Piotr Jerzy Durka

Wstęp do współczesnej statystyki

Wydawnictwo Adamantan

Redakcja: Beata Bednarczuk, Witold Mizerski

Projekt okładki: Paweł Łukomski

Ilustracja na okładce:
© zefa/Henryk T. Kaiser

Książkę złożył Autor za pomocą programu LaTeX, czcionką *antykwa Półtawskiego*. Czcionkę tę zaprojektował w latach 1923–1928 grafik i typograf Adam Półtawski, a polskiej typografii komputerowej przywróciła Grupa Użytkowników Systemu TeX `http://www.gust.org.pl`

Wydawca:
Wydawnictwo Adamantan
ul. Powstańców Śląskich 106B/33
PL 01-466 Warszawa
tel. (0-22) 4361955, fax 4361965
internet: www.adamantan.com.pl

ISBN 83-85655-82-4

Warszawa 2003

Druk:
OpolGraf S.A.
ul. Niedziałkowskiego 8–12
45-085 Opole

Spis treści

1 Wstęp **7**
1.1 Co znajdziemy w tej książce i jak z niej korzystać 8

I Statystyka — z komputerem zamiast wzorów **11**

2 Monte Carlo **13**
2.1 Hazard symulowany 13
2.2 Jak zmusić komputer do rzucania kostką 14
2.3 Oceniamy szanse 17
2.4 „Prawdziwe" Monte Carlo 21

3 Bootstrap **22**

4 Testy permutacyjne **25**
4.1 Histogram 27
4.2 Weryfikacja hipotez statystycznych — terminologia 28
4.3 Poziom istotności testu 28
4.4 Testy permutacyjne dla większych liczebności 30
4.5 Test jedno- i dwustronny 32

5 Zastosowania **33**
5.1 Przykłady 33
5.1.1 Urodziny tego samego dnia 33
5.1.2 „Trzy wśród czworga" 33
5.1.3 Awaria 2 z 12 34
5.1.4 Sondy przedwyborcze 34
5.1.5 Poprawa wyniku 35
5.2 Ile razy losować? 36
5.3 Co dalej? 37

II Podstawy teorii klasycznej **39**

6 Prawdopodobieństwo . . . **42**
6.1 Definicje prawdopodobieństwa . . . 42
6.2 Prawdopodobieństwo warunkowe i zdarzenia niezależne . . . 45
6.3 Twierdzenie Bayesa . . . 45

7 Rozkłady prawdopodobieństwa . . . **49**
7.1 Rozkłady ciągłe — gęstość prawdopodobieństwa . . . 50
7.2 Wariancja, korelacja, mediana . . . 51
7.2.1 Wartość oczekiwana . . . 52
7.2.2 Mediana . . . 52
7.2.3 Wariancja . . . 52
7.2.4 Kowariancja i współczynnik korelacji . . . 53
7.3 Rozkład równomierny . . . 53
7.4 Rozkład dwumianowy . . . 55
7.5 Rozkład Poissona . . . 57
7.6 Rozkład Gaussa . . . 59
7.7 Centralne Twierdzenie Graniczne . . . 60

8 Statystyki i estymatory . . . **63**
8.1 Estymator wartości oczekiwanej . . . 63
8.2 Prawo wielkich liczb . . . 64
8.3 Estymator wariancji . . . 65

9 Weryfikacja hipotez statystycznych . . . **66**
9.1 Test Z: rozkład normalny, znane σ i μ . . . 67
9.1.1 Poziom istotności i moc testu . . . 69
9.1.2 Schemat Weryfikacji Hipotez Statystycznych raz jeszcze . . . 72
9.2 Rozkład t (Studenta) . . . 73
9.2.1 Test t (Studenta) różnicy średnich . . . 75
9.3 Rozkład χ^2 . . . 77
9.3.1 Za dokładnie? . . . 79
9.3.2 Test χ^2 Pearsona . . . 81

10 Testy nieparametryczne . . . **86**
10.1 Test serii Walda–Wolfowitza . . . 88
10.1.1 Testowanie, czy próba jest wynikiem niezależnych losowań . . . 92
10.1.2 Test zgodności rozkładów w dwóch populacjach . . . 92
10.2 Test rang Wilcoxona–Manna–Whitneya . . . 93
10.2.1 Statystyka Wilcoxona . . . 93
10.2.2 Statystyka Manna–Whitneya . . . 94
10.2.3 Równoważność statystyk $W_{m,n}$ i $M_{m,n}$ . . . 94

III Dodatki **97**

A Czego się nie da obliczyć **99**
A.1 Problem stopu 100
A.2 Notacja $\mathcal{O}(\cdot)$ 100
A.3 Problem komiwojażera 101

B Programy w języku *Matlab* **104**

Rozdział 1

Wstęp

Nigdy w historii matematyki tak wielu nie popełniało tak licznych błędów w tak niewielu zastosowaniach[1] — chodzi oczywiście o statystykę. Dlaczego?

- Zdecydowana większość ludzi korzystających z metod statystycznych to specjaliści w zupełnie innych dziedzinach, względem których statystyka pełni rolę służebną.

- Klasyczna teoria statystyki powstawała ponad pół wieku temu i z braku podówczas komputerów opiera się na zaawansowanych metodach analitycznych (czytaj: długich i skomplikowanych wzorach) oraz koniecznych do ich wyprowadzenia założeniach, nie zawsze spełnianych w praktyce.

- Próba wyjaśnienia tej złożonej teorii na kursie lub w podręczniku dla nie-statystyków kończy się zwykle katalogiem przepisów „kiedy stosować który test". Niestety, żaden katalog nie uwzględni wszystkich przypadków, z którymi możemy mieć do czynienia, i nie zastąpi *zrozumienia* podstaw.[2]

- Główną konsekwencją rozpowszechnienia komputerów jest ułatwienie dostępu do tych skomplikowanych metod: z wczytaniem danych do specjalizowanego pakietu statystycznego jakoś sobie poradzimy, potem tylko trzeba „doklikać się" do jakiegoś testu i... komputer zawsze „wyrzuci" jakiś wynik. Ale komputer nie przyjmie odpowiedzialności za dobór metody do problemu i poprawne sformułowanie hipotezy.

---

[1] Parafraza wypowiedzi Winstona Churchilla w brytyjskim parlamencie 20 sierpnia 1940 roku, dotyczącej pilotów — w dużej części Polaków — walczących w Bitwie o Anglię: *Never in the field of human conflict was so much owed by so many to so few.* (Nigdy na polu ludzkich konfliktów tak wielu nie zawdzięczało tak wiele tak nielicznym).

[2] Na przykład studium 50 artykułów w najbardziej prestiżowym czasopiśmie medycznym (New England Journal of Medicine), w których wykorzystano do analizy wyników test t (rozdział 9.2.1) wykazało, że w ponad połowie z nich użycie tego testu było nieprawidłowe — cytat za [16].

Na szczęście komputery niosą tu również dobrą nowinę. Są nią nowe[3], rewolucyjnie proste i intuicyjne metody oparte na idei repróbkowania (ang. *resampling*) — testy permutacyjne i bootstrap — oraz możliwość szerokiego stosowania symulacji Monte Carlo. Uwalniając użytkownika od skomplikowanej teorii i wzorów pozwalają skupić się na istocie pytania, na które statystyka ma odpowiedzieć. Ponadto działają często w sytuacjach, w których tradycyjne metody analityczne zawodzą (jak np. bootstrap w szacowaniu błędów złożonych funkcji).

Nowe metody oparte są na „brutalnej mocy" obliczeniowej. Kilkadziesiąt lat temu fakt ten uniemożliwiał ich praktyczne zastosowanie (pewnie dlatego nie zawracano sobie podówczas głowy ich wymyślaniem). Kilkanaście lat temu stanowiło to poważną przeszkodę w ich rozpowszechnieniu. Dzisiaj stosowanie tych metod może Ci co najwyżej uświadomić, że komputer na Twoim biurku ma w sobie więcej mocy obliczeniowej niż maszyna do pisania, którą na co dzień zastępuje.

Wreszcie niekwestionowanym walorem tych metod jest ich ogromna wartość dydaktyczna, umożliwiająca zrozumienie podstaw *przed* zmierzeniem się z komplikacjami matematycznymi i ideowymi statystyki klasycznej.

1.1 Co znajdziemy w tej książce i jak z niej korzystać

Pierwsza część to luźne i bezstresowe (bez użycia wzorów) wprowadzenie podstawowych pojęć statystyki, dzięki którym dochodzimy — wciąż bez żadnego wzoru — do całkiem poważnych i przydatnych zastosowań metody Monte Carlo i repróbkowania (testów permutacyjnych i botstrapu). Celem tej części jest:

- zapoznanie Czytelnika z najnowszymi trendami w statystyce,
- umożliwienie samodzielnego i poprawnego rozwiązywania wielu problemów statystycznych w sposób intuicyjny drogą symulacji komputerowych, co pozwala na skoncentrowanie się na poprawnym sformułowaniu problemu (hipotezy) i znacząco zmniejsza szansę popełnienia grubego błędu metodycznego,
- wprowadzenie w sposób intuicyjny i na konkretnych przykładach pojęć z klasycznej teorii statystyki (jak np. poziom istotności i moc testu), co ułatwi zrozumienie części drugiej.

[3]Ideę testów permutacyjnych po raz pierwszy zaproponował R. A. Fischer w latach 1930-tych jako teoretyczny argument za testem t Studenta (rozdział 9.2); symulacje Monte Carlo (Stanisław Ulam, rozdział 2.4 na stronie 21) zaczęto stosować po II wojnie światowej, gdy pojawiły się pierwsze komputery. Idee repróbkowania (Julian L. Simon) i bootstrapu (Bradley Efron) w dzisiejszej postaci sformułowano w latach 80. XX wieku, jednak praktyczne możliwości wykorzystania tych metod na szerszą skalę pojawiły się dopiero w latach 90. dzięki rozwojowi technologii komputerowej.

W **części drugiej** całki itp. są już nie do uniknięcia; jednak liczba wzorów podawanych bez dowodu ograniczona jest do koniecznego minimum. Znajdziemy tam:

- podstawy wystarczające do zrozumienia klasycznej — i wciąż najbardziej powszechnej — metodologii weryfikacji hipotez statystycznych; stanowi ona podstawę większości zastosowań wnioskowania statystycznego, czyli „testów statystycznych",
- dokładne i poparte przykładami omówienie najczęściej stosowanych testów: t Studenta, χ^2 dla tabel i dopasowania rozkładu, testu serii Walda–Wolfowitza i testu rang Wilcoxona–Manna–Whitneya,
- wyprowadzenie od podstaw statystyki testu serii (rozdział 10.1), co pozwala na prześledzenie kompletnej drogi powstawania metody statystycznej również w podejściu klasycznym,
- oparty na wielu dokładnie analizowanych przykładach opis podstawowego schematu weryfikacji hipotez statystycznych, na którym opierają się wszystkie powszechnie stosowane testy statystyczne.

Do zrozumienia części pierwszej nie jest wymagane praktycznie żadne przygotowanie matematyczne. Dla samodzielnego zastosowania opisywanych w niej metod konieczne jest zastosowanie dowolnego języka programowania bądź specjalizowanego pakietu statystycznego.

Przyswojenie podstawowych pojęć wprowadzonych w części pierwszej znacznie ułatwia zrozumienie części drugiej, w której korzysta się już z pojęcia całki i podstaw kombinatoryki.

Pierwsza część książki oparta jest na intensywnym wykorzystaniu komputerów. **Dodatek A** opisuje ogólne ograniczenia, którym podlegają wszelkie rozwiązania problemów za pomocą maszyn liczących.

Na koniec **Dodatek B** zawiera oryginalne teksty wszystkich programów, wykorzystanych do tworzenia rysunków i wykonywania obliczeń prezentowanych w tej książce, w języku *Matlab*. Jest to język wysokiego poziomu o stosunkowo intuicyjnej składni, dzięki czemu teksty te mogą stanowić uzupełnienie opisywanych algorytmów również dla osób nie korzystających z pakietu *Matlab*. Studiowanie tych programów nie jest bynajmniej konieczne do zrozumienia prezentowanych w książce zagadnień. Programy te, jak również inne związane z książką materiały i ewentualne uaktualnienia, znaleźć można w Internecie pod adresem `http://statystyka.durka.info`.

bootstrap, 22

Centralne Twierdzenie Graniczne, 60

estymator
  wariancji, 65
  wartości oczekiwanej, 63

hipoteza
  alternatywna, 70
  zerowa, 28, 68, 72, 92

korelacja *zob.* współczynnik korelacji 53
krzywa dzwonowa *zob.* rozkład Gaussa 59

mediana, 52
moc testu, 69
Monte Carlo, 13, 21

obszar krytyczny, 29, 32
odchylenie standardowe, 53
  wartości średniej, 65

poziom istotności, 28–30, 32, 35, 69, 72
prawdopodobieństwo
  aksjomaty, 42
  definicje, 42
  pozorny paradoks, 43
  warunkowe, 45
prawo wielkich liczb, 64
przykłady:
  awarie ciężarówek, 34
  braki w produkcji, 25, 30
  hazard z kostką, 14
  nieuczciwy ankieter (serie), 87
  obliczanie prawdopodobieństwa, 43
  poprawa wyniku, 35, 57
  prawd. warunkowe, 45
  problem komiwojażera, 101
  problem nierozstrzygalny, 100
  problem stopu, 100
  przyspieszenie ziemskie, 66
  rozkład dwumianowy, 56
  sondy przedwyborcze, 34
  test na HIV, 47
  test Studenta, 77
  Titanic, 83
  trzy dziewczynki, 33, 57
  urodziny tego samego dnia, 33
  za dokładnie, 79

rozkład
  χ^2, 77
  t, 73
  dwumianowy, 55
  Gaussa, 59
  jednostajny *zob.* równomierny 53
  normalny *zob.* Gaussa 59
  płaski *zob.* równomierny 53
  Poissona, 57
  prostokątny *zob.* równomierny 53
  równomierny, 53
  Studenta *zob.* rozkład t 73

statystyka, 63

test
  Z, 67
  χ^2 Pearsona, 81
  t, 75
  rang *zob.* test Wilcoxona–Manna–Whitneya 93
  serii, 88
  Studenta *zob.* test t 75
  Wilcoxona–Manna–Whitneya, 93
  zgodności rozkładów (nieparametryczny), 92
twierdzenie
  Bayesa, 45
  centralne *zob.* Centralne Twierdzenie Graniczne 60
  Lindeberga–Levy'ego *zob.* Centralne Twierdzenie Graniczne 60

von Münchhausen, 22

wariancja, 52
wartość oczekiwana, 52
współczynnik korelacji, 53