statystycznych



2.3. Pojęcie dystrybuanty empirycznej, jej własności.

Dystrybuanta empiryczna to skumulowana częstość względna dla wartości $x \leq x_i$.

Dla cechy skokowej można zapisać:

$G(x_i) = w\ (x \leq x_i)$.

Jest funkcją niemalejącą, przedziałami stałą, przyjmującą wartości z przedziału: 0 – 1.

2.4. Źródła informacji statystycznej (bieżąca sprawozdawczość, badanie statystyczne, materiały wtórne).

Do podstawowych źródeł informacji statystycznej zalicza się:

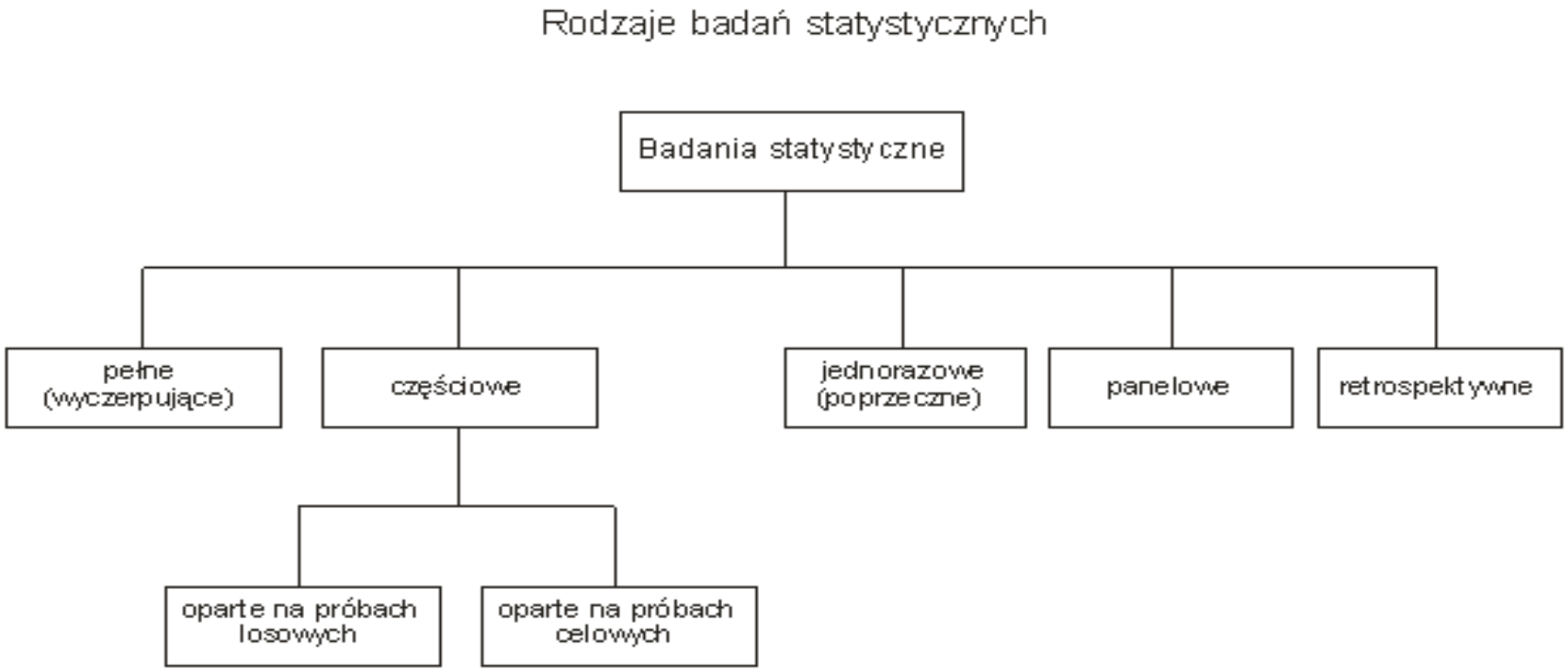
- ***bieżącą sprawozdawczość i ewidencję*** (np. sprawozdawczość poprzez system sprawozdań rejestrowany w Głównym Urzędzie Statystycznym: **www.stat.gov.pl**). Rejestracja bieżąca polega na systematycznej ewidencji ściśle określonych faktów, które traktowane są jako przedmiot badania (przykładem jest ewidencja faktów demograficznych: urodzeń, zgonów, migracji).

- ***badania statystyczne,*** które mogą być badaniami pełnymi (całkowitymi, wyczerpującymi) wtedy, kiedy obserwacji poddana jest każda jednostka obserwacji statystycznej lub badaniami częściowymi (opartymi na próbach celowych lub losowych).

Rodzaje badań statystycznych zamieszcza Schemat 2.

- ***materiały wtórne*** **–** to dane gromadzone do określonych celów, których wykorzystanie może być różne w zależności od potrzeb.

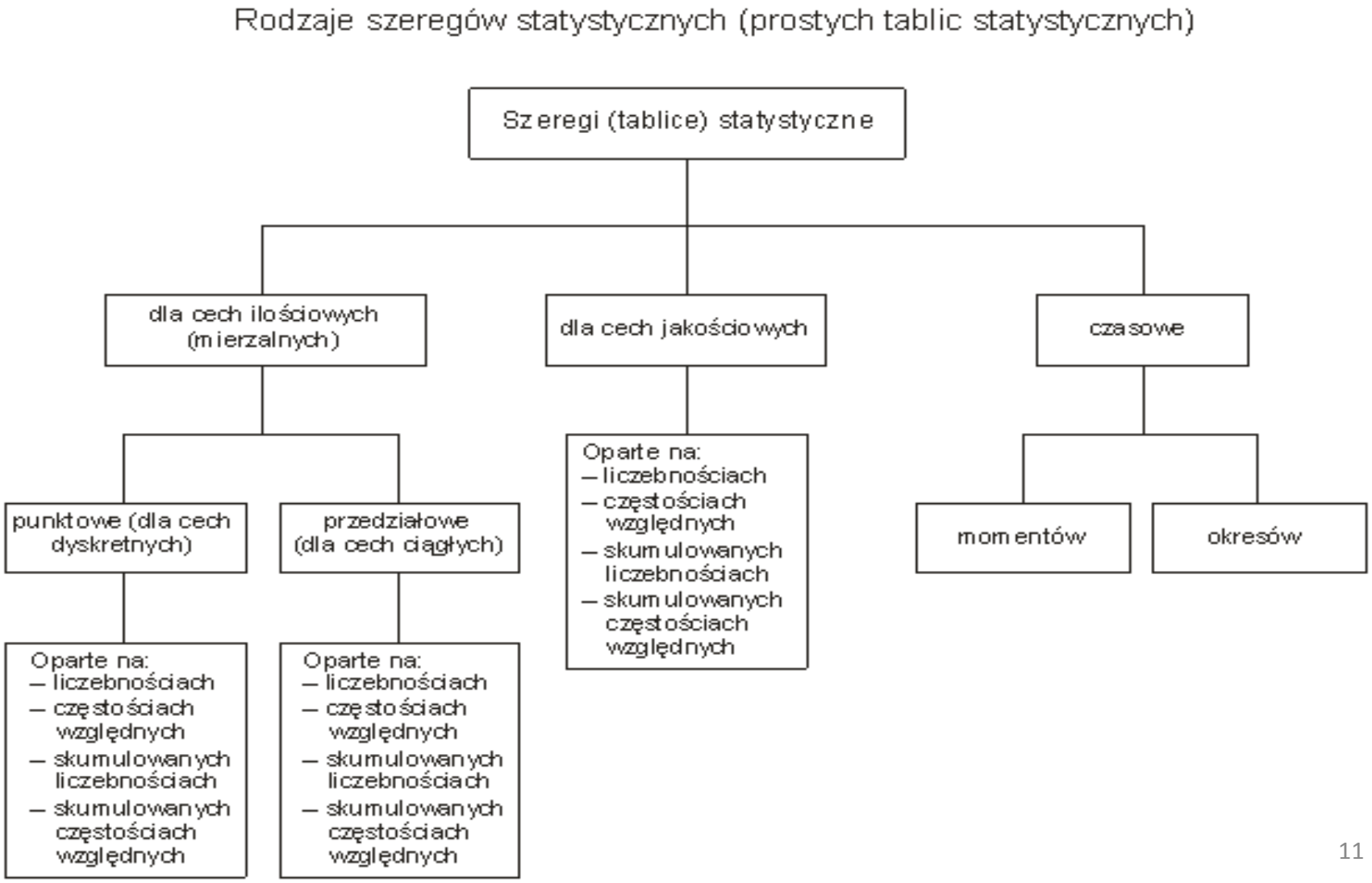
Schemat 2. Rodzaje badań statystycznych



2.5. Podstawowe schematy szeregów empirycznych opartych na: liczebnościach, częstościach względnych, skumulowanych liczebnościach i skumulowanych częstościach względnych.

Szeregi statystyczne to proste tablice statystyczne – zawierają informacje o jednej cesze. Stosownie do podziału cech statystycznych dzielą się na szeregi dla cech skokowych i ciągłych. Ich podstawowy podział zamieszcza Schemat 3. Przykładowe schematy szeregów zapisane na symbolach zamieszczono po Schemacie 3.

Schemat 3. Podstawowe rodzaje szeregów statystycznych



1. Schemat szeregu empirycznego dla cechy skokowej opartego na liczebnościach

$X = x_i$	n_i
x_1	n_1
x_2	n_2
x_3	n_3
.....	.....
.....	.....
x_k	n_k
$\sum$	n (N)

2. Schemat szeregu empirycznego dla cechy skokowej opartego na częstościach względnych

$X = x_i$	w_i
x_1	w_1
x_2	w_2
x_3	w_3
.....	.....
.....	.....
x_k	w_k
$\sum$	1.0 (100 %)

$w_i = n_i / n$ lub $w_i = n_i / N$

w_i *C (100) interpretacja w %

$\Sigma w_i = 1.0$ lub 100 %

3. Schemat szeregu empirycznego dla cechy skokowej opartego na skumulowanych liczebnościach

$X = x_i$	$n_{i,\,sk}$
x_1	n_1
x_2	$n_1 + n_2$
x_3	$n_1 + n_2 + n_3$
.....	.....
.....	.....
x_k	$n_1+n_2+ \ldots+n_k = n\ (N)$
$\sum$	x

4. Schemat szeregu empirycznego dla cechy skokowej opartego na skumulowanych częstościach względnych

$X = x_i$	$w_{i,sk}$
x_1	w_1
x_2	$w_1 + w_2$
x_3	$w_1 + w_2 + w_3$
.....	.....
.....	.....
x_k	$w_1+w_2+ ...+w_k=1.0$ lub 100%
$\sum$	x

5. Schemat szeregu empirycznego dla cechy ciągłej opartego na liczebnościach

x_{0i} - x_{1i}	n_i
x_{o1} - x_{11}	n_1
x_{02} - x_{12}	n_2
x_{03} - x_{13}	n_3
.....	.....
.....	.....
x_{0k} - x_{1k}	n_k
$\sum$	n (N)

6. Schemat szeregu empirycznego dla cechy ciągłej opartego na częstościach względnych

$x_{0i} - x_{1i}$	w_i
$x_{01} - x_{11}$	w_1
.....x_{02}..-..x_{12}	w_2
x_{03}..-..x_{13}	w_3
.....	.....
.....	.....
$x_{0k} - x_{1k}$	w_k
$\sum$	1.0 (100 %)

$wi = n_i / n$ lub $w_i = n_i / N$

w_i * C (100) interpretacja w %

$\Sigma w_i = 1.0$ lub 100 %

7. Schemat szeregu empirycznego dla cechy ciągłej opartego na skumulowanych liczebnościach.

$x_{0i} - x_{1i}$	$n_{i, sk}$
$x_{01} - x_{11}$	n_1
$x_{02} - x_{12}$	$n_1 + n_2$
$x_{03} - x_{13}$	$n_1 + n_2 + n_3$
.....	.....
.....	.....
$x_{0k} - x_{1k}$	$n_1+n_2+ ...+n_k = n$ (N)
$\sum$	x

8. Schemat szeregu empirycznego dla cechy ciągłej opartego na skumulowanych częstościach względnych.

$x_{0i} - x_{1i}$	$w_{i,sk}$
$x_{01} - x_{11}$	w_1
$x_{02} - x_{12}$	$w_1 + w_2$
$x_{03} - x_{13}$	$w_1 + w_2 + w_3$
.....	.....
.....	.....
$x_{0k} - x_{1k}$	$w_1+w_2+ ...+w_k=1.0$ lub 100%
$\sum$	x

STOSOWANE ZNAKI W TABLICACH STATYSTYCZNYCH

(-) - zjawisko nie wystąpiło

(0,0) – zjawisko istniało, jednakże w ilościach mniejszych od liczb, które mogły być wyrażone uwidocznionymi w tablicy znakami cyfrowymi

(.) - zupełny brak informacji, albo brak informacji wiarygodnych

x - wypełnienie pozycji, ze względu na układ tablicy jest niemożliwe i niecelowe

„ w tym „ – oznacza, ze nie podaje się wszystkich składników sumy

- oznacza, ze dane nie mogą być opublikowane ze względu na konieczność zachowania tajemnicy statystycznej w rozumieniu ustawy o statystyce publicznej.

2.6. Podstawowe formy prezentacji materiału (tabelaryczna: tablice proste i złożone; graficzna: różne typy wykresów) - PRZYKŁADY TABLIC I WYKRESÓW

Example 1 COCA –COLA (1)

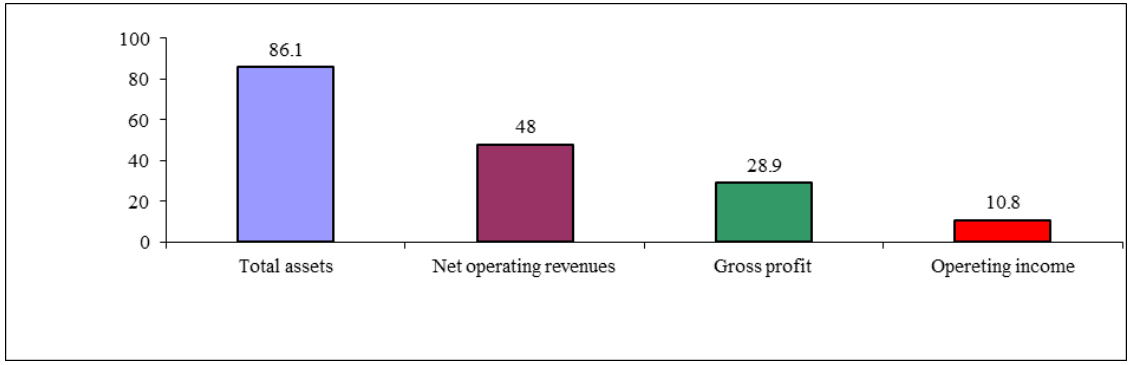
The following information taken from the annual report of Coca-Cola Company, given in the form of tables (Table 1,2,3,4). Present the information in graphical form.

Table 1. General information

Industry	Beverage
Founded	1886
Founder	Asa Griggs Candler
CEO	Muhtar Kent
Traded	NYSE: KO S&P 500 Component Dow Jones Industrial Average Component
Employees	150,900
Total assets	$ 86,174,000,000
Net operating revenues	$ 48,017,000,000
Gross profit	$ 28,964,000,000
Operating income	$ 10,779,000,000

Source: The Coca-Coll Company Annual Report 2012
http://www.coca-colacompany.com/annual-review/2012/pdf/form_10K_2012.pdf

Figure 1. General information (in billion USD)



Source: based on the data from Table 1 (origin source as in table 1).

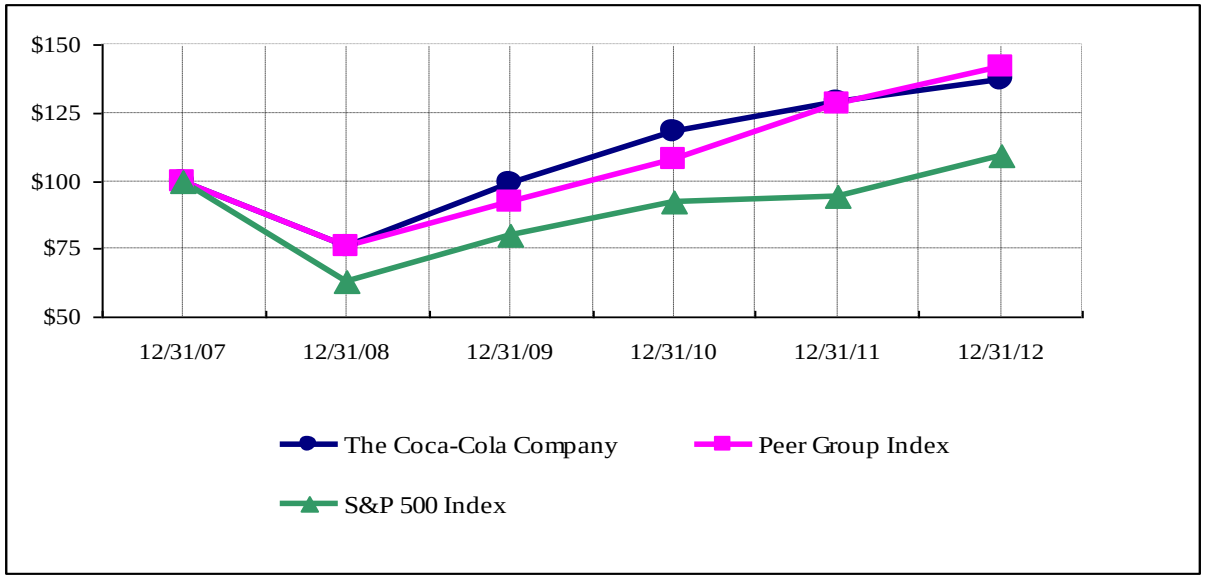
Example 1 COCA –COLA (2)

Table 2. Data on cumulative total return

Items, December 31,		2007	2008	2009	2010	2011	2012
The Coca-Cola Company	$	100	76	99	118	129	137
Peer Group Index	$	100	76	92	108	128	142
S&P 500 Index	$	100	63	80	92	94	109

Source: as in Table 1.

Figure 2. The cumulative total return



Source: based on the data from Table 2.

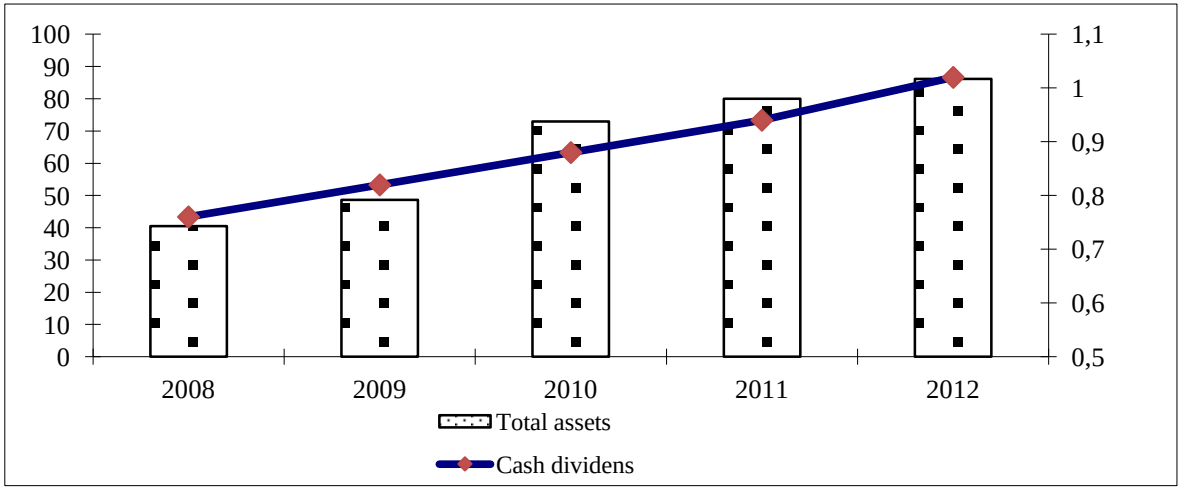
Example 1 COCA –COLA (3)

Table 3. Total assets and cash dividends

Items, December 31	2008	2009	2010	2011	2012
Total assets	40,519	48,671	72,921	79,974	86,174
Cash dividends	0.76	0.82	0.88	0.94	1.02

Source: as in Table 1.

Figure 3. Total assets and cash dividends



Source: Source: based on the data from Table 3

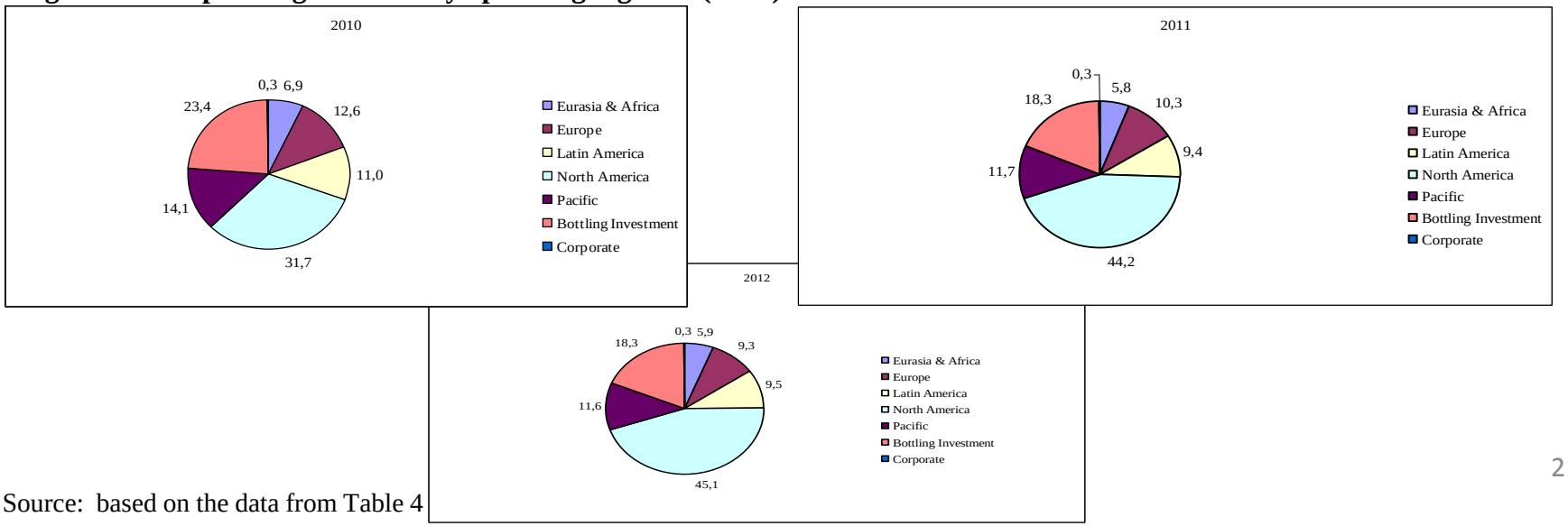
Example 1 COCA –COLA (4)

Table 4. Net operating revenues by operating segment (in %)

Year ended December 31	2012	2011	2010
Eurasia & Africa	5,9	5,8	6,9
Europe	9,3	10,3	12,6
Latin America	9,5	9,4	11,0
North America	45.1	44,2	31,7
Pacific	11.6	11,7	14,1
Bottling Investment	18.3	18,3	23,4
Corporate	0.3	0,3	0,3
Total	100,0%	100,0	100,0

Source: as in Table 1.

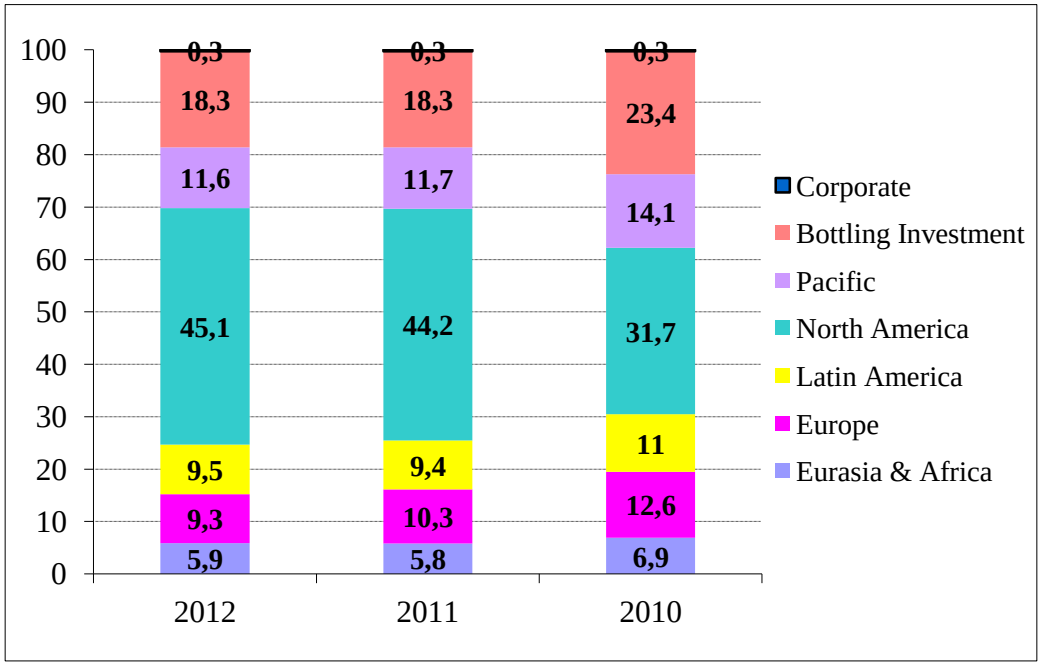
Figure 4. Net operating revenues by operating segment (in %). Panel A.



Source: based on the data from Table 4

Example 1 COCA –COLA (5)

Figure 4. Net operating revenues by operating segment (in %). Panel A.



Source: based on the data from Table 4

Example 2 CEO (1)

The data and graph presenting the results of the original surveys from the Forbes (2012) website (source given below): “Two Decades of CEO Pay” (the source table and graph presented below). The data covering the years 1989 – 2012 include information on 10 variables. The graph presented below (Figure 1), originally downloaded from the Forbes website, shows only a part of the information included in the table. Using the data, prepare a graphical report presenting information on “Two Decades of CEO Pay” in different graphical forms. The constructed graphs can be supplemented with verbal comments.

The dates on the next slide represent the issue years for the Forbes Executive Compensation reports. For years 1989 through 1999, our universe was the 800 biggest companies in the U.S. For years 2000 through the present, our universe was the 500 biggest companies in the U.S.

Example 2 CEO (2)

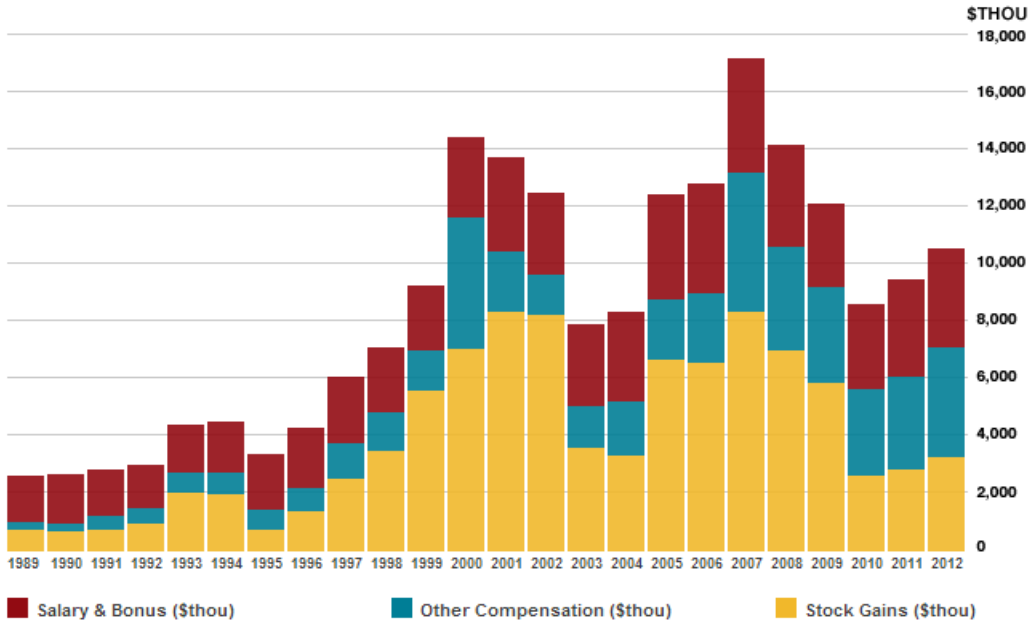
Table 1. Two Decades of CEO – selected characteristics

Year	Age	Years with company	Years as CEO	Percent owned (%)	Median value stock owned ($Mil)	Salary + bonus ($thous)	Other compen-sation ($thous)	Stock gains ($thous)	Total compensation ($thous)
1989	56,6	23,4	8,3	2,1	4	1615	226	708	2548
1990	56,4	23,2	8,3	2,5	4	1683	250	689	2622
1991	56,3	22,7	8,1	2,5	4	1638	548	611	2797
1992	56,4	22,6	8	2,3	6	1571	476	898	2945
1993	56,4	22,7	8,1	2,2	6	1654	726	1949	4330
1994	56,3	22	8,1	2,2	7	1780	791	1870	4441
1995	56,4	22	8,3	2,2	7	1980	700	655	3336
1996	56	21,1	8	2	8	2073	811	1309	4193
1997	56,4	21,1	8,2	2	10	2361	1247	2434	6043
1998	56,3	20,8	8	2	14	2309	1377	3384	7070
1999	55,9	20,1	8	2,1	13	2279	1405	5502	9186
2000	55,1	19,9	7,3	2,3	19	2856	4529	6955	14340
2001	55,4	20	7,2	1,7	14	3363	2005	8300	13668
2002	55,6	19,9	7,6	1,5	15	2856	1424	8142	12422
2003	55,2	19,6	7,2	1,4	11	2840	1459	3514	7813
2004	55,7	19,7	7,3	1,3	15	3158	1853	3263	8274
2005	55,7	19,2	7,1	1,5	17	3663	2112	6622	12397
2006	55,5	19,3	7,7	1,5	19	3835	2390	6515	12740
2007	55,7	18,7	7,3	1,4	19	4005	4879	8231	17116
2008	55,8	18,7	7,1	1,3	15	3637	3564	6924	14110
2009	56	18,9	7,3	1,2	9	2978	3274	5834	12086
2010	56,3	19	7,4	1,2	12	3034	2963	2546	8543
2011	56,6	19	7,6	1,1	14	3391	3211	2824	9426
2012	57	19,3	7,9	1	17	3508	3760	3233	10502

Sources: *Forbes Annual Executive Compensation Reports; U.S. Bureau of Labor Statistics(www.bls.gov)*

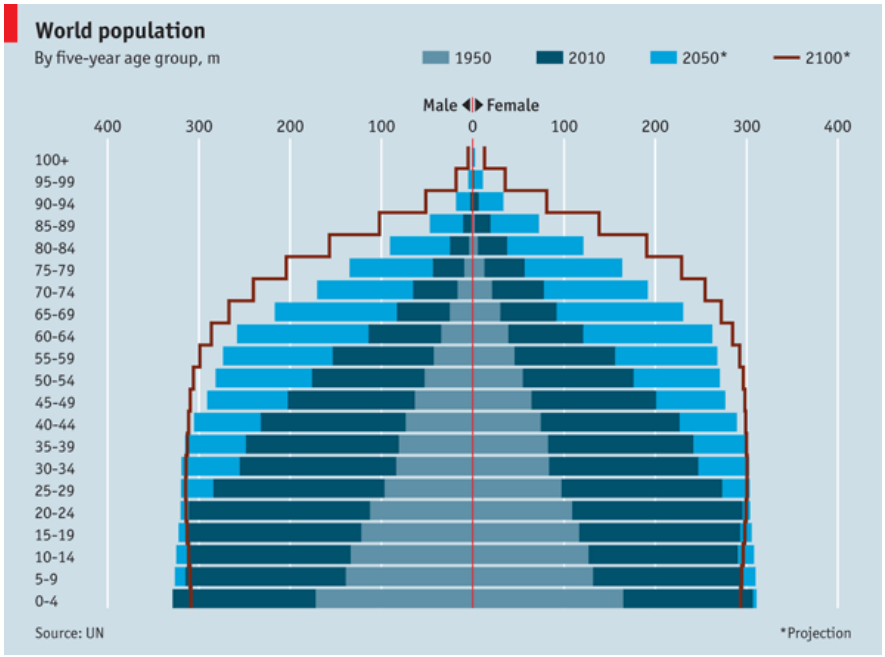
Example 2 CEO (2)

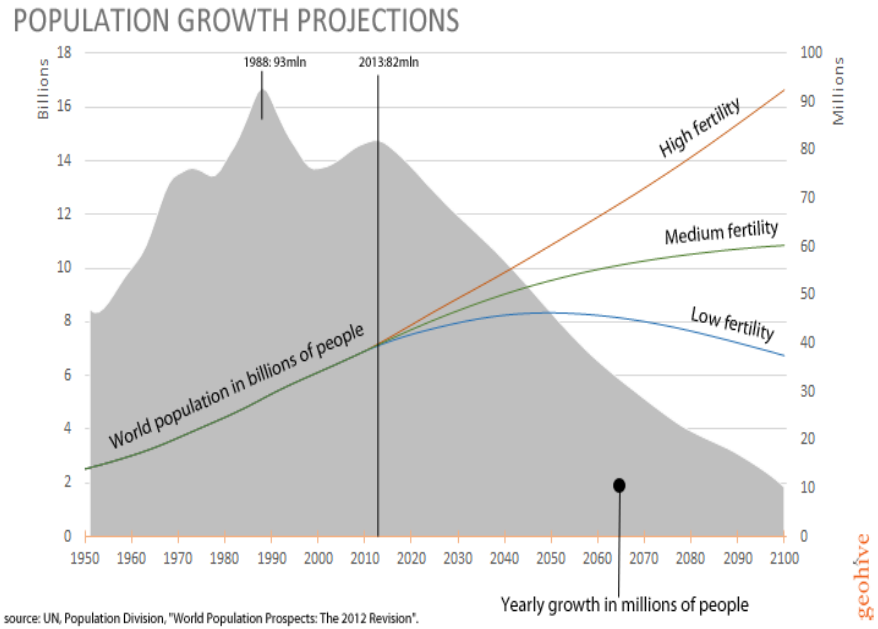
Figure 1. Two Decades of CEO Pay

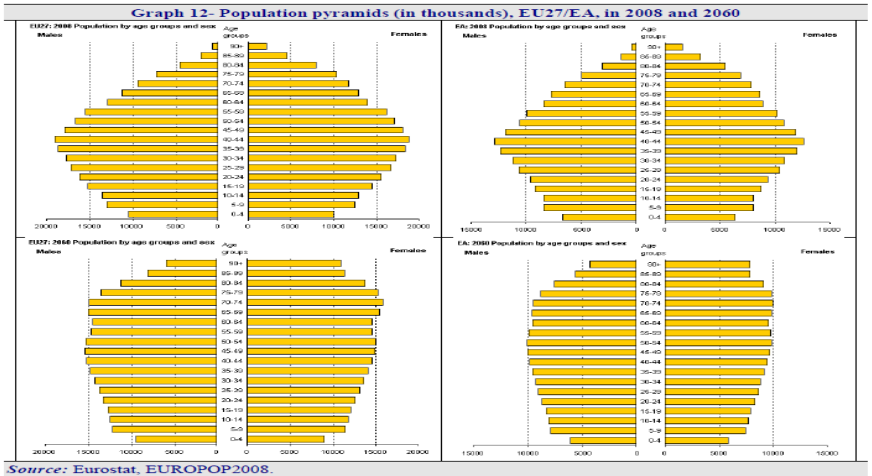


Source: Source: http://www.forbes.com/lists/2012/12/ceo-compensation-12-historical-pay-chart.html**Solution : several types of graphs (**without verbal comments)

Example 3 Population







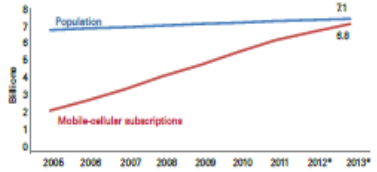
Example 4 Business (1)

The World in 2013 **ICT FACTS AND FIGURES**

"I am pleased to present the latest ICT Facts and Figures which show continued and almost universal growth in ICT uptake. Every day we are moving closer to having almost as many mobile- cellular subscriptions as people on earth. This is exciting news. The mobile revolution is m-powering people in developing countries by delivering ICT applications in education, health, government, banking, environment and business. Let us all celebrate this mobile miracle that I have no doubt will hasten our pace towards sustainable development."
Brahima Sanou, Director of the ITU Telecommunication Development Bureau

6.8 BILLION MOBILE-CELLULAR SUBSCRIPTIONS

As the number of subscriptions approaches global population figures mobile-cellular growth slows



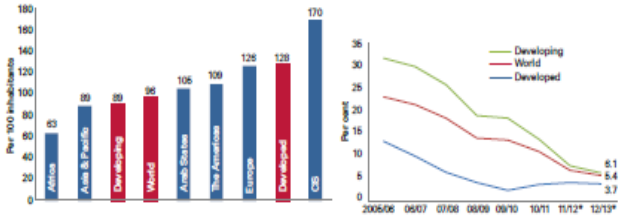
Source: ITU World Telecommunication ICT Indicators database
Note: * Estimate

In 2013, there are almost as many mobile-cellular subscriptions as people in the world, with more than half in the Asia-Pacific region (3.5 billion out of 6.8 billion total subscriptions).

As global mobile-cellular penetration approaches 100% and market saturation is reached, growth rates have fallen to their lowest levels in both developed and developing countries.

Mobile-cellular penetration rates stand at 96% globally; 128% in developed countries; and 89% in developing countries.

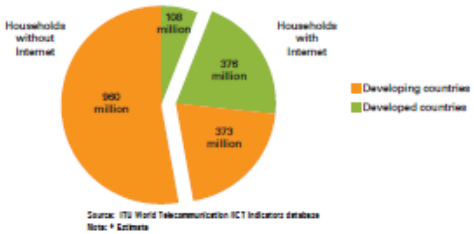
Mobile-cellular penetration, 2013*, and mobile-cellular subscription growth rates, 2005-2013*



Source: ITU World Telecommunication ICT Indicators database
Note: * Estimate

750 MILLION HOUSEHOLDS – 41% GLOBALLY – CONNECTED TO THE INTERNET

Households with Internet access, 2013*



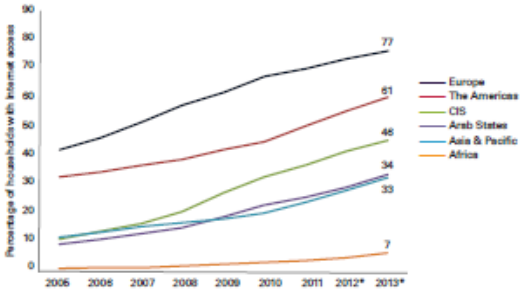
Source: ITU World Telecommunication ICT Indicators database
Note: * Estimate

In 2013, 41% of the world's households are connected to the Internet. Half of them are in the developing world, where household Internet penetration has reached 28%.

In the developed world, 78% of all households are connected to the Internet.

90% of the 1.1 billion households not connected to the Internet are in the developing world.

Households with Internet access, by region



Source: ITU World Telecommunication ICT Indicators database
Note: * Estimate

Europe and Africa are the regions with the highest and the lowest levels of household Internet penetration respectively: 77% in Europe, compared with 7% in Africa.

The majority of households in the Americas are online (61%), compared with around one third of households in the Arab States and Asia and the Pacific.

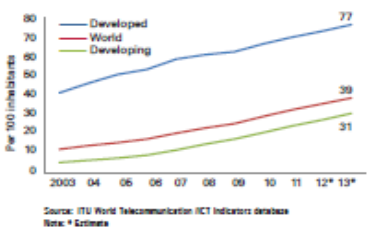
Between 2009 and 2013, Internet penetration in households has grown fastest in Africa, with annual growth of 27%, followed by 15% annual growth in Asia and the Pacific, the Arab States and the CIS.

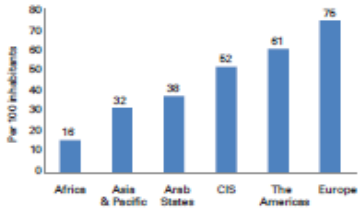
Example 4 Business (2)

2.7 BILLION PEOPLE – ALMOST 40% OF THE WORLD´S POPULATION – ARE ONLINE

In developing countries, 16% fewer women than men use the Internet

Internet users by development level, 2003-2013*, and by region, 2013*





Source: ITU World Telecommunication /ICT Indicators database
Note: * Estimate

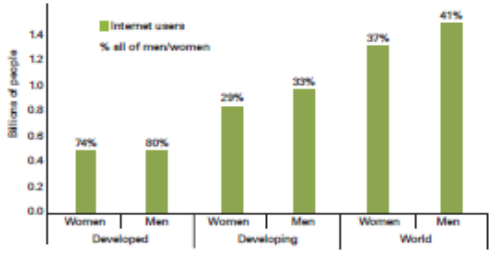
In 2013, over 2.7 billion people are using the Internet, which corresponds to 39% of the world's population.

In the developing world, 31% of the population is online, compared with 77% in the developed world.

Europe is the region with the highest Internet penetration rate in the world (75%), followed by the Americas (61%).

In Africa, 16% of people are using the Internet – only half the penetration rate of Asia and the Pacific.

The gender gap: men and women online, totals and penetration rates, 2013*



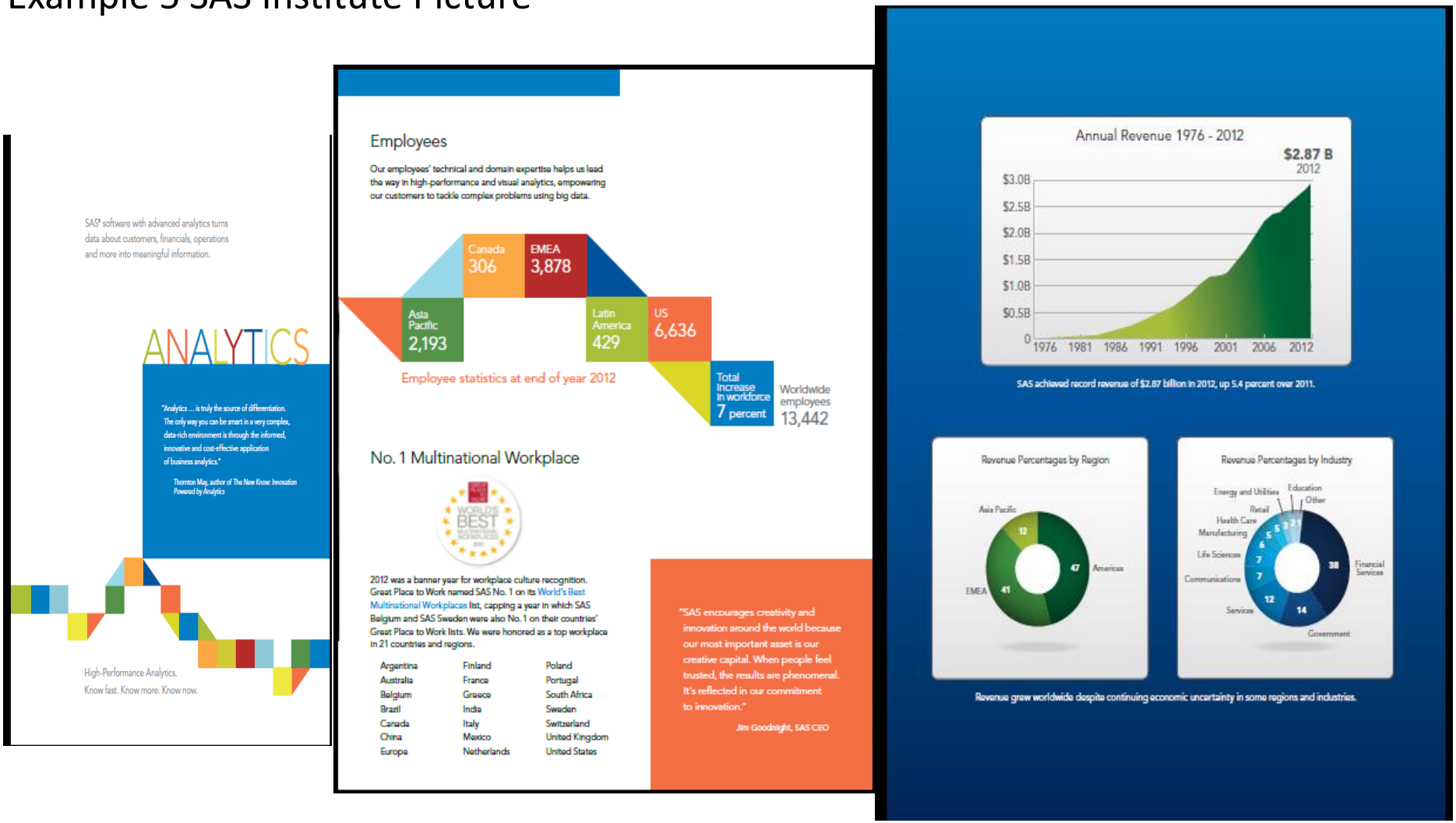
Source: ITU World Telecommunication /ICT Indicators database
Note: * Estimate

More men than women use the Internet: globally, 37% of all women are online, compared with 41% of all men. This corresponds to 1.3 billion women and 1.5 billion men.

The developing world is home to about 826 million female Internet users and 980 million male Internet users. The developed world is home to about 475 million female Internet users and 483 million male Internet users.

The gender gap is more pronounced in the developing world, where 16% fewer women than men use the Internet, compared with only 2% fewer women than men in the developed world.

Example 5 SAS Institute Picture



2.7. Zakres analiz statystycznych

Generalnie analizy statystyczne obejmują następujące zakresy (przekroje):

- opis i wnioskowanie statystyczne w analizie struktury zjawisk
- opis i wnioskowanie statystyczne w analizie współzależności zjawisk
- opis i wnioskowanie statystyczne w analizie dynamiki zjawisk

2.8. Analiza struktury

2.8.1. Miary przeciętnego poziomu

Miary przeciętnego poziomu (zwane inaczej miarami położenia) dzielą się na klasyczne i pozycyjne. Klasyczne to: średnia arytmetyczna, średnia harmoniczna, średnia geometryczna. Stosownie do rodzaju danych statystycznych stosuje się różne formuły (postacie wzorów) na wyznaczenie każdej z wartości miar. Każda z miar posiada określone własności i kryteria jej stosowania.

Miary przeciętnego poziomu – podstawowe formuły

1. Miary klasyczne.

1.1 Średnia arytmetyczna

– dane indywidualne:

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j \qquad \bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j$$

– dane pogrupowane, cecha skokowa:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \qquad \bar{x} = \sum_{i=1}^{k} x_i w_i \qquad \text{i =1,2,...,k}$$

– dane pogrupowane, cecha ciągła:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \dot{x}_i n_i \qquad \bar{x} = \sum_{i=1}^{k} \dot{x}_i w_i \qquad \text{i =1,2,...,k}$$

– średnia ze średnich cząstkowych:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} \bar{x}_i n_i \qquad \bar{x} = \sum_{i=1}^{k} \bar{x}_i w_i \qquad \text{i =1,2,...,k}$$

1.2 Średnia harmoniczna

– dane indywidualne:

$$\bar{x}_h = \frac{n}{\sum_{j=1}^{n} \frac{1}{x_j}}$$

– dane pogrupowane, cecha skokowa:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \qquad \bar{x}_h = \frac{1}{\sum_{i=1}^{k} \frac{w_i}{x_i}}$$

– dane pogrupowane, cecha ciągła:

$$\bar{x}_h = \frac{n}{\sum_{i=1}^{k} \frac{n_i}{\dot{x}_i}} \qquad \bar{x}_h = \frac{1}{\sum_{i=1}^{k} \frac{w_i}{\dot{x}_i}}$$

1.3 Średnia geometryczna

$$\bar{x}_g = \sqrt[n-1]{\frac{y(t=1)}{y(t=0)} \frac{y(t=2)}{y(t=1)} \cdots \frac{y(t=n-1)}{y(t=n-2)}}$$

$$\bar{x}_g = \sqrt[n-1]{\frac{y(t=n-1)}{y(t=0)}}$$

– Miary pozycyjne:

1.4 mediana i kwartyle

mediana – me = kwartyl drugi $k_{0,5}$

– dane indywidualne

$$me = x_{(n+1)/2}$$ dla n nieparzystych

$$me = \frac{x_{n/2} + x_{(n+2)/2}}{2}$$ dla n parzystych

dane pogrupowane

(wzory z wykorzystaniem liczebności i częstości względnych), dla mediany, kwartyla pierwszego i trzeciego.

$$me = x_{0m} + \left(\frac{n}{2} - n(x_{0m})\right)\frac{h_m}{n_m} \qquad Q_1(x) = x_{0Q_1} + \left(\frac{n}{4} - n(x_{0Q_1})\right)\frac{h_{Q_1}}{n_{Q_1}} \qquad Q_1(x) = x_{0Q_1} + \left(0{,}25 - F_n(x_{0Q_1})\right)\frac{h_{Q_1}}{w_{Q_1}}$$

$$me = x_{0m} + \left(0{,}5 - F_n(x_{0m})\right)\frac{h_m}{w_m} \qquad Q_3(x) = x_{0Q_3} + \left(\frac{3n}{4} - n(x_{0Q_3})\right)\frac{h_{Q_3}}{n_{Q_3}} \qquad Q_3(x) = x_{0Q_3} + \left(0{,}75 - F_n(x_{0Q_3})\right)\frac{h_{Q_3}}{w_{Q_3}}$$

1.5 Dominanta

(wzory z wykorzystaniem liczebności i częstości względnych)

– dane indywidualne

do - wartość cechy występująca najczęściej

– dane pogrupowane, cecha skokowa

do – wartość cechy, której odpowiada najwyższa liczebność (bądź najwyższa wartość wskaźnika struktury).

– dane pogrupowane, cecha ciągła (wzór interpolacyjny)

$$do = x_{0d} + \frac{n_d - n_{d-1}}{(n_d - n_{d-1}) + (n_d - n_{d+1})} h_d$$

$$do = x_{0d} + \frac{w_d - w_{d-1}}{(w_d - w_{d-1}) + (w_d - w_{d+1})} h_d$$

2.8.2. Miary zróżnicowania

Istotą miar zróżnicowania jest danie odpowiedzi na następujące pytanie: jak zaobserwowane wartości cechy różnią się względem przeciętnej. Podobnie jak miary położenia miary zróżnicowania (inaczej: dyspersji, zmienności, rozproszenia) dzielą się na: klasyczne i pozycyjne. Poniżej podstawowe formuły ich wyznaczania.

Miary zróżnicowania (dyspersji)

	Klasyczne	**Pozycyjne**
Absolutne	wariancja s^2 odchylenie standardowe s odchylenie przeciętne d_x typowy obszar zmienności x_{typ}	rozstęp R(x) odchylenie ćwiartkowe Q typowy obszar zmienności x_{typ}
Względne	Współczynniki zmienności V(x) $V(x) = \frac{s}{\bar{x}} c$ $V(x) = \frac{d_x}{\bar{x}} c$	$V(x) = \frac{Q}{me} c$

c-stała (np. c = {100; 1000})

2.1 Wariancja

- dane indywidualne

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 \qquad s^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \bar{x})^2$$

- dane pogrupowane, cecha skokowa

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i \qquad s^2 = \sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 w_i$$

- dane pogrupowane, cecha ciągła

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 n_i \qquad s^2 = \sum_{i=1}^{k}(\dot{x}_i - \bar{x})^2 w_i$$

- formuły uproszczone

$$s^2 = \bar{x}^2 - (\bar{x})^2 \qquad s^2 = \frac{1}{n}\sum_{i=1}^{k}(x_j)^2 - (\bar{x})^2$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(x_i^2 n_i) - (\bar{x})^2 \qquad s^2 = \sum_{i=1}^{k}(x_i^2 w_i) - (\bar{x})^2 \quad \text{i =1,...,k}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{k}(\dot{x}_i^2 n_i) - (\bar{x})^2 \qquad s^2 = \sum_{i=1}^{k}(\dot{x}_i^2 w_i) - (\bar{x})^2$$

2.2. Odchylenie standardowe

$$s = \sqrt{s^2}$$

2.3. Odchylenie przeciętne

– dane indywidualne

$$d_x = \frac{1}{n}\sum_{j=1}^{n}\left|x_j - \bar{x}\right| \qquad d_x = \frac{1}{N}\sum_{j=1}^{N}\left|x_j - \bar{x}\right|$$

– dane pogrupowane, cecha skokowa

$$d_x = \frac{1}{n}\sum_{i=1}^{k}\left|x_i - \bar{x}\right|n_i \qquad d_x = \sum_{i=1}^{k}\left|x_i - \bar{x}\right|w_i$$

– dane pogrupowane, cecha ciągła

$$d_x = \frac{1}{n}\sum_{i=1}^{n}\left|\dot{x}_i - \bar{x}\right|n_i \qquad d_x = \sum_{i=1}^{k}\left|\dot{x}_i - \bar{x}\right|w_i$$

2.4. Typowy obszar zmienności

(wyznaczony w oparciu o miary klasyczne)

$$\bar{x} - s \le x_{typ} \le \bar{x} + s$$

2.5. Rozstęp

$R(x)=x_{max} - x_{min}$

2.6. Odchylenie ćwiartkowe

$$Q = \frac{Q_3 - Q_1}{2}$$

2.7. Typowy obszar zmienności
(wyznaczony w oparciu o miary pozycyjne).

$me - Q \leq x_{typ} \leq me + Q$

2.8.3. Miary asymetrii (skośności), spłaszczenia rozkładu

3.1. Współczynnik asymetrii (miara klasyczna)

(M'_3 – moment centralny rzędu 3)

- dane indywidualne

$$A(x)=\frac{\sum_{j=1}^{n}\left(x_j-\bar{x}\right)^3}{(n-1)s^3} \qquad A(x)=\frac{\sum_{j=1}^{N}\left(x_j-\bar{x}\right)^3}{Ns^3}$$

- dane pogrupowane, cecha skokowa

$$A(x)=\frac{\sum_{i=1}^{k}\left(x_i-\bar{x}\right)^3 n_i}{(n-1)s^3} \qquad A(x)=\frac{\sum_{i=1}^{k}\left(x_j-\bar{x}\right)^3 w_i}{s^3}$$

- dane pogrupowane, cecha ciągła

$$A(x)=\frac{\sum_{i=1}^{k}\left(\dot{x}_i-\bar{x}\right)^3 n_i}{(n-1)s^3} \qquad A(x)=\frac{\sum_{i=1}^{k}\left(\dot{x}_i-\bar{x}\right)^3 w_i}{s^3}$$

3.2. Wskaźnik asymetrii (skośności) oparty na miarach pozycyjnych

$$A_2 = \frac{(Q_3 - me) - (me - Q_1)}{2Q}$$

3.3. Wskaźnik asymetrii (skośności) mieszany

$$A_1 = \frac{\bar{x} - do}{s}$$

Miary spłaszczenia (spiczastości, skupienia) rozkładu - kurtoza.

$$C_x = \frac{M_4^{/}}{s^4}$$ $M^{/}_4$ - moment centralny rzędu czwartego,

– dane indywidualne:

$$C_x = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^4}{(n-1)s^4} \qquad C_x = \frac{\sum_{j=1}^{N}(x_j - \bar{x})^4}{Ns^4}$$

– dane pogrupowane, cecha skokowa:

$$C_x = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{(n-1)s^4} \qquad C_x = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 w_i}{s^4}$$

– dane pogrupowane, cecha ciągła:

$$C_x = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 n_i}{(n-1)s^4} \qquad C_x = \frac{\sum_{i=1}^{k}(\dot{x}_i - \bar{x})^4 w_i}{s^4}$$

2.8.4. Koncentracja zjawiska i jej pomiar

Miary koncentracji (nierównomierności rozdziału sum wartości cechy)
4.1. Współczynnik koncentracji

$$K = \frac{T}{0{,}5}$$

$$K = 1 - \sum_{i=1}^{k} \left[z\left(x < x_{1i}\right) + z\left(x < x_{1i-1}\right) \right] w_i$$

gdzie:

$$w_i = \frac{n_i}{N}; \quad z_i = \frac{m_i}{M}; \quad M = \sum x_i n_i \quad M = \sum \dot{x}_i n_i; \quad m_i = x_i n_i \quad m_i = \dot{x}_i n_i$$

lub: $$K = \frac{T}{5000}$$

$\frac{z_i + z_{i-1}}{2} w_i$ - pole pojedynczej figury

$\sum_{i=1}^{k} \frac{z_i + z_{i-1}}{2} w_i$ - suma pól figur.

Pole T oblicza się ze wzoru:

$$T = 5000 - \sum_{i=1}^{k} \left(\frac{z_i + z_{i-1}}{2} \right) w_i$$

Momenty zwykłe i centralne

Moment zwykły rzędu k (k = 1,2,…) w rozkładzie empirycznym:

- dane indywidualne

$$M_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k$$

- dane pogrupowane:

$$M_k = \frac{1}{n}\sum_{i=1}^{r} x_i^k n_i \qquad M_k = \sum_{i=1}^{r} x_i^k w_i$$

Średnia arytmetyczna jest pierwszym momentem zwykłym w rozkładzie empirycznym ($M_1 = \bar{x}$)

Moment centralny rzędu k (k =1,...) w rozkładzie empirycznym

– dane indywidualne

$$M_k' = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - x)^k$$

$$M_k' = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - x)^k n_i \qquad M_k' = \frac{1}{n-1}\sum_{j=1}^{n}(\dot{x}_j - x)^k n_i$$

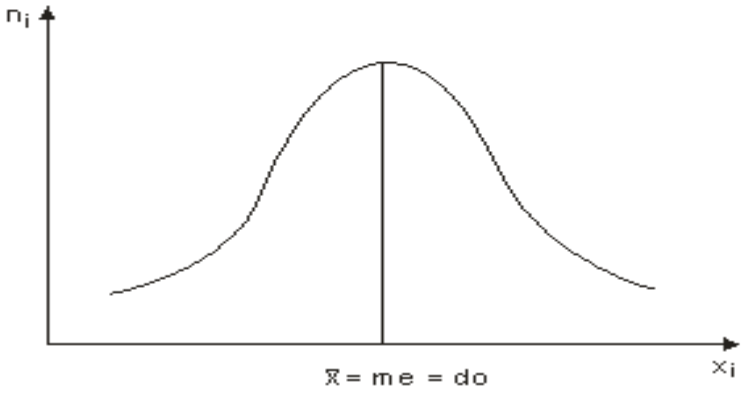
$$M_k' = \sum_{j=1}^{n}(x_j - x)^k w_i \qquad M_k' = \sum_{j=1}^{n}(x_j - x)^k w_i$$

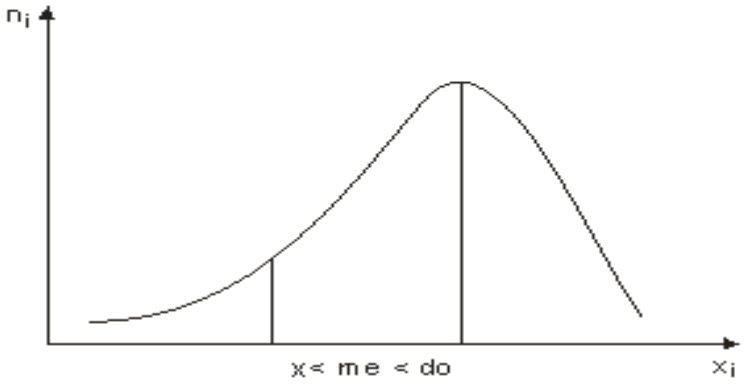
Pomiędzy wariancją s^2 a średnią arytmetyczną $\bar{x}$ zachodzi związek:

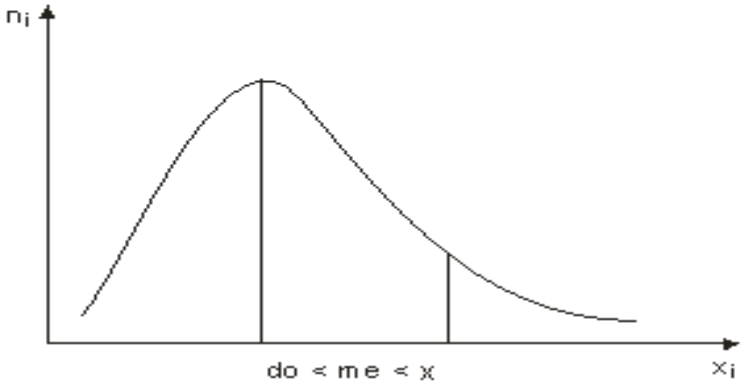
$$s^2 = M_2 - (M_1)^2 = \overline{x^2} - (\bar{x})^2$$

2.8.5. Położenie miar przeciętnego poziomu, podstawowe typy rozkładów

Położenie miar przeciętnego poziomu (średnich) w rozkładach empirycznych

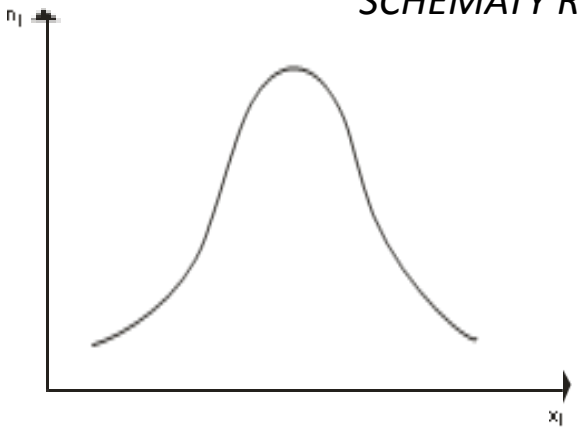






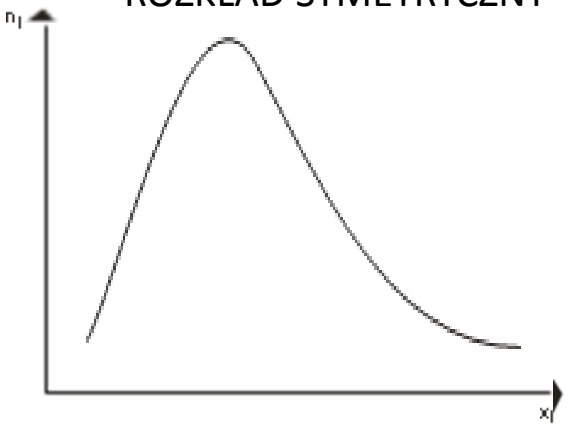
2.8.6. Przykład empiryczny

SCHEMATY ROZKŁADÓW (SZEREGÓW EMPIRYCZNYCH)

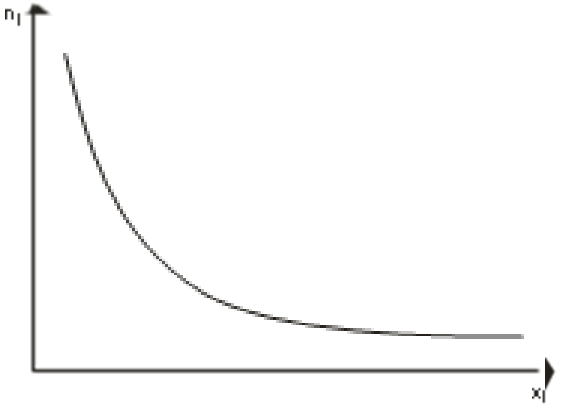


ROZKŁAD SYMETRYCZNY

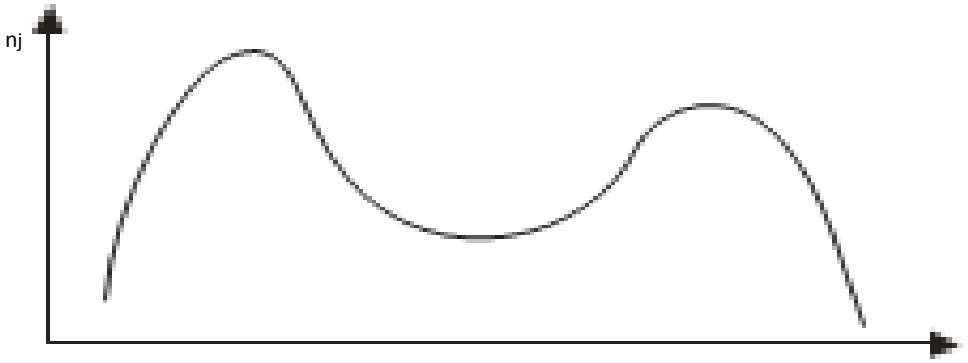
ROZKŁAD ASYMETRYCZNY



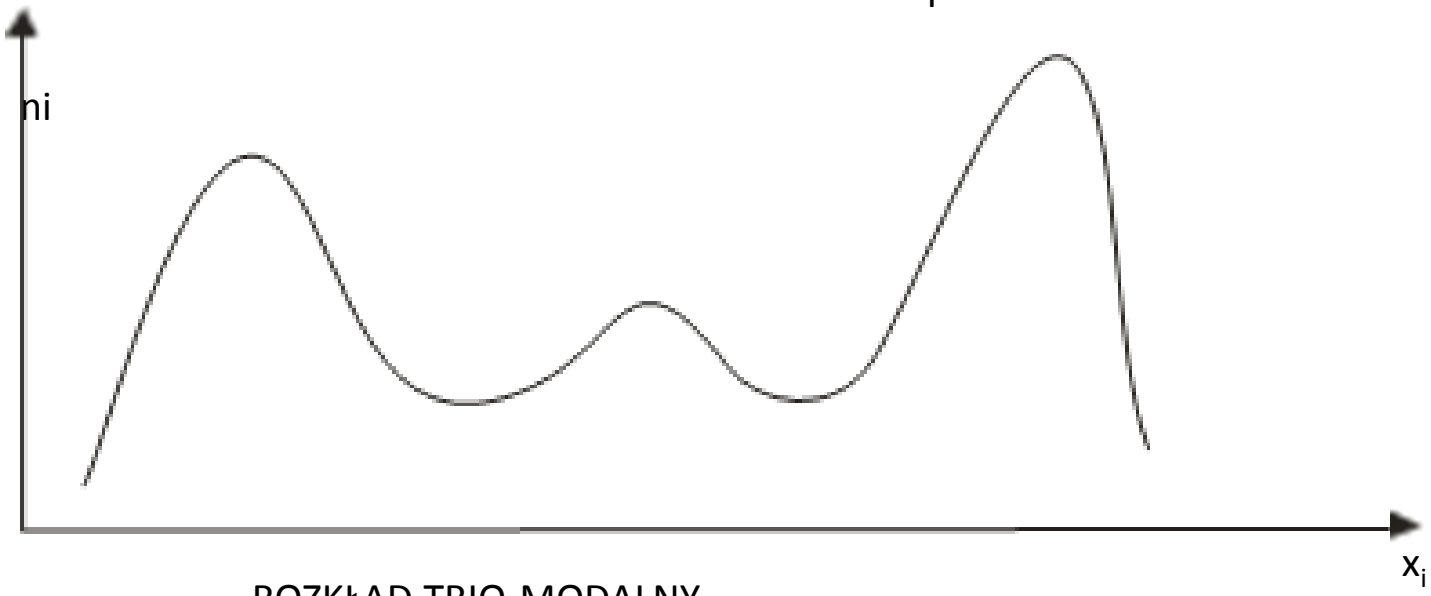
ROZKŁAD ASYMETRYCZNY



ROZKŁAD UCIĘTY



ROZKŁAD DWUMODALNY



ROZKŁAD TRIO-MODALNY

Zadanie 1. W losowej grupie klientów banku BIGG_RES zebrano informacje o wysokości kredytów świątecznych (krótkoterminowych) w tys. zł w grudniu 2014. Kwoty kredytów były następujące:

1,5; 3,5; 2; 2; 1,5; 4,5; 3; 3; 1; 8.

Określ wartości miar: przeciętnego poziomu, zróżnicowania, asymetrii, spłaszczenia i podaj ich interpretację.

Rozwiązanie:

Zadanie 2. Rozkład kredytobiorców kredytów hipotecznych (losowa próba n =454 klientów) będących w wieku 30 - 40 lat według liczby posiadanych dzieci jest następujący:

Liczba dzieci $X = x_i$	Liczba klientów n_i	w_i	$G(x_i)$	$x_i \cdot n_i$	n_{is}	$(x_i - \bar{x})^2 \cdot n_i$
0	168					
1	181					
2	83					
3	17					
4	5					
	454					

Należy:

1. Przedstawić graficznie podany rozkład.
2. Wyznaczyć częstości względne i zinterpretować wartości w1 i w3.
3. Wyznaczyć wartości dystrybuanty empirycznej $G(x_i)$ oraz zinterpretować jej wartości dla $x_i = 1$ oraz dla $x_i = 3$.
4. Narysować wykres dystrybuanty empirycznej.
5. Wyznaczyć i zinterpretować następujące miary położenia:
- średnią arytmetyczną i medianę
- kwartyl pierwszy i kwartyl trzeci.
6. Wyznaczyć i zinterpretować następujące miary zróżnicowania:
- Wariancję i odchylenie standardowe
- obszar zmienności dla typowych jednostek.
7. W oparciu o wykres rozkładu ustalić kierunek asymetrii.

Rozwiązanie

Zadanie 3. Struktura dochodów miesięcznych w tys. zł w losowo dobranych firmach rodzinnych (n = 200) posiada następujący rozkład.

Dochód w tys. złl x_{0i} - x_{1i}	[%]
15 – 25	16,5
25 – 35	23,5
35 – 45	21,0
45 – 55	19,5
55 – 65	17,0
pow. 65	2,5

Należy:

1. Opracować histogram i wielobok częstości.
2. Wyznaczyć wartości dystrybuanty empirycznej $G(x_{1i})$ i zinterpretować jej wartość dla 35 tys zł.
3. Narysować histogram i wielobok częstości skumulowanych.
4. Wyznaczyć i zinterpretować następujące miary położenia:
   - średnią arytmetyczną
   - dominantę
   - medianę
   - kwartyl pierwszy
   - kwartyl trzeci.
5. Wyznaczyć i zinterpretować następujące miary zróżnicowania:
   - wariancję
   - odchylenie standardowe
   - współczynnik zmienności
   - obszar zmienności dla typowych jednostek.
6. Ocenić skośność i spłaszczenie rozkładu.

Rozwiązanie