1. Wstęp

Aby zrozumieć myśli Boga, musimy studiować statystykę, ponieważ jest ona miarą jego celu.

Florence Nightingale[1]

Statystyka jest gramatyką nauki.

Karl Pearson[2]

Jestem pod wrażeniem elegancji definicji *statystyki* zamieszczonej w angielskiej wersji Wikipedii (*signum temporis*, nawiasem mówiąc). Określa się ją tam jako naukę „o efektywnym wykorzystywaniu danych liczbowych odnoszących się do grup osobników lub eksperymentów", obejmującą zarówno metody planowania eksperymentów, pozyskiwania danych, jak i ich opisu, analizy oraz interpretacji. Statystykę można także traktować jako pewną formę sztuki, gdyż wiele różnych decyzji jest pozostawionych samemu badaczowi.

Już od pewnego czasu matematyka (przede wszystkim statystyka) jest nowym mikroskopem biologii, ta zaś stanowi „następną fizykę" dla „królowej nauk" (Cohen 2004). Ukuto nawet odpowiednie określenia uwzględniające specyfikę metodologii, z których najpopularniejszym jest *biostatystyka*. Czy tego chcemy, czy nie – nie uciekniemy od stosowania technik z repertuaru matematyki w celu poprawy jakości opisu i pełniejszego zrozumienia praw rządzących naturą. Akceptacja takiego stanu rzeczy nie powinna być trudna, gdyż (wierzcie lub nie) na poziomie podstawowym i średnio zaawansowanym statystyka wcale nie jest specjalnie skomplikowana. Niniejszą książkę polecam przede wszystkim studentom i dokto-

[1] Florence Nightingale (1820-1910) – Angielka, twórczyni współczesnego pielęgniarstwa, pionierka technik wizualnej prezentacji danych.

[2] Karl Pearson (1857-1936) – angielski matematyk, filozof i biolog, jeden z twórców współczesnej statystyki.

rantom biologii, ochrony środowiska, medycyny i kierunków pokrewnych. Niewątpliwie będzie też źródłem przydatnej wiedzy dla pracowników nauki, gdyż prezentowane treści wykraczają w wielu miejscach poza zakres podstawowego kursu statystyki. Obecna postać tekstu różni się nieznacznie od wersji początkowej – zmiany (poprawki i uzupełnienia) wprowadzone w czerwcu 2011 r. można prześledzić na stronie http://pjadw.tripod.com/errata.htm.

Chciałbym gorąco podziękować Kasi (mojej kochanej żonie) za cierpliwość i zrozumienie. Agnieszkę przepraszam za notoryczny brak czasu; masz rację, Maleństwo – tata zbyt dużo czasu spędza przy komputerze...

Zapraszam do lektury.

2. Prawdopodobieństwo i okolice

Za każdym razem, gdy mówimy studentom: „oto czym naprawdę jest prawdopodobieństwo", jesteśmy w błędzie. Prawdopodobieństwo znaczy wiele rzeczy.

Glenn Shafer (1991)

2.1 Wprowadzenie

Stadion Narodowy, godzina 20.15. Za chwilę rozpocznie się „mecz o wszystko". Główny arbiter spotkania prosi kapitanów drużyn o podejście, po czym wyjmuje monetę. Po krótkiej wymianie zdań srebrzysty krążek zostaje wyrzucony w górę – jesteśmy świadkami... *doświadczenia losowego*. Jego rezultat zależy od przypadku (stąd nazwa), gdyż zakładamy, że moneta jest „uczciwa", podobnie zresztą jak sędzia, który będąc profesjonalistą, wprawił monetę w ruch obrotowy.

Opisane doświadczenie losowe ma tylko dwa możliwe niepodzielne wyniki, czyli mogło zajść jedno z dwóch *zdarzeń elementarnych* (ω)[3] – wyrzucony został orzeł lub reszka. W tym przypadku *przestrzeń zdarzeń elementarnych* (Ω)[4], zbiór wszystkich zdarzeń elementarnych rozpatrywanego doświadczenia losowego, jest dwuelementowa. Przestrzeń może być albo zbiorem skończonym (jak powyżej), a przynajmniej przeliczalnym (*przestrzeń skokowa*, inaczej *dyskretna*), albo zbiorem nieprzeliczalnym (*przestrzeń ciągła*).

Przykłady z przyrodniczego podwórka:

- Potomek niebieskookiej kobiety (genotyp homozygoty aa) i brązowookiego mężczyzny (genotyp heterozygoty Aa) będzie miał genotyp Aa (ω_1) albo aa (ω_2) (szanse na każdy

[3] Omega – mała litera z greckiego alfabetu.
[4] Omega – duża litera z greckiego alfabetu.

z dwóch układów są takie same) – Ω jest zbiorem skończonym, dwuelementowym.

- Poszukiwania ostatniego wspólnego przodka człowieka i szympansa: zdarzeniem elementarnym (ω_i) jest każde znalezisko budzących nadzieję szczątków, któremu przypisze się liczbę naturalną od 1 do n, gdzie n symbolizuje sukces. Uzyskany ciąg stanowi zbiór przeliczalny (w praktyce nieskończony).

- Dobowy zapis pracy serca (ω_i) ma postać funkcji ciągłej. Ω jest zbiorem nieprzeliczalnym, ponieważ istnieje nieskończenie wiele możliwych kształtów elektrokardiogramu.

Każdy podzbiór przestrzeni zdarzeń elementarnych jest określany terminem *zdarzenie losowe*[5]. Może on zawierać jeden lub większą liczbę elementów. W przypadku rzutu sześcienną kostką do gry (doświadczenie losowe) zdarzeniem losowym jest zarówno wyrzucenie trzech oczek, jak i nieparzystej liczby oczek, a także liczby oczek większej od dwóch. Zbiór zdarzeń losowych związanych z tym samym doświadczeniem losowym tworzy *rodzinę zdarzeń losowych* (S). Zdarzenia losowe mogą być: pewne, niemożliwe lub *prawdopodobne*. Z punktu widzenia statystyki interesujące są te ostatnie. Zdarzeniami losowymi zajmuje się *rachunek prawdopodobieństwa* stanowiący, bez żadnej przesady, matematyczny fundament statystyki. No dobrze, ale czym jest *prawdopodobieństwo*?

2.2 Koncepcje prawdopodobieństwa

Istnieje przynajmniej kilkanaście definicji prawdopodobieństwa, ale na szczęście nie ma potrzeby zapoznawania się z każdą z nich. W najbardziej ogólnym ujęciu prawdopodobieństwo jest matematycznym sposobem radzenia sobie z problemem niepewności. Główne koncepcje tego pojęcia można przyporządkować do dwóch kategorii: *obiektywistycznej* i *subiektywistycznej*. Pierwsza

[5]Operacje na zdarzeniach losowych są więc operacjami na zbiorach.

z nich jest najbardziej popularna i zakłada, że prawdopodobieństwo można przypisać jedynie zdarzeniom powtarzalnym (takim jak rzut kostką do gry lub monetą). Jest ona reprezentowana m.in. przez intuicyjną *definicję klasyczną* autorstwa Laplace'a[6] z 1812 r., według której prawdopodobieństwo wystąpienia zdarzenia A, czyli $P(A)$, jest równe ilorazowi liczby zdarzeń mu sprzyjających (moc zbioru A) i liczby możliwych przypadków (moc zbioru Ω). Możemy to zapisać w następujący sposób:

$$P(A) = \frac{\overline{\overline{A}}}{\overline{\overline{\Omega}}} \tag{2.2.1}$$

Zakłada się, że zdarzenia są jednakowo możliwe i wzajemnie się wykluczają.

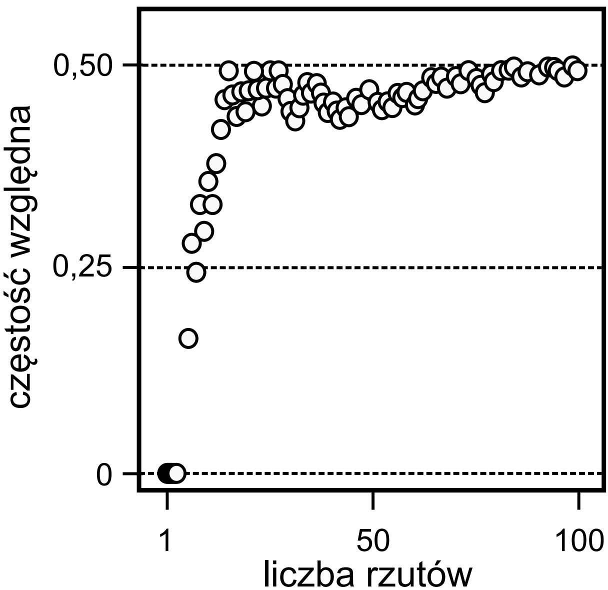
Powiedzmy, że interesuje nas prawdopodobieństwo zdarzenia polegającego na wyrzuceniu orła przy jednokrotnym rzucie symetryczną monetą. Zbiór możliwych wyników jest dwuelementowy (orzeł i reszka), zaś naszemu zdarzeniu sprzyja wyłącznie wyrzucenie orła. Po podstawieniu otrzymujemy $P(O) = 0{,}5$. Prosty problem i proste rozwiązanie. Niestety, zakres stosowalności tego podejścia ogranicza się właśnie do prostych przypadków. Główny problem z definicją klasyczną polega na tym, że wykorzystuje ona pojęcie definiowane (błąd logiczny) – „możliwe" jest synonimem „prawdopodobne". Podobną niedogodność ma *definicja geometryczna*, która za to rozwiązuje inny problem podejścia klasycznego – niemożność stosowania w sytuacji, gdy A i Ω są zbiorami nieskończonymi; liczebność tych zbiorów jest zastępowana polem powierzchni lub długością.

Definicja częstościowa von Misesa[7] (1931 r.), będąca kolejną próbą określenia, na gruncie obiektywizmu, czym jest prawdopodobieństwo, utożsamia je z granicą (*limes*) ciągu częstości. O ile racjonalizm prezentowany przez podejście Laplace'a był oparty na myśleniu w kategoriach matematyki i filozofii, o tyle koncepcja

[6] Pierre Simon de Laplace (1749–1827) – francuski matematyk, fizyk i astronom.

[7] Richard Edler von Mises (1883–1953) – amerykański matematyk urodzony we Lwowie.

częstościowa jest ze swojej natury empiryczna (oparta na obserwacjach). Wyobraźmy sobie długą serię doświadczeń losowych, polegających na rzucie symetryczną monetą. Interesuje nas prawdopodobieństwo wyrzucenia orła, więc po każdym rzucie odnotowujemy względną częstość tego zdarzenia, czyli iloraz liczby wyrzuconych do tej pory orłów i liczby rzutów. Już po wykonaniu kilkudziesięciu rzutów monetą powinniśmy zauważyć, że wspomniana wartość zbliża się do pewnej liczby. Jeśli wahania częstości zdarzenia wykazują tendencję malejącą, to liczba, ku której dążą, jest szukanym prawdopodobieństwem (ryc. 1). Oczywistą (dla matematyka) wadą tej definicji jest to, że nic nie mówi ona o warunku istnienia granicy. Ponadto po każdej serii doświadczeń (w przypadku, gdy była krótka) otrzymamy nieco inne wartości prawdopodobieństwa.



Ryc. 1. Zapis przykładowych zmian częstości wystąpienia orła w miarę wzrostu liczby wykonanych rzutów symetryczną monetą; ilustracja częstościowej definicji prawdopodobieństwa (von Misesa).

Przedstawiciele drugiego głównego nurtu, *szkoły subiektywistycznej*, utrzymują, że prawdopodobieństwo reprezentuje subiektywny osąd (miarę poziomu ufności), nie zaś obiektywnie mierzalną cechę. W związku z tym możemy je stosować **także** do zdarzeń „jednorazowych”, co nie było możliwe w przypadku stosowania koncepcji obiektywistycznej. Przykładem takiego zdarzenia jest planowana operacja konkretnego pacjenta – interesowałaby nas szansa na powodzenie **tej** operacji.

Żadna z zaprezentowanych definicji nie jest pozbawiona wad, natomiast niektóre z nich są w pewnych okolicznościach bardziej użyteczne. Korzenie dwóch dominujących obecnie szkół statystycznych, *częstościowej i bayesowskiej*[8], tkwią (upraszczając: każda osobno) w omawianych głównych koncepcjach prawdopodobieństwa. Podejściu bayesowskiemu do prawdopodobieństwa[9] poświęcony jest kolejny podrozdział.

Ostatecznie pojęcie prawdopodobieństwa zostało sformalizowane przez A. Kołmogorowa[10], który w 1933 r. podał *aksjomatykę teorii prawdopodobieństwa* (zestaw aksjomatów, czyli twierdzeń przyjmowanych bez dowodów, i definicji). Wynika z niej, że prawdopodobieństwo zdarzenia (pomijam pewne założenia) jest liczbą rzeczywistą, dla której zachodzą następujące zależności:

- Zakres wartości prawdopodobieństwa zdarzenia losowego A:

  $$0 \leqslant P(A) \leqslant 1, \tag{2.2.2}$$

  przy czym prawdopodobieństwu zdarzenia niemożliwego przyporządkowujemy 0, zaś zdarzenia pewnego 1.

- Prawdopodobieństwo zdarzenia przeciwnego:

  $$P(\bar{A}) = 1 - P(A). \tag{2.2.3}$$

[8] Thomas Bayes (1702–1761) – angielski matematyk i teolog.

[9] Nowoczesna interpretacja bayesowska (prawdopodobieństwo jako subiektywny stopień wiary w zdarzenie) powstała w latach trzydziestych XX wieku.

[10] Andriej N. Kołmogorow (1903–1987) – rosyjski matematyk.

- Prawdopodobieństwo sumy (alternatywy) np. dwóch zdarzeń:

  $$P(A{\cup}B) = P(A) + P(B) - P(A{\cap}B), \qquad (2.2.4)$$

  jeśli A i B są zdarzeniami wzajemnie się wykluczającymi, to równanie 2.2.4 przyjmuje postać:

  $$P(A{\cup}B) = P(A) + P(B). \qquad (2.2.5)$$

- Prawdopodobieństwo iloczynu (koniunkcji) np. dwóch zdarzeń:

  $$P(A{\cap}B) = P(A|B)P(B) = P(B|A)P(A), \qquad (2.2.6)$$

  jeśli zdarzenia A i B są niezależne, to równanie 2.2.6 możemy uprościć do postaci:

  $$P(A{\cap}B) = P(A)P(B). \qquad (2.2.7)$$

- Prawdopodobieństwo warunkowe zdarzenia A (jeśli zaszło zdarzenie B; przekształcone równanie 2.2.6):

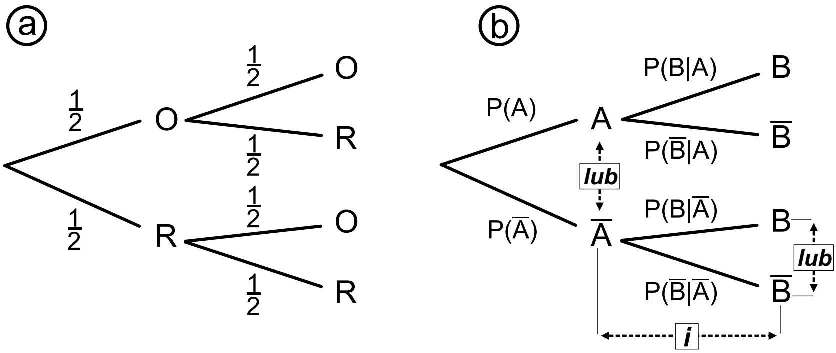
  $$P(A|B) = P(A{\cap}B) \;/\; P(B), \text{ dla } P(B) > 0, \qquad (2.2.8)$$

  ale gdy zdarzenia A i B są niezależne, to

  $$P(A|B) = P(A). \qquad (2.2.9)$$

Co ważne, propozycja Kołmogorowa jest niezależna od przyjętej przez badacza interpretacji pojęcia prawdopodobieństwa. Na koniec garść przykładów:

- Ile wynosi prawdopodobieństwo wyrzucenia orła (O) i reszki (R), w dowolnej kolejności, przy dwukrotnym rzucie monetą? Zwróćmy uwagę , że mamy tutaj do czynienia z *alternatywą dwóch wzajemnie się wykluczających koniunkcji*: (O i R) lub (R i O). Wykorzystując definicję Laplace'a (2.2.1), otrzymujemy $P(O) = P(R) = 0{,}5$. Podstawiając do równania 2.2.7 (O i R są w przypadku dwukrotnego rzutu monetą *zdarzeniami niezależnymi*), otrzymujemy $P(O \text{ i } R) = P(R \text{ i } O) = 0{,}5 \times 0{,}5 = 0{,}25$. Na koniec rozwiązujemy równanie 2.2.5: $P((O \text{ i } R) \text{ lub } (R \text{ i } O)) = 0{,}25 + 0{,}25 = 0{,}5$. Analizowaną sytuację możemy także przedstawić graficznie (ryc. 2).



Ryc. 2. Graf ilustrujący prawdopodobieństwa wyrzucenia orła i reszki przy rzucie dwiema symetrycznymi monetami (a) oraz jego forma uogólniona (b).

- Rzucamy sześcienną kostką do gry. Załóżmy, że zdarzenie A polega na wyrzuceniu pięciu oczek, zaś zdarzenie B na wyrzuceniu liczby oczek większej od trzech. Jak łatwo policzyć: $P(A) = 0{,}17$, czyli 1/6, $P(B) = 0{,}5$ (znów wykorzystaliśmy definicję Laplace'a). Jakie jest prawdopodobieństwo, że wypadło nam pięć oczek, jeśli wyrzucona liczba oczek jest większa od trzech? Ponieważ działania na zdarzeniach to działania na zbiorach (była o tym mowa), $A \cap B = A$. Korzystamy ze wzoru 2.2.8: $P(A|B) = 0{,}17 \;/\; 0{,}5 = 0{,}34$.

W obu przypadkach posiłkowaliśmy się klasyczną definicją prawdopodobieństwa, gdyż jest ona dla tak prostych sytuacji najwygodniejsza. Nic nie stoi na przeszkodzie uzyskania $P(O)$, $P(R)$,

$P(A)$ i $P(B)$ w sposób zgodny z duchem i literą definicji częstościowej.

2.3 Twierdzenie Bayesa

Twierdzenie Bayesa stanowi podstawę subiektywistycznej koncepcji prawdopodobieństwa. Stosujemy je wtedy, gdy znając wynik zdarzenia, chcemy oszacować prawdopodobieństwo możliwych przyczyn. Twierdzenie głosi, że jeśli A_1, A_2, ..., A_n są wzajemnie się wykluczającymi hipotezami, z których jedna jest prawdziwa, to

$$P(A_i|B) = KP(B|A_i)P(A_i), \tag{2.3.1}$$

gdzie K jest stałą niezależną od A, $P(A_i|\mathrm{B})$ symbolizuje prawdopodobieństwo *a posteriori* („po fakcie”, szukane) prawdziwości hipotezy A_i w świetle danych B, $P(A_i)$ – prawdopodobieństwo *a priori*, czyli *zaczątkowe* (niezależne od eksperymentu; cecha charakterystyczna dla szkoły bayesowskiej – element subiektywistyczny), $P(B|A_i)$ – prawdopodobieństwo danych w świetle hipotezy A_i (w literaturze anglojęzycznej określane też terminem *likelihood*). Proszę zwrócić uwagę na kierunkowość prawdopodobieństw warunkowych występujących we wzorze 2.3.1.

K może przyjąć postać odwrotności prawdopodobieństwa zaczątkowego zdarzenia B (danych). Jeżeli rozpatrujemy pojedynczą hipotezę, równanie 2.3.1 przyjmuje następującą postać:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \text{ dla } P(B) > 0. \tag{2.3.2}$$

Przykład

Zatoka Perska jest najważniejszym obszarem eksploatacji małży z grupy perłopławów (87% światowej „produkcji” pereł[11]). Pozyskiwane są dwie odmiany barwne pereł: różowa i białokremowa. Załóżmy, że kamienie różowe stanowią 17% pereł z tego regionu, zaś w skali globalnej udział pereł różowych wynosi 20%. Jakie jest

[11] Według Wikipedii.