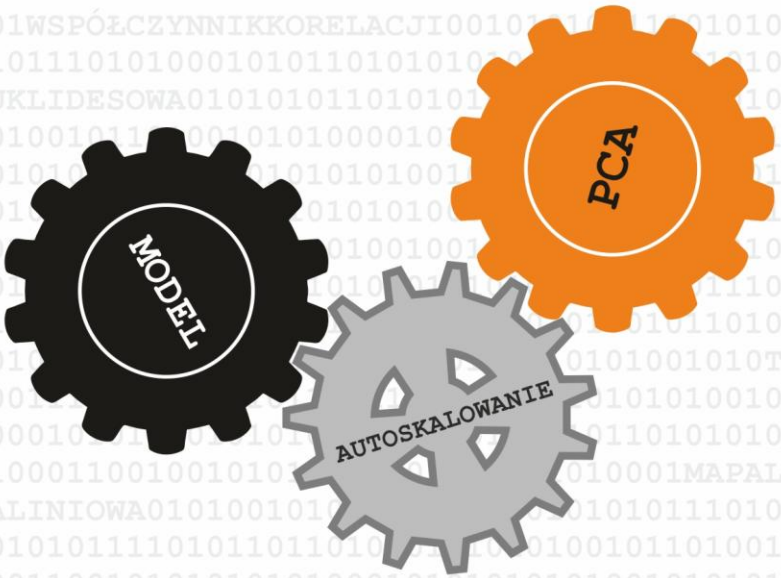
TOMASZ LASKOWSKI
JAN MAZERSKI

CHEMOMETRIA
w praktyce

ĆWICZENIA LABORATORYJNE



WYDAWNICTWO POLITECHNIKI GDAŃSKIEJ

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Podręcznik akademicki współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego, Program Operacyjny Kapitał Ludzki, nr umowy UDA-POKL 04.01.02.-00-033/09-00 „*Zwiększenie liczby absolwentów kierunków o kluczowym znaczeniu dla gospodarki opartej na wiedzy*".

Korekta językowa: .....

Wprowadzenie

Niniejszy skrypt jest przeznaczony głównie dla studentów Wydziału Chemicznego Politechniki Gdańskiej, którzy podjęli wysiłek i ryzyko zaliczania przedmiotu "Podstawy chemometrii". Zachęcamy do korzystania z niego również wszystkich tych, którzy chcieliby zastosować metody chemometryczne do opracowania uzyskanych przez siebie wyników doświadczeń i pomiarów. Jest to w pewnym sensie kontynuacja zagadnień omawianych przez jednego z nas w opracowaniu pt. *"Statystyczna analiza wyników doświadczalnych"*.

Skrypt ten powstał jako odpowiedź na zapotrzebowanie zgłaszane przez studentów poprzednich kursów tego przedmiotu. O ile od roku 2000 dysponowali oni podręcznikiem chemometrii[1], o tyle odczuwali wyraźny brak zestawu instrukcji do ćwiczeń laboratoryjnych. Liczymy, że niniejsze opracowanie w znacznym stopniu wypełni tę lukę. Należy jednak z całą mocą podkreślić, że niniejszy skrypt nie jest w żadnym znaczeniu ani podręcznikiem chemometrii (brakuje w nim szczegółowego opisu podstaw teoretycznych stosowanych metod), ani próbą popularyzacji tej dziedziny wiedzy (zawiera zbyt wiele szczegółów technicznych). Jest to **zbiór instrukcji do określonego zestawu ćwiczeń laboratoryjnych z chemometrii**, realizowanych na naszym Wydziale.

W trakcie pisania niniejszego skryptu kierowało nami przekonanie, że mając do niego dostęp Student będzie przychodził na zajęcia przygotowany. Dlatego na początku każdego spotkania Prowadzący ćwiczenie będzie podawał wyłącznie informacje niezbędne do rozpoczęcia zajęć. Oczywiście będzie również odpowiadał na indywidualne pytania i wyjaśniał wątpliwości.

***

Podczas ćwiczeń będziemy posługiwali się programem *Microsoft Excel*[2] – powszechnie znanym i popularnym arkuszem kalkulacyjnym. Jedynie w kilku przypadkach niezbędne będzie skorzystanie z bardziej specjalistycznego oprogramowania, które udostępni Prowadzący zajęcia. Oczywiście, profesjonalne zastosowanie metod chemometrycznych wymaga korzystania z odpowiednich, komercyjnych pakietów oprogramowania - świadomie jednak zrezygnowaliśmy z ich używania podczas zajęć laboratoryjnych. Zastosowanie oprogramowania tego typu opiera się bowiem na podejściu "wybierz z menu"; tj. użytkownik wybiera, mniej lub bardziej świadomie, jedną z możliwości proponowanych przez program i otrzymuje jakiś wynik. Nie da się ukryć, iż jest to doskonała zabawa - chcielibyśmy jednak, aby podczas wykonywania ćwiczeń Student **zrozumiał**, na czym polega dana metoda chemometryczna i jak powstaje wynik. Aby to osiągnąć, nie ma lepszego sposobu niż obliczenia "na piechotę", wspomagane co najwyżej niewielką pomocą typowego arkusza kalkulacyjnego.

[1] Jego nowa, poszerzona wersja ukazała się w roku 2009 nakładem *Wydawnictwa Malamut* pod tytułem: *"Chemometria praktyczna – zinterpretuj wyniki swoich pomiarów"*.
[2] Warto zatem zapoznać się z książką *"Excel dla chemików... I nie tylko"*, autorstwa W. Ufnalskiego i K. Mądrego (WNT, Warszawa 2000).

Zestaw ćwiczeń został zaplanowany w taki sposób, aby nauczyć Studenta systematyczności w pracy z danymi - w większości ćwiczeń będą bowiem wykorzystywane wyniki uzyskane w jednym z ćwiczeń wcześniejszych. W związku z powyższym, każde ćwiczenie powinno zostać starannie i poprawnie wykonane - aby to ułatwić, Prowadzący będzie oczekiwał od Studenta przygotowania z każdego ćwiczenia odpowiedniego sprawozdania. Sprawozdania te będą na bieżąco sprawdzane, co umożliwi poprawę pomyłek i uniknięcie propagacji błędów.

Sugerujemy również, aby gromadzić wszystkie dane i wyniki pośrednie w jednym pliku programu *Excel*, przeznaczając na każde ćwiczenie oddzielny arkusz. Ułatwia to wykorzystywanie wyników jednego ćwiczenia jako danych do kolejnego. Używając odpowiednich odwołań pomiędzy arkuszami można ponadto stworzyć układ, w którym poprawa błędu w jednym miejscu zostanie automatycznie uwzględniona we wszystkich kolejnych arkuszach. Jest to jednak broń obosieczna: każdy niewykryty błąd również będzie się przenosił do kolejnych arkuszy. Dlatego, w układzie ćwiczeń przewidziane są odpowiednie "punkty kontrolne" - stwarza to możliwość porównania wyników uzyskanych różnymi metodami.

Gromadzenie wszystkich danych i wyników pośrednich w jednym pliku jest bardzo wygodne, wymaga jednak zastosowania dodatkowych środków bezpieczeństwa. Oto najważniejsze z nich:

- ✓ **Zawsze należy posiadać co najmniej dwie kopie swojego pliku bazowego.** Zalecamy, aby jedną z nich przechowywać na swoim komputerze, zaś drugą na mobilnym nośniku danych (*pendrive*).
- ✓ **Wykonywanie ćwiczenia warto rozpocząć od utworzenia nowej kopii pliku bazowego** – pliku tymczasowego.
- ✓ **Bieżące obliczenia najlepiej wykonywać tylko w pliku tymczasowym**.
- ✓ **Po zakończeniu ćwiczenia warto najpierw utworzyć kopię pliku tymczasowego,** a dopiero potem ewentualnie usunąć poprzednią wersję pliku bazowego.
- ✓ **Po powrocie do domu koniecznie należy utworzyć kopię nowego pliku bazowego na swoim komputerze.**

Postępowanie takie może wydawać się przesadnie ostrożne (lub oczywiste), ale chroni przed całkowitą utratą danych i wyników. Nawet w przypadku awarii komputera w laboratorium lub utraty nośnika danych, poprzednia wersja pliku bazowego będzie do dyspozycji na komputerze Studenta, co umożliwi szybkie odtworzenie uzyskanych wyników.

***

Drogi Czytelniku - trzymasz w ręku książkę kucharską, która umożliwi Ci przygotowanie wspaniałych potraw dla ducha i umysłu; złożonych z liczb i subtelnych zależności pomiędzy nimi - nie wyjaśni ona jednak w pełni "metafizyki" ich powstawania. Lektura skryptu powinna obudzić (lub pogłębić) w Tobie ciekawość charakterystyczną dla inżyniera: "*Patrzcie Państwo, TO działa. Ja się pytam: w jaki sposób?*". Po odpowiedź na to pytanie zapraszamy na wykład.

Oddając w Twoje ręce niniejsze opracowanie mamy nadzieję, że korzystaniu z niego będzie towarzyszyła satysfakcja z poznania nowego, bardzo silnego narzędzia do wydobywania użytecznych informacji ze zbioru danych liczbowych. Gdybyśmy bowiem podeszli do sprawy filozoficznie i zacytowali Księgę Mądrości: "(…) *aleś Ty wszystko [Panie] urządził według miary i liczby, i wagi!*"[3], moglibyśmy wysnuć wniosek, iż metody chemometryczne mogą okazać się przydatne nie tylko w trakcie studiowania danych chemicznych. Wniosek, który – jako pokażą kolejne ćwiczenia – wcale nie jest mocno przesadzony.

Na koniec mamy jeszcze gorącą prośbę: jeżeli zauważysz w instrukcjach jakiekolwiek błędy (edytorskie lub merytoryczne), zgłoś to koniecznie do Prowadzącego ćwiczenie lub bezpośrednio do Autorów. *Errare humanum est.*[4]

[3] Ks. Mądrości 11:20 (*Biblia Tysiąclecia*).
[4] *"Błądzić jest rzeczą ludzką"* - Seneka Starszy.

Ćwiczenie nr 1:
ZEBRANIE DANYCH

W ciągu wielu lat zajęć laboratoryjnych z chemometrii kolejne roczniki Studentów przygotowały wiele spektakularnych zestawów danych, począwszy od danych czysto chemicznych, a skończywszy na parametrach i osiągach *strongman'ów* oraz szczegółowych wymiarach starożytnych amfiteatrów greckich.

Nie chcemy w tym miejscu wyliczać wszystkich, ciekawych problemów postawionych przez młodych adeptów chemometrii, aby nie prezentować utartych szlaków i nie odbierać następnym rocznikom szansy wykazania się pomysłowością. Chcemy jedynie powiedzieć, że charakter danych, które należy przygotować, może być zasadniczo dowolny. Istnieją jednak pewne wymagania, które zebrane dane powinny spełniać, aby zaproponowany problem był możliwy do rozwiązania. Wymagania te zostały zaprezentowane poniżej.

I. WYMAGANIA DOTYCZĄCE DANYCH.

Aby każde kolejne ćwiczenie laboratoryjne dostarczało satysfakcji, prowadziło przy tym do rozwiązania postawionego problemu (o którym za chwilę), a jednocześnie ułatwiało opanowanie podstaw chemometrii, zebrane dane powinny spełniać następujące warunki:

1) **Dane powinny składać się z 20-30 obiektów o dowolnym charakterze, opisywanych przez 6-10 cech.**
2) **Cechy, opisujące obiekty, powinny być możliwe do przedstawienia, w sposób jednoznaczny, w postaci liczb.** W związku z powyższym, cechy takie jak: *kolor farby*, *smak owocu*, *przystojność aktora* i *funkcjonalność telefonu* będą eliminowane na starcie przez Prowadzącego. Możliwe jest, co prawda, uwzględnienie zmiennych o charakterze zero-jedynkowym (*0 = telefon nie posiada Bluetooth, 1 = telefon posiada Bluetooth*), nie polecamy ich jednak z uwagi na potencjalnie niekorzystny wpływ na wyniki późniejszych analiz.
3) **Wartości wszystkich cech, opisujących obiekty, muszą być sprecyzowane dla każdego obiektu.** Oznacza to, iż niedopuszczalna jest nieznajomość nawet jednej wartości cechy dla pojedynczego obiektu.

Jak wspomnieliśmy wyżej, zebranym danym powinien towarzyszyć przeznaczony do rozwiązania problem, który przedstawimy w sekcji drugiej.

II. SFORMUŁOWANIE PROBLEMU.

Problem, którego próba rozwiązania zostanie podjęta w trakcie zajęć laboratoryjnych, a który dotyczy przygotowanych danych, może zostać wybrany spośród następujących propozycji:

1. **Modelowanie zależności wybranej cechy od pozostałych zmiennych** (nazywanych wówczas zmiennymi objaśniającymi).
2. **Analiza podobieństwa zmiennych i obiektów** (poznanie wewnętrznej struktury zbioru danych).
3. **Analiza skupień, pozwalająca na obiektywny podział niejednorodnego zbioru obiektów na jednorodne podgrupy.**

Poszczególne propozycje oferują przedstawione poniżej możliwości.

Ad. 1. Rozwiązanie problemu tego rodzaju sprowadza się do odpowiedzi na pytanie, **czy istnieje matematyczna zależność jednej, wybranej cechy** (opisującej obiekty) **od pozostałych cech**, oraz **czy możliwe jest wyrażenie tej zależności w postaci modelu liniowego**.

*Na przykład: czy istnieje **zależność wagi trzydziestu sąsiadów z bloku** od **ilości zjadanych w ciągu roku warzyw, owoców, czekolad, kebabów, lodów** oraz **wypitej coli i kawy**, czy też nie ma takiej zależności? (Jeżeli zależność zostanie wykryta, będzie również możliwe ustalenie, które smakołyki i w jaki sposób mają wpływ na wagę sąsiadów.)*

Ad. 2. Rozwiązanie problemu tego rodzaju rozpoczyna się od ustalenia, czy zaproponowany zbiór danych jest jednorodny. Ustalenie takie sprowadza się do odpowiedzi na pytania: **i)** *czy poszczególne zmienne pochodzą z tej samej populacji generalnej?*; oraz: **ii)** *czy wszystkie obiekty pochodzą z tej samej populacji generalnej?*. Uzyskanie odpowiedzi pozytywnej na obydwa pytania kończy analizę chemometryczną, w związku z czym istnieje niebezpieczeństwo, iż po kilku ćwiczeniach Student zostanie bezrobotny do końca semestru.

Dużo ciekawsza sytuacja zaistnieje wówczas, gdy chociaż na jedno z powyższych pytań odpowiedź będzie **negatywna**. Należy wtedy **wykazać, jaki jest charakter obserwowanej niejednorodności zbioru, czyli określić wewnętrzną strukturę danych**.

Ad. 3. Poczynione zostaje założenie (lub istnieje uzasadnione przypuszczenie), że zbiór obiektów nie jest jednorodny. W takiej sytuacji, **celem Studenta jest możliwie obiektywne** (ze względu na wartości wybranych zmiennych) **podzielenie go na wewnętrznie jednorodne podzbiory**.

Do tego typu analizy można podejść dwojako: **i)** *z uprzednią znajomością liczby i rodzaju podzbiorów oraz z wiedzą dotyczącą przynależności poszczególnych obiektów do tych podzbiorów* (wtedy można uzyskać (lub nie) potwierdzenie, ze wybrane zmienne zawierają informację niezbędną do podziału obiektów na takie właśnie podzbiory); lub: **ii)** *bez znajomości struktury wewnętrznej zbioru obiektów*. Wykazanie istnienia wewnętrznie jednorodnych podzbiorów będzie wtedy "wartością dodaną" analizy i nagrodą za dociekliwość naukową.

*Na przykład: **zmierzono: długość całkowitą; długość ogona; długość tylnej, prawej łapy; średnią długość wąsów; prędkość maksymalną w biegu do miski na odcinku 50 metrów po dniu postu; rozstaw oczu** oraz **zważono 11 kotów i 12 kotek** w wieku 2 lat[5]. Postawiono następujący problem: **czy te zmienne pozwalają na odróżnienie samców i samic w grupie***

[5] Pomiary takie nie zostały przeprowadzone (przynajmniej przez Autorów).

***dwuletnich kotów**, czy też należy poszukać innych cech ilościowych, które są zależne od płci tych zwierząt?*

W trakcie ćwiczenia nr 2 Prowadzący indywidualnie przedyskutuje ze Studentem zaproponowany przez Niego problem i spróbuje określić, czy możliwe jest jego rozwiązanie na podstawie zgromadzonych danych. Dopiero po tej rozmowie należy podjąć ostateczną decyzję w kwestii wyboru problemu.

III. SPRAWOZDANIE z tego ćwiczenia sprowadza się do zebrania danych zgodnych z podanymi wymaganiami oraz zaproponowania problemu do rozwiązania.

Ćwiczenie nr 2:
PRZYGOTOWANIE DANYCH DO ANALIZY

W trakcie pierwszych zajęć laboratoryjnych Student został poproszony o zebranie danych, na których, w trakcie całego semestru, będzie dokonywał najrozmaitszych operacji matematycznych w celu wyekstrahowania informacji niewidocznych gołym (czytaj: *w chemometryczne narzędzia nieuzbrojonym*) okiem. Charakter tych danych mógł być zasadniczo dowolny, z zastrzeżeniem konieczności wyboru takich cech (dalej nazywanych **zmiennymi**), które da się jednoznacznie przedstawić w postaci liczb.

Celem niniejszego ćwiczenia jest przygotowanie takiej formy zebranych danych, która sprawi, że operacje matematyczne i statystyczne, które będą na nich przeprowadzane w trakcie kolejnych ćwiczeń, dostarczały możliwie dużo informacji. Dla przejrzystości dalszych wyjaśnień, wszystkie omawiane w kolejnych instrukcjach operacje (numeryczne, statystyczne i graficzne) będą przeprowadzane - w charakterze przykładu - na konkretnym, przeciętnym i nietendencyjnie dobranym zestawie danych. Zestaw ten zostanie zaprezentowany poniżej.

Jego prezentacja będzie jednocześnie stanowiła **instrukcję, jak poprawnie przygotować tabelę danych do dalszej analizy**.

I. PREZENTACJA PRZYKŁADOWYCH DANYCH.

W sklepie internetowym, który w równie tajemniczych okolicznościach rozpoczął jak i zakończył swoją działalność, oferowano swego czasu repliki broni białej. Dla celów niniejszego opracowania, wybraliśmy z nich dwadzieścia obiektów należących do klasy mieczy, którym nadaliśmy, możliwie adekwatnie do rozmiarów, wyjątkowo legendarne nazwy zaczerpnięte z historii i literatury. Każdy z mieczy został opisany przez siedem cech: **długość całkowita, DC, [cm]; długość głowni, DG, [cm]; długość rękojeści, DR, [cm]; masa, M, [g]; odległość środka masy od rękojeści, SM, [cm]; typ miecza**[6]**, T;** oraz **cena repliki, CR, [PLN].**

Cecha przedstawiona jako **typ miecza** nie jest typem zmiennej rekomendowanym do uwzględniania w trakcie proponowanych analiz, ponieważ przyjmuje tylko trzy różne wartości. **Takie zmienne często odgrywają dużą rolę w analizie podobieństwa obiektów oraz w analizie skupień**. Jeżeli jednak zostanie podjęta decyzja o uwzględnieniu zmiennych tego rodzaju w analizowanych danych, warto pamiętać, iż nie powinno być ich więcej, niż jedna.

Zgodziwszy się zatem na uwzględnienie **typu miecza** jako zmiennej, prezentujemy poniżej tabelę danych wejściowych zestawu **MIECZE:**

[6] Mamy tu na myśli typ: jednoręczny, półtoraręczny oraz dwuręczny. Cecha ta powinna być silnie skorelowana zarówno z rozmiarami broni, jak i jej masą. Czy tak będzie rzeczywiście, pokażą dalsze analizy.

Obiekt\Zmienna	DC	DG	DR	M	SM	T	CR
AER	119	92	15	1900	5	1,5	500
AND	152	100	32	2500	12	2	260
AZU	88	71	14	1200	7	1	380
BAL	95	75	13	1400	7	1	320
DUR	102	81	14	1400	8	1	342
EXC	120	90	18	1800	10	1,5	450
GLA	120	95	12	1900	10	1,5	419
GOL	100	69	26	1100	6	1	600
GRA	106	83	15	1600	10	1	350
GUR	104	81	15	1800	10	1,5	406
GWY	103	81	15	1450	5	1	400
HER	85	60	14	1500	8	1	340
HUR	90	65	16	1600	7	1	380
JOY	100	80	14	1500	8	1	320
LOD	92	80	10	1800	10	1	375
ORK	130	97	18	1800	10	1,5	450
SIH	123	95	14	2200	8	1,5	390
UMB	180	125	40	3200	15	2	600
URI	160	120	25	2700	12	2	650
ZAD	68	54	13	800	5	1	375

Należy teraz zwrócić uwagę na kilka ważnych elementów składowych tabeli danych. Powinna ona zawierać:

1. możliwie krótkie nazwy obiektów (optymalnie: 3-4 literowe) kojarzące się z rzeczywistymi nazwami obiektów (np. skrót **EXC** reprezentuje miecz nazwany *Excalibur*);
2. możliwie krótkie nazwy zmiennych, również kojarzące się z rzeczywistymi nazwami cech (np. skrót **DG** dla zmiennej *długość głowni*);
3. wartości liczbowe odpowiednich cech dla **WSZYSTKICH** obiektów. **Niedopuszczalne są braki w tabeli!**

Ponadto, w bezpośredniej bliskości tabeli powinny znajdować się następujące informacje:

- objaśnienia krótkich nazw obiektów i zmiennych;
- odsyłacz do źródła danych;
- imię i nazwisko autora;
- data utworzenia i ostatniej modyfikacji tabeli;
- sformułowany problem, dotyczący przygotowanych danych.

II. SPRECYZOWANIE PROBLEMU odbędzie się w trakcie dyskusji z Prowadzącym.

III. SPRAWOZDANIE stanowi prawidłowo przygotowana tabela danych w arkuszu kalkulacyjnym *Excel*, gotowa do dalszych analiz.

Ćwiczenie nr 3:
KONTROLA POJEDYNCZYCH ZMIENNYCH

Celem ćwiczenia jest **kontrola przygotowanych danych liczbowych za pomocą zestawu testów statystycznych**. Kontrola ta pomoże w uzyskaniu odpowiedzi na następujące pytania:

- *jaki jest charakter rozkładu poszczególnych zmiennych?*;
- *czy istnieją przesłanki o konieczności dokonania transformacji zmiennych?*;
- *czy wśród zestawu obiektów znajdują się punkty odbiegające?*.

I. WYZNACZANIE WARTOŚCI LICZBOWYCH CHARAKTERYSTYK ROZKŁADU.

Kontrolę zestawu danych rozpoczyna się od **obliczenia wartości liczbowych kilku charakterystyk rozkładu zmiennych**. Najwygodniej jest uczynić to w formie tabeli, znajdującej się bezpośrednio pod tabelą danych wejściowych. I tak, pod każdą z kolumn z wartościami zmiennych powinny znajdować się wiersze zawierające:

wartość najmniejszą w obrębie zmiennej (**MIN**) - funkcja w Excelu: **=MIN(zakres_komórek_z_wartościami_zmiennej); np. =MIN(A05:A30)**
wartość największą w obrębie zmiennej (**MAX**) - funkcja w Excelu: **=MAX(zakres_komórek_z_wartościami_zmiennej)**
stosunek **MIN/MAX**
rozstęp rozkładu zmiennej (**r = MAX-MIN**)
środek rozkładu zmiennej (**d = (MAX+MIN)/2**)
wartość średnia zmiennej (**m**) - funkcja w Excelu: **=ŚREDNIA(zakres_komórek_z_wartościami_zmiennej)**
odchylenie standardowe zmiennej (**s**) - funkcja w Excelu: **=ODCH.STANDARDOWE.POPUL(zakres_komórek_z_wartościami_zmiennej)**
indeks skośności rozkładu zmiennej (**q**) - funkcja w Excelu: **=SKOŚNOŚĆ(zakres_komórek_z_wartościami_zmiennej)**

UWAGA! Jeżeli jedna ze zmiennych ma charakter zero-jedynkowy, bądź przyjmuje tylko dwie-trzy różne wartości - nie trzeba wyznaczać dla niej powyższych charakterystyk rozkładu ani nie należy jej transformować. Ma ona, z definicji, rozkład odbiegający od rozkładu normalnego i nic na to nie można poradzić.

Otrzymane dla każdej zmiennej charakterystyki należy teraz poddać następującym testom:

1) **czy wartość MIN/MAX > 0,1 ?**
2) **czy |d-m| < s ?**
3) **czy wartość r/s należy do przedziału <3;5> ?**
4) **czy |q| < 2 ?**

Jeżeli dla danej zmiennej odpowiedzi na cztery powyższe pytania brzmią TAK, zmienna ma prawdopodobnie rozkład zbliżony do normalnego i - przynajmniej do czasu następnego ćwiczenia - przestaje być "interesująca".

Jeżeli zaś, dla danej zmiennej, odpowiedź na przynajmniej jedno powyższe pytanie brzmi NIE, zmienna staje się "podejrzana". Przyczyny takiego stanu rzeczy mogą być dwie: **i)** wśród wartości zmiennej występuje punkt lub punkty odbiegające; **ii)** rozkład zmiennej jest silnie asymetryczny lub wielomodalny.

Aby ustalić, dlaczego rozkład danej zmiennej odbiega od rozkładu normalnego, należy wykonać **histogram wartości tej zmiennej**[7].

II. CHARAKTER ROZKŁADU POSZCZEGÓLNYCH ZMIENNYCH.

Po wykonaniu histogramów rozkładu wszystkich "podejrzanych" zmiennych, należy przyjrzeć im się i odpowiedzieć na następujące pytania (dla każdej ze zmiennych):

1) czy rozkład zmiennej jest **wielomodalny**?;
2) jeżeli rozkład zmiennej jest jednomodalny - czy jest symetryczny lub zbliżony do symetrycznego?;
3) jeżeli rozkład zmiennej jest jednomodalny - czy jest silnie lewo- lub prawoskośny?;
4) czy na histogramie widoczny jest **punkt odbiegający**?

Jeżeli odpowiedź na pytanie 1) brzmi TAK - należy zostawić zmienną w spokoju. Zmienna taka może odegrać dużą rolę w analizie podobieństwa obiektów lub w analizie skupień.

Jeżeli odpowiedź na pytanie 2) brzmi TAK - należy zostawić zmienną w spokoju. Pomimo, iż jej rozkład nie jest normalny, można ją z powodzeniem stosować praktycznie we wszystkich metodach chemometrycznych.

Jeżeli odpowiedź na pytanie 4) brzmi TAK - należy przejść do sekcji III.

Jeżeli odpowiedź na pytanie 3) brzmi TAK (rozkład jest silnie lewo- lub prawoskośny) - należy dokonać transformacji zmiennej. Transformacja zmiennej polega na przekształceniu wszystkich wartości danej zmiennej za pomocą odpowiedniej funkcji matematycznej. Po dokonaniu transformacji należy **ponownie wykonać histogram z otrzymanych wartości danej zmiennej** i ocenić, czy jej rozkład stał się przynajmniej symetryczny.

Poniższa tabela zawiera przykłady funkcji transformujących, znajdujących zastosowanie w najczęściej występujących sytuacjach:

[7] "Ręczne" wykonywanie histogramów rozkładu zmiennych nie jest zajęciem szybkim, łatwym, ani przyjemnym. Aby ułatwić Studentowi życie, zapraszamy do lektury **Dodatku A**, znajdującego się na końcu niniejszej instrukcji.

Charakter zmiennej	Przykłady funkcji transformujących
stosunek MIN/MAX<0,1; jest silnie prawoskośna	$x^* = \log_{10}(x)$, $x^* = \log_{10}(x+a)$; $x+a > 0$
zmienna jest silnie lewoskośna	$x^* = \log_{10}(a-x)$; $a > x_{MAX}$
zmienna ma postać % i x < 15%	$x^* = \log_{10}(x)$
zmienna ma postać % i x > 85%	$x^* = \log_{10}(a-x)$, a = 100
inne	$x^* = \log_{10}(x/(a-x))$, $x^* = 1/x$, inne

Funkcje transformujące można oczywiście dobierać dosyć swobodnie. **Wartość liczbową parametru a należy dobierać metodą prób i błędów.** *Na przykład:* jeżeli zmienna jest silnie lewoskośna, lecz dla funkcji transformującej **x*= log(a-x)** zostanie dobrana zbyt duża wartość **a**, zmienna po transformacji stanie się prawoskośna - należy zatem spróbować wartości mniejszej.

Należy również pamiętać, aby wartości zmiennej transformowanej, otrzymane po zastosowaniu odpowiedniej funkcji, przedstawić w należytym formacie. *Na przykład:* **niewłaściwe jest podawanie wartości liczbowych transformowanej zmiennej z dokładnością do sześciu miejsc po przecinku**. Zwykle stosuje się następujący format komórek**: *liczbowy, z trzema miejscami po przecinku*** (oczywiście *Excel* zapamiętuje te wartości z pełną dokładnością). Nieuzasadnione podanie zbyt szczegółowych wartości będzie traktowane jako błąd w sztuce.

Po dokonaniu transformacji zmiennych należy przygotować nową tabelę danych, w której wartości zmiennych transformowanych zastąpią wartości "oryginalne". Należy również zaznaczyć, które zmienne zostały poddane transformacji (najczęściej czyni się to poprzez dodanie * do etykiet zmiennych), a także odnotować - blisko tabeli - postaci funkcji transformujących.

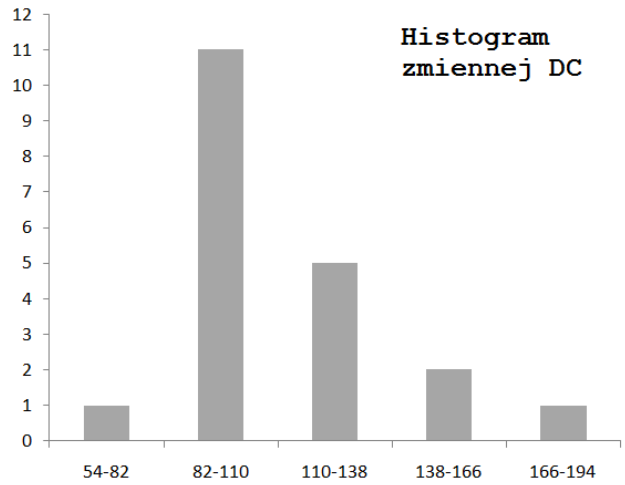
Przykład:

Weźmy na warsztat dwie zmienne z omawianego zbioru danych **MIECZE,** np. **DC** oraz **DR.** Wartości liczbowe charakterystyk rozkładu dla tych zmiennych prezentują się następująco:

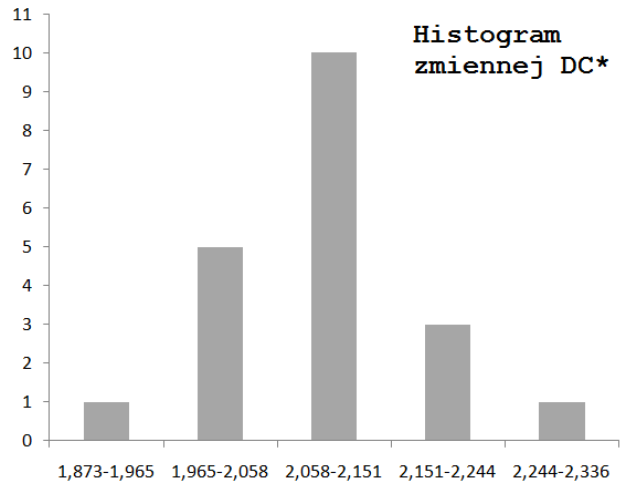
Obiekt\Zmienna	DC	DR
AER	119	15
AND	152	32
AZU	88	14
BAL	95	13
DUR	102	14
EXC	120	18
GLA	120	12
GOL	100	26
GRA	106	15
GUR	104	15
GWY	103	15
HER	85	14
HUR	90	16
JOY	100	14
LOD	92	10
ORK	130	18
SIH	123	14
UMB	180	40
URI	160	25

ZAD	68	13
MIN	68	10
MAX	180	40
MIN/MAX	0,38	0,25
r	112	30
d	124	25
m	111,85	17,65
s	26,58	7,28
q	1,03	1,96
r/s	4,21	4,12
|d-m|	12,15	7,35
|q|	1,03	1,96

Przyjrzyjmy się najpierw zmiennej **DC**. Stosunek **MIN/MAX** wynosi w jej przypadku **0,38**, jest więc zdecydowanie większy niż wartość krytyczna, równa **0,1**. Również odległość średniej arytmetycznej od środka przedziału zmienności jest mniejsza niż odchylenie standardowe. Jedynie względnie duży indeks skośności, wynoszący **1,03**, może budzić pewne wątpliwości. Ponieważ w świecie chemometrii panuje pogląd: "*jeżeli masz wątpliwości – zrób wykres*", wykonaliśmy histogram rozkładu zmiennej **DC**. Wygląda on w sposób następujący:



Histogram zmiennej ujawnia wyraźną skośność rozkładu. Postanowiliśmy sprawdzić (chociaż nie ma takiego wymogu), czy nie można poprawić symetrii rozkładu tej zmiennej na drodze prostej transformacji. Ponieważ mamy do czynienia ze zmienną prawoskośną, zastosowaliśmy funkcję transformującą $\mathbf{x^* = \log_{10}(x+a)}$. Po kilku próbach okazało się, że optymalna wartość parametru **a = 15**, i tym samym funkcja transformująca ma postać: $\mathbf{DC^* = \log_{10}(DC+15)}$. Histogram wykonany z wartości zmiennej **DC*** zaczął przypominać rozkład normalny:

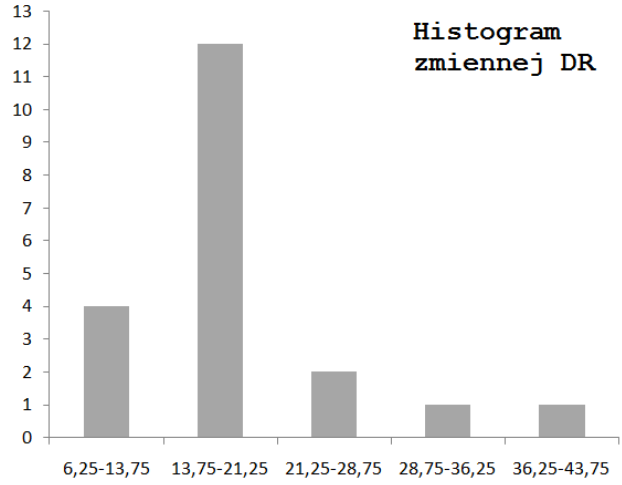


Per analogiam i z podobnych przyczyn transformowane zostały zmienne **DG** oraz **CR** przy pomocy następujących funkcji:

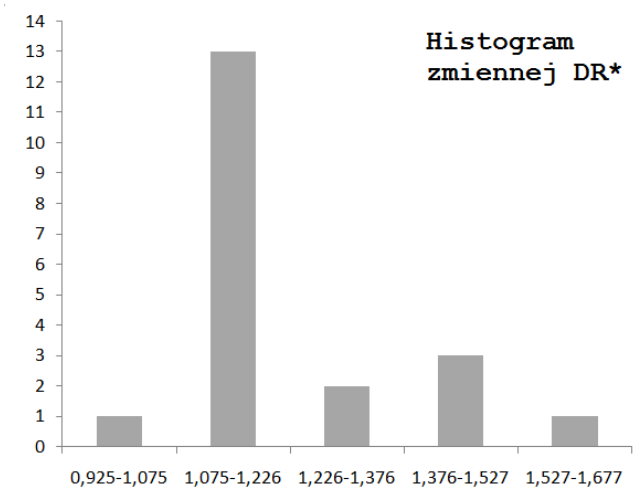
$$\mathbf{DG^* = \log_{10}(DG)}$$
$$\mathbf{CR^* = \log_{10}(CR-200)}$$

Dokonajmy teraz kontroli zmiennej **DR**. Naszą uwagę natychmiast przykuwają wartości dwóch charakterystyk: **i)** odległość średniej arytmetycznej od środka przedziału zmienności (**|d-m| = 7,35**) jest porównywalna z wartością odchylenia standardowego (**s = 7,28**); **ii)** indeks skośności **q** ma dość dużą, w zasadzie krytyczną, wartość i wynosi **1,96**.

Nie ma wątpliwości, że zmienna ta ma rozkład zdecydowanie odbiegający od rozkładu normalnego. Wykonanie histogramu w pełni potwierdza nasze przypuszczenia:



Z pozoru mamy do czynienia z rozkładem silnie prawoskośnym, podobnym do rozkładu zmiennych **DC, DG** oraz **CR.** Prosta transformacja **DR* = $\log_{10}$(DR)** ujawnia jednak interesującą własność zmiennej **DR:**



Po transformacji rozkład zmiennej stał się **dwumodalny.** W związku z powyższym, porzucamy projekt transformacji zmiennej **DR.**

Tabela danych dla zestawu **MIECZE,** po dokonaniu omówionych powyżej transformacji, prezentuje się następująco:

Obiekt\Zmienna	DC*	DG*	DR	M	SM	T	CR*
AER	2,127	1,964	15	1900	5	1,5	2,477
AND	2,223	2,000	32	2500	12	2	1,778
AZU	2,013	1,851	14	1200	7	1	2,255
BAL	2,041	1,875	13	1400	7	1	2,079
DUR	2,068	1,908	14	1400	8	1	2,152
EXC	2,130	1,954	18	1800	10	1,5	2,398
GLA	2,130	1,978	12	1900	10	1,5	2,340
GOL	2,061	1,839	26	1100	6	1	2,602
GRA	2,083	1,919	15	1600	10	1	2,176
GUR	2,076	1,908	15	1800	10	1,5	2,314
GWY	2,072	1,908	15	1450	5	1	2,301
HER	2,000	1,778	14	1500	8	1	2,146
HUR	2,021	1,813	16	1600	7	1	2,255
JOY	2,061	1,903	14	1500	8	1	2,079
LOD	2,029	1,903	10	1800	10	1	2,243
ORK	2,161	1,987	18	1800	10	1,5	2,398
SIH	2,140	1,978	14	2200	8	1,5	2,279
UMB	2,290	2,097	40	3200	15	2	2,602
URI	2,243	2,079	25	2700	12	2	2,653
ZAD	1,919	1,732	13	800	5	1	2,243

III. PUNKTY ODBIEGAJĄCE.

Po wykonaniu histogramu dla "podejrzanej" zmiennej może się okazać, iż z lewej bądź prawej strony rozkładu znajduje się punkt odbiegający. **Punkt odbiegający może się pojawić, gdy:**

1. **podczas wykonywania pomiarów lub przygotowywania danych doszło do pomyłki.** Mamy wtedy do czynienia z tzw. "błędem grubym" i należy go, w miarę możliwości, poprawić; jeżeli jednak jest to niemożliwe - obiekt, dla którego wystąpił, należy permanentnie usunąć z tabeli danych.
2. **obiekt, opisywany przez wartość odbiegającą, pochodzi z innej niż pozostałe obiekty populacji generalnej** (np. jeden chomik w populacji myszy). Wartość odbiegająca nie jest, w takim przypadku, wynikiem błędu; tym niemniej obiekt, dla którego występuje, należy usunąć z tabeli danych.
3. **silnie asymetryczny charakter rozkładu w połączeniu z małą liczebnością zestawu danych wywołuje złudzenie punktu odbiegającego.** W takim przypadku, po odpowiedniej transformacji zmiennej wartość odbiegająca powinna utracić swój wyjątkowy status.

Należy teraz podjąć decyzję, czy obiekt, który jest charakteryzowany przez odbiegającą wartość danej zmiennej, powinien pozostać w tabeli danych, czy też należy go usunąć. Jeżeli nie wiadomo, z którą z opisanych powyżej sytuacji mamy do czynienia, decyzję o ewentualnym usunięciu obiektu należy podjąć w oparciu o podany poniżej algorytm postępowania:

1) **Należy tymczasowo usunąć wartość odbiegającą zmiennej** i wykonać **nowy histogram** tej zmiennej.

2) Jeżeli rozkład zmiennej (po usunięciu wartości odbiegającej) stał się **zbliżony do normalnego bądź przynajmniej symetryczny**, **metodą przedziału ufności** (o niej za chwilę) należy **ocenić**, czy obiekt opisywany przez tę wartość usunąć z tabeli, czy też nie.

3) Jeżeli po usunięciu wartości odbiegającej **rozkład zmiennej nie uległ "poprawie"**, **należy przywrócić usuniętą wartość** i **dokonać transformacji** zmiennej.

4) Jeżeli **po dokonaniu transformacji zmiennej jej rozkład stał się symetryczny**, **nie należy usuwać "podejrzanego" obiektu** z tabeli.

5) Jeżeli **po dokonaniu transformacji zmiennej na histogramie w dalszym ciągu widoczny jest punkt odbiegający**, należy **tymczasowo usunąć wartość odbiegającą transformowanej zmiennej** i wykonać **nowy histogram** transformowanej zmiennej.

6) Jeżeli **rozkład transformowanej zmiennej** (po usunięciu wartości odbiegającej) **stał się symetryczny**, **metodą przedziału ufności należy ocenić**, czy usunąć "podejrzany" obiekt, czy też nie.

IV. METODA PRZEDZIAŁU UFNOŚCI.

Aby ułatwić Czytelnikowi poprawne zastosowanie **metody przedziału ufności**, jej założenia zaprezentujemy odwołując się do konkretnej liczby obiektów.

Załóżmy, iż nasza "podejrzana" zmienna przyjmuje **25** wartości, przy czym jedna z nich jawi się na histogramie jako wartość odbiegająca. Tymczasowo usuwamy ją z zestawu danych - pozostaną **24** wartości. Dla tych **24** wartości **obliczamy wartość średnią (m)** i **odchylenie standardowe średniej (s)**, oraz odczytujemy z tabeli[8] **wartość testu t Studenta** dla poziomu istotności **0,05** oraz **n-1** stopni swobody (w tym przypadku **n = 24** - jest to liczba wartości po odrzuceniu "podejrzanego" obiektu; zatem **n-1 = 23**). Następnie, obliczamy krańce przedziału ufności:

$$x_{min} = m - t \cdot s;$$
$$x_{max} = m + t \cdot s.$$

Jeżeli "podejrzana" wartość mieści się w przedziale wyznaczonym przez te granice - nie należy usuwać z tabeli obiektu przez nią opisywanego; jeżeli zaś nie mieści się - obiekt ten można usunąć z zestawu danych.

Przykład:

Ponieważ kontrola zmiennych przeprowadzona dla zestawu **MIECZE** nie wykazała istnienia punktów odbiegających, w celu zilustrowania tego zjawiska posłużymy się wartościami wybranej zmiennej pochodzącymi z innego zbioru danych.

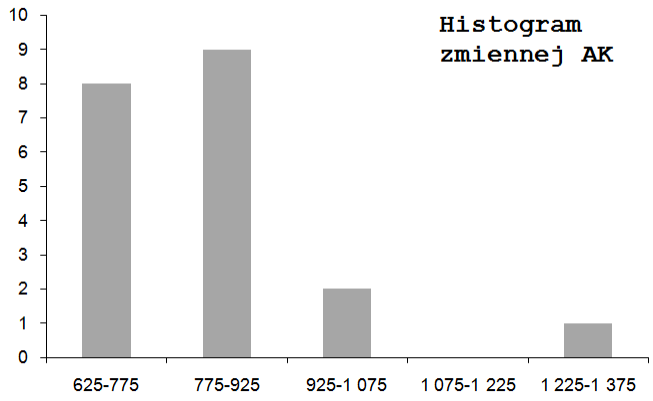
Zestaw 20 telefonów komórkowych został opisany 6 zmiennymi, w tym zmienną odpowiadającą pojemności akumulatora (**AK**), wyrażoną w [mAh]. Wartości tej zmiennej, wraz z obliczonymi wartościami liczbowymi charakterystyk rozkładu, prezentują się następująco:

Model:	AK
N6810	1000
N6260	760
N7710	1300
N7380	700
N2652	760
N7600	850
N7260	760
N6680	900
N6610	720
N6270	900
N7280	700
N5100	720
N3100	850
N2600	820

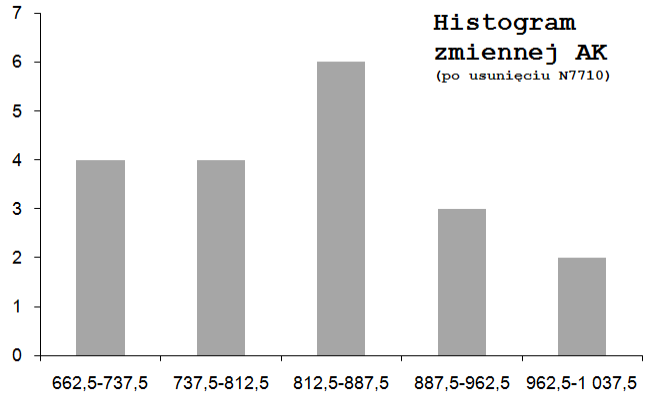
[8] Wartość krytyczną **testu t Studenta** można również uzyskać w *Excelu*, korzystając z funkcji **=ROZKŁAD.T.ODW(α;stopnie_swobody)**; gdzie: **α=0,05** dla testu dwustronnego (np. dla *przedziału ufności*) lub **0,10** dla testu jednostronnego (np. w wersji *statystyki t*).

NN90	**760**
N2610	**970**
N3120	**820**
N6103	**820**
N6630	**900**
N5070	**820**
MIN	**700**
MAX	**1300**
MIN/MAX	**0,54**
r	**600**
d	**1000**
m	**841,5**
s	**137,7**
q	**2,05**
r/s	**4,35**
|d-m|	**158,5**
|q|	**2,05**

Otrzymujemy negatywną odpowiedź na dwa z czterech postawionych w **sekcji I** pytań. Wartość bezwzględna indeksu skośności rozkładu **|q|** jest większa niż **2;** a ponadto **|d-m| > s**. Jesteśmy zatem zmuszeni do wykonania histogramu rozkładu badanej zmiennej:



Histogram rozkładu zmiennej **AK** wyraźnie sugeruje istnienie punktu odbiegającego w postaci modelu **N7710.** Zgodnie z algorytmem postępowania, podanym w **sekcji III,** tymczasowo usuwamy obiekt **N7710** z tabeli danych i wykonujemy nowy histogram zmiennej **AK:**



Histogram zmiennej **AK** zaczął wyglądać na tyle przyzwoicie, iż dalsze modyfikacje tej zmiennej nie są konieczne. Pozostaje zatem ocenić, np. metodą przedziału ufności, czy wolno nam z czystym sumieniem usunąć obiekt **N7710** z tabeli danych.

Po tymczasowym usunięciu modelu **N7710**, pozostało **n = 19** obiektów. Za pomocą *Excela* obliczamy wartość **testu t Studenta** dla omawianego przypadku (poziom istotności **0,05** i **n-1 = 18** stopni swobody):

=ROZKŁAD.T.ODW(0,05;18)

otrzymując wartość **t = 2,101**. Następnie, obliczamy nową wartość średnią zmiennej **AK** (**m = 817**) i jej nowe odchylenie standardowe (**s = 88**). Na koniec, określamy granice przedziału ufności:

$$x_{min} = m - t \cdot s = 633$$
$$x_{max} = m + t \cdot s = 1002$$

Wartość zmiennej **AK** dla obiektu **N7710** wynosi **1300**. Wartość ta nie mieści się w obliczonym przedziale ufności, dlatego obiekt ten wolno nam, z czystym sumieniem, permanentnie usunąć z tabeli danych.

V. SPRAWOZDANIE.

W sprawozdaniu Student powinien wykonać wszystkie, wyżej opisane, a konieczne dla Jego danych operacje oraz dołączyć komentarz dotyczący ewentualnych "podejrzanych" zmiennych i postępowania z nimi. Mile widziany będzie zwięzły, acz treściwy komentarz, na przykład: "*zmienna jest podejrzana, albowiem wartość bezwzględna indeksu skośności jest większa od 2, więc dokonuję transformacji funkcją: ...; przyjmując za optymalną wartość* ***a*** *równą ...; po wykonaniu histogramu transformowanej zmiennej jej rozkład stał się symetryczny*".

Dodatek A:

Automatyzacja tworzenia histogramów w *Excelu.*

W przypadku posiadania choć jednej "podejrzanej" zmiennej, konieczne jest wykonanie histogramu rozkładu tej zmiennej. Jeżeli zmienna ta wymaga transformacji, w celu dobrania optymalnej funkcji transformującej należy (niestety) wykonać kolejne histogramy (patrz: **ćwiczenie nr 3, sekcja II**). Podobna sytuacja ma miejsce w trakcie podejmowania decyzji o usunięciu punktów odbiegających (patrz: **ćwiczenie nr 3, sekcja III**).

Powiedzmy wprost, że tworzenie histogramów *per pedes* jest zajęciem czasochłonnym, uciążliwym i nudnym. W związku z powyższym, proponujemy Czytelnikowi zapoznanie się z autorskim przykładem automatyzacji tworzenia histogramów.

Oto podstawowe założenia tworzenia histogramów:

1. **Liczba przedziałów** histogramu nie powinna być większa, niż **1/4 liczby wartości danej zmiennej**. Zatem: dla 20 zmiennych idealną liczbą przedziałów jest 5, dla 25 - 6, *etc*.
2. Przedziały histogramu powinny mieć jednakową szerokość.
3. Skrajne wartości danej zmiennej powinny znajdować się w środkach skrajnych przedziałów histogramu.

Na bazie powyższych założeń stworzyliśmy prosty automat do tworzenia histogramów. Został on zbudowany dla 20 wartości danej zmiennej oraz 5 przedziałów histogramu.

	A	B	C	D
1	*Zmienna:*	*Zmienna trans.:*	*Określenie przedziałów:*	
2	wart. 1	wart. I	*MIN*	=MIN(B2:B21)
3	wart. 2	wart. II	*MAX*	=MAX(B2:B21)
4	wart. 3	wart. III	*MAX-MIN*	=(D3-D2)
5	...	...	*c*	=D4/8
6	...	...	*g1=MIN-c*	=D2-D5
7	...	...	*g2=g1+2c*	=D6+2*D5
8	...	...	*g3=g1+4c*	=D6+4*D5
9	...	...	*g4=g1+6c*	=D6+6*D5
10	...	...	*g5=g1+8c*	=D6+8*D5
11	...	...	*g6=g1+10c*	=D6+10*D5
12	...	...		
13	...	...	*Granice (g_n):*	*Wart. > (g_n)*
14	...	...	=ZAOKR.DO.TEKST(D6;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C14)
15	...	...	=ZAOKR.DO.TEKST(D7;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C15)
16	...	...	=ZAOKR.DO.TEKST(D8;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C16)
17	...	...	=ZAOKR.DO.TEKST(D9;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C17)
18	...	...	=ZAOKR.DO.TEKST(D10;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C18)
19	...	...	=ZAOKR.DO.TEKST(D11;3;FAŁSZ)	=LICZ.JEŻELI(B2:B21;">"&C19)
20	wart. 19	wart. XIX	-	
21	wart. 20	wart. XX	*Przedziały:*	*Liczba wart. w przedziałach:*
22			=ZŁĄCZ.TEKSTY(C14;C20;C15)	=D14-D15
23			=ZŁĄCZ.TEKSTY(C15;C20;C16)	=D15-D16
24			=ZŁĄCZ.TEKSTY(C16;C20;C17)	=D16-D17
25			=ZŁĄCZ.TEKSTY(C17;C20;C18)	=D17-D18
26			=ZŁĄCZ.TEKSTY(C18;C20;C19)	=D18-D19

Jak to działa?

Kursywą oznaczone są elementy opisowe w tabeli.

W kolumnach A i B znajdują się, odpowiednio: wartości "oryginalnej" zmiennej i jej ewentualne wartości po transformacji. Jeżeli wykonujemy histogram zmiennej bez transformacji, rolę kolumny **B** gra kolumna **A**. Możemy również przekopiować kolumnę **A** w miejsce kolumny **B**.

W komórkach D2:D11 znajdują się formuły obliczające graniczne wartości przedziałów. Ponieważ założyliśmy 5 równych przedziałów, a wartości **MIN** i **MAX** mają znajdować się w środku przedziałów **I** oraz **V**, to między wartością **MIN** i **MAX** powinno znajdować się 8 równych odcinków[9], których długość oznaczyliśmy jako **c**. Ponadto, jeden odcinek o długości **c** będzie znajdował się po lewej od wartości **MIN** oraz po prawej od wartości **MAX**; sumarycznie otrzymamy zatem - dla 5 przedziałów - 10 odcinków. W grupach po dwa tworzą kolejno odpowiednie przedziały. Można to przedstawić w sposób następujący:

```
przedział I: g1 --c-- MIN --c-- g2
przedział II: g2 --c-- --c-- g3
przedział III: g3 --c-- --c-- g4
przedział IV: g4 --c-- --c-- g5
przedział V: g5 --c-- MAX --c-- g6
```

W komórkach C14:C19 znajdują się funkcje transformujące wartości graniczne przedziałów do postaci tekstu, aby mogły posłużyć do opisu wykresu. Drugi argument funkcji **=ZAOKR.DO.TEKST()** (w omawianym przykładzie jest to wartość **3**) reguluje długość rozwinięcia dziesiętnego granic przedziałów histogramu. Należy dobrać jego wartość wg własnych potrzeb.

W komórkach D14:D19 znajdują się funkcje obliczające liczby wartości zmiennej, które są większe od danej granicy przedziału w histogramie. W naszym przypadku: w **D14** powinno wyjść 20, ponieważ wszystkie wartości muszą być większe od granicy **g1**, zaś w **D19** musi wyjść 0 - ponieważ wszystkie wartości zmiennej muszą być mniejsze od granicy **g6**.

W komórce C20 znajduje się myślnik. Stanowi on, wbrew wszelkim przypuszczeniom, istotny element całej układanki.

W komórkach C22:C26 znajdują się funkcje łączące odpowiednie teksty. Komórki te stanowią gotowy opis przedziałów histogramu.

W komórkach D22:D26 znajdują się funkcje obliczające liczby wartości zmiennej w danych przedziałach dzięki odejmowaniu elementów zawartych w komórkach **D14:D19**, np.: w przedziale **I** znajduje się tyle wartości, ile zostaje po odjęciu wartości większych od **g2** (**D15**) od wartości większych od **g1** (**D14**).

Wykres kolumnowy, którego opisami serii są komórki C22:C26, wartościami zaś - komórki D22:D26, tworzy najwyższej elegancji histogram rozkładu danej zmiennej. Operacje modyfikacji, transformacji oraz kasowania danych w kolumnach A i B są

[9] Gdyby przedziałów było 6 - odcinków pomiędzy **MIN** i **MAX** powinno być 10, stąd w **D5** figurowałoby **=D4/10**. Nieodzowne byłoby również dodanie po jednym nowym wierszu do obliczeń w komórkach **C14:C19**, **D14:D19**, **C22:C26** oraz **D22:D26**.

natychmiast widoczne na histogramie, ponieważ wszystkie dane wejściowe do wykresu są automatycznie przeliczane na nowo.

Ćwiczenie nr 4:
KORELACJE POMIĘDZY ZMIENNYMI

Celem ćwiczenia jest sprawdzenie, czy pomiędzy zaproponowanymi i skontrolowanymi w trakcie poprzedniego ćwiczenia zmiennymi nie występują wyraźne korelacje, to znaczy: czy poszczególne zmienne nie niosą jakiejś wspólnej informacji. Można tego dokonać na dwa, uzupełniające się sposoby:

1) obliczając współczynniki korelacji liniowej (**r**) i determinacji (**d**) dla poszczególnych par zmiennych;

2) wykonując wykresy korelacyjne dla wszystkich, możliwych par zmiennych.

Dodatkowo, wykresy korelacyjne wstępnie pomogą wychwycić tendencję obiektów do formowania grup (na razie – w układzie współrzędnych jedynie dwóch zmiennych) oraz ewentualne punkty odbiegające we wszystkich możliwych układach współrzędnych.

I. OBLICZENIE WSPÓŁCZYNNIKÓW KORELACJI LINIOWEJ ORAZ DETERMINACJI.

I.1. Współczynnik korelacji liniowej (r).

Matematyczny wzór na współczynnik korelacji liniowej (**r**) jest skrajnie przerażający i znajduje się w literaturze. Aby wyznaczyć **r** dla wybranej pary zmiennych, warto skorzystać z funkcji *Excela*:

```
=WSP.KORELACJI(zakres_wartości_pierwszej_zmiennej;
zakres_wartości_drugiej_zmiennej)
```

Ponieważ wymagane jest obliczenie wartości **r** dla wszystkich możliwych par zmiennych, najwygodniejsze będzie zbudowanie tzw. **macierzy współczynników korelacji liniowej**. Wygląda ona następująco:

	W	X	Y	Z
W	$r_{W,W} = 1$	$r_{X,W}$	$r_{Y,W}$	$r_{Z,W}$
X	$r_{W,X}$	$r_{X,X} = 1$	$r_{Y,X}$	$r_{Z,X}$
Y	$r_{W,Y}$	$r_{X,Y}$	$r_{Y,Y} = 1$	$r_{Z,Y}$
Z	$r_{W,Z}$	$r_{X,Z}$	$r_{Y,Z}$	$r_{Z,Z} = 1$

Skoro $r_{I,J} = r_{J,I}$, oczywiste jest, że górny trójkąt macierzy powtarza informację zawartą w trójkącie dolnym (i *vice versa*). W związku z powyższym, wystarczy obliczyć dowolną połowę macierzy (górny lub dolny trójkąt) oraz przekątną. Przekątna zawsze składa się z samych jedynek, należy ją jednak obliczyć dla porządku i spokoju sumienia.

Przykład:

Dla analizowanego zestawu **MIECZE,** macierz współczynników korelacji prezentuje się następująco:

	DC*	DG*	DR	M	SM	T	CR*
DC*	1,000	0,959	0,737	0,916	0,754	0,899	0,319
DG*	0,959	1,000	0,563	0,886	0,730	0,838	0,333
DR	0,737	0,563	1,000	0,679	0,620	0,693	0,275
M	0,916	0,886	0,679	1,000	0,821	0,886	0,250
SM	0,754	0,730	0,620	0,821	1,000	0,724	0,105
T	0,899	0,838	0,693	0,886	0,724	1,000	0,276
CR*	0,319	0,333	0,275	0,250	0,105	0,276	1,000

II.2. Współczynnik determinacji (d).

Współczynnik determinacji (**d**) dla pary zmiennych dany jest mało skomplikowanym wzorem:

$$d_{I,J} = (r_{I,J})^2$$

Zbudowanie **macierzy współczynników determinacji** sprowadza się zatem do podniesienia do kwadratu wartości zawartych w macierzy współczynników korelacji liniowej.

	W	X	Y	Z
W	$d_{W,W} = 1$	$d_{X,W}$	$d_{Y,W}$	$d_{Z,W}$
X	$d_{W,X}$	$d_{X,X} = 1$	$d_{Y,X}$	$d_{Z,X}$
Y	$d_{W,Y}$	$d_{X,Y}$	$d_{Y,Y} = 1$	$d_{Z,Y}$
Z	$d_{W,Z}$	$d_{X,Z}$	$d_{Y,Z}$	$d_{Z,Z} = 1$

Przykład, c.d.:

Dla analizowanego zestawu **MIECZE,** macierz współczynników determinacji prezentuje się następująco:

	DC*	DG*	DR	M	SM	T	CR*
DC*	1,000	0,920	0,543	0,840	0,569	0,808	0,101
DG*	0,920	1,000	0,317	0,786	0,533	0,702	0,111
DR	0,543	0,317	1,000	0,461	0,384	0,481	0,075
M	0,840	0,786	0,461	1,000	0,675	0,786	0,062
SM	0,569	0,533	0,384	0,675	1,000	0,525	0,011
T	0,808	0,702	0,481	0,786	0,525	1,000	0,076
CR*	0,101	0,111	0,075	0,062	0,011	0,076	1,000

II. WYKRESY KORELACYJNE DLA PAR ZMIENNYCH.

Dla **n** zmiennych, możliwe jest stworzenie **n nad 2** ich par, a co za tym idzie – tyleż samo wykresów korelacyjnych. Ich wykonanie oraz interpretację zaprezentujemy na przykładzie.

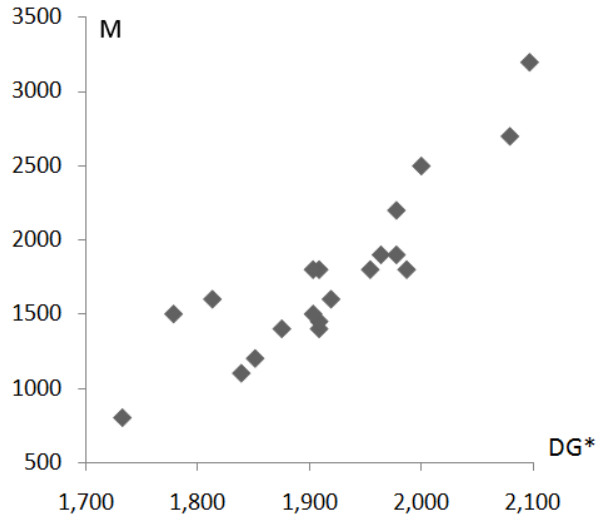
Przykład:

Wybierzmy zmienne opisujące parametry miecza – długość głowni (**DG***) oraz masę całkowitą (**M**). Zmienne te przyjmują wartości:

Obiekt\Zmienna	DG*	M
AER	1,964	1900
AND	2,000	2500
AZU	1,851	1200
BAL	1,875	1400
DUR	1,908	1400
EXC	1,954	1800
GLA	1,978	1900
GOL	1,839	1100
GRA	1,919	1600
GUR	1,908	1800
GWY	1,908	1450
HER	1,778	1500
HUR	1,813	1600
JOY	1,903	1500
LOD	1,903	1800
ORK	1,987	1800
SIH	1,978	2200
UMB	2,097	3200
URI	2,079	2700
ZAD	1,732	800

Dysponujemy zatem dwudziestoma punktami o współrzędnych odpowiadających wartościom wybranych zmiennych. Wartości te należy nanieść na zwyczajny wykres punktowy (X,Y), skonfrontować z wartościami **r** i **d** dla danej pary cech, a następnie dokonać interpretacji uzyskanego obrazu.

Przykład, c.d.:



DG* , M

r = 0,886
d = 0,786

Wyraźnie widoczna zależność liniowa. Masa miecza wzrasta liniowo wraz z długością głowni (ostrza), co w zupełności zgadza się ze zdrowym rozsądkiem.

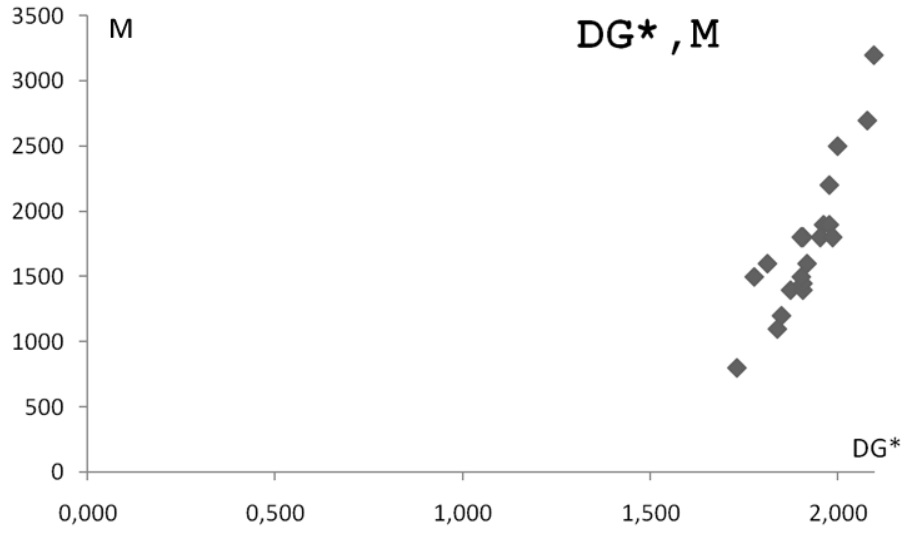
Poprawnie wykonany wykres korelacyjny powinien być zbudowany na planie kwadratu i nie zawierać pustych przedziałów na osiach.

Zastosowanie się do wymogu pierwszego (wykres na planie kwadratu) ułatwia wychwycenie nawet niewielkich tendencji analizowanych zmiennych do współliniowości. Poszukiwana linia trendu układa się wtedy pod kątem ok. 45°, a taki trend - zgodnie z wynikami badań psychologicznych - jest dla badacza najłatwiejszy do wychwycenia.

Spełnienie wymogu drugiego (brak pustych przedziałów na osiach) prowadzi do optymalnego wykorzystania całej, dostępnej przestrzeni wykresu. Należy zwrócić na to szczególną uwagę w przypadku wykonywania wykresów korelacyjnych w *Excelu*, gdyż program ten ma tendencję do automatycznego nadawania osiom wykresu wartości minimalnych (lub maksymalnych) równych 0. W efekcie, często nawet więcej niż połowa obszaru wykresu nad osią może nie zawierać ani jednego punktu.

Przykład, c.d.:

Poniżej prezentujemy wykres korelacyjny dla pary zmiennych **DG*** i **M**, wykonany absolutnie nieprawidłowo:



O tempora, o mores![10]

III. SPRAWOZDANIE.

W sprawozdaniu Student powinien umieścić:

- macierz współczynników korelacji liniowej;
- macierz współczynników determinacji;

[10] *"Co za czasy! Co za obyczaje!"* - Cyceron.

- wszystkie możliwe wykresy korelacyjne dla par zmiennych, zawierające dodatkowo: wartości **`r`** oraz **`d`**, a także krótki komentarz dotyczący informacji, jaką niesie wykres. Oto pytania pomocnicze:
  - ✓ *czy widoczna jest liniowa zależność pomiędzy zmiennymi?*
  - ✓ *czy widoczna jest zależność nieliniowa?*
  - ✓ *czy wysoka wartość współczynnika korelacji/determinacji rzeczywiście odpowiada liniowej zależności?*
  - ✓ *czy obiekty mają tendencję do tworzenia grup?*
  - ✓ *czy widoczne są wyraźne punkty odbiegające?*