Marek Cieciura, Janusz Zacharski

PODSTAWY PROBABILISTYKI Z PRZYKŁADAMI ZASTOSOWAŃ W INFORMATYCE

CZĘŚĆ I
WPROWADZENIE

Na prawach rękopisu

Warszawa, wrzesień 2011

Data ostatniej aktualizacji: czwartek, 20 października 2011, godzina 17:50

Statystyka jest bardziej sposobem myślenia lub wnioskowania niż pęczkiem recept na młócenie danych w celu odsłonięcia odpowiedzi - Calyampudi Radhakrishna Rao

Podręcznik:

PODSTAWY PROBABILISTYKI Z PRZYKŁADAMI ZASTOSOWAŃ W INFORMATYCE

publikowany jest w częściach podanych poniżej

Nr	Tytuł
I.	Wprowadzenie
II.	Statystyka opisowa
III.	Rachunek prawdopodobieństwa
IV.	Statystyka matematyczna
V.	Przykłady zastosowań w informatyce
VI.	Wybrane twierdzenia z dowodami
VII.	Tablice statystyczne

Autorzy proszą o przesyłanie wszelkich uwagi i propozycji dotyczących zawartości podręcznika z wykorzystaniem formularza kontaktowego zamieszczonego w portalu http://cieciura.net/mp/

Publikowane części będą na bieżąco poprawiane, w każdej będzie podawana data ostatniej aktualizacji.

Podręcznik udostępnia się na warunku licencji Creative Commons (CC): **Uznanie Autorstwa – Użycie Niekomercyjne – Bez Utworów Zależnych** (CC-BY-NC-ND),co oznacza:

- **Uznanie Autorstwa** (ang. Attribution - BY): zezwala się na kopiowanie, dystrybucję, wyświetlanie i użytkowanie dzieła i wszelkich jego pochodnych pod warunkiem umieszczenia informacji o twórcy.
- **Użycie Niekomercyjne** (ang. Noncommercial - NC): zezwala się na kopiowanie, dystrybucję, wyświetlanie i użytkowanie dzieła i wszelkich jego pochodnych tylko w celach niekomercyjnych..
- **Bez Utworów Zależnych** (ang. No Derivative Works - ND): zezwala się na kopiowanie, dystrybucję, wyświetlanie tylko dokładnych (dosłownych) kopii dzieła, niedozwolone jest jego zmienianie i tworzenie na jego bazie pochodnych.

Podręcznik i skorelowany z nim portal, są w pełni i powszechnie dostępne, stanowią więc Otwarte Zasoby Edukacyjne - OZE (ang. Open Educational Resources – OER).

SPIS TREŚCI

1. WPROWADZENIE

1.1. Populacja i jej cechy

Populacja jest to zbiór elementów podlegających badaniu statystycznemu.

Elementy populacji charakteryzują się:

- Właściwością wspólną, pozwalającą odróżnić elementy populacji od innych elementów, które nie należą do danej populacji.
- Właściwościami różniącymi je między sobą.

Aby można było odróżnić elementy populacji od innych elementów, populacja powinna być określona pod względem:

- rzeczowym,
- terytorialnym (przestrzennym),
- czasowym.

Zatem określenie populacji powinno zawierać odpowiedzi na pytania:

- Kto? Co?
- Gdzie?
- Kiedy?

Przykład 1.1

Populacja: Zbiór studentów pewnej uczelni (oznaczmy ją U), w Warszawie, wg stanu na 15.10. 2005.

- Kto?: Student uczelni U.
- Gdzie?: W Warszawie.
- Kiedy?: 15.10. 2005

Populacja jest określona pod względem rzeczowym, terytorialnym i czasowym. ■

Cecha populacji jest to właściwość, ze względu na którą elementy populacji mogą się różnić.

Przykład 1.2

Populacja: Zbiór studentów uczelni U, w Warszawie, wg stanu na 15.10. 2005.
Cechy populacji: wiek, płeć, stan cywilny, liczba zaległych egzaminów, kolor oczu, ocena ze statystyki. ■

1.1.1. Warianty cechy

Warianty cechy są to możliwe wartości tej cechy.

Cecha populacji	**Warianty cechy**
Płeć	Kobieta, mężczyzna
Kolor oczu	Czarny, niebieski, zielony, szary, piwny.
Ocena ze statystyki	ndst, dst, db, bdb
Liczba ukończonych lat	0, 1, 2, 3, ...
Czas świecenia żarówki	Dowolna liczba z przedziału $< 0 ; \infty)$

1.1.2. Typy cech. Skale cech

Wyróżniające jednostki wchodzące w skład badanej zbiorowości nazywamy cechami statystycznymi. Populacja statystyczna może mieć dużo rozmaitych cech. W zależności od celu badania wybieramy tylko niektóre z nich, najważniejsze w odniesieniu do interesującego nas problemu.

Rozróżniamy trzy zasadnicze typy cech: jakościowe, porządkowe i ilościowe (rys. 1.1).

Rysunek 1.1.

Cechy jakościowe (niemierzalne) to takie, których nie można jednoznacznie scharakteryzować za pomocą liczb (czyli nie można zmierzyć). Możemy je tylko opisać słowami. Możliwa jest zatem jedynie zupełna i rozłączna klasyfikacja zbioru wyników. Podstawową operacją pomiarową jest identyfikacja kategorii, do której należy zaliczyć wynik. Prowadzi to do podziału zbioru wyników na podzbiory rozłączne. Synonimem cechy jakościowej jest cecha w skali nominalnej.

Przykłady: kolor oczu, stan cywilny, zawód, adres.

W każdym z powyższych przykładów nie można stwierdzić, który wariant jest wcześniejszy od drugiego, o ile te warianty się różnią (nie jest określona różnica wariantów), ile razy jeden wariant jest większy od drugiego (nie jest określony stosunek) wariantów.

Cechy porządkowe (mierzalne) umożliwiają porządkowanie (lub uszeregowanie) wszystkich elementów zbioru wyników. Cechy takie najlepiej określa się przymiotnikami i ich stopniowaniem. Każdemu ze stanów można również przypisać liczbę według wzrostu natężenia. Proces ten nazywa się rangowaniem. Na przykład, badając wzrost osoby, możemy użyć określeń: "niski", "średni" lub "wysoki". Synonimem cechy porządkowej jest cecha w skali porządkowej.

Możliwe jest stwierdzenie dla dowolnych dwóch wariantów, czy są one równe, a jeśli nie to, który jest mniejszy od drugiego, czyli w zbiorze wariantów wprowadzona jest relacja uporządkowania.

Przykłady

Wykształcenie. Warianty w_1 - wykształcenie podstawowe, w_2 - wykształcenie średnie, w_3 wykształcenie wyższe. Naturalne jest przyjąć, że $w_1 < w_2 < w_3$. Zatem cecha jest w skali porządkowej.

W każdym z powyższych przykładów nie można stwierdzić o ile warianty się różnią (nie jest określona różnica wariantów), oraz ile razy jeden wariant jest większy od drugiego (nie jest określony stosunek) wariantów.

Cechy ilościowe (mierzalne) to takie, które dadzą się wyrazić za pomocą jednostek miary w pewnej skali. Cechami mierzalnymi są na przykład: wzrost (w cm), waga (w kg), wiek (w latach) itp. Wśród cech mierzalnych wyróżniamy dwie podgrupy: cechy ciągłe i cechy skokowe.

Cecha ciągła to zmienna, która może przyjmować każdą wartość z określonego skończonego przedziału liczbowego, np. wzrost, masa ciała czy temperatura.
Cechy skokowe mogą przyjmować wartości ze zbioru skończonego lub przeliczalnego (zwykle całkowite), na przykład: liczba posiadanych dzieci, czy wysokość zarobków.

Wyróżnia się tutaj dwie skale: przedziałową i ilorazową.

Skala przedziałowa jest to skala, w której warianty są liczbami wraz z jednostkami, przy czym określone jest odejmowanie wariantów (w tym sensie, że różnica wariantów ma sens zważywszy na znaczenia wariantów), czyli można stwierdzić o ile jednostek jeden wariant jest większy lub mniejszy od drugiego.

Przykłady

Skala temperatury Celsjusza, dni roku, miejsce zajęte przez kierowcę w wyścigu Formuły1.

W każdym z powyższych przykładów nie można stwierdzić ile razy jeden wariant jest większy od drugiego (nie jest określony stosunek wariantów), można wprawdzie podzielić liczby wyrażające warianty, lecz otrzymany stosunek nie ma sensu, gdy uwzględni się znaczenie wariantów.

Skala ilorazowa jest skala, w której warianty są liczbami wraz z jednostkami, przy czym określone jest odejmowanie i dzielenie wariantów.

Przykłady

Temperatura w skali bezwzględnej (w skali K), czas zawodnika na mecie w sekundach, masa towaru w kg.

W każdym z powyższych przykładów można stwierdzić ile razy dany wariant jest większy od drugiego.

Skale można uporządkować następująco:

Skala ilorazowa, Skala przedziałowa, Skala porządkowa, Skala nominalna

w tym sensie, że występująca w powyższym ciągu skala jest zarazem każą ze skal po niej następującej. Inaczej powyższy fakt wyrażamy mówiąc, że skala nominalna jest najniższego poziomu, po niej kolejno występują skale porządkowa, przedziałowa i najwyższego poziomu skala ilorazowa.

Możliwa jest transformacja skali wyższego poziomu na skalę niższego poziomu (patrz przykłady 17.1 oraz 17.9 – 17.11).

Pojęcie skali wprowadza się dlatego, iż w zależności od jej poziomu można stosować właściwe metody statystyczne.

Przykład 1.3
Cechy z przykładu 8.2: wiek, liczba zaległych egzaminów są cechami mierzalnymi (ilościowymi), natomiast cechy: płeć, stan cywilny, kolor oczu oraz ocena ze statystyki są cechami niemierzalnymi (jakościowymi).

Wiek i liczba zaległych egzaminów to cechy w skali ilorazowej.

Ocena ze statystyki to cecha w skali porządkowej.

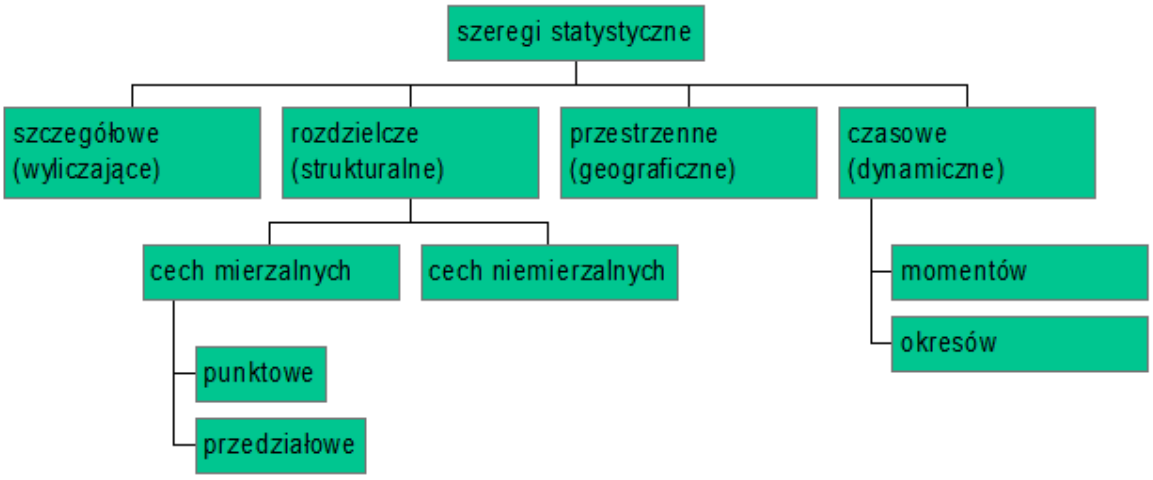
Płeć i stan cywilny są wyrażone w skali nominalnej. ■

Podsumowanie cech podano w tabeli 1.1.

W statystyce matematycznej cechy traktuje się jako zmienne losowe.

1.2. Szeregi statystyczne

Szereg statystyczny to zbiór wartości liczbowych badanej cechy uporządkowany według określonych kryteriów. Rozróżnimy kilka rodzaj szeregów statystycznych.



Rysunek 1.2.

Szeregi szczegółowe i rozdzielcze (punktowe, przedziałowe) charakteryzują stan badanej zbiorowości w określonym momencie (np. w danym miesiącu, roku). Przedstawiają więc populacje w układzie statycznym i służą do analizy jej struktury.

Szeregi przestrzenne przedstawiają rozmieszczenie wielkości statystycznych według podziału administracyjnego (gmina, powiat, województwo, krajów, regionów geograficznych).

Szeregi dynamiczne (czasowe, chronologiczne) przedstawiają rozwój zjawiska w czasie.

Szeregi czasowe momentów prezentują zjawisko w ściśle określonym momencie, zaś szeregi czasowe okresów w ściśle określonym przedziale czasowym.

Poniżej podano zasady grupowania danych statystycznych w szereg rozdzielczy przedziałowy

1. Ustalamy liczbę klas.
   Liczbę klas (oznaczenie r) wyznaczamy wg tabeli

Liczba danych statystycznych n	Liczba klas r
30-60	6-8
60-100	7-10
100-200	9-12
200-500	11-17
>500	16-25

2. Wyznaczamy długość klasy.

   Zakładamy, że wszystkie klasy mają równe długości. Długość klasy b wyznaczamy wg wzoru

$$b = \frac{r_o}{r}$$

   gdzie: $r_0 = x_{max} - x_{min}$ rozstęp

   Wynik dzielenia zaokrąglamy zawsze w górę do dokładności danych statystycznych.

Zaokrąglenie w górę zapewnia zmieszczenie się wszystkich danych statystycznych w wyznaczonych przedziałach.

3. Wyznaczamy końce klas.

Przyjmujemy, że klasy są przedziałami lewostronnie domkniętymi i prawostronnie otwartymi.

$$A_1=<a_1;a_2),\ A_2=<a_2;a_3),\ \ldots,\ A_r=<a_r;a_{r+1})$$

Wtedy przyjmujemy, że lewy koniec pierwszej klasy jest równy

$$a_1 = x_{min}$$

Zatem

$$a_2 = a_1 + b, a_3 = a_2 + b, \ldots$$

4. Wyznaczamy liczebności klas.
W tym celu wygodnie jest dane statystyczne posortować. Wyznaczone przedziały i ich liczebności przedstawiamy w tabeli

Klasa $A_i =< a_i ; a_{i+1})$	Liczebność n_i
$< a_1 ; a_2)$	n_1
$< a_2 ; a_3)$	n_2
.....	...
$< a_r ; a_{r+1})$	n_r
Razem	n

Przykład 1.4

Badano dodatek do wynagrodzenia (w zł.) 40 pracowników pewnego przedsiębiorstwa. Otrzymano następujące dane

405, 420, 411, 427, 479, 440, 378, 468, 437, 452, 421, 414, 402, 422, 462, 431, 414, 437, 405, 390, 425, 425, 400, 432, 447, 385, 419, 400, 425, 458, 439, 360, 405, 369, 406, 431, 412, 387, 416, 415.
Przedstawimy powyższe dane w szeregu rozdzielczym przedziałowym.

Rozwiązanie

Przyjmiemy, że klas jest 6 (co jest zgodne z tabelą z punktu 1). Obliczmy długość klasy.
W tym celu sortujemy dane statystyczne
360, 369, 378, 385, 387, 390, 400, 400, 402, 405, 405, 405, 406, 411, 412, 414, 414, 415, 416, 419,420, 421, 422, 425, 425, 425, 427, 431, 431, 432, 437, 437, 439, 440, 447, 452, 458, 462, 468, 479

$x_{max} = 479,\ \ x_{min} = 360,$

Rozstęp $r_o = 479 - 360 = 119,$

Liczba klas $r = 6,$

Długość klasy $b = 119/6 = 19{,}83 \approx 20$ (zaokrąglono w górę do dokładności danych statystycznych, która wynosi w tym przykładzie 1).

Wyznaczamy końce klas

$$a_1=x_{min}=360,\ \ a_2=a_1+b=380,\ \ a_3=a_2+b=400,\ \ a_4=a_3+b=420,$$
$$a_5=a_4+b=440,\ \ a_6=a_5+b=460,\ \ a_7=a_6+b=480$$

Klasy

$A_1=<360;380)$, $A_2=<380;400)$, $A_3=<400;420)$, $A_4=<420;440)$, $A_5=<440;460)$, $A_6=<460;480)$

Wyznaczamy liczebności klas. Korzystamy z posortowanych danych statystycznych. Wyniki zapisujemy w szeregu rozdzielczym przedziałowym.

Klasa $A_i=<a_i;a_{i+1})$	Liczebność n_i
$<360;380)$	3
$<380;400)$	3
$<400;420)$	14
$<420;440)$	13
$<440;460)$	4
$<460;480)$	3
Razem	40

■

W powyższym przykładzie przyjęliśmy, że przedziały są lewostronnie domknięte i prawostronnie otwarte. Można było przyjąć inaczej, że są lewostronnie otwarte i prawostronnie domknięte lub są obustronnie otwarte.

Jeśli przyjąć, że są lewostronnie otwarte i prawostronnie domknięte, to najpierw wyznaczamy prawy koniec ostatniej klasy wg wzoru $a_{r+1}=x_{max}$, a następnie przez odejmowanie długości klasy otrzymujemy końce klas poprzednich.

Jeśli przyjąć, że klasy są obustronnie otwarte, to wyznaczamy najpierw lewy koniec pierwszej klasy wg wzoru $a_1=x_{min}-\frac{\alpha}{2}$, α – dokładność danych statystycznych, a następnie przez dodawanie długości przedziału otrzymujemy końce pozostałych klas.

Zaletą klas obustronnie otwartych jest fakt, że żadna dana statystyczna nie jest równa końcowi jakiejkolwiek klasy, możemy więc w szeregu rozdzielczym przedziałowym zapisywać te klasy bez podania informacji czy końce przedziałów należą do klasy czy też nie należą.

Przykład 1.5

Przedstawimy dane statystyczne z poprzedniego przykładu w szeregu rozdzielczym przedziałowym przyjmując, że klasy są obustronnie otwarte.

Rozwiązanie

$$a_1=x_{min}-\frac{\alpha}{2}=360\text{-}0{,}5=359{,}5;\quad a_2=a_1+b=359{,}5+20=379{,}5;\ a_3=a_2+b=399{,}5 \text{ itd}$$

Szereg rozdzielczy przedziałowy

Klasa $A_i=(a_i;a_{i+1})$	Liczebność n_i
359,5 ; 379,5	3
379,5 ; 399,5	3
399,5 ; 419,5	14
419,5 ; 439,5	13
439,5 ; 459,5	4
459,5 ; 479,5	3
Razem	40

■

Tabela 1.1. Podsumowanie skal pomiarowych[1]

Rodzaj skali pomiarowej[2]	Nazwa skali pomiarowej	Właściwości skali	Przykłady	Możliwe operacje
JAKOŚCIOWA	Skala nominalna (*nominal scale*)	Najprostsza skala pomiarowa. Pozwala na identyfikację, klasyfikowanie i nazywanie poczynionych przez badacza obserwacji. Pozwala na rozróżnianie jakości. Odzwierciedla symbole wskazujące przynależność przedmiotów do pewnych klas jakościowych wyrażonych słownie za pomocą nazw i symboli np. liter lub numerycznie, tj. za pomocą liczb.	Imię i nazwisko, płeć, kolor oczu, data urodzenia, numery tramwajów, numery telefonów, symbole grupy krwi, miejsce urodzenia, miejsce zamieszkania, wyznanie religijne	Zmienne mierzone na skali nominalnej można zdefiniować jako wyszczególnienie występujących przypadków. Jedyną dozwoloną relacją porównującą dwie wartości na skali nominalnej jest równość. Tylko pewne wyniki można grupować, a uporządkowanie ich jest ryzykowne.
	Skala porządkowa (*ordinal scale*)	Jest to skala mająca właściwości porządkowe ujawniające się uszeregowaniem obserwacji badacza w obrębie jakiejś dymensji. Jej celem jest ustalenie hierarchii wartości zmiennej. Składa się z symboli - rang odnoszących się do przedmiotów uporządkowanych pod pewnym względem. Rangi określają pozycję danego przedmiotu w zbiorze przedmiotów o charakterze rosnącym lub malejącym. Pozwala na porównywanie przedmiotów między sobą, ale nie można za jej pomocą ustalić wielkości różnic między obiektami.	Stopnie wojskowe, pozycja zajmowana w tabeli przez drużyną piłkarską, ranking szkół wyższych, wyniki turnieju szachowego.	Zmienne mierzone na skali porządkowej można zdefiniować jako uszeregowanie poszczególnych przypadków ze względu na jakąś właściwość. Oprócz równości możliwe są relacje porządku ($<$ $>$ $\leq$ $\geq$)

[1] Wykorzystano http://pedagogikaspecjalna.tripod.com/notes/pdscales.html
[2] Podział zaproponowany przez Stevensa w 1946 roku

Rodzaj skali pomiarowej[2]	Nazwa skali pomiarowej	Właściwości skali	Przykłady	Możliwe operacje
ILOŚCIOWA	Skala przedziałowa/ interwałowa (*interval, additive scale*)	Składa się z symboli, których pary obrazują różnice między przedmiotami, wyrażone w jednostkach miary. Punkt zerowy zwykle jest umowny (np. temperatura topnienia lodu w skali temperatur Celsjusza). Pozwala na stwierdzenie o ile natężenie zmiennej X dla obiektu A jest większe (mniejsze) od natężenia zmiennej dla obiektu B.	Długość i szerokość geograficzna w stopniach, skale do mierzenia temperatury powietrza (Celsjusza, Fahrenheita), wyniki uzyskane w testach	Jest skalą o wysokim stopniu użyteczności dla różnorodnych pomiarów. Może być dodatkowo wyrażona normami np. w postaci skali stenowej. Różnice pomiędzy wartościami mają sensowną interpretację, ale ich iloraz nie.
	Skala stosunkowa/ ilorazowa (*ratio, absolute scale*)	Składa się z symboli, których pary przedstawiają stosunki wartości przedmiotów. Skala ta ma bezwzględne zero wartości zmiennej. Bywa nazywana skalą metryczną. Pozwala dodatkowo na stwierdzenie, że natężenie zmiennej X dla obiektu A jest k razy większe niż natężenie tej zmiennej dla obiektu B.	Długość, szerokość, wysokość przedmiotów wyrażona w jednostkach miar SI (m, cm, mm, km) lub innych (mila, cal); testy szybkości, wiek wyrażony w dniach życia, liczba dzieci w rodzinie	Skala o najwyższym stopniu użyteczności dla różnorodnych pomiarów. Pozwala dostrzec bardziej precyzyjnie różnice. Nie tylko różnice, ale także ilorazy wielkości mają interpretację. Przykładem jest masa (coś może być dwa razy cięższe). Wielkości na skali ilorazowej można dodawać odejmować i dzielić przez siebie.

1.2. Metody badań statystycznych

Dwie podstawowe metody badań statystycznych:
- Badanie pełne;
- Badanie częściowe.

1.2.1. Badanie pełne

Badanie pełne polega na wyznaczeniu wartości badanej cechy wszystkich jednostek populacji.

Zalety: Badanie pełne dostarcza kompletnych informacji o strukturze badanej cechy, a więc pozwala wyznaczyć w pełni jej rozkład prawdopodobieństwa.

Wady: Dla populacji o dużej liczbie elementów badanie pełne jest:
- ✓ Technicznie skomplikowane;
- ✓ Bardzo drogie;
- ✓ Opracowanie wyników trwa długo;
- ✓ Niemożliwe do wykonania, gdy badanie jest niszczące lub, gdy liczba wariantów populacji jest nieskończona.

W niektórych przypadkach, np. przy spisie ludności, badania pełne są obligatoryjne.

1.2.2. Badanie częściowe

Badanie częściowe polega na wyznaczeniu wartości cechy X tylko niektórych, specjalnie dobranych jednostek populacji.

Badanie częściowe stosujemy, gdy badanie:
- ✓ jest niszczące;
- ✓ pełne jest zbyt drogie;
- ✓ musi być przeprowadzone i opracowane w krótkim czasie.

1.3. Próba losowa

Badamy cechę X populacji. Losujemy z populacji n elementów.
Oznaczenia:

x_1 - wartość cechy X pierwszego wylosowanego elementu,
x_2 - wartość cechy X drugiego wylosowanego elementu,
.....
x_n - wartość cechy X n-tego wylosowanego elementu.

Ciąg

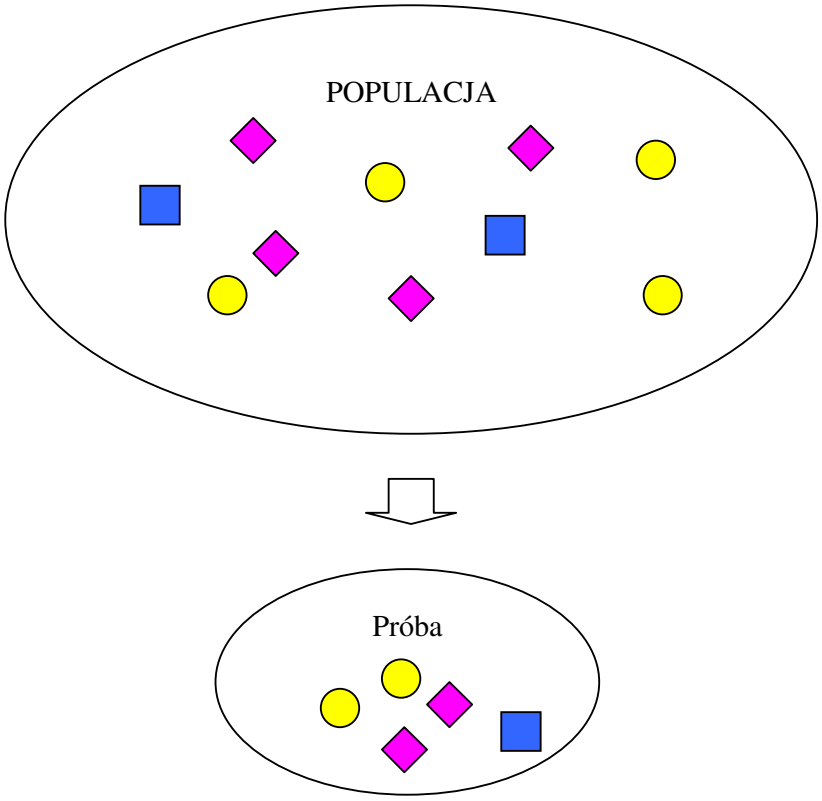
$$x_1, x_2, \ldots, x_n$$

wartości cechy X wylosowanych elementów z populacji to tzw. *próba losowa n-elementowa,* zaś n nazywamy *liczebnością próby.*

Próba reprezentatywna jest to próba, w której struktura cechy X mało różni się od struktury tej cechy w populacji – patrz rys. 1.3. Inny słowy średni poziom cech elementów próby powinien być taki sam jak w populacji.

Aby próba była reprezentatywna powinna być dostatecznie liczna i elementy populacji powinny być w odpowiedni sposób losowane.

Analizowane próby mogą być uzyskane z tych samych elementów - nazywane są one w tym wypadku **próbami powiązanymi**. Przy uzyskaniu prób z różnych elementów nazywane są one **próbami niepowiązanymi**.



Rysunek 1.3. Idea reprezentatywności próby

Przykład 1.4

Załóżmy, że chcemy przeprowadzić badania ankietowe studentów posiadających zaległości egzaminacyjne, dotyczące np. przyczyn powstawania takich zaległości. Ustaliliśmy wcześniej, że istotne będą następujące cechy populacji: wiek, płeć, stan cywilny, liczba zaległych egzaminów, wydział, rodzaj studiów, semestr. Znając procentowy rozkład dla każdej z tych cech można wygenerować próbę reprezentatywną. Pokażemy to na przykładzie dwóch cech.

Załóżmy, że w populacji występują następujące rozkłady:

Wydział	Liczba studentów z zaległymi egzaminami				Razem	Razem %
	1	2	3	4 i więcej		
Wydział A	194	33	31	27	285	28,5
Wydział B	180	43	10	45	278	27,8
Wydział C	251	65	45	76	437	43,7
Razem	625	141	86	148	1000	
Razem %	62,5	14,1	**8,6**	14,8		100

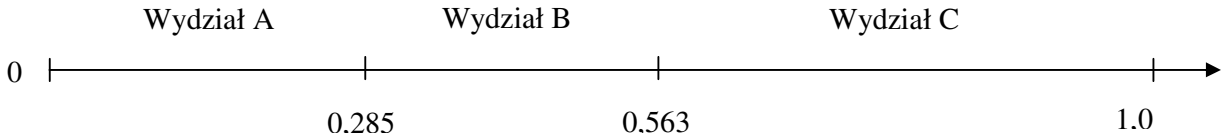
Podamy teraz algorytm wyboru próby:

1. Określamy wydział zgodnie z rozkładem prawdopodobieństwa:

$$p_1=285/1000,\ p_2=278/1000,\ p_3=437/1000$$

W tym celu generujemy liczbę losową zgodnie z rozkładem równomiernym w przedziale $< 0 ; 1>$.

W zależności od wielkości tej liczby określamy wydział, z którego będzie losowany student. Ilustruje to poniższy rysunek



Rysunek 1.4. Dane do losowania wydziału

Jeżeli wylosowana liczba jest $\leq 0{,}285$ to student będzie losowany z wydziału A, jeżeli wylosowana liczba jest zawarta między 0,285, a 0,563, to student będzie losowany z wydziału B, natomiast jeśli wylosowana liczba jest $\geq 0{,}563$, z wydziału C.

2. Generujemy drugą liczbę losową (oznaczmy ją a) zgodnie z rozkładem równomiernym w przedziale $< 0 ; 1>$.

   Przypuśćmy, że w pierwszym kroku ustalono, że student będzie losowany z wydziału A. Wtedy, jeśli $a \leq 194/285$, to losujemy studenta spośród studentów mających jeden zaległy egzamin, jeśli $194/285 < a \leq (194 + 33)/285$, to losujemy studenta spośród studentów mających dwa zaległe egzaminy itd.

   Podobnie postępujemy, gdy w pierwszym kroku ustalono, że studenta należy losować z wydziału B lub z wydziału C. ■

1.4. Zakres przedmiotu

W przedmiocie występują 4 części:

1) Statystyka opisowa
2) Rachunek prawdopodobieństwa
3) Statystyka matematyczna
4) Przykłady zastosowań rachunku prawdopodobieństwa i statystyki w informatyce.

Statystyka jest nauką zajmującą się tworzeniem metod badania rozkładu cechy populacji lub kilku cech populacji oraz badaniem współzależności tych cech na podstawie próby.

Umownie statystyka dzieli się na statystykę opisową i statystykę matematyczną.

Statystyka opisowa to dział statystyki zajmujący się metodami opisu danych statystycznych uzyskanych podczas badania statystycznego – analizie podlegają wszystkie elementy populacji bądź próby.

Celem stosowania metod statystyki opisowej jest podsumowanie zbioru danych i wyciągnięcie pewnych podstawowych wniosków i uogólnień na temat zbioru.

Statystyka opisowa podaje metody badania rozkładu cechy lub cech populacji bez metod wnioskowania statystycznego. W ramach statystyki opisowej następuje prezentacja (tabelaryczna i graficzna) rozkładu cechy.

W ramach statystyki opisowej następuje wstępne opracowanie wyników pomiarów (próby) bez posługiwania się pojęciami rachunku prawdopodobieństwa. Jeśli badana jest próba to nie są wyciągane wnioski dotyczące populacji generalnej.

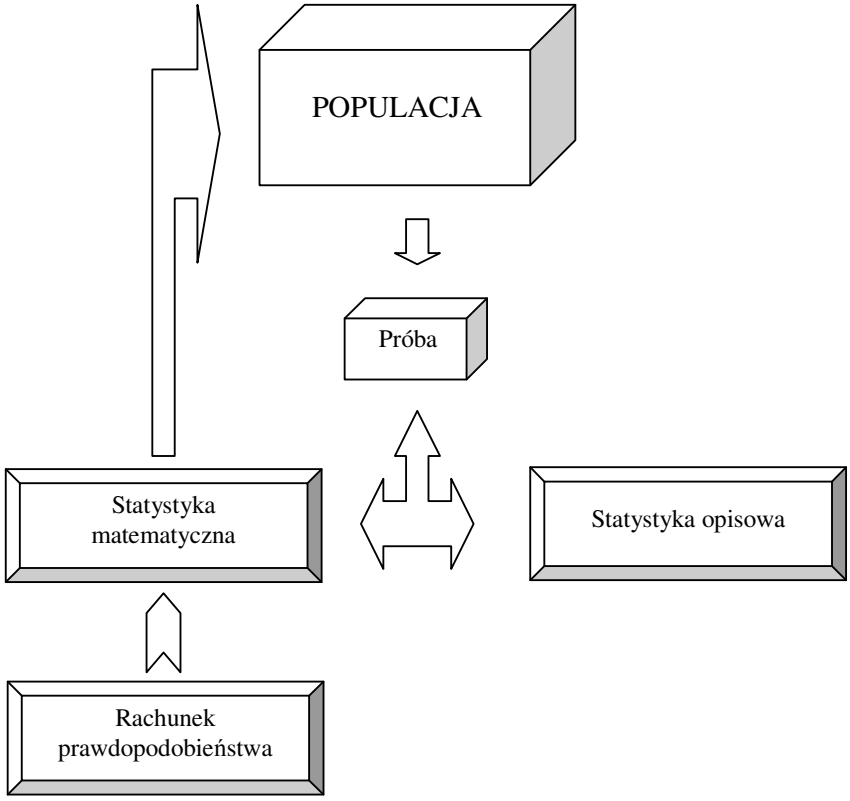
Rachunek prawdopodobieństwa to dział matematyki, zwany także Teorią prawdopodobieństwa, czy też Probabilistyką, zajmujący się tzw. zdarzeniami losowymi lub inaczej badaniem zjawisk losowych. Rachunek prawdopodobieństwa zajmuje się badaniem abstrakcyjnych pojęć matematycznych stworzonych do opisu zjawisk, które nie są

deterministyczne: zmiennych losowych w przypadku pojedynczych zdarzeń oraz procesów stochastycznych w przypadku zdarzeń powtarzających się (w czasie).

Statystyka matematyczna to dział statystyki, używający teorii prawdopodobieństwa i innych działów matematyki. Zajmuje się metodami wnioskowania statystycznego, które polegają na tym, że na podstawie wyników uzyskanych z próby formułujemy wnioski o całej zbiorowości.

Przyjmuje się, że modele badanych cech populacji są zmiennymi losowymi. Statystyka matematyczna zajmuje się budowaniem i wykorzystywaniem reguł wnioskowania statystycznego. *Wnioskowanie statystyczne* jest to wnioskowanie o rozkładzie cechy populacji lub kilku cech oraz o ich współzależności na podstawie próby.

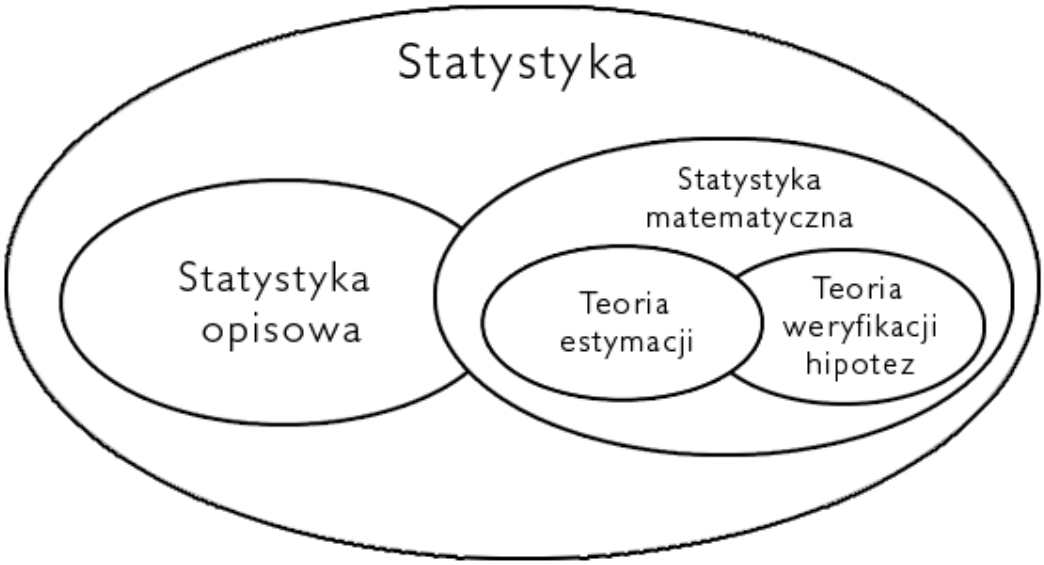
Statystykę matematyczną można umownie podzielić na dwa podstawowe działy: **teorię estymacji i teorię weryfikacji hipotez.** Umowność podziału wynika z faktu, że przy rozwiązywaniu konkretnych problemów z reguły wykorzystuje się łącznie metody z obu w/w działów.



Rysunek 1.5. Zakres przedmiotu

W ramach statystyki opisowej podano szereg charakterystyk liczbowych danych statystycznych o postaciach wynikających ze „zdrowego rozsądku". Określały one rozkład analizowanych elementów populacji czy też próby pobranej z populacji – bez żadnych uogólnień na populację.

Z kolei w przypadku estymacji, prowadzonej w ramach statystyki matematycznej, oszacowania na podstawie próby są uogólniane na populację i stąd w naturalny sposób pojawia się pytanie o dokładność takiego uogólniania.



Rysunek 1.6. Zakres statystyki

1.5. Analiza statystyczna z wykorzystaniem arkusza Excel[3]

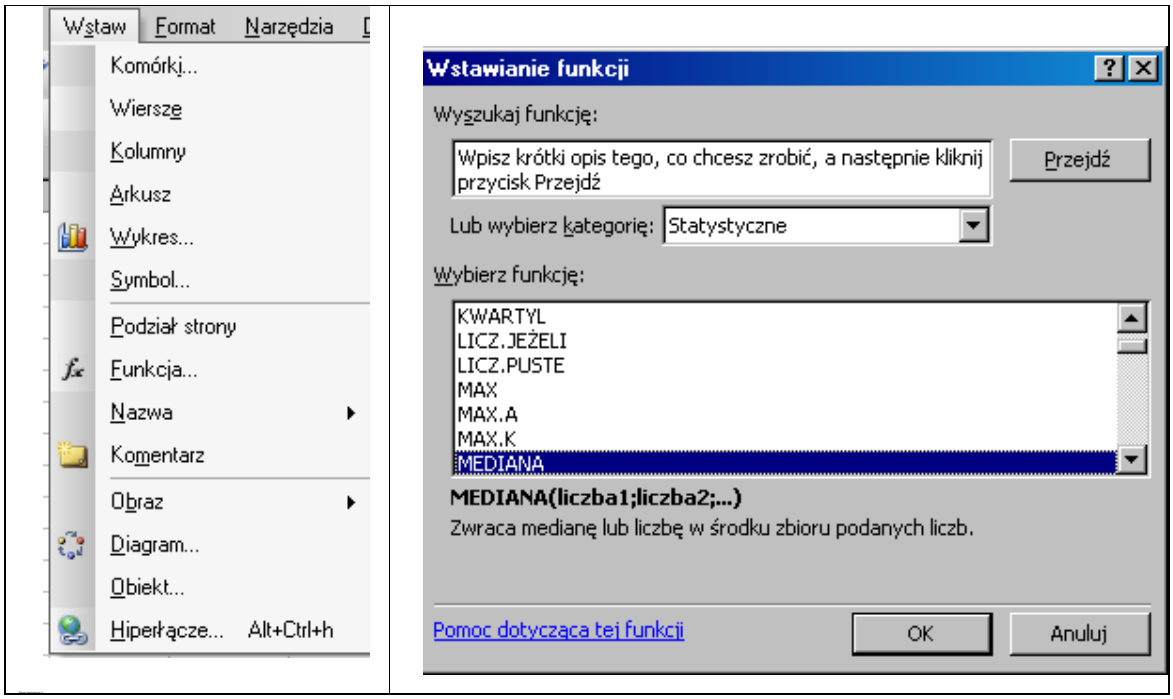
1.5.1. Uwagi wstępne

Analiza może być wykonana na trzy sposoby:

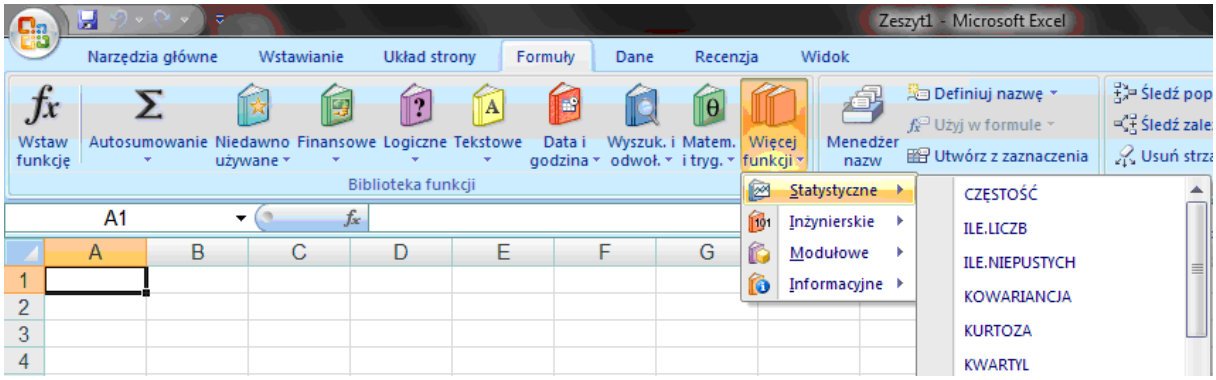
1. Utworzenie formuły obliczeniowej na podstawie operatorów: dodawania, odejmowania, mnożenia, dzielenia i potęgowania.
2. Utworzenie formuły obliczeniowej z wykorzystaniem dostępnych funkcji statystycznych i matematycznych.
3. Wykorzystanie dostępnych narzędzi statystycznych zgrupowanych w pakiecie do pakietu Analysis ToolPak.

1.5.2. Funkcje statystyczne

W pakiecie Excel dostępnych jest wiele funkcji statystycznych i matematycznych. Dostępne są one w menu Excela 2003 – pozycja **Wstaw**, po czym wybiera się kategorię **Statystyczne** lub **Matematyczne** i z tej kategorii potrzebną funkcję.



W Excelu 2007 funkcje dostępne są w sposób przedstawiony poniżej.



[3] Zamieszczone informacje pochodzą z Helpów arkusza Excwl

Tabela 1. WYKAZ FUNKCJI STATYSTYCZNYCH

Lp	Formuła	Przeznaczenie
1.	CZĘSTOŚĆ	Zwraca rozkład częstotliwości jako tablicę pionową
2.	ILE.LICZB	Zlicza liczby znajdujące się na liście argumentów
3.	ILE.NIEPUSTYCH	Zlicza wartości znajdujące się na liście argumentów
4.	KOWARIANCJA	Zwraca kowariancję, czyli średnią wartość iloczynów odpowiednich odchyleń
5.	KURTOZA	Zwraca kurtozę zbioru danych
6.	KWARTYL	Wyznacza kwartyl zbioru danych
7.	LICZ.PUSTE	Zwraca liczbę pustych komórek w pewnym zakresie
8.	MAX	Zwraca maksymalną wartość listy argumentów
9.	MAX.A	Zwraca maksymalną wartość listy argumentów z uwzględnieniem liczb, tekstów i wartości logicznych
10.	MAX.K	Zwraca k-tą największą wartość ze zbioru danych
11.	MEDIANA	Zwraca medianę podanych liczb
12.	MIN	Zwraca minimalną wartość listy argumentów
13.	MIN.A	Zwraca najmniejszą wartość listy argumentów z uwzględnieniem liczb, tekstów i wartości logicznych
14.	MIN.K	Zwraca k-tą najmniejszą wartość ze zbioru danych
15.	NACHYLENIE	Zwraca nachylenie linii regresji liniowej
16.	NORMALIZUJ	Zwraca wartość znormalizowaną
17.	ODCH.KWADRATOWE	Zwraca sumę kwadratów odchyleń
18.	ODCH.STANDARD.POPUL	Oblicza odchylenie standardowe na podstawie całej populacji
19.	ODCH.STANDARD.POPUL.A	Oblicza odchylenie standardowe na podstawie całej populacji z uwzględnieniem liczb, teksów i wartości logicznych
20.	ODCH.STANDARDOWE	Szacuje odchylenie standardowe na podstawie próbki
21.	ODCH.STANDARDOWE.A	Szacuje odchylenie standardowe na podstawie próbki z uwzględnieniem liczb, tekstów i wartości logicznych
22.	ODCH.ŚREDNIE	Zwraca średnią wartość odchyleń absolutnych punktów danych od ich wartości średniej
23.	ODCIĘTA	Zwraca punkt przecięcia osi pionowej z linią regresji liniowej
24.	PEARSON	Zwraca współczynnik korelacji momentu iloczynu Pearsona
25.	PERCENTYL	Wyznacza k-ty percentyl wartości w zakresie
26.	PERMUTACJE	Zwraca liczbę permutacji dla danej liczby obiektów

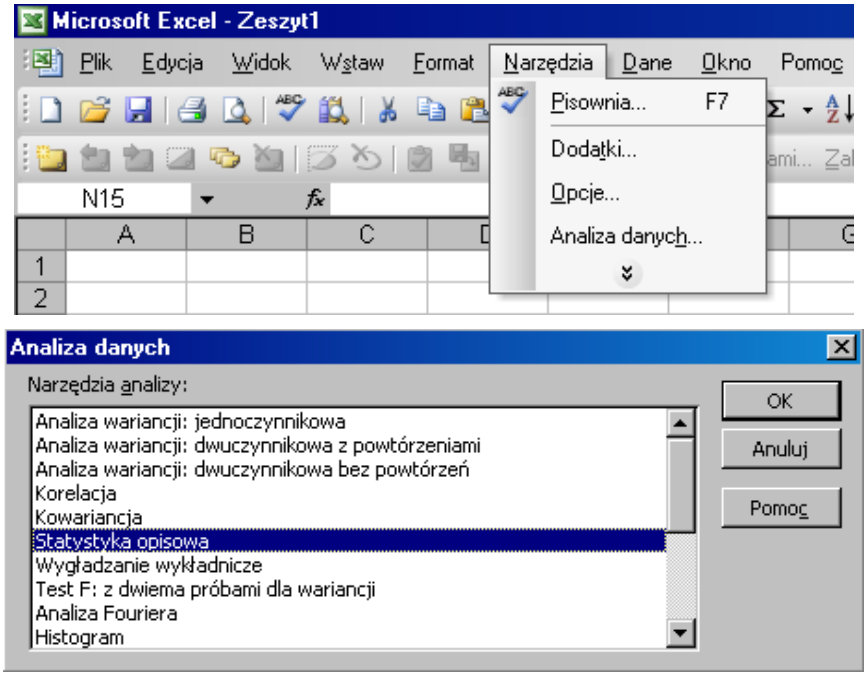
Lp	Formuła	Przeznaczenie
27.	POZYCJA	Zwraca pozycję liczby na liście liczb
28.	PRAWDPD	Zwraca prawdopodobieństwo, że wartości w zakresie leżą pomiędzy dwoma ograniczeniami
29.	PROCENT.POZYCJA	Zwraca procentową pozycję wartości w zbiorze danych
30.	PRÓG.ROZKŁAD.DWUM	Zwraca najmniejszą wartość, dla której skumulowany rozkład dwumianowy jest mniejszy lub równy wartości kryterium
31.	R.KWADRAT	Zwraca kwadrat współczynnika korelacji momentu iloczynu Pearsona
32.	REGBŁSTD	Zwraca błąd standardowy prognozowanej wartości y dla każdego x w regresji
33.	REGEXPP	Zwraca parametry trendu wykładniczego
34.	REGEXPW	Zwraca wartości trendu wykładniczego
35.	REGLINP	Oblicza statystykę dla linii, korzystając z metody najmniejszych kwadratów do obliczania linii prostej, która najlepiej pasuje do danych i zwraca tablicę opisującą tę linię
36.	REGLINW	Zwraca wartości trendu liniowego
37.	REGLINX	Zwraca wartość trendu liniowego
38.	ROZKŁAD.BETA	Zwraca skumulowaną funkcję gęstości prawdopodobieństwa beta
39.	ROZKŁAD.BETA.ODW	Zwraca odwrotność skumulowanej funkcji gęstości prawdopodobieństwa beta
40.	ROZKŁAD.CHI	Zwraca wartość prawdopodobieństwa z jednym śladem dla rozkładu chi-kwadrat
41.	ROZKŁAD.CHI.ODW	Zwraca odwrotność wartości prawdopodobieństwa z jednym śladem dla rozkładu chi-kwadrat
42.	ROZKŁAD.DWUM	Zwraca pojedynczy człon dwumianowego rozkładu prawdopodobieństwa
43.	ROZKŁAD.DWUM.PRZEC	Zwraca ujemny rozkład dwumianowy
44.	ROZKŁAD.EXP	Zwraca rozkład wykładniczy
45.	ROZKŁAD.F	Zwraca rozkład prawdopodobieństwa F
46.	ROZKŁAD.F.ODW	Zwraca odwrotność rozkładu prawdopodobieństwa F
47.	ROZKŁAD.FISHER	Zwraca transformację Fishera
48.	ROZKŁAD.FISHER.ODW	Zwraca odwrotność transformacji Fishera
49.	ROZKŁAD.GAMMA	Zwraca rozkład gamma
50.	ROZKŁAD.GAMMA.ODW	Zwraca odwrotność skumulowanego rozkładu gamma
51.	ROZKŁAD.HIPERGEOM	Zwraca rozkład hipergeometryczny
52.	ROZKŁAD.LIN.GAMMA	Zwraca logarytm naturalny funkcji gamma, Γ(x)
53.	ROZKŁAD.LOG	Zwraca skumulowany rozkład logarytmu naturalnego

Lp	Formuła	Przeznaczenie
54.	ROZKŁAD.LOG.ODW	Zwraca odwrotność rozkładu logarytmu naturalnego
55.	ROZKŁAD.NORMALNY	Zwraca rozkład normalny skumulowany
56.	ROZKŁAD.NORMALNY.ODW	Zwraca odwrotność rozkładu normalnego skumulowanego
57.	ROZKŁAD.NORMALNY.S	Zwraca standardowy rozkład normalny skumulowany
58.	ROZKŁAD.NORMALNY.S.ODW	Zwraca odwrotność standardowego rozkładu normalnego skumulowanego
59.	ROZKŁAD.POISSON	Zwraca rozkład Poissona
60.	ROZKŁAD.T	Zwraca rozkład t Studenta.
61.	ROZKŁAD.T.ODW	Zwraca odwrotność rozkładu t Studenta
62.	ROZKŁAD.WEIBULL	Zwraca rozkład Weibulla
63.	SKOŚNOŚĆ	Zwraca skośność rozkładu
64.	ŚREDNIA	Zwraca wartość średnią argumentów
65.	ŚREDNIA.A	Zwraca wartość średnią argumentów z uwzględnieniem liczb, tekstów i wartości logicznych
66.	ŚREDNIA.GEOMETRYCZNA	Zwraca średnią geometryczną
67.	ŚREDNIA.HARMONICZNA	Zwraca średnią harmoniczną
68.	ŚREDNIA.WEWN	Zwraca średnią wartość dla wnętrza zbioru danych
69.	TEST.CHI	Zwraca test niezależności
70.	TEST.F	Zwraca wynik testu F
71.	TEST.T	Zwraca prawdopodobieństwo związane z testem t Studenta
72.	TEST.Z	Zwraca wartość prawdopodobieństwa o jednym śladzie dla testu z
73.	UFNOŚĆ	Zwraca interwał ufności dla średniej populacji
74.	WARIANCJA	Szacuje wariancję na podstawie próbki
75.	WARIANCJA.A	Szacuje wariancję na podstawie próbki z uwzględnieniem liczb, tekstów i wartości logicznych
76.	WARIANCJA.POPUL	Oblicza wariancję na podstawie całej populacji
77.	WARIANCJA.POPUL.A	Oblicza wariancję na podstawie całej populacji, z uwzględnieniem liczb, tekstów i wartości logicznych
78.	WSP.KORELACJI	Zwraca współczynnik korelacji dwóch zbiorów danych
79.	WYST.NAJCZĘŚCIEJ	Zwraca wartość najczęściej występującą w zbiorze danych
80.	ZLICZ.JEŻELI	Zlicza liczbę niepustych komórek w zakresie zgodnych z podanym kryterium

Każda z funkcji statystycznych posiada Help. W przykładach zamieszczonych w podręczniku zademonstrowano wykorzystywanie wybranych funkcji statystycznych..

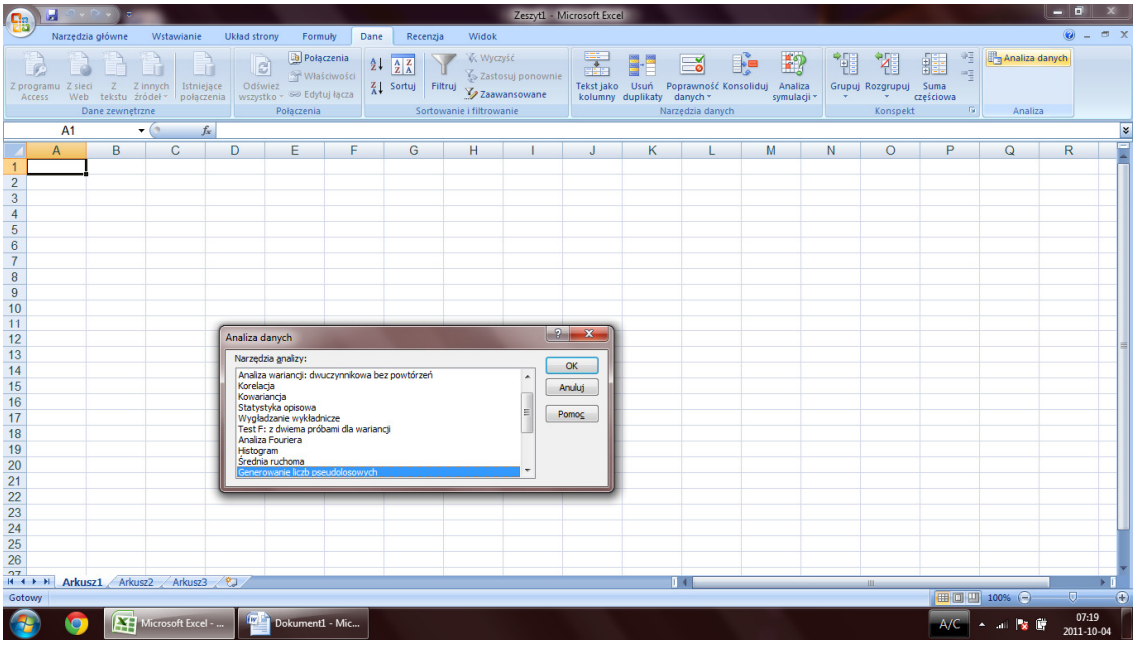
1.5.3. Pakiet Analysis ToolPak

Wykorzystana powyżej narzędzie analizy STATYSTYKA OPISOWA jest jednym z wielu narzędzi statystycznych zgrupowanych w pakiecie do pakietu Analysis ToolPak. Dostępne są one w menu Excela 2003 – pozycja **Narzędzia**, po czym wybiera się kategorię **Analiza danych** i z tej kategorii potrzebne narzędzie, np. **Statystyka opisowa**.

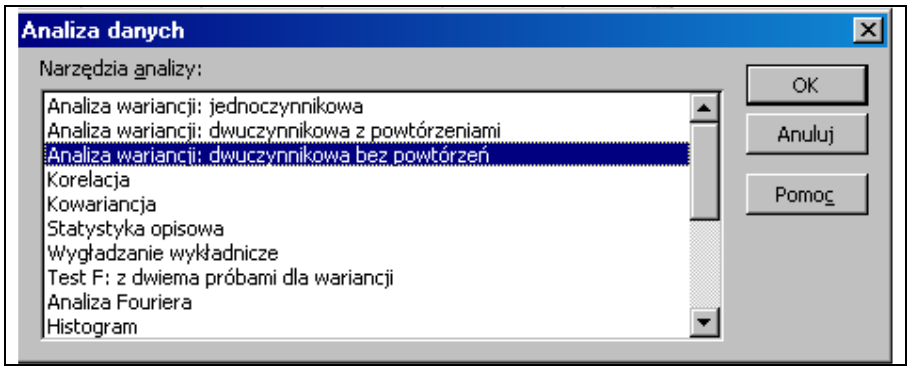


W Excelu 2007 aby uzyskać dostęp do pakietu Analysis ToolPak należy kliknąć przycisk **Analiza danych** w grupie **Analiza** na karcie **Dane,** a następnie wybrać potrzebne narzędzi, **np. Statystyka opisowa.**

.



Gdy moduł jest niedostępny, załaduj dodatek (dodatek: Program uzupełniający, który dodaje niestandardowe polecenia lub funkcje do pakietu Microsoft Office) Analysis ToolPak.



WYKAZ NARZĘDZI STATYSTYCZNYCH

1. ANOVA
2. ANOVA: POJEDYNCZY CZYNNIK
3. ANOVA: DWA CZYNNIKI Z REPLIKACJĄ
4. ANOVA: DWA CZYNNIKI BEZ REPLIKACJI
5. KORELACJA
6. KOWARIANCJA
7. STATYSTYKI OPISOWE
8. WYGŁADZANIE WYKŁADNICZE
9. TEST F: DWIE PRÓBKI DLA WARIANCJI
10. ANALIZA FOURIERA
11. HISTOGRAM
12. ŚREDNIA RUCHOMA
13. GENEROWANIE LICZB LOSOWYCH
14. RANGA I PERCENTYL
15. REGRESJA
16. PRÓBKOWANIE
17. TEST T
18. TEST T: DWIE PRÓBY, PRZY ZAŁOŻENIU RÓWNYCH WARIANCJI
19. TEST T: DWIE PRÓBY, PRZY ZAŁOŻENIU NIERÓWNYCH WARIANCJI
20. TEST T: SPAROWANY, DWIE PRÓBY DLA ŚREDNICH

Każda z narzędzi statystycznych posiada Help. W przykładach zamieszczonych w podręczniku zademonstrowano wykorzystywanie wybranych narzędzi statystycznych.

1.5.4. Wykorzystywane piśmiennictwo

[1] Joanna Kisielińska, Urszula Skórnik-Pokarowska: Podstawy statystyki z przykładami w Excelu, Wydawnictwo SGGW, Warszawa 2005
[2] Grzegorz Kończak, Grażyna Trzpiot: Analizy statystyczne z arkuszem kalkulacyjnym Microsoft EXCEL, Wydawnicwo Akademii Ekonomicznej w Katowicach, Katowice 2002
[3] Mirosława Kopertowska, Witold Sikorski: Funkcje w EXCELU w praktyce, Wydawnictwo Naukowe PWN, Warszawa 2006
[4] Maria Parlińska, Jacek Parliński: Badania statystyczne z Excelem, Wydawnictwo SGGW, Warszawa 2007
[5] Wiesława Regel: Podstawy Statystyki w Excelu, Wydawnictwo MIKOM, Warszawa 2003
[6] Agnieszka Snarska: Statystyka, Ekonometria Prognozowanie Ćwiczenia z Excelem, Wydawnictwo Placet, Warszawa 2007