Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Оренбургский государственный университет»

Кафедра математических методов и моделей в экономике

О.С. Чудинова

МЕТОДЫ РОБАСТНОГО ОЦЕНИВАНИЯ

Методические указания

Рекомендовано к изданию редакционно-издательским советом федерального государственного бюджетного образовательного учреждения высшего образования «Оренбургский государственный университет» для обучающихся по образовательным программам высшего образования по направлениям подготовки 01.03.04 Прикладная математика, 38.03.05 Бизнес-информатика, 38.04.01 Экономика

Оренбург
2019

УДК 519.22(076.5)
ББК 22.172я7
Ч 84

Рецензент – доцент, кандидат экономических наук О.И. Бантикова

Чудинова, О.С.
Ч84 Методы робастного оценивания: методические указания / О.С. Чудинова; Оренбургский гос. ун-т.– Оренбург: ОГУ, 2018. – 33 с.

Методические указания предназначены для подготовки к лабораторным и практическим занятиям по курсу «Анализ данных», содержат рекомендации к выполнению лабораторной работы и задания к практическому занятию на тему «Методы робастного оценивания». Могут использоваться для самостоятельной работы студентов, в том числе для выполнения индивидуальных заданий и курсовых работ, связанных с многомерным статистическим анализом данных.

Методические указания предназначены для обучающихся по образовательным программам высшего образования по направлениям подготовки 01.03.04 Прикладная математика, 38.03.05 Бизнес-информатика, 38.04.01 Экономика (профиль «Математические и инструментальные методы анализа социальных и экономических процессов») всех формы обучения. Методические указания могут использоваться студентами других направлений подготовки для проведения многомерного статистического анализа данных.

УДК 519.22(076.5)
ББК 22.172я7

Содержание

Введение

В методических указаниях рассматриваются теоретические аспекты методов робастного оценивания, а также практические аспекты реализации методов с помощью статистического пакета прикладных программ Statistica, математического пакета Mathcad и табличного процессора Excel.

Выполнение лабораторной работы и проведение практического занятия на тему «Методы робастного оценивания» по дисциплине «Анализ данных», относящейся к обязательным дисциплинам (модулям) вариативной части блока 1 «Дисциплины (модули)», направлено на формирование у обучающихся по направлению подготовки 01.03.04 Прикладная математика следующих общепрофессиональных и профессиональных компетенций:

ОПК-1 – готовность к самостоятельной работе;

ОПК-2 – способность использовать современные математические методы и современные прикладные программные средства и осваивать современные технологии программирования;

ПК-1 – способность использовать стандартные пакеты прикладных программ для решения практических задач на электронных вычислительных машинах, отлаживать, тестировать прикладное программное обеспечение;

ПК-9 – способность выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности, готовность использовать для их решения соответствующий естественнонаучный аппарат;

ПК-10 – готовность применять математический аппарат для решения поставленных задач, способность применить соответствующую процессу математическую модель и проверить ее адекватность, провести анализ результатов моделирования, принять решение на основе полученных результатов;

ПК-11 – готовность применять знания и навыки управления информацией;

ПК-12 – способность самостоятельно изучать новые разделы фундаментальных наук.

Выполнение лабораторной работы и проведение практического занятия на тему «Методы робастного оценивания» по дисциплине «Анализ данных» базовой части блока 1 «Дисциплины (модули)» направлено на формирование у обучающихся по направлению подготовки 38.03.05 Бизнес-информатика следующих общепрофессиональных и профессиональных компетенций:

ОПК-1 способностью решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности;

ОПК-2 способностью находить организационно-управленческие решения и готов нести за них ответственность; готов к ответственному и целеустремленному решению поставленных профессиональных задач во взаимодействии с обществом, коллективом, партнерами;

ОПК-3 способностью работать с компьютером как средством управления информацией, работать с информацией из различных источников, в том числе в глобальных компьютерных сетях;

ПК-17 способность использовать основные методы естественнонаучных дисциплин в профессиональной деятельности для теоретического и экспериментального исследования;

ПК-18 способность использовать соответствующий математический аппарат и инструментальные средства для обработки, анализа и систематизации информации по теме исследования;

ПК-19 умение готовить научно-технические отчеты, презентации, научные публикации по результатам выполненных исследований.

Использование методических указаний при выполнении курсовой работы по дисциплине «Методы анализа данных (продвинутый курс)», относящейся к обязательным дисциплинам (модулям) вариативной части блока 1 «Дисциплины (модули)», направлено на формирование у обучающихся по направлению подготовки 38.04.01 Экономика (профиль «Математические и инструментальные

методы анализа социальных и экономических процессов») следующих общепрофессиональных и профессиональных компетенций:

ОК-1 способность к абстрактному мышлению, анализу, синтезу;

ОК-3 готовность к саморазвитию, самореализации, использованию творческого потенциала;

ОПК-1 готовность к коммуникации в устной и письменной формах на русском и иностранном языках для решения задач профессиональной деятельности;

ПК-1 способность обобщать и критически оценивать результаты, полученные отечественными и зарубежными исследователями, выявлять перспективные направления, составлять программу исследований;

ПК-2 способность обосновывать актуальность, теоретическую и практическую значимость избранной темы научного исследования;

ПК-3 способность проводить самостоятельные исследования в соответствии с разработанной программой;

ПК-4 способность представлять результаты проведенного исследования научному сообществу в виде статьи или доклада;

ПК-8 способность готовить аналитические материалы для оценки мероприятий в области экономической политики и принятия стратегических решений на микро- и макроуровне;

ПК-9 способность анализировать и использовать различные источники информации для проведения экономических расчетов.

1 Теоретические аспекты методов робастного оценивания

1.1 Грубые ошибки и причины их возникновения

Выборка из генеральной совокупности может содержать значения, существенно отклоняющиеся от основного массива данных. Такие наблюдения называются аномальными (неправдоподобными, резко выделяющимися) или грубыми ошибками («выбросами») [1]. Наличие аномальных объектов искажает структуру совокупности (является одной из причин неоднородности данных), приводит к смещению оценок параметров распределения и к отклонению закона распределения от теоретического.

Основными причинам наличия аномальных наблюдений являются:

– специфические особенности отдельных элементов изучаемой совокупности;

– неправильное причисление данных к исследуемой совокупности, например, ошибки группировки, типологической классификации и пр.;

– ошибки при регистрации и обработки данных.

Решение проблем нахождения, устранения выбросов и получения адекватных оценок параметров распределения занимается специальных раздел статистики – робастное (устойчивое) оценивание. Методы робастного (устойчивого) оценивания – это статистические методы, которые позволяют получать надежные оценки параметров с учетом неизвестного закона распределения генеральной совокупности и наличия существенных отклонений в значениях данных [2].

Прежде чем использовать методы робастного оценивания, необходимо установить, существуют ли аномальные наблюдения в выборке. Для формулировки предположений относительно аномальных объектов можно использовать графическую процедуру «Ящик с усами».

1.2 Графическая процедура «Ящик с усами»

Для построения диаграммы по выборочным данным рассчитывают следующие характеристики:

1) оценки медианы (*Me*), нижней (C_1) и верхней (C_2) квартилей;

2) межквартильный размах $\Delta C = C_2 - C_1$;

3) шаг $h = 1{,}5 \cdot \Delta C$;

4) внутренние барьеры $\delta_1 = C_1 - h$, $\delta_2 = C_2 + h$;

5) наружные барьеры $Б_1 = C_1 - 2h$, $Б_2 = C_2 + 2h$.

Значения, выходящие за границы наружных барьеров, называются неправдоподобными. Значения, находящиеся между внутренними и наружными барьерами, называются внешними. Результаты расчетов удобно представлять в виде диаграммы, которая называется «ящик с усами» [1].

Сформулированные на основе графической процедуры предположения о наличии аномальных наблюдений проверяются с помощью специальных статистических критериев.

1.3 Обнаружение аномальных наблюдений в скалярном случае

В скалярном случае для выявления аномальных наблюдений используются критерии Смирнова-Граббса, Граббса, Титьена-Мура [1, 2]. Пусть $\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(n)}$ – априорный вариационный ряд случайной выборки $\xi_{1,n}$ из генеральной совокупности ξ, имеющей математическое ожидание m и дисперсию σ^2; $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ – апостериорный вариационный ряд, построенный на основе выборки $x_{1,n}$, являющейся реализацией $\xi_{1,n}$. Рассмотрим критерий Граббса. Он предназначен для проверки одного максимального и одного минимального наблюдения.

Гипотезы для проверки максимального наблюдения имеют вид:

$H_0: M\xi_{(n)} = m$ (максимальное наблюдение не является аномальным);

$H_1^+: M\xi_{(n)} = m + d$, $d > 0$ (максимальное наблюдение является аномальным).

Статистика Граббса для проверки максимального наблюдения имеет вид:

$$G_n = \frac{\sum_{i=1}^{n-1} (\xi_{(i)} - \bar{x}(\xi_{(1),(n-1)}))^2}{\sum_{i=1}^{n} (\xi_{(i)} - \bar{x}(\xi_{1,n}))^2}, \quad (1.1)$$

где $\bar{x}(\xi_{(1),(n-1)}) = \frac{1}{n-1} \sum_{i=1}^{n-1} \xi_{(i)}$.

Гипотезы для проверки минимального наблюдения имеют вид:

$H_0: M\xi_{(1)} = m$ (минимальное наблюдение не является аномальным);

$H_1^-: M\xi_{(1)} = m + d$, $d < 0$ (минимальное наблюдение является аномальным).

Статистика Граббса для проверки минимального наблюдения имеет вид:

$$G_1 = \frac{\sum_{i=2}^{n} (\xi_{(i)} - \bar{x}(\xi_{(2),(n)}))^2}{\sum_{i=1}^{n} (\xi_{(i)} - \bar{x}(\xi_{1,n}))^2}, \quad (1.2)$$

где $\bar{x}(\xi_{(2),(n)}) = \frac{1}{n-1} \sum_{i=2}^{n} \xi_{(i)}$.

Смысл критерия Граббса заключается в сравнении двух дисперсий: исходной и усеченной. Статистика Граббса может принимать значения от 0 до 1. Чем ближе значение статистики к единице, тем меньше отличие между усеченной и исходной выборками. Если наблюдаемое значение статистики меньше критического значения $G_{\alpha,n}$, то нулевая гипотеза отвергается.

Если выборка содержит более одного аномального наблюдения, то можно действовать следующим образом. Применить статистику Граббса к выборке объема *n*. Если нулевая гипотеза отвергается, то отбросить аномальное наблюдение и применить эту же статистику к выборке объема *n*-1. Процедуру повторять до тех

пор, пока в выборке не будут выявлены все аномальные объекты. При применении такой процедуры может встретиться ситуация, при которой аномальные наблюдения группируются близко друг к другу, образуя скопления точек вдали от остальных наблюдений. В такой ситуации последовательная процедура выявления аномальных наблюдений может быть нечувствительна. В этом случае рекомендуется использовать критерии Титьена-Мура [1, 2].

1.4 Обнаружение аномальных наблюдений в многомерном случае

В многомерном случае аномальный объект характеризуется вектором признаков, поэтому $\xi_i = (\xi_{i1}, \xi_{i2}, ..., \xi_{ik})^T$. В этом случае можно применять одномерные критерии к отдельным признакам, а можно использовать комплексный подход, предполагающий анализ одновременно по всем признакам. Нулевая гипотеза формулируется следующим образом:

H_0: $M\xi_i = m_\xi$ (наблюдение не является аномальным).

Если m_ξ и Σ_ξ не известны, то для проверки нулевой гипотезы используется следующая статистика [1]:

$$F(\xi_{1,n}) = \frac{(n-k-1)(n-1)}{((n-1)^2-1)k}\left(\xi_i - \bar{x}(\xi_{1,n-1})\right)^T \hat{\Sigma}^{-1}_{исп}(\xi_{1,n-1})\left(\xi_i - \bar{x}(\xi_{1,n-1})\right) \in F(\nu_1 = k, \nu_2 = n-k-1), \quad (1.3)$$

где оценки $\bar{x}(\xi_{1,n-1})$ и $\hat{\Sigma}_{исп}(\xi_{1,n-1})$ построены по случайной выборке, из которой исключено наблюдение ξ_i;

$d(x_i, \bar{x}) = \left(x_i - \bar{x}\right)^T \hat{\Sigma}^{-1}_{исп}\left(x_i - \bar{x}\right)$ – квадрат расстояния Махаланобиса от *i*-го объекта наблюдение до центра выборочной совокупности.

В случае значительного засорения многомерная совокупность подвергается проверке итерационным способом:

1) с помощью *F*-статистики проверяется каждое наблюдение ξ_i, $i = \overline{1,n}$;

2) если для всех ξ_i, $i=\overline{1,n}$, гипотеза H_0 не отвергается, то аномальных наблюдений нет. Если для некоторых ξ_i, $i=\overline{1,n}$, гипотеза H_0 отвергается, то наблюдение с наибольшим значением статистики исключается из выборки и процедура повторяется для оставшихся n-1 наблюдений.

1.5 Методы робастного оценивания

В случае наличия ошибок говорят, что имеется смесь двух распределений – основного и засоряющего. Рассмотрим методы робастного оценивания, применяемые, когда основное и засоряющее распределения симметричны.

Методы робастного оценивания реализуют два подхода к оцениванию параметров распределения. Первый подход ориентирован на устранение аномальных объектов и оценивание параметров по оставшейся (усеченной) совокупности. Этот подход реализует метод оценивания, предложенный Пуанкаре. Второй подход демонстрирует метод Винзора, предполагающий модификацию аномальных наблюдений [1, 2].

Пусть $\xi_{(1)} \le \xi_{(2)} \le \ldots \le \xi_{(n)}$ – вариационный ряд случайной выборки.

Американский ученый Пуанкаре рассматривал нормально распределенную совокупность, в которой некоторая доля элементов α $(0<\alpha<0{,}5)$ является грубыми ошибками. Для оценки математического ожидания он предложил вместо средней арифметической использовать α -урезанную среднюю:

$$T(\xi_{1,n};\alpha)=\frac{1}{n-2k}\sum_{i=k+1}^{n-k}\xi_{(i)}, \qquad (1.4)$$

где k – число грубых ошибок в начале и конце вариационного ряда.

Согласно Пуанкаре, $k \le [\alpha n]$, где n – объем выборки; α – некоторая функция вероятности засорения совокупности ε, значение которой находят по специальной таблице в зависимости от ε [1, 2].

При расчете оценки Винзора наблюдения, содержащие ошибки, заменяются на модифицированные (винзорированные) значения с устраненными или уменьшенными ошибками. Для оценки математического ожидания используется статистика:

$$W(\xi_{1,n};\alpha)=\frac{1}{n}\left[\sum_{i=k+1}^{n-k}\xi_{(i)}+k\cdot\left(\xi_{(k+1)}+\xi_{(n-k)}\right)\right]. \qquad (1.5)$$

Оценка Винзора отличается от оценки Пуанкаре тем, что $2k$ значений не удаляются из совокупности, а проецируются в ближайшую точку оставшейся совокупности данных, т.е. при расчете средней арифметической участвуют n наблюдений.

По аналогии с оценками $T(\xi_{1,n};\alpha)$ и $W(\xi_{1,n};\alpha)$ могут быть найдены оценки и других числовых характеристик (дисперсии, среднего квадратического отклонения, коэффициентов асимметрии и эксцесса и т.д.)

2 Реализация методов робастного оценивания в пакетах прикладных программ

Для анализа показателей строительной деятельности в Российской Федерации отобраны следующие показатели:

ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство», в фактически действовавших ценах, млн. руб.;

ξ_2 – ввод в действие зданий жилого и нежилого назначения, тыс. квадратных метров;

ξ_3 – общая площадь жилых помещений, приходящаяся в среднем на одного жителя, квадратных метров;

ξ_4 – ввод в действие мощностей дошкольных образовательных организаций, квадратных метров;

ξ_5 – число предприятий и организаций строительной деятельности.

Значения показателей за 2015 год приведены в таблице А.1 [3]. Провести анализ выборочных данных на наличие аномальных наблюдений:

1) для каждого признака реализовать графическую процедуру «ящик с усами», по диаграммам выдвинуть предположения о наличии аномальных наблюдений;

2) проверить выдвинутые для каждого признака предположения о наличии аномальных наблюдений с помощью статистических критериев;

3) учитывая результаты выявления аномальных наблюдений в скалярном случае, сформулировать и проверить предположения о наличии аномальных наблюдений в многомерном случае;

4) для каждого признака рассчитать оценки математического ожидания и дисперсии по методу Пуанкаре и Винзора, сравнить полученные результаты с оценками соответствующих числовых характеристик по исходной совокупности данных. Рассчитать оценки вектора математических ожиданий и ковариационной

матрицы по исходным данным и по методу Пуанкаре. Сравнить полученные результаты.

1) Для реализации графической процедур «ящик с усами» воспользуемся пакетом прикладных программ Statistica. После ввода исходных данных необходимо выбрать пункт меню Basic Statistics/Tables (рисунок 2.1) и запустить модуль Descriptive statistics (рисунок 2.2).

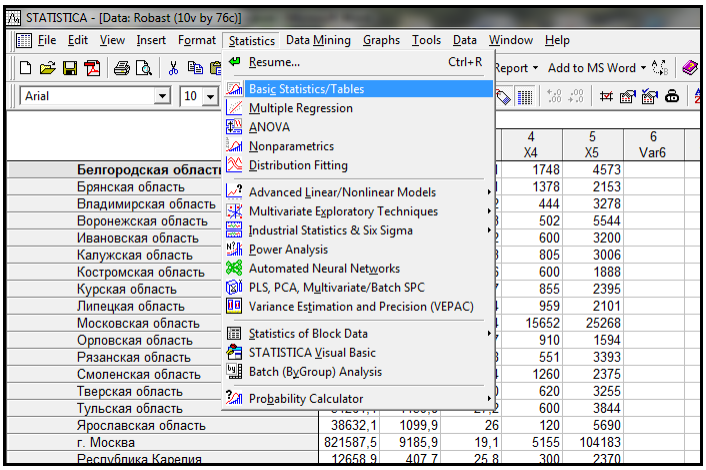


Рисунок 2.1 – Выбор пунктов меню для построения диаграммы «ящик с усами»

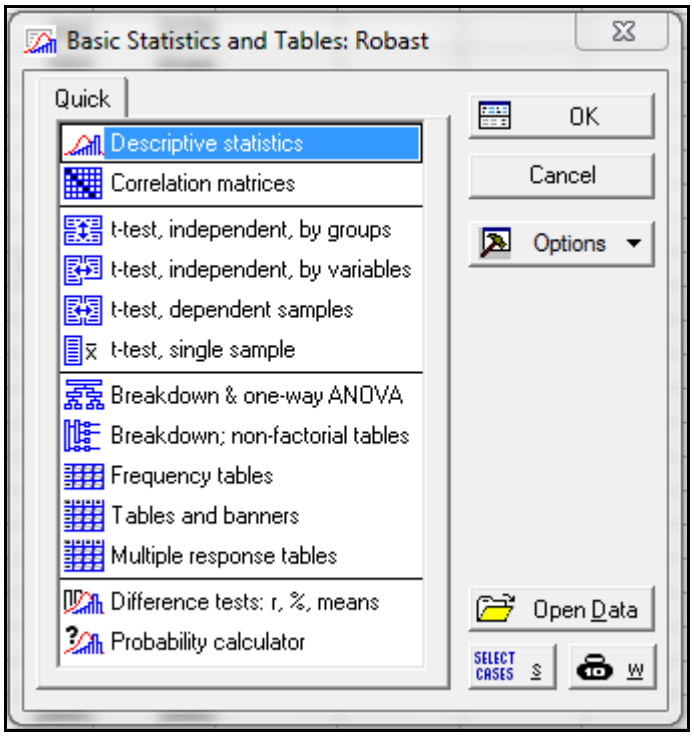


Рисунок 2.2 – Форма для запуска модуля Descriptive statistics

В появившейся на экране форме с помощью кнопки Variables выбрать признак, по которому будет строиться «ящик с усами» и на странице Options в свойствах диаграммы (Options for Box-Whisker plots) указать по каким характеристикам будет строиться график (Median/Quartiles/Range) (рисунок 2.3). После этого на странице Quick нажать кнопку Box & whisker plot for all variables (рисунок 2.4).

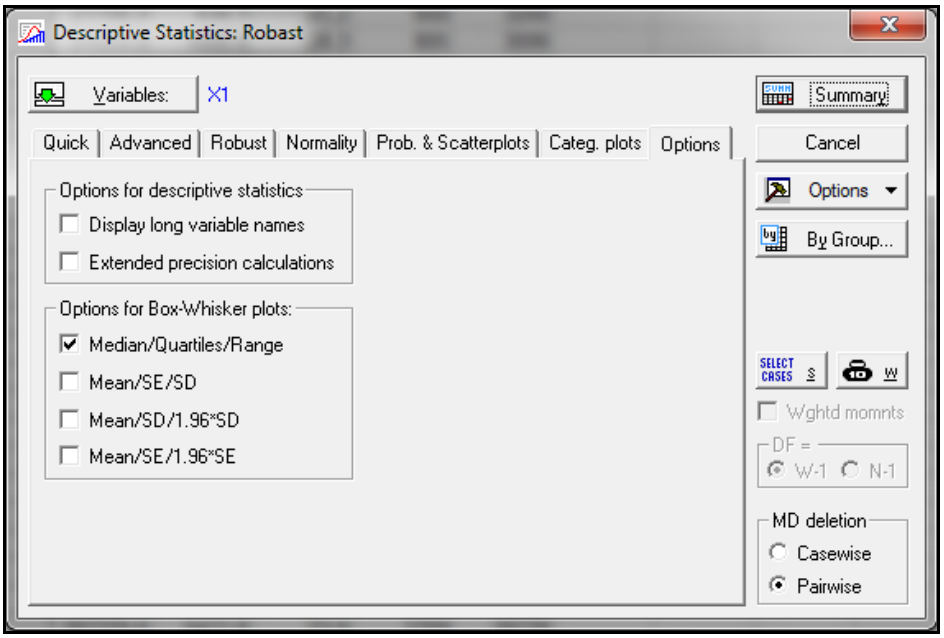


Рисунок 2.3 – Установка параметров построения диаграммы «ящик с усами»

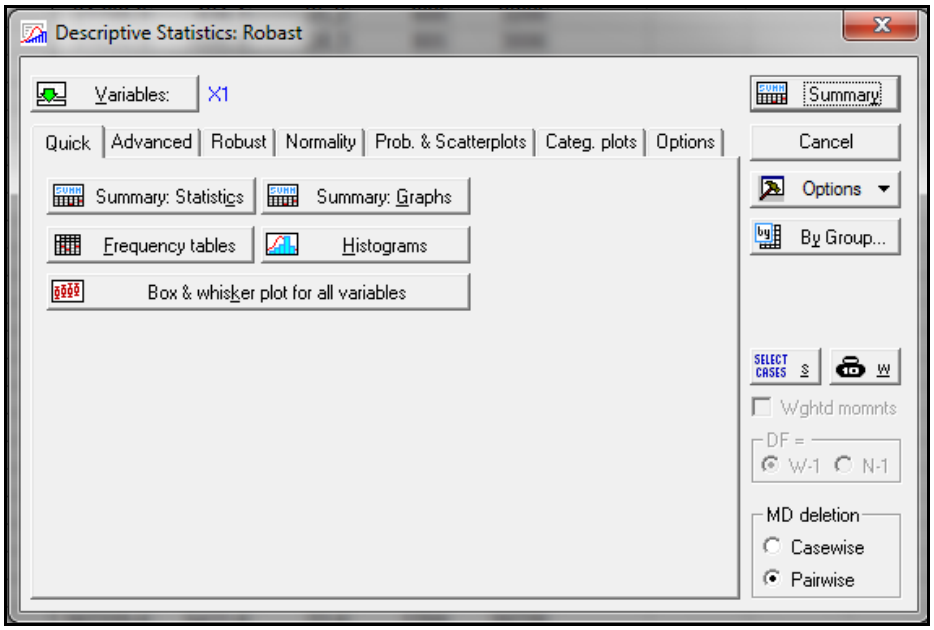


Рисунок 2.4 – Вид страницы Quick формы Descriptive statistics

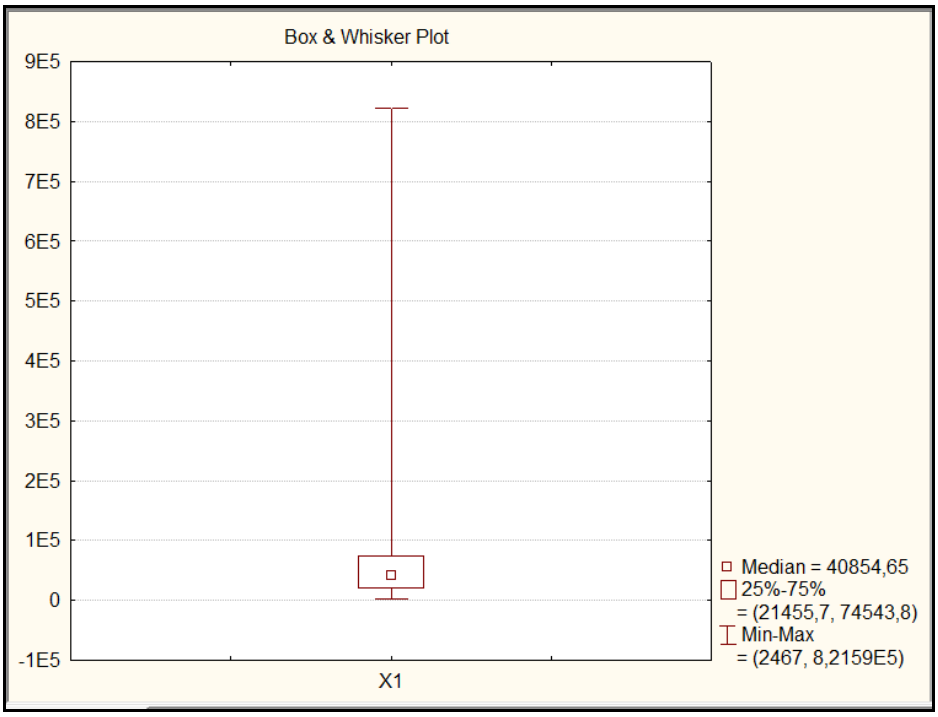


Рисунок 2.5 – Результаты построения диаграммы «ящик с усами» в пакете Statistica

Аналогичным образом стоятся графики по остальным признакам. В правом нижнем углу графика выводятся оценки медианы, нижней и верхней квартилей. Эти значения используются для расчета остальных характеристик, необходимых для построения барьеров. Результаты расчетов сведены в таблицу 2.1.

Таблица 2.1 – Результаты расчета основных числовых характеристик для построения диаграммы «ящик с усами»

Числовая характеристика	Признак				
	x_1	x_2	x_3	x_4	x_5
$\hat{Me}$	40854,65	1059,4	25	835	3103
C_1	21455,7	469,9	23,1	500	1756
C_2	74543,8	1929,5	26,9	1950	5623
ΔC	53088,1	1459,6	3,8	1450	3867
h	79632,15	2189,4	40,35	2175	5800,5
δ_1	-58176,45	-1719,5	-17,5	-1675	-4044,5
δ_2	154175,95	4118,9	67,25	4125	11423,5
$Б_1$	-137808,6	-3908,9	-57,6	-3850	-9845
$Б_2$	233808,1	6308,3	107,6	6300	17224

Для каждого признака на диаграммах, построенных в пакете Statistica, отложим рассчитанные барьеры и укажем объекты, выходящие за наружные барьеры. Диаграмма «ящик с усами» для признака ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство», представлена на рисунке 2.6.

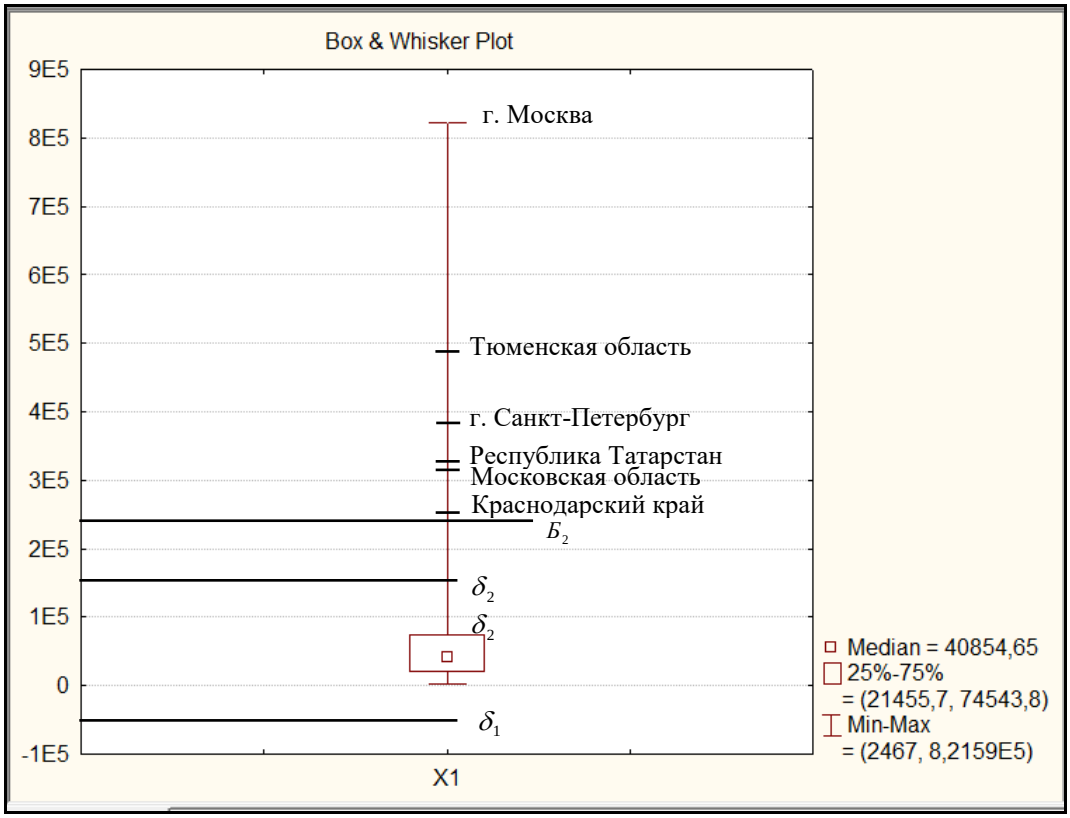
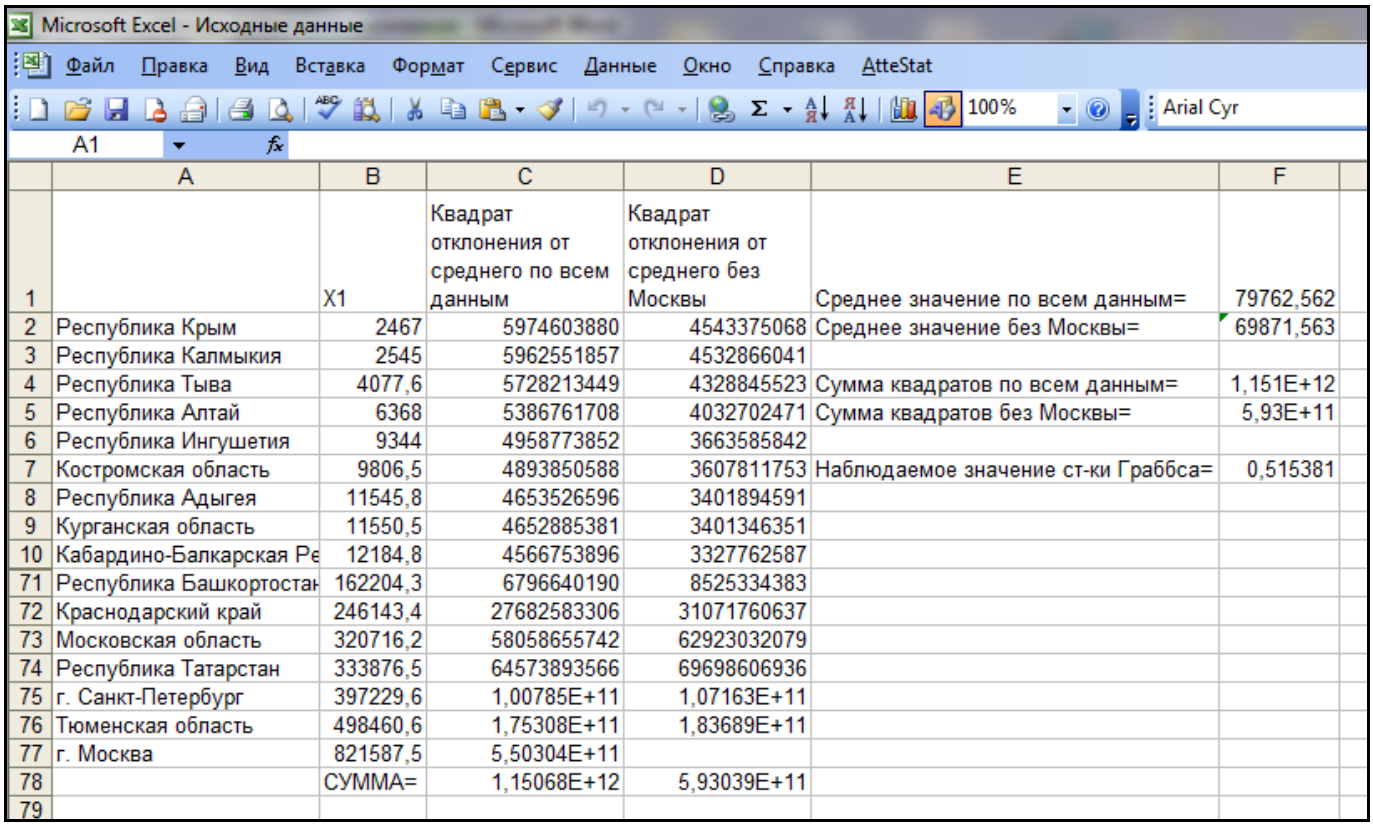


Рисунок 2.6 – Диаграмма «ящик с усами» для признака ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство»

На основе диаграммы «Ящик с усами» можно выдвинуть предположение об аномальных значениях признака ξ_1 в следующих субъектах РФ: г. Москва (821587,5 млн.руб.), Тюменская область (498460,6 млн.руб.), г. Санкт-Петербург (397229,6 млн.руб.), Республика Татарстан (333876,5 млн.руб.), Московская область (320716,6 млн.руб.) и Краснодарский край (246143,6 млн.руб.).

Аналогичным образом сформулированы предположения о наличии аномальных наблюдений по остальным признакам.

2) Сформулированные на основе диаграмм предположения о наличии аномальных наблюдений по каждому признаку проверены с помощью критерия Граббса. Данный критерий более чувствителен к наличию аномальных наблюдений по сравнению с критерием Смирнова-Граббса и может использоваться при отсутствии маскирующего эффекта. Оснований предполагать возникновение маскирующего эффекта нет, поскольку аномальные наблюдения не располагаются близко друг с другом. Поскольку аномальных значений несколько, то проверка осуществлялась с помощью итерационного алгоритма. Расчеты проводились в пакете Ecxel. Фрагмент расчетов для проверки Москвы по значению признака ξ_1 представлены на рисунке 2.7. Результаты проверки выдвинутых для каждого признака предположений о наличии аномальных наблюдений представлены в таблице 2.2.


Microsoft Excel - Исходные данные

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка AtteStat

100% Arial Cyr

A1

	A	B	C	D	E	F
1		X1	Квадрат отклонения от среднего по всем данным	Квадрат отклонения от среднего без Москвы	Среднее значение по всем данным=	79762,562
2	Республика Крым	2467	5974603880	4543375068	Среднее значение без Москвы=	69871,563
3	Республика Калмыкия	2545	5962551857	4532866041		
4	Республика Тыва	4077,6	5728213449	4328845523	Сумма квадратов по всем данным=	1,151E+12
5	Республика Алтай	6368	5386761708	4032702471	Сумма квадратов без Москвы=	5,93E+11
6	Республика Ингушетия	9344	4958773852	3663585842		
7	Костромская область	9806,5	4893850588	3607811753	Наблюдаемое значение ст-ки Граббса=	0,515381
8	Республика Адыгея	11545,8	4653526596	3401894591		
9	Курганская область	11550,5	4652885381	3401346351		
10	Кабардино-Балкарская Ре	12184,8	4566753896	3327762587		
71	Республика Башкортостан	162204,3	6796640190	8525334383		
72	Краснодарский край	246143,4	27682583306	31071760637		
73	Московская область	320716,2	58058655742	62923032079		
74	Республика Татарстан	333876,5	64573893566	69698606936		
75	г. Санкт-Петербург	397229,6	1,00785E+11	1,07163E+11		
76	Тюменская область	498460,6	1,75308E+11	1,83689E+11		
77	г. Москва	821587,5	5,50304E+11			
78		СУММА=	1,15068E+12	5,93039E+11		
79						

Рисунок 2.7 – Результаты расчета наблюдаемого значения статитики Граббса в пакете Excel

Таблица 2.2 – Результаты проверки предположений о наличии аномальных наблюдений с помощью критерия Граббса

Субъект РФ	Наблюдаемое значение статистки	Объем выборки	Критическое значение статистики	Вывод
ξ_1 - Объем работ, выполненных по виду экономической деятельности «Строительство»				
г.Москва	0,515	76	0,82	аномальное
Тюменская область	0,689	75		аномальное
г.Санкт-Петербург	0,727	74		аномальное
Республика Татарстан	0,755	73		аномальное
Московская область	0,762	72		аномальное
Краснодарский край	0,84	71		не аномальное
ξ_2 - Ввод в действие зданий жилого и нежилого назначения				
Московская область	0,559	76	0,82	аномальное
г.Москва	0,716	75		аномальное
ξ_4 - Ввод в действие мощностей дошкольных образовательных организаций				
Московская область	0,635	76	0,82	аномальное
Свердловская область	0,639	75		аномальное
Республика Татарстан	0,675	74		аномальное
ξ_5 - Число предприятий и организаций в строительной деятельности				
г.Москва	0,237	76	0,82	аномальное
г.Санкт-Петербург	0,596	75		аномальное
Московская область	0,761	74		аномальное
Краснодарский край	0,842	73		не аномальное

Критическое значение статистики Граббса после $n = 20$ увеличивается незначительно (рисунк Б.1). Поэтому в таблице 2.2 для разного числа наблюдений указано одно и тоже критичекое значение, равное 0,82.

По результатам проверки гипотез можно сделать сделать следующие выводы:

– аномально высокими значениями признака ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство», характеризуются следующие субъекты РФ: г.Москва (821587,5 млн.руб), Тюменская область (498460,6 млн.руб.), г. Санкт-Петербург (397229,6 млн.руб.), Республика Татарстан (333876,5 млн.руб.) и Московская область (320716,6 млн.руб.);

– по признаку ξ_2 – ввод в действие зданий жилого и нежилого назначения аномальными являются Московская область (14421 тыс. кв. метров) и г. Москва (9185,9 тыс. кв. метров);

– аномально высокими значениями признака ξ_4 – ввод в действие мощностей дошкольных образовательных организаций характеризуются следующие субъекты РФ: Московская область (15652 кв. метров), Свердловская область (12599 кв. метров) и Республика Татарстан (9802 кв. метров);

– по признаку ξ_5 – число предприятий и организаций в строительной деятельности аномальными являются г. Москва (104183 ед.), г. Санкт-Петербург (39728 ед.), Московская область (25268 ед.).

3) Проверим, являются ли аномальными Московская область, г. Москва и г. Санкт-Петербург в многомерном случае. Расчет наблюдаемого значения *F*-статистики проводился с помощью пакетов Excel и Mathcad. Фрагмент расчетов в пакете Excel приведен на рисунке 2.8, в Mathcad – на рисунке 2.9. Результаты представлены в таблице 2.3.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		X1	X2	X3	X4	X5								
2	Белгородская область	61981.2	2308	29.1	1748	4573		Вектор средних значений (без Московской области)=	76549.85	1643.512	24.40666667	1701.773	6433.187	
3	Брянская область	24563.8	1147	28.1	1378	2153								
4	Владимирская област	33119.3	900.7	27.2	444	3278				*Столбец 1*	*Столбец 2*	*Столбец 3*	*Столбец 4*	*Столбец 5*
5	Воронежская область	73114.4	2675	28.3	502	5544			Столбец 1	14557970322				
6	Ивановская область	21708.9	434.1	25.2	600	3200		Оценка ковариационной матрицы (без Моск_обл)=	Столбец 2	170509860.3	2624499.772			
7	Калужская область	51533.9	1687	28.3	805	3006			Столбец 3	-52492.34844	-138.005947	11.42542		
8	Костромская область	9806.5	469.9	26.6	600	1888			Столбец 4	127783698.2	2328169.447	-393.381	4333991	
9	Курская область	44569.9	1030	28.7	855	2395			Столбец 5	1382435437	16534960.48	-6679.48	11334034	162467459
10	Липецкая область	40865.5	1678	28.4	959	2101								
11	Орловская область	19207	789.5	27	910	1594		Критическое значение статистики Fкр=	2.345586					
12	Рязанская область	26500.9	1132	28.8	551	3393								

Рисунок 2.8 – Фрагмент расчетов, необходимых для вычисления наблюдаемого значения статистики Фишера для Московской области, в пакете Excel

В результате выявления аномальных наблюдений в многомерном случае можно сделать вывод, что «г. Москва», «г. Санкт-Петербург», «Московская область» являютя аномальными. Следует отметить, что причина выделения этих наблюдениий на фоне остальных субъектов РФ связана со специфическими особенностями этих субъектов. Москва, Санкт-Петербург, Московская область характеризуются более высоким уровнем социально-экономического развития по

сравнению с другии регионами и, как следствие, – более высокими темпами развития строительства.

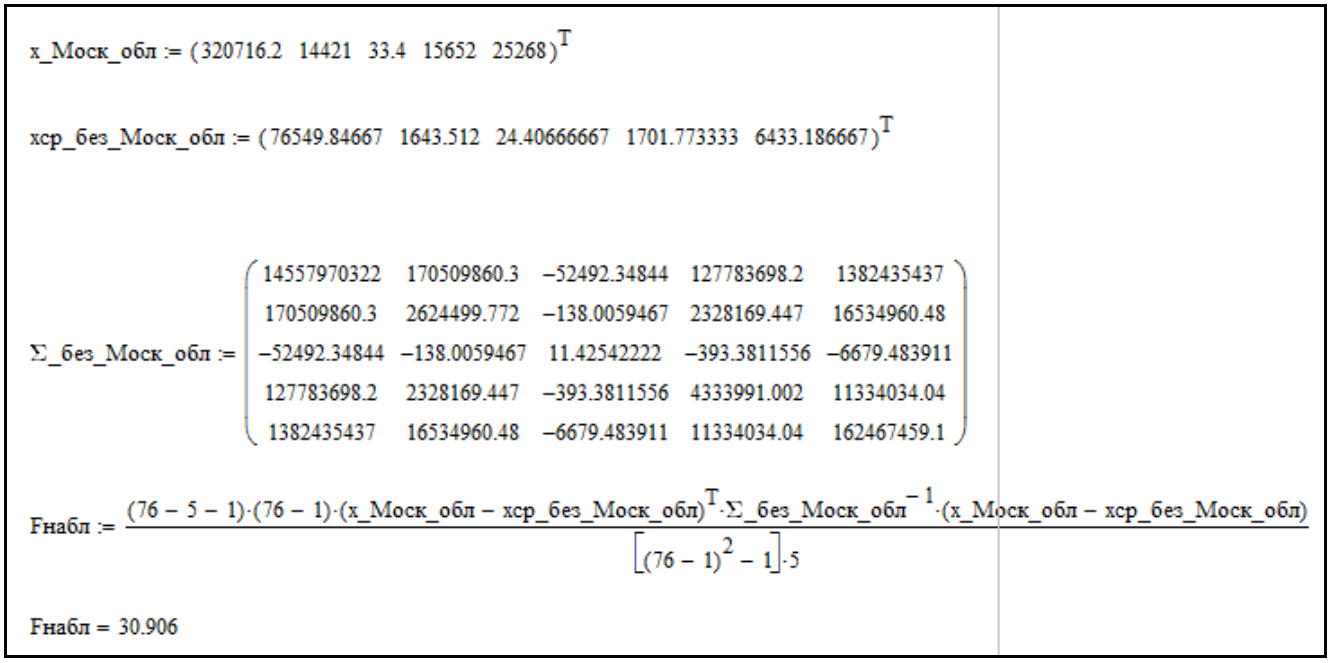


Рисунок 2.9 – Вычисление наблюдаемого значения статистики Фишера для Московской области в пакете Mathcad

Таблица 2.3 – Результаты выявления аномальных наблюдений в многомерном случае

Субъект РФ	Объем выборки	Наблюденное значение статистики Фишера	Критическое значение статистики Фишера	Вывод
Итерация №1				
Московская область	76	30,91	2,346	Наблюдние «г. Москва» является аномальным
г. Москва	76	97,11		
г.Санкт-Петербург	76	39,23		
Итерация №2				
Московская область	75	35,16	2,347	Наблюдение «г. Санкт-Петербург» является аномальным
г.Санкт-Петербург	75	44,43		
Итерация №3				
Московская область	74	41,93	2,349	Наблюдение «Московская область» является аномальным

4) Учитывая результаты выявления аномальных наблюдений в скалярном случае, для каждого признака рассчитаем оценки математического ожидания и дисперсии, используя принципиальные особенности методов Пуанкаре и Винзора. Результаты расчетов в пакете Excel по признаку ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство», представлены на рисунке 2.10.

	A	B	C	D
1		X1		
2	Республика Крым	2467	2467	2467
3	Республика Калмыкия	2545	2545	2545
4	Республика Тыва	4077.6	4077.6	4077.6
5	Республика Алтай	6368	6368	6368
6	Республика Ингушетия	9344	9344	9344
68	Самарская область	140638.4	140638.4	140638.4
69	Нижегородская область	154729.8	154729.8	154729.8
70	Ростовская область	155182.1	155182.1	155182.1
71	Республика Башкортостан	162204.3	162204.3	162204.3
72	Краснодарский край	246143.4	246143.4	246143.4
73	Московская область	320716.2		246143.4
74	Республика Татарстан	333876.5		246143.4
75	г. Санкт-Петербург	397229.6		246143.4
76	Тюменская область	498460.6		246143.4
77	г. Москва	821587.5		246143.4
78				
79	Среднее=	79762.56	51973.02	64747.39
80	Оценка дисперсии=	15342407271	2097428855	4305715822
81				
82		По исходным данным	По методу Пуанкаре	По методу Винзора

Рисунок 2.10 – Результаты расчета оценок математического ожидания и дисперсии для признака ξ_1 – объем работ, выполненных по виду экономической деятельности «Строительство»

Расчеты проводились по упорядоченным данным. По методу Пуанкаре аномальные наблюдения (г. Москва, Тюменская область, г. Санкт-Петербург, Республика Татарстан, Московская область) исключены из рассмотрения. По методу Винзора значение признака ξ_1 для этих объектов становится равным 246143,4 (значению признака ξ_1 для Краснодарского края). Аналогичным образом

рассчитаны оценки числовых характеристик для остальных признаков. Результаты представлены в таблице 2.4.

Таблица 2.4 – Результаты расчета оценок математического ожидания и дисперсии с учетом наличия аномальных наблюдений

Метод	Признак									
	ξ_1		ξ_2		ξ_3		ξ_4		ξ_5	
	$\bar{x}$	S^2	$\bar{x}$	S^2	$\bar{x}$	S^2	$\bar{x}$	S^2	$\bar{x}$	S^2
По исходным данным	79763	15342407271	1812	4772713	24.5	12.5	1885	6894633	6681	167135225
По методу Пуанкаре	51973	2097428855	1542	1906591	-	-	1442	1885309	4638	17864534
По методу Винзора	64747	4305715822	1656	2343757	-	-	1627	2658786	5203	25018705

Для третьего признака робастные оценки не рассчитывались, поскольку аномальные наблюдений по этому признаку отсутствуют. Поскольку в исходной выборке имеются аномально высокие значения признаков ξ_1, ξ_2, ξ_4, ξ_5, то оценки математического ожидания и дисперсии по исходным данным для этих признаков существенно превышают робастные оценки соответствующих числовых характеристик. Так, среднее значение первого признака по методу Пуанкаре меньше среднего значения по исходным данным на 35%, по методу Винзора – меньше на 19%.

Результаты оценивания вектора математических ожиданий и ковариационной матрицы по исходным данным и по методу Пуанкаре представлены в таблице 2.5.

После удаления из выборки Москвы, Санкт-Петербурга и Московской области оценки математических ожиданий для всех признаков уменьшились, абсолютные значения элементов матрицы $\hat{\Sigma}_\xi$ также уменьшились. Следует отметить, что оценки ковариаций для некоторых пар признаков (ξ_1 и ξ_3, ξ_3 и ξ_4, ξ_3 и ξ_5) поменяли знаки.

Таблица 2.5 – Результаты расчета оценок вектора математических ожиданий и ковариационной матрицы с учетом наличия аномальных наблюдений

Оценка вектора математических ожмданий	Оценка ковриационной матрицы
По исходным данным	
$\bar{x} = \begin{pmatrix} 79763 \\ 1812 \\ 24{,}53 \\ 1885 \\ 6681 \end{pmatrix}$	$\hat{\Sigma}_\xi = \begin{pmatrix} 1{,}5\times10^{10} & 2{,}1\times10^{8} & -23289 & 1{,}7\times10^{8} & 1{,}4\times10^{9} \\ 2{,}1\times10^{8} & 4{,}7\times10^{6} & 1356 & 4{,}6\times10^{6} & 1{,}9\times10^{7} \\ -23289 & 1356 & 12{,}33 & 1241 & -4392 \\ 1{,}7\times10^{8} & 4{,}6\times10^{6} & 1241 & 6{,}8\times10^{6} & 1{,}5\times10^{7} \\ 1{,}4\times10^{9} & 1{,}9\times10^{7} & -4392 & 1{,}5\times10^{7} & 1{,}6\times10^{8} \end{pmatrix}$
По методу Пуанкаре	
$\bar{x} = \begin{pmatrix} 6195 \\ 1489 \\ 24{,}49 \\ 1654 \\ 4638 \end{pmatrix}$	$\hat{\Sigma}_\xi = \begin{pmatrix} 5{,}7\times10^{9} & 7{,}9\times10^{7} & 4995 & 9{,}5\times10^{7} & 2{,}5\times10^{8} \\ 7{,}9\times10^{7} & 1{,}7\times10^{6} & 461 & 2{,}0\times10^{6} & 4{,}9\times10^{6} \\ 4995 & 461 & 11{,}34 & -148 & 762 \\ 9{,}5\times10^{7} & 2{,}0\times10^{6} & -148 & 4{,}2\times10^{6} & 6{,}9\times10^{6} \\ 2{,}5\times10^{8} & 4{,}9\times10^{6} & 762 & 6{,}9\times10^{6} & 1{,}8\times10^{7} \end{pmatrix}$

3 Задание, требования к оформлению и защите отчета по лабораторной работе

Выполнение лабораторной работы по теме «Методы робастного оценивания» состоит из следующих этапов:

1) ознакомление с формулировкой задания к лабораторной работе и порядком её выполнения в пакетах прикладных программ;

2) выполнение расчетов по своим данным;

3) анализ полученных результатов;

4) подготовка письменного отчета по лабораторной работе;

5) защита отчета по лабораторной работе.

Задание к лабораторной работе:

1) выбрать предмет исследования и набор показателей, характеризующих данное явление или процесс;

2) собрать статистические данные по выбранным показателям, используя сайты Федеральная служба государственной статистики РФ (http://www.gks.ru), Российского мониторинга экономического положения и здоровья населения НИУ-ВШЭ (RLMS-HSE) (http://www.cpc.unc.edu/projects/rlms и http://www.hse.ru/rlms), единого архива экономических и социальных данных Высшей Школы Экономики (http://sophist.hse.ru) и другие информационные ресурсы;

3) провести анализ выборочных данных на наличие аномальных наблюдений.

Отчет по лабораторной работе оформляется в соответствии с требованиями стандарта организации СТО 02069024.101 – 2015 [4]. Отчет должен сдержать титульный лист; задание к лабораторной работе; краткие теоретические сведения по теме лабораторной работы; результаты выполнения лабораторной работы, их анализ и интерпретацию; приложения с исходными данными и отчетами, полученными в пакетах прикладных программ.

Для защиты отчета по лабораторной работе необходимо подготовиться к ответу на вопросы и задания, приведенные ниже.

1) Назовите причины появления грубых ошибок.
2) Расскажите алгоритм реализации графической процедуры «ящик с усами».
3) Какие критерии используются для обнаружения аномальных наблюдений в скалярном случае?
4) Для чего предназначены критерий Смирнова-Граббса и критерий Граббса? В чем их суть?
5) Для чего предназначены критерии Титьена-Мура? В чем их суть?
6) Каким образом осуществляется обнаружение аномальных наблюдений в многомерном случае?
7) Расскажите, в чем состоит метод оценивания Пуанкаре?
8) Расскажите, в чем состоит метод оценивания Винзора?

4 Вопросы и задания к практическому занятию

1) По данным о количестве покупателей за день в 19 магазинах сети

1 2 6 6 7 7 8 8 8 10 11 13 14 16 16 17 18 30 32

построена диаграмма «ящик с усами». Чему равно количество внешних (V) и неправдоподобных (NP) наблюдений?

2) По данным наблюдений среднесуточная температура ($°C$) 25 июля за последние 15 лет в некоторой местности составляла: 22 20 22 22 23 14 27 25 22 23 33 13 21 20 22. Какие значения среднесуточной температуры согласно диаграмме «ящик с усами» относятся к неправдоподобным?

3) По ниже приведенным наблюдаемым значениям признака, используя критерий Смирнова-Граббса, на 5% уровне значимости определите, какие наблюдения являются грубыми ошибками: 1 2 6 6 7 7 8 8 8 10 11 13 14 16 16 17 18 30 35. Известно, что математическое ожидание признака составляет 12,1, а среднее квадратическое отклонения равно 8,2. Ниже приведен фрагмент таблицы процентных точек Смирнова-Граббса:

n	Уровень значимости α		
	0,1	0,05	0,01
…	…	…	…
18	2,577	2,728	3,017
19	2,601	2,754	3,047
20	2,623	2,779	3,079
21	2,644	2,801	3,106

4) По ниже приведенным наблюдаемым значениям признака, используя критерий Смирнова-Граббса, на 5% уровне значимости определите, какие наблюдения являются грубыми ошибками: 1 2 6 6 7 7 8 8 8 10 11 13 14 16 16 17 18 30 32. Рассчитано выборочное значение среднего квадратического отклонения, составившее 8,2. Ниже приведен фрагмент таблицы процентных точек Смирнова-Граббса:

n	Уровень значимости α		
	0,1	0,05	0,01
…	…	…	…
18	2,577	2,728	3,017
19	2,601	2,754	3,047
20	2,623	2,779	3,079
21	2,644	2,801	3,106

5) По ниже приведенным наблюдаемым значениям признака, используя критерий Граббса, на 5% уровне значимости определите, какие наблюдения являются грубыми ошибками:

№ наблюдения	x_i	$(x_i - x)^2$	x_i	$(x_i - x)^2$	x_i	$(x_i - x)^2$
1	1	123,3	-	-	1	100
2	2	102,1	2	115,0	2	81
3	6	37,3	6	45,2	6	25
4	6	37,3	6	45,2	6	25
5	7	26,1	7	32,7	7	16
6	7	26,1	7	32,7	7	16
7	8	16,9	8	22,3	8	9
8	8	16,9	8	22,3	8	9
9	8	16,9	8	22,3	8	9
10	10	4,4	10	7,4	10	1
11	11	1,2	11	3,0	11	0
12	13	0,8	13	0,1	13	4
13	14	3,6	14	1,6	14	9
14	16	15,2	16	10,7	16	25
15	16	15,2	16	10,7	16	25
16	17	24,0	17	18,3	17	36
17	18	34,7	18	27,9	18	49
18	30	320,2	30	298,5	30	361
19	32	395,8	32	371,6	-	-
Сумма	-	1217,8	-	1087,6	-	800

Критическое значение статистики Граббса $G_{0.05,19} = 0{,}624$.

6) Вычислите расстояние Махаланобиса от i-го объекта $x_i = (10,\ 25)^T$ до центра выборочной совокупности, если $\bar{x} = (5,\ \ 10)^T$, $\hat{\Sigma}^{-1}_{исп} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$.

7) Чему равно значение оценки математического ожидания по методу Винзора, если известно, что в выборке 50 55 500 51 55 57 50 53 5 50 одно минимальное и одно максимальное наблюдения являются аномальными?

8) Реализация случайной выборки имеет вид: 10 10 11 12 13 15 16. Чему равна после первой итерации оценка математического ожидания по методу Хубера, если в качестве начальной оценки параметра использовалась медиана, а шаг модификации k равен 2?

9) Реализация случайной выборки имеет вид: 10 10 10 11 12 12 13 13 15 16. Сколько наблюдений подлежит модификации на первой итерации расчета оценки математического ожидания по методу Хубера, если в качестве начальной оценки параметра использовать медиану и считать, что в данных имеется 1 выброс? Фрагмент таблицы зависимости шага модификации от вероятности засорения выборки имеет вид:

Вероятность засорения выборки	Шаг модификации
0	0
0,001	2,630
0,01	1,945
0,05	1,399
0,10	1,140
0,20	0,862

Список использованных источников

1 Большаков, А. А. Методы обработки многомерных данных и временных рядов: учеб. пособие для вузов / А. А. Большаков, Р. Н. Каримов. – М.: Горячая линия – Телеком, 2007. – 522 с.

2 Сошникова, Л.А. Многомерный статистический анализ в экономике: учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г.Е. Уебе, М. Шефер. – М.: ЮНИТИ, 1999. – 598 с.

3 Регионы России. Социально-экономические показатели. 2016: Стат. сб. / Росстат. – М., 2016. – 1326 с.

4 СТО 02069024.101–2015 Работы студенческие. Общие требования и правила оформления. – Оренбург: ОГУ, 2015. – Режим доступа: http://www.osu.ru/docs/official/standart/standart_101-2015_.pdf.

Приложение А

(обязательное)

Исходные данные для демонстрации выполнения лабораторной работы по робастному оцениванию

Таблица А.1 – Значения показателей строительной деятельности в Российской Федерации за 2015 год

Субъект РФ	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
(1)	(2)	(3)	(4)	(5)	(6)
Белгородская область	61981,2	2307,5	29,1	1748	4573
Брянская область	24563,8	1146,7	28,1	1378	2153
Владимирская область	33119,3	900,7	27,2	444	3278
Воронежская область	73114,4	2675,1	28,3	502	5544
Ивановская область	21708,9	434,1	25,2	600	3200
Калужская область	51533,9	1686,6	28,3	805	3006
Костромская область	9806,5	469,9	26,6	600	1888
Курская область	44569,9	1030,4	28,7	855	2395
Липецкая область	40865,5	1678,1	28,4	959	2101
Московская область	320716,2	14421	33,4	15652	25268
Орловская область	19207	789,5	27	910	1594
Рязанская область	26500,9	1132,4	28,8	551	3393
Смоленская область	23570,1	773,9	27,4	1260	2375
Тверская область	21202,5	778,4	30	620	3255
Тульская область	31261,1	1139,8	27,2	600	3844
Ярославская область	38632,1	1099,9	26	120	5690
г. Москва	821587,5	9185,9	19,1	5155	104183
Республика Карелия	12658,9	407,7	25,8	300	2370
Республика Коми	73209,2	404,9	26,7	995	2369
Архангельская область	39605,6	757,5	26,6	1245	2731
Вологодская область	40843,8	1088,4	28,5	500	5945
Калининградская область	48977,4	1632,3	26,7	2715	6852
Ленинградская область	92722	3467,7	25,7	2845	3265
Мурманская область	33644,5	190,1	25,1	360	2185
Новгородская область	55908,8	516,1	29,9	220	1762
Псковская область	12547,2	892,2	29,4	65	1511
г. Санкт-Петербург	397229,6	5411,6	23,6	1769	39728
Республика Адыгея	11545,8	506,8	25,8	120	780
Республика Калмыкия	2545	178	24	767	450
Краснодарский край	246143,4	5876,9	24,4	6142	18949
Астраханская область	32698,3	776,4	22,8	340	1918
Волгоградская область	69578	1591,4	23,3	1242	5354
Ростовская область	155182,1	3900,3	23,6	3850	8679

Продолжение таблицы А.1

(1)	(2)	(3)	(4)	(5)	(6)
Республика Дагестан	123921	2125,8	18,2	755	4388
Республика Ингушетия	9344	409,9	14,1	560	1101
Кабардино-Балкарская Республика	12184,8	445,7	18,9	820	1129
Карачаево-Черкесская Республика	12252,9	347	20,6	2040	626
Чеченская Республика	25828,1	1433,5	17,5	800	1756
Ставропольский край	59573,5	2402	23,7	1950	4524
Республика Башкортостан	162204,3	3879,1	24,2	3938	11695
Республика Марий Эл	14952	615,1	24,7	815	1492
Республика Мордовия	23896,3	691,9	26,4	795	1578
Республика Татарстан	333876,5	4638,3	25,2	9802	14732
Удмуртская Республика	37359	932,8	21,6	666	4699
Чувашская Республика	34580,1	1201,6	26	924	2972
Пермский край	95043,1	1929,5	23	2984	9693
Кировская область	31933,8	1255	25	1836	3333
Нижегородская область	154729,8	2421,6	26,1	1171	10183
Оренбургская область	49163,6	1809,7	24,7	865	4285
Пензенская область	38057,4	1983,3	27,3	1696	2385
Самарская область	140638,4	2797,3	25	730	12225
Саратовская область	64928,5	1782	27,8	1408	4393
Ульяновская область	50129,5	1414,8	26,1	670	2568
Курганская область	11550,5	397,1	24	335	1704
Свердловская область	116375,5	4263,8	24,9	12599	17618
Тюменская область	498460,6	5630,3	22,9	4683	17041
Челябинская область	86044,5	3904,9	25	2939	11782
Республика Алтай	6368	209,1	19,8	730	610
Республика Бурятия	19119,4	671,9	21,4	450	2227
Республика Тыва	4077,6	172,4	13,5	420	251
Республика Хакасия	13285,3	404,5	22,7	260	1126
Алтайский край	35758,9	1474,8	23,1	2082	4644
Забайкальский край	18559	412	20,4	1046	1682
Красноярский край	140009,4	2315,8	23,9	4530	8810
Иркутская область	56670,5	1583,9	23,1	3310	6850
Кемеровская область	111619,6	1733,1	23,7	1034	4884
Новосибирская область	59589	4314,4	23,7	5740	13458
Омская область	56168,7	1313,2	23,6	2082	5346
Томская область	35964,7	1013,6	23,4	2450	3403
Республика Саха (Якутия)	87190,5	802,8	21,5	2378	4094
Камчатский край	17104,9	161,9	25,2	220	1275
Приморский край	44207	765	22,4	1866	6109
Хабаровский край	52139,1	919,7	23,1	830	5623
Амурская область	49769,6	532,1	24,3	482	2017
Сахалинская область	75878,4	602,6	25,2	840	2224
Республика Крым	2467	321,4	16,3	520	2634

Приложение Б

(обязательное)

Критические значения статистики Граббса

	A	C
1	n	G(0,05;n)
2	3	0,003
3	4	0,049
4	5	0,127
5	6	0,203
6	7	0,269
7	8	0,326
8	9	0,374
9	10	0,415
10	11	0,451
11	12	0,482
12	13	0,51
13	14	0,534
14	15	0,556
15	16	0,575
16	17	0,593
17	18	0,609
18	19	0,624
19	20	0,638
20	25	0,692
21	30	0,731
22	35	0,761
23	40	0,784
24	45	0,802
25	50	0,818

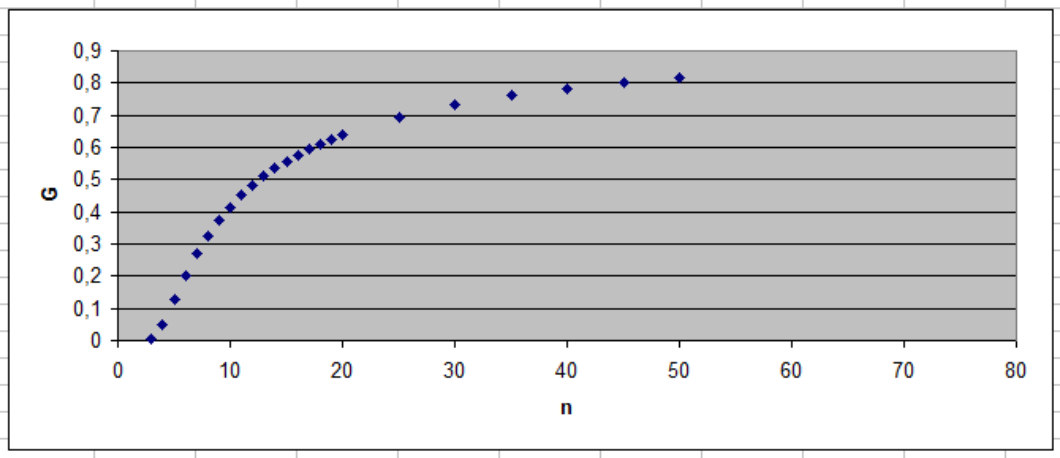


Рисунок Б.1 – Критические значения статистики Граббса $G_{0,05;n}$