Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

—

Мегафакультет трансляционных информационных технологий

**Институт дизайна и урбанистики**

**ОТЧЕТ ПО ЗАДАНИЮ №3**

Выполнил

Студент гр. **C4110** Егоров П.Н.

Преподаватель Леоненко В.Н.

«\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 г.

Санкт-Петербург

2019

**Задание:**

1. Загрузить выборку в скрипт и найти основные характеристики выборки (математическое ожидание, дисперсия, среднеквадратичное отклонение)
2. Выбрать три гипотетических вероятностных распределения для выборки и оценить их параметры на основе найденных по выборке м.о. и дисперсий
3. Нарисовать полученные плотности вероятности вместе с гистограммой (оценкой плотности по выборке)
4. Подписать на рисунке м.о. и дисперсии
5. Составить отчёт из двух файлов (скрипт в формате .py, файл .docx с описанием работы, таблицей оценок параметров и итоговым графиком).

**Выполнение**:

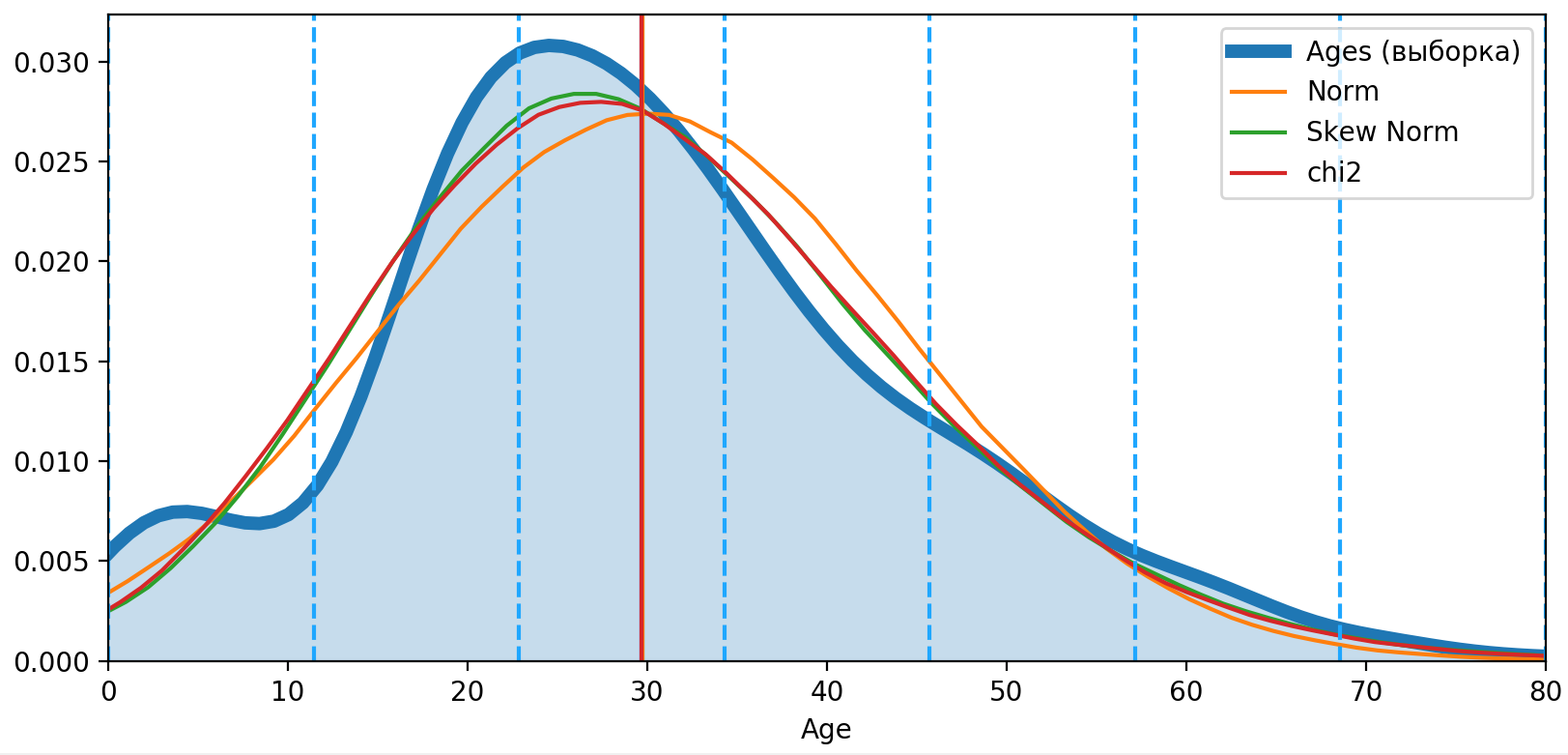
1. Была выбрана выборка возрастов пассажиров корабля «Титаник» из файла **«train.csv»** из сайта:[**https://www.kaggle.com/c/titanic/data**](https://www.kaggle.com/c/titanic/data).

Характеристики выборки:

Mean = 29.70

Variance = 211.02

Std = 14.53

1. Были выбраны следующие гипотетические распределения: симметричное нормальное распределение, несимметричное нормальное распределение, распределение хи-квадрат.
2. Был получен следующий график:

Где красной линией указан Mean, а также:

Norm - симметричное нормальное распределение

Skew Norm - несимметричное нормальное распределение

chi2 - распределение хи-квадрат

1. Таблица полученных параметров:

Если вывести параметры с помощью команд:

print(stats.skewnorm.fit(ages))

print(stats.norm.fit(ages))

print(stats.chi2.fit(ages))

То будут следующие параметры:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Parameter | Mean | Std |
| Skew Norm | 1,77 | 15,71 | 20,16 |
| Norm | - | 29,70 | 14,52 |
| Chi2 | 53,50 | -45,37 | 1,40 |

Но если через другие команды:

print(mean\_chi2.mean())

print(mean\_norm.mean())

print(mean\_skewnorm.mean())

print(mean\_chi2.var())

print(mean\_norm.var())

print(mean\_skewnorm.var())

Были получены параметры:

|  |  |  |
| --- | --- | --- |
|  | Mean | Std |
| Skew Norm | 29,73 | 210,66 |
| Norm | 29,68 | 210,16 |
| Chi2 | 29,68 | 210,34 |

В первом случае сначала генерировались случайные значения заданных распределений по подобранным параметрам (\*stats.chi2.fit(ages)), а потом на основе этих точек строились функции плотности вероятности (sns.distplot).

Во втором случае сразу строилась функция плотности (stats.skewnorm.pdf) с подобранными параметрами (\*stats.chi2.fit(ages)).

Оба варианты приведены в исходном коде (второй вариант закомментирован).

**Вторая часть – Хи-квадрат**

Сначало было разделение на интервалы всего диапазона значений выборки на оптимальное количество интервалов:

intervals = np.linspace(0,80, N\_int)

Далее производился подсчет наблюдений в каждом интервале:

for i in range(N\_int - 1):

n\_j[i] = len(ages[(ages > intervals[i]) & (ages <= intervals[i + 1])])

Как было указано в задании, необходимо провести проверку:

if np.prod(n\_j >= 5):

print('Ожидаемое число попаданий не меньше 5')

flag = 1

else:

print('Ожидаемое число попаданий меньше 5')

flag = 0

Далее высчитываются вероятности попадания наблюдения в j-ый интервал при выполнении гипотезы:

p\_j\_skew\_norm[i] = stats.skewnorm.cdf(intervals[i + 1], \*stats.skewnorm.fit(ages)) - stats.skewnorm.cdf(intervals[i], \*stats.skewnorm.fit(ages))

p\_j\_norm[i] = stats.norm.cdf(intervals[i + 1], \*stats.norm.fit(ages)) - stats.norm.cdf(intervals[i], \*stats.norm.fit(ages))

p\_j\_chi2[i] = stats.chi2.cdf(intervals[i + 1], \*stats.chi2.fit(ages)) - stats.chi2.cdf(intervals[i], \*stats.chi2.fit(ages))

И наконец, высчитывается хи-квадрат:

chi2\_skew\_norm += ((n\_j[i] - n\*p\_j\_skew\_norm[i])\*\*2)/(n\*p\_j\_skew\_norm[i])

chi2\_norm += ((n\_j[i] - n\*p\_j\_norm[i])\*\*2)/(n\*p\_j\_norm[i])

chi2\_chi2 += ((n\_j[i] - n\*p\_j\_chi2[i])\*\*2)/(n\*p\_j\_chi2[i])

Были получены следующие результаты:

'skew\_norm':  8.335104443717391,

'norm': 25.995414488764045,

'chi2': 8.926888168444053}

Находим по таблице значение хи-квадрат при уровне значимости α = 0,05 , что больше чем у 'skew\_norm' и 'chi2'.

То есть выборочное распределение соответствует либо «скошенному» нормальному распределению, либо распределению хи-квадрат.