РАЗВИТИЕ НА СТАТИСТИЧЕСКИЯ АНАЛИЗ НА ОТДАЛЕЧЕНИ НАБЛЮДЕНИЯ (OUTLIERS) ОТ КРАЯ НА ПЪРВАТА СВЕТОВНА ВОЙНА ДО СРЕДАТА НА 80-ТЕ ГОДИНИ НА XX ВЕК

Любомир Иванов[*]

Въведение

Отдалечените наблюдения (outliers), наричани още екстремални, са емпиричен феномен, който се наблюдава толкова често в практическите изследвания, че се възприема за нещо неизбежно (Barnett, Lewis, 1978, p. 2), (Beckman, Cook, 1983, p. 120). Най-общо казано, отдалечени наблюдения са тези, които се отличават в значителна степен от останалите (Kendal, Buckland, 1957, p. 209), (Everitt, Skrondal, 2010, p. 313) и поради това предизвикват съществени изменения в свойствата на статистическите оценки, по-специално по отношение на тяхната неизместеност и ефективност (Rousseeuw, Leroy, 1987, pp. 3 - 8).

Този проблем има над 250-годишна документирана история и в развитието на изследванията влиза в съприкосновение с различни области на статистическия анализ - статистическата теория на заключенията, теорията на разпределенията, Бейсовия анализ, анализа на динамика и зависимости, иконометричното моделиране и други. От съдържателна гледна точка в развитието на научната проблематика по въпросите на отдалечените наблюдения можем да обособим четири етапа:

[*] Д-р, доцент в катедра „Математика и статистика“, Стопанска академия „Д. А. Ценов“ - Свищов; e-mail: lubomir.ivanov@uni-svishtov.bg.

- първи стъпки - от средата на XVIII век до Първата световна война;
- разширяване и задълбочаване на анализа - от края на Първата световна война до средата на 80-те години на XX век;
- систематизиране на изследванията - от средата на 80-те години до края на XX век;
- алгоритмично-информационен бум - от началото на XXI век.

Първия етап вече разгледахме подробно в предходна публикация (Иванов, 2019), поради което фокусът на настоящото изследване е насочен към втория етап. Целта на статията е да се даде характеристика на развитието на теоретичните и емпиричните изследвания за отдалечените наблюдения в историческа перспектива, като се изведат основните му черти.

1. Обща характеристика на периода

Вторият период от развитието на анализа на отдалечените значения обхваща времето от края на Първата световна война приблизително до средата на 80-те години на XX век. През това време значително нараства броят на публикациите, засягащи въпросите за екстремалните значения. През този почти полувековен период анализът се задълбочава, като продължават изследванията в областите на идентификацията и устойчивите методи за оценка, но и се разширява, тъй като се развиват нови проблемни области, свързани с: анализ на ефекта от присъствието на екстремални значения, спецификата при наличие на екстремални значения в динамичните редове, развитие на теорията на разпределенията, Бейсовия анализ, редуцирането на съвкупността и автоматизираното откриване на екстремалните значения. През този период постепенно се изоставя понятието **екстремално значение** (extreme value), което все по-рядко се среща в научните изследвания. Мястото му се заема от ново понятие - **отдалечено наблюдение** (outlier), което се налага като общоприето[1].

Основен принос за развитието на научните изследвания по въпросите на отдалечените наблюдения през периода имат публикациите на Andrews, Anscombe, Barnett, Box, Dixon, Ferguson, Grubbs, Guttman, Hawkins, Huber, Tiao. В същото време в анализа на екстремалните значения се включват и класиците на модерната статистическа

[1] Вж. например в Речник на статистическите термини от 1957 г. на Kendall и Buckland: „In a sample of n observations it is possible for a limited number to be so far separated in value from the remainder that they give rise to the question whether they are not from a different population, or that the sampling technique is at fault. **Such values are called outliers“** (Kendal & Buckland, 1957, p. 209).

теория. Student стои на позицията, че екстремалните значения трябва да се премахват, като основният му аргумент е, че това намалява величината на стандартната грешка и - следователно - повишава точността на оценяването на параметрите на съвкупността (в частност на средната аритметична величина).[2] Fisher, от друга страна, се придържа към мнението, че премахването на наблюдения не бива да се основава само на тяхната отдалеченост, а трябва да има и някаква друга причина. В противен случай те трябва да се приемат като индикатор за отклонение на разпределението на грешката от нормалното.[3]

Тези две позиции продължават да се развиват и да намират нови и нови последователи, като около първата се съсредоточава направлението за устойчивите методи за оценяване (robust estimation methods), където акцентът се поставя върху получаването на точен резултат, а около другата - направлението на идентификацията (outlier detection), което се свързва с изолирането на екстремалните наблюдения, за да бъдат подложени на детайлен последващ анализ или да бъдат премахнати, за да се подобрят точността и ефективността на оценките. Антагонизмът на двете позиции намира израз в различието между целите и средствата им, като води през този период до постепенното им раздалечаване и - в крайна сметка - до обособяването им в напълно самостоятелни направления на статистическия анализ. В известен смисъл устойчивото оценяване остава свързано с отдалечените значения само в това, че разработените методи за оценка не се влияят от присъствие на екстремални стойности в анализираните данни.

2. Устойчиви методи за оценяване (Robust estimation methods)

Най-общо казано, в **първото направление** - устойчивите методи за оценяване, попадат методи, които са приспособени[4] да оценяват търсените параметри на съвкупността дори и при наличие на екстремални значения[5], като в същото време запазват по отношение на тяхната точност и изчерпателност всички свойства на

[2] Student, 1927: „Many if not most routine analyses may have a leptokurtic error system… and in such cases rejection of outlying observations improves the accuracy of the mean".
[3] Fisher, 1922: „As a statistical measure, however, the rejection of observations is too crude to be defended: and unless there are other reasons for rejection than mere divergence from the majority; it would be more philosophical to accept these extreme values, not as gross errors, but as indications that the distribution of errors is not normal".
[4] Например Beckman и Cook (Beckman & Cook, 1983, p. 127) наричат процедурата на използването на такива методи **приспособяване** (accommodation).
[5] Barnett, Lewis, 1978, p. 26: „From the robustness standpoint, we are thus aiming to devise statistical procedures which do not directly examine the outliers, but seek to accommodate them and render them less serious in their influence on estimation or tests of summary measures of the underlying distribution".

оценките, получени чрез класическите методи за оценяване. Тези методи могат да се обособят в четири основни групи[6]:

Първата група методи е свързана с използването на позиционни статистически величини, в т.ч. медианата - за оценка на центъра на разпределението, и квартилното отклонение - за оценка на разсейването. Тъй като са свързани с позицията (location) на значенията на признака в съвкупността, оценките се наричат ***L***-оценки (***L***-estimators). Подобни оценки се използват в статистическата практика и преди, но едва след началото на XX век започва изследването им във връзка с екстремалните значения.

Daniell (Daniell, 1920) използва средната квадратична грешка, за да определи относителната ефективност на различни ***L***-оценки, в т.ч. на средната от редуцирани данни и на различни комбинации от тегла. Dixon (Dixon, 1960) изследва ефективността на ***L***-оценки, получени след цензуриране на определен брой значения в двата края на емпиричното разпределение на признака, а Tukey (Tukey, 1960) изследва ефективността на медианата за оценка на централната тенденция и определя условията, при които тя дава по-точни резултати от средната аритметична величина.

Hodges и Lehmann (1963) предлагат оценка на средната величина на основата на медианата на всички усреднени двойки значения на признака, а Gastwirth (1966) предлага оценка на средната на основата на претеглена линейна комбинация на терцилите и медианата. Bickel (1965) извършва сравнителен анализ на три метода за оценка на средната: Hodges-Lehmann, оценка след редуциране и оценка след цензуриране. Той препоръчва да се използва оценката на Hodges-Lehmann, когато видът на разпределението или относителният дял на отдалечените значения в съвкупността е неизвестен. По-късно Antille (1974) предлага коригиран вариант на оценката на Hodges-Lehmann, който е свързан със значително опростяване на изчислителните процедури.

Barnett (1966) изследва свойствата на ***L***-оценките при условие, че разпределението на съвкупността се различава от нормалното и е близко до разпределението на Коши. Bickel и Hodges (1967), Dixon и Tukey (1968) и Jaeckel (1971) изследват асимптотичната теория на разпределенията, приложена по отношение на различните ***L***-оценки. Ansell (1973) анализира влиянието на асиметрията при използването на ***L***-оценките. Birnbaum и Laska (1967) и Birnbaum, Laska и Meisner (1971) анализират различните ***L***-оценки от гледна точка на тяхната оптималност.

---

[6] Класификацията на методите за устойчиво оценяване в четири групи е предложено от Beckman и Cook (1983, p. 127) и в настоящото изложение се придържам към тяхната позиция.

Втората група методи се основава на остатъчните елементи или по-точно на тяхното разпределение, изразено чрез моментите си или чрез функцията на разпределение на вероятностите. Тъй като при оценката на параметрите на съвкупността се извършва минимизиране на специална целева функция, получените оценки се наричат ***M***-оценки (***M***-estimators). Въпреки че методите на най-малките квадрати и на максималното правдоподобие могат да се разглеждат като специални случаи на ***M***-оценките, в практиката се е наложил възгледът за начало на тези методи да се приема работата на Huber (1964). Той установява, че основният проблем при оценката по метода на най-малките квадрати се корени в чувствителността на сумата на квадратите на отклоненията спрямо присъствието на отдалечени значения, и предлага вместо минимизиране на тази сума да се минимизира сумата на отклоненията, но претеглена със специални тегла. Hogg (1967) предлага при конструирането на теглата да се използва информация за ексцеса на разпределението на остатъците. Hampel (1974) изследва влиянието на различни видове претегляния върху свойствата на получените ***M***-оценки.

Третата група методи се основава на развитието на Бейсовия анализ. Първите идеи, основани на различието между априорните и апостериорните разпределения на параметрите на изучаваната съвкупност, и то във връзка с екстремалните значения, са спорадични и се появяват през 30-те години на XX в. в работите на Ogrodnikoff (1928) и Jeffreys (1932). Минават почти 30 години, преди научните търсения отново да се насочат в тази област.

De Finetti (1961) извежда оценка на средната величина, наречена от него **апостериорна средна** (posterior mean), като използва принципите на Бейсовия анализ за оценка на параметрите при наличие на екстремални наблюдения. Той обръща внимание на обстоятелството, че от гледна точка на Бейсовия анализ няма необходимост да се отхвърлят наблюдения, а точно обратното - разпределението на оценяваната средна величина се определя на базата на всички наблюдения.

Box и Tiao (1962) използват методите на Бейсовия анализ за извършването на оценка на средната величина при наличие на две отдалечени наблюдения, като посочват сериозния обем на изчислителната работа дори и при малки извадки (до 20 наблюдения). Те разглеждат данните като получени от две разпределения с еднакви математически очаквания, но с различни дисперсии.

Gebhardt (1964) изследва свойствата на Бейсовото оценяване при наличие на едно екстремално наблюдение с предварително зададена вероятност за реализация. Разгледани са вариантите на смесени разпределения с различни стойности на

математическото очакване и/или дисперсията. Той предлага да се използва квадратична минимизираща функция.

В по-късна разработка Box и Tiao (1968) разглеждат ефектите от наличието на отдалечени наблюдения в рамките на линейна връзка между параметрите на съвкупността и случайните грешки. Те приемат, че грешките произлизат от две различни разпределения, като всяко може да се реализира с определена вероятност, и въз основа на редица допускания извеждат апостериорните разпределения на търсените параметри. Важен момент в тяхната работа е, че крайното апостериорно разпределение на параметрите е независимо от вида на двете разпределения на случайните грешки.

Четвъртата група обхваща методи за оценка, основани на модификация на наличните данни. Най-простият вариант на подобна модификация е да се премахне екстремалното наблюдение и оценките да се изчисляват без него, като тази практика навлиза масово в научните изследвания на емпирични данни особено след Втората световна война. Други често използвани методи за получаване на устойчиви оценки са: **редуцирането на извадката** (trimming), при което определен брой (или дял) от най-малките и най-големите значения на признака не се вземат под внимание при оценка на търсените параметри; и **цензурирането на извадката** (Winsorization), когато определен брой от най-малките и най-големите значения на признака се заменят със следващото, респективно предходното значение (Everitt, Skrondal, 2010, p. 12).

Важен момент в развитието на правилата за отхвърляне на екстремални наблюдения е работата на Anscombe (1960). Той подлага на критика премахването на наблюдения само на основата на статистически тестове, тъй като при 5% риск от грешка винаги се отхвърлят наблюдения, и то в съотношение 1 на 20, без оглед на това дали действително са екстремални. При процедурите за цензуриране и редуциране дори не се правят тестове, а директно се премахва част от наблюденията. Според него не се прави разлика между **правило за отхвърляне** (rejection rule) и **тест за статистическа значимост** (significance test). Той издига идеята, че всяко правило за отхвърляне трябва да се разглежда от две страни - като защита срещу наличие на екстремални наблюдения (ползата от прилагането му) и като цена, която се плаща за тази полза - намаляването на точността при отхвърлянето на наблюдения, които не са реално екстремални. При съпоставката на ползата и цената, наречени от него „защита“ (protection) и „премия“ (premium) по аналогия със застраховането, се прави оценка за приложимостта на правилото за отхвърляне. Anscombe предлага три правила, като изследва оценяването на средната величина след прилагането им. Tiao и Guttman (1967) предлагат коригирана

оценка на средната величина, основана на идеите на Anscombe. При корекцията се предполага, че дисперсията на съвкупността е известна, а екстремалното наблюдение е само едно.

Guttman и Smith (Guttman, Smith, 1969, 1971) изследват ситуацията, при която екстремалното наблюдение има разпределение с различно математическо очакване и дисперсия. Анализирани са ефектите от използването на правилата на Anscombe, цензурирането (Winsorization) и частичното цензуриране (semi-Winsorization) при малки извадки. По-късно Guttman (1973a) анализира същите правила на базата на симулации по метода „Монте Карло“ при големи извадки и повече от едно екстремално наблюдение.

3. Идентификация на отдалечените наблюдения (Outlier detection)

Второто направление в анализа на отдалечените значения се свързва с развитието на методите за идентификация на отдалечените значения, което става в две посоки едновременно: на първо място, създават се нови тестове за откриване на отдалечените значения, а на второ – разработват се процедури за откриване на повече от едно отдалечено наблюдение.

Идентификацията на едно отделно взето отдалечено наблюдение се осъществява въз основа на статистически критерии, които можем да обособим в три групи. При **първата група** тестове се използва големината на различията между най-големите и съседните им значения на признака. Още през 1925 г. Irwin (1925) предлага да се използва отношението на разстоянията между трите най-големи (респективно трите най-малки) наблюдения. При условие, че дисперсията в съвкупността е известна, Irwin извежда критичните стойности за тестовата характеристика за проверка на нулевата хипотеза, която гласи, че значенията принадлежат на една хомогенна група. По същото време Tippett (1925) използва размаха на разсейването за откриване на екстремалните наблюдения, отново при предположение за известна дисперсия в съвкупността. Dixon (1951) разработва тестова процедура, при която се съпоставят различни оценки на размаха на разсейването - на основата на всички наблюдения и при премахване на най-малките или най-големите. По-късно той извършва сравнителен анализ на ефективността на различните процедури, основани на подреждането на значенията на признака (Dixon, 1962). Shapiro и Wilk (1965) разработват тест за нормалност, който се основава на претеглена сума на отклоненията между минималните и максималните значения на признака. Този тест дава възможност да се проверява за различни отклонения от нормалността, едно от които е присъствието на екстремални наблюдения.

Втората група тестове за откриване на едно отдалечено наблюдение се основава на отклонението му от центъра на разпределението и в по-общ план – на моментите на разпределението. McKay (1935) предлага за целта да се използват стандартизираните отклонения на екстремалното значение на признака от средната аритметична величина при условие, че дисперсията на разпределението е известна. Критичните значения на тестовата характеристика са табулирани от Nair (1948) както за анализирания от McKay вариант, така и когато дисперсията е оценена независимо от извадката.

Thompson (1935) също използва стандартизирано отклонение, но разглежда варианта, когато дисперсията е оценена по данни от извадката, тъй като това е най-често срещаният в практиката случай. Разпределението на тестовата характеристика е нестандартно, но е свързано с ***t***-разпределението на Стюдънт. Thompson извежда критичните значения при фиксиран брой налични екстремални наблюдения в извадката. Pearson и Chandra Sekar (1936) обръщат внимание, че на практика се проверяват не всички, а само най-голямата стойност на стандартизираното отклонение. В този случай може да се използва модификация на таблиците на Thompson със съответните корекции за различни величини на обема на извадката.

Grubbs (1950) извежда точното разпределение на стандартизираното отклонение по данни само от извадката и табулира критичните стойности за едностранна критична област. King (1953) дава тестовата характеристика и критичните стойности при двустранна критична област. По-късните изследвания на Murphy (1951), Paulson (1952), Kudo (1956) и Ferguson (1961) показват, че когато в извадката има само едно екстремално наблюдение, тестът на Grubbs дава най-висока вероятност за вярно решение при идентификацията на отдалеченото наблюдение - т.е. притежава най-висока мощност. В същото изследване Ferguson разглежда и възможностите на коефициентите на асиметрия и ексцес за идентифициране на отдалечените наблюдения. Това е реализирано в последващите разработки, като Pearson и Hartley (1966) дават критичните значения за коефициента на ексцес при големи (над 50 наблюдения), а D'Agostino и Tietjen (1971) - при малки (между 7 и 50 наблюдения) извадки.

David, Hartley и Pearson (1954) разработват тест за идентификация на отдалечени значения на основата на съотношението между размаха на разсейването и стандартното отклонение. По-късно David и Paulson (1965) изследват мощността на този тест в сравнение с други алтернативни тестове. Devlin, Gnanadesikan и Kettenring (1975) използват коефициентите на корелация за целите на идентификацията на отдалечените значения.

Третата група тестове за единично отдалечено значение се основава на Бейсовия анализ. Guttman (1973b) разработва процедура, при която се анализира възможността точно едно наблюдение да произлиза от разпределение със същата дисперсия, но с изместена средна. Проверката става на три стъпки: първо се извежда вероятностното разпределение на изместването на средната, на второ място се оценява вероятността в извадката да има точно едно отдалечено наблюдение и на трето място се установява кое е отдалеченото наблюдение.

Процедурите за откриване на едно екстремално наблюдение дават добри резултати, когато в извадката има само едно такова, но когато са повече, резултатите стават неточни. Поради това успоредно с тестовете за идентификация на едно наблюдение се разработват и процедури за откриване на повече екстремални наблюдения. **Първият начин** за това е чрез последователно идентифициране. Cacoullos (1968) разработва схема за последователно идентифициране на отдалечените значения. McMillan (1971), McMillan и David (1971) и Hawkins (1973) анализират последователното приложение на теста на Grubbs за откриване на повече от едно отдалечено наблюдение и установяват, че неговата мощност е ниска. За повишаване на точността на тестовете Rosner (1975) предлага при последователното приложение на теста на Grubbs да се извършва преизчисляване на средната аритметична величина след отстраняването на всяко отдалечено наблюдение.

Вторият начин за откриване на повече отдалечени наблюдения е едновременното идентифициране, при което отдалечените наблюдения могат да са както най-малките, така и най-големите. Dempster и Rosner (1971) използват Бейсовия анализ и резултатите на Guttman при идентификацията само на едно отдалечено наблюдение. Те разглеждат възможността в извадката да има две или повече отдалечени наблюдения, всяко с различна амплитуда. На основата на предположение за равномерно априорно разпределение на грешките те извеждат апостериорната вероятност за реализацията на точно определен брой отдалечени наблюдения в изследваната извадка. Предложената от тях процедура се осъществява на основата на решаване на модели с различен брой отдалечени значения и избор на модела с най-добро съвпадение. Tiku (1975) предлага да се използва тестова характеристика, основана на отношението на две оценки на дисперсията - от цензурираната и от нецензурираната извадка.

Третият начин е чрез едновременното идентифициране на определен брой отдалечени наблюдения, които формират група - само най-големите или само най-малките. Нарича се „блоково идентифициране“. Walsh (1950) предлага непараметричен

тест, който използва отдалечеността на най-големите наблюдения от медианата на извадката. Вариант на теста на Grubbs за тази цел е разработен от Murphy (1951) при едностранна критична област, а Tietjen и Moor (1972) го разширяват при двустранна алтернатива.

4. Разширяване на полето на анализа на отдалечени наблюдения

През периода след Първата световна война постепенно започват да се разглеждат и някои нови, концептуални въпроси, свързани с процеса на анализа на отдалечените значения. Първоначално се разглеждат само едномерни данни, но постепенно започва да се обръща внимание и на възможностите да съществуват и многомерни екстремални наблюдения. Тук се отнасят работите на Kudo (1957) и Wilks (1963), които разглеждат наличието на отдалечени значения, когато данните са многомерни - т.е. когато са регистрирани два или повече признака за наблюдаваните единици.

Друг нов момент е насочването на интереса към наличието на отдалечени наблюдения не само във вариационните редове, но и в динамичните, както и при приложението на методите за анализ на зависимости. Fox (1972) разглежда наличието на отдалечени наблюдения в динамичните редове и анализира редица въпроси, свързани с тяхната специфика. Srikantan (1961) и Tietjen, Moore и Beckman (1973) анализират проблемите при откриването на отдалечени наблюдения в остатъчните елементи от регресионните уравнения.

Полето на анализа се разширява и чрез изследване на ефектите от наличието на отдалечени значения. Joshi (1972) изучава оценяването на средната величина, когато данните следват експоненциално разпределение. Sinha (1973a, 1973b) разглежда въпросите за оценката на средната аритметична величина и на параметрите на експоненциалното разпределение при условие, че в данните са налични отдалечени наблюдения. Andrews (1973, 1974) разглежда въпросите за оценката на параметрите на множествената линейна регресия при наличие на екстремални значения.

Dixon (1953) анализира ефектите от премахването на отдалечените наблюдения. За водещ критерий той приема величината на средната квадратична грешка и разглежда различни видове и варианти на реализация на екстремалните значения. Интересен резултат от неговото изследване е, че премахването на екстремалното наблюдение води до по-точна оценка, отколкото използването на медианата с началните данни. Basu (1965) изследва въпросите за проверката на хипотези при експоненциално разпределение на съвкупността и наличие на отдалечени значения. Neyman и Scott (1971) и Green (1974)

разглеждат идеята, че някои разпределения са предразположени към пораждане на екстремални значения за разлика от други.

Наличието на големи съвкупности, в които са наблюдавани повече признаци за отделните единици, води до изследвания за възможностите на графичните изображения при откриването на екстремални наблюдения. Те се използват не толкова като заместители на тестовете, а по-скоро като допълнително, неформално средство, което да подпомага анализа на отдалечените значения. Wilk, Gnanadesikan и Huyett (1962) и Wilk и Gnanadesikan (1964) използват квантилите на гама-разпределението, за да идентифицират отдалечените многомерни наблюдения. Gnanadesikan и Wilk (1968) и Healy (1968) предлагат да се използва първо многомерната техника „анализ на основните компоненти“ (principal components analysis), след което да се визуализират разстоянията от първия основен компонент, за да се установят потенциалните екстремални наблюдения. Andrews (1972) и Gnanadesikan (1973) използват графично представяне на многомерни наблюдения в средна по обем извадка на основата на трансформация на Фурие на отделните значения. Графичното изображение дава визуална индикация за тези наблюдения, които се различават съществено от останалите.

През този период се установява наличието на проблем, наречен „маскиране“ (*masking*), при използването на различните тестове за откриване на отдалечени наблюдения. Първите коментари за способността на едно отдалечено наблюдение да увеличава дисперсията на съвкупността и така да прикрива себе си и други отдалечени наблюдения за редица тестови процедури са направени от Pearson и Chandra Sekar (1936). Самият термин е въведен от Murphy (1951). Подробности за ефективността на различни тестове привеждат в своите изследвания McMillan (1971) и Tietjen и Moore (1972).

В самия край на периода се появяват и първите разработки, насочени към автоматизираното откриване на екстремалните наблюдения - в изследванията на Swaroop, West и Lewis (1969) и Swaroop и Winter (1971). Това е предизвикано, от една страна, от натрупването на огромни количества данни за големи съвкупности, които не могат да се анализират своевременно от изследователите с традиционните методи, а от друга - от появата на изчислителните машини, които могат да извършват с голяма бързина множество последователни и еднотипни операции.

Заключение

Като цяло можем да направим **извод,** че вторият период в развитието на анализа на отдалечените наблюдения е своеобразен преходен период между спорадичните

изследвания на предходния първи период и систематизирането на изучаването, характерно за третия период. Най-важните постижения през втория период могат да се обобщят в следното:

На **първо** място, установяват се важни проблеми, свързани с влиянието и ефекта от присъствието на отдалечените наблюдения, и се появяват първите идеи за тяхното разрешаване по начин, който е обоснован от гледна точка на статистическата теория.

Второ, обособяват се в отделни направления на анализа методите за устойчиво оценяване и методите за идентификация. Всяко от тези направления развива свои методи, техники и процедури, специфично приспособени за решаване на поставените различни цели, в единия случай - отстраняване на възможни ефекти без оглед на това дали в данните има, или не отдалечени наблюдения, а в другия - установяване на наличието или отсъствието на отдалечени наблюдения и тяхната локализация.

На **трето** място, през периода има изобилие на нови и разнопосочни идеи, което спомага за бързото разрастване на научната литература, посветена на въпросите на отдалечените наблюдения. В повечето разработки се правят стъпки напред, в неизвестното и неизследваното, в резултат на което често се разкриват нови проблеми и се повдигат повече въпроси, отколкото се дават отговори.

Четвърто, за втория период от анализа на екстремалните значения е характерна хаотичност на изследванията, които се насочват към много различни проблемни области, но в същото време не придобиват завършеност и систематичност. В полето на изучаване на отдалечените значения остават много нерешени въпроси както от концептуално, така и от чисто техническо естество. Това обуславя необходимостта от създаването на концептуална рамка за изучаването; от изследвания, които да извършат систематизиране на разработките и структуриране на научната проблематика, което напълно естествено се реализира през следващия, трети период от развитието на анализа на отдалечените наблюдения.

ЦИТИРАНА ЛИТЕРАТУРА:

Иванов, Л. (2019). Първи период в развитието на статистическия анализ на „отдалечени наблюдения“ („Outliers“). Научно-практическа конференция „Икономиката на България - 30 години след началото на промените“, посветена на 75 години Съюз на учените в България. Свищов, АИ „Ценов“.

Andrews, D. F. (1972). Plots of high-dimentional data. Biometrics(28), 125 - 136.

Andrews, D. F. (1973). Robust estimation for multiple regression models. Bulletin of International Statistical Institute(45), 105 - 111.

Andrews, D. F. (1974). A robust method for multiple linear regression. Technometrics(16), 523 - 531.

Anscombe, F. (1960). Rejection of outliers. Technometrics(2), 123 - 147.

Ansell, M. (1973). Robustness of location estimators to asymmetry. Applied Statistics(22), 249 - 254.

Antille, A. (1974). A linearized version of the Hodges-Lehmann estimator. Annals of Statistics(2), 1308 - 1313.

Barnett, V. D. (1966). Order statistics estimators of the location of the Cauchy distribution. Journal of American Statistical Association(61), 1205 - 1218.

Barnett, V., T. Lewis (1978). Outliers in Statistical Data. Chichester: Wiley.

Basu, A. P. (1965). On some tests of hypotheses relating to the exponential distribution when some outliers are present. Journal of Americal Statistical Assiciation(60), 1249.

Beckman, R. J., R. D. Cook (1983). Outliers. Technometrics, *25*, 119 - 149.

Bickel, P. J. (1965). On Some Robust Estimates of Location. Annals of Mathematical Statistics(36), 847 - 858.

Bickel, P. J., J. L. Hodges (1967). The Asymptotic Theory of Galton’s Test and a Related Sample Estimate of Location. Annals of Mathematical Statistics(38), 73 - 89.

Birnbaum, A., E. M. Laska (1967). Optimal robustness: A general method with applications to linear estimators of location. Journal of American Statistical Association(62), 1230 - 1240.

Birnbaum, A., E. M. Laska, M. Meisner (1971). Optimally robust linear estimators of location. Journal of American Statistical Association(66), 302 - 310.

Box, G. E., G. C. Tiao (1962). A further look at robustness via Bayes’s theorem. Biometrika(49), 419 - 432.

Box, G. E., G. C. Tiao (1968). A Bayesian approach to some outlier problems. Biometrika(55), 119 - 129.

Cacoullos, T. (1968). A sequential scheme for detecting outliers. Bulletin de la Societe Mathematique de Grece(9), 113 - 123.

D'Agostino, R. B., G. L. Tietjen (1971). Simalution Probability Points of b2 for Small Samples. Biometrika(58), 669 - 672.

Daniell, P. J. (1920). Observations Weighted According to Order. Americal Journal of Mathematics(42), 222 - 236.

David, H. A., A. S. Paulson (1965). The performance of several tests for outliers. Biometrika(52), 429 - 436.

David, H. A., H. O. Hartley, E. S. Pearson (1954). The dustribution of the ratio, in a sigle normal sample, of range to standart deviation. Biometrika(41), 482 - 493.

De Finetti, B. (1961). The Bayesian approach to the rejection of outliers. Proceedings of the Forth Berkeley Symposium on Mathematical Statistics and Probability, 1, 199 - 210.

Dempster, A. P., B. Rosner (1971). Detection of Outliers. От S. S. Gupta, Statistical Theory and Related Topics I (161 - 180). New York: Academic Press.

Devlin, S. J., R. Gnanadesikan, J. R. Kettenring (1975). Robust estimation and outlier detection with correlation coefficients. Biometrika(62), 531 - 545.

Dixon, W. J. (1951). Ratios Involving Extreme Values. Annals of Mathematical Statistics(22), 68 - 78.

Dixon, W. J. (1953). Processing Data for Outliers. Biometrics(9), 74 - 89.

Dixon, W. J. (1960). Simplified Estimation From Censored Normal Samples. Annals of Mathematical Statistics(31), 385 - 391.

Dixon, W. J. (1962). Rejection of Observations. In A. E. Sarhan, B. G. Greenberg, Contributions to Order Statistics (299 - 321). New York: John Wiley.

Dixon, W. J., J. W. Tukey (1968). Approximate Behavior of the Distribution of Winsorized t (Trimming/Winsorization 2). Technometrics(10), 83 - 98.

Everitt, B. S., A. Skrondal, (2010). The Cambridge Dictionary of Statistics 4th Edition. Cambridge: Cambridge University Press.

Ferguson, T. S. (1961). On the Rejection of Outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1) (253 - 287). Los Angeles: University of California Press.

Fisher, R. A. (1922). On the Mathematical Foundations of Theoretical Statistics. Philosophical Transactions of the Royal Society of London, Series A(222), 309 - 368.

Fox, A. J. (1972). Outliers in time series. Journal of the Royal Statistical Society, B(43), 350 - 363.

Gastwirth, J. L. (1966). On Robust Procedures. Journal of American Statistical Association(61), 929 - 948.

Gebhardt, F. (1964). On the Risk of Some Strategies for Outlying Observations. Annals of Mathematical Statistics(35), 1524 - 1536.

Gnanadesikan, R. (1973). Graphical methods for informal inference in multivariate data analysis. Bulletin of International Statistical Institute, Book 4(45), 195 - 206.

Gnanadesikan, R., M. B. Wilk (1968). Probability plotting methods for the analysis of data. Biometrika(55), 1 - 17.

Green, R. F. (1974). A note on outlier-prone families of distributions. Annals of Statistics(2), 1293 - 1295.

Grubbs, F. E. (1950). Sample Criteria for testing Outlying Observations. Annals of Mathematical Statistics(21), 27 - 58.

Guttman, I. (1973a). Premium and Protection of Several Procedures for Dealing with Outliers When Sample Sizes Are Moderate to Large. Technometrics(15), 385 - 404.

Guttman, I. (1973b). Care and Handling of Univariate or Multivariate Outliers in Detecting Spuriosity - A Bayesian Approach. Technometrics(15), 723 - 738.

Guttman, I., D. E. Smith (1969). Investigation of Rules for Dealing With Outliers in Small Samples From the Normal Distribution I: Estimation of the Mean. Technometrics(11), 527 - 550.

Guttman, I., D. E. Smith (1971). Investigation of Rules for Dealing With Outliers in Small Samples From the Normal Distribution II: Estimation of the Variance. Techometrics(13), 101 - 111.

Hampel, F. R. (1974). The Influence Curve and Its Role in Robust Estimation. Journal of the American Statistical Association(69), 383 - 393.

Hawkins, D. M. (1973). Repeated testing for outliers. Statistica Neerlandica(27), 1 - 10.

Healy, M. J. (1968). Multivariate normal plotting. Applied Statistics(17), 157 - 161.

Hodges, J. L., E. L. Lehmann (1963). Estimates of Location Based on Rank Tests. Annals of Mathematical Statistics(34), 598 - 611.

Hogg, R. V. (1967). Some Observations on Robust Estimation. Journal of the American Statistical Association(62), 1179 - 1186.

Huber, P. J. (1964). Robust Estimation of a Location Parameter. Annals of Mathematical Statistics(35), 73 - 101.

Irwin, J. O. (1925). On a Criterion for the Rejection of Outlying Observations. Biometrika(17), 238 - 250.

Jaeckel, L. A. (1971). Some Flexible Estimates of Location. Annals of Mathematical Statistics(42), 1540 - 1552.

Jeffreys, H. (1932). An Alternative to the Rejection of Observations. Proceedings of the Royal Society(Ser. A, 137), 78 - 87.

Joshi, P. C. (1972). Efficient estimation of the mean of an exponential distribution when an outlier is present. Technometrics(14), 137 - 144.

Kendal, M. G., W. R. Buckland (1957). A Dictionary of Statistical Terms. New York: Hafner.

King, E. P. (1953). On Some Procedures for the Rejection of Suspected Data. Journal of the American Statistical Association(48), 531 - 533.

Kudo, A. (1956). On Testing Outlying Observations. *Sankhya*(17), 67 - 76.

Kudo, A. (1957). The extreme value in a multivariate normal sample. Memoirs, Faculty of Science, Kyushu University A(11), 143 - 156.

McKay, A. T. (1935). The Distribution of the Difference Betwen the Extreme Observation and the Smaple Mean on Samples of n From a Normal Universe. Biometrika(27), 466 - 471.

McMillan, R. G. (1971). Tests for One or Two Outliers in Normal Samples With Unknown Variance. Technometrics(13), 87 - 100.

McMillan, R. G., H. A. David (1971). Tests for One of Two Outliers in Normal Samples With Known Variance. Technometrics(13), 75 - 85.

Murphy, R. B. (1951). On Tests for Outlying Observations. Princeton University: Ph.D. thesis.

Nair, K. R. (1948). The Distribution of the Extreme Deviate From the Sample Mean and Its Studenized Form. Biometrika(35), 118 - 144.

Neyman, J., E. L. Scott (1971). Outlier proneness of phenomena and of related distribution. In J. Rustagi, Optimising Methods in Statistics (413 - 430). New York: Academic Press.

Ogrodnikoff, K. (1928). On the Occurence of Discordant Observations and a New Method of Treating Them. Monthly Notices of the Royal Astronomical Society(88), 523 - 532.

Paulson, E. (1952). An Optimum Solution to the K-Sample Slippage Problem for the Normal Distribution. Annals of Mathematical Statistics(23), 610 - 616.

Pearson, E. S., C. Chandra Sekar (1936). The Efficiency of Statistical Tools and a Criterion for the Rejection of Outlying Observations. Biometrika(28), 308 - 320.

Pearson, E. S., H. O. Hartley (1966). Biometrika Tables for Statisticians (3rd Edition, Vol. 1). London: Cambridge University Press.

Rosner, B. (1975). On the Detection of Many Outliers. Technometrics(17), 221 - 227.

Rousseeuw, P. J., A. M. Leroy (1987). Robust Regression and Outlier Detection. New York: Wiley.

Shapiro, S. S., M. B. Wilk (1965). An Analysis of Variance Test for Normality (complete samples). Biometrika(52), 591 - 611.

Sinha, S. K. (1973a). Distribution of order statistics and estimation of mean life when an outlier may be present. Canadian Journal of Statistics(1), 119 - 121.

Sinha, S. K. (1973b). Estimation of the parameters of a two-parameter exponential distributionn when an outlier may be present. Utilitas Mathematica(3), 75 - 82.

Srikantan, K. S. (1961). Testing for the single outlier in a regression model. *Sankhya A*(23), 251 - 260.

Student (1927). Errors in Routine Analysis. Biometrika(19), 151 - 164.

Swaroop, R., W. R. Winter (1971). A statistical tecjnique for computer identification of outliers in multivariate data. Washington, D.C.: NASA TN D-6472.

Swaroop, R., K. A. West, C. E. Lewis (1969). A simple technique for automatic computer editing of biodata. Wahsington, D.C.: NASA TN D-5275.

Thompson, W. R. (1935). On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standart Deviation. Biometrika(32), 214 - 219.

Tiao, G. C., I. Guttman (1967). Analysis of Outliers With Adjusted Residuals. Technometrics(9), 541 - 559.

Tienjen, G. L., R. H. Moore, R. J. Becjman (1973). Testing for a single outlier in simple linear regression. Technometrics(15), 717 - 721.

Tietjen, G. L., R. H. Moore (1972). Some Grubbs-Type Statistics for the Detection of Several Outliers. Technometrics(14), 583 - 597.

Tiku, M. L. (1975). A New Statistics for Testing Suspected Outliers. Communications in Statistics. A. Theory and Methods(4), 737 - 752.

Tippett, L. H. (1925). On the Extreme Individuals and the Range of Samples Taken From a Normal Population. Biometrika(17), 364 - 387.

Tukey, J. W. (1960). A Survey of Sampling From Contaminated Distributions. In I. Olkin, S. G. Ghurey, W. Hoeffdine, W. G. Madow & H. B. Mann, Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling (448 - 485). Stanford: Stanford University Press.

Walsh, J. E. (1950). Some Nonparateric Tests of Whether the Largest Observations of a Set are Too Large or too Small. Annals of Mathematical Statistics(21), 583 - 592.

Wilk, M. B., R. Gnanadesikan (1964). Graphical methods for internal comparisons in multiresponse experiments. Annals of Mathematical Statistics(35), 613 - 631.

Wilk, M. B., R. Gnandesikan, M. J. Huyett (1962). Probability plots for the gamma distribution. Technometrics(4), 1 - 20.

Wilks, S. S. (1963). Multivariate statistical outliers. Sankhya A(25), 407 - 426.