Тема 6. Описательные статистики для разных типов данных

Практические задания для самостоятельного выполнения

Задания выполняется по вариантам. Формулировка задания общая для всех вариантов; конкретные условия, указанные в общей формулировке, выбираются согласно номеру вашего варианта и описанию задачи.

Результаты выполнения задания необходимо представить в виде двух файлов:

1) ноутбук в формате *ipynb*, содержащий программный код, результаты его выполнения, а также все необходимые пояснения, выводы и комментарии (в текстовых ячейках);
2) файл в формате *pdf*, полученный путем экспорта (или вывода на печать) ноутбука из п. 1).

Внимание: в названии файлов должна обязательно присутствовать фамилия автора. Безымянные работы проверяться не будут. Все выводимые числовые значения на консоли должны быть подписаны. Все графики должны иметь название и названия осей.

Задание 1 (максимум 0,5 балла)

Создать выборку из генеральной совокупности, образованной значениями дискретно распределенной случайной величины (дискретное распределение и его параметры выбрать самостоятельно) малого объема ($n \leq 30$). Выполнить исследование полученной выборки. Для этого:

1. Построить полигон частот.
2. Построить полигон относительных частот и теоретический многоугольник распределения на одном графике.
3. Написать функцию для вычисления значений эмпирической функции распределения.
4. На одном графике построить эмпирическую и теоретическую функцию распределения.
5. Вычислить выборочное среднее, моду и медиану.
6. Вычислить выборочную дисперсию и исправленную дисперсию.
7. Вычислить выборочное среднее квадратическое отклонение и исправленное среднее квадратическое отклонение.

Задание 2 (максимум 0,5 балла)

Создать выборку из генеральной совокупности, образованной значениями непрерывно распределенной случайной величины (непрерывное распределение и его параметры выбрать самостоятельно) большого объема ($n \gg 30$). Выполнить исследование полученной выборки. Для этого:

1. Построить гистограмму частот.
2. Построить гистограмму относительных частот и теоретическую плотность распределения на одном графике.
3. На одном графике построить эмпирическую и теоретическую функцию распределения.
4. Вычислить выборочное среднее, медиану.
5. Вычислить выборочную дисперсию и исправленную дисперсию.
6. Вычислить выборочное среднее квадратическое отклонение и исправленное среднее квадратическое отклонение.

Задание 3 (максимум 1 балл)

Используя набор данных, выполнить исследование имеющихся в нем признаков. Наборы данных по вариантам представлены в *csv*-файлах. Имя файла: *Вариант N.3*, где N – номер варианта.

Для решения задания необходимо выполнить следующие шаги:

1. Импортировать данные наблюдений из файла. Вывести несколько первых записей для проверки корректности импорта и получения первого представления о значениях признаков.
2. Для каждого из признаков по описанию, данному в варианте, и по наблюдаемым значениям определить:
   - является ли признак категориальным (с указанием вида – номинальный, бинарный, порядковый) или количественным,
   - тип шкалы измерения значений признака,
   - для количественных признаков – является ли признак дискретным или непрерывным.

   Результаты вместе с объяснениями записать в текстовой ячейке.
3. Выполнить визуализацию статистического распределения признака, соответствующую типу этого признака.
4. Для каждого признака вычислить те из статистических оценок, которое допустимы для шкалы измерений этого признака:
   - выборочная мода,
   - выборочные первый и третий квартили,
   - выборочная медиана,
   - выборочная средняя,
   - выборочная дисперсия и/или исправленная дисперсия,
   - выборочное среднее квадратическое отклонение и/или исправленное среднее квадратическое отклонение.
5. Для каждого признака объяснить выбор оценок и дать интерпретацию полученным оценкам (записать в текстовой ячейке).

Задание 4 (максимум 1 балл)

Выполнить исследование двух признаков X и Y, о которых известно, что они распределены нормально. Данные наблюдений представлены в *csv*-файле (по вариантам). Имя файла: *Вариант N.4*, где N – номер варианта. В ходе решения задания необходимо выполнить следующие шаги:

1. Импортировать данные наблюдений из файлов. Извлечь выборки, избавившись от `NaN`, если они присутствовали среди значений признака.
2. Для каждого признака вычислить объем выборки, выборочную среднюю.
3. Для каждого признака получить доверительный интервал с надежностью 0,95 для параметра m нормального распределения двумя способами: непосредственно (метод `interval`) и используя точность оценки.

   Указание: при выборе формулы доверительного интервала для параметра m учитывайте данную в варианте информацию.

4. Для признака Y вычислить исправленное выборочное среднее квадратическое отклонение.
5. Для признака Y получить доверительный интервал для параметра σ нормального распределения.
6. Для каждого полученного доверительного интервала в тестовой ячейке записать сам интервал в виде двойного неравенства.
7. Для признака X получить доверительные интервалы для параметра m для еще не менее чем трех значений надежности. Для каждого интервала записать в текстовой ячейке сам интервал, его точность и надежность.
8. Основываясь на полученных результатах, ответить на вопрос:
   - как изменяется точность доверительного интервала с уменьшением/увеличением надежности оценки?

   Ответ записать в текстовой ячейке