Metody a řešené úlohy
Kompendium
statistického
zpracování dat
Milan Meloun
Jiří Militký
ACADEMIA

MILAN MELOUN, JIŘÍ MILITKÝ

Kompendium statistického zpracování dat

Metody a řešené úlohy v chemometrii, biometrii, a v dalších oborech přírodních, technických a společenských věd včetně CD

(2. doplněné a rozšířené vydání o 1100 úlohách)

Academia 2006

Obsah

Předmluva autorů 11
O autorech 16

1 CHYBY, VARIABILITA A NEJISTOTY INSTRUMENTÁLNÍCH MĚŘENÍ 21
1.1 Chyby měřicích přístrojů 21
Vzorová úloha 1.1 *Absolutní a relativní chyba pH-metru* 25
Vzorová úloha 1.2 *Třída přesnosti a práh citlivosti ampérmetru* 26
Vzorová úloha 1.3 *Mezní absolutní a relativní chyba ampérmetru* 26
1.2 Způsoby vyjádření odhadů chyb měření 26
1.2.1 Momentové odhady chyb 27
Vzorová úloha 1.4 *Relativní a absolutní systematická chyba pipety* 29
1.2.2 Kvantilové odhady chyb 30
Vzorová úloha 1.5 *Kvantilové odhady chyb přístroje* 31
1.2.3 Nepravděpodobnostní intervalové odhady chyb 32
1.3 Šíření chyb a nejistot 32
1.3.1 Metoda Taylorova rozvoje 32
Vzorová úloha 1.6 *Šíření chyb v metodě izotopového zřeďování* 33
Vzorová úloha 1.7 *Korelace chyb objemů v laboratorních operacích* 34
Vzorová úloha 1.8 *Výpočet jemnosti vlákna z hmotností a délek vláken* 35
Vzorová úloha 1.9 *Určení střední hodnoty jemnosti vláken* 36
1.3.2 Metoda dvoubodové aproximace 36
Vzorová úloha 1.10 *Určení chyby viskozity metodou dvoubodové aproximace* 37
1.3.3 Metoda simulací Monte Carlo 37
Vzorová úloha 1.11 *Hromadění chyb při určení rozpustnosti stříbrné soli* 38
Vzorová úloha 1.12 *Korelace v hromadění chyb* 39
1.4 Nejistoty výsledků měření 39
1.4.1 Porovnání přístupů k výpočtu nejistot 40
1.4.2 Kritické poznámky k výpočtu nejistot 42
Vzorová úloha 1.13 *Nejistota aritmetických operací přibližných čísel* 43
1.4.3 Přístup intervalové analýzy k nejistotám 43
Vzorová úloha 1.14 *Výpočet nejistoty teploty měřené rtuťovým teploměrem* 44
1.4.4 Zaokrouhlování čísel 45
Vzorová úloha 1.15 *Zaokrouhlování čísel na 2, 3 a 4 platná místa* 45
1.5 Úlohy 45
1.5.1 Analýza farmakologických a biochemických dat 46
1.5.2 Analýza chemických a fyzikálních dat 46
1.5.3 Analýza environmentálních, potravinářských a zemědělských dat 50
1.5.4 Analýza hutnických a mineralogických dat 51
1.5.5 Analýza fyzikálních dat 51
1.6 Kontrolní hodnoty (použitý software ADSTAT, NCSS2000) 52
1.6.1 Analýza farmakologických a biochemických dat 52
1.6.2 Analýza chemických a fyzikálních dat 52
1.6.3 Analýza environmentálních, potravinářských a zemědělských dat 53
1.6.4 Analýza hutnických a mineralogických dat 53
1.6.5 Analýza fyzikálních dat 53
1.7 Doporučená literatura 53

2 PRŮZKUMOVÁ ANALÝZA JEDNOROZMĚRNÝCH DAT 55
Vzorová úloha 2.1 *Analýza dat normálního a logaritmicko-normálního rozdělení* 56
2.1 Průzkumová (exploratorní) analýza dat EDA 57
2.2 Ověření předpokladů o datech 68
2.3 Transformace dat 71
2.4 Průběh průzkumové analýzy dat 73
Vzorová úloha 2.2 *Průzkumová analýza velkého výběru* 75
2.5 Úlohy 82
2.5.1 Analýza farmakologických a biochemických dat 82
2.5.2 Analýza chemických a fyzikálních dat 90

2.5.3 Analýza environmentálních, potravinářských a zemědělských dat 101
2.5.4 Analýza hutnických a mineralogických dat 108
2.5.5 Analýza ekonomických a sociologických dat 116
2.6 Kontrolní hodnoty (ADSTAT, NCSS2000) 122
2.6.1 Analýza farmakologických a biochemických dat 122
2.6.2 Analýza chemických a fyzikálních dat 124
2.6.3 Analýza environmentálních, potravinářských a zemědělských dat 126
2.6.4 Analýza hutnických a mineralogických dat 127
2.6.5 Analýza ekonomických a sociologických dat 129
2.7 Doporučená literatura 132

3 STATISTICKÁ ANALÝZA JEDNOROZMĚRNÝCH DAT 133
3.1 Bodový odhad parametrů polohy, rozptýlení a tvaru 133
A. Momentové míry polohy a rozptýlení 133
B. Kvantilové a robustní míry polohy a rozptýlení 136
C. Odhady parametrů polohy a rozptýlení pro důležitá rozdělení 140
3.2 Intervalový odhad parametrů polohy a rozptýlení 145
3.3 Analýza malých výběrů 146
3.4 Statistické testování 147
A. Postup testování statistické hypotézy 147
B. Testy střední hodnoty ("testy správnosti") 148
C. Testy shody středních hodnot ("testy shodnosti") 149
Vzorová úloha 3.1 *Analýza velkého výběru* 153
Vzorová úloha 3.2 *Analýza malého výběru* 156
Vzorová úloha 3.3 *Test střední hodnoty (test správnosti)* 158
Vzorová úloha 3.4 *Test shodnosti středních hodnot* 159
Vzorová úloha 3.5 *Párový test* 160
3.5 Úlohy 160
3.5.1 Analýza farmakologických a biochemických dat 161
3.5.2 Analýza chemických a fyzikálních dat 167
3.5.3 Analýza environmentálních, potravinářských a zemědělských dat 178
3.5.4 Analýza hutnických a mineralogických dat 185
3.5.5 Analýza ekonomických a sociologických dat 190
3.6 Kontrolní hodnoty (ADSTAT, NCSS2000) 194
3.6.1 Analýza farmakologických a biochemických dat 194
3.6.2 Analýza chemických a fyzikálních dat 197
3.6.3 Analýza environmentálních, potravinářských a zemědělských dat 203
3.6.4 Analýza hutnických a mineralogických dat 206
3.6.5 Analýza ekonomických a sociologických dat 209
3.7 Doporučená literatura 210

4 STATISTICKÁ ANALÝZA VÍCEROZMĚRNÝCH DAT 213
4.1 Popis vícerozměrných dat 214
4.2 Obecný postup analýzy vícerozměrných dat 216
4.3 Charakteristiky vícerozměrných náhodných veličin 218
Vzorová úloha 4.1 *Popisné charakteristiky vícerozměrných náhodných veličin* 221
4.4 Exploratorní analýza struktury objektů (EDA) 221
4.5 Určení struktury a vazeb v proměnných a objektech 227
4.5.1 Analýza hlavních komponent (PCA) 228
Vzorová úloha 4.2 *Postup metody hlavních komponent* 235
4.5.2 Faktorová analýza (FA) 240
Vzorová úloha 4.3 *Vyčíslení faktorů z korelační matice* 241
Vzorová úloha 4.4 *Ukázka pojmů a podstaty faktorové analýzy* 242
4.5.3 Kanonická korelační analýza 245
Vzorová úloha 4.5 *Ukázka pojmů a podstaty kanonické korelační analýzy* 246
Vzorová úloha 4.6 *Postup kanonické korelační analýzy* 248
4.6 Klasifikace objektů 252

4.6.1 Diskriminační analýza DA 252
Vzorová úloha 4.7 *Užití lineární diskriminační funkce* 254
Vzorová úloha 4.8 *Užití logistické diskriminace* 258
Vzorová úloha 4.9 *Užití postupu diskriminační analýzy* 260
4.6.2 Analýza shluků CLU 269
(a) Hierarchické shlukování 271
Vzorová úloha 4.10 *Nalezení shluků hráčů podobných vlastností* 272
Vzorová úloha 4.11 *Vytvoření dendrogramu objektů neuroleptika* 277
(b) Shlukování metodou nejbližších středů (K-Means) 279
Vzorová úloha 4.12 *Klasifikace objektů do shluků* 280
(c) Shlukování metodou středů-medoidů 283
Vzorová úloha 4.13 *Odhalení struktury objektů rozličnými metodami shlukování* 285
(d) Fuzzy shlukování 289
Vzorová úloha 4.14 *Klasifikace objektů barev fuzzy shlukováním* 290
4.7 Vícerozměrné škálování MDS 292
Vzorová úloha 4.15 *Vícerozměrné škálování u analýzy podobnosti* 295
4.8 Vícerozměrná kalibrace (V. Centner) 296
4.8.1 Klasická vícerozměrná kalibrace 297
4.8.2 Inverzní vícerozměrná kalibrace 298
Vzorová úloha 4.16 *Postup vícerozměrné kalibrace* 306
4.9 Úlohy 315
4.9.1 Analýza farmakologických a biochemických dat 316
4.9.2 Analýza chemických a fyzikálních dat 324
4.9.3 Analýza environmetálních, potravinářských a zemědělských dat 330
4.9.4 Analýza hutnických a mineralogických dat 340
4.9.5 Analýza ekonomických a sociologických dat 344
4.10 Doporučená literatura 350

5 ANALÝZA ROZPTYLU 353
5.1 Jednofaktorová analýza rozptylu 353
Vzorová úloha 5.1 *Zkrácený postup jednofaktorové analýzy rozptylu* 360
Vzorová úloha 5.2 *Podrobný postup v jednofaktorové analýze rozptylu* 362
5.2 Dvoufaktorová analýza rozptylu bez opakování v cele 368
Vzorová úloha 5.3 *Dvoufaktorová analýza rozptylu bez opakovaní* 371
5.3 Vyvážená dvoufaktorová analýza rozptylu 377
Vzorová úloha 5.4 *Vyvážená dvoufaktorová analýza rozptylu* 380
5.4 Nevyvážená dvoufaktorová analýza rozptylu 382
Vzorová úloha 5.5 *Nevyvážená dvoufaktorová analýza rozptylu* 384
5.5 Opakovatelnost a reprodukovatelnost (O&R analýza) 385
Vzorová úloha 5.6 *Schéma O&R analýzy* 386
5.6 Úlohy 392
5.6.1 Analýza farmakologických a biochemických dat 393
5.6.2 Analýza chemických a fyzikálních dat 401
5.6.3 Analýza environmetálních, potravinářských a zemědělských dat 408
5.6.4 Analýza hutnických a mineralogických dat 413
5.6.5 Analýza ekonomických a sociologických dat 420
5.7 Kontrolní hodnoty (ADSTAT, NCSS2000) 427
5.7.1 Analýza farmakologických a biochemických dat 427
5.7.2 Analýza chemických a fyzikálních dat 427
5.7.3 Analýza environmetálních, potravinářských a zemědělských dat 428
5.7.4 Analýza hutnických a mineralogických dat 428
5.7.5 Analýza ekonomických a sociologických dat 429
5.8 Doporučená literatura 429

6 LINEÁRNÍ REGRESNÍ MODELY 431
6.1 Jednorozměrné lineární regresní modely 437
Vzorová úloha 6.1 *Postup výstavby modelu a regresní diagnostika* 437

6.1.1 Úlohy na jednorozměrné lineární regresní modely 447
6.2 Validace nové analytické metody 455
Vzorová úloha 6.2 *Postup validace a regresní diagnostika* 455
6.2.1 Úlohy na validaci nové analytické metody 465
6.3 Lineární a nelineární kalibrace 472
Vzorová úloha 6.3 *Postup nelineární kalibrace spline funkcí* 473
6.3.1 Úlohy na lineární a nelineární kalibraci 475
6.4 Polynomické regresní modely 492
Vzorová úloha 6.4 *Optimální stupeň polynomu a snížení multikolinearity* 492
6.4.1 Úlohy na polynomické regresní modely 506
6.5 Vícerozměrné lineární regresní modely 514
Vzorová úloha 6.5 *Regresní triplet ve výstavbě vícerozměrného regresního modelu.* 514
6.5.1 Úlohy na vícerozměrné lineární regresní modely 530
6.6 Kontrolní hodnoty (ADSTAT, NCSS2000) 543
6.6.1 Jednorozměrné lineární regresní modely 543
6.6.2 Validace nové analytické metody 544
6.6.3 Úlohy na lineární a nelineární kalibraci 545
6.6.4 Úlohy na polynomické regresní modely 546
6.6.5 Vícerozměrné lineární regresní modely 547
6.7 Doporučená literatura 549

7 KORELACE 553
7.1 Druhy korelačních koeficientů 553
7.1.1 Párový korelační koeficient 553
7.1.2 Parciální korelační koeficient 554
7.1.3 Vícenásobný korelační koeficient 556
7.2 Pořadový korelační koeficient 556
7.3 Cronbachův korelační koeficient γ spolehlivosti výsledku 557
Vzorová úloha 7.1 *Postup vyšetření korelace* 559
7.4 Úlohy na korelaci 561
7.4.1 Analýza farmakologických a biochemických dat 561
7.4.2 Analýza chemických a fyzikálních dat 565
7.4.3 Analýza environmetálních, potravinářských a zemědělských dat 568
7.4.4 Analýza hutnických a mineralogických dat 571
7.4.5 Analýza ekonomických a sociologických dat 572
7.5 Kontrolní hodnoty (ADSTAT, NCSS2000) 576
7.5.1 Analýza farmakologických a biochemických dat 576
7.5.2 Analýza chemických a fyzikálních dat 577
7.5.3 Analýza environmetálních, potravinářských a zemědělských dat 577
7.5.4 Analýza hutnických a mineralogických dat 578
7.5.5 Analýza ekonomických a sociologických dat 578
7.6 Doporučená literatura 578

8 NELINEÁRNÍ REGRESNÍ MODELY 579
8.1 Tvorba nelineárního regresního modelu 579
Vzorová úloha 8.1 *Odhad tří parametrů rozšířeného Debyeova-Hückelova vztahu* 583
8.2 Úlohy 586
8.2.1 Analýza farmakologických a biochemických dat 587
8.2.2 Analýza chemických a fyzikálních dat 588
8.2.3 Analýza environmetálních, potravinářských a zemědělských dat 597
8.2.4 Analýza hutnických a mineralogických dat 603
8.2.5 Analýza matematických modelů a fyzikálních dat 603
8.3 Kontrolní hodnoty (ADSTAT, NCSS2000) 609
8.3.1 Analýza farmakologických a biochemických dat 609
8.3.2 Analýza chemických a fyzikálních dat 609
8.3.3 Analýza environmetálních, potravinářských a zemědělských dat 610
8.3.4 Analýza hutnických a mineralogických dat 611

8.3.5 Analýza matematických modelů a fyzikálních dat . . . 611
8.4 Doporučená literatura . . . 612

9 INTERPOLACE A APROXIMACE . . . 615
9.1 Klasické interpolační postupy . . . 616
Vzorová úloha 9.1 *Náhrada funkce exp(x)* . . . 617
9.1.1 Lagrangeova a Newtonova interpolační formule . . . 617
Vzorová úloha 9.2 *Náhrada funkce exp(x)* . . . 621
Vzorová úloha 9.3 *Aproximace racionální funkce* . . . 621
9.1.2 Hermitovská interpolace . . . 622
Vzorová úloha 9.4 *Hermitovská interpolace funkce exp(x)* . . . 623
9.1.3 Racionální interpolace . . . 623
Vzorová úloha 9.5 *Racionální interpolace funkce exp(x)* . . . 624
9.2 Spline interpolace . . . 625
Vzorová úloha 9.6 *Lineární B-spline* . . . 627
9.2.1 Lokální Hermitovská interpolace . . . 629
Vzorová úloha 9.7 *Lokální kubická interpolace stupňovité závislosti* . . . 631
Vzorová úloha 9.8 *Akimova interpolace schodovité závislosti* . . . 633
9.2.2 Kubické spline . . . 633
Vzorová úloha 9.9 *Spline interpolace schodovité závislosti* . . . 635
Vzorová úloha 9.10 *Interpolace pomocí spline pod napětím* . . . 637
9.3 Aproximace funkcí . . . 638
Vzorová úloha 9.11 *Aproximace funkce exp(x)* . . . 641
9.4 Aproximace tabelárních závislostí . . . 642
9.4.1 Polynomická aproximace . . . 642
Vzorová úloha 9.12 *Čebyševova aproximace funkce exp(x)* . . . 643
Vzorová úloha 9.13 *Hledání nejlepšího poměru polynomů* . . . 644
9.4.2 Úseková regrese . . . 647
Vzorová úloha 9.14 *Aproximace píku* . . . 650
Vzorová úloha 9.15 *Aplikace postupu úsekové polynomické regrese* . . . 654
Vzorová úloha 9.16 *Určení bodu ekvivalence u dvou větví titrační křivky* . . . 656
9.5 Numerické vyhlazování . . . 658
9.5.1 Spline vyhlazování . . . 659
Vzorová úloha 9.17 *Vyhlazování píku algoritmem SPÄTH* . . . 663
Vzorová úloha 9.18 *Vyhlazování píku algoritmem REINSCH* . . . 664
Vzorová úloha 9.19 *Optimální vyhlazení píku* . . . 667
9.5.2 Neparametrická regrese . . . 668
Vzorová úloha 9.20 *Neparametrická regrese píku* . . . 669
9.5.3 Číslicová filtrace . . . 670
Vzorová úloha 9.21 *Porovnání vlastností lineárních a nelineárních filtrů* . . . 672
Vzorová úloha 9.22 *Vliv délky regresního filtru na vyhlazující vlastnosti* . . . 676
Vzorová úloha 9.23 *Filtrace absorpčního spektra fenolové červeně* . . . 677
Vzorová úloha 9.24 *Výpočet hustoty kyseliny fosforečné* . . . 678
Vzorová úloha 9.25 *Určení chybějící hodnoty v infračerveném spektru* . . . 679
9.6 Postup při interpolaci a aproximaci . . . 679
9.7 Úlohy . . . 680
9.7.1 Analýza chemických a fyzikálních dat . . . 681
9.7.2 Analýza ekonomických a ostatních dat . . . 683
9.8 Doporučená literatura . . . 685

10 KONTROLA A ŘÍZENÍ JAKOSTI . . . 687
10.1 Podstata úloh řízení jakosti . . . 687
10.2 Regulační diagramy . . . 694
10.2.1 Regulační diagramy pro dílčí výběry . . . 695
10.2.2 Regulační diagramy typu "x s pruhem" . . . 695
10.2.3 Regulační diagramy pro posouzení variability . . . 700
10.2.4 Regulační diagramy kumulativních součtů, CUSUM . . . 701

10.2.5 Regulační diagramy na bázi lokálního vyhlazení 705
10.2.6 Regulační diagramy pro jednotlivé hodnoty 707
10.2.7 Regulační diagramy pro distrétní znaky 709
10.2.8 Regulační diagramy pro více proměnných 711
10.2.9 Používání regulačních diagramů (K. Kupka) 714
10.2.10 Konstrukce regulačních diagramů (K. Kupka) 715
Vzorová úloha 10.1 *Aplikace regulačního diagramu pro průměry a směrodatné odchylky* . . 715
Vzorová úloha 10.2 *Aplikace diagramu R* 717
10.2.11 Pravidla pro určování zvláštních případů 720
Vzorová úloha 10.3 *Aplikace regulačního diagramu pro jednotlivé hodnoty x* 722
10.2.12 Porušení předpokladů o datech 723
10.2.13 Pomůcky diagramů kumulativních součtů CUSUM 727
Vzorová úloha 10.4 *Lucasova modifikace regulačního diagramu CUSUM* 729
Vzorová úloha 10.5 *Aplikace diagramů EWMA* 730
Vzorová úloha 10.6 *Kontrola tavby v metalurgickém provozu regulačním diagramem* 732
Vzorová úloha 10.7 *Chemická analýza složení plynných splodin* 733
Vzorová úloha 10.8 *Aplikace Hotellingova regulačního diagramu* 735
10.3 Indexy způsobilosti procesu 738
10.4 Software pro řízení jakosti 741
10.5 Úlohy 745
10.6 Kontrolní výsledky (ADSTAT, NCSS2000) 751
10.7 Doporučená literatura 752
Rejstřík 755

Předmluva autorů

Aplikace počítačově orientovaných statistických metod v přírodovědeckých, lékařských, technických a gnozeologických oborech umožňuje nejen využití informací skrytých v datech a zobecnění získané kombinací výsledků z různých zdrojů ale také tvorbu modelů, optimalizaci a případné řízení. Jde o multidisciplinární směr na pomezí vědních disciplín statistiky a informatiky, který vedl ke vzniku nových oborů chemometrie, biometrie, psychometrie, ekonometrie, technometrie a další.

Zaměření knihy

Statistická analýza dat nabývá na stále větším významu a stává se nezbytným přístupem v řadě přírodovědných, lékařských, technických a sociálních věd. Na rozdíl od klasické matematické statistiky je zde kladen důraz nikoliv na konkrétní metody ale na jejich vhodné kombinace umožňující posouzení kvality dat a výběr vhodného statistického modelu včetně jeho ověření. Rozhodující je vždy cíl zpracování dat a úroveň znalostí o řešených problémech, které ovlivňují postup analýzy a volbu použité metody. To se projevuje zejména při statistické analýze příkladů dat z praxe. Na jedné straně je samostatné diagnostikování a odkrývání tajů skrytých v datech účinný způsob studia. Řešení zajímavých a nevšedních praktických úloh umožňuje lépe pochopit meze a možnosti různých metod a na základě analogie volit způsob zpracování vlastních úloh. Na druhé straně však u řady příkladů dat z praxe obvykle chybí cíl zpracování, který bývá definován jen vágně a nejsou uvedeny znalosti o dosavadním stavu poznání, které zná často pouze omezený okruh specialistů. Praktické úlohy analýzy dat obsahují také řadu nečekaných překážek:

- rozsahy zpracovávaných dat nejsou obyčejně velké a z důvodů čistě registračních jsou obsaženy i proměnné nevýznamné pro vlastní řešení (redundantní informace),
- v datech existují výrazné nelinearity, multikolinearita, neaditivity a skryté vazby, anomálie a vlivné body, které komplikují analýzu dat,
- rozdělení experimentálních dat jen zřídka odpovídá normálnímu,
- v datech se vyskytují vybočující měření resp. vysoká heterogenita,
- parametry hledaných modelů mají definovaný fyzikální význam a jejich numerické odhady musí proto vyhovovat jak svou velikostí tak i znaménkem.

Při řešení úloh v této knize byly voleny typové způsoby zpracování zaměřené především na statistické zvláštnosti dat. Jde o průzkumovou analýzu dat a ověřování základních předpokladů o datech s ohledem na základní schéma "data-model-statistická metoda". Konstrukce statistických modelů je zde spíše ilustrací a interpretace výsledků je pouze obecná. Odborníci v daném vědním oboru budou zřejmě schopni v řadě případů doporučit

i alternativní cíle řešení respektive alternativní interpretaci výsledků.

Okruh čtenářů

Předložené kompendium představuje kolekci metod s 1100 úlohami počítačového zpracování experimentálních dat. Je určeno především studentům vysokých škol technického, přírodovědného, ale i humanitního a ekonomického směru. Snadno poslouží čtenáři i k samostudiu. Představuje návod, dle kterého si každý může vyhodnotit také svá data. Neobsahuje definice, věty a důkazy matematického pojetí procvičované látky. Kompendium bylo sepsáno tak, aby bylo srozumitelné i středoškolsky vzdělaným čtenářům. Cílem je dát návod k logickému vyhodnocení dat každodenní praxe. Spolu s doporučeným statistickým softwarem tj. s využitím některého z programů jako ADSTAT, QC-Expert, NCSS2004, STATGRAPHICS, STATISTICA, MINITAB, SCAN, S-Plus a dalších, tvoří účinnou pomůcku analýzy experimentálních dat v laboratořích výzkumných ústavů, podnikových a státních kontrolních laboratořích a zkušebnách kontroly kvality. Kompendium je také vítanou pomůckou manažerů zdravotnických, veterinárních a vodohospodářských laboratoří, potravinářské a zemědělské inspekce, chemické a farmaceutické výroby. Uplatnění najde i u pracovníků kontroly životního prostředí všech odvětví průmyslu, energetiky a zemědělství, u technologů, pracovníků řízení jakosti, a především u vedoucích pracovníků. S daty na doprovodném CD je Kompendium bezprostředně využitelné v chemické laboratoři čili v chemometrii, ale také v ekonomice a ekonometrii, sociologii, medicíně, biologii a při sledování kvality životního prostředí. Je určeno lidem, které láká počítačově podporované odkrývání informací uložených v experimentálních datech.

Struktura kapitol

Úlohy jsou částečně převzaty z literatury a internetu avšak naprostá většina jich pochází z laboratorní a průmyslové praxe účastníků postgraduálního studia statistického zpracování dat shromážděných v posledních 15 letech na Univerzitě Pardubice. Úlohy se vesměs týkají netradičního pojetí statistické analýzy dat, které je založeno na *interaktivní analýze*.

Kompendium je organizováno podle jednotného schématu: v první části každé kapitoly je stručně pojednáno o technice analýzy dat. Tam, kde jde o klasické a známé statistické postupy, jde pouze o stručný úvod do řešené problematiky. U méně známých postupů je pak výklad podrobnější, aby postačoval k pochopení výstupů vzorových úloh. Řešení vzorových úloh včetně detailního postupu z výpočetních programů je doplněné také jejich interpretací. Poslední částí každé kapitoly jsou praktické úlohy z různých vědních oborů s uvedením mezivýsledků, indikujících správnost postupu z hlediska především numerického. Jednotlivé kapitoly lze studovat samostatně, i když členění kapitol je zvoleno tak, aby se čtenář postupně seznamoval se všemi ve statistice navazujícími technikami.

Je třeba zdůraznit, že Kompendium se týká především kolekci praktických úloh a problémů každodenní laboratorní praxe, zásadně odlišných od zjednodušených školních úloh v jiných učebnicích. Cílem je zde naučit čtenáře vyšetřovat praktická data i v případech, kdy výpočty nevedou k jednoznačným závěrům. V každé kapitole jsou úlohy rozděleny dle svého obsahového zaměření do pěti vědních skupin: úlohy s kódem **B** obsahují *biochemická, biologická a lékařská data,* **C** *chemická a fyzikální data,* **E** *environmentální, potravinářská a zemědělská data,* **H** *hutní, metalurgická a mineralogická data*, **S** *sociologická, ekonomická a ostatní data.* Pouze v 6. kapitole (lineární regresní

modely) je třídění úloh jiné: úlohy s kódem **J** obsahují *data jednoduchých lineárních regresních modelů*, **V** *data k validaci nové analytické metody,* **L** *data k polynomům*, **K** *data pro kalibraci,* **M** *data pro vícerozměrné lineární regresní modely.*

Postup analýzy dat

Jádrem výkladu každé metody je rozbor vzorové úlohy. Za principem metody následuje výklad vzorové úlohy se strukturou: 1. *Zadání:* formulace problému a cíle řešení. 2. *Data:* zadání vstupních dat. 3. *Program:* volba vhodného počítačového programu. 4. *Řešení:* detailní popis výstupu z počítače a vysvětlení výsledků. 5. *Závěr:* formulace dosažených závěrů ze získaných výsledků.

V první kapitole je k praktickému procvičování uvedeno 31 úloh, v druhé 157 úloh, ve třetí 208 úloh, ve čtvrté 125 úloh, v páté 151 úloh, v šesté 260 úloh, v sedmé 77 úloh, v osmé 63 úloh, v deváté 13 úloh a v desáté 13 úloh celkem 1098 úloh. Ke všem úlohám jsou k dispozici úplná data na doprovodném kompaktním disku, zatímco v knize jsou z úsporných důvodů data zkrácená na první a poslední řádek. Pokud by čtenáři nestačil stručný popis užité statistické metody, najde detailní výklad matematického pozadí v učebnici M. Meloun, J. Militký: STATISTICKÉ ZPRACOVÁNÍ EXPERIMENTÁLNÍCH DAT, Plus, Praha 1994, nebo East Publishing, Praha 1998 nebo v posledním vydání Academia Praha 2004. Náročné regresní metody (6. a 8. kapitola) jsou v Kompendiu pro svou rozsáhlost uvedeny velice stručně a čtenář je odkázán na uvedenou učebnici.

K řešení příkladů lze použít řady různých programových systémů. Řada z nich však neobsahuje všechny v citované učebnici probírané a potřebné metody a neužívá ani testování regresního tripletu "data - model - statistická metoda". Námi navržený postup komplexnější statistické analýzy dat je proto standardně užíván ve statistickém balíku ADSTAT. Statistická analýza dat je iterativní proces, kdy výsledkem bývá návrh dalšího zkoumání a případně i návrh nových experimentů. Tento proces však nebylo možné v Kompendiu v plném rozsahu ukázat. Snahou autorů bylo uvést alespoň výsledky jednoho typického kroku takové statistické analýzy, které jsou zajímavé a nevyžadují ani externí informace ani detailní znalost oboru. Tento *datově orientovaný postup* má však nevýhodu v tom, že informace v datech mohou být neúplné, a proto nepostačující pro statistické závěry s požadovanou mírou nejistoty. U řady převzatých příkladů nebylo také možné ověřit kvalitu měřícího procesu ani správnost měření a přístroje. Některé závěry souvisejí proto pouze s uvedenými a řešenými daty.

Zadávání dat

V laboratoři získáváme data přímo z experimentu ve formě databáze, tabulky nebo ASCII souboru. Z dat se snažíme získat maximum užitečné informace a odhalit jejich specifické rysy. Na doprovodném kompaktním disku jsou data k jednotlivým kapitolám jednak v ASCII kódu a jednak ve formě MS Excel. Většina statistického software je schopna načíst data přímo či importovat data v excelovské podobě. Data lze také "přenést" z MS Excelu do libovolného software kopírováním přes schránku (clipboard).

Úlohy jsou v knize a na disku CD číslovány jednoduchým způsobem: příkladně označení úlohy ***C201a*** znamená první úlohu ve druhé kapitole, kde **C** značí *chemická a fyzikální data* s použitou částí dat *a*. Takto deklarovaný výběr dat představuje jediný sloupec v tabulce MS Excelu, uvedený záhlavím ***C201a*** a prvky tohoto sloupce jsou pak jednotlivé řádky.

(1) Analýza jednorozměrného výběru značí analýzu jednoho sloupce: zadáme proto sloupec

čísel ***C201a***.

(2) Analýza matice dat ve vícerozměrné analýze dat nebo v regresi: obdélníková matice obsahuje stejně dlouhé sloupce prvků, představujících *znaky* (čili *proměnné*), zatímco řádky v této matici představují *objekty*. Data zadáváme po sloupcích, např. sloupec ***B402i*** značí index objektu 2. úlohy čtvrté kapitoly biochemických dat, ***B402x1*** značí první znak x_1 této úlohy, ***B402x2*** druhý znak x_2, ***B402x3*** třetí znak x_3 atd.

(3) Matice u vícefaktorové analýzy rozptylu, ANOVA: prvky matice u dvoufaktorové ANOVA (či 3D-pole u trojfaktorové) je třeba popsat jejich faktory. Sloupce ***B526f1, B526f2*** představují faktorové souřadnice prvků y a ***B526y*** hodnoty prvků sloupce y.

Diagnóza analýzou dat

Diagnóza interaktivní analýzou dat znamená hledání všech souvislostí a zvláštností ukrytých v datech. To je standardním přístupem bez počítačové podpory nemožné. Jde především o zkoumání těchto problémů:

(1) Skryté problémové údaje, jako např. hrubé chyby, systematické chyby, odlehlé hodnoty nebo extrémy, kdy musíme rozhodnout, zda problémové hodnoty z analýzy odstraníme, opravíme, či je ponecháme beze změny a zvolíme k další statistické analýze metody méně citlivé, tj. robustní. Efektivní analýza dat musí předpokládat i zvláštní hodnoty, které jsou v datech velmi vlivné. Vlivným bodům je třeba věnovat zvláštní péči již při sběru dat.

(2) Závislost dat, kdy jsou prvky analyzovaného výběru spojeny skrytými vztahy a nebyly získány nezávisle na sobě. To pochopitelně připouští jiné než pouze náhodné ovlivnění člověkem, přístrojem nebo postupem odběru dat. Oddělení vlivu jednotlivých proměnných lze chápat jako základní problém regrese a vícerozměrných metod.

(3) Chybějící data,kdy tabulka dat vykazuje "díry". Data je pak třeba analyzovat tak, aby přes tento nedostatek bylo využito a získáno maximum informace.

(4) Zvláštnosti dat se týkají symetrie rozdělení, typu rozdělení, lokální koncentrace dat, homogenity dat, anomálií a velikosti šumu.

(5) Malý počet dat ovlivňuje statistické závěry. Když se v laboratoři získá málo dat, je třeba vždy nejdříve doporučit především další měření. Existují také metody (např. Hornův postup), které za cenu mírné ztráty informací poskytnou řešení i v takovýchto případech menších výběrů.

Kontrolní hodnoty u úloh

Kontrolní hodnoty řešeních úloh jsou z pedagogických důvodů uvedeny u většiny úloh kromě těch, kde je řešení buď vícestupňové, nebo kde se na základě získaných informací provádí výběr metody či následuje složitá úprava dat nebo modelu. Kontrolní hodnoty nejsou uváděny tam, kde jsou výstupy značně obsáhlé nebo grafické (např. vícerozměrná data, interpolace), resp. závislé na použitém softwaru (např. nelineární regresi) nebo kde jsou výsledky komplexního charakteru (např. metody vícerozměrné analýzy dat, interpolace a aproximace). Čtenář by je měl chápat jako kontrolní mezivýsledky a měl by se ve svém řešení dopracovat buď přímo těchto nebo podobných a lepších odhadů.

Doprovodný disk CD

K usnadnění procvičování úloh je ke sbírce přiložen kompaktní disk, který obsahuje data ke všem úlohám. Vstupní data jsou jednak v ASCII kódu a jednak v matici pro MS Excel. Z

něho se pak dají importovat data do většiny programů. Data jsou dostupná i na internetu na osobní stránce autorů *http://meloun.upce.cz*.

Poděkování

Není možné poděkovat všem spolupracovníkům, studentům licenčního studia statistického zpracování dat a doktorandům, kteří nám pomáhali či přispěli praktickými úlohami, radami či konstruktivní kritikou. Zvláštní dík patří lektorům za celkové přehlédnutí rukopisu a řadu podnětných praktických připomínek a výpočtů. Výroba této knihy byla umožněna finanční podporou vědeckého záměru Ministerstva školství mládeže a tělovýchovy č. MSM253100002, a projektu 1M4674788501 Ministerstva školství mládeže a tělovýchovy za což autoři vyslovují svůj dík.

Prosinec 2005.

Milan Meloun
Jiří Militký

O autorech

MILAN MELOUN (*1943), prof. RNDr. DrSc. Ke studiu chemie byl přiveden už na střední škole znamenitým pedagogem RNDr. Františkem Březinou, který ho dovedl pro obor nadchnout natolik, že se mu věnuje celý dosavadní život. Vystudoval přírodovědeckou fakultu Univerzity J. E. Purkyně (dnešní Masarykova) v Brně 1965. Celá jeho vědecká kariéra vyústila do současné pozice profesora analytické chemie a chemometrie na katedře analytické chemie Univerzity Pardubice. Vedle odborné práce se věnoval cizím jazykům, což mu umožnilo udržovat krok se světovou vědou a navíc i později pedagogicky působit dva roky na Bagdádské univerzitě v Iráku a tři semestry na Královské technice ve Stockholmu. Už v počátcích své profesní kariéry pochopil význam a trend kompjuterizace a se vší vervou prosazoval moderní metody zpracování a vyhodnocování chemických dat. Nezůstalo jen u chemie čili chemometrie ale všechny ty metrie se prosazují bez nadsázky v experimentálních i na první pohled v iracionálních disciplinách. Může se považovat za jednoho z hlavních "prosazovatelů" chemometrie u nás, o čemž svědčí jeho rozsáhlá publikační aktivita, přijímaná v prestižních světových vědeckých časopisech nemluvě o vlastní škole chemometrie, doložené vědeckými učebnicemi a monografiemi. Je autorem a spoluautorem 150 originálních sdělení, 16 monografií a 12 vysokoškolských učebnic, řady patentů a zlepšovacích návrhů a na konferencích přednesl více než 200 přednášek. Je členem redakčních rad zahraničních odborných časopisů tajemníkem sekce Chemometrie při České společnosti chemické.

Jeho publikace jsou zaměřeny především na počítačovou analýzu instrumentálních dat. Knižně se uvedl spolu s Josefem Havlem dvoudílnou monografií *Computation of Solution Equilibria*, která po doplnění o extrakční rovnováhy vyšla v roce 1988 v anglickém nakladatelství Ellis Horwood, Chichester. V jeho práci představuje počítač spojovací článek mezi statistikou a analytickou chemií. Výsledkem je analytická chemometrie, předmět, který přednáší od roku 1978 podle svých učebnic. V Bagdádu napsal učebnici *Data Analysis by Statistical and Computing Technique*, University Baghdad Press, 1980, a na Královské

technice ve Stockholmu pak sbírku příkladů *Introduction to Chemometrics*, která obsahuje analýzu dat programovým systémem STATGRAPHICS. Vlastní pojetí analýzy experimentálních dat se promítá i do kapitoly *Chemometrics in the Instrumental Laboratory* uvedené v monografii, editované Jaroslavem Churáčkem *Advanced Instrumental Methods of Chemical Analysis*, Academia, Praha 1993, nebo v kapitole *Hodnocení analytických výsledků* ve Vláčilově sbírce *Příklady z chemické a instrumentální analýzy*, SNTL, Praha 1983.

Milan Meloun pochopil, že vlastní bádání by nepřineslo kýžený účinek, nebylo-li by užitečné i v běžné praxi, ať už v průmyslu, aplikovaném výzkumu a takových disciplinách , které dlouho naštěstí marně odolávaly náporu kompjuterizace. Praktickým využíváním metodologie statistického zpracování dat se tak otevřely mnohdy tajemstvím zahalené "šuplíky" uložených dat, které najednou odhalují dříve netušené zákonitosti a souvislosti. Zkušenosti spoluautora Jiřího Militkého přinesly řadu postupů ze statistické analýzy dat a především interaktivní přístup k analýze dat na osobním počítači. Společně tak vzniklo první vydání učebnice *Chemometrie - Zpracování experimentálních dat na IBM PC.* Text byl přeložen do angličtiny a po doplnění o kapitoly vícerozměrné statistiky Michelem Forinou z italského Janova vyšel postupně jako dvojdílná učebnice u nakladatele Ellis Horwood, Chichester 1991, pod titulem Milan Meloun, Jiří Militký a Michele Forina: *Chemometrics for Analytical Chemistry - Volume I. PC-Aided Statistical Data Analysis,* a *Volume II. PC-Aided Regression and Related Methods.* V českém překladu vyšla pod titulem Milan Meloun a Jiří Militký: *Statistické zpracování experimentálních dat,* v několika vydáních v nakladatelstvích PLUS Praha 1994, East Publishing Praha 1998 a Academia Praha 2004. Oba autoři vydali pro studenty také cvičebnici 800 počítačem podporovaných úloh *Kompendium statistického zpracování dat* v nakladatelství Academia Praha 2002 a poslední době monografii *Počítačová analýza vícerozměrných dat v příkladech*, Academia Praha 2005. Autoři považují za svůj největší přínos pro čtenáře a uživatele právě radost z objevování systému, vztahů a tajemství v člověkem nashromážděných údajích a zúročení vynaloženého úsilí na získání experimentálních dat. Ukazuje se, že praktická statistika není strašák ale pomocník v rukou poučeného uživatele.

Na Univerzitě v Pardubicích přednáší Milan Meloun statistické zpracování experimentálních dat, práci s daty v tabulkových procesorech, organizuje dvouleté licenční studium a krátkodobé intenzivní kurzy chemometrie pro aplikaci v průmyslu. Do kolekce počítačových znalostí každého studenta přírodních nebo technických věd patří vedle zpracování experimentálních dat i zvládnutí solidního textového editoru a tabulkového procesoru. Mezeru v literatuře ve své době vyplnily dvě učebnice, spíše sbírky řešených úloh, velice rozšířeného textového editoru WordPerfect (*Učíme se WordPerfect 5.1,* ELVIRA, Praha 1993*,* a druhá *Učíme se WordPerfect 6.0*, ELVIRA, Praha 1994) a tabulkového procesoru QUATTRO (*Učíme se QUATTRO PRO 3.0,* VŠCHT, Pardubice 1992 a *Učíme se QUATTRO PRO 4.0 CZ,* Multisys, Pardubice 1993). Oblibu těchto učebnic přinesl jejich styl a účinnost, vyučují totiž software formou řešených příkladů. Po nastudování desítek, stovek povinných příkladů a vyřešení velkého počtu úloh zvládne student ve velmi krátké době aktivně potřebnou látku a je dokonale připraven k užívání softwaru.

JIŘÍ MILITKÝ (*1949), prof. Ing. CSc., ukončil fakultu textilní, specializace textilní chemie na VŠST v Liberci roku 1973. V letech 1974 až 1976 pracoval ve Státním výzkumném ústavu textilním Liberec v oddělení matematického modelování textilních struktur. V letech 1976 až 1989 pracoval ve Výzkumném ústavu zušlechtovacím Dvůr Králové n. L., kde se věnoval převážně zpracování experimentálních dat s využitím výpočetní techniky. Od roku 1990 je vedoucím katedry textilních materiálů na Technické univerzitě Liberec. V roce 1993 byl jmenován řádným profesorem. Je členem několika vědeckých a odborných společností, The Textile Institute, JČMF, FEANI. Pracuje ve výboru sekce Chemometrie při České společnosti chemické . Je akademikem ukrajinské akademie inženýrských věd a nositelem Smithovy medaile (The Textile Institute Manchester) za vědecký přínos v oblasti textilních vláken.

Jeho publikační činnost zahrnuje oblasti textilního inženýrství, modelování kinetických procesů v pevné fázi a zpracování experimentálních dat. Je autorem nebo spoluautorem 606 vědeckých příspěvků (publikací, monografií, referátů a posterů). Jeho první kniha *Modifikovaná PES vlákna* (spoluatoři Jiří Kryštůfek, Jiří Vaníček a Oldřich Hartych) vyšla v SNTL v roce 1984. Zcela přepracované a rozšířené vydání bylo publikováno nakladatelstvím Elsevier v roce 1991. S Jiřím Kryštůfkem zpracoval knihu *Barvení akrylových vláken a směsí*, která vyšla v SNTL Praha v roce 1987. Ve spolupráci s Milanem Melounem publikoval učebnice a monografie z oblasti využití interaktivních statistických metod v chemometrii. Jiří Militký publikoval celkem 12 knih, z nichž šest je zaměřeno do oblasti zpracování experimentálních dat s využitím výpočetní techniky. Moderní metody interaktivní statistické analýzy dat zpracoval do rozsáhlého seriálu příruček *Statistické metody v textilní praxi I - IV*, vydaného v letech 1982 až 1985 v Domě techniky Pardubice. Přehled metod regrese a matematického modelování publikoval v seriálu skript *Tvorba matematických modelů I - VI*, vydaných v letech 1983 až 1989 v Domě techniky Ostrava. Vytvořil systém programů pro zpracování experimentálních dat v jazyce HPL. Tyto programy jsou charakteristické tím, že kromě stránky *statistické* vycházející vždy nejdříve z ověřování předpokladů o modelech, datech a použité metodě, využívají také progresivních *numerických* postupů (zejména v oblasti lineární a nelineární regrese). Tyto algoritmy se později staly jádrem originálního programového systému ADSTAT.

Prezentoval příspěvky na řadě konferencí o *počítačové statistice* (Edinburg, Řím, Kodaň, Dubrovník, Vídeň, Neuchatel, Tampere atd.), *chemometrii* (Montreal, Boloňa, Taormina atd.) a *souvisejících vědních disciplínách* (Nice, Perugia, Ithaca, Honolulu, Kyoto, Mt Fuji, Interlaken, Bukurešt, Lodž, Budapešt, Stockholm, Norimberk, Hakone, Bolton, Espoo, Melbourne, Hong Kong, Shanghai, Hannoi, Teneriffe, Madeira atd.).
Je organizátorem řady odborných mezinárodních konferencí a členem redakčních rad několika časopisů.

KAREL KUPKA (*1962), Ing., PhD, vystudoval specializaci analytické a fyzikální chemie na VŠCHT Pardubice a později obhájil doktorát z analytické chemie. V roce 1990 byl spoluzakladatelem společnosti pro statistickou analýzu a software TriloByte, později TriloByte Statistical Software, kde pracuje dodnes. Je spoluautorem statistického systému ADSTAT a autorem nového statistického systému QC-Expert (pro MS Windows). Jeho tvůrčí činnost je zaměřena především na pokročilé metody analýzy dat a aplikace pro výzkum a řízení jakosti v technologii. Vedle řady odborných článků a knihy *Statistické řízení jakosti* je autorem příspěvků a přednášek na řadě konferencí v ČR, SR, Evropě, USA, Japonsku, Austrálii a v roce 2003 jako hostující lektor chemometrie a statistických metod na Wits University v Johannesburgu. Od roku 2000 je členem komise TC-69 pro statistické metody v mezinárodní organizaci ISO. Dále je aktivním členem Americam Society for Quality, International Statistical Institute, International Association for Statistical Computing. V České republice podporuje rozvoj statistických metod a její aplikaci a implementaci na mnoha výzkumných a průmyslových pracovištích, je pořadatelem pravidelných konferencí Analýza dat v Pardubicích.