Санкт-Петербургский политехнический университет Петра Великого
Институт компьютерных наук и технологий
Высшая школа программной инженерии

Отчет по расчетному заданию

по дисциплине «Теория вероятностей и математическая статистика»

Вариант 62

Выполнил
Студент гр. 5130904/10103 Бендрышев С.А

Руководитель Зайцев И.В

Санкт-Петербург
2023

Оглавление

Таблица нумерованных денормализованных чисел

1521.0	1538.3	1528.8	1501.9
1501.3	1570.7	1528.3	1553.3
1530.7	1435.1	1485.8	1464.7
1449.6	1488.4	1480.9	1441.3
1463.0	1492.6	1514.0	1507.0
1555.0	1526.7	1551.3	1498.2
1497.5	1509.2	1453.4	1500.0
1553.0	1467.7	1535.4	1504.1
1463.2	1465.0	1496.7	1482.9
1514.1	1513.2	1523.6	1467.8
1541.8	1462.7	1521.4	1491.5
1493.8	1517.2	1570.7	1505.7
1539.3	1537.8	1471.4	1486.8
1538.8	1493.4	1510.7	1465.3
1472.4	1496.4	1524.3	1471.8
1498.6	1480.8	1508.6	1570.6
1511.9	1536.1	1497.9	1504.0
1465.1	1540.5	1510.1	1550.6
1543.9	1503.1	1477.6	1494.6
1459.6	1453.0	1537.4	1538.2
1533.6	1502.2	1489.5	1544.8
1557.4	1478.8	1507.0	1502.7
1480.6	1516.8	1535.9	1495.0
1444.5	1540.1	1570.3	1490.9
1530.0	1551.2	1505.1	1539.9
1527.3	1512.8	1504.4	1541.9
1528.1	1557.6	1542.7	1493.9
1519.7	1526.3	1485.4	1557.9
1520.1	1490.5	1524.9	1475.5
1513.0	1587.3	1481.1	1542.4
1519.5	1501.9	1520.6	1509.5
1551.0	1507.8	1500.5	1533.5
1488.9	1503.6	1529.0	1479.7
1482.1	1554.8	1514.9	1506.4
1518.6	1532.6	1480.7	1479.7
1490.6	1472.9	1544.7	1505.7
1559.1	1462.9	1544.9	1503.8
1513.3	1465.8	1503.4	1522.6
1488.9	1548.8	1451.2	1544.1
1528.5	1506.0	1482.3	1567.8
1501.5	1512.4	1497.8	1478.9
1456.9	1511.0	1478.9	1447.0
1554.1	1504.4	1545.7	1503.6
1485.6	1553.1	1494.6	1499.1
1403.0	1516.8	1555.4	1505.3
1473.3	1511.3	1495.5	1479.5
1544.0	1539.4	1513.7	1566.8
1493.5	1470.4	1542.5	1551.4
1511.4	1517.9	1470.4	1534.7
1521.7	1516.8	1460.5	1485.2

Таблица 1.

Вариационный ряд

Вариационный ряд – совокупность значений признака, записанных в порядке возрастания. При этом сам признак называют вариантой.

1403	1486.8	1507	1534.7
1435.1	1488.4	1507.8	1535.4
1441.3	1488.9	1508.6	1535.9
1444.5	1488.9	1509.2	1536.1
1447	1489.5	1509.5	1537.4
1449.6	1490.5	1510.1	1537.8
1451.2	1490.6	1510.7	1538.2
1453	1490.9	1511	1538.3
1453.4	1491.5	1511.3	1538.8
1456.9	1492.6	1511.4	1539.3
1459.6	1493.4	1511.9	1539.4
1460.5	1493.5	1512.4	1539.9
1462.7	1493.8	1512.8	1540.1
1462.9	1493.9	1513	1540.5
1463	1494.6	1513.2	1541.8
1463.2	1494.6	1513.3	1541.9
1464.7	1495	1513.7	1542.4
1465	1495.5	1514	1542.5
1465.1	1496.4	1514.1	1542.7
1465.3	1496.7	1514.9	1543.9
1465.8	1497.5	1516.8	1544
1467.7	1497.8	1516.8	1544.1
1467.8	1497.9	1516.8	1544.7
1470.4	1498.2	1517.2	1544.8
1470.4	1498.6	1517.9	1544.9
1471.4	1499.1	1518.6	1545.7
1471.8	1500	1519.5	1548.8
1472.4	1500.5	1519.7	1550.6
1472.9	1501.3	1520.1	1551
1473.3	1501.5	1520.6	1551.2
1475.5	1501.9	1521	1551.3
1477.6	1501.9	1521.4	1551.4
1478.8	1502.2	1521.7	1553
1478.9	1502.7	1522.6	1553.1
1478.9	1503.1	1523.6	1553.3
1479.5	1503.4	1524.3	1554.1
1479.7	1503.6	1524.9	1554.8
1479.7	1503.6	1526.3	1555
1480.6	1503.8	1526.7	1555.4
1480.7	1504	1527.3	1557.4
1480.8	1504.1	1528.1	1557.6
1480.9	1504.4	1528.3	1557.9
1481.1	1504.4	1528.5	1559.1
1482.1	1505.1	1528.8	1566.8
1482.3	1505.3	1529	1567.8
1482.9	1505.7	1530	1570.3
1485.2	1505.7	1530.7	1570.6
1485.4	1506	1532.6	1570.7
1485.6	1506.4	1533.5	1570.7
1485.8	1507	1533.6	1587.3

Оценка математического ожидания, дисперсии, медианы

Оценка математического ожидания – среднее арифметическое.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = 1508.734$$

Смещенная оценка дисперсии по всей выборке.

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$$

$$s^2 = 1007.921$$

$$s = 31.748$$

Несмещенная оценка дисперсии

$$\widehat{s^2} = \frac{ns^2}{n-1}$$

$$\widehat{s^2} = 1012.986$$

$$\widehat{s} = 31.907$$

Оценка медианы – значение варианты, которое делит вариационный ряд на две равные по числу членов части. При четном числе членов ($n = 2k$) в качестве медианы принимают

$$\widetilde{Me} = \frac{x_k + x_{k+1}}{2}$$

$$\widetilde{Me} = 1507$$

Размах варьирования (широта распределения) – разность между наибольшим и наименьшим значениями варианты

$$R = x_{max} - x_{min} = 184.300$$

Группирование, графики статистических распределений

Сгруппируем данные в 11 интервалов. Возьмем ширину интервала равной 17.
Таблица подсчета частот и частотностей по интервалам вариационного ряда:

N	Begin	End	X_i	F	$\frac{F_i}{N}$	acc($\frac{F_i}{N}$)
1	1403	1419.8	1411.4	1	0.005	0.005
2	1419.8	1436.7	1428.25	1	0.005	0.01
3	1436.7	1453.5	1445.1	7	0.035	0.045
4	1453.5	1470.4	1461.95	14	0.07	0.115
5	1470.4	1487.2	1478.8	28	0.14	0.255
6	1487.2	1504.1	1495.65	39	0.195	0.45
7	1504.1	1520.9	1512.5	40	0.2	0.65
8	1520.9	1537.8	1529.35	25	0.125	0.775
9	1537.8	1554.6	1546.2	31	0.155	0.93
10	1554.6	1571.5	1563.05	13	0.065	0.995
11	1571.5	1588.3	1579.9	1	0.005	1.0

Таблица 3. Группировка

Полигон

На оси абсцисс откладываются интервалы значений величины x, в серединах интервалов строятся ординаты, пропорциональные частотностям, и концы ординат соединяются отрезками прямых линий.

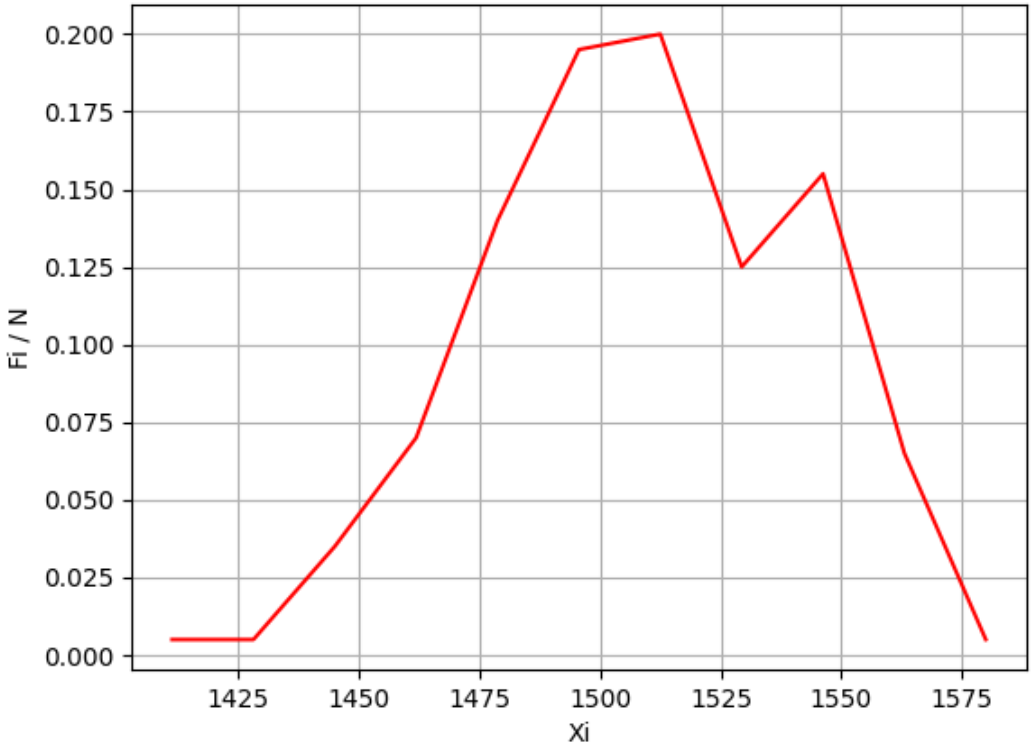

Рис. 1. Полигон

Гистограмма

Над каждым отрезком оси абсцисс, изображающим интервал значений x, строится прямоугольник, площадь которого пропорциональна частотности в данном интервале.

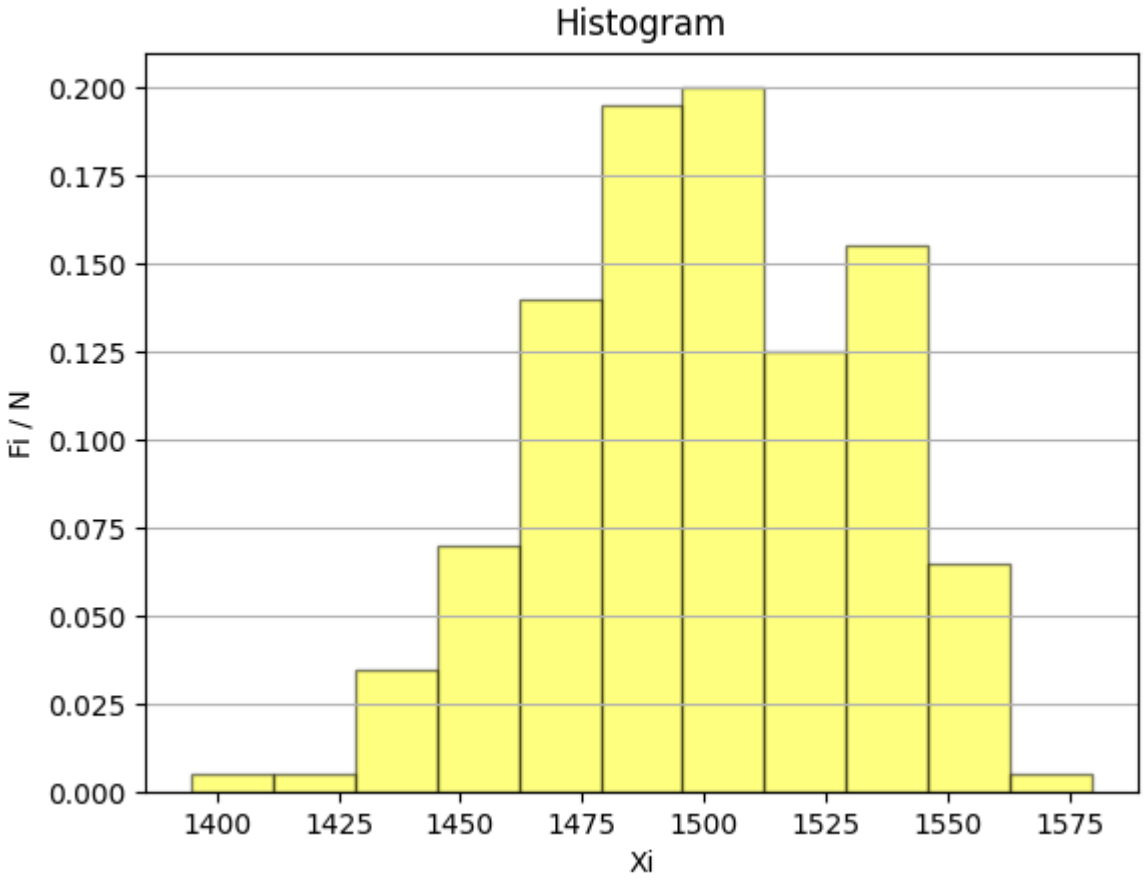

Рис. 2. Гистограмма

Ступенчатая кривая

Над каждым отрезком оси абсцисс, изображающим расстояние между серединами интервалов значений x, проводится отрезок горизонтальной прямой на высоте, пропорциональной накопленной частости в данном интервале. Концы отрезков соединяются. Накопленной частостью в данном интервале называется сумма всех частостей, начиная с первого интервала до данного интервала включительно.

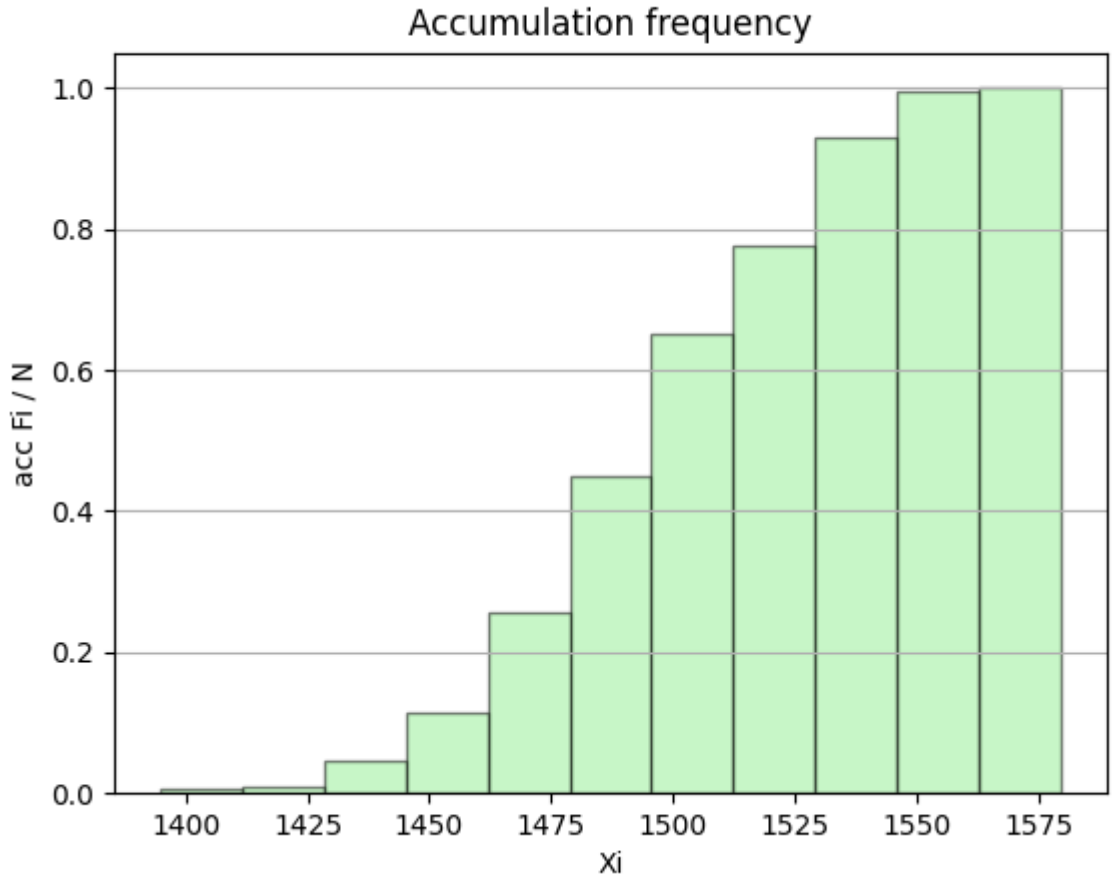

Рис. 3. Ступенчатая кривая

Точечная оценка характеристик распределения по сгруппированным данным

Эмпирические числовые характеристики случайных величин подобно теоретическим характеристикам подразделяются на характеристики положения центра группирования и характеристики рассеивания.

Характеристиками положения являются среднее арифметическое $\bar{x}$, оценки медианы $\tilde{Me}$ и моды $\tilde{Mo}$, а эмпирическими характеристиками рассеивания – дисперсия s^2, коэффициент вариации $\vartheta = \frac{s}{\bar{x}}$(при $\bar{x} \neq 0$), размах варьирования R и положение крайних (экстремальных) членов выборки.

Среднее арифметическое

$$\bar{x} = \frac{\sum_{i=1}^{l} x_i v_i}{n}$$

Здесь x_i- середина интервала, l - число интервалов, v_i - частота в интервале, n - число элементов в выборке

$$\bar{x} = 1508.62$$

Мода - середина самого многочисленного интервала

$$\tilde{Mo} = 1512.5$$

Эмпирическая дисперсия и среднее квадратическое отклонение:

$$s^2 = \frac{1}{n}\sum_{i=1}^{l} v_i (x_i - \bar{x})^2$$

$$s^2 = 1024.14$$

$$s = 32.002$$

Коэффициент вариации

$$\vartheta = \frac{s}{\bar{x}} = 0.021$$

Оценка дисперсии, полученная по сгруппированным данным, оказывается смещенной (несколько увеличенной). Исправляют это смещение с помощью поправки Шеппарда.

$$s_*^2 = s^2 - \frac{(\Delta x)^2}{l} == 1002.27$$

Асимметрия:

$$\tilde{S}_k = \frac{m_3}{s^3} = -0.142$$

$$m_3 = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^3 v_i}{\sum_{i=1}^{l} v_i}$$

Эксцесс:

$$m_4 = \frac{\sum_{i=1}^{l} (x_i - \bar{x})^4 v_i}{\sum_{i=1}^{l} v_i}$$

$$\tilde{Ex} = \frac{m_4}{s^4} - 3 = -0.220$$

Оценка параметров распределения методом квантилей

Предположим, что выборка подчиняется нормальному закону распределения. Определим оценки параметров этого закона математического ожидания m и среднего квадратичного отклонения σ, используя метод квантилей.

Для определения оценок этих двух неизвестных составим два уравнения, используя формулу

$$\Phi(\frac{x_p - m}{\sigma}) + 0.5 = p$$

Для этого возьмем два элемента из выборки с порядковыми номерами $\frac{n}{4}$ и $\frac{3n}{4}$. Их вероятности равны соответственно 0.25 и 0.75. Получим систему из двух уравнений

$$\Phi(\frac{x_{p1} - m}{\sigma}) + 0.5 = 0.25$$
$$\Phi(\frac{x_{p2} - m}{\sigma}) + 0.5 = 0.75$$

Решив систему, получим
m = 1510.750
σ = 35.508

Построение доверительных интервалов

Первые двадцать значений таблицы нумерованных денормализованных чисел, справа их вариационный ряд

1	1521.0	1449.6
2	1501.3	1459.6
3	1530.7	1463.2
4	1449.6	1465.1
5	1463.0	1472.4
6	1555.0	1493.8
7	1497.5	1497.5
8	1553.0	1498.6
9	1463.2	1501.3
10	1514.1	1511.9
11	1541.8	1514.1
12	1493.8	1530.7
13	1539.3	1538.8
14	1538.8	1539.3
15	1472.4	1541.8
16	1498.6	1543.9
17	1511.9	1463.0
18	1465.1	1521.0
19	1543.9	1553.0
20	1459.6	1555.0

Доверительный интервал для математического ожидания

В предположении нормального распределения отклонения значений от номинала построим доверительный интервал для математического ожидания при неизвестном σ по первым двадцати значениям (n = 20) данной выборки по формуле

$$\bar{x} - t_{q,n-1}\frac{s}{\sqrt{n-1}} < m_x < \bar{x} + t_{q,n-1}\frac{s}{\sqrt{n-1}}$$

Здесь $\bar{x}$ - среднее арифметическое первых 20 чисел, s - их среднее квадратичное отклонение

$$\bar{x} = 1505.680$$
$$s = 33.586$$

Определим доверительные интервалы для математического ожидания m_x при различных значимостях

$$q = 0.01, m_x \in (1483.636,\ 1527.724)$$
$$q = 0.05,\ m_x \in (1489.553,\ 1521.807)$$
$$q = 0.1,\ m_x \in (1492.357,\ 1519.003)$$

Доверительный интервал для среднеквадратичного отклонения

$$\frac{\sqrt{n}s}{\chi_2} < \sigma_x < \frac{\sqrt{n}s}{\chi_1}$$

$$q = 0.01, \chi_1 = 2.616, \chi_2 = 6.211, \sigma_x \in (24.181, 57.415)$$

$$q = 0.05, \chi_1 = 2.984, \chi_2 = 5.731, \sigma_x \in (26.206, 50.329)$$

$$q = 0.1, \chi_1 = 3.181, \chi_2 = 5.49, \sigma_x \in (27.358, 47.223)$$

Критерий хи-квадрат для проверки статистических гипотез

Проверим гипотезу о том, что выборка, данная в таблице 1, не противоречит нормальному закону распределения. Для проверки этой гипотезы воспользуемся критерием

$$\chi^2 = \sum_{i=1}^{l} \frac{(v_i - n\widetilde{p}_l)^2}{n\widetilde{p}_l}$$

Если численное значение критерия χ^2 попадает в критическую область $\chi^2 > \chi_q^2$, то гипотеза отвергается.

Оценку вероятности $\widetilde{p}_l$ находим по формуле

$$\widetilde{p}_l = \Phi\left(\frac{\beta - \bar{x}}{s}\right) - \Phi\left(\frac{\alpha - \bar{x}}{s}\right)$$

Где α и β – границы интервалов, $\bar{x}$ и s вычислены ранее по данной выборке.

Если число оцениваемых по выборке параметров равно с, то на уклонения $v_i - n\widetilde{p}_i$ накладываются тем самым ещё c связей, поэтому число независимых между собой уклонений в этом случае будет $l - c - 1$. Так как по данным выборки мы оценили параметры m_x и σ_x нормального закона, то в нашем случае число степеней свободы будет равно

$$k = l' - c - 1 = 10 - 2 - 1 = 7$$

Для нашей выборки $\chi^2 = 11.276$

$$q = 0.01,\ \chi_q^2 = 18.475$$

$$q = 0.05,\ \chi_q^2 = 14.067$$

$$q = 0.1,\ \chi_q^2 = 12.017$$

Как мы видим, на всех этих уровнях значимости нет оснований отвергнуть гипотезу о нормальности выборки

Проверка гипотезы об однородности выборки с помощью критериев знаков и Вилкоксона

Введем нулевую гипотезу H_0 о том, что выборка, данная в таблице 1, является однородной. Для проверки этой гипотезы возьмем двадцать первых и двадцать последних значений таблицы 1 (две выборки). Воспользуемся непараметрическими (независимыми от формы распределения) критериями: критерием знаков и критерием Вилкоксона.

Критерий знаков

Составим разность $z_i = x_i - y_i$, где $i = 1, 2, \ldots, 20$ – порядковые номера первых x_i и последних y_i двадцати значений выборки. Подсчитаем число положительных $k_n(+)$ и отрицательных $k_n(-)$ знаков разностей z_i (n = 20).

Затем, выбрав уровень значимости q, находим по q и n критическое значение $\bar{m}_n$ меньшего из чисел положительных и отрицательных знаков z_i. Если теперь меньшее из чисел знаков разностей окажется меньше $\bar{m}_n$, то гипотеза об однородности выборки отвергается, а если меньшее из чисел знаков разностей окажется больше $\bar{m}_n$, то следует признать, что гипотеза не противоречит данным выборки.

Для данных из таблицы 1 имеем:

$$k_n(+) = 10$$
$$k_n(-) = 10$$

Из таблицы по n = 20 получаем $\bar{m}_n$ для различных q

При $q = 5\%$: $\bar{m}_n = 6$
При $q = 1\%$: $\bar{m}_n = 4$
При $q = 10\%$: $\bar{m}_n = 6$

Так как минимальное из чисел $k_{20}(+)$ и $k_{20}(-)$ – 10 больше $\bar{m}_n$ для всех выбранных уровней значимости, нулевая гипотеза H_0 об однородности выборки не противоречит данным выборки.

Критерий Вилкоксона

Критерий Вилкоксона основан на числе инверсий. Введем нулевую гипотезу H_0 о том, что выборка, данная в таблице 1, является однородной. Эта гипотеза отвергается, если число u (число инверсий) превосходит выбранную в соответствии с уровнем значимости границу, определяемую из того, что при объемах $n > 10$ и $m > 10$ выборок число инверсий u распределено приблизительно нормально с центром:

$$M[u] = \frac{mn}{2} = 200$$

Дисперсией

$$D[u] = \frac{mn}{12}(m + n + 1) = 1366.667$$

И средним квадратическим отклонением

$$\sigma(u) = \sqrt{D[u]}$$
$$\sigma(u) = 36.968$$

Число инверсий:

$$u = 184$$

Инверсии считались с помощью python

```
def inversion_check(data_sample: np.array, count=20) -> tuple[np.float64, np.float64,
int]:
	first_sample = data_sample[:count]
	last_sample = data_sample[-count:]


	def inv_calculate(arr1, arr2):
	arr1 = sorted(arr1)
	total = 0
	for i in range(len(arr1)):
		for j in range(len(arr2)):
		if arr1[i] > arr2[j]:
			total += 1
	return total


	inv_count = inv_calculate(first_sample, last_sample)
	M = first_sample.size * last_sample.size / 2
	D = first_sample.size * last_sample.size / 12 *
	  (first_sample.size + last_sample.size + 1)
	return np.float64(M), np.float64(D), inv_count
```

Задавшись уровнем значимости q, построим критическую область, используя соотношение

$$q = 1 - 2\Phi(t)$$

Выражения для нахождения границ критической области *H*0 :

$$u \leq M[u] - t \star \sigma(u)$$

$$u \geq M[u] + t \star \sigma(u)$$

$q = 0.01$

$t = 2.576$

$u \leq 104.776$

$u \geq 295.224$

$q = 0.05$

$t = 1.960$

$u \leq 127.543$

$u \geq 272.457$

$$q = 0.1$$

$$t = 1.645$$

$$u \leq 139.192$$

$$u \geq 260.808$$

Критерий Колмогорова-Смирнова

Критерий Колмогорова – Смирнова (критерий согласия) предназначен для сопоставления двух распределений:

• эмпирического с теоретическим, например, равномерным или нормальным

• одного эмпирического распределения с другими эмпирическим распределением

Критерий позволяет найти точку, в которой сумма накопленных расхождений между двумя распределениями является наибольшей, и оценить достоверность этого расхождения.

Введем нулевую гипотезу $H0$: различия между двумя распределениями недостоверны (судя по точке максимального накопленного расхождения между ними). Альтернативная гипотеза $H1$: различия между двумя распределениями достоверны (судя по точке максимального накопленного расхождения между ними).

Заполним таблицу. Расход – разность по модулю между накопленной частотностью эмпирического и нормального распределения.

N	Begin	End	acc($\frac{F_i}{N}$)	norm acc($\frac{F_i}{N}$)	diff
1	1403	1419.8	0.005	0.002	0.003
2	1419.8	1436.7	0.01	0.011	0.001
3	1436.7	1453.5	0.045	0.041	0.004
4	1453.5	1470.4	0.115	0.113	0.002
5	1470.4	1487.2	0.255	0.248	0.007
6	1487.2	1504.1	0.45	0.442	0.008
7	1504.1	1520.9	0.65	0.649	0.001
8	1520.9	1537.8	0.775	0.82	0.045
9	1537.8	1554.6	0.93	0.925	0.005
10	1554.6	1571.5	0.995	0.976	0.019
11	1571.5	1588.3	1.0	0.993	0.007

$$D_n = sup_x \left| F_n(x) - F(x) \right| = 0.045$$

Критические значений критерия Колмогорова-Смирнова при сопоставлении эмпирического распределения с теоретическим:

При $p = 0.05$: $D_{max} = 0.09616652$

При $p = 0.01$: $D_{max} = 0.11525841$

При $p = 0.10$: $D_{max} = 0.08626703$

Если $D_n < D_{max}$, то принимается нулевая гипотеза $H0$ на выбранном уровне значимости, иначе – принимается альтернативная гипотеза.

При всех рассмотренных уровнях значимости p нет оснований отвергнуть нулевую гипотезу о нормальном законе распределения