Математическая статистика. Пример решения задачи.

Условие.

1. Сгруппировать выборку и записать статистические ряды абсолютных и относительных частот.
2. Представить выборку графически: построить полигон абсолютных частот; полигон относительных частот; нормированные гистограммы.
3. Найти оценки вариации: выборочное среднее, дисперсию, среднее квадратическое отклонение, моду, медиану.
4. Выдвинуть и проверить с уровнем значимости α=0,05 гипотезу о нормальном законе распределения генеральной совокупности, построить график подобранной функции плотности (вместе с гистограммой)
5. Построить доверительные интервалы для параметров распределения генеральной совокупности.
6. Сформулировать статистические выводы. Они должны содержать сводные результаты по каждому пункту исследования.

Данные.

48	39	43	44	34	34	32	43	40	46
25	31	34	49	39	37	45	49	31	49
43	46	34	35	42	32	41	34	42	42
38	40	46	47	34	42	38	40	38	36
30	43	41	40	40	35	35	41	38	45
37	42	38	36	44	39	32	48	43	39
43	30	32	36	42	34	49	48	49	50
37	30	44	48	44	35	45	34	33	41
43	45	50	34	33	39	41	39	46	31
40	52	44	39	35	45	33	42	42	36
44	51	45	39	34	44	40	37	43	32
33	42	40	35	37	43	48	48	50	32
40	48	45	43	36	36	42	40	37	30
44	50	46	39	41	48	44	42	36	51
44	50	47	37	33	34	42	43	43	47

Решение.

Нам дана выборка размера (объема) 150 чисел. Для начала найдем минимальное и максимальное числа в этой выборке. Минимум равен 25, максимум равен 52. Тогда размах выборки (разница между максимальным и минимальным числом) равен $R = 52 - 25 = 27$. Далее определим число интервалов, на которые мы будем разбивать выборку. По формуле это число равно $k = 1 + 3.32 * lg150 \approx 8$. Длина одного интервала определяется по формуле $h = R/k = 3{,}375$.

Теперь, зная на сколько (8) интервалов какой длины (3,375) нужно разбивать выборку, сделаем это. И подсчитав количество чисел в каждом интервале, запишем результаты в таблицу:

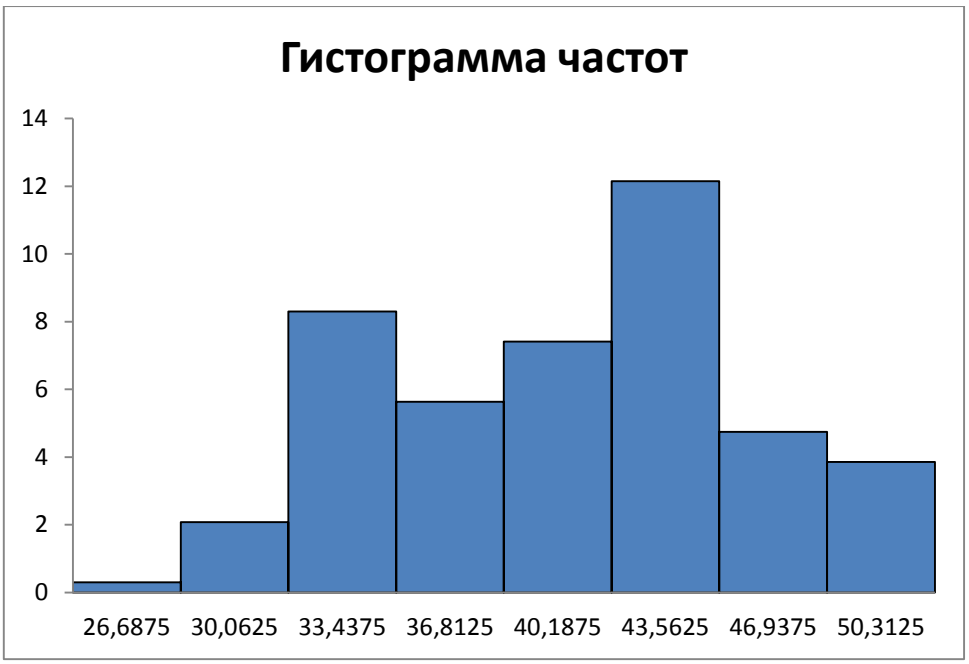
Номер интервала	Интервал	Частоты	Середины интервалов	Относительные частоты
1	25,0-28,375	1	26,6875	0,007
2	28,375-31,75	7	30,0625	0,047
3	31,75-35,125	28	33,4375	0,187
4	35,125-38,5	19	36,8125	0,127
5	38,5-41,875	25	40,1875	0,167
6	41,875-45,25	41	43,5625	0,273
7	45,25-48,625	16	46,9375	0,107
8	48,625-52,0	13	50,3125	0,087
Всего		150		1,000

Частоты -- это и есть количества чисел в каждом интервале. Относительные частоты -- это частоты, деленные на 150 (количество чисел в выборке). Середины интервалов нам понадобятся в дальнейших расчетах, поэтому их мы также внесли в таблицу.

Итак, мы выполнили первый пункт задания -- сгруппировали выборку и записали ряды частот. Переходим ко второму пункту: рисуем графики.

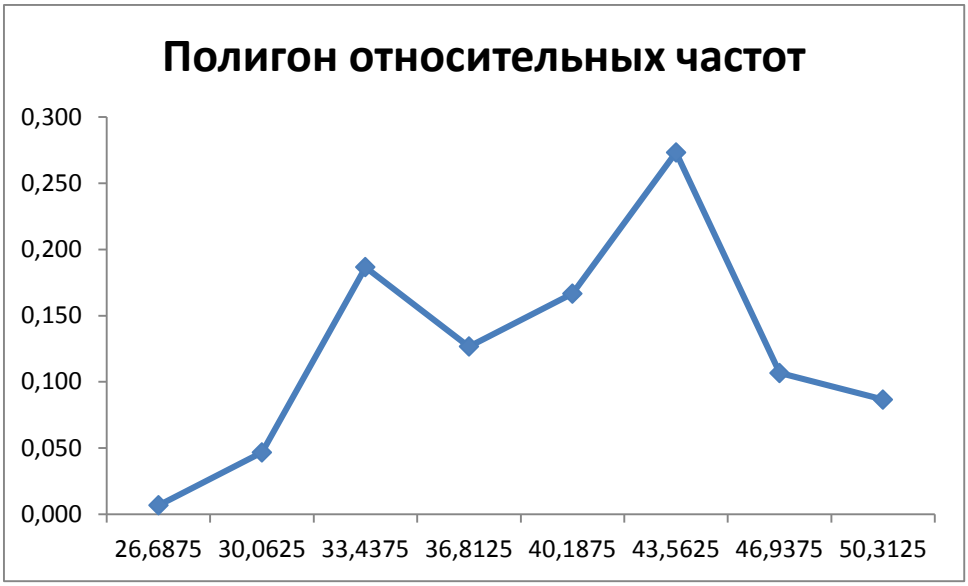
1) Полигон абсолютных частот -- это график, где по оси OX идут наши интервалы (на графике подписаны середины интервалов), а по оси OY -- частоты:

2) Гистограмма -- это график, представленный в виде диаграммы со столбцами вместо точек, высоты которых равны $\frac{m_i}{h}$, где m_i -- соответствующие частоты, а h -- длина интервала разбиения (см. выше). Подписи по горизонтальной оси снова середины отрезков для удобства построения:

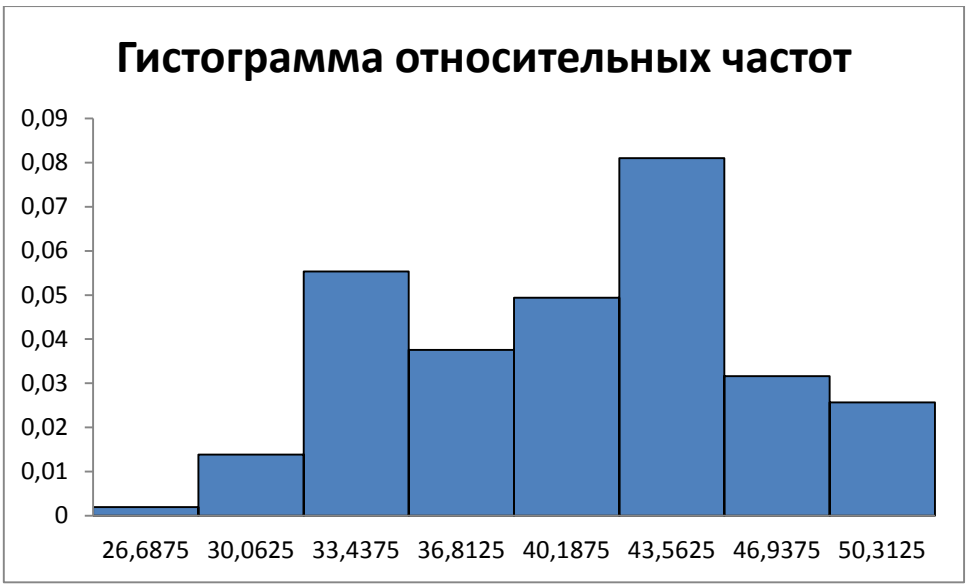


Сумма площадей всех прямоугольников в гистограмме равна размеру (объему) выборки (150 в нашем случае).

3) Полигон относительных частот строится аналогично обычному, только по оси OY отмечаются уже не частоты, а относительные частоты (легко проверить себя так: в нормированном полигоне по оси OY все числа должны быть меньше 1):



Нормированная гистограмма относительных частот -- строится аналогично обычной, только все значения по оси OY делятся на размер (объем) выборки (150 в нашем случае). Сумма площадей всех прямоугольников нормированной гистограммы относительных частот будет равна 1.



Итак, все необходимые графики построены, тем самым мы завершили выполнение пункта 2.

Далее необходимо найти некоторые числовые характеристики.

$\bar{x} = \frac{m_i * x_i}{n}$; $D_{\text{выб}} = \frac{m_i * (x_i - \bar{x})^2}{n}$; $s^2 = D_{\text{исправл}} = D_{\text{выб}} * \frac{n}{n-1}$; $s = \sqrt{D_{\text{исправл}}}$.

Для этого построим вспомогательную таблицу:

Номер интервала	Середины интервалов(x_i)	Частоты (m_i)	$m_i * x_i$	$m_i * (x_i - \bar{x})^2$
1	26,6875	1	26,6875	189,6129
2	30,0625	7	210,4375	756,392175
3	33,4375	28	936,25	1379,8512
4	36,8125	19	699,4375	252,434475
5	40,1875	25	1004,6875	1,8225
6	43,5625	41	1786,0625	395,282025
7	46,9375	16	751	671,8464
8	50,3125	13	654,0625	1262,573325
Всего		150	6068,625	4909,815

Первые три столбца в этой таблицы взяты из предыдущей. Далее идет столбец произведений середины интервала на частоту в этом интервале. Сложив все числа в этом столбце и поделив сумму на размер (объем) выборки (150), мы получаем выборочное среднее: $\bar{x} = 6068{,}625/150 = 40{,}4575$. Далее, применяя полученное число, рассчитываем следующий столбец согласно формулам в заглавии (вычитаем из середины отрезка выборочное среднее, возводим в квадрат и умножаем на частоту в этом отрезке). Поделим сумму из пятого столбца на размер выборки (150), получаем выборочную дисперсию: $D_{\text{выб}} = 4909{,}815/150 = 32{,}7321$. Исправленная выборочная дисперсия считается по формуле $D_{\text{исправл}} = D_{\text{выб}} * 150/149 = 32{,}9517$. Исправленное среднее квадратическое отклонение (далее СКО) равно корню из исправленной дисперсии $s = \sqrt{D_{\text{исправл}}} = 5{,}7404$.

Медианой называется такое значение признака, которое делит весь вариационный ряд пополам. Для ее расчета нам нужно сначала найти медианный интервал. Это 5-ый интервал (частота $m_i = 25$). Его начало -- это число $X_0 = 38{,}5$. Сумма частот всех предыдущих интервалов равна $m_{i-1}{}' = 1 + 7 + 28 + 19 = 55$. Тогда медиана вычисляется по формуле:

$$Me = X_0 + h * \frac{0{,}5 * n - m_{i-1}{}'}{m_i} = 38{,}5 + 3{,}375 * \frac{0{,}5 * 150 - 55}{25} = 41{,}2$$

где h -- это длина одного интервала (3,375), а n -- это размер выборки (150).

Модой называется наиболее часто встречающееся в выборке значение. Для ее расчета нам снова понадобится интервал с наибольшей частотой. Формула для расчета моды:

$$Mo = X_0 + h * \frac{m_i - m_{i-1}}{m_i - m_{i-1} + m_i - m_{i+1}} = 41{,}875 + 3{,}375 * \frac{41 - 25}{41 - 25 + 41 - 16} = 43{,}192$$

где m_i -- это частота в нашем 6-ом интервале, а m_{i-1} и m_{i+1} -- это частоты в 5-ом и 7-ом интервалах соответственно.

На этом завершается выполнение заданий третьего пункта.

Выдвинем гипотезу о том, что распределение генеральной совокупности подчиняется нормальному закону. Для расчета значения критерия Пирсона заполняем таблицу:

Номер интервала	Начало (a_{i-1})	Конец (a_i)	Частота (n_i)	$b_i = \frac{a_i - \bar{x}}{s}$	$\Phi(b_i)$	$p_i = \Phi(b_i) - \Phi(b_{i-1})$	$n * p_i$	$\frac{(n_i - np_i)^2}{np_i}$
1	-2E+16	28,37	1	-3E+15	0	0,018	2,648	1,026
2	28,375	31,75	7	-2,105	0,018	0,047	7,049	3E-04
3	31,75	35,12	28	-1,517	0,065	0,112	16,77	7,517
4	35,125	38,5	19	-0,929	0,176	0,19	28,51	3,174
5	38,5	41,87	25	-0,341	0,367	0,231	34,65	2,685
6	41,875	45,25	41	0,2469	0,598	0,201	30,09	3,958
7	45,25	48,62	16	0,8349	0,798	0,124	18,67	0,383
8	48,625	2E+16	13	1,4228	0,923	0,077	11,61	0,167

Где Ф – нормированная функция Лапласа (значения этой функции указаны в таблице 1). Все остальные обозначения указаны в таблице.

Вычисляя сумму чисел в последнем столбце получаем $\chi^2 = 18{,}9103$.

По таблице квантилей распределения хи-квадрат найдем значение, с которым нужно сравнить полученное. В наших условиях число степеней свободы равно $r = 8 - 2 - 1 = 5$. Для заданного уровня значимости $\alpha = 0{,}05$ получаем $\chi^2{}_{1-\alpha} = \chi^2{}_{0{,}95} = 11{,}0705$. Поскольку полученное нами значение χ^2 больше ($18{,}9103 > 11{,}0705$), то гипотеза о нормальном распределении не согласуется с имеющимися данными.

Поскольку гипотеза не подтвердилась, то для оценки математического ожидания генеральной совокупности используем формулу:

$$P\{\bar{x} - u_{\frac{1+\gamma}{2}} \frac{s}{\sqrt{n}} < m < \bar{x} + u_{\frac{1+\gamma}{2}} \frac{s}{\sqrt{n}}\} = \gamma$$

Из таблицы 2 (квантили нормального распределения) имеем $u_{\frac{1+\gamma}{2}} = u_{0,975} = 1,96$. Поэтому

$$P\{40,4575 - 1,96\frac{5,7404}{\sqrt{150}} < m < 40,4575 + 1,96\frac{5,7404}{\sqrt{150}}\} = 0,95$$

т.е.

$$P\{39,5386 < m < 41,3761\} = 0,95$$

т.е. с вероятностью 0,95 генеральное среднее значение признака попадет в интервал (39,5386 ; 41,3761).

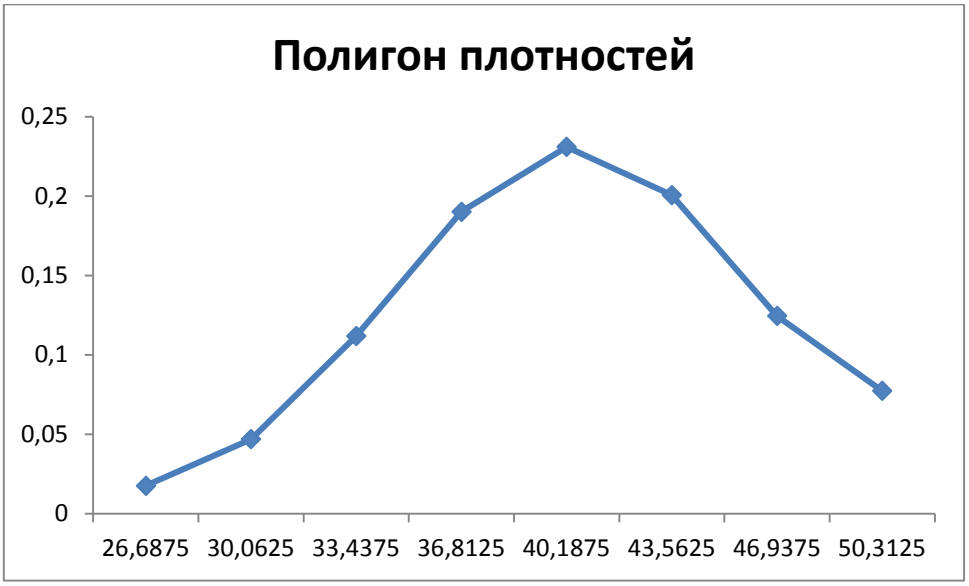
(Если гипотеза о нормальном распределении генеральной совокупности подтвердилась, то для построения доверительного интервала используется следующая формула:

$P\{\bar{x} - t_{\frac{1+\gamma}{2}}\frac{s}{\sqrt{n-1}} < m < \bar{x} + t_{\frac{1+\gamma}{2}}\frac{s}{\sqrt{n-1}}\} = \gamma$, где $t_{\frac{1+\gamma}{2}}$ находится по таблице 3 (Квантили распределения Стьюдента) с n-1 степенями свободы. В этом случае надо написать выражение для плотности)

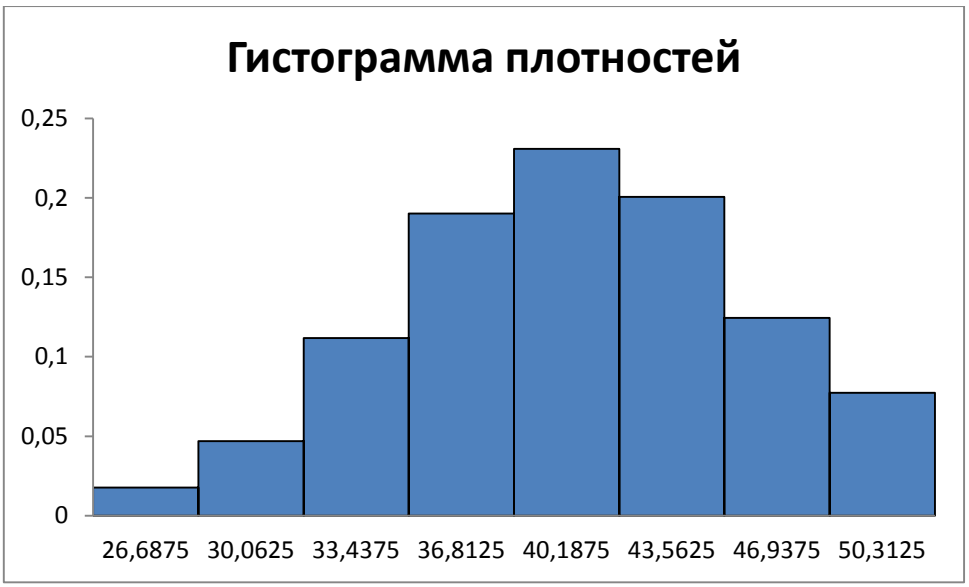
Таблицу 1 (значения нормированной функции Лапласа), таблицу 2 ((квантили нормального распределения), таблицу 3 (Квантили распределения Стьюдента) можно найти в учебниках по математической статистике или в учебном пособии «Математика. Математическая статистика» Камартина Н.М. ГУТ 2013г.

Итого, мы практически закончили решение пунктов 4 и 5. Осталось только построить графики подобранной плотности. Плотности указаны в столбце с названием $p_i = \Phi(b_i) - \Phi(b_{i-1})$ в последней таблице.

График плотностей выглядит так:



А гистограмма вот так:



Подведем итоги.

Мы разбили нашу выборку (150 значений) на 8 интервалов длиной 3,375 от (25,0 ; 28,375) до (48,625 ; 52,0), посчитали абсолютные и относительные частоты попаданий значений в эти интервалы и построили их графики. Затем нашли выборочное среднее (40,4575), дисперсию, исправленные дисперсию и СКО (5,7404), медиану (41,2) и моду (43,192). Еще самый первый из графиков (полигон частот) заставлял нас усомниться в нормальности распределения из-за наличия двух вершин с явным провалом между ними. А проверка гипотезы по критерию Пирсона окончательно убедила, что мы были правы и данные не подчинены нормальному закону распределения.