PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1117 — 2006
Dydaktyka matematyki

Ludomir M. Laudański
Politechnika Rzeszowska

JAK UCZYĆ STATYSTYKI? – PROPOZYCJA

1. Wstęp

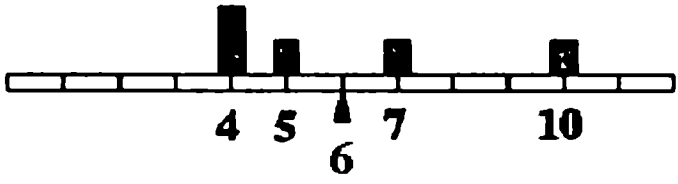
Proponowana koncepcja została opracowana na podstawie dwuczęściowego podręcznika autora zatytułowanego *Statystyka nie tylko dla licencjatów* – wydanego przez Oficynę Wydawniczą Politechniki Rzeszowskiej: cz.1 – jesienią 2004, a cz. 2 wiosną roku 2005. Przedstawiane podejście opiera się na dwu założeniach. Po pierwsze – każda z wymienionych części odpowiada ministerialnym założeniom programowym znanym w postaci tzw. minimów programowych. A więc cz. 1 – wiąże się ze statystyką opisową, a cz. 2 ze statystyką matematyczną. Posiadają one w przybliżeniu tę samą objętość i tę samą liczbę rozdziałów. Przy tym każdy rozdział o nieparzystej numeracji zawiera materiał wykładowy przeznaczony na dwie godziny lekcyjne, a następujący po nim rozdział – zadania ilustrujące materiał poprzedzającego rozdziału. Pierwsza część podręcznika zawiera 20 rozdziałów, a druga 19. Numeracja rozdziałów obu części jest ciągła dla podkreślenia ścisłej więzi pomiędzy obydwiema częściami. Drugim rysem przedkładanej propozycji jest schematyczny podział całego materiału na większe jednostki, które nazywane są statystykami pierwszego, drugiego, trzeciego i czwartego rodzaju. Rozumiemy przez te określenia, co następuje. Pierwsze dwa typy statystyk są skończenie elementowe, a tradycyjne ich nazwy to statystyki deskryptywne oraz dane zgrupowane. Za istotę postępowania uznaje się kompresję materiału statystycznego. Kompresja danych statystyk deskryptywnych sprowadza się do wyznaczenia ich miar położenia i rozproszenia. Funkcję tę spełnia przede wszystkim wartość średnia oraz wariancja. Pozwalają one na zwięzły zapis zawartości takiej statystyki. Na gruncie danych zgrupowanych ujawnia się dzięki procedurze grupowania nowy element – częstościowy obraz takiej statystyki – w graficznej postaci jest nim histogram częstości – histogram skumulowany. Przekształcenie surowych danych do postaci

zgrupowanej oznacza umiejętność wyznaczenia podstawowej średniej i wariancji (standardowego odchylenia) – a nadto obu histogramów. Na bazie histogramu skumulowanego demonstruje się operowanie percentylami. Statystyki trzeciego rodzaju stanowią analityczne przepisy unormowanych histogramów częstości – czyli tradycyjne rozkłady. Ograniczamy się do rozkładu Gaussa, dwumianowego i Poissona-Bortkiewicza. Wreszcie statystyki czwartego rodzaju to popularne rozkłady z próby: Gaussa, chi-kwadrat, *t*-Studenta i *F*-Snedecora. Na koniec dodamy, że schemat ten w sposób istotny wzbogaca się na bazie pojęcia 'wymiaru statystyk'.

Dość kuriozalne jest to, że tytułowe pojęcie 'statystyka' ani nie posiada jednej wspólnej definicji, ani nie można wskazać na jeden powszechnie przyjęty jego historyczny rodowód – mimo iż ten ważny przedmiot nauczany jest na całym świecie. Usprawiedliwia to autora niniejszego referatu, że tak trudny aspekt pozostawia poza kręgiem swoich zainteresowań. W omawianym podręczniku podjęto jednak i ten wątek, lecz uczyniono to w sposób daleki od wyczerpania. Uwagę ograniczymy do tytułowego problemu rozwijając krok po kroku wątki wskazane wyżej. Wykorzystamy podany podział tematyczny; czytelnik omawianego podręcznika jednak zauważy łatwo, że podjęte tu sprawy stanowią tylko jedną z propozycji znajdujących się na łamach omawianej książki, która naturalnie daje szersze perspektywy. W tym miejscu warto powiedzieć, że mamy na uwadze użytkowników statystyki, a nie tych, którzy ją rozwijają – a więc przede wszystkim ekonomistów, a po części także inżynierów; warto też zaznaczyć, że autor niniejszej propozycji jest inżynierem lotnictwa.

2. Statystyki pierwszego rodzaju – miary położenia i rozproszenia

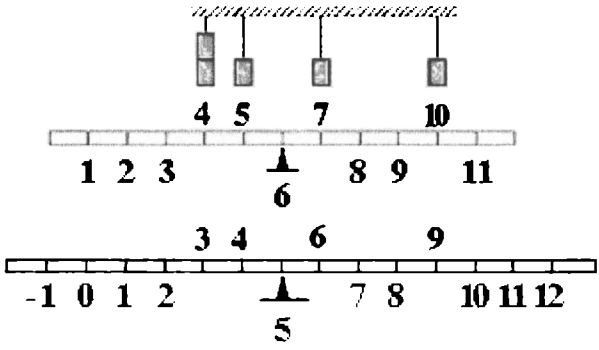
Przykładem statystyki, od której zaczynamy niniejszą propozycję, są następujące dane: 4, 4, 5, 7, 10. Na rys. 1 przedstawiamy tę statystykę graficznie, jednak w takiej postaci, która wskazuje jej miarę – położenie: w postaci podstawowej średniej – o własności nawiązującej do pojęcia równowagi dźwigni – znanej dobrze z kursu fizyki i z codziennych zastosowań.



Rys. 1. Przykład statystyki pierwszego rodzaju oraz własność podstawowej średniej

Źródło: opracowanie własne.

Ta sama analogia mechaniczna pozwala dostrzec własności wszystkich statystyk przesuniętych utworzonych na bazie statystyki wyjściowej. *Statystyka przesunięta* widoczna na rys. 2 utworzona jest przez zbiorowość liczbową: 3, 3, 4, 6, 9 – jej wartość średnią pozwala odczytać rys. 2 jako 5.



Rys. 2. Statystyka wyjściowa i statystyka przesunięta

Źródło: opracowanie własne.

Tego rodzaju podejście prowadzi w sposób naturalny do formalnej definicji wartości średniej $\overline{x}$ w postaci następującej:

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

Przekształcona lub 'przesunięta' statystyka y_i względem statystyki wyjściowej x_i pozostaje z nią w formalnym związku o postaci:

$$y_i = x_i - a .$$

Własność, jaką ukazuje rys. 2, ma postać:

$$\overline{y} = \overline{x} - a .$$

Wartość średnia kwadratowych odchyłek od podstawowej średniej jest miarą rozproszenia:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N} \quad \rightarrow \quad \sigma_x^2 = \frac{\sum_{i=1}^{N} x_i^2}{N} - \overline{x}^2 .$$

Ta nowa średnia, zwana wariancją $\sigma_x^2 = \overline{x^2} - \overline{x}^2$, wyznaczana jako różnica pomiędzy średnią kwadratową i kwadratem średniej podstawowej dostarcza przykładu prostej procedury na wyznaczenie drugiej co do hierarchii wielkości charakteryzującej każdą statystykę. Jeśli chodzi o statystyki przesunięte – ich rozproszenie pozostaje niezmienione względem statystyki wyjściowej, co również można alternatywnie odczytać z rys. 2. Podstawowa reguła teorii błędów służąca do przedstawiania wyników pomiarów:

$$\overline{x} \pm \sigma_x$$

może być rozumiana na gruncie statystyki jako prosta i pożyteczna droga kompresji danych. W odniesieniu do statystyki 4, 4, 5, 7, 10 reguła ta daje następujący rezultat:

$$6 \pm 2.28.$$

Łatwo sprawdzić, że rezultat ten pokrywa wszystkie wartości statystyki – z wyjątkiem wartości największej (czyli wartości '10'). Zamknięcie omawianego materiału stanowią pojęcia percentyli oraz *z-not*. Materiał ilustrujący pojęcie *z-noty* zawiera tab. 1, a niezbędna definicja ma postać:

$$z_i = \frac{x_i - \overline{x}}{\sigma_x}.$$

Tabela 1. Statystyka siły trzęsień Ziemi (skala Richtera)

x_i	x_i^2	*z-nota*	x_i	x_i^2	*z-nota*
3.7	13.69	–0.917887421	5.3	28.09	0.389261158
3.9	15.21	–0.754493848	5.6	31.36	0.634351517
4.2	17.64	–0.50940349	3.7	13.69	–0.917887421
7.1	50.41	1.85980331	2.9	8.41	–1.571461711
6.3	39.69	1.20622902	5.1	26.01	0.225867586
3.7	13.69	–0.917887421	5.0	25.00	0.144170799
3.8	14.44	–0.836190634	4.6	21.16	–0.182616345
4.5	20.25	–0.264313131	5.1	26.01	0.225867586
7.5	56.25	2.186590455	–	–	–
			$\sum 82$	$\sum 421$	$\sum 0$

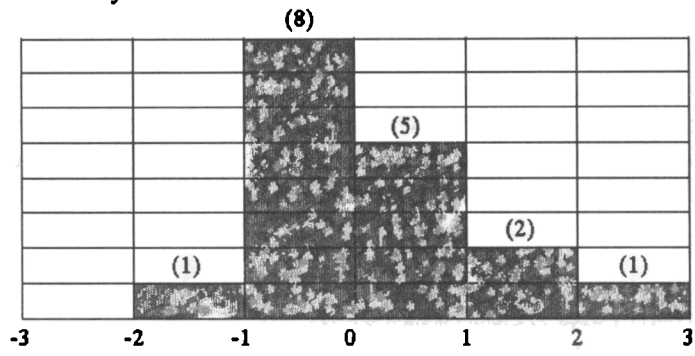
Źródło: opracowanie własne.

Statystyka *z-not* ma dwie własności: jej wartość średnia jest zerowa, a wariancja jednostkowa. Niniejsze rozważania doprowadziły nas w sposób naturalny do pojęcia histogramu częstości – jako jedynego składnika różniącego statystyki *z-not* pomiędzy sobą. Tymczasem pojęcie histogramu częstości jest podstawowym pojęciem statystyk drugiego rodzaju, których dyskusję zawiera następny paragraf. Z tego powodu – na razie – poprzestaniemy na uwadze, że rys. 3 stanowi graficzne uzupełnienie tab. 1, przedstawiając histogram częstości rozważanej statystyki. Prosimy czytelnika o wyrozumiałość, że rys. 3 znajdzie już w treści punktu 3.

3. Statystyki drugiego rodzaju

Proponujemy – aby miejsce formalnych definicji zajęło stwierdzenie, że rozważane statystyki cechują objętości wyrażające się dziesiątkami, setkami, a również i tysiącami elementów. Materiał statystyczny o tak wielkiej objętości uniemożliwia zapoznawanie się z nim w dotychczasowy sposób. Statystyki te wymaga-

ją już od pierwszego kontaktu zasadniczej kompresji – procedury takie noszą nazwę grupowania danych.

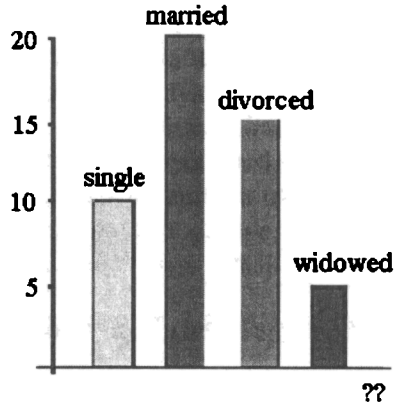


Rys. 3. Histogram częstości danych z tab. 1

Źródło: opracowanie własne.

Przy tworzenia danych zgrupowanych na podstawie tzw. surowych danych statystycznych, których sama już nazwa sugeruje konieczność ich rewaluacji, korzystamy z pewnych reguł. Zanim przedstawimy wspomniane reguły, omówimy pewien typ danych statystycznych, które proponowalibyśmy określać mianem 'zdefektowanych' (mimo że zdajemy sobie sprawę, iż nazwa taka oddaje tylko pewien ich aspekt). Do przedstawienia takich danych służą diagramy prążkowe oraz diagramy kołowe. Nie będziemy ich bliżej analizowali, ograniczając się do dwóch przykładów.

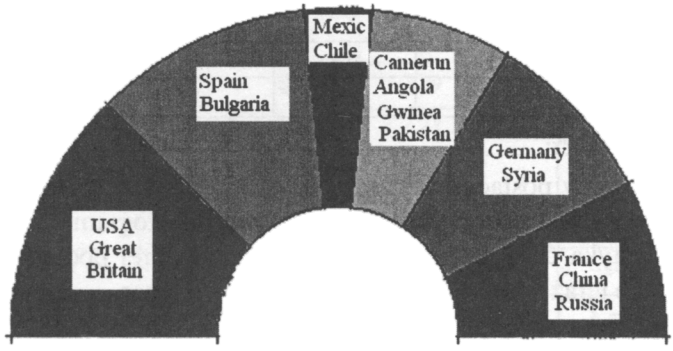
Stan cywilny załogi pewnego przedsiębiorstwa przedstawiony graficznie na rys. 4 stanowi przykład diagramu prążkowego, który – co podkreślamy – jednak nie jest histogramem częstości.



Rys. 4. Pseudohistogram częściej zwany diagramem prążkowym

Źródło: opracowanie własne.

Z kolei dane przedstawione na rys. 5 są przykładem nadużycia istoty diagramu kołowego. W danym wypadku wystarczy zauważyć, że za rezolucją w sprawie wojny w Iraku głosowały jedynie cztery kraje: USA, Wielka Brytania, Hiszpania i Bułgaria, wszystkie pozostałe nie poparły rezolucji. Nie bacząc na to, rys. 5 sugeruje równowagę głosów. Poprawne użycie takiego diagramu wymaga zachowania rzeczywistych proporcji pomiędzy polami stosownych wycinków.



Rys. 5. Diagram kołowy wyników głosowania na forum Rady Bezpieczeństwa ONZ

Źródło: opracowanie własne.

Przejdźmy do prezentacji zasadniczego materiału tego paragrafu. Stosowny przykład oparty jest na danych zamieszczonych w tab. 2.

Tabela 2. Statystyka wygenerowana za pomocą generatora RND, $N = 50$

.993	.953	.982	.835	.327	.746	.564	.039	.029	.222
.521	.704	.126	.180	.459	.055	.186	.779	.714	.768
.152	.270	.724	.165	.333	.000	.276	.987	.709	.889
.229	.443	.898	.027	.360	.397	.778	.465	.489	.298
.586	.412	.063	.628	.556	.506	.998	.825	.450	.131

Źródło: opracowanie własne.

Dane z tab. 2 poddamy procedurze grupowania, zachowując następującą kolejność:

1. Wyznaczamy zakres statystyki –

zakres = $(x_{max} - x_{min})$ w danym przypadku stanowi przedział (0, 1).

2. Określimy n liczbę przedziałów grupowania N, przestrzegając reguły:

$$2^n \leq N$$

wobec $N = 50$ – powyższy warunek spełnia $n = 5$.

3. Wyznaczamy Δ szerokość przedziałów grupowania w postaci:

$$\Delta = \frac{x_{max}}{n} \quad \rightarrow \quad \Delta = \frac{1}{5} \quad \rightarrow \quad \Delta = 0.2 .$$

4. Wyznaczamy wszystkie przedziały grupowania w postaci: [0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1.0) oraz ich punkty środkowe: 0.1, 0.3, 0.5, 0.7, 0.9. Zauważmy, że przedziały grupowania spełniają warunek lewostronnej ciągłości.

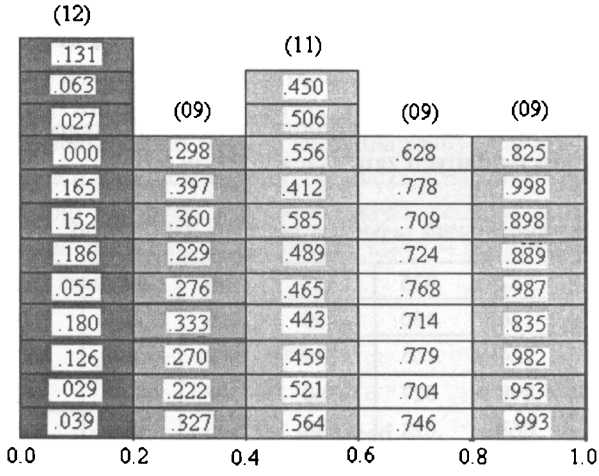
5. Histogram częstości zawiera przyporządkowanie każdemu punktowi środkowemu liczby elementów wyjściowej statystyki przypadających na odpowiadający punktowi środkowemu przedział liczbowy – zwanych częstościami (tab. 3).

Tabela 3. Dane zgrupowane statystyki z tab. 2

Punkty środkowe	0.1	0.3	0.5	0.7	0.9
Częstości	12	9	11	9	9

Źródło: opracowanie własne.

Jak to widzimy w postaci graficznej na rys. 6 – ujawniony został sekret wyjściowej statystyki – jak możemy określić otrzymany histogram częstości. Ponadto wyjściowa statystyka doznała kompresji, w wyniku której początkowy *chaos* 50 liczb reprezentuje teraz 5 uporządkowanych liczbowych par.



Rys. 6. Histogram częstości wyjściowej statystyki z tab. 2

Źródło: opracowanie własne.

Odpowiedź na pytanie, jak wyznaczyć wartość średnią i wariancję danych zgrupowanych (statystyk drugiego rodzaju), zawiera zawartość tab. 4.

Metoda bezpośrednia przedstawiona jest w postaci wyników wypełniających trzy kolumny tab. 4 – te, które następują bezpośrednio po pierwszych trzech kolumnach. W tym miejscu trzeba powiedzieć, że spotykamy tu nową definicję wartości średniej:

$$\mu_x = \frac{\sum FX}{\sum F}.$$

Stanowi ona uogólnienie definicji podanej dla statystyk pierwszego rodzaju. Wykorzystanie tej nowej definicji prowadzi do poniższych wyników:

$$\mu_x = \frac{0.1 \cdot 12 + 0.3 \cdot 9 + 0.5 \cdot 11 + 0.7 \cdot 9 + 0.9 \cdot 9}{12 + 9 + 11 + 9 + 9} = \frac{23.8}{50} = 0.476 .$$

Tabela 4. Dwie metody ewaluacji danych zgrupowanych

Przedział	X	F	FX	X^2	FX^2	U	FU	U^2	FU^2
0.8 – 1.0	0.9	9	8.1	0.81	7.29	2	18	4	36
0.6 – 0.8	0.7	9	6.3	0.49	4.41	1	9	1	9
0.4 – 0.6	0.5	11	5.5	0.25	2.75	0	0	0	0
0.2 – 0.4	0.3	9	2.7	0.09	0.81	–1	–9	1	9
0.0 – 0.2	0.1	12	1.2	0.01	0.12	–2	–24	4	48
		$\sum$50	$\sum$23.8		$\sum$15.38		$\sum$–6		$\sum$102

Źródło: opracowanie własne.

Również wariancję wyznacza się na gruncie statystyk drugiego rodzaju na podstawie nowej definicji:

$$\sigma_x^2 = \frac{\sum_{i=1}^{N} FX^2}{\sum F} - \left(\frac{\sum_{i=1}^{N} FX}{\sum F} \right)^2 ,$$

co prowadzi do wyników liczbowych jak niżej:

$$\sigma_x^2 = \frac{15.38}{50} - \left(\frac{23.8}{50} \right)^2 = 0.3076 - 0.226576 = 0.0801024 .$$

Druga metoda – zwana przez nas *metodą kodowej punktacji* – jako punkt wyjścia wykorzystuje statystykę U otrzymaną w sposób następujący:

$$U = \frac{X - R}{i}, \quad \text{gdzie:} \quad R = 0.5, \quad i = 0.2 .$$

Zawartość tab. 2 wyjaśnia pochodzenie wartości R, i. Aby otrzymać wartość średnią, należy więc skorzystać z formuły:

$$\mu = R + i \cdot \frac{\sum FU}{\sum F} .$$

Po podstawieniu wartości liczbowych prowadzi to do wyniku:

$$\mu = 0.5 - 0.2 \cdot \frac{6}{50} = 0.476 .$$

Wariancję wyznaczamy w dwu etapach. Najpierw definiujemy wariancję σ^2, którą można określić mianem *pomocniczej*:

$$\sigma_U^2 = \frac{\sum FU^2}{\sum F} - \left(\frac{\sum FU}{\sum F}\right)^2 .$$

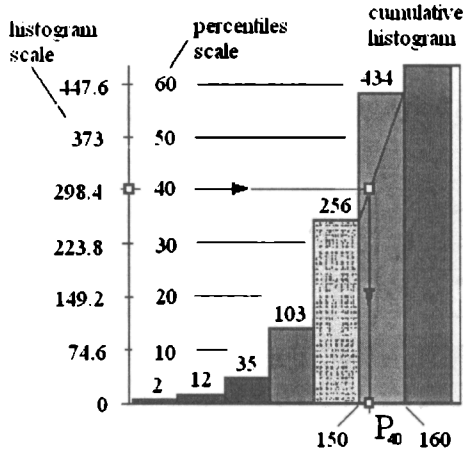
Jej wartość liczbowa wyniesie:

$$\sigma_U^2 = \frac{102}{50} - 0.456^2 = 2.0256 .$$

Poszukiwaną wariancję σ_x^2 wyznaczymy w następujący sposób:

$$\sigma_x^2 = i^2 \cdot \sigma_U^2 ,$$

co – jak łatwo sprawdzić – prowadzi do rezultatu liczbowego identycznego z otrzymanym wyżej. Zamykający składnik stanowi demonstracja, w jaki sposób adaptuje się na potrzeby statystyk drugiego rodzaju pojęcie percentyli. Ilustrujemy to na przykładzie wykorzystującym rys. 7.



Rys. 7. Wyznaczanie 40-tego percentyla

Źródło: opracowanie własne.

$$P_{40} = 150 + \frac{746 \cdot 0.4 - 256}{434 - 256}(160 - 150) = 152.382022471 .$$

4. Statystyki trzeciego rodzaju – rozkłady

4.1. Rozkład Gaussa

Przejście do statystyk trzeciego rodzaju oznacza odwołanie się do aparatu probabilistyki. Można to jednak zrobić w sposób *niejawny*. Taką *furtkę* stanowi generalizacja pojęcia histogramu częstości wykorzystująca jego *ewolucję* wraz ze wzro-

stem liczby elementów danych zgrupowanych. Prowadzi to do pojęcia 'rozkładu' jako analitycznego ujęcia 'kształtu' unormowanego histogramu częstości. Aby wprowadzić rozkład Gaussa – będący najważniejszym rozkładem statystyki – trzeba na domiar uwzględnić fakt, że jest on zdefiniowany za pomocą funkcji ciągłej:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\bar{x})^2/2\sigma^2}, \quad \text{gdzie:} \quad -\infty < x < \infty .$$

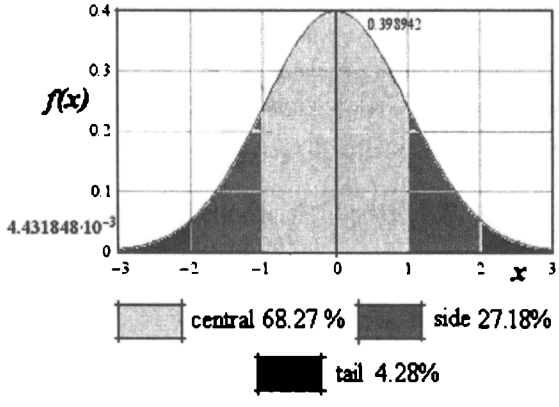
Natychmiast staje się widoczny dwuparametryczny charakter tego rozkładu, a nadto fakt, że rolę tych dwu parametrów odgrywa podstawowa średnia oraz wariancja. Tak więc, mając na względzie zastosowania, sytuacja się niejako odwraca: wielkości, których poszukiwaliśmy na gruncie obu początkowych rodzajów statystyk, tutaj muszą być wielkościami znanymi na samym początku operowania tymi statystykami. Ponadto dotychczasowe definicje podstawowej średniej i wariancji wymagają dalszego uogólnienia. Nowe definicje muszą uwzględniać 'kontynualny' charakter danych statystycznych trzeciego rodzaju. Wszystko to prowadzi do następujących definicji podstawowej średniej oraz wariancji:

$$\mu = \int_{-\infty}^{+\infty} x \cdot f(x)dx , \quad \sigma^2 = \int_{-\infty}^{+\infty} (x-\bar{x})^2 \cdot f(x)dx .$$

Definicje te łatwo się generalizuje, definiując tak zwane wyższe momenty. Sądzimy, że jest oczywiste, dlaczego jednak zatrzymujemy się w tym miejscu. W ten sposób dochodzimy – w końcowej sekwencji tej analizy – do pojęcia gaussowskich *z-notes*, których rozkład $N(0, 1)$ podaje:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

razem z graficznym obrazem tego rozkładu uwidocznionym na rys. 8.



Rys. 8. Standaryzowany rozkład Gaussa $N(0, 1)$

Źródło: opracowanie własne.

Frakcje danych statycznych opisywanych rozkładem Gaussa wymagają użycia tablic statystycznych, choć przy dzisiejszych środkach obliczeniowych mogłyby również być wyznaczane bezpośrednio – np. za pośrednictwem pakietu MathCad albo narzędzi podobnego rodzaju.

4.2. Rozkład dwumianowy

Jakkolwiek nie wydawałoby się to dziwne, następny krok – mimo iż od danych typu ciągłego przechodzimy do analizowania danych dyskretnych (aczkolwiek nieskończonej liczności) – wymaga jawnego użycia aparatu probabilistycznego. Z tego powodu właśnie w tym obszarze tematycznym jesteśmy zmuszeni przekroczyć Rubikon, wspominając słowa Cezara: *alea iacta est – kości zostały rzucone*, choć mamy jedynie na uwadze zwykły rzut monetą. Właśnie rzut monetą wiedzie do prób Bernoulliego, a one otwierają drogę do rozkładu dwumianowego – poprzez bramy kombinatoryki i definicję Laplace'a.

Powtarzające się, niezależne próby o alternatywnych wynikach cechujących się niezmiennym prawdopodobieństwem noszą nazwę prób Bernoulliego. Ich głównym przykładem są rzuty moentą. Zwyczajowo przyjmuje się, że prawdopodobieństwa obydwu wyników oznacza się symbolami p – na oznaczenie sukcesu oraz q – na oznaczenie niepowodzenia. Odtąd jeśli symbolem $b(k; n, p)$ oznaczymy prawdopodobieństwo, że w trakcie n prób Bernoulliego uzyskaliśmy k sukcesów oraz $n - k$ porażek – będzie ono wyznaczone przez rozkład dwumianowy o postaci:

$$b(k;n,p) = \binom{n}{k} \cdot p^k \cdot q^{n-k},$$

gdzie $0 < p < 1$ oraz $0 \le k \le n$.

Ponadto symbol dwumianu Newtona stanowi skrótowy zapis wyrażenia:

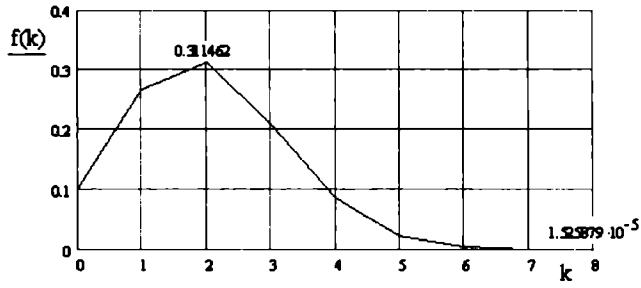
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

będącego liczbą kombinacji zdarzeń określonych mianem *sukces*. Tym razem, podając definicje średniej oraz wariancji, *cofniemy* się w ogólne porównanie z analogicznymi definicjami podanymi wyżej, pisząc:

$$\mu = \sum_{k=0}^{n} k \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = n \cdot p,$$

$$\sigma^2 = \sum_{k=0}^{n} (k - n \cdot p)^2 \cdot \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k} = n \cdot p \cdot (1-p).$$

Na rys. 9 widzimy przykład rozkładu dwumianowego, który jest powszechnie uznawany za rozkład *numer dwa* na gruncie statystyki.



Rys. 9. Rozkład dwumianowy dla $n = 8$, $p = ¼$

Źródło: opracowanie własne.

Waga i znaczenie tego rozkładu może najgłębiej uwidacznia się w parze twierdzeń de Moivre'a-Laplace'a, których tu z uwagi na skąpość miejsca już nie przytaczamy.

4.3. Rozkład Poissona-Bortkiewicza

Jedność spuścizny Gödla, Eschera i Bacha D. Hofstadter określił mianem *eternal golden braid.* Nikt nie powinien tego uznać za przesadę, jeśli trzy rozkłady statystyki – Gaussa, dwumianowy oraz Poissona-Bortkiewicza określimy podobnym mianem. Trzeci z tych rozkładów – odkryty przez S.D. Poissona, dopiero przez L. Bortkiewicza został spopularyzowany na gruncie zastosowań. Tytułowy rozkład można uzyskać z rozkładu dwumianowego podwójnym przejściem granicznym, które można zapisać w następującej niestandardowej postaci:

$$\lim\ n\,p_n \xrightarrow{n\to\infty} \lambda\,.$$

Uzyskamy na tej drodze wyjątkowo elegancką postać analityczną jednoparametrycznego rozkładu:

$$\lim\ b(k;n,p_n) \xrightarrow{n\to\infty} \frac{\lambda^k}{k!}\,e^{-\lambda}\,.$$

Nowy rozkład dyskretny ma wiele interesujących własności, z których dwie poniższe mają na gruncie statystyki znaczenie zasadnicze:

$$\overline{k} = \sum_{k=0}^{\infty} k\cdot p(k;\lambda) = \lambda\,,$$

$$\overline{k}^2 = \lambda^2 + \lambda\,, \qquad \text{a zatem} \qquad \sigma_k^2 = \lambda^2 + \lambda - \lambda^2 = \lambda\,.$$

5. Statystyki czwartego rodzaju – rozkłady z próby

Brak miejsca nie pozwala na wyczerpujące przedstawienie wspomnianych we wstępie rozkładów. Naszą uwagę poświęcimy jednemu z nich – rozkładowi średnich z próby, który jest rozkładem Gaussa. W tym celu rozważymy $x_1^{(j)}$, $x_2^{(j)}$,..., $x_N^{(j)}$

próbkę objętości N otrzymaną w sposób analogiczny do rezultatów umieszczonych w tab. 2 – czyli za pośrednictwem generatora RND liczb pseudolosowych o rozkładzie równomiernym na odcinku [0, 1). W następnym kroku utworzymy statystykę z_i-*not*, stosując następujące przekształcenie:

$$z_i = \frac{x_i - \overline{x}}{\sigma_x} = \frac{x_i - 0.5}{\sqrt{1/12}}.$$

Wykorzystaliśmy tu dobrze znaną prawidłowość wyjściowego rozkładu równomiernego statystyki x_i w postaci:

$$\overline{x} = 0.5 \quad \text{oraz} \quad \sigma_x^2 = 1/12.$$

Z kolei zdefiniujemy statystykę y_i w postaci następującej:

$$y_j = x_1^{(j)} + x_2^{(j)} + \ldots + x_N^{(j)}.$$

W ten sposób znaleźliśmy się tuż przed dokonaniem ostatniego kroku, jakim jest wykorzystanie następującej *reguły tuzina*, która prowadzi nas do rozkładu Gaussa o pożądanej postaci $N(0, 1)$:

$$\lambda_j = \sum_{i=1}^{N=12} x_i^{(j)} - 0.5.$$

Daje to nam okazję do przedstawienia jeszcze jednego zestawu *symulacyjnego* należącego do *żelaznego repertuaru* metod Monte Carlo. Końcową sekwencję numeryczną zawierają tab. 5 oraz tab. 6.

Tabela 5. Statystyka symulowana λj ($N = 12$, $M = 50$)

0.10	0.77	1.17	–0.59	–0.99	–0.45	0.40	–0.63	–0.56	–0.61
–1.46	–0.59	–0.58	–2.42	1.13	1.40	–0.98	–0.20	1.22	0.41
–0.68	1.02	–0.27	–1.09	0.26	0.20	–1.11	–0.03	0.21	1.22
–0.86	0.29	–0.10	1.04	0.12	0.08	–1.81	0.07	0.39	–0.33
–0.21	–0.80	–0.87	–0.08	–0.28	1.73	–2.04	0.83	–0.06	0.70

Źródło: opracowanie własne.

W wyniku zastosowanej procedury oczekujemy, że dane statystyczne przedstawione w tab. 5 cechuje rozkład Gaussa $N(0, 1)$. Dokonamy ich testowania po uprzednim zgrupowaniu do sześciu przedziałów o jednostkowej długości. Wynik tego grupowania znajduje się w tab. 6.

Tabela 6. Dane zgrupowane

Wartości środkowe	–2.5	–1.5	–0.5	0.5	1.5	2.5
Częstości	0.04	0.12	0.54	0.82	0.98	0

Źródło: opracowanie własne.

Ostateczny etap przedstawionej procedury jest przedstawiony na rys. 10.

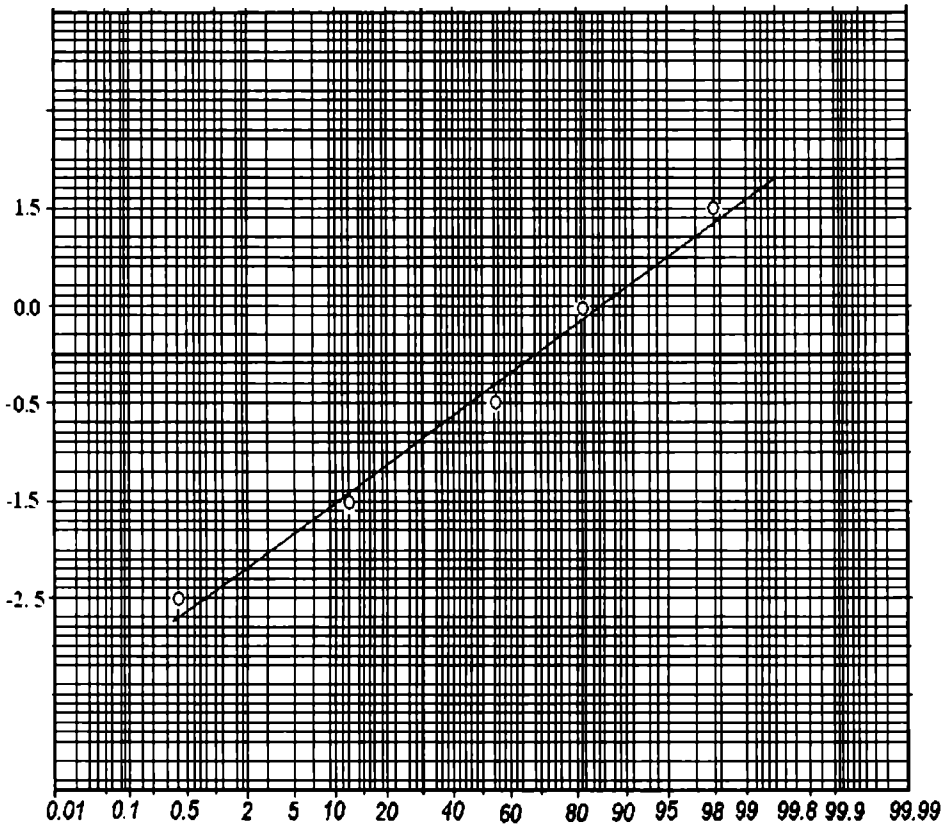
Test graficzny nie odrzucił hipotezy, że rozkład statystyki znajdującej się w tab. 5 jest rozkładem Gaussa.

Zamykając te rozważania, odnotujmy praktyczną drogę do podwyższenia dokładności symulowanego rozkładu Gaussa na podstawie reguły tuzina:

$$\lambda_j = \sum_{i=1}^{N=12} x_i^{(j)} - 0.5 .$$

Można dowieść, że biorąc $N = 48$, otrzymamy dokładniejsze odtworzenie wartości *ogonów rozkładu*. Nowy schemat symulacji przyjmie postać:

$$\zeta_j = 0.5 \sum_{i=1}^{N=48} x_i^{(j)} - 0.5 .$$



Rys. 10. Papier probabilistyczny rozkładu Gaussa

Źródło: opracowanie własne.

6. Uwagi końcowe

Jest zrozumiałe, że odpowiedź na postawione w tytule pytanie byliśmy w stanie przedstawić tutaj najzupełniej szkicowo – zważywszy, że na streszczenie ponad pół tysiąca stron omawianego podręcznika dysponowaliśmy 14 stronicami. Oralna prezentacja zawierała niektóre jeszcze aspekty czysto dydaktyczne – takie których tutaj nie znajdziemy. Dla porządku powiemy, że nie udało się nam niczego powiedzieć ani o statystykach 2-wymiarowych, ani o rozkładach 2-wymiarowych, ani o korelacji i regresji. Pozostawiliśmy bez komentarzy tematykę indeksów statystycznych, ważne rozkłady z próby, a także problematykę testowania hipotez statycznych, problemy estymacji – nie byliśmy w stanie podać żadnych własnych komentarzy odnośnie do znanych nam rynkowych pakietów softwareowych.

Literatura

C. Iuli Caesaris (MDCCCXCIII). *Belli Gallici. Libri VII cum A. Hirti Libro Octavo. In usum scholarum iterum recognovit, adiecit Galliam Antiquam tabula descriptam Bernardus Dinter.* Lipsiae. In Aedibus B.G.Teubneri.

D.R. Hofstadter (1979). *Gödel, Escher, Bach. An Eternal Golden Braid. A Metaphorical Fugue on Minds and Machines in the Spirit of Lewis Carroll.* Penguin Books. Harmondsworth.

L.M. Laudański (2004). *Statystyka nie tylko dla licencjatów. Cz.* 1. Oficyna Wydawnicza Politechniki Rzeszowskiej. Rzeszów.

L.M. Laudański (2005). *Statystyka nie tylko dla licencjatów. Cz.* 2. Oficyna Wydawnicza Politechniki Rzeszowskiej. Rzeszów.

HOW TO TEACH STATISTICS? – A PROPOSAL

Summary

The presented idea is based on a two-volume book entitled *Statistics not only for Undergraduates* which is addressed first of all to the students of economics. The first volume develops the course meant for the undergraduate students and is devoted to the so-called general statistics, that is statistics without probability. The second volume closer conforms with the popular courses of mathematical statistics. The first book contains twenty, while the second one – nineteen chapters. They are organized having in mind the assumption restraining the all odd-numbered chapters to the theoretical matters though their successors – even-numbered chapters – are devoted to the problems supporting the theory. Numbering of all the chapters goes in a single succession what underlines the formal union of both volumes. The second feature which probably differentiates this concept from the concepts popular in the literature of the subject becomes the pattern used to group the entire field into four subdivisions called *orders*. Due to this idea we commence here with the statistics of the first order called traditionally *descriptive* statistics: they always contain only a small amount of members and we may

say – all of them we can see at one glance – what even makes their compression redundant. Nevertheless, on this ground we speak about measures of the position and the measure of the dispersion – defining the basic mean and the variance. The second order statistics have a form of grouping data. Here we find the new statistical entity that is the histogram or more precisely speaking – the frequency histogram. We trace how the raw statistical data – which we may count in hundreds – cannot be seen at glance and via the procedure of grouping they loose their identity on behalf of the members of the intervals represented by the interval medium. These statistics are further analyzed by using percentiles and other similar tools. To find their mean and variance special methods should be developed – stemming from the definitions given for the first order statistics. By the third order statistics we understand distributions – the Gaussian, binomial and Poisson–Bortkiewicz. They describe the infinite number of members – countable – discrete – or continuous, uncountable. By the fourth order statistics we call the sample statistics – Gaussian, Chi-square, t-Student and finally F-Snedecor. To operate with those statistics one has to resort to the theory of probability – as particularly evaluation of the fourth-order statistics needs to use advanced tools of this theory. And in fact – such a statistical theory becomes a branch of the theory of probability – what very often becomes a *secret* of its users – not expressed in an open way, rather hidden behind as something bashful, unnamed. This concept is a vote against the impoverishment of the subject by teaching how to use fewer concepts and presenting them in depth rather than dealing with a lot of them in a shallow exposition, as suggested by computer software sold by clever and unscrupulous traders.