ДОМАШНЕЕ ЗАДАНИЕ
по МАТЕМАТИЧЕСКОЙ СТАТИСТИКЕ

Исходные данные

Задана большая выборка, объем которой n =100.

2.49	3.548	4.409	5.028	0.3911	2.096	5.242	4.586	4.149	2.678
4.208	3.993	4.32	6.29	-2.482	5.118	5.107	3.889	2.113	5.59
9.377	2.644	6.819	3.294	6.091	8.041	2.577	7.486	-1.553	2.246
6.054	0.7189	6.614	7.823	2.331	5.322	2.033	6.87	7.682	0.6396
8.007	4.824	6.995	9.007	1.359	7.471	1.455	9.796	1.051	3.364
5.393	8.342	3.71	5.33	0.8848	0.7033	1.888	4.802	1.994	8.223
3.873	7.276	8.192	1.896	8.903	3.658	5.064	4.829	8.793	3.079
3.248	-0.5263	2.747	6.493	4.397	4.705	8.667	4.91	7.09	-1.592
0.7742	2.11	4.148	4.936	2.857	0.09494	4.508	5.815	2.85	5.311
10.35	13.49	4.87	3.424	5.508	5.407	8.291	0.7415	6.166	5.146

Две малые выборки:

1) 2.49 3.548 4.409 5.028 0.3911 2.096 5.242 4.586 4.149 2.678
2) 4.208 3.993 4.32 6.29 -2.482 5.118 5.107 3.889 2.113 5.59

Первичная обработка статистической информации

Для малых выборок найдем точечные оценки

1. Выборочное среднее $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

2. Выборочная дисперсия $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}$

Исправленная выборочная дисперсия является состоятельной и несмещенной оценкой генеральной дисперсии и вычисляется по формуле $s^2 = S^2 \frac{n}{n-1}$.

3. Исправленное выборочное среднее квадратическое отклонение $s = \sqrt{s^2}$.

Для первой выборки получаем:

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} = \frac{1}{10}34,6171 \approx 3,4617;$$

$$S_1^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{1i} - \bar{x}_1)^2 = \frac{1}{10}21,1161 \approx 2,1116;$$

$$s_1^2 = S_1^2 \frac{n}{n-1} = \frac{10}{9}2,1116 \approx 2,3462,$$

$$s_1 = \sqrt{2,3462} \approx 1,532.$$

Для второй выборки получаем:

$$\overline{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i} = \frac{1}{10}38,146 \approx 3,8146;$$

$$S_2{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_{2i} - \overline{x}_2\right)^2 = \frac{1}{10}55,639 \approx 5,5639;$$

$$s_2{}^2 = S_2{}^2\frac{n}{n-1} = \frac{10}{9}5,5639 \approx 6,1821,$$

$$s_2 = \sqrt{6,1821} \approx 2,486.$$

Расчеты в таблице:

											Сумма
x_{1i}	2,49	3,548	4,409	5,028	0,3911	2,096	5,242	4,586	4,149	2,678	**34,6171**
x_{2i}	4,208	3,993	4,32	6,29	-2,482	5,118	5,107	3,889	2,113	5,59	**38,1460**
$\left(x_{1i} - \overline{x}_1\right)^2$	0,94422	0,00745	0,89736	2,45326	9,42865	1,86516	3,16943	1,26403	0,47237	0,6142	**21,1161**
$\left(x_{2i} - \overline{x}_2\right)^2$	0,15476	0,03183	0,25543	6,12761	39,6472	1,69885	1,6703	0,00554	2,89544	3,15205	**55,6390**

Для большой выборки сначала составим группированный статистический ряд.

- Найдем крайние элементы выборки: $x_{\min} = -2,482$ и $x_{\max} = 13,49$
- Разобьем полученный промежуток на k равных интервалов, вычислив k по формуле Стерджесса $k = 1 + 3,3\lg n$. Для n =100 получится k =8.
- Найдем длину каждого интервала $h = \frac{x_{\max} - x_{\min}}{k} = \frac{15,972}{8} = 1,9965$. Границы интервалов определим по формуле $a_i = x_{\min} + i \cdot h$.
- Подсчитаем интервальные частоты: n_i - число элементов выборки, попавших в интервал $\Delta_i = [a_{i-1}; a_i]$. Элемент выборки, находящийся на границе интервалов, будем относить к правому интервалу. $\sum_{i=1}^{k} n_i = n$.
- Значения всех элементов выборки, попавших в интервал $\Delta_i = [a_{i-1}; a_i]$, будем считать равными координате середины интервала $x_i^* = \frac{a_{i-1} - a_i}{2}$.

В таблице приведем результаты первичной обработки статистических данных.

Номер интервала i	a_{i-1}	a_i	Частоты n_i	Относительные частоты n_i/n	Приведенные частоты n_i/nh	Середина интервала x_i^*	Ординаты точек кривой Гаусса $f\left(x_i^*\right)$
1	-2,482	-0,4855	4	0,040	0,020	-1,48	0,016

2	-0,4855	1,511	11	0,110	0,055	0,513	0,054
3	1,511	3,5075	21	0,210	0,105	2,509	0,111
4	3,5075	5,504	32	0,320	0,160	4,506	0,14
5	5,504	7,5005	16	0,160	0,080	6,502	0,108
6	7,5005	9,497	13	0,130	0,065	8,499	0,051
7	9,497	11,494	2	0,020	0,010	10,5	0,015
8	11,494	13,49	1	0,010	0,005	12,49	0,003

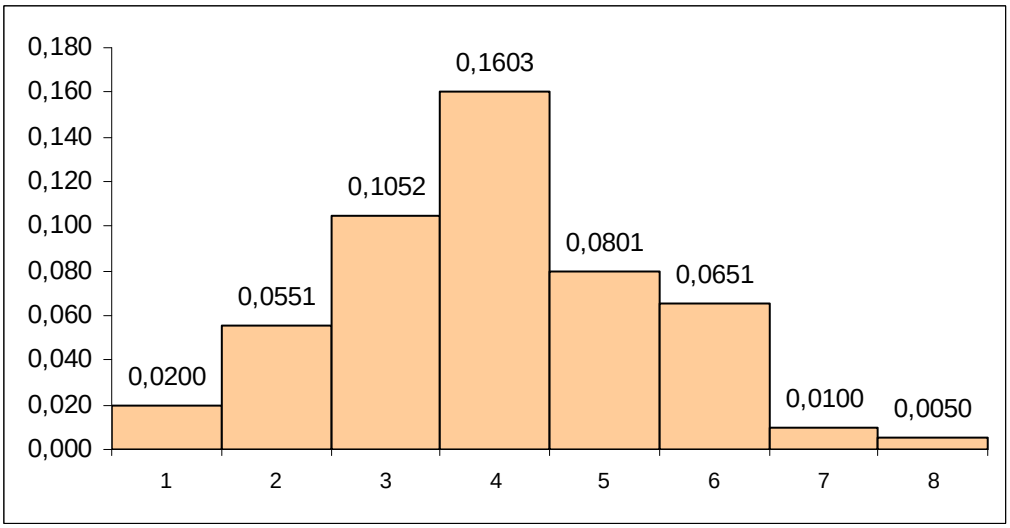
Далее вычислим:

- Выборочное среднее: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} n_i x_i^* = \frac{1}{100} 444,586 = 4,446$
- Исправленное выборочное среднее квадратическое отклонение:
$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} n_i \left(x_i^* - \overline{x}\right)^2} = \sqrt{\frac{1}{99} 808,8} \approx \sqrt{8,17} \approx 2,858$

Расчеты в таблице:

x_i^*	n_i	$n_i x_i^*$	$n_i \left(x_i^* - \overline{x}\right)^2$
-1,4838	4	-5,935	140,6409
0,5128	11	5,64025	170,1625
2,5093	21	52,6943	78,75922
4,5058	32	144,184	0,114797
6,5023	16	104,036	67,66017
8,4988	13	110,484	213,5375
10,495	2	20,9905	73,19036
12,492	1	12,4918	64,73643
>72,4			
Сумма	**100**	**444,586**	**808,80**

Построим гистограмму приведенных частот $n_i / nh = n_i / (100 \cdot 1,9965) = n_i / 199,65$

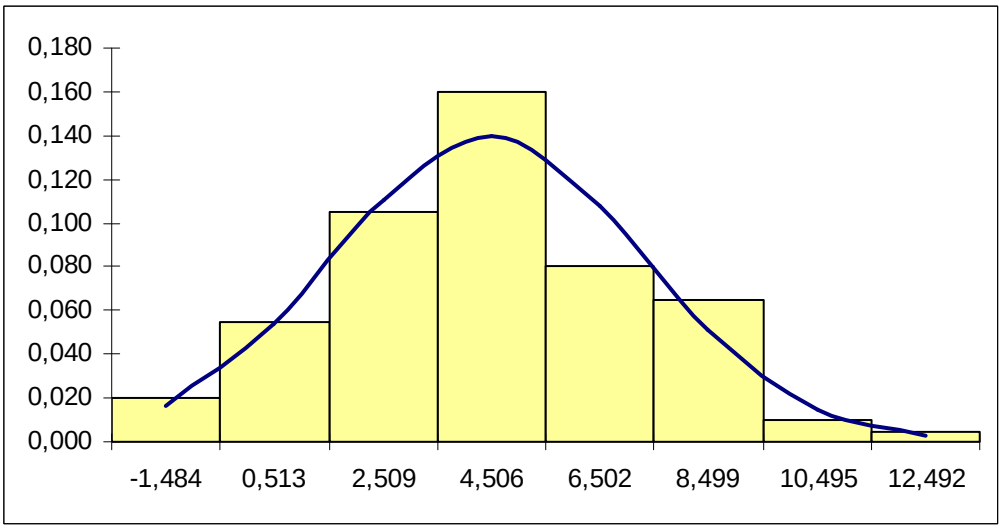


На одном чертеже с гистограммой построим кривую Гаусса для генеральной совокупности $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}$, заменив неизвестные значения генерального среднего m и генерального среднего квадратического отклонения σ их оценками $\bar{x}$ и s.

Ординаты точек кривой Гаусса $f\left(x_i^*\right)$ можно вычислить, используя таблицу значений функции $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ (Гмурман, приложение 1).

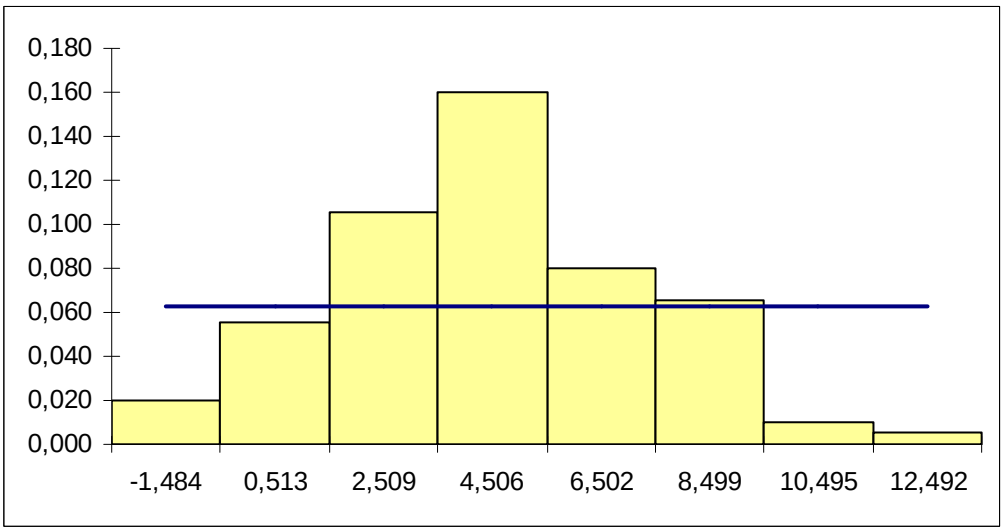
В нашем случае: $f\left(x_i^*\right)=\frac{1}{s}\varphi\left(\frac{x_i^*-\bar{x}}{s}\right)=\frac{1}{2{,}858}\varphi\left(\frac{x_i^*-4{,}446}{2{,}858}\right)$.

Получаем:



На том же чертеже построим график плотности равномерного распределения

$$f(x)=\begin{cases}1/(x_{\max}-x_{\min}), & x\in[x_{\max};x_{\min}]\\ 0, & x\notin[x_{\max};x_{\min}]\end{cases}=\begin{cases}1/15{,}972, & x\in[-2{,}482;13{,}49]\\ 0, & x\notin[-2{,}482;13{,}49]\end{cases}$$



Видно, что нормальное распределение подходит больше (лучше выравнивает гистограмму).

Построим доверительные интервалы $I_\gamma = \left(\overline{x} - \varepsilon_\gamma; \overline{x} + \varepsilon_\gamma\right)$ для неизвестного значения генерального среднего m, $\varepsilon_\gamma = \frac{s}{\sqrt{n}} t(\gamma; n)$ - точность интервальной оценки. $t(\gamma; n) = t_{\frac{1+\gamma}{2}}(n-1)$ - квантиль порядка $\frac{1+\gamma}{2}$ для распределения Стьюдента с $n-1$ степенью свободы.

Исследуем зависимость интервальных оценок от объема выборки n и доверительной вероятности γ.

Сравним доверительные интервалы для одинаковой доверительной вероятности $\gamma = 0{,}95$ и разных объемов выборки n=10 и n=100. Занесем результаты в таблицу (последние два столбца – искомые границы интервала):

γ	n	$t(\gamma; n)$	$\varepsilon_\gamma = \frac{s}{\sqrt{n}} t(\gamma; n)$	$\overline{x} - \varepsilon_\gamma$	$\overline{x} + \varepsilon_\gamma$
0,95	10	2,262	2,045	2,401	6,491
0,95	100	1,984	0,567	3,879	5,013

Видно, что при увеличении объема выборки, ширина доверительного интервала сильно уменьшается.

Сравним доверительные интервалы для различных доверительных вероятностей $\gamma = 0{,}95$; $\gamma = 0{,}99$ и $\gamma = 0{,}999$ и одинаковых объемов выборки n=100. Занесем результаты в таблицу:

γ	n	$t(\gamma;n)$	$\varepsilon_\gamma = \frac{s}{\sqrt{n}}t(\gamma;n)$	$\bar{x}-\varepsilon_\gamma$	$\bar{x}+\varepsilon_\gamma$
0,95	100	1,984	0,567	3,879	5,013
0,99	100	2,626	0,751	3,695	5,197
0,999	100	3,392	0,969	3,476	5,415

Видно, что с увеличением доверительной вероятности ширина доверительного интервала постепенно увеличивается.

Проверим гипотезу H_0: генеральная совокупность имеет нормальное распределение с параметрами $\bar{x}$ и s. В качестве альтернативной гипотезы принять гипотезу H_1: генеральная совокупность имеет иное распределение. Уровень значимости примем $\alpha = 0{,}05$.

Пронормируем случайную величину X, то есть перейдем к величине $Z = \frac{x-\bar{x}}{s}$, вычислим концы интервалов по формулам $z_i = \frac{x_i-\bar{x}}{s}$, $z_{i+1} = \frac{x_{i+1}-\bar{x}}{s}$. Вычислим теоретические (гипотетические частоты) $n_i' = nP_i$, где $n = 100$, $P_i = \Phi(z_{i+1}) - \Phi(z_i)$ - вероятность попадания в интервал (z_i, z_{i+1}), $\Phi(z)$ - функция Лапласа. Для нахождения значений составим расчетную таблицу:

x_i	x_{i+1}	n_i	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	P_i	n_i'
-2,482	-0,4855	4	-2,4238	-1,7253	-0,5000	-0,4578	0,0422	4,2237
-0,4855	1,511	11	-1,7253	-1,0268	-0,4578	-0,3477	0,1100	11,0021
1,511	3,5075	21	-1,0268	-0,3283	-0,3477	-0,1287	0,2191	21,9086
3,5075	5,504	32	-0,3283	0,3702	-0,1287	0,1444	0,2730	27,3040
5,504	7,5005	16	0,3702	1,0687	0,1444	0,3574	0,2130	21,3013
7,5005	9,497	13	1,0687	1,7672	0,3574	0,4614	0,1040	10,4004
9,497	11,494	2	1,7672	2,4657	0,4614	0,4932	0,0318	3,1760
11,494	13,49	1	2,4657	3,1642	0,4932	0,5000	0,0068	0,6837

Присоединим интервалы с малыми интервальными частотами $(n_i < 5)$ к соседним интервалам. Наблюдаемое значение критерия вычислим по формуле $\chi^2_{\text{íàáë}} = \sum_{i=1}^{5} \frac{(n_i - n_i')^2}{n_i'}$ и вычисления представим в виде таблицы.

n_i	n_i'	$\frac{(n_i-n_i')^2}{n_i'}$
15	15,226	0,003
21	21,909	0,038
32	27,304	0,808
16	21,301	1,319
16	14,260	0,212

Сумма		**2,380**

По таблице критических значений $\chi^2_{кр}$ при уровне значимости $\alpha = 0,05$ и числе степеней свободы $k = l - 3 = 5 - 3 = 2$ найдем $\chi^2_{êð} = 5,99$. Так как $\chi^2_{íàáë} = 2,38 < \chi^2_{êð} = 5,99$, можно принять нулевую гипотезу при данном уровне значимости. Опытные данные не противоречат гипотезе о нормальном распределении. Это же подтверждается видом гистограммы и кривой теоретической плотности распределения (см. выше чертеж).