

Nie można się oprzeć wrażeniu, że formuły matematyczne mają niezależny od nas byt i inteligencję, że są mądrzejsze niż my sami, nawet mądrzejsze niż ich odkrywcy i że możemy wywnioskować z nich więcej niż poprzednio w nich zawarto.

Heinrich Rudolph Hertz, fizyk, 1857 - 1894

## Statystyka opisowa

Wykład ten w zasadzie poświęcony jest statystyce, elementom statystyki w nauczaniu szkolnym. Jest napisany trochę jak fragment podręcznika. Chcę jednak naświetlić i inne, szersze zagadnienie – problem interpretacji danych liczbowych i podejmowania decyzji na ich podstawie. Jak czytać dane statystyczne, jak (nie dać się) oszukiwać za ich pomocą, jak je obrabiać, żeby uzyskać korzystny dla nas efekt – i jak poznać, że ktoś przygotował nam taki spreparowany pakiet. Uważam to za bardzo ważny motyw nauczania statystyki w szkole. Najbardziej bowiem dokuczliwą cechą statystyki jest to, że świetnie nadaje się do ... propagandy. Łatwo jest oszukiwać za pomocą statystyki, a więc jakby w majestacie matematyki. Znane jest powiedzenie, że dane statystyczne to jak ludzie: wystarczy je dostatecznie długo pomęczyć, a powiedzą wszystko. Myśl, która przewijać się będzie przez ten rozdział, mogę wyrazić tak:

Jak (nie dać się) oszukiwać za pomocą statystyki? Jak zmusić liczby, by pracowały dla nas?

Wątki te przewijały się i w poprzednich wykładach.

Elementy statystyki znajdują się od kilku lat w programach nauczania dla gimnazjów i dla liceów. Tych treści jest niewiele. Uczymy tylko o średnich (praktycznie: arytmetyczna, mediana i moda) i odchyleniu standardowym. Dotykamy też sztuki właściwej prezentacji danych statystycznych. Niewielka liczba godzin, które możemy przeznaczyć na statystykę, spowoduje, że wiedza naszych uczniów będzie bardzo powierzchowna. Tkwi to jednak w założeniu obecnie obowiązującej podstawy programowej.

W szkole nie uczymy *statystyki*, tylko *statystyki opisowej*. Sama statystyka jest bowiem konkretną dyscypliną matematyczną. *Statystyka opisowa* jest zaś sztuką stosowaną. Uczymy tylko stosowania metod matematycznych do opisu zjawisk życia społecznego, gospodarczego, do analizy wielkich zbiorów danych.

Nie jest dla mnie wcale jasne, że elementy statystyki opisowej mają stanowić część programu matematyki. Konieczność wykonywania stosunkowo skomplikowanych obliczeń (a raczej: prostych rachunków w dużej ilości) może być argumentem za umieszczeniem statystyki na lekcjach informatyki. Z kolei powiązanie z prawdziwymi problemami życia codziennego może skłaniać do umieszczenia elementów statystyki w wielu innych przedmiotach: przede wszystkim geografii i elementach przedsiębiorczości. Przeciw temu rozwiązaniu przemawia

jednak smutna rzeczywistość. W statystyce jest trochę matematyki, a fobii antymatematycznej ulega wraz z dużą częścią społeczeństwa nie mniejsza frakcja ... nauczycieli. Nauczanie statystyki w wykonaniu nauczyciela, który nie rozumie matematyki, mogłoby być ... powiem łagodnie, niewłaściwe.<sup>1</sup>

Statystyka, jak wiemy, opiera się na dużych zbiorach danych. To pierwsza zasada. Jeśli chirurg wykona dwie trudne operacje i powiedzie się jedna, to nie może twierdzić, że ma pięćdziesięcioprocentową skuteczność. Jaki zbiór uznajemy za „duży”, zależy od konkretnej sytuacji. Preferencje wyborcze Polaków badane są zwykle na próbie kilkusetosobowej. Do badań dydaktycznych też potrzebne są duże zbiorowości. Można powiedzieć, że statystyka polega na badaniu dużych zbiorowości za pomocą badania wyselekcjonowanych próbek.

Trochę terminologii. Wyobraźmy sobie, że chcę coś wiedzieć o długości stóp Polaków (niech to będą na przykład dorośli mężczyźni). Chcę znać średnią długość, odchylenie standardowe i tak dalej. Chcę znać *parametry* rozkładu zmiennej losowej „długość stopy”. Jak mam poznać te parametry? Wybieram losową, reprezentatywną próbkę, czyli na przykład 150 mężczyzn. Wyznaczam średnią długość stopy tej grupy. Ta średnia jest jedną ze *statystyk* rozkładu. Słowo *statystyka* ma zatem dwa znaczenia. Może to być dyscyplina matematyczna, ale tak samo nazywamy wielkość, opisującą *próbki*. Gdy *próbka* zmienia się w *populację*, *statystyka* przechodzi w *parametr*. Dopiero po takim wyjaśnieniu możemy (pamiętając, że *estymacja* to *oszacowanie*) zrozumieć pozornie bezsensowne zdanie:

Statystyka polega na estymacji parametrów na podstawie znajomości statystyk.

Drugą cechą badań statystycznych jest to, że stale balansujemy tam między dokładnością a wiarygodnością. Nieustannie musimy o tym pamiętać. Jeśli mówię, że 58,13 % Polaków popiera działania rządu, jestem dokładny, ale mało wiarygodny – skąd akurat te trzynaście setnych procenta? Jeśli mówię, że odsetek ten jest między 10 a 90 procent, jestem wiarygodny, ale mało dokładny. Jedną z najważniejszych spraw w badaniach statystycznych jest właśnie osiągnięcie pewnego kompromisu między dokładnością a wiarygodnością.

Czy nam się to podoba, czy nie, statystyka coraz bardziej wchodzi w nasze życie<sup>2</sup>. Jest to częściowo uboczny efekt komputeryzacji. Obliczenia statystyczne są bardzo żmudne, czasochłonne i zniechęcające. Trzeba jednak powiedzieć, że nie *są*, lecz że *były*. Komputery zmieniły sposób nauczania statystyki. Teraz wszystko możemy mieć szybko. Niekiedy za szybko i ...za dużo. W czasie ostatnich

---

<sup>1</sup> W czasach, w których ja chodziłem do liceum, był w ostatniej klasie przedmiot *Astronomia*. Uczyli go (w różnych klasach) pan od fizyki i pani od geografii. Były to dwa zupełnie inne przedmioty. Fizyk uczył o budowie gwiazd, kosmologii, teoriach powstania Wszechświata, zahaczył o geometrię nieeuklidesową i teorię względności, omawiał procesy zachodzące wewnątrz Słońca i jak będzie przebiegać podróż na Księżyc. Mnie uczyła – wdzięcznie zresztą wspomiana przez nas wszystkich – wychowawczyni, będąca nauczycielką geografii. Uczyla nas o zmienności pór roku, o krzywej zwanej *analemma* (proszę znaleźć w Internecie, co to jest!!! A ja wiem ze szkoły!!!), zmienności w pozornym ruchu Słońca, ekliptyce, precesji, wyznaczaniu odległości do gwiazd, orientowaniu się za pomocą gwiazd, o kalendarzu, i opowiadała mity o gwiazdozbiorach. Nie wartościuję tych sposobów wypełnienia treścią hasła *Astronomia*. Nie mówię, który lepszy.

<sup>2</sup> W książce Johna Allena Paulosa *Analfabetyzm matematyczny i jego skutki* (Gdańskie Wydawnictwo Oświatowe, 1999) niemal wszystkie negatywne skutki niezajomości matematyki odnoszą się do nieumiejętnego posługiwania się rachunkiem prawdopodobieństwa i statystyką.

mistrzostw świata w piłce nożnej byliśmy bombardowani informacjami statystycznymi: ile razy gracz drużyny X kopnął piłkę lewą, a ile razy prawą nogą i ile karnych obronił w ostatnich piętnastu latach bramkarz drużyny Y. A gdy Agnieszka Radwańska przegrywała w czwartej rundzie turnieju tenisowego w Wimbledonie – a był to jej pierwszy start w dużej imprezie – do samego końca w jej *dossier* na ekranie pojawiał się napis: procent wygranych meczów w turniejach wielkoszlemowych: 100.

Rachunek prawdopodobieństwa jest bezlitosny. Totalizator Sportowy ma prawo zakładać, że zarobi na czysto tyle a tyle złotych. Nie wiadomo, czy państwo Kowalscy będą mieli synka, czy córeczkę, ale w ciągu roku w Polsce urodzi się trochę więcej chłopców niż dziewczynek. Fabryka pieluszek Pampers-Boy i Pampers-Girl może produkować stale trochę więcej pieluszek dla chłopców, a trochę mniej dla dziewczynek. Nie wiem, czy akurat *ja* będę chodził w gipsie, ale w Zakopanem wiedzą, ile gipsu potrzeba na sezon. Przypadkowe, powtarzalne zdarzenia podlegają bardzo ścisłym prawom i my wszyscy planujemy naszą przyszłość stosując intuicyjnie rachunek szans, matematycznie nazywany właśnie rachunkiem prawdopodobieństwa, albo probabilistyką. Profesjonalnie stosują ten rachunek towarzystwa ubezpieczeniowe. Gdy ubezpieczam się na życie, proponują mi składkę zależną na ogół od bardzo wielu czynników: rachują moje szanse na przeżycie.

Matematyka przypadku jest bezlitosna. Z jednej strony powiada: śpij spokojnie, dyrektorze kasyna: na 99,99 procent *to* się nie zdarzy. Zrelaksuj się, trenerze – tytuł już wasz: prowadzicie w karnych już 4:0 – przecież oni by musieli strzelić pięć a wy żadnego! Nie licz, Zosiu, że życie nam się odmieni, gdy tylko wypełnimy ten kupon Totka. Premierze: nie musisz się liczyć z tym, że woda przekroczy stan alarmowy o 8 metrów.

Ale z drugiej strony rachunek szans ostrzega: nie można wykluczyć, że ktoś w kasynie trafi numer nawet 100 razy pod rząd. Jeśli szansa zdobycia bramki z karnego jest nawet 10 do 1, to *i* tak jeszcze oni mogą wygrać. Zosiu, ktoś przecież wygrywa w Totka ... więc może w tym tygodniu my ... Panie premierze, gdyby te wały były przewidziane nie tylko na powódź stulecia, ale tysiąclecia ...

Przyzwyczajiliśmy się już do tego, że przed każdymi wyborami instytuty badania opinii prognozują wyniki i że czasami zdarzają się wpadki, jak chociażby ta z wyborów parlamentarnych we wrześniu 2005 roku, kiedy już dwa dni po wyborach OBOP szacował frekwencję na 59 procent, a faktyczna wyniosła 48. Najtrudniej jest w takich przewidywaniach dobrać reprezentatywną próbkę, na przykład 1000 osób tak, żeby było w nich odpowiednio dużo rolników, przedsiębiorców, robotników - ale także młodzieży i mieszkańców małych miast i dużych metropolii i tak dalej i tak dalej. Wybór złej próbki zafałszuje prognozy. Dobór takiej próbki to sztuka, wobec której błędnie nawet matematyka.

## Właściwa prezentacja danych

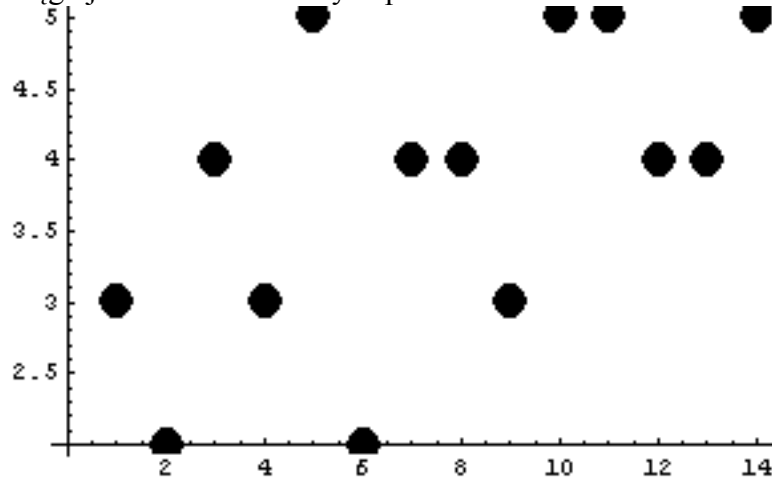
Nie będę omawiać różnych typów diagramów, wykresów itp. Zwrócę uwagę na coś istotnego: na konieczność doboru właściwej prezentacji graficznej do danych.

Przypuśćmy na przykład, że chcemy zilustrować graficznie taki szereg statystyczny:

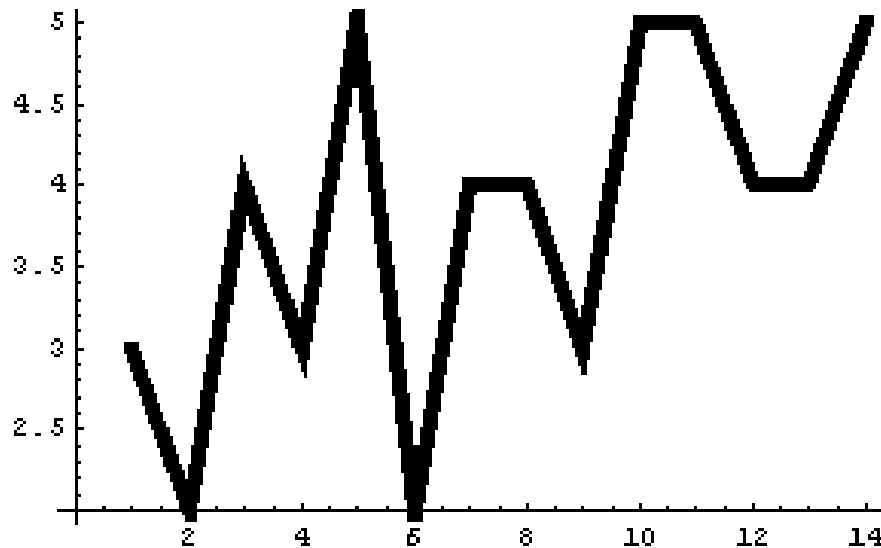
3, 2, 4, 3, 5, 2, 4, 4, 3, 5, 5, 4, 4, 5 .

Jaki wykres wybrać?

Nie ma jednej odpowiedzi. Typ wykresu zależy od typu danych. Zauważmy najpierw, że na lekcjach matematyki przyzwyczajamy uczniów, że wykresem tego ciągu jest zbiór izolowanych punktów:



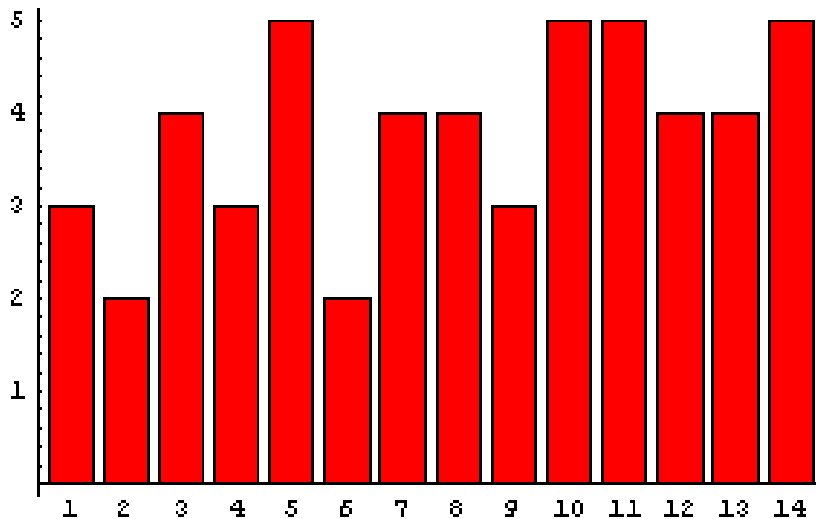
Uczniowi, który wykona wykres na przykład taki:



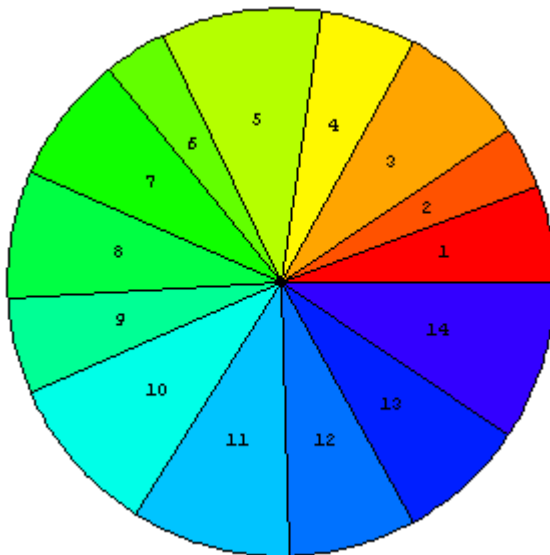
zwracamy uwagę, że jest to niepoprawne, bo funkcja nie jest określona dla niecałkowitych wartości argumentu. Tymczasem każdy z nas „w życiu codziennym” narysowałby właśnie tak.

Wyobraźmy sobie teraz, że podane liczby są stopniami matematyki, jakie Ela otrzymywała w ciągu semestru. Właśnie z wykresu drugiego widać, że Ela uczyła się coraz lepiej. Miała wzloty i upadki, ale „linia trendu” była zdecydowanie do góry. Drugi wykres ilustruje to znacznie lepiej. Możemy sobie poza tym wyobrazić, że między jednym stopniem a drugim wiedza Eli przyrastała w sposób liniowy (albo malała). Umysł nasz lubi ciągłość. Już Pitagorejczycy wpadli w kryzys ideologiczny właśnie dlatego, że nie potrafili „uciąglić” prostej. Pomówimy o tym w wykładzie 15.

Ładne i efektowne są wykresy słupkowe, takie jak ten:



Nie nadają się jednak do prezentacji ocen szkolnych. Kiedy zaś możemy zastosować wykres kołowy, taki jak na rysunku poniżej? Na przykład, gdyby były to dobrowolne składki członków klubu na jakiś cel: ktoś dał 3 złote, inny 4 albo 5. Nie jest wtedy ważny czas, a tylko procentowy wkład wpłat członków klubu na konkretny cel.



## Miary wartości centralnej

Tym skomplikowanym mianem w obecnej podstawie programowej określamy po prostu średnią. Ale jak wiemy, średnich jest bardzo dużo. Omówię tylko najczęściej używaną średnią – arytmetyczną. Jest ona najprostsza i jest to jej zarówno jej zaleta, jak i wada.

W statystyce średnią arytmetyczną oznaczamy często przez postawienie kreseczki nad zmienną. Niech na przykład  $x$  oznacza długość stopy człowieka. Jeżeli zmierzmy długości stóp pewnej liczbie osób, powiedzmy  $n$ , to matematyk powie, że otrzymaliśmy ciąg skończony  $x_1, x_2, \dots, x_n$ . Średnia arytmetyczna wyrazów takiego ciągu to liczba

$$s = \frac{x_1 + x_2 + \dots + x_n}{n}$$

Statystyk zapisze to inaczej. Niech długości stóp 11 mężczyzn będą równe odpowiednio 26, 27, 26, 25, 29, 30, 24, 28, 29, 31 cm. Statystyk zapisze, że

$$x = 26, 27, 26, 25, 29, 30, 24, 28, 29, 31, 27, 31$$

a wartością średnią (przeciętną) jest  $\bar{x} = \frac{\sum x}{11} = \frac{306}{11} = 27 \frac{9}{11}$  cm.

Czy możemy powiedzieć, że przeciętny zawodnik tej drużyny piłkarskiej ma stopę długości  $27 \frac{9}{11}$  cm? Przecież wyliczyliśmy tę przeciętną? No, to *kto* ma taką stopę? Nikt! Do tego, pierwszego paradoksu związanego z pojęciem średniej przywykamy bardzo szybko: nikt nie jest średni, nikt nie jest przeciętny. Nie dziwią nas także ułamkowe części ludzi i samochodów: jeżeli przez pierwszą godzinę przez most przejechało 157 samochodów a przez drugą 144 to średnio .... sto pięćdziesiąt i pół samochodu. Jeśli w sobotę restaurację odwiedziło 55 osób, a w niedzielę 50, to średnia frekwencja wyniosła 52 i pół osoby.....Czy wyobrażamy sobie połówkę samochodu jadącą przez most, albo pół osoby w restauracji<sup>3</sup>? A czy zdajesz sobie sprawę, Czytelniku, że w każdym państwie przeciętny obywatel ma mniej niż dwie nogi? Nie? A przecież to jasne: nikt nie ma trzech, a niektórzy są bez nogi....

Kontynuujmy ogólne rozważania na temat wielkości średniej. Obliczanie wartości średniej nie zawsze ma sens. Jeśli słoń waży dwie tony, a komar 2 miligramy, to średnio ważą tonę i miligram. Matematycznie się zgadza. Ale co to znaczy? Nic. Ale są i całkiem naturalne sytuacje, w których dopiero po chwili zastanowienia się dojdziemy do wniosku, że wykonane przez nas poprawne obliczenie wartości średniej nie ma większego sensu.

A co to znaczy, że średni, przeciętny Amerykanin jest wyższy od Japończyka, a przeciętny mężczyzna od kobiety? To wydaje się jasne i zrozumiałe. Wiemy, że to nie znaczy, że każdy Amerykanin jest wyższy od każdego Japończyka, tylko, że ..... no właśnie, na to proste pytanie jest wiele odpowiedzi, co zaraz zobaczymy.

W wykładzie o pieniądzach bardzo wiele pojęć ilustrowałem na przykładzie jednego zadania, z matury 1881. W tym wykładzie o statystyce sporą część materiału można odnosić do pewnego zadania Hugona Steinhausa. Oto ono<sup>4</sup>:

*Aki i beki.* Uczniów klasy A nazwiemy *akami*, uczniów klasy B – *bekami*. Aki chwala się, że są wyższego wzrostu niż beki, a beki uchodzą za lepszych matematyków. Gdy raz jeden z aków patrzył z góry na beka, ten zapytał: Co to właściwie znaczy, że jesteście wyżsi od nas? Czy to znaczy, że

- 1) każdy ak jest wyższy od każdego beka?
- 2) największy ak jest wyższy od największego beka?
- 3) każdy ak ma beka, od którego jest wyższy?
- 4) każdy bek ma aka, od którego jest niższy?
- 5) każdy ak ma beka, i to każdy innego, od którego jest wyższy?

<sup>3</sup> Można sobie wyobrazić: po prostu ktoś stoi albo i siedzi tak, że połowę ciała ma w pokoju, a połowę na zewnątrz! Ale „być w restauracji” połowicznie nie można.

<sup>4</sup> Hugo Steinhaus, *Sto zadań*, PWN, Warszawa 1958, zad. 61.

- 6) każdy bek ma aka, i to każdy innego, od którego jest niższy?
- 7) najmniejszy bek jest niższy od najmniejszego aka?
- 8) najmniejszy ak przewyższa więcej beków niż największy bek aków?
- 9) suma wzrostu aków jest większa niż suma wzrostu aków?
- 10) średni wzrost aków jest większy od średniego wzrostu beków?
- 11) więcej jest takich aków, którzy przewyższają jakiegoś beka, niż beków, którzy przewyższają jakiegoś aka?
- 12) więcej jest aków wzrostu wyższego od średniego wzrostu beków niż beków wzrostu wyższego od średniego wzrostu aków?
- 13) środkowy co do wzrostu ak jest wyższy od środkowego beka (w przypadku, gdy liczba uczniów w klasie jest parzysta, za wzrost środkowy uznaje się średnią arytmetyczną wzrostów środkowej pary uczniów)?

Oblany potokiem pytań ak zmał... My zaś pytamy Czytelnika: czy i które spośród 13 kwestyj są zależne od siebie? Innymi słowy: trzeba znaleźć takie pary pytań, że odpowiedź „tak” na pierwsze zmusza do odpowiedzi „tak” na drugie. Czy są pytanie równoważne, to znaczy, czy są takie pary, że odpowiedzi na oba pytania muszą być jednakowe? Czy są pary zależne, ale nierównoważne?

Rozwiązanie zadania o *akach i bekach* zostawię Czytelnikom. Nie jest trudne, a należy do takiego gatunku zadań, które łatwiej rozwiązać samodzielnie, niż przeczytać rozwiązanie. Zamieściłem je dla pokazania, jak różną interpretację matematyczną może mieć pytanie, które wydaje się proste i naturalne. Poniżej, w rozdziale o wyborach, omawiam podobny problem, ale już bardzo konkretny, wpływający na życie każdego z nas.

Przypomnę niektóre proste, szybsze sposoby obliczania średniej.

**Obliczenie średniej metodą odchyień.** Ustalamy przybliżoną średnią, „na oko”. W poniższym przykładzie wybiorę 13. Obliczamy następnie odchylenia od średniej prowizorycznej. Suma tych odchylenia, podzielona przez liczbę liczb, których średnią obliczamy, jest poprawką, którą trzeba dodać do (lub odjąć od) przyjętej przybliżonej średniej. A zatem dla liczb 13, 17, 15, 11, 13, 11, 17, 13, 11, 11 zrobimy tak:

13	17	15	11	13	11	17	13	11	11	$\Sigma$
0	+4	+2	-2	0	-2	+4	0	-2	-2	+2

Zatem prawdziwą średnią arytmetyczną jest:

$$\text{średnia prowizoryczna} + (\text{suma poprawek}) / N = 13,2 .$$

### Średnia z danych pogrupowanych.

Bardzo często zdarza się, że mamy obliczyć średnią, gdy znamy tylko dane posegregowane, ujęte w przedziały. Nie obliczymy wtedy średniej dokładnie, możemy tylko podać jej przybliżoną wartość.

Przedziały	Środek	Częstość	Suma danych z przedziału
55-59	57	1	57
50-54	52	1	52
45-49	47	3	141
40-44	42	4	168
35-39	37	6	222
30-34	32	7	224
25-29	27	12	324
20-24	22	6	132
15-19	17	8	136
10-14	12	2	24
Łączna suma danych = 1480			
Średnia = $1480/50 = 29,60$			

**Średnia ważona.**

Średni wzrost Niemca wynosi 180 cm, Polaka 178 cm, Czecha 176 cm (dane fikcyjne).

Obliczyć średni wzrost (mężczyzny) z tych krajów.

Musimy wiedzieć, ile jest Niemców, Polaków i Czechów. Przyjmijmy (dane fikcyjne):

Niemców 30 mln, Polaków 15 mln, Czechów 5 mln.

Zatem średnia to

$$\frac{30 \cdot 180 + 15 \cdot 178 + 5 \cdot 176}{50}$$

$$= 0,6 \cdot 180 + 0,3 \cdot 178 + 0,1 \cdot 176 = 179 \text{ [cm]}.$$

**Średnia z procentów.**

Wszyscy wiemy, że nie można obliczać średniej procentów. Pasztet, do którego bierzemy jednego konia i jednego zająca, nie jest złożony w 50% z mięsa końskiego i w 50% z zającego. O obliczeniach procentowych mówiłem dokładniej w wykładzie 5. W poniższym przykładzie (obrazującym odsiew w szkołach nauki jazdy) obliczenie średniego odsiewu musimy wykonać, sumując liczbę uczniów na poszczególnych kursach, liczby uczniów, którym się nie powiodło i obliczając iloraz tych dwóch wielkości.

Szkoła	Stan uczniów	Odsiew	Procent odsiewu
A	243	55	22,6
B	63	7	11,1
C	196	43	21,9
D	61	2	3,3
E	125	34	27,2
Ogółem	688	141	$141/688 = 20,5 \%$

**Ćwiczenie.** Średnia roczna temperatura w Zakopanem, na wysokości 844 m n.p.m., wynosi  $4,9^{\circ}\text{C}$ , na poziomie Doliny Gąsienicowej (1520 m n.p.m.) już tylko  $2,4^{\circ}$ , a na Kasprowym Wierchu tylko  $-0,8^{\circ}\text{C}$ <sup>5</sup>. Wyjaśnij, jak obliczono te średnie.

<sup>5</sup> Źródło: Ryszard Jakubowski, *Tatry. Przewodnik turystyczny*, wyd. Sport i Turystyka, Muza S.A., Warszawa 2002.

**Ćwiczenie.** W swoim przewodniku po Tatrach z 1891 roku Walery Eljasz podaje, że średnia wysokość szczytu w głównym paśmie Tatr Wysokich to 2340 m, a w głównym paśmie Tatr Zachodnich 2040 metrów. Następnie podaje, że po pewnych poprawkach wypadną zupełnie inne wartości, a mianowicie 2564 m i 2166 m. Jak myślisz, skąd wzięła się taka różnica? Co to za poprawki? Czy obliczona średnia to to samo, co średnia wysokość grani tatrzańskiej? Jak można wyznaczyć średnią wysokość grani? Zaproponuj metodę oszacowania tej średniej wysokości. Weź do ręki mapę.

Walery Eljasz podaje też następującą tabelkę i komentuje dane w ten sposób: „w Tatrach Zachodnich przełęcze są głębiej zakłęśte niż we Wschodnich. Z powodu jednak większej położystości szczytów, mimo że przełęcze są głębokie, zarysy grzbietu nie występują tak wybitnie, jakby oczekiwać należało. Następujące zestawienie dowodnie objaśnia wymienione uwagi oraz daje jeszcze podstawę do obszerniejszych poglądów na pionową budowę zrębu tatrzańskiego”.

	Średnia wysokość			Średnie zagłębienie grzbietu	Stosunek średniego zagłębienia do średniej wysokości grzbietu
	Szczytów	Przełęczy	Grzbietu		
Tatry Wysokie	2340	2200	2270	140	1:16,2
Tatry Zachodnie	2040	1873	1957	167	2:11,7

Średnia arytmetyczna nie powinna być stosowana dla danych znacznie różniących się między sobą. Jeżeli w firmie 10 osób zarabia po 1500 złotych miesięcznie, a jedna sto tysięcy, to średnia jest absurdalnie wysoka. Dlatego niekiedy lepiej używać mediany. Jak wiemy, jest to środkowa liczba uporządkowanego ciągu liczb – ewentualnie średnia arytmetyczna dwóch liczb środkowych. Taka *miara tendencji centralnej* (mówiąc prościej: *średnia*) jest już niewrażliwa na skrajne zmiany. Pokażę typowe sytuacje, w których średnia arytmetyczna jest niebezpiecznym narzędziem.

**Dziwny klocek.** Mam pięć klocków sześciennych o krawędziach 2, 4, 6, 8 i 10 centymetrów. Przeciętna wymiarów to 6 cm. Objętości klocków są równe odpowiednio 8, 64, 216, 512 i 1000 centymetrów sześciennych. Średnia tych liczb to  $\frac{8 + 64 + 216 + 512 + 1000}{5} = 360$ . A zatem przeciętny klocek w mojej kolekcji ma

krawędź 6 cm, a objętość 360 centymetrów sześciennych!

**Pozornie uczciwa gra.** Na przyjęciu imieninowym zabawmy gości taką grą. Należy oszacować powierzchnię naszego mieszkania. Ale wygrywa nie ten, kto poda liczbę najbliższą rzeczywistej, ale ten, czyj wynik będzie najbliższy średniej wszystkich odpowiedzi. Jeżeli zaoferujemy wysoką nagrodę, dwóch graczy może się umówić: jeden napisze, że powierzchnia mieszkania wynosi 50000 metrów kwadratowych, a

drugi ze 500 metrów. Ten drugi wygra, jeżeli goście nie znali już tego triku. Nagrodą podzielą się po połowie<sup>6</sup>.

**Drobne manipulowanie opinią publiczną.** Wyobraź sobie, że masz ze współnikiem firmę, w której zatrudniasz 10 osób. Firma prosperuje bardzo dobrze. W maju osiągnęliście łączny zysk 120000 złotych. Zapłaciliście pracownikom po 2400 złotych pensji. Dla was zostało zatem po 36000. Ujęliście to w sprawozdaniu tak:

Zysk firmy: 120000 złotych.  
Przeciętna pensja pracownika 2400 złotych.  
Na płace firma przeznaczona 5% zysku.

Ale potem spojrzeliście. To nie wygląda dobrze. Zatem część zysku zaksięgowaliście jako pensje dla siebie – po 16 tysięcy złotych. I w sprawozdaniu napisaliście:

Przeciętna pensja w firmie: 5 tysięcy złotych  
Przeciętny dochód właściciela: 16 tysięcy złotych.  
Przeciętna pensja wynosi ponad 31% dochodu właściciela.

**Jak poprawić wyniki nauczania?** Dyrektor szkoły stwierdza, że w klasach *A* i *B* oceny semestralne różnią się znacznie. Przesuwa kilku uczniów z klasy *A* do *B* i osiąga cel: w obu klasach wzrosła średnia ocen.

Nietrudno zrozumieć, jak to jest możliwe. Jeśli w klasie *A* jest uczeń, który ma średnią ocen niższą od średniej klasy *A*, ale wyższą od średniej klasy *B*, to przesunięcie go do *B* spowoduje ten właśnie skutek. Na tym efekcie oparte jest przekonanie autorów Encyklopedii Galicyjskiej (wyd. Anabasis, Kraków 1998), że w dniu, w którym Zygmunt III Waza wraz z dworem przeniósł się do Warszawy, w obu tych miastach wzrósł średni poziom inteligencji.

**Jak można dać każdemu podwyżkę**, powodując jednocześnie spadek średniego wynagrodzenia? Proste: dać niewielką podwyżkę tym, którzy już pracują, i przyjąć do pracy wiele nisko płaconych osób. Średnia spadnie .... a w warunkach zadania nie było przecież mowy o globalnym funduszu płac. Podobno przed 1989 rokiem pewien dyrektor państwowego zakładu pracy tak się zachował.

**Regularne pływanie.** W czasie pobytu w sanatorium wróciłem do regularnego pływania. Pływam wolno, bo już nie te lata.... Wiem, że przez pierwsze pół godziny mogę przepływać w 3 minuty dwie długości basenu. Przez następne pół godziny płynę już wolniej: na każdą długość basenu zużywam dwie minuty. Ale – pomyślałem sobie – skoro najpierw dwie długości w trzy minuty, a potem jedną w dwie minuty, to średnio daje to trzy długości basenu na 5 minut, a zatem 36 na godzinę.

Zrobiłem tak, jak sobie wyliczyłem. Na ścianie pływalni wisiał duży zegar i mogłem precyzyjnie kontrolować swój czas. Najpierw dwie długości na 3 minuty,

---

<sup>6</sup> Polecam natomiast grę, która początkowo ze statystyką ani z matematyką nie ma nic wspólnego. Prosimy, by każdy z gości napisał na karteczce dowolną liczbą naturalną. Wygrywa ten, czyja liczba jest najmniejsza, pod warunkiem, że nikt inny jej nie napisał. Gra udaje się przy co najmniej kilkunastu osobach. Wciąża. Statystyka wejdzie do akcji, gdy będziemy mieli do dyspozycji wynik np. 100 takich gier.

potem jedna długość na dwie minuty. Okazało się jednak, że przepłynąłem nie 36, a 35 długości basenu.

Po kilku dniach pływałem już nieco szybciej. Najpierw 5 długości basenu w 7 minut, a po pewnym czasie (i zmęczeniu) 3 długości w 5 minut. Pomyślałem sobie, że średnio da to 8 długości na 12 minut, czyli 40 na godzinę. Zatem pierwsze 20 długości popłynę szybciej, a potem zwolnię, żeby średnia była taka, jak zaplanowałem. Potem następne 20 długości w wolniejszym tempie. Kontrolowałem czas ... i znów nie wyszło. Stojąc pod prysznicem po wyjściu z pływalni, zrozumiałem, dlaczego. Po prostu dodawałem ułamki tak, jak zły uczeń: licznik do licznika, mianownik do mianownika.

### Jeszcze jeden przykład manipulacji średnią

Omówię dokładniej paradoks, o jakim wspominałem w wykładzie 10. W Europie zużycie paliwa przez samochód opisujemy liczbą litrów, jakie samochód zużywa na 100 km. W USA jest inna zasada: podajemy, ile mil przejedziemy na jednym galonie. Nie wchodząc w amerykańskie miary, obliczmy „po europejsku” i „po amerykańsku” średnie zużycie paliwa w samochodzie moim i mojej żony. Mój samochód zużywa przeciętnie 8 litrów na 100 kilometrów. Żona ma cięższy samochód, większej mocy i jeździ energiczniej. Przeciętne zużycie paliwa ma 12,5 litra na 100 km. Jeździmy tak samo często. Jakie mamy przeciętne zużycie paliwa?

Oczywiście  $\frac{8+12,5}{2} = 10,25$  litra na 100 kilometrów. Obliczmy to jednak inaczej, sposobem amerykańskim. Ja na jednym litrze przejadę aż 12,5 kilometra, żona przejedzie tylko 8. Przypominam, że jeździmy równie często. A zatem średnio przejeżdżamy na jednym litrze paliwa  $\frac{12,5+8}{2} = 10,25$  kilometra. Zgadza się? Nie!

Dziesięć i ćwierć kilometra na jednym litrze paliwa to zużycie mniejsze niż 10,25 litrów benzyny na 100 kilometrów!

Rozwiązanie paradoksu łatwo zrozumieć. Znajdźmy wzór na konwersję amerykańskiego sposobu obliczania średniej na europejski. Jeśli przejeżdżam  $a$  kilometrów na jednym litrze paliwa, to na sto kilometrów zużyję  $\frac{100}{a}$  litrów.

Wykresem funkcji  $f(a) = \frac{100}{a}$  jest hiperbola. Niech  $p$  i  $q$  będą dwiema liczbami, reprezentującymi zużycie paliwa w litrach na 100 kilometrów. Średnie zużycie obliczone po europejsku to oczywiście  $\frac{p+q}{2}$  litrów na 100 kilometrów. A po

amerykańsku? To średnia liczb  $\frac{100}{a}$  i  $\frac{100}{b}$ , czyli  $\frac{1}{2} \cdot \left( \frac{100}{a} + \frac{100}{b} \right)$ . Gdy chcemy ten wyniki „przerobić na Europę”, musimy do wyniku zastosować funkcję odwrotną do  $f$ , a więc funkcję  $g(y) = \frac{1}{100y}$ . Ale wartością tej funkcji dla

$\frac{1}{2} \cdot \left( \frac{100}{a} + \frac{100}{b} \right)$  nie jest średnia arytmetyczna liczb  $a, b$ .

**Jeszcze raz o paradoksie Simsona.** O miejsce w reprezentacji Polski w piłce nożnej kandydowali dwaj obywatele naszego kraju: Mugabe Burunda z klubu Liverpool i Nugat Pereira da Silva Corto y Derecho (grający na co dzień w Pogoni Szczecin). Trener Michał Kojonkowski poddał ich testowi. W poniedziałek i wtorek strzelali do pustej bramki. Mugabe strzelał 50 razy, z czego 15 strzałów było celnych. Pereira strzelał 25 razy i trafił 7 razy. We wtorek Mugabe strzelał 25 razy. Dziesięć strzałów ugrzęzło w siatce. Dla równowagi Pereira więcej: strzelał 50 razy i trafił 19 razy. W środę zebrał się zarząd. „No cóż, sprawa jest jasna. Mugabe miał w poniedziałek skuteczność 30%, a Pereira 28 %. We wtorek sprawa się powtórzyła: Mugabe 40%, Pereira 38%. Każdego dnia Mugabe był lepszy. Powołujemy Mugabe!” . Sprawa by się zakończyła, ale wstał znany były piłkarz Zdzisław Coniek. „Zaraz, zaraz! Obydwaj strzelali 75 razy. Mugabe trafił  $15+10 = 25$  razy, a Pereira  $7 + 19 = 26$  razy. Nie jest to duża różnica, ale jednak Pereira jest górą. Powołujemy Pereirę!”

Taka manipulacja procentami zdarza się bardzo często. Proszę przeczytać jeszcze raz ten przykład. Zrozumienie uodporni Państwa na podobne manipulacje, na które dają się nabrać nawet profesorowie matematyki (może nie profesorowie statystyki!). Ze swej strony nie sądzę, aby przyczyną niepowodzeń naszych piłkarzy na mistrzostwach świata 2006 roku była nieznamość paradoksu Simsona w zarządzie Polskiego Związku Piłki Nożnej.

**Przykład autentyczny!** Oto inny, konkretny przykład manipulacji danymi statystycznymi, aby uzyskać z góry założoną tezę. Omawiają to podręczniki statystyki, ale nie przypuszczałem, żeby ktoś naprawdę dał się na to nabrać. Tymczasem....

W artykule „Kto się boi czarnej wołgi?” , poświęconym bezpieczeństwu na drogach, w dodatku stołecznym do Gazety Wyborczej (numer sylwestrowo/noworoczny 2005/2006) czytamy:

*Wydawałoby się, że najniebezpieczniej jest po zmroku. Albo kiedy siecze deszcz. Nic bardziej mylnego. Otóż najniebezpieczniej jest w dzień i przy dobrej pogodzie – dwie trzecie wypadków i dwie trzecie ofiar. Wydawałoby się również, że najbardziej niebezpieczne są ostre zakręty. Błąd! To na długich prostych dochodzi do 60 procent wypadków, to tam ginie blisko 70 procent ofiar.*

Pójdźmy za logiką autora tego tekstu. Autostrady powinny być kręte, a gdy chcemy zmniejszyć szanse wypadku, powinniśmy się wybierać w podróż we mgle i marznącej mżawce! Czyż nie tak?

Gdzie jest błąd? W tym samym miejscu, co przy argumentacji: kobiety są lepszymi kierowcami niż mężczyźni, bo – jak pokazują statystyki – powodują znacznie mniej wypadków. Mniej więcej taki sam błąd popełniamy argumentując: jeździmy coraz gorzej, bo liczba samochodów wzrosła dwukrotnie, a liczba wypadków trzykrotnie.

Osobiście wierzę, że kobiety jeżdżą lepiej niż mężczyźni. Ale nie dlatego, że mniej wypadków jest powodowanych przez kobiety. Panie powodują łącznie mniej wypadków po prostu dlatego, że na drogach jest mniej kierowców kobiet niż mężczyzn! Natomiast kobiety bardziej biorą do siebie zasadę *defensive driving*. Tego uczą się kierowcy na kursach w USA. Bądź *defensywny*. Ustępuj, uważaj. Wczuj się w sytuację tego drugiego. Dlaczego o tym piszę w książce o nauczaniu matematyki? *Matematyk robi to lepiej*, bo matematyk myśli. Jest dla mnie oczywiste, że lepsza znajomość trygonometrii w społeczeństwie byłaby jedną z przyczyn zmniejszenia liczby wypadków komunikacyjnych, a także przerażającej

liczby zabitych młodych ludzi, którzy kibicowali *innemu* klubowi piłkarskiemu. Proszę pomyśleć, czy nie mam racji...

Jeżeli liczba samochodów wzrasta dwukrotnie, to – zakładając losowość zdarzenia – teoretyczne prawdopodobieństwo kolizji blisko czterokrotnie. Jeżeli zatem przy dwukrotnym zwiększeniu liczby samochodów zdarza się trzy razy więcej wypadków, to znaczy że jeździmy lepiej, a nie gorzej! To nie jest do końca tak – nie wiemy, na ile uproszczony model prawdopodobieństwa kolizji sprawdza się w życiu.

A zatem: wypadków w nocy i w trudnych warunkach zdarza się mniej dlatego, że wtedy jest w ogóle mniejszy ruch. Zakręty nie są bardziej bezpieczne niż prosta droga – tylko przypada na nie mniejsza część długości dróg. Poza tym na zakrętach jednak bardziej uważamy.

A oto inny przykład nonsensownych „badań” statystycznych. 17 stycznia 2006 roku w portalu Interii można było znaleźć artykuł, w którym autorka napisała, że z badań sondażowych wynika, iż obowiązkowej matury z matematyki najbardziej obawiają się ludzie młodzi.

Wzruszające, głębokie, niespodziewane odkrycie, prawda?

**Ćwiczenie.** Poniższa tabela pokazuje liczbę ludności w województwach w Polsce w roku 2003. Województwa oznaczone są według pierwszych liter na samochodowych tablicach rejestracyjnych. Sprawdź, że średnia arytmetyczna była większa od mediany i że 10 województw (na 16) miało liczbę ludności poniżej średniej (arytmetycznej). Zatem jedno lub kilka województw musiało mieć znacznie więcej ludności niż pozostałe. Które to województwa?

B	1,224
C	2,096
D	2,987
E	2,675
F	1,018
G	2,173
K	3,206
L	2,241
N	1,457
O	1,092
P	3,34
R	2,104
S	4,876
T	1,33
W	5,068
Z	1,729

Trudnym problemem w nauczaniu statystyki w szkole jest ubóstwo tematów, które można poddać analizie. Z reguły bywa tak, że zagadnienie, które można w pełni zanalizować na lekcji, jest sztuczne, a zadania pochodzące z życia są zbyt skomplikowane obliczeniowo i z reguły nieciekawe dla ucznia. Jednym z wyjątków jest następujące ćwiczenie, chętnie robione przez uczniów. Możemy je zapowiedzieć

tak: „zajmiemy się różnicą między chłopcami i dziewczynkami”. Wiele gazet lokalnych publikuje dane, dotyczące dzieci urodzonych w ostatnim tygodniu w danej miejscowości. W wersji warszawskiej ćwiczenie wygląda następująco:

**Ćwiczenie.** Kup wydanie Twojej lokalnej gazety, gdzie znajdziesz fotografie ludzi<sup>7</sup>, którzy przyszedli na świat w ostatnim tygodniu. Policz, ilu jest chłopców, ile dziewczynek. Wypisz wagę i wzrost wszystkich chłopców, wszystkich dziewczynek. Oblicz średnią wagę chłopców, średnią wagę dziewczynek, średni wzrost chłopców, średni wzrost dziewczynek, odchylenie standardowe wagi i wzrostu dla obu płci. Zinterpretuj otrzymane wyniki.

Zadanie ma sens tylko wtedy, gdy użyjemy kalkulatora, a jeszcze lepiej arkusza kalkulacyjnego. Dlatego lepszym miejscem na to zadanie są lekcje informatyki. Oto, jakie wyniki mogą wyjść. Gazeta Wyborcza z 31 grudnia 2005 publikuje zdjęcia i dane (imię, nazwisko, wzrost i wagę) 267 noworodków warszawskich: 153 chłopców i 114 dziewczynek. Można z nich wyliczyć, że średnią wagą chłopca było 3473 g, a odchylenie standardowe wyniosło 535,35 g. Średni wzrost chłopca wyniósł 54,38 cm z odchyleniem standardowym 3,14. Dla dziewczynek dane były następujące: średnia waga 3414 g., z odchyleniem standardowym 459,22 g, Wzrost 53,9 cm z odchyleniem 2,62.

Są to za małe dane statystyczne, by wnioskować coś z nich odpowiedzialnie na szerszą skalę. „Coś” jednak widać. Potwierdza się znany fakt, że chłopców rodzi się trochę więcej niż dziewczynek, że są ciężsi i dłużsi. Jak interpretować fakt, że odchylenie standardowe wagi i wzrostu jest większe u chłopców? To proste. Bez znajomości statystyki wiemy, że wśród chłopców (i dorosłych mężczyzn) występują większe różnice osobnicze, większa rozpiętość wzrostu, wagi, a także większe różnice intelektualne. Wśród mężczyzn jest więcej geniuszy ... i więcej debili, Nie sposób wyjaśnić to tylko przyczynami kulturowymi<sup>8</sup>.

W każdym razie, to drobne zagadnienie wzrostu i wagi noworodków dobrze nadaje się na dyskusję wykraczającą poza matematykę. W czasach szkolnych autora tej książki *takich* tematów nie poruszano w szkole. Teraz czasy są inne.

### **Przykład obliczeń statystycznych bez znaczenia.**

Nie tylko obliczenie średniej arytmetycznej wszystkich zwierząt w Zoo nie ma sensu. Niektóre poważnie wyglądające obliczenia też. W 2005 roku maturę z matematyki zdawało 88044 uczniów. Ładna liczba, prawda? Podzielmy liczbę zdających maturę z matematyki w poszczególnych województwach przez liczbę ludności (w tysiącach) w tym województwie. Otrzymamy dziwny wynik: na wschód od Wisły (plus całe województwo pomorskie) ten wskaźnik jest większy niż na zachód od niej. Ciekawe? Tak, tylko ... zupełnie nie wiadomo, co ten wskaźnik mierzy i jakie ma znaczenie. Można (po przeprowadzeniu stosowanych obliczeń) powiedzieć tak: prawdopodobieństwo, że losowo napotkana osoba jest uczniem zdającym matematykę, jest największe w województwie pomorskim, a najmniejsze w kujawsko-pomorskim. Tylko, że ... to nie ma sensu, choć można się dziewic, dlaczego sąsiadujące województwa leżą na dwóch końcach tej skali.

<sup>7</sup> Dziecko też człowiek.

<sup>8</sup> Jest taki pogląd, że natura eksperymentuje na samcach. Jeśli to prawda, to jednak jest to właściwy wybór Natury.

**Miary rozproszenia**

Odchylenie standardowe obciążone (zwane też odchyleniem z populacji) to

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2}, \text{ natomiast odchylenie standardowe nieobciążone wyraża się}$$

$$\text{wzorem } s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n-1} (x_i - \bar{x})^2} \text{ i zwane jest też odchyleniem z próby. W szkole}$$

mówimy tylko o odchyleniu z populacji. Różnica między próbką a populacją jest zrozumiała. Jeżeli chcę wyznaczyć średni wzrost uczniów mojej szkoły, to mogę zmierzyć ich wszystkich i w zwykły sposób obliczyć średnią arytmetyczną stosownych liczb. Wtedy wszyscy uczniowie szkoły stanowić będą populację. Jeżeli jednak na podstawie znajomości wzrostu uczniów mojej szkoły władze miasta czy województwa albo i kraju będą chciały wiedzieć, jaki jest średni wzrost wszystkich uczniów, to ci moi uczniowie będą stanowić próbkę.

Miary rozproszenia wprowadzamy choćby po to, by nie zdarzały się takie sytuacje jak opisana poniżej. Wyobraźmy sobie oto, że kierownik ośrodka wczasowego telefonuje do biura turystycznego:

– Jesteśmy przygotowani na przyjęcie grupy, którą nam przysyłacie. Jaka jest średnia wieku uczestników? Chcemy jak największej liczbie gości zapewnić rozrywki stosowne do ich wieku.

– Zaraz obliczę ... mówi sekretarka – Uczestników jest 40, a średnia to suma wieku podzielona przez liczbę uczestników. Już sumuję i dzielę ... 35 lat. Bardzo dziękujemy za troskę o naszych klientów.

W dwa dni później z autokaru w ośrodku wysiadają ... sami dziadkowie z wnukami. Dziadek ma 60 lat, wnuczek 10. Istotnie,  $60 + 10$  dzielone przez 2 to rzeczywiście 35, ale starannie przygotowanym programem zajęć dla 35-latków nikt nie jest jakoś zainteresowany. Średnia wieku obliczona była poprawnie, jednak nie dawała żadnej sensownej informacji.

Jak wiemy, w statystyce ważne są jeszcze dwie inne średnie, noszące nazwy *mediana* i *moda*. Jeżeli zarobki trzynastu osób w dziale sprzedaży przedsiębiorstwa wynoszą kolejno 0,5, 0,5, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 3 tysiące złotych, to środkową z napisanych tu liczb jest 1. To jest właśnie mediana – liczba środkowa ciągu uporządkowanego w kolejności rosnącej (lub, co na jedno wychodzi, malejącej). Gdy liczba wyrazów ciągu jest parzysta, medianą jest średnia arytmetyczna dwóch liczb „środkowych”. W powyższym przykładzie liczba 1 jest też *modą*. Tak nazywa się najczęściej występująca liczba w danym ciągu. Właśnie mediana i moda w połączeniu ze średnią arytmetyczną lepiej opisują drabinę zarobków pracowników niż sama średnia arytmetyczna. Gdy zarząd spółki zacznie pobierać trzykrotnie większe pensje niż dotychczas, wzrośnie średnia arytmetyczna wynagrodzeń, ale nie drgnie moda, ani mediana. Gdyby kierownik tego ośrodka wczasowego, o którym pisaliśmy na początku (do którego przyjechali dziadkowie z wnukami), poprosił o trzy liczby: średnią, modę i medianę, to liczby

średnia 35, moda 11, mediana 11

dałyby mu do myślenia i może nawet zgadłyby, że prawdopodobnym rozkładem będzie coś w rodzaju: ośmiu panów 65 letnich, sześciu 60 letnich, sześciu dziesięciolatków i dziesięciu jedenastolatków. Jednak najlepszy pogląd na sprawę wyrobiłby sobie, gdyby poprosił jeszcze o podanie *wariancji* wieku uczestników. Właśnie ta wielkość mierzy rozproszenie ciągu liczbowego  $a_1, a_2, \dots, a_n$ . Wariancja oznaczana jest symbolem  $\sigma^2$  i określana wzorem

$$\sigma^2 = \frac{(a_1 - E)^2 + (a_2 - E)^2 + \dots + (a_n - E)^2}{n},$$

gdzie  $E$  jest średnią arytmetyczną. Wariancja jest zatem średnią arytmetyczną kwadratów odchyłeń wyrazów ciągu od wartości przeciętnej. W naszym przykładzie z dziadkami i wnukami mamy

$$\sigma^2 = \frac{8 \cdot (65 - 35)^2 + 6 \cdot (60 - 35)^2 + 6 \cdot (10 - 35)^2 + 10 \cdot (11 - 35)^2}{30} = \frac{20460}{30} = 682,$$

a gdyby wycieczka miała się składać z w połowie z 30-latków a w połowie z 40-latków, to wariancja wyniosłaby „zaledwie”

$$\sigma^2 = \frac{15 \cdot (30 - 35)^2 + 15 \cdot (40 - 35)^2}{30} = \frac{750}{30} = 25$$

i wtedy założenie, że grupa składa się z osób w podobnym wieku, miałoby więcej sensu.

W opisywanym przykładzie najprościej byłoby zapytać wprost, kto przyjeżdża. Jeżeli jednak nie ma możliwości pełnego opisu jakiejś sytuacji, to trzeba zdać się na opis liczbowy i uśrednianie. Widzieliśmy, że można to zrobić lepiej lub gorzej.

Pierwiastek kwadratowy wariancji nazywa się odchyleniem standardowym. Oczywiście oznaczamy go literą  $\sigma$ . Jest to najczęściej używana miara rozproszenia. Im większe odchylenie standardowe, tym mniej dany ciąg jest skupiony wokół swojej średniej. Można sobie wyobrazić, że jeżeli zakład komunikacji chce premiować kierowców za regularność jazdy, powinien przyznać premie tym, dla których odchylenie standardowe od rozkładów jazdy jest jak najmniejsze.

Odchylenie standardowe jest liczbą mianowaną, wyrażoną w tych samych jednostkach, co dane wyjściowe. Nie jest jednak „wyskalowane”: nie możemy z góry narzucić, jakie odchylenie będziemy uważać za duże, a jakie za małe. Parametr ten służy do porównywania dwóch ciągów, dwóch szeregów liczbowych: ten bardziej rozstrzelony ma większe *sigma*.

Przy obliczaniu odchylenia standardowego bez kalkulatora albo za pomocą prostego kalkulatora (takiego jak w telefonie komórkowym) warto posłużyć się pewnym wzorem, który upraszcza rachunki. Niech  $x_1, x_2, x_3, \dots, x_n$  będzie szeregiem statystycznym (ciągiem liczbowym),  $\bar{x}$  jego średnią arytmetyczną. Wtedy

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \cdot \left( \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \sum_{i=1}^n \bar{x}^2 \right) = \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - 2\bar{x} \frac{\sum_{i=1}^n x_i}{n} + \bar{x}^2 = \\ &= \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n} \cdot \sum_{i=1}^n x_i^2 - \bar{x}^2. \end{aligned}$$

W oznaczeniach używanych przez statystyków wzór ten ma prostszą postać:

$$\sigma^2 = \frac{n \sum x^2 - (\sum x)^2}{n^2}.$$

Wyjaśnijmy, jak się nim posługiwać przy „ręcznych” obliczeniach. Układamy tabelkę, w której wpisujemy kolejno dane, ich kwadraty, sumę danych i sumę ich kwadratów.

Numer kolejny	$x$	$x^2$
1	18	324
2	13	169
3	15	225
4	15	225
5	13	169
6	14	196
7	11	121
8	12	144
9	14	196
10	15	225
<b>Suma</b>	<b>140</b>	<b>1994</b>

Obliczamy więc sumę liczb w kolumnie i sumę ich kwadratów. Z przytoczonego wzoru obliczamy następnie, że

$$\sigma^2 = \frac{10 \cdot 1994 - 140^2}{100} = \frac{19940 - 19600}{100} = \frac{340}{100} = 3,4$$

Odchylenie standardowe jest zatem równe  $\sqrt{3,4} \approx 1,844$ .

Przykład. Obliczymy wariancję i odchylenie standardowe wieku uczestników czterdziestoosobowej wycieczki, złożonej z 20 panów w wieku  $m = 36$  lat i 20 pań w wieku  $k = 34$  lat. Średnią wieku jest, jak i przedtem, 35 lat.

	$x$	$x^2$
wiek m	36	1296
wiek k	34	1156
Suma	70	2452
... razy 20	1400	49040

Zatem  $\sigma^2 = \frac{40 \cdot 49040 - 1400^2}{1600} = \frac{1600}{1600} = 1$ . Odchylenie standardowe jest oczywiście też równe 1. Z tych kilku przykładów widoczne jest, że odchylenie standardowe dobrze mierzy to, co ma mierzyć – rozproszenie danych.

### Jak nauczać statystyki?

Sprawa jest delikatna. Mamy tu do czynienia z dużą ilością obliczeń. Nie można już udawać, że kalkulatory i komputery nie istnieją. To, co 30 lat temu zajmowało kilkanaście czy kilkadziesiąt minut, albo wręcz i dni pracy, teraz jest do wykonania w sekundy. Kalkulator, w którym są podstawowe funkcje statystyczne można kupić za 30 złotych, a przy pewnym

szczęściu nawet za 10 (wiem z własnego doświadczenia)<sup>9</sup>. W takich kalkulatorach mamy na ogół, w pakiecie statystycznym, możliwość szybkiego obliczania sumy danych, średniej arytmetycznej i sumy kwadratów. Jak widzieliśmy, wystarcza to do znalezienia podstawowej miary rozproszenia: odchylenia standardowego.

W prawie każdej szkole jest teraz pracownia komputerowa, a w skład oprogramowania wchodzi arkusz kalkulacyjny (Excel bądź jedna z jego bezpłatnych, okrojonych wersji). Widoczne powyżej tabelki układałem za pomocą Excela i nie będę udawał przed Czytelnikiem, że robiłem to tak, jak bym to robił czterdzieści lat temu. Uważam, że takie posługiwanie się arkuszem kalkulacyjnym jest właściwe: uczy i techniki komputerowej i matematyki.

Wreszcie można skorzystać z bardziej zaawansowanych metod – choć nawet trudno to określić tym mianem. Bardziej wyrafinowane kalkulatory mają gotowe funkcje statystyczne. Ma je każdy arkusz kalkulacyjny. Nie chcę tu robić kursu Excela – ale wszystko jest w nim bardzo proste. Jak się nauczyć? W okienku „pomoc” wpisać na przykład „odchylenie” i przeczytać, jak używa się odpowiedniej procedury.

Jest tu miejsce na dygresję, sięgającą do wykładu 6 a dotyczącą zasady dydaktycznej o kierowniczej roli nauczyciela w procesie nauczania. Pokolenia nauczycieli górowały nad uczniami wiedzą merytoryczną. To dodawało nauczycielom powagi. Dziś, zwłaszcza tam, gdzie używamy komputera, nauczyciel często ma mniejsze umiejętności od ucznia. W latach osiemdziesiątych i dziewięćdziesiątych zeszłego stulecia nauczyciele bali się komputerów jak diabeł święconej wody, a lęk wypływał z obawy przed – jak im się wydawało – kompromitacją przed uczniami. Teraz jej nieco inaczej; większość z nas posługujemy się komputerami w miarę sprawnie, ale ważniejsze jest, że uczymy się, jak przyznać się do własnej niewiedzy i nie stracić autorytetu u uczniów. To jest osobny, bardzo trudny temat, nie do dyskusji w tym momencie. Chociaż ogólna zasada jest prosta. Od 50-letniego nauczyciela w nikt nie będzie oczekiwał, że nadąży za swoimi uczniami w sprincie. Zamiast denerwować się na uczniów, że sprawniej posługują się komputerem, zachwyćmy się ich młodością ale dyskretnie przekonajmy nich, że to nie oni są nadzwyczajni, tylko właśnie my, nauczyciele.

Największy kłopot jest oczywiście z ocenianiem rozwiązań zadań statystycznych, w szczególności z kryteriami maturalnymi. Należy oczywiście przewidywać najgorszą możliwość: dozwolone będą tylko proste kalkulatory. W takiej jednak sytuacji zadania statystyczne będą musiały być bardzo powierzchowne. I to jest dodatkowa trudność w nauczaniu statystyki (opisowej) w szkole. Zakres i układ materiału rozkład godzin pozwalają uczyć wyłącznie bardzo powierzchownie. Zadania maturalne o treści „oblicz odchylenie standardowe” źle świadczą o układających. Zadania takie sprawdzają bowiem tylko, czy uczeń umie podstawić dane do wzoru i ewentualnie nacisnąć kilka przycisków kalkulatora.

### Ciekawostka statystyczna sprzed 100 lat<sup>10</sup>

Bolesław Prus jest autorem, jak sam przyznaje, quasimatematycznej formuły charakteryzującej popularność pisma. Otóż popularność pisma jest według Prusa równa

$$\frac{L_p + L_o}{P_r (C_p + C_o)},$$

przy czym:

$L_p$  to średnia liczba np. dziennie rozchodzących się numerów danego pisma,

$L_o$  to średnia dzienna ogłoszeń,

$P_r$  to liczba egzemplarzy wszystkich pism rozchodzących się dziennie w danym kraju,

<sup>9</sup> Studentom mówię, że za trzy piwa w pubie.

<sup>10</sup> Wg: Julian Tuwim, *Cicer Cum Caule*, tom 2, Czytelnik, Warszawa 1959.

$C_p$  to cena prenumeraty,  
 $C_o$  to cena przeciętna ogłoszeń (np. za  $\text{cm}^2$ ).

Prus pisze, że za pomocą tej formuły można porównywać popularność pism tego samego kraju albo i różnych krajów. Nadto można ocenić zmiany, jakie zachodzą w popularności pisma skutkiem podniesienia lub obniżenia ceny prenumeraty.

Możemy ten – nie całkiem może przystający do współczesnej rzeczywistości – wzór wykorzystać do sprawdzenia, czy nasi uczniowie rozumieją wzory matematyczne. Oto przykładowe pytania:

1. Zwiększamy cenę prenumeraty. Czy popularność pisma spadnie, czy zmaleje?
2. Zwiększamy cenę ogłoszeń. Czy popularność pisma spadnie, czy zmaleje?
3. Załóżmy, że dla pewnego pisma mamy takie dane:  $L_p = 100000$ ,  $L_o = 1000$ ,  $C_p = 1$  (np. 1 euro),  $C_o = 1$  (np. 1 euro). Podnosimy cenę prenumeraty do 2 euro. Ile ogłoszeń powinniśmy zamieszczać, żeby według formuły Prusa utrzymać popularność pisma na tym samym poziomie?
4. Udało nam się zdobyć rynek i rozprowadzamy dwa razy tyle egzemplarzy, co przedtem. O ile możemy zmienić cenę prenumeraty, by utrzymać popularność pisma na tym samym poziomie?
5. Liczba ogłoszeń spadła o połowę. Jak skalkulować cenę ogłoszeń, żeby utrzymać popularność pisma na tym samym poziomie?

Ułóż inne pytania związane ze wzorem Bolesława Prusa. Ułóż podobne pytania do formuły Blacka-Scholesa, którą omówiłem na końcu wykładu 13..

## Oszustwa w majestacie matematyki

Najczęściej podawanym paradoksem jest przykład, że jeżeli zakład pracy najpierw obniży mi pensję o 20 procent, a potem – w wyniku protestów moich związków zawodowych – podniesie o 20 %, to nie będę zarabiał tyle, ile przedtem. Czytelnikom tej książki nie trzeba wyjaśniać, dlaczego. Zauważmy, też, że nie jest to żaden paradoks, tylko po prostu nieznaną procentów. Dozorczyni domu, w którym mieszkam, też nie mogła zrozumieć, dlaczego dostała mniejszą premię niż jej koleżanka, chociaż administracja obiecała wszystkim po równo: po 20 procent.

Nieco podobny błąd popełnił Kazimierz Marcinkiewicz, premier koalicji rządzącej w 2005 i 2006 roku w naszym kraju. Otóż tuż przed świętami Bożego Narodzenia 2005 przeżywaliśmy negocjacje na temat budżetu Unii Europejskiej. Tuż przed Wigilią premier ogłosił, że osiągnęliśmy sukces. W wystąpieniu w Sejmie 28 grudnia premier powiedział mniej więcej tak (cytuję dość dokładnie):

*Gdybyśmy nie zawarli w Brukseli porozumienia 17 grudnia w sprawie budżetu Unii, nasze wydatki na budowę autostrad musiałyby być o 40 procent wyższe. Można powiedzieć, że zaoszczędziliśmy 40 procent.*

Zanalizujmy ten fragment wypowiedzi. Rzeczywiście, jeżeli by premier powiedział tak:

*Gdybyśmy nie zawarli w Brukseli porozumienia 17 grudnia w sprawie budżetu Unii, nasze wydatki na budowę autostrad musiałyby być o 40 miliardów złotych wyższe.*

to mógłby dokończyć:

*..... a więc zaoszczędziliśmy 40 miliardów.*

Z procentami jest inaczej – i nigdy nie dość tłumaczenia! Jeśli już-już chcę kupić komuś prezent za 140 złotych, ale zobaczę, że w sąsiednim sklepie ten sam towar kosztuje 100 złotych, to zaoszczędzę 40 złotych. Gdybym nie zobaczył tego drugiego sklepu, to istotnie wydałbym o 40 złotych (i o 40 procent) więcej. Ale zaoszczędziłem tylko  $\frac{40}{140} = 0,29$ , niecałe 30 procent<sup>11</sup>.

Najlepiej zrozumieć to wszystko, zamieniając „czterdzieści” na „sto”. Wyobraźmy sobie, że zobaczyłem towar dwukrotnie tańszy. Gdybym go nie zobaczył, to rzeczywiście wydałbym dwa razy więcej. Ale czy to znaczy, że zaoszczędziłem 100 procent???? Zaoszczędzić 100 procent – to dostać towar za darmo!!

Na tym tle korzystnie wypadła .... pewna reklama pewnego banku, który obiecywał obniżkę o ponad 100% (p. wykład 7, str. 68).

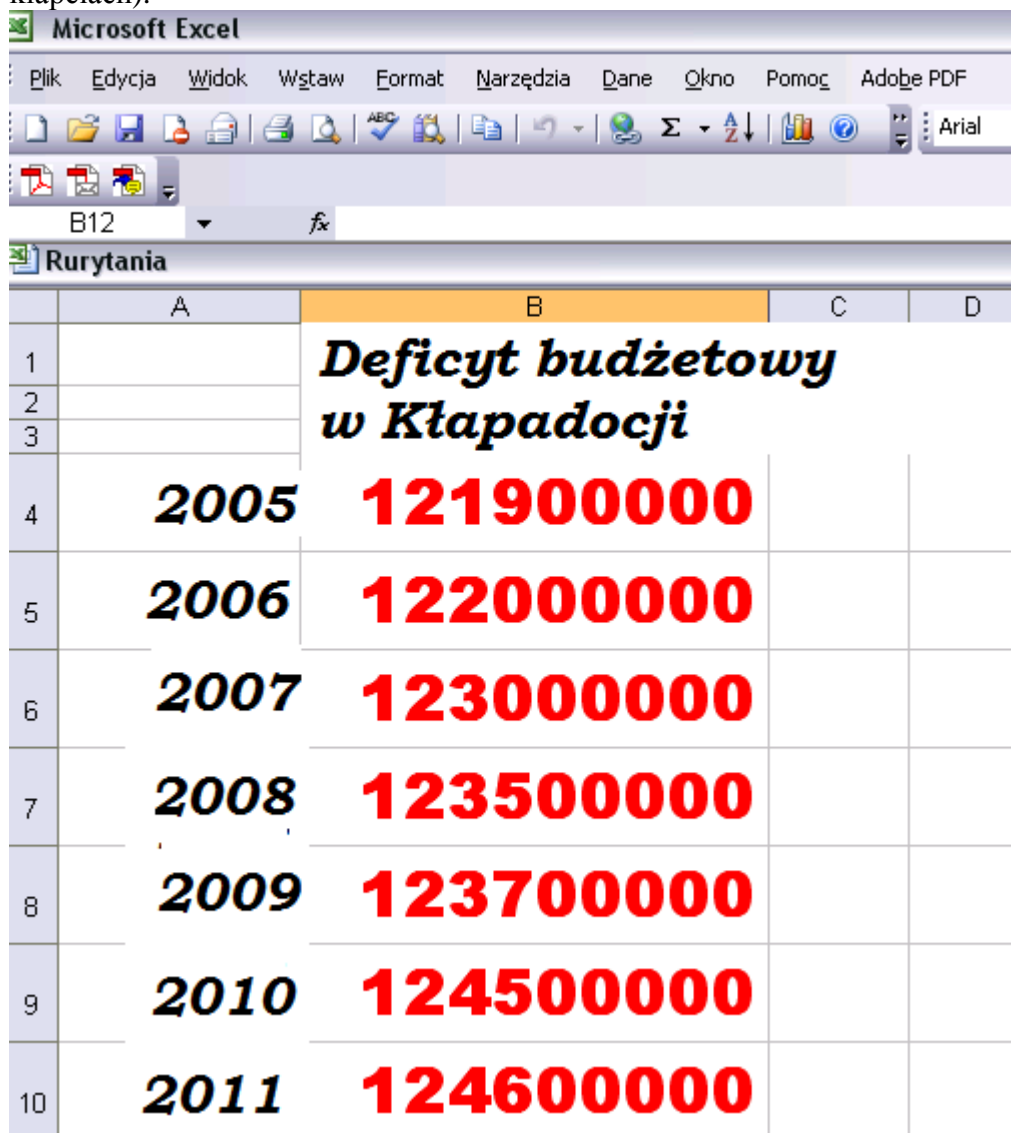
Wracajmy do matematyki i polityki. Wątpię, czy premier chciał świadomie zrobić lepsze wrażenie, mówiąc o 40 procentowych oszczędnościach, gdzie naprawdę są znacznie mniejsze. Najprawdopodobniej nie zwrócił uwagi na różnicę między zaoszczędzeniem 40 milionów złotych a 40 procent. Tu czterdzieści i tam czterdzieści.

Natomiast często używanym chwytem propagandowym jest tendencyjne sporządzanie wykresów. „Pomagają” nam w tym typowe programy obliczeniowe, z popularnym Excelem na czele. Oto przykład, sztuczny, ale bliski prawdziwemu. W pewnym państwie, nazwijmy je Kłapadocją, waluta nazywa się kłapeć. Oto deficyt budżetowy państwa (w

---

<sup>11</sup> Oto stary, jeszcze przedwojenny, dowcip. „Tato, zaoszczędziłem dziś 20 groszy!” „To bardzo dobrze, synku! A jak?” „Nie pojechałem do szkoły tramwajem, tylko biegłem za nim!” „Oj, synku, na drugi raz biegnij za taksówką – zaoszczędzisz 5 złotych!”

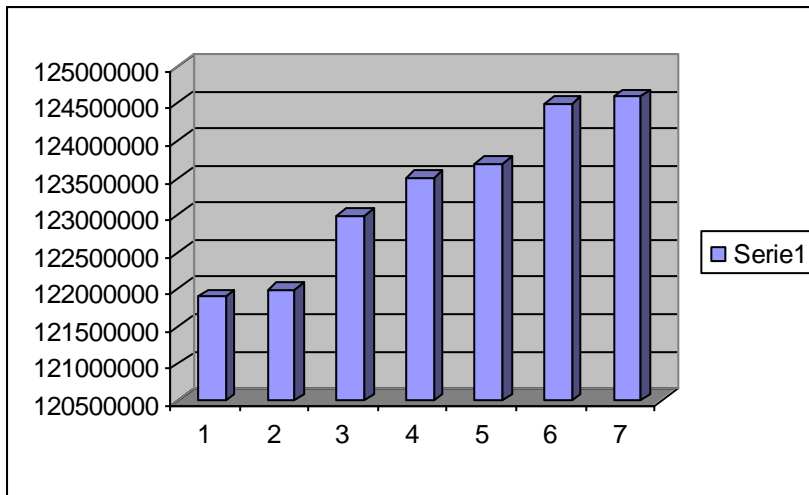
kłapciach).



	A	B	C	D
1		<b><i>Deficyt budżetowy w Kłapadocji</i></b>		
2				
3				
4	<b>2005</b>	<b>121900000</b>		
5	<b>2006</b>	<b>122000000</b>		
6	<b>2007</b>	<b>123000000</b>		
7	<b>2008</b>	<b>123500000</b>		
8	<b>2009</b>	<b>123700000</b>		
9	<b>2010</b>	<b>124500000</b>		
10	<b>2011</b>	<b>124600000</b>		

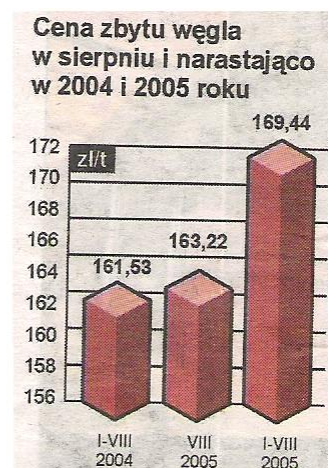
Jak to skomentuje premier? „Deficyt w zasadzie nie zwiększa się, w 2011 roku wyniósł 124 milionów kłapciów, a w 2005 prawie 122. To wzrost tylko o około 2 procent”, (w rzeczywistości 2,21 % , bo trzeba obliczyć iloraz  $\frac{124600000 - 121900000}{121900000}$  ).

Na to opozycja zrywa się z ław: „Premier mówi, że deficyt się zwiększa tylko nieznacznie. Kłamie, jak zwykle. Prawdę mówimy tylko my, nasza partia!!” Proszę, oto wykres” .... i przewodniczący partii rzuca na ścianę wykres z Excela. Pokazuje, że ostatni słupek jest ponad dwa razy wyższy niż pierwszy. Obraz zawsze robi większe wrażenie niż liczby. Posłowie są porażeni i obrazem, i matematyką. Premier zostaje napiętnowany jako oszust.



Czytelniku: zrób ten sam wykres. Weź kolumnę danych i w Excelu wybierz opcję „wykres kolumnowy”. Zobacz, że „Excel nie kłamie!”. A oto dowód, że takie wykresy można zobaczyć wszędzie. Na ogół autorzy nie wiedzą, że mogą wprowadzić nas w

## Jak oszukiwać za pomocą wykresów?



błąd.

Można oszukiwać i mniej subtelnie, wykorzystując analfabetyzm matematyczny wielu kręgów społeczeństwa. Oto wymaginowany tekst (ale oparty na prawdziwej publikacji)

Pielęgniarkom będzie lepiej. Dwa lata temu przeciętne wynagrodzenie pielęgniarki w powiecie sochaczewskim wynosiło netto 500 zł. W roku ubiegłym rząd zwiększył nakłady na służbę zdrowia o 3,1 mld zł. Jest to dwa razy więcej niż w latach ubiegłych. Hermenegilda Kociubińska, pielęgniarka z Centralnego Szpitala Klinicznego mówi: moja pensja w zeszłym miesiącu

wyniosła aż 2000 zł. Oznacza to olbrzymi, czterokrotny wzrost zarobków w służbie zdrowia.

Czy naprawdę nikt nie da się oszukać? Jeśli nawet liczby się zgadzają, to widać, że porównujemy tu średnie wynagrodzenie w prowincjonalnym szpitalu z wynagrodzeniem jednej osoby w wybranym miesiącu. Może Hermenegilda jest szefową pielęgniarek, może w tym miesiącu miała dużo dodatkowych dyżurów, a w dodatku Centralny Szpital Kliniczny ma specjalną siatkę płac? Poza tym wymienione 500 złotych jest płacą netto, a nie jest podane, czy płaca pani Kociubińskiej jest netto, czy brutto. Nie jest podane, na co poszło 3,1 mld złotych (może na biurka dla dyrektorów?). Nie wiadomo, od czego owe 3,1 miliarda złotych jest większe dwa razy.

### Nie chcesz się przyznać? Załatwimy Cię statystyką!

Jest i druga strona statystyki. Może ona pomóc wykryć prawdę.

Ot pierwszy przykład. Ludzie niechętnie przyznają się do popełnienia czynów, które w duszy uważają za naganne. Jeżeli robimy ankietę, nawet anonimową, na temat np. „czy uderzył Pan kiedyś swoją żonę?”, to otrzymamy wyniki zaniżone. Nie uznamy też za wiarygodne wyników takiej ankiety przeprowadzonej przez Internet. Czy można się wobec tego dowiedzieć statystycznej prawdy na superdrażliwe tematy? Metody idealnej nie ma, ale pewien prosty sposób zwiększenia zaufania jest. Nie pytamy ankietowanego wprost, czy bił żonę, ale wręczamy mu kostkę do gry i wydajemy instrukcję: proszę rzucić kostką. Jeżeli wypadnie coś innego niż szóstka, to ma Pan odpowiedzieć „tak”. Jeżeli wypadnie szóstka, to proszę odpowiedzieć szczerze.

Co jest takiego rewelacyjnego w tej metodzie? Jest ona podobna do najlepszego sposobu schowania liścia: należy go ukryć w lesie. Odbierający ankietę nie wie, ile oczek wypadło na kostce. Dlatego pytany ma pewność, że jego odpowiedź pozytywna nie będzie świadczyła o skłonnościach do rękoczynów, a „raczej”, że po prostu nie wyrzucił szóstki. Czuje się bezpieczny. I rzeczywiście, rzecz jest nie do odkrycia – tylko statystyka może sobie z tym poradzić. Gdyby nikt z ankietowanych nie uderzył żony, to w dużej próbie odpowiedzi „tak” będzie 5/6, czyli 83,33%. Gdyby zaś co drugi mąż bił żonę, to statystyka ta podskoczy do .... Obliczmy to. Pięć szóstych odpowiedzi „tak” pochodzą będzie z wyników rzutu kostką (wynik różny od 6), jedna dwunasta zaś od tych osób, które wyrzuciły szóstkę i były swoje połowice. Pięć szóstych plus jedna dwunasta to jedenaście dwunastych, 91,67 %. Jeśli 9167 osób na 10000 odpowie „tak”, to mamy dobre prawo postawić hipotezę, że co drugi mąż bije żonę. Gdyby bił „tylko” co dziesiąty, odpowiedni procent wyniósłby 85. Wszystko to jest bardzo łatwym zadaniem matematycznym (nawet dla uczniów nowej, zreformowanej i zamerykanizowanej szkoły). Statystyka jest bezlitosna. Nie powie, jakie numerki obstawić w Totolotku, nie powie, kto zginie w wypadku i kto jest czułym mężem, ale potrafi z dużą dokładnością przewidzieć zysk przedsiębiorstwa loteryjnego, ile osób straci życie w weekend i ile jest bijących mężów.

„Czarno na białym” widać było potęgę statystyki w zestawieniu zbiorczym wyników egzaminu maturalnego w 2005 roku. Liczba osób, które otrzymały dokładnie 15 punktów (czyli 30 procent, próg zaliczenia), była o wiele większa niż liczba tych z czternastoma i szesnastoma punktami. O czym to świadczy, wszyscy zdajemy sobie sprawę. Podobnie sprawa fałszerstwa list wyborczych w jednym z

okręgów wyborczych wyszła na jaw, gdy odkryto, że jest na niej za dużo pomyłek w numerach PESEL osób popierających pewną kandydatkę<sup>12</sup>.

## Co od czego zależy, czyli korelacja

Żadna inna metoda statystyczna nie jest tak efektywna jak korelacja. Jest to zrozumiałe, bo przecież postęp naukowy w każdej dziedzinie polega w dużej mierze na znajdowaniu związków, między którymi istnieje zależność. Współczynnik korelacji jest liczbą, która mierzy, w jakim stopniu zjawiska są powiązane. Dodatnia korelacja oznacza, że wzrost jednej wielkości powoduje proporcjonalny wzrost drugiej. Zatem wielkości wprost proporcjonalne mają największą możliwą korelację dodatnią. Odwrotnie, wielkości odwrotnie proporcjonalne mają największą możliwą korelację ujemną.

Tych tematów nie ma w szkole, będą w planowanym poziomie docelowym nauczania matematyki. O współczynniku korelacji Pearsona wspominałem w tej książce w wykładzie 12 przy okazji mocy różnicującej zadania, a w tym rozdziale rozwinę temat. Współczynnik ten wyraża wzorem, który na pierwszy rzut oka nie wygląda zachęcająco, jest jednak stosunkowo prosty i zrozumiały. Jeżeli mam próbkę dwóch cech  $(x_i, y_i)$  dla  $i = 1, 2, \dots, n$ , czyli dwa ciągi tej samej długości (albo: wektory z tej samej przestrzeni  $\mathbf{R}^n$ ), to ich współczynnikiem korelacji nazywa się liczba

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Wzór stanie się prostszy, jeśli wprowadzimy oznaczenia, używane powszechnie w podręcznikach statystyki,  $X_i = x_i - \bar{x}$ ,  $Y_i = y_i - \bar{y}$ . Wielkości  $X_i$ ,  $Y_i$  mierzą zatem odchylenia kolejnych danych  $x_i$ ,  $y_i$  od wielkości średnich. Mamy zatem:

$$r = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \cdot \sqrt{\sum_{i=1}^n Y_i^2}} = \frac{\sum XY}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}} = \frac{(X, Y)}{\|X\| \cdot \|Y\|},$$

gdzie środkowa formuła jest stosowanym często przez statystyków skrótowym zapisem pełnego wzoru, a ostatnia równość będzie zrozumiana przez tych Czytelników, którzy pamiętają trochę algebry liniowej. Mianowicie  $(X, Y)$  to iloczyn skalarny wektorów  $[X_1, X_2, \dots, X_n]$  i  $[Y_1, Y_2, \dots, Y_n]$ , a podwójne pionowe kreski oznaczają długości tych wektorów. Dalej, Czytelnicy Ci spostrzegą, że jest to wzór wyznaczający kosinus kąta (niezorientowanego) między wektorami  $[X_1, X_2, \dots,$

<sup>12</sup> W czerwcu 2006 roku posłanka ta została skazana grzywnę i karę więzienia (w zawieszeniu). Oczywiście konsekwencji politycznych nie poniosła, a dlatego, to już naprawdę nie jest temat do książki o nauczaniu matematyki.

$X_n$ ] i  $[Y_1, Y_2, \dots, Y_n]$  przestrzeni  $\mathbf{R}^n$ . Z tej obserwacji wynika, że współczynnik korelacji jest zawsze co do modułu nie większy niż 1. Również z tej samej zależności można zrozumieć, że współczynnik korelacji +1 oznacza doskonałą korelację dodatnią (wektory odchyleń od średniej  $X$ ,  $Y$  są współliniowe i o tym samym zwrocie), a współczynnik równy minus 1 oznacza doskonałą korelację ujemną (wektory o tym samym kierunku, ale przeciwnym zwrocie). Wreszcie, jeżeli współczynnik ten wynosi zero, to wektory są prostopadłe, czyli dane statystyczne pochodzące z tych dwóch serii są „najbardziej niezależne”.

Przykład. Zbadajmy, czy między wysokością terenu a średnią roczną temperaturą na Podhalu Spiszu, Orawie i Liptowie jest dodatnia korelacja. Posłużmy się danymi z drugiej połowy XIX wieku, zebranymi przez Stanisława Eljasza-Radzikowskiego:

### Średnia ciepłota miesięczna i roczna po płu stronie Tatr 1871 - 1885

	Zamek					
	Kraków	Orawski	Poronin	Zakopane	Kuźnice	Jaworzyna Spiska
wysokość	<b>220</b>	<b>501</b>	<b>742</b>	<b>830</b>	<b>1000</b>	<b>1019</b>
miesiąc						
styczeń	-3,4	-5,3	-5,6	-5,6	-5,4	-6
luty	-1,6	-3,6	-4,6	-4	-5,1	-4,4
marzec	2,2	0	-1,3	-1	-1,1	-2,6
kwiecień	8	6,4	5	4,9	4,3	3,1
maj	12,4	10,5	9,4	8,7	7,5	7
czerwiec	17,3	15	14	13,7	13,2	12,2
lipiec	19	16,3	15,3	14,8	14,3	13,8
sierpień	17,6	15,2	14,2	14	13,8	11,9
wrzesień	14	11,9	10,7	10,9	9,7	8,4
październik	8,3	7,1	6,1	6	5,5	3,9
listopad	2,3	1,1	-0,3	-0,4	-0,4	-1,4
grudzień	-2	-4,1	-4,4	-4,1	-4,5	-4,9
rok	<b>7,8</b>	<b>5,9</b>	<b>4,9</b>	<b>4,8</b>	<b>4,3</b>	<b>3,4</b>

### Średnia ciepłota po południowej stronie Tatr

	Spiska		Lipt.			
	Nowa	Stare.Hory	Kieżmark	Hradek	Smokowiec	Stary Smokowiec
	Wieś					
wysokość	<b>465</b>	<b>486</b>	<b>631</b>	<b>652</b>	<b>1000</b>	<b>1005</b>
miesiąc						

styczeń	<b>-5,7</b>	<b>-5</b>	<b>-5,5</b>	<b>-6,2</b>	<b>-6,2</b>	<b>-5,4</b>
luty	<b>-3,1</b>	<b>-3,4</b>	<b>-3,1</b>	<b>-4,6</b>	<b>-3,9</b>	<b>-3,9</b>
marzec	<b>1,4</b>	<b>0,5</b>	<b>0,6</b>	<b>0,3</b>	<b>-1,5</b>	<b>-0,7</b>
kwiecień	<b>7,8</b>	<b>6,4</b>	<b>6,5</b>	<b>6,4</b>	<b>3,9</b>	<b>4,3</b>
maj	<b>12,2</b>	<b>10,9</b>	<b>10,5</b>	<b>10,6</b>	<b>8,2</b>	<b>9</b>
czerwiec	<b>16,7</b>	<b>15,3</b>	<b>15,5</b>	<b>14,9</b>	<b>13,1</b>	<b>13,7</b>
lipiec	<b>18,2</b>	<b>16,8</b>	<b>17,1</b>	<b>16,2</b>	<b>14,8</b>	<b>15,3</b>
sierpień	<b>17</b>	<b>15,7</b>	<b>15,8</b>	<b>15,6</b>	<b>13,5</b>	<b>14,1</b>
wrzesień	<b>13,1</b>	<b>11,6</b>	<b>12,3</b>	<b>11,9</b>	<b>10,7</b>	<b>10,9</b>
październik	<b>7,8</b>	<b>6,9</b>	<b>7,5</b>	<b>6,8</b>	<b>5,7</b>	<b>5,2</b>
Listopad	<b>1,3</b>	<b>1,2</b>	<b>1</b>	<b>0,4</b>	<b>0,1</b>	<b>-0,1</b>
Grudzień	<b>-4,1</b>	<b>-3,9</b>	<b>-4,3</b>	<b>-4,5</b>	<b>-4,1</b>	<b>-4,7</b>
Rok	<b>6,9</b>	<b>6,1</b>	<b>6,1</b>	<b>5,6</b>	<b>4,5</b>	<b>4,8</b>

Obliczmy współczynnik korelacji między wysokością (nad poziom morza) miejscowości a średnią roczną temperaturą. Układamy tabelkę z elementami występującymi we wzorze na  $r$ . W niej przez  $X$  i  $Y$  oznaczamy odchylenia kolejnych wartości od średniej, czyli  $x - \bar{x}$ ,  $y - \bar{y}$ .

Kolejne kroki są następujące:

1. Wypisujemy w kolumnach pary wyników  $x$ ,  $y$ , dbając o to umieszczenie razem odpowiadających sobie wyników.
2. Obliczamy obie średnie arytmetyczne i wpisujemy na dole, w kolumnach odpowiadających obu zmiennym.
3. Obliczamy dla obu zmiennych odchylenia od średniej i wpisujemy je w kolumnach  $X$ ,  $Y$ .
4. Podnosimy do kwadratu odchylenia  $X$  i  $Y$  i wpisujemy wyniki w następnych kolumnach.
5. Obliczamy iloczyny  $XY$  i wpisujemy kolejne iloczyny w ostatniej kolumnie.
6. Obliczamy sumy kolumn  $X^2$ ,  $Y^2$ ,  $XY$ .

7. Szukany współczynnik korelacji jest równy 
$$\frac{\sum(XY)}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}}$$

Zastosujemy to teraz do zbadania zależności między temperaturą a wysokością na północ od Tatr według danych Stanisława Eljasza-Radzikowskiego. W kolumnie  $x$  wpiszemy wysokości, w kolumnie  $y$  średnie roczne temperatury.

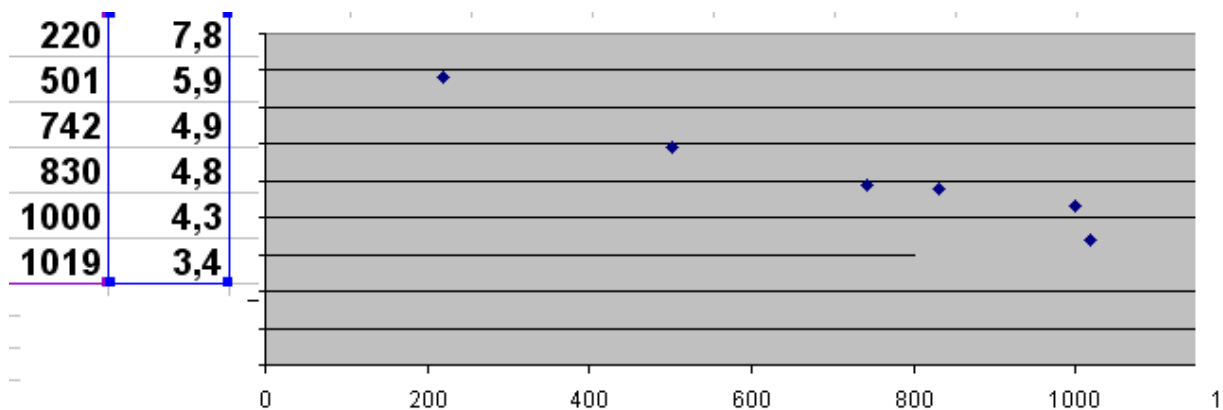
$x$	$y$	$X$	$Y$	$X^2$	$Y^2$	$XY$
220,00	7,80	-498,67	2,62	248668,44	6,85	-1304,84
501,00	5,90	-217,67	0,72	47378,78	0,51	-155,99
742,00	4,90	23,33	-0,28	544,44	0,08	-6,61
830,00	4,80	111,33	-0,38	12395,11	0,15	-42,68
1000,00	4,30	281,33	-0,88	79148,44	0,78	-248,51
1019,00	3,40	300,33	-1,78	90200,11	3,18	-535,59
średnia	<b>718,67</b>	<b>5,18</b>				
sumy				<b>478335,33</b>	<b>11,55</b>	<b>-2294,23</b>

Zgodnie z wzorem na  $r$  mamy

$$r = \frac{\sum_{i=1}^n X_i Y_i}{\sqrt{\sum_{i=1}^n X_i^2} \cdot \sqrt{\sum_{i=1}^n Y_i^2}} = \frac{\sum(XY)}{\sqrt{\sum X^2} \cdot \sqrt{\sum Y^2}} = \frac{-2294,93}{\sqrt{478335,33 \cdot 11,55}} \approx -0,97$$

Oplącił się trud. Współczynnik korelacji jest bardzo wysoki. Odkryliśmy ciekawe prawo: silną zależność średniej temperatury od wysokości. Na wykresie widać, że spadek temperatury przebiega niemal liniowo wraz ze wzrostem wysokości. Prosta, która „najlepiej pasuje” do takiego wykresu, nazywamy prostą regresji, ale omówienie jej wykracza poza ramy tej książki.

Prostą regresji widać na wykresie.... który proszę wykonać jako ćwiczenie. W Excelu podajemy dwie serie danych:  $x$  oraz  $y$  wzięte z powyższej tabeli. Wybieramy „kreator wykresów” i wykres punktowy. Niezłą – choć nie idealną – zależność liniową „psuje” tylko anomalia na samym końcu. Istotnie, porównywaliśmy tam Kuźnice z Jaworzyną Spiską, a choć oba miejsca są rzeczywiście po północnej stronie Tatr, to znający realia mogą się domyślić, że klimat tu i tam może być trochę inny.



Ćwiczenie. Przeanalizować powyższe dane o temperaturze na stokach Tatr. Obliczyć współczynnik korelacji dla stoków południowych. Za pomocą arkusza kalkulacyjnego

obliczyć współczynniki korelacji w różnych miesiącach. Obliczyć współczynniki korelacji między temperaturą na stokach północnych a południowych. Wyciągnąć wnioski.

Ćwiczenie. Zastanów się, jakie mogą być ciekawe, a łatwo dostępne dane statystyczne do sprawdzenia korelacji między nimi. Poszukaj tych danych w Internecie. Oblicz współczynnik korelacji.<sup>13</sup>

Obliczanie współczynnika korelacji ma duże walory dydaktyczne (organizacja pracy, sumienność, rzetelność) i jest jednym z niewielu wskaźników, rzeczywiście używanych w statystyce, które z powodzeniem można omówić w szkole, co więcej: wytłumaczyć jego znaczenie. A raczej *możnaby*, gdyby nie skromna liczba godzin matematyki.

Pomówmy o trudnościach filozoficznych związanych z pojęciem przyczynowości. Trudności te wychodzą również przy współczynniku korelacji. Wbrew powszechnej opinii, korelacja bliska 1 nie oznacza jakiegokolwiek związku przyczynowego. Przekonają nas o tym dwa proste przykłady. Jeszcze pół wieku temu dało się wykazać silną dodatnią korelację między liczbą narodzin dzieci w danym obszarze a liczbą bocianów na tym terenie. We Włoszech można zaś zaobserwować, że wraz ze wzrostem liczby spożywanego lodów, rośnie liczba kradzieży kieszonkowych. Czy oznacza to, że lody są kryminogenne, a może odwrotnie: okradzeni ludzie chętniej sięgają po lody? Nie, ani jedno, ani drugie. Jemy więcej lodów w czasie ciepłych miesięcy. Wtedy na ulicach włoskich miast jest więcej turystów. Są lżej ubrani i łatwiej ich okraść! A z bocianami? Też oczywiste. Bocianów jest zdecydowanie więcej na wsi, niż w miastach, a tradycyjnie na wsi ludzie mieli więcej dzieci niż w miastach (obecnie tendencja się nawet odwróciła). Pamiętajmy jednak: korelacja a relacja przyczyna-skutek to zupełnie dwie różne sprawy.

---

<sup>13</sup> Mam takie powiedzenie, którego może nadużywam. Otóż zdarza się nader często, że sztuką jest nie rozwiązanie zadania, a ułożenie. Mówię do studentów, że zadanie „ułoż zadanie” jest bardzo dobrym zadaniem.