Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего образования

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

Факультет информационных технологий и программирования

прикладная математика

**Лабораторная работа № 1**

**Статистические методы обработки данных.  
Определение основных статистических характеристик**

Выполнили студенты группы M3306:  
*Шакирова Владислава Эдуардовна  
Залкин Виктор Михайлович*

Проверила:  
*Москаленко Мария Александровна*

Теоретическая основа

Статистические оценки

Среднее значение выборки

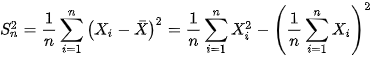
Среднее значение выборки — это сумма элементов выборки, делённая на их количество. Математическое ожидание — это более формализованное понятие для среднего значения выборки, но которое в отличие от последнего, может применяться для непрерывных случайных величин.

Среднее значение выборки:

Математическое ожидание дискретной случайной величины:

Выборочная дисперсия

Выборочная дисперсия — это оценка теоретической дисперсии распределения, рассчитанная на основе данных выборки. Она бывает: смещенная и исправленная. В этой работе мы пользовались значениями смещенной дисперсии.



Стандартная ошибка

Стандартная ошибка (стандартная ошибка среднего) — показывает, стандартное отклонение среднего выборки. Стандартная ошибка среднего это мера разброса средних выборки возле среднего всей выборки.

Стандартная ошибка:

Мода

Мода — значение во множестве наблюдений, которое встречается наиболее часто. (Мода = типичность.) Иногда в совокупности встречается более чем одна мода (например: 6, 2, 6, 6, 8, 9, 9, 9, 0; мода — 6 и 9). В этом случае можно сказать, что совокупность мультимодальна. Из структурных средних величин только мода обладает таким уникальным свойством. Как правило, мультимодальность указывает на то, что набор данных не подчиняется нормальному распределению.

Медиана

Медиана — это такое значение выборки, для которого половина выборки меньше или равна ему него, а другая половина больше или равна ему. Если отсортированная выборка содержит нечётное количество элементов, то медианой является значение в середине выборки, в ином случае значение медианы рассчитывается как среднее значение между двумя значениями выборки в середине выборки.

Медиана для выборки с нечётным количеством элементов:

Медиана для выборки с чётным количество элементов:

Первый, второй и третий квартили

Квантиль — значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Если вероятность задана в процентах, то квантиль называется процентилем или перцентилем.

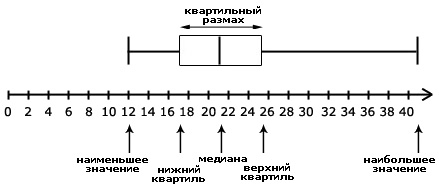
* 0,25-квантиль называется первым (или нижним) квартилем;
* 0,5-квантиль называется медианой или вторым квартилем;
* 0,75-квантиль называется третьим (или верхним) квартилем.

Интерквартильным размахом называется разность между третьим и первым квартилями. Интерквартильный размах является характеристикой разброса распределения величины и является робастным аналогом дисперсии. Вместе, медиана и интерквартильный размах могут быть использованы вместо математического ожидания и дисперсии в случае распределений с большими выбросами, либо при невозможности вычисления последних.

Для расчёта квартили надо поделить вариационный ряд медианой на две равные части, а затем в каждой из них найти медиану. В случае, если вариационный ряд состоит, к примеру, из 9 элементов, тогда за верхнюю квартиль принимают арифметическое среднее 2-го и 3-го элементов, а за нижнюю арифметическое среднее 7-го и 8-го элементов.

Ящик с усами

Ящик с усами (диаграмма размаха) — графическое представление некоторых характеристик выборки: минимум, первый квартиль (квантиль 25%), второй квартиль (квантиль 50%), третий квартиль (квантиль 75%), максимум.



Стандартное отклонение

Среднеквадратическое отклонение (стандартное отклонение) — наиболее распространённый показатель рассеивания значений случайной величины относительно её математического ожидания.



Эксцесс

Коэффициент эксцесса (Kurtosis) — это мера остроты пика (вероятность самых вероятных значений) графика распределения случайной величины по сравнению с графиком нормального распределения.

Положительный коэффициент эксцесса (островершинный график) говорит о том, что пик рассматриваемого распределения выше пика нормального распределения. Отрицательный коэффициент эксцесса (плосковершинный график) говорит об обратном.

Коэффициент эксцесса:

Ассиметричность

Коэффициент асимметрии распределения случайной величины x определяется формулой:

\gamma_1 = \frac{\mu_3}{\sigma^3},где

\mu_3 = \mathbb{E}\left[(x - \mathbb{E}x)^3\right] — третий [центральный момент](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%BE%D0%BC%D0%B5%D0%BD%D1%82%D1%8B_%D1%81%D0%BB%D1%83%D1%87%D0%B0%D0%B9%D0%BD%D0%BE%D0%B9_%D0%B2%D0%B5%D0%BB%D0%B8%D1%87%D0%B8%D0%BD%D1%8B) случайной величины x;

\sigma = \sqrt{\mathbb{D}[x]}— [стандартное отклонение](http://www.machinelearning.ru/wiki/index.php?title=%D0%A1%D1%82%D0%B0%D0%BD%D0%B4%D0%B0%D1%80%D1%82%D0%BD%D0%BE%D0%B5_%D0%BE%D1%82%D0%BA%D0%BB%D0%BE%D0%BD%D0%B5%D0%BD%D0%B8%D0%B5&action=edit) случайной величины x;

Если плотность распределения симметрична, то \gamma_1 = 0.

Если левый хвост распределения тяжелее, то \gamma_1 > 0.

Если правый хвост распределения тяжелее, то \gamma_1 < 0.

Представление данных

Эмпирическая функция распределения

Эмпирической функцией распределения, построенной по выборке называется случайная функция , принимающая значения от 0 до 1 и равная:

Эмпирическая функция распределения — это приближение теоретической функции распределения, построенное с помощью выборки из него.

Гистограмма

Другой характеристикой распределения является таблица для дискретных распределений или плотность — для абсолютно непрерывных. Эмпирическим аналогом таблицы или плотности является так называемая гистограмма.

Гистограмма строится по группированным данным. Предполагаемую область значений случайной величины *x* (или область выборочных данных) делят на некоторое количество не обязательно одинаковых интервалов. Пусть *A1, …, Ak* — интервалы на прямой, называемые интервалами группировки. Обозначим для *j = 1, …, k* через *nj* число элементов выборки, попавших в интервал *Aj*. Обычно берут число интервалов порядка .

На каждом из интервалов *Aj* строят прямоугольник, площадь которого пропорциональна *nj*: Общая площадь всех прямоугольников должна равняться единице. Если *lj* — длина интервала *Aj*; то высота *fj* прямоугольника над этим интервалом равна:



Полученная фигура, состоящая из объединения прямоугольников, называется гистограммой.

Критерии согласия

Критерии согласия проверяют, согласуется ли заданная выборка с заданным фиксированным распределением, с заданным параметрическим семейством распределений, или с другой выборкой.

* Критерий Колмогорова-Смирнова
* Критерий хи-квадрат (Пирсона)
* Критерий омега-квадрат (фон Мизеса)

Критерий согласия Пирсона — статистический критерий для проверки гипотезы о соответствии эмпирического распределения предполагаемому теоретическому распределению (в нашем случае, нормальному распределению) при размере выборки больше 100 элементов.

Использование критерия Пирсона предусматривает разбиение выборки на интервалы и определение числа наблюдений для каждого из интервалов. Для удобства интервалы выбирают одинаковой длины. Интервалы, содержащие менее пяти наблюдений, объединяют с соседними.

Критерий Пирсона:

Полученное значение критерия Пирсона сравнивается с табличным, которое берётся с учётом количества степеней свободы и уровня значимости . Если вычисленное значение меньше табличного, то гипотеза принимается (эмпирическое распределение соответствует теоретическому (сравниваемому) распределению с уровнем значимости ). Иначе, гипотеза отвергается.

Количество степеней свободы:

Интервальное оценивание

Интервал , называется доверительным интервалом для неизвестного параметра , если, с заданной доверительной вероятностью (надежностью) можно утверждать, что неизвестный параметр находится внутри этого интервала (накрывается интервалом).

Доверительный интервал для математического ожидания

Доверительный интервал для математического ожидания имеет вид:

,

где – математическое ожидание, S – стандартное отклонение, а t – определяется по таблице значений критерия Стьюдента.

Для использования этой таблицы, необходимо знать количество степеней свободы. В нашем случае оно составляет *n-1*. А также доверительную надежность (которая дана по условию).

В нашей программе мы пользовались не просто значением *t*, полученным с помощью специальной библиотеки, а уже готовой реализацией нахождения доверительного интервала для математического ожидания из библиотеки scipy.

Доверительный интервал для среднего квадратичного отклонения

Доверительный интервал для среднего квадратичного отклонения имеет вид:

,

где S – стандартное отклонение,  *– уровень значимости , – коэффициент, определяемый по таблице значений.*

Для использования этой таблицы также, необходимо знать количество степеней свободы. В нашем случае оно составляет *n-1*. И уровень значимости (или доверительную надежность).

В нашей программе мы пользовались функцией из библиотеки scipy для нахождения . Эта функция, так же, как и функция для вычисления доверительного интервала для математического ожидания, принимает на вход значение , т. е. доверительную надежность, что отличается от представлений большинства бумажных таблиц. Соответственно, вычисление окончательных значений производились относительно , а не .

Демонстрация работы программы

Гистограмма

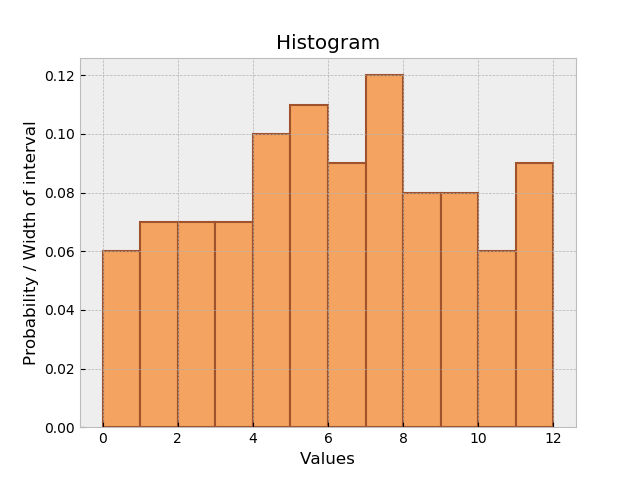


Рисунок 1. Гистограмма выборки

График функции распределения

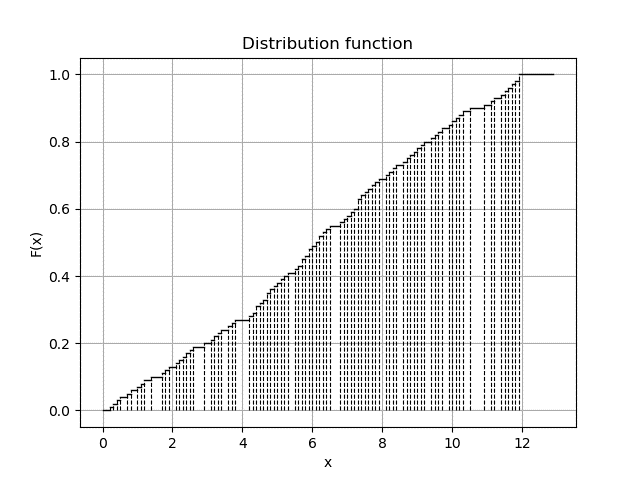


Рисунок 2. Функция распределения выборки

Ящик с усами

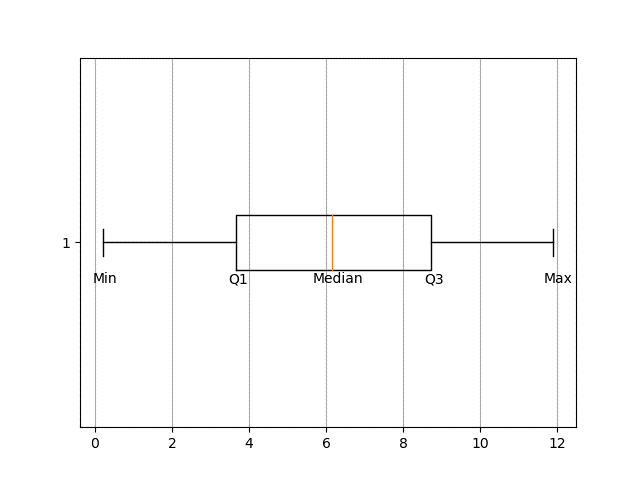


Рисунок 3. Ящик с усами для выборки

Статистические характеристики

Mean: 6.16

Sample variance: 10.53

Standart Error: 0.32

Modus: 7.30

Median: 6.15

First quartile (0.25 quantile): 3.70

Second quartile (0.5 quantile): 6.15

Third quartile (0.75 quantile): 8.80

Mustached Box parameters: 0.20|---3.70|6.15|8.80---|11.90

Standard deviation: 3.24

Excess: -0.95

Asymmetry index: -0.04

Min: 0.20

Max: 11.90

Pearson criterions (X^2 < X^2 table): 10.27 < 19.02 True

Confidence interval for mathematical expectation: (4.18; 8.15)

Confidence interval for standard deviation: (2.85; 3.77)