Задачи по математической статистике

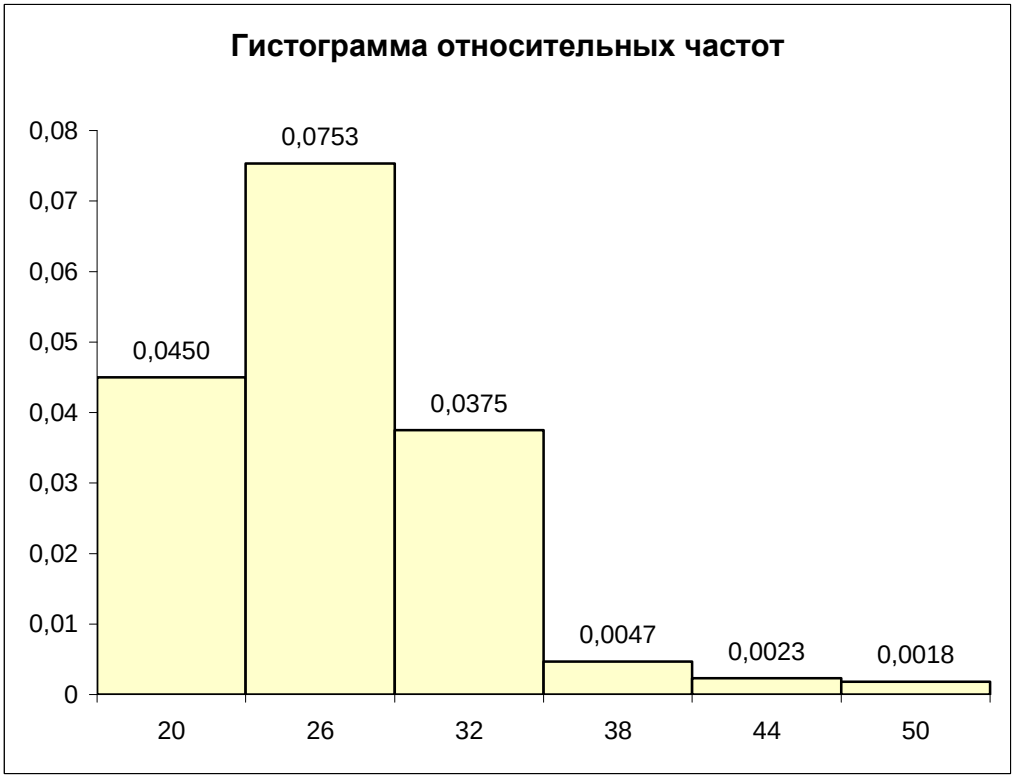
Задача 1. По данным распределения возрастного состава участников революционного движения в России 70-х годов 19-го века была построена следующая таблица

Возраст	17-23	23-29	29-35	35-41	41-47	47-53
Процент на 1000 участников	27	45,2	22,5	2,8	1,4	1,1

По имеющимся данным:
а) построить гистограмму относительных частот;
б) оценить плотность распределения случайной величины с помощью гистограммы;
в) проверить гипотезу о законе распределения с уровнем значимости 0,01.

Решение. Построим гистограмму относительных частот, для чего дополнительно вычислим плотность относительных частот $p_i = \frac{n_i}{nh} = \frac{n_i}{100 \cdot 6}$. Получаем:

начало	конец	частота	p_i
17	23	27	0,045
23	29	45,2	0,07533
29	35	22,5	0,0375
35	41	2,8	0,00467
41	47	1,4	0,00233
47	53	1,1	0,00183
		100	



По виду гистограммы можно предположить, что исследуемая величина имеет нормальный закон распределения.

Найдем точечные оценки параметров распределения. Для этого перейдем к простому вариационному ряду, выбирая в качестве вариант середины интервалов, составим расчетную таблицу:

x_i	n_i	$x_i n_i$	$(x_i-\overline{x})^2 n_i$
20	27	540	1169,71
26	45,2	1175,2	15,3103
32	22,5	720	660,481
38	2,8	106,4	365,038
44	1,4	61,6	424,741
50	1,1	55	603,243
Сумма	**100**	**2658,2**	**3238,53**

Выборочное среднее:

$$\overline{x}=\frac{1}{n}\sum x_i n_i=\frac{1}{100}2658,2=26,582.$$

Выборочная исправленная дисперсия:

$$S^2=\frac{1}{n-1}\sum(\overline{x}-x_i)^2 n_i=\frac{1}{99}3238,53\approx 32,712.$$

Выборочное исправленное среднее квадратическое отклонение: $S=\sqrt{32,712}\approx 5,719$.

Таким образом, предполагаем, что исследуемая величина имеет нормальный закон распределения с параметрами $a=26,582$ и $\sigma=5,719$.

С помощью критерия согласия Пирсона проверим, согласуется ли гипотеза с опытными данными на уровне значимости $\alpha=0,01$.

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 (выравнивающие 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'
17	23	27	-1,675	-0,626	-0,500	-0,234	0,266	26,557
23	29	45,2	-0,626	0,423	-0,234	0,164	0,398	39,820
29	35	22,5	0,423	1,472	0,164	0,429	0,266	26,570
35	41	2,8	1,472	2,521	0,429	0,494	0,065	6,468
41	47	1,4	2,521	3,570	0,494	0,500	0,006	0,567
47	53	1,1	3,570	4,619	0,500	0,500	0,000	0,018
Сумма		**100**						**100,000**

Последние три интервала объединим как малочисленные:

n_i	n_i'	$\frac{(n_i - n_i')^2}{n_i'}$
27	26,557	0,007
45,2	39,820	0,727
22,5	26,570	0,623
5,3	7,054	0,436
	Сумма	**1,794**

Сравним эмпирические и теоретические частоты, используя критерий Пирсона:

$\chi^2 = \sum \frac{(n_i - n_i')^2}{n_i'} = 1,794$.

По таблице критических точек распределения χ^2 по уровню значимости $\alpha = 0,01$ и числу степеней свободы k = 4 - 3 = 1, находим $\chi^2_{\text{кр.}} = 6,6$. Так как $\chi^2_{\text{набл.}} = 1,794 < \chi^2_{\text{кр.}} = 6,6$, то можно принять гипотезу о нормальном распределении данной величины.

Задача 2. Среди 500 молодых семей, живущих с родителями, было зарегистрировано 38 разводов в течение первых трех лет совместной жизни. Построить приближенный доверительный интервал для вероятности развода в таких семьях с уровнем доверия 0,9.

Решение. Доверительный интервал для вероятности развода p найдем по формуле

$w - t_{кр}\sqrt{\frac{w(1-w)}{n}} < p < w + t_{кр}\sqrt{\frac{w(1-w)}{n}}$, где $w = \frac{38}{500} = 0,076$ - выборочная доля, $t_{кр} = \Phi^{-1}(0,9/2) = \Phi^{-1}(0,45) = 1,645$. Подставляем:

$$0,076 - 1,645\sqrt{\frac{0,076(1-0,076)}{500}} < p < 0,076 + 1,645\sqrt{\frac{0,076(1-0,076)}{500}},$$

$0,0565 < p < 0,0955$.

Ответ: от 5,65% до 9,55%.

Задача 3. Известно, что случайная величина X имеет нормальное распределение с неизвестным математическим ожиданием a и известной дисперсией σ^2=144. По выборке объема n=90 вычислено выборочное среднее $x_в$=120. Определить доверительный интервал для неизвестного параметра a, отвечающий заданной надежности γ=0,9.

Решение. Найдем доверительный интервал для математического по формуле:

$$x_B - t\frac{\sigma}{\sqrt{n}} < a < x_B + t\frac{\sigma}{\sqrt{n}},$$

где $n = 90$, $x_B = 120$, $\sigma = \sqrt{144} = 12$, t определяется по доверительной вероятности из таблицы распределения Лапласа $t(0,9) = \Phi^{-1}(0,9/2) = \Phi^{-1}(0,45) = 1,645$. Получаем:

$$120 - 1,645\frac{12}{\sqrt{90}} < a < 120 + 1,645\frac{12}{\sqrt{90}}$$

$117,919 < a < 122,081$.

Ответ: от 117,919 до 122,081.

Задача 4. При обработке исторических материалов профессиональной переписи 1914 года были получены следующие данные: из 329 рабочих фабрики Тамбовской губернии на полевые работы уходило 146 человек, а из 494 рабочих фабрики Ярославской губернии уходило 263 человек. Проверить гипотезу о равенстве вероятности ухода рабочих на полевые работы для двух губерний при уровне значимости 0,1.

Решение. Нулевая гипотеза: $H_0 : p_1 = p_2$. В качестве конкурирующей гипотезы выберем $H_1 : p_1 \neq p_2$.

Вычислим наблюдаемое значение критерия по формуле:

$$U_{набл} = \frac{\frac{m_1}{n_1} - \frac{m_2}{n_2}}{\sqrt{\frac{m_1 + m_2}{n_1 + n_2}\left(1 - \frac{m_1 + m_2}{n_1 + n_2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}},$$

Где

$n_1 = 329,\ n_2 = 494,$

$m_1 = 146,\ m_2 = 263.$

Получаем:

$$U_{набл} = \frac{\frac{146}{329} - \frac{263}{494}}{\sqrt{\frac{146+263}{329+494}\left(1-\frac{146+263}{329+494}\right)\left(\frac{1}{329}+\frac{1}{494}\right)}} \approx -2,75$$

Вычисляем критическое значение из равенства $\Phi\left(u_{кр}\right) = \frac{1-\alpha}{2} = \frac{1-0,1}{2} = 0,45$, откуда $u_{кр} = 1,645$. Так как $|U_{набл}| = 2,75 > 1,645 = u_{кр}$, нулевую гипотезу следует отвергнуть.

Вероятности ухода рабочих на фабриках разных губерний отличаются значимо.

Задача 5. При уровне значимости α=0,1 проверить гипотезу о равенстве дисперсии двух нормально распределенных случайных величин X и Y на основе выборочных данных, приведенных в следующих таблицах:

X	x_i	35	37	39	40	41
	n_i	1	3	5	4	4
Y	y_i	36	37	38	44	42
	m_i	3	5	2	1	4

Решение. Вычислим по данным выборок исправленные выборочные дисперсии S_x^2, S_y^2.

x_i	n_i	$x_i n_i$	$(x_i - \bar{x})^2 n_i$
35	1	35	16,955
37	3	111	13,453
39	5	195	0,069
40	4	160	3,114
41	4	164	14,173
Сумма	**17**	**665**	**47,765**

Выборочная средняя $\bar{x} = \frac{1}{n}\sum x_i n_i = \frac{1}{17}665 \approx 39,118$.

Исправленная (несмещенная) выборочная дисперсия

$$S_x^2 = \frac{1}{n-1}\sum (x_i - \bar{x})^2 n_i = \frac{1}{16}47,765 = 2,985$$

y_i	m_i	$y_i m_i$	$(y_i - \bar{y})^2 m_i$
36	3	108	22,413
37	5	185	15,022
38	2	76	1,076
44	1	44	27,738
42	4	168	42,684
Сумма	**15**	**581**	**108,933**

Выборочная средняя $\bar{y} = \frac{1}{m}\sum y_i m_i = \frac{1}{15}581 \approx 38,733$.

Исправленная (несмещенная) выборочная дисперсия

$$S_y^2 = \frac{1}{m-1}\sum (y_i - \bar{y})^2 m_i = \frac{1}{14}108,933 = 7,781$$

Вычислим наблюдаемое значение критерия $F_{набл} = \frac{S_y^2}{S_x^2} = \frac{7,781}{2,985} \approx 2,606$

Найдем критическую точку при уровне значимости $\alpha / 2 = 0,05$ и числам степеней свободы $k_1 = m - 1 = 14$, $k_2 = n - 1 = 16$, $F_{кр} = 2,373$.

Так как $F_{набл} = 2,606 > 2,373 = F_{кр}$, следует отвергнуть нулевую гипотезу.

Дисперсии различаются значимо.

Задача 6. Результаты наблюдений переменных X и Y приведены в таблице. Найти выборочный коэффициент линейной корреляции и уравнение прямой регрессии Y по X.

X / Y	10	12	14	16	18	20	22
20		2	6	5			4
40	4			5	1		7
60	4	2	8	10		4	
80		3			10	2	5
100	3		4		6	5	

Решение. Построим ряды распределений для X и Y , вычислим их характеристики (выборочное среднее и выборочное среднее квадратическое отклонение).

x_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$
10	11	110	456,2096
12	7	84	137,9952
14	18	252	107,1648
16	20	320	3,872
18	17	306	41,3712
20	11	220	139,4096
22	16	352	494,6176
Сумма	**100**	**1644**	**1380,64**
Среднее		**16,44**	**13,8064**

Выборочная средняя $\bar{x} = \frac{1}{n}\sum x_i n_i = \frac{1}{100}1644 = 16,44$

Выборочная дисперсия $\overline{D}_x = \frac{1}{n}\sum (x_i - \bar{x})^2 n_i = \frac{1}{100}1380,64 = 13,8064$

Выборочное квадратическое отклонение $\sigma_x = \sqrt{\overline{D_x}} = 3,716$

y_i	n_i	$y_i \cdot n_i$	$(y_i - \overline{y})^2 \cdot n_i$
20	17	340	28577,000
40	17	680	7497,000
60	28	1680	28,000
80	20	1600	7220,000
100	18	1800	27378,000
Сумма	**100**	**6100**	**70700,000**
Среднее		**61**	**707,000**

Выборочная средняя $\overline{y} = \frac{1}{n}\sum y_i n_i = \frac{1}{100} 6100 = 61,0$

Выборочная дисперсия $\overline{D}_y = \frac{1}{n}\sum (y_i - \overline{y})^2 n_i = \frac{1}{100} 70700 = 707,0$

Выборочное квадратическое отклонение $\sigma_y = \sqrt{\overline{D_y}} = 26,589$

Коэффициент линейной корреляции вычислим по формуле $r = \frac{\sum n_{xy} x_i y_i - n\overline{x}\,\overline{y}}{n\sigma_x \sigma_y}$.

Найдем сумму $\sum n_{xy} x_i y_i = 100840$. Расчеты в таблице

$Y \setminus X$	10	12	14	16	18	20	22	$n_{xy} x_i$	$n_{xy} x_i y_i$
20	0	2	6	5	0	0	4	276	5520
40	4	0	0	5	1	0	7	292	11680
60	4	2	8	10	0	4	0	416	24960
80	0	3	0	0	10	2	5	366	29280
100	3	0	4	0	6	5	0	294	29400
									100840

Тогда коэффициент корреляции:

$r = \frac{100800 - 100 \cdot 16,44 \cdot 61}{100 \cdot 3,716 \cdot 26,589} \approx 0,056$.

Связь очень слабая, прямая по направлению.

Напишем уравнение регрессии Y на X. Оно имеет вид $\overline{y_x} - \overline{y} = r\frac{\sigma_y}{\sigma_x}(x - \overline{x})$. Подставляем все величины:

$\overline{y_x} - 61 = 0,056\frac{26,589}{3,716}(x - 16,44)$,

$\overline{y_x} = 0,402x + 54,379$.

Задача 7. При уровне значимости α=0,05 методом дисперсионного анализа проверить нулевую гипотезу о влиянии фактора на качество объекта на основании пяти измерений для трех уровней фактора. Данные измерений приводятся в таблице:

Номер измерения	Φ_1	Φ_2	Φ_3
1	34	38	28
2	36	30	24
3	26	34	22
4	25	36	20
5	30	38	23

Решение. Составим дисперсионную таблицу:

	Φ_1		Φ_2		Φ_3		
i	y_{i1}	y_{i1}^2	y_{i2}	y_{i2}^2	y_{i3}	y_{i3}^2	**Сумма**
1	34	1156	38	1444	28	784	
2	36	1296	30	900	24	576	
3	26	676	34	1156	22	484	
4	25	625	36	1296	20	400	
5	30	900	38	1444	23	529	
$T_j = \sum y_{ij}$	**151**		**176**		**117**		**444**
$S_j = \sum y_{ij}^2$		**4653**		**6240**		**2773**	**13666**
T_j^2	**22801**		**30976**		**13689**		**67466**

Найдем общую и факторную суммы квадратов отклонений, учитывая, что число уровней фактора $p=3$, число испытаний на каждом уровне $q=5$.
Получаем:

$$S_{общ} = \sum_{j=1}^{p} S_j - \frac{1}{pq}\left(\sum_{j=1}^{p} T_j\right)^2 = 13666 - \frac{1}{15}444^2 = 523,6$$

$$S_{факт} = \frac{1}{q}\sum_{j=1}^{p} T_j^2 - \frac{1}{pq}\left(\sum_{j=1}^{p} T_j\right)^2 = \frac{1}{5}67466 - \frac{1}{15}444^2 = 350,8$$

Найдем остаточную сумму квадратов отклонений

$$S_{ост} = S_{общ} - S_{факт} = 523,6 - 350,8 = 172,8$$

Найдем дисперсии

$$s_{факт}^2 = \frac{S_{факт}}{p-1} = \frac{350,8}{2} = 175,4$$

$$s_{ост}^2 = \frac{S_{ост}}{p(q-1)} = \frac{172,8}{3\cdot 4} = 14,4$$

Сравним факторную и остаточную дисперсию с помощью критерия Фишера-Снедекора.
Найдем наблюдаемое значение критерия

$$F_{набл} = \frac{s^2_{факт}}{s^2_{ост}} = \frac{175,4}{14,4} = 12,18.$$

По числу степеней свободы $k_1 = 2$, $k_2 = 12$ и по уровню значимости $\alpha = 0,05$ находим критическую точку $F_{крит} = 3,88$. Так как $F_{набл} > F_{крит}$, следует отвергнуть гипотезу, влияние фактора значимо.