Spis treści

Przedmowa

Przedstawiamy Państwu drugą część książki „Modele i metody biologii matematycznej", poświęconą modelom probabilistycznym. Obszerne fragmenty książki powstały na podstawie wykładów prowadzonych przez autora w latach 2004–2015 dla studentów i doktorantów matematyki Uniwersytetu Śląskiego w Katowicach i Uniwersytetu Jagiellońskiego w Krakowie. Na treść i kształt książki wpłynęły również wykłady prowadzone w African Institute for Mathematical Sciences w Muizenbergu w RPA w roku 2010 oraz serie wykładów podczas szkół naukowych: „Stochastic Differential Equations" w ramach szkoły „Équations Différentielles Ordinaires et Systèmes Dynamiques" w Algierze w roku 2006; „Models of Population Dynamics and Their Applications in Genetics" w ramach szkoły „From Genetics to Mathematics" w Zbąszyniu w roku 2007; „Stochastic Semigroups and their Applications in Physics and Biology" w trakcie szkoły „Evolutionary Equations with Applications in Natural Sciences" w Muizenbergu w roku 2013 oraz podczas szkół „Evolution Equations: Theory and Applications" w Besançon w roku 2015 i „From Individual Based Models to Structured Population Level Description" w Będlewie w roku 2018. Autor pragnie podziękować organizatorom wspomnianych szkół: Jackowi Banasiakowi, Mirosławowi Lachowiczowi, Jackowi Miękiszowi, Mustaphie Mokhtar-Kharroubi, Nadii El Saadi oraz dyrektorowi AIMS Fritzowi Hahne za zaproszenie do wygłoszenia wykładów.

Pragnę też podziękować studentom i doktorantom z Polski i zagranicy za cierpliwość i aktywność w czasie zajęć. Dzięki nim mogłem przetestować część prezentowanego materiału i poprawić go. Zadania znajdujące się na końcu każdego rozdziału są ściśle związane z tematyką wykładu i w sposób istotny go poszerzają. Na ogół dołączone są do nich wskazówki ułatwiające rozwiązanie. Świadomie nie zamieszczamy pełnych rozwiązań, a jedynie obszerne wskazówki, gdy rozwiązanie zadania może nastręczać trudności. Część zadań ma charakter otwarty – na przykład skonstruować model opisujący jakiś proces biologiczny. Zadania takie nie mają jednoznacznych rozwiązań i ich celem jest uaktywnienie czytelnika w zakresie doboru metod matematycznych do bada-

nia obiektów pozamatematycznych. Zachęcamy czytelnika do rozwiązywania takich zadań.

Książka przeznaczona jest przede wszystkim dla studentów matematyki oraz pracowników naukowych zainteresowanych zastosowaniami matematyki w biologii. Celem wykładu jest zaprezentowanie różnych zastosowań probabilistycznych w biologii. Ze względu na rozmiary książki pominięte zostały niektóre długie i techniczne rozumowania. Pominięte zostały też dowody znanych faktów z teorii prawdopodobieństwa i procesów stochastycznych, które można znaleźć w dostępnych książkach z tych dziedzin.

Książka podzielona jest na rozdziały według klasyfikacji matematycznej, a rozpatrywane modele biologiczne są tej klasyfikacji podporządkowane. Takie podejście może prowadzić do pewnych trudności, jeżeli czytelnik jest zainteresowany konkretnym zagadnieniem biologicznym, które może występować w kilku miejscach książki ze względu na inne metody matematyczne użyte w konstrukcji i analizie modelu. Zaletą takiego układu jest jednak fakt, że nie musimy powtarzać tych samych, często dość zaawansowanych rozumowań matematycznych. Gdy jakieś zagadnienie biologiczne, np. ekspresja genów, jest modelowane przy użyciu różnych obiektów matematycznych, starałem się podać przy jego omawianiu odnośniki do innych fragmentów książki, poświęconych podobnym problemom.

Na polskim rynku dostępnych jest kilka książek związanych z modelowaniem matematycznym w biologii, między innymi książki Murraya [270], Foryś [123], Uchmańskiego [373] oraz Czerniawskiego i innych [78]. Nie ma w nich praktycznie modeli probabilistycznych. Sporo ciekawych przykładów zastosowań rachunku prawdopodobieństwa w biologii i medycynie można znaleźć w klasycznych książkach Fellera [113] i Neymana [272]. Druga z nich zawiera również przykłady zastosowania statystyki matematycznej. Z obu książek intensywnie korzystałem przy przygotowaniu wykładów i ćwiczeń. Zamieszczone w nich modele są jednak oparte na stosunkowo prostych metodach probabilistycznych.

Naszym celem jest przedstawienie całego spektrum potencjalnych zastosowań metod probabilistycznych w biologii. Już na etapie wprowadzania podstawowych pojęć probabilistycznych, takich jak prawdopodobieństwo warunkowe czy reguła Bayesa, pojawiają się nietrywialne ich zastosowania w demografii, biologii i medycynie. Przedstawimy zastosowania tych pojęć do opisu i prognozowania wielkości populacji oraz jej struktury wiekowej, wyznaczania współczynników ryzyka oraz skuteczności testów medycznych. Będziemy się starali posługiwać danymi rzeczywistymi, zaczerpniętymi np. z *Rocznika demograficznego*. Przedstawimy również zastosowania elementów teorii prawdopodobieństwa w prostych modelach genetyki. Następnie przypomnimy pojęcia związane ze zmiennymi i wektorami losowymi oraz podstawowe twierdzenia rachunku prawdopodobieństwa. Podamy ich zastosowania do opisu

i badania funkcji przeżycia, modeli cyklu komórkowego, procesów fragmentacji i szacowania wielkości populacji. Rozdział pierwszy zakończymy rozważaniami dotyczącymi entropii i informacji oraz zastosowaniem tych pojęć w genetyce i dynamice populacyjnej.

Modele dotyczące dziedziczenia prowadzą do teorii skończonych łańcuchów Markowa, omawianych w drugim rozdziale, a te z kolei do teorii operatorów stochastycznych (Markowa) oraz półgrup stochastycznych. Dodatek A poświęcony jest prezentacji teorii operatorów i półgrup stochastycznych, ze szczególnym uzwględnieniem twierdzeń dotyczących ich asymptotyki. Z twierdzeń tych będziemy korzystać w całej książce. W ramach teorii łańcuchów Markowa przedstawimy klasyfikację stanów i twierdzenia o zbieżności do stanów stacjonarnych, quasi-stacjonarnych i pochłaniających. Wyznaczymy też średnie czasy powracania i pochłaniania. Wyniki teoretyczne zastosujemy do badania łańcuchów Markowa opisujących teorię dziedziczenia, błądzenie losowe, procesy urodzin i śmierci, rozwój epidemii w małej populacji, dryf genetyczny, sieci regulatorowe genów oraz automaty komórkowe. Dużo miejsca poświęcimy procesom gałązkowym, których badanie zapoczątkowało rozwój zastosowań metod probabilistycznych w demografii i naukach przyrodniczych. Będziemy również analizować modele ewolucji DNA, odgrywające istotną rolę w genetyce populacyjnej. Niemal we wszystkich rozdziałach książki będziemy poznawać modele ekspresji genów, kluczowe dla zrozumienia zagadnień biologii molekularnej.

W następnym rozdziale przedstawimy fragment teorii ergodycznej, która zajmuje się stochastycznymi własnościami układów dynamicznych. Przedstawimy podstawowe pojęcia tej teorii: miarę niezmienniczą, ergodyczność, mieszanie i dokładność. Podamy charakteryzacje tych pojęć, używając operatorów Frobeniusa–Perrona. Udowodnimy twierdzenie Poincarégo o powracaniu oraz lemat Kaca i przedstawimy indywidualne twierdzenie ergodyczne Birkhoffa–Chinczyna. Dowody tego twierdzenia oraz jego uogólnień dla procesów stochastycznych przedstawione są w dodatku B. Omówimy też związek między własnościami ergodycznymi i chaotycznymi układów dynamicznych. W ostatniej części rozdziału przedstawimy modele biologiczne prowadzące do układów dynamicznych i zbadamy ich własności. Będziemy się zajmować modelami rozwoju populacji, w których liczba osobników w kolejnych pokoleniach zmienia się według wzoru rekurencyjnego $x_{n+1} = S(x_n)$. Będziemy też badać modele wytwarzania erytrocytów, prowadzące do układów dynamicznych na przestrzeniach nieskończenie wymiarowych, opisujących ewolucję rozwiązań pewnych równań różniczkowych cząstkowych. W badaniu modeli biologicznych będziemy używać różnorodnych metod teorii ergodycznej: operatorów Frobeniusa–Perrona, mocnych wersji twierdzenia Kryłowa–Bogolubowa o istnieniu miary niezmienniczej dla ciągłych układów dynamicznych, a także miar gaussowskich w przestrzeniach nieskończenie wymiarowych.

W rozdziale czwartym przedstawiamy modele opisywane z użyciem kawałkami deterministycznych procesów Markowa. Procesy kawałkami deterministyczne są naturalnym uogólnieniem zarówno łańcuchów Markowa, jak i układów dynamicznych (potoków). Pojawiają się one w opisie niemal wszystkich zagadnień przyrodniczych, w których zjawiskom deterministycznym towarzyszą losowe zmiany skokowe. Rozpoczniemy od podania podstawowych informacji o procesach stochastycznych, ze szczególnym uwzględnieniem procesów Markowa. Następnie poznamy zastosowania procesów Markowa z czasem dyskretnym. Procesy tego typu można przedstawić jako iteracje stochastyczne i są one wersjami stochastycznymi wielu modeli iteracyjnych dynamiki populacyjnej. Większość zastosowań dotyczy procesów kawałkami deterministycznych z czasem ciągłym. Trajektorie takich procesów mają przeliczalną liczbę skoków w losowych momentach, a w przedziałach między skokami mają opis deterministyczny. Rozpoczniemy od stosunkowo prostych modeli czysto skokowych lub ze skokową zmianą prędkości, wykorzystywanych w opisie ruchu niektórych obiektów biologicznych oraz układów dynamicznych z losowymi skokami trajektorii, z których korzystamy w modelowaniu pojedynczych linii genealogicznych i w modelach dynamiki populacyjnej z katastrofami. Następnie przedstawimy model, w którym występuje kilka układów dynamicznych z przełączaniem dynamiki. Taki model świetnie nadaje się do opisu ekspresji genów. Przedstawimy również modele cyklu komórkowego i aktywności neuronu, w których skok następuje w momentach nielosowych, np. na brzegu obszaru, w którym proces stochastyczny jest określony. Ostatnią grupę modeli stanowią indywidualne modele strukturalne. W modelach tych każdy osobnik ma pewne cechy, np. wiek, dojrzałość w przypadku komórek czy określony fenotyp. Między osobnikami występują różnego rodzaju interakcje, które zmieniają populację, a sama populacja to zespół cech wszystkich osobników.

W rozdziale piątym zapoznamy się z równaniami stochastycznymi i ich zastosowaniami w modelach biologicznych. Przedstawimy pojęcie całki Itô oraz równania stochastycznego, omówimy związki równań stochastycznych z procesami dyfuzji i równaniami cząstkowymi, a także przedstawimy twierdzenia dotyczące asymptotyki rozwiązań i ich rozkładów. Spróbujemy też wyjaśnić, dlaczego pewne zagadnienia biologiczne warto modelować za pomocą równań stochastycznych, w jaki sposób wprowadzać stochastykę do modelu i czym się różni zaburzenie demograficzne od zaburzenia środowiskowego. Należy podkreślić, że równania stochastyczne pojawiające się w modelach biologicznych znacznie różnią się od typowych równań stochastycznych pochodzących z fizyki lub ekonomii. Na przykład po przejściu granicznym od prostego procesu urodzin i śmierci do odpowiedniego procesu dyfuzji, uzyskane równanie stochastyczne ma pierwiastkowy współczynnik dyfuzji, a więc nie spełnia on warunku Lipschitza, zwykle zakładanego w twierdzeniach o istnieniu i jednoznaczności rozwiązań. Współczynniki występujące w równaniach rosną szybciej niż

liniowo; często też rozważa się procesy dyfuzji zdegenerowanej. Pokażemy zastosowania równań stochastycznych w typowych modelach populacyjnych: drapieżca-ofiara i konkurencji, a również w modelu chemostatu i w modelach epidemiologicznych.

W ostatnim rozdziale przedstawimy bardziej zaawansowane modele matematyczne. Rozpoczynają go twierdzenia o przejściach granicznych dla procesów skokowych i dyfuzji. Takie przejścia graniczne pozwalają połączyć ze sobą różne modele podobnych procesów biologicznych. Pokazujemy np., że procesy gałązkowe używane w dyskretnym opisie wzrostu populacji po odpowiednim przejściu granicznym zmieniają się w dyfuzję Fellera. Kluczową rolę w rozdziale VI pełni fragment poświęcony granicom makroskopowym modeli indywidualnych. Oprócz granic deterministycznych mogą się pojawiać granice w postaci superprocesów, czyli procesów o wartościach w przestrzeni miar. Dość dokładnie omawiamy dwa superprocesy: Dawsona–Watanabe i Fleminga–Viota. Procesy te są również rozwiązaniami stochastycznych równań cząstkowych, i to dość specjalnej postaci. Sporą część rozdziału zajmują modele fenotypowe i modele wzrostu fitoplanktonu. Wybór tych zagadnień nie jest przypadkowy. Stanowiły one istotną część zainteresowań autora, ale pokazują również, jak różnorodne metody matematyczne można stosować przy ich badaniu oraz że otrzymywane w wyniku modelowania obiekty matematyczne mogą prowadzić do interesujących zagadnień matematycznych.

W książce nie omawiamy zastosowań statystyki matematycznej w biologii i medycynie. Współcześnie jest to tak szeroka i zaawansowana gałąź wiedzy, że próby pobieżnego lub wyrywkowego jej omówienia nie miałyby sensu.

W przygotowaniu wykładu korzystaliśmy z różnorodnej literatury. Poza wspomnianymi już książkami Fellera [113] i Neymana [272] korzystaliśmy między innymi z fragmentów książek Allena [9], Baileya [22], Borowkowa [53], Etheridge [106], Ethiera i Kurtza [107], Evansa [108], Foguela [121], Fomina, Kornfelda i Sinaja [122], Iosifescu [165], Jacoda i Sziriajewa [166], Kallenberga [180], Lasoty i Mackeya [222], Samorodnickiego [339], Sziriajewa [349], Wentzella [386] oraz oryginalnych prac cytowanych w tekście.

Miło mi podziękować moim współpracownikom Krzysztofowi Argasińskiemu, Adamowi Bobrowskiemu, Katarzynie Pichór, Marcie Tyran-Kamińskiej i Radosławowi Wieczorkowi za cenne uwagi, które przyczyniły się do ulepszenia pierwotnego tekstu. Książka powstała w wyniku realizacji projektu badawczego nr 2017/27/B/ST1/00100, finansowanego ze środków Narodowego Centrum Nauki.

Ryszard Rudnicki

I Podstawy teorii prawdopodobieństwa

1. Wstęp

1.1. Uwagi historyczne

Nowoczesna teoria prawdopodobieństwa to dziedzina o szerokim spektrum zagadnień i zaawansowanym aparacie pojęciowym. Wraz z wyrosłą na jej bazie statystyką matematyczną jest głównym narzędziem zastosowań matematyki. Nawet szybki rozwój technologii komputerowych, który umożliwił zastąpienie skomplikowanych obliczeń symulacjami numerycznymi, nie zmniejszył jej znaczenia w innych dziedzinach nauki, wręcz przeciwnie – pozwolił na wprowadzenie metod probabilistycznych do rozwiązywania zaawansowanych zagadnień analizy matematycznej, czego przykładem jest *metoda Monte Carlo* przybliżonego obliczania całek i rozwiązywania równań różniczkowych cząstkowych. Biorąc pod uwagę złożoność problematyki teorii prawdopodobieństwa i jej znaczenie dla zastosowań, czytelnik może być zaskoczony faktem, że w ramach *AMS Mathematics Subject Classification* cała matematyka podzielona jest na kilkadziesiąt sekcji, z czego zaledwie dwie dotyczą bezpośrednio probabilistyki, mianowicie sekcja 60: *teoria prawdopodobieństwa i procesy stochastyczne* i sekcja 62: *statystyka*. Przyczyną tego stanu rzeczy jest fakt, że teoria prawdopodobieństwa stała się usystematyzowaną dziedziną matematyki stosunkowo późno. Pewne próby systematyzacji rachunku prawdopodobieństwa podjęte zostały w latach dwudziestych XX wieku, głównie przez von Misesa, który wprowadził pojęcie przestrzeni próbek; dopiero jednak aksjomatyczne ujęcie przedstawione w książce Kołmogorowa z roku 1933 pozwoliło ujednolicić klasyczną teorię i przyspieszyło dalsze, bardziej zaawansowane badania.

Samo pojęcie „prawdopodobieństwa" ma znacznie dłuższą i ciekawą historię [150]. Współcześnie kojarzy się ono ze słowem „losowy" lub „niepewny", ale jego pochodzenie jest inne. Człowiek od niepamiętnych czasów miał do czy-

nienia ze zjawiskami losowymi, a nawet sam wykonywał eksperymenty, które współcześnie można nazwać losowymi. Wszelkiego rodzaju wróżby oparte były na losowości, ale uznawano ich wynik jako boską wskazówkę, legitymującą wybór właściwej decyzji. Pierwotnie słowo „prawdopodobny" oznaczało „zatwierdzony" lub „godny aprobaty". W tym swoistym determinizmie był wyjątek – już w starożytnym Egipcie bardzo dbano o to, aby kości do gry były dobrze wyważone. Świadczyć to może o tym, że już wtedy ludzie mieli podobne do nas wyobrażenie o losowości, a koncepcja zbioru zdarzeń jednakowo prawdopodobnych, która legła u podstaw współczesnej probabilistyki, nie była im obca. Zagadnienia związane z grami hazardowymi stały się główną inspiracją do rozwoju rachunku prawdopodobieństwa w XVII wieku. Warto jednak wspomnieć, że już około roku 220 rzymski prawnik i mąż stanu Ulpian opracował na potrzeby ówczesnego systemu rent tablicę, podającą oczekiwany pozostały czas życia człowieka o ustalonym wieku. Wątpliwe, aby Ulpian znalazł te zależności, korzystając z wiedzy aktuarialnej, ale jest to zagadnienie, które można rozwiązać, używając metod probabilistycznych oraz danych dotyczących rozkładu śmiertelności. Tak więc początkowy rozwój teorii prawdopodobieństwa był inspirowany nie tylko zagadnieniami z kręgu gier hazardowych, ale również problemami dotyczącymi demografii.

Nie będziemy tu szczegółowo przedstawiać historii rozwoju teorii prawdopodobieństwa, której początki związane są z nazwiskami Fermata i Pascala, a wkład w jej rozwój wniosło wielu wybitnych uczonych. Podamy jedynie kilka ważnych przykładów z historii teorii prawdopodobieństwa, związanych z zastosowaniami w biologii i medycynie. W 1662 roku John Graunt opracował na podstawie londyńskich danych dotyczących śmiertelności (*Bills of mortality*) tablice funkcji przeżycia, jak również oszacował inne wielkości związane z demografią i chorobami zakaźnymi. Następnie Christiaan Huygens wprowadził do rachunku prawdopodobieństwa wartość oczekiwaną i korzystając z obliczeń Graunta, określił średni czas życia człowieka. Daniel Bernoulli, syn Johanna i bratanek Jacoba, autora pierwszego twierdzenia granicznego, w swojej przełomowej pracy z 1766 roku [37, 91] dowodzi, że szczepienia przeciw czarnej ospie zwiększają szanse przeżycia. Bernoulli, korzystając z tablic Halleya funkcji przeżycia (patrz rozdz. I, punkt 6.1) i zakładając, że ryzyko zarażenia się ospą w ciągu jednego roku wynosi 1/8 (osoby, które przeżyły, nabierają odporności i nie mogą ponownie zachorować), oraz przyjmując, że śmiertelność wynosi 12,5%, wyznacza funkcję przeżycia, gdyby wyeliminowano zachorowalność na ospę. W pracy pojawia się też pierwszy zaawansowany model matematyczny opisujący zależność funkcji przeżycia od śmiertelności w wyniku zachorowania na ospę.

Kolejne ważne zastosowania rachunku prawdopodobieństwa w biologii dotyczą teorii procesów gałązkowych, a ich geneza związana jest z badaniem wymierania rodzin arystokratycznych. Problem zbadał po raz pierwszy francu-

ski statystyk Irénée-Jules Bienaymé w pracy z roku 1845, która uległa zapomnieniu. W roku 1873 wszechstronny brytyjski uczony Francis Galton ponownie sformułował ten problem, ale nie był w stanie go rozwiązać. W następnym roku Henry Watson opublikował niepełne rozwiązanie. Mimo iż interesowało się problemem kilku znanych matematyków, jego pełne rozwiązanie podał dopiero w roku 1930 duński matematyk Johan Steffensen. Należy tu wspomnieć protegowanego Galtona – Karla Pearsona. Ten znakomity statystyk wywarł znaczący wpływ na badania statystyczne w biologii i medycynie; był między innymi współzałożycielem czasopism *Biometrika* i *Annals of Human Genetics*.

W początkach XX wieku otwarło się nowe pole zastosowań teorii prawdopodobieństwa – genetyka. Godfrey Hardy [153] w roku 1908 udowodnił, że przy założeniu losowego łączenia się w pary, częstości genotypów w populacji diploidalnej nie zmieniają się z pokolenia na pokolenie (prawo Hardy'ego–Weinberga). Przełomowe znaczenie w rozwoju statystyki, genetyki i biologii ewolucyjnej odegrały prace Ronalda Fishera [137]. Praca [116] stworzyła podwaliny rozwoju genetyki i wyjaśniała w oparciu o model genetyczny utrzymanie różnorodności fenotypowej w populacji, a w książce [117] sformułowane zostało między innymi podstawowe prawo Fishera selekcji naturalnej: *Tempo wzrostu przystosowania dowolnego organizmu jest równe jego genetycznej zmienności*. Rozliczne zastosowania teorii prawdopodobieństwa w biologii, również te dotyczące genetyki, pojawiły się wraz z rozwojem teorii łańcuchów Markowa [165]. Jednym z najważniejszych przykładów jest tu proces urodzin i śmierci, wprowadzony przez Williama Fellera [112].

Należy jednak zaznaczyć, że używanie metod probabilistycznych w biologii miało też wielu przeciwników. Jednym z powodów było przeświadczenie o deterministycznym charakterze procesów zachodzących w przyrodzie, ugruntowane w fizyce i chemii, zanim pojawiła się mechanika statystyczna i kwantowa. Uważano, że teorie oparte na losowości nie mogą być elementem nauk ścisłych. Szczególnie atakowana, i to przez przedstawicieli skrajnie różnych światopoglądów, była naukowa teoria ewolucji, bazująca na zdarzeniach losowych. Do przeciwników należeli kreacjoniści i zwolennicy powolnej ukierunkowanej ewolucji, zwani lamarkistami (teoria transmutacji). Skrajnym przykładem odrzucenia ewolucji opartej na losowości był „łysenkizm", stworzony w ZSRR przez Trofima Łysenkę w latach trzydziestych XX wieku; odrzucał on prawa dziedziczności, przypisując nieograniczone możliwości przekształcania organizmów właściwemu doborowi środowiska. Absurdalne teorie Łysenki zostały zaaprobowane przez Stalina jako jedyne dopuszczalne i przetrwały do połowy lat sześćdziesiątych, powodując olbrzymi regres rolnictwa i biologii w ZSRR.

W drugiej połowie XX wieku, wraz z rozwojem teorii prawdopodobieństwa, znacznie szybciej rośnie liczba modeli probabilistycznych biologii matematycznej. Oprócz standardowych już modeli opartych na metodach bayesowskich

oraz łańcuchach Markowa szeroko korzysta się z procesów stochastycznych typu dyfuzyjnego, procesów kawałkami deterministycznych, a nawet tzw. superprocesów, czyli procesów o wartościach w przestrzeni miar, wygodnych w opisie zaawansowanych modeli dynamiki populacyjnej. Część modeli wprowadza się, używając stochastycznych równań różniczkowych, zarówno zwyczajnych, jak i cząstkowych, a ich badanie wymaga użycia wyrafinowanych twierdzeń granicznych i zaawansowanego aparatu analizy funkcjonalnej.

1.2. Stochastyczny charakter zjawisk przyrodniczych

Jak wspomnieliśmy, używanie metod teorii prawdopodobieństwa do opisu rzeczywistości budziło wiele kontrowersji, głównie ze względu na zbyt mało ścisły charakter tej teorii, zwłaszcza w początkowym okresie jej rozwoju. Charakterystyczny jest tu zarzut Alberta Einsteina przeciwko mechanice kwantowej: „Bóg nie gra w kości". Wypowiedź ta jest o tyle zaskakująca, że prace Einsteina przyczyniły się do wyjaśnienia stochastycznej natury procesów dyfuzji.

Okazuje się, że nawet procesy całkowicie deterministyczne mogą mieć przebiegi praktycznie nie do odróżnienia od procesów losowych. Duża część rozdziału III poświęcona jest takim rozważaniom. Przykładem są tu ciągi liczb zadane transformacją logistyczną

$$x_{n+1} = 4x_n(1 - x_n), \quad n = 1, 2, \ldots$$

(patrz rys. I.1). Łatwo zauważymy tu dużą zależność od warunku początkowego. O ile obserwując kolejne wartości tych liczb, bylibyśmy w stanie odtworzyć regułę deterministyczną określającą ten ciąg, to obserwując np. ciąg

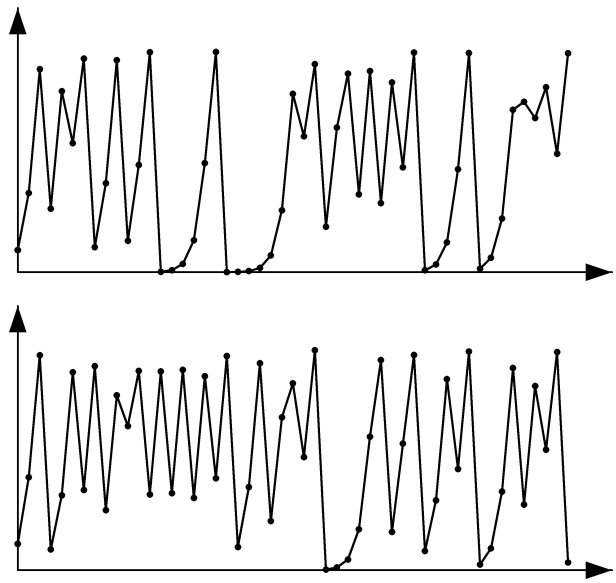
$$x_0, x_{10}, x_{20}, x_{30}, x_{40}, x_{50}, x_{60}, \ldots,$$

trudno byłoby nie tylko odtworzyć wzór na kolejne wyrazy ciągu, ale nawet odróżnić ten ciąg od wartości ciągu niezależnych zmiennych losowych ξ_n o tym samym rozkładzie i o gęstości

$$f(x) = \frac{1}{\pi\sqrt{x(1-x)}}.$$

Przykład ten pokazuje, że nawet proste i całkowicie deterministyczne procesy stają się losowe, jeżeli w zbyt długich odstępach czasu dokonujemy ich obserwacji – wtedy opis probabilistyczny staje się jak najbardziej uprawniony. Można byłoby odpowiedzieć Einsteinowi, że może i Bóg nie gra w kości, ale my często nie jesteśmy w stanie ustalić, czy proces jest losowy, czy deterministyczny.

Innym istotnym powodem wprowadzenia elementów probabilistycznych w modelach zjawisk o charakterze deterministycznym jest brak możliwości



Rysunek I.1. Kolejne iteracje transformacji $S(x) = 4x(1-x)$, startujące z $x_0 = 0{,}1$ na górnym rysunku i z $x_0 = 0{,}12$ na dolnym

w miarę pełnego opisu deterministycznego. Jeżeli na przebieg procesu wpływa wiele różnych czynników, to niektóre z nich można uznać za nieprzewidywalne i zastąpić zaburzeniem stochastycznym. Oznacza to, że w opisie oprócz części deterministycznej powinniśmy uwzględnić jeszcze zaburzenie stochastyczne (szum), które zastępuje nieznane nam składniki zjawiska. Problem dodawania szumu nie jest trywialny, zwłaszcza dla modeli opisanych równaniami różniczkowymi. Również gdy wyznaczamy wartość mierzoną w doświadczeniach eksperymentalnych, pojedynczy pomiar obarczony jest pewnym błędem i zastosowanie metod probabilistycznych (głównie centralnego twierdzenia granicznego) pozwala na dokładniejsze określenie mierzonej wielkości i oszacowanie błędu.

Prawie wszystkie procesy biologiczne można opisywać, używając zarówno modeli deterministycznych, jak i probabilistycznych (stochastycznych). Gdy np. badamy zmienność w czasie niewielkiej populacji, właściwy wydaje się model probabilistyczny, bo zmiana liczby osobników zależy od czynników losowych, jak narodziny lub śmierć osobnika. W takim modelu wielkość populacji opisana jest przez proces stochastyczny, a więc w tym wypadku funkcję zależną od czasu o wartościach w zbiorze liczb naturalnych z losowymi skokami wartości w momentach narodzin i śmierci osobników. Proces taki nazywany jest *procesem urodzin i śmierci*. Konkretny przebieg procesu, a więc gdy znamy momenty i typ skoku, nazywamy *realizacją procesu*.

Realizacja procesu nie jest wyznaczona jednoznacznie: momenty skoku, przemieszczenie i rodzaj skoku są wybierane losowo, a więc funkcja ta zależy też od nieznanego nam parametru losowego. Jeżeli zaś zastąpimy funkcję losową jej wartością oczekiwaną, a więc uśrednimy ją względem parametru losowego, otrzymamy model deterministyczny opisany przez „zwykłą" funkcję czasu, ale o wartościach w zbiorze liczb rzeczywistych nieujemnych. Można się spierać o to, czy tak otrzymany model deterministyczny ma sens, bo co to na przykład znaczy, że populacja liczy 157,63 osobników – ale takie podejście w modelowaniu jest jak najbardziej uprawnione. Na dodatek, te dwa modele mogą prowadzić do pozornie różnych wniosków biologicznych. Jeżeli np. założymy, że współczynniki urodzeń i śmierci są takie same, to w modelu deterministycznym wielkość populacji będzie stała, natomiast w modelu probabilistycznym będzie podlegać fluktuacjom, a prawie wszystkie realizacje będą dążyć do zera, gdy czas zmierza do nieskończoności. Model probabilistyczny sugeruje zatem wymieranie populacji, a deterministyczny jej stabilizację. Zauważmy jeszcze, że model probabilistyczny dostarcza więcej informacji niż model deterministyczny, bo oprócz średniej wielkości populacji możemy jeszcze podać rozkład wielkości populacji w ustalonych momentach czasu, a więc prawdopodobieństwa, że populacja liczy ustaloną liczbę osobników.

Dla dużej populacji punktem wyjścia może być model deterministyczny. Zwykle operujemy wtedy względną liczbą osobników $x(t)$. Niech $N(t)$ będzie liczbą osobników w populacji w chwili t, a N_0 pewną ustaloną liczbą naturalną rzędu $N(t)$. Względną wielkość populacji definiujemy wzorem $x(t) = N(t)/N_0$. Mimo iż formalnie funkcja $x(t)$ nie jest ciągła, zwykle można ją dobrze przybliżać funkcjami różniczkowalnymi; dlatego przyjmujemy, że funkcja $x(t)$ jest różniczkowalna oraz spełnia równanie różniczkowe postaci

$$x'(t) = \lambda(t, x(t))x(t). \tag{1.1}$$

Liczba $\lambda(t, x(t))$, zwana *współczynnikiem wzrostu*, wyraża się wzorem

$$\lambda(t, x(t)) = b(t, x(t)) - d(t, x(t)), \tag{1.2}$$

gdzie b i d nazywamy odpowiednio *współczynnikiem urodzeń* i *współczynnikiem śmiertelności* (lub *śmierci*, lub *umieralności*). Model ten różni się od wcześniejszego modelu deterministycznego jedynie skalowaniem wielkości populacji. Korzystając z modelu deterministycznego, opisanego równaniem różniczkowym (1.1), możemy budować model probabilistyczny, dodając element losowości, np. przyjmując, że współczynnik wzrostu nie jest ściśle określony dla ustalonych t i x, ale jest modyfikowany przez zaburzenie stochastyczne. W ten sposób otrzymamy model probabilistyczny opisany za pomocą równania stochastycznego (np. typu Itô).

Inny sposób otrzymania modelu wzrostu dużej populacji polega na tym, że rozpatrujemy ciąg procesów urodzin i śmierci, w którym zmieniamy skok dla n-tego procesu przy pojedynczym zdarzeniu (urodzeniu lub śmierci) np. na $1/n$ i odpowiednio modyfikujemy współczynniki urodzeń i śmiertelności. Następnie znajdujemy proces graniczny, gdy $n \to \infty$. W wyniku tej operacji, w zależności od skalowania, proces graniczny może być np. funkcją deterministyczną opisaną równaniem (1.1), ale może też być procesem stochastycznym. Widzimy więc, że praktycznie to samo zagadnienie może być modelowane na wiele sposobów, zarówno deterministycznie, jak i stochastycznie, a użycie narzędzi probabilistycznych prowadzi do ciekawych interakcji między różnymi modelami.

1.3. Zasady modelowania probabilistycznego w biologii

Podstawy modelowania matematycznego są takie same dla modeli deterministycznych i probabilistycznych; omówiliśmy je w [324]. Przypomnijmy, że modelowanie to proces wieloetapowy, obejmujący sformułowanie przesłanek biologicznych za pomocą modelu matematycznego, następnie zbadanie jego własności i biologiczną interpretację wyników, a na koniec weryfikację poprzez porównanie wyników z danymi empirycznymi i ewentualną korektę modelu. W modelu należy uwzględnić te elementy, które mają istotny wpływ na przebieg opisywanego zjawiska. Dobry model powinien dawać jasno sformułowane wnioski, tak aby można było je analizować. Zwykle rozpatrujemy zestaw modeli opisujących dane zjawisko, od najprostszych (z angielskiego zwanych *toy models*) do modeli złożonych, uwzględniających wszystkie znane nam istotne czynniki. W tej książce zajmujemy się głównie tzw. modelami modułowymi, które przedstawiają pewne elementy procesów przyrodniczych i mogą służyć do budowy modeli złożonych, w pełni opisujących dane zjawisko.

Modele probabilistyczne pojawiają się w sposób naturalny w opisie wielu zagadnień biologicznych i medycznych. Na przykład badanie skuteczności lekarstw lub testów medycznych wymaga użycia narzędzi statystycznych lub co najmniej prostych metod bayesowskich, a większość zagadnień z genetyki lub biologii molekularnej można dobrze modelować, używając łańcuchów Markowa lub bardziej skomplikowanych procesów stochastycznych. Gdy model deterministyczny słabo oddaje rzeczywisty proces, często zastępujemy go wersją probabilistyczną, dodając zaburzenie stochastyczne. W takiej sytuacji najlepiej jest, gdy samo zaburzenie jest wyprowadzone z przesłanek biologicznych. Dopuszczalna jest również prosta modyfikacja modelu przez sztuczne wprowadzenie szumu; takie postępowanie wymaga jednak pewnej ostrożności i zrozumienia natury zaburzenia, aby nie popełnić prostych błędów dyskwalifikujących model, np. prowadzących do ujemnej wielkości populacji.

2. Przestrzeń zdarzeń i prawdopodobieństwo warunkowe

2.1. Przestrzeń probabilistyczna

Rozpoczniemy od formalnej definicji przestrzeni probabilistycznej, a następnie przedstawimy pewne interpretacje wprowadzonych pojęć.

Przestrzenią probabilistyczną nazywamy trójkę $(\Omega, \Sigma, \mathrm{P})$ złożoną ze zbioru Ω zwanego *zbiorem zdarzeń elementarnych* lub *przestrzenią próbek*, σ-algebry Σ, której elementy nazywamy *zdarzeniami*, i miary probabilistycznej $\mathrm{P}\colon \Sigma \to [0,1]$, zwanej krótko *prawdopodobieństwem*. Przypominamy, że zbiór $\Sigma \subseteq 2^{\Omega}$ nazywamy *σ-algebrą* (lub *σ-ciałem*), jeżeli spełnia następujące warunki:

(i) $\emptyset \in \Sigma$,

(ii) jeżeli $A \in \Sigma$, to $X \setminus A \in \Sigma$,

(iii) dla dowolnych zbiorów $A_n \in \Sigma$, $n \geq 1$, mamy $\bigcup_{n=1}^{\infty} A_n \in \Sigma$.

Z warunków (i)–(iii) wynika, że dla dowolnych zbiorów $A_n \in \Sigma$, $n \geq 1$, mamy $\bigcap_{n=1}^{\infty} A_n \in \Sigma$. Funkcję $\mu\colon \Sigma \to [0,\infty]$ spełniającą warunki

(i) $\mu(\emptyset) = 0$,

(ii) $\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n)$, jeśli $A_n \in \Sigma$ dla $n \geq 0$ oraz $A_i \cap A_j = \emptyset$ dla $i \neq j$,

nazywamy *miarą*. Miarę P nazywamy *miarą probabilistyczną*, jeśli $\mathrm{P}(\Omega) = 1$. Przyjmujemy oznaczenie $A' = \Omega \setminus A$. Wtedy $\mathrm{P}(A') = 1 - \mathrm{P}(A)$.

Pojęcie zdarzenia elementarnego jest pojęciem pierwotnym i nie definiujemy go. Zwykle przestrzeń probabilistyczna związana jest z pewnym doświadczeniem losowym (wykonanym lub wymyślonym). Często przyjmuje się, że zbiór zdarzeń elementarnych składa się ze wszystkich możliwych wyników eksperymentu lub obserwacji [40]. Tak przyjęta definicja dobrze odpowiada eksperymentom losowym, ale też mogłaby zbyt mocno narzucać wybór zbioru Ω. Na przykład przy n-krotnym rzucie monetą, gdy interesuje nas liczba uzyskanych „orłów", automatycznie przestrzenią byłby zbiór $\{0, 1, \ldots, n\}$, ale nic nie stoi na przeszkodzie, aby przyjąć, że Ω jest zbiorem wszystkich ciągów n-elementowych „orłów" i „reszek". W praktyce możemy swobodnie badać eksperymenty losowe, używając teorii prawdopodobieństwa, bez znajomości zbioru zdarzeń elementarnych. Do definicji konkretnego zdarzenia niepotrzebna jest zwykle znajomość zbioru zdarzeń elementarnych, z których się składa; wystarczy podać, jakich wyników eksperymentu zdarzenie to dotyczy.

Intuicyjnie prawdopodobieństwo zdarzenia odpowiada częstości występowania tego zdarzenia w eksperymencie losowym. Określenie prawdopodobieństwa ustalonego zdarzenia może polegać na tym, że wielokrotnie wykonujemy doświadczenie i sprawdzamy, jak często dane zdarzenie się pojawiło; wtedy przyjmujemy, że prawdopodobieństwo zdarzenia jest równe częstości jego występowania. Zwykle jednak nie trzeba wykonywać żadnych doświadczeń losowych, aby zdefiniować prawdopodobieństwo zdarzenia. Tak jest dla pojedynczego rzutu kostką, gdzie przyjmujemy, że każdy z wyników jest jednakowo prawdopodobny i pojawia się z prawdopodobieństwem 1/6. Prawdopodobieństwa bardziej skomplikowanych zdarzeń, dotyczących np. kilkukrotnych rzutów kostką, obliczamy już, opierając się na tym założeniu i używając metod probabilistycznych.

Celem klasycznej teorii prawdopodobieństwa było wyznaczenie prawdopodobieństw pewnych zdarzeń na podstawie znajomości prawdopodobieństw innych zdarzeń [272]; podstawowe problemy zastosowań teorii prawdopodobieństwa dotyczą tego typu zagadnień. Ważną rolę w badaniu takich zagadnień odgrywa pojęcie prawdopodobieństwa warunkowego i związany z nim wzór na prawdopodobieństwo całkowite i reguła Bayesa.

2.2. Prawdopodobieństwo warunkowe i niezależność zdarzeń

Niech $(\Omega, \Sigma, \mathrm{P})$ będzie przestrzenią probabilistyczną. Niech A i B będą zdarzeniami losowymi, przy czym $\mathrm{P}(A) > 0$. Wyrażenie

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \tag{2.1}$$

nazywamy *prawdopodobieństwem warunkowym* zdarzenia B pod warunkiem A.

Korzystając z definicji prawdopodobieństwa warunkowego, można wykazać, że dla dowolnych zdarzeń $A_1, \ldots, A_n$ spełniających nierówność

$$\mathrm{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$$

zachodzi *wzór multiplikatywny*

$$\begin{aligned} &\mathrm{P}(A_1 \cap A_2 \cap \cdots \cap A_n) \\ &\quad = \mathrm{P}(A_1)\,\mathrm{P}(A_2|A_1)\,\mathrm{P}(A_3|A_1 \cap A_2) \cdots \mathrm{P}(A_n|A_1 \cap \cdots \cap A_{n-1}). \end{aligned} \tag{2.2}$$

Niech $A_1, A_2, \ldots$ będzie skończonym lub nieskończonym ciągiem zdarzeń parami rozłącznych (tj. $A_i \cap A_j = \varnothing$ dla $i \neq j$), których sumą jest cała przestrzeń, a B niech będzie dowolnym zdarzeniem. Wtedy zachodzi *wzór na*

prawdopodobieństwo całkowite

$$\mathrm{P}(B) = \sum_i \mathrm{P}(B|A_i)\,\mathrm{P}(A_i). \tag{2.3}$$

Sumowanie $\sum_i$ jest po wszystkich możliwych i, dla których $\mathrm{P}(A_i) > 0$.

Niech C będzie dowolnym zdarzeniem. Zakładamy, że $\mathrm{P}(B) > 0$ oraz $\mathrm{P}(C) > 0$. Wtedy spełniona jest *reguła Bayesa*

$$\mathrm{P}(C|B) = \frac{\mathrm{P}(B|C)\,\mathrm{P}(C)}{\sum_i \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}. \tag{2.4}$$

Aby to udowodnić, korzystamy dwukrotnie ze wzoru na prawdopodobieństwo warunkowe oraz ze wzoru na prawdopodobieństwo całkowite:

$$\mathrm{P}(C|B) = \frac{\mathrm{P}(B \cap C)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|C)\,\mathrm{P}(C)}{\sum_i \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}.$$

W szczególności dla dowolnego k mamy

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(B|A_k)\,\mathrm{P}(A_k)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B|A_k)\,\mathrm{P}(A_k)}{\sum_i \mathrm{P}(B|A_i)\,\mathrm{P}(A_i)}. \tag{2.5}$$

Wzór ten pozwala na wyznaczenie „odwrotnego" prawdopodobieństwa warunkowego i często występuje we wnioskowaniu statystycznym.

Jednym z kluczowych pojęć teorii prawdopodobieństwa jest niezależność (zdarzeń, σ-algebr, zmiennych losowych, procesów stochastycznych). Mówimy, że zdarzenia A i B są *niezależne*, jeżeli $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$.

Zauważmy, że jeżeli zdarzenia A i B mają prawdopodobieństwa dodatnie, to warunek ten jest równoważny warunkowi $\mathrm{P}(A|B) = \mathrm{P}(A)$ (lub $\mathrm{P}(B|A) = \mathrm{P}(B)$). Niezależność zdarzeń oznacza więc, że zajście zdarzenia B nie wpływa na prawdopodobieństwo zdarzenia A i na odwrót.

Niech $\{A_\lambda\}_{\lambda\in\Lambda}$ będzie dowolnym zbiorem zdarzeń. Mówimy, że zbiór $\{A_\lambda\}_{\lambda\in\Lambda}$ jest *zbiorem zdarzeń niezależnych*, jeżeli dla dowolnego skończonego ciągu indeksów $\lambda_1, \ldots, \lambda_n$ mamy

$$\mathrm{P}(A_{\lambda_1} \cap \cdots \cap A_{\lambda_n}) = \mathrm{P}(A_{\lambda_1}) \cdots \mathrm{P}(A_{\lambda_n}). \tag{2.6}$$

Warto wspomnieć, że trzy zdarzenia mogą być parami niezależne, ale nie spełniać warunku (2.6).

Uwaga I.1. Pojęcie prawdopodobieństwa warunkowego i niezależności zdarzeń można interpretować w następujący sposób. Niech $(\Omega, \Sigma, \mathrm{P})$ będzie przestrzenią probabilistyczną i niech A będzie zdarzeniem losowym, przy czym $\mathrm{P}(A) > 0$. Możemy teraz wprowadzić nową przestrzeń probabilistyczną $(A, \Sigma_A, \mathrm{P}_A)$, która

jest „obcięciem" wyjściowej przestrzeni do zbioru A, gdzie $\Sigma_A = \{A \cap B \colon B \in \Sigma\}$ oraz

$$\mathrm{P}_A(A \cap B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)}.$$

W mianowniku pojawiło się $\mathrm{P}(A)$, aby nowa przestrzeń była probabilistyczna, tj. $\mathrm{P}_A(A) = 1$. Stąd prawdopodobieństwo warunkowe $\mathrm{P}(B|A)$ wynosi $\mathrm{P}_A(A \cap B)$, jest więc prawdopodobieństwem zdarzenia B obciętego do przestrzeni $(A, \Sigma_A, \mathrm{P}_A)$. Zdarzenia A i B są niezależne, jeżeli obcięcie zdarzenia B do przestrzeni $(A, \Sigma_A, \mathrm{P}_A)$ nie zmienia jego prawdopodobieństwa.

2.3. Czynniki i dane demograficzne

W analizie różnorodnych zagadnień demograficznych, medycznych i biologicznych będziemy się starali posługiwać danymi rzeczywistymi, tak aby czytelnik mógł się przekonać o użyteczności metod matematycznych w tych naukach. Współcześnie dane takie są zwykle dostępne w internecie, ale bardzo często są one albo cząstkowe, albo już tak przetworzone, że trudno z nich bezpośrednio skorzystać. Również w takich razach będziemy się starali tak dobierać hipotetyczne dane w analizowanych problemach, aby były bliskie danym rzeczywistym.

Stosunkowo łatwo można znaleźć rozmaite dane dotyczące demografii; na przykład Główny Urząd Statystyczny publikuje obszerny *Rocznik demograficzny* [133], z którego korzystaliśmy. Część danych pochodzi ze strony internetowej GUS: http://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/ludnosc-piramida/. Dane dotyczące urodzeń i zgonów są wystarczająco dokładne i obszerne. Wśród czynników wpływających bezpośrednio na demografię jest również migracja ludności, a więc emigracja i imigracja. Dane dotyczące migracji są również dostępne w *Roczniku demograficznym*, ale z jednej strony mogą być obarczone błędem ze względu na niejasny status emigrantów i imigrantów (czy np. pobyt za granicą jest stały, czy czasowy), a co ważniejsze, trudno jest prognozować wielkość migracji, która zależy od wielu czynników wewnętrznych i zewnętrznych.

Poniżej zebraliśmy podstawowe dane demograficzne dotyczące Polski w roku 2015. Porównujemy również niektóre wskaźniki demograficzne z lat 1990 i 2015. Do porównania wybraliśmy rok 1990 ze względu na dostępność danych, a również dlatego, że 25 lat odpowiada w przybliżeniu różnicy między kolejnymi pokoleniami, a więc np. wyże i niże demograficzne występują w tych samych grupach wiekowych.

W tabeli I.1 przedstawiona jest struktura wiekowa ludności Polski (obu płci łącznie) na dzień 31 grudnia 2015 r., śmiertelność w poszczególnych przedziałach wieku oraz iloraz współczynników śmiertelności mężczyzn i kobiet.

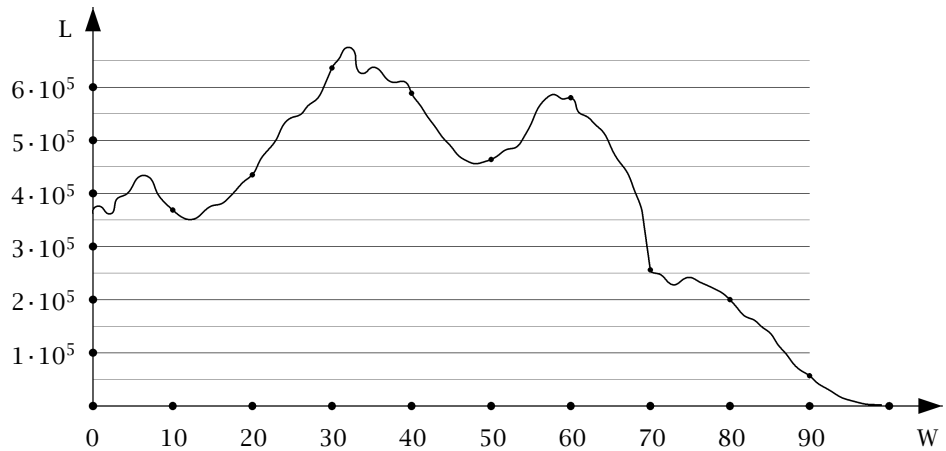
W	0–4	5–9	10–14	15–19	20–24	25–29
L	1891,6	2065,6	1797,3	1977,7	2411,3	2832,5
P	4,92	5,37	4,68	5,15	6,27	7,37
Z	400/17*	9	13	39	59	65
I	1,22/1,27	1,11	1,45	2,62	4,09	4,61
W	30–34	35–39	40–44	45–49	50–54	55–59
L	3245,5	3102,8	2730,6	2334,0	2408,3	2837,1
P	8,44	8,07	7,10	6,07	6,27	7,38
Z	85	130	206	352	581	918
I	3,8	3,23	2,86	2,72	2,64	2,5
W	60–64	65–69	70–74	75–79	80–84	85+
L	2726,5	2161,8	1208,2	1139,3	862,7	704,4
P	7,09	5,62	3,14	2,96	2,24	1,83
Z	1384	1960	2790	4249	7210	15400
I	2,37	2,28	2,1	1,85	1,53	1,18

Tabela I.1. Struktura wiekowa ludności Polski w roku 2015. Oznaczenia: W – przedział wieku, L – liczba mieszkańców w tysiącach, P – udział procentowy w całej populacji, Z – liczba zgonów na 100 tys. ludności (obu płci) w danej grupie wiekowej, I – iloraz współczynników śmiertelności mężczyzn i kobiet. *Dane dotyczące śmiertelności w przedziale wieku 0–4 zostały podzielone na pierwszy rok życia (400 zgonów w pierwszym roku życia na 100 tys. urodzeń żywych) oraz na wiek 1–4 (17 zgonów na 100 tys. dzieci w tej grupie wiekowej).

Zgodnie z danymi GUS liczba ludności wynosiła ogółem 38 mln 437 tys. W roku 2015 urodziło się 362,1 tys. dzieci, co stanowiło 0,94% całej populacji, a zmarło 394,9 tys. osób (1,03%).

Podając dane dotyczące wielu chorób występujących głównie u osób starszych, będziemy często dzielić całą populację na trzy przedziały wieku: osoby w wieku 0–44, 45–64 i 65+. W poszczególnych przedziałach było: 22 mln 55 tys. osób (57% populacji), 10 mln 306 tys. (27%) i 6 mln 76 tys. (16%).

Rysunek I.2 przedstawia wykres rozkładu wiekowego ludności Polski (bez podziału na płeć) według stanu na dzień 31 grudnia 2015 r. Rozkład wieku miał trzy wyraźne maksima lokalne, charakterystyczne dla roczników wyżu demograficznego. Maksima osiągane są dla osób w wieku 32 lat (urodzonych w roku 1983) – 675 tys. (1,8% populacji); w wieku 58 lat (urodzonych w roku 1957) – 586 tys. (1,5%) i w wieku 6 lat (urodzonych w roku 2009) – 433 tys. (1,1%). Minima lokalne występują w wieku 12 lat (urodzonych w roku 2003) – 351 tys. (0,9%) i w wieku 48 lat (urodzonych w roku 1967) – 456 tys. (1,2%).

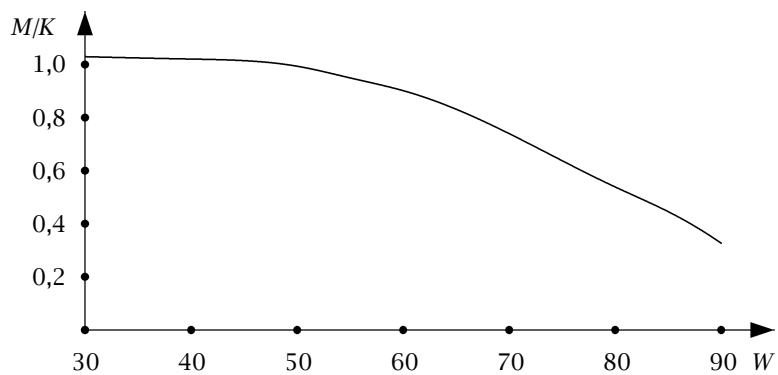


Rysunek I.2. Rozkład wieku ludności Polski w roku 2015. W oznacza wiek w latach, a L liczbę ludności w danym wieku.

Rysunek I.3. Rozkład wieku ludności Polski w roku 1990

Jak zmieniła się struktura wiekowa, można zaobserwować, porównując rozkłady z lat 1990 i 2015. Rozkład z roku 1990 przedstawiony jest na rysunku I.3. Całkowita liczba ludności w tym roku wynosiła 38 mln 144 tys. i była zbliżona do tej z roku 2015. Maksima lokalne były w wieku 7 lat – 699 tys., 35 lat – 675 tys. i 60 lat – 409 tys., a minima w wieku 23 lat – 481 tys., 48 lat – 338 tys. i 73 lat – 130 tys. Jeżeli pominiemy fluktuacje związane z występowaniem cyklu demograficznego poprzez uśrednienie rozkładu na dostatecznie długich przedziałach, to liczebność kolejnych grup wiekowych maleje liniowo, podczas gdy w rozkładzie z roku 2015 dominuje ludność w wieku 25–65 lat.

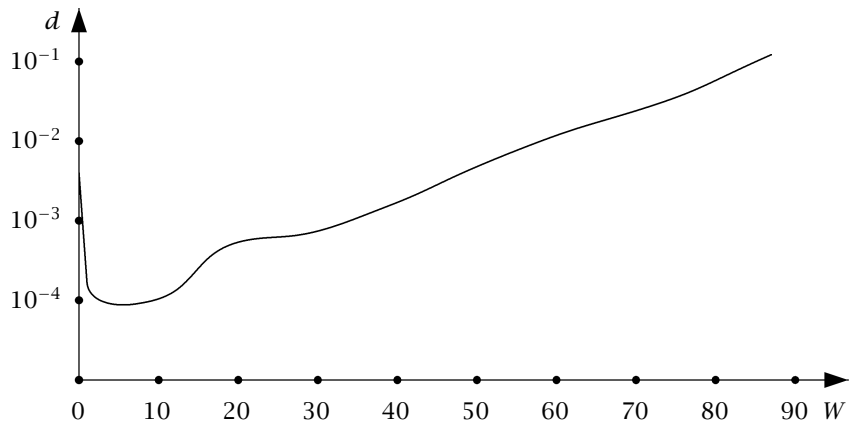
Do 59. roku życia występują niewielkie różnice w liczbie kobiet i mężczyzn. Dokładniej mówiąc, do 41. roku życia jest 51% mężczyzn, potem do 51. roku życia mamy równowagę, a następnie proporcje się odwracają. W wieku 60 lat mamy 47% mężczyzn, w wieku 70 lat 43%, 75 lat 39%, 80 lat 35%, 85 lat 31% i 90 lat tylko 25%. Dane dotyczące proporcji płci są dość ważne, pokazują bowiem istotne różnice w śmiertelności kobiet i mężczyzn. Wykres na rysunku I.4 ilustruje, jak zmienia się stosunek liczby mężczyzn do liczby kobiet w zależności od wieku, według danych z roku 2015.



Rysunek I.4. Proporcje płci w zależności od wieku

W prognozowaniu demograficznym ważną rolę odgrywa *współczynnik śmiertelności d*, wyrażony ilorazem liczby osób zmarłych w danym roku do łącznej liczby osób w danej populacji na początku roku. Współczynnik ten można obliczyć na podstawie liczby zgonów na 100 tys. ludności w danej grupie wiekowej (patrz tabela I.1). Wykres na rysunku I.5 przedstawia zależność współczynnika śmiertelności od wieku. Współczynnik śmiertelności rośnie w przybliżeniu w tempie wykładniczym, dlatego dla ilustracji tej zależności przyjęliśmy skalę logarytmiczną na osi Od. Na rysunku I.5 ograniczyliśmy się do przedziału wiekowego 0–87 lat, warto jednak dodać, że tendencja do wzrostu wykładniczego d utrzymuje się również powyżej tego wieku.[1] Na przykład w wieku 96 lat współczynnik ten wynosi około 0,25, a dla ogółu osób w wieku 100 lat i więcej – około 0,44.

[1]Wykresy wykonano zgodnie z danymi dostępnymi w tabeli 1 w następujący sposób. Dane przedstawiają średnią śmiertelność w przedziałach wieku długości 5 lat. Np. w przedziale wieku 35–39 mamy 130 zgonów w ciągu roku na 100 tys. osób. Jako współrzędną x bierzemy środek przedziału, więc $x = 37$, a jako współrzędną y bierzemy $y = \ln 130$, bo na osi Oy jest skala logarytmiczna. Tak otrzymane punkty łączymy krzywą, wykorzystując *krzywe Béziera* trzeciego stopnia i korzystając z programu do produkcji czcionek METAFONT.



Rysunek I.5. Zależność współczynnika śmiertelności d od wieku

Współczynniki śmiertelności kobiet i mężczyzn różnią się istotnie. W tabeli I.1 podany jest iloraz współczynników śmiertelności mężczyzn i kobiet dla różnych przedziałów wieku. W przedziale wieku 15–74 umiera ponad dwa razy więcej mężczyzn niż kobiet, a w przedziale 20–34 aż czterokrotnie więcej. Czytelnik może być zdziwiony, dlaczego w takim razie udziały procentowe kobiet i mężczyzn w populacji zmieniają się w niewielkim stopniu aż do pięćdziesiątego roku życia. Wiąże się to z faktem, że w świetle aktualnych danych średni współczynnik śmiertelności w przedziale wieku 1–50 wynosi w przybliżeniu 10^{-3}, co oznacza, że około 95% osób dożywa wieku 50 lat, a więc wpływ śmiertelności na zmianę proporcji obu płci w tym przedziale wieku nie jest istotny.

W tabeli I.2 porównano liczbę zgonów w latach 1990 i 2015 w różnych przedziałach wieku. Widzimy wyraźną tendencję do zmniejszenia się współ-

W	0–4	5–9	10–14	15–19	20–24	25–29
2015	400/17	9	13	39	59	65
1990	1934/59	27	29	69	104	119
W	30–34	35–39	40–44	45–49	50–54	55–59
2015	85	130	206	352	581	918
1990	174	259	398	620	928	1395
W	60–64	65–69	70–74	75–79	80–84	85+
2015	1384	1960	2790	4249	7210	15400
1990	2044	2981	4468	7282	11704	20155

Tabela I.2. Liczba zgonów na 100 tys. ludności (obu płci) w latach 1990 i 2015 w różnych przedziałach wieku

czynnika śmiertelności we wszystkich przedziałach. Podobnie jak w roku 2015, współczynnik śmiertelności w roku 1990 rośnie w tempie zbliżonym do wykładniczego, ale jego wartości są przesunięte w czasie 6–7 lat (np. w roku 1990 współczynnik śmiertelności dla wieku 40 lat miał zbliżoną wartość do tej w roku 2015 dla wieku 47 lat). Porównując ilorazy współczynnika śmiertelności mężczyzn i kobiet w różnych przedziałach wieku w latach 1990 i 2015, można zauważyć pewne fluktuacje losowe, ale bez wyraźnej tendencji.

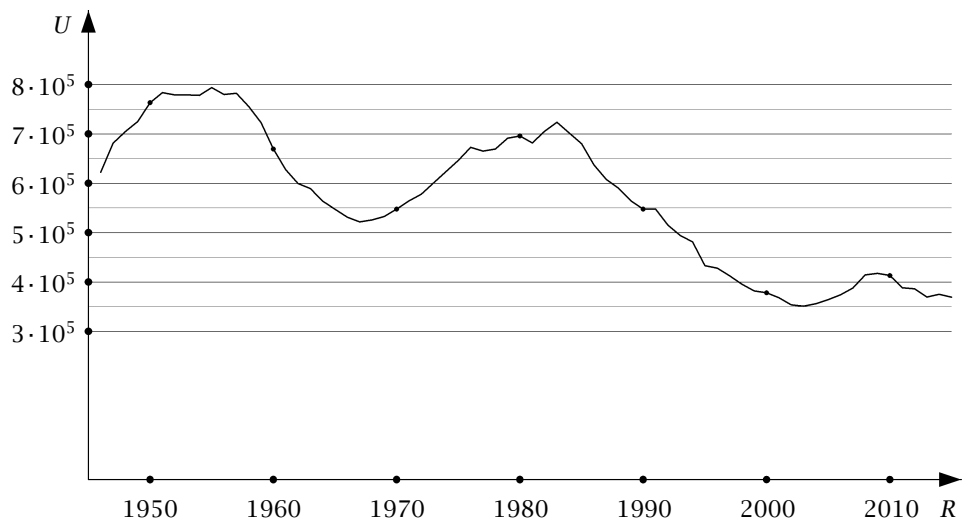
Uwaga I.2 (Superstulatkowie – maksymalna długość życia człowieka)**.** Jak już wspomnieliśmy, wyznaczony przez nas współczynnik śmiertelności wynosił 0,25 w wieku 96 lat, a dla ogółu osób w wieku 100 lat i więcej – około 0,44. Nasuwają się dwa pytania: jak będzie się zmieniał współczynnik śmiertelności dla osób powyżej stu lat i czy istnieje granica długości życia człowieka? Dożycie wieku 100 lat nie jest już czymś wyjątkowym. W Polsce, kraju stosunkowo młodym demograficznie, mamy około 3,5 tys. osób, które ukończyły 100 lat. Według danych ONZ aktualnie na świecie żyje 573 tys. stulatków, z czego tylko w USA 97 tys. i w Japonii 79 tys. (1 osoba na 2 tys. mieszkańców Japonii). Gerontolodzy szacują, że tzw. *superstulatków*, a więc osób, które mają 110 lat i więcej, żyje na świecie od 300 do 450, jest to więc bardzo nieliczna grupa wśród stulatków. Nie ma udokumentowanego przypadku osoby, która żyłaby dłużej niż biblijne 120 lat.

Ze względu na to, że grupa superstulatków jest stosunkowo nieliczna, wyznaczenie w pełni wiarygodnego współczynnika śmiertelności w tej grupie wiekowej jest niemożliwe, choć pewne szacunkowe wartości można podać [28, 132]. Należy zwrócić uwagę na fakt, że do obliczenia współczynnika śmiertelności używaliśmy modelu z czasem dyskretnym, a więc d mówi nam, jaka część osób w danym wieku zmarła w ciągu roku. Często współczynnik śmiertelności μ definiujemy na podstawie modelu z czasem ciągłym, przyjmując, że wielkość populacji maleje według wzoru $x' = -\mu x$; jeśli więc współczynnik μ jest stały, to $d = 1 - e^{-\mu}$ jest prawdopodobieństwem śmierci w ciągu jednego roku, a $1/\mu$ to oczekiwana średnia długość życia. Dla stulatków współczynniki d i μ są stosunkowo duże i nie obowiązuje tu wzór przybliżony $\mu \approx d$, dlatego nie można ich używać wymiennie. Według różnych szacunków od 0,15% do 0,25% stulatków dożywa wieku 110 lat, co odpowiada współczynnikom śmiertelności $\mu = 0{,}62$ i $d = 0{,}45$.

W przedziale wieku 30–90 oba współczynniki rosną wraz z wiekiem w przybliżeniu wykładniczo. Oczywiście współczynnik d nie może przekroczyć wartości 1, ale nieograniczony wzrost współczynnika μ jest teoretycznie możliwy. W pracy [28], na podstawie analizy danych dotyczących dość dużej grupy stulatków, stwierdzono, że współczynnik μ ma wzrost ograniczony. Mówimy wtedy o zjawisku *wypłaszczenia śmiertelności* (ang. *mortality plateau*). Taka graniczna wartość μ wynosi około 0,65 i jest osiągana w wieku 105 lat. Zgodnie

z otrzymanymi wynikami, szansa przeżycia przez superstulatka kolejnych 10 lat wynosi 0,15%. Jeżeli teoria ta się potwierdzi, to biorąc pod uwagę, że liczba stulatków i superstulatków rośnie, możemy się spodziewać, iż granica wieku 120 lat zostanie przekroczona, choć przekroczenie wieku 130 lat jest zupełnie nieprawdopodobne.

Drugim obok śmiertelności ważnym czynnikiem demograficznym jest dynamika urodzeń. Wykres na rysunku I.6 przedstawia, jak zmieniała się liczba dzieci urodzonych w latach 1946–2015. Kolejne lokalne maksima urodzeń (odpowiadające wyżom demograficznym) były w latach 1955 – 793,8 tys., 1983 – 723,6 tys. i 2009 – 417,6 tys., a minima lokalne w latach 1967 – 521,8 tys. i 2003 – 351,1 tys. Liczba dzieci urodzonych w danym okresie zależy od struktury wiekowej populacji i dzietności kobiet.



Rysunek I.6. Urodzenia żywe w latach 1946–2015, gdzie R oznacza rok, a U liczbę dzieci urodzonych w danym roku

O ile strukturę wiekową populacji, a w szczególności liczbę kobiet w wieku rozrodczym i rozkład ich wieku, można dość dokładnie przewidywać na wiele lat naprzód, to trudno jest określić, jakie będą tendencje dotyczące liczby dzieci i rozkładu urodzeń według wieku matki. W tabeli I.3 porównaliśmy dane dotyczące urodzeń w latach 1990 i 2015. Podana jest w niej liczba urodzeń żywych w zależności od grupy wiekowej i ich procentowy udział w poszczególnych przedziałach oraz średnia liczba matek noworodków na tysiąc kobiet w danej grupie wiekowej – $10^3 b$. Współczynnik b nazywamy *współczynnikiem urodzeń* i oznacza on iloraz liczby matek noworodków do ogólnej liczby kobiet w danej grupie. Należy jeszcze wspomnieć, że możemy definiować inne współ-

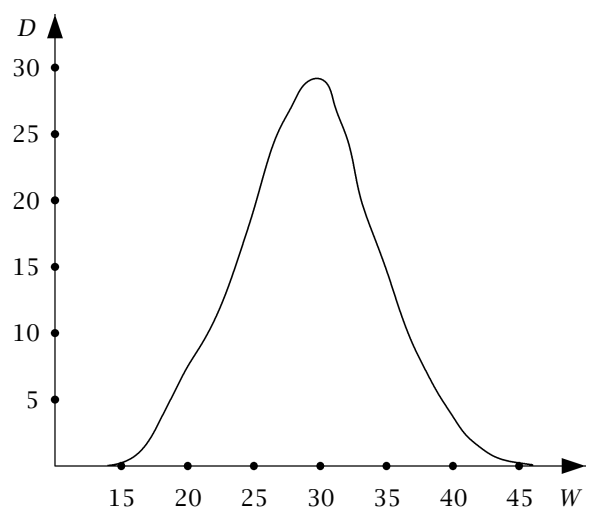
	Ogółem	–19	20–24	25–29	30–34	35–39	40–44	45+
1990	547720	44074	199575	160520	95052	39847	8374	278
%	100	8,047	36,437	29,307	17,354	7,275	1,529	0,051
$10^3 b$		31,73	166,04	122,07	58,97	24,40	6,26	0,31
2015	369308	12030	57107	123994	119140	47972	8725	340
%	100	3,256	15,463	33,574	32,260	12,990	2,363	0,0921
$10^3 b$		12,48	48,30	89,06	74,45	31,31	6,45	0,29

Tabela I.3. Urodzenia żywe wg wieku matki w latach 1990 i 2015 w różnych przedziałach wieku

czynniki urodzeń, np. iloraz liczby noworodków dziewczynek do liczby kobiet lub iloraz liczby noworodków do liczby kobiet i mężczyzn w danej grupie. Zakładając, że kobiety stanowią około 49% populacji osób w wieku poniżej 45 lat, przyjmujemy, że te dwa ostatnie współczynniki urodzeń wynoszą $0{,}49b$. Jeżeli populacja podzielona jest na n grup, a w danej grupie i liczba kobiet wynosi K_i, liczba ludności L_i, a współczynnik urodzeń b_i, to łączną liczbę dzieci urodzonych w danym roku znajdujemy ze wzoru

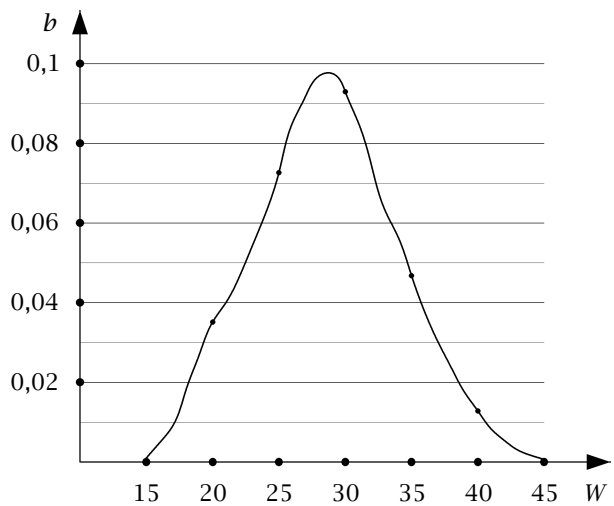
$$D = \sum_{i=1}^{n} K_i b_i \approx \sum_{i=1}^{n} 0{,}49 L_i b_i. \tag{2.7}$$

Wykres na rysunku I.7 przedstawia zależność urodzeń w liczbach bezwzględnych od wieku matki. Dane pochodzą z roku 2015. Najwiekszą liczbę



Rysunek I.7. Urodzenia według wieku matki w roku 2015: W – wiek matki, D – liczba dzieci w tysiącach

urodzeń (29 tys.) odnotowano dla kobiet w wieku 30 lat. Zależność współczynnika urodzeń od wieku matki przedstawiona jest na rysunku I.8. Współczynnik urodzeń miał największą wartość dla kobiet w wieku 29 lat i wynosił 0,0975. Porównując dane z tabeli I.3, możemy zauważyć, że nastąpiła istotna zmiana w rozkładzie wieku matek w ciągu ostatnich 25 lat. Najwięcej urodzeń w roku 1990 przypadało na matki w wieku 24 lat, podczas gdy w roku 2015 – na matki w wieku 30 lat, a największy współczynnik urodzeń miały kobiety 29-letnie.

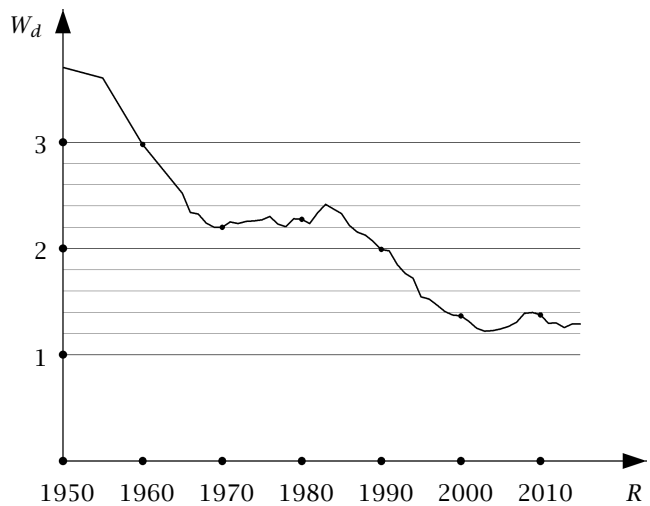


Rysunek I.8. Współczynnik urodzeń według wieku matki w roku 2015; W – wiek matki, b – współczynnik urodzeń

Ważnym wskaźnikiem demograficznym jest *dzietność kobiet* – definiowana jako przeciętna liczba dzieci, które urodziłaby kobieta w ciągu całego okresu rozrodczego (wiek 15–49 lat) przy założeniu, że w poszczególnych fazach tego okresu rodziłaby z intensywnością obserwowaną wśród kobiet w badanym roku. Zwykle przyjmuje się, że do utrzymania populacji na stałym poziomie dzietność powinna wynosić 2,1, a więc nieco przekraczać 2, ponieważ kobiety stanowią około 49% osób w wieku do 49 lat, a trzeba też uwzględnić śmiertelność dziewczynek w wieku do 14 lat. Dzietność możemy stosunkowo łatwo wyznaczyć, korzystając z rozkładu współczynnika urodzeń (patrz tabela I.3 i rys. I.8). Jeżeli populacja podzielona jest na n przedziałów wieku, a przedział i obejmuje t_i roczników kobiet o średnim współczynniku urodzeń b_i, to współczynnik dzietności wynosi

$$W_d = \sum_{i=1}^{n} t_i b_i. \tag{2.8}$$

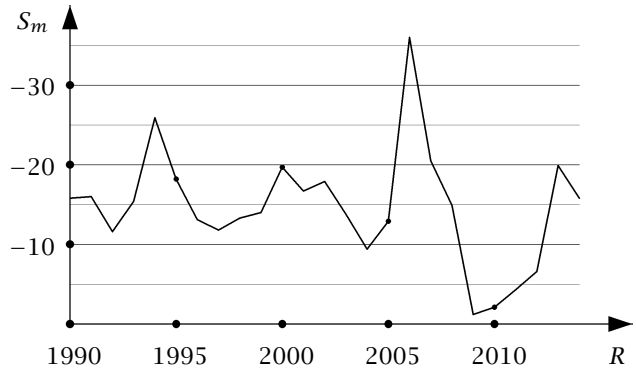
W roku 1990 dzietność wynosiła 1,991, a w roku 2015 tylko 1,289. Dla porównania, w latach pięćdziesiątych dzietność była w przedziale 3,6–3,7, w roku 1965 wynosiła 2,5, w latach siedemdziesiątych na poziomie 2,2–2,3, w roku 1983 osiągnęła lokalne maksimum 2,416 i od tego roku systematycznie spadała aż do najniższego poziomu 1,222 w roku 2003. W następnych latach utrzymywała się średnio na poziomie 1,3, osiągając maksimum lokalne 1,398 w roku 2009. Wykres na rysunku I.9 przedstawia, jak zmieniała się dzietność kobiet w latach 1950–2015.



Rysunek I.9. Dzietność kobiet w latach 1950–2015, gdzie R oznacza rok, a W_d współczynnik dzietności

Ostatnim elementem decydującym o strukturze demograficznej jest migracja. Jak już wspomnieliśmy, wielkość migracji zależy od wielu czynników wewnętrznych i zewnętrznych. Z demograficznego punktu widzenia najistotniejszym wskaźnikiem jest *saldo migracji*, wyrażone różnicą w liczbach bezwzględnych między imigracją i emigracją. Wykres na rysunku I.10 przedstawia bilans migracji na pobyt stały w latach 1990–2014. W całym rozważanym okresie saldo migracji było ujemne i cechowało się dużą zmiennością. Największy ubytek ludności Polski z tego powodu odnotowano w roku 2006 – 36 tys., a najmniejszy w latach 2009–2011 – średnio 2,5 tys. na rok. W latach 1990–2014 wyemigrowało 603 tys. osób, imigrowało 243 tys., saldo migracji wynosiło więc -360 tys.

W badaniach struktury wiekowej populacji i w jej prognozowaniu istotne są dane dotyczące wieku i płci emigrantów i imigrantów. W tabeli I.4 przedstawiamy dane dotyczące roku 2014. Należy zwrócić uwagę na fakt, że w pewnych przedziałach wieku czynniki migracyjne mają większy wpływ na zmianę liczeb-



Rysunek I.10. Saldo migracji w latach 1990–2014, gdzie R oznacza rok, a S_m saldo migracji w tysiącach

	Imigracja			Emigracja		
	Łącznie	M	K	Łącznie	M	K
Ogółem	12330	6777	5553	28080	13803	14277
0–4	3783	1926	1857	1290	647	643
5–9	759	380	379	1946	1028	918
10–14	287	140	147	1531	771	760
15–19	394	207	187	1876	1243	633
20–24	537	279	258	1834	1005	829
25–29	1081	710	371	3090	1445	1645
30–34	1322	848	474	4300	1936	2364
35–39	933	588	345	3559	1654	1905
40–44	611	371	240	2450	1185	1265
45–49	482	273	209	1632	792	840
50–54	490	238	252	1382	646	736
55–59	492	250	242	1223	590	633
60–64	518	238	280	921	452	469
65–69	338	199	139	423	192	231
70+	303	130	173	623	217	406

Tabela I.4. Migracja w roku 2014 z podziałem na imigrację, emigrację i różne grupy wiekowe

ności tych grup niż śmiertelność. Na przykład w grupie wiekowej 25–29 lat saldo migracji w roku 2014 wynosiło −2009, a zmarło 1840 osób.

2.4. Prognozy demograficzne

Mając do dyspozycji dane dotyczące demografii Polski, rozpoczniemy od modelowania prognoz demograficznych. Będą nas interesowały prognozy wieloletnie, które w dużym stopniu zależą od przyjętych założeń. Najprostsze modele oparte są na założeniu, że wskaźniki demograficzne, takie jak współczynnik urodzeń i śmiertelności czy bilans migracji, będą stałe i takie same jak w roku bazowym. Modele te można modyfikować tak, aby uwzględnić trendy wieloletnie. Na przykład współczynnik śmiertelności od wielu lat maleje, i to we wszystkich przedziałach wieku. Można również prowadzić symulacje, przyjmując, że niektóre wskaźniki ulegną istotnym zmianom ze względu na decyzje państwa. Na przykład zmiana w polityce socjalnej może wpłynąć na odpowiedni wzrost współczynnika urodzeń. Będziemy się głównie posługiwać danymi dotyczącymi przedziałów wieku (tabele I.1–I.4). Dokładniejsze prognozy uzyskalibyśmy, wykorzystując dane dotyczące roczników; przy odpowiedniej metodologii uzyskane różnice będą jednak niewielkie w stosunku do błędów, które mogą się pojawić w wyniku rozminięcia się założeń modelu z przyszłymi wskaźnikami.

Przykład I.3 (Statyczny model demograficzny I)**.** Rozpoczniemy od najprostszego modelu, w którym mamy dane dotyczące liczebności wszystkich roczników w populacji w wyjściowym roku (który będziemy nazywać *rokiem zerowym*) oraz znamy współczynniki śmiertelności i urodzeń dla wszystkich roczników. Będziemy zakładać, że współczynniki te będą stałe w prognozowanym okresie. Niech L_i będzie liczbą ludności w wieku i w roku zerowym (powiedzmy na koniec 31 grudnia roku zerowego). Przyjmijmy, że b_i i d_i są odpowiednio współczynnikiem urodzeń i śmiertelności w wieku i. Współczynnik śmiertelności d_i traktujemy jako prawdopodobieństwo zgonu w ciągu roku pod warunkiem, że osoba jest z danej grupy wiekowej. Zatem $1 - d_{i+1}$ będzie prawdopodobieństwem przeżycia następnego roku.

Na początek zauważmy, że z populacji liczącej L_i osób następny rok może przeżyć z pewnym prawdopodobieństwem p_l dokładnie l osób, gdzie l jest liczbą naturalną z przedziału od zera do L_i. Prawdopodobieństwo to można obliczyć z *rozkładu Bernoulliego*, nie będziemy się jednak tym zajmować. Interesuje nas spodziewana (nieco precyzyjniej, *średnia*) liczba osób, które przeżyją następny rok. Intuicja podpowiada, że będzie ich $L_i(1 - d_{i+1})$. Intuicja ta ma głębokie uzasadnienie w *prawie wielkich liczb*, które poznamy później. Szansa przeżycia kolejnych k lat przez osobę w wieku i wynosi

$$p_{i\,i+k} = \prod_{j=i+1}^{i+k} (1 - d_j), \tag{2.9}$$

a prognozowana liczba osób w wieku $i \geq r$ w r-tym roku wynosi $L_i^r = L_{i-r}p_{i-r\,i}$. Jeżeli chcielibyśmy w tym modelu uwzględnić migrację (przy założeniu, że jest ona również na stałym poziomie zależnym tylko od wieku, co jest dużym uproszczeniem), to można we wzorze (2.9) składnik $1-d_j$ zastąpić składnikiem $1-d_j+m_j$, gdzie m_j jest *względnym saldem migracji*, a więc $m_i = S_m(i)/L_i$, gdzie $S_m(i)$ jest saldem migracji dla osób w wieku i lat. Osoby w wieku $i < r$ urodziły się już po roku zerowym, a więc ich prognozowaną liczbę znajdujemy, korzystając ze wzoru (2.7). Osoby te urodziły się w roku $r-i$. Ich liczba w tym roku określona jest przybliżonym wzorem

$$D_{r-i} \approx \sum_j 0{,}49 L_j^{r-i} b_j,$$

gdzie zmienna j oznacza wiek matki i zmienia się w zakresie przyjętym w dostępnych danych dotyczących współczynnika urodzeń. Przyjmując, że okres prognozy jest na tyle krótki, że matki urodziły się do roku zerowego włącznie (np. $r \leq 15$), otrzymujemy

$$D_{r-i} \approx \sum_j 0{,}49 L_{j-r+i} p_{j-r+i\,j} b_j$$

i ostatecznie

$$L_i^r \approx \sum_j 0{,}49 L_{j-r+i} p_{j-r+i\,j} b_j p_{0i}. \tag{2.10}$$

Przyjmując, że śmiertelność wśród kobiet w wieku rozrodczym oraz dzieci jest na tyle mała, że można ją zaniedbać, ostatecznie otrzymujemy

$$L_i^r \approx \begin{cases} \sum_j 0{,}49 L_{j-r+i} b_j & \text{dla } i < r, \\ L_{i-r} p_{i-r\,i} & \text{dla } i \geq r. \end{cases} \tag{2.11}$$

Jeżeli interesuje nas prognoza w dłuższym okresie, to należy rozbić go na krótsze okresy, np. 15-letnie, tak, aby można było pominąć sytuacje, w których dziewczynka rodzi się i zostaje matką w tym samym okresie.

Uwaga I.4. Obliczając liczbę nowo narodzonych dzieci w danym roku, skorzystaliśmy ze wzoru (2.7), który jest formalnie zgodny z przyjętą definicją współczynnika urodzeń. W tym wypadku najpierw znajdujemy liczbę ludności w przedziale wieku od jeden wzwyż, a następnie na podstawie tych danych obliczamy liczbę dzieci urodzonych w danym roku. Procedura wyznaczania rozkładu wiekowego w ustalonym roku na podstawie rozkładu z roku poprzedzającego jest więc dwuetapowa. Możemy ją nieco uprościć, przyjmując, że jeżeli L_j jest liczbą ludności w wieku j w danym roku, to $D = \sum_j 0{,}49 L_j b_j$ jest liczbą dzieci w następnym roku. Wyniki z użyciem obu wzorów niewiele się od

siebie różnią, o ile nie ma zbyt dużych różnic tego samego znaku w liczebności kolejnych roczników kobiet w wieku rozrodczym. W roku 2015 różnice te były znaczne – w kolejnych rocznikach było przeciętnie 10 tys. mniej kobiet w grupie wiekowej 19–32 lat, co spowodowało, że liczba dzieci obliczona według drugiej metody była o 7,5 tys. mniejsza niż według schematu z przykładu I.3.

Przykład I.5 (Statyczny model demograficzny II)**.** Ponieważ dane w tabelach I.1–I.4 dotyczą przedziałów wieku, a nie pojedynczych roczników, wyjaśnimy, jak należy zmodyfikować poprzedni model, aby można było korzystać z tych danych. Załóżmy, że przedział wieku obejmuje n roczników oraz r jest krotnością n. Przyjmijmy, że w ramach tej samej grupy wiekowej współczynniki urodzeń i śmiertelności oraz bilans migracji są identyczne oraz że roczniki w tej samej grupie są równoliczne. Będzie nas interesować stan populacji po n latach. Mając taki model, będzie można go zastosować wraz ze zmodyfikowanym modelem poprzednim do wyznaczenia prognozy długoletniej.

Niech k-ty przedział wieku obejmujący roczniki od $n(k-1)$ do $nk-1$ liczy L_k osób i ma współczynnik urodzeń b_k, współczynnik śmiertelności d_k i względne saldo migracji m_k. Rozważmy rocznik $nk-i$. Wtedy osoba z tego rocznika będzie pozostawać w tej grupie przez $i-1$ lat, a następnie przejdzie do grupy $k+1$, w której będzie przez $n-i+1$ lat. Szansa przeżycia kolejnych n lat wynosi zatem

$$(1-d_k)^{i-1}(1-d_{k+1})^{n-i+1} \approx 1-(i-1)d_k-(n-i+1)d_{k+1}.$$

Po n latach wszystkie osoby z grupy wiekowej k utworzą grupę wiekową $k+1$, która będzie liczyć

$$\begin{aligned} L_{k+1}^1 &= \frac{L_k}{n}\sum_{i=1}^{n}(1-d_k)^{i-1}(1-d_{k+1})^{n-i+1} \\ &\approx \frac{L_k}{n}\sum_{i=1}^{n}\big(1-(i-1)d_k-(n-i+1)d_{k+1}\big) \\ &\approx L_k\left(1-\frac{n-1}{2}d_k-\frac{n+1}{2}d_{k+1}\right) \end{aligned}$$

osób. Można więc przyjąć, że *n-letni współczynnik śmiertelności* $\bar{d}_k$ dla k-tej grupy wiekowej wynosi

$$\bar{d}_k = \frac{(n-1)d_k+(n+1)d_{k+1}}{2}. \tag{2.12}$$

Jeżeli pominiemy stosunkowo niską śmiertelność wśród kobiet w wieku rozrodczym i migrantów, to w analogiczny sposób otrzymujemy *n-letni współczynnik*

urodzeń $\bar{b}_k$ oraz *n-letnie względne saldo migracji* $\bar{m}_k$:

$$\bar{b}_k = \frac{(n-1)b_k + (n+1)b_{k+1}}{2}, \qquad \bar{m}_k = \frac{(n-1)m_k + (n+1)m_{k+1}}{2}. \tag{2.13}$$

Jeśli wyznaczymy nowe wskaźniki dotyczące okresu n-letniego, możemy nadal posłużyć się poprzednim modelem, zastępując krok roczny krokiem n-letnim. Ponieważ dane, którymi dysponujemy, dotyczą zwykle pięcioletnich przedziałów wieku, więc $n = 5$ i wtedy $\bar{d}_k = 2d_k + 3d_{k+1}$, $\bar{b}_k = 2b_k + 3b_{k+1}$ oraz $\bar{m}_k = 2m_k + 3m_{k+1}$.

Uwaga I.6. Ze względów poglądowych w tabelach I.1 i I.2 podaliśmy dane dotyczące śmiertelności w roku 2015 rozbite na wiek 0 i wiek 1–4. Dla uproszczenia obliczeń, prognozując strukturę wiekową, przyjmujemy, że średnia śmiertelność w całym przedziale wieku 0–4 wynosiła $(400 + 4 \cdot 17)/5 \approx 94$ na sto tysięcy mieszkańców, a współczynnik śmiertelności wynosił 0,00094.

Przykład I.7 (Dynamiczny model demograficzny)**.** W wyniku wzrostu świadomości zdrowotnej, podniesienia poziomu życia i poprawy opieki zdrowotnej stale wydłuża się przeciętna długość życia i zmniejsza się śmiertelność we wszystkich przedziałach wieku. Przyjmując, że jest to tendencja trwała, należy zrezygnować ze współczynnika śmiertelności zależnego tylko od wieku i wprowadzić do modelu współczynnik śmiertelności zmienny w czasie. Jeżeli przyjmiemy, że d_{jr} jest współczynnikiem śmiertelności w wieku j i w r-tym roku od roku zerowego, to szansa przeżycia kolejnych k lat przez osobę w wieku i wynosi

$$p_{i\,i+k} = \prod_{r=1}^{k} (1 - d_{i+r\,r}) \tag{2.14}$$

i zastępując wzór (2.9) w przykładzie I.3, otrzymamy model dynamiczny. Podobną procedurę można zastosować w przykładzie I.5, modyfikując wskaźniki w okresie n-letnim.

Pojawia się pytanie, jak prognozować współczynniki śmiertelności w następnych latach, znając dynamikę ich zmienności w latach poprzednich. Pewnych wskazówek dostarcza nam tabela I.2, na podstawie której możemy porównać współczynniki śmiertelności w różnych przedziałach wieku w latach 1990 i 2015, a także wykres zależności współczynnika śmiertelności od wieku (patrz rys. I.5). Po pierwsze widzimy, że prawie pięciokrotnie zmalała umieralność niemowląt, a mniej niż 0,4% dzieci umiera w wieku 1–19 lat przy średnim współczynniku śmiertelności w tym okresie około $2 \cdot 10^{-4}$; śmiertelność w tej grupie wiekowej jest więc niewielka. Współczynnik śmiertelności w roku 2015 wynosił 0,49–0,68 współczynnika śmiertelności z roku 1990 zależnie od grupy wiekowej w przedziale wieku 15–84. Jeżeli tendencja ta się utrzyma, to możemy przyjąć,

że w tym przedziale wieku współczynnik śmiertelności zmaleje o około 10% w ciągu pięciu lat. W zadaniach będziemy rozważać wersję dynamiczną modelu z przykładu I.5 i przyjmować, że współczynnik śmiertelności w grupie wiekowej k w okresie od $5s$ do $5s+4$ lat od roku zerowego wynosi $\bar{d}_{ks} = 0{,}9^s \bar{d}_k$. W wieku powyżej 84 lat zmiany współczynnika śmiertelności w latach 1990–2015 były mniejsze, głównie ze względu na szybki wzrost współczynnika śmiertelności wraz z wiekiem. Prognozując współczynnik śmiertelności w wieku powyżej 84 lat, możemy przyjąć, że rośnie on wykładniczo wraz z wiekiem, i skorzystać z wcześniej obliczonych współczynników dla młodszych roczników.

Państwo poprzez swoją politykę lub też sytuacja międzynarodowa mogą wpływać na zachowania demograficzne ludności. Przykładem mogą tu być programy społeczne, które mają za zadanie zwiększenie współczynnika urodzeń. Również polepszenie lub pogorszenie koniunktury gospodarczej może wpłynąć na migrację zagraniczną. Prognozowanie wpływu takich czynników na strukturę demograficzną kraju może pomóc w wypracowaniu odpowiednich decyzji strategicznych. Prognozy liczby ludności zależą od przyjętych hipotez, modele są jednak budowane na tych samych zasadach co poprzednie z uwzględnieniem innych wskaźników. W zadaniach I.5–I.9 prognozujemy strukturę demograficzną Polski w latach 2030 i 2045 w zależności od wybranego modelu i przyjętych założeń. Szczególnie ciekawe może być porównanie prognoz demograficznych z różnymi współczynnikami urodzeń.

2.5. Model McKendricka

W modelach demograficznych rozważanych do tej pory czas zmieniał się w odstępach roku, a populacja była podzielona według roczników lub przedziałów wieku oraz według płci. Czas zmieniał się więc w sposób dyskretny, a także struktura podziału populacji na mniejsze grupy (podpopulacje) była dyskretna. Modele takie nazywamy modelami z czasem i strukturą dyskretną lub krótko *modelami Lesliego* [229]. Do opisu struktury wiekowej ludności wybraliśmy model dyskretny ze względu na typ dostępnych danych i rodzaj rozpatrywanych problemów, a także z powodu jego prostoty. Jeżeli ograniczymy się do rozważania rozkładu wiekowego wyłącznie populacji kobiet (rozkład wiekowy mężczyzn można wyznaczyć, znając proporcje liczby mężczyzn do liczby kobiet w zależności od wieku), to ewolucja tego rozkładu określona jest wzorem rekurencyjnym

$$K_i^r = \begin{cases} K_{i-1}^{r-1}(1-d_i^r), & \text{gdy } i \geq 1, \\ \sum_j K_j^r b_j^r, & \text{gdy } i = 0, \end{cases} \tag{2.15}$$

gdzie K_i^r jest liczbą kobiet w wieku i w r-tym roku, d_i^r jest współczynnikiem śmiertelności w wieku i wyznaczonym w r-tym roku, a b_i^r jest współczynnikiem urodzeń dziewczynek przez kobiety w wieku i w r-tym roku.

Zamiast modeli dyskretnych można rozważać model z czasem ciągłym, a także z wiekiem opisywanym zmienną rzeczywistą. Modele takie są dość często używane i mają wielorakie zastosowania, zarówno dla populacji ludzi i zwierząt (gdzie ograniczamy się do kobiet lub samic), jak i dla populacji komórkowych. Ważną rolę odgrywa tu model McKendricka [257] z roku 1926, wprowadzony przy okazji jego badań epidemiologicznych. Należy zaznaczyć, że wcześniej podobny model zaproponowali Sharpe i Lotka [348], a w roku 1959 stosował go von Foerster [120] w badaniach populacji komórkowych – i to jemu czasami przypisuje się jego wprowadzenie.

Będziemy używać standardowych angielskich oznaczeń t na czas oraz a na wiek osobnika (lub komórki), natomiast $u(t,a)$ jest gęstością rozkładu wieku, a więc $\int_0^{a_1} u(t,a)\,da$ jest liczbą (lub biomasą) osobników, których wiek w chwili t nie przekracza a_1. Należy zaznaczyć, że termin „gęstość rozkładu" nie jest tu używany w znaczeniu probabilistycznym – całka z gęstości po całej przestrzeni nie musi wynosić 1 i może się zmieniać w czasie.

Podobnie jak w modelu dyskretnym wprowadza się współczynniki urodzeń i śmiertelności. Przez $\mu(t,a)$ oznaczamy współczynnik śmiertelności; zakładamy, że prawdopodobieństwo, że osobnik w wieku a w chwili t nie przeżyje do chwili $t+\Delta t$, wynosi $\mu(t,a)\Delta t + o(\Delta t)$, gdzie $o(x)$ oznacza taką funkcję zmiennej x, że $\lim_{x\to 0^+} o(x)/x = 0$. Zatem, zgodnie z definicją $\mu(t,a)$, liczba osobników w wieku a, które zmarły w czasie $(t, t+\Delta t)$, wynosi

$$u(t,a) - u(t+\Delta t, a+\Delta t) = (\mu(t,a)\Delta t + o(\Delta t))u(t,a).$$

Dzieląc obie strony równości przez $-\Delta t$ i przechodząc do granicy przy $\Delta t \to 0^+$, otrzymujemy

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = -\mu(t,a)u(t,a). \tag{2.16}$$

Podobnie definiujemy współczynnik urodzeń $b(t,a)$. Liczba nowych osobników jest proporcjonalna do współczynnika urodzeń i liczebności populacji, wynosi więc $b(t,a)u(t,a)$. Po zsumowaniu po całym przedziale $[0, a_m]$ otrzymujemy wzór

$$u(t,0) = \int_0^{a_m} b(t,a)u(t,a)\,da, \tag{2.17}$$

gdzie $a_m \le \infty$ jest maksymalnym rozważanym wiekiem.

Równania (2.16) i (2.17) tworzą model McKendricka i wraz z warunkiem początkowym $u(0,a) = u_0(a)$ pozwalają ustalić, jak zmienia się rozkład wieku

osobników w czasie. O funkcjach $\mu(t,a)$ i $b(t,a)$ będziemy zakładać, że są ciągłe. Przez $N(t)$ oznaczamy całkowitą liczbę osobników w populacji w chwili t, tj. $N(t) = \int_0^{a_m} u(t,a)\,da$. Funkcję $p(t,a) = u(t,a)/N(t)$ nazywamy *profilem wiekowym* populacji.

Dla populacji komórkowych przyjmujemy, że dzieląc się, komórka ginie. Jeżeli $d(t,a)$ jest współczynnikiem naturalnej śmiertelności, a $r(t,a)$ jest współczynnikiem podziału, to

$$\mu(t,a) = d(t,a) + r(t,a), \quad b(t,a) = 2r(t,a).$$

2.6. Zastosowania w zagadnieniach ochrony zdrowia

Podamy teraz kilka przykładów zastosowania prawdopodobieństwa warunkowego i reguły Bayesa przy wyznaczaniu współczynników ryzyka oraz badaniu skuteczności testów.

Przykład I.8 (Współczynniki surowe i standaryzowane)**.** W tablicach danych dotyczących ochrony zdrowia pojawiają się dwa rodzaje współczynników: surowe i standaryzowane. Na przykład *surowy współczynnik zachorowalności* na nowotwory określa liczbę zgłaszanych po raz pierwszy w danym roku kalendarzowym przypadków zachorowań na nowotwory złośliwe w przeliczeniu na 100 tys. mieszkańców; wynosi on więc $W_{\mathrm{su}} = \frac{Z}{L} \cdot 10^5$, gdzie Z jest liczbą nowych zachorowań zgłoszonych w danym roku kalendarzowym, a L jest liczebnością badanej populacji w tym roku. Podobnie można określać surowe współczynniki umieralności. Pomijając czynnik 10^5, surowy współczynnik zachorowalności jest prawdopodobieństwem zachorowania w danym roku losowo wybranej osoby z całej populacji.

Współczynniki surowe nie uwzględniają struktury wiekowej populacji, dlatego nie należy ich używać do porównywania danych z różnych populacji. Z tego powodu wprowadza się *standaryzowane* (wg wieku) współczynniki zachorowalności i umieralności, które określają, ile zachorowań wystąpiłoby w danej populacji (w przeliczeniu na 100 tys. mieszkańców), gdyby struktura wieku tej populacji była taka sama jak struktura wieku populacji przyjętej za standard. Jako populację standardową przyjmujemy tu populację świata. Współczynniki standaryzowane obliczamy według wzoru

$$W_{\mathrm{st}} = \frac{x_1 y_1 + x_2 y_2 + \cdots + x_n y_n}{y_1 + y_2 + \cdots + y_n},$$

gdzie $x_1, \ldots, x_n$ są współczynnikami surowymi dla poszczególnych 5-letnich grup wieku, a $y_1, \ldots, y_n$ jest liczebnością standardowej populacji w tych samych przedziałach wieku. Zauważmy, że $x_i \cdot 10^{-5}$ jest prawdopodobieństwem

W	0–4	5–9	10–14	15–19	20–24	25–29
%	8,855	8,687	8,597	8,474	8,222	7,928
W	30–34	35–39	40–44	45–49	50–54	55–59
%	7,605	7,145	6,590	6,038	5,371	4,547
W	60–64	65–69	70–74	75–79	80–84	85+
%	3,723	2,955	2,210	1,515	0,905	0,632

Tabela I.5. Standardowa populacja świata *World (WHO 2000–2025) Standard*

warunkowym zachorowania w danym roku, gdy osoba należy do grupy wiekowej i. Przyjmiemy, że y_i' jest względną liczebnością grupy wiekowej i w standardowej populacji, tj. $y_i' = y_i/(y_1 + \cdots + y_n)$. Wtedy $W_{\mathrm{st}} = x_1 y_1' + \cdots + x_n y_n'$, a więc $W_{\mathrm{st}} 10^{-5}$ jest prawdopodobieństwem (całkowitym) zachorowalności w populacji standardowej, jeżeli miałaby takie same współczynniki zachorowań w poszczególnych przedziałach wieku jak w danej populacji. Przy porównywaniu danych będziemy przyjmować jako standardową populację świata *World (WHO 2000-2025) Standard*, o rozkładzie przedstawionym w tabeli I.5.

Przykład I.9 (Czysty współczynnik ryzyka)**.** Przy badaniu skuteczności terapii nowotworowych ważne są informacje, jak często po zastosowaniu terapii i powrocie do zdrowia następuje powrót choroby lub zgon pacjenta w wyniku nawrotu choroby w określonym czasie T. Pomijając trudności ze zbieraniem takich danych, pojawia się jeszcze inny ważny aspekt zagadnienia. Można rozważać różne współczynniki ryzyka nawrotu choroby (lub śmierci). Pierwszy to *surowy współczynnik ryzyka* nawrotu choroby, $W_{\mathrm{su}} = \frac{n_1}{n}$, gdzie n jest całkowitą liczbą pacjentów, którzy powrócili do zdrowia, a n_1 jest liczbą pacjentów z nawrotem choroby (lub zgonów w wyniku nawrotu choroby) w określonym czasie T. Wadą tego współczynnika jest to, że część pacjentów może umrzeć w czasie T z innych przyczyn. Współczynnik ten jest więc obarczony istotnym błędem, często związanym np. ze środowiskiem, w którym żyją pacjenci.

Okazuje się, że wyznaczenie (zdefiniowanie) *czystego współczynnika ryzyka* W_c, nieobarczonego wspomnianym błędem, nie jest proste. Można byłoby przyjąć, że czysty współczynnik ryzyka wynosi $\frac{n_1}{n-n_2}$, gdzie n_2 jest liczbą pacjentów zmarłych w czasie T z innych przyczyn. Współczynnik ten również nie jest w pełni adekwatny, ponieważ u osób zmarłych z innych przyczyn obserwacja, czy nastąpił nawrót choroby, była prowadzona w czasie krótszym niż T. Jeżeli przyjęlibyśmy, że zarówno śmiertelność z innych przyczyn, jak i nawrót choroby rozkładają się jednostajnie w rozpatrywanym przedziale czasu oraz

że zdarzenia te są niezależne, to otrzymalibyśmy zależność

$$W_c = \frac{n_1 + \frac{1}{2}n_2 W_c}{n}, \tag{2.18}$$

gdzie składnik $\frac{1}{2}n_2 W_c$ uwzględnia potencjalne przypadki nawrotu choroby u pacjentów zmarłych z innych przyczyn. Przekształcając wzór (2.18), otrzymujemy

$$W_c = \frac{n_1}{n - \frac{1}{2}n_2}. \tag{2.19}$$

W praktyce pozostaje jeszcze poważniejszy problem braku danych od części pacjentów, który należy uwzględnić przy ocenie ryzyka. Pełniejszą analizę przedstawionego problemu można znaleźć w [272].

Przykład I.10 (Skuteczność diagnostyki)**.** W wypadku wielu chorób przeprowadza się wstępne testy (lub np. badania radiologiczne) mające stwierdzić, czy pacjent jest chory. Testy te nie są zwykle w pełni skuteczne, a metody probabilistyczne pozwalają ocenić ich skuteczność.

Przypuśćmy, że grupę badanych podzieliliśmy na trzy kategorie A_1, A_2 i A_3 – osób zdrowych, w początkowej fazie choroby i w zaawansowanym stadium choroby. Przyjmujemy, że $q_i = \mathrm{P}(A_i)$, $i = 1, 2, 3$, oznacza względną liczebność poszczególnych kategorii badanych. Przypuśćmy, że na podstawie wcześniej przeprowadzonych badań klinicznych ustalono, że test daje wynik pozytywny (stwierdza, że badany jest chory) z prawdopodobieństwami p_1, p_2, p_3 w kolejnych kategoriach. Jeżeli B oznacza zdarzenie, że test dał wynik pozytywny, to $p_i = \mathrm{P}(B|A_i)$. Ze wzoru (2.3) wynika, że prawdopodobieństwo, że u losowo wybranej osoby test wypadnie pozytywnie, wynosi

$$\begin{aligned} \mathrm{P}(B) &= \mathrm{P}(B|A_1)\,\mathrm{P}(A_1) + \mathrm{P}(B|A_2)\,\mathrm{P}(A_2) + \mathrm{P}(B|A_3)\,\mathrm{P}(A_3) \\ &= p_1 q_1 + p_2 q_2 + p_3 q_3. \end{aligned} \tag{2.20}$$

Na podstawie reguły Bayesa (2.5) prawdopodobieństwo warunkowe, że badany należy do kategorii A_k, jeśli wiemy, że test wypadł pozytywnie, wynosi

$$\mathrm{P}(A_k|B) = \frac{\mathrm{P}(B|A_k)\,\mathrm{P}(A_k)}{\mathrm{P}(B)} = \frac{p_k q_k}{p_1 q_1 + p_2 q_2 + p_3 q_3}, \tag{2.21}$$

a prawdopodobieństwo tego zdarzenia pod warunkiem, że wynik testu był negatywny, wynosi

$$\mathrm{P}(A_k|B') = \frac{\mathrm{P}(B'|A_k)\,\mathrm{P}(A_k)}{\mathrm{P}(B')} = \frac{(1 - p_k) q_k}{1 - (p_1 q_1 + p_2 q_2 + p_3 q_3)}. \tag{2.22}$$

Rozważmy następujący przykład. Zakładamy, że przeprowadzono badania kontrolne na grupie osób, wśród których 98% jest zdrowych, a po 1% w początkowej fazie choroby i w zaawansowanym stadium choroby. Załóżmy, że test daje wynik pozytywny w kolejnych kategoriach w 2%, 90% i 97% przypadków. Zatem $p_1 = 0{,}02$, $p_2 = 0{,}9$, $p_3 = 0{,}97$, $q_1 = 0{,}98$, $q_2 = 0{,}01$ oraz $q_3 = 0{,}01$. Wtedy

$$\mathrm{P}(B) = 0{,}02 \cdot 0{,}98 + 0{,}9 \cdot 0{,}01 + 0{,}97 \cdot 0{,}01 = 0{,}0383,$$

a na podstawie wzorów (2.21) i (2.22) obliczamy odpowiednie prawdopodobieństwa warunkowe:

$$\begin{aligned} &\mathrm{P}(A_1|B) \approx 0{,}5117, \quad \mathrm{P}(A_2|B) \approx 0{,}2350, \quad \mathrm{P}(A_3|B) \approx 0{,}2533, \\ &\mathrm{P}(A_1|B') \approx 0{,}9986, \quad \mathrm{P}(A_2|B') \approx 0{,}0010, \quad \mathrm{P}(A_3|B') \approx 0{,}0003. \end{aligned}$$

Zauważmy, że ponad połowa osób z pozytywnym wynikiem testu była zdrowa, co wiąże się z faktem, że osoby zdrowe stanowiły znaczną większość badanej grupy. Wynika stąd, że jeśli wynik był pozytywny, to konieczne są dalsze badania; za to wynik negatywny dość dobrze wyklucza chorobę.

Przykład I.11 (Testy wielokrotne)**.** Jak pokazaliśmy w przykładzie I.10, pozytywny wynik testu wcale nie musi oznaczać, że osoba jest chora. Szczególnie dotyczy to *badań przesiewowych*, które przeprowadza się wśród osób nieposiadających objawów choroby w celu jej wykrycia i wczesnego leczenia, aby zapobiec poważnym następstwom choroby w przyszłości. W tym wypadku grupa osób zdrowych jest znacznie liczniejsza od osób chorych i pozytywny wynik testu ze stosunkowo dużym prawdopodobieństwem nie wyklucza, że osoba jest zdrowa. Dlatego należy przeprowadzić dokładniejsze badania lub inny, możliwie niezależny test.

Rozważmy najprostszy model, w którym badanych dzieli się na osoby zdrowe i chore, a A jest zdarzeniem, że osoba jest zdrowa. Niech $q = \mathrm{P}(A)$. Mamy dwa niezależne testy i oznaczamy przez B_j, $j = 1, 2$, zdarzenie, że wynik j-tego testu jest pozytywny. Zakładamy, że j-ty test daje wynik pozytywny z prawdopodobieństwem p_j w grupie zdrowych i p'_j w grupie chorych. Na podstawie reguły Bayesa (2.5) mamy

$$\mathrm{P}(A|B_1) = \frac{\mathrm{P}(B_1|A)\,\mathrm{P}(A)}{\mathrm{P}(B_1|A)\,\mathrm{P}(A) + \mathrm{P}(B_1|A')\,\mathrm{P}(A')} = \frac{p_1 q}{p_1 q + p'_1(1-q)}. \tag{2.23}$$

Rozważmy teraz nową przestrzeń probabilistyczną $(B_1, \Sigma_{B_1}, \tilde{\mathrm{P}})$ odpowiadającą grupie osób, które przeszły pierwszy test pozytywnie. Niech $\tilde{A}$ oznacza zdarzenie, że osoba z tej grupy jest zdrowa. Wtedy $\tilde{\mathrm{P}}(\tilde{A}) = \mathrm{P}(A|B_1)$ i zakładając, że testy są niezależne, prawdopodobieństwo warunkowe zdarzenia, że zdrowa

osoba przeszła oba testy pozytywnie, wynosi

$$\mathrm{P}(A|B_1\cap B_2)=\tilde{\mathrm{P}}(\tilde{A}|B_2)=\frac{\tilde{\mathrm{P}}(B_2|\tilde{A})\tilde{\mathrm{P}}(\tilde{A})}{\tilde{\mathrm{P}}(B_2|\tilde{A})\tilde{\mathrm{P}}(\tilde{A})+\tilde{\mathrm{P}}(B_2|\tilde{A}')\tilde{\mathrm{P}}(\tilde{A}')}$$
$$=\frac{p_2\tilde{\mathrm{P}}(\tilde{A})}{p_2\tilde{\mathrm{P}}(\tilde{A})+p_2'(1-\tilde{\mathrm{P}}(\tilde{A}))}=\frac{p_1p_2q}{p_1p_2q+p_1'p_2'(1-q)}. \tag{2.24}$$

Ten sam wynik otrzymamy, przyjmując, że testy są niezależne, gdy

$$\begin{aligned}\mathrm{P}(B_1\cap B_2|A)&=\mathrm{P}(B_1|A)\,\mathrm{P}(B_2|A),\\ \mathrm{P}(B_1\cap B_2|A')&=\mathrm{P}(B_1|A')\,\mathrm{P}(B_2|A'),\end{aligned} \tag{2.25}$$

a więc gdy zdarzenia B_1 i B_2 obcięte do przestrzeni A i A' są niezależne. Czytelnik porówna skuteczność pojedynczych i dwukrotnych testów, rozwiązując zadanie I.17. Założenie (2.25) jest dość silne, ale nawet przy pewnej zależności wyników testów dwukrotne badanie dość skutecznie eliminuje przypadki wyniku pozytywnego u osób zdrowych. W wypadku testów zależnych, aby wyznaczyć prawdopodobieństwo, że zdrowa osoba przeszła oba testy pozytywnie, musimy znać bezpośrednio prawdopodobieństwa warunkowe $\mathrm{P}(B_1\cap B_2|A)$ i $\mathrm{P}(B_1\cap B_2|A')$. Czytelnik zainteresowany medycznymi problemami diagnostyki znajdzie ciekawe przykłady w opracowaniu [282].

3. Elementarne modele genetyki populacyjnej

Wyjaśnienie, dlaczego dzieci dziedziczą cechy rodziców, pozostawało zagadką aż do roku 1944, kiedy to grupa uczonych pracujących pod kierunkiem Oswalda Avery'ego odkryła, że nośnikiem przenoszącym informacje są tu mikroskopijne cząsteczki DNA. W istocie rzeczy na długo przed tym odkryciem było oczywiste, że takie nośniki informacji zawiera każda komórka, i nazwano je genami. Jeszcze wcześniej Gregor Mendel, obserwując cechy roślin, np. kolor kwiatów grochu, odkrył podstawowe prawa dziedziczenia.

Przedstawimy teraz modele probabilistyczne teorii Mendla oparte na współczesnej wiedzy genetycznej. Będziemy w nich korzystać jedynie z prawdopodobieństwa warunkowego. Modele dotyczące stopnia pokrewieństwa przedstawione są w następnym rozdziale w punkcie 3.1. Modele te można również wprowadzić, używając pojęć elementarnych, ale w pewnym zakresie analizy tych modeli wygodnie jest posługiwać się terminologią i własnościami łańcuchów Markowa. W rozdziale drugim przedstawimy też modele dryfu genetycznego Wrighta–Fishera i Morana.

3.1. Krzyżowanie i dziedziczenie

Za dziedziczenie odpowiadają fragmenty DNA zwane genami. We wszystkich komórkach osobnika występują kopie tego samego DNA, a więc te same geny. Geny rozmieszczone są na chromosomach, a chromosomy w komórkach człowieka i wielu innych gatunków występują parami. Wyjątek stanowią komórki rozrodcze, zwane gametami, z pojedynczymi chromosomami. Para chromosomów jest praktycznie identyczna (jednym z wyjątków jest para XY chromosomów płci u osobnika płci męskiej), co oznacza, że w tych samych miejscach mamy te same geny. Mówiąc dokładniej, każdy gen może występować w różnych formach (mutacjach) zwanych allelami i chromosomy tej samej pary mogą się różnić allelami.

Rozważmy najprostszą sytuację, gdy gen ma dwa allele; zwyczajowo oznaczamy je A i a. Mamy więc trzy pary AA, Aa i aa, zwane genotypami (nie rozróżniamy par Aa i aA). W klasycznym modelu Mendla gen A odpowiada za kwiaty czerwone, a gen a za białe. Zatem w przypadku pary AA kwiaty są czerwone, w przypadku pary aa białe, a w przypadku pary Aa różowe.

Występują również allele dominujące (oznaczane zwyczajowo wielką literą) lub recesywne. Wtedy kwiaty mogą być tylko albo koloru czerwonego, dla pary AA (homozygota dominująca) i Aa (heterozygota), albo koloru białego, dla pary aa (homozygota recesywna). Recesywność polega więc na tym, że cecha kodowana przez allel a ujawni się tylko dla pary aa.

Komórki rozrodcze mają pojedyncze allele. Zakładamy, że jest to jeden z pary alleli wybrany z prawdopodobieństwem $1/2$. Zatem w organizmach, w których występują pary chromosomów (np. u człowieka), mamy pary alleli pochodzące od przodka męskiego i żeńskiego.

Rozważmy teraz procesy dziedziczenia. Mamy pewną populację i zakładamy, że osobniki w tej populacji mogą się łączyć w pary, przy czym wszystkie wybory pary rodziców są wzajemnie niezależne. Jest to model wyidealizowany, ale dość dobrze opisuje wiele naturalnych populacji, a przynajmniej przekazywanie niektórych cech. Załóżmy, że w danym pokoleniu trzy genotypy AA, Aa i aa występują w populacji z odpowiednimi częstościami x, $2y$ i z. Zatem $x + 2y + z = 1$ oraz allel A występuje z prawdopodobieństwem $p = x + y$, zaś allel a z prawdopodobieństwem $q = 1 - p = y + z$. Komórki rozrodcze z prawdopodobieństwem p będą zawierać allel A, a z prawdopodobieństwem q allel a. Konsekwencją losowego kojarzenia się par będzie rozkład par alleli AA, Aa i aa w następnym pokoleniu z prawdopodobieństwami

$$x_1 = p^2, \quad 2y_1 = 2pq, \quad z_1 = q^2. \tag{3.1}$$

Zauważmy, że allel A występuje w nowym pokoleniu z częstością

$$x_1 + y_1 = p^2 + pq = p(p + q) = p,$$

a więc rozkład alleli się nie zmienił. W kolejnych pokoleniach rozkład genotypów, czyli par alleli, jest również postaci (3.1). Możemy, więc stwierdzić, że rozkład (3.1) jest stabilny i ustabilizował się po jednym pokoleniu. Fakt ten został po raz pierwszy zauważony w roku 1908 przez G. H. Hardy'ego oraz niezależnie przez W. Weinberga i nosi nazwę *prawa Hardy'ego–Weinberga*. Możemy zatem przyjąć, że rozkład genotypów w dużej populacji jest bliski rozkładowi stabilnemu i zależy jedynie od p, a więc częstości występowania alleli w populacji.

Uwaga I.12. Zauważmy, że podobny rezultat otrzymamy, jeżeli gen ma więcej niż dwa allele. Jeżeli mamy n alleli $A_1, \ldots, A_n$ tego samego genu, to będziemy mieli $\binom{n+1}{2}$ różnych par alleli. Jeżeli p_i jest częstością występowania i-tego allela w populacji, to w następnym pokoleniu mamy p_i^2 par A_iA_i oraz $2p_ip_j$ par A_iA_j dla $i \neq j$.

Uwaga I.13. Niektóre cechy zależą od dwóch lub więcej par genów. Ponieważ dla jednej pary mamy trzy genotypy, więc przy dwóch parach mamy $9 = 3 \times 3$ genotypów $AABB$, $AABb, \ldots, aabb$, a przy n-parach 3^n genotypów. Gdy wszystkie geny leżą na tej samej parze chromosomów, pojawia się następujący problem. Genotyp $AaBb$ może być realizowany na dwa sposoby: albo na chromosomach mamy pary alleli AB i ab, albo Ab i aB. Jeżeli populacja w pierwszym pokoleniu składa się wyłącznie z osobników o parach alleli AB i ab, to w następnych pokoleniach będą tylko genotypy $AABB$, $AaBb$ i $aabb$; dla populacji o parach alleli Ab i aB w następnych pokoleniach będą genotypy $AAbb$, $aaBB$ i $AaBb$. Rozkład genotypów w następnych pokoleniach może więc zależeć od sposobu rozmieszczenia alleli na chromosomach.

Problem badania rozkładu genotypów, gdy wszystkie geny leżą na tej samej parze chromosomów, można sprowadzić do rozważania pojedynczego genu, ale z większą liczbą alleli. Gdy np. rozważamy dwie pary genów, zastępujemy je pojedynczymi genami z allelami $A_1 = AB$, $A_2 = Ab$, $A_3 = aB$ i $A_4 = ab$, mamy więc formalnie 10 genotypów A_iA_j, $1 \leq i \leq j \leq 4$. W szczególności osobniki o genotypie $AaBb$ formalnie dzielimy na dwa genotypy A_1A_4 i A_2A_3. Na podstawie prawa Hardy'ego–Weinberga rozkład genotypów ustabilizuje się już w następnym pokoleniu, a częstość występowania genotypu $AaBb$ w kolejnych pokoleniach jest sumą częstości genotypów A_1A_4 i A_2A_3. Również gdy geny występują na różnych chromosomach, rozkłady genotypów dążą do stanu równowagi (stabilnego), choć nie następuje to w jednym pokoleniu (patrz zad. I.20).

Uwaga I.14. Nawet przy całkowicie losowym kojarzeniu par następują małe losowe zmiany częstości występowania alleli, a więc wielkości p i q. Zmiany te bez dodatkowych innych czynników mogą doprowadzić po wielu pokoleniach do całkowitej eliminacji jednego z alleli. Dodatkowe zjawiska związane z mutacją i selekcją często jednak zapobiegają takim zmianom.

3.2. Cechy związane z płcią

Niektóre cechy i choroby wiążą się z płcią człowieka. Płeć determinowana jest występowaniem pary chromosomów XX u kobiety i XY u mężczyzny. Chromosom Y jest znacznie krótszy od chromosomu X i nie ma wielu genów występujących na chromosomie X, więc genotypy związane z chromosomami płci mogą być u kobiet i mężczyzn inne. Geny występujące tylko na chromosomie X nazywamy *genami skojarzonymi płciowo*.

Jeżeli gen występuje na chromosomie X, a nie występuje na chromosomie Y, to osobniki płci żeńskiej mają genotypy AA, Aa i aa, a męskiej A i a. Załóżmy losowe kojarzenie się par. Przyjmijmy, że genotypy AA, Aa i aa występują z odpowiednimi częstościami x, $2y$ i z. Allel A występuje więc z prawdopodobieństwem $p = x + y$, a allel a z prawdopodobieństwem $q = 1 - p = y + z$. Oznaczmy przez p' i q' częstości występowania genotypów męskich A i a. W następnym pokoleniu osobnik płci męskiej otrzymuje swój chromosom X od matki, a więc prawdopodobieństwa genotypów męskich wynoszą odpowiednio

$$p_1' = p, \quad q_1' = q.$$

Ponieważ zakładamy niezależne dobieranie się w pary, genotypy żeńskie AA, Aa i aa w następnym pokoleniu występują z prawdopodobieństwami

$$x_1 = pp', \quad 2y_1 = pq' + qp', \quad z_1 = qq'.$$

Stąd

$$p_1 = x_1 + y_1 = \frac{1}{2}(pp' + pq' + pp' + qp') = \frac{1}{2}(p + p'),$$
$$q_1 = y_1 + z_1 = \frac{1}{2}(pq' + qq' + qp' + qq') = \frac{1}{2}(q' + q).$$

Rozważając kolejne pokolenia, otrzymamy następujące wzory rekurencyjne:

$$p_n = \frac{1}{2}(p_{n-1} + p'_{n-1}), \ q_n = \frac{1}{2}(q_{n-1} + q'_{n-1}), \ p'_n = p_{n-1}, \ q'_n = q_{n-1}. \tag{3.2}$$

Przez podstawienie wykazujemy, że

$$p_n = \alpha + (-1)^n \frac{p - p'}{3 \cdot 2^n}, \quad q_n = \beta + (-1)^n \frac{q - q'}{3 \cdot 2^n}, \tag{3.3}$$

gdzie

$$\alpha = \frac{1}{3}(2p + p'), \quad \beta = \frac{1}{3}(2q + q').$$

Zatem $p_n \to \alpha$, $p'_n \to \alpha$, $q_n \to \beta$, $q'_n \to \beta$ i możemy przyjąć, że w dużej ustabilizowanej populacji genotypy męskie A i a występują z częstościami α i β, a genotypy żeńskie AA, Aa i aa z częstościami α^2, $2\alpha\beta$ i β^2, gdzie $\alpha + \beta = 1$.

Cecha związana z genem recesywnym, np. daltonizm, występuje u mężczyzn dla genotypu a, a u kobiet dla genotypu aa. Częstość takiej cechy u mężczyzn wynosi zatem β, a u kobiet β^2, co tłumaczy, dlaczego daltonizm występuje znacznie częściej u mężczyzn niż u kobiet.

3.3. Selekcja

Często procesom krzyżowania towarzyszą procesy selekcji, polegające na tym, że niektóre genotypy są z pewną częstością eliminowane z populacji. Selekcja może być naturalna, jeżeli np. osobniki o ustalonym genotypie nie mogą się rozmnażać, ale może też być prowadzona w celu otrzymania osobników o pożądanych cechach. Inny rodzaj selekcji, polegający na tym, że genotypy łączą się w pary z niejednakowym prawdopodobieństwem, może występować u ludzi z przyczyn kulturowych lub religijnych.

Rozważymy teraz przykład procesu selekcji. Przyjmijmy, że genotyp aa jest eliminowany z populacji. Jeżeli genotypy AA, Aa i aa są rozłożone z częstościami x, $2y$ i z, to proces selekcji prowadzi do rozkładu tych genotypów w procesie krzyżowania z częstościami

$$x^* = \frac{x}{1-z}, \quad 2y^* = \frac{2y}{1-z}, \quad z^* = 0.$$

Wtedy częstość występowania gamet z allelami A i a wynosi odpowiednio

$$p = \frac{x+y}{1-z}, \quad q = \frac{y}{1-z}.$$

Załóżmy losowe kojarzenie się par z uwzględnieniem wcześniejszej selekcji. Wtedy w następnym pokoleniu genotypy AA, Aa i aa są rozłożone z częstościami

$$x_1 = p^2, \quad 2y_1 = 2pq, \quad z_1 = q^2.$$

W ten sposób otrzymujemy wzory rekurencyjne na prawdopodobieństwa występowania genotypów oraz alleli w kolejnych pokoleniach:

$$p_n = \frac{x_n + y_n}{1-z_n}, \quad q_n = \frac{y_n}{1-z_n},$$
$$x_{n+1} = p_n^2, \quad 2y_{n+1} = 2p_nq_n, \quad z_{n+1} = q_n^2.$$

Stąd

$$p_{n+1} = \frac{x_{n+1} + y_{n+1}}{1-z_{n+1}} = \frac{p_n}{1-q_n^2} = \frac{1}{1+q_n}$$

oraz

$$q_{n+1} = \frac{y_{n+1}}{1 - z_{n+1}} = \frac{q_n}{1 + q_n}.$$

Z ostatniego wzoru otrzymujemy

$$\frac{1}{q_{n+1}} = 1 + \frac{1}{q_n},$$

a więc

$$q_n = \frac{q}{1 + nq}, \quad z_{n+1} = \left(\frac{q}{1 + nq}\right)^2.$$

Zauważmy, że w ten sposób następuje eliminacja allelu a, ale w dość wolnym tempie. Prawdopodobieństwo występowania allelu a jest w przybliżeniu proporcjonalne do n^{-1}, podczas gdy dla genotypu aa mamy proporcjonalność do n^{-2}.

Jeżeli procesowi selekcji towarzyszy proces mutacji, powodujący zmiany allelu A na a, to nawet tak drastyczna selekcja nie musi prowadzić do eliminacji allelu a. Procesy selekcji przebiegają znacznie szybciej, jeżeli niepożądana cecha związana jest z płcią (patrz zad. I.22–I.25).

3.4. Słownik terminów z genetyki

Zbierzemy teraz w postaci słownika niezbędne informacje dotyczące DNA.

allel jedna z form genu występująca w określonym miejscu na chromosomie. Allele tego samego genu różnią się jednym lub kilkoma nukleotydami. Rozróżniamy allele dominujące (oznaczane zwyczajowo wielką literą, np. A) lub recesywne (oznaczane małą literą, np. a). W organizmach, w których występują pary chromosomów (np. u człowieka), mamy pary alleli pochodzące od przodka męskiego i żeńskiego, a więc trzy kombinacje: aa (tzw. homozygota recesywna), aA lub Aa (heterozygota) oraz AA (homozygota dominująca). Recesywność polega na tym, że cecha kodowana przez allel a ujawni się tylko w przypadku aa.

chromosom forma organizacji materiału genetycznego wewnątrz komórki. Chromosomy dzielą się na autosomy związane z dziedziczeniem cech niesprzężonych z płcią oraz chromosomy płciowe (allosomy lub heterosomy), które występują u konkretnej płci. Człowiek ma 23 pary chromosomów, w tym jedną parę chromosomów płciowych (u kobiet złożoną z dwóch chromosomów X, u mężczyzn z chromosomu X i chromosomu Y). Z reguły chromosomy są przekazywane jako odrębne jednostki z całym zestawem genów (alleli).

DNA (skrót od *kwas deoksyrybonukleinowy*) nośnik informacji genetycznej organizmów żywych, zbudowany jest z liniowo ułożonych *nukleotydów*. Najczęściej występuje DNA dwuniciowe, zbudowane z połączonych ze sobą łańcuchów nukleotydów owijających się wokół wspólnej osi i tworzących tzw. prawoskrętną podwójną helisę.

eukarioty organizmy zbudowane z komórek posiadających jądro komórkowe z chromosomami, co odróżnia je od **prokariotów**, których komórki nie zawierają jądra komórkowego.

fenotyp zespół cech organizmu. Oddziaływanie między genotypem a środowiskiem daje fenotyp.

gen fragment DNA o zdolności tworzenia jakiegoś RNA lub białka. Gen jest podstawową jednostką dziedziczenia.

genom materiał genetyczny zawarty w pojedynczym zespole chromosomów. Na przykład komórki człowieka posiadają dwa genomy, z wyjątkiem komórek rozrodczych, które zawierają po jednym genomie. W przypadku eukariotów genom odnosi się do DNA zawartego w jądrze komórki.

genotyp układ alleli danego osobnika warunkujący jego właściwości dziedziczne.

kodon jednostka w sekwencji mRNA (*informacyjne RNA*) składająca się z trzech nukleotydów (odpowiadających zasadom A, C, G, U) kodujących określony aminokwas. Istnieją $4^3 = 64$ kodony, z czego 60 określa 20 aminokwasów (różne kodony mogą określać ten sam aminokwas); kodon AUG (metionina) inicjuje translację, a trzy kodony kończą translację. Dzięki temu, że występuje kodon inicjujący, każdy nukleotyd wchodzi w skład jednego kodonu, nie trzeba oddzielać kodonów „przecinkami", kodonom przyporządkowane są jednoznacznie aminokwasy, a ich kolejność ułożenia w białku odpowiada kolejności w mRNA. Również DNA zbudowane jest z trójek nukleotydowych (zawierających zasady A, C, G, T).

nukleotyd podstawowy składnik DNA i RNA. Nukleotyd zbudowany jest z cukru – w przypadku DNA *deoksyrybozy*, a w przypadku RNA *rybozy* – oraz jednej z *zasad azotowych*.

RNA (skrót od *kwasy rybonukleinowe*) cząsteczki organiczne zbudowane zwykle z jednej nici rybonukleotydów, używane między innymi w transkrypcji DNA i w syntezie białek.

zasady azotowe w przypadku DNA: adenina A i guanina G (zasady purynowe) oraz cytozyna C i tymina T (zasady pirymidynowe); w RNA i w DNA niektórych wirusów zamiast tyminy T występuje uracyl U. W DNA dwuniciowym zasady leżące naprzeciwko siebie połączone są według wzoru A-T lub G-C.

4. Zmienne losowe i ich własności

4.1. Pojęcie zmiennej losowej i jej rozkładu

W teorii prawdopodobieństwa kluczową rolę odgrywa pojęcie zmiennej losowej. Niech $(\Omega, \Sigma, \mathrm{P})$ będzie dowolną przestrzenią probabilistyczną. *Zmienną losową* nazywamy funkcję mierzalną $\xi\colon \Omega \to \mathbb{R}$ (symbol $\mathbb{R}$ oznacza zbiór liczb rzeczywistych), gdzie zbiór $\mathbb{R}$ rozpatrujemy wraz z σ-algebrą zbiorów borelowskich w $\mathbb{R}$.

Przypominamy, że jeżeli $(X, \mathcal{A})$ i $(Y, \mathcal{B})$ są przestrzeniami mierzalnymi z σ-algebrami $\mathcal{A}$ i $\mathcal{B}$, to odwzorowanie $f\colon X \to Y$ nazywamy *funkcją mierzalną*, jeżeli $f^{-1}(B) \in \mathcal{A}$ dla dowolnego zbioru $B \in \mathcal{B}$. Ponadto dla każdej przestrzeni topologicznej można rozpatrywać najmniejszą σ-algebrę zawierającą zbiory otwarte; jej elementy nazywamy *zbiorami borelowskimi*, a σ-algebrę zbiorów borelowskich w przestrzeni X oznaczamy przez $\mathcal{B}(X)$.

Zatem funkcja $\xi\colon \Omega \to \mathbb{R}$ jest zmienną losową, jeżeli $\xi^{-1}(B) \in \Sigma$ dla dowolnego zbioru $B \in \mathcal{B}(\mathbb{R})$.

Oprócz pojęcia zmiennej losowej wprowadza się ogólniejsze pojęcia wektora losowego i elementu losowego. Niech $(X, \mathcal{A})$ będzie przestrzenią mierzalną. Wtedy odwzorowanie mierzalne $\xi\colon \Omega \to X$ nazywamy *elementem losowym*. Zwykle rozważamy elementy losowe o wartościach w przestrzeni topologicznej X wyposażonej w σ-algebrę $\mathcal{B}(X)$. Jeżeli $X = \mathbb{R}^n$, to ξ nazywamy n-wymiarowym *wektorem losowym* lub n-wymiarową zmienną losową. Wektor losowy ma postać $\xi(\omega) = (\xi_1(\omega), \dots, \xi_n(\omega))$, gdzie ξ_i są zmiennymi losowymi dla $i = 1, \dots, n$.

W praktyce często posługujemy się zmiennymi losowymi bez znajomości przestrzeni probabilistycznej, na której są określone. Istotna jest znajomość rozkładu zmiennej losowej, a więc z jakim prawdopodobieństwem zmienna losowa przyjmuje wartości w interesujących nas zbiorach.

Rozkładem zmiennej losowej nazywamy miarę probabilistyczną μ określoną na σ-algebrze $\mathcal{B}(\mathbb{R})$ wzorem $\mu(B) = \mathrm{P}(\xi \in B)$ dla $B \in \mathcal{B}(\mathbb{R})$. Zapis $\mathrm{P}(\xi \in B)$ jest skróconą formą wzoru $\mathrm{P}(\{\omega \in \Omega\colon \xi(\omega) \in B\})$. Podobnie definiujemy rozkład dla wektorów i elementów losowych. Formalnie pojęcie rozkładu prawdopodobieństwa może funkcjonować bez odwoływania się do zmiennej losowej, jako miara probabilistyczna na $\mathbb{R}$ lub na ogólniejszej przestrzeni, np. $\mathbb{R}^n$.

Rozkład zmiennej losowej można wyznaczać, korzystając z pojęcia dystrybuanty. *Dystrybuantą* zmiennej losowej nazywamy funkcję $F_\xi\colon \mathbb{R} \to [0, 1]$ określoną wzorem

$$F_\xi(x) = \mathrm{P}(\xi \le x).$$

Niech F będzie dystrybuantą zmiennej losowej ξ. Funkcja F ma następujące własności:

(i) F jest niemalejąca: jeżeli $x_1 \le x_2$, to $F(x_1) \le F(x_2)$,

(ii) $\lim_{x\to-\infty} F(x) = 0$ i $\lim_{x\to\infty} F(x) = 1$,

(iii) F jest prawostronnie ciągła: $\lim_{x\to x_0^+} F(x) = F(x_0)$ dla dowolnego $x_0 \in \mathbb{R}$.

Na odwrót, jeżeli funkcja F spełnia warunki (i)–(iii), to jest dystrybuantą zmiennej losowej $\xi(\omega) = \omega$ określonej na przestrzeni probabilistycznej $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathrm{P})$, gdzie P jest miarą zdefiniowaną wzorem $\mathrm{P}((-\infty, x]) = F(x)$ dla $x \in \mathbb{R}$, z którego jednoznacznie można wyznaczyć $\mathrm{P}(B)$ dla $B \in \mathcal{B}(\mathbb{R})$.

Uwaga I.15. W niektórych podręcznikach dystrybuantę definiuje się wzorem $F_\xi(x) = \mathrm{P}(\xi < x)$. Definicja ta różni się od poprzedniej, jeżeli zmienna losowa przyjmuje pewną wartość z dodatnim prawdopodobieństwem.

Wśród rozkładów wyróżniamy rozkłady ciągłe i dyskretne. Mówimy, że zmienna losowa ξ ma *rozkład ciągły*, jeżeli jej dystrybuanta jest funkcją ciągłą. Jeżeli natomiast istnieje taki przeliczalny zbiór $S = \{x_1, x_2, \dots\}$, że $\mathrm{P}(\xi \in S) = 1$, to mówimy, że zmienna losowa jest *dyskretna* oraz że ma dystrybuantę i rozkład dyskretny. W tym wypadku rozkład ξ jest w pełni wyznaczony przez wartości $p_1 = \mathrm{P}(\xi = x_1)$, $p_2 = \mathrm{P}(\xi = x_2)$, ... następująco:

$$\mathrm{P}(\xi \in B) = \sum_{\{i:\, x_i \in B\}} p_i \quad \text{dla } B \in \mathcal{B}(\mathbb{R}). \tag{4.1}$$

Wśród zmiennych losowych o rozkładach ciągłych wyróżniamy zmienne mające gęstości. Mówimy, że zmienna losowa, jak również jej rozkład, ma *gęstość* względem miary Lebesgue'a, lub krótko: ma gęstość f, jeśli f jest taką nieujemną funkcją mierzalną określoną na prostej $\mathbb{R}$, że

$$\mathrm{P}(\xi \in B) = \int_B f(x)\,dx \quad \text{dla } B \in \mathcal{B}(\mathbb{R}). \tag{4.2}$$

Podamy teraz kilka przykładów rozkładów dyskretnych oraz rozkładów mających gęstości. Najprostszym rozkładem dyskretnym jest rozkład *jednostajny*, gdy zmienna losowa przyjmuje skończoną liczbę wartości z tym samym prawdopodobieństwem. Do podstawowych rozkładów dyskretnych należą rozkład dwumianowy (Bernoulliego), rozkład Poissona, rozkład geometryczny, rozkład potęgowy oraz rozkład logarytmiczny.

Przykład I.16 (rozkład dwumianowy (Bernoulliego))**.** Wykonujemy serię n niezależnych doświadczeń, w których prawdopodobieństwo sukcesu wynosi $p \in (0,1)$. Zmienna losowa ξ wyraża liczbę sukcesów. Wtedy ξ ma *rozkład dwumianowy* (zwany też *rozkładem Bernoulliego*):

$$\mathrm{P}(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{dla } k = 0, 1, \dots, n. \tag{4.3}$$

Przykład I.17 (rozkład dwumianowy mieszany)**.** W diagnostyce niektóre testy wykonywane są kilkakrotnie, gdy np. mamy kilka różnych próbek badanej tkanki lub płynów ustrojowych. Rozważmy najprostszy model, w którym badanych dzieli się na osoby zdrowe i chore; niech q będzie prawdopodobieństwem, że losowo wybrana osoba z rozpatrywanej próby jest zdrowa. Test daje wynik pozytywny z prawdopodobieństwem p_1 dla osób zdrowych i p_2 dla osób chorych. Wybieramy osobę i wykonujemy serię n niezależnych testów. Zmienna losowa ξ wyraża liczbę wyników pozytywnych. Wtedy ξ ma *rozkład dwumianowy mieszany* (lub *ważony*):

$$\mathrm{P}(\xi = k) = q\binom{n}{k} p_1^k (1-p_1)^{n-k} + (1-q)\binom{n}{k} p_2^k (1-p_2)^{n-k} \text{ dla } k = 0, 1, \ldots, n.$$

Korzystając z reguły Bayesa (2.5), możemy wyznaczać prawdopodobieństwo, że osoba jest zdrowa lub chora, w zależności od liczby wyników pozytywnych (patrz zad. I.26).

Pojęcie rozkładu dwumianowego mieszanego można uogólnić, rozważając większą liczbę rozłącznych grup osób (warunków itp.); wtedy

$$\mathrm{P}(\xi = k) = \sum_{i=1}^{s} q_i \binom{n}{k} p_i^k (1-p_i)^{n-k} \quad \text{dla } k = 0, 1, \ldots, n, \tag{4.4}$$

gdzie $q_1, \ldots, q_s$ są nieujemnymi liczbami rzeczywistymi spełniającymi warunek $q_1 + \cdots + q_s = 1$, a $p_i \in [0,1]$ dla $i = 1, \ldots, s$. Można też rozpatrywać rozkład dwumianowy mieszany, w którym występuje przeliczalna liczba rozkładów dwumianowych.

Przykład I.18 (rozkład Poissona)**.** Przypuśćmy, że podobnie jak dla rozkładu dwumianowego wykonujemy serię n niezależnych doświadczeń, w których prawdopodobieństwo sukcesu wynosi $p_n = \lambda/n$, gdzie λ jest ustaloną liczbą dodatnią. Można łatwo wykazać, że

$$\lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda} \tag{4.5}$$

dla dowolnego $k \in \mathbb{N}$ (przez $\mathbb{N}$ oznaczamy zbiór liczb naturalnych, tzn. nieujemnych liczb całkowitych). Wzór po prawej stronie wyraża zatem przybliżone prawdopodobieństwo k sukcesów w n niezależnych doświadczeniach, gdy n jest dostatecznie wielkie, a prawdopodobieństwo sukcesu w pojedynczym doświadczeniu jest małe i wynosi $p = \lambda/n$. Wyrażenie po prawej stronie wzoru (4.5) wyznacza rozkład prawdopodobieństwa na zbiorze liczb naturalnych, zwany *rozkładem Poissona*.

Przykład I.19 (rozkład geometryczny)**.** Znów wykonujemy serię niezależnych doświadczeń, w których prawdopodobieństwo sukcesu wynosi $p \in (0,1)$. Serię

przerywamy, jeżeli doświadczenie zakończyło się porażką. Niech zmienna losowa ξ oznacza liczbę sukcesów, przy czym jeżeli pierwsze doświadczenie zakończyło się porażką, to przyjmujemy $\xi = 0$. Zmienna losowa ξ ma *rozkład geometryczny* wyrażony wzorem

$$\mathrm{P}(\xi = k) = (1-p)p^k \quad \text{dla dowolnego } k \in \mathbb{N}. \tag{4.6}$$

Przykład I.20 (rozkład hipergeometryczny)**.** W wielu zastosowaniach występuje rozkład hipergeometryczny, który wygodnie jest wprowadzić, rozpatrując następujące zadanie kombinatoryczne. W urnie mamy N kul, wśród których jest B kul białych i C kul czarnych, przy czym $B + C = N$. Z urny wylosowano próbkę liczącą K kul. Należy wyznaczyć prawdopodobieństwo, że w wylosowanej próbie jest dokładnie n kul białych.

Zauważmy, że różnych próbek K-elementowych jest $\binom{N}{K}$. Ze zbioru kul białych możemy wybrać n kul na $\binom{B}{n}$ sposobów, a ze zbioru kul czarnych możemy wybrać $K-n$ kul na $\binom{C}{K-n}$ sposobów. Stąd wśród próbek K-elementowych mamy $\binom{B}{n}\binom{C}{K-n}$ różnych próbek zawierających n kul białych. Rozkład zmiennej losowej ξ wyznaczającej liczbę kul białych w próbce K-elementowej określony jest więc wzorem

$$\mathrm{P}(\xi = n) = \frac{\binom{B}{n}\binom{C}{K-n}}{\binom{N}{K}}. \tag{4.7}$$

Przyjmujemy tu konwencję $\binom{m}{k} = 0$, gdy $k < 0$ lub $k > m$. Rozkład (4.7) nazywamy *rozkładem hipergeometrycznym*.

Uwaga I.21. Rozkłady, w których występuje symbol silni (np. rozkład Bernoulliego, Poissona i hipergeometryczny), są niewygodne do obliczeń i często korzysta się ze wzorów asymptotycznych, wystarczających w wielu zastosowaniach.

Dwa ciągi (a_n) i (b_n) o wyrazach różnych od zera nazywamy *równoważnymi* i piszemy $a_n \sim b_n$, jeżeli $\lim_{n\to\infty} a_n/b_n = 1$. Wygodnie jest używać następującego *wzoru Stirlinga*:

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n}. \tag{4.8}$$

Wykażemy, że jeżeli $k = cn$, gdzie $c \in (0,1)$, to

$$\begin{aligned} &\binom{n}{k} \sim \frac{1}{\sqrt{2\pi n c(1-c)}}\, e^{nH(c)}, \\ &H(c) = -c\ln c - (1-c)\ln(1-c). \end{aligned} \tag{4.9}$$

Istotnie, korzystając ze wzoru Stirlinga, otrzymujemy

$$\begin{aligned}\binom{n}{k} &= \frac{n!}{k!(n-k)!} \sim \frac{\sqrt{2\pi n}}{\sqrt{2\pi k}\sqrt{2\pi(n-k)}} \frac{n^n}{k^k(n-k)^{n-k}} \frac{e^{-n}}{e^{-k}e^{-(n-k)}} \\ &= \frac{1}{\sqrt{2\pi nc(1-c)}} \exp\{n\ln n - k\ln k - (n-k)\ln(n-k)\} \\ &= \frac{1}{\sqrt{2\pi nc(1-c)}} \exp\{n\ln n - cn\ln(cn) - (1-c)n\ln((1-c)n))\} \\ &= \frac{1}{\sqrt{2\pi nc(1-c)}} \exp\{-cn\ln c - (1-c)n\ln(1-c)\}.\end{aligned}$$

Przykład I.22 (rozkład potęgowy)**.** Ustalmy liczbę rzeczywistą $a > 1$ i liczbę naturalną $n > 0$. Niech $c_{a,n}$ będzie taką stałą, że $c_{a,n}\sum_{k=n}^{\infty} k^{-a} = 1$. Jeżeli zmienna losowa ξ o wartościach w zbiorze $\{n, n+1, \dots\}$ spełnia warunek

$$\mathrm{P}(\xi = k) = c_{a,n}k^{-a} \quad \text{dla dowolnego } k \geq n, \tag{4.10}$$

to mówimy, że ξ ma *rozkład potęgowy*. Trudno jest podać prostą interpretację rozkładu potęgowego, tak jak w przypadku rozkładu dwumianowego, Poissona, czy geometrycznego. Mimo to rozkład potęgowy wydaje się być najlepszym przybliżeniem wielu rozkładów empirycznych, między innymi w sieciach biologicznych i w biologii genomu [201].

Przykład I.23 (rozkład logarytmiczny)**.** Ustalmy $p \in (0,1)$. Jeżeli zmienna losowa ξ o wartościach w zbiorze $\{1, 2, \dots\}$ spełnia warunek

$$\mathrm{P}(\xi = k) = c_p\frac{p^k}{k} \quad \text{dla dowolnego } k \geq 1, \tag{4.11}$$

gdzie $c_p = -\frac{1}{\ln(1-p)}$, to mówimy, że ξ ma *rozkład logarytmiczny*. Rozkład logarytmiczny pojawił się w pracy [118] jako model opisu rozkładu względnej liczebności kolejnych gatunków na danym obszarze. Opisuje on również rozkład rodzin paralogów w genomie [330, 355].

Podamy teraz kilka przykładów rozkładów ciągłych.

Przykład I.24 (rozkład jednostajny)**.** Niech a i b będą liczbami rzeczywistymi i $a < b$. Rozkład o gęstości

$$f(x) = \frac{1}{b-a}\mathbf{1}_{[a,b]}(x) \tag{4.12}$$

nazywamy *rozkładem jednostajnym* na przedziale $[a, b]$. Rozkład ten ma

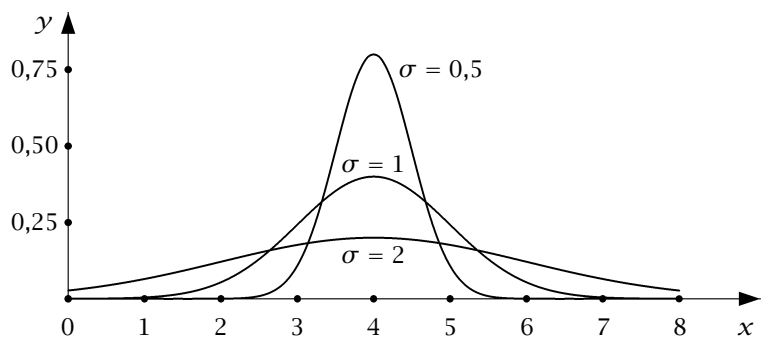
dystrybuantę

$$F(x) = \begin{cases} 0, & \text{gdy } x < a, \\ \dfrac{x-a}{b-a}, & \text{gdy } a \le x \le b, \\ 1, & \text{gdy } x > b. \end{cases} \tag{4.13}$$

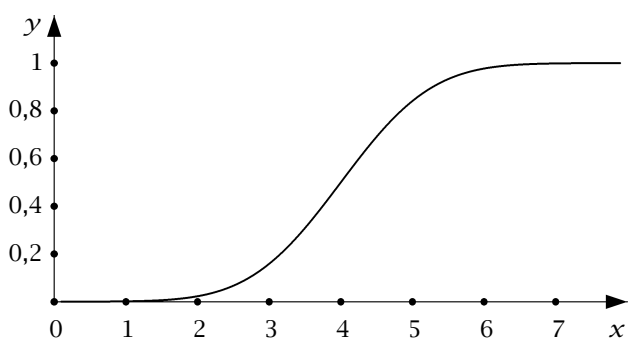
Przykład I.25 (rozkład normalny (Gaussa))**.** Niech m będzie liczbą rzeczywistą, a σ liczbą dodatnią. Rozkład o gęstości

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right) \tag{4.14}$$

nazywamy *rozkładem normalnym* (lub *gaussowskim*) o parametrach m, σ i oznaczamy przez $\mathcal{N}(m, \sigma^2)$. Jeżeli $m = 0$ i $\sigma = 1$, to rozkład normalny nazywamy *standardowym*, a jego dystrybuantę oznaczamy przez $\Phi(x)$. Na rysunku I.11 przedstawione są wykresy gęstości rozkładu normalnego dla różnych parametrów σ. Rysunek I.12 przedstawia wykres dystrybuanty rozkładu normalnego o parametrach $m = 4$ i $\sigma = 1$.



Rysunek I.11. Gęstość rozkładu normalnego dla $m = 4$ i $\sigma = 0{,}5$, $\sigma = 1$, $\sigma = 2$



Rysunek I.12. Dystrybuanta rozkładu normalnego o parametrach $m = 4$ i $\sigma = 1$

Przykład I.26 (rozkład gamma)**.** Rozkład o gęstości

$$f(x) = \begin{cases} \dfrac{\alpha^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\alpha x}, & \text{gdy } x \geq 0, \\ 0, & \text{gdy } x < 0, \end{cases} \tag{4.15}$$

nazywamy *rozkładem gamma* o parametrach $\alpha > 0$ i $\lambda > 0$, gdzie Γ jest *funkcją gamma* określoną wzorem

$$\Gamma(\lambda) = \int_0^\infty x^{\lambda-1} e^{-x}\, dx, \quad \lambda > 0. \tag{4.16}$$

Należy wspomnieć, że

$$\Gamma(x+1) = x\Gamma(x) \quad \text{dla } x > 0 \text{ oraz } \Gamma(n+1) = n! \text{ dla } n \in \mathbb{N}. \tag{4.17}$$

Rozkład gamma z $\lambda = 1$ nazywamy *rozkładem wykładniczym*. Rozkład wykładniczy ma więc gęstość

$$f(x) = \begin{cases} \alpha e^{-\alpha x}, & \text{gdy } x \geq 0, \\ 0, & \text{gdy } x < 0, \end{cases} \tag{4.18}$$

i dystrybuantę

$$F(x) = \begin{cases} 0, & \text{gdy } x < 0, \\ 1 - e^{-\alpha x}, & \text{gdy } x \geq 0. \end{cases} \tag{4.19}$$

Wykresy gęstości rozkładu gamma dla różnych λ przedstawione są na rysunku I.13.

Przykład I.27 (rozkład potęgowy (Pareto))**.** Niech x_m i k będą dodatnimi liczbami rzeczywistymi. Rozkład o dystrybuancie

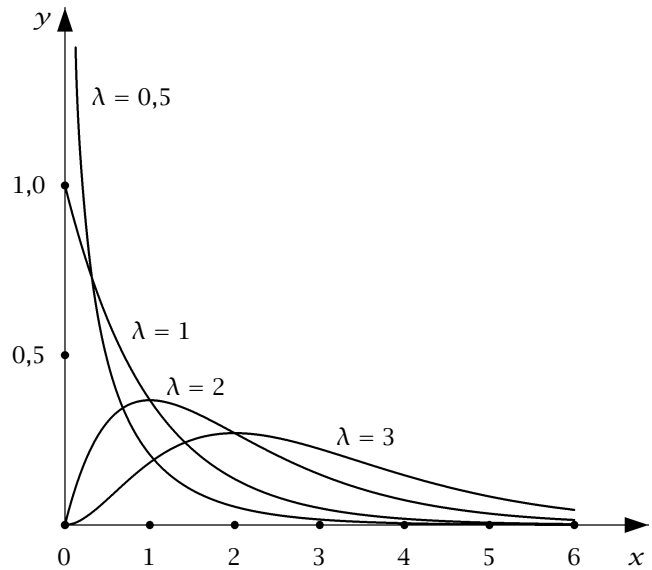
$$F(x) = \begin{cases} 1 - \left(\dfrac{x_m}{x}\right)^k, & \text{gdy } x \geq x_m, \\ 0, & \text{gdy } x < x_m, \end{cases} \tag{4.20}$$

nazywamy *rozkładem potęgowym* lub *rozkładem Pareto* o parametrze skali x_m i parametrze kształtu k.

Przykład I.28 (rozkład Cauchy'ego)**.** Rozkład o gęstości

$$g(x) = \frac{1}{\pi(1+x^2)} \tag{4.21}$$

nazywamy standardowym *rozkładem Cauchy'ego*. Jego dystrybuanta wyraża się wzorem $F(x) = \frac{1}{2} + \frac{1}{\pi} \operatorname{arctg} x$.



Rysunek I.13. Rozkład gamma dla $\alpha = 1$ i $\lambda = 0{,}5$, $\lambda = 1$, $\lambda = 2$, $\lambda = 3$

4.2. Charakterystyki zmiennych losowych

Wartością oczekiwaną lub *momentem pierwszego rzędu* zmiennej losowej ξ nazywamy liczbę

$$\mathrm{E}\,\xi = \int_{\Omega} \xi(\omega)\,\mathrm{P}(d\omega), \tag{4.22}$$

o ile całka po prawej stronie istnieje. Definicja wartości oczekiwanej przenosi się automatycznie na zmienne losowe o wartościach zespolonych i wektory losowe. Wartość oczekiwana wektora losowego $\xi = (\xi_1, \ldots, \xi_n)$ wynosi $\mathrm{E}\,\xi = (\mathrm{E}\,\xi_1, \ldots, \mathrm{E}\,\xi_n)$.

Wartość oczekiwaną zmiennej losowej możemy wyznaczyć, korzystając z jej dystrybuanty:

$$\mathrm{E}\,\xi = \int_{-\infty}^{\infty} x\,dF(x). \tag{4.23}$$

Całka występująca po prawej stronie wzoru (4.23) jest całką Stieltjesa względem funkcji F. Jeżeli zmienna losowa jest dyskretna, to

$$\mathrm{E}\,\xi = \sum_{i \in I} x_i p_i, \tag{4.24}$$

gdzie liczby x_i, $i \in I$, są wszystkimi wartościami, które przyjmuje zmienna ξ, a $p_i = \mathrm{P}(\xi = x_i)$. Jeśli zbiór I jest skończony, to wartość oczekiwana istnieje,

a jeżeli jest nieskończony, to wartość oczekiwana istnieje, o ile szereg po prawej stronie wzoru (4.24) jest bezwzględnie zbieżny. Jeśli zmienna losowa ξ ma gęstość f, to

$$\mathrm{E}\,\xi = \int_{-\infty}^{\infty} x f(x)\,dx. \tag{4.25}$$

Wzory (4.23)–(4.25) można uogólnić na inne funkcjonały od zmiennych losowych. Niech $g\colon \mathbb{R} \to \mathbb{R}$ będzie funkcją mierzalną. Wtedy

$$\mathrm{E}\,g(\xi) = \int_{-\infty}^{\infty} g(x)\,dF(x), \tag{4.26}$$

o ile ta całka istnieje, a jeżeli ξ ma gęstość f, to

$$\mathrm{E}\,g(\xi) = \int_{-\infty}^{\infty} g(x) f(x)\,dx. \tag{4.27}$$

Wariancją $\mathrm{D}^2\,\xi$ zmiennej losowej ξ nazywamy liczbę

$$\mathrm{D}^2\,\xi = \mathrm{E}(\xi - \mathrm{E}\,\xi)^2, \tag{4.28}$$

którą możemy także obliczać ze wzoru

$$\mathrm{D}^2\,\xi = \mathrm{E}\,\xi^2 - (\mathrm{E}\,\xi)^2, \tag{4.29}$$

o ile $\mathrm{E}\,\xi^2$ istnieje. Niekiedy wariancję oznacza się przez $\mathrm{Var}\,\xi$. Liczbę $\sigma = \sqrt{\mathrm{D}^2\,\xi}$ nazywamy *odchyleniem standardowym*.

Jeżeli $\mathrm{E}\,\xi = 0$ i $\mathrm{D}^2\,\xi = 1$, to zmienną losową nazywamy *standaryzowaną*. Każda zmienna losowa o skończonej wariancji wyznacza zmienną standaryzowaną za pomocą przekształcenia liniowego

$$\xi_1 = \frac{\xi - \mathrm{E}\,\xi}{\sqrt{\mathrm{D}^2\,\xi}}.$$

W analizie i prezentacji danych statystycznych ważną rolę odgrywa pojęcie mediany. Mówimy np., że mediana zarobków jakiejś grupy wynosi m, jeżeli połowa osób w grupie zarabia mniej niż m, a druga połowa więcej niż m. Pojęcie to można przenieść na zmienne losowe, przy czym definicja może nie być jednoznaczna. Jeżeli zmienna losowa ξ ma ciągłą dystrybuantę F, to *medianą* ξ nazywamy taką liczbę m, że $F(m) = 0{,}5$. Zauważmy, że m może nie być określone jednoznacznie, jeżeli funkcja F przyjmuje wartość 0,5 na pewnym odcinku. W tym wypadku można przyjąć, że mediana jest osiągana w środku tego odcinka. Jeszcze większy problem stwarza definicja mediany dla dyskretnych zmiennych losowych. Wtedy taki punkt m, że $F(m) = 0{,}5$, może nie istnieć.

Przyjmujemy, że dla dowolnej zmiennej losowej ξ *mediana* jest definiowana jako taki punkt m, że $\mathrm{P}(\xi \le m) \ge 0{,}5$ i $\mathrm{P}(\xi \ge m) \ge 0{,}5$, co w języku dystrybuanty wyraża się zależnościami $F(m^-) \le 0{,}5$ i $F(m) \ge 0{,}5$. Jeżeli rozkład jest symetryczny względem punktu $x = m$, to m jest jednocześnie wartością oczekiwaną (o ile istnieje) i medianą tego rozkładu.

Przykład I.29. Dla rozkładu dwumianowego z parametrem p i liczbą prób n mamy $\mathrm{E}\,\xi = pn$, $\mathrm{D}^2\,\xi = p(1-p)n$, a mediana jest jedną z liczb $[np]-1$, $[np]$, $[np]+1$.

Przykład I.30. Wartość oczekiwana i wariancja dla rozkładu Poissona z parametrem λ wynoszą λ.

Przykład I.31. Dla rozkładu geometrycznego z parametrem p, a więc gdy $\mathrm{P}(\xi = k) = (1-p)p^k$ dla $k \in \mathbb{N}$, mamy $\mathrm{E}\,\xi = p(1-p)^{-1}$, $\mathrm{D}^2\,\xi = p(1-p)^{-2}$, a mediana wynosi $m = -\big[\frac{1}{\log_2 p}\big] - 1$. Zauważmy, że różnica między wartością oczekiwaną i medianą jest duża dla $p \approx 1$ (patrz zad. I.33).

Przykład I.32. Dla zmiennej losowej o rozkładzie hipergeometrycznym (patrz przykład I.20) z parametrami B, $C = N - B$, K mamy

$$\mathrm{E}\,\xi = \frac{BK}{N}, \quad \mathrm{D}^2\,\xi = \frac{KB(N-B)(N-K)}{N^2(N-1)},$$

a mediana wynosi

$$m = \left[\frac{(B+1)(K+1)}{N+2}\right].$$

Przykład I.33. Rozkład jednostajny na przedziale $[a,b]$ ma wartość oczekiwaną $(a+b)/2$ i wariancję $(b-a)^2/12$.

Przykład I.34. Rozkład normalny z parametrami m i σ ma wartość oczekiwaną m i wariancję σ^2.

Przykład I.35. Rozkład gamma z parametrami α i λ ma wartość oczekiwaną λ/α i wariancję λ/α^2.

Przykład I.36. Rozkład potęgowy o parametrze skali x_m i parametrze kształtu k ma medianę $x_m \sqrt[k]{2}$. Wartość oczekiwana istnieje dla $k > 1$ i wynosi $\frac{x_m k}{k-1}$. Wariancja istnieje dla $k > 2$ i wynosi $\frac{x_m^2 k}{(k-1)^2(k-2)}$.

Przykład I.37. Standardowy rozkład Cauchy'ego ma medianę równą zeru, ale nie ma wartości oczekiwanej.

Ważną rolę w teorii prawdopodobieństwa odgrywa pojęcie funkcji charakterystycznej i funkcji tworzącej zmiennej losowej. Przedstawimy teraz definicje

obu pojęć, podamy bez dowodu ich własności oraz sformułujemy podstawowe twierdzenia dotyczące związku między funkcją charakterystyczną i rozkładem zmiennej losowej.

Funkcją charakterystyczną zmiennej losowej ξ o dystrybuancie F nazywamy funkcję o dziedzinie $\mathbb{R}$ i o wartościach zespolonych, określoną wzorem

$$\varphi_\xi(t) = \mathrm{E}\, e^{it\xi} = \int_{-\infty}^{\infty} e^{itx}\, dF(x). \tag{4.30}$$

Funkcja charakterystyczna jest więc transformatą Fouriera rozkładu zmiennej ξ, zgodnie z jedną z wielu definicji transformacji Fouriera. Jeżeli $\eta = \sigma\xi + m$, to wprost ze wzoru (4.30) otrzymujemy zależność

$$\varphi_\eta(t) = e^{itm}\varphi_\xi(\sigma t). \tag{4.31}$$

Przykład I.38. Jeżeli ξ jest zmienną dyskretną o wartościach x_k, $k \in I$, oraz $p_k = \mathrm{P}(\xi = x_k)$, to

$$\varphi_\xi(t) = \sum_{k\in I} p_k e^{itx_k}. \tag{4.32}$$

W szczególności dla rozkładu dwumianowego z parametrem p i liczbą prób n mamy $\varphi_\xi(t) = (1-p+pe^{it})^n$, zaś $\exp(\lambda e^{it} - \lambda)$ jest funkcją charakterystyczną rozkładu Poissona z parametrem λ.

Przykład I.39. Jeżeli zmienna losowa ξ ma gęstość

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

to $f'(x) = -xf(x)$, a wobec tego $\widehat{f'} = -\widehat{xf}$, gdzie $\hat{g}$ oznacza transformatę Fouriera funkcji g. Ponieważ $\widehat{f'}\,(t) = -it\hat{f}\,(t)$ i $\widehat{xf} = -i\hat{f}\,'(t)$, więc transformata funkcji f spełnia równanie różniczkowe

$$\hat{f}\,'(t) = -t\hat{f}(t)$$

z warunkiem początkowym $\hat{f}(0) = \mathrm{E}\,1 = 1$. Rozwiązaniem tego równania jest funkcja $\exp(-t^2/2)$, a więc $\varphi_\xi(t) = \exp(-t^2/2)$. Jeżeli zmienna losowa η ma rozkład $\mathcal{N}(m, \sigma^2)$, to ze wzoru (4.31) otrzymujemy równość $\varphi_\eta(t) = \exp(itm - (t\sigma)^2/2)$.

Jeżeli $\mathrm{E}\,|\xi|^k < \infty$ dla pewnego $k \in \mathbb{N}$, to korzystając z własności transformacji Fouriera, stwierdzamy, że istnieje ciągła k-ta pochodna funkcji φ_ξ oraz

$$\varphi_\xi^{(k)}(0) = i^k\, \mathrm{E}\,\xi^k. \tag{4.33}$$

Ponieważ funkcja charakterystyczna jest transformatą Fouriera rozkładu zmiennej losowej, więc jeśli zmienna losowa ma gęstość f, to ze wzoru na odwrotną transformację Fouriera otrzymujemy

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_\xi(t)\, dt. \tag{4.34}$$

W ogólnym przypadku mamy następujące twierdzenie pozwalające wyrazić dystrybuantę zmiennej losowej za pomocą funkcji charakterystycznej.

Twierdzenie I.40 (Lévy'ego)**.** *Jeżeli F jest dystrybuantą, a φ funkcją charakterystyczną zmiennej losowej ξ, to dla dowolnych punktów ciągłości x i y funkcji F mamy*

$$F(y) - F(x) = \frac{1}{2\pi} \lim_{T\to\infty} \int_{-T}^{T} \frac{e^{-itx} - e^{-ity}}{it} \varphi(t)\, dt. \tag{4.35}$$

Zauważmy, że funkcja $\varphi(t)/t$ może być niecałkowalna w $t = 0$ i wtedy całkę $\int_{-T}^{T}$ po prawej stronie wzoru (4.35) rozumiemy jako wartość główną całki niewłaściwej, czyli jako granicę $\lim_{\varepsilon\to 0^+} (\int_{-T}^{-\varepsilon} + \int_{\varepsilon}^{T})$.

Z funkcji charakterystycznych możemy korzystać przy badaniu zbieżności rozkładów. Mówimy, że ciąg (ξ_n) zmiennych losowych (lub ich rozkładów) jest *słabo zbieżny* lub *zbieżny według rozkładu* do zmiennej losowej ξ, jeżeli dla dowolnej funkcji ciągłej i ograniczonej g mamy $\lim_{n\to\infty} \mathrm{E}\, g(\xi_n) = \mathrm{E}\, g(\xi)$. W ten sam sposób definiujemy zbieżność słabą (lub według rozkładu) dla wektorów losowych i elementów losowych.

Słabą zbieżność zmiennych losowych można wygodnie sformułować w języku dystrybuant. Niech F_n i F oznaczają dystrybuanty zmiennych losowych ξ_n i ξ. Wtedy słaba zbieżność ciągu (ξ_n) do ξ oznacza zbieżność ciągu $(F_n(x))$ do $F(x)$ w każdym punkcie ciągłości x funkcji F. Zbieżność słabą będziemy oznaczać symbolem $\Rightarrow$ (zarówno dla zmiennych losowych, dystrybuant, jak i rozkładów). Następujące twierdzenie podaje związek między zbieżnością słabą a zbieżnością funkcji charakterystycznych.

Twierdzenie I.41. *Niech (φ_n) będzie ciągiem funkcji charakterystycznych zmiennych losowych ξ_n. Jeżeli $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ dla każdego $t \in \mathbb{R}$ i funkcja φ jest ciągła w $t = 0$, to $\varphi(t)$ jest funkcją charakterystyczną pewnej zmiennej losowej ξ oraz $\xi_n \Rightarrow \xi$. Na odwrót, jeżeli $\xi_n \Rightarrow \xi$, to $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ dla każdego $t \in \mathbb{R}$.*

Korzystając z funkcji charakterystycznych, można udowodnić następujące twierdzenie o zbieżności wektorów losowych.

Twierdzenie I.42. *Ciąg wektorów losowych $\xi_n = (\xi_n^1, \dots, \xi_n^d)$ jest słabo zbieżny do wektora $\xi = (\xi^1, \dots, \xi^d)$ wtedy i tylko wtedy, gdy dla każdego punktu $(a_1, \dots, a_d) \in \mathbb{R}^d$ zmienne losowe $a_1\xi_n^1 + \dots + a_d\xi_n^d$ są słabo zbieżne do zmiennej losowej $a_1\xi^1 + \dots + a_d\xi^d$.*

Dla zmiennych losowych o wartościach w zbiorze liczb całkowitych zamiast funkcji charakterystycznych wygodnie jest używać funkcji tworzących. *Funkcją tworzącą* zmiennej losowej ξ przyjmującej wartości całkowite nazywamy funkcję zmiennej zespolonej $\psi_\xi(z) = \mathrm{E}\, z^\xi$. Funkcja ψ_ξ nie musi być określona w całym zbiorze liczb zespolonych $\mathbb{C}$, ale na pewno jest poprawnie określona na okręgu jednostkowym, przy czym $\varphi_\xi(t) = \psi_\xi(e^{it})$. Mamy następujące zależności:

$$\psi_\xi(z) = \sum_k z^k \,\mathrm{P}(\xi = k), \qquad \mathrm{P}(\xi = k) = \frac{1}{2\pi i} \int_{|z|=1} \frac{\psi_\xi(z)}{z^{k+1}}\, dz. \tag{4.36}$$

Jeżeli $\mathrm{E}\,|\xi|^k < \infty$ dla pewnego $k \in \mathbb{N}$, to korzystając ze wzoru (4.33) i zależności $\psi_\xi(z) = \mathrm{E}\, z^\xi$, stwierdzamy, że

$$\mathrm{E}\,\xi^k = \left(z\frac{\partial}{\partial z}\right)^k \psi_\xi(z)\Big|_{z=1}, \tag{4.37}$$

a wyrażenie po prawej stronie wzoru obliczamy w ten sposób, że najpierw k razy działamy operatorem $z\frac{\partial}{\partial z}$ na $\psi_\xi(z)$, a następnie wyznaczamy wartość otrzymanego wyrażenia w $z = 1$.

Na koniec tego punktu wspomnimy o *metodzie momentów*, wygodnej przy rozwiązywaniu wielu praktycznych zagadnień. Nieco szersze omówienie tej metody można znaleźć w książce [40]. Będziemy rozpatrywać zmienne losowe o skończonych momentach $m_k = \mathrm{E}\,\xi^k$, $k = 1, 2, \dots$, dowolnego rzędu.

Pojawiają się tu dwa naturalne pytania: po pierwsze, czy momenty wyznaczają rozkład zmiennej losowej ξ jednoznacznie; po drugie, jeżeli mamy ciąg zmiennych losowych, których momenty dowolnego rzędu są zbieżne, to czy i w jakim sensie ciąg tych zmiennych losowych jest zbieżny do pewnej zmiennej losowej. Następujące twierdzenia podają odpowiedzi na te pytania.

Twierdzenie I.43. *Jeżeli zmienna losowa ξ o rozkładzie μ ma skończone momenty $m_k = \mathrm{E}\,\xi^k$ dla każdego $k = 1, 2, \dots$ i szereg potęgowy*

$$\sum_{k=1}^{\infty} m_k r^k / k!$$

ma dodatni promień zbieżności, to miara μ jest jednoznacznie wyznaczona przez ciąg (m_k).

Twierdzenie I.44. *Załóżmy, że rozkład zmiennej losowej ξ jest jednoznacznie określony przez jej momenty. Jeżeli zmienne losowe ξ_n mają skończone momenty dowolnego rzędu oraz $\lim_{n\to\infty} \mathrm{E}\,\xi_n^k = \mathrm{E}\,\xi^k$ dla dowolnego $k \geq 1$, to $\xi_n \Rightarrow \xi$.*

4.3. Niezależność i warunkowe wartości oczekiwane

Rozpoczniemy od wprowadzenia pojęcia dystrybuanty wektora losowego, co pozwoli nam na podanie prostej definicji niezależności zmiennych losowych.

Niech $\xi_1, \ldots, \xi_n$ będą zmiennymi losowymi na przestrzeni probabilistycznej $(\Omega, \Sigma, \mathrm{P})$. Funkcję $F_{\xi_1 \ldots \xi_n} \colon \mathbb{R}^n \to [0,1]$ określoną wzorem

$$F_{\xi_1 \ldots \xi_n}(x_1, \ldots, x_n) = \mathrm{P}(\xi_1 \leq x_1, \ldots, \xi_n \leq x_n) \tag{4.38}$$

nazywamy *dystrybuantą wektora losowego* $\xi = (\xi_1, \ldots, \xi_n)$.

Podobnie jak w przypadku jednowymiarowym wyróżniamy wektory losowe o rozkładzie absolutnie ciągłym i dla takiego wektora ξ definiujemy jego *gęstość* $f \colon \mathbb{R}^n \to [0, \infty)$ jako funkcję spełniającą warunek

$$\mathrm{P}(\xi \in B) = \int \cdots \int_B f(t_1, \ldots, t_n)\, dt_1 \ldots dt_n \quad \text{dla } B \in \mathcal{B}(\mathbb{R}^n).$$

Jeżeli rozkład wektora losowego jest absolutnie ciągły, to jego dystrybuanta F i gęstość f spełniają zależność

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(t_1, \ldots, t_n)\, dt_1 \ldots dt_n \quad \text{dla } x \in \mathbb{R}^n. \tag{4.39}$$

Zmienne losowe $\xi_1, \ldots, \xi_n$ nazywamy *niezależnymi*, jeżeli

$$F_{\xi_1 \ldots \xi_n}(x_1, \ldots, x_n) = F_{\xi_1}(x_1) \ldots F_{\xi_n}(x_n) \quad \text{dla } x \in \mathbb{R}^n. \tag{4.40}$$

Definicję tę można uogólnić na dowolną rodzinę zmiennych losowych, przyjmując, że zmienne losowe z tej rodziny są *niezależne*, jeżeli niezależne są zmienne z każdego skończonego podzbioru tej rodziny. Jeżeli zmienne losowe $\xi_1, \ldots, \xi_n$ są niezależne i mają gęstości $f_1, \ldots, f_n$, to wektor losowy $\xi = (\xi_1, \ldots, \xi_n)$ ma rozkład absolutnie ciągły o gęstości

$$f(x_1, \ldots, x_n) = f_1(x_1) \ldots f_n(x_n).$$

Niezależność zmiennych losowych możemy również zdefiniować w języku σ-algebr zdarzeń związanych z tymi zmiennymi. Niech $\mathcal{A}_1, \ldots, \mathcal{A}_n$ będą podzbiorami σ-algebry Σ dla $i = 1, \ldots, n$. Jeżeli dowolne zdarzenia $A_1 \in \mathcal{A}_1$, $\ldots, A_n \in \mathcal{A}_n$ są niezależne, to mówimy, że rodziny zdarzeń $\mathcal{A}_1, \ldots, \mathcal{A}_n$ są *niezależne*. Definicję tę można przenieść na dowolny zbiór rodzin zdarzeń, zakładając niezależność rodzin ze skończonych podzbiorów tego zbioru.

Każda zmienna losowa ξ *generuje* σ-algebrę zdarzeń określoną wzorem

$$\mathcal{F}_\xi = \{\xi^{-1}(B)\colon B \in \mathcal{B}(\mathbb{R})\}.$$

Niezależność zmiennych losowych $\xi_1, \dots, \xi_n$ jest równoważna niezależności σ-algebr $\mathcal{F}_{\xi_1}, \dots, \mathcal{F}_{\xi_n}$. Implikacja w jedną stronę jest oczywista, bo zbiór $\{\omega \in \Omega\colon \xi_i(\omega) \le x_i\}$ z definicji dystrybuanty F_{ξ_i} należy do σ-algebry $\mathcal{F}_{\xi_i}$. Bezpośredni dowód implikacji odwrotnej jest trudniejszy. Implikacja ta wynika m.in. z następującego twierdzenia.

Twierdzenie I.45. *Załóżmy, że σ-algebry $\mathcal{A}_1, \dots, \mathcal{A}_n$ są generowane przez rodziny zdarzeń $\mathcal{A}'_1, \dots, \mathcal{A}'_n$ i każda rodzina $\mathcal{A}'_i$ jest π-układem. Jeżeli rodziny $\mathcal{A}'_1, \dots, \mathcal{A}'_n$ są niezależne, to σ-algebry $\mathcal{A}_1, \dots, \mathcal{A}_n$ są również niezależne.*

Mówimy, że rodzina $\mathcal{P}$ jest *π-układem*, jeżeli jest zamknięta względem skończonych iloczynów, tj. jeżeli $A, B \in \mathcal{P}$, to $A \cap B \in \mathcal{P}$. Rodzina $\mathcal{P}$ *generuje* σ-algebrę $\mathcal{A}$, jeżeli $\mathcal{A}$ jest najmniejszą σ-algebrą zawierającą $\mathcal{P}$.

Podamy teraz definicje różnych wartości warunkowych dla zmiennych losowych, zdarzeń i σ-algebr zdarzeń. Rozpoczniemy od definicji $\mathrm{E}(\xi|A)$, gdzie ξ jest zmienną losową, a A zdarzeniem o dodatnim prawdopodobieństwie. O zmiennej losowej ξ będziemy stale zakładać, że $\mathrm{E}\,|\xi| < \infty$.

Jak już wspomnieliśmy przy okazji definicji prawdopodobieństwa warunkowego dwóch zdarzeń (patrz uwaga I.1), definicja wartości warunkowej względem zdarzenia A sprowadza się do zacieśnienia przestrzeni zdarzeń elementarnych do zbioru A i rozpatrywania nowego prawdopodobieństwa $\mathrm{P}_A(B) = \mathrm{P}(B)/\mathrm{P}(A)$ dla $B \subseteq A$. Podobnie definiujemy *warunkową wartość oczekiwaną zmiennej losowej ξ względem zdarzenia A* jako

$$\mathrm{E}(\xi|A) = \int_A \xi(\omega)\,\mathrm{P}_A(d\omega) = \frac{1}{\mathrm{P}(A)} \int_A \xi(\omega)\,\mathrm{P}(d\omega).$$

Wzór na prawdopodobieństwo całkowite można przenieść na zmienne losowe. Niech $A_1, A_2, \dots$ będzie skończonym lub nieskończonym ciągiem zdarzeń parami rozłącznych (tj. $A_i \cap A_j = \emptyset$ dla $i \ne j$), których sumą jest cała przestrzeń, a ξ – dowolną zmienną losową. Zachodzi następujący *wzór na prawdopodobieństwo całkowite* dla wartości oczekiwanej:

$$\mathrm{E}\,\xi = \sum_i \mathrm{E}(\xi|A_i)\,\mathrm{P}(A_i). \tag{4.41}$$

Sumowanie $\sum_i$ jest po wszystkich możliwych i, dla których $\mathrm{P}(A_i) > 0$.

Niech $\mathcal{A}$ będzie σ-algebrą zawartą w Σ. *Warunkowa wartość oczekiwana* $\mathrm{E}(\xi|\mathcal{A})$ zmiennej losowej ξ *względem σ-algebry $\mathcal{A}$* jest taką zmienną losową η mierzalną względem $\mathcal{A}$, że

$$\int_A \xi(\omega)\,\mathrm{P}(d\omega) = \int_A \eta(\omega)\,\mathrm{P}(d\omega) \quad \text{dla } A \in \mathcal{A}. \tag{4.42}$$

Istnienie i jednoznaczność $\mathrm{E}(\xi|\mathcal{A})$ wynika z twierdzenia Radona–Nikodyma. Będziemy korzystać z następujących własności warunkowej wartości oczekiwanej:

(a) jeżeli zmienna losowa η jest mierzalna względem σ-algebry $\mathcal{A}$, to

$$\mathrm{E}(\eta\xi|\mathcal{A}) = \eta\,\mathrm{E}(\xi|\mathcal{A}),$$

(b) $\mathrm{E}(\mathrm{E}(\xi|\mathcal{A})) = \mathrm{E}\,\xi$,

(c) jeśli zmienna losowa ξ i σ-algebra $\mathcal{A}$ są niezależne, to $\mathrm{E}(\xi|\mathcal{A}) = \mathrm{E}\,\xi$.

Udowodnienie powyższych wzorów pozostawiamy czytelnikowi jako ćwiczenie (patrz zad. I.40–I.42).

Przykład I.46. Niech σ-algebra $\mathcal{A}$ będzie generowana przez zdarzenia parami rozłączne $A_1, \ldots, A_m$. Zakładamy, że $\bigcup_{i=1}^m A_i = \Omega$ i definiujemy $\eta = \mathrm{E}(\xi|\mathcal{A})$. Ponieważ zmienna losowa η jest mierzalna względem $\mathcal{A}$, więc jest postaci $\eta = \sum_{i=1}^m c_i \mathbf{1}_{A_i}$. Pozostaje nam wyznaczyć takie stałe $c_1, \ldots, c_m$, że

$$\int_{A_j} \xi(\omega)\,\mathrm{P}(d\omega) = \int_{A_j} \sum_{i=1}^m c_i \mathbf{1}_{A_i}(\omega)\,\mathrm{P}(d\omega).$$

Ma to miejsce, gdy

$$c_i = \frac{1}{\mathrm{P}(A_i)} \int_{A_i} \xi(\omega)\,\mathrm{P}(d\omega) = \mathrm{E}(\xi|A_i).$$

Niech $C \in \Sigma$ i niech $\mathcal{A}$ będzie σ-algebrą zawartą w Σ. Zmienną losową

$$\mathrm{P}(C|\mathcal{A}) = \mathrm{E}(C|\mathcal{A}) := \mathrm{E}(\mathbf{1}_C|\mathcal{A})$$

nazywamy *prawdopodobieństwem warunkowym* zdarzenia C względem σ-algebry $\mathcal{A}$. Jeżeli σ-algebra $\mathcal{A}$ jest taka jak w przykładzie I.46, to

$$\mathrm{P}(C|\mathcal{A}) = \sum_{i=1}^m c_i \mathbf{1}_{A_i}, \quad \text{gdzie} \quad c_i = \frac{\mathrm{P}(C \cap A_i)}{\mathrm{P}(A_i)} = \mathrm{P}(C|A_i).$$

Niech ξ i η będą zmiennymi losowymi i niech $\mathcal{F}_\eta$ będzie σ-algebrą generowaną przez zmienną losową η. *Warunkową wartość oczekiwaną zmiennej losowej ξ względem zmiennej losowej η* definiujemy wzorem

$$\mathrm{E}(\xi|\eta) = \mathrm{E}(\xi|\mathcal{F}_\eta).$$

Ponieważ zmienna losowa $\mathrm{E}(\xi|\eta)$ jest mierzalna względem σ-algebry generowanej przez η, można wykazać, że istnieje taka funkcja mierzalna $g\colon \mathbb{R} \to \mathbb{R}$, że

$$\mathrm{E}(\xi|\eta) = g(\eta).$$

Przyjmujemy, że $\mathrm{E}(\xi \mid \eta = y) = g(y)$. Funkcję $y \mapsto \mathrm{E}(\xi \mid \eta = y)$ nazywamy *regresją zmiennej losowej ξ względem η*. W klasie wszystkich funkcji mierzalnych $h\colon \mathbb{R} \to \mathbb{R}$ wartość oczekiwana $\mathrm{E}(\xi - h(\eta))^2$ osiąga minimum dla $h = g$.

Dla dowolnego zbioru $C \in \Sigma$ definiujemy *prawdopodobieństwo warunkowe* zdarzenia C względem zmiennej losowej η wzorem

$$\mathrm{P}(C|\eta) = \mathrm{E}(\mathbf{1}_C|\mathcal{F}_\eta);$$

podobnie jak poprzednio definiujemy też $\mathrm{P}(C \mid \eta = y) = g(y)$, gdzie $g(\eta) = \mathrm{P}(C|\eta)$.

Uwaga I.47. Niech $\zeta = \mathrm{E}(\xi|\eta)$. Wiemy, że $\zeta = g(\eta)$ dla pewnej funkcji $g\colon \mathbb{R} \to \mathbb{R}$. Nasuwa się naturalne pytanie, jak wyznaczyć funkcję g. Jeżeli para (ξ, η) ma gęstość rozkładu $f_{\xi,\eta}(x, y)$, to

$$g(y) = \frac{\int_{-\infty}^{\infty} x f_{\xi,\eta}(x,y)\,dx}{\int_{-\infty}^{\infty} f_{\xi,\eta}(x,y)\,dx} = \frac{\int_{-\infty}^{\infty} x f_{\xi,\eta}(x,y)\,dx}{f_\eta(y)}.$$

Dowód tego faktu pozostawiamy czytelnikowi (patrz zad. I.44). Wyrażenie

$$f_\xi(x \mid \eta = y) := \frac{f_{\xi,\eta}(x,y)}{f_\eta(y)}$$

nazywamy *warunkową gęstością prawdopodobieństwa*; wtedy $g(y)$ jest wartością oczekiwaną warunkowej gęstości prawdopodobieństwa, tzn.

$$g(y) = \int_{-\infty}^{\infty} x f_\xi(x \mid \eta = y)\,dx.$$

Niech C będzie zdarzeniem o prawdopodobieństwie dodatnim i niech $g(y) = \mathrm{P}(C \mid \eta = y)$. Wtedy $g(y) = \mathrm{P}(C)F_C(y)/F(y)$, gdzie F jest dystrybuantą zmiennej losowej η, a F_C jest jej *dystrybuantą warunkową* przy warunku C, określoną wzorem $F_C(y) = \mathrm{P}(\eta \le x \mid C)$.

4.4. Rodziny σ-algebr. Prawa zero-jedynkowe

Ustalmy nieskończony ciąg zdarzeń $A_1, A_2, \ldots$. Interesuje nas prawdopodobieństwo zdarzenia A polegającego na tym, że zaszło nieskończenie wiele spośród zdarzeń $A_1, A_2, \ldots$. Zdarzenie A określone jest wzorem

$$A = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

Twierdzenie I.48. *Jeśli $\sum_{k=1}^{\infty} \mathrm{P}(A_k) < \infty$, to $\mathrm{P}(A) = 0$. Jeśli zdarzenia $A_1, A_2, \ldots$ są niezależne i $\sum_{k=1}^{\infty} \mathrm{P}(A_k) = \infty$, to $\mathrm{P}(A) = 1$.*

Twierdzenie I.48 nosi nazwę *lematu Borela–Cantellego* i jest jednym z tzw. praw zero-jedynkowych. Aby sformułować kolejne prawo zero-jedynkowe, musimy wprowadzić pomocnicze σ-algebry.

Przez *σ-algebrę generowaną przez rodzinę zmiennych losowych* ξ_α rozumiemy σ-algebrę generowaną przez wszystkie zdarzenia postaci $\{\xi_\alpha \in B\}$, gdzie $B \in \mathcal{B}(\mathbb{R})$. Jest to również najmniejsza σ-algebra zawarta w Σ, względem której wszystkie zmienne ξ_α, jak również wektory losowe z nich utworzone, są mierzalne.

Rozważmy teraz ciąg zmiennych losowych $\xi_1, \xi_2, \ldots$. Z ciągiem (ξ_n) będziemy wiązać następujące σ-algebry: $\mathcal{F}_{k,n}$ – generowaną przez zmienne losowe $\xi_k, \ldots, \xi_n$; $\mathcal{F}_{\leq n}$ – generowaną przez $\xi_1, \ldots, \xi_n$; oraz $\mathcal{F}_{\geq n}$ – generowaną przez $\xi_n, \xi_{n+1}, \ldots$. Niech

$$\mathcal{F}_\infty = \bigcap_{n=1}^{\infty} \mathcal{F}_{\geq n}.$$

Wtedy $\mathcal{F}_\infty$ nazywamy *σ-ciałem ogonowym* lub *σ-algebrą ogonową*; zdarzenia należące do $\mathcal{F}_\infty$ nazywamy *resztkowymi* albo *ogonowymi*. Ogólnie o σ-ciele ogonowym i zdarzeniach resztkowych możemy mówić, gdy zastąpimy ciąg $\mathcal{F}_{\geq n}$ dowolnym zstępującym ciągiem σ-algebr.

Twierdzenie I.49 (Prawo zero-jedynkowe Kołmogorowa)**.** *Jeżeli zmienne losowe $\xi_1, \xi_2, \ldots$ są niezależne i A jest zdarzeniem resztkowym, to* $\mathrm{P}(A) = 0$ *lub* $\mathrm{P}(A) = 1$.

Podamy teraz *twierdzenie o zbieżności warunkowej*, zwane też *prawem zero-jedynkowym Lévy'ego*. W tym celu wprowadzimy pojęcie zbieżności prawie na pewno. Ciąg (ξ_n) jest *zbieżny do ξ prawie na pewno* (albo *prawie wszędzie*, albo *z prawdopodobieństwem* 1), jeżeli $\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)$ dla prawie wszystkich ω (tj. z wyjątkiem ω ze zbioru o zerowym prawdopodobieństwie). Zbieżność prawie na pewno oznaczamy $\xi_n \to \xi$ p.n.

Twierdzenie I.50. *Niech ξ będzie zmienną losową o skończonej wartości oczekiwanej określoną na przestrzeni probabilistycznej $(\Omega, \Sigma, \mathrm{P})$ i niech $(\mathcal{F}_n)_{n\in\mathbb{N}}$ będzie ciągiem σ-algebr zawartych w Σ, przy czym spełniony jest jeden z dwóch warunków:*

- *ciąg $(\mathcal{F}_n)_{n\in\mathbb{N}}$ jest rosnący, a $\mathcal{F}_\infty$ jest najmniejszą σ-algebrą zawierającą wszystkie σ-algebry $\mathcal{F}_n$;*
- *ciąg $(\mathcal{F}_n)_{n\in\mathbb{N}}$ jest malejący, a $\mathcal{F}_\infty$ jest częścią wspólną σ-algebr $\mathcal{F}_n$.*

Wtedy

$$\lim_{n\to\infty} \mathrm{E}(\xi|\mathcal{F}_n) = \mathrm{E}(\xi|\mathcal{F}_\infty) \quad p.n.$$

Nazwa „prawo zero-jedynkowe" bierze się stąd, że jeżeli A jest zdarzeniem z σ-algebry $\mathcal{F}_\infty$, to $\lim_{n\to\infty} \mathrm{P}(A|\mathcal{F}_n) = \mathbf{1}_A$ p.n. Z otrzymanego wzoru wynika prawo zero-jedynkowe Kołmogorowa. Niech zmienne losowe $\xi_1, \xi_2, \dots$ będą niezależne, załóżmy, że $\mathcal{F}_n = \mathcal{F}_{\leq n} = \sigma(\xi_1, \dots, \xi_n)$, i niech $A \in \mathcal{F}_\infty$. Wtedy zdarzenie A jest niezależne od σ-ciała $\mathcal{F}_n$, więc $\mathrm{P}(A|\mathcal{F}_n) = \mathrm{E}(\mathbf{1}_A|\mathcal{F}_n) = \mathrm{P}(A)$. Zatem $\mathrm{P}(A) = \mathbf{1}_A$ p.n., a więc $\mathrm{P}(A) = 1$ lub $\mathrm{P}(A) = 0$.

Znane są jeszcze inne prawa zero-jedynkowe, np. twierdzenie Hewitta–Savage'a dla zdarzeń symetrycznych. W dalszej części książki poznamy prawa zero-jedynkowe związane z układami dynamicznymi, martyngałami i procesem Wienera.

Na koniec tego punktu wprowadzimy pojęcia momentu zatrzymania i zmiennej niezależnej od przyszłości. Zmienną losową ν przyjmującą wartości w zbiorze liczb naturalnych nazywamy *momentem zatrzymania*, jeżeli $\{\nu \leq n\} \in \mathcal{F}_{\leq n}$ dla każdego n. Definicję momentu zatrzymania przenosi się natychmiast na przypadek, gdy σ-algebry $\mathcal{F}_{\leq n}$ zastąpimy dowolną rosnącą rodziną σ-algebr.

Jeżeli zmienne losowe $\xi_1, \xi_2, \dots$ są niezależne, to σ-algebry $\mathcal{F}_{\leq n}$ i $\mathcal{F}_{\geq n+1}$ są też niezależne, w szczególności zdarzenie $\{\nu \leq n\}$ nie zależy od $\mathcal{F}_{\geq n+1}$. O takiej zmiennej ν mówimy, że *nie zależy od przyszłości*.

Następujące twierdzenie pozwala łatwo wyznaczać wartości oczekiwane sum zmiennych losowych o losowej liczbie składników. Przyjmujemy oznaczenie $S_\nu = \xi_1 + \dots + \xi_\nu$.

Twierdzenie I.51 (Tożsamość Walda)**.** *Jeżeli zmienne losowe $\xi_1, \xi_2, \dots$ są niezależne i mają taki sam rozkład o skończonej wartości oczekiwanej, a zmienna losowa ν nie zależy od przyszłości oraz $\mathrm{E}\,\nu < \infty$, to*

$$\mathrm{E}\,S_\nu = \mathrm{E}\,\xi_1 \cdot \mathrm{E}\,\nu.$$

4.5. Wektory losowe; działania na nich i ich charakterystyki

Niech ξ i η będą zmiennymi losowymi o skończonych wartościach oczekiwanych. *Kowariancją* zmiennych ξ i η nazywamy liczbę

$$\mathrm{Cov}(\xi, \eta) = \mathrm{E}((\xi - \mathrm{E}\,\xi)(\eta - \mathrm{E}\,\eta)).$$

Jeżeli ξ i η mają skończone wariancje, to ich kowariancja istnieje. Jeżeli ponadto wariancje zmiennych losowych ξ i η są dodatnie, to *współczynnikiem korelacji* nazywamy liczbę

$$\rho(\xi, \eta) = \frac{\mathrm{Cov}(\xi, \eta)}{\sigma_\xi \sigma_\eta} = \frac{\mathrm{E}((\xi - \mathrm{E}\,\xi)(\eta - \mathrm{E}\,\eta))}{\sqrt{\mathrm{D}^2\,\xi}\sqrt{\mathrm{D}^2\,\eta}}.$$

Korelacja jest więc wartością oczekiwaną iloczynu zestandaryzowanych zmiennych losowych ξ i η. Jeżeli zmienne losowe ξ i η są niezależne, to kowariancja i współczynnik korelacji są zerowe, co natychmiast wynika ze wzoru $\mathrm{E}(\xi\eta) = \mathrm{E}\,\xi \cdot \mathrm{E}\,\eta$. Współczynnik korelacji jest liczbą z przedziału $[-1,1]$ i w zależności od jego znaku mówimy o skorelowaniu dodatnim, ujemnym lub nieskorelowaniu zmiennych losowych. Jeżeli $\rho(\xi,\eta) = \pm 1$, to istnieją takie stałe a i b, że $\mathrm{P}(\eta = a\xi + b) = 1$, przy czym $a > 0$, gdy $\rho(\xi,\eta) = 1$, oraz $a < 0$, gdy $\rho(\xi,\eta) = -1$.

Jeżeli mamy wektor losowy $\xi = (\xi_1,\dots,\xi_n)$, to $(n \times n)$-wymiarowe macierze $[\mathrm{Cov}(\xi_i,\xi_j)]$ i $[\rho(\xi_i,\xi_j)]$ nazywamy odpowiednio *macierzą kowariancji* i *macierzą korelacji*. Macierze te są symetryczne i nieujemnie określone. Przypominamy, że macierz rzeczywista $A = [a_{ij}]$ jest *nieujemnie określona*, jeżeli dla dowolnego $x \in \mathbb{R}^n$ mamy $Ax \cdot x \geq 0$. W szczególności $\det A \geq 0$.

Jeżeli $A = [\mathrm{Cov}(\xi_i,\xi_j)]$ lub $A = [\rho(\xi_i,\xi_j)]$, to $\det A = 0$ wtedy i tylko wtedy, gdy wektor losowy ξ jest *zdegenerowany*, co oznacza istnienie takich stałych $c_1,\dots,c_{n+1}$, że $\mathrm{P}(c_1\xi_1 + \cdots + c_n\xi_n = c_{n+1}) = 1$.

Przykład I.52. Niech $\xi_1,\dots,\xi_n$ będzie ciągiem niezależnych wektorów o rozkładzie $\mathcal{N}(0,1)$. Niech $C = [c_{ij}]$ będzie macierzą rzeczywistą o wymiarach $n \times n$, $m \in \mathbb{R}^n$ i niech $\eta_i = \sum_{r=k}^{n} c_{ik}\xi_k + m_i$. Wtedy zmienne losowe η_i mają wartości oczekiwane m_i, a wektor $\eta = (\eta_1,\dots,\eta_n)$ ma macierz kowariancji $A = CC^T$, gdzie C^T jest macierzą transponowaną względem C. Wektor η nazywamy *n-wymiarową zmienną gaussowską*. Jeżeli $\det C \neq 0$, to wektor η jest niezdegenerowany i ma rozkład o gęstości

$$f(x_1,\dots,x_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det A}} \exp\bigl(-\tfrac{1}{2}(x-m)A^{-1}(x-m)^T\bigr). \tag{4.43}$$

Dla dowolnych zmiennych losowych $\xi_1,\dots,\xi_n$ mamy

$$\mathrm{E}(\xi_1 + \cdots + \xi_n) = \mathrm{E}\,\xi_1 + \cdots + \mathrm{E}\,\xi_n,$$

o ile wartości oczekiwane tych zmiennych istnieją. Jeżeli zmienne losowe $\xi_1,\dots,\xi_n$ są niezależne, to

$$\mathrm{D}^2(\xi_1 + \cdots + \xi_n) = \mathrm{D}^2\,\xi_1 + \cdots + \mathrm{D}^2\,\xi_n,$$
$$\varphi_{\xi_1+\cdots+\xi_n}(t) = \varphi_{\xi_1}(t)\dots\varphi_{\xi_n}(t).$$

Funkcję charakterystyczną definiujemy również dla wektora losowego $\xi = (\xi_1,\dots,\xi_n)$ jako odwzorowanie $\varphi_\xi\colon \mathbb{R}^n \to \mathbb{C}$ określone wzorem

$$\varphi_\xi(t_1,\dots,t_n) = \mathrm{E}\,e^{it\cdot\xi} = \mathrm{E}\Bigl(\exp\sum_{k=1}^{n} it_k\xi_k\Bigr). \tag{4.44}$$

Znając funkcję charakterystyczną wektora losowego, możemy wyznaczać inne charakterystyki wektorów losowych, np. jeżeli istnieje $\mathrm{E}\,\xi_1^{k_1}\dots\xi_n^{k_n}$, to

$$\mathrm{E}\,\xi_1^{k_1}\dots\xi_n^{k_n} = i^{-(k_1+\dots+k_n)}\frac{\partial^{k_1+\dots+k_n}\varphi_\xi(t)}{\partial t_1^{k_1}\dots\partial t_n^{k_n}}\Big|_{t=0}.$$

W wielu zagadnieniach praktycznych znamy rozkład pewnego wektora losowego $\xi = (\xi_1,\dots,\xi_n)$, a interesuje nas rozkład zmiennej losowej $\eta = S(\xi_1,\dots,\xi_n)$, gdzie $S\colon \mathbb{R}^n \to \mathbb{R}$ jest pewną funkcją. Ograniczymy się do przypadku, gdy wektor ξ ma gęstość $f_\xi(x_1,\dots,x_n)$, a S jest funkcją ciągłą. Zauważmy, że dystrybuantę F_η można wyznaczyć ze wzoru

$$F_\eta(x) = \mathrm{P}(\eta \le x) = \int\cdots\int_{\{t\colon S(t)\le x\}} f_\xi(t_1,\dots,t_n)\,dt_1\dots dt_n. \tag{4.45}$$

Jeżeli zmienna losowa η ma gęstość, to można ją wyznaczyć, korzystając ze wzoru (4.45).

Przykład I.53. Rozpoczniemy od przykładu jednowymiarowego. Niech $\eta = \xi^2$. Wtedy dla $x \ge 0$ mamy

$$F_\eta(x) = \mathrm{P}(\eta \le x) = \mathrm{P}(|\xi| \le \sqrt{x}\,) = \int_{-\sqrt{x}}^{\sqrt{x}} f_\xi(t)\,dt,$$

a stąd $f_\eta(x) = \frac{1}{2}x^{-1/2}\big(f_\xi(-\sqrt{x}) + f_\xi(\sqrt{x})\big)$ dla $x \ge 0$. Jeżeli ξ ma standardowy rozkład normalny, to $f_\eta(x) = (2\pi x)^{-1/2}e^{-x/2}$, jest to więc rozkład gamma z oboma parametrami $\frac{1}{2}$. Jeżeli $\eta = \xi_1^2 + \dots + \xi_n^2$, gdzie $\xi_1,\dots,\xi_n$ są niezależnymi zmiennymi losowymi o standardowym rozkładzie normalnym, to η ma rozkład gamma z parametrami $\alpha = \frac{1}{2}$ i $\lambda = \frac{n}{2}$. Rozkład ten nazywamy *rozkładem χ^2 o n stopniach swobody*.

Przykład I.54. Niech ξ będzie zmienną losową przyjmującą wartości w pewnym przedziale Δ, tzn. $\mathrm{P}(\xi \in \Delta) = 1$, a $S\colon \Delta \to \mathbb{R}$ niech będzie funkcją różniczkowalną o dodatniej pochodnej. Wtedy dystrybuanta zmiennej losowej $\eta = S(\xi)$ określona jest wzorem

$$F_\eta(x) = \mathrm{P}(\eta \le x) = F_\xi(S^{-1}(x)) = \int_{-\infty}^{S^{-1}(x)} f_\xi(t)\,dt.$$

Różniczkując obie strony równości względem x, otrzymujemy

$$f_\eta(x) = \frac{d}{dx}(S^{-1}(x))f_\xi(S^{-1}(x)) = \frac{1}{S'(S^{-1}(x))}f_\xi(S^{-1}(x)). \tag{4.46}$$

Przykład I.55. Niech $\eta = \xi_1 + \cdots + \xi_n$. Korzystając ze wzoru (4.45), otrzymujemy

$$F_\eta(x) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ razy}} \int_{-\infty}^{x-t_2-\cdots-t_n} f_\xi(t_1,\ldots,t_n)\,dt_1\ldots dt_n. \tag{4.47}$$

Korzystając teraz ze wzoru na różniczkowanie pod znakiem całki, otrzymujemy wzór na gęstość zmiennej η:

$$f_\eta(x) = \underbrace{\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty}}_{n-1 \text{ razy}} f_\xi(x - t_2 - \cdots - t_n, t_2, \ldots, t_n)\,dt_2\ldots dt_n. \tag{4.48}$$

Jeżeli zmienne $\xi_1,\ldots,\xi_n$ są niezależne, to $f_\xi(t_1,\ldots,t_n) = f_{\xi_1}(t_1)\ldots f_{\xi_n}(t_n)$ i wtedy

$$f_\eta = f_{\eta_1} * \cdots * f_{\eta_n}, \tag{4.49}$$

gdzie symbol $*$ oznacza *splot* (*konwolucję*) funkcji i dla dowolnych funkcji całkowalnych g i h określony jest wzorem

$$g * h(x) = \int_{-\infty}^{\infty} g(t)h(x-t)\,dt = \int_{-\infty}^{\infty} g(x-t)h(t)\,dt.$$

Przykład I.56. Niech $\eta = \max(\xi_1,\ldots,\xi_n)$. Wtedy

$$F_\eta(x) = \int_{-\infty}^{x} \cdots \int_{-\infty}^{x} f_\xi(t_1,\ldots,t_n)\,dt_1\ldots dt_n.$$

Jeżeli zmienne losowe $\xi_1,\ldots,\xi_n$ są niezależne, to

$$F_\eta(x) = F_{\xi_1}(x)\ldots F_{\xi_n}(x),$$

podczas gdy

$$f_\eta(x) = \Big(\frac{f_{\xi_1}(x)}{F_{\xi_1}(x)} + \cdots + \frac{f_{\xi_n}(x)}{F_{\xi_n}(x)}\Big)F_{\xi_1}(x)\ldots F_{\xi_n}(x).$$

5. Podstawowe twierdzenia rachunku prawdopodobieństwa

5.1. Prawo wielkich liczb

Rozpoczniemy od przypomnienia różnych rodzajów zbieżności zmiennych losowych. W punkcie 4.2 wprowadziliśmy pojęcie zbieżności słabej. Przypominamy, że ciąg (ξ_n) zmiennych losowych jest słabo zbieżny (albo zbieżny

według rozkładu) do zmiennej losowej ξ, jeżeli dla dowolnej funkcji ciągłej i ograniczonej g mamy $\lim_{n\to\infty} \mathrm{E}\, g(\xi_n) = \mathrm{E}\, g(\xi)$; w języku dystrybuant oznacza to, że $F_{\xi_n}(x) \to F_\xi(x)$ w każdym punkcie ciągłości x funkcji F_ξ. Zbieżność słabą oznaczaliśmy przez $\Rightarrow$.

Ciąg (ξ_n) jest *zbieżny według prawdopodobieństwa* (albo *stochastycznie*), jeżeli dla każdego $\varepsilon > 0$ zachodzi równość

$$\lim_{n\to\infty} \mathrm{P}(|\xi_n - \xi| > \varepsilon) = 0,$$

co oznaczamy $\xi_n \xrightarrow{\mathrm{P}} \xi$. W punkcie 4.4 wprowadziliśmy zbieżność prawie na pewno. Przypominamy, że ciąg (ξ_n) jest zbieżny do ξ prawie na pewno, jeżeli $\lim_{n\to\infty} \xi_n(\omega) = \xi(\omega)$ dla prawie wszystkich ω. Ze zbieżności prawie na pewno wynika zbieżność według prawdopodobieństwa, a z niej zbieżność według rozkładu. Implikacje odwrotne nie zachodzą.

Uwaga I.57. Pojęcie zbieżności słabej nie jest bezpośrednio związane z przestrzenią probabilistyczną i możemy mówić o zbieżności według rozkładu dla zmiennych określonych w różnych przestrzeniach probabilistycznych. Następujące *twierdzenie Skorochoda o reprezentacji* (patrz [39, str. 70]) pokazuje, że mając słabą zbieżność rozkładów, można tak dobrać przestrzeń probabilistyczną i zmienne losowe, aby uzyskać zbieżność prawie na pewno.

Twierdzenie I.58. *Niech $(\mu_n)_{n\in\mathbb{N}}$ będzie ciągiem miar probabilistycznych na przestrzeni metrycznej (X, ρ), słabo zbieżnym do pewnej miary probabilistycznej μ. Załóżmy, że nośnik topologiczny miary μ jest przestrzenią ośrodkową. Wtedy istnieją: przestrzeń probabilistyczna $(\Omega, \Sigma, \mathrm{P})$, ciąg $(\xi_n)_{n\in\mathbb{N}}$ zmiennych losowych oraz zmienna losowa ξ określone na tej przestrzeni, o rozkładach odpowiednio $(\mu_n)_{n\in\mathbb{N}}$ i μ, dla których $\xi_n \to \xi$ p.n.*

Nośnik topologiczny miary borelowskiej m definiujemy jako najmniejszy zbiór domknięty F spełniający warunek $m(X \setminus F) = 0$ lub, równoważnie,

$$F = \{x \in X\colon m(B(x, \varepsilon)) > 0 \text{ dla każdego } \varepsilon > 0\},$$

gdzie $B(x, \varepsilon)$ jest kulą otwartą o środku x i promieniu ε.

W teorii prawdopodobieństwa i teorii ergodycznej kluczową rolę odgrywają twierdzenia o zbieżności średnich. Rozważmy ciąg (ξ_n) zmiennych losowych i niech $S_n = \xi_1 + \cdots + \xi_n$. Interesuje nas, kiedy ciąg (S_n/n) jest zbieżny i jaką ma granicę. Odpowiedź zależy od przyjętych założeń dotyczących ciągu (ξ_n).

Twierdzenie I.59 (Mocne prawo wielkich liczb)**.** *Załóżmy, że (ξ_n) jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie. Jeżeli zmienne losowe mają skończoną wartość oczekiwaną m, to $\frac{S_n}{n} \to m$ p.n.*

Prawo wielkich liczb zostało po raz pierwszy sformułowane przez Jakuba Bernoulliego w roku 1713 dla zmiennych losowych binarnych (o wartościach 0 i 1) i było wielokrotnie poprawiane. Twierdzenie I.59 zostało udowodnione przez Kołmogorowa, a wcześniej przez Chinczyna w wersji ze zbieżnością według prawdopodobieństwa (słabe prawo wielkich liczb).

Podamy teraz wersję prawa wielkich liczb dla elementów losowych. Niech $\xi_1, \xi_2, \dots$ będzie ciągiem elementów losowych o wartościach w przestrzeni mierzalnej $(X, \mathcal{A})$. *Miarę empiryczną* μ_n na σ-algebrze $\mathcal{A}$ określamy wzorem

$$\mu_n(A) = \frac{1}{n} \sum_{k=1}^{n} \mathbf{1}_A(\xi_k).$$

Jeżeli $\xi_1, \xi_2, \dots$ jest ciągiem niezależnych elementów losowych o tym samym rozkładzie μ, to dla każdego ustalonego zbioru $A \in \mathcal{A}$ zmienne losowe $\mathbf{1}_A(\xi_1), \mathbf{1}_A(\xi_2), \dots$ są niezależne i o tym samym rozkładzie z wartością oczekiwaną $\mu(A)$. Z prawa wielkich liczb wynika zbieżność p.n. ciągu $(\mu_n(A))$ do $\mu(A)$. To proste stwierdzenie ma ważne implikacje praktyczne. Przeprowadzając np. niezależne eksperymenty, możemy wyznaczać rozkłady parametrów badanego obiektu. Inne zastosowania poznamy w dalszej części książki, a teraz ograniczymy się do prostego przykładu. Przyjmijmy, że model stochastyczny opisuje rozkład liczby pewnych molekuł w komórce. Przy założeniu, że procesy w różnych komórkach zachodzą niezależnie, możemy wyznaczyć rozkład empiryczny liczby molekuł w całej populacji komórkowej, a więc podać, jaka część komórek ma ustaloną liczbę molekuł.

Można podać wersje mocnego prawa wielkich liczb, gdy zmienne losowe (ξ_n) mają różne rozkłady, ale wtedy potrzebne są dodatkowe założenia.

Twierdzenie I.60. *Załóżmy, że* (ξ_n) *jest ciągiem niezależnych zmiennych losowych o skończonych wariancjach. Wtedy*

$$\textit{jeżeli} \quad \sum_{k=1}^{\infty} \frac{D^2\xi_k}{k^2} < \infty, \quad \textit{to} \quad \frac{S_n - \mathrm{E}\, S_n}{n} \to 0 \ \textit{p.n.}$$

Na podstawie twierdzenia I.59 średnie ciągu niezależnych zmiennych losowych dążą prawie na pewno do ich wartości średniej m. Nasuwa się pytanie, o ile n-ta średnia może się różnić od m. Odpowiedź wynika z następującego twierdzenia.

Twierdzenie I.61 (Prawo iterowanego logarytmu)**.** *Niech* (ξ_n) *będzie ciągiem niezależnych zmiennych losowych o tym samym rozkładzie. Załóżmy, że* $\mathrm{E}\,\xi_n = 0$ *i* $\mathrm{D}^2\,\xi_n = \sigma^2$. *Wtedy*

$$\liminf_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = -\sigma, \quad \limsup_{n\to\infty} \frac{S_n}{\sqrt{2n \ln\ln n}} = \sigma \quad \textit{p.n.} \tag{5.1}$$

5.2. Centralne twierdzenie graniczne

Do jednych z najczęściej stosowanych twierdzeń rachunku prawdopodobieństwa należy centralne twierdzenie graniczne. Rozpoczniemy od następującej klasycznej wersji.

Twierdzenie I.62 (Centralne twierdzenie graniczne)**.** *Załóżmy, że* (ξ_n) *jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie o wartości oczekiwanej* $\mathrm{E}\,\xi_k = a$ *i skończonej wariancji* $\mathrm{D}^2\,\xi_k = \sigma^2 > 0$*. Niech* $S_n = \xi_1 + \cdots + \xi_n$ *oraz*

$$\eta_n = \frac{S_n - an}{\sigma\sqrt{n}}.$$

Wtedy $\mathrm{P}(\eta_n \le x) \to \Phi(x)$ *jednostajnie dla* $x \in \mathbb{R}$*, gdy* $n \to \infty$*, gdzie* $\Phi(x)$ *jest dystrybuantą standardowego rozkładu normalnego.*

Prostym przykładem zastosowania twierdzenia I.62 jest wyznaczanie przybliżonej dystrybuanty rozkładu dwumianowego. Zmienna losowa S_n o rozkładzie dwumianowym dla n prób i z prawdopodobieństwem sukcesu p w pojedynczej próbie jest sumą niezależnych zmiennych losowych $\xi_1, \ldots, \xi_n$ o jednakowym rozkładzie $\mathrm{P}(\xi_k = 1) = p$ i $\mathrm{P}(\xi_k = 0) = 1 - p$. Zgodnie z centralnym twierdzeniem granicznym dystrybuantę S_n można przybliżać dystrybuantą rozkładu normalnego z parametrami $m = pn$ i $\sigma^2 = p(1-p)n$. Wykresy na rysunkach I.14 i I.15 przedstawiają przybliżenie rozkładu dwumianowego z $n = 9$ oraz $p = 1/2$ i $p = 2/3$.

Sformułujemy teraz dwie wersje centralnego twierdzenia granicznego dla niezależnych ciągów zmiennych losowych o niejednakowym rozkładzie. Będziemy rozważać następujący schemat tworzenia zmiennych losowych η_n. Załóżmy, że dla każdego n naturalnego istnieje ciąg niezależnych zmiennych losowych $\xi_{1,n}, \xi_{2,n}, \ldots, \xi_{n,n}$ spełniający warunek

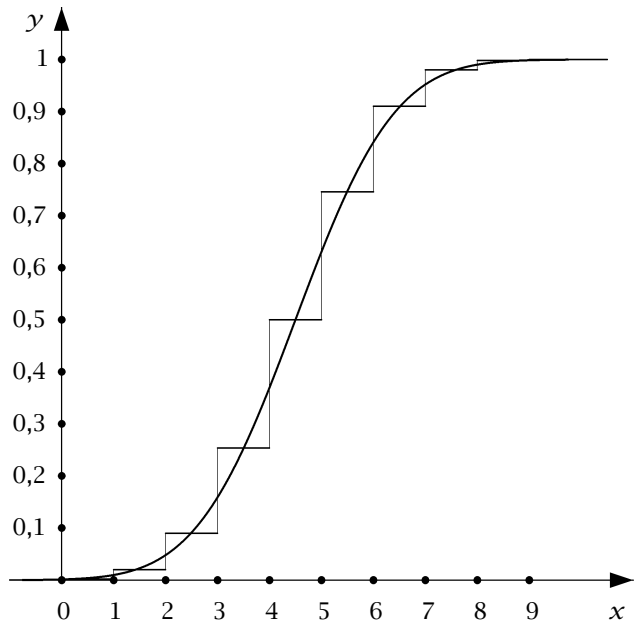
$$\mathrm{E}\,\xi_{k,n} = 0, \quad \sum_{k=1}^{n} \mathrm{D}^2\,\xi_{k,n} = 1, \tag{5.2}$$

i niech $\eta_n = \xi_{1,n} + \cdots + \xi_{n,n}$. Będziemy rozważać następujący warunek Lindeberga:

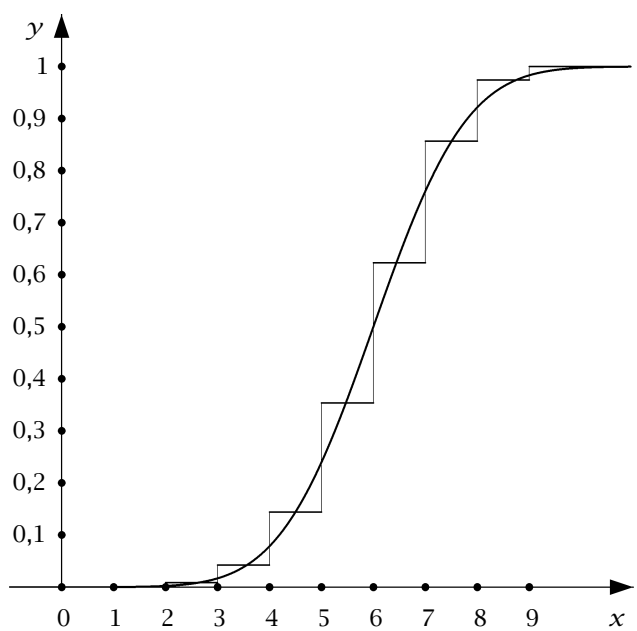
$$\lim_{n\to\infty} \sum_{k=1}^{n} \int_{\{|\xi_{n,k}|>\varepsilon\}} \xi_{n,k}^2(\omega)\,\mathrm{P}(d\omega) = 0 \quad \text{dla każdego } \varepsilon > 0. \tag{L}$$

Twierdzenie I.63 (Lindeberga–Lévy'ego)**.** *Jeśli spełnione są warunki* (5.2) *i* (L)*, to* $\mathrm{P}(\eta_n \le x) \to \Phi(x)$ *jednostajnie względem* x*.*

Z twierdzenia I.63 można uzyskać wersję centralnego twierdzenia granicznego w klasycznej postaci dla zmiennych losowych o różnych rozkładach. Niech



Rysunek I.14. Dystrybuanta rozkładu Bernoulliego dla $p = 0{,}5$ i $n = 9$ oraz jej przybliżenie dystrybuantą rozkładu normalnego z $m = 4{,}5$ i $\sigma = 1{,}5$



Rysunek I.15. Dystrybuanta rozkładu Bernoulliego dla $p = 2/3$ i $n = 9$ oraz jej przybliżenie dystrybuantą rozkładu normalnego z $m = 6$ i $\sigma = \sqrt{2}$

(ξ_n) będzie ciągiem niezależnych zmiennych losowych o wartości oczekiwanej $\mathrm{E}\,\xi_k = a_k$ i skończonej wariancji $\mathrm{D}^2\,\xi_k = \sigma_k^2 > 0$. Niech

$$B_n^2 = \sum_{k=1}^{n} \sigma_k^2, \qquad \eta_n = \frac{S_n - \sum_{k=1}^{n} a_k}{B_n}.$$

Zauważmy, że $\sum_{k=1}^{n} a_k$ jest wartością oczekiwaną, a B_n^2 wariancją zmiennej S_n, a więc zmienna losowa η_n jest standaryzowana – ma zerową wartość oczekiwaną i wariancję równą 1. Jednocześnie η_n jest sumą niezależnych zmiennych losowych $\xi_{1,n}, \ldots, \xi_{n,n}$ określonych wzorem

$$\xi_{k,n} = \frac{\xi_k - a_k}{B_n},$$

dla których sprawdzamy warunek Lindeberga (L). W szczególności z twierdzenia I.63 wynika następujące:

Twierdzenie I.64 (Lapunowa)**.** *Załóżmy, że dla pewnego $\delta > 0$ istnieją stałe $c_k^{2+\delta} = \mathrm{E}\,|\xi_k - a_k|^{2+\delta}$, i niech $C_n^{2+\delta} = \sum_{k=1}^{n} c_k^{2+\delta}$. Jeżeli $C_n/B_n \to 0$, gdy $n \to \infty$, to $\mathrm{P}(\eta_n \leq x) \to \Phi(x)$.*

Zauważmy, że w twierdzeniu Lapunowa pojawił się warunek zakładający istnienie momentów zmiennych losowych ξ_k rzędu większego niż 2, a więc samo istnienie wariancji nie wystarcza do jego sformułowania.

Należy wspomnieć o wersjach centralnego twierdzenia granicznego, w których warunek niezależności zastępuje się innymi warunkami, jak słaba niezależność lub mieszanie; istnieją również wersje dla procesów stochastycznych oraz wersje ze zbieżnością do tzw. rozkładów α-stabilnych, jeżeli drugi moment nie istnieje [40, 53, 113].

Gdy zmienne losowe (ξ_n) mają gęstości, pojawia się naturalne pytanie, czy twierdzenie I.62 można wzmocnić i zamiast zbieżności dystrybuant uzyskać zbieżność gęstości ciągu (η_n) do gęstości rozkładu normalnego. Tak jest istotnie, przy dodatkowym założeniu.

Twierdzenie I.65 (Centralne twierdzenie graniczne dla gęstości)**.** *Załóżmy, że (ξ_n) jest ciągiem niezależnych zmiennych losowych o tym samym rozkładzie absolutnie ciągłym i funkcji charakterystycznej φ. Jeżeli $|\varphi|^r$ jest funkcją całkowalną dla pewnego $r > 0$, to zmienne losowe*

$$\eta_n = \frac{S_n - n\,\mathrm{E}\,\xi_1}{\sqrt{n\,\mathrm{D}^2\,\xi_1}}$$

mają gęstości f_n, które zmierzają jednostajnie do gęstości standardowego rozkładu normalnego. Jeżeli gęstość zmiennej ξ_n jest ograniczona, to również gęstości zmiennych η_n zmierzają jednostajnie do gęstości standardowego rozkładu normalnego.

Centralne twierdzenie graniczne można również sformułować dla wektorów losowych.

Twierdzenie I.66 (CTG dla wektorów losowych)**.** *Załóżmy, że* (ξ_n) *jest ciągiem niezależnych wektorów losowych w* $\mathbb{R}^d$ *o tym samym rozkładzie z macierzą kowariancji* A*. Wtedy ciąg wektorów losowych*

$$\eta_n = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} (\xi_k - \mathrm{E}\,\xi_k)$$

jest zbieżny według rozkładu do wektora normalnego o zerowej wartości oczekiwanej i macierzy kowariancji A*.*

5.3. Nierówności dla zmiennych losowych

Na koniec tej części książki przedstawimy kilka ważnych nierówności, z których korzysta się zarówno w dowodach różnych twierdzeń, jak i w zastosowaniach do szacowania prawdopodobieństw zdarzeń.

Jeżeli $\xi \geq 0$, to dla każdego ε mamy

$$\mathrm{P}(\xi \geq \varepsilon) \leq \frac{\mathrm{E}\,\xi}{\varepsilon}. \tag{5.3}$$

Niech ξ będzie zmienną losową ze skończoną wariancją $\sigma^2 = \mathrm{D}^2\,\xi$. Wtedy dla każdej liczby rzeczywistej $c > 0$ mamy

$$\mathrm{P}(|\xi - \mathrm{E}\,\xi| \geq c\sigma) \leq \frac{1}{c^2}. \tag{5.4}$$

Dla dowolnej zmiennej losowej ξ oraz $p > 0$ i $\varepsilon > 0$ mamy

$$\mathrm{P}(|\xi| \geq \varepsilon) \leq \frac{\mathrm{E}\,|\xi|^p}{\varepsilon^p}. \tag{5.5}$$

Nierówności (5.3) i (5.4) noszą nazwę *nierówności Czebyszewa*, a (5.5) to *nierówność Markowa*. Nierówności (5.4) i (5.5) są prostymi wnioskami z nierówności (5.3), którą też można łatwo sprawdzić.

Niech $\xi_1, \ldots, \xi_n$ będą niezależnymi zmiennymi losowymi o zerowych wartościach oczekiwanych. Jeżeli

$$S_k = \xi_1 + \cdots + \xi_k \quad \text{oraz} \quad M_n = \max\{|S_1|, \ldots, |S_n|\},$$

to spełniona jest *nierówność Kołmogorowa*

$$\mathrm{P}(M_n \geq x) \leq \frac{\mathrm{D}^2\,S_n}{x^2} \quad \text{dla } x > 0. \tag{5.6}$$

Nierówność (5.6) to istotne wzmocnienie nierówności Czebyszewa (5.4).

Niech teraz $f\colon \mathbb{R} \to \mathbb{R}$ będzie funkcją wypukłą, a ξ zmienną losową. Wtedy zachodzi *nierówność Jensena*

$$\mathrm{E}\, f(\xi) \geq f(\mathrm{E}(\xi)). \tag{5.7}$$

Nierówność (5.7) można stosunkowo łatwo udowodnić, korzystając ze wskazówki do zadania I.48. Będziemy też korzystać z nierówności Jensena w następujacej postaci. Niech (X, Σ, μ) będzie przestrzenią probabilistyczną. Jeżeli F jest rzeczywistą funkcją wypukłą, a $h\colon X \to \mathbb{R}$ funkcją całkowalną, to

$$F\Big(\int_X h(x)\,\mu(dx)\Big) \leq \int_X F(h(x))\,\mu(dx). \tag{5.8}$$

6. Przykłady zastosowań

6.1. Funkcja przeżycia i oczekiwany czas życia

Rozpoczniemy od modeli dyskretnych z punktu 2.4. W statycznym modelu demograficznym z przykładu I.3 szansa przeżycia n lat wynosi

$$q_n = \prod_{j=0}^{n} (1 - d_j), \tag{6.1}$$

gdzie d_j to współczynnik śmiertelności w wieku j. Przypominamy, że przyjęliśmy tu konwencję, że dziecko w roku urodzenia jest w wieku zerowym, a przeżycie n lat oznacza, że osoba dożyła do końca roku kalendarzowego, w którym miała n-te urodziny. Wzór ten ulega tylko drobnym modyfikacjom w przypadku dynamicznego modelu demograficznego z przykładu I.7, w którym współczynnik d_j zależy nie tylko od wieku osobnika j, ale również od roku r, w którym jest obliczany. Model dyskretny można przenieść na inne populacje, często z sensowną zmianą skali czasu, np. zamiast roku rozważamy dni lub godziny w przypadku mikroorganizmów. Współczynnik śmiertelności może zależeć od innych czynników niż wiek osobnika, np. od fenotypu, dojrzałości lub rozmiarów organizmu dla populacji komórkowych.

Korzystając ze wzoru (6.1), można wyznaczyć inne charakterystyki demograficzne. Prawdopodobieństwo, że osoba przeżyje dokładnie n lat, wynosi

$$q_{n-1} - q_n = \prod_{j=0}^{n-1} (1 - d_j) - \prod_{j=0}^{n} (1 - d_j) = d_n \prod_{j=0}^{n-1} (1 - d_j),$$

a stąd *średnią długość życia* S wyznaczymy ze wzoru

$$S = \sum_{n=1}^{n_{\max}} n(q_{n-1} - q_n) = \sum_{n=1}^{n_{\max}} \Big(n d_n \prod_{j=0}^{n-1} (1 - d_j)\Big), \tag{6.2}$$

gdzie $n_{\max}$ jest maksymalnym wiekiem w populacji. Podobnie można wyznaczyć wariancję i odchylenie standardowe średniej długości życia. W danych statystycznych posługujemy się *medianą długości życia*, czyli chcemy wyznaczyć taki wiek, do którego dożyjemy z prawdopodobieństwem 0,5. Dokładniej, mediana będzie taką liczbą naturalną n, że $q_n \geq 0{,}5 \geq q_{n+1}$.

Możemy również wyznaczyć rozkład pozostałej długości życia pod warunkiem, że osoba jest w wieku i, oraz średnią, wariancję i medianę tego rozkładu. Przypominamy, że zgodnie ze wzorem (2.9) szansa przeżycia kolejnych k lat dla osoby w wieku i wynosi

$$p_{i\,i+k} = \prod_{j=i+1}^{i+k} (1 - d_j),$$

a więc prawdopodobieństwo, że przeżyje ona dokładnie k kolejnych lat, wynosi

$$p_{i\,i+k-1} - p_{i\,i+k} = \prod_{j=i+1}^{i+k-1} (1 - d_j) - \prod_{j=i+1}^{i+k} (1 - d_j) = d_{i+k} \prod_{j=i+1}^{i+k-1} (1 - d_j).$$

Znalezienie wartości oczekiwanej, wariancji i mediany tego rozkładu pozostawiamy czytelnikowi jako zadanie.

Uwaga I.67. Korzystając z danych w tabeli I.1, możemy w sposób przybliżony wyznaczyć wymienione wyżej charakterystyki demograficzne. W tabeli podano liczbę zgonów Z w ciągu roku na 100 tys. ludności w 5-letnich przedziałach wieku. Przyjmiemy, że współczynniki śmiertelności są takie same dla wszystkich roczników z tego samego przedziału wieku i wynoszą $d_i = 10^{-5} Z$. Tabela podaje tylko sumaryczne dane do przedziału wieku 85+; do obliczenia wartości oczekiwanej przyjmujemy, że współczynniki śmiertelności w następnych przedziałach wieku wynoszą: $d = 0{,}12$ dla 85–89, $d = 0{,}19$ dla 90–94 i $d = 0{,}28$ dla 95–99, a starszych osób nie uwzględniamy. Ponieważ śmiertelność w wieku 1–49 nie przekracza 0,004, wygodnie jest używać w tym przedziale wieku wzoru przybliżonego $\prod_{i=1}^{n}(1 - d_i) = 1 - \sum_{i=1}^{n} d_i$.

Rozważmy teraz model McKendricka z punktu 2.5. Przypominamy, że był to model z czasem ciągłym, w którym prawdopodobieństwo, że osobnik w wieku a w chwili t nie dożyje do chwili $t + \Delta a$, wynosi $\mu(t, a)\Delta a + o(\Delta a)$. Funkcję $\mu(t, a)$ nazwaliśmy współczynnikiem śmiertelności. Niech zmienna losowa T oznacza długość życia osobnika i niech $G(a) = \mathrm{P}(T > a)$. Funkcję G nazywamy *funkcją przeżycia* i wprost z jej definicji otrzymujemy

$$\begin{aligned} G(a) - G(a + \Delta a) &= \mathrm{P}(a < T \leq a + \Delta a) \\ &= \mathrm{P}(a < T \leq a + \Delta a \mid T > a)\, \mathrm{P}(T > a). \end{aligned}$$

Zakładamy, że osobnik urodził się w chwili t_0, więc zgodnie z definicją współczynnika śmiertelności mamy

$$\mathrm{P}(a < T \le a + \Delta a \mid T > a) = \mu(t_0 + a, a)\Delta a + o(\Delta a),$$

a więc

$$\frac{G(a) - G(a + \Delta a)}{G(a)} = \mu(t_0 + a, a)\Delta a + o(\Delta a).$$

Po podzieleniu obu stron przez Δa i przejściu z Δa do granicy w zerze otrzymujemy

$$-\frac{G'(a)}{G(a)} = \mu(t_0 + a, a).$$

Ostatnią równość możemy zapisać w postaci

$$\frac{d}{da} \ln G(a) = -\mu(t_0 + a, a),$$

skąd po scałkowaniu względem a otrzymujemy

$$\ln G(a) - \ln G(0) = -\int_0^a \mu(t_0 + s, s)\, ds.$$

Ponieważ $G(0) = \mathrm{P}(T > 0) = 1$, ostatecznie

$$G(a) = \exp\Big(-\int_0^a \mu(t_0 + s, s)\, ds\Big). \tag{6.3}$$

Funkcja G jest dopełnieniem dystrybuanty F_T zmiennej losowej T, tzn. $G(a) + F_T(a) = 1$, przy czym dla $a < 0$ formalnie przyjmujemy, że $G(a) = 1$.

Zauważmy, że zmienna losowa T wyrażająca czas życia osobnika ma gęstość f_T określoną wzorem $f_T(a) = 0$ dla $a < 0$ oraz

$$f_T(a) = \mu(t_0 + a, a) \exp\Big(-\int_0^a \mu(t_0 + s, s)\, ds\Big) \quad \text{dla } a \ge 0. \tag{6.4}$$

Wzór (6.4) pozwala na wyznaczenie średniej i wariancji długości życia.

Niech $G_x(a)$ będzie prawdopodobieństwem, że *pozostała długość życia* osobnika w wieku x jest większa niż a. Wtedy

$$G_x(a) = \mathrm{P}(T > x + a \mid T \ge x) = \frac{\mathrm{P}(T > x + a)}{\mathrm{P}(T \ge x)} = \frac{G(x + a)}{G(x)}.$$

Oznacza to, że *średnia pozostała długość życia* osobnika w wieku x jest określona wzorem

$$\begin{aligned}\mathrm{E}(T - x \mid T \ge x) &= -\int_0^{a_m - x} a G_x'(a)\, da = \int_0^{a_m - x} G_x(a)\, da \\ &= \int_0^{a_m - x} \frac{G(x + a)}{G(a)}\, da.\end{aligned}$$

W szczególności *średnia długość życia* wynosi

$$\mathrm{E}\,T = \int_0^{a_m} G(a)\,da.$$

Rozważmy teraz populację, w której współczynniki urodzeń b i śmiertelności μ nie zależą od czasu, a jedynie od wieku osobnika. Rozwiązania układu równań (2.16)–(2.17) dążą do rozwiązania postaci $u(t,a) = Ce^{\lambda_0 t}p_*(a)$, tzn. $\lim_{t\to\infty} e^{-\lambda_0 t}u(t,a) = Cp_*(a)$, gdzie $p_*(a)$ jest rozkładem wieku osobników (profilem wiekowym) określonym wzorem

$$p_*(a) = C_1 e^{-\lambda_0 a - \int_0^a \mu(s)\,ds},$$

liczba λ_0 jest jedynym rozwiązaniem równania

$$\int_0^{a_m} b(a)e^{-\lambda a}\exp\Big\{-\int_0^a \mu(s)\,ds\Big\}\,da = 1,$$

a C_1 jest stałą normującą (patrz [324, uwaga V.16 oraz rozdz. V, punkt 3.4]). Ponieważ funkcja przeżycia G zdefiniowana jest wzorem

$$G(a) = e^{-\int_0^a \mu(s)\,ds},$$

mamy interesujący związek między funkcją przeżycia G, profilem wiekowym $p_*(a)$ oraz wykładnikiem wzrostu populacji λ_0:

$$p_*(a) = C_1 e^{-\lambda_0 a}G(a). \tag{6.5}$$

Niech $Q(a) = \int_0^a \mu(s)\,ds$ i niech ξ będzie zmienną losową o rozkładzie wykładniczym z parametrem $\alpha = 1$. Ponieważ

$$\mathrm{P}(T > a) = e^{-Q(a)} = \mathrm{P}(\xi > Q(a)),$$

rozkłady zmiennych losowych $Q(T)$ i ξ są takie same. Mówimy wtedy, że zmienne te są *równe względem rozkładu*, i piszemy $Q(T) \stackrel{\mathrm{d}}{=} \xi$. Jeżeli śmiertelność μ nie zależy od wieku, to $Q(a) = \mu a$ i wtedy $T \stackrel{\mathrm{d}}{=} \xi/\mu$.

Na koniec tych dość teoretycznych rozważań dotyczących funkcji przeżycia przedstawimy dwa przykłady.

Przykład I.68. Jeżeli $\mu(a) = ca^{k-1}$, gdzie $c, k > 0$, to

$$G(a) = \exp\Big(-\int_0^a \mu(s)\,ds\Big) = \exp\Big(-\int_0^a cs^{k-1}\,ds\Big) = e^{-ck^{-1}a^k}.$$

Często dobieramy takie $\lambda > 0$, że $c = k\lambda^{-k}$; rozkład z funkcją przeżycia $G(a) = e^{-(a/\lambda)^k}$ nazywamy *rozkładem Weibulla* z parametrem skali λ i parametrem kształtu k.

Przykład I.69. Badając dane demograficzne, zaobserwowaliśmy, że w dość dużym przedziale wieku współczynnik śmiertelności rośnie wykładniczo (patrz rysunek I.5), a więc spełnia równanie $\mu'(a) = c\mu(a)$, gdzie c jest pewną stałą dodatnią. Zjawisko to można wyjaśnić w ten sposób, że w wyniku procesu starzenia się rośnie liczba negatywnych zmian w organizmie, a współczynnik śmiertelności wyraża sumę tych zmian. Zmiany te działają kaskadowo i powodują pojawienie się nowych zmian negatywnych, co daje zależność $\mu'(a) = c\mu(a)$. Wobec tego $\mu(a) = \mu_0 e^{ca}$, a stąd

$$G(a) = \exp\Big(-\int_0^a \mu_0 e^{cs}\,ds\Big) = \exp\Big(\frac{\mu_0}{c}(1-e^{ca})\Big).$$

Rozkład z tak wyznaczoną funkcją przeżycia nazywamy *rozkładem Gompertza*.

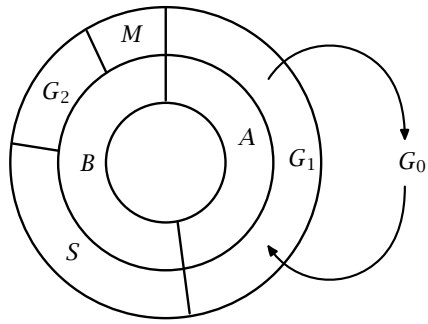
Wyznaczenie podstawowych wielkości demograficznych dla modeli o rozkładach Weibulla i Gompertza pozostawiamy czytelnikowi jako zadania I.53–I.55.

6.2. Modele pokoleniowe cyklu komórkowego

Cyklem komórkowym nazywamy ciąg procesów zachodzących w komórce i prowadzących do jej podziału. Procesy te regulowane są przez skomplikowaną sieć interakcji między białkami – np. uproszczony model cyklu komórkowego u ssaków opisany jest za pomocą układu złożonego z osiemnastu równań różniczkowych [275].

Rozróżnia się cztery podstawowe fazy cyklu komórkowego [7, 175, 269]. Pierwsza faza, oznaczana zwykle przez G_1, jest *fazą wzrostu* komórki. W tej fazie komórka dojrzewa do podziału i syntetyzuje różnorodne enzymy. Ponieważ komórki bezpośrednio po podziale mogą się istotnie różnić pod względem dojrzałości mierzonej np. ich wielkością lub zawartością materiału genetycznego potrzebnego do rozpoczęcia procesu podziału komórki, więc długość fazy G_1 jest bardzo zróżnicowana nawet dla komórek tego samego rodzaju i w dużym stopniu zależy od stanu dojrzałości komórki w chwili początkowej. W kolejnej fazie cyklu komórkowego, oznaczanej przez S, następuje synteza DNA i można przyjąć, że od tej fazy zaczynają się procesy związane z podziałem komórki. W następnej fazie G_2 w komórce wytwarzane są białka niezbędne w procesie mitozy, tj. podziału komórkowego. W ostatniej fazie M (mitozy lub mejozy) następuje podział jądra komórkowego i podział komórki.

Łączna długość faz S, G_2 i M jest praktycznie jednakowa dla wszystkich komórek tego samego rodzaju/gatunku. Rozważana jest również faza G_0 (spoczynkowa). Komórki wielokomórkowych organizmów eukariotycznych zwykle wchodzą w fazę G_0 z fazy G_1 i mogą pozostawać w fazie spoczynkowej przez



Rysunek I.16. Schematyczny model cyklu komórkowego

długi okres, ewentualnie aż do śmierci, lub po pewnym czasie wrócić do fazy G_1. Schematyczny model cyklu komórkowego przedstawiony jest na rysunku I.16.

W literaturze można spotkać różne modele cyklu komórkowego [296], np. model czterofazowy [30], ale najbardziej popularne są modele jedno lub dwufazowe. W modelu jednofazowym łączymy fazy G_1, S, G_2 i M w jedną, a zaniedbujemy fazę G_0. Drugą kategorią są modele dwufazowe. Biolodzy zwykle łączą fazy G_1, S, G_2 w tzw. interfazę, drugą fazą jest faza M. Z matematycznego punktu widzenia lepszy jest podział na fazę wzrostu $A = G_1$ o losowej długości t_A i fazę podziału B składającą się z faz S, G_2 i M o praktycznie stałej długości t_B. W tych modelach również nie uwzględnia się fazy G_0. Można spotkać też modele dwufazowe, w których łączy się fazy G_1, S, G_2 i M w jedną, a drugą fazą jest faza spoczynkowa G_0 [16, 310]. Niektóre modele cyklu komórkowego i ich własności badane są w [324, rozdz. V], dlatego nasze rozważania ograniczymy do pewnych zagadnień probabilistycznych.

W modelach cyklu komórkowego stan komórki opisany jest parametrem lub zespołem parametrów opisujących *dojrzałość komórki*. Dojrzałość utożsamiana jest z wiekiem fizjologicznym komórki (zegarem komórkowym), decydującym o przechodzeniu komórki przez kolejne fazy rozwoju. Rozpoczniemy od stosunkowo prostego modelu Lasoty i Mackeya [221], w którym cały cykl składa się z jednej fazy, a dojrzałością jest wielkość komórki $x > 0$ (objętość, masa). Będzie nas interesować zależność między dojrzałością komórek na początku cyklu komórkowego w kolejnych pokoleniach komórek. Niech $g(x)$ będzie współczynnikiem wzrostu dojrzałości, a więc dojrzałość x rośnie według równania

$$\frac{dx}{dt} = g(x). \tag{6.6}$$

O funkcji g załóżmy, że ma ciągłą i ograniczoną pochodną oraz $g(x) > 0$ dla $x > 0$.

Przyjmijmy, że prawdopodobieństwo, iż komórka o dojrzałości x podzieli się w czasie Δt, wynosi $\Delta \mathrm{P} = \psi(x)\Delta t + o(\Delta t)$. O funkcji ψ będziemy zakładać, że jest ciągła i istnieje $m_0 > 0$ spełniające warunek $\psi(x) = 0$ dla $x \le m_0$ i $\psi(x) > 0$ dla $x > m_0$. Zgodnie z przyjętym założeniem minimalna dojrzałość komórki wynosi $m_0/2$. Ponieważ

$$\Delta \mathrm{P} = \psi(x)\,\Delta t + o(\Delta t), \qquad \Delta x = g(x)\,\Delta t + o(\Delta t),$$

więc

$$\Delta \mathrm{P} = \frac{\psi(x)}{g(x)}\,\Delta x + o(\Delta x).$$

Jeżeli przyjmiemy, że $\mu(x) = \frac{\psi(x)}{g(x)}$, i utożsamimy x z wiekiem a, to możemy mówić o funkcji przeżycia G zależnej od dojrzałości x i w tym wypadku wynoszącej

$$G(x) = \exp\Big\{-\int_{x_0}^{x} \frac{\psi(s)}{g(s)}\,ds\Big\},$$

gdzie x_0 jest dojrzałością komórki na początku cyklu komórkowego, a więc w chwili $t = 0$. Ponieważ minimalna dojrzałość komórki wynosi $m_0/2$, więc przyjmując oznaczenie

$$Q(x) = \int_{m_0/2}^{x} \frac{\psi(s)}{g(s)}\,ds,$$

otrzymujemy

$$G(x) = e^{Q(x_0)-Q(x)}.$$

Będziemy zakładać, że $\lim_{x\to\infty} Q(x) = \infty$, co gwarantuje, że komórka podzieli się z prawdopodobieństwem 1.

Niech teraz X będzie dojrzałością komórki w momencie podziału i niech η będzie zmienną losową o rozkładzie wykładniczym z parametrem $\alpha = 1$. Ponieważ

$$\mathrm{P}(X > x) = e^{Q(x_0)-Q(x)} = \mathrm{P}(\eta > Q(x) - Q(x_0)),$$

więc $\mathrm{P}(Q(X) > Q(x)) = \mathrm{P}(Q(x_0) + \eta > Q(x))$, skąd $X \overset{\mathrm{d}}{=} Q^{-1}(Q(x_0) + \eta)$. Podsumowując, stwierdzamy, że jeżeli komórka miała dojrzałość x_0 na początku cyklu komórkowego, to dojrzałość komórek potomnych na początku ich cyklu komórkowego określona jest za pomocą zmiennej losowej

$$\xi = \frac{1}{2}Q^{-1}(Q(x_0) + \eta). \tag{6.7}$$

Korzystając ze wzoru (6.7), możemy znajdować różne charakterystyki związane z modelem, które z kolei umożliwiają jego weryfikację na podstawie eksperymentów. Ze wzoru (6.7) można np. wyznaczyć zmienną losową T opisującą długość cyklu komórkowego, a więc czas życia komórki do momentu jej

podziału. Niech $\pi_t x_0$ będzie rozwiązaniem równania (6.6) w chwili t, spełniającym warunek początkowy $x(0) = x_0$. Jeżeli komórka w chwili początkowej miała dojrzałość x_0, to T spełnia zależność

$$\pi_T x_0 \stackrel{\mathrm{d}}{=} \frac{1}{2} Q^{-1}(Q(x_0) + \eta). \tag{6.8}$$

Prawdopodobieństwo podziału komórki w czasie $[t, t + \Delta t]$ pod warunkiem, że podział ten nie nastąpił wcześniej, wynosi $\Delta \mathrm{P} = \psi(\pi_t x_0)\Delta t + o(\Delta t)$, więc rozkład długości cyklu komórkowego można wyrazić bezpośrednio wzorem

$$\mathrm{P}(T > a) = \exp\Big(-\int_0^a \psi(\pi_t x_0)\, dt\Big). \tag{6.9}$$

Szczególnie interesujące są zależności między długością cyklu komórkowego komórek siostrzanych (powstających w wyniku podziału tej samej komórki) oraz między dojrzałością ich komórek potomnych [221]. Jeżeli np. komórki siostrzane miały początkową dojrzałość x_0, to początkowa dojrzałość komórek potomnych ξ_1 i ξ_2 wynosi $\xi_i = \frac{1}{2} Q^{-1}(Q(x_0) + \eta_i)$, $i = 1, 2$, przy czym zmienne losowe η_1 i η_2 są niezależne i mają standardowy rozkład wykładniczy.

Bardziej zaawansowany jest dwufazowy model cyklu komórkowego Tyrchy [371]. W modelu tym występują dwie fazy: faza wzrostu A i faza podziału B. Podobnie jak w poprzednim modelu dojrzałość rośnie według równania (6.6), a funkcja ψ w tym wypadku jest intensywnością przejścia z fazy A do B. Komórka przebywa w fazie B ustalony okres czasu t_B, a następnie dzieli się. Niech x_0 będzie początkową dojrzałością komórki. Ponieważ w momencie wejścia do fazy B dojrzałość komórki jest zmienną losową określoną wzorem $Q^{-1}(Q(x_0) + \eta)$, a długość drugiej fazy wynosi t_B, więc dojrzałość komórki potomnej na początku cyklu komórkowego wynosi

$$\xi = \frac{1}{2} \pi_{t_B}(Q^{-1}(Q(x_0) + \eta)). \tag{6.10}$$

6.3. Fragmentacja

W punkcie 6.2 rozważaliśmy cykl komórkowy, w którym komórka dzieliła się na dwie o tej samej wielkości. Można rozważać modele, w których komórki potomne mogą mieć różne rozmiary. Również w innych procesach biologicznych można rozpatrywać podziały niesymetryczne. Komórki fitoplanktonu często występują w grupach zwanych *agregatami*; w badaniu dynamiki agregatów ważną rolę odgrywają procesy fragmentacji, a więc podziału na mniejsze agregaty.

Ograniczymy rozważania do podziałów na dwie części. Niech x będzie wielkością agregatu przed podziałem i niech ξ_1 i ξ_2 będą wielkościami agregatów

powstałych po podziale. Wtedy $\xi_1 + \xi_2 = x$ oraz obie zmienne losowe mają ten sam rozkład $\mathrm{P}_x(A)$, tj. $\mathrm{P}_x(A) = \mathrm{P}(\xi_1 \in A) = \mathrm{P}(\xi_2 \in A)$ dla podzbiorów borelowskich A przedziału $[0,x]$. Z przyjętych założeń o zmiennych losowych ξ_1 i ξ_2 wynika, że $\mathrm{P}_x([0,y]) = \mathrm{P}_x([x-y,x])$ dla $y \in [0,x]$. Ponieważ $\mathrm{E}\,\xi_1 + \mathrm{E}\,\xi_2 = x$ oraz $\mathrm{E}\,\xi_1 = \mathrm{E}\,\xi_2$, więc $\mathrm{E}\,\xi_1 = \mathrm{E}\,\xi_2 = \frac{1}{2}$. Kowariancja zmiennych losowych ξ_1 i ξ_2 wynosi

$$\begin{aligned}\operatorname{Cov}(\xi_1,\xi_2) &= \mathrm{E}((\xi_1 - \mathrm{E}\,\xi_1)(\xi_2 - \mathrm{E}\,\xi_2)) = \mathrm{E}((\xi_1 - \tfrac{1}{2})(\xi_2 - \tfrac{1}{2}))\\ &= \mathrm{E}((\xi_1 - \tfrac{1}{2})(x - \xi_1 - \tfrac{1}{2})) = \mathrm{E}((\xi_1 - \tfrac{1}{2})x) - \mathrm{E}((\xi_1 - \tfrac{1}{2})(\xi_1 + \tfrac{1}{2}))\\ &= -\mathrm{E}(\xi_1^2 - (\tfrac{1}{2})^2) = -\mathrm{E}(\xi_1^2) + (\mathrm{E}\,\xi_1)^2 = -\mathrm{D}^2\,\xi_1,\end{aligned}$$

a współczynnik korelacji wynosi -1, co ściśle wiąże się z zależnością $\xi_2 = x - \xi_1$.

Załóżmy dodatkowo, że miara $\mu(A) = \mathrm{P}_x(A)$ ma gęstość, którą będziemy oznaczać przez $p(y|x)$, a więc $\int_A p(y|x)\,dy = \mathrm{P}_x(A)$ dla dowolnego zbioru borelowskiego $A \subset [0,x]$, i niech $p(y|x)$ będzie funkcją mierzalną zmiennych (x,y). Jeżeli ξ jest zmienną losową opisującą rozkład wielkości agregatu przed podziałem, a f_{ξ_1} jest gęstością zmiennej losowej ξ_1, to

$$f_{\xi_1}(y \mid \xi = x) = p(y|x),$$

a jeżeli zmienna losowa ξ ma gęstość f, to

$$f_{\xi_1}(y) = \int_y^\infty p(y|x) f(x)\,dx. \tag{6.11}$$

Niech funkcja $u(t,x)$ opisuje rozkład wielkości agregatów w chwili t, tj. $\int_A u(t,x)\,dx$ jest liczbą agregatów o wielkości ze zbioru A w chwili t. Załóżmy, że prawdopodobieństwo podziału agregatu wielkości x w czasie Δt wynosi $\psi(x)\Delta t + o(\Delta t)$. Wtedy w czasie Δt z dokładnością do $o(\Delta t)$:

(a) $\psi(x)u(t,x)\Delta t$ agregatów wielkości x podzieli się,

(b) powstanie $\int_x^\infty p(x|z)\psi(z)u(t,z)\,dz\,\Delta t$ agregatów o wielkości x w wyniku podziału większych agregatów.

Korzystając z (a) i (b), łatwo znajdujemy *równanie fragmentacji*

$$\frac{\partial u}{\partial t}(t,x) = \int_x^\infty p(x|z)\psi(z)u(t,z)\,dz - \psi(x)u(t,x). \tag{6.12}$$

Wzór (6.11), w połączeniu ze wzorami (6.7) i (6.10), pozwala na wyznaczenie gęstości rozkładów wielkości komórek potomnych na początku cyklu komórkowego dla modeli z punktu 6.2 z podziałem niesymetrycznym.

6.4. Szacowanie wielkości populacji

Szacowanie wielkości populacji należy do podstawowych zagadnień ekologii. Mimo iż literatura dotycząca tej tematyki jest obszerna (np. [11, 203, 301, 347]), dominują w niej opracowania poświęcone rozwiązaniom praktycznym, z gotowymi wzorami i algorytmami, bez precyzyjnego matematycznego uzasadnienia nawet uproszczonych modeli. Ze względu na złożoność problemu trudno wybrać model stosunkowo łatwy do przeanalizowania, a jednocześnie użyteczny praktycznie. Jest to zagadnienie statystyczne, więc metody, których możemy użyć, są ograniczone. Przy szacowaniu wielkości populacji używana jest *metoda wielokrotnych złowień* (ang. *capture-recapture* lub *mark and recapture*), polegająca na tym, że po „złowieniu" osobników znakujemy je, a po ponownych złowieniach, na podstawie liczebności osobników wcześniej oznakowanych, szacujemy wielkość całej populacji.

Rozpoczniemy od najprostszej metody opartej na dwóch złowieniach, która wymaga spełnienia dość rygorystycznych założeń dotyczących zachowania populacji:

- prawdopodobieństwo schwytania każdego osobnika jest takie samo,
- populacja jest zamknięta, a więc nie zachodzi migracja osobników między kolejnymi złowieniami,
- prawdopodobieństwo ponownego schwytania osobnika już złowionego i oznakowanego jest takie samo jak przy pierwszym złowieniu.

Niech N będzie nieznaną wielkością populacji, N_1 i N_2 liczbą schwytanych i oznakowanych osobników w kolejnych złowieniach, a n – liczbą osobników złowionych dwukrotnie. Przyjmując, że udział osobników oznakowanych w całej populacji jest taki sam jak w grupie osobników powtórnie schwytanych, otrzymujemy zależność

$$\frac{N_1}{N} = \frac{n}{N_2},$$

skąd uzyskujemy oszacowanie (estymator) wartości N w postaci *wzoru Lincolna–Petersena*

$$\hat{N} = \frac{N_1 N_2}{n}. \tag{6.13}$$

Wzór (6.13) można też wyprowadzić w inny sposób. Załóżmy, że mamy ustalone wielkości N, N_1 i N_2. Wtedy n ma rozkład hipergeometryczny (4.7) z parametrami $B = N_1$, $C = N - N_1$, $K = N_2$, a więc określony wzorem

$$\mathrm{P}(\xi = n) = \frac{\binom{N_1}{n}\binom{N-N_1}{N_2-n}}{\binom{N}{N_2}}. \tag{6.14}$$

Ustalmy N_1, N_2, n i zbadajmy, jak zmienia się $\mathrm{P}(\xi = n)$ przy zmianie N. Oznaczmy to prawdopodobieństwo przez $q(N)$. Wtedy

$$\frac{q(N+1)}{q(N)} = \frac{\binom{N}{N_2}\binom{N+1-N_1}{N_2-n}}{\binom{N+1}{N_2}\binom{N-N_1}{N_2-n}} = \frac{(N+1-N_1)(N+1-N_2)}{(N+1)(N+1-N_1-N_2+n)}$$
$$= 1 + \frac{N_1N_2 - n(N+1)}{(N+1)(N+1-N_1-N_2+n)}.$$

Ustalmy $\bar{N}$ tak, aby $n\bar{N} \le N_1N_2 < n(\bar{N}+1)$, a więc

$$\bar{N} = \left[\frac{N_1N_2}{n}\right].$$

Jeżeli N_1N_2/n nie jest liczbą całkowitą, to $\frac{q(N+1)}{q(N)} > 1$ dla $N < \bar{N}$ oraz $\frac{q(N+1)}{q(N)} < 1$ dla $N \ge \bar{N}$, a więc ciąg $(q(N))$ ma maksimum dla $N = \bar{N}$. Jeżeli N_1N_2/n jest liczbą całkowitą, to ciąg $(q(N))$ ma maksimum dla $N = \bar{N} - 1$ i $N = \bar{N}$. Używając języka statystyki, wykazaliśmy, że $\bar{N}$ jest *estymatorem największej wiarygodności* dla N. Estymator $\bar{N}$ różni się nieznacznie od estymatora $\hat{N}$ obliczonego według wzoru Lincolna–Petersena.

Okazuje się, że wyznaczona w ten sposób wielkość populacji jest przeszacowana, w szczególności gdy wyrazy występujące we wzorze (6.13) są stosunkowo małe. Chapman [71] podał znacznie dokładniejszy estymator dla N:

$$\hat{N}_C = \frac{(N_1+1)(N_2+1)}{n+1} - 1. \tag{6.15}$$

Wyrażenie $\hat{N}_C$ nosi nazwę *estymatora Chapmana*. Ponieważ $\hat{N}_C$ nie musi być liczbą całkowitą, przyjmujemy, że szacowana wielkość populacji wynosi $[\hat{N}_C]$.

Przykład I.70. Aby porównać oba estymatory dla N, rozważmy następujący przykład. Niech $N_1 = 10$, $N_2 = 10$ i $n = 2$. Wtedy $\hat{N} = 10 \cdot 10/2 = 50$, podczas gdy $\hat{N}_C = 11 \cdot 11/3 - 1 = 39\frac{1}{3}$. Różnica oszacowań wielkości populacji za pomocą obu estymatorów jest duża w stosunku do szacunkowej wielkości populacji, nawet przy spełnieniu dość rygorystycznych założeń dotyczących rozpatrywanego modelu. Przy dużych wartościach n różnice oszacowań nie są już tak znaczne.

Z praktycznego punktu widzenia należałoby jeszcze powiedzieć, z jakim błędem wyznaczyliśmy wartość N, lub używając terminologii statystycznej, określić przedział ufności dla ustalonych współczynników ufności. Ograniczymy się do podania wariancji:

$$\operatorname{Var} N_C = \frac{(N_1+1)(N_2+1)(N_1-n)(N_2-n)}{(n+1)^2(n+2)}. \tag{6.16}$$

Różnice w przyjętych oszacowaniach wielkości N związane są z wyborem właściwego modelu – czy np. przy oznakowaniu lub powtórnym odłowie staramy się schwytać maksymalną liczbę osobników, czy też liczba ta jest ustalona z góry. Interesujące rozważania na ten temat można znaleźć w pracy [385], w której podano różne estymatory w zależności od przyjętego modelu i wyznaczono momenty wielkości N. Pominiemy wyprowadzenie wzorów (6.15) i (6.16), a zainteresowanych czytelników odsyłamy do książki [347]. Korzystając z pracy [385], przedstawimy podobne modele i dla jednego z nich znajdziemy $\mathrm{E}\,N$ i $\mathrm{Var}\,N$.

Rozpoczniemy od modelu, w którym z góry ustalamy liczbę osobników do odłowu, a więc N_1 i N_2 są znane przed eksperymentem. W analizie modelu będziemy korzystać z pewnych faktów związanych z funkcjami hipergeometrycznymi, nie wchodząc głębiej w ich teorię. Dla liczb naturalnych a, b, c spełniających nierówność $c \geq a + b + 2$ definiujemy funkcję

$$f(a,b,c) = \sum_{n=0}^{\infty} \frac{(n+a)!(n+b)!}{n!(n+c)!}. \tag{6.17}$$

Funkcja f spełnia *tożsamość Gaussa*

$$f(a,b,c) = \frac{a!b!(c-a-b-2)!}{(c-a-1)!(c-b-1)!}. \tag{6.18}$$

Stąd wnioskujemy, że

$$\begin{aligned} \frac{f(a+1,b+1,c+1)}{f(a,b,c)} &= \frac{(a+1)!(b+1)!(c-a-b-3)!}{a!b!(c-a-b-2)!} \\ &= \frac{(a+1)(b+1)}{c-a-b-2}. \end{aligned} \tag{6.19}$$

Aby uniknąć długich wzorów, będziemy stosować uproszczony zapis prawdopodobieństw warunkowych i np. oznaczać przez $\mathrm{P}(n \mid N_1, N_2, N)$ prawdopodobieństwo, że liczba osobników złowionych dwukrotnie wynosi n pod warunkiem, że w pierwszym i drugim odłowie i w całej populacji jest odpowiednio N_1, N_2, N osobników. Zgodnie ze wzorem (6.14),

$$\mathrm{P}(n \mid N_1, N_2, N) = \frac{N_1!N_2!(N-N_1)!(N-N_2)!}{n!(N_1-n)!(N_2-n)!(N-N_s)!N!}, \tag{6.20}$$

gdzie $N_s = N_1 + N_2 - n$ jest liczbą osobników schwytanych co najmniej raz.

Chcemy obliczyć $\mathrm{P}(N \mid N_1, N_2, n)$. W tym celu skorzystamy z następującej ogólnej obserwacji:

$$\begin{aligned} \mathrm{P}(A|B\cap C) &= \frac{\mathrm{P}(A\cap B\cap C)}{\mathrm{P}(B\cap C)} = \frac{\mathrm{P}(C|A\cap B)\,\mathrm{P}(A\cap B)}{\mathrm{P}(B\cap C)} \\ &= \frac{\mathrm{P}(C|A\cap B)\,\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(C|B)\,\mathrm{P}(B)} = \frac{\mathrm{P}(C|A\cap B)\,\mathrm{P}(A|B)}{\mathrm{P}(C|B)}. \end{aligned}$$

Stąd

$$\mathrm{P}(N \mid N_1, N_2, n) = \frac{\mathrm{P}(n \mid N_1, N_2, N)\,\mathrm{P}(N \mid N_1, N_2)}{\mathrm{P}(n \mid N_1, N_2)} \tag{6.21}$$

i zgodnie ze wzorem (6.20) mamy

$$\mathrm{P}(N \mid N_1, N_2, n) = \frac{N_1!N_2!(N-N_1)!(N-N_2)!\,\mathrm{P}(N \mid N_1, N_2)}{n!(N_1-n)!(N_2-n)!(N-N_s)!N!\,\mathrm{P}(n \mid N_1, N_2)}.$$

Składnik

$$\frac{N_1!N_2!}{n!(N_1-n)!(N_2-n)!\,\mathrm{P}(n \mid N_1, N_2)}$$

nie zależy od N, a ponieważ liczby N_1 i N_2 były ustalone z góry niezależnie od N, przyjmujemy, że wyrażenie $\mathrm{P}(N \mid N_1, N_2)$ nie zależy od $N \geq N_s$. Należy tu zaznaczyć, że przyjęty rozkład jest *a priori* niewłaściwy, bo $\sum_N \mathrm{P}(N \mid N_1, N_2) = \infty$. Z przyjętych założeń wynika, że

$$\mathrm{P}(N \mid N_1, N_2, n) = \frac{C(N-N_1)!(N-N_2)!}{(N-N_s)!N!}, \tag{6.22}$$

gdzie stałą normującą C dobieramy tak, aby $\sum_{N=N_s}^{\infty} \mathrm{P}(N \mid N_1, N_2, n) = 1$.

Niech $K = N - N_s$, $n_1 = N_1 - n$ oraz $n_2 = N_2 - n$. Wtedy $N - N_1 = K + n_2$, $N - N_2 = K + n_1$, $N - N_s = K$ oraz $N = K + N_s$. Zgodnie ze wzorami (6.22) i (6.17),

$$\mathrm{P}(N = N_s + K \mid N_1, N_2, n) = \frac{C(K+n_1)!(K+n_2)!}{K!(K+N_s)!}, \quad K = 0, 1, \ldots, \tag{6.23}$$

gdzie $C = \big(f(n_1, n_2, N_s)\big)^{-1}$. Wyznaczymy $\mathrm{E}N$ i $\operatorname{Var} N$. Zauważmy, że $\mathrm{E}N = N_s + \mathrm{E}K$ oraz $\operatorname{Var} N = \operatorname{Var} K$. Mamy

$$\begin{aligned}
\mathrm{E}K &= \sum_{k=0}^{\infty} \frac{C(k+n_1)!(k+n_2)!k}{k!(k+N_s)!} = \sum_{k=0}^{\infty} \frac{C(k+n_1+1)!(k+n_2+1)!k}{k!(k+N_s+1)!} \\
&= \frac{f(n_1+1, n_2+1, N_s+1)}{f(n_1, n_2, N_s)},
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}K^2 &= \sum_{k=0}^{\infty} \frac{C(k+n_1)!(k+n_2)!k^2}{k!(k+N_s)!} \\
&= \sum_{k=0}^{\infty} \frac{C(k+n_1)!(k+n_2)!k}{k!(k+N_s)!} + \sum_{k=0}^{\infty} \frac{C(k+n_1)!(k+n_2)!k(k-1)}{k!(k+N_s)!} \\
&= \sum_{k=0}^{\infty} \frac{C(k+n_1+1)!(k+n_2+1)!}{k!(k+N_s+1)!} + \sum_{k=0}^{\infty} \frac{C(k+n_1+2)!(k+n_2+2)!}{k!(k+N_s+2)!} \\
&= \frac{f(n_1+1, n_2+1, N_s+1)}{f(n_1, n_2, N_s)} + \frac{f(n_1+2, n_2+2, N_s+2)}{f(n_1, n_2, N_s)}.
\end{aligned}$$

Korzystając ze wzoru (6.19), otrzymujemy

$$\begin{aligned}
\mathrm{E}K &= \frac{(n_1+1)(n_2+1)}{N_s-n_1-n_2-2} = \frac{(n_1+1)(n_2+1)}{n-2},\\
\mathrm{E}K^2 &= \mathrm{E}K + \frac{f(n_1+2,n_2+2,N_s+2)}{f(n_1+1,n_2+1,N_s+1)}\,\frac{f(n_1+1,n_2+1,N_s+1)}{f(n_1,n_2,N_s)}\\
&= (\mathrm{E}K)\left(1+\frac{(n_1+2)(n_2+2)}{N_s-n_1-n_2-3}\right) = (\mathrm{E}K)\left(1+\frac{(n_1+2)(n_2+2)}{n-3}\right).
\end{aligned}$$

Stąd

$$\begin{aligned}
\operatorname{Var}K &= \mathrm{E}K^2 - (\mathrm{E}K)^2\\
&= \left(1+\frac{(n_1+2)(n_2+2)}{n-3} - \frac{(n_1+1)(n_2+1)}{n-2}\right)\mathrm{E}K\\
&= \frac{(n_1+1)(n_2+1)(N_1-1)(N_2-1)}{(n-2)^2(n-3)}.
\end{aligned}$$

Korzystając ze wzorów $\mathrm{E}N = N_s + \mathrm{E}K$ oraz $\operatorname{Var}N = \operatorname{Var}K$, ostatecznie otrzymujemy

$$\begin{aligned}
\mathrm{E}N &= \frac{(N_1-1)(N_2-1)}{n-1},\\
\operatorname{Var}N &= \frac{(N_1-n+1)(N_2-n+1)(N_1-1)(N_2-1)}{(n-2)^2(n-3)}
\end{aligned} \tag{6.24}$$

dla $n > 3$.

W pewnych modelach rozkład $\mathrm{P}(N \mid N_1, N_2, n)$ można wyznaczyć, korzystając ze wzoru

$$\mathrm{P}(N \mid N_1, N_2, n) = \frac{\mathrm{P}(n \mid N_1, N_2, N)\,\mathrm{P}(N_1|N)\,\mathrm{P}(N_2|N)\,\mathrm{P}(N)}{\mathrm{P}(N_1, N_2, n)}. \tag{6.25}$$

Dowód wzoru (6.25) jest podobny do dowodu (6.21) i pozostawiamy go czytelnikowi jako ćwiczenie.

Rozważmy teraz eksperyment, w którym staramy się złowić maksymalną liczbę osobników w obu odłowach. Przyjmijmy, że N_1 i N_2 są wybierane losowo spośród liczb $0, 1, \ldots, N$ z jednakowym prawdopodobieństwem, a więc $\mathrm{P}(N_1|N) = 1/(N+1)$ i $\mathrm{P}(N_2|N) = 1/(N+1)$. Jeżeli teraz *a priori* ustalimy rozkład wielkości N, czyli wartości $\mathrm{P}(N)$, otrzymamy

$$\mathrm{P}(N \mid N_1, N_2, n) = \frac{C(N-N_1)!(N-N_2)!\,\mathrm{P}(N)}{(N-N_s)!N!(N+1)^2}, \tag{6.26}$$

gdzie C jest stałą normującą. Jeżeli np. rozkład zmiennej N jest stały, to

$$\begin{aligned}
\mathrm{E}N &\approx \frac{(N_1+1)(N_2+1)}{n} - 2,\\
\operatorname{Var}N &\approx \frac{(N_1-n+1)(N_2-n+1)(N_1+1)(N_2+1)}{n^2(n-1)}
\end{aligned} \tag{6.27}$$

dla $n > 3$.

Znacznie lepsze oszacowania wielkości populacji uzyskujemy, wykonując więcej odłowień. Ograniczymy się do szkicowego przedstawienia metody [342] dla populacji zamkniętej. W tym wypadku dokonujemy m odłowień i za każdym razem znakujemy te osobniki, które nie zostały wcześniej oznakowane, po czym je wypuszczamy. Niech N_i oznacza liczbę osobników złowionych w i-tej próbie, M_i – liczbę osobników oznakowanych przed i-tą próbą, zaś n_i – liczbę osobników oznakowanych złowionych w i-tej próbie. Wielkość populacji szacujemy, stosując wyrażenie

$$\hat{N} = \frac{\sum_{i=2}^{m} M_i N_i}{\sum_{i=2}^{m} n_i}, \tag{6.28}$$

zwane *estymatorem Schnabel*. Można również wyznaczyć wartość $\mathrm{Var}(1/\hat{N})$, która podaje nam informację o błędzie:

$$\mathrm{Var}\left(\frac{1}{\hat{N}}\right) = \frac{\sum_{i=2}^{m} n_i}{(\sum_{i=2}^{m} M_i N_i)^2}. \tag{6.29}$$

Nie będziemy wyprowadzać wzorów (6.28) i (6.29), a jedynie zwrócimy uwagę na związek wzoru (6.28) ze wzorem Lincolna–Petersena (6.13). Zauważmy, że dla $m = 2$ wzory (6.13) i (6.28) są identyczne, bo $M_2 = N_1$. Ponieważ przed i-tym połowem oznakowaliśmy M_i osobników, iloraz $M_i N_i / n_i$ odpowiada wzorowi Lincolna–Petersena w przypadku, gdy pierwsze $i - 1$ prób traktujemy łącznie, a sprawdzamy, jaka jest liczba osobników oznakowanych w i-tym odłowie. Wyrażenie po prawej stronie wzoru (6.28) można więc traktować jako średnią ważoną z m prób.

Przykład I.71. W kolejnych pięciu dniach odłowiono i oznakowano pewną liczbę żab określonego gatunku. W tabeli I.6 podano liczbę N_i osobników odłowionych i liczbę n_i oznakowanych w każdej próbie oraz przedstawiono metodę obliczania M_i. Zgodnie ze wzorem (6.28) mamy

$$\hat{N} = \frac{52 \cdot 44 + 82 \cdot 38 + 108 \cdot 40 + 130 \cdot 32}{14 + 12 + 18 + 17} = \frac{13884}{61} \approx 227.$$

i	N_i	n_i	$N_i - n_i$	M_i
1	52	0	52	0
2	44	14	30	52
3	38	12	26	52 + 30 = 82
4	40	18	22	82 + 26 = 108
5	32	17	15	108 + 22 = 130

Tabela I.6. Przykład I.71

Stosując wzór Lincolna–Petersena do pierwszej i drugiej próby, otrzymujemy

$$\hat{N} = \frac{N_1 N_2}{n_2} = \frac{52 \cdot 44}{14} \approx 163,$$

a dla ostatniej próby

$$\hat{N} = \frac{M_5 N_5}{n_5} = \frac{130 \cdot 32}{17} \approx 245.$$

Wynik otrzymany za pomocą metody Schnabel jest bliższy ostatniego wyniku otrzymanego metodą Lincolna–Petersena niż wtedy, gdy uwzględnialiśmy tylko dwie pierwsze próby, co można łatwo wytłumaczyć tym, że przed ostatnim odłowem liczba oznakowanych osobników była dostatecznie duża (ponad połowa populacji). Oszacowane odchylenie standardowe zmiennej $1/N$ wynosi

$$\sigma = \sqrt{\frac{\sum_{i=1}^{m} n_i}{(\sum_{i=1}^{m} M_i N_i)^2}} = \frac{\sqrt{61}}{13884} \approx 0{,}00056.$$

Błąd wyznaczenia $\hat{N}$ możemy oszacować, korzystając z nierówności Czebyszewa

$$\mathrm{P}\left(\left| \frac{1}{N} - \frac{1}{\hat{N}} \right| \geq c\sigma \right) \leq \frac{1}{c^2}.$$

Przyjmując np. $c = \sqrt{10}$, otrzymujemy

$$\mathrm{P}\left(\left| \frac{1}{N} - \frac{1}{\hat{N}} \right| \geq 0{,}00177 \right) \leq 0{,}1.$$

Ponieważ $1/\hat{N} \approx 0,00441$, więc co najmniej z prawdopodobieństwem 0,9 mamy

$$0{,}00441 - 0{,}00177 \leq \frac{1}{N} \leq 0{,}00441 + 0{,}00177,$$

a stąd stwierdzamy, że $162 \leq N \leq 379$ z prawdopodobieństwem $> 0{,}9$. Widzimy, że oszacowanie błędu uzyskane za pomocą nierówności Czebyszewa nie jest zbyt dokładne. Znacznie mniejszy błąd uzyskamy, przyjmując, że $1/N$ dobrze przybliża się w otoczeniu wartości $1/\hat{N}$ rozkładem normalnym. Wtedy już dla $c = 1{,}7$ mamy $\mathrm{P}(|1/N - 1/\hat{N})| \geq c\sigma) < 0{,}1$ i $187 \leq N \leq 289$.

Przedstawione metody szacowania wielkości populacji stanowią zaledwie wstęp do znacznie bardziej zaawansowanej teorii wyznaczania różnych parametrów populacyjnych na podstawie metody wielokrotnych złowień. Założenia, że populacja jest zamknięta i kolejne odłowy są od siebie niezależne, są bardzo ograniczające, dlatego stosuje się inne metody, niewymagające tych założeń. Ponadto za ich pomocą można badać takie parametry, jak współczynniki śmiertelności lub urodzeń. Metody te stosuje się nie tylko w ekologii, ale również

w innych działach biologii i w medycynie. Oprócz literatury wspomnianej na początku rozważań, istnieje dobrze rozbudowane oprogramowanie uwzględniające różne założenia ekologiczne, np. program *MARK*.

7. Entropia i informacja

7.1. Intuicje matematyczne i podstawowe definicje

Pojęcie entropii wywodzi się z termodynamiki; zamiast jednak odwoływać się do pojęć fizycznych, przedstawimy proste intuicje związane z teorią informacji, prowadzące do tzw. entropii Shannona.

Niech ω będzie pewnym nieznanym nam elementem przestrzeni probabilistycznej $(\Omega, \Sigma, \mathrm{P})$. Ile informacji uzyskaliśmy o ω, jeżeli dowiedzieliśmy się, że ω należy do pewnego zbioru $A \in \Sigma$? Jeżeli podzielimy przestrzeń na dwa zbiory A i $\Omega \setminus A$, przy czym $\mathrm{P}(A) = 1/2$, to wiadomość, że $\omega \in A$, przynosi jednostkową informację o ω (1 bit), bo zdarzenia A i $\Omega \setminus A$ są tak samo prawdopodobne. Do zakodowania tej wiadomości w systemie binarnym, a więc za pomocą cyfr $0, 1$, wystarczy jeden znak. Jeżeli $\mathrm{P}(A) = 1/8$, to wiadomość, że $\omega \in A$, daje trzy bity informacji, bo skoro przestrzeń jest podzielona na $8 = 2^3$ zdarzeń jednakowo prawdopodobnych, to w systemie binarnym możemy je zakodować za pomocą ciągów trójelementowych cyfr $0, 1$.

Ogólnie przyjmujemy, że jeżeli $\mathrm{P}(A) > 0$, to *ilość informacji* otrzymanej przy zajściu zdarzenia A wynosi

$$I(A) = -\log_2 \mathrm{P}(A).$$

(Jednostkę informacji (bit) można zamienić na inną, zastępując $\log_2$ w definicji ilości informacji logarytmem o innej podstawie, np. logarytmem naturalnym). Tak zdefiniowana ilość informacji ma następującą własność addytywności:

$$I(A \cap B) = I(A) + I_A(B), \tag{7.1}$$

gdzie $I_A(B)$ jest ilością informacji, którą uzyskaliśmy, traktując B jako zdarzenie w przestrzeni probabilistycznej $(A, \Sigma_A, \mathrm{P}_A)$. Istotnie, $\mathrm{P}_A(B) = \mathrm{P}(B|A)$, więc

$$\begin{aligned} I(A) + I_A(B) &= -\log_2 \mathrm{P}(A) - \log_2 \mathrm{P}(B|A) = -\log_2 \mathrm{P}(A) - \log_2 \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} \\ &= -\log_2 \mathrm{P}(A \cap B) = I(A \cap B). \end{aligned}$$

Jeżeli w szczególności A i B są zdarzeniami niezależnymi, to $\mathrm{P}_A(B) = \mathrm{P}(B|A) = \mathrm{P}(B)$, a więc $I_A(B) = I(B)$, skąd

$$I(A \cap B) = I(A) + I(B). \tag{7.2}$$

Podzielmy przestrzeń na n zdarzeń rozłącznych $A_1, \dots, A_n$. Niech $p_i = \mathrm{P}(A_i)$ dla $i = 1, \dots, n$. Jeżeli w rezultacie jednego doświadczenia otrzymujemy informację, że ω należy do jednego z tych zdarzeń, to średnia ilość informacji wynosi

$$H(\mathbf{p}) = \mathrm{E}\, I = -\sum_{i=1}^{n} p_i \log_2 p_i, \qquad \mathbf{p} = (p_1, \dots, p_n). \tag{7.3}$$

Wyrażenie $H(\mathbf{p})$ nosi nazwę *entropii doświadczenia*. Przy $p_i = 0$ przyjmujemy, że $p_i \log_2 p_i = 0$. Entropię można też utożsamiać z miarą niepewności lub nieprzewidywalności. Im większa entropia, tym większa niepewność wyniku doświadczenia. Wykażemy, że największa entropia jest wtedy, gdy $p_1 = \dots = p_n = 1/n$ (patrz wzór (7.31)), a więc gdy wszystkie wyniki doświadczenia są jednakowo prawdopodobne; jeżeli zaś entropia wynosi zero, to wynik doświadczenia jest w pełni przewidywalny.

Pojęcie entropii związane jest też z oszczędnym kodowaniem. Mamy przekazać wiadomość zapisaną za pomocą n liter $A_1, \dots, A_n$. Litery traktujemy jako rozłączne zdarzenia, których suma jest całą przestrzenią Ω. Litery kodujemy binarnie. Chcemy to zrobić tak, aby użyć jak najmniej znaków. Jeżeli np. są cztery litery i kodujemy je kolejno 00, 01, 10, 11, to do zakodowania jednej litery używamy średnio dwóch znaków. Jeżeli prawdopodobieństwa występowania liter wynoszą 1/2, 1/4, 1/8 i 1/8, to oszczędniejsze kodowanie jest postaci $0 := A_1$, $10 := A_2$, $110 := A_3$ i $111 := A_4$. Wtedy do zakodowania jednej litery potrzeba średnio

$$\tfrac{1}{2} \cdot 1 + \tfrac{1}{4} \cdot 2 + \tfrac{1}{8} \cdot 3 + \tfrac{1}{8} \cdot 3 = H\big(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}\big) = \tfrac{7}{4}$$

znaków. Przy długim tekście zawierającym N liter i oszczędnym kodowaniu, tekst zapiszemy za pomocą ciągu binarnego zawierającego około $\frac{7}{4}N$ znaków, podczas gdy przy poprzednim kodowaniu mamy $2N$ znaków. W praktyce problem jest jednak bardziej skomplikowany.

W definicji entropii można zamiast $\log_2$ używać logarytmów o innych podstawach > 1, np. logarytmu naturalnego. Tak zmodyfikowana entropia będzie się różnić od wyjściowej stałą dodatnią przed znakiem sumy. W dalszych definicjach podstawa logarytmu nie odgrywa roli, piszemy więc log, przyjmując, że podstawa logarytmu jest ustalona.

Entropia ma następującą własność addytywności, analogiczną do (7.1). Niech $\mathbf{p} = (p_1, \dots, p_i)$ i $\mathbf{q} = (q_1, \dots, q_j)$ będą takimi ciągami liczb nieujemnych, że $p+q = 1$, $p = p_1 + \dots + p_i$, $q = q_1 + \dots + q_j$ i $p, q > 0$. Jeżeli $\mathbf{p}^* = (\frac{p_1}{p}, \dots, \frac{p_i}{p})$ i $\mathbf{q}^* = (\frac{q_1}{q}, \dots, \frac{q_j}{q})$, to

$$H(\mathbf{p}, \mathbf{q}) = H(p, q) + pH(\mathbf{p}^*) + qH(\mathbf{q}^*). \tag{7.4}$$

Wzór (7.4) mówi nam o tym, że średnia informacja uzyskana poprzez rozbicia na coraz mniejsze zbiory sumuje się. Istotnie, $H(p,q)$ jest średnią informacją uzyskaną z rozbicia na dwa zbiory o miarach p i q; następnie wybieramy losowo jeden z tych zbiorów i obliczamy informację uzyskaną z jego rozbicia. Wzór (7.4) łatwo uogólnia się na przypadek, gdy pierwsze rozbicie jest na więcej niż dwa zbiory.

Dla dalszych celów wygodnie jest zdefiniować pojęcie entropii za pomocą zmiennej losowej. Niech ξ będzie zmienną losową przyjmującą skończoną liczbę wartości $x_1,\dots,x_n$. Wtedy *entropią zmiennej losowej* ξ nazywamy liczbę

$$H(\xi) = -\sum_{i=1}^{n} \mathrm{P}(\xi = x_i)\log \mathrm{P}(\xi = x_i). \tag{7.5}$$

Definicja (7.5) pokrywa się z (7.3), gdy $A_i = \{\xi = x_i\}$ dla $i = 1,\dots,n$.

Jeżeli mamy dwie zmienne losowe ξ i η przyjmujące skończoną liczbę wartości, to *entropia łączna* (albo *produktowa*) zmiennych ξ i η określona jest wzorem

$$H(\xi,\eta) = -\sum_{i,j} \mathrm{P}(\xi = x_i, \eta = y_j)\log \mathrm{P}(\xi = x_i, \eta = y_j), \tag{7.6}$$

gdzie x_i, y_j są wartościami przyjmowanymi przez ξ i η, a sumowanie jest po wszystkich wskaźnikach. Gdy zmienne losowe ξ i η są niezależne, łatwo z definicji entropii łącznej wnioskujemy, że

$$H(\xi,\eta) = H(\xi) + H(\eta). \tag{7.7}$$

Znacznie trudniejszy jest dowód nierówności

$$H(\xi,\eta) \le H(\xi) + H(\eta) \tag{7.8}$$

dla dowolnych zmiennych losowych o skończonych zbiorach wartości (patrz zad. I.63).

Uwaga I.72. Entropię łączną można bezpośrednio zdefiniować dla rozbić przestrzeni probabilistycznej na zbiory. Niech $A_1,\dots,A_m$ i $B_1,\dots,B_n$ będą dwoma rozbiciami przestrzeni Ω na zbiory mierzalne i rozłączne. Wtedy rodzina zbiorów postaci $C_{ij} = A_i \cap B_j$, gdzie $i = 1,\dots,m$ oraz $j = 1,\dots,n$, jest też rozbiciem przestrzeni Ω. Takie rozbicie będziemy nazywać *produktowym*. Jeżeli ξ jest zmienną losową odpowiadającą rozbiciu $A_1,\dots,A_m$, tj. $A_i = \{\xi = x_i\}$ dla $i = 1,\dots,m$, a η jest zmienną losową odpowiadającą rozbiciu $B_1,\dots,B_n$, to entropia rozbicia produktowego C_{ij}, gdzie $i = 1,\dots,m$ oraz $j = 1,\dots,n$, jest równa $H(\xi,\eta)$.

Dalej wprowadzimy pojęcia entropii warunkowej i informacji wspólnej. Pojęcia te można też formalnie zdefiniować dla rozbić przestrzeni probabilistycznej.

Entropię η pod warunkiem ξ określamy wzorem

$$H(\eta|\xi) = -\sum_i \mathrm{P}(\xi = x_i) \sum_j \mathrm{P}(\eta = y_j \mid \xi = x_i) \log \mathrm{P}(\eta = y_j \mid \xi = x_i). \tag{7.9}$$

Entropia warunkowa spełnia równanie

$$H(\eta|\xi) = H(\xi, \eta) - H(\xi), \tag{7.10}$$

skąd i z nierówności (7.8) otrzymujemy

$$H(\eta|\xi) \le H(\eta); \tag{7.11}$$

także na odwrót, z (7.11) wynika (7.8). Zgodnie ze wzorem (7.10) entropia η pod warunkiem ξ mówi nam, ile dodatkowo informacji dostarczy zmienna η, jeżeli już znamy informację dostarczoną przez ξ.

Wprowadzamy również *informację wspólną* (lub *wzajemną*) *zmiennych losowych* ξ i η, określoną wzorem

$$I(\xi, \eta) = \sum_{i,j} \mathrm{P}(\xi = x_i, \eta = y_j) \log \frac{\mathrm{P}(\xi = x_i, \eta = y_j)}{\mathrm{P}(\xi = x_i)\,\mathrm{P}(\eta = y_j)}. \tag{7.12}$$

Natychmiast z definicji informacji wspólnej otrzymujemy wzór

$$I(\xi, \eta) = H(\eta) + H(\xi) - H(\xi, \eta), \tag{7.13}$$

który wyjaśnia nazwę tego pojęcia.

Gdy zmienne losowe ξ, η przyjmują wartości w tym samym zbiorze, można zdefiniować *entropię względną* rozkładu ξ względem rozkładu η (zwaną również *odległością Kullbacka–Leiblera* lub *dywergencją Kullbacka–Leiblera*):

$$H_{KL}(\eta|\xi) = \sum_i \mathrm{P}(\eta = x_i) \log \frac{\mathrm{P}(\eta = x_i)}{\mathrm{P}(\xi = x_i)}. \tag{7.14}$$

Zauważmy, że w definicji entropii względnej nie występuje wspólny rozkład zmiennych losowych ξ i η, a jedynie rozkłady każdej z tych zmiennych. Entropia względna mierzy odległość między rozkładami zmiennych losowych ξ i η, w odróżnieniu od entropii warunkowej, związanej z niezależnością i korelacją zmiennych losowych (patrz zad. I.64). Zauważmy, że

$$H_{KL}(\eta|\xi) = -\sum_i \mathrm{P}(\eta = x_i) \log \mathrm{P}(\xi = x_i) - H(\eta). \tag{7.15}$$

Wyrażenie $H_c(\eta,\xi) = -\sum_i \mathrm{P}(\eta = x_i)\log \mathrm{P}(\xi = x_i)$ nosi nazwę *entropii krzyżowej* η względem ξ i opisuje średnią liczbę bitów potrzebną do zakodowania zmiennej η przy użyciu kodowania odpowiadającego zmiennej ξ.

7.2. Przypadek zmiennych losowych mających gęstości

Pojęcia wprowadzone w punkcie 7.1, dotyczące zmiennych losowych dyskretnych, można łatwo przenieść na przypadek, gdy zmienne losowe mają gęstości.

Niech ξ i η będą zmiennymi losowymi lub ogólniej elementami losowymi o wartościach w przestrzeniach X i Y z miarami σ-skończonymi m_X i m_Y o gęstościach $f_\xi(x)$, $f_\eta(y)$; załóżmy też, że ich wspólny rozkład ma gęstość $f_{\xi\eta}(x,y)$. Wtedy *entropią zmiennej losowej* ξ nazywamy liczbę

$$H(\xi) = -\int_X f_\xi(x)\log f_\xi(x)\, m_X(dx). \tag{7.16}$$

Entropia łączna (albo *produktowa*) zmiennych ξ i η określona jest wzorem

$$H(\xi,\eta) = -\int_Y\int_X f_{\xi\eta}(x,y)\log f_{\xi\eta}(x,y)\, m_X(dx)\, m_Y(dy). \tag{7.17}$$

Entropię zmiennej η pod warunkiem zmiennej ξ określamy wzorem

$$\begin{aligned} H(\eta|\xi) &= -\int_X f_\xi(x)\int_Y f(y|x)\log f(y|x)\, m_Y(dy)\, m_X(dx) \\ &= -\int_Y\int_X f_{\xi\eta}(x,y)\log\frac{f_{\xi\eta}(x,y)}{f_\xi(x)}\, m_X(dx)\, m_Y(dy), \end{aligned} \tag{7.18}$$

gdzie $f(y|x) = \frac{f_{\xi\eta}(x,y)}{f_\xi(x)}$. *Informację wspólną* (lub *wzajemną*) *zmiennych losowych* ξ i η określamy wzorem

$$I(\xi,\eta) = \int_Y\int_X f_{\xi\eta}(x,y)\log\frac{f_{\xi\eta}(x,y)}{f_\xi(x)f_\eta(y)}\, m_X(dx)\, m_Y(dy). \tag{7.19}$$

Jeżeli ξ i η przyjmują wartości w tej samej przestrzeni X, to *entropię względną* rozkładu η względem rozkładu ξ definiujemy jako

$$H_{KL}(\eta|\xi) = \int_X f_\eta(x)\log\frac{f_\eta(x)}{f_\xi(x)}\, m_X(dx). \tag{7.20}$$

Możemy również określić *entropię krzyżową* zmiennej η względem ξ wzorem

$$H_c(\eta,\xi) = -\int_X f_\eta(x)\log f_\xi(x)\, m_X(dx).$$

Entropia, entropia względna i entropia krzyżowa powiązane są równością

$$H_{KL}(\eta|\xi) = H_c(\eta,\xi) - H(\eta). \tag{7.21}$$

Zdefiniowanie entropii, entropii łącznej i warunkowej, informacji wzajemnej oraz entropii względnej dla elementów losowych ma szereg zalet. Po pierwsze, definicja entropii łącznej dla ξ i η pokrywa się z definicją entropii wektora losowego (ξ,η). Po drugie, definicje te obejmują też dyskretne wersje entropii i informacji, ponieważ jako przestrzenie X i Y można przyjąć zbiory wartości tych zmiennych losowych, a jako miary m_X i m_Y przyjąć miary liczące (tzn. $m_X(A)$ jest liczbą elementów zbioru A); miary te są dyskretne, więc zmienne losowe ξ i η mają gęstości. Takie podejście pozwala na ujednolicenie dowodów. Po trzecie, widzimy, że wprowadzone definicje dotyczą gęstości rozkładów zmiennych losowych ξ i η albo gęstości ich łącznego rozkładu, można więc wprowadzić te definicje w języku samych gęstości. W szczególności, jeżeli f jest gęstością rozkładu elementu losowego ξ, to $H(f) := H(\xi)$ nazywamy *entropią gęstości* f.

Niech X będzie przestrzenią z miarą m. Funkcję mierzalną $f\colon X \to [0,\infty)$ nazywamy *gęstością*, jeżeli $\int_X f(x)\,m(dx) = 1$. Entropię gęstości f możemy zdefiniować wzorem $H(f) = -\int_X f(x)\log f(x)\,m(dx)$. Jeśli zaś f i g są gęstościami, to

$$H_{KL}(f|g) = \int_X f(x)\log\frac{f(x)}{g(x)}\,m(dx) \tag{7.22}$$

nazywamy *entropią* gęstości f *względem* g. Podobnie *entropię krzyżową gęstości* f *względem* g określamy wzorem

$$H_c(f,g) = -\int_X f(x)\log g(x)\,m(dx).$$

Własności (7.7), (7.8), (7.10) i (7.13) entropii łącznej i warunkowej oraz informacji wspólnej przenoszą się na przypadek gęstości; dowody z wyjątkiem dowodu nierówności (7.8) są natychmiastowe. Ponieważ warunki (7.8) i (7.11) są równoważne, udowodnimy (7.11). Skorzystamy z nierówności Jensena (5.8):

$$F\Big(\int_X h(x)\,\mu(dx)\Big) \le \int_X F(h(x))\,\mu(dx),$$

gdzie F jest funkcją wypukłą, μ miarą probabilistyczną, a $h(x)$ funkcją całkowalną. W tym wypadku ustalamy y i przyjmujemy, że $F(x) = -\log x$, $\mu(dx) = \frac{f_{\xi\eta}(x,y)}{f_\eta(y)}\,m_X(dx)$ i $h(x) = \frac{f_\xi(x)}{f_{\xi\eta}(x,y)}$. Z nierówności Jensena otrzymujemy

$$-\log\Big(\int_X \frac{f_\xi(x)}{f_{\xi\eta}(x,y)}\,\frac{f_{\xi\eta}(x,y)}{f_\eta(y)}\,m_X(dx)\Big) \le -\int_X \log\Big(\frac{f_\xi(x)}{f_{\xi\eta}(x,y)}\Big)\frac{f_{\xi\eta}(x,y)}{f_\eta(y)}\,m_X(dx),$$

więc

$$\log\Big(\int_X \frac{f_\xi(x)}{f_\eta(y)}\, m_X(dx)\Big) \geq -\int_X \frac{f_{\xi\eta}(x,y)}{f_\eta(y)} \log \frac{f_{\xi\eta}(x,y)}{f_\xi(x)}\, m_X(dx).$$

Po pomnożeniu obu stron nierówności przez $f_\eta(y)$ i scałkowaniu względem y otrzymujemy

$$-\int_Y f_\eta(y)\log f_\eta(y)\, m_Y(dy) \geq -\int_Y \int_X f_{\xi\eta}(x,y)\log\frac{f_{\xi\eta}(x,y)}{f_\xi(x)}\, m_X(dx)\, m_Y(dy),$$

a zatem

$$H(\eta) \geq H(\eta|\xi).$$

7.3. Entropia uogólniona

Korzystając z pracy [247], wprowadzimy teraz ogólniejszą definicję entropii względnej, wygodną w wielu zastosowaniach, i przedstawimy jej własności. Zauważmy, że wzór (7.22) można zapisać w postaci

$$H_{KL}(f|g) = \int_X g(x)\eta\Big(\frac{f(x)}{g(x)}\Big)\, m(dx), \quad \text{gdzie} \quad \eta(x) = x\log x.$$

Ponieważ $\eta(x) = x\log x$ jest funkcją wypukłą, naturalne uogólnienie entropii względnej H_{KL} uzyskamy, podstawiając w miejsce tej funkcji inną funkcję wypukłą. Niech $\eta\colon [0,\infty) \to \mathbb{R}$ będzie dowolną funkcją ciągłą i wypukłą. Dla dowolnych gęstości f i g definiujemy *η-entropię gęstości f względem g* wzorem

$$H_\eta(f|g) = \int_X g(x)\eta\Big(\frac{f(x)}{g(x)}\Big)\, m(dx). \tag{7.23}$$

Ponieważ f i g mogą przyjmować wartości 0, wzór (7.23) wymaga doprecyzowania. Wprowadzamy funkcję pomocniczą $\varphi\colon [0,\infty)\times[0,\infty) \to \mathbb{R}\cup\{\infty\}$, przyjmując

$$\varphi(u,v) = \begin{cases} v\eta(u/v), & v>0,\ u\geq 0,\\ 0, & v=0,\ u=0,\\ u\eta'(\infty), & v=0,\ u>0, \end{cases}$$

gdzie

$$\eta'(\infty) = \lim_{v\to\infty} \eta'(v) = \lim_{v\to\infty} \eta(v)/v.$$

Przyjmujemy, że

$$H_\eta(f|g) = \int_X \varphi(f(x), g(x))\, m(dx) \tag{7.24}$$

dla dowolnych gęstości f i g.

Dobierając różne funkcje η, można uzyskać inne η-entropie niż odległość Kullbacka–Leiblera. Podamy dwa najbardziej znane przykłady.

(i) Jeżeli $\eta_1(u) = |1-u|$, to $H_{\eta_1}(f|g) = \|f-g\|$, gdzie $\|\cdot\|$ jest normą w L^1.

(ii) Niech $\alpha \in (0,1)$ i $\eta_\alpha(u) = -u^\alpha$. Wtedy

$$H_{\eta_\alpha}(f|g) = -\int_X f^\alpha(x) g^{1-\alpha}(x)\, m(dx).$$

Przedstawimy teraz kilka własności η-entropii. Dla dowolnych gęstości f i g spełniona jest nierówność

$$H_\eta(f|g) \geq \eta(1). \tag{7.25}$$

Aby to udowodnić, skorzystamy z nierówności Jensena (5.8), przyjmując $F = \eta$, $\mu(dx) = g(x)m(dx)$ i $h(x) = f(x)/g(x)$. Wtedy

$$H_\eta(f|g) = \int_X \eta\Big(\frac{f}{g}\Big) g\, dm \geq \eta\Big(\int_X \frac{f}{g} g\, dm\Big) = \eta\left(\int_X f\, dm\right) = \eta(1),$$

co dowodzi (7.25).

Entropii warunkowej można używać do badania zbieżności w L^1.

Twierdzenie I.73. *Załóżmy, że $\eta\colon [0,\infty) \to \mathbb{R}$ jest funkcją ciągłą i wypukłą oraz $\eta''(1) > 0$. Niech (f_n) i (g_n) będą ciągami gęstości. Wtedy*

$$\lim_{n\to\infty} H_\eta(f_n|g_n) = \eta(1) \implies \lim_{n\to\infty} \|f_n - g_n\| = 0,$$

gdzie $\|\cdot\|$ jest normą w przestrzeni L^1.

Stosunkowo prosty dowód twierdzenia I.73 podany jest w pracy [247]. Założenie $\eta''(1) > 0$ można zastąpić słabszym założeniem ścisłej wypukłości funkcji $\eta(x)$ w $x = 1$. Jeżeli $\eta''(1) > 0$, to z twierdzenia I.73 i z (7.25) wynika nierówność $H_\eta(f|g) > \eta(1)$ dla $f \neq g$.

Kolejna ważna własność η-entropii wiąże się z pojęciem operatora Markowa. Odwzorowanie liniowe $P\colon L^1 \to L^1$ przeprowadzające gęstości na gęstości nazywamy *operatorem Markowa* (lub *operatorem stochastycznym*).

Twierdzenie I.74. *Jeżeli P jest operatorem Markowa, to*

$$H_\eta(f|g) \geq H_\eta(Pf|Pg) \tag{7.26}$$

dla dowolnych gęstości f i g.

Dowód twierdzenia I.74 podany w pracy [247] przebiega następująco. Ponieważ η jest funkcją wypukłą, istnieją takie ciągi (a_n) i (b_n) liczb rzeczywistych, że

$$\eta(u) = \sup\{a_n + b_n u\colon\ n \in \mathbb{N}\}. \tag{7.27}$$

Wynika stąd natychmiast, że

$$\varphi(u,v) = \sup\{a_n v + b_n u\colon\ n \in \mathbb{N}\}. \tag{7.28}$$

Ponieważ operator Markowa jest monotoniczny, z (7.28) wynika, że

$$P\varphi(f,g) \geq P(a_n g + b_n f) = a_n Pg + b_n Pf.$$

Stosując jeszcze raz (7.28) dla $u = Pf$ i $v = Pg$, otrzymujemy

$$P\varphi(f,g) \geq \varphi(Pf, Pg),$$

a ponieważ operator Markowa zachowuje całkę, więc

$$\begin{aligned} H_\eta(f|g) &= \int_X \varphi(f,g)\,dm = \int_X P\varphi(f,g)\,dm \geq \int_X \varphi(Pf,Pg)\,dm \\ &\geq H_\eta(Pf|Pg). \end{aligned}$$

Entropię $H(f)$ można również uogólnić, zastępując funkcję $x\log x$ inną funkcją wypukłą. Niech $\eta\colon X \to \mathbb{R}$ będzie funkcją ciągłą i wypukłą. Wtedy wyrażenie

$$H_\eta(f) = -\int_X \eta(f(x))\,m(dx) \tag{7.29}$$

nazywamy *η-entropią gęstości f*. Zauważmy, że jeżeli $m(X) = 1$, to $g \equiv 1$ jest gęstością oraz $H_\eta(f) = -H_\eta(f|1)$ (patrz (7.23)). Tożsamość tę można uogólnić: jeśli f i g są gęstościami względem miary m, to f/g jest gęstością względem miary μ określonej wzorem $\mu(dx) = g(x)\,m(dx)$ oraz

$$H_{\eta,\mu}\left(\frac{f}{g}\right) = -H_{\eta,m}(f|g),$$

gdzie miary μ i m w oznaczeniach entropii i entropii względnej podkreślają, z jakimi miarami związane są te entropie.

Jeżeli $m(X) < \infty$, to spełniona jest nierówność

$$H_\eta(f) \leq -m(X)\eta\left(\frac{1}{m(X)}\right) \tag{7.30}$$

dla dowolnej gęstości f. Aby to wykazać, skorzystamy ponownie z nierówności Jensena (5.8), przyjmując tym razem $F = m(X)\eta$, $\mu = \frac{m}{m(X)}$ i $h = f$. Wtedy

$$m(X)\eta\Big(\int_X \frac{f(x)\,m(dx)}{m(X)}\Big) \le \int_X \eta(f(x))\,m(dx),$$

więc

$$m(X)\eta\Big(\frac{1}{m(X)}\Big) \le -H_\eta(f),$$

co dowodzi (7.30). W (7.30) zachodzi równość, gdy $f \equiv 1/m(X)$. Jeżeli $\eta(x) = x\log x$, to

$$H(f) \le -m(X)\cdot\frac{1}{m(X)}\log\Big(\frac{1}{m(X)}\Big) = \log m(X).$$

Ponieważ entropia doświadczenia (7.3) jest entropią gęstości $(p_1,\dots,p_n)$ określonej na przestrzeni n-elementowej z *miarą liczącą* m (tzn. $m(A)$ jest liczbą elementów zbioru A), więc

$$H(p_1,\dots,p_n) \le \log_2 m(X) = \log_2 n, \tag{7.31}$$

a maksimum entropii osiągane jest dla $\mathbf{p} = (\frac{1}{n},\dots,\frac{1}{n})$.

Jeżeli P jest operatorem Markowa, możemy go rozszerzyć na dowolne funkcje mierzalne nieujemne, przyjmując $Pf(x) = \lim_{n\to\infty} Pf_n(x)$, gdzie (f_n) jest dowolnym ciągiem rosnącym funkcji całkowalnych zbieżnym prawie wszędzie do f (definicja $Pf(x)$ nie zależy od wyboru ciągu (f_n) dla p.w. x). Tak więc ma sens wyrażenie $P\mathbf{1}_X$. Operator Markowa P nazywamy *podwójnie stochastycznym*, jeżeli $P\mathbf{1}_X = \mathbf{1}_X$.

Twierdzenie I.75. *Jeżeli P jest operatorem podwójnie stochastycznym, to*

$$H_\eta(f) \le H_\eta(Pf) \tag{7.32}$$

dla dowolnych gęstości f.

Dowód twierdzenia I.75 przebiega analogicznie do dowodu twierdzenia I.74, przy czym zamiast wzoru (7.28) stosujemy bezpośrednio wzór (7.27), otrzymując $P\eta(f) \ge \eta(Pf)$. Ponieważ operator Markowa zachowuje całkę, więc

$$-H_\eta(f) = \int_X \eta(f)\,dm \ge \int_X P\eta(f)\,dm \ge \int_X \eta(Pf)\,dm \ge -H_\eta(Pf)$$

i otrzymujemy nierówność (7.32).

Uwaga I.76. Twierdzenie I.75 nie zachodzi, jeżeli opuścimy założenie $P\mathbf{1}_X = \mathbf{1}_X$. Niech P będzie operatorem Markowa na przestrzeni $L^1[0,\infty)$ z miarą

Lebesgue'a, określonym wzorem $Pf(x) = 2f(2x)$. Jeżeli $f = \mathbf{1}_{[0,1]}$, to $Pf = 2 \cdot \mathbf{1}_{[0,1/2]}$ oraz

$$H(f) = -\int_0^\infty \mathbf{1}_{[0,1]}(x) \ln \mathbf{1}_{[0,1]}(x)\, dx = 0,$$
$$H(Pf) = -\int_0^\infty 2\cdot\mathbf{1}_{[0,1/2]}(x) \ln(2\cdot\mathbf{1}_{[0,1/2]}(x))\, dx = -\int_0^{1/2} 2\ln 2\, dx = -\ln 2,$$

a więc $H(Pf) < H(f)$.

Przykład I.77. Niech $X = \{1,\dots,n\}$, $\Sigma = 2^X$, a m niech będzie miarą liczącą. Przestrzeń $L^1(X,\Sigma,m)$ utożsamiamy z $\mathbb{R}^n$, a gęstości w tej przestrzeni to wektory $\mathbf{p}$ o wyrazach nieujemnych, spełniające warunek $p_1 + \cdots + p_n = 1$. W tym wypadku operator podwójnie stochastyczny jest odwzorowaniem liniowym $P\colon \mathbb{R}^n \to \mathbb{R}^n$ o macierzy A o wyrazach nieujemnych, w której suma wyrazów w każdym wierszu i w każdej kolumnie jest równa 1. Dla takiej macierzy A, gęstości $\mathbf{p}$ i funkcji wypukłej η spełniona jest zatem nierówność

$$H_\eta(p) \le H_\eta(Ap) \le -n\,\eta(\tfrac{1}{n}).$$

7.4. Informacja i entropia w genetyce

Kiedy w latach sześćdziesiątych XX wieku poznano kod genetyczny, a w następnym dziesięcioleciu zaczęto sekwencjonować DNA, pojawiły się duże nadzieje, że matematyczna teoria informacji pozwoli odpowiedzieć na wiele pytań dotyczących informacji zmagazynowanej w kodzie genetycznym. Oczekiwania te spełniły się tylko częściowo. Opierając się na pracy [2, 3] oraz książce [200], przedstawimy kilka ogólnych wniosków wynikających z analizy kodu w cząsteczkach DNA i RNA przy użyciu entropii i teorii informacji.

Rozpoczniemy od przypomnienia podstawowych faktów z biologii. Nośniki informacji genetycznej DNA i RNA zbudowane są z nukleotydów, a te z kolei z cukrów i zasad azotowych. W wypadku DNA zasadami są adenina A, guanina G, cytozyna C i tymina T, podczas gdy w RNA zamiast tyminy występuje uracyl U. Z punktu widzenia teorii informacji cząsteczki te kodowane są za pomocą czterech liter A, G, C, T lub A, G, C, U. Mimo iż cząsteczka DNA jest dwuniciowa, znajomość kolejności położenia zasad na jednej nici w pełni ją opisuje, bo zasady na obu niciach połączone są według ustalonych wzorów. Możemy zatem patrzeć na cząsteczki DNA i RNA jak na odpowiednio długie wyrazy w alfabecie czteroliterowym. Trójki nukleotydów tworzą kodony, które odgrywają kluczową rolę w syntezie białek. Kodonów jest $4^3 = 64$, z czego 60 służy do budowy 20 aminokwasów. Z punktu widzenia syntezy białek informacja genetyczna może więc być zapisana za pomocą wyrazów w alfabecie dwudziestoliterowym.

Ograniczymy się do kodowania czteroliterowego i będziemy używać entropii z logarytmem o podstawie 4. Na ustalonym miejscu w nici DNA (lub RNA) może występować jedna z liter A, G, C, T. Niech p_A, p_G, p_C, p_T będą prawdopodobieństwami wystąpienia tych liter. Wtedy entropia pojedynczego miejsca, a więc zmiennej losowej przyjmującej wartości A, G, C, T, wynosi

$$H = -\sum_{i=A,G,C,T} p_i \log_4 p_i.$$

Rozważmy teraz ustaloną sekwencję kolejnych nukleotydów. Oznaczmy entropię tej sekwencji przez H_L, gdzie L jest liczbą nukleotydów w sekwencji. Niech $H(j)$ będzie entropią j-tego miejsca w sekwencji. Wtedy H_L jest entropią łączną i spełnia nierówność

$$H_L \leq \sum_{j=1}^{L} H(j).$$

Jeżeli korelacja nukleotydów jest mała, to

$$H_L \approx \sum_{j=1}^{L} H(j). \tag{7.33}$$

Korelacja może być jednak stosunkowo duża, szczególnie w populacjach nieustabilizowanych genetycznie, ponieważ mutacje zwykle zachodzą jednocześnie na dłuższych sekwencjach nukleotydów. Mimo to do obliczenia H_L używamy wzoru przybliżonego (7.33) z dość prostych powodów. Jeżeli np. sekwencja składa się z dziesięciu nukleotydów, to liczba różnych ciągów wynosi 4^{10}, musielibyśmy więc znać prawdopodobieństwa ponad miliona różnych wariantów ustalonej sekwencji, czyli zbadać tę sekwencję u co najmniej kilku milionów osobników tego samego gatunku.

Pojawia się pytanie, jaką informację można uzyskać, jeżeli wyznaczymy wartość H_L. Maksymalna wartość entropii H_L równa się L i jest osiągana, gdy na każdym miejscu w sekwencji występuje dowolny z symboli A, G, C, T z prawdopodobieństwem $1/4$; minimalna wartość H_L równa się 0, gdy cała sekwencja jest wyznaczona jednoznacznie. Załóżmy, że populacja jest ustabilizowana, a więc upłynęło dostatecznie dużo czasu od ostatniej zmiany ewolucyjnej. Wtedy można przyjąć, że sekwencje nukleotydów o niskiej entropii odpowiadają miejscom, w których zakodowana jest istotna informacja genetyczna, ponieważ mutacje na tych miejscach powodowałyby szkodliwy wpływ na przystosowanie osobnika. Miejsca, które nie kodują informacji genetycznej, mogą swobodnie mutować i poszczególne symbole będą się pojawiać losowo, ich entropia będzie więc duża. Można przyjąć, że informacja genetyczna zakodowana w ustalonej sekwencji jest różnicą między entropią maksymalną dla sekwencji ustalonej

długości a entropią rzeczywistą, czyli w przybliżeniu wynosi

$$I = L - H(L).$$

Podejście to nie dostarcza informacji o tym, co jest zakodowane, ale może wskazywać na miejsca, gdzie informacja genetyczna jest przechowywana.

Jeżeli rozważymy cały genom (lub np. cały kod dla cząsteczki RNA, białka), to wielkość

$$C(N) = N - H(N) = N - \sum_{i=1}^{k} H(L_i) \tag{7.34}$$

nazywamy *złożonością biologiczną* lub *ewolucyjną* genomu. We wzorze (7.34), N jest liczbą nukleotydów w genomie (lub innych biomolekułach), k liczbą segmentów, w których dokonuje się pomiaru, np. genów, a L_i ich długością. Możemy również zdefiniować *biologiczną* (lub *ewolucyjną*) *gęstość informacji* wzorem

$$D(N) = C(N)/N = 1 - H(N)/N = 1 - \sum_{i=1}^{k} H(L_i)/N. \tag{7.35}$$

Interesujące jest porównanie wielkości $C(N)$, $H(N)$ i $D(N)$ dla różnych organizmów. Wraz z ewolucyjnym rozwojem organizmów i towarzyszącym mu wzrostem genomów wielkości $C(N)$, $H(N)$ również rosną, co nie jest zaskakujące, ale $D(N)$ szybko maleje. Oznacza to, że genomy organizmów wysoko zorganizowanych w stosunku np. do prokariotów (np. bakterii) mają małą gęstość informacji. Mówiąc kolokwialnie, ewolucja nie dąży do optymalizacji kodu genetycznego, a wręcz zwiększa udział „śmieciowych" lokalizacji. Ten pozorny paradoks (związany z większym wysiłkiem energetycznym przy replikacji DNA) można wytłumaczyć następująco. Przy kolejnych etapach ewolucji zwiększał się genom, ponieważ organizm stawał się bardziej skomplikowany, ale niektóre lokalizacje, które wcześniej niosły informację genetyczną, przestawały być funkcjonalnie istotne, a więc stawały się podatne na mutacje. W ten sposób entropia rosła szybciej w porównaniu do wzrostu informacji. Również istnienie dużych sekwencji zdolnych do mutacji może ułatwiać przystosowanie do środowiska.

Zawartość informacji w różnych sekwencjach DNA lub RNA i złożoność biologiczna to nie jedyne zagadnienia genetyczne, które można badać, korzystając z teorii entropii i informacji. Do ważnych problemów należy korelacja poszczególnych sekwencji kodu genetycznego (lub kodów różnych organizmów); można ją badać, korzystając z entropii warunkowej i informacji wspólnej. Wysoki współczynnik korelacji może oznaczać powiązanie funkcjonalne wybranych sekwencji. Ciekawe jest również zagadnienie, w jaki sposób informacja związana z adaptacją do środowiska (np. lekoodpornością) jest umiejscowiona w genomie.

7.5. Zastosowania entropii względnej w dynamice populacyjnej

Entropia względna może służyć do opisu odległości między różnymi rozkładami oraz do badania zbieżności, dlatego jest często używana do badania asymptotyki w różnych modelach fizycznych i biologicznych. Praca [21] zawiera przegląd zastosowań klasycznej entropii względnej (tj. odległości Kullbacka–Leiblera) w systemach biologicznych, m.in. w dynamice replikatorów, badaniu łańcuchów Markowa i reakcji biochemicznych. Ponieważ wspomniane zastosowania oparte są na podobnych metodach, ograniczymy się tu do prostego modelu typu Lotki–Volterry.

Rozważmy populację złożoną z k podpopulacji. Przyjmijmy, że podpopulacje reprezentują różne gatunki, w szczególności osobniki rozmnażają się w ramach swojej podpopulacji, ale istnieje konkurencja między podpopulacjami. Wektor $\mathbf{x}(t) = [x_1(t), \dots, x_k(t)]$ opisuje stan populacji w chwili t, gdzie $x_i(t)$ jest rozmiarem (liczebnością) i-tej podpopulacji w chwili t. Załóżmy, że

$$x_i'(t) = f_i(\mathbf{x}(t))x_i(t), \quad i = 1, \dots, k, \tag{7.36}$$

gdzie $f_i(\mathbf{x}(t))$ jest współczynnikiem wzrostu i-tej podpopulacji, czyli różnicą między współczynnikiem urodzeń i śmierci. Niech

$$x(t) = x_1(t) + \cdots + x_k(t).$$

Wektor $\mathbf{p}(t) = [p_1(t), \dots, p_k(t)]$ z $p_i(t) = x_i(t)/x(t)$ przedstawia rozkład całej populacji na podpopulacje. Korzystając z równania (7.36), otrzymujemy

$$x'(t)p_i(t) + x(t)p_i'(t) = f_i(\mathbf{x}(t))x(t)p_i(t)$$

dla $i = 1, \dots, k$ oraz

$$x'(t) = \sum_{j=1}^{k} f_j(\mathbf{x}(t))x(t)p_j(t),$$

a stąd

$$p_i'(t) = f_i(\mathbf{x}(t))p_i(t) - \Big[\sum_{j=1}^{k} f_j(\mathbf{x}(t))p_j(t)\Big]p_i(t), \quad i = 1, \dots, k. \tag{7.37}$$

Wyrażenie w nawiasie kwadratowym jest średnim współczynnikiem wzrostu populacji i często zapisuje się je w postaci $\langle \mathbf{f}(\mathbf{x}(t)) \rangle$. Otrzymujemy w ten sposób *równanie replikatorowe*

$$p_i'(t) = (f_i(\mathbf{x}(t)) - \langle \mathbf{f}(\mathbf{x}(t)) \rangle)p_i(t), \quad i = 1, \dots, k. \tag{7.38}$$

Niech $\mathbf{q}$ będzie ustalonym rozkładem prawdopodobieństwa. Interesuje nas, jak zmienia się w czasie funkcja

$$H_{KL}(\mathbf{q}, \mathbf{p}(t)) = \sum_{i=1}^{k} q_i \ln\Big(\frac{q_i}{p_i(t)}\Big).$$

Wtedy

$$\frac{d}{dt} H_{KL}(\mathbf{q}, \mathbf{p}(t)) = -\sum_{i=1}^{k} q_i \frac{p_i'(t)}{p_i(t)} = -\sum_{i=1}^{k} (f_i(\mathbf{x}(t)) - \langle \mathbf{f}(\mathbf{x}(t))\rangle) q_i.$$

Ponieważ $q_1 + \cdots + q_k = 1$, więc

$$\frac{d}{dt} H_{KL}(\mathbf{q}, \mathbf{p}(t)) = \sum_{i=1}^{k} f_i(\mathbf{x}(t))(p_i - q_i) = (\mathbf{p}(t) - \mathbf{q}) \cdot \mathbf{f}(\mathbf{x}(t)). \tag{7.39}$$

Zatem entropia warunkowa nie rośnie, jeżeli spełniony jest warunek

$$(\mathbf{p}(t) - \mathbf{q}) \cdot \mathbf{f}(\mathbf{x}(t)) \le 0. \tag{7.40}$$

Jeżeli podamy warunek wystarczający na to, aby przynajmniej w pewnym otoczeniu wektora $\mathbf{q}$ spełniony był warunek (7.40), to odległość Kullbacka–Leiblera między stanami $\mathbf{p}(t)$ i $\mathbf{q}$ nie zwiększa się, a więc stan $\mathbf{q}$ jest stabilny.

Rozważmy przypadek, gdy $\mathbf{f}$ jest funkcją liniową, a więc $\mathbf{f}(\mathbf{x}) = A\mathbf{x}$, gdzie A jest pewną macierzą kwadratową o wymiarach $k \times k$. Wtedy warunek (7.40) można zapisać w postaci

$$(\mathbf{p}(t) - \mathbf{q}) \cdot A\mathbf{p}(t) \le 0. \tag{7.41}$$

Pokażemy, jak wyznaczyć $\mathbf{q}$ o współrzędnych dodatnich tak, aby nierówność (7.41) była spełniona w pewnym otoczeniu wektora $\mathbf{q}$. Wtedy

$$(\mathbf{p} - \mathbf{q}) \cdot A\mathbf{p} \le 0 \tag{7.42}$$

dla wszystkich wektorów $\mathbf{p}$ postaci $\mathbf{p} = \mathbf{q} + \varepsilon\mathbf{r}$, gdzie wektor $\mathbf{r}$ spełnia warunki $\|\mathbf{r}\| \le 1$ i $r_1 + \cdots + r_k = 0$, a ε jest dostatecznie małe. Poprzednią nierówność możemy zapisać w postaci

$$\varepsilon\mathbf{r} \cdot A(\mathbf{q} + \varepsilon\mathbf{r}) \le 0, \tag{7.43}$$

skąd wnioskujemy, że

$$\mathbf{r} \cdot A\mathbf{q} = 0 \quad \text{oraz} \quad \mathbf{r} \cdot A\mathbf{r} \le 0 \tag{7.44}$$

dla dowolnego wektora $\mathbf{r}$ spełniającego warunek $r_1 + \cdots + r_k = 0$. Podstawiając za $\mathbf{r}$ wektory o jednej współrzędnej 1, innej -1, a pozostałych 0, wnioskujemy,

że poszukiwany wektor $\mathbf{q}$, o ile istnieje, spełnia warunek $A\mathbf{q} = [c, c, \ldots, c]^T$ dla pewnej stałej c. Z (7.39) otrzymujemy również

$$\frac{d}{dt} H_{KL}(\mathbf{q}, \mathbf{q}(t)) \leq 0,$$

gdzie $\mathbf{q}(t)$ jest rozwiązaniem równania replikatorowego z $\mathbf{x}(0) = x(0)\mathbf{q}$. Stąd $H_{KL}(\mathbf{q}, \mathbf{q}(t)) \leq 0$, a ponieważ entropia względna nie może być ujemna, więc $\mathbf{q}(t) = \mathbf{q}$. Podsumujmy uzyskane wyniki. Jeżeli macierz A jest niedodatnio określona na przestrzeni $E = \{\mathbf{r}\colon r_1 + \cdots + r_k = 0\}$ oraz rozkład wektora $\mathbf{q}$ spełnia równanie $A\mathbf{q} = [c, c, \ldots, c]^T$ dla pewnej stałej c, to rozkład ten jest stacjonarny i stabilny.

Zadania

I.1 Udowodnić wzór (2.2).

I.2 Udowodnić wzór (2.3).

I.3 Korzystając z tabeli I.1, wyznaczyć rozkład zgonów według wieku w Polsce w roku 2015.
Wskazówka. Skorzystać ze wzoru Bayesa, a z tabeli I.1 odczytać współczynniki śmiertelności i wielkości populacji względem wieku.

I.4 Korzystając z tabel I.1 i I.3, wyznaczyć rozkład liczby urodzeń według wieku matki w roku 2015. Przyjąć, że kobiety w wieku rozrodczym stanowiły 49% populacji w każdym przedziale wieku.

I.5 Korzystając z modelu statycznego z przykładu I.5 bez uwzględniania migracji, wyznaczyć strukturę wiekową Polski w roku 2030.
Wskazówka. W tym i następnych zadaniach skorzystać z uwagi I.6 dotyczącej współczynnika śmiertelności w przedziale wieku 0–4.

I.6 Korzystając z modelu statycznego z przykładu I.5 bez uwzględniania migracji, wyznaczyć strukturę wiekową Polski w roku 2045.

I.7 Korzystając z modelu statycznego z przykładu I.5 z uwzględnieniem migracji na poziomie roku 2014, wyznaczyć strukturę wiekową Polski w roku 2030.

I.8 Korzystając z modelu dynamicznego z przykładu I.7 z uwzględnieniem migracji na poziomie roku 2014, wyznaczyć strukturę wiekową Polski w latach 2030 i 2045.

I.9 Korzystając z modelu dynamicznego z przykładu I.7 z uwzględnieniem migracji na poziomie roku 2014, wyznaczyć strukturę wiekową Polski w latach 2030 i 2045 przy założeniu, że współczynnik urodzeń w każdej grupie wiekowej będzie (a) o 10% większy, (b) o 20% większy niż w roku 2015.

I.10 W Polsce ryzyko zawału serca w ciągu najbliższych pięciu lat u mężczyzny w wieku 65–69 lat wynosi około 11%. Wysokie skurczowe ciśnienie tętnicze wśród mężczyzn z tego przedziału wieku występuje 2,5 razy częściej u osób, które przeszły zawał w ciągu pięciu lat, niż w pozostałej grupie. Wyznaczyć ryzyko zawału u mężczyzn z tego przedziału wieku z wysokim skurczowym ciśnieniem tętniczym.

Uwaga I.78. W wielu publikacjach popularnonaukowych można znaleźć wskazówki, jak wyznaczyć prawdopodobieństwo wystąpienia choroby w zależności od różnych czynników ryzyka. Na przykład w poradniku [302] podano ocenę ryzyka wystąpienia zawału mięśnia sercowego lub nagłego zgonu wieńcowego w ciągu 10 lat. W skali *PROCAM* wyróżnia się kilka czynników ryzyka: wiek, stężenie cholesterolu (wysokie LDL i niskie HDL) i trójglicerydów, palenie papierosów, cukrzyca, wystąpienie zawału u najbliższych krewnych przed 60. rokiem życia i wysokie skurczowe ciśnienie tętnicze. Za czynniki ryzyka przydzielane są punkty, które następnie sumujemy i w zależności od otrzymanej sumy odczytujemy z tabeli ocenę ryzyka. Na przykład 26 punktów otrzymują mężczyźni w przedziale wieku 60–65 lat, za stężenie cholesterolu LDL w zakresie 1,6–1,89 g/l przydziela się 14 punktów, a 8 punktów za skurczowe ciśnienie tętnicze powyżej 160. W przypadku 40 pkt i 48 pkt ryzyko wynosi odpowiednio 6,1% i 12,8%.

I.11 Wyznaczyć surowe i standaryzowane ryzyko zgonu w Polsce na podstawie danych z z tabel I.1 i I.5.

I.12 W tabeli I.7 podaliśmy dane dotyczące nowych przypadków zachorowalności na nowotwory złośliwe. Wyznaczyć surowe współczynniki zachorowań na nowotwory w różnych przedziałach wieku i według płci. Wyznaczyć standaryzowane współczynniki zachorowań na nowotwory w różnych przedziałach wieku dla obu płci łącznie.

	Mężczyźni			Kobiety		
W	0–44	45–64	65+	0–44	45–64	65+
L	11485	5149	2020	11080	5499	3305
Z	3985	28478	39323	6464	31462	34624

Tabela I.7. Zachorowania na nowotwory złośliwe w roku 2011 według wieku. L – liczebność danej grupy w tys., Z – liczba przypadków nowych zachorowań na nowotwory złośliwe (źrodło: Krajowy Rejestr Nowotworów).

I.13 Korzystając z surowych współczynników zachorowań na nowotwory wyznaczonych w zadaniu I.12, obliczyć dla kolejnych przedziałów wieku prawdopodobieństwa, że na nowotwór zachorował mężczyzna.

I.14 Przypuśćmy, że ryzyko zachorowania na nowotwór w ciągu roku wynosi 0,05% u ludzi w wieku 0–44, 0,6% w wieku 45–64 lat, a u starszych 1,4%. Wyznaczyć prawdopodobieństwo, że losowo wybrany mieszkaniec Polski zachoruje na nowotwór w tym roku. Przyjmujemy, że proporcje kolejnych grup wiekowych wynoszą 57%, 27% i 16%.

I.15 Przyjmując dane takie, jak w zadaniu I.14, wyznaczyć prawdopodobieństwa, że osoba, która zachorowała na nowotwór, jest w wieku 0–44, w wieku 45–64 lat oraz w wieku 65+.

I.16 Załóżmy, że przeprowadzono badania kontrolne na grupie osób, wśród których 97% jest zdrowych, 2% w początkowej fazie choroby i 1% w zaawansowanym stadium choroby. Załóżmy, że test daje wynik pozytywny (czyli sugeruje, że osoba jest chora) dla 1%, 95% i 99% osób w kolejnych grupach. Wyznaczyć prawdopodobieństwa, że test dał wynik pozytywny u osoby zdrowej i wyniki negatywne u osób w początkowym i zaawansowanym stadium choroby.

I.17 Załóżmy, że w badanej grupie jest 2% osób chorych. Pierwszy test daje wynik pozytywny u 98% chorych i 2% zdrowych, a drugi test odpowiednio 99% i 3%. Załóżmy, że testy są niezależne. Obliczyć prawdopodobieństwa, że osoba jest zdrowa, gdy wykonano jeden z tych testów i wynik był pozytywny, a także gdy wykonano oba testy i dwukrotnie wynik był pozytywny.

I.18 Załóżmy, że w grupie kontrolnej jest 2% osób chorych, a test daje wynik pozytywny u 99% chorych i 2% zdrowych. Osoba badana ma wysoki poziom leukocytów we krwi, a wiadomo, że ryzyko wystąpienia choroby u takiej osoby wynosi 10%; z drugiej strony, 90% chorych ma wysoki poziom leukocytów. Jakie jest prawdopodobieństwo, że osoba jest chora, jeżeli test dał wynik pozytywny?
Wskazówka. Potraktować badanie krwi jako wstępny test niezależny od testu właściwego.

I.19 Załóżmy, że w grupie kontrolnej jest 2% osób chorych, a test daje wynik pozytywny u 99% chorych i 2% zdrowych. Osoba badana ma wysoki poziom leukocytów we krwi, a wiadomo, że ryzyko choroby wśród takich osób jest trzy razy większe niż u pozostałych. Jakie jest prawdopodobieństwo, że osoba jest chora, jeżeli test dał wynik pozytywny?

I.20 Rozważmy cechę zależną od dwóch par genów leżących na różnych chromosomach. Załóżmy, że w wyjściowej populacji genotypy występują z następującymi częstościami: $AABB$ – x_{11}, $AABb$ – $2x_{12}$, $AAbb$ – x_{13}, $AaBB$ – $2x_{21}$,

$AaBb$ – $4x_{22}$, $Aabb$ – $2x_{23}$, $aaBB$ – x_{31}, $aaBb$ – $2x_{32}$ oraz $aabb$ – x_{33}. Niech

$$\begin{aligned} p_{AB} &= x_{11} + x_{12} + x_{21} + x_{22}, \\ p_{Ab} &= x_{12} + x_{13} + x_{22} + x_{23}, \\ p_{aB} &= x_{21} + x_{22} + x_{31} + x_{32}, \\ p_{aa} &= x_{22} + x_{23} + x_{32} + x_{33}, \\ I &= p_{aB} \cdot p_{Ab} - p_{AB} \cdot p_{ab} \end{aligned}$$

oraz

$$p^{\infty}_{AB} = p_{AB} + I, \quad p^{\infty}_{Ab} = p_{Ab} - I, \quad p^{\infty}_{aB} = p_{aB} - I, \quad p^{\infty}_{ab} = p_{ab} + I.$$

Wykazać, że prawdopodobieństwa genotypów w kolejnych pokoleniach stabilizują się, gdy liczba pokoleń dąży do nieskończoności, a rozkład graniczny jest postaci

$$\begin{array}{lll} p_{AABB} = (p^{\infty}_{AB})^2, & p_{AABb} = 2p^{\infty}_{AB} \cdot p^{\infty}_{Ab}, & p_{AAbb} = (p^{\infty}_{Ab})^2, \\ p_{AaBB} = 2p^{\infty}_{AB} \cdot p^{\infty}_{aB}, & p_{AaBb} = 2p^{\infty}_{AB} \cdot p^{\infty}_{ab} + 2p^{\infty}_{Ab} \cdot p^{\infty}_{aB}, & p_{Aabb} = 2p^{\infty}_{Ab} \cdot p^{\infty}_{ab}, \\ p_{aaBB} = (p^{\infty}_{aB})^2, & p_{aaBb} = 2p^{\infty}_{aB} \cdot p^{\infty}_{ab}, & p_{aabb} = (p^{\infty}_{ab})^2. \end{array}$$

Wykazać, że $p^{\infty}_{AB} + p^{\infty}_{Ab} + p^{\infty}_{aB} + p^{\infty}_{ab} = 1$ oraz $p^{\infty}_{AB} \cdot p^{\infty}_{ab} = p^{\infty}_{Ab} \cdot p^{\infty}_{aB}$. Niech $x = p_{AB} + p_{Ab}$ i $y = p_{AB} + p_{aB}$. Wykazać, że

$$p^{\infty}_{AB} = xy, \quad p^{\infty}_{Ab} = x(1-y), \quad p^{\infty}_{aB} = (1-x)y, \quad p^{\infty}_{ab} = (1-x)(1-y).$$

Wskazówka. Uzasadnić, że w wyjściowej populacji pary genów AB, Ab, aB, ab w komórkach rozrodczych występują z następującymi prawdopodobieństwami: p_{AB}, p_{Ab}, p_{aB}, p_{ab}. Następnie wyznaczyć rozkłady genotypów w pierwszym pokoleniu i wykazać, że pary genów w komórkach rozrodczych po pierwszym pokoleniu występują z prawdopodobieństwami: $p^1_{AB} = p_{AB} + \frac{1}{2}I$, $p^1_{Ab} = p_{Ab} - \frac{1}{2}I$, $p^1_{aB} = p_{aB} - \frac{1}{2}I$, $p^1_{ab} = p_{ab} + \frac{1}{2}I$. Obliczyć $I^1 = p^1_{aB} \cdot p^1_{Ab} - p^1_{AB} \cdot p^1_{ab}$ i wykazać, że $I^1 = \frac{1}{2}I$. Rozumując indukcyjnie, wyznaczyć graniczne wzory na p^{∞}_{AB}, p^{∞}_{Ab}, p^{∞}_{aB} oraz p^{∞}_{ab}, a następnie obliczyć graniczne rozkłady genotypów.

I.21 Rozważmy proces selekcji, w którym genotyp aa jest eliminowany z populacji z prawdopodobieństwem $\lambda \in (0,1)$. Wyprowadzić wzory rekurencyjne na rozkłady alleli A i a oraz genotypów AA, Aa i aa w kolejnych pokoleniach.

I.22 Niech a będzie skojarzonym płciowo genem recesywnym, a selekcja polega na eliminacji osobników płci męskiej z genotypem a. Opisać, jak zmieniają się rozkłady genotypów żeńskich AA, Aa, aa w kolejnych pokoleniach.

I.23 Niech a będzie skojarzonym płciowo genem recesywnym, a selekcja polega na eliminacji osobników płci żeńskiej z genotypem aa. Opisać, jak zmieniają się rozkłady genotypów żeńskich AA, Aa, aa w kolejnych pokoleniach.

I.24 Niech a będzie skojarzonym płciowo genem recesywnym, a selekcja polega na eliminacji osobników płci żeńskiej z genotypem aa oraz osobników płci męskiej z genotypem a. Opisać, jak zmieniają się rozkłady genotypów męskich i żeńskich w kolejnych pokoleniach.

I.25 Rozważmy teraz selekcję polegającą na eliminacji osobników płci męskiej z genotypem a z prawdopodobieństwem $\lambda \in (0,1)$. Opisać, jak zmieniają się rozkłady genotypów męskich A, a i żeńskich AA, Aa, aa w kolejnych pokoleniach.

I.26 Załóżmy, że w grupie kontrolnej jest 2% osób chorych, a test daje wynik pozytywny u 99% chorych i 2% zdrowych. Test wykonano trzykrotnie i dał k razy wynik pozytywny. Wyznaczyć prawdopodobieństwo, że osoba jest chora w zależności od k.
Wskazówka. Zastosować rozkład dwumianowy mieszany i regułę Bayesa. W tym wypadku wyznaczamy $\mathrm{P}(A|B)$, gdzie A jest zdarzeniem, że osoba jest chora, zaś B, że test wypadł pozytywnie k razy.

I.27 Udowodnić wzór (4.5).

I.28 Wykazać, że jeżeli ξ jest zmienną losową o rozkładzie Bernoulliego, to dla $k = cn$, gdzie $c \in (0,1)$, zachodzi wzór asymptotyczny

$$\mathrm{P}(\xi = k) \sim \frac{1}{\sqrt{2\pi n c(1-c)}} \exp\left\{-n\left(c \ln \frac{c}{p} + (1-c)\ln\frac{1-c}{1-p}\right)\right\}. \tag{7.45}$$

Wskazówka. Skorzystać ze wzoru (4.9).

I.29 Wyznaczyć $\Gamma(\frac{1}{2})$.
Wskazówka. Podstawić $x = y^2/2$ we wzorze (4.16).

I.30 Wykazać, że

$$\begin{aligned}\mathrm{E}(\xi - E\xi)^3 &= \mathrm{E}\,\xi^3 - 3\,\mathrm{E}\,\xi^2\,\mathrm{E}\,\xi + 2(\mathrm{E}\,\xi)^3,\\ \mathrm{E}(\xi - E\xi)^4 &= \mathrm{E}\,\xi^4 - 4\,\mathrm{E}\,\xi^3\,\mathrm{E}\,\xi - 6\,\mathrm{E}\,\xi^2(\mathrm{E}\,\xi)^2 - 3(\mathrm{E}\,\xi)^4,\end{aligned}$$

o ile momenty zwykłe po prawej stronie wzorów istnieją. Wyrażenie typu $\mathrm{E}(\xi - E\xi)^k$ nazywamy *momentem centralnym* rzędu k.

I.31 Rozważmy zmienną losową ξ o rozkładzie dwumianowym z parametrem p i liczbą prób n. Korzystając ze wzorów $\varphi_\xi(t) = (q + pe^{it})^n$, $q = 1-p$, oraz $\varphi_\xi^{(k)}(0) = i^k\,\mathrm{E}\,\xi^k$, wyznaczyć $\mathrm{E}\,\xi$, $\mathrm{E}\,\xi^2$, $\mathrm{E}\,\xi^3$ i $\mathrm{E}\,\xi^4$, a następnie wykazać, że

$$\mathrm{E}(\xi - \mathrm{E}\,\xi)^4 = (3(n-2)pq + 1)npq. \tag{7.46}$$

I.32 Wykazać, że wartość oczekiwana i wariancja dla rozkładu Poissona z parametrem λ wynoszą λ.

I.33 Niech ξ_p będzie zmienną losową o rozkładzie geometrycznym z parametrem p. Wyznaczyć granicę

$$\lim_{p\to 1^-} \frac{\mathrm{E}\,\xi_p}{m_p},$$

gdzie m_p jest medianą zmiennej ξ_p.

I.34 Wyznaczyć wartość oczekiwaną dla rozkładu logarytmicznego.

I.35 Wykazać, że rozkład gamma z parametrami α i λ ma wartość oczekiwaną λ/α i wariancję λ/α^2.

I.36 Wyznaczyć funkcje charakterystyczne rozkładu Poissona i rozkładu geometrycznego.

I.37 Niech ξ_1 i ξ_2 będą niezależnymi zmiennymi losowymi o rozkładzie Poissona z parametrami λ_1 i λ_2. Wykazać, że zmienna losowa $\xi = \xi_1 + \xi_2$ ma rozkład Poissona z parametrem $\lambda = \lambda_1 + \lambda_2$.
Wskazówka. Skorzystać z funkcji charakterystycznej lub ze wzoru

$$\mathrm{P}(\xi = k) = \sum_{j=0}^{k} \mathrm{P}(\xi_1 = j)\,\mathrm{P}(\xi_2 = k - j).$$

I.38 Samice wielu gatunków owadów składają stosunkowo dużą liczbę jaj, powiedzmy N, ale szanse osiągnięcia wieku dojrzałego ich potomków są małe i wynoszą p. Można więc przyjąć, że liczba potomków pojedynczej samicy, która osiągnęła wiek dojrzały, ma rozkład Poissona o parametrze $\lambda = pN$. Jaki będzie rozkład liczby dorosłych potomków n samic?

I.39 Wyznaczyć funkcje charakterystyczne rozkładu jednostajnego na przedziale $[a, b]$ i rozkładu wykładniczego.

I.40 Wykazać, że jeżeli zmienna losowa η jest mierzalna względem σ-algebry $\mathcal{A}$, to

$$\mathrm{E}(\eta\xi|\mathcal{A}) = \eta\,\mathrm{E}(\xi|\mathcal{A}).$$

I.41 Wykazać, że $\mathrm{E}(\mathrm{E}(\xi|\mathcal{A})) = E\xi$.

I.42 Wykazać, że jeśli zmienna losowa ξ i σ-algebra $\mathcal{A}$ są niezależne, to $\mathrm{E}(\xi|\mathcal{A}) = \mathrm{E}\,\xi$.

I.43 Niech ζ i η będą rzeczywistymi zmiennymi losowymi. Wykazać, że jeżeli zmienna losowa ζ jest mierzalna względem η (tzn. jest mierzalna względem σ-algebry $\mathcal{F}_\eta$), to istnieje taka funkcja mierzalna $g\colon \mathbb{R} \to \mathbb{R}$, że $\zeta = g(\eta)$.
Wskazówka. Niech $B_c \in \mathcal{B}(\mathbb{R})$ będą takimi zbiorami, że $\{\zeta < c\} = \eta^{-1}(B_c)$ dla $c \in \mathbb{R}$. Wykazać, że rodzinę $\{B_c\}$ można tak dobrać, że $B_{c_1} \subset B_{c_2}$, gdy $c_1 < c_2$, np. zamieniając ją na rodzinę $\tilde{B}_c = \bigcup\{B_q\colon q < c,\ q \in \mathbb{Q}\}$. Następnie przyjąć, że $g(x) = \alpha$ dla $x \in \bigcap_{n=1}^{\infty} B_{\alpha+1/n} \setminus B_\alpha$.

I.44 Załóżmy, że para zmiennych losowych ξ i η ma gęstość rozkładu $f(x,y)$. Wykazać, że $\mathrm{E}(\xi|\eta) = g(\eta)$, gdzie

$$g(y) = \frac{\int_{-\infty}^{\infty} x f(x,y)\,dx}{\int_{-\infty}^{\infty} f(x,y)\,dx}.$$

I.45 Niech $\xi_1, \ldots, \xi_n$ będą niezależnymi zmiennymi losowymi o tym samym rozkładzie wykładniczym z parametrem α. Wykazać, że gęstość rozkładu zmiennej losowej $\eta = \xi_1 + \cdots + \xi_n$ ma rozkład gamma z parametrami α i $\lambda = n$.

I.46 Niech $\xi_1, \ldots, \xi_n$ będą niezależnymi zmiennymi losowymi o rozkładach wykładniczych z parametrami odpowiednio $\alpha_1, \ldots, \alpha_n$. Wyznaczyć gęstość rozkładu zmiennej losowej $\eta = \min(\xi_1, \ldots, \xi_n)$.
Wskazówka. Zamiast $\mathrm{P}(\eta \le x)$ obliczyć $\mathrm{P}(\eta > x)$.

I.47 W Polsce w roku 2015 urodziło się 189 677 chłopców i 179 631 dziewczynek. Jakie jest prawdopodobieństwo, że w losowej grupie 10 000 noworodków będzie o 400 więcej chłopców niż dziewczynek?

I.48 Niech $f : \mathbb{R} \to \mathbb{R}$ będzie funkcją wypukłą, a ξ rzeczywistą zmienną losową. Udowodnić nierówność Jensena $\mathrm{E}\, f(\xi) \ge f(\mathrm{E}(\xi))$.
Wskazówka. Skorzystać z faktu, że funkcja wypukła f jest równa supremum po wszystkich takich prostych $y = ax + b$, że $ax + b \le f(x)$ dla $x \in \mathbb{R}$. Stąd $\mathrm{E}\, f(X) \ge \mathrm{E}(a + bX) = a + b\,\mathrm{E}\, X$.

I.49 W modelu demograficznym z przykładu I.3 wyznaczyć średnią, wariancję i medianę pozostałej długości życia pod warunkiem, że osoba jest w wieku i.

I.50 Korzystając z danych w tabeli I.1 i uwagi I.67, wyznaczyć średnią, odchylenie standardowe oraz medianę długości życia.

I.51 Korzystając z danych w tabeli I.1 i uwagi I.67, wyznaczyć średnią, odchylenie standardowe oraz medianę pozostałej długości życia dla osób w wieku 60 i 70 lat.

I.52 Wyznaczyć zależność wariancji pozostałej długości życia od funkcji przeżycia w modelu McKendricka.

I.53 Rozważmy model McKendricka ze współczynnikiem śmiertelności μ niezależnym od czasu i wieku. Wyznaczyć średnią i odchylenie standardowe długości życia. Jaki jest związek $G(a)$ z $G_x(a)$?

I.54 Wyznaczyć średnią długość życia w modelu McKendricka z rozkładem Weibulla, tj. dla $G(a) = e^{-(a/\lambda)^k}$.

I.55 Rozważmy model McKendricka ze współczynnikiem śmiertelności $\mu(a) = \mu_0 e^{ca}$, gdzie μ_0 i c są stałymi dodatnimi. Wyrazić zmienną losową T opisującą długość życia jako funkcję standardowego rozkładu wykładniczego.

I.56 Niech T_1 i T_2 będzie długością cyklu komórkowego dwóch komórek siostrzanych o początkowej dojrzałości x_0. Załóżmy, że komórki rozwijają się niezależnie. Wyznaczyć rozkład wielkości $|T_1 - T_2|$.

I.57 Niech f będzie gęstością rozkładu dojrzałości komórek w chwili rozpoczęcia cyklu komórkowego. Korzystając ze wzorów (6.7) i (6.10), wyznaczyć gęstość rozkładu dojrzałości początkowej ich komórek potomnych.

I.58 Rozważmy model cyklu komórkowego z niesymetrycznym podziałem. Załóżmy, że jeżeli x jest dojrzałością komórki w momencie podziału, to rozkład początkowej dojrzałości y komórek potomnych wynosi $p(y|x)$. Załóżmy, że komórka ma dojrzałość x_0 na początku cyklu komórkowego. Korzystając ze wzorów (6.7) i (6.10), wyznaczyć gęstość rozkładu dojrzałości komórek potomnych.

I.59 Udowodnić wzór (6.25).

I.60 Badamy liczebność pewnej populacji zamkniętej. W pierwszym odłowie oznakowano 150 osobników, a w drugim schwytano 130, z czego 13 było już oznakowanych. Wyznaczyć wielkość populacji, używając oszacowań (6.13), (6.15), (6.24) i (6.27). Wyznaczyć odchylenie standardowe dla estymatora Chapmana. Korzystając z nierówności Czebyszewa (5.4), wyznaczyć taki przedział $\Delta = (\hat{N}_C - a, \hat{N}_C + a)$, że $\mathrm{P}(N \in \Delta) > 0{,}8$. Jaki będzie przedział Δ, jeżeli założymy, że w otoczeniu wartości $\hat{N}_C$ rozkład liczby N jest dobrze przybliżany rozkładem normalnym?

I.61 Korzystając z metody Schnabel, oszacować wielkość populacji, jeżeli w kolejnych odłowach liczba N_i osobników odłowionych i liczba n_i osobników oznakowanych w każdej próbie przedstawione są w tabeli I.8. Porównać te wyniki z uzyskanymi za pomocą wzoru Lincolna–Petersena.

i	1	2	3	4	5	6
N_i	30	28	25	30	26	33
n_i	0	8	9	12	12	17

Tabela I.8. Dane do zadania I.61

I.62 Udowodnić wzór (7.7).

I.63 Udowodnić wzór (7.8).
Wskazówka. Niech $p_i = \mathrm{P}(\xi = x_i)$, $q_j = \mathrm{P}(\eta_j = y_j)$, $p_{ij} = \mathrm{P}(\xi = x_i, \eta_j = y_j)$. Wtedy $\sum_i p_{ij} = q_j$, $\sum_j p_{ij} = p_i$ oraz $H(\xi, \eta) = -\sum_{i,j} p_{ij} \log p_{ij}$. Bez straty

ogólności można przyjąć, że log oznacza logarytm naturalny. Wyznaczymy maksymalną wartość $H(\xi,\eta)$ przy założeniu, że p_i oraz q_j są ustalone. W tym celu korzystamy z metody czynników nieoznaczonych Lagrange'a i badamy, kiedy

$$\frac{\partial}{\partial p_{ij}}\Big[-\sum_{i,j} p_{ij}\ln p_{ij} - \mu_j\Big(\sum_i p_{ij} - q_j\Big) - \lambda_i\Big(\sum_j p_{ij} - p_i\Big)\Big] = 0$$

dla dowolnych i,j. Wykazujemy, że $p_{ij} = e^{-1-\lambda_i-\mu_j}$. Z naszych warunków wynika istnienie takich stałych c_1, c_2, że $e^{-\lambda_i} = p_i c_1$, $e^{-\mu_j} = q_j c_2$. Pozostaje wywnioskować, że $p_{ij} = p_i q_j$ w maksimum wartości $H(\xi,\eta)$.

I.64 Wykazać, że $H(\eta|\xi) = 0$ wtedy i tylko wtedy gdy $\eta = \varphi(\xi)$, gdzie φ jest pewną funkcją.

I.65 Niech $\alpha_1,\dots,\alpha_n$ będą pewnymi liczbami dodatnimi i niech m spełnia nierówność $\min_i \alpha_i < m < \max_i \alpha_i$. Wykazać, że istnieje liczba rzeczywista λ spełniająca równanie

$$\sum_{i=1}^n \alpha_i e^{-\lambda\alpha_i} = m\sum_{i=1}^n e^{-\lambda\alpha_i}.$$

Wykazać, że wśród wszystkich rozkładów $\mathbf{p} = (p_1,\dots,p_n)$ spełniających warunek

$$\sum_{i=1}^n \alpha_i p_i = m$$

największą entropię ma *rozkład Boltzmanna*

$$q_i = \frac{1}{c}e^{-\lambda\alpha_i}, \quad \text{gdzie} \quad c = \sum_{i=1}^n e^{-\lambda\alpha_i}.$$

Wskazówka. Wykazać, że $H(\mathbf{q}) = -\sum_{i=1}^n p_i \ln q_i$. Stąd

$$0 \le H_{KL}(\mathbf{p}|\mathbf{q}) = -H(\mathbf{p}) + H(\mathbf{q}).$$

I.66 Ustalmy $a < 0$ i $\mathbf{b} \in \mathbb{R}^3$. W przestrzeni $X = \mathbb{R}^3$ z miarą Lebesgue'a dobieramy $c > 0$ tak, aby funkcja $M(\mathbf{x}) = c\exp(a\|\mathbf{x}\|^2 + \mathbf{b}\cdot\mathbf{x})$ była gęstością. Wykazać, że wśród wszystkich gęstości f spełniających warunek

$$\int_{\mathbb{R}^3} g(\mathbf{x})f(\mathbf{x})\,d\mathbf{x} = \int_{\mathbb{R}^3} g(\mathbf{x})M(\mathbf{x})\,d\mathbf{x}$$

dla $g(\mathbf{x}) = x_1$, dla $g(\mathbf{x}) = x_2$, dla $g(\mathbf{x}) = x_3$ i dla $g(\mathbf{x}) = \|\mathbf{x}\|^2$ entropia $H(f)$ osiąga największą wartość dla $f = M$.
Wskazówka. Wykazać, że $H_{KL}(f|M) = H(M) - H(f)$.