ЛАБОРАТОРНАЯ РАБОТА

Систематизация и графическое представление малой и большой выборок

§1. Цель работы

Овладение приемами первичной обработки статистических данных.

§2. Содержание работы

Лабораторная работа состоит из двух заданий.

Задание 1. Каждому студенту предлагается вариант выборки объемом $n = 20$ элементов (малая выборка), равных количеству выходов из строя в сутки автобусов предприятия, например

№ элемента	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
элемент	2	1	5	4	5	2	6	4	3	6	5	3	7	4	9	8	7	3	6	4

Требуется:

1.1 Построить вариационный ряд.

1.2 Найти x_{min}, x_{max}, $z_{1/4}$, $z_{3/4}$, med и построить «ящик с усами».

1.3 Построить статистический ряд и полигон.

1.4 Вычислить выборочные числовые характеристики $\overline{x}$ и s.

Задание 2. Для обработки каждому студенту предлагается большая выборка, объемом 100 элементов, которые могут интерпретироваться как отклонения результатов измерения изучаемой случайной величины X от ее среднего значения. (Большая выборка представляет собой таблицу из пяти полос. Каждая полоса содержит 20 чисел.)

Данные выборки являются фрагментами таблицы случайных чисел, распределённых нормально $N(0;1)$.

Требуется:

2.1 построить группированный статистический ряд;

2.2 построить гистограмму приведенных частот $n_i/(nh)$ (n_i – частоты попадания элементов выборки в промежутки, n – объем выборки, h – длина каждого промежутка);

2.3 построить полигон приведенных частот $n_i/(nh)$ для средних точек z_i промежутков группированного статистического ряда;

2.4 эмпирическую функцию распределения $F^*(x)$ и её график – *кумуляту* для средних

точек z_i промежутков группированного статистического ряда;

2.5 относительную частоту p_b^* события $X < 0$;

2.6 для сравнения вместе с гистограммой в том же масштабе строится стандартная нормальная кривая $y = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Значения функции $\varphi(x)$ (Гаусса) приведены в таблице 1.1

Таблица 1.1. Значения функции Гаусса $\varphi(x)$.

x	0	0,5	1,0	1,5	2,0	2,5	3,0
$\varphi(x)$	0,40	0,35	0,24	0,13	0,05	0,02	0,004

2.7 для сравнения вместе с кумулятой в том же масштабе строится график функции Лапласа $y = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$. Ее значения приведены в таблице 1.2

Таблица 1.2. Значения функции Лапласа $\Phi(x)$.

x	$-\infty$	–2,0	–1,5	–1,0	–0,5	0,0	0,5	1,0	1,5	2,0	2,5	$+\infty$
$\Phi(x)$	0,00	0,02	0,07	0,16	0,31	0,50	0,69	0,84	0,93	0,98	0,99	1,00

2.8 Вычислить $\overline{x}$ и s – оценки генерального математического ожидания m_X и среднего квадратического отклонения σ_X, соответственно.

2.9 Вычислить точность ε оценки $\overline{x}$ при заданной надежности $\gamma = 0.95$.

2.10 Построить доверительный интервал для математического ожидания m нормальной генеральной совокупности при $\gamma = 0.95$.

§3. Методика проведения работы

3.1 Студенты знакомятся с содержанием Лабораторной работы. Затем они разбирают образцы Лабораторной работы.

3.2 Каждый студент получает свои статистические данные, записывает содержание задания и самостоятельно начинает его выполнять. Оформляет работу в тетради в клетку.

3.3 Преподаватель проверяет выполненные работы и проводит её обсуждение. Цель обсуждения – сделать выводы из полученных результатов.

Замечание. Так как в задании 2 для обработки дана выборка из нормальной генеральной совокупности с известными параметрами $m = 0$ и $\sigma = 1$, то фактически даны задачи с известными ответами, а именно:

- Гистограмма и полигон являются графическими оценками графика генеральной плотности, т.е. кривой Гаусса. На сделанных чертежах это должно быть видно.

- Кумулята является графической оценкой графика генеральной функции распределения $y = \mathbf{P}(X < x) = \Phi(x)$, т.е. функции Лапласа. На сделанных чертежах факт близости графиков должен быть заметен.

§4. Краткие теоретические сведения (глоссарий)

Обработка выборок позволяет сделать обоснованные суждения о свойствах генеральной совокупности.

4.1 *Вариационным рядом* называется последовательность элементов выборки, записанная в неубывающем порядке: $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$. Равные элементы выборки повторяются. Элементы $x_{(i)}$ вариационного ряда называются порядковыми статистиками; $x_{min} = x_{(1)}, x_{max} = x_{(n)}$ – крайние порядковые статистики.

4.2 *Медианой выборки* med называется средний элемент вариационного ряда, если объем выборки n – число нечетное, или полусумма двух средних элементов, если n – четное:

$$med = \begin{cases} x_{(l+1)} & \text{при } n = 2l+1, \\ \frac{x_{(l)} + x_{(l+1)}}{2} & \text{при } n = 2l. \end{cases} \qquad (1.1)$$

При $n = 20$ $med = \frac{1}{2}(x_{(10)} + x_{(11)})$.

В общем случае выборочная медиана med является оценкой генеральной медианы **Me**, которая определяется как корень уравнения $F(x) = 1/2$, где $F(x) = \mathbf{P}(X < x)$ непрерывная строго возрастающая генеральная функция распределения. В случае симметричного непрерывного распределения медиана *med* оценивает центр симметрии распределения, совпадающий с **Me**.

4.3 *Выборочными квартилями* $z_{1/4}, z_{3/4}$ называются элементы вариационного ряда, на четверть отстоящие от краев. Их точное определение дается формулами:

$$z_{1/4} = x_{(i)};\ i = \begin{cases} [n/4] + 1 & \text{при } n/4 \text{ дробном,} \\ n/4 & \text{при } n/4 \text{ целом.} \end{cases} \qquad (1.2)$$

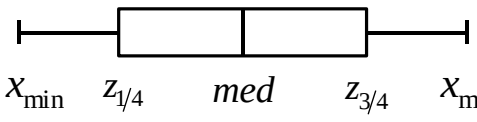
Здесь $[a]$ – целая часть числа a, т.е. наибольшее целое, не превосходящее a.

$$z_{3/4} = x_{(n-i+1)}. \qquad (1.3)$$

При $n = 20$ $z_{1/4} = x_{(5)}$, $z_{3/4} = x_{(16)}$.

Выборочные квартили $z_{1/4}$, $z_{3/4}$ являются статическими оценками соответствующих генеральных квартилей $x_{1/4}$ и $x_{3/4}$, определяемых как корни уравнений $F(x) = 1/4$ и $F(x) = 3/4$. Здесь $F(x)$ – непрерывная строго возрастающая генеральная функция распределения.

4.4 Ящик с усами (box plot) дает сжатое обобщенное представление о выборке.



Относительная частота p_x события $X < x$ определяется формулой

$$p_x = \frac{m(x)}{n}, \qquad (1.4)$$

где n – объем выборки, $m(x)$ – количество элементов вариационного ряда, находящихся левее точки x.

$p_b = \frac{m(0)}{n}$ – для большой выборки.

Для симметричного генерального распределения $p_b \approx 1/2$. Величина отступления p_b от $1/2$ численно характеризует асимметричность выборки.

4.6 Группированный статистический ряд строится для большой выборки следующим образом. Промежуток $[x_{min},\ x_{max}]$ разбивается на k равных промежутков $\Delta_1, \Delta_2, \ldots, \Delta_k$. Число k находится по одной из формул:

$$k \approx 1{,}72n^{1/3}; \quad k \approx 1 + 3{,}3\lg n \quad \text{(формула Старджесса)} \qquad (1.5)$$

Для $n = 100 \Rightarrow k = 8$, для $n = 200 \Rightarrow k = 10$.

Длина каждого промежутка

$$h = \frac{R}{k} = \frac{x_{\max} - x_{\min}}{k}.$$

Пусть $a_0, a_1, \ldots, a_k$ – граничные точки промежутков. Тогда $\Delta_1 = [x_{\min}; a_1)$, $\Delta_2 = [a_1; a_2), \ldots, \Delta_i = [a_{i-1}; a_i), \ldots, \Delta_k = [a_{k-1}; x_{\max}]$.

Пусть n_i – число элементов выборки, попавших в промежуток Δ_i. Числа $n_1, n_2, \ldots, n_k$ называются *частотами попадания элементов выборки* в рассматриваемые промежутки.

Группированным статистическим рядом называется совокупность промежутков $\Delta_1, \Delta_2, \ldots, \Delta_k$ и соответствующих им частот $n_1, n_2, \ldots, n_k$. Группированный статистический ряд оформляется в виде таблицы (см. ниже образец расчета).

4.7 *Гистограммой приведенных частот* выборки называется фигура, образованная прямоугольниками с основаниями Δ_i и высотами $n_i/(nh)$. Числа n_i/n называются относительными, а числа $n_i/(nh)$ – приведенными частотами группированной выборки.

Ступенчатая ломаная, ограничивающая гистограмму сверху, при увеличении n сближается с кривой генеральной плотности вероятности, поэтому является ее оценкой.

Ординаты гистограмм приведенных относительных частот, относительных частот и просто частот пропорциональны, поэтому дают одинаковое представление о выборочном распределении.

4.8 *Полигоном* приведенных частот группированной выборки называется ломаная с вершинами в точках $\left(x_i^*,\ \frac{n_i}{nh}\right)$ $(i = 1, \ldots, k)$. Здесь x_i^* – середина промежутка Δ_i. С помощью полигона также оценивается кривая генеральной плотности вероятности.

4.9 *Эмпирической функцией распределения* называется относительная частота события $X < x$. Для группированного статистического ряда она строится для средних точек x_i^* промежутков Δ_i, как представителей всех элементов выборки, попавших в Δ_i

$$F_n^*(x) = \frac{1}{n}\sum_{x_i^*<x} n_i \qquad (1.6)$$

$F_n^*(x)$ является функцией распределения дискретной случайной величины X^*, заданной таблицей распределения

X^*	x_1^*	x_2^*	…	x_k^*
P	n_1/n	n_2/n	…	n_k/n

Ее графиком является восходящая ступенчатая линия, называемая кумулятой (линия накопленных относительных частот). Кумулята является оценкой графика генеральной функции распределения $y = F(x)$.

§5. Образец выполнения лабораторной работы

Задание 1. Дана выборка объемом $n = 20$ элементов, равных количеству выходов из строя в сутки автобусов автопредприятия.

№ элемента	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
элемент	2	1	5	4	5	2	6	4	3	6	5	3	7	4	9	8	7	3	6	4

▶ 1.1. Строим вариационный ряд, упорядочивая элементы выборки в неубывающем порядке:

1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9.

1.2. Находим $x_{\min}$, $x_{\max}$, $z_{1/4}$, $z_{3/4}$, med.

$x_{\min} = 1$ – первый элемент вариационного ряда.

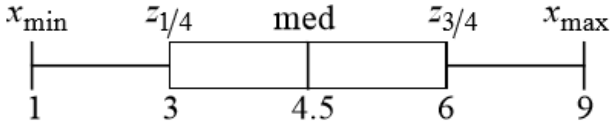
$x_{\max} = 9$ – последний элемент вариационного ряда.

$\text{med} = \frac{1}{2}(x_{(10)} + x_{(11)}) = \frac{1}{2}(4+5) = 4.5$.

Нижнюю квартиль $z_{1/4}$ находим по формулам

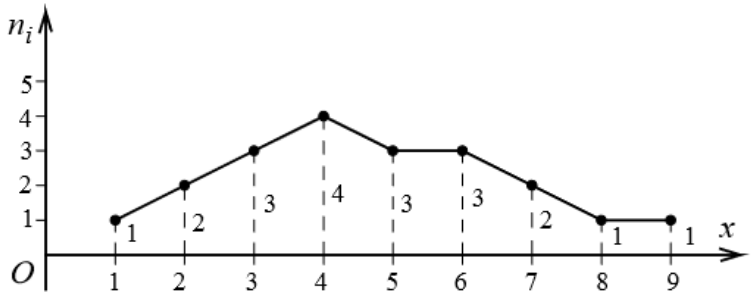
$$z_{1/4} = x_{(i)}; \quad i = \left[\frac{n+3}{4}\right].$$

Здесь $n = 20$; $i = \left[\frac{20+3}{4}\right] = \left[\frac{23}{4}\right] = 5$; $z_{1/4} = x_{(5)} = 3$. Верхнюю квартиль $z_{3/4}$ находим по формуле: $z_{3/4} = x_{(n-i+1)} = x_{(16)} = 6$. «Ящик с усами»:



1.3. Строим статистический ряд и его графическое изображение – полигон.

Элементы z_i	1	2	3	4	5	6	7	8	9	Σ
Частоты n_i	1	2	3	4	3	3	2	1	1	20



По конфигурации этот полигон близок к полигону распределения Пуассона с параметром $a = 5$.

1.4. Вычисляем $\bar{x}$ и s.

$$\bar{x} = \frac{1}{n}\sum_i n_i z_i = \frac{1}{20}(1\cdot 1 + 2\cdot 2 + 3\cdot 3 + 4\cdot 4 + 5\cdot 3 + 6\cdot 3 + 7\cdot 2 + 8\cdot 1 + 9\cdot 1) =$$

$$= \frac{1}{20}(1 + 4 + 9 + 16 + 15 + 18 + 14 + 8 + 9) = 94/20 = 4.7;$$

$$s^2 = \frac{1}{n}\sum_i n_i z_i^2 - \bar{x}^2 =$$

$$= \frac{1}{20}(1\cdot 1^2 + 2\cdot 2^2 + 3\cdot 3^2 + 4\cdot 4^2 + 3\cdot 5^2 + 3\cdot 6^2 + 2\cdot 7^2 + 1\cdot 8^2 + 1\cdot 9^2) - 4.7^2 =$$

$$= \frac{1}{20}(1 + 8 + 27 + 64 + 75 + 108 + 98 + 64 + 81) - 22.09 = 4.21;$$

$$s = \sqrt{s^2} = \sqrt{4.21} \approx 2.05.$$

Заключение. Вычисленные значения $\text{med} = 4.5$, $\bar{x} = 4.7$, $s^2 = 4.21$ близки к указанному в ответе значению параметра $a = 5$ закона распределения Пуассона, что является грубой проверкой ответов. ◀

Задание 2. Дана большая выборка объема $n = 100$ элементов из нормальной генеральной совокупности $N(0; 1)$, записанная в виде таблицы 1.3.

Таблица 1.3

№ п/п	1	2	3	4	5	6	7	8	9	10
1	1,49	–0,35	–0,63	–0,70	0,93	1,39	0,77	–0,96	–0,85	–1,86
2	1,02	–0,47	1,28	**2,52**	0,57	–1,85	0,19	–0,50	–0,27	1,19
3	–0,29	–0,45	0,44	0,25	–1,38	–1,81	–0,60	–1,63	–2,53	0,19
4	0,08	–0,24	–0,29	–0,54	0,32	–0,77	0,40	–0,02	1,03	0,12
5	1,78	–0,07	1,73	–1,49	–2,36	1,29	1,12	0,70	–0,90	0,37
6	–1,09	1,03	–0,06	–0,08	–0,99	–0,64	–1,03	0,66	1,66	–0,93

7	–0,70	0,43	–0,59	–1,52	–1,04	–1,31	0,09	–0,07	**–2,76**	–0,10
8	0,47	–0,66	1,44	–1,21	0,28	–1,04	1,02	0,19	–0,15	0,19
9	–0,51	0,34	–1,27	–1,16	–0,26	1,56	–1,44	1,18	–0,70	–0,31
10	1,42	–0,43	0,63	0,82	0,41	–0,87	0,66	–1,05	0,61	0,62

2.1 Образуем группированный статистический ряд.

Находим $x_{\min} = -2{,}76$; $x_{\max} = 2{,}52$.

Вычисляем размах $R = x_{\max} - x_{\min} = 2{,}52 + 2{,}76 = 5{,}28$.

Число промежутков группированного статистического ряда принимаем равным $k = 8$.

Длина каждого промежутка $h = \dfrac{R}{k} = \dfrac{5{,}28}{8} = 0{,}66$.

Образуем промежутки группированного статистического ряда:

$$\Delta_1 = [-2{,}76;\ -2{,}10), \qquad \Delta_2 = [-2{,}10;\ -1{,}44),$$
$$\Delta_3 = [-1{,}44;\ -0{,}78), \qquad \Delta_4 = [-0{,}78;\ -0{,}12),$$
$$\Delta_5 = [-0{,}12;\ 0{,}54), \qquad \Delta_6 = [0{,}54;\ 1{,}20),$$
$$\Delta_7 = [1{,}20;\ 1{,}86), \qquad \Delta_8 = [1{,}86;\ 2{,}52].$$

Распределяем элементы выборки по образованным промежуткам Δ_i и подсчитываем частоты n_i ($i = 1, \ldots, 8$). Технически это делается следующим образом. Просматриваем выборку по порядку и каждый элемент относим в соответствующий промежуток, ставя при этом палочку в графе (таблица 1.4) рядом. Когда накапливается четыре палочки ||||, их перечеркиваем после появления в этом промежутке следующего элемента. Получается пяток  . Затем образуем следующий пяток и т.д. Все результаты, оформляем в виде таблицы 1.4.

№ промежутка	Границы промежутков		Подсчет частот	n_i	Средняя точка промежутка
	a_{i-1}	a_i			
1	–2,76	–2,10	|||	3	–2,43
2	–2,10	–1,44	𝍸 |	6	–1,77
3	–1,44	–0,76	𝍸 𝍸 𝍸 |||	18	–1,11
4	–0,76	–0,12	𝍸 𝍸 𝍸 𝍸 ||	22	–0,45
5	–0,12	0,54	𝍸 𝍸 𝍸 𝍸 |||	23	0,21
6	0,54	1,20	𝍸 𝍸 𝍸 ||	17	0,87
7	1,20	1,86	𝍸 𝍸	10	1,53
8	1,86	2,52	|	1	2,19
Σ	–	–	–	100	–

2.2 Строим гистограмму приведенных частот (рис.1).

2.3 Строим полигон (рис. 1).

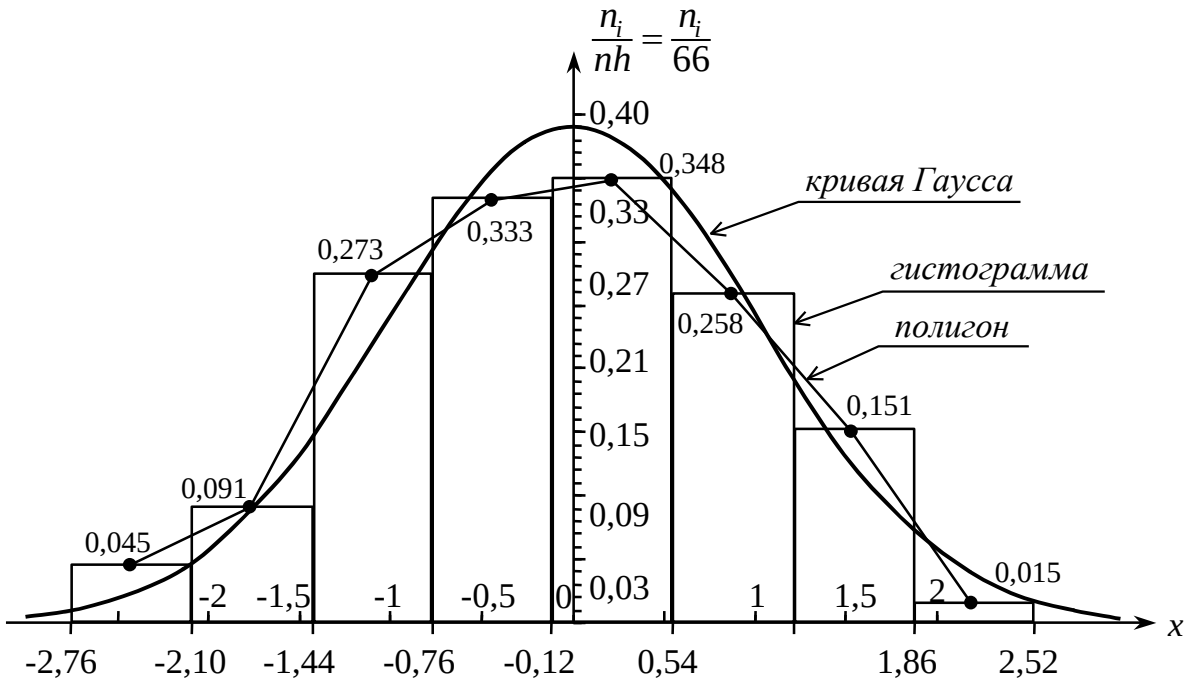


Рис. 1. Гистограмма приведенных частот, полигон, кривая Гаусса.

2.4 Строим кумуляту. Предварительно составляем аналитическое выражение для эмпирической функции распределения:

$$y = F^*(x) = \begin{cases} 0, & -\infty < x < -2{,}43; \\ 0{,}03, & -2{,}43 \le x < -1{,}77; \\ 0{,}09, & -1{,}77 \le x < -1{,}11; \\ 0{,}27, & -1{,}11 \le x < -0{,}45; \\ 0{,}49, & -0{,}45 \le x < 0{,}21; \\ 0{,}72, & 0{,}21 \le x < 0{,}87; \\ 0{,}89, & 0{,}87 \le x < 1{,}53; \\ 0{,}99, & 1{,}53 \le x < 2{,}19; \\ 1, & 2{,}19 \le x < +\infty. \end{cases}$$

График приведён на рис 2.

2.5 С помощью таблицы 1.3 вычисляем относительную частоту события $X < 0$:

$$p_b^* = \frac{m(0)}{n} = \frac{55}{100} = 0{,}55.$$

2.6 Для сравнения вместе с гистограммой в том же масштабе строится стандартная нормальная кривая $y = \varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. Значения функции $\varphi(x)$ (Гаусса) приведены в таблице 1.1.

2.7 Для сравнения вместе с кумулятой в том же масштабе строится график функции

Лапласа $y = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$. Ее значения приведены в таблице 1.2.

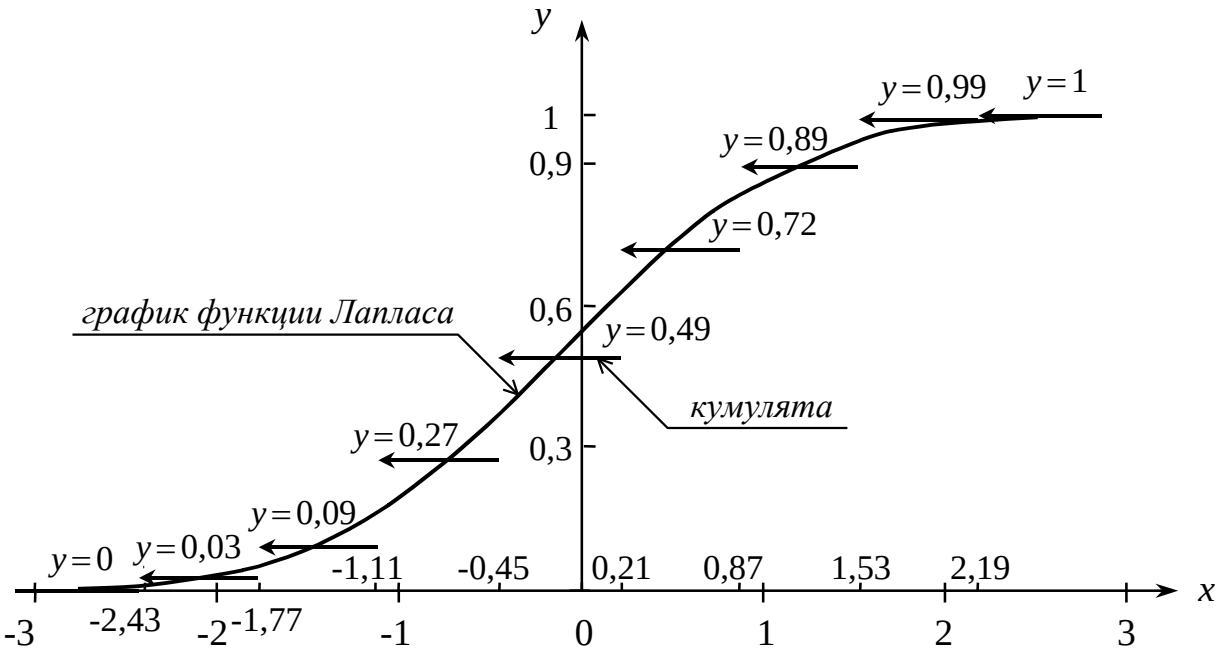


Рис. 2. Кумулята и график функции Лапласа.

2.8 Вычисляем выборочные числовые характеристики $\overline{x}$ и s

$$\overline{x} = \frac{1}{100} \sum_{i=1}^{8} n_i \cdot x_i^* = \frac{1}{100} (3 \cdot (-2.43) + 6 \cdot (-1.77) + 18 \cdot (-1.11) +$$

$$+22 \cdot (-0.45) + 23 \cdot 0.21 + 17 \cdot 0.87 + 10 \cdot 1.53 + 1 \cdot 2.19) = -0.1068 \approx -0.11;$$

$$+22 \cdot (-0.45) + 23 \cdot 0.21 + 17 \cdot 0.87 + 10 \cdot 1.53 + 1 \cdot 2.19) = -0.1068 \approx -0.11;$$

$$s^2 = \frac{1}{100} \sum_{i=1}^{8} n_i \cdot (x_i^*)^2 - \overline{x}^2 = 1.0409;$$

$$S = \sqrt{s^2} = \sqrt{1.0409} \approx 1.02.$$

2.9 Вычисляем точность ε оценки $\overline{x}$ при заданной надежности $\gamma = 0.95$. Для этого применим формулы

$$\mathbf{P}(|\overline{x} - m| < \varepsilon) \approx 2\Phi(x) - 1 = \gamma, \qquad \varepsilon = \frac{sx}{\sqrt{n}}$$

Отсюда $\Phi(x) = \frac{1+\gamma}{2} = \frac{1+0.95}{2} = 0.975$. С помощью таблицы квантилей нормального распределения $N(0;1)$ (в конце книги) решаем это уравнение, т. е. находим квантиль $x = u_{\frac{1+\gamma}{2}} = u_{0.975} = 1.96$. Тогда

$$\varepsilon = \frac{s \cdot u_{\frac{1+\gamma}{2}}}{\sqrt{100}} = \frac{1.02 \cdot 1.96}{10} \approx 0.2.$$

Итак, $\mathbf{P}(|\overline{x} - m| < 0.2)$. Иначе $m = -0.11 \pm 0.2$ с надежностью $\gamma = 0.95$.

2.10 Строим доверительный интервал для m при $\gamma = 0.95$. Для этого используем результат предыдущего пункта 2.6.

$\mathbf{P}(\overline{x} - \varepsilon < m < \overline{x} + \varepsilon) = \gamma$; $\mathbf{P}(-0.11 - 0.2 < m < -0.11 + 0.2) = 0.95$.

Таким образом, $-0.31 < m < -0.09$ с надёжностью $\gamma = 0.95$.

Выводы.

1. График выборочных распределений – гистограмма – ступенчатая линия, полигон, кумулята хорошо аппроксимирует теоретические кривые – Гаусса и Лапласа, поэтому можно выдвинуть гипотезу о нормальности генерального распределения.
2. Относительная частота $p_b^* = 0{,}55$ в сравнении с истинной вероятностью $\mathbf{P}(X < 0) = 0{,}5$ свидетельствуют в пользу симметричности генерального распределения.
3. Заметим попутно, что по кумуляте тоже можно приближенно найти относительную частоту p_b. Таким образом, получаем $p_b \approx 0,49$.