Jest to zapis odczytu wygłoszonego na XLIV Szkole Matematyki Poglądowej *Do czego to się przydaje?*, Sulejów, styczeń 2010.

Statystyka użyteczna prawie wszędzie?

Andrzej DĄBROWSKI, Wrocław

W końcu XVIII wieku rozważano problem, czy komety należą do układu słonecznego. Pierre Simon Laplace (1749–1827) francuski matematyk, astronom, geodeta i fizyk na pytanie to odpowiedział negatywnie w roku 1812.

A nawet minister spraw wewnętrznych. Napoleon, wspominał, że usunął go z tego stanowiska po sześciu tygodniach ponieważ *wniósł do rządu ducha nieskończenie małych.*

Laplace twierdził, że orbity komet przecinają się pod losowym kątem z ekliptyką, co jest sprzeczne z ich przynależnością do Układu Słonecznego. Twierdzenie to oparł na obserwacji, że średni kąt przecięcia orbit komet z ekliptyką w przybliżeniu pokrywa się z wielkością oczekiwaną kąta przy założeniu losowości orbit. Wcześniej, w roku 1796, w podobny sposób Laplace uzasadnił hipotezę mgławicową. Pokazał, że ekliptyki planet nie są losowe i koncentrują się w pobliżu jednej płaszczyzny, wyznaczającej położenie dysku, z którego powstały planety.

Hipoteza mgławicowa przypisywana jest również Emanuelowi Swedenborgowi i Immanuelowi Kantowi. Głosi ona, że planety Układu Słonecznego powstały z płaskiego dysku utworzonego z materii, która odrywała się stopniowo od masy w centrum tworzącej Słońce.

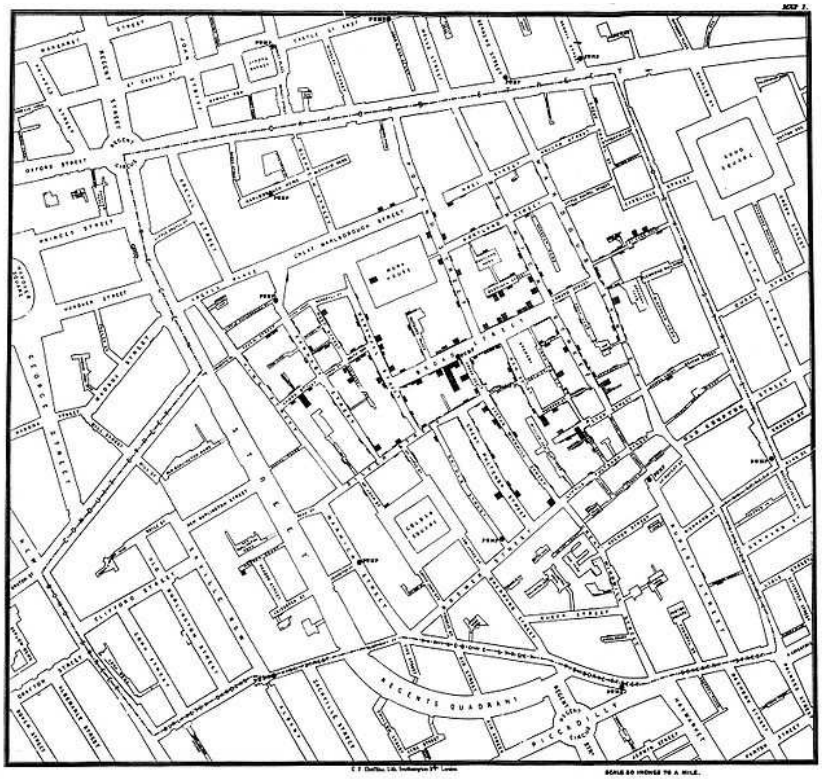
W obu przypadkach Laplace posłużył się metodą argumentacji podobną do rozumowania niewprost. Gdyby komety należały do Układu Słonecznego to kąt przecięcia ich orbit z ekliptyką winien być nieprzypadkowy. Skoro jest przypadkowy – to nie należą do Układu. Gdyby hipoteza mgławicowa nie była prawdziwa, to ekliptyki planet musiałyby być losowe – a nie są. Jest to argument za przyjęciem tej hipotezy.

Taki sposób postępowania, w którym dostateczne nagromadzenie faktów stanowi silny argument na rzecz rozważanej hipotezy jest charakterystyczny dla rozumowania statystycznego. Teoria statystyki, rozwinięta w XX wieku dostarcza kryteriów, zwanych testami, które pozwalają ocenić, jak silne są argumenty zgromadzone w danych.

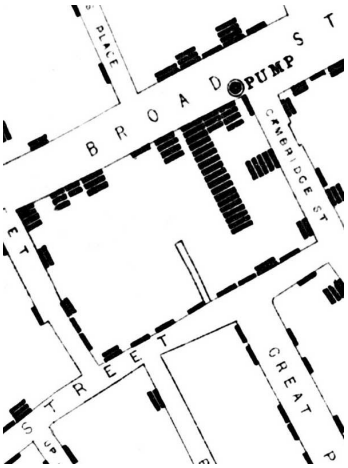
Krótko mówiąc, statystyka pozwala ocenić, czy można już zakrzyknąć: *To nie może być przypadek!* czy: *To czysty przypadek!* Czasami wypracowanie tej oceny wymaga pomysłowości i wyrafinowanych metod matematycznych.

Historie, które opowiem, zdarzyły się naprawdę. Pokazują, jak powstają nowe metody statystyczne i jak statystyka umożliwia rozstrzygnięcie ważnych problemów.

Cholera



Latem 1854 roku w Londynie wybuchła wielka epidemia cholery. Takie wybuchy epidemii zdarzały się w Londynie dość często. Przyczynami powstawania i rozprzestrzeniania się choroby od dawna interesował się znany lekarz dr John Snow. Już podczas poprzedniej epidemii cholery dr John Snow przypuszczał, że choroba spowodowana jest przez substancje, które dostały się z wodą przez przewód pokarmowy, a nie przez oddychanie miazmatami unoszącymi się w powietrzu, jak twierdziła ówczesna medycyna. Sporządził on mapę dzielnicy Soho (w której było zachorowało 3/4 mieszkańców), na której zaznaczył miejsca poboru wody i lokalizację mieszkań zmarłych mieszkańców (czarne punkty na mapie). W środku obszaru znalazła się pompa na Broad Street. Woda do pompy na Broad Street czerpana była z Tamizy, która była tak zanieczyszczona, że okna Parlamentu wychodzące na rzekę, mimo upałów, były tego lata zamknięte.



Spośród 89 zmarłych tylko 10 mieszkało w pobliżu innej pompy. Co dziwne, żaden z 70 pracowników browaru przy Broad Street nie zachorował. Również 535 mieszkańców hotelu robotniczego przy Poland Street, mieszczącego się w samym środku obszaru epidemii, nie zachorowało. Okazało się, że zarówno pracownicy browaru jaki i mieszkańcy hotelu robotniczego czerpali wodę z własnego ujęcia.

Dr Snow sporządził zestawienia częstości zachorowań na cholerę w różnych punktach Londynu i pomiarów czystości wody w pobliżu tych miejsc. Była to pierwsza w historii medycyny tak precyzyjnie przeprowadzona analiza epidemiologiczna.

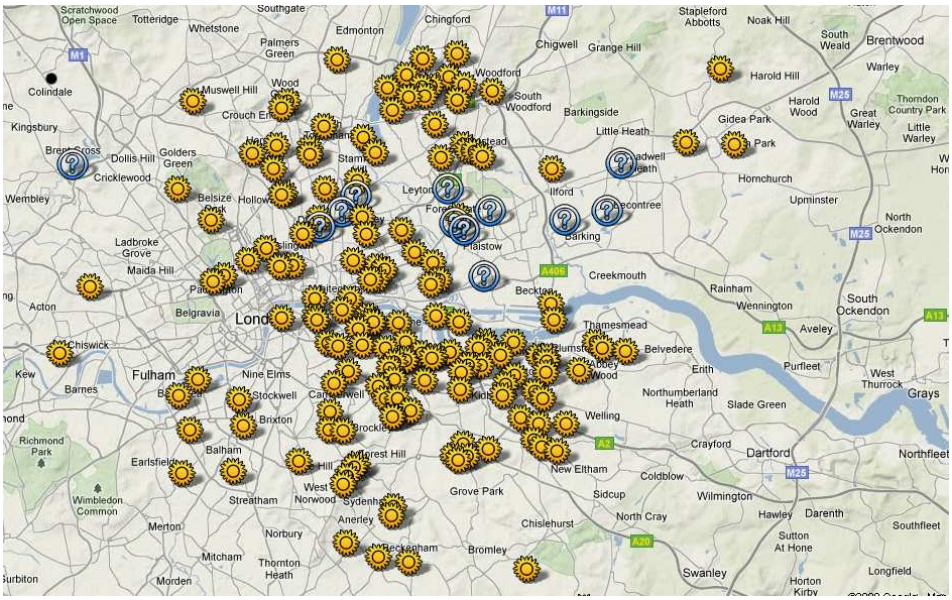
Na wniosek dr Snowa usunięto rączkę pompy na Broad Street, co uniemożliwiło pobieranie stąd wody. Liczba zachorowań w dzielnicy Soho ze 142 w dniu 1 września zmalała do 14 w dzień po usunięciu rączki 7 września.

Ustalenie przyczyn epidemii cholery przez Snowa w dużej mierze zależało od analizy miejsc występowania ognisk choroby. Taką analizę skupień charakterystycznych obiektów przeprowadza się we współczesnej medycynie – w badaniach profilaktycznych, epidemiologii, w diagnostyce używającej różnych metod obrazowania (rentgen, mammografia, tomograf, rezonans magnetyczny).

Wskazanie od kiedy lub na jakim obszarze występuje nieprzypadkowo duże skupienie niekorzystnych zjawisk jest zadaniem nowego działu statystyki – statystyki skaningowej. Statystyka skaningowa ma zastosowanie również w genetyce, astronomii (badanie skupisk gwiazd i galaktyk) i archeologii.

V2

Jesienią 1944 roku Londyn został zaatakowany przez nową, niezwykle groźną broń niemiecką. Rakiety V2 (V od *Vergeltungswaffe* – broń zemsty) ważyły 13 ton i rozwijały czterokrotną prędkość dźwięku (4800 km/godz). Ulokowane głównie na wybrzeżu Holandii, średnio w odległości 250 km od Londynu, lecąc na wysokości 80 km docierały do stolicy Wielkiej Brytanii po 3 minutach od startu. Obrona przeciwlotnicza nie mogła więc podjąć skutecznej obrony. Żadna z rakiet nie została zniszczona, a miejsca upadku rakiety były trudne do przewidzenia.



Mapa upadku pocisków V2, zrekonstruowana na stronie internetowej Londynu
`http://londonist.com/2009/01/london_v2_rocket_sitesmapped.php`

Nic dziwnego, że w tej sytuacji zaczęły się mnożyć spekulacje. Mieszkańcy Londynu podejrzewali, że Niemcy atakowali szczególnie zaciekle tylko niektóre miejsca. Powiadano też, że w oszczędzonych przez V2 dzielnicach na pewno mieszkają sympatycy nazistów. Rozpoczęły się paniczne przeprowadzki do przypuszczalnie bardziej bezpiecznych miejsc. Zapadła decyzja o weryfikacji, czy miejsca upadku rakiet tworzą nieprzypadkowe skupienia.

Sprawdzenie, czy obserwacje pojawiają się całkowicie losowo było jednym z pierwszych zagadnień statystycznych, rozwiązanych przez Karla Pearsona (1857-1936). Skonstruował on test χ^2, który pozwala ocenić stopień zgodności danych z postulowanym rozkładem prawdopodobieństwa, będącym modelem losowości. Obszar Londynu o rozmiarach 12 na 12 km został podzielony na 576 kwadratów o boku 500 m.

Za pomocą testu χ^2 Pearson w 1900 rozstrzygnął problem, czy liczby pojawiają się przypadkowo na kole ruletki w Monte Carlo.

Policzono liczbę rakiet V2, które spadły na każdy kwadrat.

liczba rakiet	0	1	2	3	4	7
liczba kwadratów	229	211	93	35	7	1

Łącznie spadło na ten obszar 537 rakiet. Średnio na kwadrat przypadło więc $537/576 \approx 0{,}93$ rakiety. Przyjmując założenie, że naturalnym modelem losowości w tym przypadku jest rozkład Poissona, oczekiwana liczba rakiet w kwadratach wynosi

liczba rakiet	0	1	2	3	4	?5
oczekiwana liczba kwadratów	226,7	211,4	98,5	30,6	7,1	1,6

Te dwie tabele są zadziwiająco zgodne. Można oszacować, że prawdopodobieństwo pojawienia się takiej zgodności przy założeniu losowości, wynosi 0,88. Jest więc dostatecznie duże, aby przyjąć, że pociski spadały przypadkowo na Londyn.

Wyniki tych analiz zostały opublikowane po wojnie, w 1946 roku w *Journal of the Institute of Actuaries* przez pracownika firmy ubezpieczeniowej Prudential R.D. Clarka w pracy pod tytułem *An Application of the Poisson Distribution.*

Problem seriacji

Wybitny egiptolog brytyjski Flinders Petrie (1853–1942), prowadząc u schyłku XIX wieku wykopaliska w Diospolis Parva w Egipcie napotkał groby, zawierające fragmenty ceramiki. Niestety, warstwy zostały tak przemieszczone, że datowanie znalezisk metodą stratygrafii okazało się niemożliwe. Jedynym sposobem było uporządkowanie w czasie kształtów i motywów występujących na ceramice. Petrie wypisał zawartość każdego grobu na kartce papieru zaznaczając typy ceramiki w kolumnach (1 oznacza występowanie motywu, 0 – brak):

Petrie przyjął założenie, że różne rodzaje ceramiki pojawiały się w grobach w określonym czasie – stąd uporządkowanie według motywów daje możliwość uporządkowania w czasie. Takie datowanie nazwano datowaniem względnym, w przeciwieństwie do metod datowania bezwzględnego, takiego jak metoda węgla C-14.

Groby	puchar	czarny brzeg	butla	płaskie	rączka	punkty	spirale
g1	0	1	1	0	1	0	0
g2	0	0	0	1	0	1	0
g3	0	0	0	0	0	1	0
g4	0	0	0	1	1	1	1
g5	1	1	0	0	0	0	0
g6	0	0	0	0	1	0	1

Egiptolog przestawiał karteczki z grobami i zmieniał kolejność kolumn tak, aż uzyskał satysfakcjonującą konfigurację:

Groby	puchar	czarny brzeg	butla	rączka	spirale	płaskie	punkty
g5	1	1	0	0	0	0	0
g1	0	1	1	1	0	0	0
g6	0	0	0	1	1	0	0
g4	0	0	0	1	1	1	1
g2	0	0	0	0	0	1	1
g3	0	0	0	0	0	0	1

Zagadnieniem przekształcenia macierzy kontyngencji do postaci Petriego zainteresował się pierwszy w historii profesor statystyki na Uniwersytecie Oksfordzkim David Kendall (1918–2007). W pracy *Incidence Matrices, Interval Graphs and Seriation in Archaeology* , opublikowanej w Pacific Journal of Matematics w 1969, podał warunki konieczne i dostateczne na to, aby macierz można było przekształcić do postaci Petriego. Warunki podane przez Kendalla nieco ułatwiły poszukiwanie seriacji, ale nie rozwiązały tego zagadnienia. Jeśli macierz nie da się sprowadzić do postaci Petriego, to z praktycznego punktu widzenia satysfakcjonujące jest doprowadzenie jej do postaci bliskiej macierzy Petriego.

Więcej na ten temat można przeczytać w artykule Dąbrowski, A. *Obrazy zależności*, Matematyka, Społeczeństwo, Nauczanie 27 (2001) str. 26–34.

Uznał przy tym (co zostało zaakceptowane przez środowisko archeologów), że tak uporządkowane sekwencje motywów wskazują na sekwencje czasowe grobów – wspólne motywy wskazują na sąsiedztwo grobów w czasie.

Ustalenie porządku wierszy i kolumn tak aby jedynki i zera występowały w macierzy w zwartych grupach nazwano seriacją. Macierze o tej własności, nazwano macierzami Petriego. Przekształcenie macierzy kontyngencji do postaci Petriego jest trudnym i atrakcyjnym problemem matematycznym.

Do rozwiązania tego zagadnienia zastosowano z pozoru bezsensowną metodę: położeniom i motywom przypisano wartości punktowe.

Jako przykład niech posłuży prosta tabela kontyngencji M z czterema lokalizacjami P1, P2, P3 i P4 oraz trzema motywami A, B i C.

	A	B	C
P1	1	0	1
P2	0	0	1
P3	1	0	0
P4	0	1	1

Na przykład, położeniom P1, P2, P3 i P4 można przypisać arbitralnie 1, 2, -3 i 3 punkty. Podobnie, motywom A, B i C można przypisać 4, 2 i -2 punkty. Systemy te generują punktację wtórną.

Punktacja wtórna dla wierszy jest generowana przez punktację kolumn. Oceniając motywy występujące w P1, można zgromadzić 2 punkty $(1 * 4 + 0 * 2 + 1 * (-2) = 2)$, co oznacza średnio 1 punkt za znalezisko w lokalizacji P1. Podobnie, za jedno znalezisko w lokalizacji P2 należałoby przyznać -2 punkty, a w lokalizacjach P3 i P4 po 4 i 0 punktów. Punktacja wtórna za wiersze wynosi $(1, -2, 4, 0)$. W analogiczny sposób można wyznaczyć punktację wtórną dla kolumn.

		A	B	C	punktacja wierszy	
					pierwotna	wtórna
P1		1	0	1	1	1
P2		0	0	1	2	-2
P3		1	0	0	-3	4
P4		0	1	1	3	0
punktacja kolumn	pierwotna	4	2	-2		
	wtórna	-1	3	2		

Punktacja zaproponowana dla tej macierzy nie jest stabilna. Punktacja jest stabilna, gdy punktacje pierwotna i wtórna wierszy i kolumn są proporcjonalne. Dla punktacji stabilnej skala wierszy (kolumn) to współczynnik proporcjonalności punktacji wtórnej do pierwotnej dla wierszy (kolumn). Waga stabilnej punktacji to iloczyn skal wierszy i kolumn. Każda macierz kontyngencji ma stabilną punktację. Stabilnych i istotnie różnych (nieproporcjonalnych) systemów punktacji jest $\min(w, k)$, gdzie w jest liczbą wierszy, k – liczbą kolumn.

Jedna ze stabilnych punktacji jest trywialna: gdy każdy wiersz i każda kolumna ma taką samą liczbę punktów. Waga trywialnej punktacji wynosi 1.

Dwie nietrywialne punktacje dla macierzy M są postaci:

punktacja o wadze 2/3

		A	B	C	punktacja wierszy	
					pierwotna	wtórna
P1		1	0	1	-3	-1
P2		0	0	1	3	1
P3		1	0	0	-9	-3
P4		0	1	1	6	2
punktacja kolumn	pierwotna	-3	3	1		
	wtórna	-6	6	2		

punktacja o wadze 1/4

		A	B	C	punktacja wierszy	
					pierwotna	wtórna
P1		1	0	1	-2	-1
P2		0	0	1	-8	-4
P3		1	0	0	4	2
P4		0	1	1	4	2
punktacja kolumn	pierwotna	2	8	-4		
	wtórna	1	4	-2		

Macierze Petriego i punktacje stabilne łączy ważny fakt:

> *Uporządkowanie wierszy i kolumn tworzące macierz najbardziej zbliżoną do macierzy Petriego wyznacza stabilna, nietrywialna punktacja o największej wadze.*

Macierz Petriego można uzyskać porządkując wiersze i kolumny tak, aby punktacja wierszy i kolumn była monotoniczna (malejąca lub rosnąca).

Po takim uporządkowaniu wierszy, a potem kolumn otrzymana macierz M jest w postaci Petriego.

	A	C	B	punktacja
P3	1	0	0	−9
P1	1	1	0	−3
P2	0	1	0	3
P4	0	1	1	6
punktacja	−3	1	3	

Stabilna, nietrywialna punktacja o maksymalnej wadze 0,86 dla przykładu archeologicznego jest postaci

Groby	puchar	czarny brzeg	butla	płaskie	rączka	punkty	spirale	punktacja pierwotna	punktacja wtórna
g1	0	1	1	0	1	0	0	0,77	0,71
g2	0	0	0	1	0	1	0	−0,93	−0,86
g3	0	0	0	0	0	1	0	−0,97	−0,90
g4	0	0	0	1	1	1	1	−0,60	−0,56
g5	1	1	0	0	0	0	0	1,74	1,61
g6	0	0	0	0	1	0	1	−0,28	−0,26
pierwotna punktacja	1,87	1,35	0,83	−0,82	−0,039	−0,90	−0,47	skala	0,93
wtórna	1,74	1,25	0,77	−0,76	−0,036	−0,83	−0,44	0,93	0,86

Po uporządkowaniu według punktacji otrzymać można macierz, uzyskaną metodą prób i błędów przez Petriego.

Analiza korespondencji stworzona została w 1964, przez profesora Instytutu Statystyki Uniwersytetu Paryskiego, Jeana Paula Benzécriego (1932–) na potrzeby badań lingwistycznych.

Algorytm poszukiwania optymalnej punktacji jest częścią metody statystycznej, zwanej analizą korespondencji (analizą odpowiedniości). Za pomocą analizy korespondencji bada się zależności między cechami. Stosuje się ją w naukach społecznych, wspomnianej tu archeologii i medycynie.

Brak kolorów praktycznie uniemożliwia zademonstrowanie tutaj niesłychanie efektywnego i efektownego przykładu wykorzystania macierzy Petriego w genetyce przedstawionego w pracy Olena Morozova i inni, *A Seriation Approach for Visualization-Driven Discovery of Co-Expression Patterns in Serial Analysis of Gene Expression (SAGE) Data* `http://www.plosone.org/article/info:doi%2F10.1371%2Fjournal.pone.0003205`. Polecam obejrzenie tej strony.