

МІЖРЕГІОНАЛЬНА
АКАДЕМІЯ УПРАВЛІННЯ ПЕРСОНАЛОМ



МАУП

П. І. Бідюк, Б. П. Ткач, Т. Харрінгтон

МАТЕМАТИЧНА СТАТИСТИКА

Навчальний посібник

Київ
ДП «Видавничий дім «Персонал»
2016

Рецензенти: *Ф. Г. Гаращенко*, д-р техн. наук, проф. (Київ. нац. ун-т імені Тараса Шевченка);
М. В. Андреев, д-р фіз.-мат. наук, проф. (Ін-т прикладного системного аналізу НТУУ «КПІ»);
І. І. Дахно, д-р екон. наук, проф.

Схвалено Вченою радою Міжрегіональної Академії управління персоналом (протокол № 5 від 27.05.15)

Математична статистика: навч. посіб. / П. І. Бідюк, Б. П. Ткач, Т. Харрінгтон. — К.: ДП «Вид. дім «Персонал», 2016. — 348 с. — Бібліогр.: с. 346.

ISBN 978-617-02-0234-5

Висвітлено сучасні методи статистичного аналізу даних, представлених часовими перерізами або часовими рядами. Розглянуто методику статистичного аналізу даних у вигляді чотирьох основних етапів, методи збору та попередньої обробки даних. Наведено основи побудови регресійних моделей процесів довільної природи. Запропоновано методику застосування розв'язків дискретних рівнянь типу авторегресії та авторегресії з ковзним середнім до аналізу поведінки випадкових процесів з детермінованими складовими та знаходження оцінок коротко- і середньострокових прогнозів динаміки їх розвитку. Розглянуто елементи теорії прийняття статистичних рішень, моделювання ризику та перевірки гіпотез. Наведено методику факторного і дискримінантного аналізу. Окремий розділ присвячено оптимальному статистичному оцінюванню станів динамічних систем за допомогою фільтра Калмана з використанням дискретних моделей у просторі станів, що враховують збурення станів та похибки вимірів. Подано процедуру статистичного моделювання фільтра, яка надає можливість формувати прогнозуючі розподіли, а також приклади застосування методів обробки статистичних даних до розв'язування реальних задач.

Для студентів, аспірантів та викладачів, а також для інженерів, які спеціалізуються у галузі розв'язання задач статистичного аналізу даних, математичного моделювання і прогнозування динаміки розвитку фінансово-економічних процесів та процесів іншої природи, представлених статистичними або експериментальними даними.

© П. І. Бідюк, Б. П. Ткач, Т. Харрінгтон,
2016

© Міжрегіональна Академія управління
персоналом (МАУП), 2016

ISBN 978-617-02-0234-5

© ДП «Видавничий дім «Персонал», 2016

ЗМІСТ

Вступ	8
Перелік скорочень	12
Розділ 1. ЧОТИРИ ЕТАПИ СТАТИСТИЧНОГО АНАЛІЗУ	14
1.1. Задачі, які розв'язують методами математичної статистики	14
1.2. Планування збору даних (перший етап статистичного аналізу)	17
1.3. Попередня обробка і дослідження даних (другий етап статистичного аналізу)	20
1.4. Оцінювання параметрів статистичних і математичних моделей (третій етап статистичного аналізу)	31
1.5. Перевірка сформульованих гіпотез (четвертий етап статистичного аналізу)	34
1.6. Суцільний та вибірковий методи збору даних	39
1.7. Достовірність статистичних досліджень	42
1.8. Контрольні питання і справи	44
Розділ 2. МОДА, МЕДІАНА, СЕРЕДНЄ І ВАРІАЦІЯ	47
2.1. Означення	47
2.2. Дві властивості середнього	49
2.3. Вплив зміни значень ряду розподілу на середнє	50
2.4. Деякі приклади застосування середнього, медіани і моди	51
2.5. Позначення та знаходження середнього арифметичного	53
2.6. Визначення поточного середнього	53
2.7. Математичне сподівання дискретної випадкової змінної	54
2.8. Інші види середнього	56
2.9. Варіація	59
2.10. Виміри варіації: середнє відхилення та дисперсія	59
2.11. Знаходження незміщеної оцінки дисперсії	62
2.12. Властивості дисперсії	66
2.13. Стандартне або середнє квадратичне відхилення	67
2.14. Вплив зміни значень елементів ряду на дисперсію	69

2.15. Застосування дисперсії.....	70
2.16. Однаково розподілені взаємно незалежні випадкові величини	71
2.17. Початкові і центральні моменти	73
2.18. Поняття групових і загальних статистичних характеристик	75
2.19. Контрольні питання і вправи	77

Розділ 3. ОПИС ПОЛОЖЕННЯ ОКРЕМОГО СПОСТЕРЕЖЕННЯ В РЯДУ РОЗПОДІЛУ, ГРУПУВАННЯ ДАНИХ

3.1. Процентильні ранги і процентилі	79
3.2. Нормовані відхилення (z -оцінки).....	81
3.3. Середні і z -оцінки	83
3.4. Порівняння z -оцінок і процентилів.....	84
3.5. Середнє і стандартне відхилення ряду, утвореного із z -оцінок	85
3.6. Стандартні або T -оцінки.....	86
3.7. Застосування процентильних рангів і стандартних оцінок.....	87
3.8. Часові ряди і часові перерізи даних	87
3.9. Дискретні і неперервні величини.....	88
3.10. Табулювання і графічне зображення дискретних величин	89
3.11. Округлювання значень.....	90
3.12. Групування даних і побудова групової таблиці частот.....	91
3.13. Гістограма і полігон ряду розподілу	93
3.14. Контрольні питання і вправи	95

Розділ 4. СТАТИСТИЧНІ ПАРАМЕТРИ ДЛЯ ЗГРУПОВАНИХ ДАНИХ

4.1. Визначення медіани і процентилів з гістограми	97
4.2. Знаходження медіани і процентилів.....	99
4.3. Квartilі і децилі.....	100
4.4. Кумулятивна крива.....	101
4.5. Визначення моди згрупованих даних	102
4.6. Визначення середнього згрупованих даних.....	102
4.7. Знаходження дисперсії і стандартного відхилення для згрупованих даних	104
4.8. Контрольні питання і вправи.....	107

Розділ 5. НОРМАЛЬНИЙ РОЗПОДІЛ	108
5.1. Теоретичні розподіли	108
5.2. Нормальний розподіл і його графік	108
5.3. Чотири властивості нормального розподілу	110
5.4. Функція розподілу і числові характеристики неперервних величин	110
5.5. Статистичні характеристики нормального розподілу	114
5.6. Дослідження кривої нормального розподілу	116
5.7. Вплив параметрів нормального розподілу на форму нормальної кривої	117
5.8. Ймовірність попадання випадкової нормальної величини в заданий інтервал	118
5.9. Ймовірність отримання випадковою величиною заданого значення відхилення	119
5.10. Правило трьох сигм	120
5.11. Оцінка відхилення теоретичного розподілу від нормального, асиметрія і експес	121
5.12. Використання таблиці відносних площ нормального розподілу	123
5.13. Поняття про теорему Ляпунова. Формулювання центральної граничної теореми (ЦГТ)	125
5.14. Властивості ряду розподілу середніх	127
5.15. Процентильний ранг і z-оцінки вибіркового розподілу середніх	129
5.16. Ймовірність і нормальний розподіл	131
5.17. Квантилі нормованого розподілу	135
5.18. Контрольні питання і вправи	135
Розділ 6. АНАЛІЗ ПРОЦЕСУ ПРИЙНЯТТЯ РІШЕНЬ, РИЗИК І ПЕРЕВІРКА ГІПОТЕЗ	138
6.1. Вступ	138
6.2. Основні принципи перевірки гіпотез	141
6.3. Статистична перевірка гіпотез	148
6.4. Перевірка гіпотез у задачах трьох типів	157
6.5. Зауваження стосовно термінології	162
6.6. Односторонні критерії	163
6.7. Приклади застосування процедур перевірки гіпотез	168
6.8. Альтернативне формулювання процедури перевірки гіпотези	170

6.9. Метод побудови процедури перевірки гіпотез.....	172
6.10. Особливості односторонньої і двосторонньої перевірки гіпотез.....	179
6.11. Оцінювання надійності правил перевірки статистичних гіпотез	179
6.12. Контрольні питання і вправи	184
Розділ 7. ДЕЯКІ СТАНДАРТНІ ПРОЦЕДУРИ ПЕРЕВІРКИ ГІПОТЕЗ ТА ІНТЕРВАЛЬНЕ ОЦІНЮВАННЯ	187
7.1. Перевірка гіпотез щодо середніх	187
7.2. Поняття інтервальних оцінок.....	196
7.3. Інтервальна оцінка середнього значення розподілу.....	198
7.4. Контрольні питання і вправи	202
Розділ 8. ПОБУДОВА РЕГРЕСІЙНИХ МОДЕЛЕЙ	204
8.1. Аналіз процесу.....	205
8.2. Попередня обробка даних.....	206
8.3. Аналіз наявності нелінійностей.....	213
8.4. Формування інших елементів структури моделі	215
8.5. Оцінювання параметрів моделей-кандидатів.....	222
8.6. Діагностика моделей — вибір кращої з множини кандидатів.....	224
8.7. Приклади побудови моделей за статистичними даними	230
8.8. Контрольні питання і вправи	243
Розділ 9. ЗАСТОСУВАННЯ РІЗНИЦЕВИХ РІВНЯНЬ У РЕГРЕСІЙНОМУ МОДЕЛЮВАННІ.....	246
9.1. Загальні відомості про різницеві рівняння.....	246
9.2. Ітераційний метод знаходження розв'язків різницевих рівнянь	252
9.3. Знаходження загальних розв'язків однорідних рівнянь та частинних розв'язків неоднорідних.....	256
9.4. Приклади знаходження повних розв'язків різницевих рівнянь	273
9.5. Контрольні питання і вправи	280
Розділ 10. ФАКТОРНИЙ АНАЛІЗ	282
10.1. Завдання факторного аналізу	282
10.2. Основна модель факторного аналізу.....	284
10.3. Аналіз дисперсії вимірів у факторному аналізі.....	287

10.4. Знаходження матриці коефіцієнтів парної кореляції та її перетворення у факторному аналізі	295
10.5. Контрольні питання і вправи	300
Розділ 11. ДИСКРИМІНАНТНИЙ АНАЛІЗ.....	301
11.1. Завдання дискримінантного аналізу.....	301
11.2. Лінійна дискримінантна функція.....	301
11.3. Перевірка гіпотез у дискримінантному аналізі.....	310
11.4. Контрольні питання і вправи	312
Розділ 12. СТАТИСТИЧНА ОБРОБКА ДАНИХ ЗА ДОПОМОГОЮ ОПТИМАЛЬНОГО ФІЛЬТРА.....	314
12.1. Принцип рекурсивного оцінювання.....	314
12.2. Дискретний фільтр Калмана для вільної динамічної системи.....	315
12.3. Дискретний фільтр Калмана для лінійної системи з детермінованими і стохастичними входами	323
12.4. Заходи щодо підвищення якості оптимального фільтра	327
12.5. Приклади побудови оптимального фільтра	329
12.6. Оцінювання невимірюваних компонент вектора стану за допомогою оптимального фільтра.....	337
12.7. Функція прогнозування на основі оптимального фільтра.....	338
12.8. Контрольні питання і вправи	343
Список літератури.....	346
Додаток	347

ВСТУП

Статистичний аналіз даних (САД) виконується з метою побудови прогнозуючих моделей та прийняття рішень на основі оцінок прогнозів. Його застосовують у соціально-економічних, фінансових, технічних та інших системах для коротко- і середньострокового прогнозування обсягів виробництва, накопичення продукції на складах, оцінювання альтернативних економічних стратегій, формування бюджетів підприємств і держави, оцінювання, прогнозування та менеджменту ризиків різної природи, а також для розв'язання багатьох інших задач. Загалом, коло задач, що розв'язуються методами математичної статистики, дуже широке. Кількість різних підходів, методів, моделей та методик статистичного аналізу даних також надзвичайно велика, а тому сьогодні доцільно описувати в одній публікації один із напрямів розвитку САД, наприклад, методи попередньої обробки даних — перший напрям; регресійний аналіз — другий напрям; методи оцінювання параметрів математичних і статистичних моделей — ще один напрям і т. ін. Однак, такий підхід до висвітлення САД потребує досить великих матеріальних витрат на публікації. Тому ця публікація розрахована на вступний семестровий курс для студентів, окремі розділи будуть корисні також аспірантам. Побудова моделей, описаних у книзі, орієнтована на використання даних у формі часових рядів.

На сьогодні у спеціальній літературі описано досить багато методів аналізу статистичних та експериментальних даних з метою побудови математичних і статистичних моделей для прогнозування розвитку лінійних стаціонарних та нелінійних нестаціонарних процесів на основі використання даних у вигляді часових рядів. Найбільш поширеними серед них є регресійні методи — авторегресія (АР), авторегресія з ковзним середнім (АРКС), авторегресія з інтегрованим ковзним середнім (АРИКС), лінійна та нелінійна множинна регресія, квантильна регресія, регресійні дерева, метод групового урахування аргументів (МГВА), нейромережі різноманітних структур, байєсівські моделі і мережі, нечіткі множини, нечіткі нейромережі та ін. Відносно “універсальними” методами моделювання і прогнозування є МГВА і нечіткі нейромережі. Однак практика показує, що одного, навіть відносно універсального методу, не завжди достатньо для досягнення повноти аналізу досліджуваного процесу з подаль-

шою метою використання результатів аналізу для прийняття рішень.

Кожний метод аналізу даних має свої недоліки і переваги стосовно обчислювальних витрат, характеристик точності (адекватності) моделей і оцінок прогнозів. Так, висока (практично прийнятна) точність прогнозу за допомогою МГВА або нейромережі іноді досягається за рахунок високих обчислювальних витрат, які можуть виявитися неприйнятними в окремих випадках. Це особливо стосується застосування отриманої моделі у системі керування реального часу, де модель необхідна для оцінювання прогнозу керованих процесів і синтезу керуючого впливу у кожному періоді дискретизації вимірів. Суттєвий виграш стосовно обчислювальних витрат можна досягти у такому випадку, наприклад, за допомогою набагато простішої моделі авторегресії з ковзним середнім (АРКС) або авторегресії з інтегрованим ковзним середнім (АРІКС), перевагами якої є відносна простота структури та можливості її оперативної адаптації (структури і параметрів) до змінних характеристик процесу у реальному часі.

Іншим поширеним типом даних є часові перерізи, які характеризують стан процесу (об'єкта, суб'єкта) на вибраний момент часу. Наприклад, поточний стан макроекономіки країни можна характеризувати множиною змінних (350–400 змінних залежно від постановки задачі), які вимірюють на заданий момент часу, наприклад, на кінець року. Надалі ці дані можна використати для побудови статистичної моделі у формі багатовимірної розподілу або регресії спеціального типу. Якщо змінні, віднесені до часового перерізу, реєструвати через однакові проміжки часу (наприклад, кожні півроку), то далі можна перейти до представлення у формі часового ряду.

Загалом економетричні статистичні дані або експериментально отримані виміри містять множину структурних компонент, математичний опис яких потребує застосування сучасних методів ймовірно-статистичного аналізу, наведених у посібнику.

Метою даної роботи є: 1) — представлення основних етапів статистичного аналізу даних; 2) — аналіз основних описових статистик даних (моди, медіани, середнього, варіації); 3) — опис положення окремого спостереження у ряду розподілу (процентилі, z -оцінки, T -оцінки), групування даних та їх статистичні параметри; 4) — аналіз нормального розподілу, як одного із самих поширених у САД; 5) — аналіз процесу прийняття статистичних рішень і перевірка гіпотез; 6) — представлення методики побудови математичних (регресій-

них) моделей на основі статистичних даних; 7) — аналіз стаціонарності досліджуваних процесів; 8) — представлення методів факторного і дискримінантного аналізу; 9) — докладний аналіз методу оптимального оцінювання і прогнозування стану процесу (об'єкта) з урахуванням статистичних характеристик (коваріацій) випадкових зовнішніх збурень і шумів (похибок) вимірів за допомогою моделей у просторі станів (за допомогою фільтра Калмана).

Запропоновано нове визначення структури математичної моделі, яке складається з п'яти елементів і сприяє поглибленому розумінні студентами цього важливого поняття, яке використовується практично в усіх розділах навчального посібника. У посібнику запропонована оригінальна методика побудови регресійних моделей, коректне застосування якої дає можливість будувати лінійні та нелінійні математичні моделі високого ступеня адекватності процесам. Для встановлення ступеня адекватності побудованої моделі об'єкта, що моделюється, наведено належну множину статистичних критеріїв якості (коефіцієнт детермінації, t -статистика Стьюдента, статистика Дарбіна-Уотсона, інформаційний критерій Акайке, статистика Байєса-Шварца та ін.). Ще одна множина критеріїв запропонована для аналізу якості коротко- та середньострокових прогнозів, які оцінюються на основі створеної моделі (середня квадратична похибка, середня абсолютна похибка у процентах, коефіцієнт Тейла та ін.). Також представлена оригінальна методика знаходження оцінок багатокрових прогнозів за допомогою функцій прогнозування, які можна отримати на основі побудованих математичних моделей АР та АРКС.

У посібнику запропонована методика знаходження розв'язків різницевих рівнянь (АР та АРКС). Розв'язки необхідні для аналізу досліджуваних процесів, використання їх у системах керування та імітаційного моделювання, а також для знаходження оцінок прогнозів, необхідних для прийняття рішень на їх основі.

Наведена у посібнику методика побудови і застосування оптимального фільтра буде корисною для студентів, аспірантів та інженерів, які зацікавлені у створенні багатовимірних (матричних) причинно-наслідкових моделей. Моделі такого типу знаходять широке застосування у моделюванні та прогнозуванні об'єктів керування у дослідженні та менеджменті ризиків, у системах діагностики в техніці та медицині, системах підтримки прийняття рішень різного призначення тощо. Попередній досвід побудови та використання оптималь-