|  |
| --- |
| МИНОБРНАУКИ РОССИИ |
| Федеральное государственное бюджетное образовательное учреждение высшего образования **«МИРЭА − Российский технологический университет»**  **РТУ МИРЭА** |

**Институт информационных технологий (ИИТ)**

**Кафедра прикладной математики (ПМ)**

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ**

по дисциплине «Технологии и инструментарий анализа больших данных»

**Практическое занятие №3**

|  |  |  |  |
| --- | --- | --- | --- |
| Студент группы ИНБО-01-17 | *ИНБО-07-21, Веригин Никита Алексеевич* | (подпись) | |
| Преподаватель | *Царёв Роман Юрьевич* | (подпись) | |
| Отчет представлен | «\_\_\_»\_\_\_\_\_\_\_\_2024 г. | |  | |

Москва 2024 г.

Введение

Цели работы:

1. Загрузить данные из файла “insurance.csv”.
2. С помощью метода describe() посмотреть статистику по данным. Сделать выводы.
3. Построить гистограммы для числовых показателей. Сделать выводы.
4. Найти меры центральной тенденции и меры разброса для индекса массы тела (bmi) и расходов (charges). Отобразить результаты в виде текста и на гистограммах (3 вертикальные линии). Добавить легенду на графики. Сделать выводы.
5. Построить box-plot для числовых показателей. Названия графиков должны соответствовать названиям признаков. Сделать выводы.
6. Используя признак charges или imb, проверить, выполняется ли центральная предельная теорема. Использовать различные длины выборок n. Количество выборок = 300. Вывести результат в виде гистограмм. Найти стандартное отклонение и среднее для полученных распределений. Сделать выводы.
7. Построить 95% и 99% доверительный интервал для среднего значения расходов и среднего значения индекса массы тела.
8. Проверить распределения следующих признаков на нормальность: индекс массы тела, расходы. Сформулировать нулевую и альтернативную гипотезы. Для каждого признака использовать KS-тест и q-q plot. Сделать выводы на основе полученных p-значений.
9. Загрузить данные из файла “ECDCCases.csv”.
10. Проверить в данных наличие пропущенных значений. Вывести количество пропущенных значений в процентах. Удалить два признака, в которых больше всех пропущенных значений. Для оставшихся признаков обработать пропуски: для категориального признака использовать заполнение значением по умолчанию (например, «other»), для числового признака использовать заполнение медианным значением. Показать, что пропусков больше в данных нет.
11. Посмотреть статистику по данным, используя describe(). Сделать выводы о том, какие признаки содержат выбросы. Посмотреть, для каких стран количество смертей в день превысило 3000 и сколько таких дней было.
12. Найти дублирование данных. Удалить дубликаты.
13. Загрузить данные из файла “bmi.csv”. Взять оттуда две выборки. Одна выборка – это индекс массы тела людей c региона northwest, вторая выборка – это индекс массы тела людей с региона southwest. Сравнить средние значения этих выборок, используя t-критерий Стьюдента. Предварительно проверить выборки на нормальность (критерий Шопиро-Уилка) и на гомогенность дисперсии (критерий Бартлетта).
14. С помощью критерия Хи-квадрат проверить, является ли полученное распределение равномерным. Использовать функцию scipy.stats.chisquare().
15. С помощью критерия Хи-квадрат проверить, являются ли переменные зависимыми. Использовать функцию scipy.stats.chi2\_contingency(). Влияет ли семейное положение на занятость?
16. Оформить отчет о проделанной работе, написать выводы

ХОД РАБОТЫ

Код программ и результаты их выполнения представлены на Рисунках 1-20.

Изображение выглядит как текст, Шрифт, снимок экрана

Автоматически созданное описание

**Рисунок 1 *—* Задача 1**

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 2 *—* Задача 2**

* Непустых значений в каждом из столбцов одинаковое количество – 1338
* Средний возраст - 39.207025, возраст людей находится в диапазоне от 18 до 64 лет
* bmi (средний индекс массы тела) - 30.7, значения находятся в диапазоне от 16 до 53.1
* Среднее количество детей - 1.1, количество находится в диапазоне от 0 до 5
* В страховых выплатах среднее значение - 13270.4. Максимальное - 63770.43, минимальное - 1121.87

**Изображение выглядит как текст, снимок экрана, диаграмма, График

Автоматически созданное описание**

**Рисунок 3 *—* Задача 3**

У всего набора данных 7 столбцов, 3 из которых не числовые: sex, smoker и region.

**Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание**

**Рисунок 4 *—* Задача 3**

* В колонке age значение 20 встречается чаще других
* В колонке children довольно мало значений 4 и 5
* В колонке charges в основном встречаются значения до 15,000
* В колонке bmi хорошее распределение значений

**Изображение выглядит как текст, снимок экрана, программное обеспечение

Автоматически созданное описание**

**Рисунок 5 *—* Задача 4**

**Изображение выглядит как текст, снимок экрана, программное обеспечение

Автоматически созданное описание**

**Рисунок 6 *—* Задача 4**

**Изображение выглядит как текст, снимок экрана, График, диаграмма

Автоматически созданное описание**

**Рисунок 7 *—* Задача 4, гистограммы**

**Изображение выглядит как текст, снимок экрана, программное обеспечение, диаграмма

Автоматически созданное описание**

**Рисунок 8 *—* Задача 4, гистограммы**

Меры центральной тенденции для столбца bmi находятся примерно в одном значении, что нельзя сказать про charges  
Размах (разница между максимальным значением и минимальным) для bmi и charges сильно разнится со стандартным отклонением и межквартильным размахом. Для bmi стандартное отклонение отличается от межквартильного размаха сильнее, чем у столбца charges

**Изображение выглядит как снимок экрана, текст, программное обеспечение, линия

Автоматически созданное описание**

**Рисунок 9 *—* Задача 5**

Среди значений в столбце age нет выбросов, в отличие от столбца bmi. Графики показывают медиану (оранжевая вертикальная линия посередине коробки), внутри коробки 50% значений каждого из столбцов, на усах - остальные 50% значений (за исключением выбросов, они находятся за пределами усов)

**Изображение выглядит как текст, снимок экрана, карта, диаграмма

Автоматически созданное описание**

**Рисунок 10 *—* Задача 6**

Была проверена центральная предельная теорема: среднее значение при каждом распределении выборочных средних разных длин при увеличнеии длины выборки начинает стремиться к среднему значению глобальной совокупности. Также можно заметить, что при увеличении длин выборок, стандартное отклонение начинает стремиться к нулевому значению, что опять же подтверждает центральную предельную теорему.

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 11 *—* Задача 7**

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 12 *—* Код задачи 8**

**Изображение выглядит как линия, График, диаграмма, текст

Автоматически созданное описание**

**Изображение выглядит как График, линия, диаграмма, текст

Автоматически созданное описание**

**Рисунок 13 *—* Задача 8**

На основе результатов проведенного исследования в задании 8 можно сделать вывод, что нулевая гипотеза выполняется для индекса массы тела, но не выполняется для расходов, так как:

* p-значения для выборки bmi больше 0.05, что позволяет отбросить альтернативную гипотезу для индекса массы тела.
* p-значения для выборки charges сильно меньше 0.05, что позволяет отбросить нулевую гипотезу для расходов.
* На графиках, построенных при помощи Q-Q plot видно, что данные индекса массы тела имеют нормальное распределение, а данные расходов — нет.

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 14 *—* Задача 10**

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 15 — Задача 11**

Признаки, содержащие выбросы:

* cases - минимальное значение в этом столбце отрицательное, хотя случаев заболевания не должно быть отрицательное количество
* deaths - минимальное значение в этом столбце тоже отрицательное, хотя смертей не может быть отрицательное количество

Изображение выглядит как текст, программное обеспечение, Мультимедийное программное обеспечение, снимок экрана

Автоматически созданное описание

**Рисунок 16 — Задача 11**

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 17 — Задача 12**

**Изображение выглядит как текст, снимок экрана, программное обеспечение

Автоматически созданное описание**

**Рисунок 18 — Задача 13**

**Изображение выглядит как текст, снимок экрана, Шрифт

Автоматически созданное описание**

**Рисунок 19 — Задача 14**

При проверке критерием Хи-квадрат pvalue больше, чем 0.05, поэтому можно сделать вывод, что распределение является равномерным.

**Изображение выглядит как текст, снимок экрана, программное обеспечение, Мультимедийное программное обеспечение

Автоматически созданное описание**

**Рисунок 20 — Задача 15**

Используя критерий Хи-квадрат (сhi2) и таблицу сопряженности (contingency) удалось проверить, что переменные являются независимыми (pvalue меньше 0.05). Значит, семейное положение влияет на занятость.

ЗАКЛЮЧЕНИЕ

В ходе выполнения практической работы были изучены различные распределения и способы их определения (тест Бартлета, критерий Стьюдента), 95% и 99% доверительные интервалы, а также меры центральной тенденции и меры разброса, для которых были построены графики.