Министерство науки и высшего образования Российской Федерации

Новосибирский Государственный технический университет

Кафедра автоматизированных систем управления



**ОТЧЁТ**

**Отчет по лабораторной работе 1**

**по дисциплине «Методы анализа данных»**

**«Предварительный анализ данных, дисперсионный анализ, корреляционный анализ, регрессионный анализ в среде MS Excel»**

Выполнили

студенты группы АВТ-812:

Березин Дмитрий

Бородина Алина

Глинин Евгений

Преподаватель:

Ганелина Наталья Давидовна,

к.т.н., доцент кафедры АСУ

г. Новосибирск

2021 г.

Содержание

[Цель работы 3](#_Toc84280187)

[Текст задания 3](#_Toc84280188)

[Описательная статистика. Корреляционный анализ 6](#_Toc84280189)

[Пример и интерпретация ложной корреляции 12](#_Toc84280190)

[Регрессионная модель. Уравнение регрессии, график остатков, 14](#_Toc84280193)

[гистограмма остатков. Оценка качества построенной модели 14](#_Toc84280194)

[Регрессионный анализ 17](#_Toc84280195)

[Дисперсионный анализ 19](#_Toc84280196)

# Цель работы

Изучение корреляционного анализа, явлений ложной корреляции, регрессионного анализа, однофакторного дисперсионного анализа.

# Текст задания

**Описательная статистика. Корреляционный анализ**

1. Выбрать массив данных, описать параметры, зависимые и независимые переменные.

2. Провести дескриптивный анализ, оценить близость выборок к нормальной.

3. Построить гистограммы выбранных атрибутов. Длину интервала рассчитать с помощью формулы Стерджесса.

4. Изучить пример построения таблиц сопряженности (п. 4 из списка литературы). Выбрать в рассматриваемой выборке категориальные переменные или перевести числовые переменные в категориальные, провести анализ с помощью таблиц сопряженности. Описать результаты.

5. Провести корреляционный анализ. Если рассматриваемые выборки далеки от нормального распределения – провести ранговый корреляционный анализ.

6. Оценить значимость корреляции.

7. Выделить сильно- и слабокоррелированные признаки.

8. Интерпретировать результаты.

9. На сайте http://www.tylervigen.com/spurious-correlations выбрать массив данных, построить графики, оценить ложную корреляцию между параметрами.

**Регрессионный анализ**

10. Провести корреляционный анализ. Если рассматриваемые выборки далеки от нормального распределения – провести ранговый корреляционный анализ.

11. Оценить значимость корреляции.

12. Выделить сильно- и слабокоррелированные признаки. Построить диаграммы рассеивания для выбранных признаков.

13. Интерпретировать результаты.

14. Построить линейную регрессионную модель для признаков с высоким коэффициентом корреляции.

15. Вывести график остатков. Оценить постоянство среднего и дисперсии.

16. Построить гистограмму стандартизированных остатков.

17. Записать уравнение регрессии.

18. По значениям коэффициента детерминации, графику и гистограмме остатков оценить качество построенной модели.

19. Оценить значимость построенной модели.

**Дисперсионный анализ**

20. Сформулировать начальную гипотезу.

21. Для независимой переменной (фактора) определить градации, разбить значения зависимой переменной в соответствии с градациями фактора. Если фактор – не категориальная переменная, а числовая, аргументировать разбиение ее значений на интервалы.

22. Провести однофакторный дисперсионный анализ.

23. Подтвердить или опровергнуть выдвинутую гипотезу. Интерпретировать результаты.

24. Дополнительное задание. Провести двухфакторный анализ.

# Описательная статистика. Корреляционный анализ

Выбрали массив данных «Automobile Data Set». Этот набор данных состоит из трех типов объектов: (а) спецификация автомобиля с точки зрения различных характеристик, (б) присвоенный ему рейтинг страхового риска, (в) его нормализованные потери при эксплуатации по сравнению с другими автомобилями. Данные включают в следующее:

1. Рейтинг: -3, -2, -1, 0, 1, 2, 3.

2. нормализованные потери.

3. Производитель.

4. Вид топлива.

5. стремление.

6. Кол-во дверей.

7. Кузов.

8. ведущие колеса.

9. Расположение двигателя.

10. Колесная база.

11. длина.

12. ширина.

13. высота.

14. Снаряженная масса.

15. Тип двигателя.

16. Число цилиндров.

17. Объем двигателя.

18. Топливная система.

19. Диаметр цилиндра.

20. ход поршня.

21. Степень сжатия.

22. лошадиные силы.

23. пиковая частота вращения.

24. город-миль на галлон.

25. шоссе-миль на галлон.

26. Цена.

Зависимые переменные: Тип кузова автомобиля (convertible, hardtop, hatchback, sedan, wagon).

Независимые переменные: Размер двигателя автомобиля (61-258).

**Результаты дескриптивного анализа**

|  |  |
| --- | --- |
| **engine-size** | |
|  |
| Среднее | 119,2264151 |  |
| Стандартная ошибка | 2,415697881 |  |
| Медиана | 110 |  |
| Мода | 92 |  |
| Стандартное отклонение | 30,46079126 |  |
| Дисперсия выборки | 927,8598042 |  |
| Эксцесс | 2,950273357 |  |
| Асимметричность | 1,490609644 |  |
| Интервал | 197 |  |
| Минимум | 61 |  |
| Максимум | 258 |  |
| Сумма | 18957 |  |
| Счет | 159 |  |

В таблице видно, что средний размер двигателя – 119,2264151. Максимальный размер двигателя – 258. Исходя из значения моды, можно утверждать, что большинство двигателей обладают размером – 92.

Коэффициент эксцесса – положительный, значит, распределение является островершинным. Положительный знак коэффициента асимметрии говорит о наличии правосторонней асимметрии.

ВЫВОД

|  |  |
| --- | --- |
| **horsepower** | |
|  |
| Среднее | 95,83647799 |  |
| Стандартная ошибка | 2,436142067 |  |
| Медиана | 88 |  |
| Мода | 68 |  |
| Стандартное отклонение | 30,71858264 |  |
| Дисперсия выборки | 943,6313192 |  |
| Эксцесс | 0,298877231 |  |
| Асимметричность | 0,916662556 |  |
| Интервал | 152 |  |
| Минимум | 48 |  |
| Максимум | 200 |  |
| Сумма | 15238 |  |
| Счет | 159 |  |

В таблице видно, что средняя мощность двигателя – 95,83647799. Максимальная мощность – 200. Исходя из значения моды, можно утверждать, что большинство двигателей обладают мощностью – 68.

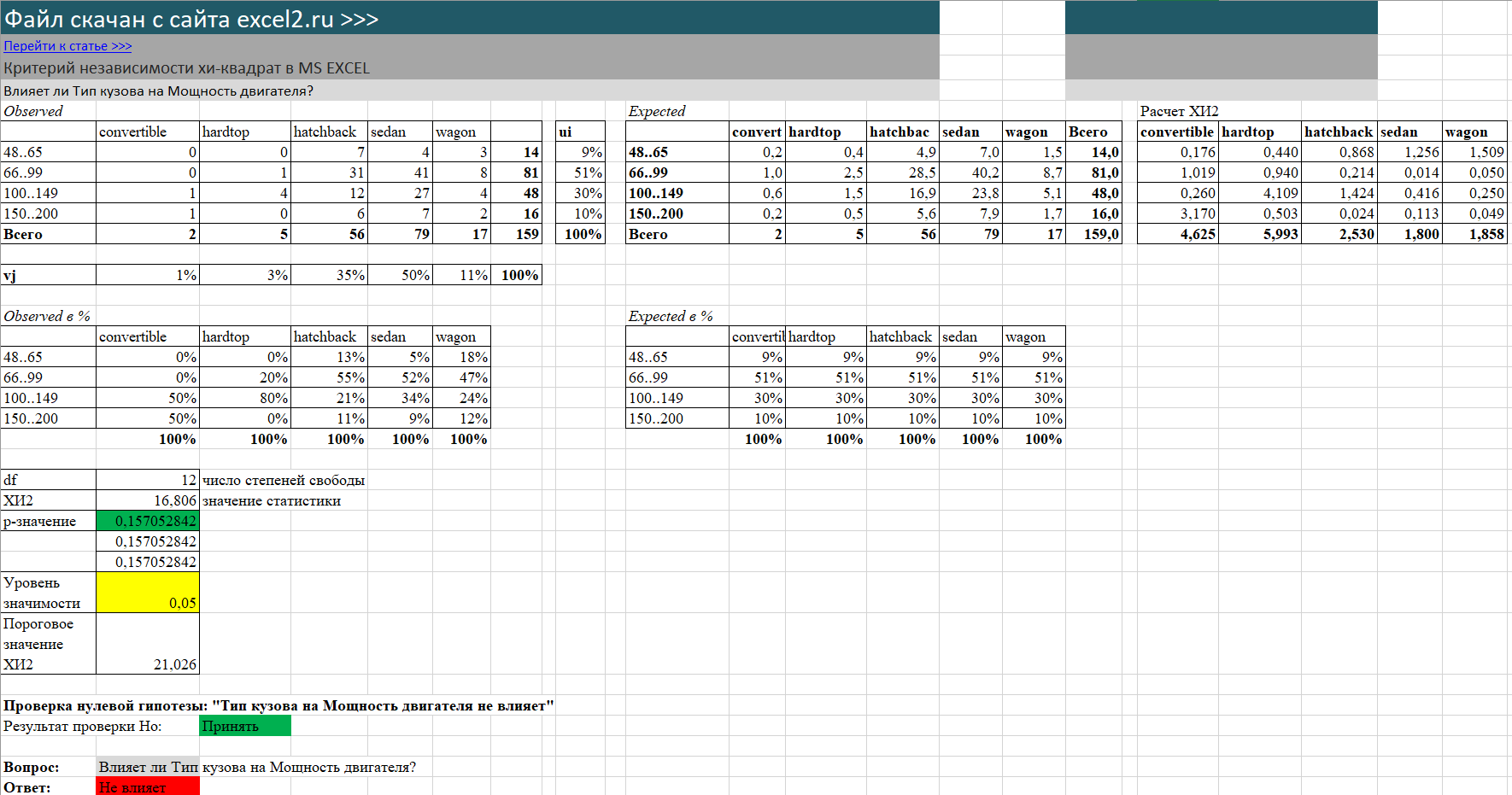
Коэффициент эксцесса – положительный, значит, распределение является островершинным. Положительный знак коэффициента асимметрии говорит о наличии правосторонней асимметрии.

**Анализ результатов таблиц сопряженности**

В качестве категориальных переменных были выбраны:

1) типы кузовов автомобилей (convertible, ardtop, hatchback, sedan, wagon), а лошадиные силы были разделены на 4 категорий (48-65, 66-99, 100-149, 150-200). Была сформулирована нулевая гипотеза H0: "Влияет ли Тип кузова на Мощность двигателя?".

Был проведен анализ с помощью таблиц сопряженности. В результате нулевая гипотеза была принята. Отсюда следует, что тип кузова автомобиля не влияет на мощность двигателя.



**Результаты корреляционного анализа**

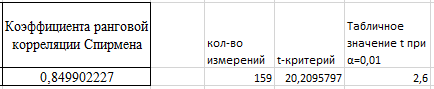
Так как рассматриваемые выборки (Размер двигателя и количество лошадиных сил) были далеки от нормального распределения, то был проведен ранговый корреляционный анализ.

|  |
| --- |
| Коэффициента ранговой корреляции Спирмена |
| 0,849902227 |

Корреляция равна 0,849902227, значит, что присутствует сильная линейная зависимость.

**Значимость**

**Таблица**



Что это значит? Модель не удовлетворительная? Мы ее не принимаем?

Пример и интерпретация ложной корреляции

На сайте http://www.tylervigen.com/spurious-correlations выбраны два динамических ряда, построены графики, проведена оценка ложной корреляций между переменными.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| People who died by falling out of their bed Deaths (US) (CDC) | 450 | 516 | 551 | 594 | 503 | 621 | 626 | 690 | 737 | 780 |
|  |
| Money spent on pets (US) Billions of dollars (Bureau of Economic Analysis) | 39,7 | 41,9 | 44,6 | 46,8 | 49,8 | 53,1 | 56,9 | 61,8 | 65,7 | 67,1 |  |
|  |

|  |
| --- |
| Коэффициент корреляции |
| 0,944768 |

Рассчитаем взаимную корреляцию по остаткам. В данном случае не удалось избавиться от ложной корреляции с помощью тренда.

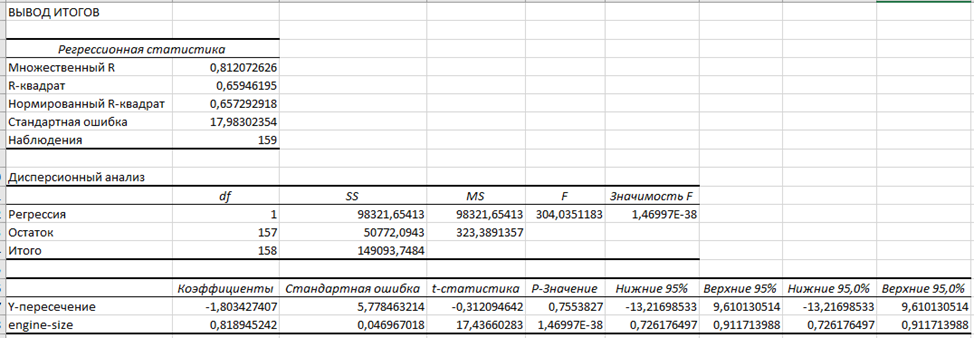
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 |
| Тренд смерти | 459 | 491,885 | 524,77 | 557,655 | 590,54 | 623,425 | 656,31 | 689,195 | 722,08 | 754,965 |
| Остатки смерти | -9 | 24,115 | 26,23 | 36,345 | -87,54 | -2,425 | -30,31 | 0,805 | 14,92 | 25,035 |
| Тренд деньги | 38,2 | 41,4291 | 44,6582 | 47,8873 | 51,1164 | 54,3455 | 57,5746 | 60,8037 | 64,0328 | 67,2619 |
| Остатки деньги | 1,5 | 0,4709 | -0,0582 | -1,0873 | -1,3164 | -1,2455 | -0,6746 | 0,9963 | 1,6672 | -0,1619 |

|  |
| --- |
| Коэффициент корреляции |
| 0,318211 |

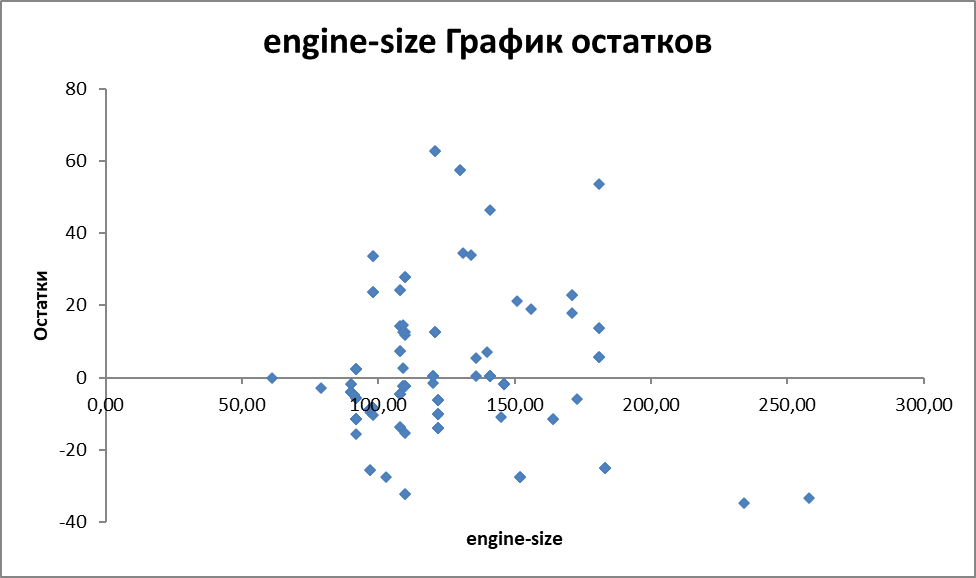
Регрессионная модель. Уравнение регрессии, график остатков,

гистограмма остатков. Оценка качества построенной модели

Построим линейную регрессионную модель (для признаков с низким коэффициентом корреляции).

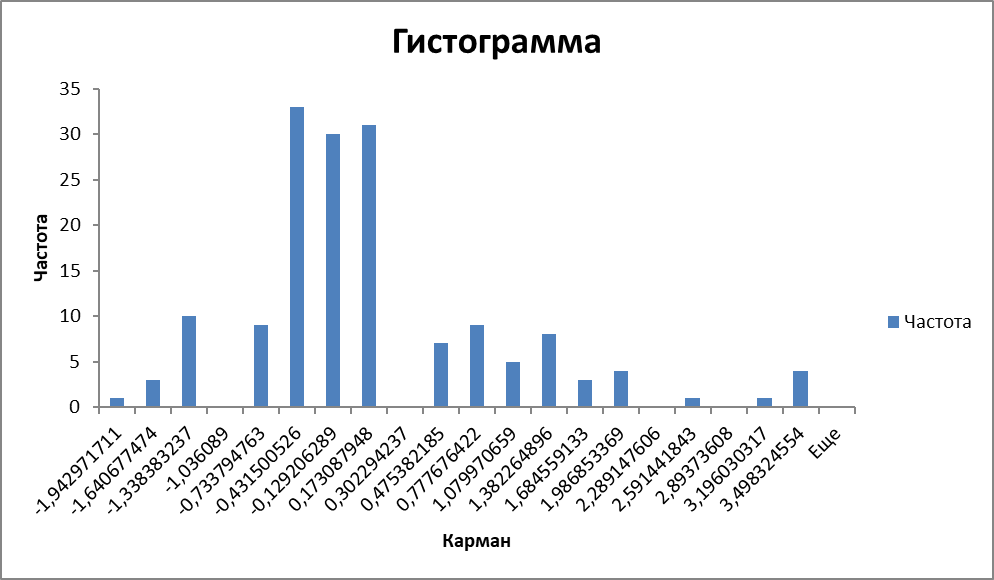
––

Выведем график остатков.



Оценим постоянство среднего и дисперсии. Среднее остатков – постоянное и равно нулю. Дисперсия (разброс) постоянна, среднеквадратическое отклонение равно примерно 18.

Построим гистограмму стандартизированных остатков.



Эта гистограмма похожа на нормальное распределение с правосторонней асимметрией.

Запишем уравнение регрессии. Horsepower и engine-size – это коэффициенты уравнения регрессии. Таким образом, уравнение будет иметь вид: y = 0.8189×x-1,8034.

Коэффициент -1,8034 показывает, каким будет Y, если все переменные в рассматриваемой модели будут равны 0. То есть на значение анализируемого параметра влияют и другие факторы, не описанные в модели.

Коэффициент 0.8189 показывает весомость переменной Х на Y. То есть, как размер двигателя в пределах данной модели влияет на количество лошадей в двигателе с весом 0.8189. Положительный знак указывает на положительное влияние: чем больше размер двигателя, тем больше лошадей в двигателе.

Это даёт нам возможность составить прогноз. Допустим, размер двигателя будет 300 дюймов, и нам нужно примерно представить, сколько будет лошадей в нём. Подставляем в уравнение регрессии данное значение и получаем результат y = 0.8189×x-1,8034. Следовательно, у двигателя размером 300 дюймов, будет примерно 243,867 лошадей.

Оценим качество построенной модели.

**График остатков и гистограмма**

R-квадрат – это коэффициент детерминации. В данном примере R-квадрат = 0,659 (66%), т. е. расчетные параметры модели объясняют вариацию зависимого параметра при изменении незавсимого параметра на 66%. Чем выше значение коэффициента детерминации, тем выбранная модель считается более применимой для конкретной задачи. Считается, что она корректно описывает реальную ситуацию при значении коэффициента детерминации выше 0,8. Следовательно, смысла в построенной модели нет.

Регрессионный анализ

Построим диаграмму рассеяния для тех же переменных, для которых была построена линейная регрессионная модель. Добавим линию тренда на диаграмму рассеяния. Был выбран полиномиальный тренд (степень полинома – 4). Отобразим уравнение и коэффициент детерминации на графике.

Были посчитаны новые значения зависимой переменной по полученному уравнению. Далее посчитаны остатки в новой модели (yисходное- yпредсказанное). Затем построен график остатков.

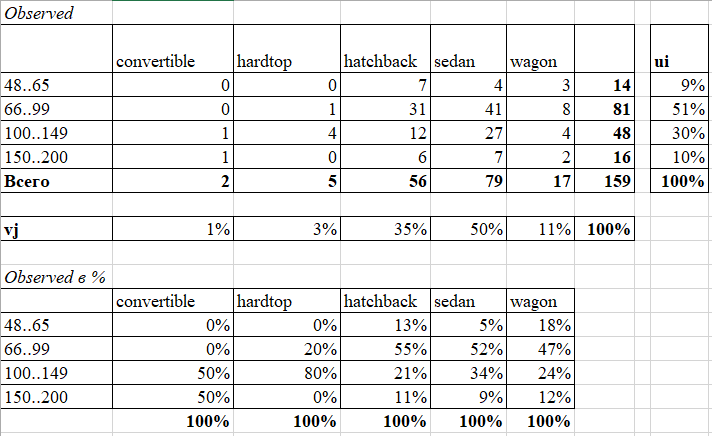
Сравним две регрессионные модели: линейную (из раздела 7) и полиномиальную. Обе модели имеют графики остатков с постоянными средним и дисперсией, однако наблюдаются незначительные выбросы. Коэффициент детерминации во второй модели получился немного лучше, чем в первой модели (0,66 и 0,69 соответственно). Значит, построенные модели получились неудачными.

# Дисперсионный анализ

Для проведения однофакторного дисперсионного анализа необходимо сформулировать начальную гипотезу. Рассмотрим связь между типом кузова автомобиля и мощностью автомобиля. В качестве нулевой гипотезы примем следующее: «Тип кузова не влияет на мощность двигателя».

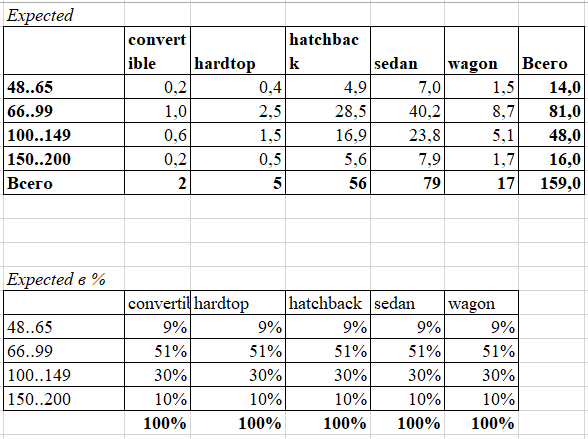
В качестве независимой переменной (фактора) примем тип кузова. Тип кузова – категориальная переменная, так как принимает одно из пяти значений, определяемых словами «convertible», «hardtop», «hatchback», «sedan», «wagon».

Ниже приведена таблица, полученная из статистических данных:

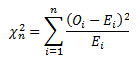


В этой таблице отражено количество автомобилей, кузов которых соответствует одному из пяти типов и обладающих некоторой мощностью двигателя. Данные показаны как количественно, так и в процентах.

Также нами была составлена аналогичная таблица, которая бы отражала ожидаемые величины для каждой ячейки:

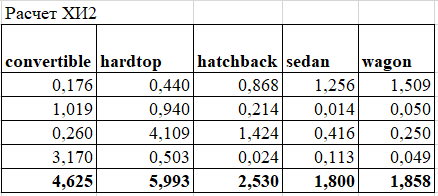


Для проведения дисперсионного анализа необходимо использовать критерий хи-квадрат. Статистика для хи-квадрат представляет собой сумму вида:

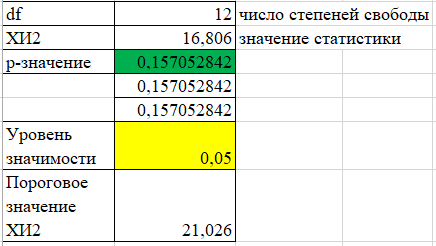


Здесь в качестве параметра Oi примем количество автомобилей в ячейке из таблицы с данными, которые были получены из статистики. Параметр Ei – значение ячейки из таблицы с ожидаемыми значениями.

Сначала мы сгруппировали слагаемые по каждому типу кузовов:



Результаты группирования слагаемых представлены в последней строке таблицы. Затем мы сложили все значения подсумм, определенных для каждого типа кузова и затем, получили статистику для критерия хи-квадрат:



Далее нами было определено значение критерия независимости, по которому мы смогли определить отсутствие связи между кузовом автомобиля и мощностью двигателя. Таким образом нами была принята нулевая гипотеза «Тип кузова не влияет на мощность двигателя».

**Двухфакторный дисперсионный анализ**

Для проведения двухфакторного дисперсионного анализа была сформулирована гипотеза «Тип потребляемого топлива и тип кузова автомобиля влияют на лошадиные силы двигателя». Для независимых переменных были определены. Значения зависимой переменной были разбиты на 4 интервала.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Лошадиные силы | Тип кузова | Тип топлива | |
| gas | diesel |
| 1 | 48..65 | convertible | 0 | 0 |
| 2 | 66..99 |  | 0 | 0 |
| 3 | 100..149 |  | 1 | 0 |
| 4 | 150..200 |  | 1 | 0 |
| 5 | 48..65 | hardtop | 0 | 0 |
| 6 | 66..99 |  | 1 | 0 |
| 7 | 100..149 |  | 3 | 1 |
| 8 | 150..200 |  | 0 | 0 |
| 9 | 48..65 | hatchback | 6 | 1 |
| 10 | 66..99 |  | 32 | 0 |
| 11 | 100..149 |  | 11 | 0 |
| 12 | 150..200 |  | 6 | 0 |
| 13 | 48..65 | sedan | 0 | 4 |
| 14 | 66..99 |  | 36 | 6 |
| 15 | 100..149 |  | 24 | 2 |
| 16 | 150..200 |  | 7 | 0 |
| 17 | 48..65 | wagon | 3 | 0 |
| 18 | 66..99 |  | 8 | 0 |
| 19 | 100..149 |  | 3 | 1 |
| 20 | 150..200 |  | 2 | 0 |

Значение F-критерия фактора А — влияния типа кузова автомобиля на тип топлива автомобиля, Fнабл равно 3,317 а Fкрит лежит в интервале (2,689; +∞). Fнабл не лежит в критической области, следовательно, принимаем, что тип кузова автомобиля влияет на тип топлива автомобиля.

Значение F-критерия фактора B — влияние мощности двигателя на тип топлива автомобиля, Fнабл равно 9,52, а Fкрит лежит в интервале (4,1708; +∞). Fнабл не лежит в критической области, следовательно, принимаем, что мощность двигателя не влияет на тип топлива автомобиля.

