Санкт-Петербургский Политехнический
Университет им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

Сводный отчёт по дисциплине "Математическая статистика"по лабораторным отчётам 5 – 8

3 курс, группа 33631/2

Студент Д. А. Плаксин

Преподаватель Баженов А. Н.

Санкт-Петербург
2019 г.

Содержание

1 Список иллюстраций

2 Список таблиц

3 Постановка задачи

3.1 Вычисление коэффициента корреляции

Необходимо построить выборки объёмом $20, 60, 100, 1000$ для двумерного нормального распределения с коэффициентами корреляции $\rho = 0, 0.5, 0.9$

Вычислить коэффициент корреляции Пирсона, Спирмана и квадрантный коэффициент корреляции для каждой выборки. Эти же вычисления повторить для смеси двумерных нормальных распределений [4]:

$$f(x, y) = 0.9N(x, y, 0, 0, 1, 1, 0.9) + 0.1N(x, y, 0, 0, 10, 10, -0.9) \tag{1}$$

На графике изобразить точки выборки и эллипс равновероятности.

3.2 Оценки линий регрессии

Необходимо найти оценки линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек отрезка $[-1.8;\ 2]$ с равномерным шагом 0.2. Ошибку e_i считать нормально распределённой с параметрами $(0,\ 1)$. В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей.

3.3 Точечная оценка параметров распределения

Необходимо сгенерировать выборку объемом 100 элементов для нормального распределения $N(x; 0, 1)$. По сгенерированной выборке оценить параметры μ и σ нормального закона методом максимального правдоподобия. В качестве основной гипотезы H_0 будем считать, что сгенерированное распределение имеет вид $N(x, \hat{\mu}, \hat{\sigma})$. Проверить основную гипотезу, используя критерий согласия χ. В качестве ровня значимости взять $\alpha = 0,05$. Привести таблицу вычислений χ^2.

3.4 Интервальные оценки параметров распределения

Для двух выборок 20 и 100 элементов, сгенерированных согласно нормальному закону $N(x, 0, 1)$, для параметров масштаба и положения построить асимптотически нормальные интервальные оценки на основе точечных оценок метода максимального правдоподобия и классические интервальные оценки на основе статистик χ^2 и Стьюдента. В качестве параметра надёжности взять $\gamma = 0.95$.

4 Теория

4.1 Вычисление коэффициента корреляции

1. Двумерное нормально распределение [5]:

$$N(x, y, 0, 0, 1, 1, \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)} \tag{2}$$

2. Коэффициент корреляции Пирсона [6]:

$$r_{xy} = \left(\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right)\left(\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{-\frac{1}{2}} \tag{3}$$

3. Коэффициент корреляции Спирмана [7]:

$$\rho_n = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n} d_i^2 \tag{4}$$

4. Квадрантный коэффициент корреляции [8]:

$$\hat{q} = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - med\ x)sign(y_i - med\ y) \tag{5}$$

4.2 Оценки линий регрессии

Простая линейная регрессия [9]:

$$y_i = ax_i + b + e_i,\ i = \overline{1,n}, \tag{6}$$

где x_i – заданные числа, y_i – наблюдаемые значения, e_i – независимы и нормально распределены, a и b – неизвестные параметры, подлежащие оцениванию.

4.2.1 Метод наименьших квадратов

Критерий – минимизация функции [10]:

$$Q(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 \to \min \tag{7}$$

Оценка $\hat{a}$ и $\hat{b}$ параметров a и b, в которых достигается минимум $Q(a,b)$, называются МНК-оценками. В случае линейной регрессии их можно вычислить из формулы [11]:

$$\begin{cases} \hat{a} = \frac{\overline{xy} - \overline{x}\overline{y}}{\overline{x^2} - \overline{x}^2} \\ \hat{b} = \overline{y} - \hat{a}\overline{x} \end{cases} \tag{8}$$

Метод наименьших квадратов является несмещённой оценкой.

МНК чувствителен к выбросам (т.к. в вычислении используется выборочное среднее значение величин крайне неустойчивое к редким и большим по величине выбросам)

4.2.2 Метод наименьших модулей

Критерий наименьших модулей – заключается в минимизации следующей функции [12]:

$$M(a,b) = \sum_{i=1}^{n}|y_i - ax_i - b| \to \min \tag{9}$$

МНМ-оценки обладают свойством робастности Но на практике решение реализуется только численно

4.3 Точечная оценка параметров распределения

4.3.1 Метод максимального правдоподобия

Метод максимального правдоподобия – метод оценивания неизвестного параметра путём максимимзации функции правдоподобия.

$$\hat{\theta}_{\text{МП}} = argmax\mathbf{L}(x_1, x_2, \ldots, x_n, \theta) \tag{10}$$

Где $\mathbf{L}$ это функция правдоподобия, которая представляет собой совместную плотность вероятности независимых случайных величин $X_1, x_2, \ldots, x_n$ и является функцией неизвестного параметра θ

$$\mathbf{L} = f(x_1, \theta) \cdot f(x_2, \theta) \cdot \dots \cdot f(x_n, \theta) \tag{11}$$

Оценкой максимального правдоподобия будем называть такое значение $\overset{\wedge}{\theta}_{\text{МП}}$ из множества допустимых значений параметра θ, для которого функция правдоподобия принимает максимальное значение при заданных $x_1, x_2, \ldots, x_n$.

Тогда при оценивании математического ожидания m и дисперсии σ^2 нормального распределения $N(m, \sigma)$ получим:

$$\ln(\mathbf{L}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - m)^2 \tag{12}$$

4.3.2 Критерий согласия Пирсона

Разобьём генеральную совокупность на k непересекающихся подмножеств $\Delta_1, \Delta_2, \ldots, \Delta_k,\ \Delta_i = (a_i, a_{i+1}],\ p_i = P(X \in \Delta_i),\ i = 1, 2, \ldots, k$ – вероятность того, что точка попала в iый промежуток.

Так как генеральная совокупность это $\mathbb{R}$, то крайние промежутки будут бесконечными: $\Delta_1 = (-\infty, a_1],\ \Delta_k = (a_k, \infty),\ p_i = F(a_i) - F(a_{i-1})$

n_i – частота попадания выборочных элементов в $\Delta_i,\ i = 1, 2, \ldots, k$.

В случае справедливости гипотезы H_0 относительно частоты $\frac{n_i}{n}$ при больших n должны быть близки к p_i, значит в качестве меры имеет смысл взять:

$$Z = \sum_{i=1}^{k}\frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 \tag{13}$$

Тогда

$$\chi^2_B = \sum_{i=1}^{k}\frac{n}{p_i}\left(\frac{n_i}{n} - p_i\right)^2 = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i} \tag{14}$$

Для выполнения гипотезы H_0 должны выполняться следующие условия [13]:

$$\chi^2_B < \chi^2_{1-\alpha}(k-1) \tag{15}$$

где $\chi^2_{1-\alpha}(k-1)$ – квантиль распределения χ^2 с $k-1$ степенями свободы порядка $1-\alpha$, где α заданный уровень значимости.

4.4 Интервальные оценки параметров распределения

Оценкой максимального правдоподобия для математического ожидания является среднее арифметическое: $\mu = \frac{1}{n}\sum\limits_{i=1}^{n} x_i$.

Оценка максимального правдоподобия для дисперсии вычисляется по формуле: $\sigma^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$.

Доверительным интервалом или интервальной оценкой числовой характеристики или параметра распределения θ с доверительной вероятностью γ называется интервал со случайными границами (θ_1, θ_2), содержащий параметр θ с вероятностью γ [15].

Функция распределения Стьюдента [16]:

$$T = \sqrt{n-1}\frac{\overline{x} - \mu}{\delta} \tag{16}$$

Функция плотности распределения χ^2 [17]:

$$f(x) = \begin{cases} 0, & x \le 0 \\ \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & x > 0 \end{cases} \tag{17}$$

Интервальная оценка математического ожидания [18]:

$$P = \left(\overline{x} - \frac{\sigma t_{1-\frac{a}{2}}(n-1)}{\sqrt{n-1}} < \mu < \overline{x} + \frac{\sigma t_{1-\frac{a}{2}}(n-1)}{\sqrt{n-1}} \right) = \gamma, \tag{18}$$

где $t_{1-\frac{a}{2}}$ – квантиль распределения Стьюдента порядка $1-\frac{a}{2}$.

Интервальная оценка дисперсии [16]:

$$P = \left(\frac{\sigma\sqrt{n}}{\sqrt{\chi^2_{1-\frac{a}{2}}(n-1)}} < \sigma < \frac{\sigma\sqrt{n}}{\sqrt{\chi^2_{\frac{a}{2}}(n-1)}} \right) = \gamma, \tag{19}$$

где $\chi^2_{1-\frac{a}{2}}$, $\chi^2_{\frac{a}{2}}$ – квантили распределения Стьюдента порядков $1-\frac{a}{2}$ и $\frac{a}{2}$ соответственно.

Асимптотическая интервальная оценка математического ожидания [16]:

$$P = \left(\overline{x} - \frac{\sigma u_{1-\frac{a}{2}}}{\sqrt{n}} < \mu < \overline{x} + \frac{\sigma u_{1-\frac{a}{2}}}{\sqrt{n}} \right) = \gamma, \tag{20}$$

где $u_{1-\frac{a}{2}}$ – квантиль нормального распределения $N(x, 0, 1)$ порядка $1-\frac{a}{2}$.

5 Реализация

Работы была выполнена на языке $Python3.7$. Для генерации выборок использовался модуль [1]. Для построения графиков использовалась библиотека matplotlib [2]. Функции распределения обрабатывались при помощи библиотеки scipy.stats [3]

6 Результаты

6.1 Вычисление коэффициента корреляции

Рис. 1: Графики двумерного нормального распределения(2) при $p = 0.0$

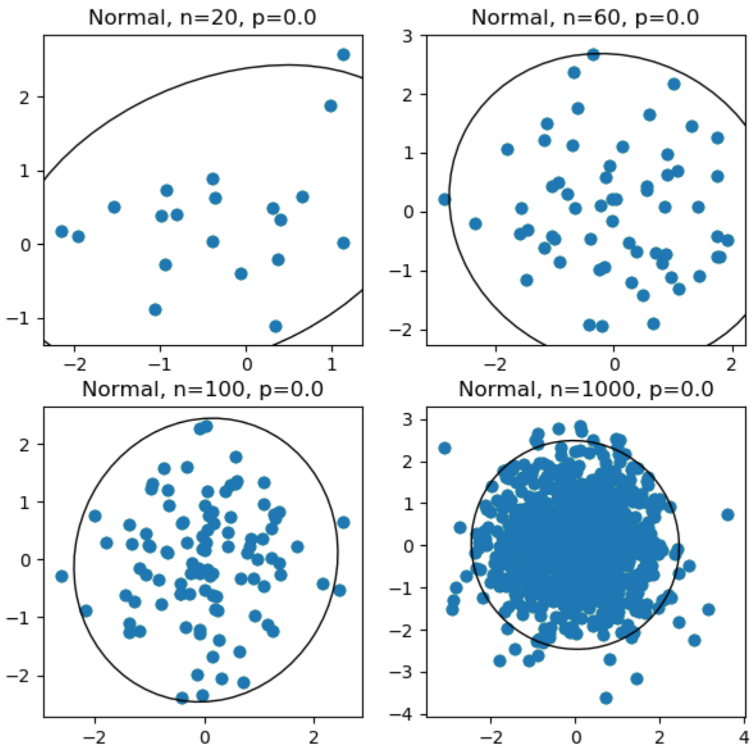

Таблица 1: Результаты для двумерного нормального распределения (2) при $p = 0.0$

Normal $n = 20,\ p = 0.0$			
	Pearson	Spearman	Quad
E	0.18892	0.14541	0.06000
E^2	0.05409	0.04186	0.02800
D	0.01840	0.02071	0.02440

Normal $n = 60,\ p = 0.0$			
	Pearson	Spearman	Quad
E	-0.04642	-0.05109	-0.03333
E^2	0.01080	0.00965	0.00667
D	0.00865	0.00704	0.00556

.

Normal $n = 100,\ p = 0.0$			
	Pearson	Spearman	Quad
E	-0.03469	-0.02805	-0.03200
E^2	0.00531	0.00539	0.00864
D	0.00411	0.00461	0.00762

Normal $n = 1000,\ p = 0.0$			
	Pearson	Spearman	Quad
E	0.00805	0.01039	0.00760
E^2	0.00094	0.00083	0.00094
D	0.00088	0.00073	0.00088

Рис. 2: Графики двумерного нормального распределения(2) при $p = 0.5$

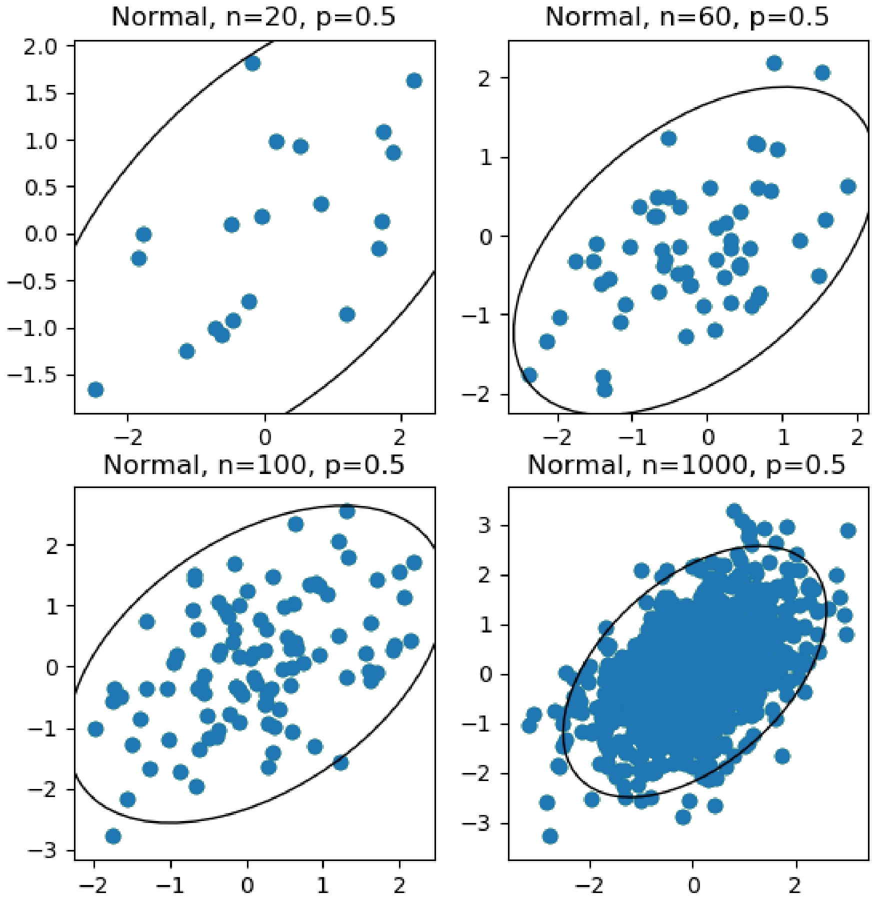

Таблица 2: Результаты для двумерного нормального распределения (2) при $p = 0.5$

Normal $n = 20,\ p = 0.5$			
	Pearson	Spearman	Quad
E	0.50363	0.52647	0.46000
E^2	0.27217	0.30152	0.22800
D	0.01853	0.02435	0.01640

Normal $n = 60,\ p = 0.5$			
	Pearson	Spearman	Quad
E	0.50847	0.47194	0.31333
E^2	0.26921	0.23710	0.12222
D	0.01067	0.01437	0.02404

.

Normal $n = 100,\ p = 0.5$			
	Pearson	Spearman	Quad
E	0.49628	0.47702	0.33200
E^2	0.25200	0.23489	0.12048
D	0.00570	0.00734	0.01026

Normal $n = 1000,\ p = 0.5$			
	Pearson	Spearman	Quad
E	0.49458	0.47938	0.33320
E^2	0.24515	0.23052	0.11207
D	0.00054	0.00071	0.00105

Рис. 3: График двумерного нормального распределения (2) при $p = 0.9$

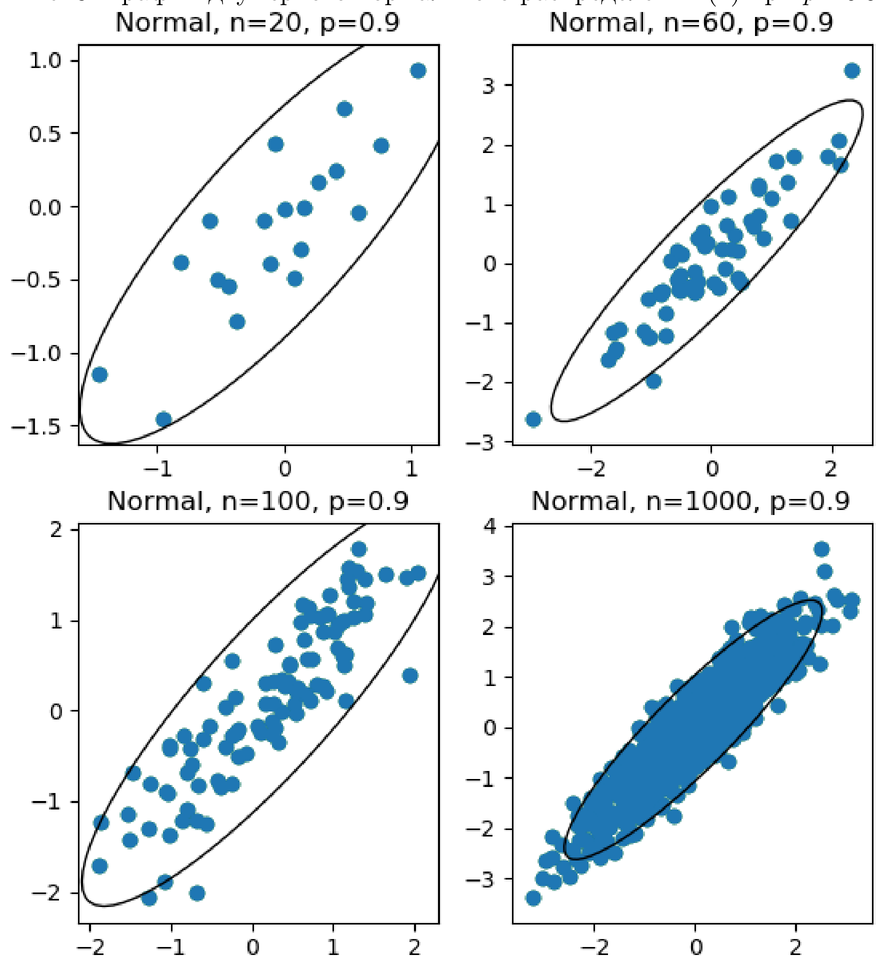

Таблица 3: Результаты для двумерного нормального распределения (2) при $p = 0.9$

Normal $n = 20,\ p = 0.9$			
	Pearson	Spearman	Quad
E	0.90154	0.85850	0.64000
E^2	0.81558	0.74275	0.44000
D	0.00281	0.00574	0.03040

Normal $n = 60,\ p = 0.9$			
	Pearson	Spearman	Quad
E	0.89761	0.88464	0.69333
E^2	0.80681	0.78457	0.48622
D	0.00112	0.00198	0.00551

.

Normal $n = 100,\ p = 0.9$			
	Pearson	Spearman	Quad
E	0.89624	0.88888	0.71600
E^2	0.80360	0.79094	0.51728
D	0.00037	0.00082	0.00462

Normal $n = 1000,\ p = 0.9$			
	Pearson	Spearman	Quad
E	0.89971	0.88953	0.71120
E^2	0.80951	0.79132	0.50603
D	0.00004	0.00005	0.00022

Рис. 4: Графики смеси двумерных нормальных распределений

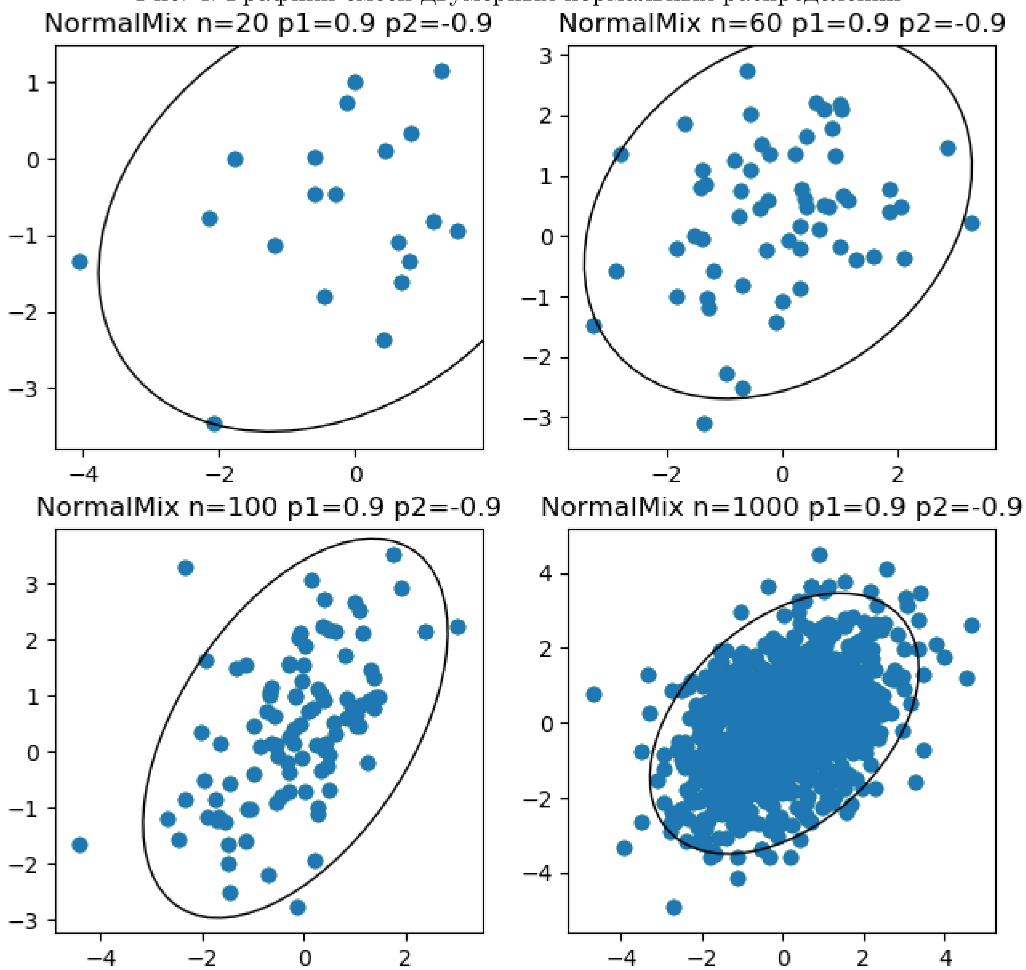

Таблица 4: Результаты для смеси двумерных нормальных распределений

NormalMix $n = 20,\ p_1 = 0.9,\ p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.90154	0.85850	0.64000
E^2	0.12784	0.14056	0.18000
D	0.03603	0.04226	0.11240

NormalMix $n = 60,\ p_1 = -0.9,\ p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.34444	0.33502	0.24000
E^2	0.13330	0.12744	0.08711
D	0.01466	0.01521	0.02951

.

NormalMix $n = 100,\ p_1 = 0.9,\ p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.42503	0.39751	0.26400
E^2	0.18615	0.16254	0.07584
D	0.00550	0.00453	0.00614

NormalMix $n = 1000,\ p_1 = 0.9,\ p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.38948	0.37427	0.25080
E^2	0.15242	0.14103	0.06380
D	0.00073	0.00095	0.00090

6.2 Оценки линий регрессии

Рис. 5: Графики линейной регрессии

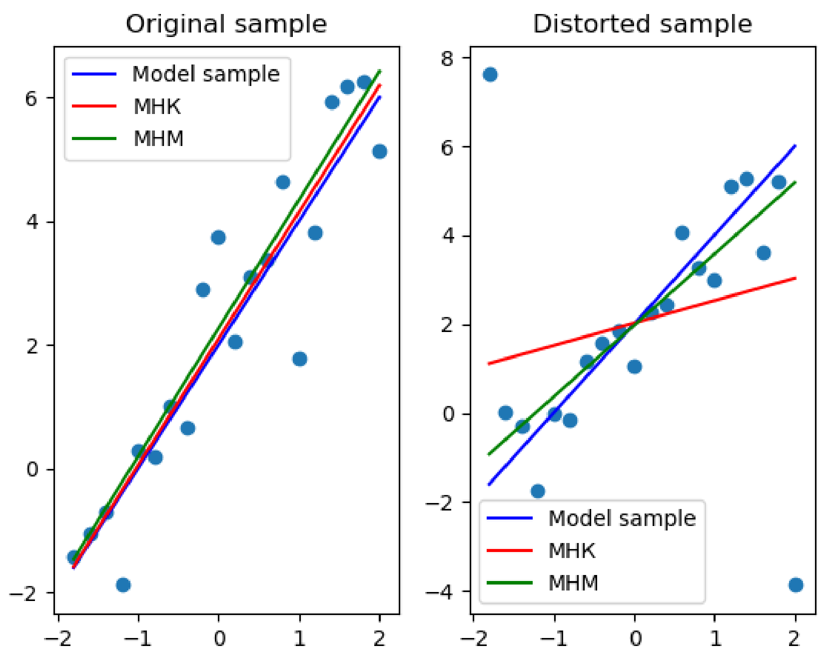

Таблица 5: Таблица оценок коэффициентов линейной регрессии без возмущёний

	$\hat{a}$	$\hat{b}$
МНК	2.000	2.000
МНМ	2.045	2.098

Таблица 6: Таблица оценок коэффициентов линейной регрессии с возмущёниями

	$\hat{a}$	$\hat{b}$
МНК	2.000	2.000
МНМ	0.505	2.023

6.3 Точечная оценка параметров распределения

6.3.1 Метод максимального правдоподобия

При подсчете оценок параметров закона нормального распределения методом максимального правдоподобия были получены следующие значения:

$$\hat{m}_{\text{МП}} = -0.1235$$
$$\hat{\sigma}^2_{\text{МП}} = 0.9877 \tag{21}$$

6.3.2 Критерий Пирсона

Таблица 7: Таблица вычислений χ^2

i	Δ_i	n_i	p_i	np_i	$n_i - np_i$	$\frac{(n_i-np_i)^2}{np_i}$
1	$(-\infty, -1.6605]$	5	0.0517	5.1670	-0.1670	0.0054
2	(-1.6605, -0.7749)	17	0.1950	19.4966	-2.4966	0.3197
3	(-0.7749, 0.1108)	40	0.3390	33.8962	6.1038	1.0991
4	(0.1108, 0.9965)	27	0.2778	27.7824	-0.7824	0.0220
5	$(0.9965, \infty)$	11	0.1284	12.8411	-1.8411	0.2640
$\sum$		100	1	100	0	1.8665

$$\chi^2_B = 1.8665$$

6.4 Интервальные оценки параметров распределения

Таблица 8: Результаты

Метод	n	μ	σ
На основе ММП	20	$[-0.33527, 0.40768]$	$[0.71165, 1.36749]$
	100	$[-0.20017, 0.16315]$	$[0.79980, 1.05748]$
Асимптотический	20	$[-0.37421, 0.44662]$	$[0.75586, 1.11704]$
	100	$[-0.19705, 0.16003]$	$[0.80188, 1.01998]$

7 Выводы

7.1 Вычисление коэффициента корреляции

По таблицам 1, 2, 3, 4, видно, что, при увеличении объёма выборки, подсчитанные коэффициенты корреляции стремятся к теоретическим.

Ближе всех к данному коэффициенту корреляции находится коэффициент Пирсона.

По графикам видно, что при уменьшении корреляции эллипс равновероятности стремится к окружности, а при увеличении растягивается.

7.2 Оценки линий регрессии

По графику 5 видно, что оба метода дают хорошую оценку коэффициентов линейной регрессии, если нет выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы мало влияют на оценку по МНМ. Ценой за это является бо́льшая по сравнению с МНК сложность вычисления.

7.3 Точечная оценка параметров распределения

В данной работе получено значение критерия согласия Пирсона $\chi^2_B = 1.8665$. Табличное значение квантиля $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(4) = 9,4877$ [14].

Значит $\chi^2_B < \chi^2_{0.95}(4)$, из этого следует, что основная гипотеза H_0 соотносится с выборкой на уровне $\alpha = 0.05$.

7.4 Интервальные оценки параметров распределения

По полученным результатам видно, что оба подхода дают лучший результат на выборках большого объема. Если рассматривать результаты для выборки объема $n = 20$ элементов, то видно, что интервал меньше и точнее в классической интервальной оценке.

8 Список литературы

[1] Модуль numpy - https://physics.susu.ru/vorontsov/language/numpy.html

[2] Модуль matplotlib - https://matplotlib.org/users/index.html

[3] Модуль scipy - https://docs.scipy.org/doc/scipy/reference/

[4] http://stu.sernam.ru/book_stat3.php?id=55

[5] Двумерное нормальное распределение: https://en.wikipedia.org/wiki/Multivariate_normal_distribution

[6] Коэффициент корреляции Пирса: http://statistica.ru/theory/koeffitsient-korrelyatsii/

[7] Коэффициент корреляции Спирмана: http://economic-definition.com/Exchange_Terminology/Koefficient_korrelyacii_Correlation_coefficient__eto.html

[8] Квадрантный коэффициент корреляции: https://www.researchgate.net/profile/Pavel_Smirnov8/publication/-316973167_Robastnye_metody_i_algoritmy_ocenivania_korrelacionnyh_harakteristik_dannyh_na_osnove_novyh_vysokoeffektivnyh_i_bystryh_robastnyh_ocenok_masstaba/links/591b019d458515695282-8a52/Robastnye-metody-i-algoritmy-ocenivania-korrelacionnyh-harakteristik-dannyh-na-osnove-novyh-vysokoeffektivnyh-i-bystryh-robastnyh-ocenok-masstaba.pdf#page=81

[9] https://en.wikipedia.org/wiki/Linear_regression

[10] http://www.cleverstudents.ru/articles/mnk.html

[11] Шевляков Г. Л. Лекции по математической статистике, 2019.

[12] Вероятностные разделы математики. Учебник для бакалавров технических направлений. //Под ред. Максимова Ю.Д. - СПб.:"Иван Федоров 2001. - 592 с.

[13] https://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test

[14] Таблица значений χ^2 - https://en.wikipedia.org/wiki/Student%27s_t-distribution

[15] https://en.wikipedia.org/wiki/Confidence_interval

[16] https://en.wikipedia.org/wiki/Student%27s_t-distribution

[17] https://en.wikipedia.org/wiki/Chi-squared_distribution

[18] Шевляков Г. Л. Лекции по математической статистике, 2019.

9 Приложения

Код отчёта: https://github.com/MisterProper9000/MatStatLabs/blob/master/MatStatLab5/MatStatLab8.tex
Код лаборатрной: https://github.com/MisterProper9000/MatStatLabs/blob/master/MatStatLab5/MatStatLab8.py