

Statystyka dla studentów
kierunków technicznych i przyrodniczych

Jacek Koronacki, Jan Mielniczuk

Spis treści

1	Wstępna analiza danych	13
1.1.	Wprowadzenie	13
1.2.	Graficzne przedstawienie danych	13
1.2.1.	Wykresy dla danych jakościowych	14
1.2.2.	Wykresy dla danych ilościowych	17
1.2.3.	Wykresy przebiegu	25
1.3.	Wskaźniki sumaryczne	27
1.3.1.	Wskaźniki położenia	27
1.3.2.	Wskaźniki rozproszenia	35
1.3.3.	Wykres ramkowy	41
1.4.	Gęstości rozkładów – wprowadzenie	49
1.4.1.	Podstawowe pojęcia	49
1.4.2.	Gęstości normalne	52
1.5.	Zadania	56
2	Od modelu probabilistycznego do wnioskowania statystycznego	61
2.1.	Model probabilistyczny – podstawy	61
2.1.1.	Doświadczenia losowe i rachunek zdarzeń losowych	62
2.1.2.	Prawdopodobieństwo	69
2.1.3.	Prawdopodobieństwo warunkowe i zdarzenia niezależne	79
2.2.	Zmienne losowe	92
2.2.1.	Zmienne dyskretne i ich rozkłady	94

2.2.2.	Wskaźniki położenia i rozproszenia dla dyskretnej zmiennej losowej	99
2.2.3.	Przykłady rozkładów dyskretnych	104
2.2.4.	Ciągłe zmienne losowe	111
2.2.5.	Wskaźniki położenia i rozproszenia dla ciągłych zmiennych losowych	114
2.2.6.	Przykłady ciągłych zmiennych losowych	116
2.2.7.	Nierówność Czebyszewa	120
2.3.	Para zmiennych losowych	122
2.4.	Wnioskowanie statystyczne – podstawy	138
2.4.1.	Podstawowe pojęcia	138
2.4.2.	Rozkład średniej w prostej próbie losowej	140
2.4.3.	Rozkład częstości	147
2.4.4.	Estymatory i ich podstawowe własności	150
2.5.	Metody zbierania danych	158
2.5.1.	Podstawowy schemat eksperymentalny	158
2.5.2.	Inne schematy eksperymentalne	162
2.6.	Zadania	165
3	Wnioskowanie statystyczne	174
3.1.	Wprowadzenie	174
3.2.	Estymacja punktowa	177
3.2.1.	Estymatory największej wiarygodności	177
3.2.2.	Estymatory oparte na metodzie momentów	189
3.2.3.	M-estymatory	193
3.3.	Estymacja przedziałowa	196
3.3.1.	Przedziały ufności dla wartości średniej rozkładu normalnego	197
3.3.2.	Przedziały ufności dla wariancji rozkładu normalnego	205
3.3.3.	Uwaga o przedziałach ufności w przypadku rozkładów ciągłych, innych niż normalny	209
3.3.4.	Przedziały ufności dla proporcji	210
3.4.	Testowanie hipotez	213
3.4.1.	Testowanie hipotez w rodzinach rozkładów normalnych i rozkładów dwupunktowych	213

3.4.2. Testowanie zgodności	238
3.5. Zadania	254
4 Analiza zależności zmiennych ilościowych	260
4.1. Wprowadzenie	260
4.2. Analiza zależności dwóch zmiennych ilościowych	260
4.2.1. Współczynnik korelacji próbkowej	263
4.2.2. Liniowa zależność między dwiema zmiennymi, prosta regresji	265
4.2.3. Model zależności liniowej	272
4.2.4. Wnioskowanie w modelu zależności liniowej	276
4.2.5. Analiza wartości resztowych	284
4.3. Analiza zależności wielu zmiennych ilościowych	291
4.3.1. Model liniowy regresji wielokrotnej	293
4.3.2. Własności estymatorów MNK	297
4.3.3. Diagnostyka modelu regresji	304
4.3.4. Analiza zależności parametrów samochodów	311
4.4. Zadania	313
5 Analiza wariancji	319
5.1. Wprowadzenie	319
5.2. Analiza jednoczynnikowa	321
5.2.1. Test F analizy wariancji	321
5.2.2. Związki z analizą regresji	331
5.2.3. Porównania wielokrotne	334
5.2.4. Zrandomizowany plan blokowy	337
5.3. Analiza dwuczynnikowa	342
5.4. Zadania	354
6 Analiza danych jakościowych	359
6.1. Wprowadzenie	359
6.2. Analiza jednej zmiennej	366
6.2.1. Uwagi wstępne	366
6.2.2. Testowanie prostej hipotezy o zgodności	367
6.2.3. Testowanie złożonej hipotezy o zgodności	372

6.3.	Testowanie jednorodności	375
6.4.	Analiza dwóch zmiennych losowych	377
6.4.1.	Testowanie niezależności	377
6.4.2.	Analiza zależności	379
6.5.	Uwagi o poprawności wnioskowania i paradoksy Simpsona	390
6.6.	Zadania	393
7	Metody wyboru prób z populacji skończonej	398
7.1.	Metoda reprezentacyjna	398
7.1.1.	Cel metody reprezentacyjnej	398
7.1.2.	Podstawowe schematy losowania prób	401
7.2.	Estymatory parametrów populacji dla różnych schematów losowania	405
7.2.1.	Estymator Horwitza–Thompsona wartości średniej cechy . . .	405
7.2.2.	Przedział ufności dla wartości średniej cechy	410
7.2.3.	Estymatory wartości średniej cechy oparte na cenie dodatkowej	412
7.2.4.	Estymator proporcji	415
7.2.5.	Estymacja ilorazu wartości średnich	417
7.2.6.	Estymatory średniej dla schematu losowania warstwowego . .	420
7.3.	Zadania	423
8	Metoda Monte Carlo	426
8.1.	Wprowadzenie	426
8.2.	Generatory liczb pseudolosowych	427
8.2.1.	Generatory liczb pseudolosowych z rozkładu jednostajnego .	427
8.2.2.	Metoda przekształcenia kwantylowego	429
8.2.3.	Metoda oparta na reprezentacji zmiennych losowych	430
8.2.4.	Metoda eliminacji	433
8.3.	Szacowanie parametrów rozkładu metodą Monte Carlo	435
8.3.1.	Estymatory parametrów rozkładu otrzymane metodą Monte Carlo	435
8.3.2.	Błędy standardowe estymatorów i przedziały ufności	436
8.3.3.	Modelowanie eksperymentów losowych metodą Monte Carlo .	438
8.4.	Testy permutacyjne	441
8.4.1.	Testowanie jednorodności	441

8.4.2.	Testowanie niezależności cech	445
8.5.	Estymacja rozkładu statystyki metodą bootstrap	445
8.5.1.	Zasada bootstrap	446
8.5.2.	Błąd standardowy typu bootstrap	449
8.5.3.	Przedziały ufności typu bootstrap	451
8.5.4.	Testowanie hipotez przy użyciu metody bootstrap	453
8.6.	Zadania	454
9	Metody rangowe	458
9.1.	Wprowadzenie	458
9.2.	Porównanie rozkładu cech w dwóch populacjach	459
9.2.1.	Test Wilcoxona	460
9.2.2.	Własności statystyki Wilcoxona	463
9.2.3.	Estymacja parametru przesunięcia Δ	466
9.2.4.	Test Kołmogorowa–Smirnowa	467
9.3.	Testy porównania rozkładów dla par obserwacji	467
9.3.1.	Test Wilcoxona dla par obserwacji	467
9.3.2.	Własności statystyki Wilcoxona dla par obserwacji	469
9.3.3.	Estymacja parametru przesunięcia Δ	471
9.3.4.	Test znaków	472
9.4.	Rangowe testy niezależności	472
9.4.1.	Współczynnik korelacji Spearmana	473
9.4.2.	Współczynnik Kendalla	474
9.5.	Porównanie rozkładów cech w wielu populacjach	475
9.5.1.	Test Kruskalla–Wallisa	476
9.5.2.	Porównania wielokrotne	479
9.6.	Metody rangowe dla modelu regresji liniowej	480
9.7.	Zadania	481

Przedmowa

Miniony wiek bywa nazywany wiekiem informacji. Moc obliczeniowa komputerów oraz pojemność ich pamięci rosły w ostatnich dziesięcioleciach nieomal z dnia na dzień. Doświadczaliśmy i nadal doświadczamy niebywałego rozwoju możliwości komunikacji w sieciach komputerowych. Oczywiście wiązał się z tym wszystkim ogromny wzrost możliwości gromadzenia informacji. W przypadku dużych baz danych mamy często do czynienia z megabajtami i nierzadko z terabajtami danych. Wielkiego znaczenia nabrała zatem potrzeba inteligentnego przetwarzania zebranych informacji.

Niejako w cieniu owej rewolucji informatycznej przez cały XX wiek trwał też niezwykle rozwój analizy danych i wnioskowania statystycznego, czyli – statystyki. Obydwa procesy nie były przy tym od siebie niezależne. Z jednej strony, bez statystyki nie ma możliwości pełnego zrozumienia i zinterpretowania wiedzy ukrytej w danych. Z drugiej zaś, techniczny rozwój komputerów umożliwił zalgorytmizowanie procedur statystycznych i rozwiązywanie zadań, z którymi człowiek sam nie mógłby sobie poradzić przez dziesiątki lat wyętej pracy.

Jednym ze skutków rozwoju informatyki i statystyki jest upowszechnienie badań statystycznych w niemal we wszystkich dziedzinach nauki i praktyki. We wszystkich sferach naszej działalności zbieramy dane, które – nie poddane odpowiedniej analizie – jawią się raczej jako niewiele mówiący chaos niż pewne uporządkowane *uniwersum*. Dzięki statystyce dostrzegamy ów ukryty w danych porządek oraz patrzymy na dane z właściwej perspektywy, tak jak autor obrazu, którego reprodukcję zamieszczamy na okładce, spojrział na ulicę Madrytu – z bliska pełną południowego zamętu i chaotycznie poruszających się ludzi, a z perspektywy będącą częścią pięknego i zrozumiałego ładu.

Podręcznik ten jest wprowadzeniem w ładu oferowany przez statystykę. Jest adresowany przede wszystkim do przyszłych techników i przyrodników, ale

uważamy, że będzie przydatny także dla studentów innych kierunków, zwłaszcza ekonomicznych, rolniczych, społecznych i medycznych. Powinien również zainteresować tych absolwentów wszystkich wymienionych kierunków, którzy uważają, że ich podstawowa wiedza statystyczna jest niedostateczna.

Oddawany do rąk Czytelnika podręcznik odbiega stylem od książek ze statystyki matematycznej już choćby dlatego, że jest adresowany do niematematyków. Swoją konstrukcją nawiązuje do anglosaskiej tradycji uczenia statystyki, którego celem jest danie dogłębnego i szerokiego, ale zarazem możliwie przystępnego wprowadzenia do przedmiotu.

Jesteśmy przekonani, że pojawienie się tego rodzaju podręcznika jest potrzebne, ponieważ przyczyni się do podniesienia poziomu zrozumienia i popularności statystyki wśród studentów. Niezbędne jest ukazanie się książki umożliwiającej wprowadzenie do statystyki w sposób całościowy, w pełni wykorzystującej nowoczesne techniki obliczeniowe. Jednocześnie Czytelnik musi nauczyć się wykorzystywania owych narzędzi w sposób odpowiedzialny i oparty na dobrym zrozumieniu przedmiotu. Nie trzeba nikogo przekonywać, że obecność na rynku licznych komputerowych pakietów statystycznych jest tyleż błogosławieństwem, co i przekleństwem, bowiem łatwo z nich korzystać bez żadnego zrozumienia oferowanych przez pakiet wyników.

Środkiem do lepszego zrozumienia przedmiotu nie może być nadmierna ścisłość formalna i przytaczanie wielu dowodów, lecz oparcie się na pouczających choć prostych przykładach i szerokiej argumentacji, odwołującej się do zdrowego rozsądku. Tak właśnie pisany jest nasz podręcznik. Jesteśmy przekonani, że w ten sposób można doskonale przekazać istotę rozumowania statystycznego. Za złożonym nawet matematycznym wywodem zawsze kryje się przejrzysta intuicja. To ją przede wszystkim powinien osiąść Czytelnik.

Zakres trzech pierwszych rozdziałów książki (rozdział zawierający wstępną analizę danych, rozdział poświęcony przejściu od modelu probabilistycznego do wnioskowania statystycznego oraz rozdział opisujący podstawy wnioskowania statystycznego) odpowiada typowym uczelnianym kursom ze statystyki; jednak różni się sposobem ujęcia materiału. W sposobie wykładu oraz wyborze tematów szczegółowych kierujemy się potrzebami praktyki oraz swymi doświadczeniami dydaktycznymi z uczelni w Warszawie (w ostatnich latach PJWSTK), The University of Michigan w Ann Arbor, Rice University w Houston i The University of New South Wales w Sydney. Niewątpliwie wpływ wywarły na nas najlepsze podręczniki anglosaskie, zwłaszcza książka Moore'a i McCabe'a „Introduction to the Practice of Statistics”, Freeman & Co 1998, którą najczęściej sami wykorzystywaliśmy w nauczaniu.

Trzy pierwsze rozdziały tego podręcznika uzupełnione o omówioną w rozdziale 4 analizę regresji są pomyślane jako podstawa kursu semestralnego,

wprowadzającego słuchacza w zagadnienia statystyki i obejmującego tygodniowo przynajmniej dwugodzinny wykład oraz dwugodzinne laboratorium. W ramach kursu semestralnego udaje się omówić tylko zasadnicze kwestie analizy regresji. W pięciu następnych rozdziałach przedstawiono wybrane, najbardziej istotne dla praktyka zagadnienia statystyki: analizę wariancji i analizę zależności cech jakościowych, metody próbkowania, zagadnienia symulacji komputerowej i metod rangowych. Na podstawie tych rozdziałów oraz zaprezentowanej obszerniej analizy regresji, wykładowca może zaplanować drugi semestr wykładu ze statystyki.

Wśród zagadnień szczegółowych nie znalazło się miejsce dla niezwykle ważnych metod statystyki wielowymiarowej, które – mamy nadzieję – staną się treścią naszego następnego podręcznika. Nie mamy przy tym wątpliwości, że ze względu na występującą obecnie złożoność danych metody wielowymiarowe staną się już wkrótce elementem podstawowych wykładów ze statystyki.

Chcielibyśmy podkreślić fakt, że książka powstała w ramach działalności statutowej Instytutu Podstaw Informatyki Polskiej Akademii Nauk. W trakcie przygotowywania kolejnych wersji manuskryptu, korzystaliśmy z wnikliwych uwag Stanisława Gnota, Andrzeja Dąbrowskiego i Andrzeja Michalskiego oraz Elżbiety Ferenstein, którym serdecznie dziękujemy. Jesteśmy bardzo wdzięczni naszemu najbliższemu współpracownikowi, Janowi Ćwikowi, który sporządził wszystkie rysunki, przygotował ostateczny skład książki, przeliczył i sprawdził wiele przykładów oraz pomagał nam w trakcie kolejnych korekt. Składamy podziękowania Polsko-Japońskiej Wyższej Szkole Technik Komputerowych za finansowe wsparcie wydania naszego podręcznika. Dziękujemy Muzeum Narodowemu w Warszawie za wyrażenie zgody na reprodukcję obrazu Józefa Pankiewicza „Ulica w Madrycie”. Z wdzięcznością myślimy o znakomitej pracy redakcyjnej Pani Lilianny Szymańskiej i o opiece Pani Redaktor Zofii Leszczyńskiej nad całością przedsięwzięcia. Bez ich wielkiego poświęcenia i zaangażowania szybkie wydanie tej książki byłoby niemożliwe.

Jacek Koronacki i Jan Mielniczuk

Warszawa, w lipcu 2001