МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

**Национальный исследовательский ядерный университет «МИФИ»**

# 

**Институт интеллектуальных кибернетических систем**

**КАФЕДРА КИБЕРНЕТИКИ**

### БДЗ

**по курсу "Математическая статистика"**

**студента группы Б20-524**

**Гаврилова Максима Сергеевича**

#### Вариант № 3

**Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

##### Подпись: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

2022 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант №\_\_\_\_\_\_\_

ФИО студента \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ группа \_\_\_

Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Результаты статистических тестов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № задания | Проверяемая гипотеза *H*0 | Критерий | Статистическое решение  (α = 0.1) | Вывод |
| 4.1 |  | Хи-квадрат |  |  |
| 4.2 |  | Харке-Бера |  |  |
| 5.1 |  | знаков |  |  |
| 5.2 |  | Хи-квадрат |  |  |

Выводы:

|  |
| --- |
| В результате проведённого в п.4 статистического анализа обнаружено, что |
| В результате проведённого в п.5 статистического анализа обнаружено, что |

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант №\_\_\_\_\_\_\_\_

ФИО студента \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ группа \_\_\_\_\_\_\_\_

Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Результаты статистических тестов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № задания | Проверяемая гипотеза *H*0 | Критерий | Статистическое решение  (α = 0.1) | Вывод |
| 6 |  | Хи-квадрат |  |  |
| 7 |  | ANOVA |  |  |

Выводы:

|  |
| --- |
| В результате проведённого в п.6 статистического анализа обнаружено, что |
| В результате проведённого в п.7 статистического анализа обнаружено, что |

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант №\_\_\_\_\_\_\_\_

ФИО студента \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ группа \_\_\_\_\_\_\_\_

Оценка: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Подпись:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Сводная таблица свойств различных регрессионных моделей:**

|  |  |  |  |
| --- | --- | --- | --- |
| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
| Точность |  |  |  |
| Значимость |  |  |  |
| Адекватность |  |  |  |
| Степень тесноты связи |  |  |  |

Выводы:

|  |
| --- |
| В результате проведённого в п.8 статистического анализа обнаружено, что |
| В результате проведённого в п.9 статистического анализа обнаружено, что |

1. Описательные статистики

*1.1. Выборочные характеристики*

Анализируемый признак 1 – Number of calories consumed per day

Анализируемый признак 2 – Grams of fat consumed per day

Анализируемый признак 3 – Grams of fiber consumed per day

## а) Привести формулы расчёта выборочных характеристик

|  |  |
| --- | --- |
| Выборочная хар-ка | Формула расчета |
| Объём выборки | n (дано) |
| Среднее |  |
| Выборочная дисперсия |  |
| Выборочное среднеквадратическое отклонение | S = |
| Выборочный коэффициент асимметрии |  |
| Выборочный эксцесс |  |

*б) Рассчитать выборочные характеристики*

|  |  |  |  |
| --- | --- | --- | --- |
| Выборочная хар-ка | Признак 1 | Признак 2 | Признак 3 |
| Среднее | 1796.65 | 77.03 | 12.79 |
| Выборочная дисперсия | 462872.63 | 1144.43 | 28.41 |
| Выборочное среднеквадратическое отклонение | 680.34 | 33.83 | 5.33 |
| Выборочный коэффициент асимметрии | 1.74 | 1.1 | 1.15 |
| Выборочный эксцесс | 7.98 | 1.96 | 2.43 |

*1.2. Группировка и гистограммы частот*

Анализируемый признак – Grams of fat consumed per day

Объём выборки – 315

*а) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 8 | Формула Стёрджеса | 27.6875 |

*б) Построить таблицу частот*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Накопл. частота | Относит. накопл. частота |
| 1 | 14.4 | 42.09 | 36 | 0.11 | 36 | 0.11 |
| 2 | 42.09 | 69.78 | 111 | 0.35 | 147 | 0.46 |
| *3* | 69.78 | 97.46 | 95 | 0.3 | 242 | 0.76 |
| 4 | 97.46 | 125.15 | 47 | 0.15 | 289 | 0.91 |
| 5 | 125.15 | 152.84 | 15 | 0.05 | 304 | 0.96 |
| 6 | 152.84 | 180.53 | 8 | 0.03 | 312 | 0.99 |
| 7 | 180.53 | 208.21 | 2 | 0.006 | 314 | 0.996 |
| 8 | 208.21 | 235.9 | 1 | 0.003 | 315 | 1 |
|  |  |  |  |  |  |  |

*в) Построить гистограммы частот и полигоны частот*

|  |  |
| --- | --- |
| Гистограмма и полигон частот | *Гистограмма и полигон относительных частот* |
| Гистограмма и полигон накопленных частот | *Гистограмма и полигон накопленных относительных частот* |

*г) Построить график эмпирической функции распределения*

|  |
| --- |
| Эмпирическая функция распределения |

**2. Интервальные оценки**

*2.1. Доверительные интервалы для мат. ожидания*

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Оцениваемый параметр – мат. ожидание

*а) Привести формулы расчёта доверительных интервалов*

|  |  |
| --- | --- |
| Граница доверительного интервала | Формула расчета |
| Нижняя граница |  |
| Верхняя граница |  |

*б) Рассчитать доверительные интервалы*

|  |  |  |  |
| --- | --- | --- | --- |
| Граница доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| Нижняя граница | 1697 | 1721 | 1733 |
| Верхняя граница | 1896 | 1872 | 1860 |

*2.2. Доверительные интервалы для дисперсии*

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Оцениваемый параметр – дисперсия

*а) Привести формулы расчёта доверительных интервалов*

|  |  |
| --- | --- |
| Граница доверительного интервала | Формула расчета |
| Нижняя граница |  |
| Верхняя граница |  |

*б) Рассчитать доверительные интервалы*

|  |  |  |  |
| --- | --- | --- | --- |
| Граница доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| Нижняя граница | 380177 | 398216 | 407891 |
| Верхняя граница | 574002 | 544751 | 530546 |

*2.3. Доверительные интервалы для разности мат. ожиданий*

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315, 315

Оцениваемый параметр – разность мат. ожиданий

*а) Привести формулы расчёта доверительных интервалов*

|  |  |
| --- | --- |
| Граница доверительного интервала | Формула расчета |
| Нижняя граница |  |
| Верхняя граница |  |

*б) Рассчитать доверительные интервалы*

|  |  |  |  |
| --- | --- | --- | --- |
| Граница доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| Нижняя граница | -453 | -444 | -439 |
| Верхняя граница | -373 | -382 | -387 |

*2.4. Доверительные интервалы для отношения дисперсий*

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315, 315

Оцениваемый параметр – отношение дисперсий

*а) Привести формулы расчёта доверительных интервалов*

|  |  |
| --- | --- |
| Граница доверительного интервала | Формула расчета |
| Нижняя граница |  |
| Верхняя граница |  |

*б) Рассчитать доверительные интервалы*

|  |  |  |  |
| --- | --- | --- | --- |
| Граница доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| Нижняя граница | 0.49 | 0.56 | 0.6 |
| Верхняя граница | 1 | 0.97 | 0.94 |

**3. Проверка статистических гипотез о математических ожиданиях и дисперсиях**

*3.1. Проверка статистических гипотез о математических ожиданиях*

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |
| --- | --- |
|  | Выражение |
| Формула расчета статистики критерия |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы | T(n-1) |
| Формулы расчета критических точек |  |
| Формула расчета *p-value* |  |

*б) Выбрать произвольные значения m0 и проверить статистические гипотезы*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *m*0 | Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 1500 | 0.1 | 7.74 | *0* |  | m |
| 1800 | 0.1 | -0.087 | *0.93* |  | m = 1800 |
| 2000 | 0.1 | -5.3 | *0* |  | m |

*3.2. Проверка статистических гипотез о дисперсиях*

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |
| --- | --- |
|  | Выражение |
| Формула расчета статистики критерия |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |
| Формулы расчета критических точек | *;* |
| Формула расчета *p-value* |  |

*б) Выбрать произвольные значения σ0 и проверить статистические гипотезы*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| σ0 | Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 400000 | 0.1 | 363.36 | *0.057* |  |  |
| 450000 | 0.1 | 322.98 | *0.703* |  |  |
| 500000 | 0.1 | 290.68 | *0.353* |  |  |

*3.3. Проверка статистических гипотез о равенстве математических ожиданий*

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |
| --- | --- |
|  | Выражение |
| Формула расчета статистики критерия |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |
| Формулы расчета критических точек |  |
| Формула расчета *p-value* |  |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | -26.39 | *0* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

*3.4. Проверка статистических гипотез о равенстве дисперсий*

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |
| --- | --- |
|  | Выражение |
| Формула расчета статистики критерия |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |
| Формулы расчета критических точек |  |
| Формула расчета *p-value* |  |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 0.767 | *0.019* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

**4. Критерии согласия**

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

*4.1. Критерий хи-квадрат*

Теоретическое распределение – нормальное.

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | - число элементов выборки, принадлежащих i-той группе  - вероятность того, что значение случайной величины попадёт в i-тую группу |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | k - число групп  r - число неизвестных параметров распределения, оцениваемых по выборке |
| Формула расчета критической точки |  | – уровень значимости |
| Формула расчета *p-value* |  |  |

*б) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 8 | Формула Стёрджеса: k 1.3ln(n) | 777.125 |

*в) Построить таблицу частот*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Вероятность попадания в интервал при условии истинности основной гипотезы |
| 1 | 445.2 | 1222.325 | 56 | 0.1778 | 0.1758 |
| 2 | 1222.325 | 1999.45 | 158 | 0.5016 | 0.4179 |
| *3* | 1999.45 | 2776.575 | 73 | 0.2317 | 0.3079 |
| 4 | 2776.575 | 3553.7 | 25 | 0.0794 | 0.07 |
| 5 | 3553.7 | 4330.825 | 1 | 0.0032 | 0.0048 |
| 6 | 4330.825 | 5107.95 | 1 | 0.0032 | 0.0001 |
| 7 | 5107.95 | 5885.075 | 0 | 0 | 0 |
| 8 | 5885.075 | 6662.2 | 1 | 0.0032 | 0 |

г) Построить гистограмму относительных частот и функцию плотности теоретического распределения на одном графике

|  |
| --- |
|  |

*д) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 12.774 | *0.005* |  | X имеет распределение, отличное от нормального |
| 0.05 | 12.774 | *0.005* |  |
| 0.1 | 12.774 | *0.005* |  |

*4.2. Проверка гипотезы о нормальности на основе коэффициента асимметрии и эксцесса (критерий Харке-Бера)*

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | S – выборочный коэффициент асимметрии  K – выборочный эксцесс |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  | – уровень значимости |
| Формула расчета *p-value* |  |  |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 995.22 | *0* |  | X имеет распределение, отличное от нормального |
| 0.05 |  |
| 0.1 |  |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.4 статистического анализа обнаружено, что количество калорий, в день потребляемых пациентами, не является нормально распределённой величиной. |

**5. Проверка однородности выборок**

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

*5.1 Критерий знаков*

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | число знаков «+» в последовательности знаков разностей соответствующих элементов выборки |
| Закон распределения статистики критерия при условии истинности основной гипотезы | N(0, 1) |  |
| Формула расчета критической точки |  |  |
| Формула расчета *p-value* |  |  |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | -146.5 | *0* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

*5.2. Критерий хи-квадрат*

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | k – число групп  – j-ые элементы выборок соответственно 1-го и 2-го признаков |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  | – уровень значимости |
| Формула расчета *p-value* |  |  |

*б) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 8 | Формула Стёрджеса: k 1.3ln(n) | 215.875 |

*в) Построить таблицу частот*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Номер интервала | Нижняя граница | Верхняя граница | Частота признака 1 | Частота признака 2 | Относит. частота признака 1 | Относит. частота признака 2 |
| 1 | 0 | 215.875 | 229 | 3 | 0.727 | 0.0095 |
| 2 | 215.875 | 431.75 | 69 | 58 | 0.219 | 0.1841 |
| *3* | 431.75 | 647.625 | 8 | 144 | 0.0254 | 0.4571 |
| 4 | 647.625 | 863.5 | 4 | 83 | 0.0127 | 0.2635 |
| 5 | 863.5 | 1079.375 | 2 | 19 | 0.0063 | 0.0603 |
| 6 | 1079.375 | 1295.25 | 1 | 5 | 0.0032 | 0.0159 |
| 7 | 1295.25 | 1511.125 | 2 | 1 | 0.0063 | 0.0032 |
| 8 | 1511.125 | 1727 | 0 | 2 | 0 | 0.0063 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

*г) Построить гистограммы относительных частот на одном графике*

|  |
| --- |
|  |

*д) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 17252.467 | *0* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

Вывод (в терминах предметной области)

|  |
| --- |
| В результате проведённого в п.5 статистического анализа обнаружено, что анализируемые признаки имеют разные функции распределения. |

**6. Таблицы сопряжённости**

Факторный признак *x* – Sex

Результативный признак *y* – Smoking status

Объёмы выборок – 315

Статистическая гипотеза –

*а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез*

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | - частота пары в выборке  - частота пары при условии, что верна |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | - число вариантов признака  - число вариантов признака |
| Формула расчета критической точки |  | – уровень значимости |
| Формула расчета *p-value* |  |  |

*б) Построить эмпирическую таблицу сопряжённости*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x y* | Never | Former | Current Smoker | Σ |
| Male | 13 | 22 | 7 | 42 |
| Female | 144 | 93 | 36 | 273 |
| Σ | 157 | 115 | 43 | 315 |

*в) Построить теоретическую таблицу сопряжённости*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *x y* | Never | Former | Current Smoker | Σ |
| Male | 20.93 | 15.33 | 5.73 | 42 |
| Female | 136.07 | 99.67 | 37.27 | 273 |
| Σ | 157 | 115 | 43 | 315 |

*г) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 7.14 | *0.0282* |  |  |
| 0.05 |  |  |
| 0.1 |  |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.6 статистического анализа обнаружено, что при уровне значимости 0.01 принимается гипотеза об отсутствии статистической зависимости между признаками 1 и 2, однако при уровнях значимости 0.05 и 0.1 эта гипотеза отклоняется. |

**7. Дисперсионный анализ**

Факторный признак *x* – Smoking status

Результативный признак *y* – Quetelet (weight/(height^2))

Число вариантов факторного признака – 3

Объёмы выборок – 315

Статистическая гипотеза –

*а) Рассчитать групповые выборочные характеристики*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Вариант факторного признака | Объём выборки | Групповые средние | Групповые дисперсии |
| 1 | Never | 157 | 26.73 | 47.12 |
| 2 | Former | 115 | 25.93 | 24.92 |
| 3 | Current smoker | 43 | 24.69 | 24.16 |
|  |  |  |  |  |
|  |  |  |  |  |

б) Привести формулы расчёта показателей вариации, используемых в дисперсионном анализе

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак |  | k - 1 |  |
| Остаточные признаки |  | n – k |  |
| Все признаки |  | n - 1 |  |

*в) Рассчитать показатели вариации, используемые в дисперсионном анализе*

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 0.2821 | 2 | 44.43075 |
| Остаточные признаки | 35.8807 | 312 | 36.2257 |
| Все признаки | 36.1628 | 314 | 36.278 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dмежгр* | *Dвнутригр* | *Dобщ* | *Dмежгр* + *Dвнутригр* |
| Значение | 0.2821 | 35.8807 | 36.1628 | 36.1628 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Эмпирический коэффициент детерминации |  | 0.0078 |
| Эмпирическое корреляционное отношение |  | 0.0883 |

*е) Охарактеризовать тип связи между факторным и результативным признаками*

|  |
| --- |
|  |

ж) Указать формулы расчёта показателей, используемых при проверке статистической гипотезы дисперсионного анализа

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  |  |
| Формула расчета *p-value* |  |  |

*з) Проверить статистическую гипотезу дисперсионного анализа*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 2.07 | *0.13* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.7 статистического анализа обнаружено, что отсутствует статистическая зависимость между факторным признаком Smoking Status и результативным признаком Quetelet. |

**8. Корреляционный анализ**

*8.1. Расчёт парных коэффициентов корреляции*

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

*а) Рассчитать точечные оценки коэффициентов корреляции*

|  |  |  |
| --- | --- | --- |
|  | Формула расчета | Значение |
| Линейный коэффициент корреляции |  | 0.07 |
| Ранговый коэффициент корреляции по Спирмену |  | 0.13 |
| Ранговый коэффициент корреляции по Кендаллу |  | 0.09 |

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

|  |  |
| --- | --- |
| Граница доверительного интервала | Формула расчета |
| Нижняя граница |  |
| Верхняя граница |  |

*в) Рассчитать доверительные интервалы для линейного коэффициента корреляции*

|  |  |  |  |
| --- | --- | --- | --- |
| Граница доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| Нижняя граница | -0.07 | -0.04 | -0.02 |
| Верхняя граница | 0.21 | 0.18 | 0.16 |

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициентов корреляции

|  |  |  |
| --- | --- | --- |
| Статистическая гипотеза | Формула расчета статистики критерия | Закон распределения статистики критерия при условии истинности основной гипотезы |
|  |  |  |
|  |  |  |
|  |  |  |

*д) Проверить значимость коэффициентов корреляции*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Статистическая гипотеза | Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
|  | 0.1 | 1.24 | *0.2* |  |  |
|  | 0.1 | 2.32 | *0.02* |  |  |
|  | 0.1 | 2.39 | *0.02* |  |  |

*8.2. Расчёт множественных коэффициентов корреляции*

Анализируемый признак 1 – Number of calories consumed per day

Анализируемый признак 2 – Grams of fat consumed per day

Анализируемый признак 3 – Grams of fiber consumed per day

Объёмы выборок – 315

*а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу*

|  |  |  |  |
| --- | --- | --- | --- |
| Признак  Признак | 1 | 2 | 3 |
| 1 | 1 | 0.72 | 0.39 |
| 2 | 0.72 | 1 | 0.22 |
| 3 | 0.39 | 0.22 | 1 |

*б) Рассчитать матрицу значений p-value для ранговых коэффициентов корреляции по Кендаллу (статистическая гипотеза* *)*

|  |  |  |  |
| --- | --- | --- | --- |
| Признак  Признак | 1 | 2 | 3 |
| 1 | – | 0 | 0 |
| 2 | 0 | – | 0 |
| 3 | 0 | 0 | – |

*в) Рассчитать точечную оценку коэффициента конкордации*

|  |  |  |
| --- | --- | --- |
|  | Формула расчета | Значение |
| Коэффициент конкордации |  | 0.726 |

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициента конкордации

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  |  |
| Формула расчета *p-value* |  |  |

*д) Проверить значимость коэффициента конкордации*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 684 | *0* |  | между признаками 1, 2 и 3 существует ранговая корреляция |
| 0.05 |  |
| 0.1 |  |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.8 статистического анализа обнаружено, что между признаками Plasma beta-carotene и Plasma Retinol отсутствует линейная связь, а между признаками Number of calories consumed per day, Grams of fat consumed per day и Grams of fiber consumed per day присутствует ранговая корреляция |

**9. Регрессионный анализ**

*9.1 Простейшая линейная регрессионная модель*

Факторный признак *x* – Number of calories consumed per day

Результативный признак *y* – Plasma beta-carotene (ng/ml)

Уравнение регрессии – 

*9.1.1. Точечные оценки линейной регрессионной модели*

*а) Рассчитать точечные оценки параметров линейной регрессионной модели*

|  |  |  |
| --- | --- | --- |
| Параметр | Формула расчета | Значение |
| β0 |  | 201 |
| β1 |  | -0.006 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = 201 – 0.006x |

в) Привести формулы расчёта показателей вариации, используемых в регрессионном анализе

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак |  | k - 1 |  |
| Остаточные признаки |  | n - k |  |
| Все признаки |  | n - 1 |  |

*г) Рассчитать показатели вариации, используемые в регрессионном анализе*

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 16 | 1 | 5040 |
| Остаточные признаки | 33367 | 313 | 33580 |
| Все признаки | 33383 | 314 | 33489 |

д) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 16 | 33367 | 33383 | 33383 |

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Коэффициент детерминации |  | 0.00048 |
| Корреляционное отношение |  | 0.022 |

ж) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

|  |
| --- |
|  |

*9.1.2. Интервальные оценки линейной регрессионной модели*

а) Привести формулы расчёта доверительных интервалов для параметров линейной регрессионной модели

|  |  |  |
| --- | --- | --- |
| Параметр | Границы доверительного интервала | Формула расчета |
| β0 | Нижняя граница |  |
| Верхняя граница |  |
| β1 | Нижняя граница |  |
| Верхняя граница |  |

б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Параметр | Границы доверительного интервала | α = 0.01 | α = 0.05 | α = 0.1 |
| β0 | Нижняя граница | 125 | 143 | 153 |
| Верхняя граница | 276 | 258 | 249 |
| β1 | Нижняя граница | -0.045 | -0.036 | -0.031 |
| Верхняя граница | 0.033 | 0.024 | 0.019 |

*в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)*

|  |  |
| --- | --- |
| Границы доверительного интервала | Формула расчета |
| Нижняя граница *flow*(*x*) |  |
| Верхняя граница *fhigh*(*x*) |  |

*г) Построить диаграмму рассеяния признаков x и y*. *Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии flow(x) и fhigh(x) на уровне значимости α = 0.1*

|  |
| --- |
|  |

*д) Построить график остатков *

|  |
| --- |
|  |

*9.1.3. Проверка значимости линейной регрессионной модели*

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  |  |
| Формула расчета *p-value* |  |  |

## б) Проверить значимость линейной регрессионной модели

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 0.15 | *0.69* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

*9.2 Линейная регрессионная модель общего вида*

Факторный признак *x* – Number of calories consumed per day

Результативный признак *y* – Plasma beta-carotene (ng/ml)

Уравнение регрессии – квадратичное по *x*: 

*9.2.1. Точечные оценки линейной регрессионной модели*

*а) Рассчитать точечные оценки параметров линейной регрессионной модели*

|  |  |  |
| --- | --- | --- |
| Параметр | Формула расчета | Значение |
| β0 | y =  F = | 156 |
| β1 | 0.037 |
| β2 | -0.000009 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = 156 + 0.037x – 0.000009 |

*в) Рассчитать показатели вариации, используемые в регрессионном анализе*

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 136.46 | 2 | 21492.45 |
| Остаточные признаки | 33246.52 | 312 | 33566.2 |
| Все признаки | 33382.98 | 314 | 33489.3 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 136.46 | 33246.52 | 33382.98 | 33382.98 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Коэффициент детерминации |  | 0.004 |
| Корреляционное отношение |  | 0.063 |

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

|  |
| --- |
|  |

*9.2.2. Интервальные оценки линейной регрессионной модели*

*а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)*

|  |  |
| --- | --- |
| Границы доверительного интервала | Формула расчета |
| Нижняя граница *flow*(*x*) |  |
| Верхняя граница *fhigh*(*x*) |  |

*б) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии flow(x) и fhigh(x) на уровне значимости α = 0.1*

|  |
| --- |
|  |

*в) Построить график остатков *

|  |
| --- |
|  |

*9.2.3. Проверка значимости линейной регрессионной модели*

Статистическая гипотеза – 

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  |  |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  |  |
| Формула расчета *p-value* |  |  |

*б)* *Проверить значимость линейной регрессионной модели*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 0.63 | *0.53* |  |  |
| 0.05 |  |  |
| 0.1 |  |  |

*9.3 Множественная линейная регрессионная модель*

Факторный признак 1 *x*1 – Number of calories consumed per day

Факторный признак 2 *x*2 – Quetelet (weight/(height^2))

Результативный признак *y* – Plasma beta-carotene (ng/ml)

Уравнение регрессии – 

*а) Рассчитать точечные оценки параметров линейной регрессионной модели*

|  |  |  |
| --- | --- | --- |
| Параметр | Формула расчета | Значение |
| β0 | y =  F = | 383 |
| β1 | -0.0058 |
| β2 | -6.98 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = 383 – 0.0058 |

*в) Рассчитать показатели вариации, используемые в регрессионном анализе*

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 1771.85 | 2 | 279066 |
| Остаточные признаки | 31611.13 | 312 | 31915 |
| Все признаки | 33382.98 | 314 | 33489 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 1771.85 | 31611.13 | 33382.98 | 33382.98 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Множественный коэффициент детерминации |  | 0.056 |
| Множественное корреляционное отношение |  | 0.237 |

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

|  |
| --- |
|  |

*9.4. Выводы*

*а) Сводная таблица показателей вариации для различных регрессионных моделей*

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
| Факторный признак | 16 | 136 | 1772 |
| Остаточные признаки | 33367 | 33247 | 31611 |
| Все признаки | 33383 | 33383 | 33383 |

*б) Сводная таблица свойств различных регрессионных моделей*

|  |  |  |  |
| --- | --- | --- | --- |
| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
| Точность | 0.048% | 0.4% | 5.6% |
| Значимость | Нет | Нет | Нет |
| Адекватность | Нет | Нет | Нет |
| Степень тесноты связи | слабая | слабая | слабая |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.9 статистического анализа обнаружено, что ни одна из предложенный регрессионных моделей не отражает реальную зависимость признака Plasma beta-carotene от признаков Number of calories consumed per day и Quetelet |