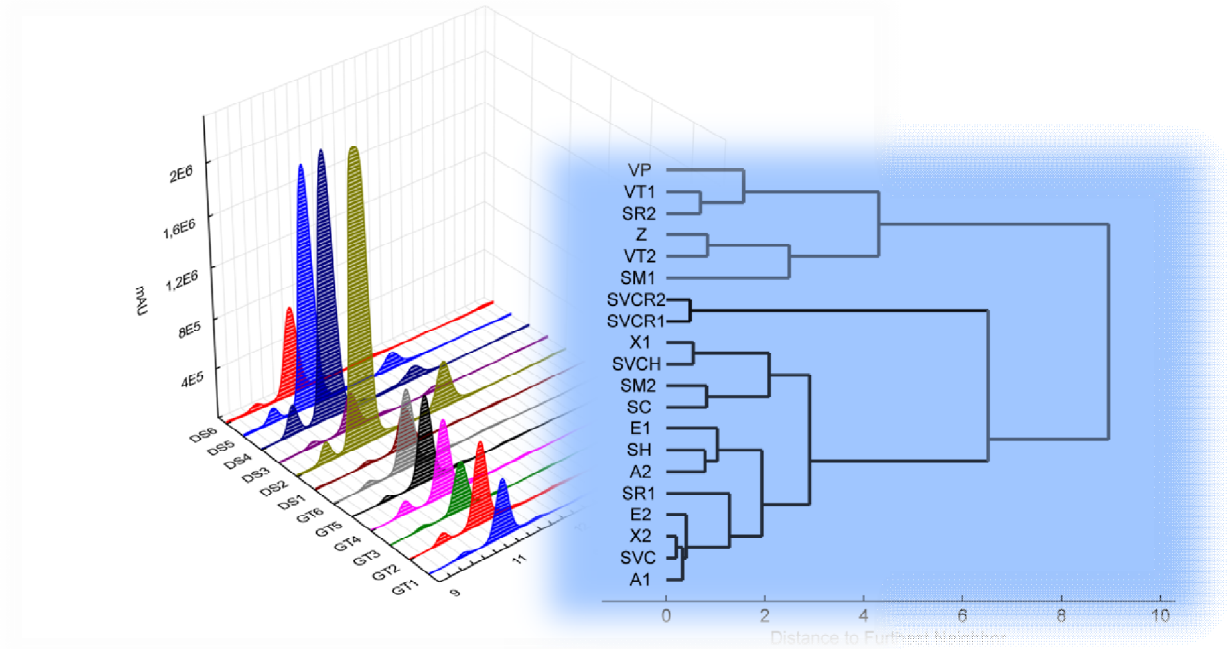
KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki

Materiały do wykładu

Statystyczne i chemometryczne metody analizy danych w chemii medycznej i biologii



dr hab. Michał Marszałł
dr Bogumiła Kupcewicz

T O R U Ń 2013

SPIS TREŚCI

1. Wprowadzenie

Chemia medyczna jest dyscypliną badawczą znajdującą się na pograniczu chemii, biologii i farmakologii, której głównym celem jest projektowanie, synteza i badanie właściwości związków chemicznych o potencjalnej aktywności farmakologicznej. Eksperyment badawczy, niezależnie od dziedziny naukowej, wiąże się z koniecznością jego zaplanowania i wykonania, a następnie analizą uzyskanych wyników i wnioskowaniem. Do poprawnej realizacji każdego z tych elementów przydatne są statystyczne i chemometryczne metody analizy danych. Celem wykładu jest przedstawienie możliwości jakie dają metody statystyczne i chemometria zarówno w planowaniu eksperymentu jak i analizie uzyskanych wyników. Wykład ma charakter interdyscyplinarny, przeznaczony jest głównie dla doktorantów realizujących badania obejmujące zagadnienia z zakresu nauk biologicznych, chemicznych, medycznych czy farmaceutycznych.
Wybór odpowiedniej metody opracowania danych wymaga zarówno merytorycznej znajomości badanych zjawisk jak i doświadczenia w stosowaniu różnych metod statystycznych lub chociaż świadomości ich możliwości i ograniczeń.
W ramach wykładu zostaną omówione aspekty teoretyczne i praktyczne poszczególnych etapów analizy danych od ich wstępnej eksploracji, przez dobór odpowiednich metod (testów statystycznych lub technik chemometrycznych) po merytoryczną interpretację otrzymanych wyników. Bardzo istotnym zagadnieniem jest znajomość kryteriów, które decydują o wyborze odpowiednich testów (parametrycznych i nieparametrycznych) oraz etapów weryfikacji hipotez statystycznych.
Szczególny nacisk położony jest na ilustrację każdego z omawianych zagadnień szeregiem odpowiednio dobranych przykładów, pozwalających na ich pełniejsze zrozumienie. Umiejętność doboru metod i samodzielnego przeprowadzenia statystycznej analizy danych lub przynajmniej jej świadome śledzenie wzbogaci warsztat metodologiczny doktoranta, co jest ważne zwłaszcza w kontekście badań realizowanych w ramach interdyscyplinarnej rozprawy doktorskiej. Dodatkowo zostaną przedstawione przykłady typowych błędów popełnianych w doborze, wykonaniu i interpretowaniu wyników metod statystycznych i chemometrycznych.

2. Elementy chemii medycznej

Projektowanie substancji leczniczych jest kluczowym etapem w poszukiwaniu nowych leków. Jednak zanim leków syntetycznych zaczęto używać jako leków, odkrycie leczniczego działania związków naturalnych było dziełem przypadku. W początkach cywilizacji źródłem „leków” były przeważnie produkty naturalne pochodzenia roślinnego lub zwierzęcego, stosowane wyłącznie na podstawie obserwacji skutków ich stosowania [1]. W przypadku odkrycia leczniczego działania np. konkretnej rośliny, opracowano różne sposoby sporządzanie surowców dla celów leczniczych np. odwary, nalewki lub maści [2].
Sposoby przyrządzania – receptury poszczególnych surowców zostały spisane przez rzymskiego lekarza – farmaceutę Galena (II w n.e.). Stąd leki wykonane z surowców roślinnych często określa się mianem leków galenowych.

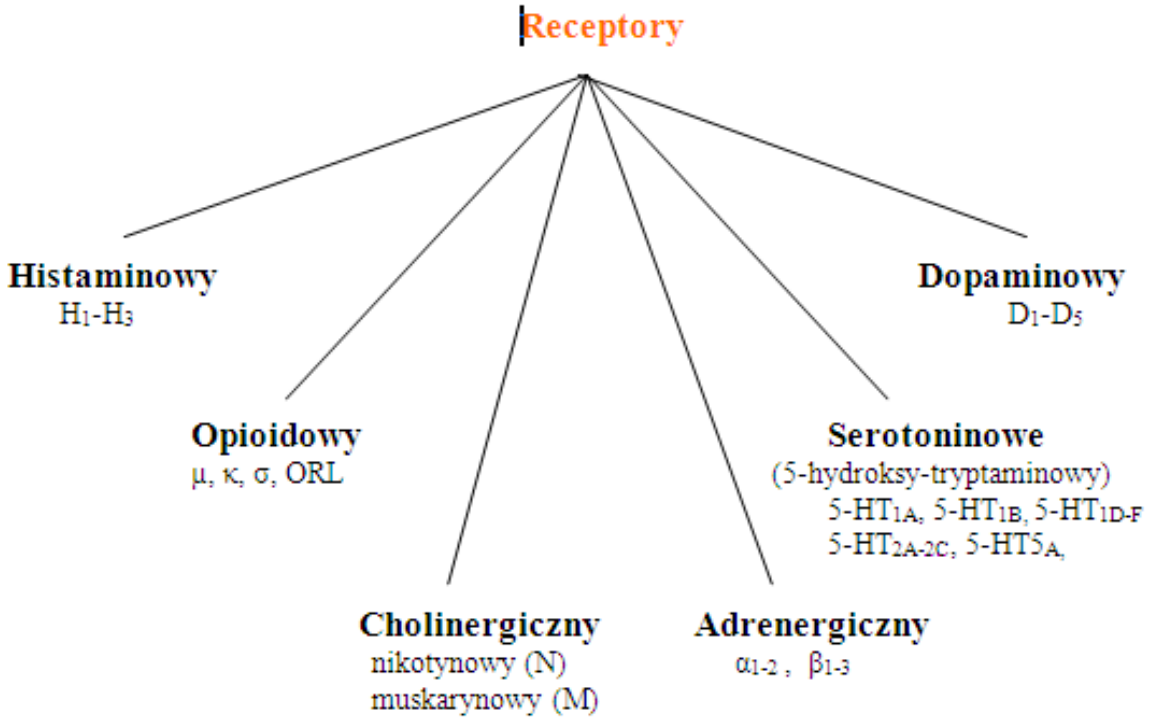
Kolejnym krokiem milowym był rozwój chemii a w szczególności jatrochemii – chemii lekarskiej, której rozwój przypisuje się Paracelsusowi na przełomie XV i XVI wieku.
Można śmiało stwierdzić, iż wysnuwając teorię, że rośliny lecznicze zawdzięczają swoje właściwości związkom leczniczym w nich zawartym, stworzył on podwaliny rozwoju chemii medycznej oraz szeregu dyscyplin naukowych jak chociażby chemia leków czy farmakognozja. Następne lata to znaczny postęp w pozyskiwaniu substancji leczniczych z surowców leczniczych. Dopiero przełom XIX i XX wieku przyniósł znaczny postęp w rozwój chemii organicznej, dzięki której opracowano szereg syntez związków aktywnych biologicznie, a część z nich stosowana jest do chwili obecnej. Początkowy okres ubiegłego wieku zaowocował min. odkryciem penicyliny, wprowadzeniem do lecznictwa barbitalu, sulfonamidów czy antybiotyków. Jednak za początek przemysłu farmaceutycznego uważa się izolację salicylanów, wprowadzenie oraz syntezę i produkcję kwasu acetylosalicylowego (rok 1897).

Naukowe podstawy farmakoterapii zostały sformułowane na początku XX wieku przez Paula Ehricha – niemieckiego chemika i lekarza, którego do dnia dzisiejszego uważa się za twórcę podstaw farmakoterapii. Na podstawie tej teorii wnioskowano, iż działanie leków polega na oddziaływaniu ze strukturami biologicznymi co w konsekwencji umożliwiło racjonalne poszukiwanie struktur chemicznych o pożądanym działaniu biologicznym [3]. Ponieważ komórka jest podstawową jednostką funkcjonowania organizmu, stąd stała się szczególnym celem działania leków. Ze względu na złożoną budowę komórki, tak naprawdę oddziałują w jej poszczególnych strukturach. Na poziomie molekularnym można powiedzieć że, są cztery główne miejsca działania leków: lipidy, węglowodany, białka oraz kwasy nukleinowe (DNA i RNA). Co prawda dla wielu z nich dokładna rola fizjologiczna nie jest jeszcze poznana, jednak znaczny rozwój biotechnologii oraz biologii molekularnej w zakresie badań *in vitro* przez ostatnie trzy dekady, jak również badania nad genomem człowieka pozwoliły na bardziej dynamiczny rozwój farmakoterapii.
Współczesna koncepcja projektowania nowych leków bazuje na określeniu miejsca działania leku (*drug target*), a następnie na poszukiwaniu substancji oddziałującej z tym miejscem. Poszukiwanie nowych leków, aż do ich zastosowania klinicznego jest procesem wieloetapowym i czasochłonnym. Głównym celem takiego postępowania jest zaprojektowanie związku o korzystniejszych właściwościach terapeutycznych niż struktura wiodąca (*lead compound*).
Najczęściej celem substancji leczniczych są białka, głównie receptory i enzymy. Specyficzność i selektywność leków do miejsca działania obejmującego enzymy, receptory

czy też różne podtypy receptora są czynnikami determinującymi ich skuteczność i bezpieczeństwo. Leki muszą być zdolne do specyficznego oddziaływania z kreślonymi cząsteczkami (target) nie zakłócając przy tym działania ważnych dla życia układów enzymatycznych. Uważa się, iż liczba potencjalnych celów wynosi około 25 000 [3].
Receptory, podobnie jak białka, cechują się zdolnością do wiązani ligandów z różną selektywnością i siłą – powinowactwem. Aktywność leku opiera się na specyficzności oddziaływania z białkami. Jeżeli powinowactwo i specyficzność jest zbyt mała to należy się liczyć z poważnymi konsekwencjami – działaniami niepożądanymi. Receptory najczęściej zbudowane są z wielu jednostek, których budowa przestrzenna i funkcje ulegają zmianie po przyłączeniu efektora allosterycznego. Wskutek oddziaływaniom z innymi cząsteczkami przekazują sygnał do wnętrza komórki. Receptory jako główne miejsce przekazywania informacji przez substancje sygnalizacyjne występują nie tylko w błonie komórkowej, ale również we wnętrzu komórki. Dlatego receptory dzieli się na związane z błoną komórkową (receptory związane z białkiem G, wykazujące aktywność kinazy tyrozynowej, serynowo-treoninowej i cyklazy guanylowej) oraz receptory jądrowe zlokalizowane w cytoplazmie lub jądrze komórkowym.
Nazewnictwo receptorów najczęściej pochodzi od typu lub nazwy specyficznego neuroprzekaźnika: np. receptor aktywowany przez adrenalinę lub noradrenalinę nazywamy receptorem adrenergicznym (tzw. adrenoreceptory), który z kolei na swoją lokalizację i niewielkie różnice strukturalne różnią się funkcją. Stąd oprócz głównych grup (typów) receptora wyróżniamy podtypy. Na Ryc. 1 podano przykładowe typy i podtypy receptorów. Wiele z podtypów receptorów zidentyfikowano stosunkowo niedawno a badania nad ich strukturą i aktywnością cały czas trwają. Dlatego w literaturze możemy spotkać niepełne informacje co do ich nazwy i mechanizmu działania, które często opierają się hipotezie i muszą być potwierdzone w badaniach zarówno *in vivo* jak i *in vitro*.
Poprzez powinowactwo leku do receptora należy rozumieć jego zdolność do oddziaływania poprzez „wiązanie” z nim. Jednak trzeba pamiętać aby działanie leku na określony receptor mogło wywołać określony efekt farmakologiczny, oprócz łączenia się z receptorem musi posiadać aktywność wewnętrzną. W tym celu podczas projektowania leków zarówno powinowactwo jak i aktywność wewnętrzna może być przewidziana wykorzystując różne modele matematyczne.

Ze względu na aktywność leków oddziałujących na receptor(y) można je podzielić na dwie grupy:

- **agoniści** – charakteryzują się dużą aktywnością wewnętrzną warunkująca pobudzenie receptora
- **antagoniści** – aktywność antagonisty jest równa 0, w konsekwencji blokuje receptor



Ryc. 1 Przykłady typów i podtypów receptorów.

Przykładem leków agonistów receptorów α-adrenergicznych są chociażby pochodne 2-imidazoliny np. ksylometazolina (Xylometazolin) czy nafazolina (Naphazolin). Leki te bezpośredni aktywując receptory α-adrenergiczne w naczyniach krwionośnych nosa, obkurczają je przez co zmniejsza się przekrwienie błony śluzowej i jednocześnie obrzęk nosa. Do często stosowanych leków – antagonistów możemy zaliczyć leki β-adrenolityczne tzw. betaadrenolityki”, hamujące czynność układu współczulnego. Przykładem są: propranolol i sotalol, które działają zarówno na receptory β_1 oraz β_2. Działając na receptory β obniżają ciśnienie śródgałkowe, hamują wydzielanie reniny i w konsekwencji obniżają ciśnienie, działają na mięśnie – mogą powodować skurcz oskrzeli, w obrębie naczyń krwionośnych – skurcz naczyń i zwiększenie oporu obwodowego. Jednak, bardziej pożądane są leki selektywnie działające na receptor β_2 w obrębie serca, którego pobudzenie powiduje zmniejszenie częstości akcji serca (efekt chronotropowy ujemny), spadek przewodnictwa przedsionkowo-komorowego (efekt dromotropowy ujemny), obniżenie kurczliwości mięśnia komór (efekt inotropowy ujemny) i zmniejszenie objętości wyrzutowej i minutowej serca, zmniejszenie zużycia tlenu przez mięsień sercowy.

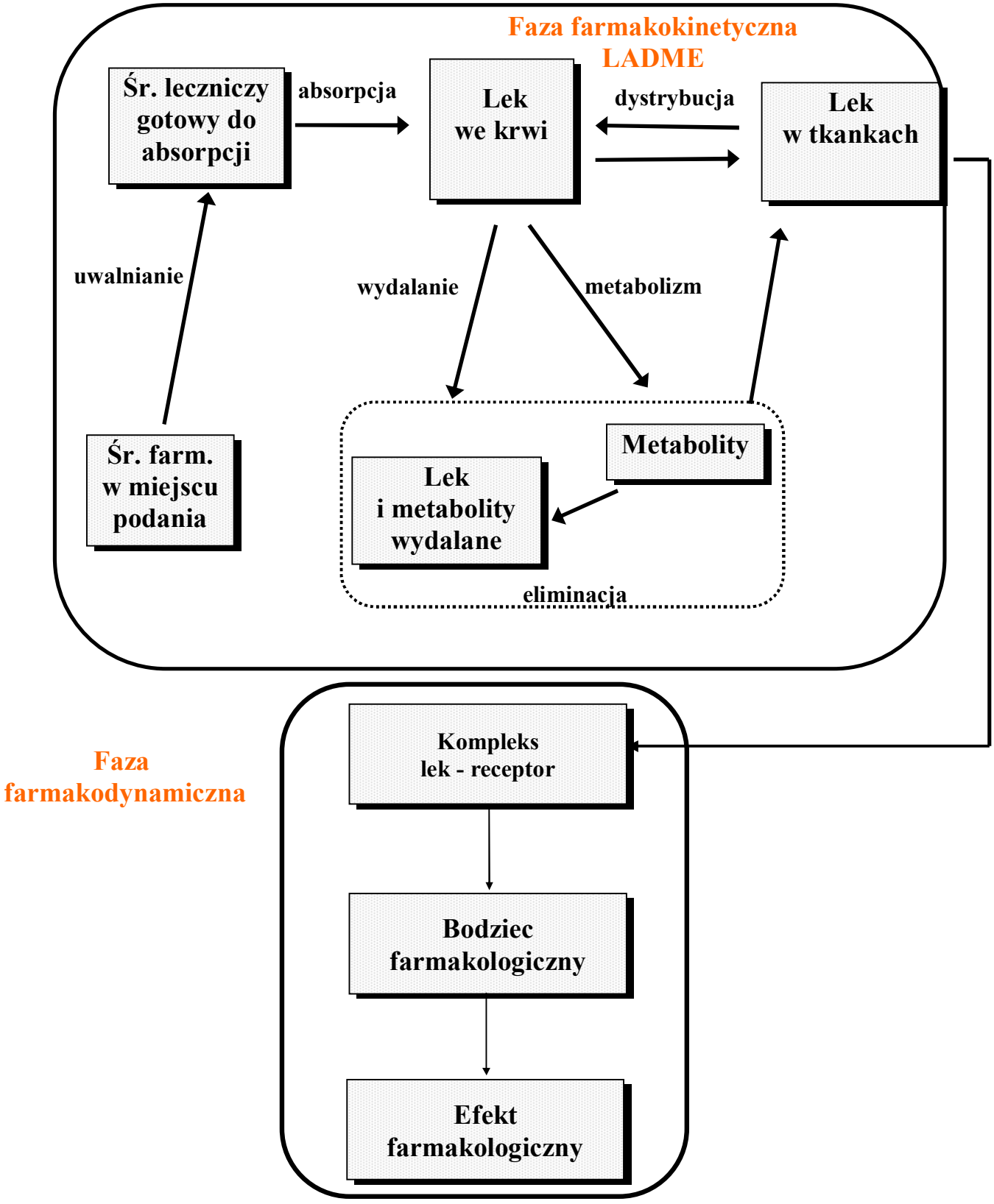
Oprócz receptorów, również ważnym miejscem działania leków są enzymy. Spełniają one w naszym organizmie funkcję katalizatorów – substancji obniżającej energię aktywacji tym samym zmieniającej kinetykę reakcji. Enzymy będące białkami o właściwościach biokatalizatorów regulują szereg istotnych procesów życiowych. Stąd stanowią ważny punkt uchwytu leków. Zdolność oddziaływania substratu (leku) z enzymem jest kluczowym

etapem w projektowaniu leków. Potencjalny lek powinien wiązać się z ze specyficznym miejscem-części enzymu zwanym miejscem aktywnym.
Na etapie projektowania leków, oprócz jego aktywności bierze się pod uwagę właściwości fizyko-chemiczne umożliwiające ich wiązanie poprzez wiązania jonowe, wodorowe oraz van der Waalsa. Dodatkowo, musi być zachowana zasada, iż wiązania takie muszą być na tyle zrównoważone, aby utrzymać lek w miejscu aktywnym, a z drugiej strony wiązanie takie powinno być odwracalne tzn. nietrwałe. W praktyce projektuje się leki strukturalnie podobne do naturalnego (endogennego) enzymu. W przypadku leków mających na celu hamowanie aktywności enzymów – są to inhibitory, które w przypadku współzawodnictwa o te samo miejsce wiążące co naturalny substrat nazywa się inhibitorami kompetycyjnymi. W przypadku kiedy inhibitory wiążą się z innymi obszarami enzymu niż miejsce aktywne mówimy wtedy o inhibitorach niekompetycyjnych. Najczęściej są to tzw. inhibitory allosteryczne, które w przypadku wiązania z enzymem zmieniają jego kształt uniemożliwiając połączenie się substancji z miejscem aktywnym.

3. Farmakokinetyka i farmakodynamika

Los leku w organizmie żywym możemy opisać za pomocą dwóch faz: kinetycznej i farmakodynamicznej. Farmakokinetyka zajmuje się przebiegiem w czasie szeregu procesów decydujących o losach leku oraz jego metabolitów tj. wchłanianie (absorpcja, *ang. absorbtion*, A), rozmieszczenie (dystrybucję, *ang. distribution*, D), metabolizm (*ang. metabolizm*, M) i wydalenie (eliminację, *ang. excretion*, E). Proces absorpcji ma miejsce w przypadku podania leku drogą pozanaczyniową (np. podanie doustne lub drogą wziewną), w której lek musi być wchłonięty do krwiobiegu aby mógł oddziaływać ogólnoustrojowo. Dopiero po wchłonięciu do krwi lek może „dotrzeć” do receptora i wykazać swoje działanie lecznicze. Wyjątkiem są leki o działaniu miejscowym (np. neutralizujące pH soku żołądkowego). Jeżeli leki mają postać stałą (np. tabletki lub kapsułki), muszą one ulec deformacji, aby zawarte w nich substancje lecznicze uległy uwolnieniu (*ang. liberation*, L). W odniesieniu do leków podanych donaczyniowo, pomija się fazę L i A, gdyż dociera on bezpośrednio do krwiobiegu. Stąd w literaturze często można spotkać określenie fazy farmakokinetycznej jako skrót LADME. Faza farmakodynamiczna opisuje wpływ leku na receptor i komórkę (Ryc. 2).

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki



Ryc. 2 Losy leku w ustroju: faza farmakokinetyczna i farmakodynamiczna

4. Struktura cząsteczki i właściwości biologiczne

Działanie leku polega na jego oddziaływaniu ze strukturami biologicznymi. Specyficzność oraz selektywność tych oddziaływań zależy od właściwości fizycznych i chemicznych danego związku – leku. Po przyjęciu leku następuje uwolnienie substancji a potem absorpcja i dystrybucja w tkankach. Procesy te wpływają również na inne jak: magazynowanie, metabolizowanie oraz wydalanie.
Właściwości fizyczne i chemiczne danego związku – leku wpływają również na końcowy proces – farmakodynamikę. Wchłanianie leków niskocząsteczkowych zależy głównie od ich polarności. Związki silnie polarne nie mogą przenikać przez błony komórkowe, więc nie są wchłanianie przez komórki. Polarność wpływa również na farmakokinetykę tych leków, gdyż po przyjmowanym doustnym leków związki te nie są wchłanianie w jelitach do krwiobiegu. Przykładem jest sorbitol, który stosowany doustnie pełni rolę leku przeczyszczającego. Jeżeli związki te podaje się dożylnie to są one wydalane przez nerki i mają działanie moczopędne. Wyjątkiem są substancje, które mogą być „przeniesione” w sposób bierny lub czynny przez białka transportowe przez barierę, która stanowi błona komórkowa.
Związki o nieznacznej biodostępności po podaniu doustnym mogą być jednak aktywne farmakologicznie [3]. Ich działanie może polegać na oddziaływaniu z pozakomórkowymi biocząsteczkami, z celem znajdującym się na powierzchni komórki lub po prostu działać miejscowo.
Wchłanianie leków do krwiobiegu jest uwarunkowane jego zdolnością do przenikania przez błony biologiczne. Leki działające w ośrodkowym układzie musza tez być zdolne do przenikania przez barierę krew-mózg. Powstaje ona po wewnętrznej stronie naczyń krwionośnych w mózgu z warstwy komórek śródbłonka uszczelnionych wyspecjalizowanymi strukturami białkowymi.

Lipofilowość

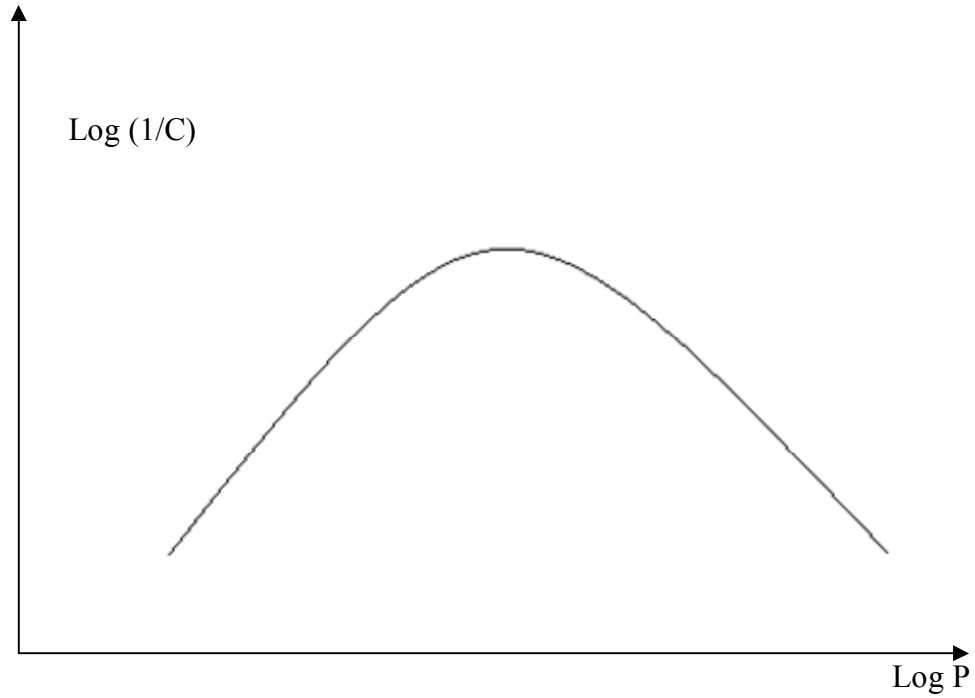
Odpowiednia lipofilowość jest warunkiem koniecznym do przenikania leku do struktur mózgu. Wyjątek stanowią substancje wchłanianie na drodze transportu aktywnego. Miarą litofilności substancji jest *współczynniki podziału* P (*ang*. partition coefficient). W praktyce oznacza on współczynnik podziału leku pomiędzy dwie fazy: 1-oktanolu i wody. Parametr ten można obliczyć za pomocą równania:

$$P = \frac{[lek]o}{[lek]w}$$

gdzie w oznacza fazę wodną, o – fazę oktanolową.

Szybkość większości etapów procesu farmakokinetyki małych cząsteczek w tkankach jest proporcjonalna do wartości logarytmu P i najczęściej opisuje się ją w skali log P.
Związki o wartości log P < 0 są lepiej rozpuszczalne w wodzie a substancje czynne wartości log P > 0 rozpuszczają się lepiej w 1-oktanolu. Można więc łatwo stwierdzić, iż wraz ze wzrostem lipofilowości związku zwiększa się jego aktywność biologiczna, co jest

odzwierciedleniem zdolności związku do pokonywania błon komórkowych. Nawet podczas łączenia się z enzymami czy receptorami, zwykle ma to miejsce poprzez struktury hydrofobowe. Tak więc wzrost lipofilowości ułatwi jego przenikanie przez bariery biologiczne oraz zwiększa jego siłę wiązania w miejscu działania
Na podstawie tego można wnioskować, iż zwiększenie wartości parametru log P prowadzi do nieograniczonego wzrostu aktywności biologicznej. Niestety nie, gdyż lek zbyt silnie lipofilny jest słabo rozpuszczalny w fazie wodnej i dodatkowo może być wychwycony przez tkankę tłuszczową. W konsekwencji może nigdy nie dotrzeć do miejsca działania.



Ryc. 3 Zależność aktywności od parametru log P

5. Projektowanie leków

Głównym etapem jest zaprojektowanie nowego związku-potencjalnego leku o korzystniejszych właściwościach terapeutycznych niż struktura wiodąca. Etap ten ma na celu uzyskanie potencjalnych leków o dużym indeksie (współczynniku) terapeutycznym tzn. o zwiększonej aktywności biologicznej przy jednoczesnym obniżeniu działań niepożądanych. Znanych jest szereg metod i rodzajów modyfikacji (ingerencji) struktury wiodącej celem uzyskania optymalnej struktury. Do najczęstszych należą:

- wymiana podstawników,
- zmiana rozmiaru cząsteczki,
- kondensacja oraz wymiana pierścieni,
- modyfikacja łańcucha bocznego.

W celu ułatwienia projektowania leków o pożądanej farmakokinetyce powstała tzw. reguła Lipińskiego (inaczej „Reguła pięciu”). Według tej reguły cząsteczka spełniająca poniższe kryteria ma szanse stać się potencjalnym lekiem o określonej aktywności:

- masa molowa **poniżej** 500 Da,
- wartość log P jest **mniejsza** niż 5,
- liczba donorów protonów jest **mniejsza** niż 5,
- liczba akceptorów jest wielokrotnością pięciu i jest **mniejsza** niż 10.

Wyjątkiem od tej reguły są leki, które charakteryzują się zdolnością do transportu aktywnego.

Znaczący krok w projektowaniu nowych leków dała chemia kombinatoryczna. Rozwój chemii kombinatorycznej dał początek metodologii określanej jako wysoko wydajne badania przesiewowe (*high throughput screening,* HTS). Ogólnie mówiąc, mianem HTS określa się każdą technikę, która pozwala na szybkie „wyizolowanie” związków o dużym powinowactwie do danego enzymu lub receptora kwalifikując je jako potencjalne leki. Współczesna bioanalityka dysponuje technikami pozwalającymi na dość precyzyjne oszacowanie powinowactwa leków do biomakromolekuł. W większości przypadków są to narzędzia, których zastosowanie wiąże się z użyciem dużej ilości odczynników oraz fluorescencyjnych i radioaktywnych znaczników, czy też wysoki koszt, ogranicza jej dostępność w jednostkach naukowych. Dlatego wyzwaniem dla współczesnych technik HTS jest stworzenie takiego narzędzia, które umożliwiłoby badania nie tylko na poziomie lek-białko, ale również badać oddziaływania typu antygen-przeciwciało, białko-białko, białko-kwas nukleinowy również w aspekcie nowych oddziaływań. Techniki HTS są często dodatkowym elementem w procesie poszukiwania leków do tradycyjnie stosowanych testów biologicznych *in vivo* i *in vitro*.
Do współczesnych metod HTS zaliczamy m. in. badania z wykorzystaniem magnetycznego rezonansu jądrowego, powierzchniowy rezonans plazmonowy (SPR) czy chromatografii biopowinowactwa. Wszystkie te techniki wspomagane są chemometrią, przy wykorzystaniu komputerów i odpowiedniego oprogramowania, które na przykład mogą posłużyć do obliczenia energii poszczególnych konformacji celem wyznaczenia minimum energetycznego danej struktury a następnie jej właściwości fizycznych i chemicznych.

6. Ilościowa zależność struktura-aktywność (QSAR)

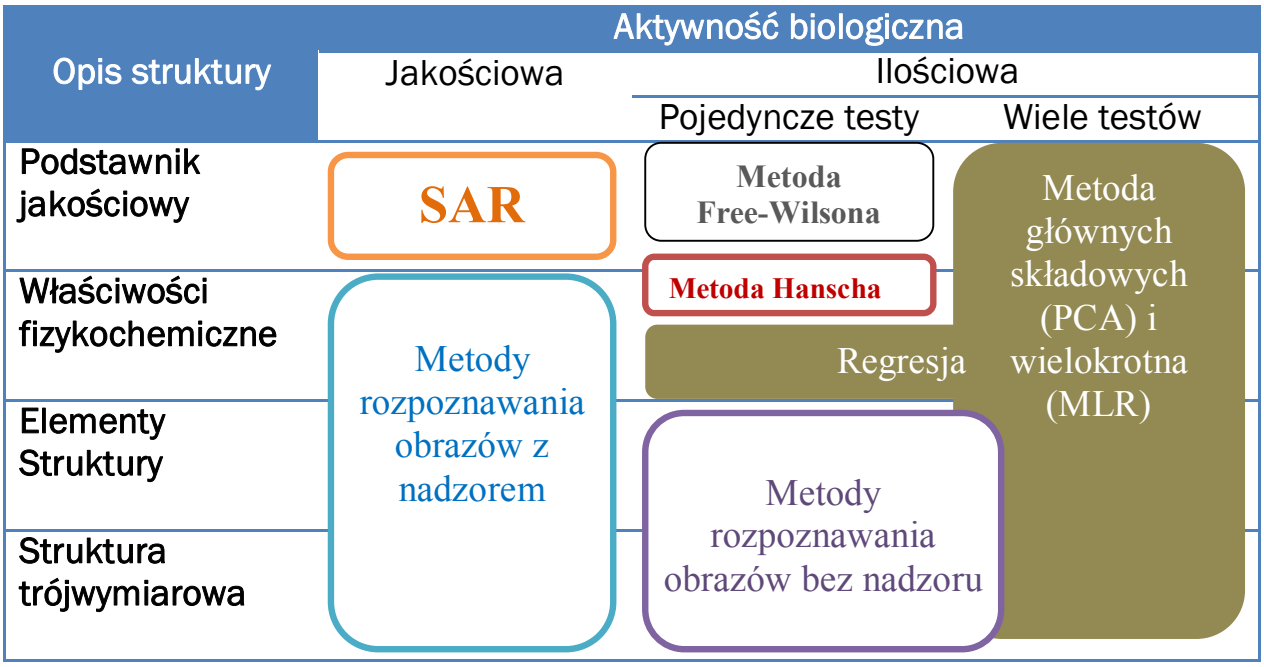
Istotne miejsce w poszukiwaniu i projektowaniu nowych leków zajmuje analiza zależności pomiędzy strukturą a aktywnością badanych związków. Jeśli zależności te wyrażone są ilościowo mówimy o badaniach QSAR (Quatitative Structure-Activity Relationships – Ilościowe zależności struktura-aktywność).

Do podstawowych sposobów opisu **aktywności biologicznej** można zaliczyć:

- podział badanych związków na dwie klasy, np. związki aktywne i nieaktywne lub większej ilości klas (skala nominalna),
- pojedynczy, ilościowy test aktywności, zwykle w formie –log(c), gdzie c oznacza stężenie wywołujące standardową odpowiedź biologiczną.
- wiele (bateria) testów ilościowych, co pozwala na pełniejszy opis zależności [4].

Podobnie **opis struktury** związków można wykonać na wiele sposobów:

- jakościowy opis podstawników,
- ilościowy opis właściwości fizykochemicznych całego związku. Można tu rozróżnić przypadek ograniczonego zestawu właściwości oraz wersję z baterią właściwości,
- ilościowe wielkości uzyskane z metod chemii obliczeniowej, np. rzędy wiązań lub ładunki cząstkowe,
- jakościowy opis elementów struktury,
- struktura trójwymiarowa cząsteczki, ewentualnie z opisem rozkładu pola elektrostatycznego generowanego przez cząsteczkę



Ryc. 4 Różne techniki opisu zależność struktura-aktywność [4].

W przypadku nominalnej skali aktywności biologicznej (klasyfikacja aktywne - nieaktywne) stosuje się zwykle tylko najprostszy sposób opisu budowy chemicznej: jakościowy opis podstawników czyli metodę SAR.

Z kolei w przypadku jednorodnej grupy związków o wspólnym szkielecie (scaffold) z różnymi podstawnikami, o takim samym mechanizmie działania, dla których dysponujemy wynikami jednego lub co najwyżej kilku testów biologicznych (z ilościowo wyrażoną aktywnością) to odpowiednie są regresyjne metody typu metoda Hanscha lub Free-Wilsona.
Częściej jednak mamy do czynienia z dużą liczbą związków i wynikami licznych testów, co daje macierz danych, której nie da się zinterpretować „ręcznie”. W analizie takich wielowymiarowych danych znajdują zastosowanie metody chemometryczne. W pierwszym etapie analizy interesuje nas często odpowiedź na dwa pytania:

- na ile stosowane testy biologiczne są do siebie podobne ze względu na badaną grupę związków? A jeśli są podobne to czy wszystkie są potrzebne do opisu aktywności badanych związków?
- na ile badane związki są do siebie podobne ze względu na stosowane testy?

Odpowiedź na te pytania daje chemometryczna analiza podobieństwa, a dokładniej analiza skupień, analiza czynnikowa czy też analiza głównych składowych. Z jednej strony w tych metodach można analizować tzw. przestrzeń zmiennych (wyników testów, wartości parametrów fizykochemicznych), a z drugiej strony przestrzeń przypadków (badanych związków). **Metody te dokładniej zostały opisane w rozdziałach 12 i 14.**

Bardzo ważnym elementem badań QSAR jest wybór parametrów fizykochemicznych i wybór związków w taki sposób, aby analizowane parametry były odpowiednio zróżnicowane. Ważna jest także liczebność badanych związków, powinno być ich odpowiednio dużo, aby otrzymane wyniki były wiarygodne statystycznie. Minimalna liczba związków dla każdego badanego parametru (zmiennej) to **pięć**.

7. Podstawowe pojęcia w statystycznej analizie danych

Skale pomiaru zjawisk

Skala nominalna: (dla cech o charakterze jakościowym), przy pomiarach w tej skali można dzielić obiekty na grupy np. podział ze względu na płeć (kobiety i mężczyźni), ze względu na grupę krwi itp.

Skala porządkowa: (dla cech o charakterze jakościowym), pomiar polega na uporządkowaniu lub uszeregowaniu obiektów według określonego kryterium. Przykładem może być uporządkowanie grupy kobiet ze względu na wzrost: „niski”, „średni”, „wysoki”.

Skala przedziałowa: (dla cech ilościowych), pomiar polega na przyporządkowaniu liczby określającej natężenie mierzonego zjawiska, przy czym początek skali jest ustalony w sposób umowny. Przykładem może być pomiar temperatury w skali Celsjusza, czas kalendarzowy.

Skala ilorazowa: (dla cech ilościowych), ma wszystkie cechy skali przedziałowej, ale różni się tym, że zero jest obiektywnym początkiem skali. W tej skali można wykonywać wszystkie obliczenia arytmetyczne. Przykładem może być pomiar temperatury w skali Kelvina.

Etapy badania statystycznego

Wyróżnia się cztery etapy badania statystycznego:

- projektowanie badania: sformułowanie celu badawczego, zdefiniowanie zbiorowości statystycznej i jednostki statystycznej, dokonanie wyboru cech statystycznych, określenie metody badania statystycznego, podanie źródeł pozyskania danych, określenie sposobu opracowania i prezentacji zebranego materiału,
- zbieranie materiału statystycznego, wykonanie eksperymentu, obserwacja statystyczna,
- opracowanie zebranego materiału statystycznego: podział uporządkowanego materiału według określonych kryteriów, zliczenie pogrupowanych danych, interpretacja zebranego materiału i pogrupowanie wartości,
- analiza zebranego materiału statystycznego i wnioskowanie statystyczne [5].

Podstawowe cechy statystyczne

Do podstawowych cech statystycznych wykorzystywanych do opisu właściwości zbioru danych należą: miary **położenia**, **zmienności** oraz miary **asymetrii** i **koncentracji**. Jednym z pierwszych etapów eksploracji danych jest obliczenie podstawowych charakterystyk, jak średnia, mediana, wariancja czy odchylenie standardowe.
Miary położenia - wskazują wartość najlepiej reprezentującą wszystkie obserwacje (wyniki), inaczej mówiąc służą do określenia miejsca rozkładu, w którym skupia się duża część obserwacji. Miary położenia dzielą się na przeciętne i kwantyle.

Najczęściej stosowane miary przeciętne to: średnia arytmetyczna, średnia geometryczna, modalna oraz mediana (kwartyl drugi). Ponadto często wykorzystuje się także kwartyl pierwszy i trzeci. Najlepszą i najczęściej stosowana miarą charakteryzującą rozkład wyników jest średnia arytmetyczna, obliczana na podstawie wszystkich wyników.

Miary położenia

Średnia arytmetyczna – suma wartości zmiennej podzielona przez ich liczbę:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Średnia geometryczna – jest pierwiastkiem n-tego stopnia z iloczynu *n* wyników:

$$\overline{x} = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots . x_n}$$

Średnia arytmetyczna ważona – stosowana jest w przypadku obliczania średniej z kilku serii wyników:

$$\overline{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Wagami są liczebności każdej serii wyników.

Mediana – wartość środkowa serii, druga bardzo często wykorzystywana miara tendencji centralnej. Jest to miara pozycyjna rozdzielająca zbiór wartości zmiennej na dwie, równe liczebnie części.

$$Me = \frac{x_{n+1}}{2}$$, gdy *n* jest nieparzyste

$$Me = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$$, gdy n jest parzyste.

Modalna (ang. *mode*) – wartość najczęściej występująca w serii wyników.

Miary rozrzutu (zmienności)

Rozstęp – różnica między największą i najmniejszą wartością zmiennej *x*:

$$R = x_n - x_1$$

Rozstęp jest miarą łatwą do obliczenia, ale posiadającą poważną wadę: jest obliczany na podstawie tylko dwóch skrajnych wyników.

Wariancja – średnia arytmetyczna sumy kwadratów odchyleń wartości poszczególnych wyników od średniej arytmetycznej:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Odchylenie standardowe – określa rozrzut wyników pomiarów wokół wartości średniej, jest pierwiastkiem kwadratowym z wariancji:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Względne odchylenie standardowe- dawniej określane jako współczynnik zmienności.

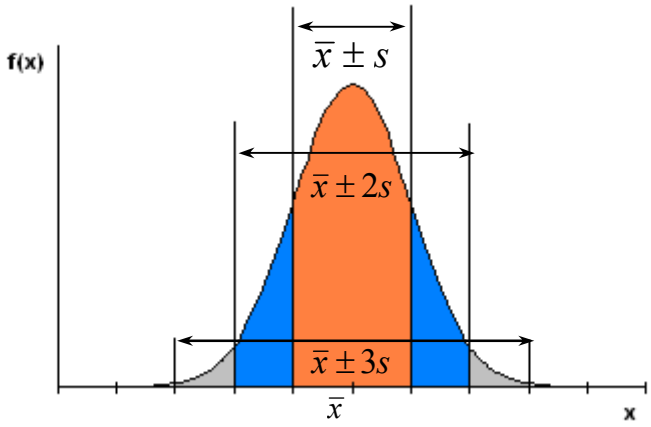
$$RSD = \frac{s}{\overline{x}} \cdot 100\%$$

Odchylenie standardowe posiada następujące właściwości:

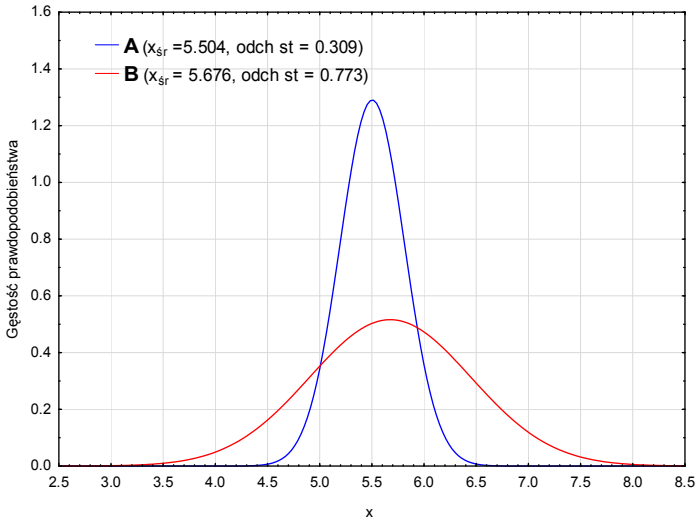
- jest wielkością obliczaną na podstawie wszystkich wyników,
- wartość odchylenia standardowego pozostaje stała jeśli do wszystkich wyników doda się pewną stałą liczbę lub wszystkie wyniki pomnoży przez stałą liczbę większą od zera.

Odchylenie standardowe i wariancja są równe zeru, gdy wszystkie wyniki są identyczne. W przypadku, gdy wyniki równoległych pomiarów różnią się, wielkości te są dodatnie i tym większe im większe rozproszenie wyników.

Stosowanie odchylenia standardowego do charakterystyki serii wyników wymaga założenia, że rozkład tych wyników jest rozkładem **normalnym.** W takim przypadku można wykorzystać tzw. **regułę trzech sigm** (odchyleń standardowych) (Ryc. 5).
Wg tej reguły w przedziale $\overline{x} \pm s$ znajduje się ponad dwie trzecie (68%) wszystkich wyników. Tylko 5 % wyników wykracza poza przedział $\overline{x} \pm 2s$, natomiast poza przedziałem $\overline{x} \pm 3s$ znajduje się zaledwie ok. 0,3% wyników. Oznacza to, że tylko jeden na 370 wyników różni się od średniej arytmetycznej o więcej niż 3s.



Ryc. 5 Właściwości rozkładu normalnego



Ryc. 6 Rozkład normalny dwóch zmiennych charakteryzujących się podobną średnią i różnym odchyleniem standardowym

Miary asymetrii

Kierunek i siłę asymetrii określa współczynnik asymetrii (skośność).

$$A_S = \frac{\sum_{i=1}^{n} (x_1 - \bar{x})^3}{(n-1)(n-2)s^3}$$

Współczynnik asymetrii jest wielkością niemianowaną, a jego znak mówi o kierunku asymetrii. Im wyższa wartość bezwzględna A_s tym rozkład bardziej niesymetryczny.

Asymetrię rozkładu można określić porównując średnią arytmetyczną z medianą i modalną:

- $\bar{x} = Me = Mo$ - dla rozkładu symetrycznego ($A_s = 0$)
- $\bar{x} > Me > Mo$ - dla rozkładu prawostronnie asymetrycznego ($A_s > 0$)
- $\bar{x} < Me < Mo$ - dla rozkładu lewostronnie asymetrycznego ($A_s < 0$).

Miary koncentracji

Najpopularniejszą miarą skupienia obserwacji wokół średniej jest **kurtoza** (**K**). Im wyższa wartość K tym smuklejsza krzywa rozkładu czy większa koncentracja wartości zmiennej wokół średniej.
Jeśli K>0 to rozkład jest smuklejszy niż normalny, natomiast dla K<0 rozkład jest bardziej spłaszczony niż normalny.

Wymienione wyżej liczbowe charakterystyki należą do tzw. statystyki opisowej. Statystyka opisowa ujmuje w syntetyczny sposób podstawowe własności rozkładów badanych zmiennych i jest pierwszym etapem statystycznego opracowywania zebranych danych.

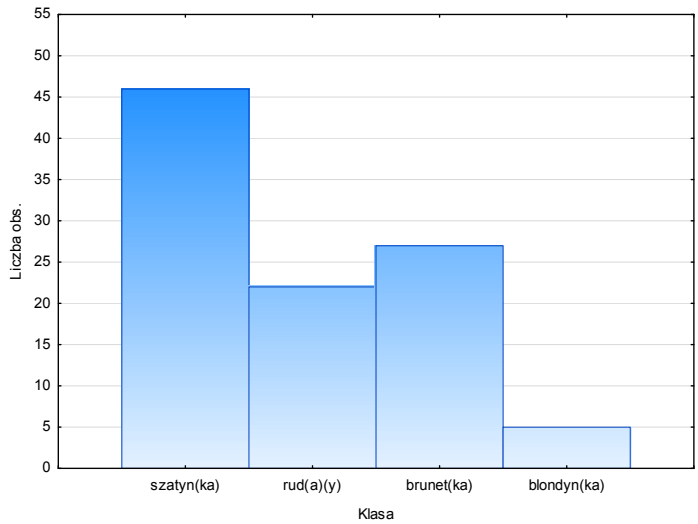
Jednym z podstawowych elementów analizy statystycznej jest określenie empirycznego (zaobserwowanego) rozkładu zmiennej. Empiryczny rozkład zmiennej to przyporządkowanie poszczególnym wartościom zmiennej (w przypadku zmiennej jakościowej) lub przedziałom wartości zmiennej (dla zmiennej ilościowej) odpowiadających im liczebności.
Do tego celu wykorzystuje się **szeregi rozdzielcze** (tabele liczności).

Tabele liczności

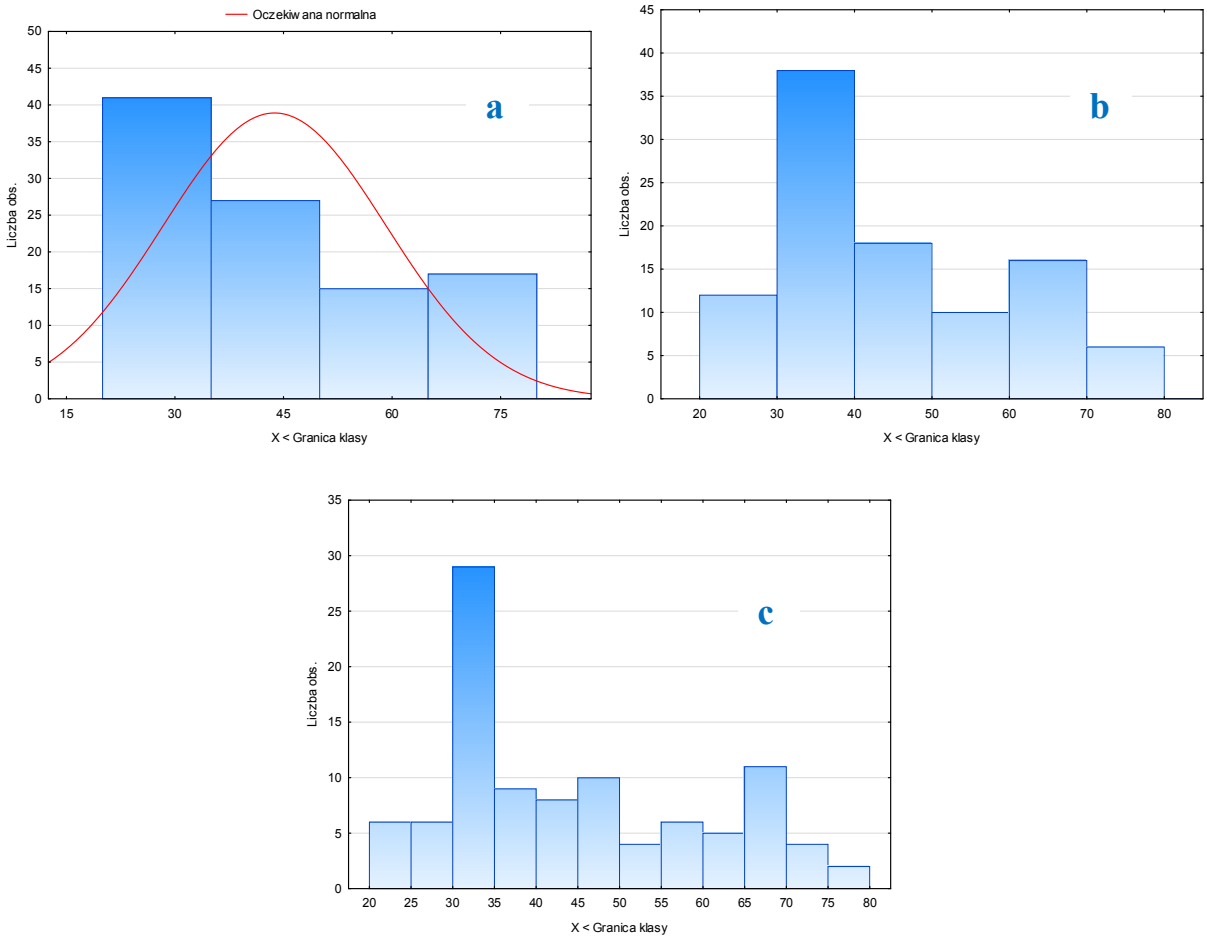
Tabele liczności należą do często stosowane do wstępnej analizy danych jakościowych, ale można je także zastosować do zmiennych ilościowych. Dla zmiennej wyrażonej w skali nominalnej (jakościowej) liczba klas (przedziałów) narzuca się w sposób dość naturalny: palący papierosy, niepalący, ewentualnie palący dużo, średnio, mało, nigdy. Podobnie będzie wyglądał podział grupy ludzi ze względu na kolor włosów (Ryc. 7). Natomiast w przypadku zmiennej ilościowej należy ustalić: liczbę przedziałów i ich granice. Wymaga to przemyślenia jak wyglądają dane i co się chce osiągnąć. Liczba klas może wpłynąć na interpretację otrzymanych wyników (Ryc. 8).
Liczba klas zależy przede wszystkim od liczby obserwacji. Poniżej podano zalecane liczby klas w zależności od liczby obserwacji:

Liczba obserwacji n	Liczba zalecanych klas
40-60	6-8
60-100	7-10
100-200	9-12
200-500	11-17

Graficzną interpretacją tabeli liczności jest histogram (patrz Ryc. 9).



Ryc. 7 Histogram dla zmiennej: kolor włosów



Ryc. 8 Histogramy rozkładu zmiennej “wiek” przy różnych szerokościach przedziałów klasowych

Jeżeli rozkład zmiennej charakteryzuje się jednym maksimum, to rozkład jest **jednomodalny** (Ryc. 8a). Dodatkowo można zaobserwować, że rozkład jest asymetryczny - prawoskośny. Zwiększenie liczby klas pozwala na stwierdzenie, że w rzeczywistości rozkład jest **dwumodalny**.
Należy pamiętać, że zarówno zbyt mała jak i zbyt duża liczba klas może prowadzić, w pierwszym wypadku do niewykrycia pewnych prawidłowości, a w drugim z kolei do zatarcia tych prawidłowości.

8. Wizualizacja danych

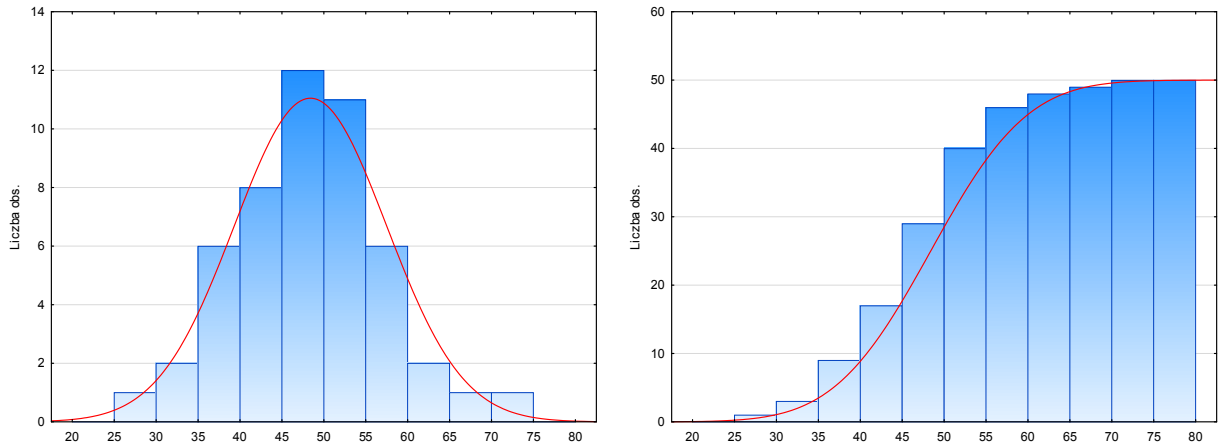
Wstępna analiza danych powinna także obejmować wykonanie odpowiednich wykresów pozwalających na ocenę zróżnicowania danych, asymetrii rozkładu, odróżnienia danych nietypowych (odstających).

Typy wykresów:

- histogram,
- wykresy rozrzutu (korelacji),
- wykresy typu średnia i błędy (ramka-wąsy, kolumnowe, słupkowe),
- wykresy normalności,
- wykresy liniowe (tzw. profile przypadków i zmiennych),
- wykresy kołowe,
- wykresy macierzowe i skategoryzowane.

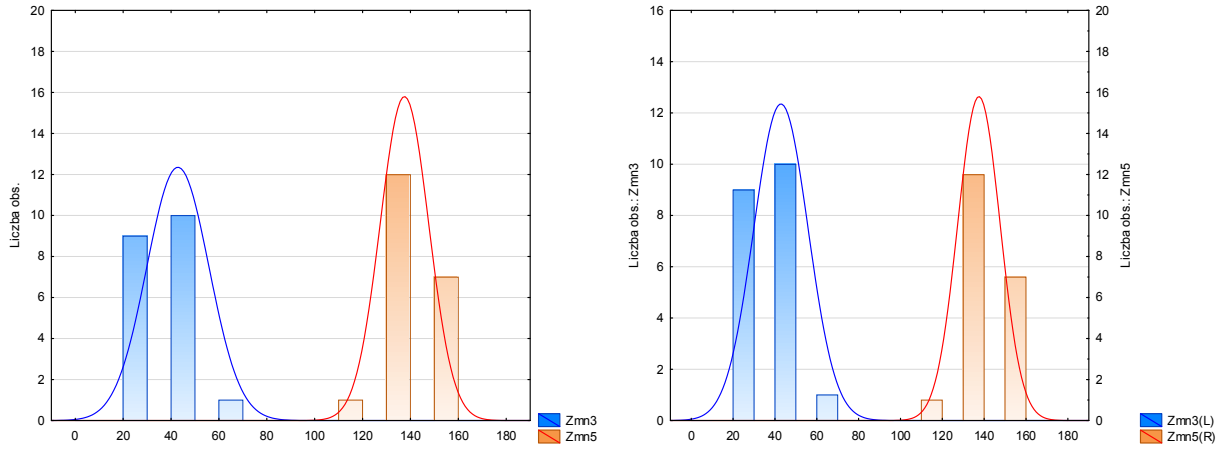
Histogramy

Histogramu służą do graficznego przedstawienia rozkładu liczebności zmiennej. Oś odciętych reprezentuje różne wartości zmiennej lub przedziały wartości, natomiast na osi rzędnych umieszczona jest liczba obserwacji (wysokość słupka).



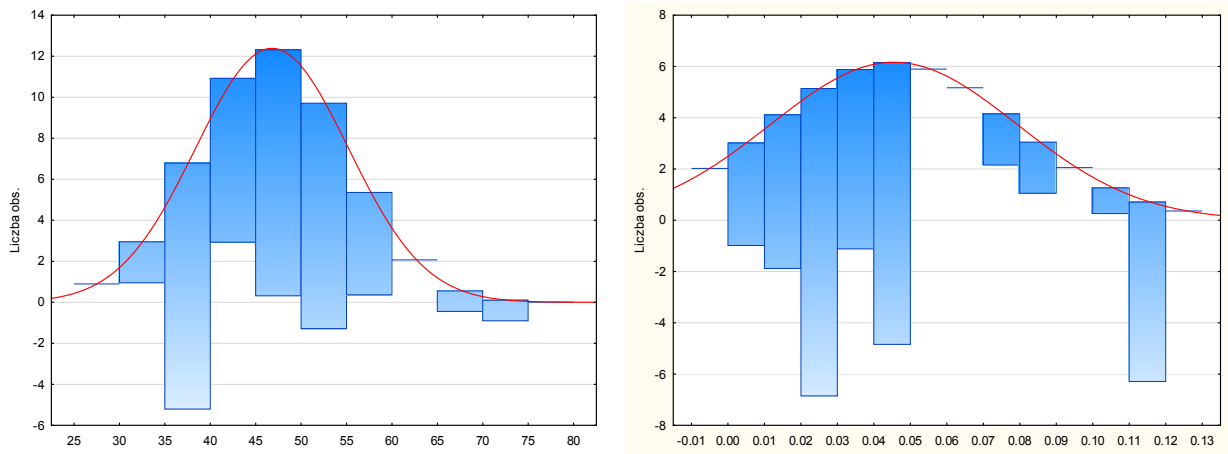
Ryc. 9 Histogram zmiennej wraz z krzywą rozkładu normalnego; histogram standardowy (lewy wykres) i skumulowany (prawy wykres)

Ryc. 10 przedstawiono histogram podwójny – kombinację dwóch histogramów, z których każdy jest osobno wyskalowany oraz histogram wielokrotny, gdzie liczebności obu zmiennych są rysowane w odniesieniu do tej samej osi Y. Histogram wielokrotny ułatwia porównanie między histogramami.



Ryc. 10 Przykładowy histogram podwójny (lewy) oraz histogram wielokrotny (prawy)

Inny rodzaj histogramu – tzw. wiszące słupki może służyć jako wizualny test normalności rozkładu.



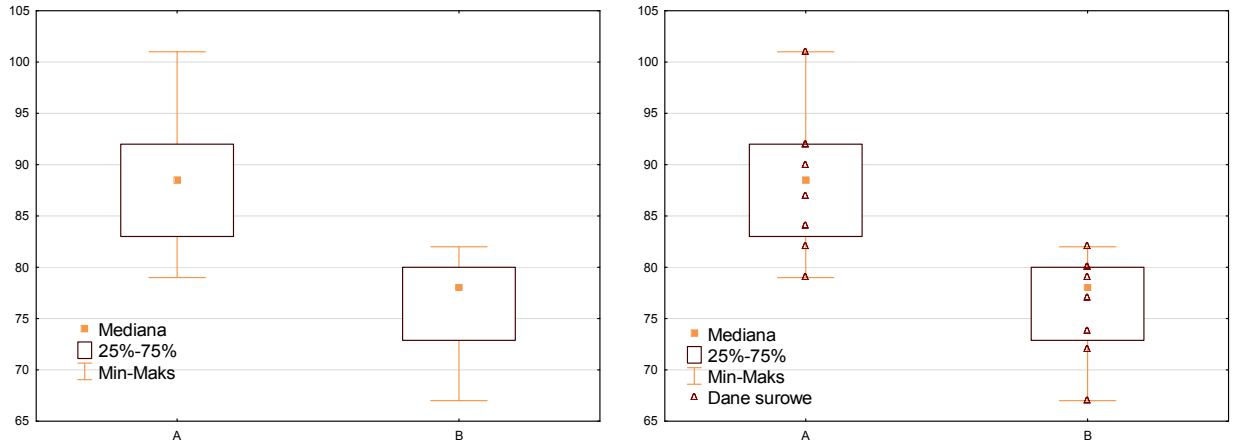
Ryc. 11 Przykłady histogramów z tzw. **wiszącymi słupkami**. Rozkład normalny - lewy wykres, brak rozkładu normalnego – prawy wykres.

Wykresy typu ramka-wąsy (box and whiskers)

Ten typ wykresu służy do pokazania zakresu wybranej zmiennej lub zmiennych oraz statystyki opisowe, jak średnia, mediana, odchylenie standardowe, kwartale itp. Na wykresie można także umieścić dane surowe oraz wyróżnić punkty odstające.
Można utworzyć kilka rodzajów wykresów ramka-wąsy, niosących różne informacje:

- punkt centralny – mediana, ramka – kwartale, wąsy – rozstęp,

- punkt centralny – średnia, ramka – błąd standardowy, wąsy – odchylenie standardowe,
- punkt centralny – średnia, ramka – odchylenie standardowe, wąsy – 1,96*odchylenie standardowe.



Ryc. 12 Przykładowy wykres **ramka-wąsy** (*box and whisker*)

Na Ryc. 12 przedstawiono przykładowy wykres typu ramka-wąsy, składający się z punktu odpowiadającego medianie, ramki wyznaczającej pozycje dolnego i górnego kwartyla oraz wąsów określających wartość najmniejszą i największą wyników (rozstęp). Analiza tego typu wykresu pozwala na:

- określenie zróżnicowania wartości zmiennej,

W tym celu należy porównać długość czterech odcinków wyznaczonych przez wąsy oraz pudełko czyli kwartyli (zawierających 25% wyników). Szerokość ramki daje informacje o zróżnicowaniu 50% najbardziej typowych wyników; im szersza tym zróżnicowanie większe.

- ocenę asymetrii rozkładu.

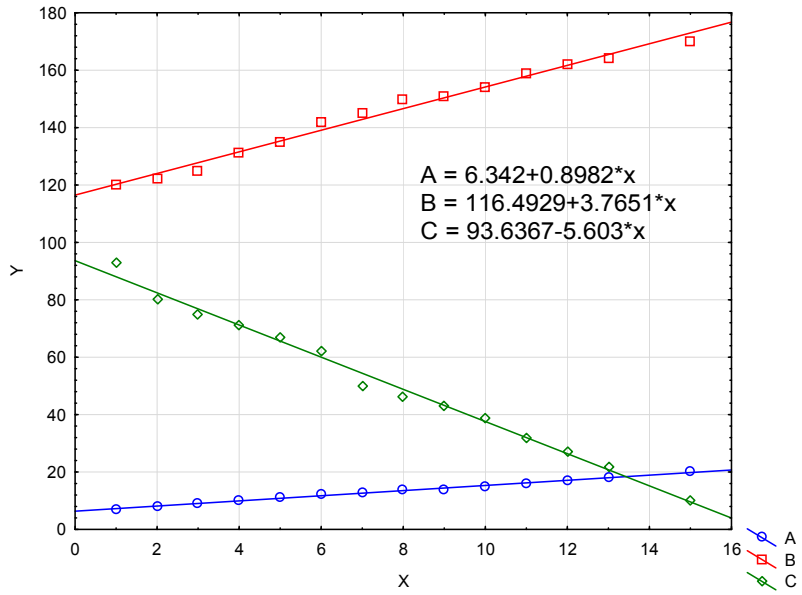
Jeśli wąsy są tej samej długości rozkład jest symetryczny, natomiast jeśli górny wąs jest dłuższy od dolnego rozkład jest prawostronnie asymetryczny Ryc. 12, zmienna A). Asymetria lewostronna występuje w odwrotnym przypadku, gdy dolny wąs jest dłuższy od górnego (Ryc. 12, zmienna B).

Wykresy typu ramka-wąsy pomocne są również do wizualnej oceny statystycznej istotności różnicy pomiędzy średnimi dwóch lub więcej zmiennych.

Wykresy rozrzutu

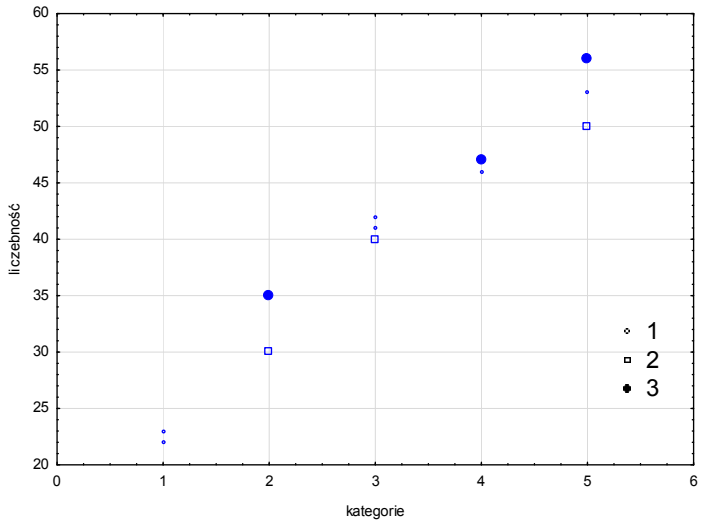
Wykresy rozrzutu umożliwiają zbadanie zależności pomiędzy zmiennymi. Poszczególnym punktom na wykresie odpowiadają dwie współrzędne (X,Y), które jednoznacznie określają ich położenie w układzie współrzędnych.
Zmienne powiązane układają się wzdłuż pewnej krzywej, niepowiązane natomiast będą tworzyć nieregularny kształt (chmurę). Na Ryc. 13 przedstawiono przykładowy wykres

rozrzutu dla trzech zmiennych zależnych od tej samej zmiennej niezależnej. Do prezentowanych zależności zastosowano dopasowanie liniowe.
Wykresy tego typu wykorzystywane są do porównywania kilku korelacji.



Ryc. 13 Wykres rozrzutu wielu zmiennych (wielokrotny)

Inny typ wykresu rozrzutu to wykres rozrzutu liczności pokrywających się punktów między dwiema zmiennymi (Ryc. 14). Zsumowana liczebność pokrywających się punktów przedstawiona jest na wykresie w postaci znaczników różnej wielkości.

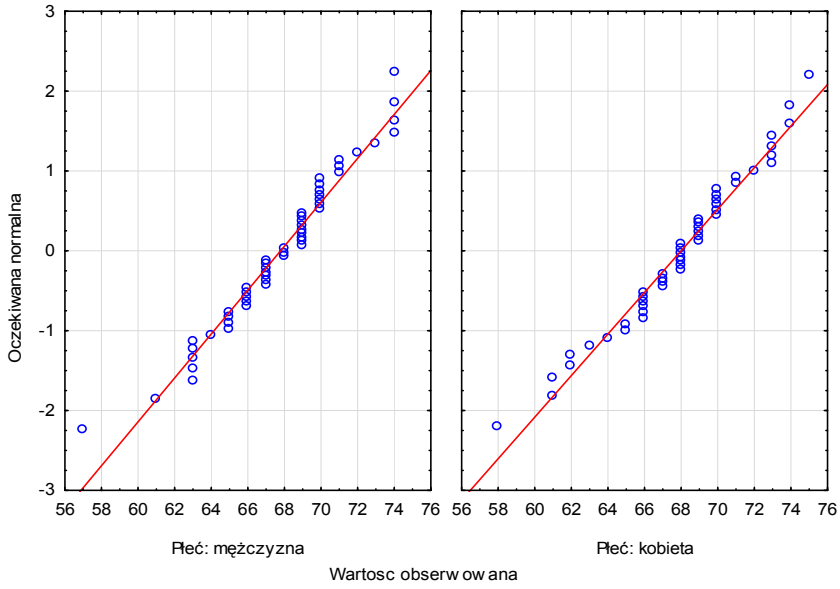


Ryc. 14 Wykres rozrzutu liczności

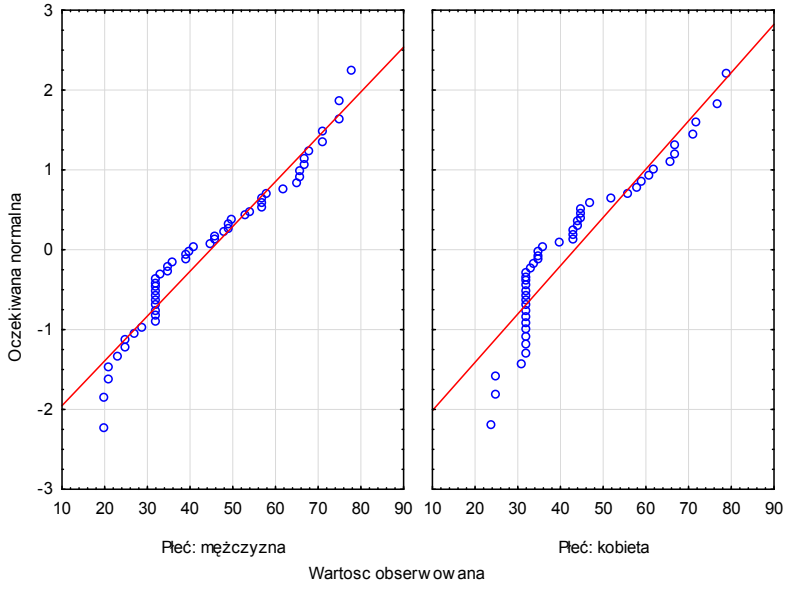
Wykresy normalności

Wykresy normalności pozwalają na wzrokową ocenę normalności rozkładu oraz odstępstwa rozkładu empirycznego od rozkładu normalnego.

Na wykresie umieszcza się skumulowane częstości względne obserwacji w układzie współrzędnych, odpowiednio wyskalowanym, aby wykres dystrybuanty rozkładu normalnego był linią prostą. Im bardziej wszystkie punkty układają się na prostej, tym bardziej prawdopodobne jest, że rozkład zmiennej jest normalny (Ryc. 15). Natomiast jeśli punkty tworzą określony wzór wokół prostej sugeruje to rodzaj przekształcenia, jakie należy zastosować żeby rozkład sprowadzić do normalnego. Np. układ punktów na Ryc. 16 sugeruje przekształcenie logarytmiczne (rozkład log-normalny).



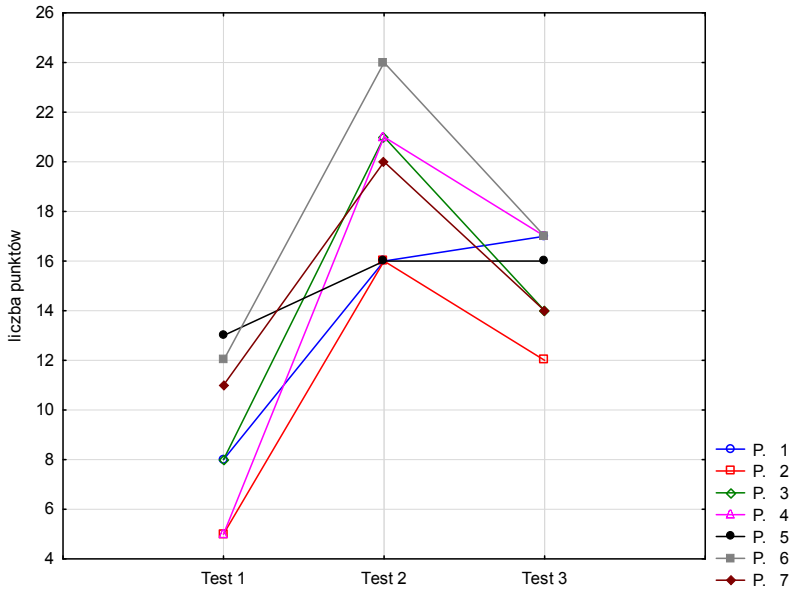
Ryc. 15 Skategoryzowane (ze względu na płeć) wykresy normalności - **zmienna ma rozkład normalny.**



Ryc. 16 Skategoryzowane (ze względu na płeć) wykresy normalności - **zmienna nie ma rozkładu normalnego**

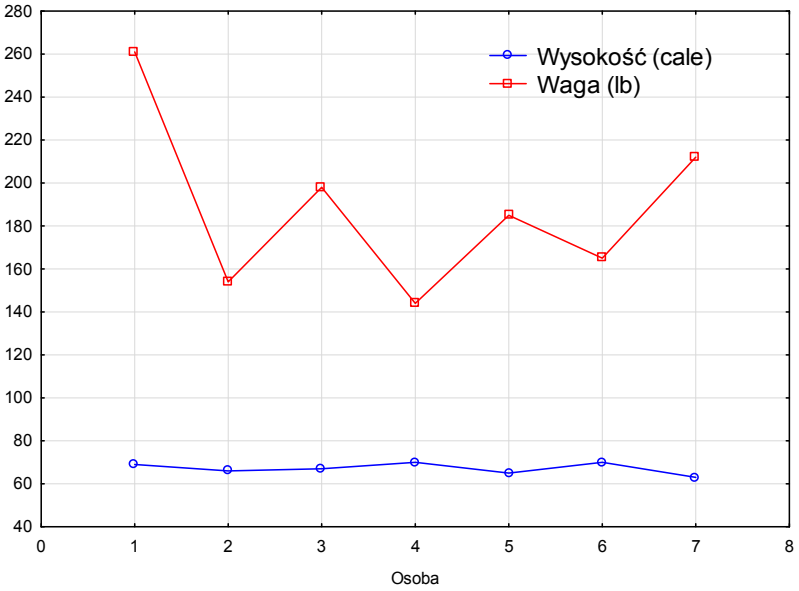
Wykresy liniowe

Wykresy liniowe można utworzyć dla zmiennej (wielu zmiennych) lub przypadków (tzw. profile przypadków). Profil przypadku prezentuje sekwencję wartości zmiennych dla określonego przypadku (obiektu) czyli wykres zmienności wartości w komórkach arkusza danych **w poziomie** (Ryc. 17).



Ryc. 17 Wykres liniowy dla przypadków (profile przypadków)

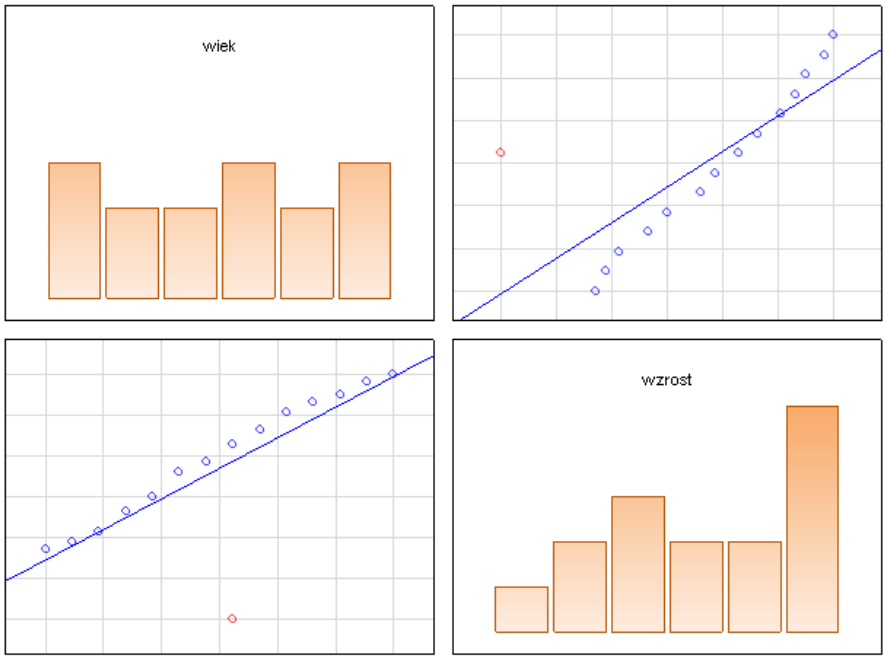
Natomiast wykresy liniowe zmiennej prezentują sekwencję wartości danej zmiennej dla poszczególnych przypadków (obiektów). Jest to wykres zmienności wartości w komórkach arkusza danych **w pionie** (Ryc. 18).



Ryc. 18 Wykres liniowy dwóch zmiennych

Wykresy macierzowe

Wykresy tego typu służą do zbiorczego przedstawiania współzależności pomiędzy wieloma zmiennymi w postaci macierzy zwykłych wykresów rozrzutu i/lub histogramów (Ryc. 19).



Ryc. 19 Macierzowy wykres rozrzutu

Tabela czy wykres?

Odpowiednia prezentacja danych może dostarczyć dużo cennych informacji ułatwiających ich interpretację. Przy wyborze sposobu prezentacji danych często można mieć dylemat co wybrać: wykres czy tabelę. Nawet jeśli już zdecydujemy że dane przedstawimy np. na wykresie, to stajemy przed kolejnym wyborem – a mianowicie rodzaju wykresu. Pozornie wybór tabeli może wydawać się prostszy, ale także tabele mogą mieć różne postacie ułatwiające lub utrudniające interpretację wyników.
Tabele, ze względu na funkcje jakie mają spełniać, można podzielić na dwie kategorie:

- tabele do „przechowywania” danych, w których zbieramy dane surowe,
- tabele zawierające elementy analizy danych pozwalające na mniej lub bardziej wnikliwą interpretację wyników.

Nawet dla stosunkowo prostych danych można skonstruować kilka form tabeli. Na przykład, dane dotyczące liczby osób w Polsce i Chinach z podziałem na płeć i trzy kategorie wiekowe, można umieścić w **ośmiu(!)** różnych tabelach.
Ponieważ liczby, w naturalny sposób, są najłatwiejsze do porównania, gdy znajdują się obok siebie, to najlepszą formą tabeli będzie taka, która pozwoli na umieszczenie obok siebie porównywanych danych.

Wybór jednego z ośmiu wariantów, będzie zależeć od tego, do porównywania których danych ma służyć konkretna tabela.

Tabelę 1 należy wybrać, jeśli chcemy zwrócić uwagę na porównanie liczebności mężczyzn w obu krajach oraz liczebności kobiet w obu krajach (porównanie „poziome” - dane będą umieszczone w sąsiadujących kolumnach).

Tabela 1

Wiek (lata)	**Mężczyźni**		**Kobiety**	
	Polska	Chiny	Polska	Chiny
0-21				
22-49				
50+				

Jeśli z kolei bardzie zależy nam na porównaniu proporcji kobiet i mężczyzn w poszczególnych krajach to odpowiedniejsza będzie Tabela 2.

Tabela 2

Wiek (lata)	**Polska**		**Chiny**	
	Mężczyźni	Kobiety	Mężczyźni	Kobiety
0-21				
22-49				
50+				

Tabela 3

Kraj	**Mężczyźni** (wiek, lata)			**Kobiety** (wiek, lata)		
	0-21	22-49	50+	0-21	22-49	50+
Polska						
Chiny						

Tabela 4

Kraj	**Polska** (wiek, lata)			**Chiny** (wiek, lata)		
	0-21	22-49	50+	0-21	22-49	50+
Mężczyźni						
Kobiety						

Jak widać w Tabeli 5 „nadrzędną” informacją jest liczba kobiet i mężczyzn w różnych kategoriach wiekowych. Z kolei w Tabeli 6 porównujemy liczebności Polaków i Chińczyków w kategoriach wiekowych.

Tabela 5

Kraj	**0-21 lat**		**22-49 lat**		**50+ lat**	
	Mężczyźni	Kobiety	Mężczyźni	Kobiety	Mężczyźni	Kobiety
Polska						
Chiny						

Tabela 6

Kraj	**0-21 lat**		**22-49 lat**		**50+ lat**	
	Polska	Chiny	Polska	Chiny	Polska	Chiny
Mężczyźni						
Kobiety						

Tabele 7 i 8 różnią się konstrukcją od pozostałych i raczej wymuszają porównania „pionowe” - danych umieszczonych w różnych wierszach.

Tabela 7

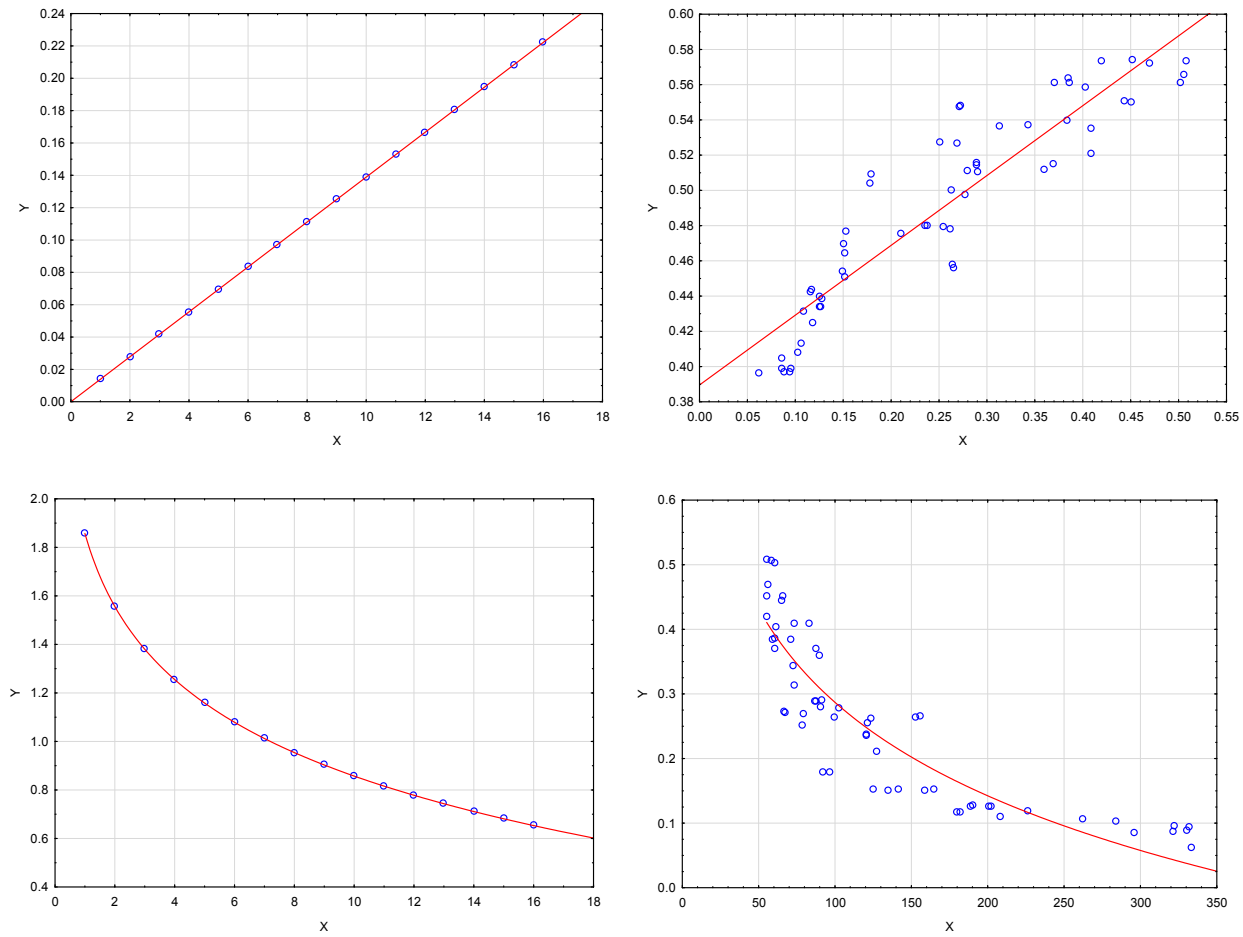
	0-21 lat	**22-49 lat**	**50+ lat**
Mężczyźni			
Polska			
Chiny			
Kobiety			
Polska			
Chiny			

Tabela 8

	0-21 lat	**22-49 lat**	**50+ lat**
Polska			
Mężczyźni			
Kobiety			
Chiny			
Mężczyźni			
Kobiety			

9. Analiza współzależności zjawisk

Jednym z celów wielu doświadczeń biologicznych i medycznych jest badanie współzależności między zmiennymi, czyli jej kształt, siła i kierunek.
Wyróżnia się dwa rodzaje współzależności zmiennych:
funkcyjną, która występuje gdy zmianie wartości jednej zmiennej towarzyszy ściśle określona zmiana wartości drugiej zmiennej inaczej mówiąc jednej wartości zmiennej niezależnej (X) odpowiada tylko jedna! wartość zmiennej zależnej (Y),
stochastyczną (**korelacyjną**), gdy wraz ze zmianą jednej zmiennej zmienia się rozkład prawdopodobieństwa czyli wartościom zmiennej niezależnej odpowiadają ściśle średnie wartości zmiennej zależnej. W przypadku zależności stochastycznej można ustalić jak zmieni się – średnio – wartość zmiennej Y przy zmianie wartości zmiennej X.
Prostym sposobem wykrywania zależności korelacyjnej miedzy badanymi zmiennymi jest wykorzystanie wykresów rozrzutu, które reprezentują graficznie związek pomiędzy zmiennymi. Wzrokowa ocena pozwala często na określenie siły i rodzaju zależności.
Na Ryc. 20 przedstawiono przykłady zależności funkcyjnych oraz stochastycznych (korelacyjnych).

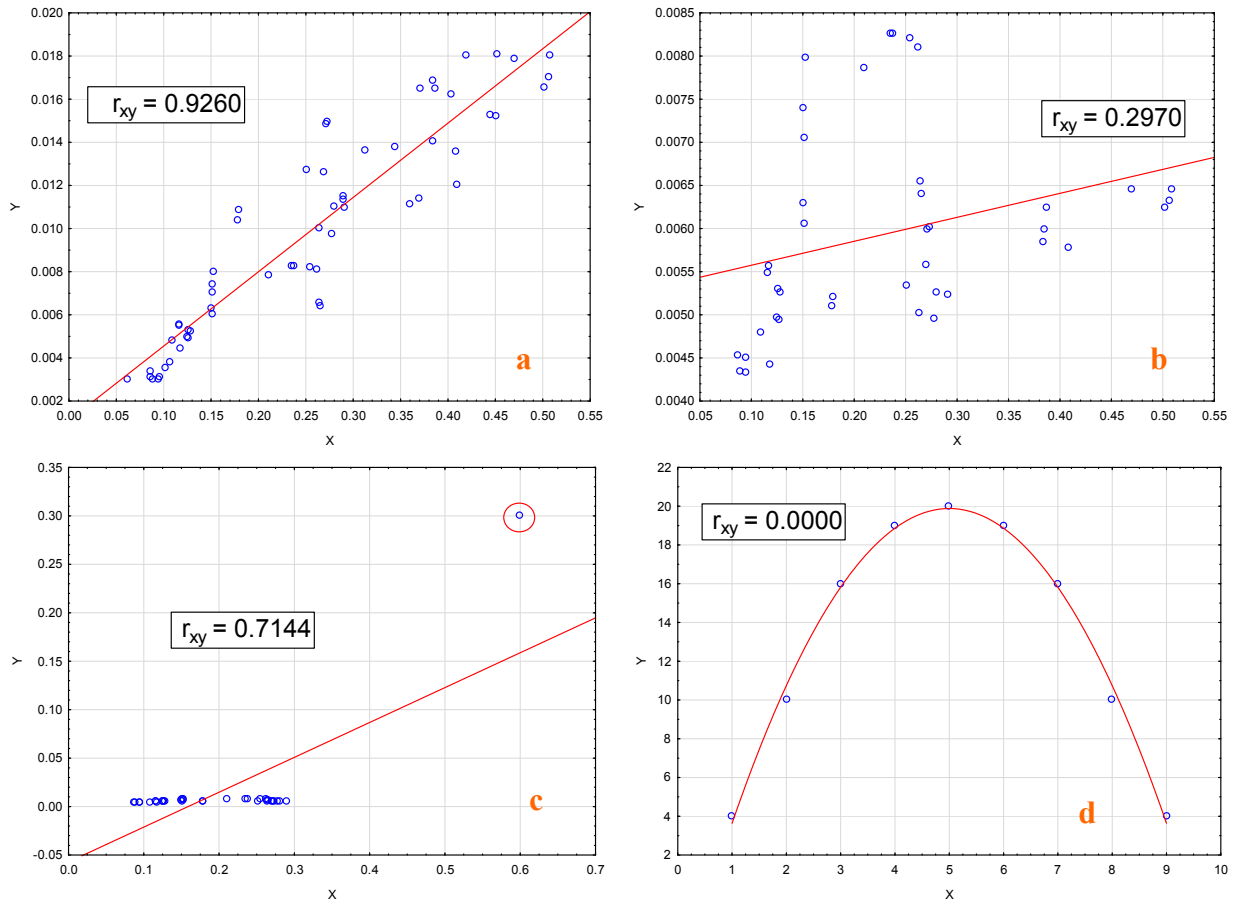


Ryc. 20 Przykłady zależności funkcyjnych - deterministycznych (lewa strona) i zależności korelacyjnych - stochastycznych (prawa strona).

Analiza korelacyjna

Analiza korelacyjna pozwala wykryć i opisać w sposób ilościowy tylko współwystępowanie zmiennych.
Uwaga: liczbowe stwierdzenie występowania współzależności nie zawsze oznacza występowanie związku przyczynowo-skutkowego między badanymi zmiennymi.
Siłę współzależności dwóch zmiennych można wyrazić liczbowo za pomocą wielu mierników, jak kowariancja czy współczynnik korelacji liniowej Pearsona (dla zmiennych ilościowych) oraz współczynnik R Spearmana czy Tau Kendalla (dla zmiennych jakościowych).

Najbardziej popularny współczynnik korelacji czyli **współczynnik korelacji liniowej Pearsona** (r_{xy}) wykorzystywany jest do pomiaru siły zależności prostoliniowej między dwiema cechami mierzalnymi (ilościowymi). Przyjmuje wartości z przedziału [-1,1]. Znak współczynnika informuje o kierunku korelacji, natomiast jego bezwzględna wartość o sile związku. Jeżeli r_{xy}=0, oznacza to zupełny brak liniowej zależności pomiędzy zmiennymi. Nie oznacza to jednak braku zależności innej niż liniowa (przykłady Ryc. 21).



Ryc. 21 Zależność współczynnika korelacji liniowej Pearsona od układu punktów.

Przy interpretacji współczynnika korelacji r_{xy} należy pamiętać, że:

- korelacja powinna być istotna,
- wartość współczynnika 0 nie oznacza braku zależności w ogóle, a tylko brak zależności liniowej (Ryc. 21 d),
- wartość współczynnika silnie zależy od wyników nietypowych (odstających), jak na Ryc. 21 c.

Analiza regresji

Celem badania związków zachodzące między zmiennymi jest określenie wpływu jaki wywiera zmienna niezależna (X) na zmienną z nią powiązaną (Y). Zmienna zależna określana jest także jako: *objaśniana* lub *prognozowana*, natomiast zmienne niezależne nazywane są także zmiennymi: *objaśniającymi* lub *predykatorami*.
Uwzględniając liczbę zmiennych można mieć do czynienia z czterema sytuacjami:

- zmienna zależna **jednowymiarowa** – **jedna** zmienna niezależna,
- zmienna zależna **jednowymiarowa** – **wiele** zmiennych niezależnych,
- zmienna zależna **wielowymiarowa** – **jedna** zmienna niezależna,
- zmienna zależna **wielowymiarowa** – **wiele** zmiennych niezależnych.

Formalnym zapisem wpływu jednej (lub wielu) zmiennych na inną zmienną są funkcje regresji. W przypadku zależności pomiędzy **jedną** zmienną zależną a **jedną** zmienną niezależną mówimy o *regresji prostej*. Natomiast jeśli badamy wpływ **wielu** zmiennych niezależnych na **jedną** zmienną zależną stosujemy *regresję wieloraką*.

Analiza regresji wykorzystywana jest do:

- badania czy zmienna (lub zmienne) X ma (mają) istotny wpływ na zmienną Y
- przewidywania wartości zmienne zależnej (Y) na podstawie znanych wartości zmiennej niezależnej (X) lub kilku zmiennych niezależnych,
- oceny efektów decyzji związanych ze zmianą X.

Ogólnie proces wyznaczania modelu funkcji regresji można podzielić na kilka etapów:

- **specyfikacja modelu** – sformułowanie modelu w postaci nadającej się do dalszej analizy i weryfikacji empirycznej,
- **estymacja parametrów modelu** –otrzymanie liczbowych wartości parametrów modelu i zastosowanie odpowiednich metod statystycznych do otrzymania jak najlepszych ocen parametrów,
- **weryfikacja modelu** – szacowanie istotności otrzymanych parametrów,
- **sprawdzenie zdolności predykcyjnej modelu** – wykorzystanie modelu do obliczenia nieznanych wartości zmiennej zależnej na podstawie danej wartości zmiennej niezależnej.

Matematyczna postać regresji liniowej:

$$Y = \beta_0 + \beta_1 \cdot x + \varepsilon$$

Parametr β_0 – oznacza wyraz wolny, β_1 – współczynnik regresji zmiennej Y względem zmiennej X, odpowiadający współczynnikowi kierunkowemu funkcji liniowej. Parametr ε oznacza składnik losowy.

W celu odpowiedniego dopasowania funkcji regresji do danych stosuje się metodę najmniejszych kwadratów, której istotą jest minimalizowanie sumy kwadratów różnic między wartościami empirycznymi (y_i) a wartościami teoretycznymi ($\hat{y}_i$) (z równania krzywej).

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min$$

Założenia modelu:

- model jest liniowy względem parametrów β_0 i β_1,
- liczba obserwacji (wyników) musi być większa lub równa liczbie oszacowanych parametrów czyli ≥ 2,
- składnik losowy (ε) ma wartość oczekiwaną równą zeru czyli czynniki nieuwzględnione w modelu nie mają istotnego wpływu na średnią wartość Y,
- tzw. **reszty modelu** mają rozkład normalny i są ze sobą nieskorelowane. Reszta odpowiadająca i-tej obserwacji wyrażą się wzorem:

$$e_i = y_i - \hat{y}_i \qquad i=1,2,\ldots,n,$$

gdzie y_i –wartość rzeczywista zmiennej, a $\hat{y}_i$ – wartość teoretyczna wyznaczona z krzywej regresji.

Wariancję składnika losowego czyli tzw. **wariancję resztową** wyraża się wzorem:

$$S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

Pierwiastek kwadratowy z wariancji resztowej czyli **odchylenie standardowe reszt** nazywane także **błędem standardowym estymacji** jest często stosowaną miarą zgodności modelu z danymi empirycznymi. Błąd standardowy estymacji wskazuje na przeciętną różnicę między wartościami rzeczywistymi zmiennej Y a wartościami teoretycznymi.
Ocen jakości modelu obejmuje także ocenę istotności parametrów regresji czyli wyrazu wolnego i współczynnika regresji. Z wykorzystaniem testu t sprawdzamy czy parametry te różnią się statystycznie istotnie od zera.

Zastosowanie analizy regresji prostej:
Analizę regresji można wykorzystać do porównania dokładności dwóch metod (np. metod diagnostycznych w medycynie, dwóch metod instrumentalnych w chemii). Jeśli wyniki otrzymane jedną metodą oznaczymy **A**, a drugą metodą **B** to można wyliczyć parametry regresji liniowej $A = b_0 + b_1 \cdot B$. Obie metody należy uznać za idealnie równoważne (dające

identyczne wyniki) jeśli $b_1=1$, a $b_0=0$. Wówczas równanie regresji przyjmuje postać: A = B czyli każdej wartości otrzymanej metodą A odpowiada taka sama wartość metody B. Jeśli parametr $b_0 \neq 0$, wówczas jedna z metod jest obarczona względem drugiej błędem systematyczny.

Regresja wieloraka

W przypadku kiedy celem analizy jest sprawdzenie wpływu **kilku** zmiennych niezależnych na jedną zmienną zależną (objaśnianą) stosuje się regresję wieloraką (wielokrotną), której równanie ma postać:

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + \varepsilon$$

Założenia modelu:

- model jest liniowy względem parametrów,
- liczba obserwacji (wyników) musi być większa lub równa liczbie parametrów w równaniu. W praktyce należy starać się, żeby liczba obserwacji przypadająca na jeden szacowany parametr była **co najmniej 5**. Im większa liczba obserwacji, tym większa precyzja oszacowania parametrów.
- żadna ze zmiennych niezależnych **nie jest** kombinacją liniową innych zmiennych. Zmienne wprowadzone do modelu powinny być silnie skorelowane ze zmienną zależną, natomiast bardzo słabo skorelowane ze sobą. W praktyce oznacza to, żeby współczynnik korelacji pomiędzy zmiennymi niezależnymi był **mniejszy** niż pomiędzy nimi a zmienną zależną.
- składnik losowy ma wartość oczekiwaną zero,
- wariancja reszt jest taka sama dla wszystkich obserwacji,
- reszty są nieskorelowane,
- reszty mają rozkład normalny.

W przypadku regresji wielorakiej do oceny dopasowania modelu do danych empirycznych stosuje się m. in. **współczynnik determinacji R^2**. Np. wartość współczynnika determinacji 0.85 oznacza, że 85% ogólnej zmienności zmiennej zależnej jest wyjaśnione przez model. Niestety nie zawsze wysoka wartość współczynnika determinacji oznacza dobry model regresji wielorakiej. Należy być szczególnie ostrożnym w interpretacji wartości R^2 gdy:

- liczba obserwacji (n) jest równa liczbie zmiennych w modelu (k), R^2 jest wówczas równe 1, co nie odzwierciedla siły rzeczywistego związku,
- n jest niewiele większe od k+1,
- model nie jest liniowy,
- w modelu nie uwzględniono wyrazu wolnego,
- występuje współliniowość zmiennych niezależnych.

Wadą współczynnika determinacji jest to, że nie dostarcza informacji o istotności zmiennych niezależnych wprowadzanych do modelu; im więcej zmiennych tym R^2 będzie wyższe. Często wzrostowi współczynnika determinacji towarzyszy brak istotności zmiennych dodawanych do modelu, co może być spowodowane ich współliniowością. Wobec tego

zamiast wartością „surowego” R^2 należy wykorzystywać **poprawiony** (**skorygowany**) **R^2,** którego wartość maleje po wprowadzeniu do modelu zmiennych nieistotnych.

Regresja grzbietowa

Regresje grzbietową stosuje się w przypadku kiedy zmienne niezależne są silnie skorelowane i oceny współczynników regresji nie mogą być uzyskane przez zastosowanie metody najmniejszych kwadratów.

Tolerancja

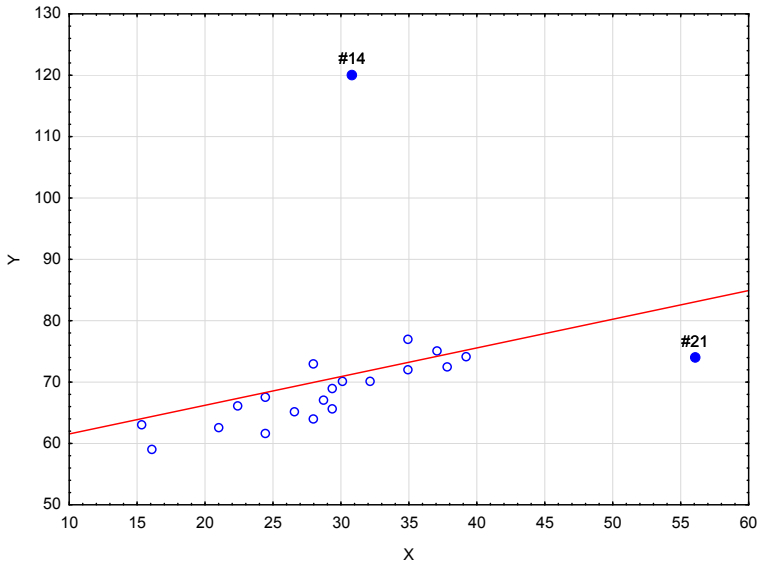
Tolerancję zmiennej definiuje się jako 1-R^2 (1 minus kwadrat korelacji wielokrotnej) tej zmiennej ze wszystkimi innymi zmiennymi niezależnymi w równaniu regresji. Im mniejsza tolerancja zmiennej, tym bardziej nadmiarowy (redundantny) jest jej wkład do regresji.

Analiza reszt, obserwacje odstające (nietypowe i wpływowe)

Po skonstruowaniu modelu (równia regresji) należy zawsze przeanalizować wartości przewidywane i wartości reszt. Analiza regresji jest czuła na pojedyncze obserwacje różniące się znacząco o pozostałych wartości, wobec czego należy unikać sytuacji, w których model jest nadmiernie uwarunkowany właśnie przez takie przypadki. Obserwacje odstające mogą być wynikiem błędu w pomiarze, pomyłki przy wprowadzaniu danych do arkusza, zanieczyszczeniem próbki itp. Z drugiej strony obserwacje odstające mogą wskazywać na braki w modelu lub dobór niewłaściwego. Obserwacji odstających nie należy pochopnie usuwać (chyba, że jest to ewidentnie pomyłka) ponieważ mogą one wnieść dodatkową informację będącą źródłem poprawy modelu lub nawet odkrycia nieznanych zależności.

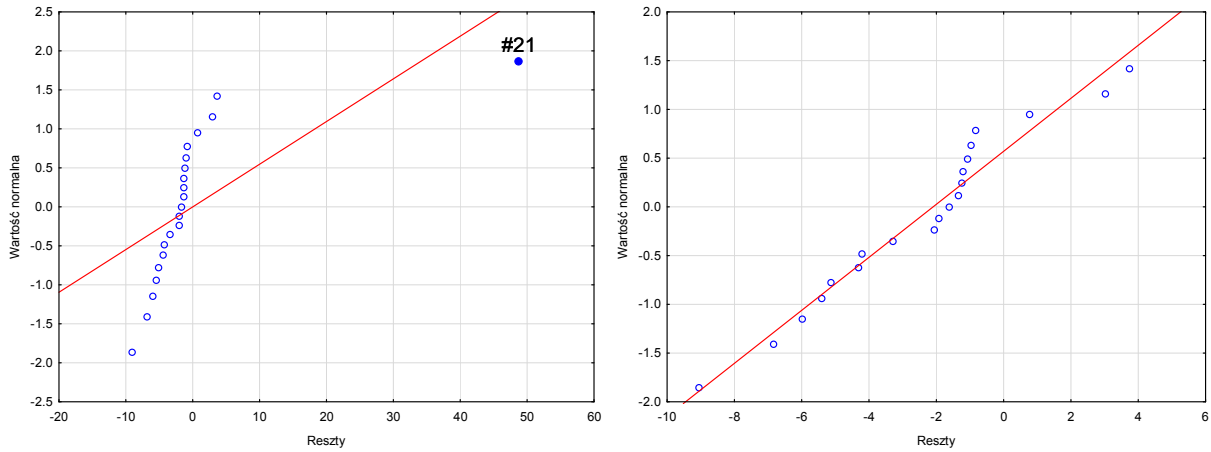
Analiza reszt musi być połączona z gruntowną znajomością badanego zjawiska czyli m.in.:

- sposobem zbierania danych,
- typem jednostki eksperymentalnej,
- jednostką w jakiej wyrażona jest każda zmienna,
- zakresem wartości oraz wartością typową dla każdej zmiennej.

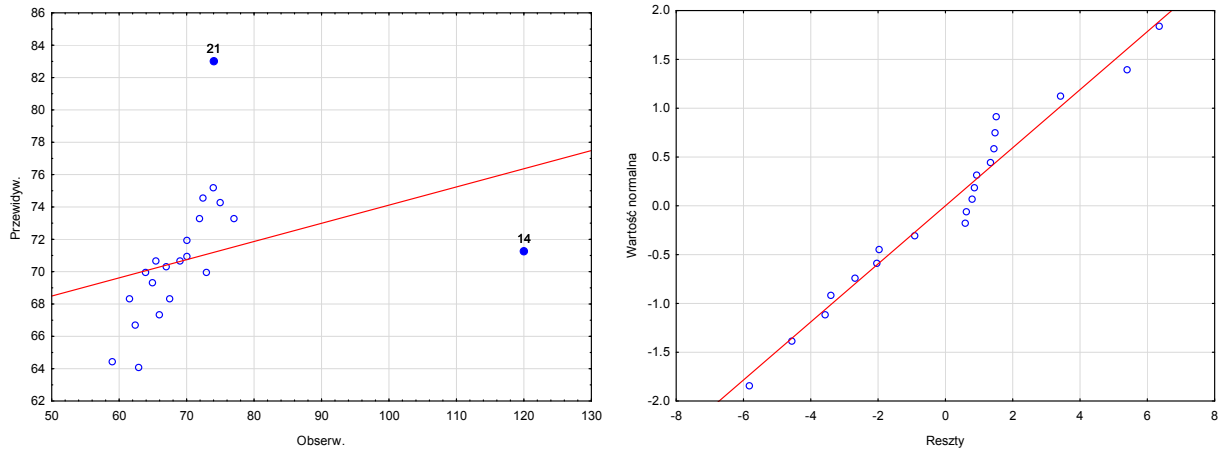


Ryc. 22 Wykres rozrzutu: przykład obserwacji **nietypowej** (#14) i **wpływowej** (21)

Na Ryc. 23 po lewej stronie założenie normalności rozkładu reszt nie jest spełnione (punkty nie układają się wzdłuż linii prostej), natomiast po usunięciu obserwacji odstającej (#21) reszty mają rozkład normalny.



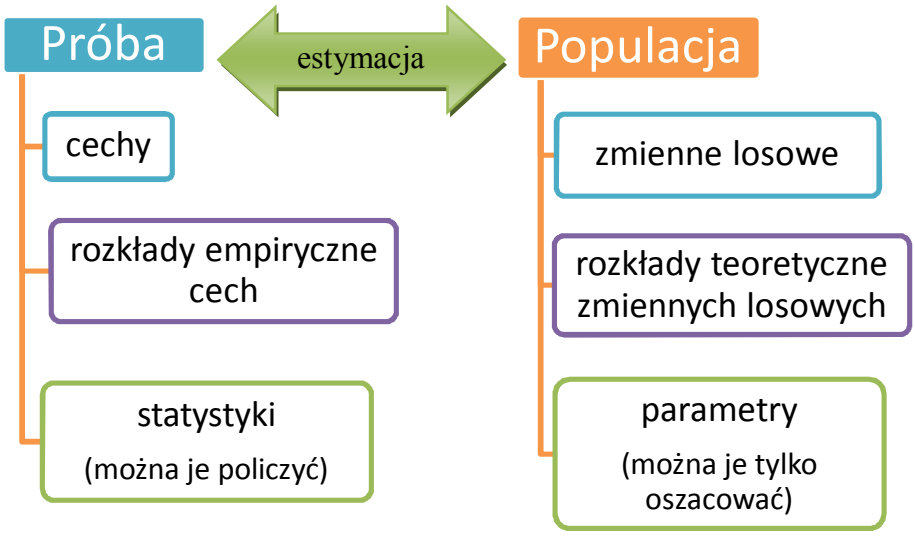
Ryc. 23 Wykres normalności reszt **przed** i **po** usunięciu obserwacji odstającej



Ryc. 24 Wykresy normalności reszt modelu zawierającego wszystkie obserwacje (po lewej stronie) oraz modelu po usunięciu obserwacji 14 (po prawej stronie).

10. Wnioskowanie statystyczne

Badacza (statystyka) zazwyczaj interesują właściwości grupy czy zespołu, a nie właściwości poszczególnych elementów do nich należących. W statystyce zbiór wszystkich elementów (jednostek), które charakteryzują się określonymi właściwościami nazywa się **populacją** lub **zbiorowością generalną**. Niestety badacz nie dysponuje danymi dotyczącymi całej populacji, wobec tego swoje badanie musi realizować na tzw. **próbie** czyli podgrupie wybranej z populacji za pomocą odpowiedniej metody.
Wnioskowanie statystyczne polega na uogólnianiu wynikających z eksperymentu wykonanego na próbie i wykracza poza informacje wynikające ze zgromadzonych danych.
Schemat na Ryc. 25 pokazuje zależności pomiędzy **próbą** a **populacją**.



Ryc. 25 Zależności pomiędzy próbą a populacją

Jeśli np. badacz sformułował twierdzenie, że pewien lek pozwala na skuteczne leczenie i wydłużenie życia chorym na określony rodzaj nowotworu, to jest to **hipoteza naukowa**.
Tak postawiona hipoteza naukowa determinuje określony sposób postępowania czyli wykonanie odpowiednio zaplanowanego eksperymentu, w wyniku którego zostaną zebrane dane liczbowe. Te z kolei muszą zostać zweryfikowane przez analizę statystyczną zgodnie z odpowiednio postawionymi **hipotezami statystycznymi**.

Praktyczne procedury stosowane do testowania hipotez noszą nazwę **testów istotności**.
Test istotności – jest to test statystyczny, w wyniku którego przyjmuje się decyzję o odrzuceniu hipotezy zerowej lub stwierdza brak podstaw do jej odrzucenia.
Uwaga: brak podstaw do odrzucenia hipotezy zerowej nie oznacza jej przyjęcia!

Przeprowadzenie testu istotności składa się z kilku etapów:

- formułowanie hipotezy zerowej i alternatywnej,
- wybór tzw. poziomu istotności,

- dobór odpowiedniego do postawionej hipotezy testu i obliczenie jego wartości w oparciu o dane pochodzące z próby, weryfikacja założeń testu,
- wyznaczenie obszaru krytycznego testu,
- interpretacja wyników - podjęcie decyzji o odrzuceniu lub braku podstaw do odrzucenia hipotezy zerowej

Hipoteza zerowa i alternatywna

W pierwszym etapie wnioskowania statystycznego formułuje się hipotezę, którą chcemy sprawdzić tzw. hipotezę **zerową** i oznacza H_0. Hipoteza zerowa zazwyczaj zakłada brak różnic i zależy nam na jej odrzuceniu (oczywiście pod warunkiem istnienia podstaw do takiej decyzji).
Przykładowo jeśli ocenie ma podlegać czas działania dwóch leków przeciwbólowych (nowego - **1** i starego - **2**), to hipoteza zerowa ma postać:

$$H_0: \mu_1 = \mu_2$$

Natomiast hipoteza alternatywna – na której przyjęciu nam zależy- powinna zakładać różnicę, a tym przypadku fakt, że lek **1** działa dłużej niż lek **2** (czyli, że jest skuteczniejszy).

$$H_0: \mu_1 > \mu_2$$

Przykłady hipotez zerowych dla różnych eksperymentów

Badanie	Hipoteza zerowa	Komentarz
Wpływ terapii lekiem na poziom cholesterolu w porównaniu z plecebo	$H_0: \mu_1 = \mu_2$ lub $H_0: \mu_1-\mu_2=0$	μ_1 – odpowiada średniemu stężeniu cholesterolu po podaniu leku, μ_2 – średnie stężenie cholesterolu po podaniu placebo
Wpływ antybiotyku na skuteczność leczenia	$H_0: p_0=0,8$	p_0 – opowiada prawdziwemu stosunkowi pacjentów wyleczonych; tak sformułowana H_0 oznacza, że hipotetyczna skuteczność leczenia wynosi 80%
Wyznaczanie średniej masy tabletki	$H_0: M=250\ mg$	Wyznaczona masa wynosi 250 mg
Porównanie dwóch procedur mieszania pod względem homogeniczności otrzymanych mieszanin		Wariancje uzyskane z prób dwóch procedur są hipotetycznie równe

W trakcie weryfikacji hipotezy zerowej (za pomocą odpowiednio dobranego testu) można popełnić dwa rodzaje błędów:

- **błąd I rodzaju:** odrzucić hipotezę zerową, mimo że jest ona prawdziwa,
- **błąd II rodzaju:** przyjąć hipotezę zerową, gdy w rzeczywistości jest ona fałszywa.

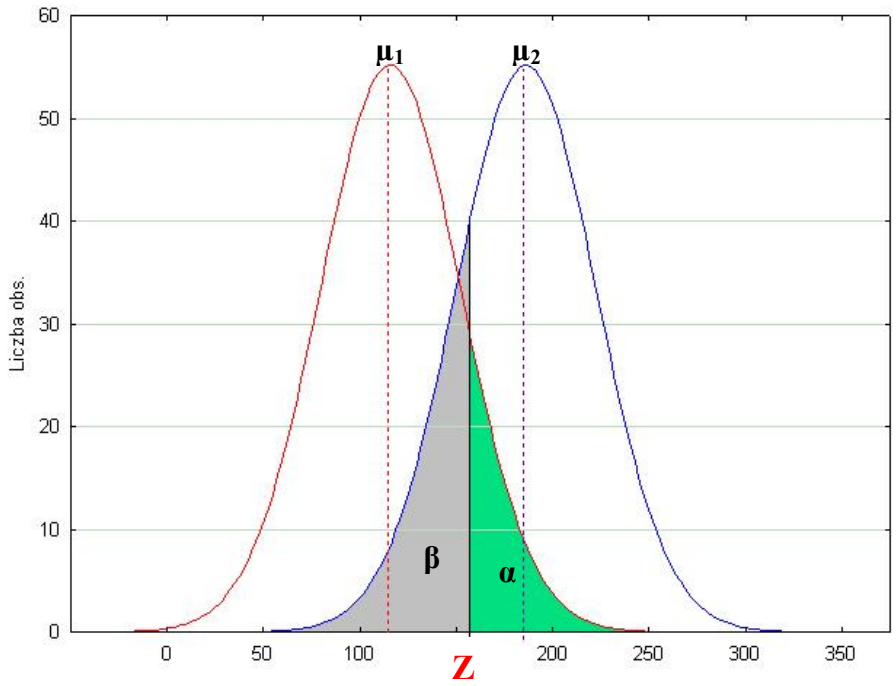
W poniższej tabeli przedstawiono zestawienie błędów popełnianych przy testowaniu hipotez.

Decyzja	H_0 prawdziwa	H_0 nieprawdziwa
Odrzucenie H_0 (= przyjęcie H_1)	**Błąd I rodzaju** (α)	**Decyzja trafna** ($1-\beta$) *Moc testu*
Brak podstaw do odrzucenia H_0	**Decyzja trafna** ($1-\alpha$) *Współczynnik trafności testu*	**Błąd II rodzaju** (β)

Prawdopodobieństwo popełnienia błędu I rodzaju określa się jako **poziom istotności** i oznacza przez **α**. Prawdopodobieństwo popełnienia błędu II rodzaju oznaczane jest symbolem **β**.
Wartości **α** i **β** są ze sobą powiązane (Ryc. 26). Przy przyjętym poziomie istotności **α** obszar krytyczny obejmuje wszystkie wartości średnie większe od **Z** (obszar zielony). Natomiast kolorem szarym zaznaczono obszar odpowiadający **β.**
Zmniejszenie **α** powoduje przesunięcie punktu Z na prawo, a tym samym zwiększenie powierzchni pola szarego czyli **β.**

Zwykle badacz przyjmuje jakiś konkretny poziom istotności. Powszechnie stosowane są poziomy 0,05 i 0,01, co w pierwszym przypadku zakłada 5-procentowe (1 raz na 20) popełnienie pomyłki przy podejmowaniu decyzji, a w drugim 1% błędu.

Decyzję o odrzuceniu lub nieodrzuceniu hipotezy zerowej na poziomie istotności **α** podejmuje się powszechnie bez odwoływania się do błędu II rodzaju. Przy dowolnej wielkości **α**, wartość **β** zależy od liczebności próby i rzeczywistej różnicy między porównywanymi wielkościami (np. średnimi). **β** jest prawdopodobieństwem nieodrzucenia H_0, gdy H_1 jest prawdziwa. Wielkość **1-β** nazywana jest mocą testu. Moc testu określa prawdopodobieństwo uznania hipotezy H_1 za prawdziwą, w sytuacji gdy jest ona rzeczywiście prawdziwa (decyzja trafna!). Czyli, na przykład uznamy, że istnieje różnica w skuteczności dwóch leków, gdy rzeczywiście ta skuteczność jest różna.



Ryc. 26 Graficzna prezentacja **b**łędów I i II rodzaju

Poziom prawdopodobieństwa p

Nie należy mylić z poziomem istotności!
Poziom prawdopodobieństwa (oznaczany jako p, p-wartość – *p-value*) jest to wyliczany w statystycznych programach komputerowych najmniejszy poziom istotności, przy którym obliczona wartość statystyki testowej prowadzi do odrzucenia hipotezy zerowej.
Na przykład jeśli wartość p wynosi 0,025, to oznacza, że hipotezę zerową odrzuca się dla każdego poziomu istotności **większego od 0,025** (np. 0.05), natomiast przyjęcie jako poziom istotności liczby mniejszej niż 0.025 (np. 0.01) nie pozwala na odrzucenie hipotezy zerowej.
Obecnie testy statystyczne wykonuje się korzystając z pakietów komputerowych i bardzo rzadko przeprowadza się je „na piechotę”. W pierwszym etapie formułuje się hipotezę zerową i przyjmuje poziom istotności **α**. Następnie wykonuje się obliczenia (z wykorzystaniem programu komputerowego) i otrzymuje wartość prawdopodobieństwa **p**.
Decyzję podejmuje się w oparciu o następujące kryteria:

- jeśli **α>p**, to na danym poziomie istotności **α** hipotezę zerową należy odrzucić (prawdziwa jest alternatywna),
- jeśli **α<p**, to na danym poziomie istotności **α** nie ma podstaw do odrzucenia hipotezy zerowej.

Ważne:
Poziom istotności **α** – to ustalona z góry (przed przeprowadzeniem testu) liczba.
Poziom prawdopodobieństwa **p** – jest zmienną losową, a nie z góry ustaloną stałą.

Istotność statystyczna a istotność praktyczna

O istotności statystycznej świadczy wartość p mniejsza niż ustalony poziom istotności α. Wynik statystycznie istotny oznacza, że różnica uzyskana w doświadczeniu jest większa do tej będącej jedynie wynikiem przypadku. Natomiast praktyczna ważność efektów otrzymanych w doświadczeniu **nie zależy!** od wielkości prawdopodobieństwa p. Bywają sytuacje, że statystycznie istotne efekty są pozbawione praktycznego znaczenia. To czy badanie ma znaczenie praktyczne zależy od wielkości efektu, możliwości jego uogólnienia czy tez innych czynników np. ekonomicznych. Na przykład zaobserwowanie statystycznie istotnie mniejszej śmiertelności po zastosowaniu określonego leku, może nie mieć praktycznego znaczenia w przypadku zbyt dużych kosztów terapii – przekraczających możliwości podmiotów odpowiedzialnych za leczenie.

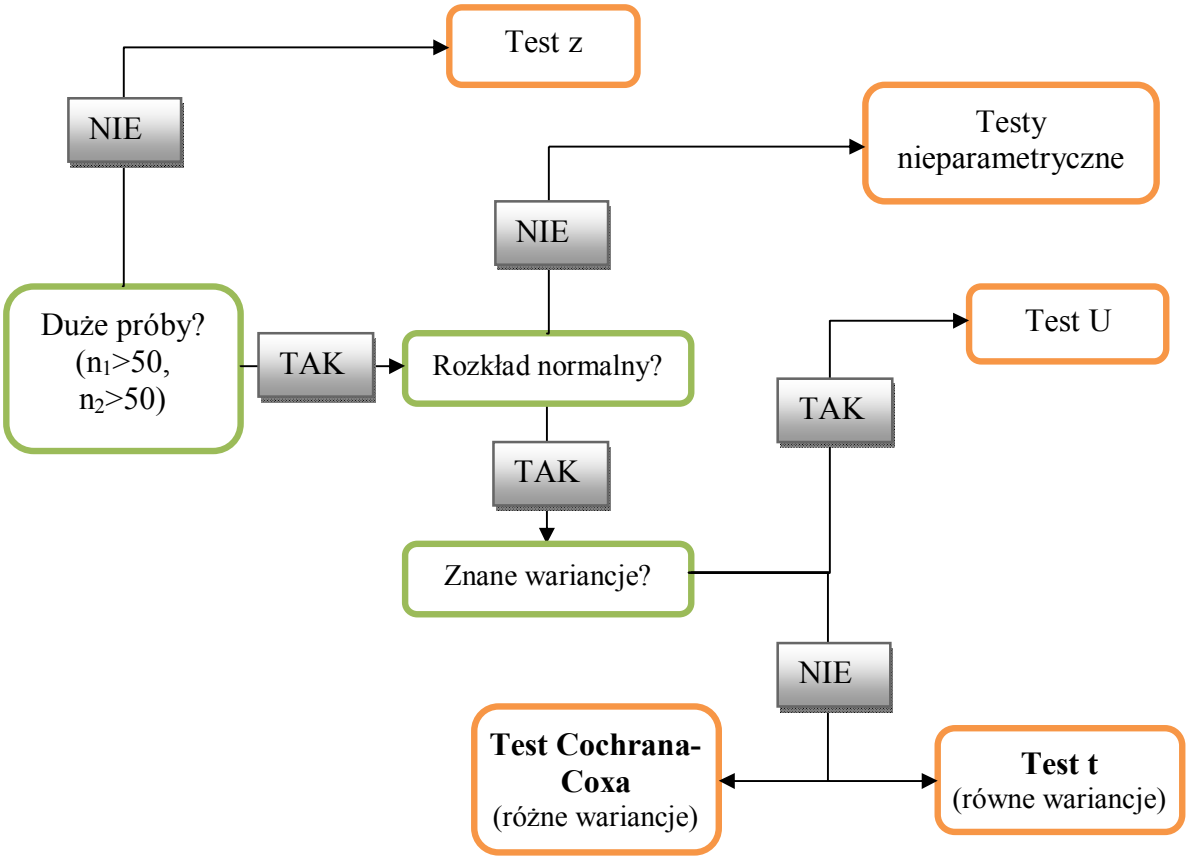
Testy istotności (testy t-Studenta)

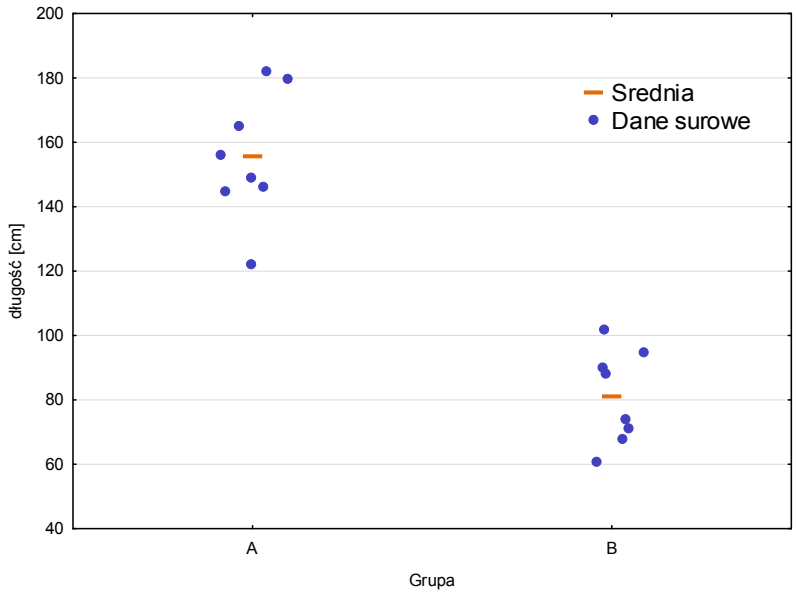
Bardzo częstym zadaniem w statystycznym testowaniu hipotez jest porównanie średnich z dwóch grup (populacji). Do tego celu można wykorzystać jeden z testów t-Studenta. Wybór odpowiedniego testu zależy od wielu czynników: liczebności próby, spełnienia założeń o normalności rozkładu i jednorodności wariancji (patrz schemat poniżej).

Kryteria wyboru testów istotności

- **Sprawdzenie założenia o normalności rozkładu.** To kryterium pozwala na decyzję czy wybierzemy któryś z testów parametrycznych (jeśli dane spełniają to założenie) lub testy nieparametryczne (jeśli dane nie podlegają rozkładowi normalnemu). Warto tutaj wspomnieć, że w wielu przypadkach proste przekształcenie danych jak np. ich zlogarytmowanie powoduje, że normalność rozkładu zostaje spełniona i można stosować testy parametryczne, które charakteryzują się większą mocą od testó○ nieparametrycznych.
- **Zmienne powiązane i niepowiązane.** Zbiór testów istotności można podzielić na dwie grupy: testy przeznaczone do testowania różnić między grupami niepowiązanymi (niezależnymi) oraz testy dla grup powiązanych (zależnych). W zależności od rozpatrywanego problemu należy wybrać odpowiedni test.
- **Założenie jednorodności wariancji.** Do sprawdzenia tego założenia można zastosować szereg testów np. test Levena, Bartletta czy test F. W przypadku braku jednorodności należy zastosować odpowiedni test – Cochrana-Coxa.

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki





Ryc. 27 Zmodyfikowany wykres ramka-wąsy

11. Analiza wariancji (ANOVA)

Zespół metod statystycznych określanych ogólnie mianem analiza wariancji (ANOVA) wykorzystuje się do testowania różnic pomiędzy co najmniej trzema grupami. Najczęściej wykorzystuje się dwa najprostsze schematy analizy wariancji:

- analiza wariancji dla klasyfikacji pojedynczej (jednoczynnikowa ANOVA),
- wieloczynnikowa analiza wariancji.

Jednoczynnikowa analiza wariancji

W tego typu analizie bada się wpływ tylko jednego czynnika klasyfikującego. Podobnie jak w przypadku testów istotności dla dwóch grup ANOVA wymaga spełnienia kilku założeń:

- analizowana zmienna jest mierzalna,
- porównywane grupy mają rozkłady normalne,
- rozkłady te mają jednakową wariancję (założenie jednorodności wariancji).

Spełnienie powyższych założeń pozwala na zastosowanie parametrycznej analizy wariancji, natomiast w przypadku niespełnienia założeń należy zastosować nieparametryczny wariant ANOVA (test Kruskala-Wallisa).

Hipotezę zerową można zapisać w postaci:

$$H_0: \mu_1 = \mu_2 = \ldots = \mu_k$$ (średnie w k-grupach są jednakowe),

wobec hipotezy alternatywnej:

$$H_1:$$ *co najmniej dwie średnie różnią się między sobą*

Prawdziwość hipotezy alternatywnej (odrzucenie hipotezy zerowej przy określonym poziomie istotności) daje nam jedynie informację, że wśród porównywanych grup co najmniej dwie różnią się statystycznie istotnie, ale nie uzyskujemy informacji ile grup faktycznie się od siebie różni i które. W celu uzyskania odpowiedzi na te pytania należy zastosować tzw. testy po fakcie (*post-hoc*).

Podstawą analizy wariancji jest możliwość rozbicia sumy kwadratów wariancji całkowitej (SS całkowita) dla wszystkich wyników obserwacji na dwie składowe:

- sumę kwadratów opisującą zmienność wewnątrz grup (SS reszta),
- sumę kwadratów opisującą zmienność między grupami (SS między grupami).

SS całkowita = SS reszta + SS między grupami

df całkowite = df grup +df reszt

df – liczba stopni swobody (*degrees of freedom)*

Źródło zmienności	Suma kwadratów	Liczba stopni swobody	Średni kwadrat	Test F
Między grupami	SS między grupami	k-1	MS pomiędzy grupami (SS/df)	**F**=MS pom. grupami/ MS reszta
Wewnątrz grup	SS reszta	n-k	MS reszta (SS/df)	
Całkowita	SS całkowita	n-1		

Jeżeli analiza wariancji nie pokaże istotności różnic między rozpatrywanymi średnimi, nie przeprowadza się już dalszych testów. Natomiast w przypadku odrzucenia hipotezy zerowej przeprowadza się testy *post-hoc.*
Do dyspozycji mamy szereg testów:

- test **NIR** (najmniejszych istotnych różnic),
- testy **Tuckeya**, **Duncana**, **Neumana-Keulsa** – oparte na studentyzowanym rozstępie, umożliwiające grupowanie średnich,
- testy **Scheffego**, **Dunneta** czy **Bonferroniego** – oparte na przedziałach ufności.

Przykład:
Poniżej zamieszczono przykładowe wyniki jednoczynnikowej analizy wariancji, z których wynika, że istnieją statystycznie istotne różnice pomiędzy zawartością lowastatyny (LOV) w różnych preparatach stosowanych w leczeniu hipercholesterolemii (p<0,05)

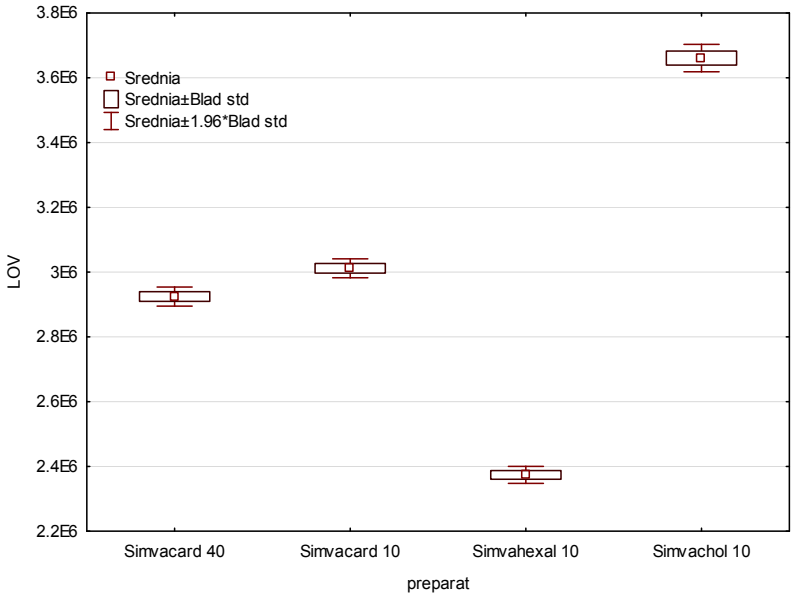
Tabela 1 Wyniki jednoczynnikowej analizy wariancji

Analiza wariancji (zaznaczone efekty są istotne z p < .05000)								
	SS - Efekt	**df - Efekt**	**MS - Efekt**	**SS - Błąd**	**df - Błąd**	**MS - Błąd**	**F**	**p**
LOV	7.11E+12	3	2.37E+12	7.59E+10	31	2.45E+09	969.27	0.0000

Kolejnym etapem jest zastosowanie któregoś z testów *post-hoc* w celu znalezienia, które preparaty różnią statystycznie istotnie. W tabeli zamieszczono wyniki testu NIR. Jak widać, wszystkie preparaty różnią się wzajemnie od siebie statystycznie istotnie.

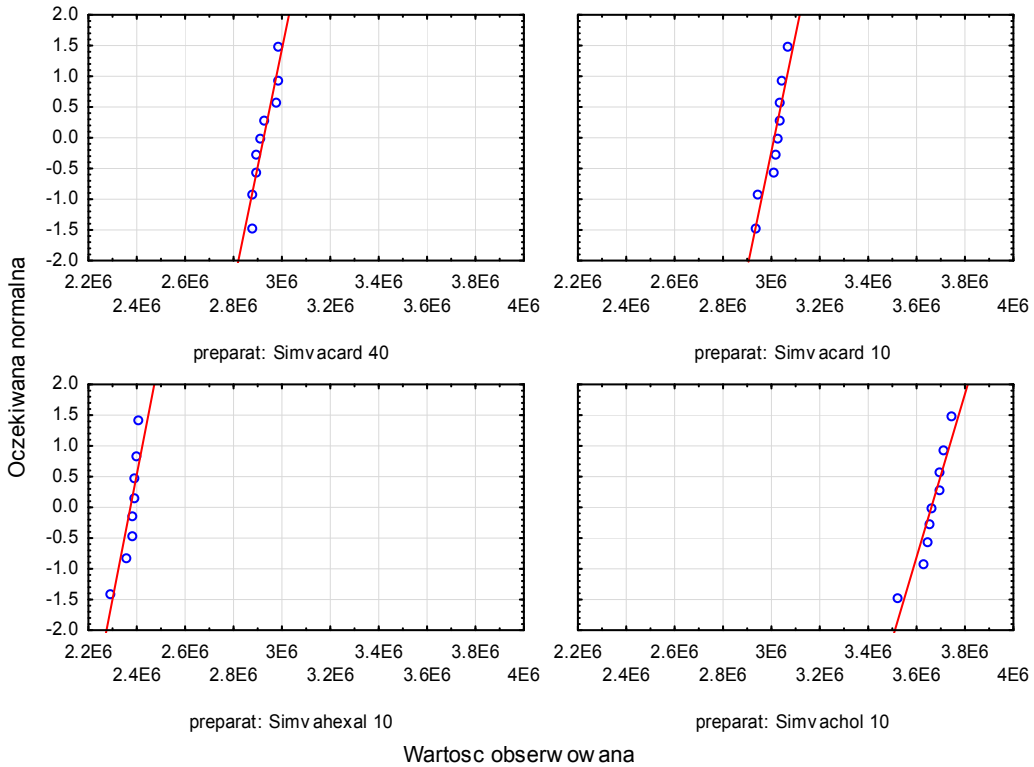
Test NIR (zaznaczone różnice są istotne z p < .05000)				
	{1}	**{2}**	**{3}**	**{4}**
Simvacard 40 {1}		0.000748	0.000000	0.000000
Simvacard 10 {2}	0.000748		0.000000	0.000000
Simvahexal 10 {3}	0.000000	0.000000		0.000000
Simvachol 10 {4}	0.000000	0.000000	0.000000	

Wniosek o odrzuceniu hipotezy zerowej potwierdza także analiza wykresu prezentującego średnie dla porównywanych grup (Ryc. 28). Wykres ten wskazuje na duże zróżnicowanie średnich i sugeruje statystycznie istotne różnice pomiędzy wszystkimi lekami.



Ryc. 28 Wykres typu ramka-wąsy porównywanych leków

Metoda ANOVA zakłada, że zmienna zależna (zawartość LOV) w obrębie grup podlega rozkładowi normalnemu. Do oceny wzrokowej spełnienia tego założenia można wykorzystać skategoryzowane wykresy normalności, na których punkty leżą bardzo blisko prostej. Należy się zatem spodziewać, że zmienna zależna w grupach podlega rozkładowi normalnemu. Można dodatkowo zastosować jeden z testów normalności np. Shapiro-Wilka.

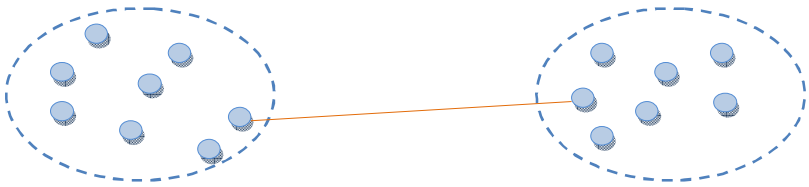


Ryc. 29 Skategoryzowany wykres normalności

12. Analiza skupień

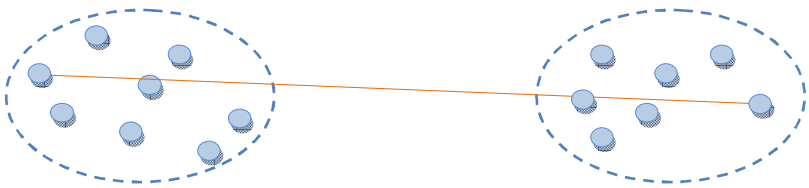
Analiza skupień jest metodą polegającą na klasyfikacji badanych obiektów do grup (skupień), tak aby powstałe skupienia zawierały w swoim obrębie obiekty jak najbardziej podobne. Liczba grup do których zostaną przyporządkowane obiekty nie jest znana *a priori* (podobnie jak w przypadku analizy głównych składowych), a przed wykonaniem analizy matematycznej nie czyni się żadnych założeń odnośnie rozkładu zmiennych.
Najpopularniejszymi metodami stosowanymi do podziału badanego zbioru na wewnętrznie jednorodne skupiska są: metody hierarchiczne oraz grupowanie metodą k-średnich. Rezultatem metod hierarchicznych jest prezentacja wyników w postaci dendrogramów, a techniki służące do ich uzyskania można podzielić na aglomeracyjne oraz podziałowe. Techniki te w zasadniczo różny sposób traktują „początkowy” zbiór obiektów. Procedury aglomeracyjne łączą ze sobą przypadki (będące pierwotnie jednoelementowymi skupieniami) leżące najbliżej siebie w hiperprzestrzenni zmiennych, aż do uzyskania monolitycznych grup. Rozpatrywanie wszystkich obiektów jako jedno skupienie (bardzo zróżnicowane), a następnie rozszczepianie go na coraz to mniejsze, bardziej homogenne, jest charakterystyczne dla technik podziałowych, są one jednak rzadziej stosowane. Jak wspomniano wyżej techniki aglomeracyjne składają się z szeregu cykli polegających na integrowaniu ze sobą najbliżej siebie położonych skupisk utworzonych w poprzednim etapie.

Metoda pojedynczego wiązania (ang. *single linkage metod*) zwana również *metodą najbliższego sąsiedztwa.* W tej metodzie odległość między skupiskami (A i B) równa jest odległości pomiędzy dwoma najbliższymi obiektami, pochodzących z różnych skupisk.



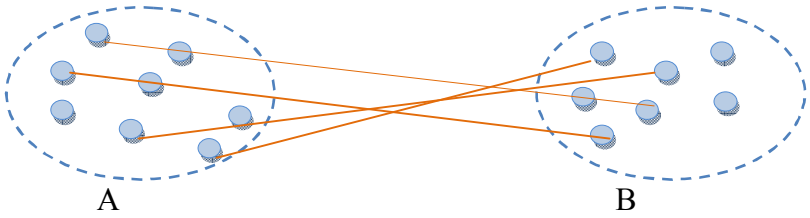
A B

Metoda pełnego wiązania (ang. *complete linkage*) zwana również *metodą najdalszego wiązania.* Jest przeciwieństwem poprzedniej metody, odległość między skupiskami (A i B) równa jest największej odległości pomiędzy dwoma obiektami leżącymi, pochodzących z różnych skupisk.



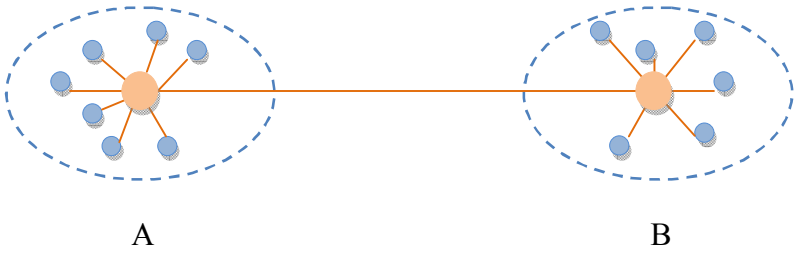
A B

Metoda średnich połaczeń (ang. *unweighted pair-group metod using arithmetic averages* – UPGMA). W metodzie tej odległość między skupiskami (A i B) równa jest średniej arytmetycznej wszystkich odległości pomiędzy obiektami należącymi do różnych skupisk.



Metoda średnich połączeń ważonych (ang. *weighted pair-group metod using arithmetic averages* – WPGMA). Stosowana jest, jeżeli istnieje podejrzenie, że skupiska różnią się liczebnością. Metoda analogiczna do metody średnich połączeń z ta różnicą, że szacuje się wagę obrazująca wielkość grupy.

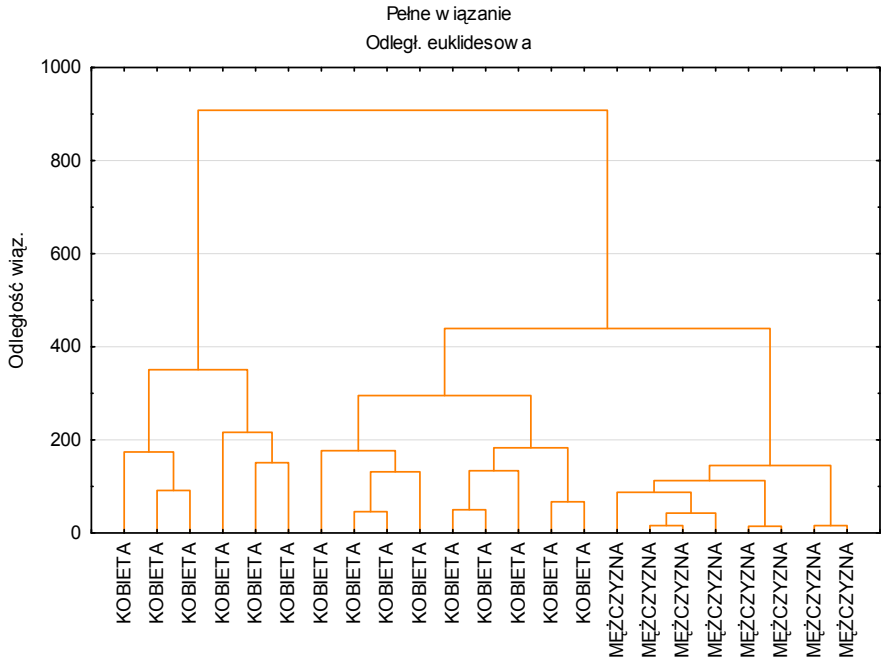
Metoda środków ciężkości (ang. *unweighted pair-group metod using the centroid average* – UPGMC). Odległość między skupieniami (A i B) równa jest odległości pomiędzy środkami ciężkości skupisk, które są ich punktami średnimi w przestrzeni wielowymiarowej.



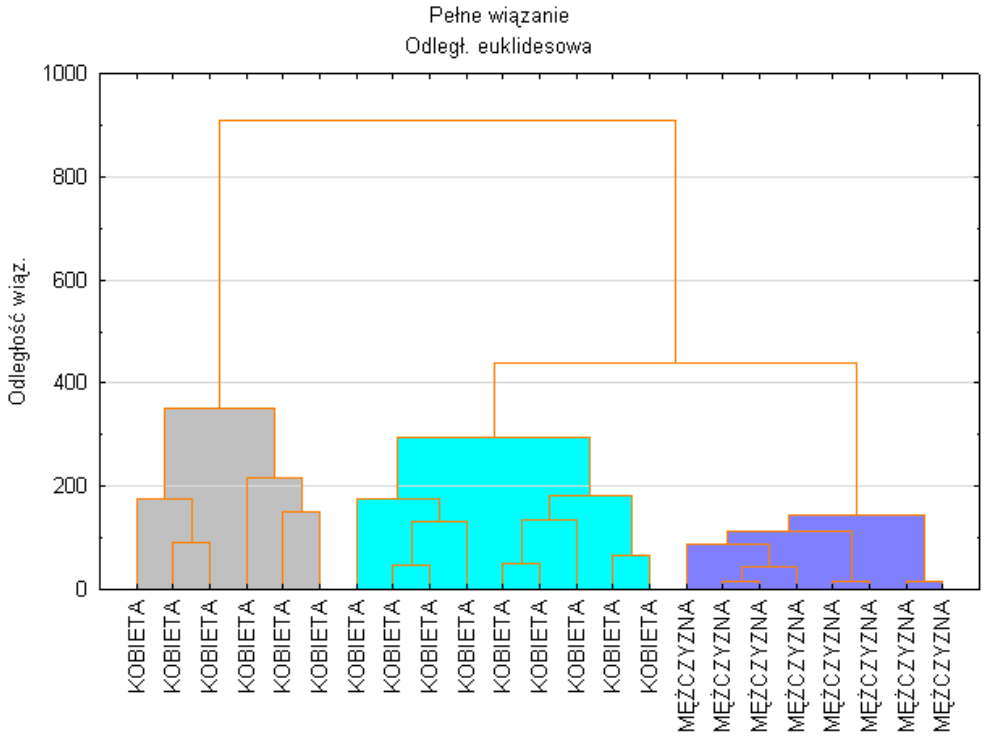
Metoda ważonych środków ciężkości (*mediany*) (ang. *weighted pair-group metod using the centroid average* – WPGMC). Metoda analogiczna z wyżej wymienioną, z tą różnicą, że w obliczeniach uwzględnia się wagę obrazującą zawartość obiektów w grupie.
Metoda Warda. Odległość między skupiskami szacowana jest na podstawie analizy wariancji. Mimo, że utworzone skupienia charakteryzują się niewielką ilością obiektów, metoda daje bardzo dobre efekty.

Przykład
Analiza danych dotyczących czasu poświęcanego przez grupę mężczyzn i kobiet na różne aktywności: pracę, opiekę nad dziećmi, transport, prace domowe, zakupy, sen, higienę osobistą, oglądanie telewizji itp.

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki

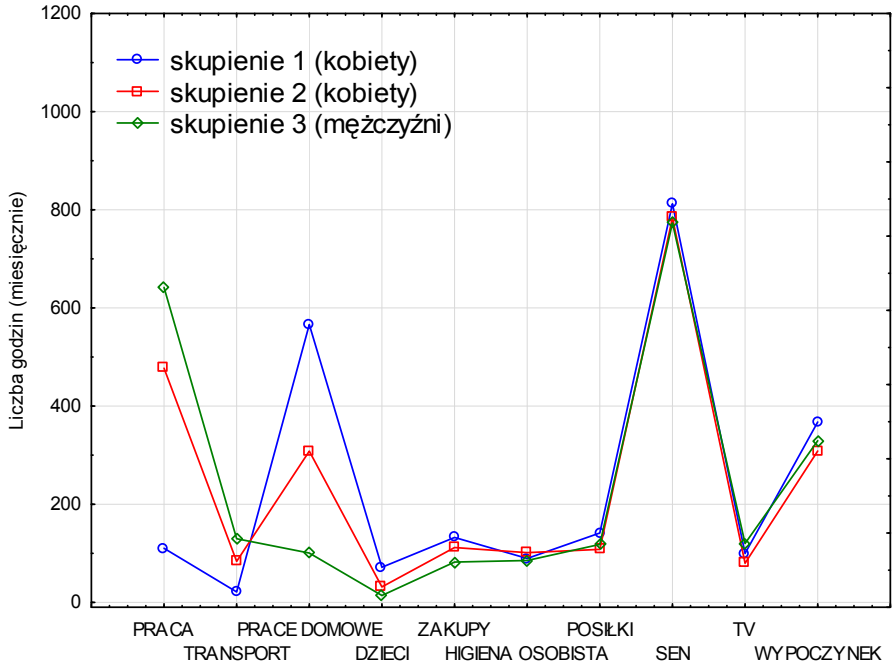


Ryc. 30 Dendrogram wygenerowany w analizie skupień metodą aglomeracyjną

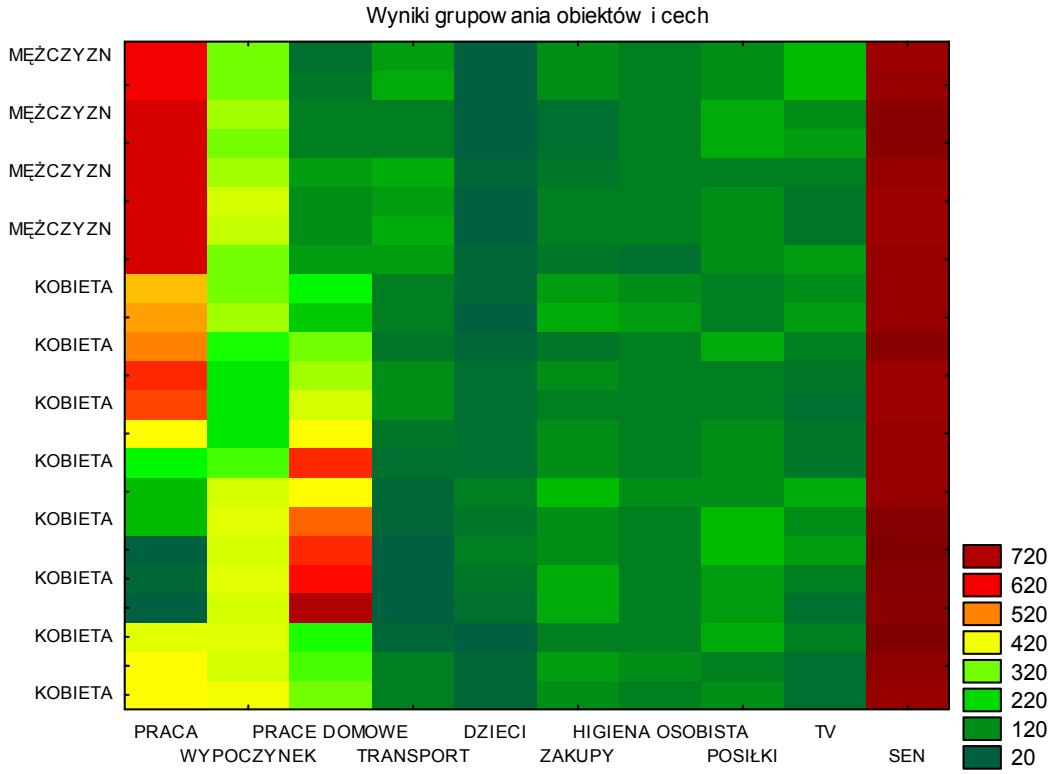


Ryc. 31 Dendrogram zmodyfikowany – zaznaczone skupienia

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki

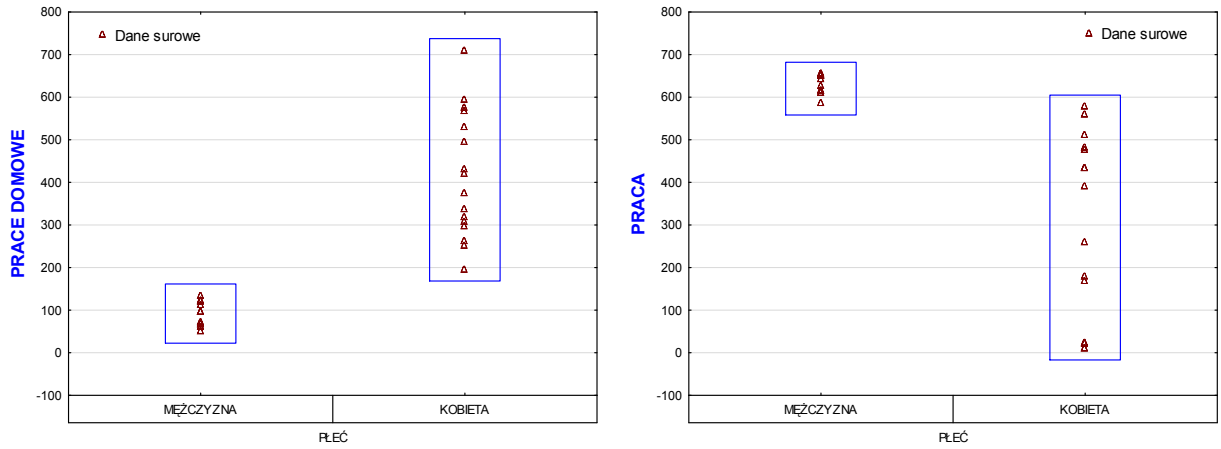


Ryc. 32 Wykres ilustrujący wyniki analizy skupień metodą k-średnich

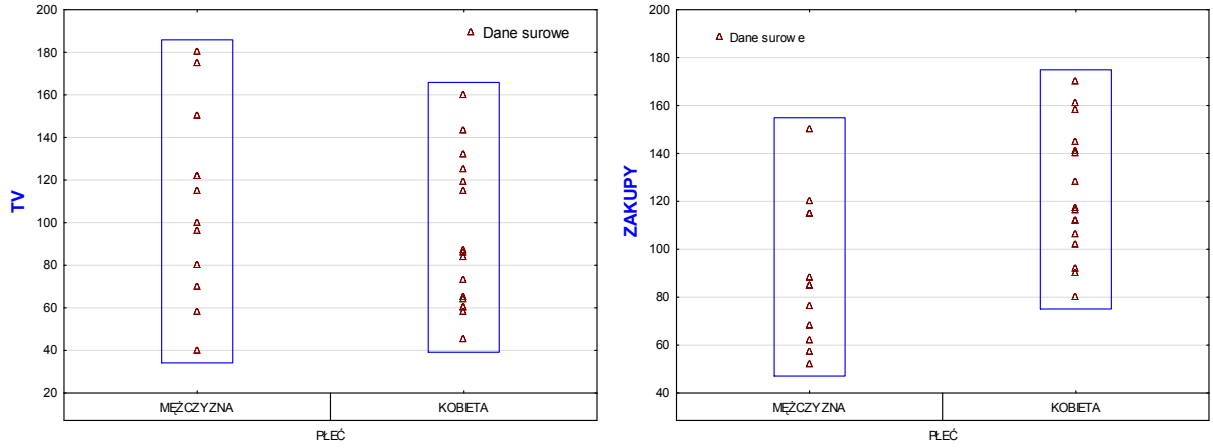


Ryc. 33 Mapa grupująca obiekty ich cechy

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki



Ryc. 34 Wykresy zmienności dla zmiennych PRACE DOMOWE i PRACA jako przykład danych o skrajnie różnym rozrzucie



Ryc. 35 Wykresy zmienności dla zmiennych TV i ZAKUPY jako przykład danych o podobnym rozrzucie

13. Analiza głównych składowych (PCA)

Zjawiska będące przedmiotem badań biologicznych i medycznych charakteryzuje bardzo duża złożoność, co oznacza, że do ich opisu wykorzystuje się mnóstwo cech, a efektem tego jest ogromna liczba zmiennych (od kilkudziesięciu, nawet do kilku tysięcy). Mamy wówczas do czynienia z wielowymiarową przestrzenią danych, których eksploracja wymaga odpowiednich metod. Jedną z częściej stosowanych metod jest analiza głównych składowych, której zadaniem jest wyrażenie wielowymiarowych obserwacji przy użyciu jak najmniejszej liczby zmiennych (redukcja zmiennych) przy zachowaniu pewnych relacji między nimi.

Analiza głównych składowych pozwala na:

- redukcję liczby zmiennych,
- wykrycie struktury i prawidłowości w związkach miedzy zmiennymi,
- opis i klasyfikację obserwacji (obiektów) w nowej przestrzeni zdefiniowanej przez nowe zmienne (tzw. czynniki lub główne składowe).

Ideę analizy czynnikowej, w tym analizy głównych składowych można wyjaśnić na bardzo prostym przykładzie p**ołączenia dwóch zmiennych w jeden czynnik.** Korelację między dwiema zmiennymi można przedstawić na wykresie rozrzutu (Ryc. 36) i dopasować linię, która "najlepiej" oddaje związek liniowy między tymi zmiennymi **X'**. Następnie dokonuje się takiego obrotu układu współrzędnych, aby jedna z osi (np. **X**) pokrywała się z linią wyrażająca maksymalną wariancję chmury punktów czyli **X'**. Pojedyncze wartości obiektów (punktów) na takiej nowej osi (nowym czynniku, składowej głównej) reprezentowanym przez linię regresji, mogą zostać wykorzystanie w zastępstwie dwóch poprzednich współrzędnych każdego punktu czyli **X** i **Y**. W pewnym sensie nastąpiła redukcja dwóch zmiennych (**X** i **Y**) do jednego czynnika (**X'**). Należy zauważyć, że nowy czynnik jest w rzeczywistości kombinacją liniową tych dwóch zmiennych.

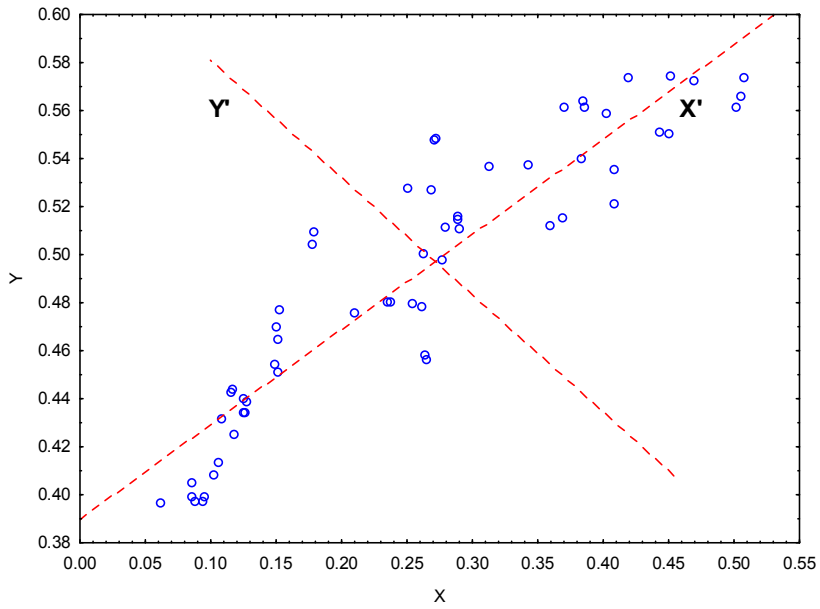
Wyodrębnianie składowych głównych jest równoznaczne z **rotacją maksymalizującą wariancję** (varimax) wyjściowej przestrzeni zmiennych. Oczywiście należy mieć świadomość, że w przykładzie pokazanym na Ryc. 36 użycie tylko jednej nowej zmiennej X' do opisu wszystkich punktów powoduje utratę części informacji (nie wszystkie punkty leżą na linii lub bardzo blisko niej). Niemniej jednak w ten sposób można dokonać redukcji dwóch zmiennych (X,Y) do jednej X'.

Opisany przykład łączenia dwóch skorelowanych zmiennych w jeden czynnik pokazuje podstawowa ideę analizy głównych składowych, który można uogólnić na wiele zmiennych. Oczywiście obliczenia komplikują się, ale zasada zastępowania dwóch lub więcej zmiennych przez jeden czynnik pozostaje taka sama.

Można zatem zbiór **n** zmiennych można zredukować, z niewielką stratą informacji, do zbioru **k** (**k<n**) zmiennych. Po wybraniu pierwszej nowej zmiennej – osi wprowadza się kolejną oś **Y'**(czynnik) nieskorelowaną z pierwszą (ortogonalną) - Ryc. 36.

Każda składowa główna (czynnik) jest zatem kombinacją liniową zmiennych wyjściowych i jest definiowana tak, aby w możliwie największym stopniu wyjaśniać zmienność oryginalnych zmiennych. Miarą zmienności w PCA jest wariancja.

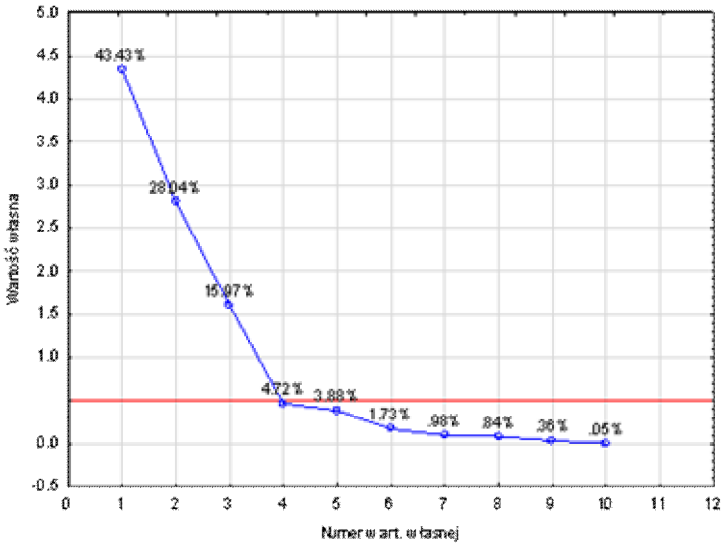
Wymiar nowej przestrzeni (czyli wybór liczby składowych głównych) zależeć będzie od tego w jakim stopniu chcemy układem nowych zmiennych wyjaśnić zmienność oryginalnych zmiennych. Każda kolejna składowa wyjaśnia coraz mniejszą część zmienności oryginalnej.



Ryc. 36 Wykres rozrzutu z nowym układem współrzednych

W praktyce stosowane są trzy kryteria wyboru optymalnej liczby głównych składowych (czynników):

- kryterium Kaisera – wykorzystuje się tylko te składowe, dla których wartości własne (eigenvalue) **są większe od 1**. W Tabela 2, podanej w opisanym niżej przykładzie, kryterium Kaisera spełniają dwie pierwsze główne składowe (PC1 i PC2),
- wykres osypiska (Ryc. 37) – na wykresie pokazującym kolejne wartości własne szuka się punktu, od którego zaczyna się łagodny spadek czyli nie ma znaczącego przyrostu informacji (tzw. osypisko czynnikowe). Istotne znaczenie mają tylko czynniki znajdujące się po lewej stronie tego punktu,



Ryc. 37 Wykres osypiska

- trzecie kryterium jest bardziej arbitralne – bierze się pod uwagę procent wariancji wyjaśnionej przez daną główną składową. Jeśli dla pierwszych dwóch czy trzech głównych składowych sumaryczna wariancja wyjaśnia co najmniej 75% wariancji wszystkich obserwowanych zmiennych to można na tym poprzestać.

Założenia analizy głównych składowych

Zmienne wykorzystywane w analizie głównych składowych powinny być mierzalne w skali co najmniej przedziałowej i pozostawać ze sobą w **liniowej zależności**. Jeśli zmienne nie są ze sobą powiązane, to każda z nich będzie stanowić osobną składową i nie uzyskamy redukcji zmiennych. Jeśli redukcja zmiennych ma być skuteczna korelacje między zmiennymi, wyrażone za pomocą współczynnika korelacji liniowej Pearsona r, powinny być >0,30. Im wyższe korelacje tym bardziej uzasadnione jest stosowanie analizy głównych składowych.

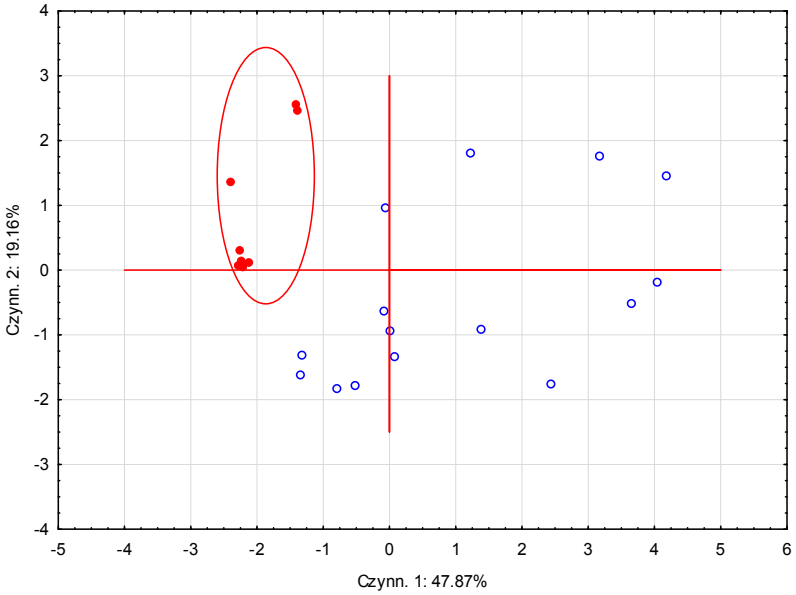
Założenie **normalności rozkładu** zmiennych nie jest konieczne jeśli PCA stosuje się do opisu struktury i powiązań w dużym zbiorze danych. Natomiast jeśli chcemy zastosować wnioskowanie statystyczne do określenia istotności składowych, konieczne jest spełnienie założenia o wielowymiarowej normalności.

W celu otrzymania wiarygodnych wyników badana próba musi mieć określoną **liczebność**. Minimum stanowi próba o liczebności 50 obserwacji (obiektów). Inna reguła mówi, że na każdą zmienną powinno przypadać co najmniej pięć obserwacji. Kolejne założenie, które musi być spełnione to **reprezentatywność** próby.

Analiza głównych składowych (jak i korelacja liniowa, na której się opiera) jest czuła na **obserwacje odstające**. Przed rozpoczęciem analizy takie punkty powinny zostać wykryte i usunięte.

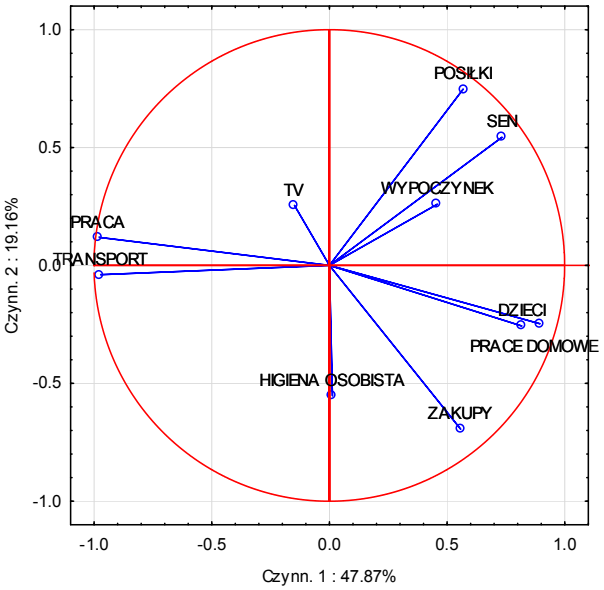
Wizualizacja wyników PCA

Wynikiem analizy głównych składowych jest utworzenie nowego zestawu zmiennych ortogonalnych (osi układu współrzędnych) i wyliczenie współrzędnych wszystkich obiektów (przypadków) w takim układzie. Są to tzw. współrzędne czynnikowe przypadków. Dla punktów o takich współrzędnych można utworzyć wykres rozrzutu (Ryc. 38) – dla pierwszych dwóch głównych składowych.



Ryc. 38 Projekcja obiektów na płaszczyznę dwóch pierwszych głównych składowych (czynników)

Podobny wykres można utworzyć na podstawie współrzędnych czynnikowych zmiennych (Ryc. 39).



Ryc. 39 Projekcja zmiennych na płaszczyznę dwóch pierwszych głównych składowych (czynników)

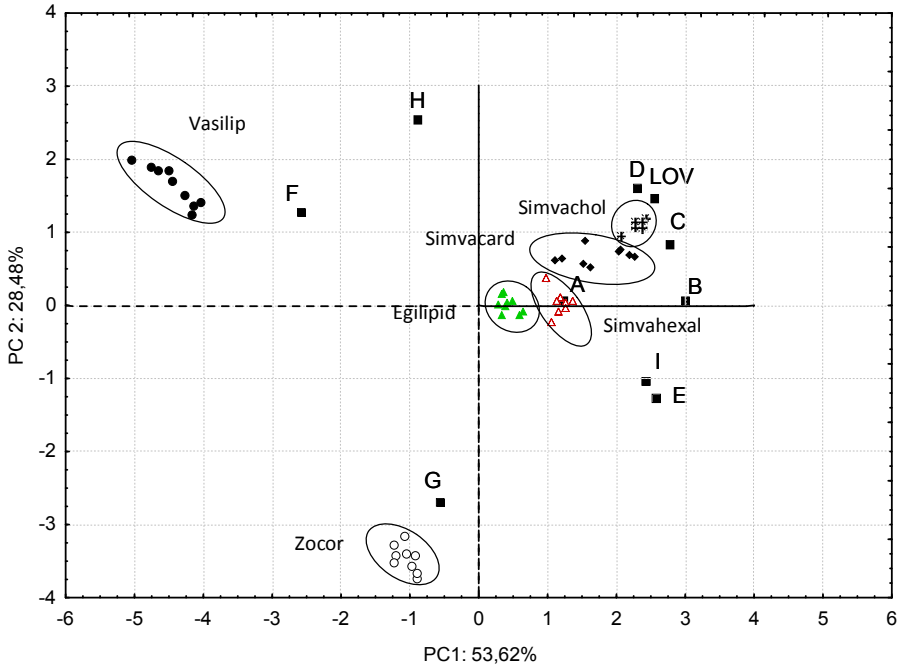
Punkty reprezentują wartości współczynników korelacji pomiędzy surowymi zmiennymi a składowymi i nie przekraczają 1 (czerwone - tzw. koło jednostkowe). Rzutowanie zmiennych na koło daje informacje, które zmienne są skorelowane z danym czynnikiem (główną składową) oraz ze sobą. Im dalej od środka (bliżej koła) znajduje się punkt (ładunek), tym wyższa korelacja danej zmiennej z osią (główną składową). Interpretację danej głównej składowej należy odnosić do zmiennych jak najsilniej z nią skorelowanych. Z kolei im bliżej

siebie położone są dwa punkty tym silniej dodatnio skorelowane są zmienne reprezentowane przez te punkty (wektory punktów są równoległe). Jeśli dwa wektory są prostopadłe to oznacza brak korelacji pomiędzy zmiennymi. A jeśli wektory leżą po przeciwnej stronie koła to zmienne skorelowane są ujemnie.

Wykres ten, w powiązaniu z poprzednim (tzw. **biplot**), pozwala na interpretację struktury zbioru danych i wzajemnego powiązania badanych zmiennych (przykład poniżej).

Przykład

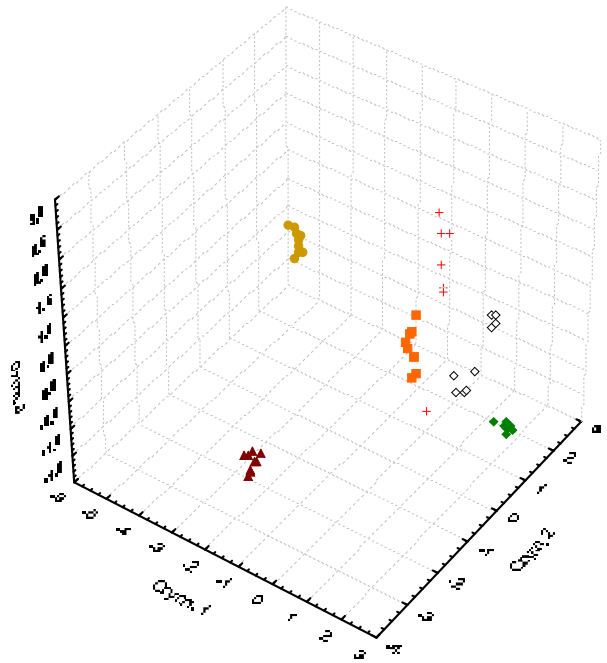
Wykonano analizę chromatograficzną (HPLC) ekstraktów z tabletek zawierających jako substancję czynną simwastatynę – lek z grupy statyn stosowany w leczeniu hipercholesterolemii. Z chromatogramów wybrano 10 pików – zanieczyszczeń badanych preparatów.



Ryc. 40 BIPLOT - projekcja **zmiennych** i **przypadków** na płaszczyznę dwóch pierwszych głównych składowych (czynników)

Tabela 2 Ładunki czynnikowe zmiennych dla trzech pierwszych głównych składowych

Składnik	PC1	PC2	PC3
A	0.398313	0.016414	0.904122
B	**0.980166**	0.017735	-0.036170
C	**0.903929**	0.288204	0.000379
LOV	**0.834214**	0.504709	-0.142125
D	**0.749452**	0.554528	-0.282577
E	**0.842202**	-0.450163	0.052250
F	**-0.845848**	0.437946	-0.034110
G	-0.185711	**-0.947760**	-0.161116
H	-0.293204	**0.880328**	0.073361
I	**0.789049**	-0.366764	-0.096244
Wartość własna (Eigenvalue)	5.36188	2.84806	0.96329
% całkowitej wariancji	53.6188	28.4806	9.6329
Wariancja skumulowana	**53.6188**	**82.0994**	**91.7323**



Ryc. 41 Wykres 3W ładunków czynnikowych w układzie współrzędnych wyznaczonych przez trzy pierwsze główne składowe.

14. Najczęściej spotykane błędy w publikacjach z zakresu nauk przyrodniczych [6]

Niepoprawne stosowanie statystyki opisowej

Dwie najczęściej stosowane statystyki opisowe dla zmiennych ciągłych to średnia arytmetyczna i odchylenie standardowe. Trzeba jednak pamiętać, że statystyki te poprawnie opisują tylko zmienne podlegające rozkładowi normalnemu. Z definicji rozkładu normalnego wynika, że w przedziale średnia +/- 1 odchylenie standardowe zawartych jest 68% wartości zmiennej, w przedziale średnia +/- 3 odchylenia standardowe 99% wartości. Przy rozkładzie zmiennej odbiegającym wyraźnie od rozkładu normalnego te zależności przestają być prawdziwe i średnia wraz z odchyleniem standardowym nie opisują prawidłowo własności takiej zmiennej. W takiej sytuacji zalecane jest stosowanie innych statystyk, jak mediana i kwartale.
Inny problem w stosowaniu statystyki opisowej to wyliczanie odchylenia standardowego z **dwóch(!)** wartości, co jest oczywiście matematycznie możliwe, natomiast nie opisuje wystarczająco dobrze prób **o małej liczności.**
Dodatkowo wiele danych biologicznych charakteryzuje się rozkładem odbiegającym od normalnego i z tego powodu mediana i kwartale powinny być stosowane częściej.

Dzielenie wartości zmiennej ciągłej na niewłaściwe kategorie

W celu uproszczenia analizy statystycznej, zmienną ciągłą np. wzrost często dzieli się na arbitralnie wybrane kategorie np. niski, średni, wysoki. W takich sytuacjach należy:

- wyjaśnić powody dlaczego utworzono taką, a nie inną liczbę klas,

Niesprawdzanie założeń testów statystycznych

Stosowanie metod parametrycznych (np. testów t, ANOVA czy regresji liniowej) w przypadku gdy dane nie były zweryfikowane pod kątem spełnienia założenia o normalności rozkładu.

Niepodawanie wartości p

Wartość p jest bardzo często mylnie interpretowana lub opisywana w sposób niewystarczający.
Przykład: ocena działanie leku obniżającego poziom cholesterolu we krwi.

- „**Działanie leku było statystycznie istotne**” – to zdanie właściwie nic nie mówi, ponieważ nie wiadomo na jakim poziomie istotności wykazano różnicę i czy efekt może mieć znaczenie praktyczne,

- „**Działanie leku obniżającego poziom cholesterolu we krwi było statystycznie istotne (p<0,05)**” – nie wiadomo nadal jak duży był efekt, np. p=0,048 daje statystycznie

istotną różnicę, ale klinicznie jest to bez znaczenia (jeśli oczywiście celem jest znalezienie leku działającego znacznie skuteczniej).

- **Średni poziom cholesterolu we krwi w grupie poddanej działaniu leku obniżył się z 220 mg do 190 mg (p=0,017)** - ten rodzaj opisu wyników analizy statystycznej jest chyba najbardziej typowy w literaturze medycznej. Podana jest wyjściowa i końcowa wartość, ale brak jest informacji o różnicy (choć można ją policzyć w pamięci) i jej niepewności np. w postaci przedziału ufności.

- **Średni poziom cholesterolu we krwi w grupie poddanej działaniu leku obniżył się o 30 mg, z 220 mg do 160 mg (95%, CI=13 do 47 mg, p=0,017)** - podany przedział ufności różnicy między średnimi daje informację, że jeżeli badany lek będzie testowany na grupie 100 osób, to u 95 z nich średni spadek stężenia cholesterolu będzie się zawierał w przedziale od 13 do 47 mg. Wartość 13 może być nieistotna z klinicznego punktu widzenia, natomiast obniżenie stężenia cholesterolu o 43 mg może mieć znaczenie praktyczne.
  Jeśli wszystkie wartości leżące w przedziale ufności są klinicznie istotne wówczas taki lek ma większe szanse osiągnąć sukces w postaci skutecznej terapii.

Interpretowanie statystycznie nieistotnych wyników jako negatywnych

Często autorzy stwierdzają, że brak statystycznie istotnych różnic jako wynik negatywny, podczas gdy jest on po prostu nierozstrzygający. Moc testu statystycznego to zdolność do wykrywania różnic w grupach o odpowiedniej liczebności, jeśli taka różnica rzeczywiście istnieje. W badaniach gdzie wykorzystuje się testy o małej mocy lub mało liczne grupy, brak statystycznie istotnych różnic nie jest wynikiem negatywnym, a tylko nie pozwala na pożądane rozstrzygnięcie.

Absence of evidence is not evidence of absence!!!

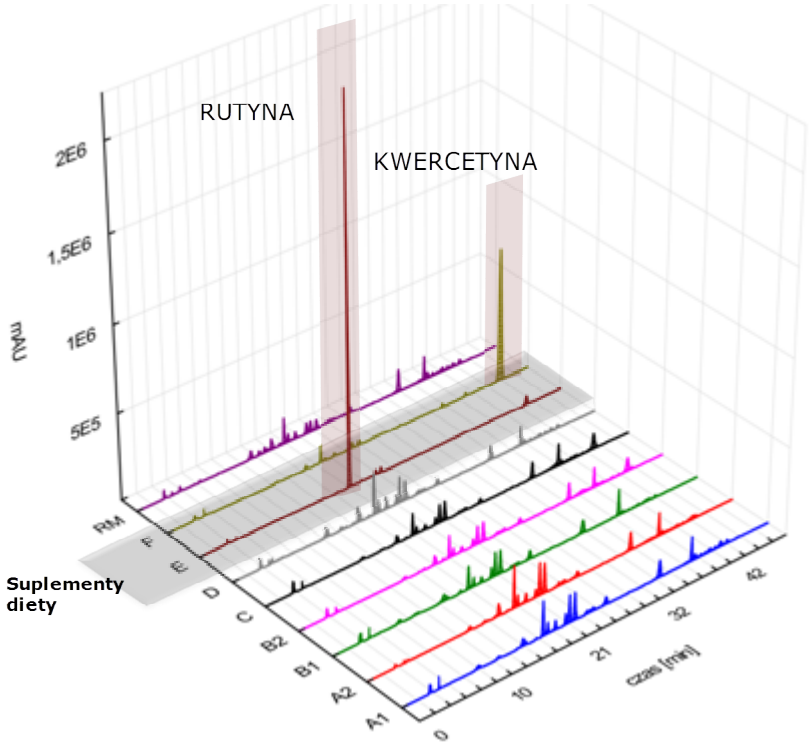
Spodziewana różnica	Wymagana liczebność grup*
5%	1450-3200
10%	440-820
20%	140-210
30%	80-100
40%	50-60

*poziom istotności 0,05; moc testu 80%

15. Przykłady analizy chemometrycznej

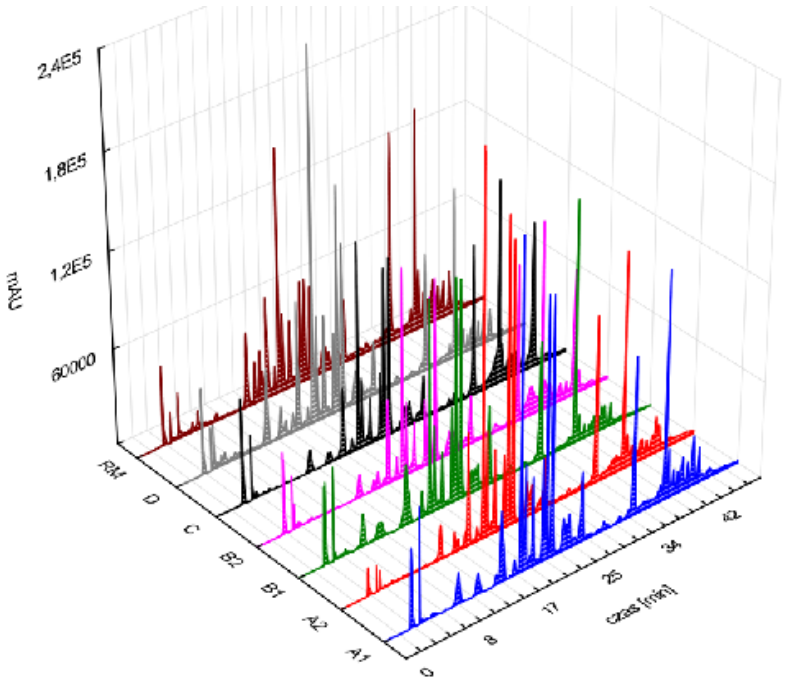
Ocena aktywności antyoksydacyjnej leków i suplementów diety zawierających wyciąg z miłorzębu japońskiego Gingko biloba [7]

Preparat	zawartość ekstraktu w 1 kaps./tabl. [mg]	kategoria produktu (postać)
A1	40	produkt leczniczy (kapsułki)
A2	80	produkt leczniczy (kapsułki)
B1	40	produkt leczniczy (tabletki)
B2	80	produkt leczniczy (tabletki)
C	40	produkt leczniczy (tabletki)
D	40	produkt leczniczy (tabletki)
E	80	suplement diety (kapsułki)
F	80	suplement diety (tabletki)
RM	-	suchy ekstrakt z liści *Ginkgo biloba* (materiał referencyjny)

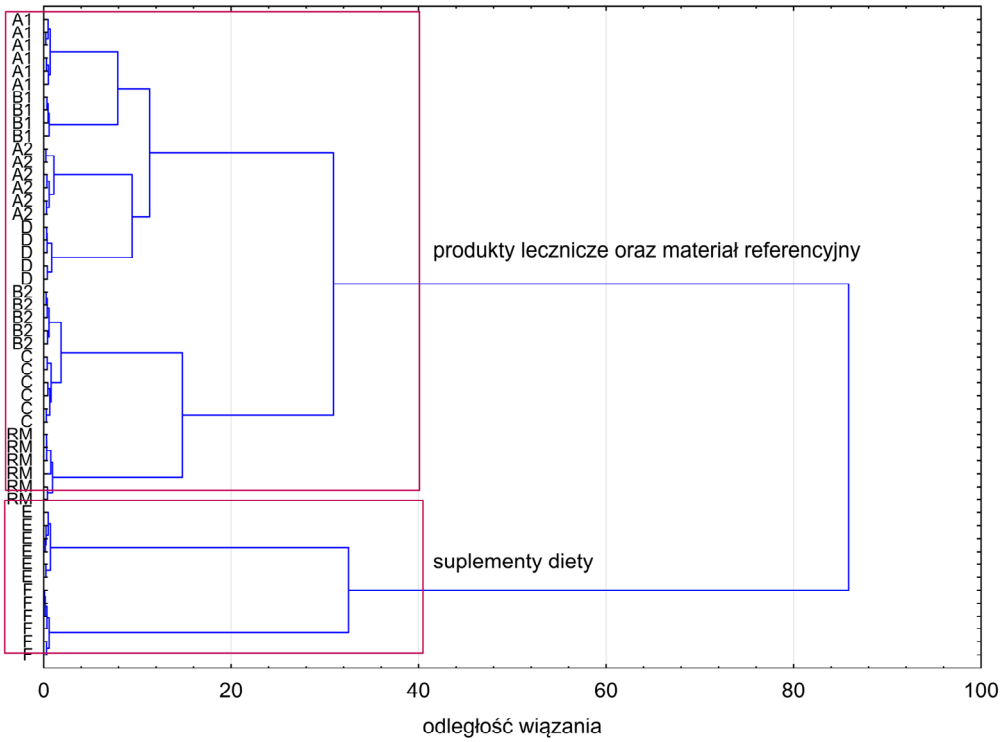


Ryc. 42 Chromatogram badanych preparatów. Dwa z nich zostały zafałszowane dodatkiem rytyny (suplement diety **E**) oraz kwercetyny (suplement diety **F**)

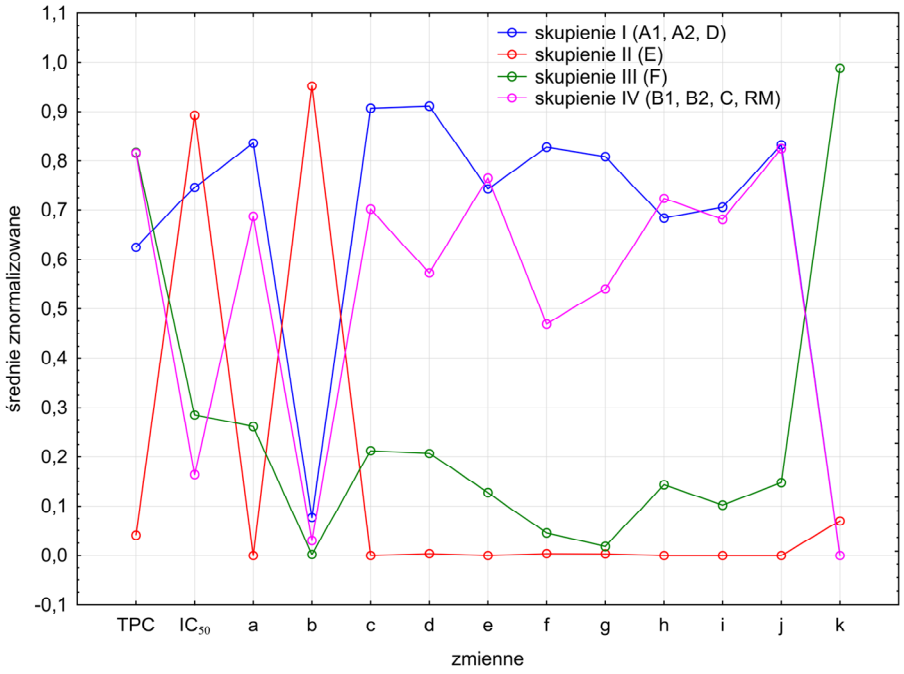
Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki



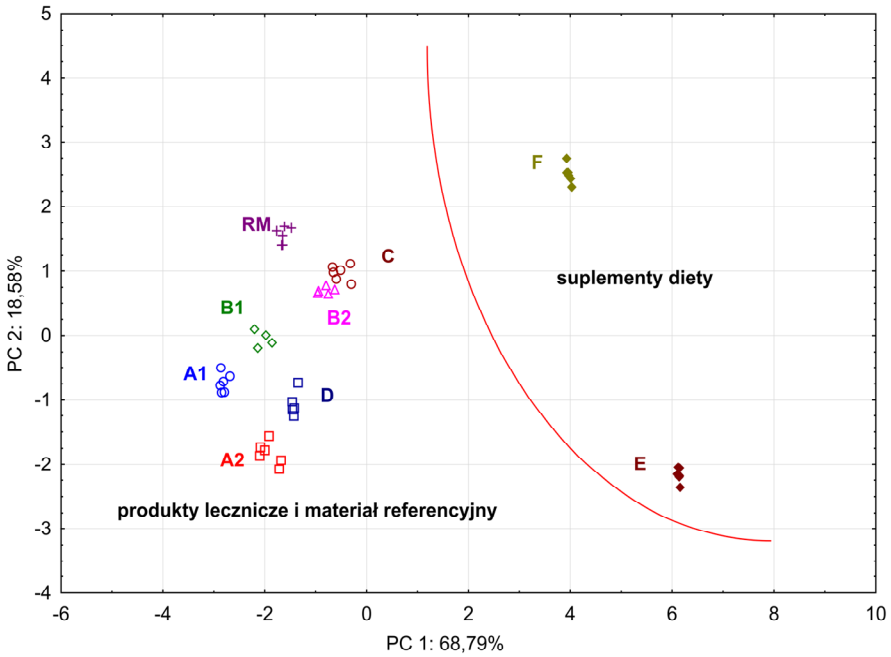
Ryc. 43 Chromatogramy ekstraktów badanych preparatów po usunięciu suplementów diety



Ryc. 44 Wynik analizy skupień metodą hierarchiczną (metoda Warda, kwadrat odległości Euklidesa) Zmienne: TPC, IC_{50} (DPPH$^{\bullet}$), pola powierzchni 11 pików wybranych z chromatogramu

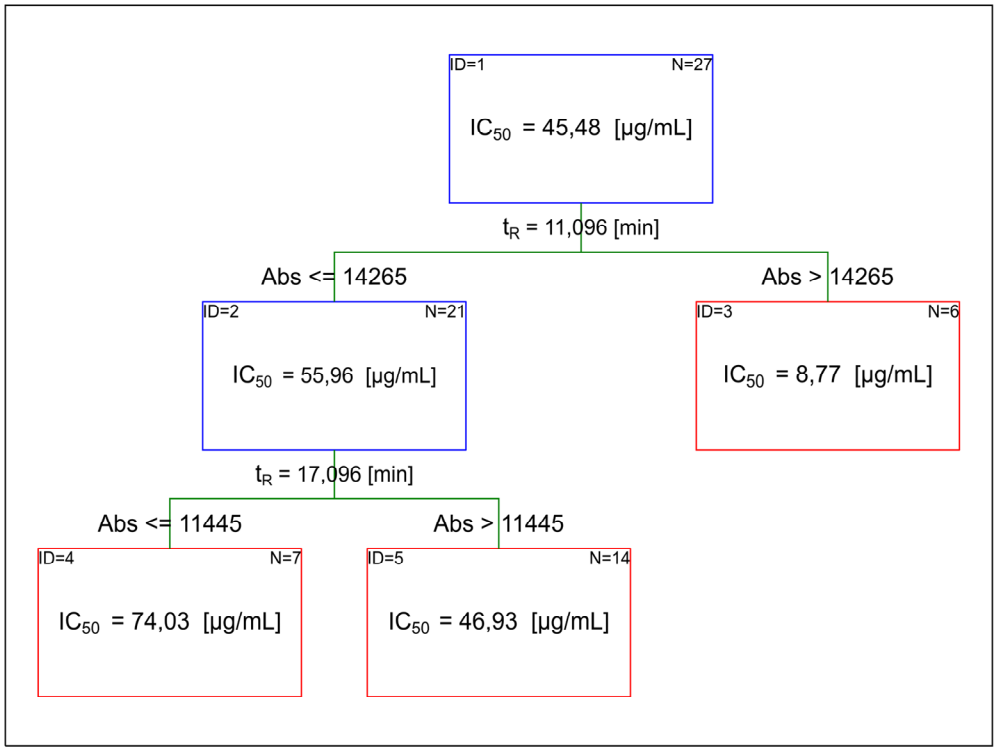


Ryc. 45 Wynik analizy skupień metodą k średnich. Zmienne: TPC, IC_{50} ($DPPH^{\bullet}$), pola powierzchni 11 pików wybranych z chromatogramu.

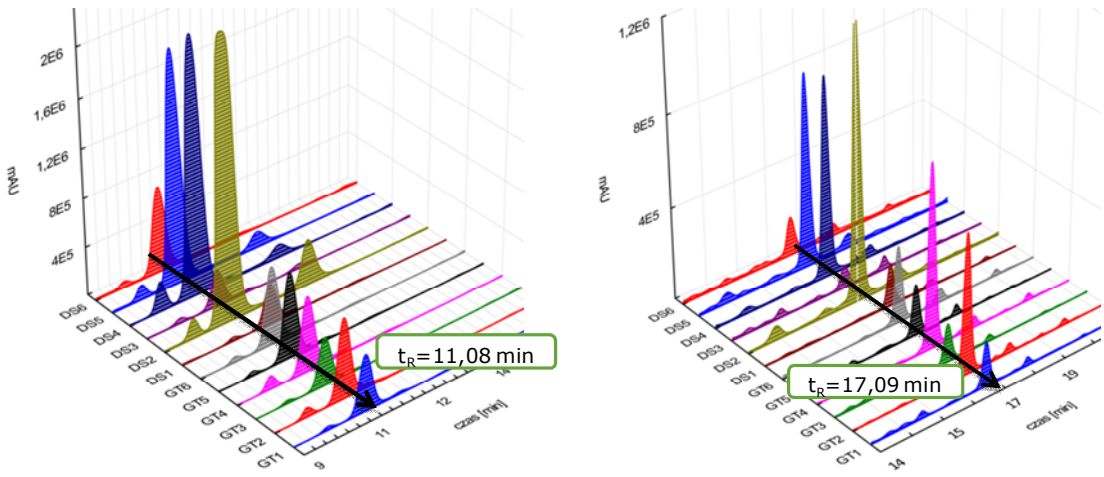


Ryc. 46 Wyniki analizy skupień – rzut przypadków na płaszczyznę wyznaczoną przez dwie pierwsze główne składowe

Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki

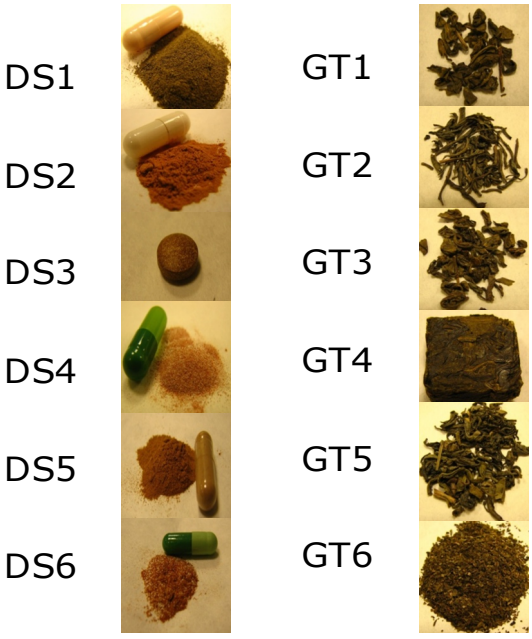


Ryc. 47 Model regresyjny Reguły decyzyjne wskazujące wartość $IC_{50(ABTS)}$ badanego ekstraktu w oparciu o chromatogram *fingerprint*. Poniżej piki na chromatogramie, na podstawie których zbudowano reguły decyzyjne.



Ocena aktywności antyoksydacyjnej zielonych herbat i suplementów diety zawierających wyciąg z zielonej herbaty

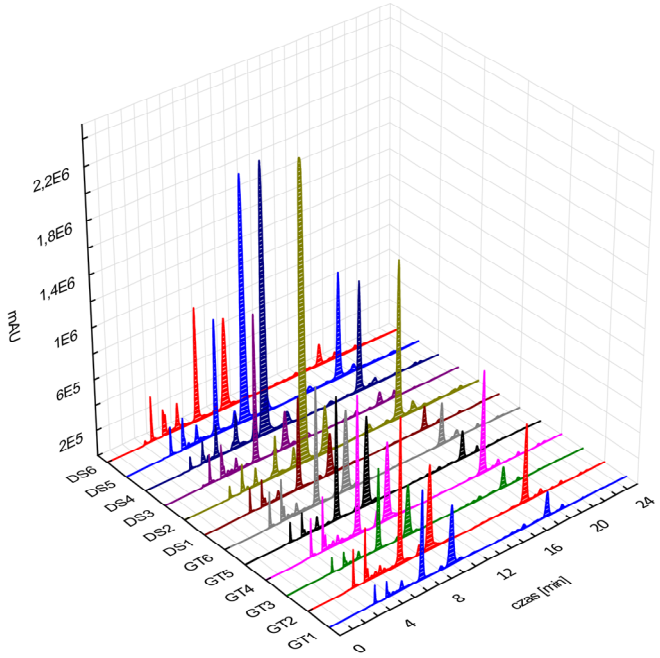
Kod	Skład produktu podany przez producenta na etykiecie
DS1	rozdrobnione liście herbaty zielonej
DS2	ekstrakt z liści zielonej herbaty (wysokoskoncentrowany)
DS3	wyciąg z zielonej herbaty
DS4	ekstrakt z zielonej herbaty (silnie skoncentrowany)
DS5	wystandaryzowany ekstrakt z zielonej herbaty
DS6	suchy wyciąg z zielonej herbaty



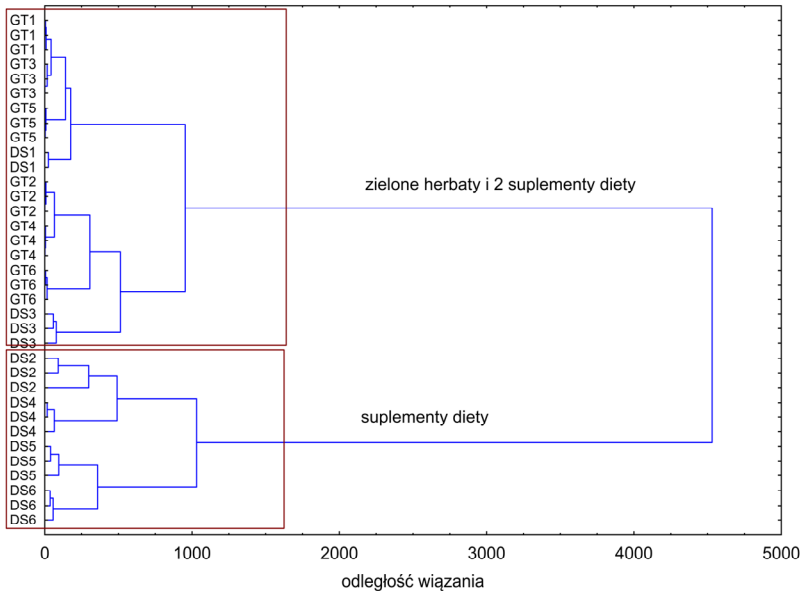
Projekt obejmował:

- wykonanie analizy chromatograficznej metodą RP-HPLC ekstraktów metanolo-wodnych badanych herbat i suplementów diety,
- oznaczenie całkowitej zawartości związków polifenolowych TPC (metoda spektrofotometryczna z odczynnikiem Folina-Ciocalteau),
- oznaczenie aktywności przeciutleniającej (metoda spektrofotometryczna – redukcja rodnika DPPH lub kationorodnika $ABTS^{\bullet+}$,
- analizę chemometryczną uzyskanych wyników:
  - eksplorację danych w celu znalezienia podobieństw między badanymi preparatami (analiza skupień, analiza głównych składowych)
  - konstrukcję modelu kalibracyjnego pozwalającego na przewidywanie TPC i aktywności antyoksydacyjnej na podstawie chromatogramu typu *fingerprint* (drzewa decyzyjne, metoda najmniejszych cząstkowych kwadratów PLS).

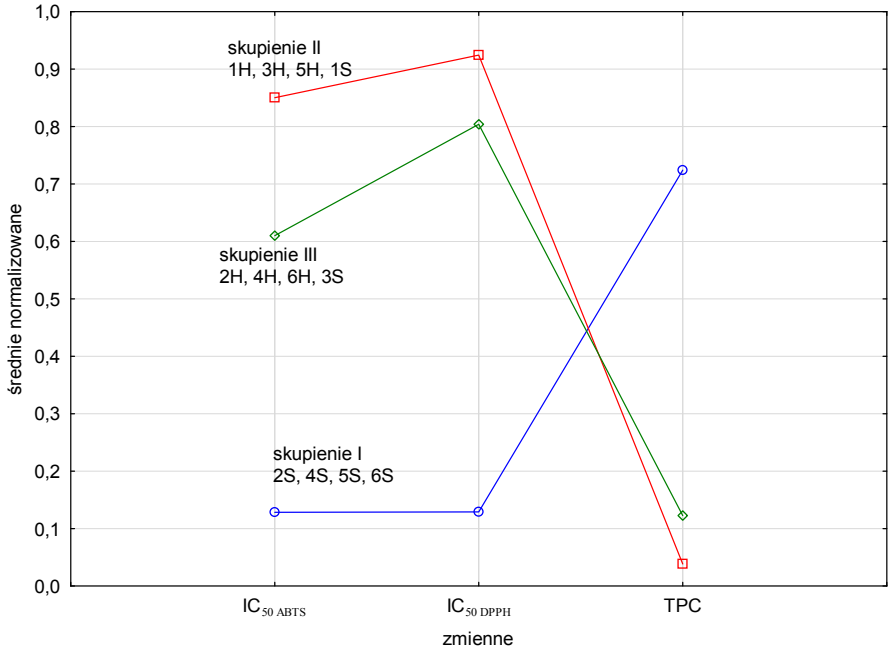
Projekt pn. „*Wzmocnienie potencjału dydaktycznego UMK w Toruniu w dziedzinach matematyczno-przyrodniczych*”
realizowany w ramach Poddziałania 4.1.1 Programu Operacyjnego Kapitał Ludzki



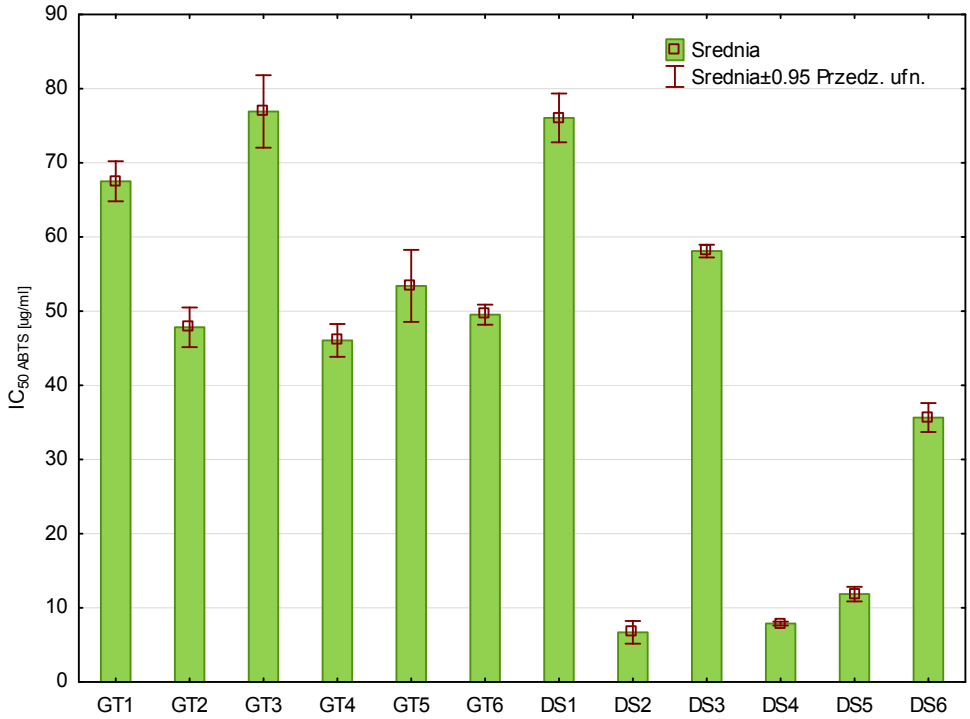
Ryc. 48 Chromatogram 3D badanych preparatów



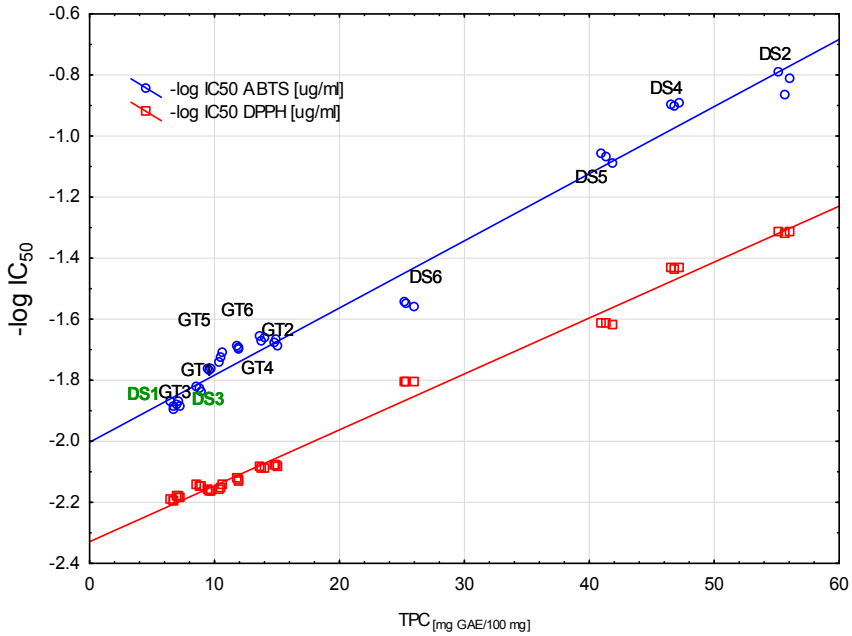
Ryc. 49 Wynik analizy skupień metodą hierarchiczną (metoda Warda, kwadrat odległości Euklidesa) Zmienne: TPC, IC_{50} ($ABTS^{\bullet+}$), IC_{50} ($DPPH^{\bullet}$), Chromat. „fingerprint” (140 punktów pomiar.)



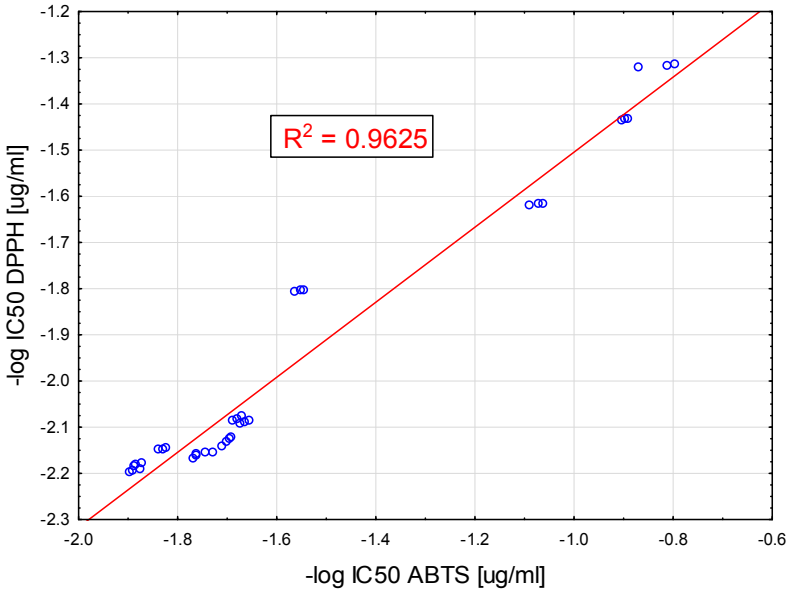
Ryc. 50 Analiza skupień metodą k-średnich. Zmienne: TPC, IC_{50} ($ABTS^{\bullet+}$), IC_{50} ($DPPH^{\bullet}$), Chromat. „fingerprint” (140 punktów pomiar.)



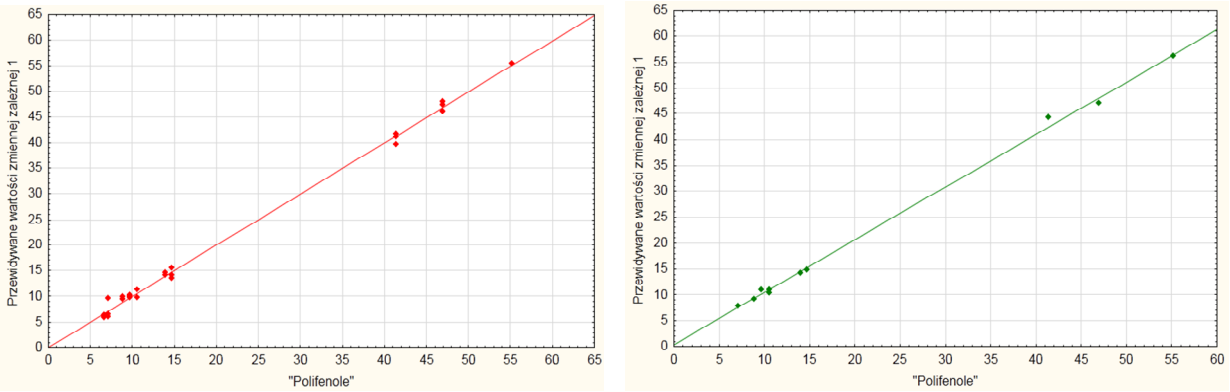
Ryc. 51 Aktywność antyoksydacyjna badanych preparatów wyrażona jako IC_{50} wyznaczona metodą spektrofotometryczną – redukcji kationorodnika $ABTS^{+\bullet}$. IC_{50} – oznacza stężenie ekstraktu powodujące dezaktywację rodnika do 50% wartości wyjściowej.



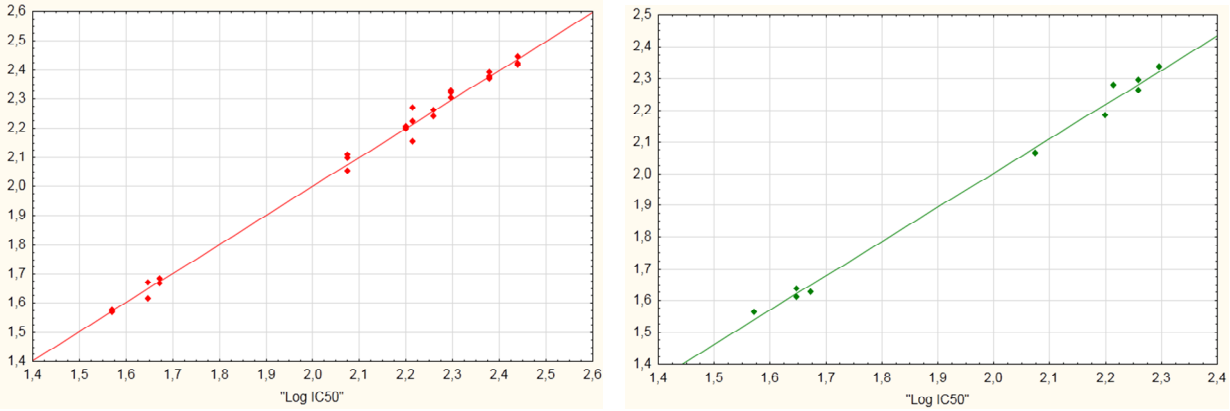
Ryc. 52 Zależność aktywności antyoksydacyjnej badanych preparatów (dwiema metodami) wyrażonej jako –log(IC_{50}) od całkowitej zawartości związków polifenolowych TPC wyrażonej w ekwiwalentach kwasu galusowego na 100 mg produktu (mg GAE/100 mg)



Ryc. 53 Korelacja pomiędzy aktywnością antyoksydacyjną wyznaczoną metodą z rodnikiem DPPH a aktywnością antyoksydacyjną oznaczoną metodą z kationorodnikiem $ABTS^{+\bullet}$. Aktywność wyrażono jako –log(IC_{50}). IC_{50} – oznacza stężenie ekstraktu powodujące dezaktywację rodnika do 50% wartości wyjściowej.



Ryc. 54 Przewidywanie całkowitej zawartości polifenoli TPC na podstawie chromatogramu typu *fingerprint.* Model kalibracyjny uzyskany metodą PLS (zbiór uczący – lewy wykres, zbiór testowy – prawy wykres). Zmienne ukryte: 5, PRESS : 0,064619 (Prediction Sum of Squares)



Ryc. 55 Przewidywanie wartości IC_{50} na podstawie chromatogramu typu *fingerprint.* Model kalibracyjny uzyskany metodą PLS (zbiór uczący – lewy wykres, zbiór testowy – prawy wykres). Zmienne ukryte: 7, PRESS : 0,064619 (Prediction Sum of Squares)

16. Literatura zalecana

1. Carlberg C., Statistical Analysis Microsoft Excel 2010, Pearson Education Inc., 2011.
2. Chemometria w analityce – wybrane zagadnienia, praca pod red. D. Zuby i A. Parczewskiego, Wydawnictwo Instytutu Ekspertyz Sądowych, 2008.
3. Elektroniczny podręcznik statystyki, http://www.statsoft.pl/textbook/stathome.html
4. Gonet. M, Excel w obliczeniach naukowych i technicznych, Helion, 2010
5. Hand D., Manilla H., Smyth P., Eksploracja danych , WNT, 2005.
6. Kala R., Statystyka dla przyrodników, Wyd. Akademia Rolnicza w Poznaniu, 2002.
7. Mazerski J., Chemometria praktyczna, Wyd. Malamut, 2009.
8. Miller J.N., Miller J.C., Statistics and chemometrics for Analitycal Chemistry, wydanie VI, Pearson, 2010
9. Patrick G. L., Chemia medyczna. Podstawowe zagadnienia, WNT, 2003.
10. Sokołowski A., Analizy wielowymiarowe, Materiały kursowe, StatSoft Polska, 2007.
11. Stanisz A., Przystępny kurs statystyki, t. I-III, StatSoft, 2006.
12. Steinhilber D., Schubert-Zsilavecz, Roth H. J., Chemia medyczna, Med Pharm Polska, 2012.
13. Watała C., Biostatystyka – wykorzystanie metod statystycznych w pracy badawczej w naukach biomedycznych, α-medica Press, 2002.
14. Wątroba J., Zastosowanie statystyki w analizie wyników badań medycznych, Materiały kursowe, StatSoft, 2013.

17. Podstawowe pojęcia statystyczne – słownik pol-ang

Nazwa w języku polskim	Nazwa w języku angielskim
Błąd przewidywania	error of prediction
CA (Analiza skupień)	Cluster Analysis
Grupa kontrolna	control group
Hipoteza statystyczna	statistical hypothesis
Hipoteza zerowa	null hypothesis
Jednorodność wariancji	homogeneity of variance
Kurtoza	curtosis
Kwadrat łaciński	latin square
MLR (Regresja wielokrotna)	Multilinear Regression
Moc testu	power of a test
Nachylenie	slope
Odchylenie przeciętne	mean deviation
Odchylenie standardowe	standard deviation
Odporność testu	robustness
PCA (Analiza głównych składowych)	Principal Component Analysis
Poziom istotności	significance level
Przeciętna	average
Przewidywanie	prediction
QSAR (Ilościowa zależność struktura-aktywność)	Quantitative Structure Activity Relationships
Randomizacja	randomization
Regresja liniowa	linear regression
Rozkład	distribution
Rozkład normalny	normal distribution
Rozstęp	range
Skośność	skewness
Stopnie swobody	degrees of freedom
Średni błąd standardowy	standard error mean
Średnia arytmetyczna	arithmetic mean
Średnia geometryczna	geometric mean
Test istotności	test of significance
Test nieparametryczny	nonparametrical test
Wariancja	variance
Wartość przewidywana	predicted value
Współczynnik determinacji	coefficient of determination
Współczynnik korelacji	correlation coefficient
Wykres rozrzutu	scatterplot
Zmienna niezależna	independent variable
Zmienna zależna	dependent variable

18. Spis rycin

19. Przypisy

[1] A. Zejc, M. Gorczyca (red.) Chemia leków, Wydawnictwo Lekarskie PZWL, Warszawa 2004.

[2] Patrick G.L., Chemia medyczna, Wydawnictwo Naukowo-Techniczne Warszawa, 2001.

[3] Steinhilber D., Schubert-Zsilavecz M., Roth H.J., Chemia medyczna , MedPharm Polska 2012.

[4] Mazerski J., Projektowanie nowych chemoterapeutyków, cz. IV Klasyczne metody QSAR, http://www.pg.gda.pl/chem/Katedry/Leki_Biochemia/dydaktyka/pnch/cz4.pdf

[5] Stanisz A., Przystępny kurs statystyki, t. I, StatSoft, 2006.

[6] Lang T., Twenty statistical errors even you can find in biomedical research articles, Croatian Medical Journal (2004) 45, 361-370

[7] Ronowicz J., Kupcewicz B., Budzisz E., Chemometric analysis of antioxidant properties of herbal products containing*Ginkgo biloba* extract, Central European Journal of Biology (2013) 8, 374-385