

ANALIZA DANYCH W AUDYCIE



Cz.II. Podstawowe metody analizy danych w audycie

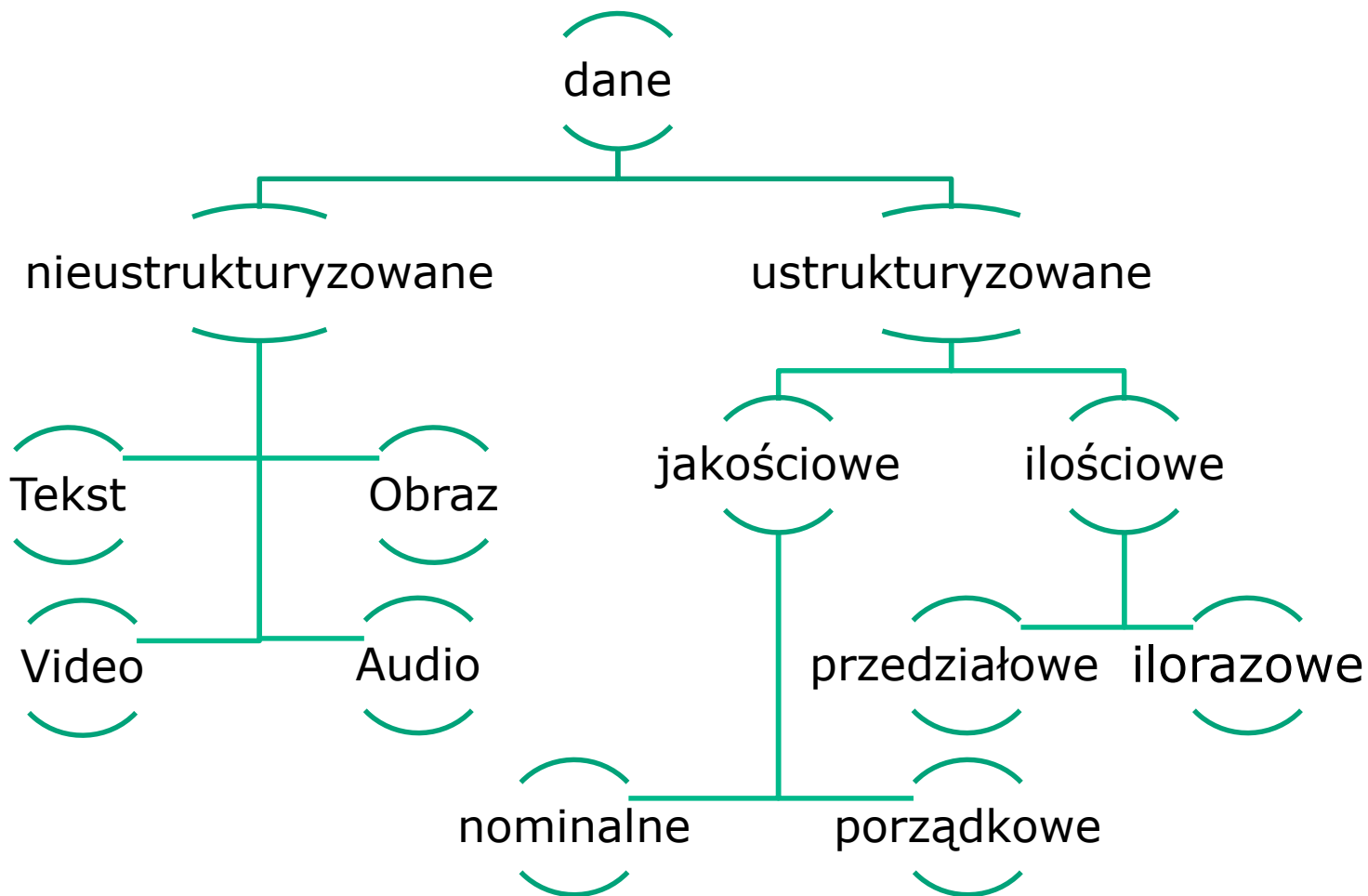
dr Wiesław KARLIŃSKI



Plan spotkania

1. Rodzaje danych
 2. Statystyka opisowa danych
 3. Elementy analizy statystycznej
-
4. Typowe audytorskie funkcje analityczne
 5. Metody analityczne stosowane w audycie
 - zgodności
 - wykonania zadań
 - finansowym
 - śledczym
 6. Podsumowanie

Klasyfikacja danych



Rodzaje danych

dane jakościowe

- **nominalne** – przyjmują określoną liczbę wartości, stąd można dokonać klasyfikacji obiektów, ale nie można ich uszeregować pod kątem nasilenia cechy (*np. kolor oczu, narodowość, płeć*)
- **porządkowe** – pozwalają na klasyfikację oraz uszeregowanie obiektów wg kategorii od najmniejszego do największego nasilenia cechy (*np. oceny, kategorie wielkości, wykształcenie*)

dane ilościowe

- **interwałowe(przedziałowe)** – pozwalają nie tylko na uszeregowanie obiektów, ale również na obliczenie wielkości różnicy między nimi (*np. temperatura w °C*)
- **ilorazowe** – pozwalają na obliczenie różnicy oraz proporcji dwóch wartości badanej cechy (*np. waga, długość, koszt*)

Jak opisać dane jednym parametrem?

Poprzez podanie wartości typowej dla zbioru, tj.:

- Średniej
 - Dominanty (wartości modalnej)
 - Mediany
 - Kwartyli
-

Najczęściej posługujemy się **średnią**, ale:

- Nie jest to dobra miara przy dużym rozrzucie danych lub danych odstających
- Może być obliczona tylko dla zmiennych ilościowych
- Nie zawsze stosuje się prostą średnią arytmetyczną

Rodzaje średnich

- **Średnia arytmetyczna** – dla typowych danych jednostkowych (średnia gm. mazowiecka=17,3 tys mieszk.)
- **Średnia arytmetyczna ważona** – jeśli dane przedziałowe
- **Średnia geometryczna** – dane w postaci szeregu czasowego i oceniamy np. średnie tempo wzrostu
- **Średnia harmoniczna** – jeśli cecha wyrażona jest w jednostkach względnych (np. km/h)
- **Średnia obcięta (Excel – wewnętrzna)** – pomija określony odsetek wartości skrajnych (średnia gm. obcięta o 5% = 10 tys mieszk.), zastosowanie praktyczne np. WIBOR

Średnie – dwa przykłady

1. Dane przedziałowe – średnia ważona

Wzrost (cm)	n_i	m_i	$m_i \times n_i$
150 – 160	10	155	1550
160 – 170	25	165	4125
170 – 180	30	175	5250
180 – 190	10	185	1850
190 – 200	1	195	195
Suma	76		12970
Średnia			170,7

2. Rowerzysta podzielił trasę na 2 równe odcinki (*po 20 km*). Pierwszy odcinek przejechał z prędkością 20 km/h, a drugi z 10 km/h.

Jaka była jego średnia prędkość?

Średnie – dwa przykłady

1. Dane przedziałowe – średnia ważona

Wzrost (cm)	ni	mi	mi x ni
150 – 160	10	155	1550
160 – 170	25	165	4125

Średnia:

zwykła $(10+20)/2 = 15$ km/h?

-> *czas = 2h 40 min*

harmoniczna = 13,3 km/h

-> *czas = 40/13,3 = 3h*

Średnia			170,7
---------	--	--	-------

2. Rowerzysta podzielił trasę na 2 równe odcinki (*po 20 km*). Pierwszy odcinek przejechał z prędkością 20 km/h, a drugi z 10 km/h.

Jaka była jego średnia prędkość?

Dominanta (modalna)

Dominanta to wartość, która występuje w zbiorze najczęściej. Często w miejsce liczby podaje się odsetek wystąpień wartości.

Stosowalność i przydatność:

- dla zm. nominalnej to jedyna pojedyncza miara rozkładu (*gminy w typy: m-35, w-226, mw-53*)
- dla zm. porządkowej, może być stosowana w zależności od uwarunkowań (*przykład dalej*)
- dla zm. ilościowej ma sens w przypadku danych pogrupowanych w klasy.

Ograniczenia stosowania:

jeśli w zbiorze występują wartości o zbliżonej częstości, wtedy lepiej pokazać wszystkie częstości niż modalną

Miary pozycyjne - mediana i kwartyle

Mediana (wartość środkowa) – wartość, której nie przekracza (co najmniej) połowa elementów danej zbiorowości. Mediana dzieli uporządkowany narastająco zbiór obserwacji na połowy.

- Nie jest wrażliwa na obserwacje odstające
- Ważna przy zmiennej porządkowej (*przykład dalej*)
- Dla zmiennej ilościowej lepsza od średniej przy danych odstających
- Nie może być stosowana dla zmiennej nominalnej

Kwartył dolny (Q1) i górny (Q3) to wartość, poniżej której leży odpowiednio 25% i 75% obserwacji w uporządkowanym szeregu danych dotyczących badanej zmiennej

Dane porządkowe - modalna czy mediana?

Ocena	częstość	%
1 bardzo źle	15	4,8
2 raczej źle	60	19,1
3 przeciętnie	98	31,2
4 raczej dobrze	101	32,2
5 bardzo dobrze	40	12,7
ogółem gmin	314	100,0

modalna

Dane porządkowe – modalna czy mediana?

Ocena	częstość	%	% narastająco
1 bardzo źle	15	4,8	4,8
2 raczej źle	60	19,1	23,9
3 przeciętnie	98	31,2	55,1
4 raczej dobrze	101	32,2	87,3
5 bardzo dobrze	40	12,7	100,0
ogółem gmin	314	100,0	

mediana

modalna

Miary rozrzutu (rozproszenia danych)

Rozstęp – różnica pomiędzy wartością max i min (max-min)

Odchylenie przeciętne – średnie odchylenie poszczególnych wartości od średniej, proste w interpretacji stąd stosowane w audycie (MAD lub AAD), raczej nie stosowane w statystyce

Odchylenie standardowe – miara rozrzutu powszechnie stosowana w statystyce, ale nieco trudna w interpretacji

Odstęp międzykwartylowy – różnica pomiędzy kwartylem górnym i dolnym ($Q3-Q1$), stosowana m.in. przy identyfikacji danych odstających

Współczynnik zmienności: $V = \text{odch. stand.} / \text{średnia} * 100\%$

Miary pozycyjne i rozproszenia – podsumowanie

Podsumowanie

- średnia jest bardzo czuła na wartości odstające,
- w statystyce nie podaje się średniej bez odchylenia standardowego,
- jeśli $V > 100\%$, to posługiwanie się średnią jest ryzykowne,
- jeśli średnia i mediana znacznie się różnią, to znaczy, że należy zastosować medianę,
- mediana wchodzi coraz bardziej do potocznego obiegu (również prawnego), tym niemniej stosując ją warto to zapisać czytelnie,
- posługiwanie się rozstępem daje czasem wyniki równie spektakularne co nieuprawnione.

Mazowsze: średnia= 17,3 tys.; $V=589\%$; mediana = 6,6 tys.

Możliwe operacje na danych – podsumowanie

Operacje/miary	Skala			
	nominalna	porządkowa	interwałowa	ilorazowa
Zliczanie (częstość)	✓	✓	✓	✓
uporządkowanie		✓	✓	✓
modalna	✓	✓	✓	✓
mediana		✓	✓	✓
średnia	✓*)		✓	✓
dodawanie/odejmowanie			✓	✓
mnożenie/dzielenie				✓
wartość zerowa				✓

*) wyłącznie dla zmiennej binarnej (0/1)

Przykład

Mamy ma dokonać oceny oddziałów firmy pod kątem warunków pracy tj. zagęszczenia pracowników (zalecana wartość: $5 \text{ m}^2 \pm 10\%$).

Dane o zagęszczeniu przedstawiono poniżej:

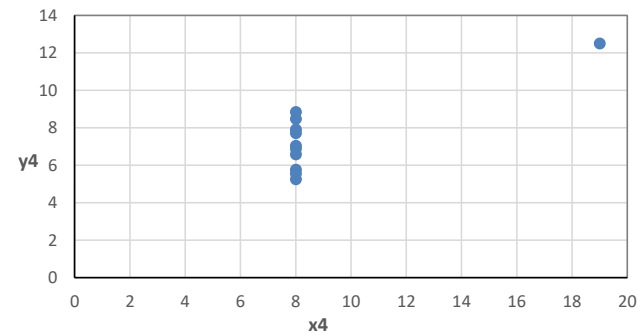
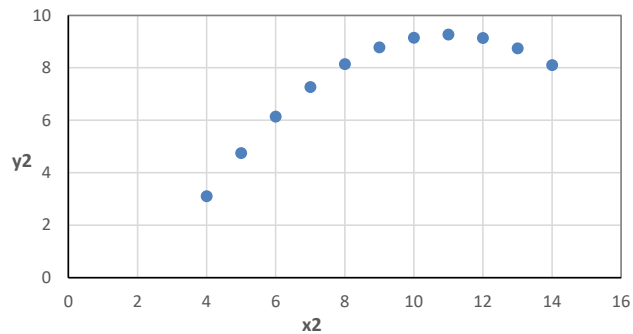
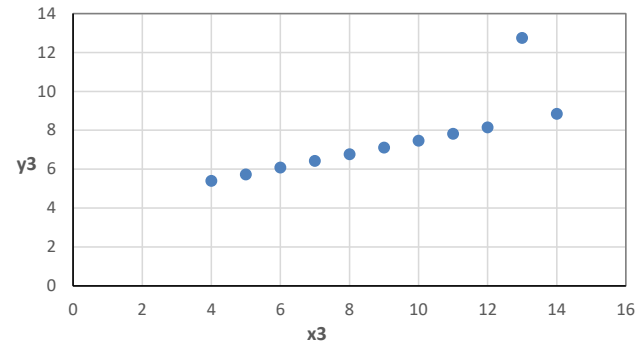
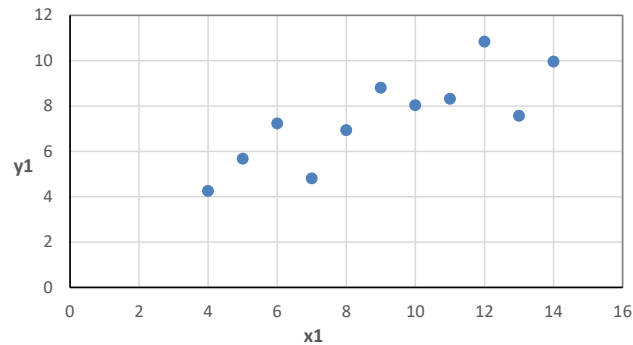
oddz.1	oddz.2	oddz.3	oddz.4	oddz.5	oddz.6	oddz.7	oddz.8	oddz.9	oddz.10
3,5	4,5	4,8	4,9	4,9	5,0	5,1	5,2	5,3	20,5

Która z poniższych konkluzji będzie najbardziej adekwatna?

- a) zagęszczenie pracowników jest zgodne z zalecaną wartością, z wyłączeniem oddz.1, gdzie wynosi 3,5 m.kw. na osobę,
- b) warunki lokalowe w firmie są bardzo dobre, gdyż średnio na pracownika przypada 6,4 m.kw. pow., przy wartości zalecanej 5 m.kw.,
- c) występuje bardzo duża dysproporcja w zagęszczeniu pracowników, gdyż jego poziom (max/min) różni się prawie 6-krotnie,
- d) zagęszczenie pracowników jest zgodne z zalecaną wartością, z wyłączeniem oddz.10, gdzie wynosi 20 m.kw. na osobę.

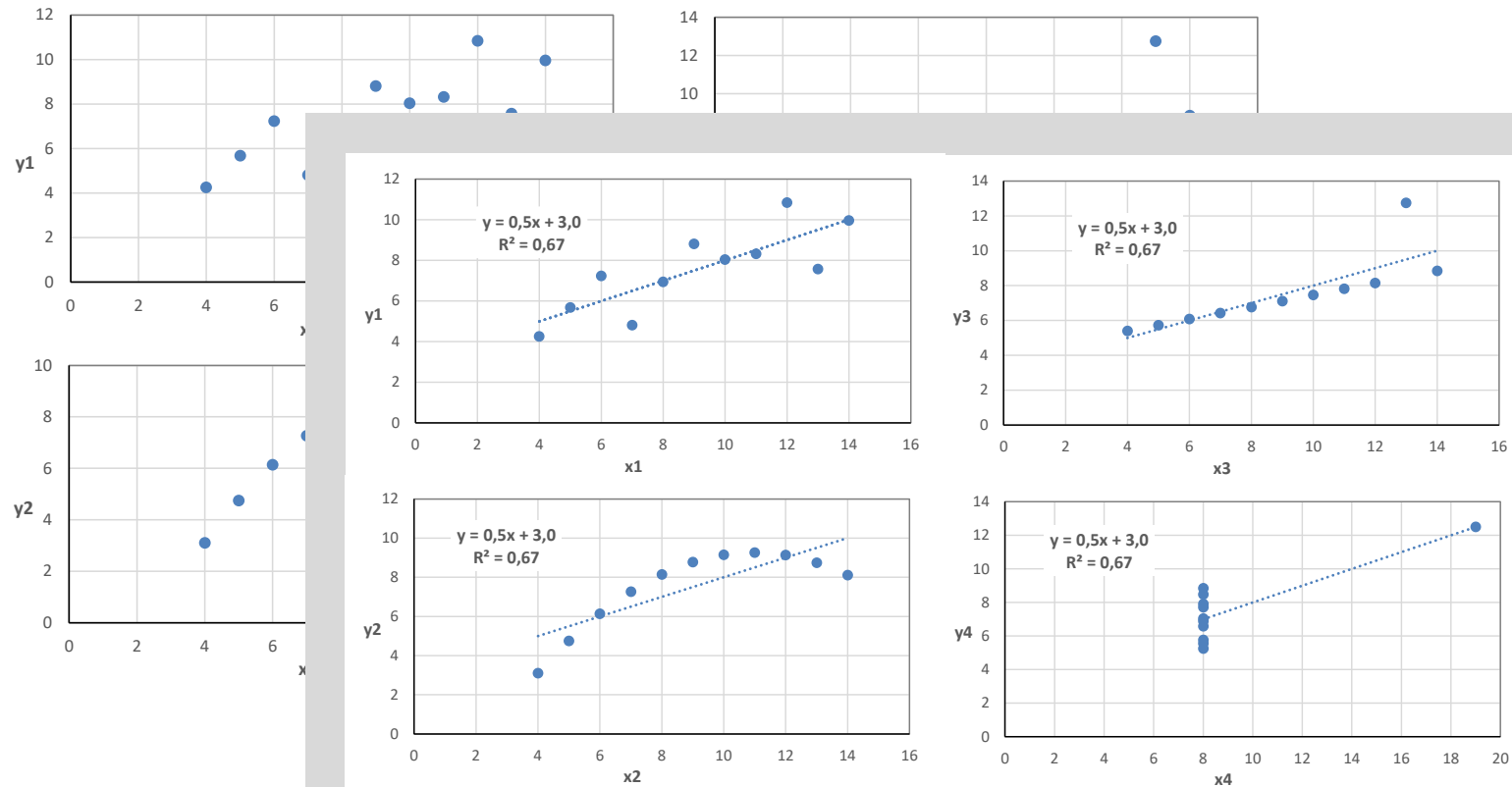
Czy zawsze wystarczą miary statystyczne?

Przykład: tzw. kwartet Anscombe'a



Czy zawsze wystarczą miary statystyczne?

Przykład: tzw. kwartet Anscombe'a



Średnie: $x=9,0$; $y=7,50$

Odch.stand: $x=3,16$; $y=1,94$

Dane odstające

Dane odstające to dane, które na tyle różnią się od reszty danych w zbiorze, że istnieje przypuszczenie, iż zostały wygenerowane przez inny mechanizm niż pozostałe dane.

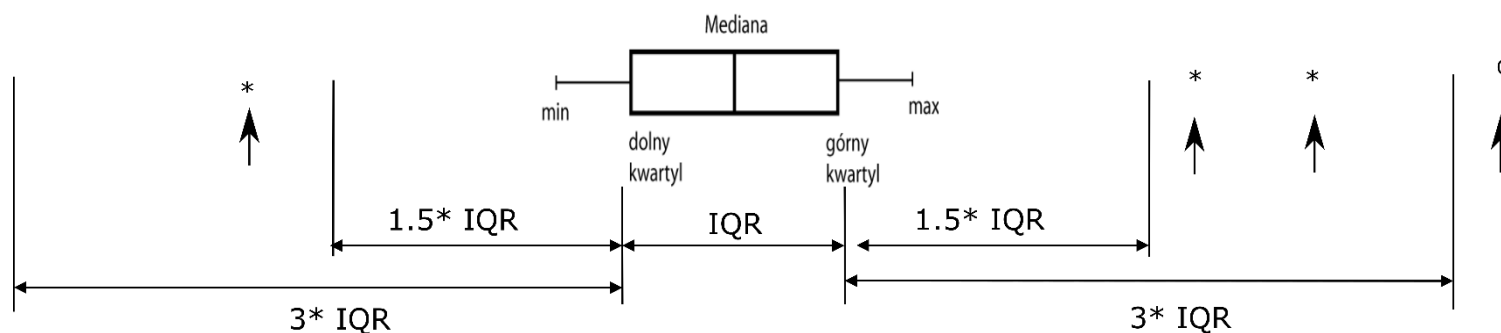
Jest to szczególnie ważne:

- w próbie, na podstawie której chcemy dokonywać ekstrapolacji,
- przy opisie zależności wielu zmiennych i budowie modeli analitycznych,
- przy korzystaniu ze średniej i rozstępu.

W audycie, w odróżnieniu od klasycznej analizy danych, występowanie danych odstających może mieć również pozytywne skutki, tzn. może być źródłem interesujących ustaleń. Dlatego musimy umieć je zidentyfikować.

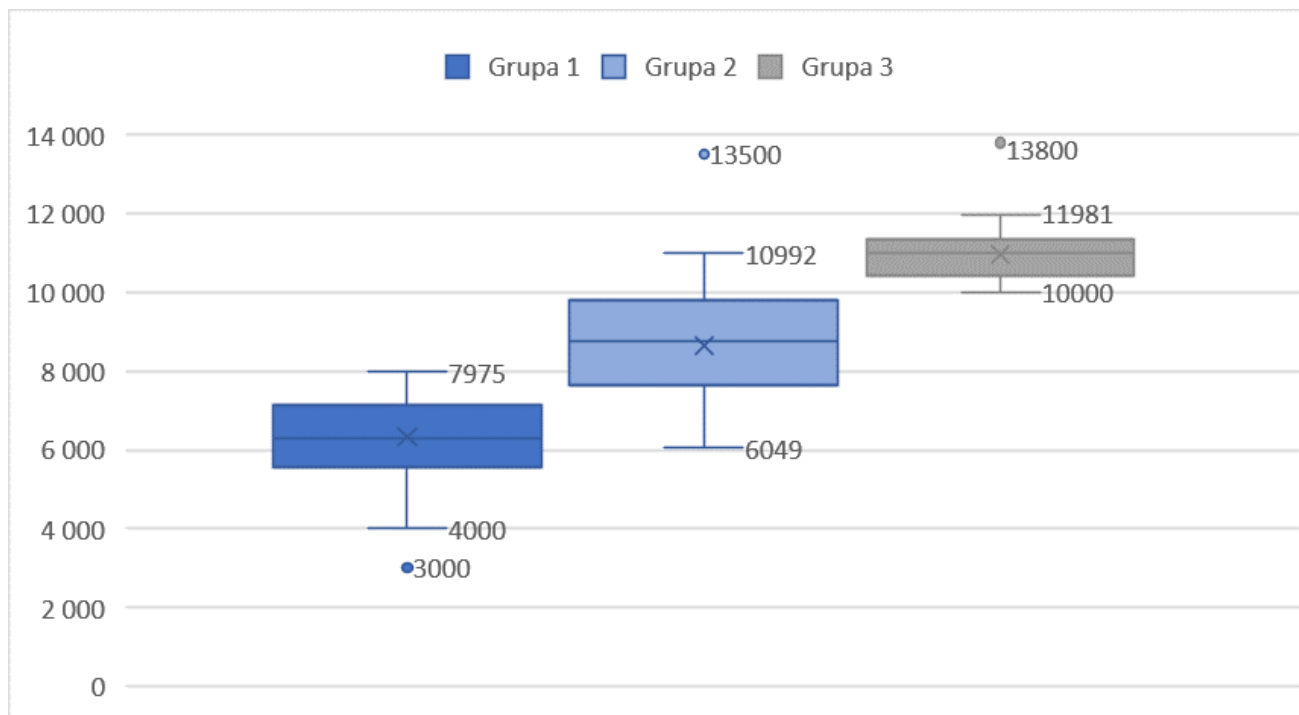
Metody identyfikacji danych odstających – 1 zmienna

- **Uproszczona:** 5-10 najmniejszych i 5-10 największych wartości (*funkcje min.k i max.k w Excelu*)
- **Parametryczna:** różnica pomiędzy wartością danej obserwacji i wartością średnią jest większa niż 2 lub 3 odchylenia standardowe
- **Nieparametryczna:** różnica pomiędzy wartością danej obserwacji i górnym/dolnym kwartylem jest większa niż:
 - 1,5 odstępu międzykwartylowego (przypadki odstające)
 - 3 odstępów międzykwartylowych (przypadki ekstremalne)

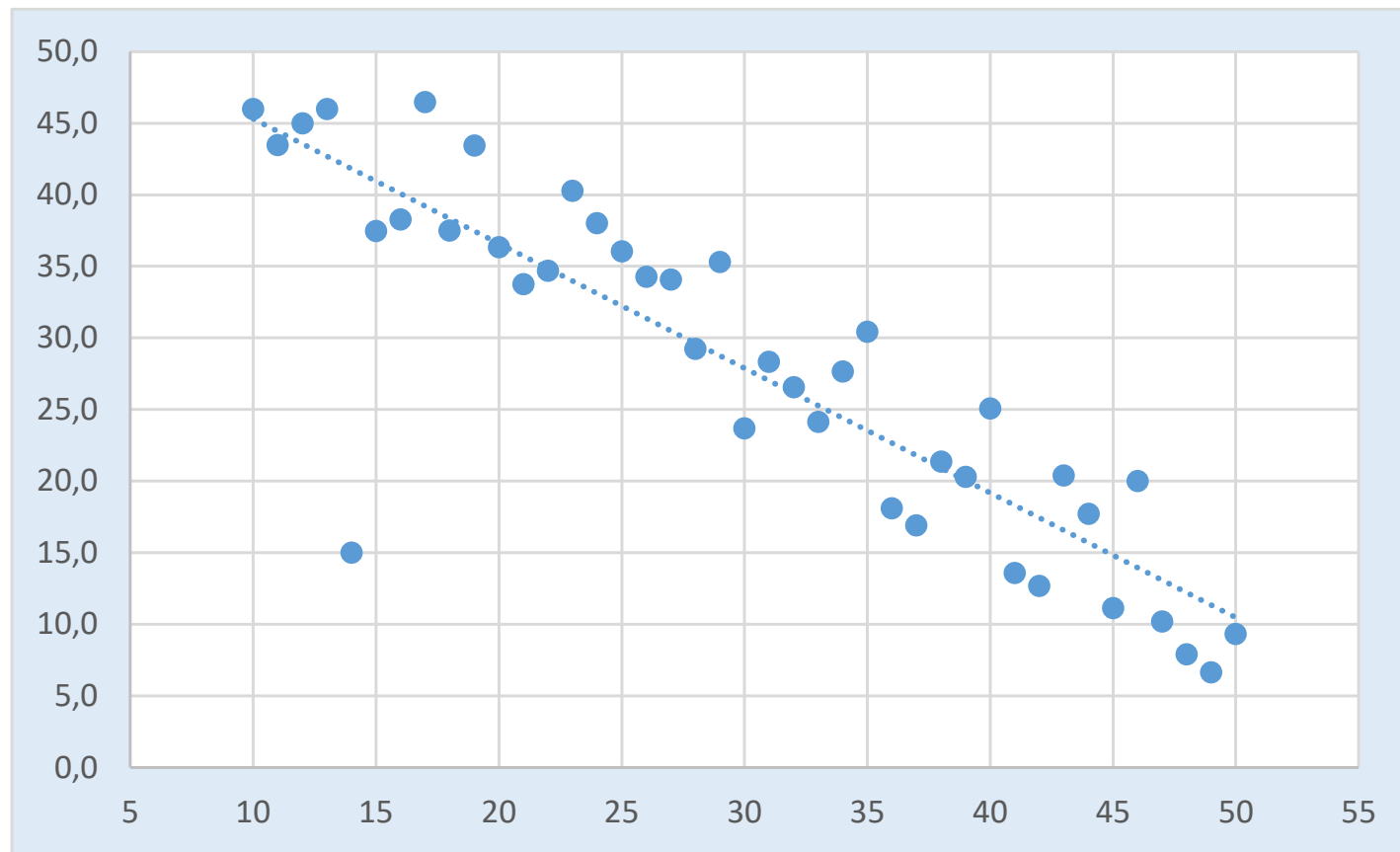


Wykres pudełkowy (skrzynka i wąsy) – przykład

Wynagrodzenie netto w trzech grupach pracowniczych



Dane odstające (2 zmienne) – ident. graficzna



Dane nietypowe (2 zmienne) – ident. analityczna

- **odstające:** istotnie różnią się od modelu opisującego zależność Y od X , choć nie muszą mieć dużego wpływu na trend (równanie regresji)

badanie tzw. studentyzowanych reszt

- **wpływowe:** mają duży wpływ na trend (usunięcie pojedynczej obserwacji zmieni wartość trendu)

odległość Cook'a i met. DFFITS

- **wysokiej dźwigni:** typowa wartość Y , a odstająca wartość X (działa jak dźwignia)

wyznaczanie wartości dźwigni

Dane nietypowe (2 zmienne) – ident. analityczna

- **odstające:** istotnie różnią się od modelu opisującego zależność Y od X, choć nie muszą mieć dużego wpływu na trend (równanie regresji)

badanie t

- **wpływowe:**

pojedynczej obs

odległość

- **wysokiej dź**

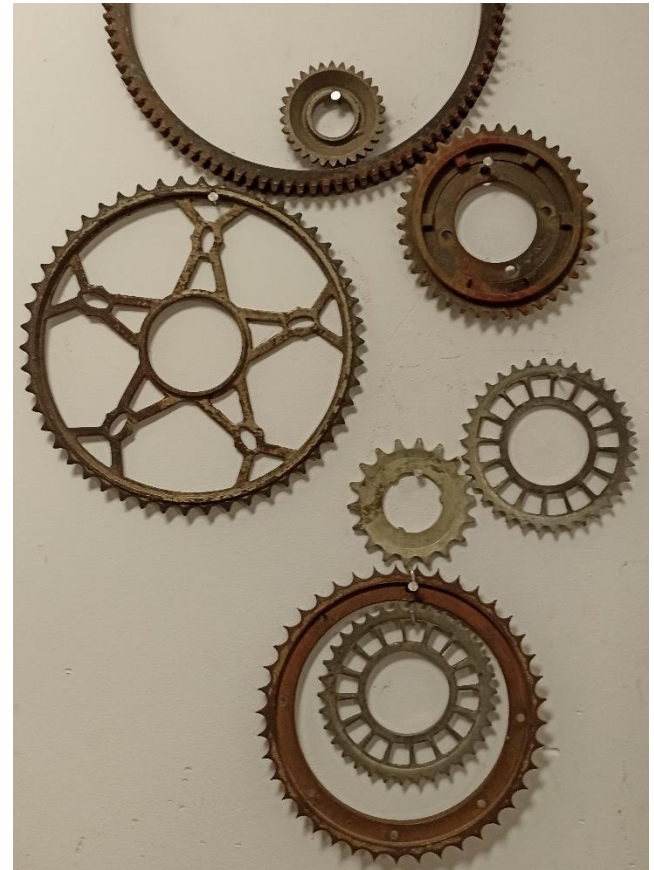
wartość X (dział

wyznaczan

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	X	Y	Pr	Re	Dźwignia	M	Rs	T-test	odl.Cooka	DFFITS	O	W	D
2	10	46,0	#	#	0,09	#	#	0,898	0,00	0,04	1	1	
3	11	43,5	#	##	0,09	#	##	0,867	0,00	- 0,05			
4	12	45,0	#	#	0,08	#	#	0,797	0,00	0,08			
5	13	46,0	#	#	0,07	#	#	0,557	0,01	0,17			
6	14	15,0	#	##	0,07	#	##	0,000	0,85	- 2,01			
7	15	37,5	#	##	0,06	#	##	0,541	0,01	- 0,16			
8	16	38,3	#	##	0,06	#	##	0,754	0,00	- 0,08			
9	17	46,5	#	#	0,05	#	#	0,200	0,05	0,31			
10	18	37,5	#	##	0,05	#	##	0,886	0,00	- 0,03			
11	19	43,4	#	#	0,05	#	#	0,295	0,03	0,23			
12	20	36,4	#	##	0,04	#	##	0,969	0,00	- 0,01			
13	21	33,8	#	##	0,04	#	##	0,735	0,00	- 0,07			
14	22	34,7	#	##	0,04	#	##	0,980	0,00	- 0,00			
15	23	40,3	#	#	0,03	#	#	0,274	0,02	0,20			
16	24	38,0	#	#	0,03	#	#	0,395	0,01	0,15			

Analiza współzależności

- czy istnieje zależność między cechami (zmiennymi) X i Y ?
- jak silna jest ta zależność?
- jaki jest kierunek zależności?
- czy można napisać wzór tej zależności?
- czy na podstawie wzoru można przewidzieć wartość cechy Y ?



Odpowiedź na powyższe pytania daje analiza korelacji i regresji

Pomiar korelacji

Do pomiaru korelacji służą współczynniki: korelacji liniowej Pearsona (r), rang Spearmana, stosunek korelacyjny, V-Cramera

r Pearsona: zakres od -1 do 1 (znak = kierunek zależności)

poziom zależności: mały ($<0,4$), umiarkowany ($0,4-0,7$),
znaczący ($0,7-0,9$), bardzo silny ($>0,9$)

Uwagi:

- Najczęściej stosowany
- Nie jest najlepszą miarą w sytuacji gdy zależność jest nieliniowa
- Wrażliwy na pojedyncze wartości odstające

Korelacja nie jest dowodem na występowanie związku przyczynowo-skutkowego.

Analiza regresji

Analiza służy do opisu zależności pomiędzy zmiennymi x i y w postaci funkcji (w Excelu – linii trendu) oraz do prognozowania y na podstawie x .

Najprostszy model to zależność liniowa: **$y = ax + b$**

W rzeczywistości mamy często do czynienia z zależnością nieliniową np. logarytmiczną (*np. krzywa produkcji, prawo Webera-Fechnera*).

W celu wyznaczenia funkcji regresji (linii trendu) można:

- posłużyć się metodą graficzną – wybieramy taki rodzaj trendu, aby uzyskać największy współczynnik determinacji R-kwadrat
- zastosować metody analityczne (spory zestaw funkcji w Excelu)

Analiza regresji

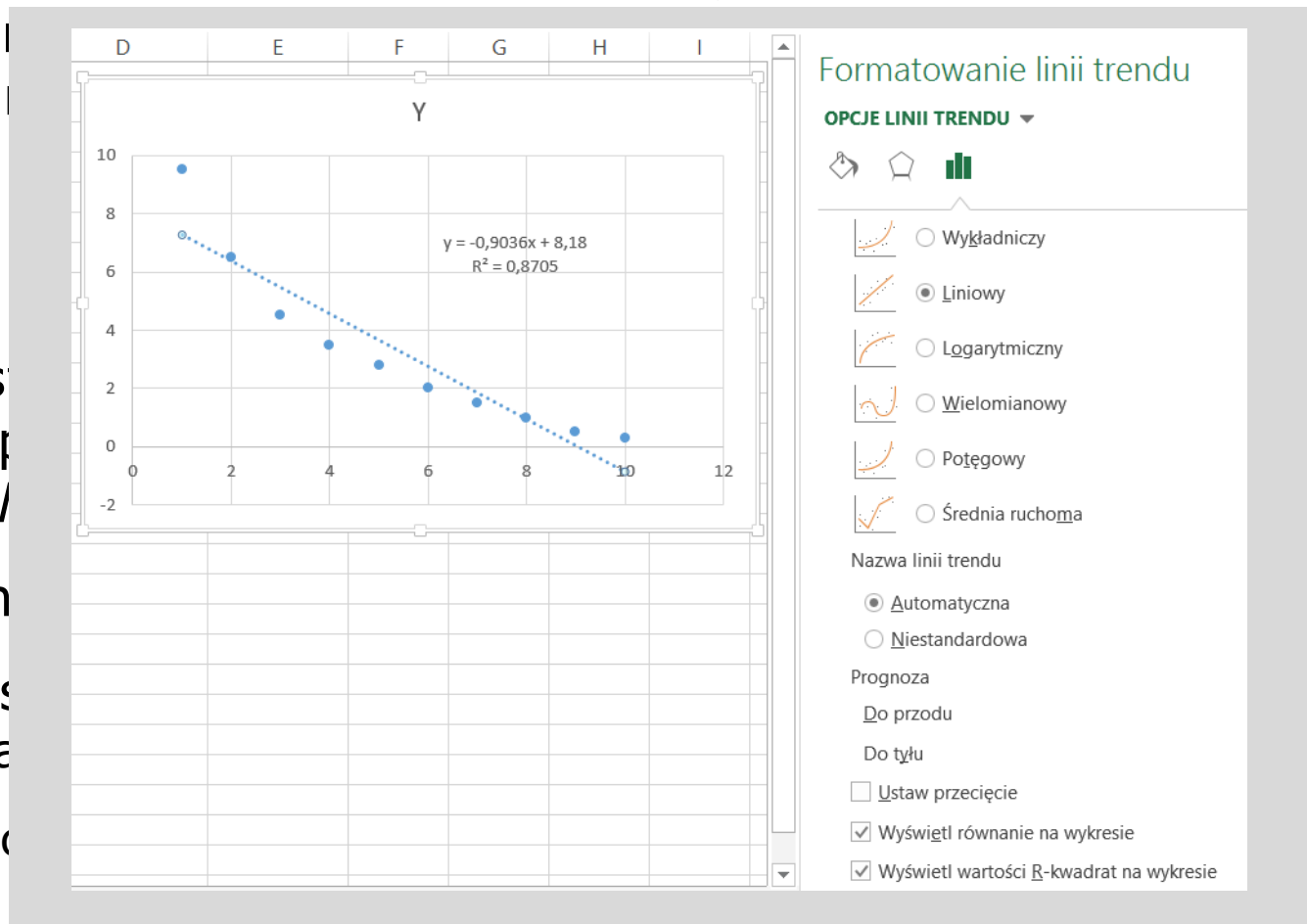
Analiza służy do opisu zależności pomiędzy zmiennymi x i y w postaci funkcji, która może być użyta do prognozowania.

Najprostszym

W rzeczywistości jest to zależność nieliniowa np. *Webera-Fechnera*.

W celu wyznaczenia

- posłużyć się
- aby uzyskać
- zastosować

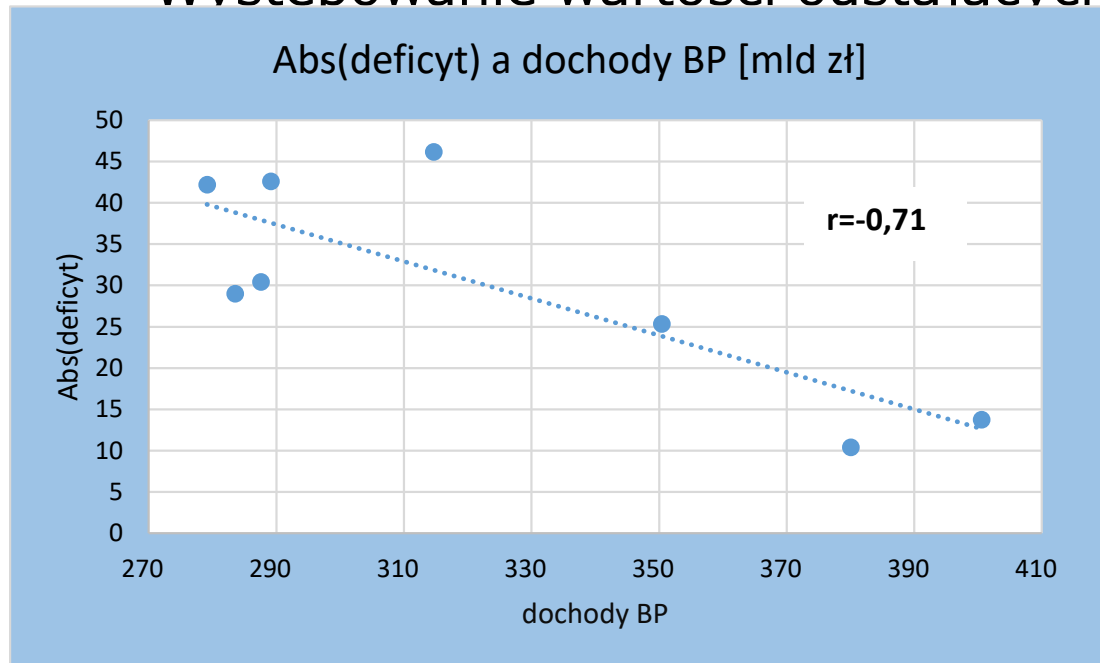


Badanie współzależności – problemy

- Występowanie wartości odstających
- Występowanie rozłącznych podgrup danych (skupisk na wykresie) – przy analizie łącznej możemy uzyskać relację odwrotną niż dla każdej z podgrup (paradoks Simpsona)
- Stosowanie analizy dla danych narastających

Badanie współzależności – problemy

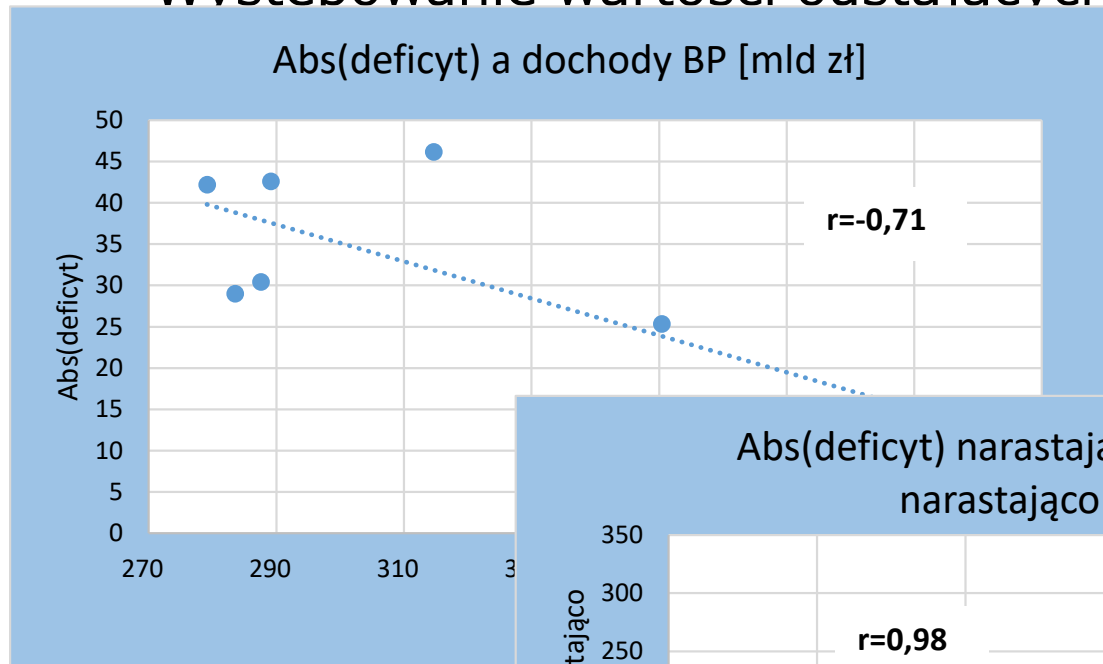
- Występowanie wartości odstających



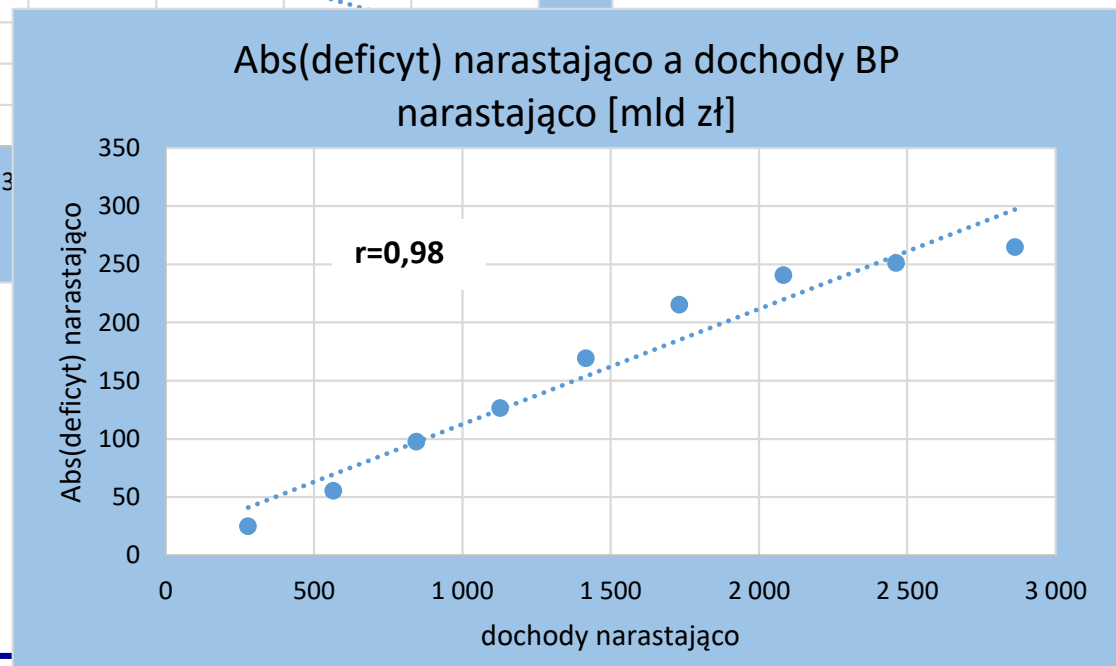
danych (skupisk na
żemy uzyskać
dgrup (paradoks
stających

Badanie współzależności – problemy

- Występowanie wartości odstających



danych (skupisk na
możemy uzyskać
podgrup (paradoks
stających



Typowe audytorskie analizy danych

1. Wyznaczenie profilu danych
2. Identyfikacja nieciągłości (luk) w numeracji danych
3. Identyfikacja powtórzeń (duplikatów)
4. Kategoryzacja danych
 - jednowymiarowa
 - dwuwymiarowa (macierz, tabela krzyżowa)
5. Stratyfikacja danych
6. Wiekowanie

Profil danych

Zestaw miar i wskaźników pozwalający na dokonanie ogólnego oglądu danych pod kątem rozmiaru, typów, zakresów, poprawności zapisu oraz podstawowych miar rozkładu.

1. Podstawowe dane informacyjne

liczba zapisów wg typu: liczbowe (lub daty), tekstowe, puste

2. Weryfikacja zakresu i poprawności danych liczbowych

wartość min i max, liczba zapisów dodatnich, ujemnych, zerowych, niskiej wartości, suma zapisów ogółem i niskiej wartości

3. Parametryczne miary rozkładu danych liczbowych

średnia, odchylenie standardowe, wsp. zmienności

4. Pozycyjne miary rozkładu danych liczbowych

mediana, kwartyle, liczba zapisów odstających

Profil danych

Zestaw miar i wskaźników po
oglądu danych pod kątem roz
poprawności zapisu oraz pods

1. Podstawowe dane inform

liczba zapisów wg typu: lic

2.Weryfikacja zakresu i pop

wartość min i max, lic
zerowych, niskiej wartoś
wartości

3.Parametryczne miary rozk

średnia, odchylenie standa

4.Pozycyjne miary rozkładu

mediana, kwartyle, liczba z

	A	B	C	D	E
1	L. wierszy (z nagłówkiem):	785			
2	nazwa zmiennej	Wartość	Data	Nr_trans	
3	dane liczbowe lub daty	784	783	0	
4	puste zapisy	0	1	1	
5	dane tekstowe	0	0	783	
6	liczba dodatnich	784	783		
7	liczba ujemnych	0	0		
8	liczba zerowych	0	0		
9	liczba niskiej wartości	1	0		
10	suma zapisów	629 701 456,3			
11	suma niskiej wartości	8,00	-		
12	suma ujemnych	-	-		
13	wartość min	8,0	20.04.2020		
14	wartość min-1	102,6	20.04.2020		
15	wartość min-2	115,8	20.04.2020		
16	wartość max-2	19 553 488,9	21.11.2020		
17	wartość max-1	27 428 524,2	21.11.2020		
18	wartość max	37 177 514,7	21.11.2020		
19	średnia	803 190,6			
20	odch.standardowe	2 373 141,8			
21	wsp. zmienności	295%			
22	mediana	154 015,4	18.07.2020		
23	kwartyl 1	13 223,9	30.05.2020		
24	kwartyl 3	599 545,2	16.09.2020		
25	I.odstających - małych	0	0		
26	I.odstających - dużych	92	0		
27					

Luki i duplikaty

- Przed analizą należy upewnić się jaki jest sposób numeracji zapisów (np. w numeracji miesięcznej należy uwzględnić dwie cechy: miesiąc i numer).
- Jeśli identyfikator ma charakter alfanumeryczny trzeba rozdzielić część tekstową i numeryczną i potraktować jako dwie cechy identyfikacyjne.
- W przypadku analizy luk należy zwymiarować rozmiar luki (liczbę brakujących pozycji).
- Realizacja w Excelu wymaga sortowania danych wg cech identyfikacyjnych. Duplikaty można w Excelu zidentyfikować również poprzez formatowanie warunkowe.

Kategoryzacja

- To klasyfikacja danych na podstawie cechy jakościowej.
- Pokazuje wykaz i liczebność kategorii, a często również sumę jakiejś cechy ilościowej w ramach każdej kategorii.
- Kategoryzacja wg dwóch cech równocześnie ma najczęściej formę macierzy (kategorie jako nagłówki wierszy i kolumn macierzy, a wewnątrz liczebności).

W Excelu:

- 1 wymiar: identyfikacja kategorii – UNIKATOWE, zliczanie i sumowanie – funkcje warunkowe
- 2 wymiary: najlepiej skorzystać z tabeli przestawnej

Kategoryzacja

- To klasyfikacja
- Pokazuje wykaz jakiejś cechy il
- Kategoryzacja formę macierz macierzy, a we

W Excelu:

- 1 wymiar: ide sumowanie – f
- 2 wymiary: naj

typ	liczba	ludność
m	35	3 016 824
w	226	1 502 826
mw	53	903 518
ogółem	314	5 423 168

powiat	typ gminy			suma
	m	w	mw	
1401		4	2	6
1402	1	7	1	9
1403	2	10	2	14
1404	1	3	1	5
1405	2	3	1	6
1406		6	4	10
1407		6	1	7
1408	1	3	1	5
1409		5	1	6
1410		5	1	6
1411	1	8	1	10
1412	2	8	3	13
1413	1	9		10
1414	1	3	2	6
1415		10	1	11
1416	1	9	1	11
1417	2	5	1	8
1418		2	4	6
1419		12	3	15

Stratyfikacja

To klasyfikacja danych na podstawie cechy ilościowej, którą dzieli się na przedziały klasowe wartości.

Dla każdego z przedziałów zlicza się liczbę (ew. odsetek) wystąpień oraz sumę wartości.

Dobre rady:

- Na wstępie należy podjąć decyzję co do sposobu definicji przedziałów: **prawostronnie** albo lewostronnie zamknięte.
- Finalną definicję przedziałów warto poprzedzić analizą wstępną (automatyczny podział zakresów na 10-20 przedziałów).
- Wyniki można zobrazować graficznie w postaci histogramu.
- Jeśli celem ma być prezentacja rozkładu zmiennej, to należy przyjąć równe wartości przedziałów klasowych.

Stratyfikacja

	A	B	C	D	E	F	G	H
1	klas	10	GG_klas_1	liczba		GG_klas2	liczba	suma
2	liczebność	314	180 547	312		2 000	2	3 388
3	suma	5 423 168	359 448	1		5 000	96	374 009
4	min	1 646	538 350	0		10 000	117	808 784
5	max	1 790 658	717 251	0		20 000	61	861 395
6	max-min	1 789 012	896 152	0		50 000	29	870 585
7	mediana	6 632	1 075 053	0		100 000	6	383 553
8	szer.klasy	178 901	1 253 954	0		200 000	1	119 425
9			1 432 856	0		500 000	1	211 371
10			1 611 757	0		powyżej	1	1 790 658
11			1 790 658	1			314	5 423 168
12			powyżej	0				
13				314				
14								
15								

lowej, którą

v. odsetek)

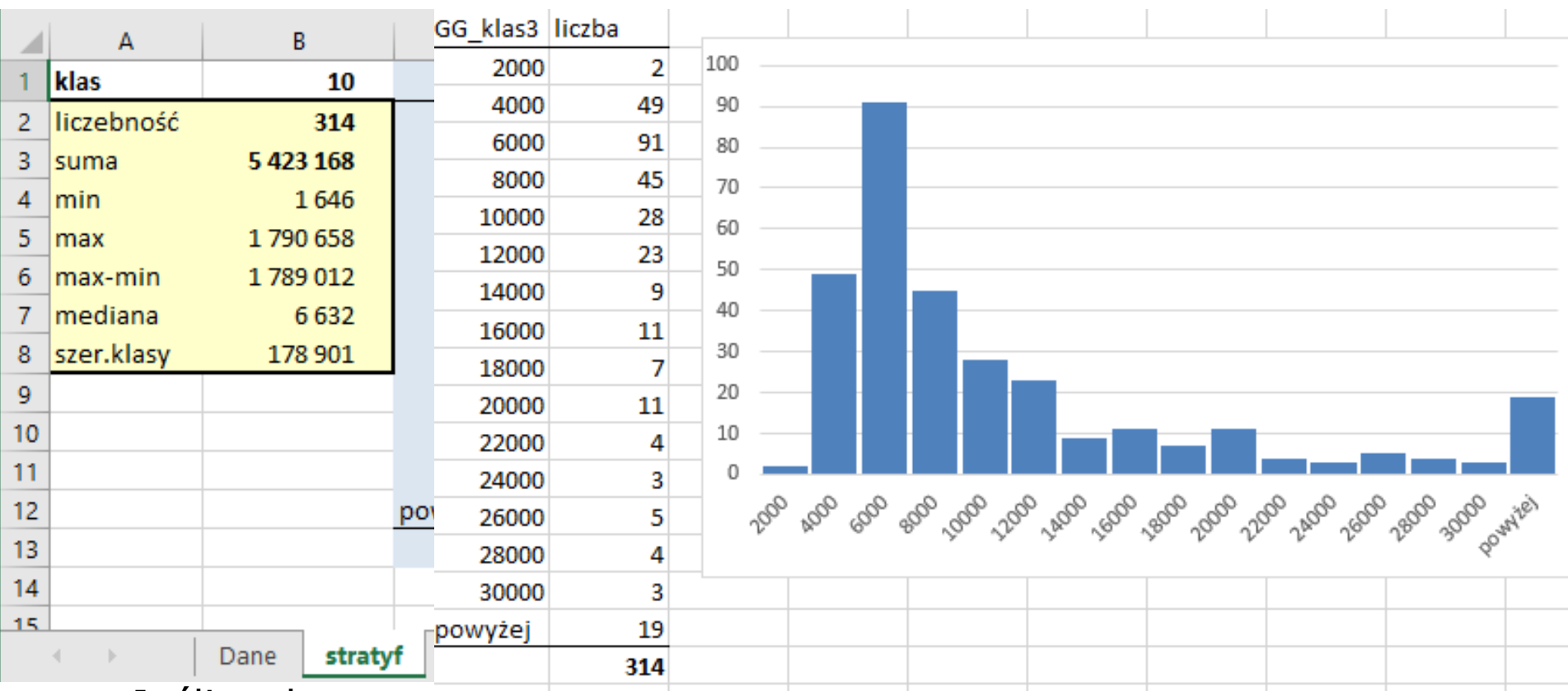
obu definicji
knięte

lizą wstępną
w)

ramu

- Jeśli celem ma być prezentacja rozkładu zmiennej, to należy przyjąć równe wartości przedziałów klasowych

Stratyfikacja



- Jeśli celem ma być prezentacja rozkładu zmiennej, to należy przyjąć równe wartości przedziałów klasowych

Wiekowanie

Obliczanie wieku danego procesu (np. data końcowa minus data początkowa) i stratyfikacja względem wieku.

Uwagi:

- Przedziały klasowe (najczęściej w dniach) wynikają z przepisów lub innych przesłanek (np. 30, 60, 180, 270, ...).
- Wyodrębnić braki dat, a w przypadku braku daty końcowej zwrócić uwagę na sprawy nie zakończone pomimo upływu wymaganego terminu.
- Zwrócić uwagę czy data zapisana jest w formacie data, czy data:czas.

Przykład:

fakturę wystawiono z datą: 2022.11.03, płatność 30 dni,
zapłatę zaewidencjonowano: 2022.12.03 13:30

Wiek = 30,56, czyli przekroczono 30 dni

Podstawowe metody analizy danych

- w audycie zgodności
- w audycie wykonania zadań
- w audycie typu finansowego
- w audycie śledczym

Analiza danych w audycie zgodności

Porównanie danych z wymogami zapisanymi w:

- przepisach prawa powszechnie obowiązującego,
- procedurach i politykach danej firmy,
- zasadach obsługi badanego procesu,
- wymaganiach dotyczących samego systemu IT,
- zasadach dobrej praktyki.

Badanie dotyczy na ogół takich cech, jak:

kompletność, poprawność formalna, terminowość,
klasyfikacja, powiązanie, jednoznaczność, ciągłość,
niesprzeczność, zachowanie sekwencji postępowania

Badania w audycie wykonania zadań (1)

1. Analizy wskaźnikowe

- proste w realizacji np. wynik/nakład
- problem w przypadku gdy mamy więcej nakładów lub wyników

2. Analiza kosztów i korzyści (CBA)

- trudność: zwymiarowanie finansowe korzyści
- w projektach wieloletnich dyskontowanie korzyści
- popularna przy planowaniu projektów, rzadziej w audycie
- dostępne szczegółowe poradniki dotyczące jej stosowania

3. Analiza efektywności kosztowej (CEA)

- koszt finansowo, efekt w jednostkach naturalnych
- metoda uproszczona – wymóg pojedynczego efektu
- modyfikacje: DCG (*Dynamic Generation Cost*), ICER (*Incremental Cost-Effectiveness Ratio*)

4. Analiza korelacji i regresji

- relacje pomiędzy różnymi cechami dot. badanego podmiotu:
 - silna korelacja tam, gdzie nie powinna występować
 - brak korelacji lub odwrotna tam, gdzie powinna występować
- relacje pomiędzy różnymi cechami dot. podmiotów zależnych (oddziały, jednostki podległe):
 - identyfikacja trendu i odchyień od trendu
 - analiza zmian współczynnika korelacji w czasie
- relacji pomiędzy cechami dot. niezależne podmioty z danej branży:
 - nasza jednostka na tle innych z branży

4. Analiza korelacji i regresji

- relacje pomiędzy cechami dot. badanej podmioty
 - silna korelacja
 - brak korelacji
- relacje pomiędzy cechami dot. różnych podmiotów (oddziały, jednostki)
 - identyfikacja
 - analiza zmian
- relacji pomiędzy cechami dot. niezależne podmioty z danej branży
 - nasza jednostka na tle innych z branży

Analiza zmian reorganizacyjnych firmy:

- miarą alokacji zasobów ludzkich jest dostosowanie liczby zatrudnionych w poszczególnych oddziałach firmy do liczby zadań,
- wzrost współczynnika korelacji tych cech po reorganizacji będzie potwierdzał właściwy kierunek zmian reorganizacyjnych,
- spadek współczynnika, stawia w wątpliwość kierunek lub skuteczność tych zmian.

5. Analiza szeregów czasowych

- Stosuje się, gdy poszczególne obserwacje badanej zmiennej przypisane są do kolejnych jednostek czasu
- Na dane w szeregu mogą mieć wpływ: wahania okresowe, wahania losowe, trend
- Zakres i cel badania:
 - identyfikacja prawidłowości, którym podlega proces w czasie
 - porównanie z wahaniami innych cech lub innych podmiotów
 - identyfikacja wartości nietypowych
 - prognozowanie wartości cechy w czasie

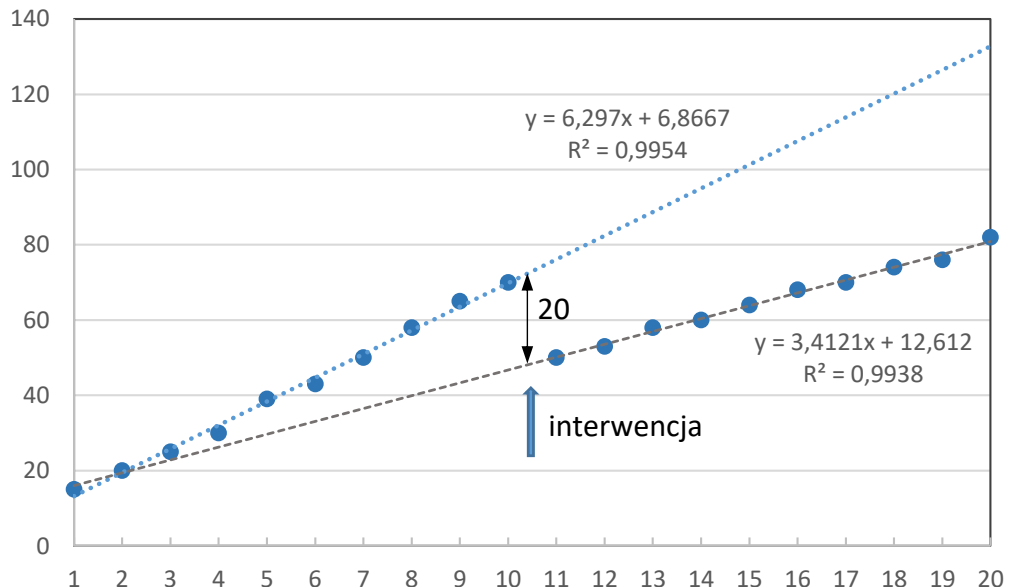
5a. Analiza przerwanych szeregów czasowych (ITS – *Interrupted Time Series*)

- Stosowana, gdy chcemy dokonać analizy wpływu danej interwencji na obszar objęty badaniem (zmiana wartości lub trendu)

Badania w audycie wykonania zadań (3)

5. Analiza szeregów czasowych

- Stosuje się, gdy poszczególne dane są przypisane do kolejnych okresów czasu
- Na dane w szeregu mogą być nałożone wahania losowe, trend
- Zakres i cel badania:
 - identyfikacja prawdy
 - porównanie z wartościami
 - identyfikacja wartości
 - prognozowanie wartości



5a. Analiza przerwanych szeregów czasowych (ITS – *Interrupted Time Series*)

- Stosowana, gdy chcemy dokonać analizy wpływu danej interwencji na obszar objęty badaniem (zmiana wartości lub trendu)

Badania w audycie wykonania zadań (4)

6. Analiza obwiedni danych – DEA (*Data Envelopment Analysis*)

- ocena efektywności przy wielu nakładach i wielu wynikach bez znajomości zależności funkcyjnej pomiędzy wynikami i nakładami

$$\frac{\text{wyniki}}{\text{nakłady}} = \frac{a_1 y_1 + a_2 y_2 + a_3 y_3 + \dots}{b_1 x_1 + b_2 x_2 + \dots}$$

- do obliczeń niezbędne są dane o nakładach i wynikach wielu podobnych podmiotów (DMU – Decision Making Unit)
- zastosowanie metody DEA pozwala na:
 - identyfikację obiektów optymalnych (granicznych),
 - wyznaczenie poziomu efektywności pozostałych obiektów względem efektywności granicznej,
 - wskazanie obiektów wzorcowych (referencyjnych) w stosunku do każdego z obiektów nieoptymalnych.

Badania w audycie wykonania zadań (4)

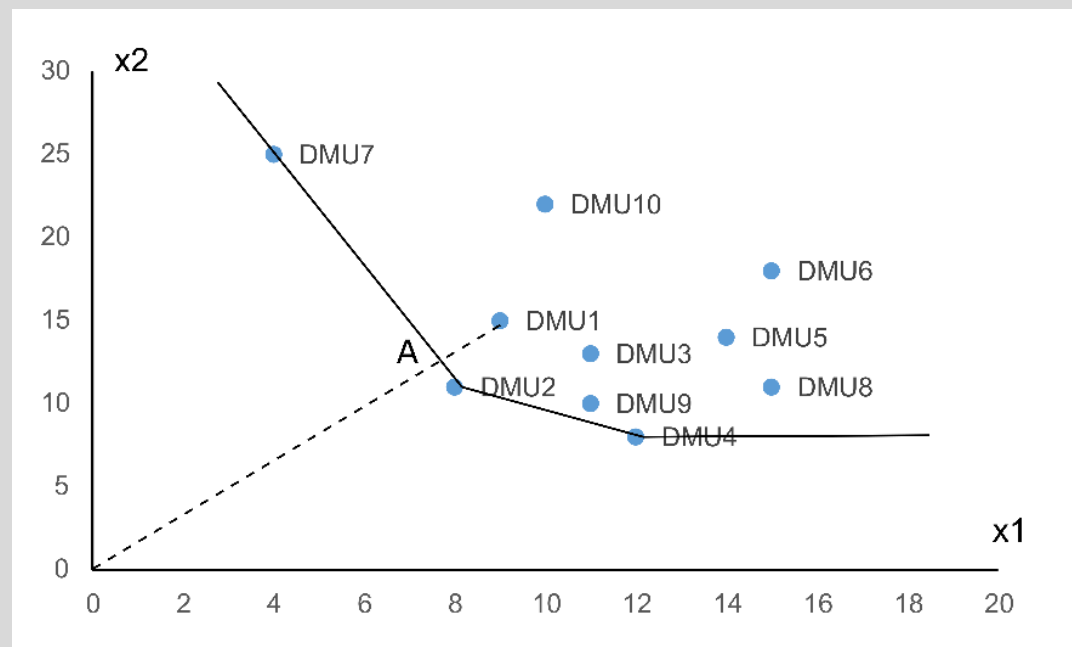
6. Analiza obwiedni danych – DEA (*Data Envelopment Analysis*)

- ocena efektywności przy wielu nakładach i wielu wynikach bez znajomości zależności funkcyjnej pomiędzy wynikami i nakładami

Obwiednia danych – model zorientowany na nakłady

$\frac{w}{na}$

- do obliczeń niezbędne są dane o podobnych podmiotach
- zastosowanie metody:
 - identyfikację obiektów efektywnych
 - wyznaczenie poziomu efektywności względem efektów
 - wskazanie obiektów, które należy poprawić



Badania w audycie wykonania zadań (5)

6. Analiza DEA c.d.

- analiza ukierunkowana jest na nakłady (I) albo na wyniki (O)
- najbardziej popularne metody CCR i BCC
- metoda stosowana szeroko w ewaluacji, benchmarkingu, przy tworzeniu list rankingowych
- szczególnie przydatna w sektorze publicznym ze względu na trudność określenia wartości docelowych mierników efektywności
- stosunkowo mało spopularyzowana w audycie
- wymaga zastosowania specjalistycznych narzędzi informatycznych

7. Badania ankietowe (patrz cz.I wykładu)

Szerzej i możliwie prosto o DEA: Kontrola Państwowa, 4/2022



CAATs w audycie typu finansowego – źródła danych

- System FK (dziennik, księga główna, księgi pomocnicze)
- Plik JPK-KR (format xml)

Zestaw danych:

Dane identyfikacyjne, Zestawienie obrotów i sald (ZOiS), Dziennik, Zapisy na kontach analitycznych (Konto_Zapis), Elementy kontrolne

Import:

najlepiej do odrębnych arkuszy, zapis w postaci tablic;
w Excelu można wykorzystać opcję **Dane-> Pobierz Dane** lub
opcję **Deweloper -> Źródło** do mapowania struktury z pliku
Schemat_JPK_KR i **Deweloper -> Import** do importu danych

CAATs w audycie typu finansowego – weryfikacja

- Przy pobieraniu z FK:
 - sprawdzenie formatów, zakresów, podsumowań
- W przypadku importu JPK-KR sprawdzenie:
 - zgodności danych z elementami kontrolnymi
 - zakresu dat z okresem, którego dotyczy raport JPK-KR
 - zgodności ZOiS z JPK-KR i ZOiS z systemu FK
 - spójności poszczególnych tablic
- Przygotowanie zmiennych pomocniczych:
 - identyfikator kontrahenta
 - paragraf klasyfikacji budżetowej

CAATs w audycie typu finansowego – księgi (1)

- Brak obligatoryjnych elementów zapisu
- Błędne daty
- Błędna klasyfikacja paragrafowa
- Wn<>Ma dla kont bilansowych
- Luki w numeracji zapisów w Dzienniku
- Operacja gospodarcza dot. innego roku lub innego okresu
- Zapis po sprawozdaniu
- Niezgodność paragrafu po stronie Wn i Ma
- Pomijanie kont rozrachunkowych (Wn-4xx – Ma-13x)
- Niedozwolone zapisy dot. środków UE
- Niedozwolone pary kont (np. 227 – 130)

CAATs w audycie typu finansowego – księgi (2)

- Nieterminowe odprowadzenie odpisów i świadczeń
- Przekroczenie planu wydatków
- Wykonanie większe niż zaangażowanie
- BO różne od BZ roku poprzedniego (konto po koncie)
- Zapisy nietypowe:
 - zapisy stornujące dot. dużych kwot
 - naruszenie naturalnej sekwencji dat
 - zapis w dniu wolnym od pracy
 - pojedyncze księgowania operatora
 - dokonywane przez pracownika IT
 - w dacie dowodu zewnętrznego
 - zmniejszające koszty

Analizy w audycie finansowym – sprawozdanie

- analizy wskaźnikowe (płynność, aktywność gospodarcza, zadłużenie, rentowność)
 - analiza dynamiki zmian (trendu)
 - analiza porównawcza (porównanie wskaźników z branżą)
 - test racjonalności
 - ocena zagrożeń dla kontynuacji działalności (różne modele)
-
- badanie pod kątem możliwych oszustw (*patrz audyt śledczy*)

Analityka śledcza (wykrywanie oszustw)*

- Analiza Benforda – test pierwszych cyfr (1 lub 2)
- Względny współczynnik wielkości ($RSF = V_{\max} / V_{\max-1}$)
- Test duplikatów wartości w podzbiorach
- Test SSS (ta sama kwota, data, faktura, ten sam dostawca)
- Test SSD (ta sama kwota, data, faktura, inny dostawca)
- Analiza korelacji i analiza szeregów czasowych
- Analizy wskaźnikowe
 - wskaźniki operacyjne
 - model Beneisha (8) i Roxasa (5): *manipulator/niemanipulator*

Patrz publikacje: Marc Nigrini i Sanders Gee

Analiza Benforda

Prawo Benforda – częstość występowania poszczególnych cyfr na pierwszych (najbardziej znaczących) pozycjach liczb pochodzenia naturalnego jest mocno zróżnicowana:

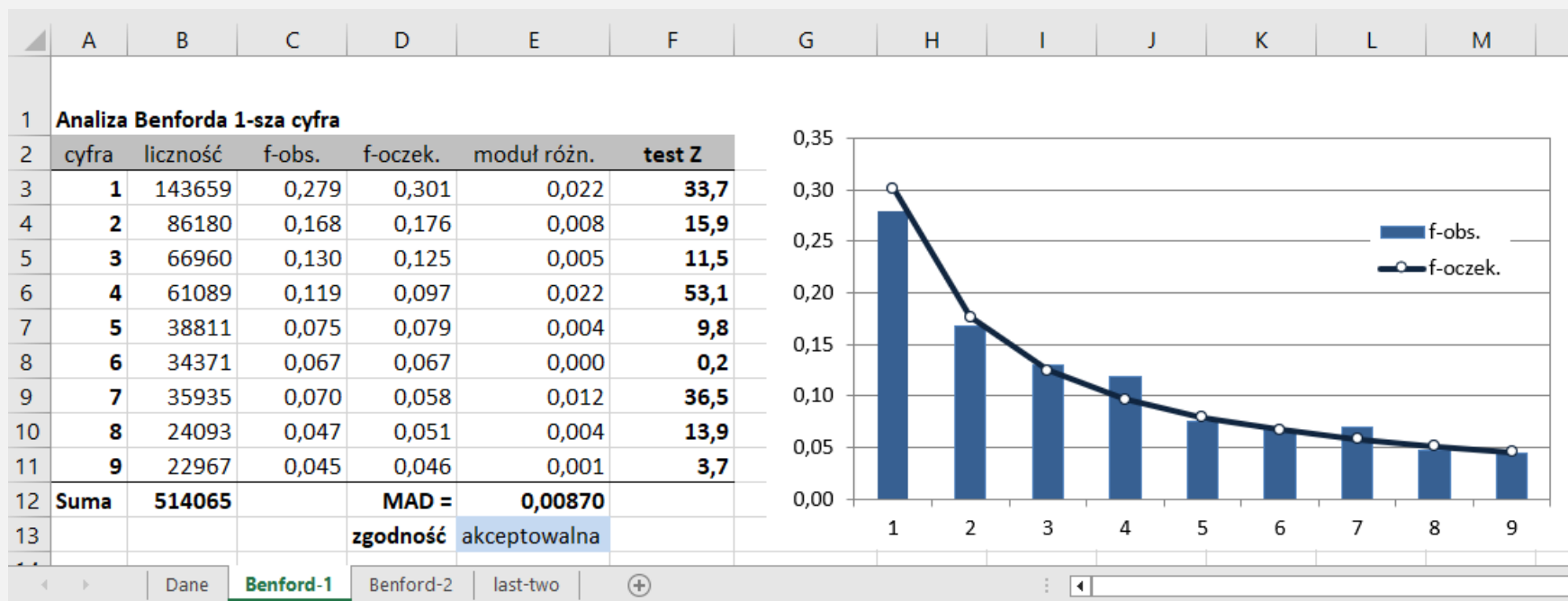
1	2	3	4	5	6	7	8	9
0,301	0,176	0,125	0,097	0,079	0,067	0,058	0,051	0,046

Istota analizy Benforda:

- Analizę prowadzimy dla 1-szej i 2-ch pierwszych cyfr
- Oceniamy zgodność obserwowanej częstości z teoretyczną (analitycznie lub graficznie)
- Miara zgodności: dla całego rozkładu – MAD, dla pozycji – test Z
- W przypadku odchyleń – selekcjonujemy zapisy i badamy przyczyny

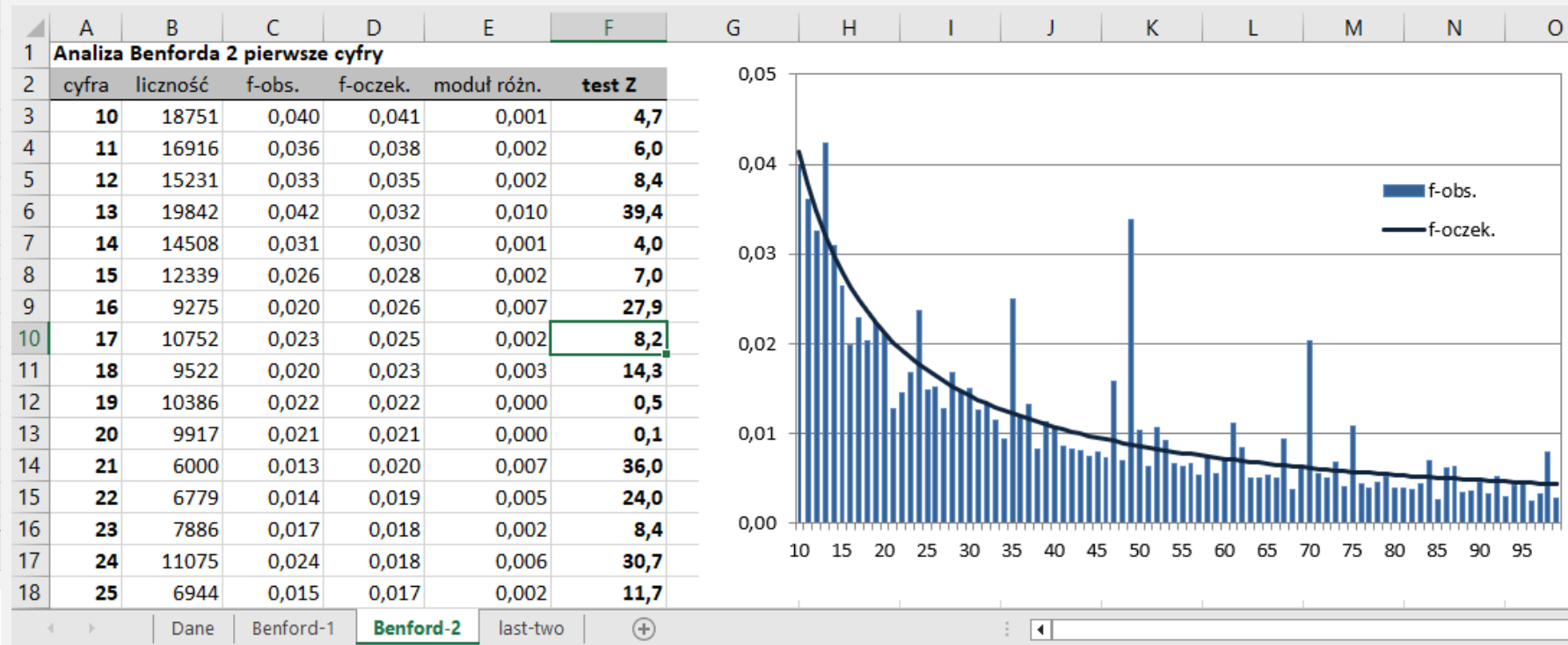
Analiza Benforda

Prawo Benforda – częstość występowania poszczególnych cyfr na pierwszych (najbardziej znaczących) pozycjach liczb pochodzenia naturalnego jest mocno zróżnicowana.



Analiza Benforda

Prawo Benforda – częstość występowania poszczególnych cyfr na pierwszych (najbardziej znaczących) pozycjach liczb pochodzenia naturalnego jest mocno zróżnicowana.



Problemy ze stosowaniem CAATs:

- brak wystarczającej wiedzy o zasobach informacyjnych na etapie przygotowania audytu
- trudność w powiązaniu procesu poddawanego audytowi z zasobami systemów IT
- brak standaryzacji raportów z systemów IT, a w efekcie duża pracochłonność importu danych
- oczekiwanie twardych dowodów
- pochopne wnioski i rozczarowanie efektami
- bariera wiedzy i umiejętności po stronie audytorów
- ryzyko niezrozumienia wyników przez audytowanego

Wydaje się, że od wykorzystania CAATs w audycie nie ma odwrotu.

Dziękuję Państwu za udział w cz. II

ws.karlinski@gmail.com