МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
Национальный исследовательский ядерный университет «МИФИ»

Институт интеллектуальных кибернетических систем

КАФЕДРА КИБЕРНЕТИКИ

БДЗ

по курсу "Математическая статистика"

студента группы Б22-534

Баранова Александра

Вариант №2

Оценка: ________________

Подпись: ________________

2024 г.

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – B7 (Weight (lbs))

Анализируемый признак 2 – B8 (Height (inches))

Анализируемый признак 3 – B9 (Neck circumference (cm))

а) Привести формулы расчёта выборочных характеристик

Выборочная хар-ка	Формула расчета
Объём выборки	n
Среднее	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Выборочная дисперсия	$D_X^* = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$
Выборочное среднеквадратическое отклонение	$\sigma_X^* = \sqrt{D_X^*} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$
Выборочный коэффициент асимметрии	$\gamma_X^* = \frac{\mu_{3,X}^*}{\left(\sigma_X^*\right)^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{\frac{3}{2}}}$
Выборочный эксцесс	$\varepsilon_X^* = \frac{\mu_{4,X}^*}{\left(\sigma_X^*\right)^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{2}}$

б) Рассчитать выборочные характеристики

Выборочная хар-ка	Признак 1	Признак 2	Признак 3
Среднее	178.92	70.15	37.99
Выборочная дисперсия	860.30	13.36	5.89
Выборочное среднеквадратическое отклонение	29.33	3.66	2.43
Выборочный коэффициент асимметрии	1.20	-5.35	0.55
Выборочный эксцесс	5.14	58.35	2.64

1.2. Группировка и гистограммы частот

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

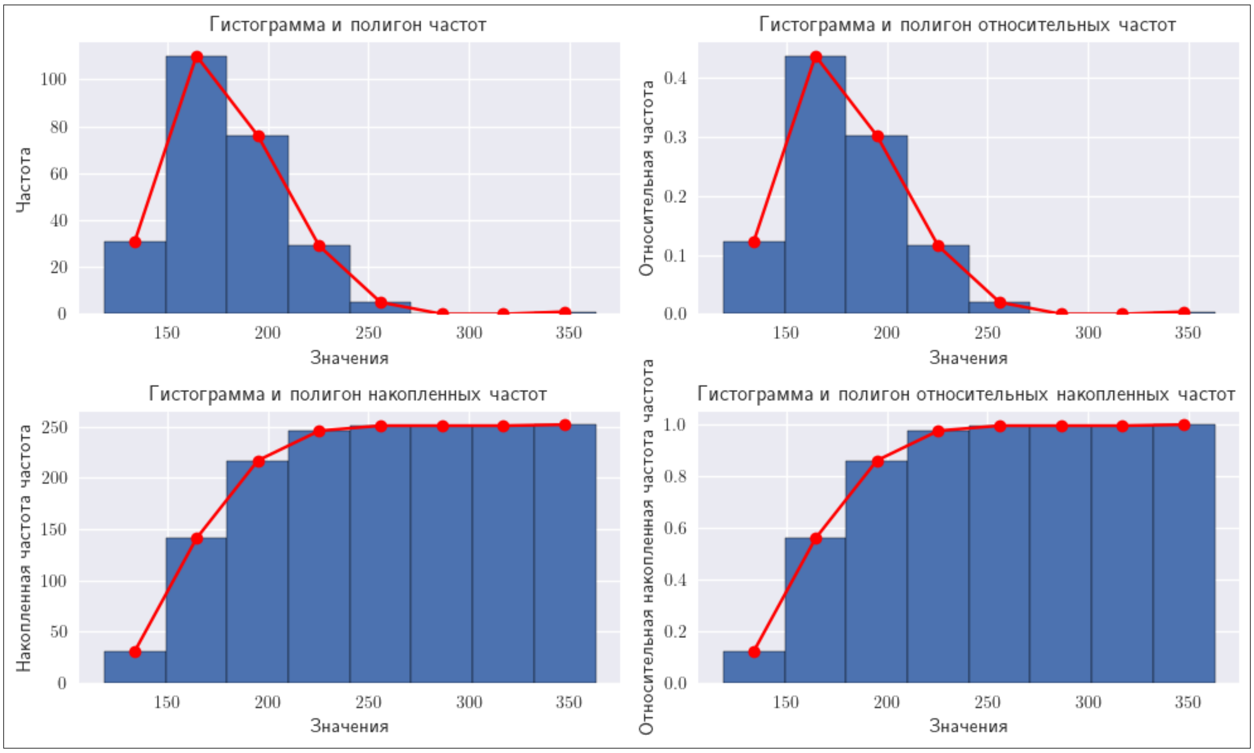
а) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
8	Формула Стерджесса: $k \approx 1 + 1{,}3 \ln n$	от 30.58 до 30.83

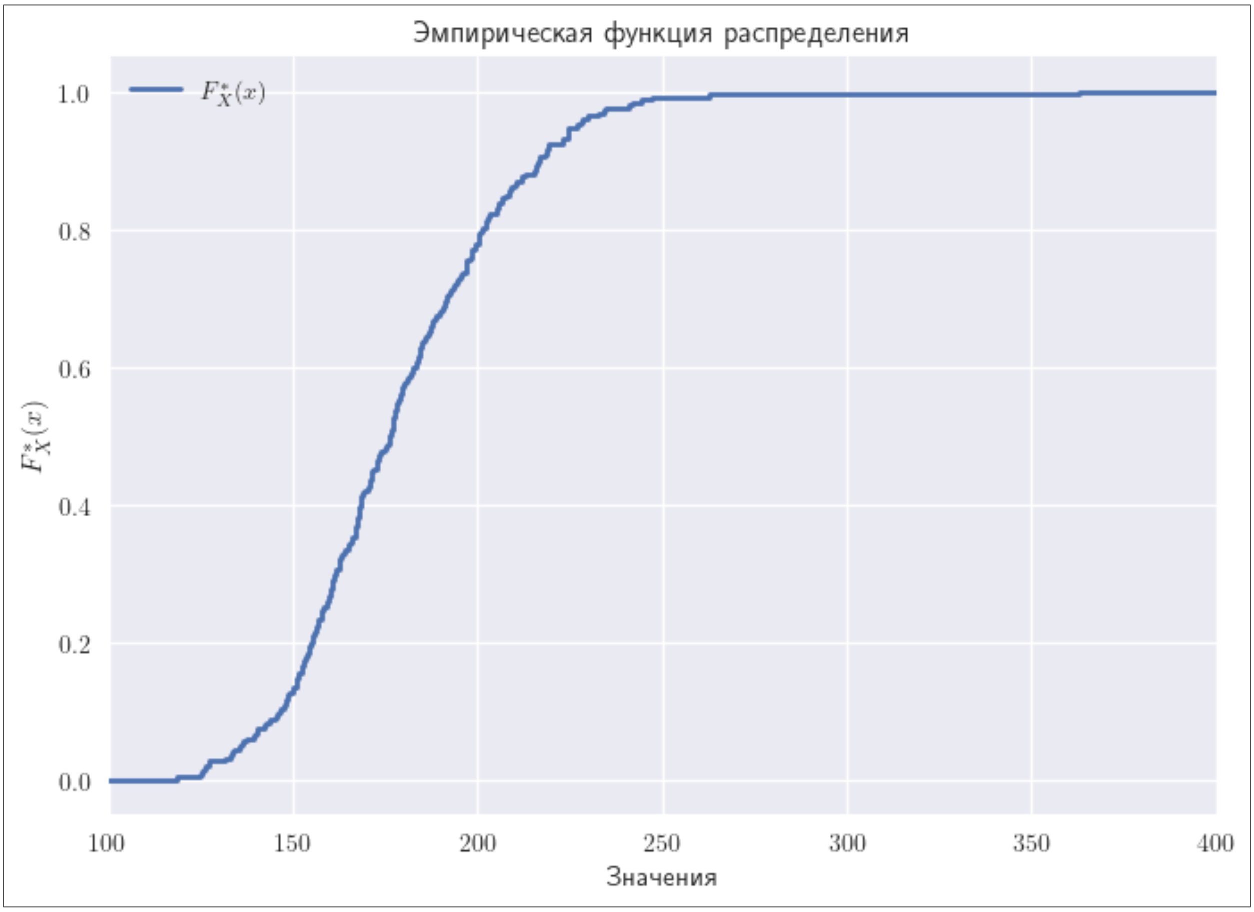
б) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Накопл. частота	Относит. накопл. частота
1	118.25	149.08	31	0.12	31	0.12
2	149.08	179.66	110	0.44	141	0.56
3	179.66	210.24	76	0.3	217	0.86
4	210.24	240.82	29	0.12	246	0.98
5	240.82	271.41	5	0.02	251	1.00
6	271.41	301.99	0	0.00	251	1.00
7	301.99	332.57	0	0.00	251	1.00
8	332.57	363.15	1	0.00	252	1.00

в) Построить гистограммы частот и полигоны частот



г) Построить график эмпирической функции распределения



2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

Оцениваемый параметр – m

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\overline{\mathbf{X}} - \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2},(n-1)}$
Верхняя граница	$\overline{\mathbf{X}} + \frac{S}{\sqrt{n}} \cdot t_{1-\frac{\alpha}{2},(n-1)}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	174.12	175.28	175.87
Верхняя граница	183.73	182.57	181.98

2.2. Доверительные интервалы для дисперсии

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

Оцениваемый параметр – σ^2

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{(n-1) \cdot S^2}{\chi^2_{1-\frac{\alpha}{2},(n-1)}}$
Верхняя граница	$\frac{(n-1) \cdot S^2}{\chi^2_{\frac{\alpha}{2},(n-1)}}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	693.83	730.50	750.27
Верхняя граница	1100.22	1037.24	1006.86

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = 252$

Оцениваемый параметр – $m_1 - m_2$

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$(\overline{\mathbf{X_1}} - \overline{\mathbf{X_2}}) - t_{1-\frac{\alpha}{2},(n_1+n_2-2)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
Верхняя граница	$(\overline{\mathbf{X_1}} - \overline{\mathbf{X_2}}) + t_{1-\frac{\alpha}{2},(n_1+n_2-2)} \cdot S \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
S	$\sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	-9.46	-8.95	-8.69
Верхняя граница	-5.24	-5.75	-6.00

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = 252$

Оцениваемый параметр – $\frac{\sigma_1^2}{\sigma_2^2}$

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{S_1^2}{S_2^2} \cdot F_{\frac{\alpha}{2},(n_1-1,n_2-1)}$
Верхняя граница	$\frac{S_1^2}{S_2^2} \cdot F_{1-\frac{\alpha}{2},(n_1-1,n_2-1)}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	1.63	1.77	1.84
Верхняя граница	3.14	2.90	2.79

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

Статистическая гипотеза – $\begin{array}{l} H_0: \ m = m_0 \\ H': \ m \neq m_0 \end{array}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \dfrac{\overline{X} - m_0}{S/\sqrt{n}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$T(n-1)$
Формулы расчета критических точек	$\pm t_{1-\frac{\alpha}{2},n-1}$
Формула расчета *p-value*	$2\min(F_Z(z_{\text{выб}} \mid H_0),\ 1 - F_Z(z_{\text{выб}} \mid H_0))$

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

m_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
120	0.1	31.83	*0.00*	H_0 отклоняется	$m \neq 120$
180	0.1	-0.58	*0.56*	H_0 принимается	$m = 180$
240	0.1	-32.99	*0.00*	H_0 отклоняется	$m \neq 240$

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

Статистическая гипотеза – $\begin{array}{l} H_0: \ \sigma = \sigma_0 \\ H': \ \sigma \neq \sigma_0 \end{array}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \frac{(n-1)S^2}{\sigma_0^2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$
Формулы расчета критических точек	$\chi^2_{\frac{\alpha}{2},n-1},\ \chi^2_{1-\frac{\alpha}{2},n-1}$
Формула расчета *p-value*	$2\min(F_Z(z_{\text{выб}} \mid H_0),\ 1 - F_Z(z_{\text{выб}} \mid H_0))$

б) Выбрать произвольные значения σ₀ и проверить статистические гипотезы

σ_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
10	0.1	2167.94	*0.00*	H_0 отклоняется	$\sigma \neq 10$
20	0.1	541.99	*0.00*	H_0 отклоняется	$\sigma \neq 10$
30	0.1	240.88	*0.67*	H_0 принимается	$\sigma \neq 10$

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = 252$

Статистическая гипотеза – $\begin{matrix} H_0: & m_1 = m_2 \\ H': & m_1 \neq m_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$Z = \dfrac{\overline{X_1} - \overline{X_2}}{S\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, где $S^2 = \dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$T(n_1 + n_2 - 2)$
Формулы расчета критических точек	$\pm t_{1-\frac{\alpha}{2}, n_1+n_2-2}$
Формула расчета *p-value*	$2\min(F_Z(z_{\text{выб}} \mid H_0),\ 1 - F_Z(z_{\text{выб}} \mid H_0))$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-9.01	*0.00*	H_0 отклоняется	$m_1 \neq m_2$
0.05			H_0 отклоняется	$m_1 \neq m_2$
0.1			H_0 отклоняется	$m_1 \neq m_2$

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = 252$

Статистическая гипотеза – $\begin{matrix} H_0: & \sigma_1 = \sigma_2 \\ H': & \sigma_1 \neq \sigma_2 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение
Формула расчета статистики критерия	$$Z = \frac{S_1^2}{S_2^2}$$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(n_1 - 1,\ n_2 - 1)$
Формулы расчета критических точек	$F_{\frac{\alpha}{2},\ n_1-1,\ n_2-1},\ F_{1-\frac{\alpha}{2},\ n_1-1,\ n_2-1}$
Формула расчета *p-value*	$2\min(F_Z(z_{\text{выб}} \mid H_0),\ 1 - F_Z(z_{\text{выб}} \mid H_0))$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	2.27	*0.00*	H_0 отклоняется	$\sigma_1 \neq \sigma_2$
0.05			H_0 отклоняется	$\sigma_1 \neq \sigma_2$
0.1			H_0 отклоняется	$\sigma_1 \neq \sigma_2$

4. Критерии согласия

Анализируемый признак – B7 (Weight (lbs))

Объём выборки – 252

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное.

Статистическая гипотеза – $\begin{matrix} H_0: & X \sim N \\ H': & X \nsim N \end{matrix}$

а) Указать формулы расчета показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \sum_{i=1}^{k} \frac{\left(n_i - n \cdot p_i\right)^2}{n \cdot p_i}$	k - число интервалов в группированном статистическом ряду. n_i - частота попадания случайной величины X в интервал Δ_i. p_i - вероятность попадания случайной величины X в интервал Δ_i в условиях H_0, то есть, если $\Delta_i = (a_{i-1}, a_i]$, то $p_i = \int_{a_{i-1}}^{a_i} g(x)dx = G(a_i) - G(a_{i-1})$.
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(k - r - 1)$	r - количество оцениваемых параметров у предполагаемого распределения G.
Формула расчета критической точки	$\chi^2_{1-\alpha,\ k-r-1}$	Малые значения Z нам также подходят, поэтому критическая область выбирается правосторонней
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

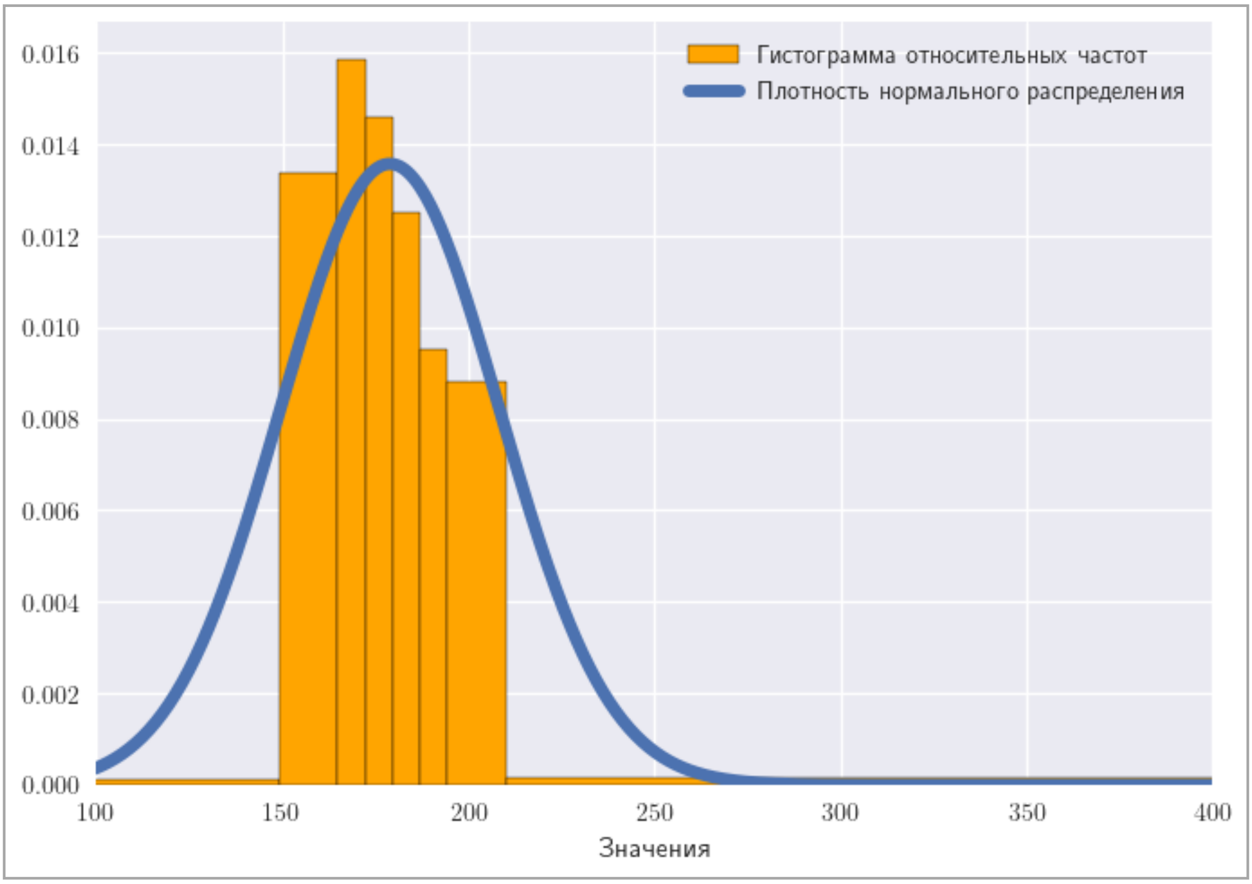
б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
8	$k \approx 1 + 1.3 \ln n$ - формула Стерджесса $n \cdot p_i \gtrsim 5$ - поправка на чувствительность критерия	От 7.30 до ∞

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Вероятность попадания в интервал при условии истинности основной гипотезы
1	$-\infty$	149.08	31	0.12	0.15
2	149.08	164.82	53	0.21	0.16
3	164.82	172.32	30	0.12	0.10
4	172.32	179.66	27	0.11	0.10
5	179.66	186.96	23	0.09	0.10
6	186.96	194.45	18	0.07	0.09
7	194.45	210.24	35	0.14	0.16
8	210.24	$+\infty$	35	0.14	0.14

г) Построить гистограмму относительных частот и функцию плотности теоретического распределения на одном графике



д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	9.07	*0.11*	H_0 принимается	$X \sim N$
0.05			H_0 принимается	$X \sim N$
0.1			H_0 принимается	$X \sim N$

4.2. Проверка гипотезы о нормальности на основе коэффициента асимметрии и эксцесса (критерий Харке-Бера)

Статистическая гипотеза – $\begin{matrix} H_0: & X \sim N \\ H': & X \nsim N \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = Z_1^2 + Z_2^2$, где $Z_1 = \frac{\gamma_X^*}{\sqrt{\frac{6}{n}}}$; $Z_2 = \frac{\varepsilon_X^*}{\sqrt{\frac{24}{n}}}$	γ_X^* - выборочный коэффициент асимметрии. ε_X^* - выборочный эксцесс. n - объём выборки.
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(2)$	
Формула расчета критической точки	$\chi^2_{1-\alpha,\ 2}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	337.89	*0.00*	H_0 отклоняется	$X \nsim N$
0.05			H_0 отклоняется	$X \nsim N$
0.1			H_0 отклоняется	$X \nsim N$

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что критерий χ^2 не отвергает гипотезу о нормальности распределения, но критерий Харке-Бера отвергает её.

Так как для этих данных критерий Харке-Бера более чувствителен, чем критерий χ^2, то можно сделать вывод, что выборка B7 не имеет нормального распределения.

5. Проверка однородности выборок

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = 252$

5.1 Критерий знаков

Статистическая гипотеза – $\begin{array}{l} H_0: \ F_X(x) = F_Y(x) \\ H': \ F_X(x) \neq F_Y(x) \end{array}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \dfrac{K_+ - \frac{n}{2}}{\sqrt{n}/2}$	K_+ - число знаков \`+\` в выборке $z_1, \ldots, z_n = x_1 - y_1, \ldots, x_n - y_n$
Закон распределения статистики критерия при условии истинности основной гипотезы	$N(0,1)$	
Формула расчета критической точки	$\pm u_{1-\frac{\alpha}{2}}$	Двусторонняя критическая область
Формула расчета *p-value*	$2\min(F_Z(z_{\text{выб}} \mid H_0),\ 1 - F_Z(z_{\text{выб}} \mid H_0))$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-12.72	*0.00*	H_0 отклоняется	$F_X(x) \neq F_Y(x)$
0.05			H_0 отклоняется	$F_X(x) \neq F_Y(x)$
0.1			H_0 отклоняется	$F_X(x) \neq F_Y(x)$

5.2. Критерий хи-квадрат

Статистическая гипотеза – $\begin{matrix} H_0: & F_X(x) = F_Y(x) \\ H': & F_X(x) \neq F_Y(x) \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = nm \sum_{i=1}^{k} \frac{1}{n_i + m_i} \left(\frac{n_i}{n} + \frac{m_i}{m} \right)$	k - число интервалов в группированном статистическом ряду. n_i - частота попадания случайной величины X в интервал Δ_i. m_i - частота попадания случайной величины Y в интервал Δ_i. n - объём выборки $x_1, \ldots, x_n$ m - объём выборки $y_1, \ldots, y_m$
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(k-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha,\ k-1}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

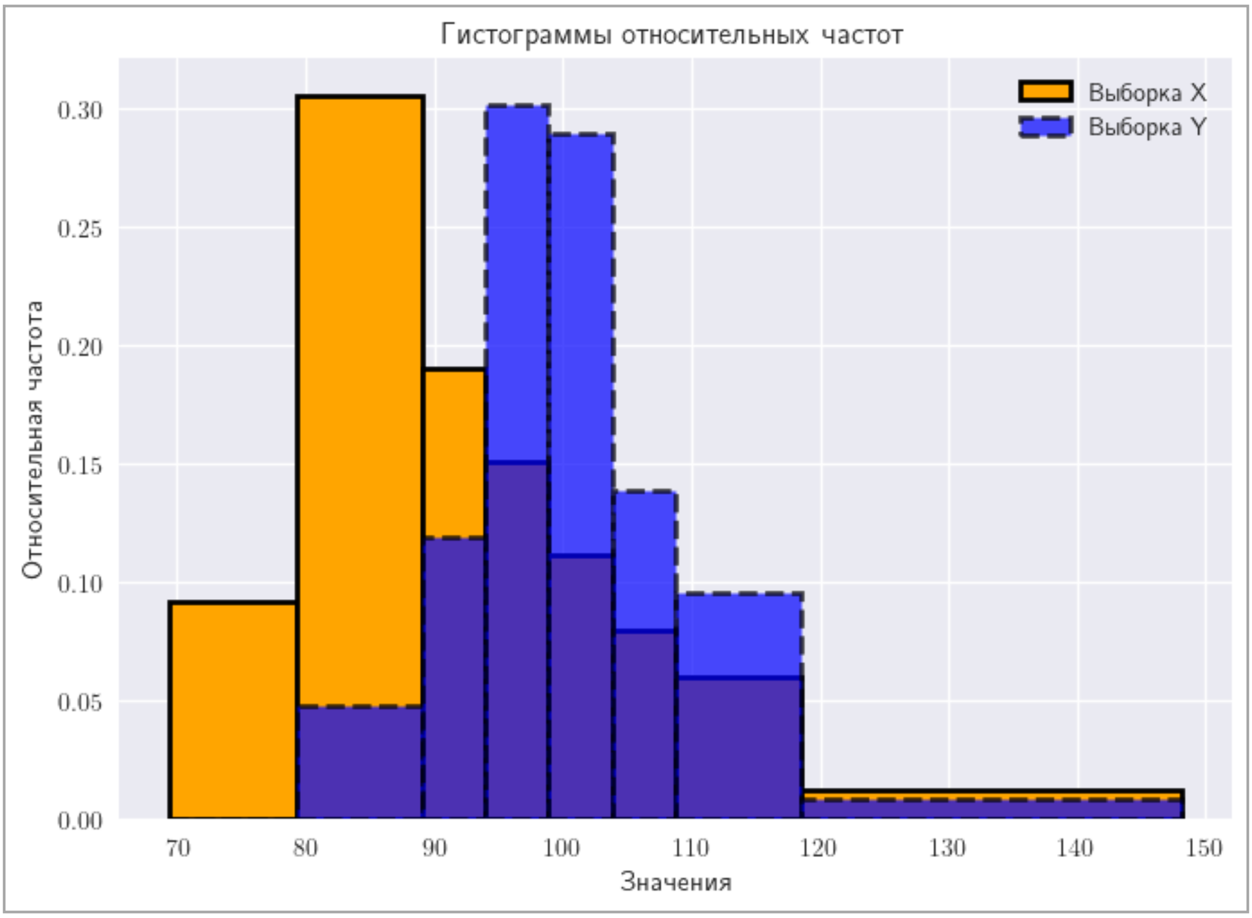
б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
8	$k \approx 1 + 1.3 \ln \max(n, m)$ - формула Стерджесса $n_i + m_i \gtrsim 5$ - поправка на чувствительность критерия	от 4.92 до 29.51

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота признака 1	Частота признака 2	Относит. частота признака 1	Относит. частота признака 2
1	69.4	79.24	23	0	0,09	0
2	79.24	89.08	77	12	0,31	0,05
3	89.08	93.99	48	30	0,19	0,12
4	93.99	98.91	38	76	0,15	0,3
5	98.91	103.83	28	73	0,11	0,29
6	103.83	108.75	20	35	0,08	0,14
7	108.75	118.59	15	24	0,06	0,1
8	118.59	148.10	3	2	0,01	0,01

г) Построить гистограммы относительных частот на одном графике



д) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	113.71	*0.00*	H_0 отклоняется	$F_X(x) \neq F_Y(x)$
0.05			H_0 отклоняется	$F_X(x) \neq F_Y(x)$
0.1			H_0 отклоняется	$F_X(x) \neq F_Y(x)$

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа обнаружено, что выборки В11 и В12 неоднородны.

6. Таблицы сопряжённости

Факторный признак *x* – B3 (Body fat)

Результативный признак *y* – B5 (Sex)

Объёмы выборок – $n_1 = n_2 = n = 252$

Статистическая гипотеза – $H_0: \; F_Y(y \mid_{X=x^{(1)}}) = F_Y(y \mid_{X=x^{(2)}}) = \ldots = F_Y(y \mid_{X=x^{(k_1)}}) = F_Y(y)$
$H': \; \exists i, j : F_Y(y \mid_{X=x^{(i)}}) \neq F_Y(y \mid_{X=x^{(j)}})$

а) Указать формулы расчёта показателей, используемых при проверке статистических гипотез

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \sum_{i=1}^{k_1} \sum_{j=1}^{k_2} \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$	$(x_1, y_1), \ldots, (x_n, y_n)$ - наблюдения случайного вектора (X, Y), где X, Y - случайные величины дискретного типа $x^{(1)}, \ldots, x^{(k_1)}$ - варианты признака X $y^{(1)}, \ldots, y^{(k_2)}$ - варианты признака Y n_{ij} - выборочная частота варианта $(x^{(i)}, y^{(j)})$ в выборке $(x_1, y_1), \ldots, (x_n, y_n)$ m_{ij} - теоретическая частота варианта $(x^{(i)}, y^{(j)})$ в выборке $(x_1, y_1), \ldots, (x_n, y_n)$ при условии истинности H_0
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2((k_1 - 1)(k_2 - 1))$	
Формула расчета критической точки	$\chi^2_{1-\alpha,\ (k_1-1)(k_2-1)}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

б) Построить эмпирическую таблицу сопряжённости

Вариант Y / *Вариант X*	$y^{(1)}$	$y^{(2)}$	$\sum$
$x^{(1)}$	65	55	120
$x^{(2)}$	8	31	39
$x^{(3)}$	20	73	93
$\sum$	93	159	252

в) Построить теоретическую таблицу сопряжённости

Вариант Y / *Вариант X*	$y^{(1)}$	$y^{(2)}$	$\sum$
$x^{(1)}$	44,29	75,71	120
$x^{(2)}$	14,39	24,61	39
$x^{(3)}$	34,32	58,68	93
$\sum$	93	159	252

г) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	29.33	*0.00*	H_0 отклоняется	$\exists i, j : F_Y(y \mid_{X=x^{(i)}}) \neq F_Y(y \mid_{X=x^{(j)}})$
0.05			H_0 отклоняется	$\exists i, j : F_Y(y \mid_{X=x^{(i)}}) \neq F_Y(y \mid_{X=x^{(j)}})$
0.1			H_0 отклоняется	$\exists i, j : F_Y(y \mid_{X=x^{(i)}}) \neq F_Y(y \mid_{X=x^{(j)}})$

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что между факторным признаком B3 и результативным признаком B5 присутствует статистическая связь. Под действием B3 оказывается влияние на распределение B5.

7. Дисперсионный анализ

Факторный признак x – B6 (Town)

Результативный признак y – B1 (Body density determined from underwater weighing)

Число вариантов факторного признака – $k = 4$

Объёмы выборок – $n_1 = n_2 = n = 252$

Статистическая гипотеза – $\begin{array}{l} H_0: \ F_Y(y \mid_{X=x_1}) = F_Y(y \mid_{X=x_2}) = \ldots = F_Y(y \mid_{X=x_k}) = F_Y(y) \\ H': \ \exists i, j: F_Y(y \mid_{X=x_i}) \neq F_Y(y \mid_{X=x_j}) \end{array}$

а) Рассчитать групповые выборочные характеристики

№ п/п	Вариант факторного признака	Объём выборки	Групповые средние	Групповые дисперсии
1	Arlington	73	1.05	0.00
2	Norwood	41	1.05	0.00
3	Revere	82	1.05	0.00
4	Somerville	56	1.06	0.00

б) Привести формулы расчёта показателей вариации, используемых в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$\tilde{D}_{\text{межгр}} = \frac{1}{n}\sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$	$k - 1$	$\frac{n}{k-1}\tilde{D}_{\text{межгр}}$
Остаточные признаки	$\tilde{D}_{\text{внутр}} = \frac{1}{n}\sum_{i=1}^{k} n_i\tilde{\sigma}_i^{\,2}$, где $\tilde{\sigma}_i^{\,2} = \frac{1}{n_i}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i^{\,2})$ — групповая дисперсия	$n - k$	$\frac{n}{n-k}\tilde{D}_{\text{внутр}}$
Все признаки	$\tilde{D}_{\text{общ}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2$	$n - 1$	$\frac{n}{n-1}\tilde{D}_{\text{общ}}$

в) Рассчитать показатели вариации, используемые в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$\tilde{D}_{\text{межгр}} = 0.00$	$k - 1 = 3$	$\frac{n}{k-1}\tilde{D}_{\text{межгр}} = 0.00$
Остаточные признаки	$\tilde{D}_{\text{внутр}} = 0.00$	$n - k = 248$	$\frac{n}{n-k}\tilde{D}_{\text{внутр}} = 0.00$
Все признаки	$\tilde{D}_{\text{общ}} = 0.00$	$n - 1 = 251$	$\frac{n}{n-1}\tilde{D}_{\text{общ}} = 0.00$

г) Проверить правило сложения дисперсий

Показатель	$\tilde{D}_{межгр}$	$\tilde{D}_{внутр}$	$\tilde{D}_{межгр}$	$\tilde{D}_{\text{межгр}} + \tilde{D}_{\text{внутр}}$
Значение	0.00	0.00	0.00	0.00

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Эмпирический коэффициент детерминации	$\tilde{\eta}^2 = \frac{\tilde{D}_{\text{межгр}}}{\tilde{D}_{\text{общ}}}$	0.02
Эмпирическое корреляционное отношение	$\tilde{\eta} = \sqrt{\tilde{\eta}^2} = \sqrt{\frac{\tilde{D}_{\text{межгр}}}{\tilde{D}_{\text{общ}}}}$	0.13

е) Охарактеризовать тип связи между факторным и результативным признаками

По шкале Чеддока наблюдается слабая степень статистической связи между факторным признаком B6 и результативным признаком B1.

ж) Указать формулы расчёта показателей, используемых при проверке статистической гипотезы дисперсионного анализа

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{n-k}{k-1}\frac{\tilde{D}_{\text{межгр}}}{\tilde{D}_{\text{внутр}}}$	k — число групп n — объём выборки $y_1, \ldots, y_n$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(k-1, n-k)$	
Формула расчета критической точки	$F_{1-\alpha, k-1, n-k}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

з) Проверить статистическую гипотезу дисперсионного анализа

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	1.34	*0.26*	H_0 принимается	$F_Y(y \mid_{X=x_1}) = F_Y(y \mid_{X=x_2}) = \ldots = F_Y(y \mid_{X=x_k}) = F_Y(y)$
0.05			H_0 принимается	$F_Y(y \mid_{X=x_1}) = F_Y(y \mid_{X=x_2}) = \ldots = F_Y(y \mid_{X=x_k}) = F_Y(y)$
0.1			H_0 принимается	$F_Y(y \mid_{X=x_1}) = F_Y(y \mid_{X=x_2}) = \ldots = F_Y(y \mid_{X=x_k}) = F_Y(y)$

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что между факторным признаком B6 и результативным признаком B1 отсутствует статистическая связь. Под действием B6 не оказывается влияние на распределение B1.

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – B11 (Abdomen circumference (cm))

Анализируемый признак 2 – B12 (Hip circumference (cm))

Объёмы выборок – $n_1 = n_2 = n = 252$

a) Рассчитать точечные оценки коэффициентов корреляции

	Формула расчета	Значение
Линейный коэффициент корреляции	$\tilde{\rho}_{XY} = \dfrac{c\tilde{o}v(X,Y)}{\tilde{\sigma}_X \cdot \tilde{\sigma}_Y} = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$	0.87
Ранговый коэффициент корреляции по Спирмену	$\tilde{\rho}_{XY}^{(sp)} = \tilde{\rho}_{RS}$, где R и S — ранги для выборок X и Y соответственно. Можно показать, что $\tilde{\rho}_{XY}^{(sp)} = 1 - \dfrac{6S}{n(n^2-1)}$, где $S = \sum_{i=1}^{n}(r_i - s_i)^2$	0.85
Ранговый коэффициент корреляции по Кендаллу	$\tilde{\tau}_{XY} = \dfrac{N_+ - N_-}{n(n-1)/2}$, где N_+, N_- — количество пар точек (x_i, y_i) таких, что $(x_i - x_j)(y_i - y_j) > 0$, $(x_i - x_j)(y_i - y_j) < 0$ соответственно. Другой вариант расчёта: $\tilde{\tau}_{XY} = \dfrac{4R}{n(n-1)} - 1$, где $R = \sum_{i=1}^{n}\sum_{j=i+1}^{n} 1[s_j > s_i]$ — число инверсий в выборке $(r_1, s_1), \ldots, (r_n, s_n)$, предварительно отсортированной по возрастанию r_i.	0.66

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

Для небольших объёмов выборок: $n < 500$

Граница доверительного интервала	Формула расчета
Нижняя граница	$th\left(\frac{1}{2}\ln\frac{1+\tilde{\rho}}{1-\tilde{\rho}} + \frac{\tilde{\rho}}{2(n-1)} - \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right)$
Верхняя граница	$th\left(\frac{1}{2}\ln\frac{1+\tilde{\rho}}{1-\tilde{\rho}} + \frac{\tilde{\rho}}{2(n-1)} + \frac{u_{1-\frac{\alpha}{2}}}{\sqrt{n-3}}\right)$

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

Граница доверительного интервала	α = 0.01	α = 0.05	α = 0.1
Нижняя граница	0.83	0.84	0.85
Верхняя граница	0.91	0.90	0.90

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициентов корреляции

Статистическая гипотеза	Формула расчета статистики критерия	Закон распределения статистики критерия при условии истинности основной гипотезы
$H_0:\ \rho_{XY} = 0$ $H':\ \rho_{XY} \neq 0$	$Z = \frac{\tilde{\rho}_{XY}\cdot\sqrt{n-2}}{\sqrt{1-\tilde{\rho}_{XY}^2}}$	$T(n-2)$
$H_0:\ \rho_{XY}^{(sp)} = 0$ $H':\ \rho_{XY}^{(sp)} \neq 0$	$Z = \frac{\tilde{\rho}_{XY}^{(sp)}\cdot\sqrt{n-2}}{\sqrt{1-\left(\tilde{\rho}_{XY}^{(sp)}\right)^2}}$	$T(n-2)$
$H_0:\ \tau_{XY} = 0$ $H':\ \tau_{XY} \neq 0$	$\sqrt{\frac{9n(n+1)}{2(2n+5)}}\cdot\tilde{\tau}_{XY}$	$N(0,1)$

д) Проверить значимость коэффициентов корреляции

Статистическая гипотеза	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
$H_0: \ \rho_{XY} = 0$ $H': \ \rho_{XY} \neq 0$	0.1	28.45	*0.00*	H_0 отклоняется	$\rho_{XY} \neq 0$
$H_0: \ \rho_{XY}^{(sp)} = 0$ $H': \ \rho_{XY}^{(sp)} \neq 0$	0.1	25.00	*0.00*	H_0 отклоняется	$\rho_{XY}^{(sp)} \neq 0$
$H_0: \ \tau_{XY} = 0$ $H': \ \tau_{XY} \neq 0$	0.1	15.71	*0.00*	H_0 отклоняется	$\tau_{XY} \neq 0$

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – B7 (Weight (lbs))

Анализируемый признак 2 – B8 (Height (inches))

Анализируемый признак 3 – B9 (Neck circumference (cm))

Объёмы выборок – $n_1 = n_2 = n_3 = n = 252$

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

Признак \ Признак	B7	B8	B9
B7	1.00	0.37	0.62
B8	0.37	1.00	0.22
B9	0.62	0.22	1.00

б) Рассчитать матрицу значений p-value для ранговых коэффициентов корреляции по Кендаллу

Статистическая гипотеза: $\begin{matrix} H_0: \ \tau = 0 \\ H': \ \tau \neq 0 \end{matrix}$

Признак \ Признак	B7	B8	B9
B7	—	0.00	0.00
B8	0.00	—	0.00
B9	0.00	0.00	—

в) Рассчитать точечную оценку коэффициента конкордации

	Формула расчета	Значение
Коэффициент конкордации	$\tilde{W} = \frac{12}{k^2(n^3 - n)} \cdot \sum_{i=1}^{n} \left(\sum_{j=1}^{k} r_{ij} - \frac{k(n+1)}{2} \right)^2$, где r_{ij} — ранг i-ого объекта в j-ой выборке. k — количество выборок. n —объём выборок.	0.70

г) Указать формулы расчёта показателей, используемых при проверке значимости коэффициента конкордации

Статистический критерий: $\begin{array}{l} H_0: \ W = 0 \\ H': \ W \neq 0 \end{array}$

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = n(k-1)\tilde{W}$	$\tilde{W}$ — точечная оценка коэффициента конкордации. k — количество выборок. n — объём каждой выборки.
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha,\, n-1}$	Правосторонняя критическая область

Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

д) Проверить значимость коэффициента конкордации

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	351.65	*0.00*	H_0 отклоняется	$W \neq 0$
0.05			H_0 отклоняется	$W \neq 0$
0.1			H_0 отклоняется	$W \neq 0$

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что между признаками B11 и B12 существует сильная положительная, в первую очередь линейная, связь. Между признаками B7, B8 и B9 также наблюдается положительная монотонная корреляционная связь. B7 и B9 имеют наиболее сильную положительную монотонную связь, а B8 с остальными имеет более слабую связь.

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак x – B1 (Body density determined from underwater weighing)

Результативный признак y – B2 (Percent body fat from Siri's equation)

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x$

9.1.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
β_0	$\overline{y} - \tilde{\rho}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X} \cdot \overline{x}$	477.65
β_1	$\tilde{\rho}_{XY} \cdot \frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X}$	-434.36

б) Записать точечную оценку уравнения регрессии

$$\tilde{f}(x) = \tilde{\beta}_0 + \tilde{\beta}_1 x = 477.65 - 434.36x$$

в) Привести формулы расчёта показателей вариации, используемых в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Регрессия	$\tilde{D}_{Y\|X} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{f}(x_i) - \overline{y})^2$	$k - 1$	$\frac{n}{k-1}\tilde{D}_{Y\|X}$
Остаточные признаки	$\tilde{D}_{Y\text{ ост}} = \frac{1}{n}\sum_{i=1}^{n}(\tilde{f}(x_i) - y_i)^2$	$n - k$	$\frac{n}{n-k}\tilde{D}_{Y\text{ ост}}$
Все признаки	$\tilde{D}_{Y\text{ общ}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2$	$n - 1$	$\frac{n}{n-1}\tilde{D}_{Y\text{ общ}}$

k — число оцениваемых параметров функции регрессии $f(x)$.

г) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Регрессия	$\tilde{D}_{Y\|X} = 68.06$	$k - 1 = 1$	$\frac{n}{k-1}\tilde{D}_{Y\|X} = 17152.07$
Остаточные признаки	$\tilde{D}_{Y\ \text{ост}} = 1.69$	$n - k = 250$	$\frac{n}{n-k}\tilde{D}_{Y\ \text{ост}} = 1.71$
Все признаки	$\tilde{D}_{Y\ \text{общ}} = 69.76$	$n - 1 = 251$	$\frac{n}{n-1}\tilde{D}_{Y\ \text{общ}} = 70.04$

д) Проверить правило сложения дисперсий

Показатель	$\tilde{D}_{Y\|X}$	$\tilde{D}_{Y\ ост}$	$\tilde{D}_{Y\ общ}$	$\tilde{D}_{Y\|X} + \tilde{D}_{Y\ ост}$
Значение	68.06	1.69	69.76	69.76

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\tilde{R}^2 = \frac{\tilde{D}_{Y\|X}}{\tilde{D}_{Y\ \text{общ}}}$	0.98
Корреляционное отношение	$\tilde{R} = \sqrt{\tilde{R}^2} = \sqrt{\frac{\tilde{D}_{Y\|X}}{\tilde{D}_{Y\ \text{общ}}}}$	0.99

ж) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Наблюдается очень сильная (почти функциональная) корреляционная связь между факторным признаком B1 и результативным признаком B2.

9.1.2. Интервальные оценки линейной регрессионной модели

а) Привести формулы расчёта доверительных интервалов для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	Формула расчета
β0	Нижняя граница	$\tilde{\beta}_0 - t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 \tilde{D}_X}}$
	Верхняя граница	$\tilde{\beta}_0 + t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 \tilde{D}_X}}$
β1	Нижняя граница	$\tilde{\beta}_1 - t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{1}{n \tilde{D}_X}}$
	Верхняя граница	$\tilde{\beta}_1 + t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{1}{n \tilde{D}_X}}$

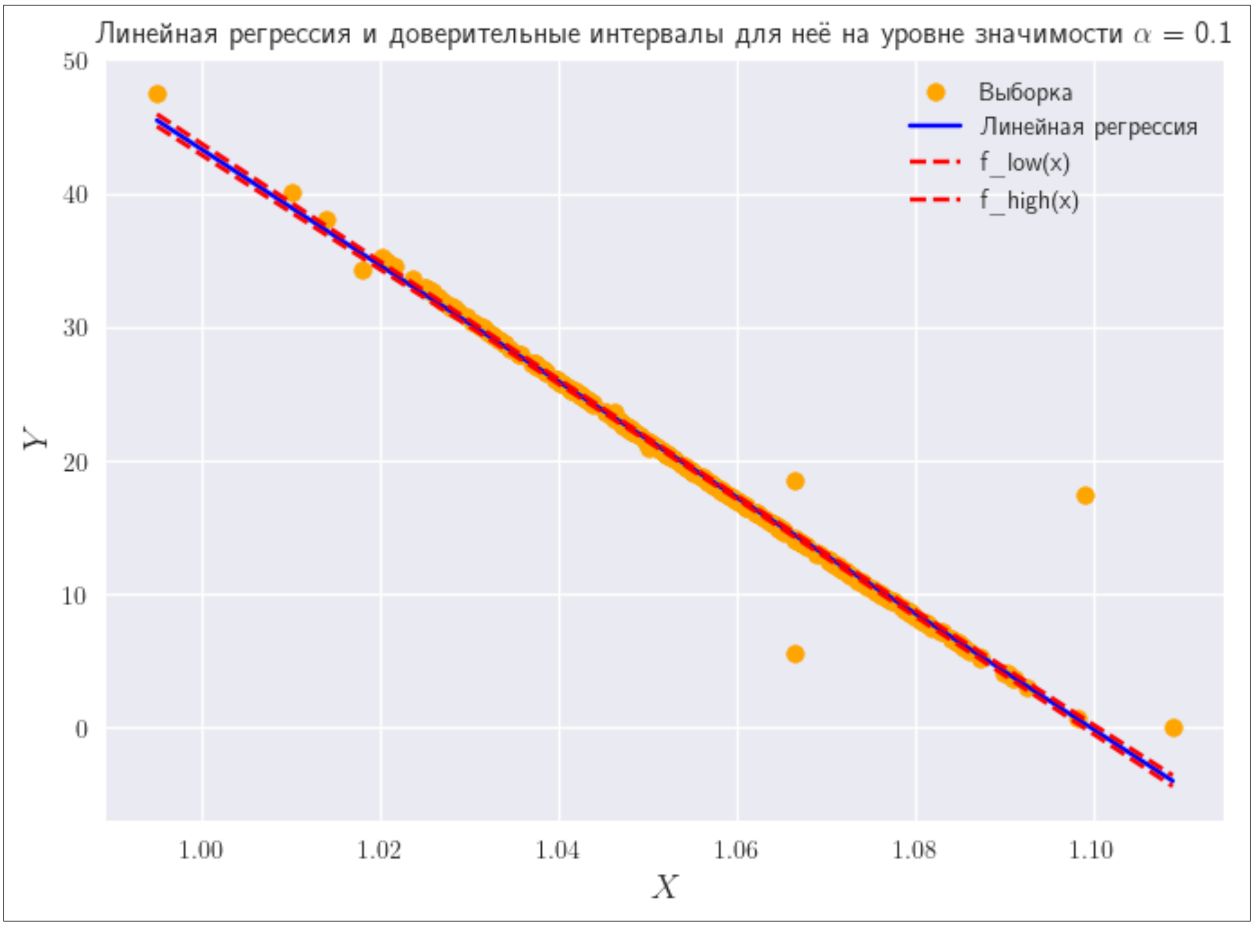
б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	α = 0.01	α = 0.05	α = 0.1
β0	Нижняя граница	465.77	468.65	470.1
	Верхняя граница	489.53	486.66	485.2
β1	Нижняя граница	-445.61	-442.9	-441.52
	Верхняя граница	-423.11	-425.82	-427.20

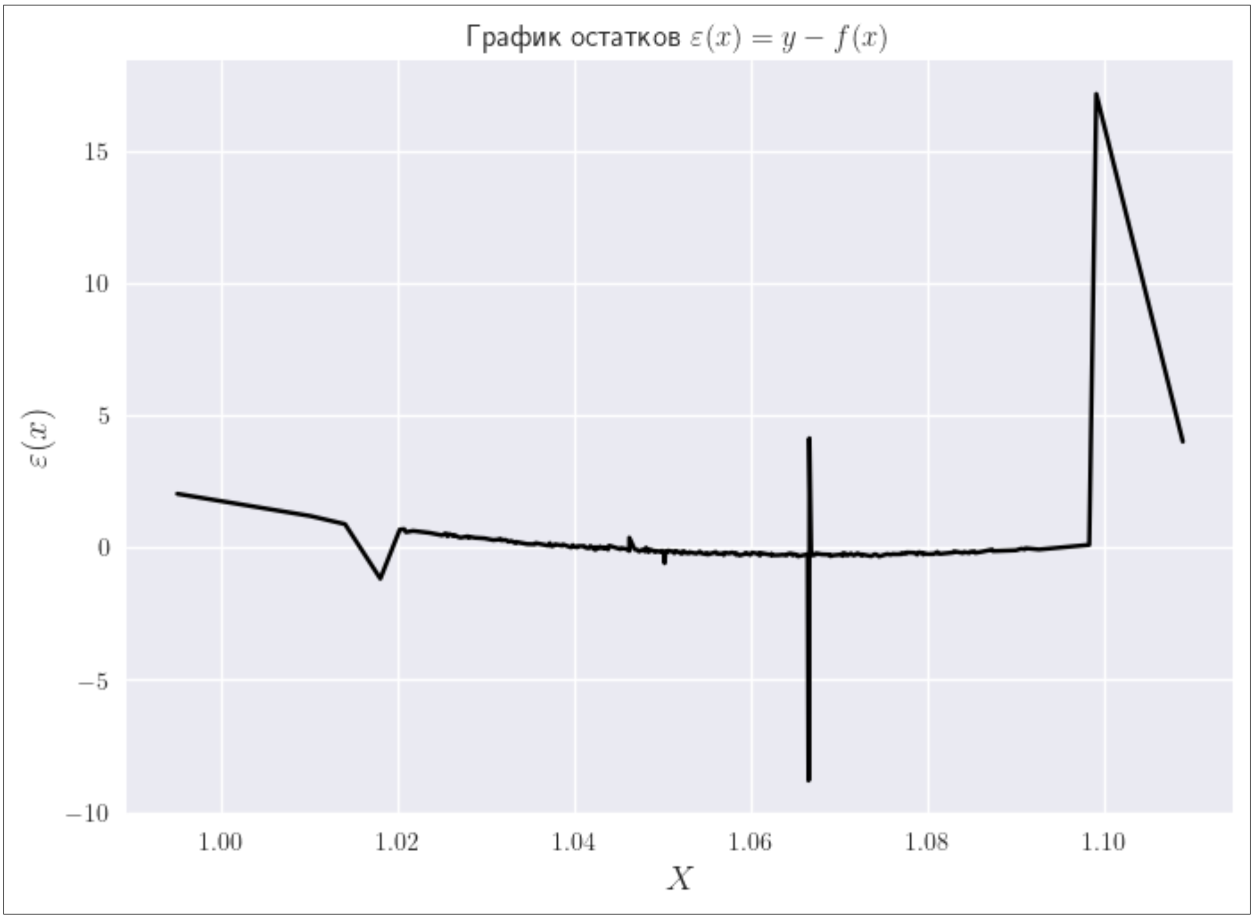
в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$(\tilde{\beta}_0 + \tilde{\beta}_1 x) - t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\,\text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{n\tilde{D}_X}}$
Верхняя граница $f_{high}(x)$	$(\tilde{\beta}_0 + \tilde{\beta}_1 x) + t_{1-\frac{\alpha}{2},\,(n-2)} \cdot \sqrt{\tilde{D}_{Y\,\text{ост}}^{\text{несмещ}}} \cdot \sqrt{\frac{1}{n} + \frac{(x-\overline{x})^2}{n\tilde{D}_X}}$

г) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости α = 0.1



д) Построить график остатков $\varepsilon(x) = y - f(x)$



9.1.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $\begin{matrix} H_0: & \beta_1 = 0 \\ H': & \beta_1 \neq 0 \end{matrix}$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \dfrac{\tilde{D}_{Y\|X}}{\tilde{D}_{Y\ \text{ост}}/(n-2)}$	n — объём выборки $(x_1, y_1), \dots, (x_n, y_n)$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(1,\ n-2)$	
Формула расчета критической точки	$F_{1-\alpha,\ 1,\ n-2}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	10044.03	*0.00*	H_0 отклоняется	$\beta_1 \neq 0$
0.05			H_0 отклоняется	$\beta_1 \neq 0$
0.1			H_0 отклоняется	$\beta_1 \neq 0$

9.2 Линейная регрессионная модель общего вида

Факторный признак *x* – B1 (Body density determined from underwater weighing)

Результативный признак *y* – B2 (Percent body fat from Siri's equation)

Уравнение регрессии – квадратичное по *x*: $f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$

9.2.1. Точечные оценки линейной регрессионной модели

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
$\tilde{\beta}_{\downarrow} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix}$	$\tilde{\beta}_{\downarrow} = (F^T F)^{-1} F^T y_{\downarrow}$, где F — регрессионная матрица, $y_{\downarrow}$ — вектор значений результативного признака.	$\begin{pmatrix} 1644.23 \\ -2645.80 \\ 1047.70 \end{pmatrix}$

б) Записать точечную оценку уравнения регрессии

$$\tilde{f}(x) = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 x^2 = 1644.23 - 2645.80x + 1047.70x^2$$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Регрессия	$\tilde{D}_{Y\|X} = 68.30$	$k - 1 = 2$	$\frac{n}{k-1}\tilde{D}_{Y\|X} = 8606.14$
Остаточные признаки	$\tilde{D}_{Y\,\text{ост}} = 1.46$	$n - k = 249$	$\frac{n}{n-k}\tilde{D}_{Y\,\text{ост}} = 1.47$
Все признаки	$\tilde{D}_{Y\,\text{общ}} = 69.76$	$n - 1 = 251$	$\frac{n}{n-1}\tilde{D}_{Y\,\text{общ}} = 70.04$

k — число оцениваемых параметров функции регрессии $f(x)$.

г) Проверить правило сложения дисперсий

Показатель	$\tilde{D}_{Y\|X}$	$\tilde{D}_{Y\ ост}$	$\tilde{D}_{Y\ общ}$	$\tilde{D}_{Y\|X} + \tilde{D}_{Y\ ост}$
Значение	68.30	1.46	69.76	69.76

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\tilde{R}^2 = \dfrac{\tilde{D}_{Y\|X}}{\tilde{D}_{Y\ \text{общ}}}$	0.98
Корреляционное отношение	$\tilde{R} = \sqrt{\tilde{R}^2} = \sqrt{\dfrac{\tilde{D}_{Y\|X}}{\tilde{D}_{Y\ \text{общ}}}}$	0.99

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

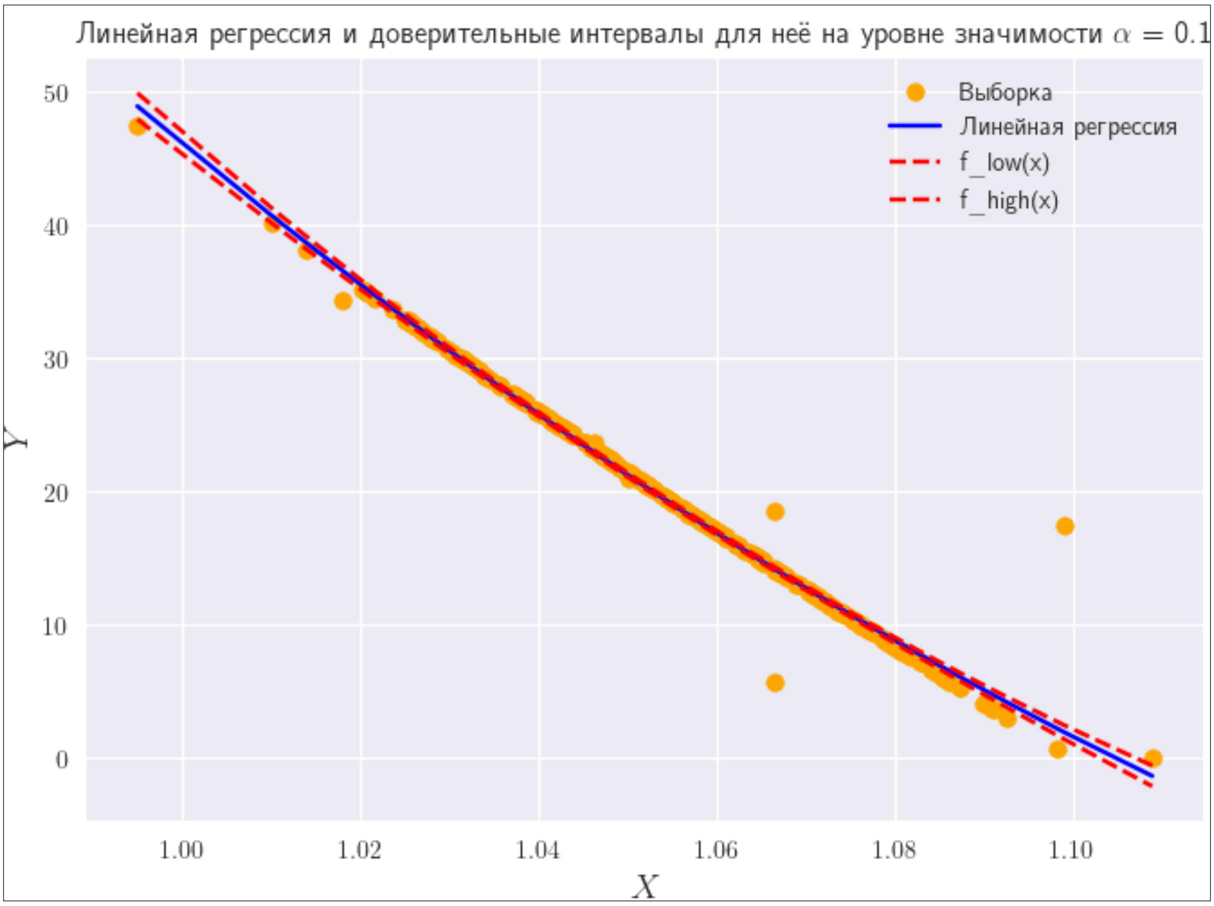
Наблюдается очень сильная (почти функциональная) корреляционная связь между факторным признаком B1 и результативным признаком B2.

9.2.2. Интервальные оценки линейной регрессионной модели

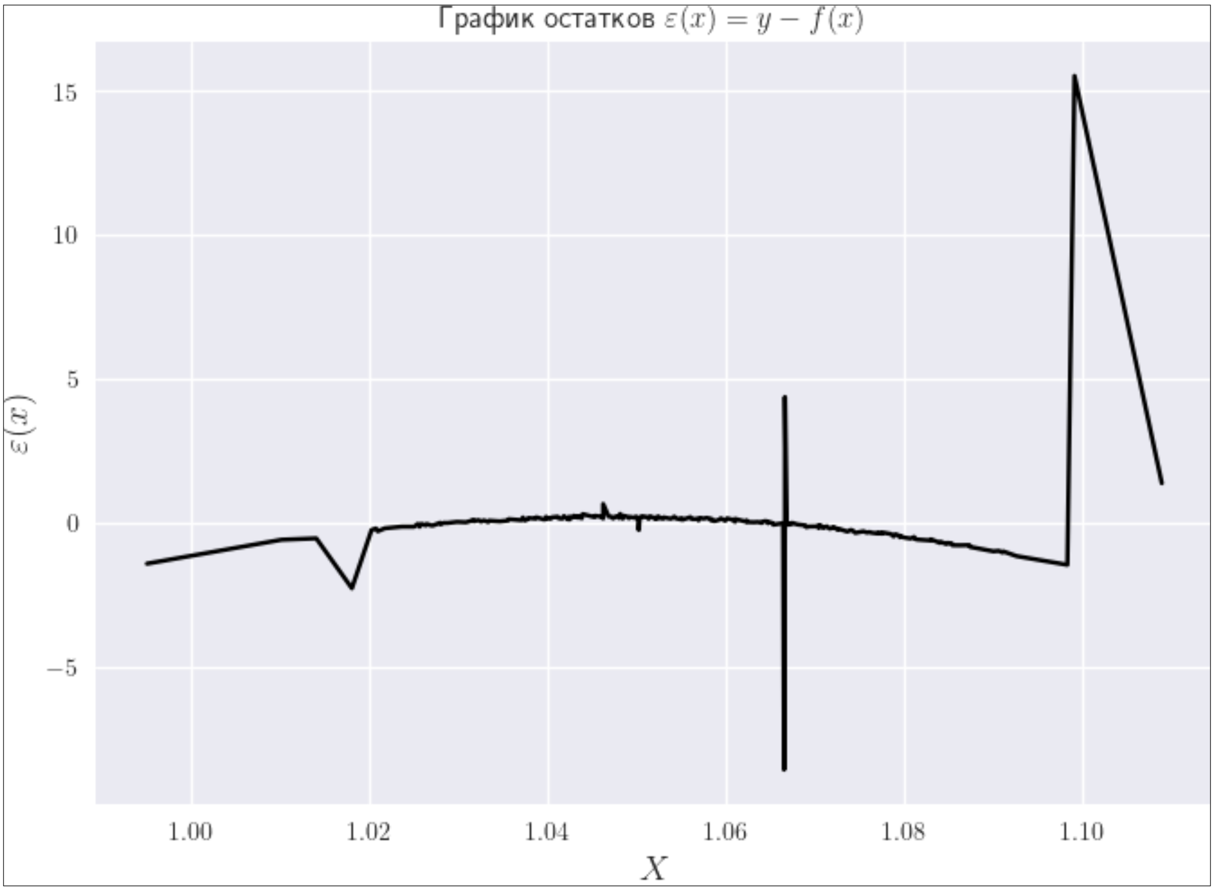
а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$\tilde{f}(x) - t_{1-\frac{\alpha}{2},\ (n-k)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\vec{\varphi}(x)\left(F^T F\right)^{-1} \varphi_{\downarrow}(x)}$
Верхняя граница $f_{high}(x)$	$\tilde{f}(x) + t_{1-\frac{\alpha}{2},\ (n-k)} \cdot \sqrt{\tilde{D}_{Y\ \text{ост}}^{\text{несмещ}}} \cdot \sqrt{\vec{\varphi}(x)\left(F^T F\right)^{-1} \varphi_{\downarrow}(x)}$

б) Построить диаграмму рассеяния признаков x и y. Нанести на диаграмму функцию регрессии f(x), а также нижние и верхние границы линии регрессии $f_{low}(x)$ и $f_{high}(x)$ на уровне значимости $\alpha = 0.1$



в) Построить график остатков $\varepsilon(x) = y - f(x)$



9.2.3. Проверка значимости линейной регрессионной модели

Статистическая гипотеза – $\begin{array}{l} H_0: \ \beta_1 = \beta_2 = 0 \\ H': \ \beta_1^2 + \beta_2^2 > 0 \end{array}$

а) Указать формулы расчёта показателей, используемых при проверке значимости линейной регрессионной модели

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \dfrac{\tilde{D}_{Y\mid X}^{\text{несмещ}}}{\tilde{D}_{Y\text{ ост}}^{\text{несмещ}}}$	n — объём выборки $(x_1, y_1), \ldots, (x_n, y_n)$
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(k-1,\ n-2)$	
Формула расчета критической точки	$F_{1-\alpha,\ k-1,\ n-2}$	Правосторонняя критическая область
Формула расчета *p-value*	$1 - F_Z(z_{\text{выб}} \mid H_0)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	5843.56	*0.00*	H_0 отклоняется	$\beta_1^2 + \beta_2^2 > 0$
0.05			H_0 отклоняется	$\beta_1^2 + \beta_2^2 > 0$
0.1			H_0 отклоняется	$\beta_1^2 + \beta_2^2 > 0$

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – B1 (Body density determined from underwater weighing)

Факторный признак 2 x_2 – B4 (Age (years))

Результативный признак y – B2 (Percent body fat from Siri's equation)

Уравнение регрессии – $f(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

а) Рассчитать точечные оценки параметров линейной регрессионной модели

Параметр	Формула расчета	Значение
$\tilde{\beta}_{\downarrow} = \begin{pmatrix} \tilde{\beta}_0 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{pmatrix}$	$\tilde{\beta}_{\downarrow} = (F^T F)^{-1} F^T y_{\downarrow}$, где F — регрессионная матрица, $y_{\downarrow}$ — вектор значений результативного признака.	$\begin{pmatrix} 474.69 \\ -432.08 \\ 0.01 \end{pmatrix}$

б) Записать точечную оценку уравнения регрессии

$$\tilde{f}(x) = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 = 474.69 - 432.08x_1 + 0.01x_2$$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Регрессия	$\tilde{D}_{Y\mid X_1,X_2} = 68.09$	$k - 1 = 2$	$\frac{n}{k-1}\tilde{D}_{Y\mid X_1,X_2} = 8578.86$
Остаточные признаки	$\tilde{D}_{Y\text{ ост}} = 1.67$	$n - k = 249$	$\frac{n}{n-k}\tilde{D}_{Y\text{ ост}} = 1.69$
Все признаки	$\tilde{D}_{Y\text{ общ}} = 69.75$	$n - 1 = 251$	$\frac{n}{n-1}\tilde{D}_{Y\text{ общ}} = 70.04$

k — число оцениваемых параметров функции регрессии $f(x)$.

г) Проверить правило сложения дисперсий

Показатель	$\tilde{D}_{Y\mid X_1,X_2}$	$\tilde{D}_{Y\ ост}$	$\tilde{D}_{Y\ общ}$	$\tilde{D}_{Y\mid X_1,X_2} + \tilde{D}_{Y\ ост}$
Значение	68.09	1.67	69.76	69.76

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\tilde{R}^2 = \frac{\tilde{D}_{Y\mid X_1,X_2}}{\tilde{D}_{Y\text{ общ}}}$	0.98
Корреляционное отношение	$\tilde{R} = \sqrt{\tilde{R}^2} = \sqrt{\frac{\tilde{D}_{Y\mid X_1,X_2}}{\tilde{D}_{Y\text{ общ}}}}$	0.99

е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

Наблюдается очень сильная (почти функциональная) корреляционная связь между факторными признаками B1 и B4 и результативным признаком B2.

9.4. Выводы

а) Сводная таблица показателей вариации для различных регрессионных моделей

Источник вариации	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Регрессия	$\tilde{D}_{Y\mid X} = 68.06$	$\tilde{D}_{Y\mid X} = 68.30$	$\tilde{D}_{Y\mid X_1,X_2} = 68.09$
Остаточные признаки	$\tilde{D}_{Y\text{ ост}} = 1.69$	$\tilde{D}_{Y\text{ ост}} = 1.46$	$\tilde{D}_{Y\text{ ост}} = 1.67$
Все признаки	$\tilde{D}_{Y\text{ общ}} = 69.76$	$\tilde{D}_{Y\text{ общ}} = 69.76$	$\tilde{D}_{Y\text{ общ}} = 69.76$

б) Сводная таблица свойств различных регрессионных моделей

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	Точная	Точная	Точная
Значимость	Значимая	Значимая	Значимая
Адекватность	Адекватная	Неадекватная	Неадекватная
Степень тесноты связи	Очень тесная	Очень тесная	Очень тесная

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что результативный признак B2 имеет очень тесную связь с факторным признаком B1. При добавлении в модель факторного признака B4 связь между результативным признаком и факторными признаками не ухудшилась.