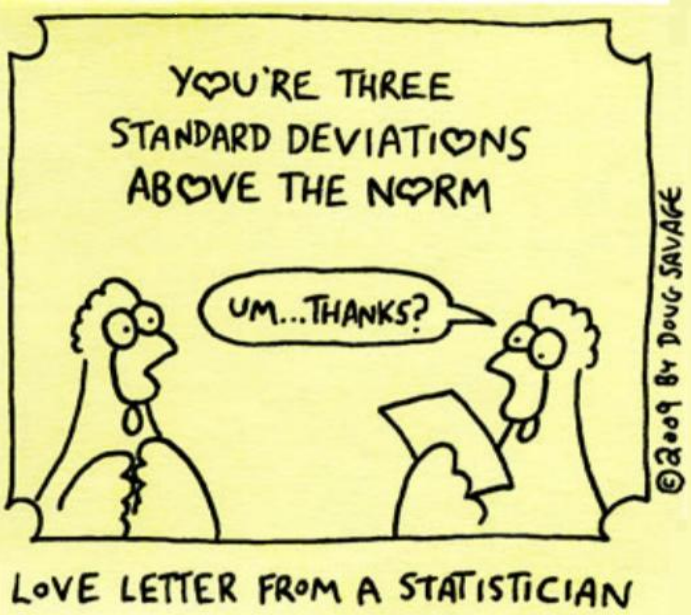


LOVE LETTER FROM A STATISTICIAN

2. Statystyka opisowa

2.1 Statystyka opisowa

Definicja 2.1 — Statystyka opisowa. Statystyka opisowa — dział statystyki zajmujący się metodami opisu danych statystycznych uzyskanych podczas badania statystycznego. Celem stosowania metod statystyki opisowej jest podsumowanie zbioru danych i wyciągnięcie pewnych podstawowych wniosków i uogólnień na temat zbioru. Statystykę opisową stosuje się zazwyczaj jako pierwszy i podstawowy krok w analizie zebranych danych.[2]

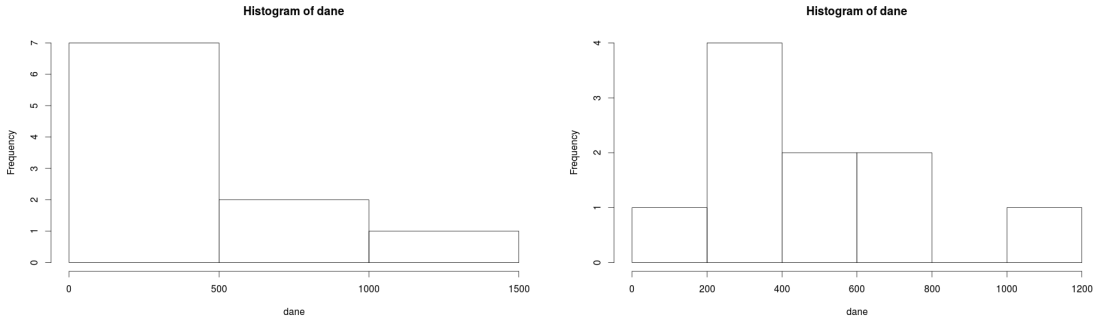
2.2 Rozkład cechy statystycznej

Jednym z podstawowych narzędzi statystyki opisowej jest szereg rozdzielczy, poznany na poprzednich zajęciach. Szereg rozdzielczy pozwala nam na analizowanie danych ilościowych w dużo wygodniejszy sposób, ale ma on jeden parametr krytyczny: liczbę przedziałów. Jeżeli przedziałów jest mało to wtedy widzimy dane w dużym przybliżeniu (nie widzimy szczegółów), z kolei gdy przedziałów jest dużo to dużo ciężej jest nam wyrobić sobie intuicję o danych przez analizę długiej tabelki.

Zauważ jednak, że gdy obok każdego przedziału dorysujemy liczbę kresek odpowiadającą liczności to zaczynają się one układać w pewien wykres. Taki wykres nazywamy histogramem i wizualizuje on empiryczną gęstość danych.

▪ **Przykład 2.1** Rozważając dane z poprzednich laboratoriów (liczba liniej kodu w programach studentów) zbudowaliśmy następujący szereg, wraz z narysowanymi kreskami dla każdej z liczności. Rysunek 2.1 prezentuje 2 histogramy sporządzone dla tych samych danych z różną liczbą przedziałów.

przedział	liczność	miejsce na kreski
(197,506]	7	IIIIIIII
(506,815]	2	II
(815,1124]	1	I



Rysunek 2.1: Histogramy dla przykładowych danych z 3 przedziałami (lewa) i z 6 przedziałami (prawa).

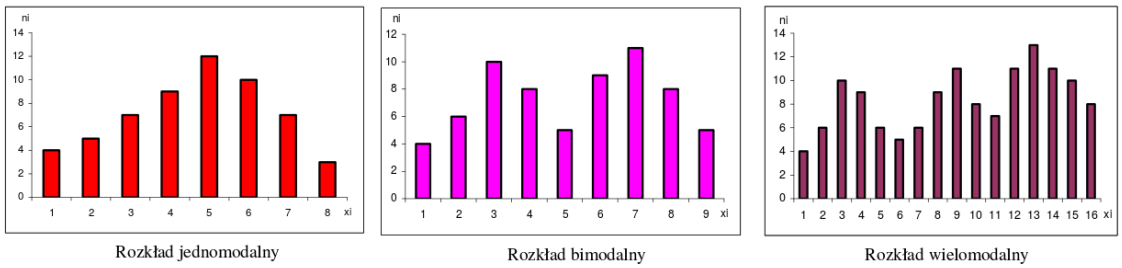
■

Poznajmy też trochę terminów statystycznych na opis rozkładów, które przedstawiają histogramy. Rozkład, który ma jeden wysoki „pik" nazywamy unimodalnym (tak jak te dwa histogramy w przykładzie), jeżeli ma ich 2 to wtedy bimodalnym, jeśli więcej to wielomodalnym[1]. Natomiast histogram, którego najwyższy „pik" jest w wartości skrajnej nazywamy antymodalnym.

Rozkład dodatkowo może być symetryczny (nasz wysoki „pik" jest otoczony w miarę symetrycznymi, coraz niższymi słupkami) lub skośny. W takiej sytuacji wyróżniamy rozkłady prawostronnie lub dodatnio skośne z długim ogonem po prawej stronie. A jeżeli sytuacja jest odwrotna to nazywamy go lewostronnie lub ujemnie skośnym.

Problem 2.1 W szkole podstawowej zmierzono wysokość wszystkich osób się tam znajdujących (w godzinach pracy szkoły). Ilu modalnego rozkładu się spodziewasz? [6]

[1] Aby być w 100% poprawnym powinniśmy rozróżniać między rozkładami uni-, bi-, wielo-modowymi a rozkładami jedno-, dwu-, wielo-wierzchołkowymi. Rozkładowi przyznajemy taką modalność ile jest najwyższych „pików" (czyli, aby był bimodalny muszą być 2 „piki" o takiej samej wysokości), natomiast jeżeli po prostu widzimy 2 wysokie „piki" otoczone przez wiele niższych to jest to rozkład dwuwierzchołkowy. Natomiast w praktyce dość często używa się słowa „bimodalny" na określenie rozkładu dwuwierzchołkowego



Rysunek 2.2: Różne typy rozkładów [1]

2.3 Miary tendencji centralnej

Pomimo tego, ze histogram czy szereg rozdzielczy dostarczają nam sporo informacji o badanej zmiennej to często chcielibyśmy wyrazić wiedzę o naszych danych w postaci jednej liczby (np. w celu łatwego porównania dwóch cech, próbek lub populacji).

Jedną z podstawowych charakterystyk rozkładu jest to gdzie leży jego centrum/środek, jaka wartość jest „typowa" dla tych danych W tym rozdziale poznamy kilka takich miar dla opisu tej właśnie cechy. Miary te nazywamy miarami tendencji centralnej.

Definicja 2.2 — Dominanta. Dominantą lub modą (ang. *mode*) nazywamy wartość (lub wartości), która występuje w próbce najczęściej.

Definicja 2.3 — Mediana w próbce. Gdy rozmiar próby n jest nieparzysty to medianą nazywamy wartość leżącą dokładnie na środku posortowanego ciągu z danymi.

$$x_{med} = x_{(n+1)/2}$$

Jeżeli rozmiar próby jest parzysty to medianą nazywamy średnią arytmetyczną dwóch elementów leżących na środku posortowanego ciągu z danymi.

$$x_{med} = \frac{x_{n/2} + x_{(n/2)+1}}{2}$$

Mediana jest więc środkową wartością danych, których dokładnie połowa jest od niej większa, a druga połowa jest od niej mniejsza. Zwróć uwagę, że często (mentalnie) taką własność przypisujemy do średniej arytmetycznej, co jest błędem. Medianę możemy także zdefiniować ogólniej dla zmiennej losowej X jako wartość x_{med} taką, że $P(X \geq x_{med}) \geq 0.5 \wedge P(X \leq x_{med}) \leq 0.5$

Definicja 2.4 — Średnia w populacji. Średnią arytmetyczną w populacji nazywamy miarę wyrażoną wzorem:

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

Definicja 2.5 — Średnia w próbce. Średnią arytmetyczną w próbce nazywamy miarę wyrażoną wzorem:

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

Zauważ, że wzór na średnią w populacji i średnią w próbce jest taki sam, różnią się tylko oznaczenia: μ oznacza średnią w populacji (która zwykle nie jest nam znana), a $\bar{x}$ oznacza średnią w próbce (która jest pewnym przybliżeniem nieznanej, prawdziwej wartości μ). Wprowadzamy takie rozróżnienie nie tylko po to aby zaznajomić Ciebie ze stosowaną w statystyce notacją, ale również dlatego, że już w następnej sekcji poznasz miary które inaczej liczysz w próbce, a inaczej w populacji.

Średnią arytmetyczną możemy interpretować jako, znany z fizyki, środek masy ciężkości rozkładu. Niestety, ma też pewne wady np. średnia arytmetyczna nie sprawdza się jako miara znajdująca „najbardziej typową" obserwację w rozkładach nie unimodalnych. Inną wadą średniej arytmetycznej jest też to, że w przeciwieństwie do mediany czy dominanty

często przyjmuje ona wartości ułamkowe, także dla wartości dyskretnych. Dla przykładu średnia liczba rąk w sali laboratoryjnej z danych 10, 14, 16 będzie wynosiła $13\frac{1}{3}$ ręki.

Problem 2.2 Zidentyfikuj wartość średniej populacji i/lub średniej próbki [6]:

- Przeciętny Amerykanin wydał średnio 52$ w 2007 roku na dekoracje Halloween'owe. Aby sprawdzić czy ta kwota się zmieni badacze przeprowadzili ankietę w 2008, która obejmowała 1500 gospodarstw domowych i wynikało z niej, że przeciętny Amerykanin wydał średnio 58$ na dekoracje Halloween'owe w 2008.
- Wśród studentów pewnej politechniki przeprowadzono ankietę w której zapytano 294 studentów o ich średnią ocen. Następnie w celu sprawdzenia reprezentatywności porównano uzyskaną wartość (3,85) z wewnętrznymi danymi uczelni (3,35).
- Średni czas biegacza w tegorocznej edycji Maratonu Poznańskiego wyniósł 3 godziny 23 minuty i 42 sekundy.

Problem 2.3 Dla jakiego typu danych możesz użyć: mediany, średniej, dominanty?

Problem 2.4 Dwóch studentów tego samego roku ma średnią ocen 3.5. Czy prawdziwe jest stwierdzenie, że ich oceny z poszczególnych przedmiotów nie muszą być takie same, ale całkowita liczba piątek (5.0) dla każdego z nich musi być taka sama?

Ćwiczenie 2.1 — Średnia i mediana. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/02/cw-2.xls` i rozwiąż ćwiczenie. ■

Problem 2.5 Średnia wynik drużyny A wynosi 9. Jak bardzo różne są te dane od danych ze średnią 120? O czym nam mówi średnia arytmetyczna?

Ćwiczenie 2.2 Do wyników obu drużyn dodaj dodatkowego uczestnika z liczbą punktów 1000. Jak zmieniła się średnia? Jak zmieniła się mediana? ■

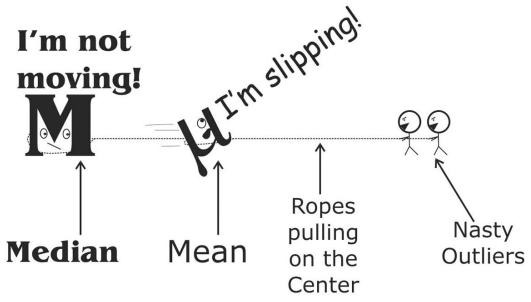
Zaobserwowaliśmy, że średnia arytmetyczna jest bardzo wrażliwa na dodanie do zbioru jednej/kilku wartości o skraje dużej lub skrajnie małej wartości. Mediana natomiast jest na takie skrajne wartości odporna, co powoduje że często nazywamy ją statystyką odporną (ang. *robust, resistant statistic*).

W statystyce skrajne obserwacje z często ekstremalnymi wartościami cechy nazywamy obserwacjami odstającymi. Powstawanie takich obserwacji może się wiązać z błędami przy zbieraniu danych (Pani przy okienku wpisało się jedno 0 za dużo), błędów technologicznych (np. przekręcony `int`) lub, co gorsza, może być pewną własnością badanego procesu. Nie ma więc jasnej odpowiedzi na pytanie co powinno się zrobić z obserwacjami odstającymi, gdyż odpowiedź ta zależy od specyfiki problemu.

Definicja 2.6 — Obserwacja odstająca. Obserwacja odstająca lub samotnicza (ang. *outlier*) to obserwacja, która przyjmuje ekstremalną wartość badanej cechy statystycznej w porównaniu z innymi obserwacjami.

Pomimo tego średnia arytmetyczna jest używana dość często w praktyce i to w bardzo ważnych zastosowaniach np. przy obliczeniu średniego kursu walut. Jak bankowcy radzą sobie z brakiem odporności na wartości odstające?

> Narodowy Bank Polski skupia się przede wszystkim na kursie euro i dolara amerykańskiego (EUR i USD). Ma swoją listę 10 banków, do których zwraca się pomiędzy godziną 10:55 a 11:00 z zapytaniem o kurs sprzedaży i zakupu



Rysunek 2.3: Która z miar położenia jest najbardziej odporna na obserwacje odstające?

> tych walut. Następnie sumuje kurs sprzedaży i zakupu danej waluty w każdym banku i dzieląc sumę przez 2, oblicza znaną nam już średnią arytmetyczną. Żeby pozbyć się wartości skrajnych (...), z tych dziesięciu średnich NBP odrzuca dwie najwyższe i dwie najniższe. A z pozostałych sześciu znowu liczy średnią arytmetyczną [3].

Taki sposób obliczania średniej nazywamy średnią ucinaną. Średnia ucinana jest też stosowana np. przy ocenie skoków narciarskich, gdzie pomija się ocenę sędziów, którzy przyznali odpowiednio najwyższą i najniższą notę.

Definicja 2.7 — Średnia ucinana. Średnią ucinaną (trymowaną) nazywamy miarę wyrażoną wzorem:

$$\bar{x} = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_i$$

gdzie x_i są posortowane od najmniejszej do największej.

Problem 2.6 Według Facebook'a (2011) średnia liczba znajomych użytkowników portalu wynosi 190, a połowa użytkowników ma więcej niż 100 znajomych. Mając do dyspozycji te dane jakiego kształtu rozkładu tej cechy statystycznej się spodziewasz [6]?

Ciekawą obserwację o poznanych miarach tendencji centralnej można znaleźć w [9]. Autor zauważa, że każda z poznanych miar: średnia, mediana i dominanta jest wartością która jest najbliżej wszystkich danych, tylko że każda z nich używa innej definicji odległości:

- średnia arytmetyczna jest najbliżej wszystkich wartości biorąc pod uwagę odległość Euklidesową ($\sqrt{\sum(x_i - \bar{x})^2}$),
- mediana jest najbliżej wszystkich wartości biorąc pod uwagę bezwzględną wartość różnicy wartości (metryka Manhattan, miejska $\sum |x_i - x_{med}|$),
- dominanta optymalizuje odległość zdefiniowaną jako „jeżeli liczby są takie same to odległość wynosi 0, w przeciwnym wypadku 1".

ADVANCED

Skoro średnia arytmetyczna jest tak nieodporna na obserwacje odstające to czy w ogóle jest sens ją liczyć? Czy daje nam ona jakąś wiedzę, poza środkiem ciężkość rozkładu? Aby lepiej to unaocznić rozwiążmy proste zadanie.

■ **Przykład 2.2** Załóżmy, że mamy 100-elementową próbę losową z cechy ilościowej która jest nieujemna. Średnia z tej próby wynosi $\bar{x} = 10$, jaka jest maksymalna liczba obserwacji większa równa $x' = 20$?

Ponieważ mamy cechę nieujemną $x_i \geq 0$, a największą liczbę uzyskamy gdy próbka będzie się składała z samych zer i dwudziestek. Oznaczmy szukaną liczbę dwudziestek jako $a \geq 0$, mamy wtedy

$$\frac{ax' + 0 \cdot (100 - a)}{100} = 10$$
$$\frac{20}{100}a = 10$$
$$a = 50$$

Odpowiedź: W tej próbce nie może być więcej niż 50 obserwacji większych równych 20. ∎

Jak pewne się spodziewasz obserwację z tego zadania możemy uogólnić do stwierdzenia dla próbki o dowolnej wielkości: w próbce nieujemnych liczb losowych o średniej $\bar{x} = 10$ nie może być więcej niż 50% obserwacji większych równych 20.

Możemy uogólnić to dalej do dowolnej wartości średniej $\bar{x}$ i dowolnego x' i dochodzimy wtedy do nierówności Markowa.

Twierdzenie 2.1 — Nierówność Markowa. Jeżeli średnia listy *nieujemnych* liczb x_i wynosi $\bar{x}$ to stosunek liczb większych lub równych x' jest mniejszy równy $\frac{\bar{x}}{x'}$.

Bardziej formalnie: Niech X będzie nieujemną zmienną losową i załóżmy, że $\mathbb{E}(X)$ istnieje. Dla każdego $x' > 0$,

$$P(X \geq x') \leq \frac{\mathbb{E}(X)}{x'}$$

Problem 2.7 Na pewnym uniwersytecie obowiązują oceny pomiędzy 0 (niedostateczny) i 4.0 (celująco). Student ma średnią ocen 2.2, jaki jest maksymalny współczynnik przedmiotów, który zaliczył celująco? [9]

Kończąc, rozważmy sytuację gdy trzy miary: dominanta, średnia arytmetyczna i mediana są sobie równe. Co to znaczy? To znaczy, że mamy do czynienia z rozkładem symetrycznym, a wszystkie wartości miar wskazują na jego najwyższy „pik”. W przypadku rozkładów mocno niesymetrycznych, koniec końców, mediana jest chyba lepszą miarą niż średnia, ponieważ nie bierze pod uwagę obserwacji odstających. Z tego powodu często stosujemy inne, dodatkowe miary, które działają bardzo podobnie jak mediana.

Definicja 2.8 — Percentyle. k-tym percentylem nazywamy wartość należącą do próby, poniżej której znajduje się $k\%$ liczb z tej próby.

Zauważ, że 50 percentylem jest właśnie mediana, 0 percentylem jest minimalna wartość, a 100 percentylem jest wartość maksimum.

Definicja 2.9 — Kwartyle. Pierwszy kwartyl to 25 percentyl, drugi kwartyl to 50 percentyl (mediana), a trzeci kwartyl to 75 percentyl.

Istnieje wiele różnych sposobów obliczania percentyli (kwartyli). Wystarczy wspomnieć, że standardowa w R funkcja `quantile()` umożliwia obliczenie percentyli na 9 różnych sposobów! A każdy ze sposobów może dawać różne wyniki w zależności od danych[2]. Jedną z możliwości formalnego zdefiniowania k-tego percentyla, którą będziemy

[2] Zobacz ciekawy test na `http://tolstoy.newcastle.edu.au/R/e17/help/att-1067/Quartiles_in_R.pdf`

stosować na laboratoriach jest

$$\arg\min_x F(x) \geq k\%$$

gdzie $F(x)$ oznacza wartość empirycznej dystrybuanty – czyli dystrybuanty którą konstruujemy na podstawie danych.

▪ **Przykład 2.3** Mamy próbkę danych o wartościach $2, 10, 10, 48, 99$. Wyznaczmy 30 percentyl tych danych. W tym celu obliczmy prawdopodobieństwo każdej z danych, a następnie obliczmy wartości dystrybuanty empirycznej.

x	2	10	48	99
P(x)	0,2	0,4	0,2	0,2
F(x)	0,2	0,6	0,8	1,0

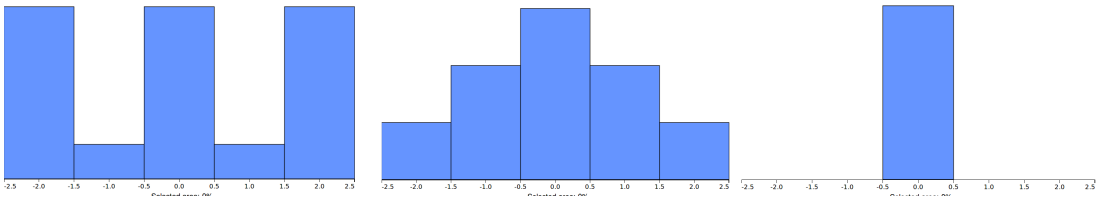
Znajdźmy teraz wartość x taką, że $\min_x F(x) \geq 30\%$. Taką wartością jest 10, ponieważ dla $F(10) = 0,6 \geq 0,3$, a kolejnym mniejszym x jest 2, którego wartość dystrybuanty jest już mniejsza niż 30%. ▪

W analizie danych często korzystamy z tak zwanego posumowania pięcioliczbowego (ang. *five-number summary*), czyli podsumowania składającego się z minimum, wartości pierwszego kwartyla, mediany, wartości trzeciego kwartyla i maksimum. W R takie podsumowanie możemy uzyskać za pomocą funkcji `fivenum()` lub przy pomocy standardowej funkcji `summary()`, która dodatkowo wyświetla średnią arytmetyczną danych.

2.4 Miary rozproszenia

Ćwiczenie 2.3 Otwórz arkusz Excela: `http://www.cs.put.poznan.pl/mlango/siad/lab2/cw-var.xls` ▪

Wiemy już jak zidentyfikować centrum rozkładu, ale centrum rozkładu to nie wszystko: dane mogą mieć np. tę samą średnią arytmetyczną i medianę, a bardzo się różnić (patrz rysunek 2.4). Innym wskaźnikiem opisującym dane jest poziom rozproszenia wartości, jeden z nich już poznaliśmy na poprzednich laboratoriach:



Rysunek 2.4: Trzy histogramy z taką samą średnią i medianą, równą 0.

Definicja 2.10 — Rozstęp. Rozstęp wyrażamy wzorem:

$$R = x_{max} - x_{min}$$

Rozstęp ma oczywistą wadę: nie jest w ogóle odporny na obserwacje odstające, ale co gorsza nie bierze pod uwagę żadnej innej obserwacji oprócz minimum i maksimum

(wyobraź sobie wszystkie możliwe histogramy które możesz narysować tak aby rozstęp się nie zmienił).

Jeżeli więc chcemy uwzględnić wszystkie wartości to może policzmy jak daleko są one od średniej? Możemy zdefiniujmy więc miarę jako średnią odległość danej od średniej:

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

Hej, coś zdefiniowaliśmy i nie ma rameczki „definicja"? No cóż, ta miara zbyt wiele nam nie powie... Dlaczego? Bo miara ta zawsze wynosi 0.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i) - \frac{1}{n}\sum_{i=1}^{n}(\bar{x}) = \bar{x} - \frac{1}{n}\sum_{i=1}^{n}(\bar{x}) = \bar{x} - \frac{1}{n}n\bar{x} = \bar{x} - \bar{x} = 0$$

Twierdzenie 2.2 Suma wszystkich odchyleń od średniej arytmetycznej jest równa 0.

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

Możemy to sobie bardzo prosto wyobrazić np. na podstawie rozkładu normalnego. Jeżeli odejmiemy od każdej z wartości średnią to liczby w prawej połowie histogramu będą dodatnie, a liczby w lewej połowie będą ujemne. Tych liczb jest tyle samo po obu stronach średniej, więc po dodaniu te liczby się zniosą. Z tego powodu musimy zapobiec znoszeniu się tych liczb, bo zarówno odległości po prawej stronie średniej jak i lewej wnoszą tyle samo informacji o zmienności danych. Z tego powodu, aby uczynić wszystkie odległości dodatnimi, wyciągnijmy z nich wartość bezwzględną.

Definicja 2.11 — Odchylenie średnie. Odchylenie średnie określone jest wzorem[a]:

$$D_m = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

[a] Czasami stosuje się „odchylenie średnie od mediany", gdzie $\bar{x}$ we wzorze zastępuje się x_{med}. Zgodnie z wcześniejszym komentarzem dotyczącym optymalizowanej przez medianę odległości, odchylenie średnie od mediany osiąga minimalną wartość tj. żadna inna liczba wpisana do wzoru zamiast $\bar{x}$ nie da niższego rezultatu niż mediana dla danej próbki .

Jednakże wartość ta ma jedną podstawową wadę: wartość bezwzględna powoduje różne trudności matematyczne np. ciężko rozwiązywać równania, albo liczyć całki czy pochodne. Z tego powodu dużo częściej spotykamy się z miarą, w której różnicę pomiędzy średnią a daną podnosi się do drugiej potęgi. Wprowadzenie potęgi ma także pozytywny efekt uboczny: małe różnice (czyli wartości blisko średniej) mają mniejszy wpływ na wartość wskaźnika niż duże różnice.

Definicja 2.12 — Wariancja w populacji. Wariancję w populacji określamy wzorem:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

Warto się zastanowić dlaczego liczymy średni kwadrat różnicy od średniej, a nie od innej wartości np. od mediany albo mody? Otóż nie zależnie jaką liczbę wstawimy

zamiast μ do wzoru zawsze minimum tej funkcji uzyskamy dla średniej (co jest zgodne z komentarzem o minimalizowanej przez nią odległością). Pokażmy to matematycznie szukając minimum wyrażenia $\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2$ oczywiście za pomocą pochodnej:

$$\frac{d}{da}\frac{1}{n}\sum_{i=1}^{n}(x_i-a)^2 = -\frac{2}{n}\sum_{i=1}^{n}(x_i-a) = -2\frac{1}{n}\left(\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}a\right) = -2(\bar{x}-a)$$

widzimy więc, że równanie przyjmuje 0 dla $a=\bar{x}$.

Pokazaliśmy już wzór dla wariancji w populacji – jaki jest wzór na wariancję w próbie? Niestety nie taki sam...

Definicja 2.13 — Wariancja w próbie. Wariancję w próbie, w sytuacji gdy średnia populacji μ nie jest znana, określamy wzorem:

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

Oczywiście pojawia się tutaj pytanie: skąd wzięło się $n-1$ w mianowniku? Stosowanie $n-1$ zamiast n we wzorze na wariancję nazywamy poprawą Bessel'a (ang. *Bessel's correction*) od nazwiska jej autora. I niestety, na to pytanie nie udzielę dzisiaj jednoznacznej odpowiedzi, ale pojawi się ona za 2 tygodnie, kiedy będziemy mówili o teorii estymacji.

Na razie spróbuj wyobrazić sobie to w ten sposób: jeżeli znałbyś prawdziwą wartość μ to na podstawie próbki mógłbyś bez problemu oszacować wariancję używając standardowo n w mianowniku. Jednak skoro tej wartości nie znasz to najpierw musisz użyć swoich danych do oszacowania $\bar{x}$. Zauważ jednak, że jeśli tak robisz to dopasowujesz średnią $\bar{x}$ do tych danych, a następnie na podstawie różnicy wartości z tą średnią wyliczasz wariancję! Dlaczego następuje dopasowanie średniej? Z tej zależności, którą pokazaliśmy wcześniej: $\sum_{i=1}^{n}(x_i-\bar{x})=0$. Zauważ, że $\bar{x}$ jest tylko pewnym przybliżeniem/oszacowaniem prawdziwej wartości μ, a więc gdybyś policzył $\sum_{i=1}^{n}(x_i-\mu)$ to najprawdopodobniej[3] uzyskasz wartość różną od zera. Wynika[4] z tego że $\sum_{i=1}^{n}(x_i-\bar{x})^2 \leq \sum_{i=1}^{n}(x_i-\mu)^2$, a więc oszacowana wartość wariancji będzie zbyt mała w stosunku do prawdziwej. Stosujemy więc poprawkę, aby trochę ją zwiększyć ;) Warto zauważyć, że dla bardzo dużych n różnica jest praktycznie nieistotna.

Problem 2.8 Jaki jest zakres wartości wariancji i rozstępu?

Problem 2.9 Jak zmieni się rozstęp i wariancja danych z chwilą gdy do całej próby dodamy stałą liczbę c?

Problem 2.10 Jak zmieni się rozstęp i wariancja danych z chwilą całą próbę pomnożymy przez stałą liczbę c?

Wariancja ma jedną dużą wadę: jej jednostka to jednostka cechy statystycznej do kwadratu. Załóżmy, że liczmy wariancję wysokości człowieka w metrach. Jak będzie wtedy jednostka wariancja? Metry kwadratowe. Jak możemy jednostkę powierzchni odnieść do wysokości człowieka? Z tego powodu najczęściej posługujemy się w analizie wartością odchylenia standardowego:

[3] Chyba, że masz idealną estymację średniej...

[4] $\sum_{i=1}^{n}(x_i-\mu)^2 = \sum_{i=1}^{n}\left((x_i-\bar{x})+(\bar{x}-\mu)\right)^2 = \sum_{i=1}^{n}\left((x_i-\bar{x})^2+2(\bar{x}-\mu)(x_i-\bar{x})+(\bar{x}-\mu)^2\right) = \sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-\mu)\sum_{i=1}^{n}(x_i-\bar{x})+\sum_{i=1}^{n}(\bar{x}-\mu)^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(\bar{x}-\mu)^2 \geq \sum_{i=1}^{n}(x_i-\bar{x})^2$

Definicja 2.14 — Odchylenie standardowe. Odchylenie standardowe odpowiednio w populacji i w próbie określamy wzorami:

$$S = \sqrt{S^2} \quad \sigma = \sqrt{\sigma^2}$$

Problem 2.11 Z 25 osobowej grupy dziekańskiej, 24 studentów napisało egzamin a jeden student z powodu choroby ma go napisać w późniejszym terminie. Profesor sprawdził pracę studentów, którzy pisali w pierwszym terminie i średnia punktów wyniosła 74 z odchyleniem standardowym $8,9$. Student, który napisał egzamin w późniejszym terminie otrzymał 64 punkty [6].

- Jak wynik studenta wpłynie na średni wynik?
- Jaka jest nowa wartość średniej?
- Jak wynik studenta wpłynie na odchylenie standardowe?

Dodatkowo często chcielibyśmy porównywać zmienność dwóch zbiorów danych z różnymi średnimi i odchyleniami standardowymi. Nie powinniśmy tego robić tylko przy użyciu samego odchylenia standardowego. Dlaczego? Wyobraźmy sobie dwa zbiory danych określające ceny pewnych produktów. W pierwszym zbiorze mamy ceny mieszkań z $\bar{x} = 250000\,zł$ i $s = 5\,zł$. Widzimy, że dane są bardzo skoncentrowane wokoło tej dużej kwoty (co jest nawet trochę podejrzane). W drugim zbiorze mamy cenę paczki zapałek z $\bar{x} = 0,99\,zł$ i $s = 5\,zł$. W tym zbiorze wahania ceny określilibyśmy jako bardzo duże, pomimo tego, że odchylenie standardowe ma taką samą wartość jak w pierwszym zbiorze! Z tego powodu do porównywania zmienności danych pomiędzy różnymi zbiorami używamy współczynnika zmienności.

Definicja 2.15 — Współczynnik zmienności. Współczynnik zmienności wyrażamy wzorem:

$$V_S = \frac{S}{\bar{x}}$$

Jak pewnie zauważyłeś odchylenie standardowe ma tą samą jednostkę jak badana cecha. Fakt ten statystycy notorycznie wykorzystują do uproszczenia sobie życia i... zapomnienia o jednostkach w której została wyrażona zmienna. Jeżeli znamy średnią i odchylenie standardowe zmiennej to możemy ją przemienić w zmienną wyrażoną w „liczbie odchyleń standardowych od średniej".

Jak uzyskać taką zmienną w nowej jednostce? Pomyśl sobie o konwersji wagi z funtów na kilogramy ($1\,kg \approx 2,2$ funta): po prostu wagę wyrażoną w funtach podzieliłbyś przez $2,2$ np. $10\,[\text{funtów}] = \frac{10}{2,2} = 4,5\,[kg]$. Tak samo jest w statystyce, jeżeli wiemy, że zmienna o odchyleniu standardowym $S = 10$ jest większa o 5 od średniej to wiemy, że jest ona większa o $0,5$ odchylenia standardowego. A wiec aby przekonwertować zmienną do nowej jednostki należy najpierw odjąć od niej średnią (aby uzyskać odległość od średniej), a potem podzielić przez odchylenie standardowe.

Definicja 2.16 — Standaryzacja Z. Standaryzacją Z (ang. *z-score*) nazywamy przetransformowanie każdej wartości cechy statystycznej następującym wzorem[a]:

$$z_i = \frac{x_i - \bar{x}}{S}$$

[a] Bardziej ogólny wzór dla zmiennej losowej $Z(X) = \frac{X - \mathbb{E}X}{\mathbb{D}X}$

Zauważ że jednostką ustandaryzowanej zmiennej jest „liczba odchyleń standardowych", a

jej średnia zawsze wynosi 0.

Fakt, że zmienna Z nie ma konkretnej jednostki (jej jednostką jest liczba odchyleń standardowych od średniej) wykorzystują np. psycholodzy [7]. Dla przykładu, chciano zbadać jak zaangażowanie rodziców wpływa na wyniki osiągane przez dziecku w grze w tenisa. Wyniki dziecka w grze można łatwo zmierzyć, ale jak zmierzyć zaangażowanie rodziców? W tym celu zadano im kilka pytań np. „Ile pieniędzy rocznie przeznaczają Państwo na sprzęt do tenisa?", „Ile pieniędzy rocznie przeznaczają Państwo na ubrania do tenisa?", „Ile kilometrów tygodniowo przejeżdżają Państwo aby dowieść dziecko na zajęcia i wydarzenia sportowe związane z tenisem?" itd. W dalszej analizie potrzebowano jednego wskaźnika do wyrażenia zaangażowania rodziców, a tutaj problem: niektóre odpowiedzi wyrażone są w km/rok, inne w zł/rok lub w zł/tygodniowo? W badaniu tym po prostu zamieniono każdy wynik pytania na wartość z, a ponieważ dla każdego pytania jest ona po prostu wyrażona w odchyleniach standardowych bez problemu można było je dodać i porównywać [10].

W przypadku średniej arytmetycznej z pomocą w jej interpretacji przychodzi nam nierówność Markowa, to dla interpretacji odchylenia standardowego mamy nierówność Czebyszewa.

Twierdzenie 2.3 — Nierówność Czebyszewa. Jeżeli średnia listy liczb x_i wynosi $\bar{x}$ a odchylenie standardowe wynosi S to stosunek liczb oddalonych od średniej o k lub więcej odchyleń standardowych jest mniejszy równy $\frac{1}{k^2}$.

Bardziej formalnie: Niech $\mu = \mathbb{E}(X)$, $\sigma^2 = \mathbb{D}^2(X)$ i $Z(X) = \frac{X-\mathbb{E}X}{\mathbb{D}X} = \frac{X-\mu}{\sigma}$ wtedy[a]

$$P(|X-\mu| \geq x') \leq \frac{\sigma^2}{x'^2}$$

$$P(|Z| \geq k) \leq \frac{1}{k^2}$$

[a] Dowód przy użyciu nierówności Markowa: $P(|X-\mu| \geq x') = P(|X-\mu|^2 \geq x'^2) \leq \frac{\mathbb{E}[(X-\mu)^2]}{x'^2} = \frac{\sigma^2}{x'^2}$

Co wynika z tego twierdzenia? Że znając średnią i odchylenie standardowo danej zmiennej mamy pewność, że maksymalnie $\frac{1}{4} = 25\%$ danych jest oddalonych od średniej o 2 odchylenia standardowe, a $\frac{1}{9} \approx 11\%$ o 3 itd.

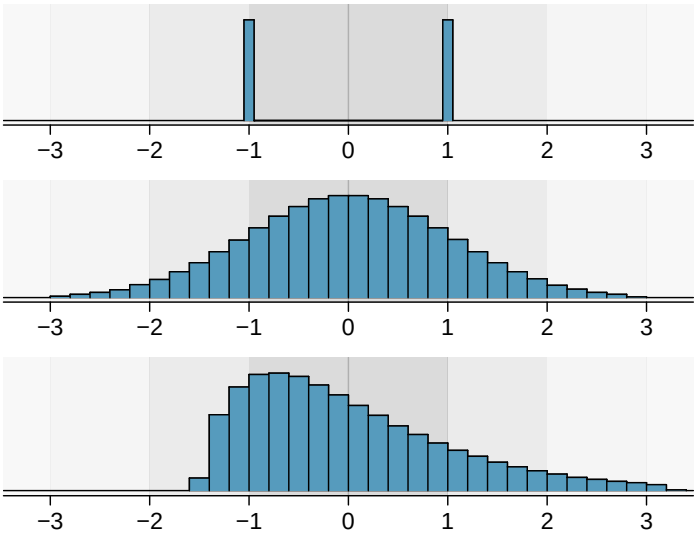
Problem 2.12 Zgodnie z nierównością Czebyszewa jaki maksymalny procent danych może być większy o 5 odchyleń standardowych?

Problem 2.13 Pewny typ żarówki ma średni czas życia równy 10 000 godzin z odchyleniem standardowym 470 godzin. Jaki procent żarówek będzie świecił dłużej niż 12491 godzin? [9]

Problem 2.14 Średnia waga zawodnika sumo to 140 kg z odchyleniem standardowym 30 kg. Jaki procent zawodników ma wagę pomiędzy 90 kg a 190 kg? (nie można policzyć dokładnie, znajdź górne ograniczenie)[9]

Problem 2.15 W klasie jest 200 studentów. Średnia liczba pieniędzy w portfelu studenta to 15 zł. Ilu z nich może mieć powyżej 75 zł w portfelu? [9]

Jak pewnie podejrzewasz, odchylenie standardowe, tak samo jak średnia jest podatne na obserwacje odstające. Czy istnieje jakaś miara rozstępu bardziej na to odporna? Tak, to



Rysunek 2.5: Trzy różne rozkłady wartości próby z tą samą średnią $\mu = 0$ i odchyleniem standardowym $\sigma = 1$. [6]

miara zbudowana analogicznie jak rozstęp ale zamiast minimum i maksimum bierzemy pod uwagę różnicę pomiędzy pierwszym i trzecim kwartylem.

Definicja 2.17 — Rozstęp międzykwartylowy. Rozstępem międzykwartylowym nazywamy miarę:

$$IQR = Q_3 - Q_1$$

Problem 2.16 Dla pewnej listy elementów podano poniżej statystyki. Ile będą one wynosiły dla tej samej listy, której każdy z elementów został pomnożony przez 9 a następnie dodano do niego 3? ($x'_i = 9x_i + 3$)

- $\bar{x} = 8$
- 11 jest dominantą
- $x_{med} = 15$
- $S = 20$
- $IQR = 23$
- $R = 25$

2.5 Skośność i kurtoza

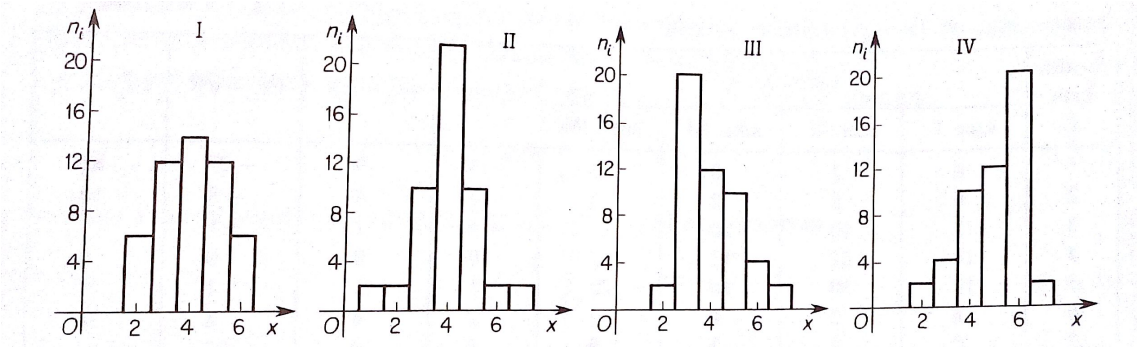
Niestety, wiedza o średniej i o odchyleniu standardowym czasami nie jest wystarczająca aby zidentyfikować prawidłowo rozkład (patrz rysunek 2.5). Potrzebne są więc nam jakieś dodatkowe miary... Jak je uzyskać? Wtedy budzi się w nas zmysł Polaka-kombinatora i zastanawiamy się a co by było gdyby w wariancji zamiast podnosić różnicę do kwadratu podnosić ją do 3 lub 4 potęgi?

Definicja 2.18 — Moment centralny. Momentem centralnym rzędu k nazywamy:

$$M_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Sprawdźmy jakie wartości uzyska M_3 i M_4 dla przykładowych danych:

■ **Przykład 2.4** Dla 4 zbiorów danych, które prezentują histogramy poniższe histogramy policzono średnią, odchylenie standardowe i M_3 oraz M_4.[8]



I $\bar{x} = 4 \quad S^2 = 1,44 \quad M_3 = 0 \quad M_4 = 4,32$
II $\bar{x} = 4 \quad S^2 = 1,44 \quad M_3 = 0 \quad M_4 = 8,16$
III $\bar{x} = 4 \quad S^2 = 1,44 \quad M_3 = 1,2 \quad M_4 = 5,76$
IV $\bar{x} = 4 \quad S^2 = 1,44 \quad M_3 = -1,2 \quad M_4 = 5,76$

■

Zauważ bardzo ciekawą właściwość nowo policzonych miar. M_3 przyjęło wartość 0 dla rozkładów symetrycznych, wartość dodatnią dla rozkładu prawostronnie skośnego oraz wartość ujemną dla rozkładu lewostronnie skośnego. Zależność M_4 nie jest już tak oczywista, ale zauważ że im wyższa wartość tym wyższy był najwyższy „pik" rozkładu.

Miary te niosą zatem cenną informację, ale mają istotną wadę: nie są znormalizowane, a więc nie jest możliwe porównywanie na nich rozkładów np. pokazujących daną w innej jednostce. Z tego powodu używamy miar znormalizowanych, które definiujemy poniżej.

Definicja 2.19 — Współczynnik asymetrii (skośności). Współczynnik asymetrii (skośności) definiujemy wzorem:

$$g_1 = \frac{M_3}{s^3}$$

Definicja 2.20 — Współczynnik koncentracji (kurtoza). Współczynnik koncentracji (skupienia), zwany też kurtozą definiujemy wzorem:

$$K = \frac{M_4}{s^4}$$

Dodatkowo często odejmujemy od kurtozy liczbę 3, ponieważ kurtoza dla rozkładu normalnego wynosi właśnie 3 stanowiąc ważny punkt referencyjny[5]. W związku z tym kiedy uzyskujemy wartość kurtozy powyżej 3 to mówimy o rozkładzie wyostrzonym (w stosunku do rozkładu normalnego), a kurtoza poniżej 3 wskazuje na rozkład spłaszczony.

Definicja 2.21 — Współczynnik wyostrzenia (eksces). Współczynnik wyostrzenia (spłaszczenia) lub inaczej eksces to

$$g_2 = K - 3 = \frac{M_4}{s^4} - 3$$

[5]Dowód można znaleźć na Wikipedii: `https://pl.wikipedia.org/wiki/Kurtoza`

Uwaga! Podane wzory na skośność, kurtozę czy eksces są wzorami na te statystyki w populacji. Dla obliczenia ich w próbie, podobnie jak w przypadku wariancji trzeba je dostosować i wzory nie wyglądają wtedy tak prosto patrz np. `https://en.wikipedia.org/wiki/Kurtosis#Estimators_of_population_kurtosis`

Jak pewnie podejrzewasz otrzymanie asymetrii równej dokładnie 0 na rzeczywistych danych jest prawie niemożliwe. Sposób interpretowania tego współczynnika zależy więc od analityka danych. Z tego powodu, aby wyrobić sobie intuicję warto dowiedzieć się jakich heurystyk używają doświadczeni statystycy. Np. w [5] autor interpretuje skośność większą od 1 lub mniejszą od -1 jako „rozkład mocno skośny", wartości skośności mniejsze (większe) od -0.5 (0.5) interpretuje się jako „rozkład umiarkowanie skośny". W końcu skośność pomiędzy -0.5 a 0.5 interpretuje jako „rozkład prawie symetryczny".

Z kolei interpretacja kurtozy dostarcza większej liczby problemów. Klasycznie interpretuje się ją podobnie jak to uczyniliśmy dwa akapity wyżej: większa (niż 3) kurtoza oznacza wyższy (niż rozkładu normalnego) „pik" histogramu, a kurtoza mniejsza niż 3 oznacza „pik" niższy. Jednakże w ostatnim czasie postuluje się inną (i bardziej poprawną) jej interpretację. Wyższa kurtoza oznacza, że więcej wariancji wynika z rzadkich, ekstremalnych obserwacji niż z częstych odchyłek małej wielkości. Wynika z tego, że im wyższa wartość, tym częściej rozkład produkuje dalekie od średniej obserwacje – jest więc więcej obserwacji odstających i są one bardziej ekstremalne.

Wartość kurtozy jest więc mocniej związana z ogonami rozkładu niż z jego centralnym „pikiem": jej wyższa wartość oznacza grubsze i dłuższe ogony rozkładu (często pik jest wtedy relatywnie wyższy i ostrzejszy), a jej niższa wartość oznacza krótsze i cieńsze ogony rozkładu (przez co pik jest często niższy). Jaka jest więc najniższa wartość kurtozy? Wynosi ona -2 i jest to wartość osiągana przez rozkład wyników rzutu sprawiedliwą monetą (rozkład Bernoulliego z $p = 0.5$). Taki rozkład ma dwa równe słupki – ma więc on „ogon" nieskończenie krótki i nieskończenie cienki (po prostu żadnego ogonu nie ma!), a o żadnych obserwacjach odstających nie może być mowy.

Ćwiczenie 2.4 — Policzenie wszystkich poznanych statystyk i ich interpretacja. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/02/cw-1.xls` i rozwiąż ćwiczenie. ■

2.6 *Jak obliczyć statystyki opisowe w dużych danych?

Liczenie statystyk opisowych w dużych danych może sprawiać problemy. Dla przykładu zauważmy, że naiwna implementacja średniej arytmetycznej wymaga aby dane w całości mieściły się w pamięci RAM, co często nie jest wykonalne.

Dla przykładu mamy czujniki rozmieszczone wzdłuż wybrzeża mierzące temperaturę wady, poziom fal, siłę wiatru itd. Czujników takich jest oczywiście bardzo, bardzo dużo a każdy z nich wysyła swoje pomiary do głównego serwera co kilka milisekund. Zauważ, że nawet gdybyśmy chcieli te wszystkie pomiary zapisywać na dysku twardym (nie mówmy nawet o pamięci RAM) to po nawet stosunkowo krótkim czasie (kilka dni) zabrakłoby nam miejsca.

Z tego powodu chcielibyśmy liczyć np. średnią arytmetyczną w sposób przyrostowy, a dane potraktować jako strumień danych (ang. *data stream*). Co to znaczy? To znaczy, że mamy w pamięci RAM jakiś bufor (kilka, kilkanaście zmiennych), który aktualizujemy z

każdą nadesłaną obserwacją, jednak po tej aktualizacji pomiar jest bezpowrotnie zapominany i nigdy nie możemy do niego wrócić. Po pewnym czasie np. jednym roku lub jednym dniu na żądanie użytkownika system jest w stanie na podstawie tych buforów policzyć żądaną statystykę.

▪ **Przykład 2.5** Jak w takiej sytuacji policzyć średnią? Średnią dla n elementów możemy wyrazić wzorem:

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

Rozpoczynając od $\bar{x}_0 = 0$ możemy przekształcić to we wzór rekurencyjny:

$$\bar{x}_n = \frac{(n-1)\bar{x}_{n-1} + x_n}{n}$$

Ten wzór ma jedną wadę: wymaga on od nas obliczenia $(n-1)\bar{x}_{n-1}$ co przy dużym n prawdopodobnie nie zmieści się nam w standardowej zmiennej! Przekształćmy więc dalej:

$$\begin{aligned}
\bar{x}_n &= \frac{(n-1)\bar{x}_{n-1} + x_n}{n} \\
&= \frac{(n-1)\bar{x}_{n-1}}{n} + \frac{x_n}{n} \\
&= \frac{(n-1)\bar{x}_{n-1}}{n} + \frac{x_n}{n} + \frac{\bar{x}_{n-1}}{n} - \frac{\bar{x}_{n-1}}{n} \\
&= \bar{x}_{n-1} + \frac{x_i - \bar{x}_{n-1}}{n}
\end{aligned} \tag{2.1}$$

▪

Problem 2.17 W jaki sposób można przyrostowo obliczyć wariancję?

2.6.1 Liczenie średniej arytmetycznej z szeregu rozdzielczego

WYKŁAD

Natomiast w jaki sposób możemy obliczyć w taki sposób medianę? Albo modę? Jednym ze sposób na poradzenie sobie z tym problemem jest zastosowanie szeregów rozdzielczych, które poznaliśmy na poprzednich zajęciach (a przecież ich przechowywanie w pamięci nie powinno być kosztowne). Jednak jak obliczyć średnią, wariancję, dominantę... z szeregu rozdzielczego?

Problem 2.18 Jak możemy zbudować szereg rozdzielczy przy danych przychodzących ze strumienia? Nie znamy np. rozstępu danych czy liczby przedziałów.

▪ **Przykład 2.6 — Średnia z szeregu rozdzielczego.** Rozważmy przykładowy szereg rozdzielczy przedziałowy:

Przedział	Liczność n_i
47,5-52,5	2
52,5-57,5	7
57,5-62,5	15
62,5-67,5	21
67,5-72,5	77
72,5-77,5	18
77,5-82,5	11
82,5-87,5	7
87,5-92,5	3
92,5-97,5	1

W jaki sposób możemy obliczyć średnią z tego szeregu? Oczywiście nie możemy tego zrobić dokładnie, bo nie mamy wszystkich danych. Ale możemy założyć, że dane w przedziałach mają rozkład jednostajny (jest to założenie, które najprawdopodobniej jest błędne, ale cóż taki life...). Skoro tak to średnią danych w przedziale jest środek tego przedziału. Wyznaczmy więc środki przedziałów, a zarazem średnie wartości danych w poszczególnych przedziałach:

Przedział	Liczność n_i	Środek przedziału $\dot{x}_i$
47,5-52,5	2	50
52,5-57,5	7	55
57,5-62,5	15	60
62,5-67,5	21	65
67,5-72,5	77	70
72,5-77,5	18	75
77,5-82,5	11	80
82,5-87,5	7	85
87,5-92,5	3	90
92,5-97,5	1	95

Zauważ, że w tej chwili mamy średnią grupy liczb $\dot{x}_i$ oraz liczność tej grupy n_i. Możemy więc bardzo prosto dowiedzieć się ile wynosi suma liczb w poszczególnych przedziałach[6] ($n_i\dot{x}_i$):

Przedział	Liczność n_i	Środek przedziału $\dot{x}_i$	Suma liczb w przedziale	
47,5-52,5	2	50	100	$(=2\cdot 50)$
52,5-57,5	7	55	385	$(=7\cdot 55)$
57,5-62,5	15	60	900	$(=15\cdot 60)$
62,5-67,5	21	65	1365	$(=21\cdot 65)$
67,5-72,5	77	70	5390	(=...)
72,5-77,5	18	75	1350	
77,5-82,5	11	80	880	
82,5-87,5	7	85	595	
87,5-92,5	3	90	270	
92,5-97,5	1	95	95	

[6] Wzór na średnią $\bar{x} = \frac{1}{n}\sum_i^n x_i$, a więc $\bar{x}n = \sum_i^n x_i$

Możemy więc policzyć całkowitą sumę liczb i podzielić ją przez liczbę obserwacji:

$$\bar{x} = \frac{1}{(2+7+15+...)}(100+385+900+...) = \frac{1}{162}11330 \approx 69,94$$

To co zrobiliśmy to tak naprawdę policzenie średniej ważonej. ▪

Definicja 2.22 — Średnia arytmetyczna z szeregu. Średnią arytmetyczną z szeregu rozdzielczego obliczamy jako średnią ważoną z jego środków przedziałów:

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{K} n_i \dot{x}_i$$

gdzie $n = \sum n_i$, K to liczba przedziałów, n_i to liczność i-tego przedziału, a $\dot{x}_i$ to jego środek.

Uwaga! Jest to wartość przybliżona, zakładająca jednorodny rozkład wartości w przedziale (tj. że średnia wartości w przedziale jest równa jego środkowi).

2.6.2 Liczenie wariancji z szeregu rozdzielczego

WYKŁAD

▪ **Przykład 2.7 — Wariancja z szeregu rozdzielczego.** Rozważmy przykładowy szereg rozdzielczy podany w poprzednim przykładzie.

W jaki sposób możemy obliczyć wariancję z tego szeregu? Oczywiście (znów) nie możemy tego zrobić dokładnie, bo nie mamy wszystkich danych. Ale zauważmy, że wariancja to jest delikatnie zmodyfikowana średnia arytmetyczna (dzielimy na próbce przez $n-1$) kwadratów odchyłek. A więc możemy ją obliczyć bardzo podobnie jak zwykłą średnią arytmetyczną!

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

Po pierwsze, aby obliczyć odchyłkę od średniej musimy wyznaczyć średnią, co zrobiliśmy w poprzednim przykładzie i wynosi ona $69,94$. Znów, zakładamy, że dane w przedziałach mają rozkład jednostajny, a więc odchyłkę od średniej możemy przybliżyć przez $\dot{x}_i - \bar{x}$.

Przedział	Liczność n_i	Środek przedziału $\dot{x}_i$	$\dot{x}_i - \bar{x}$	$(\dot{x}_i - \bar{x})^2$	
47,5-52,5	2	50	-19,94	397,6036	$(=(50-69,94)^2)$
52,5-57,5	7	55	-14,94	223,2036	$(=(55-69,94)^2)$
57,5-62,5	15	60	-9,94	98,8036	$(=(60-69,94)^2)$
62,5-67,5	21	65	-4,94	24,4036	$(=(65-69,94)^2)$
67,5-72,5	77	70	0,06	0,0036	(...)
72,5-77,5	18	75	5,06	25,6036	
77,5-82,5	11	80	10,06	101,2036	
82,5-87,5	7	85	15,06	226,8036	
87,5-92,5	3	90	20,06	402,4036	
92,5-97,5	1	95	25,06	628,0036	

Mając tak przyszykowaną kolumnę potrafimy policzyć z niej średnią arytmetyczną: każdą wartość mnożymy przez liczność przedziału, sumujemy a następnie dzielimy przez liczbę obserwacji (w próbce liczba obserwacji - 1).

$$S^2 = \frac{1}{(2+7+15+...)-1}(397,6036 \cdot 2 + 223,2036 \cdot 7 + ...) = \frac{1}{161}9349,3832 \approx 58,07$$

! Zauważ, że wariancje można wyrazić wzorem: $\mathbb{D}^2 X = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$. Możesz to wykorzystać do obliczenia wariancji z szeregu licząc po prostu 2 razy średnią: raz na zwykłych x, a drugi raz na x^2.

■

Definicja 2.23 — Wariancja z szeregu. Wariancję z szeregu obliczamy podobnie jak średnią z szeregu, poprzez zasotowanie średniej ważonej do odchyłek środków przedziałów od średniej:

$$S^2 \approx \frac{1}{n-1}\sum_{i=1}^{K} n_i(\dot{x}_i - \bar{x})^2$$

gdzie $n = \sum n_i$, K to liczba przedziałów, n_i to liczność i-tego przedziału, a $\dot{x}_i$ to jego środek.

Uwaga! Jest to wartość przybliżona, zakładająca jednorodny rozkład wartości w przedziale (tj. że średni kwadrat odchyłek w przedziale jest równy średniemu kwadratowi odchyłki jego środka)

Ćwiczenie 2.5 — Średnia i wariancja w szeregu rozdzielczym. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/02/cw-4.xls` i rozwiąż ćwiczenie. ■

2.6.3 Liczenie mediany z szeregu rozdzielczego

WYKŁAD

■ **Przykład 2.8 — Mediana z szeregu rozdzielczego.** Rozważmy przykładowy szereg rozdzielczy podany w poprzednim przykładzie. Ile wyniesie mediana? Znów, nie możemy jej wyznaczyć dokładnie, natomiast wiemy, że będzie ona leżała w połowie posortowanych wartości. Z poprzednich zadań wiemy, że liczność wynosi 162, więc pozycja mediany to $\frac{162}{2} = 81$. W którym przedziale leży ta wartość? Aby się tego dowiedzieć policzmy liczność skumulowaną.

Przedział	Liczność n_i	Liczność skumulowana	
47,5-52,5	2	2	
52,5-57,5	7	9	(=2+7)
57,5-62,5	15	24	(=2+7+15)
62,5-67,5	21	45	(=2+7+15+21)
67,5-72,5	77	122	(=2+7+15+21+77)
72,5-77,5	18	140	(=...)
77,5-82,5	11	151	
82,5-87,5	7	158	
87,5-92,5	3	161	
92,5-97,5	1	162	(=suma wszystkich liczności)

Patrząc na liczności skumulowane widzimy, że obserwacje x_1, x_2 należą do przedziału pierwszego (47,5-52,5), obserwacje $x_3, x_4, ..., x_9$ należą do drugiego przedziału itd. W jakim więc przedziale jest szukana mediana czyli x_{81}? W przedziale 67,5-72,5, który zawiera obserwacje $x_{45}, x_{46}, ..., x_{122}$.

Teraz, gdy wiemy już w którym przedziale jest mediana zastanówmy się w którym miejscu tego przedziału leży. Wiemy, że jest ona na 81 pozycji, a więc jest to $81-45=36$ pozycja w przedziale. Załóżmy, że dane w przedziale mają rozkład jednorodny. Zauważ, że przy takim założeniu wartość x_{81} leży dokładnie w $\frac{36}{77}$ szerokości przedziału (pozycja w przedziale dzielona przez liczność przedziału).

W związku z tym mediana jest oddalona od lewego brzegu tego przedziału o $\frac{36}{77}$ szerokości przedziału czyli $\frac{36}{77} \cdot 5 \approx 2,33$. Podsumowując:

$$x_{med} \approx 67,5 + 2,33 = 69,83$$

■

Definicja 2.24 — Mediana z szeregu. Medianę z szeregu obliczamy w sposób następujący:

1. Oblicz pozycję mediany tj. $i = \frac{n}{2}$
2. Znajdź przedział w którym jest mediana (możesz sobie pomóc patrząc na liczebność skumulowaną). Przedział ten ma indeks m, liczność n_m i szerokość h.
3. Oblicz pozycję mediany w wybranym przedziale. Jeżeli suma liczności poprzednich przedziałów wynosi $\sum_{i=1}^{m-1} n_i$ to pozycja mediany w tym przedziale to $i_{\text{w przedziale}} = \frac{n}{2} - \sum_{i=1}^{m-1} n_i$
4. Załóż rozkład jednorodny w przedziale. Mediana będzie więc równa lewemu krańcowi przedziału x_l dodać $i_{\text{w przedziale}}$ razy odległość przypadająca każdej wartości w przedziale ($\frac{h}{n_m}$).
   Podsumowując:
$$x_{med} = x_l + \frac{h}{n_m}\left(\frac{n}{2} - \sum_{i=1}^{m-1} n_i\right)$$

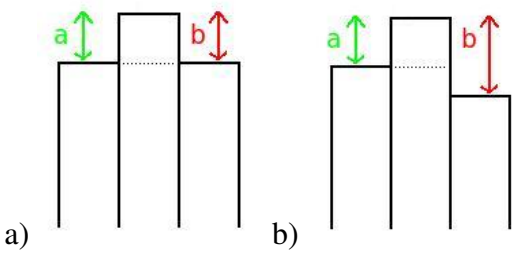
2.6.4 Liczenie dominanty z szeregu rozdzielczego

WYKŁAD

■ **Przykład 2.9 — Dominanta z szeregu rozdzielczego.** Rozważmy przykładowy szereg rozdzielczy podany w poprzednim przykładzie. Jak wyznaczyć dominantę? Po prostu sprawdźmy, który przedział ma największą liczność.

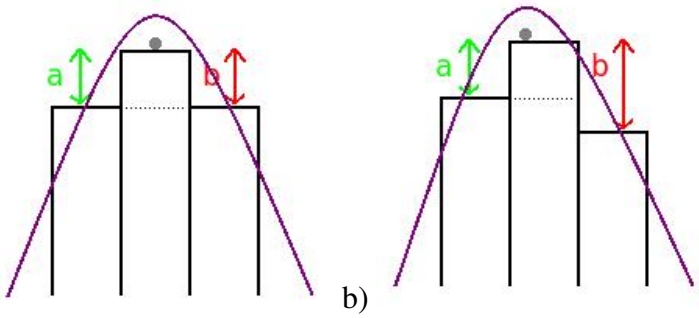
Przedział	Liczność n_i
47,5-52,5	2
52,5-57,5	7
57,5-62,5	15
62,5-67,5	21
67,5-72,5	77
72,5-77,5	18
77,5-82,5	11
82,5-87,5	7
87,5-92,5	3
92,5-97,5	1

W naszym przykładzie przedział ten to przedział $(67,5-72,5]$. Jednak jak wyznaczyć która wartość dokładnie jest dominantą/modą? W tym celu musimy spojrzeć na liczności przedziałów, które otaczają nasz wybrany przedział. Dlaczego? Jak wiemy od szeregu rozdzielczego do histogramu droga jest niedługa, więc zwizualizujmy sobie liczność naszego przedziału i przedziałów go otaczających. Ustalmy dwie sytuacje: liczności otaczających przedziałów są równe oraz jedna z otaczających przedział liczności jest większa od drugiej (w naszym szeregu mamy właśnie taki przypadek).



a) b)

Dlaczego rozważamy takie dwie sytuacje? Otóż (odpowiednio znormalizowany) histogram jest empirycznym przybliżeniem rozkładu prawdopodobieństwa badanej cechy statystycznej (zmiennej losowej). Modą funkcji gęstości jest wartość dla której przyjmuje ona największą wartość. Jak więc mogłaby wyglądać taka funkcja gęstości przybliżona tymi dwoma histogramami?



a) b)

Zauważ, że w pierwszym przypadku szczyt funkcji gęstości (czyli dominanta) leży dokładnie na środku przedziału, a w drugim leży on bardziej z jego lewej strony[7]. Miejsce to możemy wyznaczyć poprzez zastosowanie wzoru $\frac{a}{a+b}$ razy szerokość przedziału.

Dlatego, w celu wyznaczenia dominanty do lewego brzegu przedziału dodajemy $\frac{a}{a+b}$ szerokości przedziału. Zauważ, że gdy $a = b$ to trafiamy dokładnie w jego środek ($\frac{a}{a+b} = \frac{a}{2a} = 0,5$), a jeśli $a > b$ to trafiamy bardziej w lewą stronę i odpowiednio jeśli $a < b$ to bardziej w prawą stronę. Jak możesz się domyśleć z rysunku, wysokości a i b to różnice pomiędzy licznością przedziału zawierającego dominantę oraz licznościami przedziałów go otaczających. Podsumowując:

$$x_{moda} \approx 67,5 + \frac{77-21}{(77-21)+(77-18)} \cdot 5 = 67,5 + 0,49 \cdot 5 = 69,93$$

■

Definicja 2.25 — Dominanta z szeregu. Dominantę z szeregu rozdzielczego obliczamy następującym wzorem:

$$m_0 \approx x_l + \frac{n_0 - n_{-1}}{(n_0 - n_{-1}) + (n_0 - n_{+1})} h$$

gdzie n_0 to częstość przedziału klasowemu z największą częstością, x_l to jego lewy brzeg, h to jego szerokość, n_{+1} i n_{-1} to liczność przedziału następującego po nim i jego poprzedzającego.

Ćwiczenie 2.6 — Średnia, dominanta, skośność i wariancja w szeregu rozdzielczym. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/02/cw-5.xls` i rozwiąż ćwiczenie. ■

2.7 Jeszcze więcej średnich...

Definicja 2.26 — Średnia geometryczna. Średnią geometryczną wyrażamy wzorem:

$$\bar{x}_G = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Definicja 2.27 — Średnia harmoniczna. Średnią harmoniczną wyrażamy wzorem:

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

■ **Przykład 2.10 — Jaką średnią wybrać?.** Prowadzimy analizę systemu, który przewiduje prognozę pogody dla kierowców. Aby prawidłowo zmierzyć jakość takiego systemu

[7] Ponieważ liczność przedziału po prawej stronie jest dużo niższa wydaje się, że prawa część rozważanego przedziału jest rzadsza niż lewa

zdefiniowaliśmy dwa wskaźniki procentowe: skuteczność przewidywania deszczu d i skuteczność przewidywania mgły m ($m, h \in [0\%, 100\%]$). Chcielibyśmy jednak zdefiniować jeden współczynnik „jakość systemu" poprzez wyciągnięcie średniej z d i m. Którą średnią powinniśmy wybrać?

Najpierw rozważmy kilka przypadków:

- nasz system beznadziejnie przewiduje deszcz ($d = 0\%$), ale świetnie przewiduje mgłę ($m = 100\%$)

$$\bar{x} = \frac{d+m}{2} = 50\% \qquad \bar{x}_G = \sqrt{d \cdot m} = 0\% \qquad \bar{x}_H = \frac{2}{\frac{1}{m} + \frac{1}{d}} = \ldots$$

- nasz system dość dobrze przewiduje deszcz ($d = 60\%$) i trochę słabiej przewiduje mgłę ($m = 40\%$)

$$\bar{x} = \frac{d+m}{2} = 50\% \qquad \bar{x}_G = \sqrt{d \cdot m} \approx 49\% \qquad \bar{x}_H = \frac{2}{\frac{1}{m} + \frac{1}{d}} = 48\%$$

- nasz system dość dobrze przewiduje deszcz ($d = 70\%$) i słabo przewiduje mgłę ($m = 30\%$)

$$\bar{x} = \frac{d+m}{2} = 50\% \qquad \bar{x}_G = \sqrt{d \cdot m} \approx 45,8\% \qquad \bar{x}_H = \frac{2}{\frac{1}{m} + \frac{1}{d}} = 42\%$$

- nasz system dość dobrze przewiduje deszcz ($d = 80\%$) i dość dobrze przewiduje mgłę ($m = 80\%$)

$$\bar{x} = \frac{d+m}{2} = 80\% \qquad \bar{x}_G = \sqrt{d \cdot m} = 80\% \qquad \bar{x}_H = \frac{2}{\frac{1}{m} + \frac{1}{d}} = 80\%$$

Którą średnią powinniśmy wybrać? Cóż, to zależy... Czy przypadki od 1 do 3 są dla nas tak samo dobre? Jeżeli odpowiedź brzmi „nie" to nie możemy wybrać średniej arytmetycznej, która przypisała im taki sam wynik. Pozostaje nam do rozważenia średnia geometryczna i harmoniczna. Problem ze średnią harmoniczną jest oczywisty: jeżeli choć jeden ze współczynników wynosi 0 to nie możemy jej policzyć (dzielenie przez 0), więc jeżeli chcemy być odporni na taką sytuację pozostaje nam średnia geometryczna. ∎

Powyższy przykład jest też dobrą ilustracją zależności zachodzącymi między różnymi średnimi, które zdefiniujemy poniżej.

Twierdzenie 2.4 — Nierówność Cauchy'ego o średnich. Dla liczb *dodatnich* x_i średnia arytmetyczna jest większa równa średniej geometrycznej, a ta z kolei jest większa równa średniej harmonicznej[a]. Przy czym równość zachodzi tylko wtedy gdy liczby x_i są sobie równe.

$$\bar{x} \geq \bar{x}_G \geq \bar{x}_H \qquad \text{dla } x_i \geq 0$$

[a] Dowód można znaleźć na Wikipedii: `https://pl.wikipedia.org/wiki/Nier%C3%B3wno%C5%9B%C4%87_Cauchy'ego_o_%C5%9Brednich`

Ćwiczenie 2.7 — Wybór odpowiedniej średniej. Otwórz arkusz kalkulacyjny dostępny pod następującym linkiem: `https://ophelia.cs.put.poznan.pl/webdav/ad/students/02/cw-3.xls` i rozwiąż ćwiczenie. ■

Literatura

Literatura powtórkowa

Podstawowe statystyki opisowe można znaleźć w rozdziałach 1.6—1.8 darmowej książki „OpenIntro Statistics" [6]. Książka jest dostępna do ściągnięcia na stornie `www.openintro.org/stat/textbook.php`. Z literatury w języku polskim polecam pierwszy rozdział książki [4].

Literatura dla chętnych

W prasie czy telewizji dość często słyszy się pojęcia takie jak „statystyczny obywatel" albo „przeciętny człowiek". Ale kim jest ten „przeciętny człowiek"? Polecam do obejrzenia dwa 3-minutowe filmiki stworzone przez National Geographic Magazine, które wyjaśniają kim on jest, jak wygląda i czy ma telefon komórkowy?

- 7 Billion: Are You Typical? `https://www.youtube.com/watch?v=4B2xOvKFFz4`
- 7 Billion `https://www.youtube.com/watch?v=sc4HxPxNrZ0`

Pytania sprawdzające zrozumienie

Pytanie 2.1 Oblicz średnią, dominantę, medianę, wariancję, odchylenie standardowe, pierwszy i trzeci kwartyl dla podanych danych.
Pytanie 2.2 Rozpoznaj rozkład danych na podstawie podanych statystyk opisowych (np. przypisz histogram do zestawu statystyk).
Pytanie 2.3 Oblicz proste zadanie bazując na nierównościach Markowa i Czebyszewa.
Pytanie 2.4 * Oblicz średnią, dominantę, medianę, wariancję, odchylenie standardowe z szeregu rozdzielczego.

Bibliografia

[1] Statystyka opisowa. `http://ekonometria.woiz.polsl.pl/statystykaopisowa/rozklady.pdf`. Dostęp: 2016-02-10.

[2] Statystyka opisowa – Wikipedia, wolna encyklopedia. `https://pl.wikipedia.org/wiki/Statystyka_opisowa`. Dostęp: 2016-02-10.

[3] Średni kurs NBP i słów kilka o średniej trymowanej. `http://www.statystyczny.pl/sredni-kurs-nbp-i-slow-kilka-o-sredniej-trymowanej/`, 2016. Dostęp: 2016-02-10.

[4] Amir D. Aczel. *Statystyka w zarządzaniu. Pełny wykład*. PWN, 2005.

[5] M. G. Bulmer. *Principles of Statistics*. Dover, 1979.

[6] D.M. Diez, C.D. Barr, i M. Çetinkaya Rundel. *OpenIntro Statistics: Third Edition*. OpenIntro, Inc., 2015. ISBN 194345003X. URL `openintro.org`.

[7] Dennis Howitt i Duncan Cramer. *Introduction to Statistics in Psychology*. Pearson Education Limited, City, 3 edition, 2005.

[8] W. Krysicki, J. Bartos, W. Dyczka, K. Królikowska, i M. Wasilewski. *Rachunek prawdopodobienstwa i statystyka matematyczna w zadaniach. Część II: Statystyka matematyczna*. Wydawnictwo Naukowe PWN, 2002. ISBN 8301113847.

[9] Philip B. Stark. SticiGui. `http://www.stat.berkeley.edu/~stark/SticiGui`, 2013. Dostęp: 2016-02-10.

[10] H. Szostak. *Competitive performance, anxiety and perceptions of parental pressure in young tennis players*. Loughborough Universitt, 1995.