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**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

#### Вариант № 13

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

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

2022 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант № 13

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 не является нормально распределенной величиной. |
| В результате проведённого в п.5 статистического анализа обнаружено, что выборки *A5* и *A8* неоднородны, т.еимеют разные распределения. |

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант № 13

ФИО студента: группа:

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

Результаты статистических тестов:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № задания | Проверяемая гипотеза *H*0 | Критерий | Статистическое решение  (α = 0.1) | Вывод |
| 6 | y – A14  x – A9 | Хи-квадрат | *Отвергаем* | A14 зависит от A9 |
| 7 | y – A8  x – A4 | ANOVA | *Отвергаем* | A4 зависит от A8 |

Выводы:

|  |
| --- |
| В результате проведённого в п.6 статистического анализа обнаружено, что средняя зарплата Average salary (A9) оказывает влияние на Average compensation (A14), т.е между этими признаками существует статистическая связь. |
| В результате проведённого в п.7 статистического анализа обнаружено, что факторный признак x – A4 Type (I, IIA, or IIB) заметно влияет на Результативный признак y - A8 (Average salary - all ranks) |

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант № 13

ФИО студента: группа:

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

**Сводная таблица свойств различных регрессионных моделей:**

|  |  |  |  |
| --- | --- | --- | --- |
| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
| Точность | 32.7% | 41.7% | 91.5% |
| Значимость | Да | Да | Да |
| Адекватность | Нет | Нет | Да |
| Степень тесноты связи | Заметная | Заметная | Сильная |

Выводы:

|  |
| --- |
| В результате проведённого в п.8 статистического анализа обнаружено, что выборки A5 (Average salary - full professors), A6 (Average salary - associate professors) и A8 (Average salary - all ranks) зависимы друг от друга как попарно, так и все вместе. |
| В результате проведённого в п.9 статистического анализа обнаружено, что точность множественной линейной модели гораздо больше точности простейшей л.м и л.м с квадратичным членом. Все регрессионные модели значимы и адекватно (в случае множественной линейной регрессии) отражают зависимость A13 от A5 и A15, однако степень тесноты связи в множественной модели по шкале Чеддока является сильной, когда как в остальных моделях является лишь заметной. |

1. Описательные статистики

*1.1. Выборочные характеристики*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A6 Average salary - associate professors

Анализируемый признак 3 – A8 Average salary - all ranks

## а) Привести формулы расчёта выборочных характеристик

|  |  |
| --- | --- |
| Выборочная хар-ка | Формула расчета |
| Объём выборки | n |
| Среднее |  |
| Выборочная дисперсия |  |
| Выборочное среднеквадратическое отклонение |  |
| Выборочный коэффициент асимметрии |  |
| Выборочный эксцесс |  |

*б) Рассчитать выборочные характеристики*

|  |  |  |  |
| --- | --- | --- | --- |
| Выборочная хар-ка | Признак 1 | Признак 2 | Признак 3 |
| Среднее | 526.48 | 420.04 | 428.03 |
| Выборочная дисперсия | 13868.86 | 4957.76 | 8217.62 |
| Выборочное среднеквадратическое отклонение | 117.77 | 70.41 | 90.65 |
| Выборочный коэффициент асимметрии | 0.68 | 0.35 | 0.82 |
| Выборочный эксцесс | 0.53 | 0.18 | 0.98 |

*1.2. Группировка и гистограммы частот*

Анализируемый признак – A8 Average salary - all ranks

Объём выборки – 1073

*а) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 11 | Формула Стерджесса:  k = 1 + | (max-min)/k =57.63 |

*б) Построить таблицу частот*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Накопл. частота | Относит. накопл. частота |
| 1 | 232.0 | 289.63 | 30 | 0.028 | 30 | 0.028 |
| 2 | 289.63 | 347.27 | 172 | 0.16 | 202 | 0.188 |
| 3 | 347.27 | 404.91 | 290 | 0.27 | 492 | 0.458 |
| 4 | 404.91 | 462.55 | 255 | 0.238 | 747 | 0.696 |
| 5 | 462.55 | 520.18 | 145 | 0.135 | 892 | 0.831 |
| 6 | 520.18 | 577.82 | 118 | 0.11 | 1010 | 0.941 |
| 7 | 577.82 | 635.45 | 40 | 0.037 | 1050 | 0.978 |
| 8 | 635.45 | 693.09 | 13 | 0.012 | 1063 | 0.99 |
| 9 | 693.09 | 750.73 | 6 | 0.006 | 1069 | 0.996 |
| 10 | 750.73 | 808.36 | 2 | 0.002 | 1071 | 0.998 |
| 11 | 808.36 | 866.0 | 2 | 0.002 | 1073 | 1.0 |

*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 – A8 Average salary - all ranks

Объём выборки – 1073

Оцениваемый параметр – *m*

*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 | 420.89 | 422.6 | 423.48 |
| Верхняя граница | 435.17 | 433.46 | 432.59 |

*2.2. Доверительные интервалы для дисперсии*

Анализируемый признак – A8 Average salary - all ranks

Объём выборки – 1073

Оцениваемый параметр –

*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 | 7371.65 | 7564.01 | 7665.04 |
| Верхняя граница | 9210.03 | 8960.22 | 8835.87 |

*2.3. Доверительные интервалы для разности мат. ожиданий*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

Оцениваемый параметр –

*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 | 86.75 | 89.55 | 90.99 |
| Верхняя граница | 110.15 | 107.35 | 105.92 |

*2.4. Доверительные интервалы для отношения дисперсий*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

Оцениваемый параметр –

*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 | 1.44 | 1.5 | 1.53 |
| Верхняя граница | 1.98 | 1.9 | 1.87 |

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

*3.1. Проверка статистических гипотез о математических ожиданиях*

Анализируемый признак – A8 Average salary - all ranks

Объём выборки – 1073

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 m0 и проверить статистические гипотезы*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *m*0 | Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 420 | 0.1 | 2.9 | *0.004* | *Отвергаем* |  |
| 428 | 0.1 | 0.012 | *0.99* | *Принимаем* |  |
| 430 | 0.1 | -0.71 | *0.48* | *Принимаем* |  |

*3.2. Проверка статистических гипотез о дисперсиях*

Анализируемый признак – A8 Average salary - all ranks

Объём выборки – 1073

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 |
| 90.5 | 0.1 | 1075.58 | *0.93* | *Принимаем* | 90.5 |
| 95 | 0.1 | 976.1 | *0.03* | *Отвергаем* |  |
| 87 | 0.1 | 1163.86 | *0.05* | *Отвергаем* |  |

*3.3. Проверка статистических гипотез о равенстве математических ожиданий*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

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 | 21.7 |  | *Отвергаем* |  |
| 0.05 | *Отвергаем* |  |
| 0.1 | *Отвергаем* |  |

*3.4. Проверка статистических гипотез о равенстве дисперсий*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

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 | 1.69 |  | *Отвергаем* |  |
| 0.05 | *Отвергаем* |  |
| 0.1 | *Отвергаем* |  |

**4. Критерии согласия**

Анализируемый признак – A8 Average salary - all ranks

Объём выборки – 1073

*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 интервалу  - вероятность попадания в каждый интервал |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | k – количество разбиений на интервалы  r - число неизвестных параметров распределения |
| Формула расчета критической точки |  | Правосторонняя критическая область.  – уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  – закон распределения хи квадрат |

*б) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 11 | Формула Стерджесса:  k = 1 + | (max-min)/k =57.63 |

*в) Построить таблицу частот*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Номер интервала | Нижняя граница | Верхняя граница | Частота | Относит. частота | Вероятность попадания в интервал при условии истинности основной гипотезы |
| 1 | 232 | 289.63 | 30 | 0.028 | 0.048 |
| 2 | 289.63 | 347.27 | 172 | 0.16 | 0.123 |
| 3 | 347.27 | 404.91 | 290 | 0.27 | 0.212 |
| 4 | 404.91 | 462.55 | 255 | 0.238 | 0.249 |
| 5 | 462.55 | 520.18 | 145 | 0.135 | 0.197 |
| 6 | 520.18 | 577.82 | 118 | 0.11 | 0.105 |
| 7 | 577.82 | 635.45 | 40 | 0.037 | 0.038 |
| 8 | 635.45 | 693.09 | 13 | 0.012 | 0.009 |
| 9 | 693.09 | 750.73 | 6 | 0.006 | 0.002 |
| 10 | 750.73 | 808.36 | 2 | 0.002 | 0.0002 |
| 11 | 808.36 | 866 | 2 | 0.002 | 1.29 \* |

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 | 373.83 | *0* | *Отвергаем* |  |
| 0.05 | 373.83 | *0* | *Отвергаем* |  |
| 0.1 | 373.83 | *0* | *Отвергаем* |  |

*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 |  | *– коэффициент асимметрии*  *– коэффициент*  *эксцесса*  *– третий центральный момент*  *– четвертый центральный момент*  *n -число наблюдений* |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  |  |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  закон распределения хи квадрат |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 162.56 | *0* | *Отвергаем* |  |
| 0.05 | *Отвергаем* |  |
| 0.1 | *Отвергаем* |  |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.4 статистического анализа обнаружено, что выборка A8 не является нормально распределенной |

**5. Проверка однородности выборок**

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

*5.1 Критерий знаков*

Статистическая гипотеза –

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

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | n – объем выборки  K – число знаков «+» в последовательности знаков разностей x1 – y1,…,xn – yn  H - частота «успеха» |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | N – нормальное распределение |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  закон нормального распределения |

*б) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 32.76 |  | *Отвергаем* |  |
| 0.05 | *Отвергаем* |  |
| 0.1 | *Отвергаем* |  |

*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 |  | – объемы выборок  *– число элементов выборки на вероятность попадания в интервал* |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | k – число интервалов  – распределение хи квадрат |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  – закон распределения хи квадрат |

*б) Выбрать число групп*

|  |  |  |
| --- | --- | --- |
| Число групп | Обоснование выбора числа групп | Ширина интервалов |
| 11 | Формула Стерджесса:  k = 1 + | 70.63 |

*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 | 232 | 302,64 | 6 | 53 | 0,005592 | 0,049394 |
| 2 | 302,64 | 373,27 | 78 | 275 | 0,072693 | 0,256291 |
| *3* | 373,27 | 443,91 | 186 | 345 | 0,173346 | 0,321528 |
| 4 | 443,91 | 514,55 | 288 | 205 | 0,268406 | 0,191053 |
| 5 | 514,55 | 585,18 | 207 | 138 | 0,192917 | 0,128611 |
| 6 | 585,18 | 655,82 | 143 | 41 | 0,133271 | 0,038211 |
| 7 | 655,82 | 726,45 | 102 | 9 | 0,095061 | 0,008388 |
| 8 | 726,45 | 797,09 | 41 | 5 | 0,038211 | 0,00466 |
| 9 | 797,09 | 867,73 | 11 | 2 | 0,010252 | 0,001864 |
| 10 | 867,73 | 938,36 | 7 | 0 | 0,006524 | 0 |
| 11 | 938,36 | 1009 | 4 | 0 | 0,003728 | 0 |

*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 | 153.69 | *0* | *Отвергаем* |  |
| 0.05 | *Отвергаем* |  |
| 0.1 | *Отвергаем* |  |

Вывод (в терминах предметной области)

|  |
| --- |
| В результате проведённого в п.5 статистического анализа обнаружено, что выборки A5 и A8 имеют разные распределения, так как гипотеза была отвергнута критерием знаков и критерием хи квадрат. |

**6. Таблицы сопряжённости**

Факторный признак *x* – A9 Average salary - all ranks > all-college average?

Результативный признак *y* – A14 Average compensation - all ranks > all-college average?

Объёмы выборок – 1073

Статистическая гипотеза –

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

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | - частота варианта в выборке  - теоретическая частота |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | - число вариантов случайной величины  - число вариантов случайной величины  – распределение хи квадрат |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  - закон распределения хи квадрат |

*б) Построить эмпирическую таблицу сопряжённости*

|  |  |  |  |
| --- | --- | --- | --- |
| *x y* | N | Y | Σ |
| N | 570 | 31 | 601 |
| Y | 22 | 450 | 472 |
| Σ | 592 | 481 | 1073 |

*в) Построить теоретическую таблицу сопряжённости*

|  |  |  |  |
| --- | --- | --- | --- |
| *x y* | N | Y | Σ |
| N | 331.59 | 269.41 | 601 |
| Y | 260.41 | 211.59 | 472 |
| Σ | 592 | 481 | 1073 |

*г) Проверить статистические гипотезы*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 865.68 |  | *Отвергаем* | A14 зависит от A9 |
| 0.05 | *Отвергаем* | A14 зависит от A9 |
| 0.1 | *Отвергаем* | A14 зависит от A9 |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.6 статистического анализа обнаружено, что средняя зарплата Average salary (A9) оказывает влияние на Average compensation (A14), т.е. между этими признаками существует статистическая связь. |

**7. Дисперсионный анализ**

Факторный признак *x* – A4 Type (I, IIA, or IIB)

Результативный признак *y* – A8 Average salary - all ranks

Число вариантов факторного признака – 3

Объёмы выборок – 1073

Статистическая гипотеза –

*а) Рассчитать групповые выборочные характеристики*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| № п/п | Вариант факторного признака | Объём выборки | Групповые средние | Групповые дисперсии |
| 1 | I | 180 | 533.67 | 7652.25 |
| 2 | IIA | 359 | 440.82 | 4684.1 |
| 3 | IIB | 534 | 383.83 | 4975.63 |

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

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 2898.99 | 2 | 1555309.57 |
| Остаточные признаки | 5310.97 | 1070 | 5325.86 |
| Все признаки | 8209.96 | 1072 | 8217.62 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dмежгр* | *Dвнутригр* | *Dобщ* | *Dмежгр* + *Dвнутригр* |
| Значение | 2898.993 | 5310.970 | 8209.963 | 8209.963 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Эмпирический коэффициент детерминации |  | 0.353 |
| Эмпирическое корреляционное отношение |  | 0.594 |

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

|  |
| --- |
| По данным из шкалы Чеддака тип связи между A4 и A8 – **заметная**. |

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

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | n – объем выборки  кол-во групп (факторных признаков)  – межгрупповая дисперсия  – общая дисперсия |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | F – распределение Фишера |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  – закон распределения фишера |

*з) Проверить статистическую гипотезу дисперсионного анализа*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 292.03 |  | *Отвергаем* | A4 зависит от A8 |
| 0.05 | *Отвергаем* | A4 зависит от A8 |
| 0.1 | *Отвергаем* | A4 зависит от A8 |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.7 статистического анализа обнаружено, что факторный признак *x* – A4 Type (I, IIA, or IIB) заметно влияет на Результативный признак *y -* A8 (Average salary - all ranks) |

**8. Корреляционный анализ**

*8.1. Расчёт парных коэффициентов корреляции*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A8 Average salary - all ranks

Объёмы выборок – 1073

*а) Рассчитать точечные оценки коэффициентов корреляции*

|  |  |  |
| --- | --- | --- |
|  | Формула расчета | Значение |
| Линейный коэффициент корреляции |  | 0.967 |
| Ранговый коэффициент корреляции по Спирмену |  | 0.964 |
| Ранговый коэффициент корреляции по Кендаллу | - число пар наблюдений , для которых выполнено условие  - число пар наблюдений , для которых выполнено условие | 0.84 |

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 | 0.963 | 0.964 | 0.965 |
| Верхняя граница | 0.973 | 0.972 | 0.971 |

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 | 125.52 | 0 | *Отвергаем* |  |
|  | 0.1 | 119.03 | 0 | *Отвергаем* |  |
|  | 0.1 | 41.27 | 0 | *Отвергаем* |  |

*8.2. Расчёт множественных коэффициентов корреляции*

Анализируемый признак 1 – A5 Average salary - full professors

Анализируемый признак 2 – A6 Average salary - associate professors

Анализируемый признак 3 – A8 Average salary - all ranks

Объёмы выборок – 1073

*а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу*

|  |  |  |  |
| --- | --- | --- | --- |
| Признак  Признак | A5 | A6 | A8 |
| A5 | 1 | 0.82 | 0.84 |
| A6 | 0.82 | 1 | 0.81 |
| A8 | 0.84 | 0.81 | 1 |

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

|  |  |  |  |
| --- | --- | --- | --- |
| Признак  Признак | A5 | A6 | A8 |
| A5 | – | 0 | 0 |
| A6 | 0 | – | 0 |
| A8 | 0 | 0 | – |

*в) Рассчитать точечную оценку коэффициента конкордации*

|  |  |  |
| --- | --- | --- |
|  | Формула расчета | Значение |
| Коэффициент конкордации |  | 0.971 |

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

|  |  |  |
| --- | --- | --- |
|  | Выражение | Пояснение использованных обозначений |
| Формула расчета статистики критерия |  | n - объем выборки  k - число выборок  W - коэффициент конкордации |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | – распределение хи квадрат |
| Формула расчета критической точки |  | α- уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  - закон распределения хи квадрат |

*д) Проверить значимость коэффициента конкордации*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 3121.94 | *0* | *Отвергаем* | A5, A6, A8 - зависимы |
| 0.05 | *Отвергаем* | A5, A6, A8 - зависимы |
| 0.1 | *Отвергаем* | A5, A6, A8 - зависимы |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.8 статистического анализа обнаружено, что выборки A5, A6 и A8 зависимы друг от друга как попарно, так и все вместе. |

**9. Регрессионный анализ**

*9.1 Простейшая линейная регрессионная модель*

Факторный признак *x* – A15 Number of full professors

Результативный признак *y* – A13 Average compensation - all ranks

Уравнение регрессии – 

*9.1.1. Точечные оценки линейной регрессионной модели*

*а) Рассчитать точечные оценки параметров линейной регрессионной модели*

|  |  |  |
| --- | --- | --- |
| Параметр | Формула расчета | Значение |
| β0 |  | 488.99 |
| β1 |  | 0.47 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = |

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

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

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

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 4589.95 | 1 | 4925023.17 |
| Остаточные признаки | 9440.54 | 1071 | 9458.164 |
| Все признаки | 14030.49 | 1072 | 14043.58 |

д) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 9440.54 | 4589.95 | 14030.49 | 14030.49 |

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Коэффициент детерминации |  | 0.327 |
| Корреляционное отношение |  | 0.572 |

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

|  |
| --- |
| Тип связи между факторным и результативным признаками, определяемой рассчитанной простейшей линейной регрессией - **заметная**. |

*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 | 479.629 | 481.873 | 483.019 |
| Верхняя граница | 498.357 | 496.114 | 494.967 |
| β1 | Нижняя граница | 0.412 | 0.425 | 0.431 |
| Верхняя граница | 0.517 | 0.505 | 0.498 |

*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 |  | n – объем выборки  - выборочная дисперсия, обусловленная регрессией Y на X  – остаточная дисперсия |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | – распределение Фишера |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | – статистика критерия  - функция распределения хи квадрат |

## б) Проверить значимость линейной регрессионной модели

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 520.72 |  | *Отвергаем* | Регрессионная модель значима |
| 0.05 | *Отвергаем* | Регрессионная модель значима |
| 0.1 | *Отвергаем* | Регрессионная модель значима |

*9.2 Линейная регрессионная модель общего вида*

Факторный признак *x* – A15 Number of full professors

Результативный признак *y* – A13 Average compensation - all ranks

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 | ↓ ↓  ↓ | 460.11 |
| β1 | ↓ | 1.025 |
| β2 |  | -0.001 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = |

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

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 5844.45 | 2 | 3135548.6 |
| Остаточные признаки | 8186.04 | 1070 | 8193.68 |
| Все признаки | 14030.49 | 1072 | 14069.83 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 5844.45 | 8186.04 | 14030.49 | 14030.49 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Коэффициент детерминации |  | 0.417 |
| Корреляционное отношение |  | 0.645 |

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

|  |
| --- |
| Тип связи между факторным и результативными признаками, определяемой рассчитанной линейной регрессией - **заметная.** |

*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 |  | n – объем выборки  - число неизвестных параметров функции регрессии  - выборочная дисперсия, обусловленная регрессией Y на X  – остаточная дисперсия |
| Закон распределения статистики критерия при условии истинности основной гипотезы |  | – распределение Фишера |
| Формула расчета критической точки |  | - уровень значимости |
| Формула расчета *p-value* |  | Z – статистика критерия  - функция распределения хи квадрат |

*б)* *Проверить значимость линейной регрессионной модели*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Уровень значимости | Выборочное значение статистики критерия | *p-value* | Статистическое решение | Вывод |
| 0.01 | 381.965 |  | *Отвергаем* | Регрессионная модель значима |
| 0.05 | *Отвергаем* | Регрессионная модель значима |
| 0.1 | *Отвергаем* | Регрессионная модель значима |

*9.3 Множественная линейная регрессионная модель*

Факторный признак 1 *x*1 – A15 Number of full professors

Факторный признак 2 *x*2 – A5 Average salary - full professors

Результативный признак *y* – A13 Average compensation - all ranks

Уравнение регрессии – 

*а) Рассчитать точечные оценки параметров линейной регрессионной модели*

|  |  |  |
| --- | --- | --- |
| Параметр | Формула расчета | Значение |
| β0 | ↓ ↓  ↓ | 40.18 |
| β1 | ↓ | 0.037 |
| β2 |  | 0.94 |

*б) Записать точечную оценку уравнения регрессии*

|  |
| --- |
| *f*(*x*) = |

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

|  |  |  |  |
| --- | --- | --- | --- |
| Источник вариации | Показатель вариации | Число степеней свободы | Несмещенная оценка |
| Факторный признак | 12834.7 | 2 | 6885817.86 |
| Остаточные признаки | 1195.79 | 1070 | 1196.9 |
| Все признаки | 14030.49 | 1072 | 14069.83 |

г) Проверить правило сложения дисперсий

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Показатель | *Dрегр* | *Dост* | *Dобщ* | *Dрегр* + *Dост* |
| Значение | 12834.7 | 1195.79 | 14030.49 | 14030.49 |

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

|  |  |  |
| --- | --- | --- |
| Показатель | Формула расчета | Значение |
| Множественный коэффициент детерминации |  | 0.915 |
| Множественное корреляционное отношение |  | 0.956 |

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

|  |
| --- |
| Тип связи между факторным и результативными признаками, определяемой рассчитанной линейной регрессией - **сильная**. |

*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 | 4589.95 | 5844.45 | 12834.7 |
| Остаточные признаки | 9440.54 | 8186.04 | 1195.79 |
| Все признаки | 14030.49 | 14030.49 | 14030.49 |

*б) Сводная таблица свойств различных регрессионных моделей*

|  |  |  |  |
| --- | --- | --- | --- |
| Свойство | Простейшая линейная модель | Линейная модель с квадратичным членом | Множественная линейная модель |
| Точность | 32.7% | 41.7% | 91.5% |
| Значимость | Да | Да | Да |
| Адекватность | Нет | Нет | Да |
| Степень тесноты связи | Заметная | Заметная | Сильная |

*Вывод (в терминах предметной области)*

|  |
| --- |
| В результате проведённого в п.9 статистического анализа обнаружено, что точность множественной линейной модели гораздо больше точности простейшей л.м и л.м с квадратичным членом. Все регрессионный модели значимы и адекватно (в случае множественной линейной регрессии) отражают зависимость A13 от A5 и A15, однако степень тесноты связи в множественной модели по шкале Чеддока является сильной, когда как в остальных моделях является лишь заметной. |