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

Курсовая работа

по дисциплине «Теория вероятностей и математическая статистика»

Выполнил
студент гр. в3530904/00030 В.С. Баганов

Руководитель
И.В. Зайцев

«_____» _____ 202__ г.

Санкт-Петербург
2023

Содержание

1. Теория вероятностей и математическая статистика

1.1. Программа работы

1. Написать таблицу чисел с четырьмя верными знаками размером 1x200.

2. Построить нумерованный вариационный ряд1 того же размера, что и у таблицы.

3. Произвести оценку математического ожидания, дисперсии (смещенную или несмещенную), медианы.

4. Сгруппировать значения варианты в 10 или 12 интервалов. Построить таблицу разбиения на интервалы.

5. Построить полигон, гистограмму, ступенчатую кривую.

6. По сгруппированным данным построить оценки математического ожидания, дисперсии (смещенную и несмещенную), стандарта, моды, коэффициента вариации, коэффициента асимметрии, эксцесса.

7. В предположении нормальности генеральной совокупности, из которой сделана выборка, найти оценки параметров m и σ методом квантилей.

8. Считая, что первые двадцать значений таблицы представляют собой отдельную новую выборку из нормальной генеральной совокупности, построить нумерованный вариационный ряд; найти оценку математического ожидания, дисперсии и стандарта. Построить доверительные интервалы для математического ожидания, дисперсии, стандарта.

10. С помощью критерия хи-квадрат проверить гипотезы о принадлежности выборки генеральной совокупности, распределенной по законам: нормальному, равномерному и Пуассона.

11. Проверить гипотезу об однородности выборки, используя критерий знаков и критерий Вилкоксона, по первым и последним двадцати значениям выборки. [1]

[1]Вариационный ряд (упорядоченная выборка) – совокупность значений признака, записанных в порядке их возрастания. Сам признак называется вариантой (случайной величиной).

1.1.1. Выборка.

1	**465**	41	**456**	81	**469**	121	**440**	161	**470**
2	**442**	42	**479**	82	**411**	122	**448**	162	**427**
3	**444**	43	**528**	83	**512**	123	**424**	163	**485**
4	**399**	44	**442**	84	**446**	124	**498**	164	**579**
5	**420**	45	**436**	85	**516**	125	**485**	165	**393**
6	**436**	46	**451**	86	**517**	126	**516**	166	**445**
7	**544**	47	**494**	87	**428**	127	**436**	167	**567**
8	**463**	48	**503**	88	**573**	128	**483**	168	**441**
9	**500**	49	**459**	89	**426**	129	**389**	169	**464**
10	**486**	50	**494**	90	**452**	130	**435**	170	**432**
11	**488**	51	**520**	91	**531**	131	**478**	171	**475**
12	**450**	52	**413**	92	**429**	132	**455**	172	**395**
13	**521**	53	**450**	93	**465**	133	**440**	173	**424**
14	**497**	54	**470**	94	**492**	134	**441**	174	**471**
15	**431**	55	**526**	95	**550**	135	**539**	175	**425**
16	**466**	56	**377**	96	**489**	136	**513**	176	**504**
17	**529**	57	**463**	97	**538**	137	**409**	177	**496**
18	**491**	58	**362**	98	**422**	138	**542**	178	**508**
19	**510**	59	**583**	99	**445**	139	**532**	179	**433**
20	**514**	60	**493**	100	**474**	140	**503**	180	**448**
21	**436**	61	**477**	101	**474**	141	**497**	181	**477**
22	**532**	62	**418**	102	**529**	142	**480**	182	**441**
23	**482**	63	**352**	103	**457**	143	**464**	183	**546**
24	**436**	64	**455**	104	**433**	144	**542**	184	**378**
25	**518**	65	**420**	105	**455**	145	**496**	185	**451**
26	**458**	66	**408**	106	**555**	146	**528**	186	**493**
27	**404**	67	**442**	107	**501**	147	**503**	187	**366**
28	**504**	68	**483**	108	**459**	148	**509**	188	**473**
29	**435**	69	**487**	109	**457**	149	**468**	189	**456**
30	**538**	70	**458**	110	**424**	150	**405**	190	**529**
31	**508**	71	**455**	111	**473**	151	**526**	191	**406**
32	**462**	72	**425**	112	**428**	152	**534**	192	**363**
33	**564**	73	**460**	113	**425**	153	**483**	193	**487**
34	**436**	74	**536**	114	**471**	154	**550**	194	**509**
35	**462**	75	**478**	115	**400**	155	**494**	195	**444**
36	**499**	76	**543**	116	**456**	156	**439**	196	**428**
37	**435**	77	**525**	117	**459**	157	**433**	197	**377**
38	**554**	78	**464**	118	**501**	158	**396**	198	**479**
39	**534**	79	**499**	119	**487**	159	**552**	199	**510**
40	**485**	80	**462**	120	**472**	160	**511**	200	**488**

1.1.2. Нумерованный вариационный ряд.

Вариационным рядом называется последовательность всех элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются.

1	**352**	41	**433**	81	**458**	121	**485**	161	**513**
2	**362**	42	**435**	82	**458**	122	**485**	162	**514**
3	**363**	43	**435**	83	**459**	123	**485**	163	**516**
4	**366**	44	**435**	84	**459**	124	**486**	164	**516**
5	**377**	45	**436**	85	**459**	125	**487**	165	**517**
6	**377**	46	**436**	86	**460**	126	**487**	166	**518**
7	**378**	47	**436**	87	**462**	127	**487**	167	**520**
8	**389**	48	**436**	88	**462**	128	**488**	168	**521**
9	**393**	49	**436**	89	**462**	129	**488**	169	**525**
10	**395**	50	**436**	90	**463**	130	**489**	170	**526**
11	**396**	51	**439**	91	**463**	131	**491**	171	**526**
12	**399**	52	**440**	92	**464**	132	**492**	172	**528**
13	**400**	53	**440**	93	**464**	133	**493**	173	**528**
14	**404**	54	**441**	94	**464**	134	**493**	174	**529**
15	**405**	55	**441**	95	**465**	135	**494**	175	**529**
16	**406**	56	**441**	96	**465**	136	**494**	176	**529**
17	**408**	57	**442**	97	**466**	137	**494**	177	**531**
18	**409**	58	**442**	98	**468**	138	**496**	178	**532**
19	**411**	59	**442**	99	**469**	139	**496**	179	**532**
20	**413**	60	**444**	100	**470**	140	**497**	180	**534**
21	**418**	61	**444**	101	**470**	141	**497**	181	**534**
22	**420**	62	**445**	102	**471**	142	**498**	182	**536**
23	**420**	63	**445**	103	**471**	143	**499**	183	**538**
24	**422**	64	**446**	104	**472**	144	**499**	184	**538**
25	**424**	65	**448**	105	**473**	145	**500**	185	**539**
26	**424**	66	**448**	106	**473**	146	**501**	186	**542**
27	**424**	67	**450**	107	**474**	147	**501**	187	**542**
28	**425**	68	**450**	108	**474**	148	**503**	188	**543**
29	**425**	69	**451**	109	**475**	149	**503**	189	**544**
30	**425**	70	**451**	110	**477**	150	**503**	190	**546**
31	**426**	71	**452**	111	**477**	151	**504**	191	**550**
32	**427**	72	**455**	112	**478**	152	**504**	192	**550**
33	**428**	73	**455**	113	**478**	153	**508**	193	**552**
34	**428**	74	**455**	114	**479**	154	**508**	194	**554**
35	**428**	75	**455**	115	**479**	155	**509**	195	**555**
36	**429**	76	**456**	116	**480**	156	**509**	196	**564**
37	**431**	77	**456**	117	**482**	157	**510**	197	**567**
38	**432**	78	**456**	118	**483**	158	**510**	198	**573**
39	**433**	79	**457**	119	**483**	159	**511**	199	**579**
40	**433**	80	**457**	120	**483**	160	**512**	200	**583**

1.1.3. Оценка параметров распределения по всей выборке.

Математическое ожидание (среднее арифметическое):

$$\overline{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i = \frac{1}{200} * 94324 = 471.62$$

Смещенная оценка дисперсии:

$$s^2 = \frac{1}{n} * \sum_{i=1}^{n} x_i^2 - \overline{x^2} = 2157,2655$$

Несмещенная оценка дисперсии:

$$s^{*2} = \frac{n * s^2}{n-1} = s^2 + \frac{1}{n-1} * s^2 = 2168,1061$$

Среднее квадратическое отклонение (стандартное отклонение):

$$s = \sqrt{s^2} = 46.56$$

Оценка медианы*:
a) Если число членов вариационного ряда нечетное (n=2k+1):

$$\widetilde{M_e} = X_{k+1}$$

где

$$X_{k+1}$$

(k+1)-й член вариационного ряда.
b) При четном числе членов (n=2k) в качестве медианы принимают

$$\widetilde{M_e} = \frac{X_k + X_{k+!}}{2} = 470.0$$

[2]

Мат. ожидание: 469.42
Смещенная дисперсия: 3048.3536000000013 3048.353599999973
СКО: 55.21189726861414
Квадрат несмещенной дисперсии: 3063.671959798996
Смещенная дисперсия: 55.35044678951558
Медиана: 469.5

[2]Медиана – значение варианты, которое делит вариационный ряд на две равные по числу членов части.

1.1.4. Группирование данных

При большом объеме выборки для удобства вычислений прибегают к группированию данных в интервалы. Ширина интервала (шаг разбиения) вычисляется по формуле:

$$\Delta X = \frac{R}{l} = 23$$

где R – размах варьирования (широта распределения), т.е. разность между наибольшим и наименьшим значением варианты, а l – число интервалов.

$$R = X_{max} - X_{min} = 583 - 352 = 231$$

Предполагая нормальное распределение, число интервалов l = 11

№ интервала	Границы интервалов	Частота в интервале	Частототность в интревале	Середина интервала
1	340-363	3	0,015	351,55
2	363-386	4	0,02	374,65
3	386-409	11	0,055	397,75
4	409-432	20	0,1	420,85
5	432-455	37	0,185	443,95
6	455-478	38	0,19	467,05
7	478-501	34	0,17	490,15
8	501-524	21	0,105	513,25
9	524-547	22	0,11	536,35
10	547-571	7	0,035	559,45
11	571-594	3	0,015	582,55

1.1.5. Графики статистических распределений

Для наглядности распределение выборочных данных изображают графически несколькими способами.

Полигон: на оси абсцисс откладываются интервалы значений величины x, в серединах интервалов строятся ординаты, пропорциональные частотам (или частостям), и концы ординат соединяются отрезками прямых линий.

Гистограмма: над каждым отрезком оси абсцисс, изображающим интервал значений x, строится прямоугольник, площадь которого пропорциональна частоте (или частости) в данном интервале.

Ступенчатая кривая: над каждым отрезком оси абсцисс, изображающим расстояние между серединами интервалов значений x, проводится отрезок горизонтальной прямой на высоте, пропорциональной накопленной частоте (или накопленной частости) в данном интервале3. Концы отрезков соединяются.

[3]

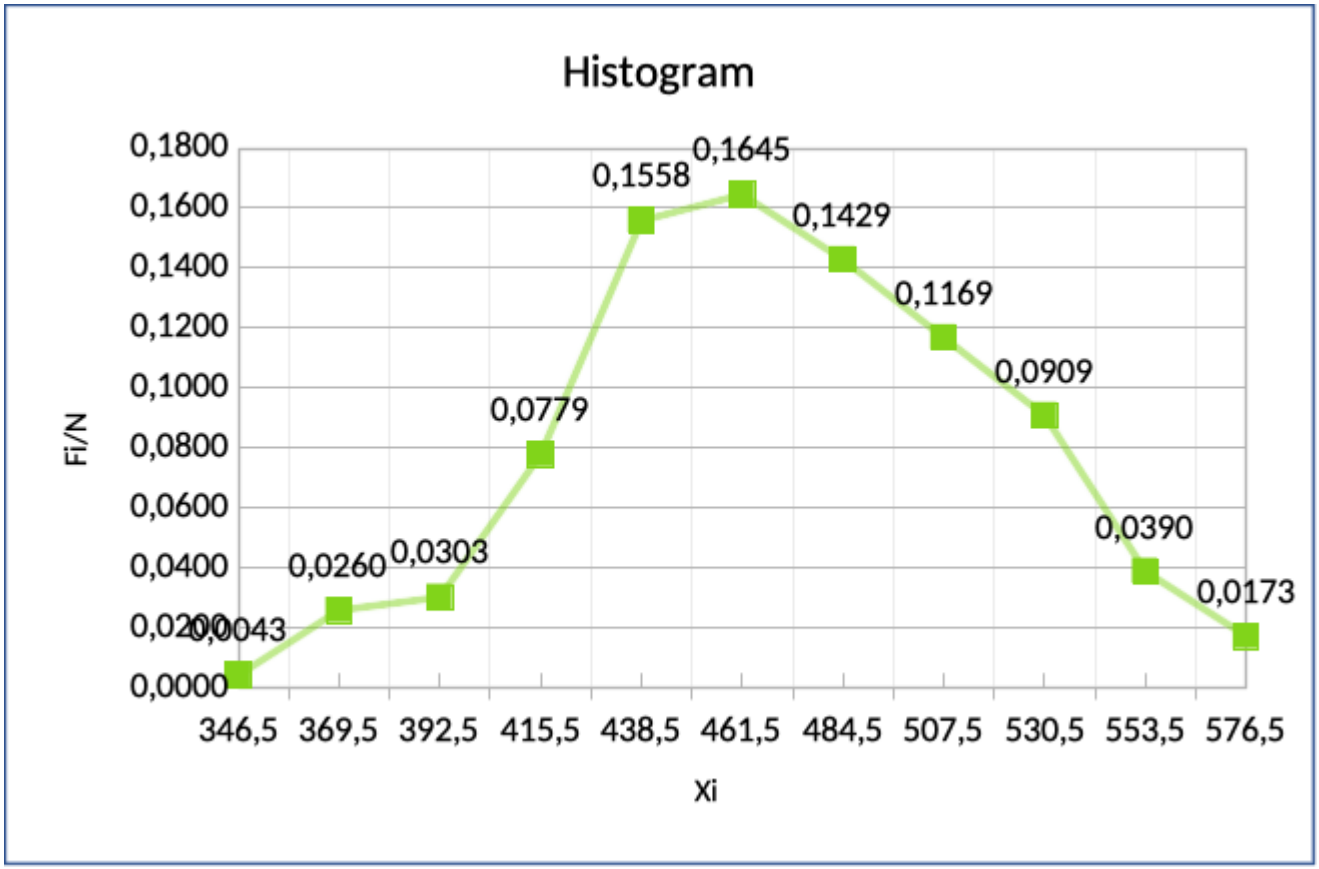

Рисунок 1.1. Полигон

[3]Накопленная частота в данном интервале – сумма всех частот, начиная с первого интервала до данного интервала включительно.

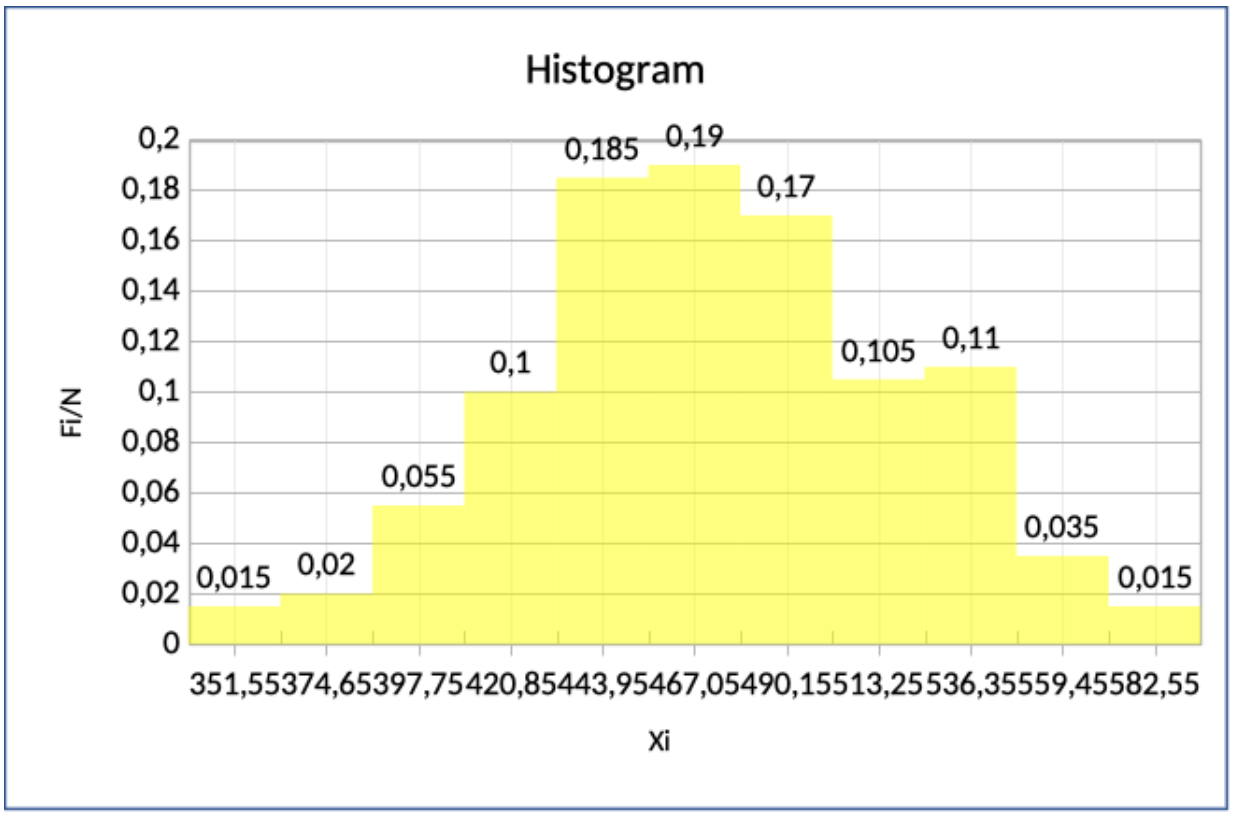

Рисунок 1.2. Гистограмма

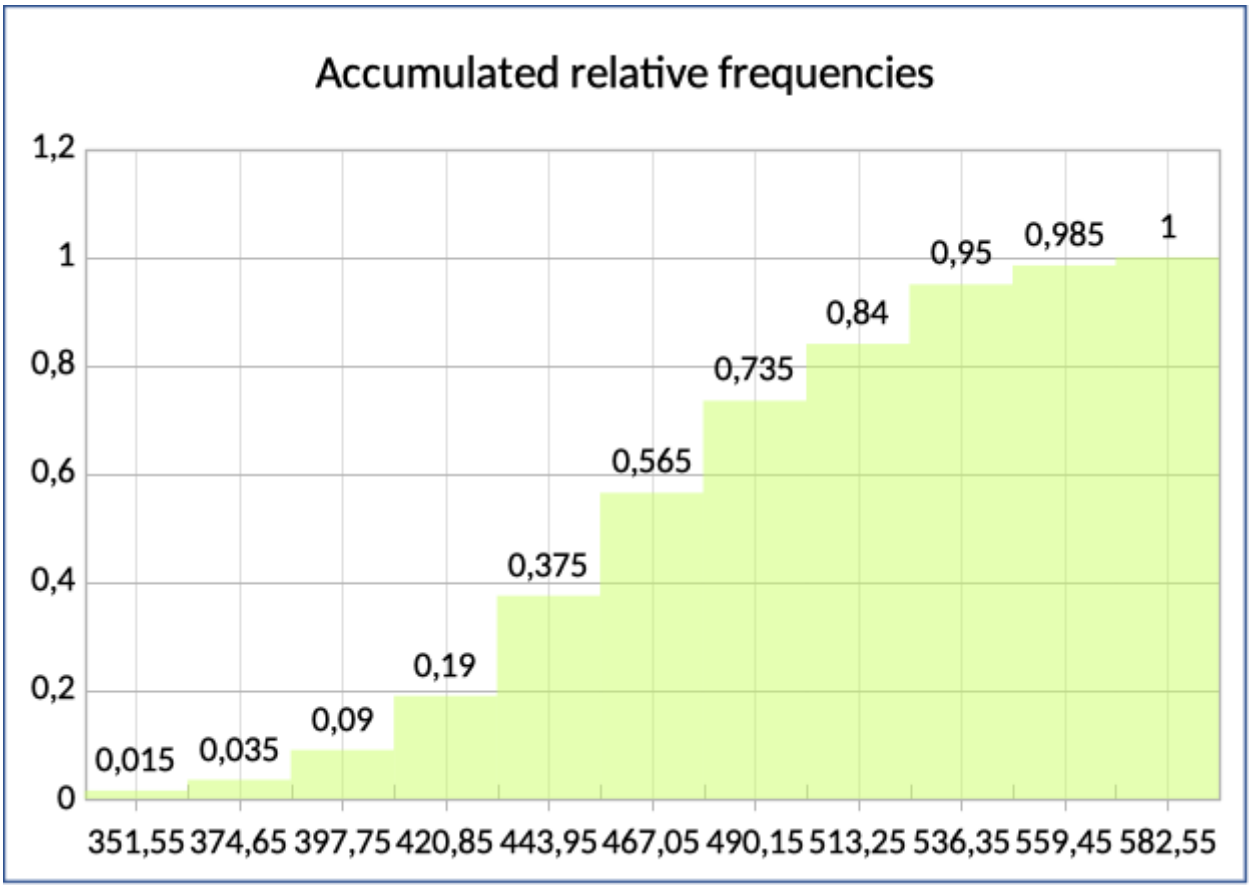

Рисунок 1.3. Ступенчатая кривая

1.1.6. Точечная оценка характеристик распределения по сгруппированным данным.

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

Меры центральной тенденции:

Математическое ожидание (среднее арифметическое):

$$\overline{x} = \frac{1}{n} * \sum_{i=1}^{11} x_i * \nu_i = 471.62$$

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

Мода: Самые многочисленные интервалы – это интервалы №5 и №6 с границами 432-455 и 455-478 соответственно.

Моды сгруппированной выборки – их середины: 443,95 и 467,05

$$\widetilde{M_0} = 467,05$$

равна середине самого многочисленного интервала

Меры изменчивости:

Смещенная оценка дисперсии:

$$\overline{x} = \frac{1}{n} * \sum_{i=1}^{11} x_i * \nu_i - \overline{x^2} =$$

Несмещенная оценка дисперсии:

$$s^{*2} = s^2 - \frac{(\Delta x)^2}{12}$$

Среднее квадратическое отклонение (стандартное отклонение):

$$s = \sqrt{s^2} = 46.56$$

Коэффициент вариации:

$$\upsilon = \frac{S}{\overline{x}} = 0,01, (\overline{x} \neq 0)$$

Коэффициент асимметрии:

$$\widetilde{S_k} = \frac{m_3}{s^3} = -0,0129$$

Эксцесс:

$$\widetilde{E_x} = \frac{m_4}{s^4} - 3 = -0,3208$$

Для нормального распределения Sk = 0 и Ex = 0 Поэтому показатели асимметрии и эксцесса, отличные от нуля, указывают на отклонение рассматриваемого распределения от нормального.

Выборочные асимметрия и эксцесс, как и все оценки, являются случайными величинами и могут не совпадать с теоретическими. Критические значения коэффициента асимметрии и коэффициента эксцесса (объем выборки n = 200, уровень значимости α = 1˜k(crit ical value) = 0.403 ˜x(crit ical value) = 0.832 ˜k < ˜k(crit ical value) и ˜x < ˜x(crit ical value), т.е. вычисленные значения коэффициента асимметрии и коэффициента эксцесса не попадают в критические области, следовательно гипотезу о том, что выборочные данные распределены нормально, нельзя считать ошибочной.

1.1.7. Оценка параметров распределения методом квантилей.

Квантилем, отвечающим заданному уровню вероятности p, называют такое значение варианты x = x_p при котором функция распределения принимает вид:

$$F(X_p) = p$$

Некоторые квантили получили особые названия. Например, медианой распределения $\widetilde{M_e}$ называют квантиль, отвечающий значению p = 0.5; квантили, соответствующие значениям p = 0.25 и p = 0.75, называют нижним и верхним квартилями.

Функция нормального распределения:

$$F(X_p) = \frac{(x_p - m)}{\delta} + 0.5 = p$$

где Φ(x) – функция Лапласа.

Чтобы определить два неизвестных параметра закона нормального распределения – математическое ожидание (m) и среднее квадратическое отклонение (δ), необходимо составить два уравнения, используя формулу, приведенную выше. Для этого из вариационного ряда возьмем два произвольных значения варианты вблизи медианы – x_{p1} и x_{p2} с соответствующими им вероятностями p_1 и p_2. Вероятности p_1 и p_2 равны порядковому номеру выбранных вариант, деленному на 200 (т.к. число элементов выборки n = 200).

Решение:

Пусть x_{p1} = 462, p_1 = 80/200 = 0.4, x_{p2} = 472, p_2 = 120/200 = 0.6. Подставляя эти значения в формулу функции нормального распределения, получаем уравнения5:

$$\begin{cases} F(X_{p1}) = \frac{(x_p - m)}{\delta} + 0.5 = 0.4, \\ \\ F(X_{p2}) = \frac{(x_p - m)}{\delta} + 0.5 = 0.6, \end{cases}$$

[4] [5]

Упорщая:

$$\begin{cases} F(X_{p1}) = \frac{(x_p - m)}{\delta} = -0.1, \\ \\ F(X_{p2}) = \frac{(x_p - m)}{\delta} = 0.1, \end{cases}$$

Функция Лапласа – нечетная, поэтому −Φ(x) = Φ(−x). Тогда

$$\begin{cases} F(X_{p1}) = \frac{(x_p - m)}{\delta} = 0.1, \\ \\ F(X_{p2}) = \frac{(x_p - m)}{\delta} = 0.1, \end{cases}$$

Воспользуемся таблицей значений функции Лапласа.

Вероятности p попадания случайной величины в интервалы (x_{p1}, M_e) и (M_e, x_{p2},) одинаковые и составляют по 0.1. Тогда согласно таблице, если Φ(x) = 0.1, то x ≈ 0.255.

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

[5] Предпочтительно брать квантиль даже более близкие к медиане (но не саму медиану).

Преобразовываем уравнения к следующему виду:

$$\begin{cases} F(X_{p1}) = \frac{(x_p - m)}{\delta} = 0.1, \\ \\ F(X_{p2}) = \frac{(x_p - m)}{\delta} = 0.1, \end{cases}$$

Затем вычитаем одно уравнение из другого:

$$\begin{cases} F(X_{p1}) = \frac{(462 - m)}{\delta} = 0.255, \\ \\ F(X_{p2}) = \frac{(472 - m)}{\delta} = 0.255, \end{cases}$$

Решая систему уравнений, находим, что δ = 45.4 , m = 470.4

Таким образом, полученные значения параметров m ˜и δ ˜(с учетом небольшой погрешности) соответствуют вычисленным ранее (в пункте 3) значениям математического ожидания (x = 471.62) и стандартного отклонения (s = 46.5) .

1.1.8. Построение доверительных интервалов.

Выборка из первых 20 значений таблицы и Нумерованный вариационный ряд:

No xi изм.	xi
1	465
2	442
3	444
4	399
5	420
6	436
7	544
8	463
9	500
10	486
11	488
12	450
13	521
14	497
15	431
16	466
17	529
18	491
19	510
20	514

No xi изм.	xi	xi^2
7	544	295936
17	529	279841
13	521	271441
20	514	264196
19	510	260100
9	500	250000
14	497	247009
18	491	241081
11	488	238144
10	486	236196
16	466	217156
1	465	216225
8	463	214369
12	450	202500
3	444	197136
2	442	195364
6	436	190096
15	431	185761
5	420	176400
4	399	159201

Математическое ожидание (среднее арифметическое):

$$\overline{x} = \frac{1}{n} * \sum_{i=1}^{n} x_i = \frac{1}{20} * 9496 = 474.8$$

Смещенная оценка дисперсии:

$$s^2 = \frac{1}{n} * \sum_{i=1}^{n} x_i^2 - \overline{x^2} = 2157,2655$$

Несмещенная оценка дисперсии:

$$s^{*2} = \frac{n * s^2}{n-1} = s^2 + \frac{1}{n-1} * s^2 = 2168,1061$$

Среднее квадратическое отклонение (стандартное отклонение):

$$s = \sqrt{s^2} = 37,34$$

Доверительный интервал для истинного математического ожидания mx6:

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

где двустороннее обратное t-распределение Стьюдента вычисляется с помощью LibreOffice*.

*На вход функции T.INV.T(q, df) подается два параметра: уровень значимости (вероятность ошибки) q, а также число степеней свободы df = n - 1 = 19.

Доверительный интервал для истинной дисперсии

$$s_x{}^2$$

и истинного среднего квадратического отклонения (стандартного отклонения)

$$s_x :$$

где пределы χ_1^{22}
вычисляются средствами LibreOffice*.
*На вход функции CHISQ.INV(p, df) подается два параметра:
вероятности

$$p_1 = 1 - \frac{1}{2} * \frac{q}{100}$$

для χ_1^2

$$p_2 = \frac{1}{2} * \frac{q}{100}$$

для χ_2^2
(где q – уровень значимости), а также число степеней свободы df = n - 1 = 19.
Определим доверительные интервалы для математического ожидания mx, задаваясь различными уровнями значимости q. Значения квантилей
$\overline{x} + t_{q,n-1}$
берутся из таблиц распределения Стьюдента по двум входам: по числу степеней свободы (n-1) и уровнями значимости q. Уровень значимости q здесь и в дальнейшем предполагается заданным.
Определим доверительные интервалы для математического ожидания mx, задаваясь различными уровнями значимости q. Значения квантилей берутся из таблиц распределения Стьюдента по двум входам: по числу степеней свободы (n-1) и уровнями значимости q. Уровень значимости q здесь и в дальнейшем предполагается заданным.
Для q=5%:
t= 2,093

$$394,125 < m_x < 597,775$$

Для q=10%:
t= 1,7295

$$394,125 < m_x < 597,775$$

Для q=1%:
t= 2,860935

$$394,125 < m_x < 597,775$$

Уровень значимости q здесь примем за 1% и получаем такие доверительные интервалы:
Для q=1%:

$$7,97 < \sigma^2 < 37,83$$

$$2,82 < \sigma < 6,15$$

Для q=10%:

$$10,61 < \sigma^2 < 24,78$$

$$3,26 < \sigma < 4,98$$

Для q=5%:

$$9,58 < \sigma^2 < 28,54$$

$$3,10 < \sigma < 5,34$$

Таким образом, определенны доверительные интервалы для параметров

$$\sigma^2 \sigma.$$

Т.к. интервалы (297.010, 884.937) и (17.234, 29.748) охватывают истинные значения дисперсии sx2 = 410.204 и стандартного отклонения sx = 20.253 соответственно, получившиеся неравенства являются верными.

1.1.9. Критерий хи-квадрат для проверки статистических гипотез.

Примем гипотезу о том, что выборка из 200 элементов подчиняется нормальному закону распределения. Для проверки этой гипотезы используется критерий χ2:

$$x^2 = \sum_{i=1}^{n} \frac{(v_i - n * \widetilde{p}_l)^2}{n * \widetilde{p}_l} = 474.8$$

Если полученное значение χ2 попадёт в область допустимых значений

$\chi 2 < \chi 2_{(criticalvalue)}$

то данные выборки не противоречат гипотезе о нормальности распределения. В противном случае гипотеза отвергается.

Если же численное значение критерия χ2 попадет в критическую область χ2 > χ2(critical value), то гипотеза отвергается.

Вычисления критерия удобно свести в табл.

Сумма, называемая критерием x2, асимптотически распределена как хи- квадрат. При практических расчетах для нахождения критического значения этой суммы можно пользоваться таблицами распределения хи-квадрат только в том случае, если для всех интервалов

Поэтому в табл. 7 интервалы из табл.6 с номерами 1, 2, 3 и 10, 11 объединены. Оценку вероятности :

$$\widetilde{P}_i = P(\alpha < X < \beta) = \phi(\frac{\beta - \overline{x}}{s}) - \phi(\frac{\beta - \overline{x}}{s})$$

где α и β – нижняя и верхняя границы интервалов, x и s вычислены по сгруппированным данным выборки (в пункте 6), а значения функции Лапласа находятся с помощью LibreOffice*.

Так как по данным выборки мы оценили два параметра mx и σx нормального закона (т.е. c = 2), то в нашем случае число степеней свободы будет равно: k = l' - c - 1 = 8 - 2 – 1 = 5, где l' = 8 – число интервалов, получившихся после объединения интервалов.

p ĩ (теоретическая частотность) – вероятность попадания в интервал при действительно нормальном распределении.

*На вход функции NORM.S.DIST(аргумент функции Лапласа, 1) подается два параметра: a также число 1, чтобы NORM.S.DIST вычисляла ss интегральную функцию Лапласа.

Следует отметить, что критерий χ2 нормально работает, когда каждый интервал содержит не менее 5 точек. Поэтому интервалы, где νi < 5, необходимо объединить с соседними интервалами.

χ2(critical value) также вычисляется средствами LibreOffice*.

*На вход функции CHISQ.INV.RT(q, df) подается два параметра: уровень значимости (вероятность ошибки) q, а также число степеней свободы df = l - 3 (или = l' - 3 в случае объединения нескольких интервалов).

Поскольку в начале и в конце таблицы имеются интервалы с частотой (νi) < 5, объединим их с соседними интервалами прежде чем начнем вычислять взвешенные квадраты уравнений

№ Интервала	Начало	Конец	Середина	Fi частота в интревале	Частотность в интервале Fi / N	acc Fi / N	Pi теор частоность	Fi теор частота	F теор (x)	Fт(x) - Fэ(x)
1	340	363,1	351,55	3	0,015	0,015	0,0075	1,5	0,0075	0,0075
2	363,1	386,2	374,65	4	0,02	0,035	0,0234	4,7	0,0309	0,0041
3	386,2	409,3	397,75	11	0,055	0,09	0,0571	11,4	0,088	0,002
4	409,3	432,4	420,85	20	0,1	0,19	0,1094	21,9	0,1975	0,0075
5	432,4	455,5	443,95	37	0,185	0,375	0,1648	33	0,3622	0,0128
6	455,5	478,6	467,05	38	0,19	0,565	0,195	39	0,5572	0,0078
7	478,6	501,7	490,15	34	0,17	0,735	0,1813	36,3	0,7385	0,0035
8	501,7	524,8	513,25	21	0,105	0,84	0,1324	26,5	0,8709	0,0309
9	524,8	547,9	536,35	22	0,11	0,95	0,076	15,2	0,947	0,003
10	547,9	571	559,45	7	0,035	0,985	0,0343	6,9	0,9812	0,0038
11	571	594,1	582,55	3	0,015	1	0,0121	2,4	0,9934	0,0066
			Sum:	200	1					**0,0309**

№ Интервала	Начало	Конец	Середина	Fi частота в интревале	Частотность в интервале Fi / N	acc Fi / N	Pi теор частоность	Fi теор частота	F теор (x)	Fт(x) - Fэ(x)
1										
2										
3	340	409	374,5	4	0,02	0,02	0,087	17,4	0,087	0,067
4	409,3	432,4	420,85	19	0,095	0,115	0,1094	21,9	0,1964	0,0814
5	432,4	455,5	443,95	36	0,18	0,295	0,1648	33	0,3612	0,0662
6	455,5	478,6	467,05	38	0,19	0,485	0,195	39	0,5562	0,0712
7	478,6	501,7	490,15	33	0,165	0,65	0,1813	36,3	0,7375	0,0875
8	501,7	524,8	513,25	31	0,155	0,805	0,1324	26,5	0,8699	0,0649
9	524,8	547,9	536,35	21	0,105	0,91	0,076	15,2	0,9459	0,0359
10	547,9	594	570,95	18	0,09	1	0,0464	9,3	0,9923	0,0077
11						1	0	0	0,9923	0,0077
			Sum:	200	1					**0,0875**

Вычислим сумму значений взвешенных квадратных отклонений:

$$\sum_{i=1}^{8} \frac{(v_i - n * \widetilde{p_l})^2}{n * \widetilde{p_l}} = 3,71257$$

Таким образом, $\chi 2 = 3{,}71257$

По таблице распределения хи-квадрат найдем значения $\chi 2$ для числа степеней свободы $k = 5$ и уровней значимости q, равным 1%, 5% и 10%.

Для $q = 1\%$:

$\chi 2 = 15{,}0862$

Поскольку $3{,}71257 < 15{,}0862$, то гипотеза о нормальности выборки не противоречит данным измерений.

Для $q = 5\%$:

$\chi 2 = 11{,}0705$

Поскольку $3{,}71257 < 11{,}0705$, то гипотеза о нормальности выборки не противоречит данным измерений.

Для $q = 10\%$:

$\chi 2 = 9{,}2363$

Поскольку $3{,}71257 < 9{,}2363$, то гипотеза о нормальности выборки не противоречит данным измерений.

1.1.10. Проверка гипотезы об однородности выборки с помощью критериев знаков и Вилкоксона.

Введем нулевую гипотезу H0 о том, что выборка из 200 элементов является однородной. Для проверки этой гипотезы возьмем двадцать первых и двадцать последних значений выборки.

Воспользуемся непараметрическими (независимыми от формы распределения) критериям: критерием знаков и критерием Вилкоксона.

Критерий знаков

Вычислим знаки разностей zi = xi – yi, где i = 1, 2, ..., 20 – порядковые номера первых xi и последних yi двадцати значений выборки, используя средства LibreOffice*.

*Функция SIGN(xi – yi).

Затем посчитаем количество положительных kn(+) и отрицательных kn(–) знаков разностей zi (n = 20) с помощью функции LibreOffice*. *На вход функции COUNTIF(range, sign) подается два параметра: весь диапазон значений zi , а также положительное число 1 (для вычисления kn(+)) или отрицательное число -1 (для вычисления kn(–)).

Зная q и n, находим в таблице “Критерий знаков” критическое значение mn для минимального числа из kn(+) и kn(–). Если меньшее из чисел знаков разностей окажется меньше mn , то гипотеза об однородности выборки отвергается, а если меньшее из чисел знаков разностей окажется больше mn, то следует признать, что гипотеза не противоречит данным выборки.

Решение:

№	Выборка No1 (первые 20 элементов)	Выборка No2 (последние 20 элементов)	zi = xi – yi
1	465	448	1
2	442	477	-1
3	444	441	1
4	399	546	-1
5	420	378	1
6	436	451	-1
7	544	493	1
8	463	366	1
9	500	473	1
10	486	456	1
11	488	529	-1
12	450	406	1
13	521	363	1
14	497	487	1
15	431	509	-1
16	466	444	1
17	529	428	1
18	491	377	1
19	510	479	1
20	514	510	1

Подсчет показывает, что положительных знаков
k20(+)=15 и отрицательных знаков k20(–)=5
Пусть уровень значимости
q = 1% (0.01). по таблице m20 = 4
q = 5% (0.05). по таблице m20 = 5
q = 10% (0.1). по таблице m20 = 5

Тогда согласно таблице “Критерий знаков” критическое значение m20 = 4. Поскольку сравнивайся минимальное из чисел k20(+) и k20(–), а именно 5, получаем 5 > 4 (т.е. k20(+) > m20). Таким образом, можно заключить, что гипотеза H0 об однородности выборки не противоречит данным выборки. Нет оснований чтобы утверждать нулевую гипотезу.

Критерий Вилкоксона

Данный критерий основан на числе инверсий. Элементы двух выборок располагаются в общую последовательность в порядке возрастания их значений.

где x1, x2, ..., x5 – элементы первой выборки, а y1, y2, ..., y5 – элементы второй выборки.

Если какому-либо значению x предшествует некоторый y, значит эта пара дает инверсию. Так, в приведенной выше последовательности x1 имеет одну инверсию с y1,ax2,x3, x4 даютпотриинверсии(cy1, y2, y3)ит.д.Общее число инверсий u=1+ 3+ 3 + 3 + 6 = 16.

Гипотеза H0 об однородности выборки из 200 элементов отвергается, если число u выходит за пределы вычисленных в соответствии с уровнем значимости q границ. Критическая область определяется исходя из того, что при объемах n > 10 и m > 10 выборок число инверсий u имеет нормальное распределение с центром

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

Дисперсией:

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

И средним квадратическим отклонением:

$$\sigma[u] = \sqrt{D[u]}$$

Чтобы найти нижнюю и верхнюю границы критической области для числа u, воспользуемся возможностями LibreOffice*: с помощью обратной функции Лапласа вычислим расстояние (выраженное в количестве стандартных отклонений σ) от среднего до каждой из двух границ.

*На вход функции NORM.S.DIST(number) подается следующий параметр: 1 q для σ1 (нижняя граница) и 1 − 1 q для σ2 (верхняя граница).

Затем преобразуем стандартные отклонения σ1 и σ2 в конкретные значения:

$$u_{lower} = M[u] + \sigma_1 * \sigma[u]$$
$$u_{upper} = M[u] + \sigma_2 * \sigma[u]$$

Если верно неравенство

$$u_{lower} < u < u_{upper}$$

, то оснований для отклонения гипотезы H0 нет, поскольку вероятность того, что данное значение u было получено не случайно и выборка из 200 элементов действительно является однородной, составляет 100 – q

Решение:

Номер выборки	Значение	Количество инверсий
2	363	0
2	366	0
2	377	0
2	378	0
1	399	4
2	406	0
1	420	5
2	428	0
1	431	6
1	436	6
2	441	0
1	442	7
1	444	7
2	444	0
2	448	0
1	450	9
2	451	0
2	456	0
1	463	11
1	465	11
1	466	11
2	473	0
2	477	0
2	479	0
1	486	14
2	487	0
1	488	15
1	491	15
2	493	0
1	497	16
1	500	16
2	509	0
1	510	17
2	510	0
1	514	18
1	521	18
1	529	18
2	529	0
1	544	19
2	546	0
	сумма	**243**

Объединим элементы выборок в одну последовательность в порядке возрастания их значений и поставим напротив всех элементов из первой выборки число “1”, а напротив всех элементов из второй – число “2”.

Затем посчитаем число инверсий ui для каждого элемента с помощью LibreOffice* и вычислим общее значение u как сумму всех ui.

*В каждой ячейке задается условие =IF(Ai=2, 0, COUNTIF($A1$:Ai, 2)), где A1 – номер ячейки первого элемента последовательности, а i – индекс текущего элемента.

Получаем:

Найдем M[u], D[u] и σ[u] , при условии m = n = 20:

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

Дисперсией:

$$D[u] = \frac{mn}{12}(m + n + 1) = 1366,667$$

И средним квадратическим отклонением:

$$\sigma[u] = \sqrt{D[u]} = 36,968$$

Построим критическую область, задавшись уровнем значимости q = 1%:

$$\sigma_1 = NORM.S.DIST(0.005) = -2.5758$$

$$\sigma_2 = NORM.S.DIST(0.095) = 2.5758$$

Границы критической области:

$$u_{lower} = M[u] + \sigma * \sigma[u] = 200 - 36,968 * 2.5758 = 104.776$$

$$u_{upper} = M[u] + \sigma * \sigma[u] = 200 + 36,968 * 2.5758 = 295.224$$

Теперь определим границы критической области для уровня значимости q = 5%:

$$\sigma_1 = NORM.S.DIST(0.05) = -1.96$$

$$\sigma_2 = NORM.S.DIST(0.95) = 1.96$$

Границы критической области:

$$u_{lower} = M[u] + \sigma * \sigma[u] = 200 - 36,968 * 1.96 = 127.54$$

$$u_{upper} = M[u] + \sigma * \sigma[u] = 200 + 36,968 * 1.96 = 275.45$$

Теперь определим границы критической области для уровня значимости q = 10%:

$$\sigma_1 = NORM.S.DIST(0.05) = -1.645$$

$$\sigma_2 = NORM.S.DIST(0.95) = 1.645$$

Границы критической области:

$$u_{lower} = M[u] + \sigma * \sigma[u] = 200 - 36,968 * 1.645 = 139.192$$

$$u_{upper} = M[u] + \sigma * \sigma[u] = 200 + 36,968 * 1.645 = 260.808$$

Поскольку в обоих случаях неравенство

$$u_{lower} < u < u_{upper}$$

верно (т.к.

$$104.776 < 243 < 295.224$$

$$127.54 < 243 < 275.45$$

$$139.192 < 243 < 260.808$$

), гипотеза H0 об однородности выборки не противоречит полученным данным.

И оснований для отклонения гипотезы H0 нет, нулевая гипотеза верна, и выборка из 200 элементов действительно является однородной.

При всех рассмотренных уровнях значимости число инверсий u не лежит в критической области, а потому нулевая гипотеза H0 об однородности выборки не противоречит данным выборки.