МИНОБРНАУКИ РОССИИ

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ

ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)

Кафедра ИС

ОТЧЁТ

по практической работе №2

по дисциплине «Статистический анализ»

Тема: Обработка выборочных данных.

Нахождение точечных оценок параметров распределения

Вариант № 8

Студент гр. 9372 Иванов Р.С.

Преподаватель Сучков А.И.

Санкт-Петербург

2021

Цель работы.

Получение практических навыков нахождения точечных статистических оценок параметров распределения

Основные теоретические положения.

Выборочным средним называется среднее арифметическое всех значений выборочной совокупности (обозначения: $\bar{x}_{\text{в}}$, $\bar{x}$, M*[X], m_x^*).

$$\bar{x}_{\text{в}} = \sum_{i=1}^{k} x_i p_i^* \tag{1}$$

Выборочной дисперсией называется среднее арифметическое квадратов отклонений вариант от выборочной средней.

$$D_{\text{в}} = \sum_{i=1}^{\kappa} (x_i - \bar{x}_{\text{в}})^2 p_i^* \tag{2}$$

Выборочное среднее квадратическое отклонение определяется по формуле:

$$\sigma_{\text{в}} = \sqrt{\text{D}_{\text{в}}} \tag{3}$$

Модой M_0^* вариационного ряда называется такое значение варианты, которой соответствует наибольшая частота

Медианой M_{e}^* называется значение признака, приходящееся на середину ранжированного ряда наблюдений.

Постановка задачи.

Для заданных выборочных данных вычислить с использованием метода моментов и условных вариант точечные статистические оценки математического ожидания, дисперсии, среднеквадратичного отклонения, асимметрии, эксцесса, моды, медианы и коэффициента вариации исследуемой случайной величины. Полученные результаты содержательно проинтерпретировать.

Выполнение работы.

Во время выполнения работы был написан код на языке Python, выполняющий поставленную задачу. Реализовано чтение и дальнейшая обработка данных из Price_Mileage.csv, одобренного преподавателем. Для этого была использована библиотека csv.

Выборка была сформирована во время выполнения практической работы 0 и использована в ходе выполнения этой практической работы. Объём выборки составлял 115 позиций.

Полученные в программе значения были записаны в *Result.xlsx* с помощью библиотеки *openpyxl* для дальнейшего представления в отчёте.

Пункт 1

Для интервального ряда, полученного в практической работе №1 были найдены середины интервалов, а также накопленные частоты. Результаты занесены в таблицу.

Таблица 1 – Середины интервалов, накопленные частоты

Интервал	(0, 27023]	(27023, 54046]	(54046, 81069]	(81069, 108092]	(108092, 135115]	(135115, 162138]	(162138, 189161]	(189161, 216184]	(216184, 243207]
Середина	13511.5	40534.5	67557.5	94580.5	121603.5	148626.5	175649.5	202672.5	229695.5
Нак. частота	0	0.304	0.687	0.809	0.922	0.957	0.965	0.991	0.991

Пункт 2

Для полученных вариант были вычислены условные варианты. Результаты занесены в таблицу.

Таблица 2 – Условные варианты

Интервал	(0, 27023]	(27023, 54046]	(54046, 81069]	(81069, 108092]	(108092, 135115]	(135115, 162138]	(162138, 189161]	(189161, 216184]	(216184, 243207]
Условная	-4	-3	-2	-1	0	1	2	3	4

Пункт 3

Вычислить условные эмпирические моменты v_r^* через условные варианты. С помощью условных эмпирических моментов вычислить центральные эмпирические моменты μ_r^*. Полученные результаты занести в таблицу.

Таблица 3 – Начальные и центральные эмпирические моменты

Номер момента	1	2	3	4
v_r^*	-2.62609	9.165217	-30.1217	113.6174
μ_r^*	0	2.268885	5.863496	5865.676

Пункт 4

Вычислены выборочные среднее и дисперсии с помощью стандартной формулы и с помощью условных вариант.

$\bar{x}_в$ через стандартную формулу	$\bar{x}_в$ через условные варианты
50638.75	50638.75

Дисперсия через стандартную формулу	Дисперсия через условные варианты
1656836092	1656836092

Пункт 5

Вычислены исправленная выборочная дисперсия и исправленное СКО. Они почти не отличаются т. к. объём выборки больше 30.

Исправленная выборочная дисперсия	Исправленное выборочное СКО
1671369743	40882.39

Смещённая оценка дисперсии	Смещённая оценка СКО
1656836093	40704.25

Пункт 6

Найдена статистическая оценка коэффициентов асимметрии и эксцесса.

Оценка асимметрии	Оценка эксцесса
1.71568666	1139.445

Оценка асимметрии позволяет нам сказать, что наше распределение отклоняется от нормального во второй половине графика после моды, она более вытянута, чем должна быть.

Оценка эксцесса позволяет нам сказать, что наше распределение имеет намного более высокую и острую вершину по сравнению с нормальным распределением

Пункт 7

Была вычислена мода и медиана для заданного распределения.

Мода	Медиана
33259	39029

Значения моды и медианы расположены в самом начале диапазона значений выборки. Зная, что представляет собой выборка, можно сделать вывод о том, наибольшее количество машин, представленных в выборке, имеют пробег около 33000 миль, а число машин с пробегом менее 39000 миль примерно равно числу машин с пробегом больше.

Пункт 8

Был вычислен коэффициент вариации.

$$V^* \approx 124\%$$

Значение коэффициента вариации больше 100 процентов, поэтому можно сказать, что выборка является неоднородной.

Вывод.

Изучены основные правила вычисления числовых характеристик выборки. Освоен метод упрощённых вычислений. Получены условные эмпирические начальные и центральные моменты до 4 порядка. Оценён график распределения выборки с помощью статистических оценок асимметрии и эксцесса. Найдена мода и медиана, сделан вывод о значениях выборки. Посчитан коэффициент вариации, сделан вывод о неоднородности выборки.

ПРИЛОЖЕНИЕ А

ИСХОДНЫЙ КОД ПРОГРАММЫ

```
import math
import csv
from openpyxl import Workbook

sample = []

with open('Price-Mileage.csv') as csv_file:  # Читаем выборку из файла 0 работы
    spam_reader = csv.reader(csv_file, quotechar='|')
    for row in spam_reader:
        x, y = row[0].split(';')
        if y.isdigit(): sample.append(int(y))

sample.sort()  # Используя встроенную функцию сортировки получаем ранжированный ряд

R = sample[len(sample) - 1] - sample[0]  # Размах
print("R =", R)

k = round(1 + math.log2(len(sample)))  # Число интервалов(Формула Стёрджеса)
print("k = ", k)

h = round(R / k)  # Длина интервала
print("h =", h)

k += 1  # Иначе интервалы не покроют выборку

x0 = sample[0] - h / 2  # Начало первого частичного интервала
if x0 < 0:
    x0 = 0

print("x0 =", x0)

interval = []
x = x0

for i in range(k):
    interval.append([(x, x + h), 0, 0])
    x += h

# Получаем интервальный ряд
for i in sample:
    for j in range(k):
        if interval[j][0][0] < i <= interval[j][0][1]:
            interval[j][1] += 1
            break

for i in interval:
    i[2] = i[1] / len(sample)

print(interval)

middle_int = []
accum_freq = []
accum_afreq = []
accum_freq.append(0)
```

```
accum_afreq.append(0)
a = 0
b = 0

# Вычисляем серидины интервалов и их накопленные частоты
for i in range(len(interval)):
    a = a + interval[i][2]
    b = b + interval[i][1]
    middle_int.append(interval[i][0][0] + h / 2)
    accum_afreq.append(a)
    accum_freq.append(b)

C = middle_int[int(k/2)]  # Число C из теории, h было вычислено ранее при построении
интервального ряда
con_var = []  # Условные варианты

for i in range(k):
    con_var.append([int((middle_int[i] - C)/h), interval[i][1], interval[i][2]])

SEM = []  # Выборочные начальные моменты до 4 порядка
CEM = []  # Выборочные центральные моменты до 4 порядка
# Вычисляем условные эмп момент 1 порядка
for i in range(4):
    SEM.append(0)
    for j in range(k):
        SEM[i] += (con_var[j][0]**(i+1))*con_var[j][2]

CEM.append(0)   # Центральный момент 1 порядка (выборочное среднее для усл вариант)
CEM.append(SEM[1]-SEM[0]**2)  # 2 порядка (выборочная дисперсия для усл вариант)
CEM.append(SEM[2]-3*SEM[1]*SEM[0]+2*SEM[0]**3)   # 3 порядка
CEM.append(SEM[3]-4*SEM[0]*SEM[3]+6*(SEM[0]**2)*SEM[3]-3*SEM[0]**4)  # 4 порядка

Xs = 0      # Выборочное среднее по обычной формуле
CXs = 0     # Выборочное среднее через моменты условных вариант
Ds = 0      # Дисперсия по обычной формуле
CXs = 0     # Дисперсия через моменты условных вариант

for i in range(k):
    Xs += middle_int[i]*interval[i][2]

for i in range(k):
    Ds += (middle_int[i] - Xs)**2*interval[i][2]

CXs = SEM[0]*h + C          # Из формулы метода упрощённых вычислений
CDs = CEM[1]*(h**2)         # Из формулы метода упрощённых вычислений
ds = math.sqrt(Ds)          # Выборочное СКО
cds = math.sqrt(CEM[1])     # Выборочное СКО для усл вариант

Asym = CEM[2]/(cds**3)       # Коэффициэнт ассиметрии для условных вариант (как я
понимаю это и есть стат оценка?)
Excess = CEM[3]/(cds**4)     # Коэффициент эксцесса для условных вариант

sDs = Ds*len(sample)/(len(sample)-1)    # Исправленная выборочная дисперсия
s = math.sqrt(sDs)                      # Исправленное выборочное СКО

Mod = 0      # Частость модального интервала
i0 = 0      # Его номер
Mod0 = 0     # Истинное значение моды
```

```
for i in range(k):
    if interval[i][1] > Mod:
        Mod = interval[i][1]
        i0 = i

Mod0 = int(interval[i0][0][0] + h*((interval[i0][2] - interval[i0-
1][2])/((interval[i0][2] - interval[i0-1][2])+(interval[i0][2] - interval[i0+1][2]))))

i0 = 0      # Номер медианного интервала
Med0 = 0     # Истинное значение медианы

for i in range(k):
    if accum_afreq[i] > 0.5:
        i0 = i
        break

px = 0      # Число нужное для линейной интерполяции медианы
for i in range(i0-1):
    px += 1/interval[i][2]

# Med0 = interval[i0][0][1] + (h/interval[i0][2])*(0.5 - h*px)
# При выполнении линейной интерполяции для медианы получилось неадекватно большое
отрицательное число
# Я не знаю с чем это связано поэтому найду её просто как серидину ранжированной ряда

Med0 = sample[int(len(sample)/2)]

CV = Xs/ds      # Коэффициент вариации

wb = Workbook()
ws = wb.create_sheet()

# filename = "output1.xlsx"
# ws = wb.active
ws.title = "Result"
ws["A1"] = "Интервальный"
ws.append([str(interval[i][0]) for i in range(len(interval))])
ws["A3"] = "Середины"
ws.append(middle_int)
ws["A5"] = "Накопленный частоты"
ws.append(accum_afreq[i] for i in range(k))
ws["A7"] = "Условные"
ws.append(con_var[i][0] for i in range(len(interval)))
ws["A9"] = "Начальные эмп моменты"
i = [1, 2, 3, 4]
ws.append(i)
ws.append(SEM)
ws["A12"] = "Центральные эмп моменты"
ws.append(CEM)

ws["A14"] = "Хв через стандартную формулу"
ws["A15"] = Xs
ws["B14"] = "Хв через условные варианты"
ws["B15"] = CXs

ws["A16"] = "Дисперсия через стандартную формулу"
ws["A17"] = Ds
```

```
ws["B16"] = "Дисперсия через условные варианты"
ws["B17"] = CDs

ws["A18"] = "Исправленная выборочная дисперсия"
ws["A19"] = sDs
ws["B18"] = "Исправленное выборочное СКО"
ws["B19"] = s
ws["A20"] = "Смещённая оценка дисперсии"
ws["A21"] = Ds
ws["B20"] = "Смещённая оценка СКО"
ws["B21"] = ds

ws["A22"] = "Оценка ассиметрии"
ws["A23"] = Asym
ws["B22"] = "Оценка эксцесса"
ws["B23"] = Excess

ws["A24"] = "Мода"
ws["A25"] = Mod0
ws["B24"] = "Медиана"
ws["B25"] = Med0

ws["A26"] = "Коэффициэнт вариации"
ws["A27"] = CV

wb.save('Result.xlsx')
```