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 Б20-524

Гаврилова Максима Сергеевича

Вариант № 3

Оценка: _____

Подпись: _____

2022 г.

ОТЧЕТ № 1

по теме «Проверка статистических гипотез»

Вариант №_____

ФИО студента _____ группа ___

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
4.1		Хи-квадрат		
4.2		Харке-Бера		
5.1		знаков		
5.2		Хи-квадрат		

Выводы:

В результате проведённого в п.4 статистического анализа обнаружено, что
В результате проведённого в п.5 статистического анализа обнаружено, что

ОТЧЕТ № 2

по теме «Анализ статистических взаимосвязей»

Вариант №_____

ФИО студента _____ группа _____

Оценка: _____ Подпись:_____

Результаты статистических тестов:

№ задания	Проверяемая гипотеза H_0	Критерий	Статистическое решение ($\alpha = 0.1$)	Вывод
6		Хи-квадрат		
7		ANOVA		

Выводы:

В результате проведённого в п.6 статистического анализа обнаружено, что
В результате проведённого в п.7 статистического анализа обнаружено, что

ОТЧЕТ № 3

по теме «Основы регрессионного анализа»

Вариант №_____

ФИО студента _____ группа _____

Оценка: _____ Подпись:_____

Сводная таблица свойств различных регрессионных моделей:

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность			
Значимость			
Адекватность			
Степень тесноты связи			

Выводы:

В результате проведённого в п.8 статистического анализа обнаружено, что
В результате проведённого в п.9 статистического анализа обнаружено, что

1. Описательные статистики

1.1. Выборочные характеристики

Анализируемый признак 1 – Number of calories consumed per day

Анализируемый признак 2 – Grams of fat consumed per day

Анализируемый признак 3 – Grams of fiber consumed per day

а) Привести формулы расчёта выборочных характеристик

Выборочная хар-ка	Формула расчета
Объём выборки	n (дано)
Среднее	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Выборочная дисперсия	$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$
Выборочное среднеквадратическое отклонение	$S = \sqrt{S^2}$
Выборочный коэффициент асимметрии	$\gamma = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^3}{S^3}$
Выборочный эксцесс	$\varepsilon = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^4}{S^4} - 3$

б) Рассчитать выборочные характеристики

Выборочная хар-ка	Признак 1	Признак 2	Признак 3
Среднее	1796.65	77.03	12.79
Выборочная дисперсия	462872.63	1144.43	28.41
Выборочное среднеквадратическое отклонение	680.34	33.83	5.33
Выборочный коэффициент асимметрии	1.74	1.1	1.15
Выборочный эксцесс	7.98	1.96	2.43

1.2. Группировка и гистограммы частот

Анализируемый признак – Grams of fat consumed per day

Объём выборки – 315

а) Выбрать число групп

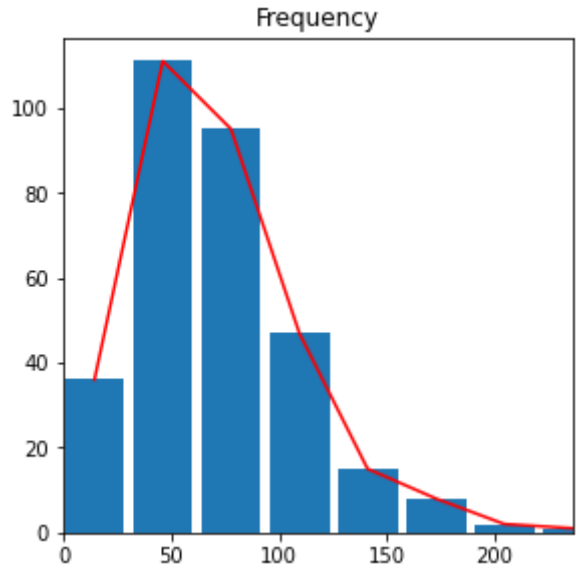
Число групп	Обоснование выбора числа групп	Ширина интервалов
8	Формула Стёрджеса	27.6875

б) Построить таблицу частот

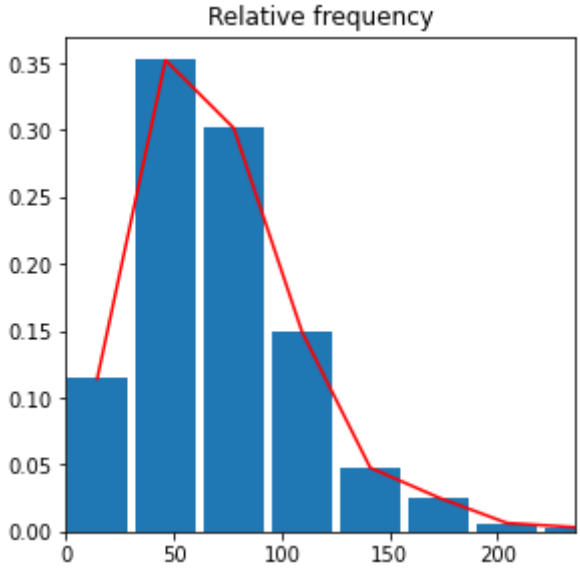
Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Накопл. частота	Относит. накопл. частота
1	14.4	42.09	36	0.11	36	0.11
2	42.09	69.78	111	0.35	147	0.46
3	69.78	97.46	95	0.3	242	0.76
4	97.46	125.15	47	0.15	289	0.91
5	125.15	152.84	15	0.05	304	0.96
6	152.84	180.53	8	0.03	312	0.99
7	180.53	208.21	2	0.006	314	0.996
8	208.21	235.9	1	0.003	315	1

в) Построить гистограммы частот и полигоны частот

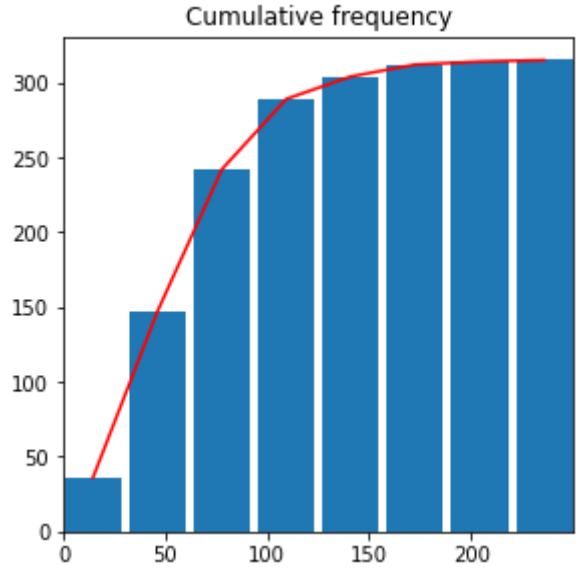
Гистограмма и полигон частот



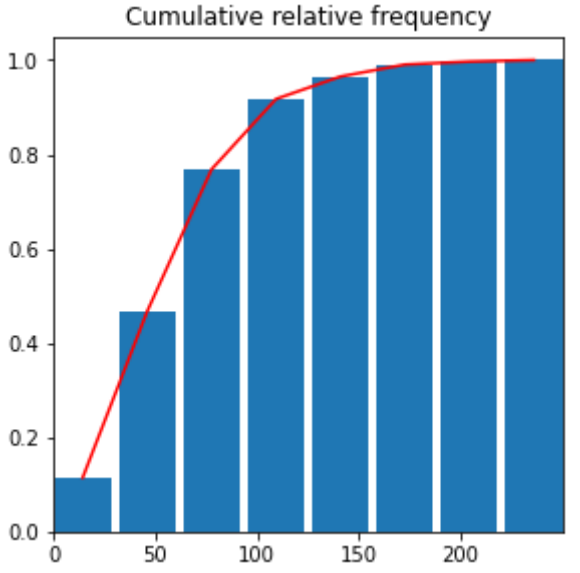
Гистограмма и полигон относительных частот



Гистограмма и полигон накопленных частот

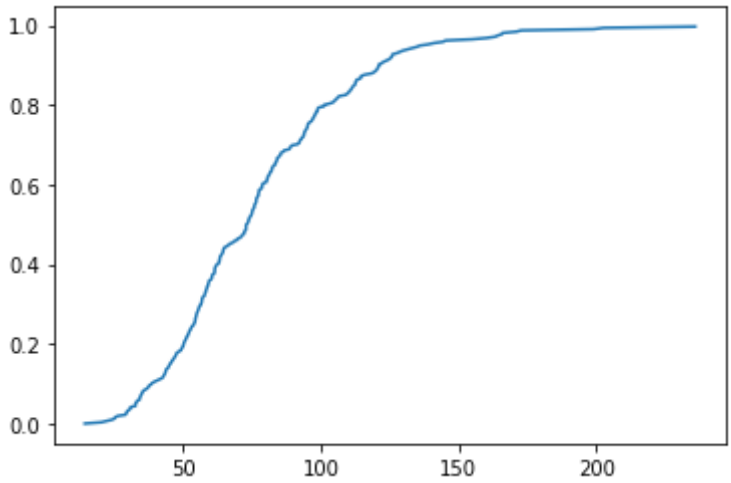


Гистограмма и полигон накопленных относительных частот



г) Построить график эмпирической функции распределения

Эмпирическая функция распределения



2. Интервальные оценки

2.1. Доверительные интервалы для мат. ожидания

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Оцениваемый параметр – мат. ожидание

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\bar{X} - \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$
Верхняя граница	$\bar{X} + \frac{s}{\sqrt{n}} t_{1-\alpha/2}(n-1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	1697	1721	1733
Верхняя граница	1896	1872	1860

2.2. Доверительные интервалы для дисперсии

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

Оцениваемый параметр – дисперсия

а) Привести формулы расчёта доверительных интервалов

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

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	380177	398216	407891
Верхняя граница	574002	544751	530546

2.3. Доверительные интервалы для разности мат. ожиданий

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315, 315

Оцениваемый параметр – разность мат. ожиданий

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$(\overline{x_1} - \overline{x_2}) - u_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Верхняя граница	$(\overline{x_1} - \overline{x_2}) + u_{1-\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	-453	-444	-439
Верхняя граница	-373	-382	-387

2.4. Доверительные интервалы для отношения дисперсий

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315, 315

Оцениваемый параметр – отношение дисперсий

а) Привести формулы расчёта доверительных интервалов

Граница доверительного интервала	Формула расчета
Нижняя граница	$\frac{s_1^2}{s_2^2} F_{\alpha/2}(n_2 - 1, n_1 - 1)$
Верхняя граница	$\frac{s_1^2}{s_2^2} F_{1-\alpha/2}(n_2 - 1, n_1 - 1)$

б) Рассчитать доверительные интервалы

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	0.49	0.56	0.6
Верхняя граница	1	0.97	0.94

3. Проверка статистических гипотез о математических ожиданиях и дисперсиях

3.1. Проверка статистических гипотез о математических ожиданиях

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

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

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

	Выражение
Формула расчета статистики критерия	$Z = \frac{\bar{X} - m_0}{\sqrt{s^2/n}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	T(n-1)
Формулы расчета критических точек	$\pm t_{1-\alpha/2}(n-1)$
Формула расчета *p-value*	$2 * min(F_Z(z), 1 - F_Z(z))$

б) Выбрать произвольные значения m_0 и проверить статистические гипотезы

m_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
1500	0.1	7.74	*0*	отклоняем H_0	m ≠ 1500
1800	0.1	-0.087	*0.93*	принимаем H_0	m = 1800
2000	0.1	-5.3	*0*	отклоняем H_0	m ≠ 2000

3.2. Проверка статистических гипотез о дисперсиях

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

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

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	$\chi^2_{\alpha/2}(n-1); \chi^2_{1-\alpha/2}(n-1)$
Формула расчета *p-value*	$2 * min(F_Z(z), 1 - F_Z(z))$

б) Выбрать произвольные значения σ_0 и проверить статистические гипотезы

σ_0	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
400000	0.1	363.36	*0.057*	отклоняем H_0	$\sigma \neq 400000$
450000	0.1	322.98	*0.703*	принимаем H_0	$\sigma = 450000$
500000	0.1	290.68	*0.353*	принимаем H_0	$\sigma = 500000$

3.3. Проверка статистических гипотез о равенстве математических ожиданий

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

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	$Z = \frac{\overline{X_1} - \overline{X_2}}{S * \sqrt{1/n_1 + 1/n_2}}$
Закон распределения статистики критерия при условии истинности основной гипотезы	$T(n_1 + n_2 - 2)$
Формулы расчета критических точек	$\pm t_{1-\alpha}(n_1 + n_2 - 2)$

Формула расчета *p-value*	$2 * min\left(F_Z(z), 1 - F_Z(z)\right)$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-26.39	*0*	отклоняем H_0	$m_1 \neq m_2$
0.05			отклоняем H_0	$m_1 \neq m_2$
0.1			отклоняем H_0	$m_1 \neq m_2$

3.4. Проверка статистических гипотез о равенстве дисперсий

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

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*	$2 * min\left(F_Z(z), 1 - F_Z(z)\right)$

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.767	*0.019*	принимаем H_0	$\sigma_1 = \sigma_2$
0.05			отклоняем H_0	$\sigma_1 \neq \sigma_2$

0.1			отклоняем H_0	$\sigma_1 \neq \sigma_2$

4. Критерии согласия

Анализируемый признак – Number of calories consumed per day

Объём выборки – 315

4.1. Критерий хи-квадрат

Теоретическое распределение – нормальное.

Статистическая гипотеза – $H_0 : F(x)$ □ N

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$	n_i- число элементов выборки, принадлежащих i-той группе p_i- вероятность того, что значение случайной величины попадёт в i-тую группу
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(k - r - 1)$	k - число групп r - число неизвестных параметров распределения, оцениваемых по выборке
Формула расчета критической точки	$\chi^2_{1-\alpha}(k - r - 1)$	α – уровень значимости
Формула расчета *p-value*	$1 - F_Z(z)$	

б) Выбрать число групп

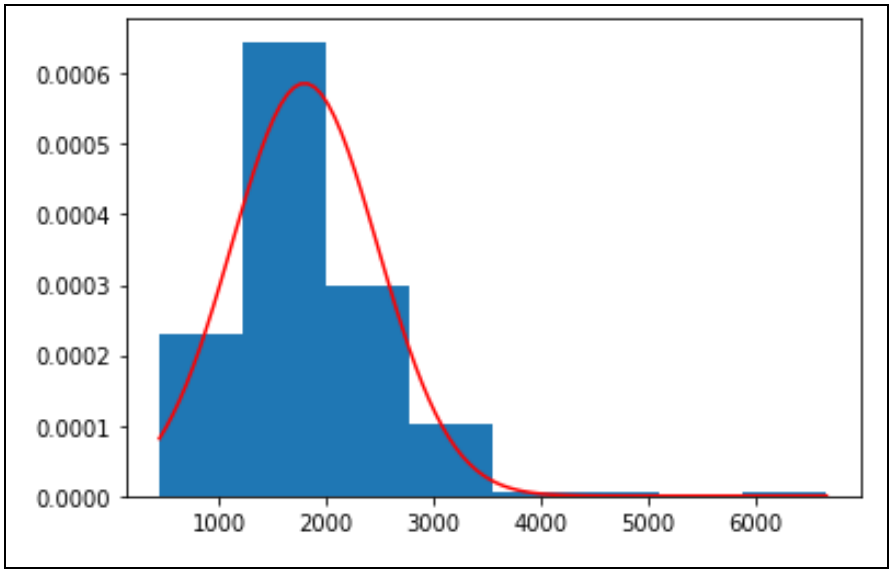
Число групп	Обоснование выбора числа групп	Ширина интервалов
8	Формула Стёрджеса: k ≈ 1 + 1.3ln(n)	777.125

в) Построить таблицу частот

Номер интервала	Нижняя граница	Верхняя граница	Частота	Относит. частота	Вероятность попадания в интервал при условии истинности основной гипотезы
1	445.2	1222.325	56	0.1778	0.1758
2	1222.325	1999.45	158	0.5016	0.4179
3	1999.45	2776.575	73	0.2317	0.3079
4	2776.575	3553.7	25	0.0794	0.07

5	3553.7	4330.825	1	0.0032	0.0048
6	4330.825	5107.95	1	0.0032	0.0001
7	5107.95	5885.075	0	0	0
8	5885.075	6662.2	1	0.0032	0

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	12.774	*0.005*	отклоняем H_0	X имеет распределение, отличное от нормального
0.05	12.774	*0.005*	отклоняем H_0	
0.1	12.774	*0.005*	отклоняем H_0	

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

Статистическая гипотеза – $H_0 : F(x) \square N$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{n}{6}(S^2 + \frac{1}{4}K^2)$	S – выборочный коэффициент асимметрии K – выборочный эксцесс
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(2)$	

Формула расчета критической точки	$\chi^2_{1-\alpha}(2)$	α – уровень значимости
Формула расчета *p-value*	$1 - F_Z(z)$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	995.22	*0*	отклоняем H_0	X имеет распределение, отличное от нормального
0.05			отклоняем H_0	
0.1			отклоняем H_0	

Вывод (в терминах предметной области)

В результате проведённого в п.4 статистического анализа обнаружено, что количество калорий, в день потребляемых пациентами, не является нормально распределённой величиной.

5. Проверка однородности выборок

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

5.1 Критерий знаков

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{2K^+ - n}{\sqrt{n}}$	K^+ – число знаков «+» в последовательности знаков разностей соответствующих элементов выборки
Закон распределения статистики критерия при условии истинности основной гипотезы	N(0, 1)	
Формула расчета критической точки	$\pm N_{1-\frac{\alpha}{2}}(0,1)$	
Формула расчета *p-value*	$2 * min\left(F_Z(z), 1 - F_Z(z)\right)$	

б) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	-146.5	*0*	отклоняем H_0	$F_1(x) \neq F_2(x)$
0.05			отклоняем H_0	$F_1(x) \neq F_2(x)$
0.1			отклоняем H_0	$F_1(x) \neq F_2(x)$

5.2. Критерий хи-квадрат

Статистическая гипотеза – $H_0 : F_1(x) = F_2(x)$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$\sum_{j=1}^{k} \frac{(n_{1j} - n_{2j})^2}{n_{1j} + n_{2j}}$	k – число групп n_{1j}, n_{2j} – j-ые элементы выборок соответственно 1-го и 2-го признаков
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(k-1)$	
Формула расчета критической точки	$\chi^2_{1-\alpha}(k-1)$	α – уровень значимости
Формула расчета *p-value*	$1 - F_Z(z)$	

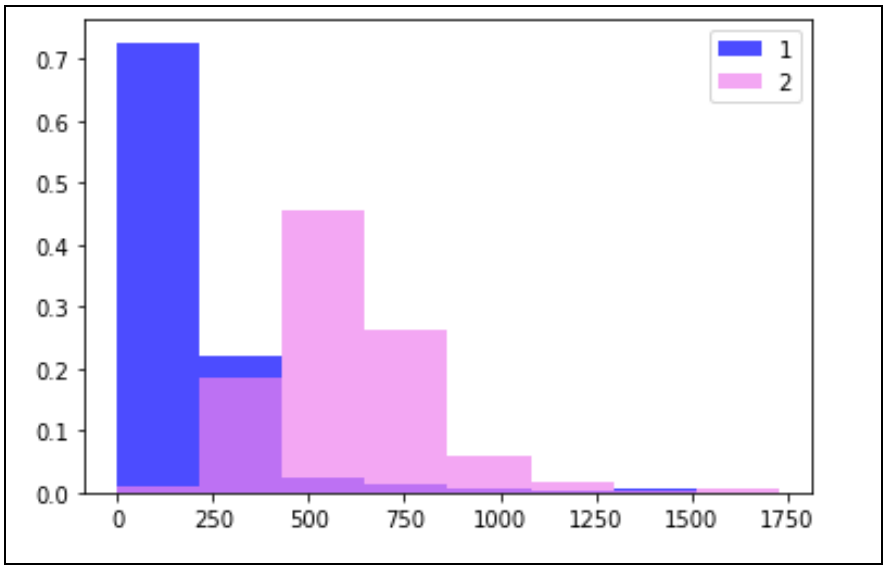
б) Выбрать число групп

Число групп	Обоснование выбора числа групп	Ширина интервалов
8	Формула Стёрджеса: $k \approx 1 + 1.3\ln(n)$	215.875

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	0	215.875	229	3	0.727	0.0095
2	215.875	431.75	69	58	0.219	0.1841
3	431.75	647.625	8	144	0.0254	0.4571
4	647.625	863.5	4	83	0.0127	0.2635
5	863.5	1079.375	2	19	0.0063	0.0603
6	1079.375	1295.25	1	5	0.0032	0.0159
7	1295.25	1511.125	2	1	0.0063	0.0032
8	1511.125	1727	0	2	0	0.0063

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	17252.467	*0*	отклоняем H_0	$F_1(x) \neq F_2(x)$
0.05			отклоняем H_0	$F_1(x) \neq F_2(x)$
0.1			отклоняем H_0	$F_1(x) \neq F_2(x)$

Вывод (в терминах предметной области)

В результате проведённого в п.5 статистического анализа обнаружено, что анализируемые признаки имеют разные функции распределения.

6. Таблицы сопряжённости

Факторный признак *x* – Sex

Результативный признак *y* – Smoking status

Объёмы выборок – 315

Статистическая гипотеза – H_0: $F_Y(y|X = \text{Male}) = F_Y(y|X = \text{Female}) = F_Y(y)$

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\frac{(n_{ij} - m_{ij})^2}{m_{ij}}$	n_{ij}- частота пары (x_i, y_j) в выборке m_{ij}- частота пары (x_i, y_j) при условии, что H_0 верна
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2((n_1 - 1)(n_2 - 1))$	n_1 - число вариантов признака *X* n_2 - число вариантов признака *Y*
Формула расчета критической точки	$\chi^2_{1-\alpha}((n_1 - 1)(n_2 - 1))$	α – уровень значимости
Формула расчета *p-value*	$1 - F_Z(z)$	

б) Построить эмпирическую таблицу сопряжённости

x \ *y*	Never	Former	Current Smoker	Σ
Male	13	22	7	42
Female	144	93	36	273
Σ	157	115	43	315

в) Построить теоретическую таблицу сопряжённости

x \ *y*	Never	Former	Current Smoker	Σ
Male	20.93	15.33	5.73	42
Female	136.07	99.67	37.27	273
Σ	157	115	43	315

г) Проверить статистические гипотезы

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	7.14	*0.0282*	принимаем H_0	$F_Y(y\|X = \text{Male}) = F_Y(y\|X = \text{Female}) = F_Y(y)$
0.05			отклоняем H_0	$F_Y(y\|X = \text{Male}) \neq F_Y(y\|X = \text{Female})$
0.1			отклоняем H_0	

Вывод (в терминах предметной области)

В результате проведённого в п.6 статистического анализа обнаружено, что при уровне значимости 0.01 принимается гипотеза об отсутствии статистической зависимости между признаками 1 и 2, однако при уровнях значимости 0.05 и 0.1 эта гипотеза отклоняется.

7. Дисперсионный анализ

Факторный признак *x* – Smoking status

Результативный признак *y* – Quetelet (weight/(height^2))

Число вариантов факторного признака – 3

Объёмы выборок – 315

Статистическая гипотеза – $H_0: m_1 = m_2 = m_3$

а) Рассчитать групповые выборочные характеристики

№ п/п	Вариант факторного признака	Объём выборки	Групповые средние	Групповые дисперсии
1	Never	157	26.73	47.12
2	Former	115	25.93	24.92
3	Current smoker	43	24.69	24.16

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

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$\widetilde{D}_{\text{меж}} = \frac{1}{n}\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	k - 1	$\frac{n}{k-1}\widetilde{D}_{\text{меж}}$
Остаточные признаки	$\widetilde{D}_{\text{внутр}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$	n – k	$\frac{n}{n-k}\widetilde{D}_{\text{внутр}}$
Все признаки	$\widetilde{D}_{\text{общ}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$	n - 1	$\frac{n}{n-1}\widetilde{D}_{\text{общ}}$

в) Рассчитать показатели вариации, используемые в дисперсионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	0.2821	2	44.43075

Остаточные признаки	35.8807	312	36.2257
Все признаки	36.1628	314	36.278

г) Проверить правило сложения дисперсий

Показатель	$D_{межгр}$	$D_{внутригр}$	$D_{общ}$	$D_{межгр} + D_{внутригр}$
Значение	0.2821	35.8807	36.1628	36.1628

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Эмпирический коэффициент детерминации	$\tilde{\eta}^2 = \frac{\widetilde{D}_{\text{меж}}}{\widetilde{D}_{\text{общ}}}$	0.0078
Эмпирическое корреляционное отношение	$\tilde{\eta} = \sqrt{\frac{\widetilde{D}_{\text{меж}}}{\widetilde{D}_{\text{общ}}}}$	0.0883

е) Охарактеризовать тип связи между факторным и результативным признаками

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{\widetilde{D}_{\text{меж}}/(k-1)}{\widetilde{D}_{\text{внутр}}/(n-k)}$	
Закон распределения статистики критерия при условии истинности основной гипотезы	$F(k-1, n-k)$	
Формула расчета критической точки	$f_{1-\alpha}(k-1, n-k)$	

Формула расчета *p-value*	$1 - F_Z(z)$	

з) Проверить статистическую гипотезу дисперсионного анализа

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	2.07	*0.13*	принимаем H_0	$m_1 = m_2 = m_3$
0.05			принимаем H_0	$m_1 = m_2 = m_3$
0.1			принимаем H_0	$m_1 = m_2 = m_3$

Вывод (в терминах предметной области)

В результате проведённого в п.7 статистического анализа обнаружено, что отсутствует статистическая зависимость между факторным признаком Smoking Status и результативным признаком Quetelet.

8. Корреляционный анализ

8.1. Расчёт парных коэффициентов корреляции

Анализируемый признак 1 – Plasma beta-carotene (ng/ml)

Анализируемый признак 2 – Plasma Retinol (ng/ml)

Объёмы выборок – 315

а) Рассчитать точечные оценки коэффициентов корреляции

	Формула расчета	Значение
Линейный коэффициент корреляции	$\tilde{\rho}_{XY} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$	0.07
Ранговый коэффициент корреляции по Спирмену	$\tilde{\rho}_{XY}^{(\text{сп})} = \frac{cov(R,S)}{\sigma_R \sigma_S}$	0.13
Ранговый коэффициент корреляции по Кендаллу	$\tilde{\tau}_{XY} = \frac{4\sum_{i=1}^{n}\sum_{j=i+1}^{n}[S_j > S_i]}{n(n-1)} - 1$	0.09

б) Привести формулы расчёта доверительного интервала для линейного коэффициента корреляции

Граница доверительного интервала	Формула расчета
Нижняя граница	$\tilde{\rho}_{XY} + \frac{\tilde{\rho}_{XY}(1-(\tilde{\rho}_{XY})^2)}{2n} - u_{1-\frac{\alpha}{2}} \frac{1-(\tilde{\rho}_{XY})^2}{\sqrt{n}}$
Верхняя граница	$\tilde{\rho}_{XY} + \frac{\tilde{\rho}_{XY}(1-(\tilde{\rho}_{XY})^2)}{2n} + u_{1-\frac{\alpha}{2}} \frac{1-(\tilde{\rho}_{XY})^2}{\sqrt{n}}$

в) Рассчитать доверительные интервалы для линейного коэффициента корреляции

Граница доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
Нижняя граница	-0.07	-0.04	-0.02
Верхняя граница	0.21	0.18	0.16

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

$H_0: \rho = 0$ $H': \rho \neq 0$	$Z = \frac{\tilde{\rho}_{XY}}{\sqrt{1-(\tilde{\rho}_{XY})^2}}\sqrt{n-2}$	$T(n-2)$
$H_0: r^{(сп)} = 0$ $H': r^{(сп)} \neq 0$	$Z = \frac{\tilde{\rho}_{XY}^{(\text{сп})}}{\sqrt{1-(\tilde{\rho}_{XY}^{(\text{сп})})^2}}\sqrt{n-2}$	$T(n-2)$
$H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$	$Z = \tilde{\tau}_{XY}\sqrt{\frac{9n(n+1)}{2(2n+5)}}$	$N(0,1)$

д) Проверить значимость коэффициентов корреляции

Статистическая гипотеза	Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
$H_0: \rho = 0$ $H': \rho \neq 0$	0.1	1.24	*0.2*	принимаем H_0	$\rho = 0$
$H_0: r^{(сп)} = 0$ $H': r^{(сп)} \neq 0$	0.1	2.32	*0.02*	отклоняем H_0	$r^{(\text{сп})} \neq 0$
$H_0: r^{(кен)} = 0$ $H': r^{(кен)} \neq 0$	0.1	2.39	*0.02*	отклоняем H_0	$r^{(\text{кен})} \neq 0$

8.2. Расчёт множественных коэффициентов корреляции

Анализируемый признак 1 – Number of calories consumed per day

Анализируемый признак 2 – Grams of fat consumed per day

Анализируемый признак 3 – Grams of fiber consumed per day

Объёмы выборок – 315

а) Рассчитать матрицу ранговых коэффициентов корреляции по Кендаллу

Признак / Признак	1	2	3
1	1	0.72	0.39
2	0.72	1	0.22
3	0.39	0.22	1

б) Рассчитать матрицу значений p-value для ранговых коэффициентов корреляции по Кендаллу (статистическая гипотеза $H_0: r^{(кен)} = 0,\ H': r^{(кен)} \neq 0$)

Признак / Признак	1	2	3
1	–	0	0
2	0	–	0
3	0	0	–

в) Рассчитать точечную оценку коэффициента конкордации

	Формула расчета	Значение
Коэффициент конкордации	$W = \frac{12}{n^3 - n}\sum_{i=1}^{n}(\frac{1}{k}\sum_{j=1}^{k} r_{ij} - \frac{n+1}{2})^2$	0.726

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = k(n-1)W$	
Закон распределения статистики критерия при условии истинности основной гипотезы	$\chi^2(n-1)$	
Формула расчета критической точки	$\pm\chi^2_{1-\frac{\alpha}{2}}(n-1)$	
Формула расчета *p-value*	$2 * min(F_Z(z), 1 - F_Z(z))$	

д) Проверить значимость коэффициента конкордации

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	684	*0*	отклоняем H_0	между признаками 1, 2 и 3 существует ранговая корреляция
0.05			отклоняем H_0	
0.1			отклоняем H_0	

Вывод (в терминах предметной области)

В результате проведённого в п.8 статистического анализа обнаружено, что между признаками Plasma beta-carotene и Plasma Retinol отсутствует линейная связь, а между признаками Number of calories consumed per day, Grams of fat consumed per day и Grams of fiber consumed per day присутствует ранговая корреляция

9. Регрессионный анализ

9.1 Простейшая линейная регрессионная модель

Факторный признак *x* – Number of calories consumed per day

Результативный признак *y* – Plasma beta-carotene (ng/ml)

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	$\widetilde{\beta_0} = \bar{y} - \widetilde{\beta_1}\bar{x}$	201
β_1	$\widetilde{\beta_1} = \tilde{\rho}_{XY}\frac{\tilde{\sigma}_Y}{\tilde{\sigma}_X}$	-0.006

б) Записать точечную оценку уравнения регрессии

$f(x) = 201 - 0.006x$

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

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	$\widetilde{D}_{\text{регр}} = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \bar{y})^2$	k - 1	$\frac{n}{k-1}\widetilde{D}_{Y\|X}$
Остаточные признаки	$\widetilde{D}_{\text{ост}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$	n - k	$\frac{n}{n-k}\widetilde{D}_{resY}$
Все признаки	$\widetilde{D}_{\text{общ}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$	n - 1	$\frac{n}{n-1}\widetilde{D}_Y$

г) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка

Факторный признак	16	1	5040
Остаточные признаки	33367	313	33580
Все признаки	33383	314	33489

д) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	16	33367	33383	33383

е) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\tilde{R}_{XY}^2 = \frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}$	0.00048
Корреляционное отношение	$\tilde{R}_{XY} = \sqrt{\frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}}$	0.022

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

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
β_0	Нижняя граница	$\widetilde{\beta_0} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2 \widetilde{D}_X}}$

	Верхняя граница	$\widetilde{\beta_0} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n^2\widetilde{D}_X}}$
β_1	Нижняя граница	$\widetilde{\beta_1} - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}}\sqrt{\frac{1}{\text{n}\widetilde{D}_X}}$
	Верхняя граница	$\widetilde{\beta_1} + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}}\sqrt{\frac{1}{\text{n}\widetilde{D}_X}}$

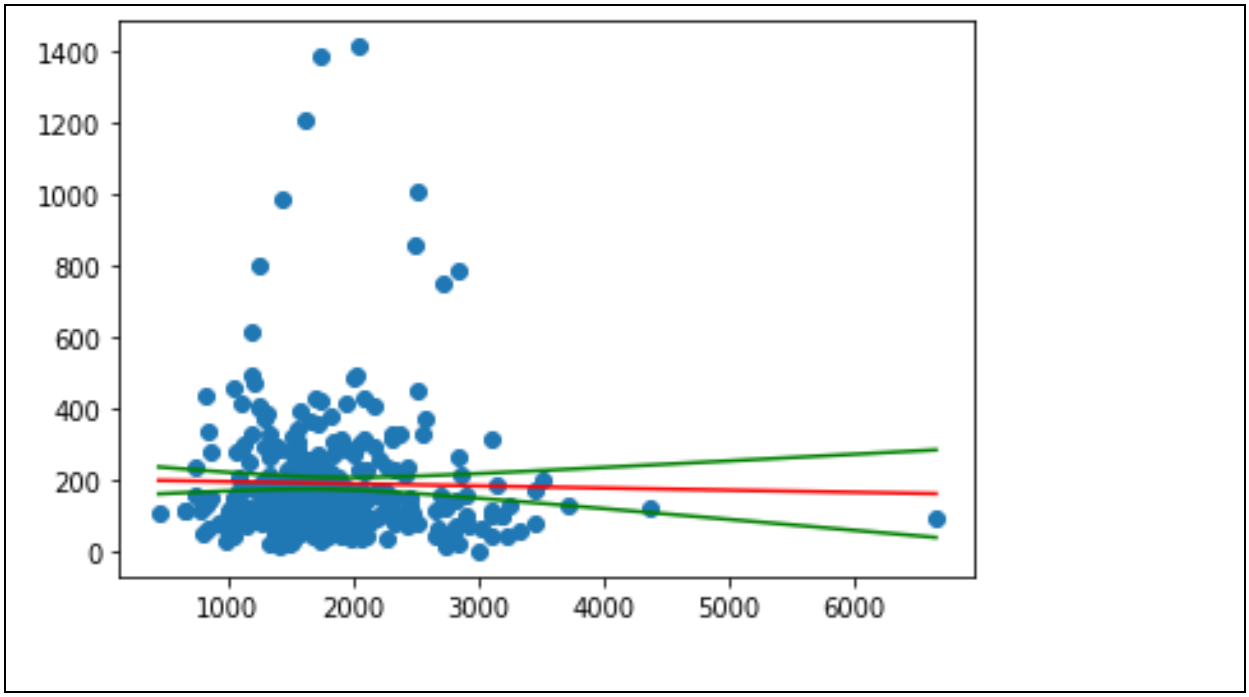
б) Рассчитать доверительные интервалы для параметров линейной регрессионной модели

Параметр	Границы доверительного интервала	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.1$
β_0	Нижняя граница	125	143	153
	Верхняя граница	276	258	249
β_1	Нижняя граница	-0.045	-0.036	-0.031
	Верхняя граница	0.033	0.024	0.019

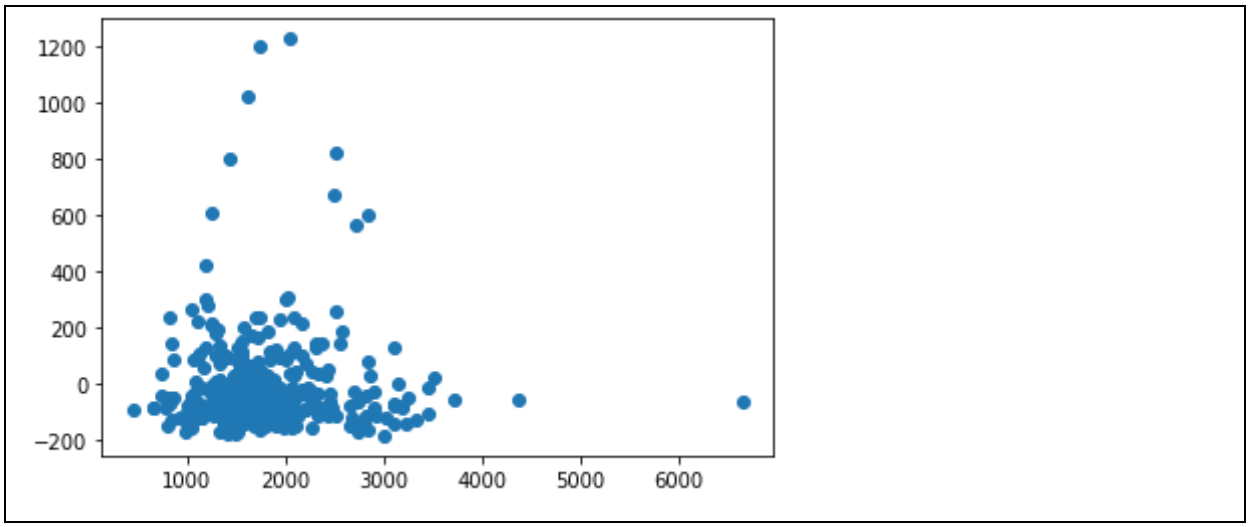
в) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

Границы доверительного интервала	Формула расчета
Нижняя граница $f_{low}(x)$	$\tilde{f}(x) - t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{n\widetilde{D}_X}\right)}$
Верхняя граница $f_{high}(x)$	$\tilde{f}(x) + t_{1-\frac{\alpha}{2}}(n-2)\sqrt{\widetilde{D}_{\text{ост}}\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{n\widetilde{D}_X}\right)}$

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 $\alpha = 0.1$



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	$Z = \frac{\tilde{R}_{XY}^2}{(1-\tilde{R}_{XY}^2)/(n-2)}$	
Закон распределения статистики критерия при условии истинности основной гипотезы	$f(1, n-2)$	
Формула расчета критической точки	$f_{1-\alpha}(1, n-2)$	
Формула расчета *p-value*	$1 - F_Z(z)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.15	*0.69*	принимаем H_0	$\beta_1 = 0$
0.05			принимаем H_0	$\beta_1 = 0$
0.1			принимаем H_0	$\beta_1 = 0$

9.2 Линейная регрессионная модель общего вида

Факторный признак *x* – Number of calories consumed per day

Результативный признак *y* – Plasma beta-carotene (ng/ml)

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
β_0	$\tilde{\beta} = (F^T F)^{-1} F^T y$ $\tilde{\beta} = (\beta_0, \beta_1, \beta_2)^T$ $y = (y_1, \dots, y_n)^T$ $F = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \dots & \dots & \dots \\ 1 & x_n & x_n^2 \end{pmatrix}$	156
β_1		0.037
β_2		-0.000009

б) Записать точечную оценку уравнения регрессии

$f(x) = 156 + 0.037x - 0.000009x^2$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	136.46	2	21492.45
Остаточные признаки	33246.52	312	33566.2
Все признаки	33382.98	314	33489.3

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	136.46	33246.52	33382.98	33382.98

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Коэффициент детерминации	$\tilde{R}_{XY}^2 = \frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}$	0.004
Корреляционное отношение	$\tilde{R}_{XY} = \sqrt{\frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}}$	0.063

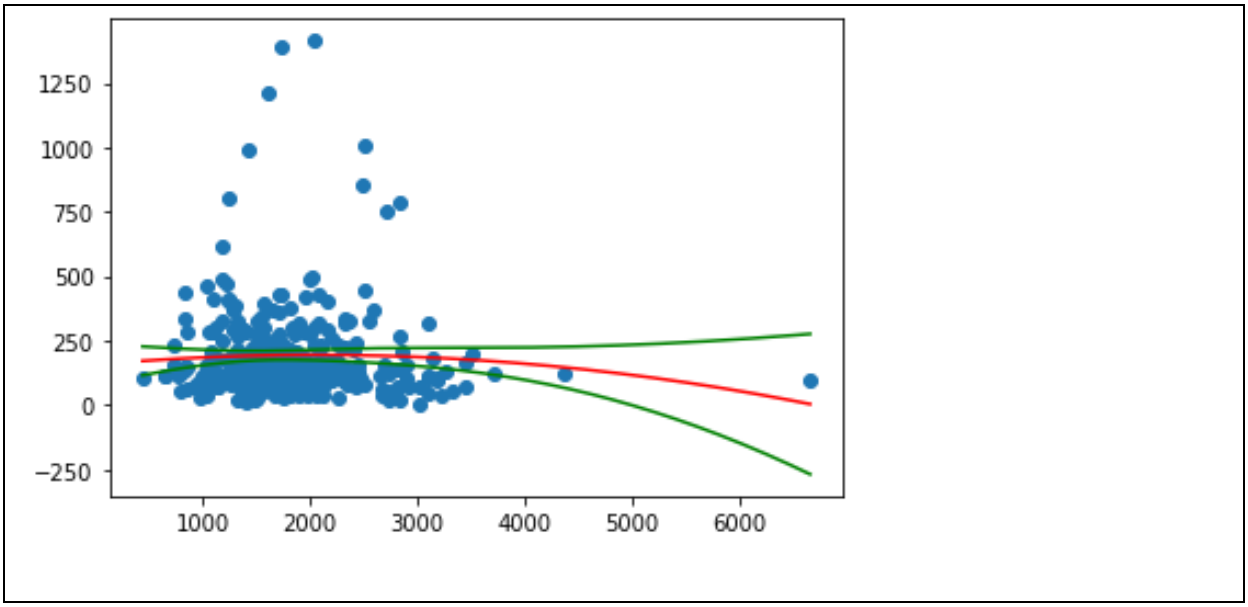
е) Охарактеризовать тип связи между факторным и результативным признаками, определяемой рассчитанной линейной регрессией

9.2.2. Интервальные оценки линейной регрессионной модели

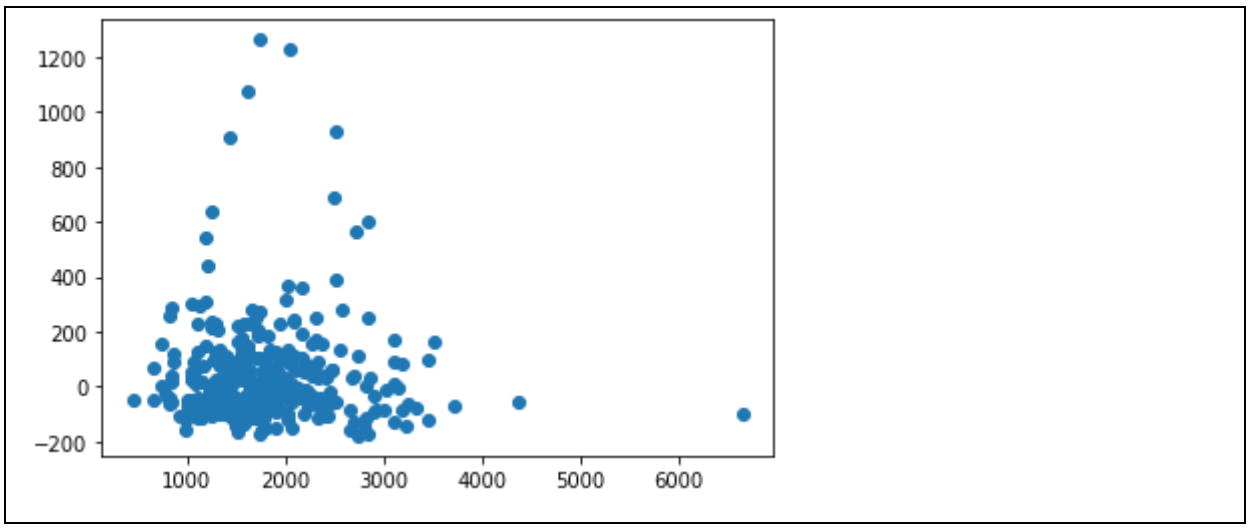
а) Привести формулы расчёта доверительного интервала для значений регрессии f(x)

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

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 – $H_0 : \beta_1 = \beta_2 = 0$
$H' :$ *не* H_0

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

	Выражение	Пояснение использованных обозначений
Формула расчета статистики критерия	$Z = \frac{\tilde{R}^2_{XY}/(k\ -\ 1)}{(1 - \tilde{R}^2_{XY})/(n - k)}$	
Закон распределения статистики критерия при условии истинности основной гипотезы	$f(k-1,\ n\ -\ k)$	
Формула расчета критической точки	$f_{1-\alpha/2}(k-1,\ n\ -\ k)$	
Формула расчета *p-value*	$1 - F_Z(z)$	

б) Проверить значимость линейной регрессионной модели

Уровень значимости	Выборочное значение статистики критерия	*p-value*	Статистическое решение	Вывод
0.01	0.63	*0.53*	принимаем H_0	$\beta_1 = \beta_2 = 0$
0.05			принимаем H_0	$\beta_1 = \beta_2 = 0$
0.1			принимаем H_0	$\beta_1 = \beta_2 = 0$

9.3 Множественная линейная регрессионная модель

Факторный признак 1 x_1 – Number of calories consumed per day

Факторный признак 2 x_2 – Quetelet (weight/(height^2))

Результативный признак y – Plasma beta-carotene (ng/ml)

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
β_0	$\tilde{\beta} = (F^T F)^{-1} F^T y$	383
β_1	$\tilde{\beta} = (\beta_0, \beta_1, \beta_2)^T$	-0.0058
β_2	$y = (y_1, \dots, y_n)^T$	-6.98
	$F = \begin{pmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} \end{pmatrix}$	

б) Записать точечную оценку уравнения регрессии

$f(x) = 383 - 0.0058x_1 - 6.98x_2$

в) Рассчитать показатели вариации, используемые в регрессионном анализе

Источник вариации	Показатель вариации	Число степеней свободы	Несмещенная оценка
Факторный признак	1771.85	2	279066
Остаточные признаки	31611.13	312	31915
Все признаки	33382.98	314	33489

г) Проверить правило сложения дисперсий

Показатель	$D_{регр}$	$D_{ост}$	$D_{общ}$	$D_{регр} + D_{ост}$
Значение	1771.85	31611.13	33382.98	33382.98

д) Рассчитать показатели тесноты связи между факторным и результативным признаками

Показатель	Формула расчета	Значение
Множественный коэффициент детерминации	$\tilde{R}_{XY}^2 = \frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}$	0.056
Множественное корреляционное отношение	$\tilde{R}_{XY} = \sqrt{\frac{\widetilde{D}_{\text{регр}}}{\widetilde{D}_{\text{общ}}}}$	0.237

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

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
Факторный признак	16	136	1772
Остаточные признаки	33367	33247	31611
Все признаки	33383	33383	33383

б) Сводная таблица свойств различных регрессионных моделей

Свойство	Простейшая линейная модель	Линейная модель с квадратичным членом	Множественная линейная модель
Точность	0.048%	0.4%	5.6%
Значимость	Нет	Нет	Нет
Адекватность	Нет	Нет	Нет
Степень тесноты связи	слабая	слабая	слабая

Вывод (в терминах предметной области)

В результате проведённого в п.9 статистического анализа обнаружено, что ни одна из предложенный регрессионных моделей не отражает реальную зависимость признака Plasma beta-carotene от признаков Number of calories consumed per day и Quetelet