****

СОДЕРЖАНИЕ

[ВВЕДЕНИЕ 3](#_Toc51647)

[1 Анализ данных в математической статистике 4](#_Toc51648)

[2 Современные виды анализа в математической статистике 8](#_Toc51649)

[2.1 Корреляционный анализ 8](#_Toc51650)

[2.2 Регрессионный анализ 9](#_Toc51651)

[2.3 Канонический анализ 10](#_Toc51652)

[2.4 Метод сравнения средних 11](#_Toc51653)

[2.5 Частотный анализ 12](#_Toc51654)

[2.6 Методы, основанные на объединении объектов исследования 12](#_Toc51655)

[2.7 Дискриминантный анализ 14](#_Toc51656)

[2.8 Факторный анализ 15](#_Toc51657)

[2.9 Анализ главных компонент и классификация 16](#_Toc51658)

[2.10 Многомерное шкалирование 17](#_Toc51659)

[2.11 Моделирование структурными уравнениями 18](#_Toc51660)

[2.12 Карты контроля качества 20](#_Toc51661)

[3 АНАЛИЗ ДАННЫХ 22](#_Toc51662)

[ЗАКЛЮЧЕНИЕ 27](#_Toc51663)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 28](#_Toc51664)

# ВВЕДЕНИЕ

Итогом научно-технической революции XX века стал неутомимый рост производственных процессов в таких сферах человеческой жизни как техника, экономика, медицина, сельское хозяйство, естественные науки, военное дело, социология, политология и т.д. Но с движением прогресса увеличивалась необходимость в учете и контроле больших объёмов информации, с целью оптимизации и улучшения бизнес-процессов.

Роль статистического анализа данных в современном мире неоценима. С помощью методов математической статистики можно рассчитать размер расходов топлива при перевозках грузов, найти зависимость между какимлибо симптомом и заболеванием, контролировать качество выпускаемой продукции.

Объектом исследования курсовой работы является автомобильный рынок Краснодарского края в 2017 году. Данные по рынку занесены в таблицу.

Целью курсовой работы является анализ автомобильного рынка Краснодарского края и прогнозирование его продаж при помощи пакета прикладного статистического анализа данных STATISTICA 6.

Основная задача работы – изучение методов статистической обработки данных и проведение анализов.

В ходе работы будут разобраны основные методы математической статистики и проведены корреляционный, регрессионный и частотный анализ.

# Анализ данных в математической статистике

Понятие статистики (от латинского status – состояние дел) появилось в XVI в. в Италии и ее функции сводились к сбору данных о государстве. Позже в 1746 году немецкий ученный Готфрид Ахенвалль ввел в науку этот термин, заменив германскую университетскую дисциплину «Государствоведение» на «Статистика» и положив начало развития её как отдельной науки.

На сегодняшний день тяжело представить сферу человеческой деятельности, не использующей статистику. Статистика – это наука, которая позволяет получить и обосновать какие-либо суждения о группе (объектов или субъектов) с некоторой внутренней неоднородностью, помогает увидеть и обосновать закономерности в беспорядочном наборе случайных данных, помогает обусловить наши действия с целью увеличения количества правильно принятых нами решений.

Целесообразно определить три вида научной и прикладной деятельности в области статистических методов анализа данных (по степени специфичности методов, сопряженной с погруженностью в конкретные проблемы):

а) разработка и исследование методов общего назначения, без учета

специфики области применения;

б) разработка и исследование статистических моделей реальных явлений

и процессов в соответствии с потребностями той или иной области деятельности;

в) применение статистических методов и моделей для статистического

анализа конкретных данных.

Математическая статистика – это наука, которая изучает методы сбора, систематизации, обработки и использования статистических данных для получения научно обоснованных выводов и принятия на их основе решений. Но стоит помнить, что прикладная статистика и математическая статистика не подразумевают одно и то же, поскольку многие методы статистического анализа данных, такие как кластерный анализ и многомерное шкалирование, выходят за рамки классической мат. статистики.

Под прикладной статистикой понимают дисциплину, разрабатывающую и систематизирующую понятия, приемы, математические модели и методы, для последующего сбора, систематизации и обработки статистических данных с целью получения научных и практических выводов.

Полученные в следствие каких-либо экспериментов или наблюдений статистические данные являются объектом изучения в прикладной статистике.

В свою очередь математическая статистика представляет собой методы статистической обработки данных, предусматривающих возможность вероятностной интерпретации обрабатываемых данных и полученных в результате обработки статистических выводов.

При работе с математической статистикой необходимо четко знать определение статистических данных, а также понимать значение переменных.

Статистические данные – это совокупность объектов (наблюдений, случаев) и признаков (переменных), их характеризующих. Например, телефоны различных брэндов и признаки, их характеризующие – страна производитель, диагональ экрана, операционная система, частота процессора, объём памяти, цена и т.д.

Переменные – это величины, которые в результате измерения могут принимать различные значения. Их разделяют на два типа:

−Независимые переменные – это переменные, значения которых в процессе эксперимента можно изменять.

−Зависимые переменные – это переменные, значения которых можно только измерять.

Переменные могут быть измерены в различных шкалах. Рассматривают следующие типы шкал, представленные в порядке возрастания их информативности: номинальная, порядковая, интервальная, шкала отношений, абсолютная. Различие шкал определяется их информативностью и количеством допустимых математических действий. Номинальная шкала самая «бедная», так как не определена ни одна арифметическая операция. Самой «богатой» является абсолютная, потому что для нее выполняются все 4 типа отношений между числами и объектами:

1. эквивалентность;
2. порядок;
3. равенство интервалов;
4. равенство отношений;

Измерение в номинальной (классификационной) шкале означает определение принадлежности объекта (наблюдения) к тому или иному классу. В ней можно лишь посчитать количество объектов в классах – частоту и относительную частоту.

Измерение в порядковой (ранговой) шкале, помимо определения класса принадлежности, позволяет упорядочить наблюдения, сравнив их между собой в каком-то отношении. Однако эта шкала определяет только какое из двух наблюдений предпочтительнее и не может установить дистанцию между классами. Поэтому порядковые экспериментальные данные, даже если они изображены цифрами, нельзя рассматривать как числа и выполнять над ними арифметические операции. В этой шкале дополнительно к подсчету частоты объекта можно вычислить ранг объекта.

При измерении в интервальной шкале упорядочивание наблюдений можно выполнить настолько точно, что известны расстояния между любыми двумя их них. Шкала интервалов единственна с точностью до линейных преобразований (*y = ax + b*). Это означает, что шкала имеет произвольную точку отсчета – условный нуль. Над переменными в данной шкале можно выполнять операцию определения расстояния между наблюдениями. Расстояния являются полноправными числами и над ними можно выполнять любые арифметические операции.

Шкала отношений похожа на интервальную шкалу, но она единственна с точностью до преобразования вида *y = ax*. Это означает, что шкала имеет фиксированную точку отсчета – абсолютный нуль, но произвольный масштаб измерения. Измерения в этой шкале – полноправные числа и над ними можно выполнять любые арифметические действия.

Абсолютная шкала имеет и абсолютный нуль, и абсолютную единицу измерения (масштаб). Примером абсолютной шкалы является числовая прямая. Эта шкала безразмерна, поэтому измерения в ней могут быть использованы в качестве показателя степени или основания логарифма.

По своей сути, конечная цель всякого исследования или научного анализа заключается в нахождении зависимостей между переменными. При помощи статистического анализа данных можно детально изучить взаимосвязи между различными переменными и спрогнозировать их изменение в будущем.

Так как вычислительные операции многомерного анализа данных являются достаточно трудоемкими при их реализации, то даже при работе с небольшим объемом данных грамотная и эффективная его обработка практически невозможна без использования вычислительной техники. На сегодняшний день неутомимо движутся вперед и развиваются компьютерные технологии обработки и анализа данных. Большой сегмент рынка прикладных программ занимают пакеты по статистической обработке данных. Одним из наиболее известных в России пакетов для прикладного статистического анализа данных является пакет STATISTICA.

# Современные виды анализа в математической статистике

## Корреляционный анализ

Корреляционный анализ – метод обработки [статистических](https://dic.academic.ru/dic.nsf/ruwiki/749) данных, заключающийся в изучении коэффициентов (корреляции) между переменными. При этом сравниваются коэффициенты корреляции между одной парой или множеством пар признаков для установления между ними статистических взаимосвязей.

Коэффициент корреляции указывает на степень зависимости одного фактора от другого, принимая при этом значения от 0 до 1. Он имеет две формы: теоретическую и выборочную. Теоретический коэффициент корреляции традиционно обозначается греческой буквой p, которая произносится как «ро» и соответствует латинской «r». Для переменных x и y этот коэффициент определяется следующим образом:

𝜎𝑥,𝑦

𝑝𝑥,𝑦 

Выборочный коэффициент корреляции r определяется путем замены теоретических дисперсий и ковариации в формуле теоретической корреляции на их несмещенные оценки, т.е. формула приобретает следующий вид:

𝑐𝑜𝑣(𝑥, 𝑦)

𝑟𝑥,𝑦 

Цель корреляционного анализа – обеспечить получение некоторой информации об одной переменной с помощью другой переменной. В случаях, когда возможно достижение цели, говорят, что переменные коррелируют. В самом общем виде принятие гипотезы о наличии корреляции означает что изменение значения переменной А, произойдет одновременно с пропорциональным изменением значения Б: если обе переменные растут, то корреляция положительная, если одна переменная растёт, а вторая уменьшается, корреляция отрицательная.

Кроме того, при проведении корреляционного анализа стоит учитывать, что некоторые признаки могут иметь абсурдную связь между друг другом. Для объяснения этого явления введено понятие ложной корреляции.

## Регрессионный анализ

Регрессионный анализ - это метод установления аналитического выражения стохастической зависимости между исследуемыми признаками. Уравнение регрессии показывает, как в среднем изменяется y (зависимая переменная, описывающая процесс, который мы пытаемся предсказать или понять) при изменении любого из xi, и имеет вид:

𝑦 = 𝛽0 + 𝛽1 ∗ 𝑥1 + 𝛽2 ∗ 𝑥2 + ⋯ + 𝛽𝑛 ∗ 𝑥𝑛 + 𝑢

Величина y, рассматриваемая как зависимая переменная, состоит из двух составляющих:

1) неслучайной составляющей β0+β1х1+…+βnхn, где хi выступает как объясняющая (или независимая) переменная, а постоянные величины β — это коэффициенты, которые рассчитываются в результате выполнения регрессионного анализа. Вычисляются величины для каждой независимой переменной, которые представляют силу и тип взаимосвязи независимой переменной по отношению к зависимой; 2) случайного члена u.

Если уравнение регрессии имеет только одну независимую переменную, то это простой регрессионный анализ. Если n≥2, то такой анализ называется многофакторным.

В ходе регрессионного анализа решаются две основные задачи:

−построение уравнения регрессии, т.е. нахождение вида зависимости между результатным показателем и независимыми факторами *x*1, *x*2, …, *xn.*

−оценка значимости полученного уравнения, т.е. определение того, насколько выбранные факторные признаки объясняют вариацию признака *у.*

В отличие от корреляционного анализа, который только отвечает на вопрос, существует ли связь между анализируемыми признаками, регрессионный анализ дает и ее формализованное выражение. Кроме того, если корреляционный анализ изучает любую взаимосвязь факторов, то регрессионный - одностороннюю зависимость, т.е. связь, показывающую, каким образом изменение факторных признаков влияет на признак результативный.

## Канонический анализ

Канонический анализ разработан для анализа зависимостей между двумя совокупностями признаков, характеризующих объекты. Канонический анализ является обобщением множественной корреляции как меры связи между одной переменной и множеством других переменных. Как известно, множественная корреляция есть максимальная корреляция между одной переменной и линейной функцией других переменных. Пусть, например, первое множество переменных состоит из признаков у1, …, ур, второе множество состоит из – х1, …, хq, тогда взаимосвязь между данными множествами можно оценить как корреляцию между линейными комбинациями a1y1 + a2y2 + ... + apyp, b1x1 + b2x2 + ... + bqxq,, которая называется канонической корреляцией. Задача канонического анализа в нахождении весовых коэффициентов таким образом, чтобы каноническая корреляция была максимальной. Веса (коэффициенты) этих линейных комбинаций могут быть найдены как решение задачи на собственные значения. Получаемые при этом решения называются каноническими корнями, каждое из которых объясняет уникальную долю изменчивости между двумя наборами переменных. Анализ находит новые собственные значения по шагам, на каждом шаге максимизируя корреляцию между каноническими переменными. Первые несколько пар канонических переменных в Каноническом Анализе обычно объясняют наибольшую долю различия между двумя множествами переменных.

## Метод сравнения средних

Сравнение средних значений различных выборок относится к наиболее часто применяемым методам статистического анализа, так как в прикладных исследованиях часто встречаются случаи, когда средний результат некоторого признака одной серии экспериментов отличается от среднего результата другой серии. Так как средние это результаты измерений, то, как правило, они всегда различаются, вопрос в том, можно ли объяснить обнаруженное расхождение средних неизбежными случайными ошибками эксперимента или оно вызвано определенными причинами. При этом всегда должен быть выяснен вопрос, можно ли объяснить имеющееся различие средних значений статистическими колебаниями или нет. Если нет, говорят о статистически значимом различии между сравниваемыми группами.

При сравнении средних значений выборок предполагается, что обе выборки подчиняются нормальному распределению. Если это не так, то вычисляются медианы и для сравнения выборок используется непараметрические тесты.

При данном анализе выделяют четыре различные тестовые ситуации и в этих ситуациях применяются следующие статистические тесты:

− Сравнение двух независимых выборок. T-тест для независимых выборок (тест Стьюдента)

− Сравнение двух зависимых (спаренных) выборок. T-тест для зависимых выборок

− Сравнение более двух независимых выборок. Однофакторный дисперсионный анализ (ANOVA)

− Сравнение более двух зависимых выборок. Однофакторный дисперсионный анализ с повторными измерениями (MANOVA)

## Частотный анализ

Частотный анализ (или метод одновходовых таблиц) представляет собой простейший метод анализа категориальных переменных. Таблицы частот могут быть использованы также для исследования количественных переменных, хотя при этом могут возникнуть трудности с интерпретацией результатов. Часто их используют как одну из процедур разведочного анализа, чтобы просмотреть, каким образом различные группы данных распределены в выборке. Для иллюстрации таблиц частот применяют гистограммы.

## Методы, основанные на объединении объектов исследования

Кластерный анализ – это метод классификации многомерных наблюдений на основе определения сходства или близости (расстояния) между объектами. Цель кластерного анализа заключается в определении однородных в некотором смысле групп, которые называются кластерами. Кластерный анализ применяется в самых различных задачах, где необходимо исследовать структуру совокупности и выполнить классификацию объектов.

Наиболее востребовано использование кластерного анализа для исследований в медицине, биологии, психологии, социологии и экономике.

Алгоритм кластерного анализа включает пять этапов:

− представление исходных данных в виде матрицы (таблицы

"объект – признак").

− определение сходства объектов.

− выбор метода объединения объектов в кластеры.

− определение оптимального числа кластеров.

− интерпретация кластеров и качества разбиения.

Деревья классификации – это метод классификационного анализа, позволяющий предсказывать принадлежность объектов к тому или иному классу в зависимости от соответствующих значений признаков, характеризующих объекты. Признаки называются независимыми переменными, а переменная, указывающая на принадлежность объектов к классам, называется зависимой.

В отличие от классического дискриминантного анализа, деревья классификации способны выполнять одномерное ветвление по переменным различных типов − категориальным, порядковым, интервальным. Не накладываются какие-либо ограничения на закон распределения количественных переменных. По аналогии с дискриминантным анализом метод дает возможность анализировать вклады отдельных переменных в процедуру классификации.

Деревья классификации могут быть, а иногда и бывают, очень сложными. Однако использование специальных графических процедур позволяет упростить интерпретацию результатов даже для очень сложных деревьев.

Возможность графического представления результатов и простота интерпретации во многом объясняют большую популярность деревьев классификации в прикладных областях, однако наиболее важные отличительные свойства деревьев классификации – их иерархичность и широкая применимость.

Структура метода такова, что пользователь имеет возможность по управляемым параметрам строить деревья произвольной сложности, добиваясь минимальных ошибок классификации. Но по сложному дереву, изза большой совокупности решающих правил, затруднительно классифицировать новый объект. Поэтому при построении дерева классификации пользователь должен найти разумный компромисс между сложностью дерева и трудоемкостью процедуры классификации.

## Дискриминантный анализ

Дискриминантный анализ - это раздел математической статистики, содержанием которого является разработка методов решения задач различения (дискриминации) объектов наблюдения по определенным признакам. Например, разбиение совокупности предприятий на несколько однородных групп по значениям каких-либо показателей производственнохозяйственной деятельности.

Методы дискриминантного анализа находят применение в различных областях: медицине, социологии, психологии, экономике и т.д. При наблюдении больших статистических совокупностей часто появляется необходимость разделить неоднородную совокупность на однородные группы (классы). Такое расчленение в дальнейшем при проведении статистического анализа дает лучшие результаты моделирования зависимостей между отдельными признаками.

Дискриминантным анализом можно воспользоваться в тех случаях, когда возникает необходимость отнесения того или иного объекта к одному из реально существующих или выделенных определенным способом классов. Все процедуры дискриминантного анализа делятся на две группы. Первая группа процедур позволяет интерпретировать различия между существующими классами, вторая - проводить классификацию новых объектов в тех случаях, когда неизвестно заранее, к какому из существующих классов они принадлежат.

## Факторный анализ

Факторный анализ представляет собой совокупность методов, которые на основе реально существующих связей анализируемых признаков, связей самих наблюдаемых объектов, позволяют выявлять скрытые (неявные, латентные) обобщающие характеристики организационной структуры и механизма развития изучаемых явлений, процессов.

Методы факторного анализа в исследовательской практике применяются главным образом с целью сжатия информации, получения небольшого числа обобщающих признаков, объясняющих вариативность (дисперсию) элементарных признаков (R-техника факторного анализа) или вариативность наблюдаемых объектов (Q-техника факторного анализа). Алгоритмы факторного анализа основываются на использовании редуцированной матрицы парных корреляций (ковариаций). Редуцированная матрица – это матрица, на главной диагонали которой расположены не единицы (оценки) полной корреляции или оценки полной дисперсии, а их редуцированные, несколько уменьшенные величины. При этом постулируется, что в результате анализа будет объяснена не вся дисперсия изучаемых признаков (объектов), а ее некоторая часть, обычно большая. Оставшаяся необъясненная часть дисперсии — это характерность, возникающая из-за специфичности наблюдаемых объектов, или ошибок, допускаемых при регистрации явлений, процессов, т.е. ненадежности вводных данных.

Главными целями факторного анализа являются:

− сокращение числа переменных (редукция данных)

− определение структуры взаимосвязей между переменными, т.е. классификация переменных.

Поэтому факторный анализ используется или как метод сокращения данных или как метод классификации.

Анализ соответствий содержит описательные и разведочные методы анализа двухвходовых и многовходовых таблиц. Эти методы по своей природе похожи на методы Факторного анализа и позволяют исследовать структуру группирующих переменных, включенных в таблицу. Одной из наиболее общих разновидностей многовходовых таблиц типа являются частотные таблицы сопряженности.

В классическом анализе соответствий частоты в таблице сопряженности стандартизуются таким образом, чтобы сумма наблюдений во всех ячейках была равна 1. Одной из целей анализа соответствий является представление содержимого таблицы относительных частот в виде расстояний между отдельными строками и/или столбцами таблицы в пространстве возможно более низкой размерности.

## Анализ главных компонент и классификация

Метод анализ главных компонент и классификация позволяет решить задачу анализа данных большой размерности и служит для достижения двух целей:

* уменьшение общего числа переменных для того, чтобы получить «главные» и «некоррелирующие» переменные;
* классификация переменных и наблюдений, при помощи строящегося факторного пространства.

Метод имеет сходство с факторным анализом в постановочной части решаемых задач, но имеет ряд существенных отличий:

* при анализе главных компонент не используются итеративные методы для извлечения факторов;
* наряду с активными переменными и наблюдениями, используемыми для извлечения главных компонент, можно задать вспомогательные переменные и/или наблюдения; затем вспомогательные переменные и наблюдения проектируются на факторное пространство, вычисленное на основе активных переменных и наблюдений;
* перечисленные возможности позволяют использовать метод как мощное средство для классификации одновременно переменных и наблюдений. Решение основной задачи метода достигается созданием векторного пространства скрытых переменных с размерностью меньше исходной. Исходная размерность определяется числом переменных для анализа в исходных данных.

## Многомерное шкалирование

Многомерное шкалирование применимо к данным, для которых непригодно большинство обычных методов факторизации. Основной тип данных в многомерном шкалировании представляет собой меру близости или различия между двумя объектами. Данные, используемые для анализа методом многомерного шкалирования, часто получают из матриц парных сравнений объектов. По ним требуется восстановить неизвестную размерность р анализируемого признакового пространства и приписать каждому объекту координаты таким образом, чтобы вычисленные попарные евклидовы расстояния между объектами по возможности совпали бы с заданными матрицей парных сравнений.

Целью метода многомерного шкалирования является отображение информации о конфигурации исходных многомерных данных, заданных матрицей парных сравнений, в виде конфигурации точек в соответствующем пространстве меньшей размерности.

Задачи многомерного шкалирования:

− построение систем координат, в которых представляются исследуемые объекты, и содержательная интерпретацию этих шкал в качестве вполне определенных характеристик, существенных для дифференциации объектов в рассматриваемом смысле. В такой постановке конечные прикладные цели многомерного шкалирования, по существу, не отличаются от задач, для решения которых используется факторный анализ.

− верификация геометрической конфигурации системы анализируемых объектов в координатном пространстве латентных переменных. Речь идет о ситуациях, в которых исходя из содержательных соображений, относящихся к механизму изучаемого явления, формулируются гипотезы о размерности пространства скрытых переменных и о типе геометрической конфигурации системы точек, представляющих анализируемые объекты в этом пространстве. Результатом применения многомерного шкалирования в задачах этого типа является статистическая проверка гипотез.

Часто в качестве основной цели многомерного шкалирования рассматривается сжатие исходного массива данных с минимальными потерями в их информативности. Если сопоставляется большое число объектов, то размерность квадратной матрицы парных сравнений может быть велика. Решение задачи многомерного шкалирования позволяет перейти от формы исходных данных типа "объект – объект" к более распространенной и удобной для статистической обработки форме исходных данных типа "объект – признак" с одновременным сокращением объема исходных данных.

Необходимо подчеркнуть одно очень важное свойство многомерного шкалирования: независимо от того, какого типа задача решается, его всегда можно использовать как инструмент визуализации исходных данных.

## Моделирование структурными уравнениями

Моделирования структурными уравнениями – это всеобъемлющая и необычайно мощная техника многомерного анализа включает большое количество методов из различных областей статистики. Кратко можно сказать, что SEPATH представляет собой мощное развитие многих методов многомерного анализа, а именно множественная регрессия и [факторный анализ](http://statsoft.ru/home/textbook/modules/stfacan.html) получили здесь естественное развитие и объединение.

Основные задачи, для решения которых используются структурные уравнения следующие:

− причинное моделирование или анализ путей, при проведении которого предполагается, что между переменными имеются причинные взаимосвязи. Возможна проверка гипотез и подгонка параметров причинной модели, описываемой линейными уравнениями. Причинные модели могут включать явные или латентные переменные, или и те и другие;

− подтверждающий факторный анализ, используемый как развитие обычного факторного анализа для проверки определенных гипотез о структуре факторных нагрузок и корреляций между факторами;

− факторный анализ второго порядка, являющийся модификацией факторного анализа, при проведении которого для получения факторов второго порядка анализируется корреляционная матрица общих факторов;

 − регрессионные модели, являющиеся модификацией

Многомерного линейного регрессионного анализа, в котором коэффициенты регрессии могут быть зафиксированы равными друг другу или каким-нибудь заданным значениям;

− моделирование ковариационной структуры, которое позволяет проверить гипотезу о том, что матрица ковариации имеет определенный вид. Например, с помощью этой процедуры вы можете проверить гипотезу о равенстве дисперсий у всех переменных;

− модели корреляционной структуры, которое позволяет проверить гипотезу о том, что матрица корреляции имеет определенный вид.

Классическим примером является гипотеза о том, что матрица корреляции имеет циклическую структуру;

− модели структуры средних, которые позволяют исследовать структуру средних, например, одновременно с анализом дисперсий и ковариаций.

Структурные уравнения, включающие только линейные связи между явными и латентными переменными, могут быть изображены в виде диаграмм путей.

## Карты контроля качества

На сегодняшний день актуальными являются вопросы разработки методов контроля качества, которые позволят своевременно выявить признаки нарушения технологического процесса или оказания услуг. Вообще говоря, существует два "врага" качества продукции:

1. отклонения от плановых спецификаций;
2. слишком большой разброс реальных характеристик изделий (относительно плановых спецификаций).

При этом, для достижения и поддержания высокого уровня качества, удовлетворяющего потребителя нужны методы, направленные не на устранение дефектов готовой продукции и несоответствий услуг, а на предупреждение и прогнозирование причин их появления. Контрольная карта – это инструмент, позволяющий отслеживать ход протекания процесса и воздействовать на него, предупреждая его отклонения от предъявленных к процессу требований. Инструментарий карт контроля качества широко использует статистические методы, основанные на теории вероятностей и математической статистики. Применение статистических методов позволяет при ограниченных объемах анализируемых изделий с заданной степенью точности и достоверности судить о состоянии качества выпускаемой продукции. Обеспечивает прогнозирование, оптимальное регулирование проблем в области качества, принятие верных управленческих решений не на основе интуиции, а при помощи научного изучения и выявления закономерностей в накапливаемых массивах числовой информации.

# АНАЛИЗ ДАННЫХ

Применяя основные методы математической статистики, проанализируем автомобильный рынок Краснодарского края в 2017 году.

Исследования будут проводиться на основе данных представленных в таблице (Приложение А).

Проведем 3 метода статистического анализа, такие как корреляционный анализ, регрессионный анализ и частотный анализ. С их помощью выявим основные зависимости между обрабатываемыми переменными и спрогнозируем возможные значения, построив уравнение регрессии.



Рисунок 1 – Фрагмент таблицы исходных данных в STATISTICA

|  |  |  |
| --- | --- | --- |
| Variable | Correlations (Rinok\_auto)Marked correlations are significant at p < ,05000 N=61 (Casewise deletion of missing data) |  |
| Means | Std.Dev. | производство | состояние | продано | стоимость |
|  | производство | 1 | 0 | 1,000000 | -0,065787 | -0,621618 | 0,202763 |
| состояние | 0 | 1 | -0,065787 | 1,000000 | 0,325038 | -0,161121 |
| продано | 150 | 288 | -0,621618 | 0,325038 | 1,000000 | -0,215832 |
| стоимость | 2192377 | 4189531 | 0,202763 | -0,161121 | -0,215832 | 1,000000 |

Рисунок 2 – Матрица корреляций

Для выявления зависимостей между анализируемыми переменными и нахождения их значимости воспользуемся корреляционным анализом. Целью корреляционного анализа является получение информации об одной переменной с помощью другой переменной. Коэффициент корреляции указывает на степень зависимости одного фактора от другого, принимая при этом значения от 0 до 1.

Сделав данный анализ, на рисунке 2 можно наблюдать, что существует статистически значимая взаимосвязь между производством и количеством проданных автомобилей, а также между состоянием авто и количеством продаж. Остальные корреляции являются незначимыми при уровне значимости p<0,05.

Так как явной корреляционной зависимости между продажами и стоимостью в проведенном анализе не оказалось, воспользуемся корреляционным непараметрическим анализом. Исследуем зависимость по коэффициенту Спирмена.

На рисунке 3 видно, что количество продаж статистически значимо зависит от стоимости.

 Spearman Rank Order Correlations (Rinok\_aut

MD pairwise deleted

 Marked correlations are significant at p <,05000

Variable продано стоимость продано1,000000 -0,880575 стоимость -0,880575 1,000000

Рисунок 3 – Матрица корреляций по критерию Спирмена

Используя такой метод математической статистики, как регрессионный анализ, можно выявить функциональную зависимость количества продаж от стоимости автомобилей.

Построим показательную регрессию. В качестве зависимой переменной выступает «продано», а независимая «стоимость».

|  |  |
| --- | --- |
| N=61 | Regression Summary for Dependent Variable: LN-V5 (Rinok\_auR= ,87941033 RІ= ,77336254 Adjusted RІ= ,76952122F(1,59)=201,33 p<0,0000 Std.Error of estimate: ,78426 |
| b\* | Std.Err. of b\* | b | Std.Err. of b | t(59) | p-value |
|  | Intercept |  |  | 22,73272 | 1,326242 | 17,1407 | 0,000000 |
| LN-V6 | -0,879410 | 0,061978 | -1,35061 | 0,095187 | -14,1890 | 0,000000 |

Рисунок 4 – Таблица линеаризованного уравнения показательной регрессии

Пояснение переменных используемых в при построении таблицы парной регрессии:

LN-V6 – натуральный логарифм переменной «стоимость»; LN-V5 – натуральный логарифм переменной «продано».

Уравнение регрессии: *продано=22,73\*стоимость-1,35 ln(продано)=3,12-1,35\*ln(стоимость)*

Исходя из таблицы регрессии можно заключить, что при увеличении уровня цен в среднем на 1% количество проданных автомобилей в Краснодарском крае уменьшается в среднем на 1,35%.

Так как F-критерий=201,33, то можно утверждать, что построенная модель регрессии имеет высокую значимость.

В нашем случае коэффициент детерминации R2 очень близок к 1, т.е. линия регрессии довольно точно описывает наблюдения, и связь между независимой и зависимой переменными очень существенна. Из этого следует, что доля дисперсии, объясненной регрессией, равна 0,773.

В свою очередь, коэффициент множественной корреляции R равный 0,879, показывает на сколько хорошо подобрано уравнение регрессии.

Стандартная ошибка оценки=0,785, что говорит о мере рассеяния наблюдаемых значений относительно регрессионной прямой.

Уровень значимости p существенно меньший 0,05 показывает о значимости всего уравнения регрессии.

Кроме того, в построенной модели найденные коэффициенты регрессии имеют высокие значения t-критерия, что говорит о значимости коэффициентов корреляции, и являются статистически значимыми при уровне значимости 0,05.

Для визуализации частот распределения переменных воспользуемся частотным анализом данных.

Проведем его, используя переменные: марка, производство, состояние.

|  |  |
| --- | --- |
| Category | Frequency table: марка (Rinok\_auto) |
| Count | Cumulative Count | Percent | Cumulative Percent |
|  | Bentley | 3 | 3 | 4,91803 | 4,9180 |
| Maserati | 4 | 7 | 6,55738 | 11,4754 |
| LADA | 11 | 18 | 18,03279 | 29,5082 |
| Ford | 5 | 23 | 8,19672 | 37,7049 |
| KIA | 6 | 29 | 9,83607 | 47,5410 |
| Hyundai | 7 | 36 | 11,47541 | 59,0164 |
| Toyota | 5 | 41 | 8,19672 | 67,2131 |
| Volkswagen | 6 | 47 | 9,83607 | 77,0492 |
| Renault | 5 | 52 | 8,19672 | 85,2459 |
| Audi | 5 | 57 | 8,19672 | 93,4426 |
| Nissan | 4 | 61 | 6,55738 | 100,0000 |
| Missing | 0 | 61 | 0,00000 | 100,0000 |

Рисунок 5 – Таблица частот по маркам автомобилей

|  |  |
| --- | --- |
| Category | Frequency table: производство (Rinok\_auto |
| Count | Cumulative Count | Percent | Cumulative Percent |
|  | отечественный | 11 | 11 | 18,03279 | 18,0328 |
| иномарка | 50 | 61 | 81,96721 | 100,0000 |
| Missing | 0 | 61 | 0,00000 | 100,0000 |

Рисунок 6 – Таблица частот по производству

|  |  |
| --- | --- |
| Category | Frequency table: состояние (Rinok\_auto) |
| Count | Cumulative Count | Percent | Cumulative Percent |
|  | новый | 32 | 32 | 52,45902 | 52,4590 |
| поддержанный | 29 | 61 | 47,54098 | 100,0000 |
| Missing | 0 | 61 | 0,00000 | 100,0000 |

Рисунок 7 – Таблица частот по состоянию автомобилей

На рисунках 5-7 можно увидеть частоту распределения и процентную составляющую значений представленных переменных.

Рисунок 8 дает возможность увидеть распределение различных марок автомобилей по их стоимости и по количествам продаж. Из данной диаграммы видно, что наибольшая цена автомобиля марки Bentley, а наибольшие продажи принадлежат марке LADA.



Рисунок 8 – Диаграмма рассеяния различных марок автомобилей по количеству проданных автомобилей и их стоимости

# ЗАКЛЮЧЕНИЕ

Изучены основные методы математической статистики. Построено уравнение регрессии, предсказывающее уровень продаж автомобилей при заданных значениях. Были найдены статистически значимые зависимости между производством и количеством проданных автомобилей, состоянием авто и количеством продаж, между стоимостью автомобилей и количеством продаж.

В результате проделанной работы при помощи прикладного программного обеспечения STATISTICA был проанализирован

автомобильный рынок Краснодарского края в 2017 году.

Проведенные анализы имеют значительное практическое применение в области автомобилестроения и во всей автоиндустрии.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Халафян А.А. STATISTICA 6 Статистический анализ данных, Москва, изд. «Бином-Пресс», 2007 – 512 с.
2. Доугерти К. Введение в эконометрику, Нью-Йорк, 1992. Пер. с англ. Е.Н. Лукаш, О.Ю. Шибанкин, О.О. Замков, под ред. О.О. Замков, изд.

«Инфра-М», 1999 – 416 с.

1. Елисеева И.И. Эконометрика, изд. «Финансы и статистика», 2003 – 344 с.
2. Рынок новых легковых автомобилей за 10 месяцев 2017 года. URL:

<https://www.autostat.ru/press-releases/32200/>(дата обращения: 29.10.2017)

1. В Краснодарском крае за 7 месяцев 2017г. продали 9 Bentley и 7 Maserati.

URL: <https://kuban.rbc.ru/krasnodar/freenews/59a55baf9a7947c7ad10cbde>

(дата обращения: 29.10.2017)

1. Продажа автомобилей в Краснодарском крае. URL: <http://auto.drom.ru/region23>(дата обращения: 29.10.2017)
2. Обзор методов статистического анализа данных. URL: <http://statlab.kubsu.ru/node/4>(дата обращения: 12.11.2017)

ПРИЛОЖЕНИЕ А

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Марка автомобиля  | Модель автомобиля  | Производство автомобиля  | Состояние втомобиля  | Продано, шт  | Стоимость, руб.  |
| Bentley  | Continental GT  | иномарка  | новый  | 5  | 16330000  |
| Bentley  | Mulsanne  | иномарка  | новый  | 1  | 22750000  |
| Bentley  | Mulsanne  | иномарка  | поддержанный  | 3  | 19900000  |
| Maserati  | Quattroporte  | иномарка  | поддержанный  | 1  | 2260000  |
| Maserati  | Quattroporte  | иномарка  | новый  | 1  | 2500000  |
| Maserati  | GranTurismo  | иномарка  | поддержанный  | 3  | 3285000  |
| Maserati  | GranTurismo  | иномарка  | новый  | 2  | 3900000  |
| LADA  | 2107  | отечественный  | поддержанный  | 1617  | 92000  |
| LADA  | 2114  | отечественный  | поддержанный  | 1190  | 177000  |
| LADA  | 2109  | отечественный  | поддержанный  | 1182  | 80000  |
| LADA  | Vesta  | отечественный  | новый  | 297  | 730000  |
| LADA  | X-Ray  | отечественный  | новый  | 85  | 815000  |
| LADA  | Largus  | отечественный  | новый  | 222  | 460000  |
| LADA  | Largus  | отечественный  | поддержанный  | 305  | 393000  |
| LADA  | Priora  | отечественный  | новый  | 143  | 423000  |
| LADA  | Priora  | отечественный  | поддержанный  | 334  | 347000  |
| LADA  | Kalina  | отечественный  | новый  | 97  | 542000  |
| LADA  | Kalina  | отечественный  | поддержанный  | 346  | 299000  |
| Ford  | Focus  | иномарка  | новый  | 43  | 894000  |
| Ford  | Focus  | иномарка  | поддержанный  | 101  | 586000  |
| Ford  | Fiesta  | иномарка  | новый  | 56  | 774000  |
| Ford  | Fiesta  | иномарка  | поддержанный  | 97  | 437000  |
| Ford  | Explorer  | иномарка  | новый  | 11  | 2840000  |
| KIA  | Sportage  | иномарка  | новый  | 29  | 1793000  |
| KIA  | Sportage  | иномарка  | поддержанный  | 44  | 1083000  |
| KIA  | Rio  | иномарка  | новый  | 51  | 634000  |
| KIA  | Rio  | иномарка  | поддержанный  | 85  | 450000  |
| KIA  | Cee'd  | иномарка  | новый  | 38  | 974000  |
| KIA  | Cee'd  | иномарка  | поддержанный  | 70  | 665000  |
| Hyundai  | i40  | иномарка  | новый  | 61  | 1475000  |
| Hyundai  | i40  | иномарка  | поддержанный  | 87  | 788000  |
| Hyundai  | Solaris  | иномарка  | новый  | 97  | 884000  |
| Марка автомобиля  | Модель автомобиля  | Производство автомобиля  | Состояние втомобиля  | Продано, шт  | Стоимость, руб.  |
| Hyundai  | Solaris  | иномарка  | поддержанный  | 228  | 575000  |
| Hyundai  | Accent  | иномарка  | поддержанный  | 320  | 480000  |
| Hyundai  | H1  | иномарка  | новый  | 39  | 2139000  |
| Hyundai  | H1  | иномарка  | поддержанный  | 98  | 1270000  |
| Toyota  | Camry  | иномарка  | новый  | 22  | 1736000  |
| Toyota  | Camry  | иномарка  | поддержанный  | 52  | 1099000  |
| Toyota  | Corolla  | иномарка  | новый  | 40  | 1368000  |
| Toyota  | Corolla  | иномарка  | поддержанный  | 115  | 815000  |
| Toyota  | Tundra  | иномарка  | новый  | 4  | 5255000  |
| Volkswagen  | Polo  | иномарка  | новый  | 68  | 837000  |
| Volkswagen  | Polo  | иномарка  | поддержанный  | 75  | 550000  |
| Volkswagen  | Tiguan  | иномарка  | новый  | 67  | 1997000  |
| Volkswagen  | Tiguan  | иномарка  | поддержанный  | 103  | 1160000  |
| Volkswagen  | Passat  | иномарка  | новый  | 53  | 2162000  |
| Volkswagen  | Passat  | иномарка  | поддержанный  | 88  | 1024000  |
| Renault  | Logan  | иномарка  | новый  | 151  | 679000  |
| Renault  | Logan  | иномарка  | поддержанный  | 234  | 410000  |
| Renault  | Kapture  | иномарка  | новый  | 62  | 1133000  |
| Renault  | Duster  | иномарка  | новый  | 89  | 1052000  |
| Renault  | Duster  | иномарка  | поддержанный  | 186  | 670000  |
| Audi  | A4  | иномарка  | новый  | 19  | 2390000  |
| Audi  | A4  | иномарка  | поддержанный  | 34  | 1100000  |
| Audi  | A7  | иномарка  | новый  | 7  | 4616000  |
| Audi  | Q5  | иномарка  | новый  | 16  | 3680000  |
| Audi  | Q5  | иномарка  | поддержанный  | 28  | 1730000  |
| Nissan  | Qashqai  | иномарка  | новый  | 42  | 1623000  |
| Nissan  | Qashqai  | иномарка  | поддержанный  | 79  | 780000  |
| Nissan  | Juke  | иномарка  | новый  | 64  | 1165000  |
| Nissan  | Juke  | иномарка  | поддержанный  | 83  | 680000  |