Санкт-Петербургский Политехнический
Университет им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

Сводный отчет по лабораторным 5-8.

3 курс, группа 3630102/70301

Студент группы 3630102/70301 Лебедев К.С.

Преподаватель Баженов А. Н.

Санкт-Петербург
2020 г.

Содержание

1 Список иллюстраций

2 Постановка задачи

1. Построить выборки 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. Повторить вычисления для смеси двумерных нормальных распределений:

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

2. Найти оценки линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек отрезка $[-1.8;\ 2]$ с равномерным шагом. 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
   - Критерий наименьших модулей

   Проделать то же самое для выборки, у которой в значении y_1 и y_{20} вносятся возмущения 10 и -10 соответственно.

3. 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.

4. 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.

3 Теория

3.1 Вычисление коэффициента корреляции

1. Двумерное нормально распределение [11]:

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

2. Коэффициент корреляции Пирсона [12]:

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

3. Коэффициент корреляции Спирмана [13]:

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

4. Квадрантный коэффициент корреляции [14]:

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

3.2 Оценка регрессии

Простая линейная регрессия [4]:

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

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

1. *Метод наименьших квадратов*
   Критерий – минимизация функции:

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

   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 :

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

   Метод наименьших квадратов является несмещённой оценкой.

2. *Метод наименьших модулей*
   Критерий наименьших модулей – заключается в минимизации следующей функции :

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

   МНМ-оценки обладают свойством робастности

3.3 Точечная оценка параметров распределения

1. *Метод максимального правдоподобия*
   Метод максимального правдоподобия – метод оценивания неизвестного параметра путём максимизации 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

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

   Оценкой максимального правдоподобия будем называть такое значение $\hat{\theta}_{\text{МП}}$ 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:

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

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 Δ_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 :

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

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

3.4 Интервальные оценки

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 [5].

Функция распределения Стьюдента :

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

Функция плотности распределения χ^2 :

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

Интервальная оценка математического ожидания :

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

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

Интервальная оценка дисперсии:

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

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 :

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

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

4 Реализация

Работы была выполнена на языке *Python*3.6. Для генерации выборок использовался модуль [1]. Для построения графиков использовалась библиотека matplotlib [2]. Функции распределения обрабатывались при помощи библиотеки scipy.stats [3]

5 Результаты

5.1 Вычисление коэффицентов кореляции

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

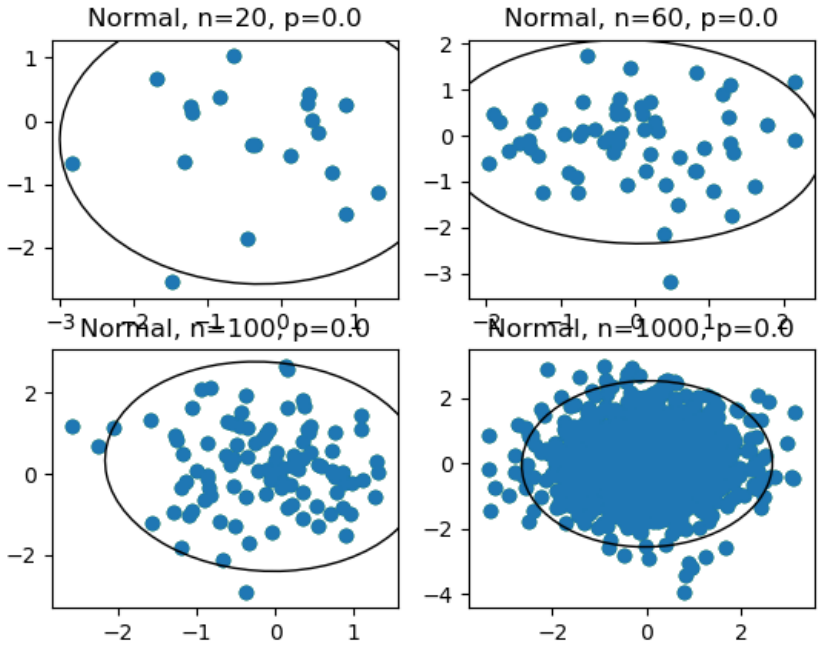

Рис. 2: Графики двумерного нормального распределения(2) при $p = 0.5$

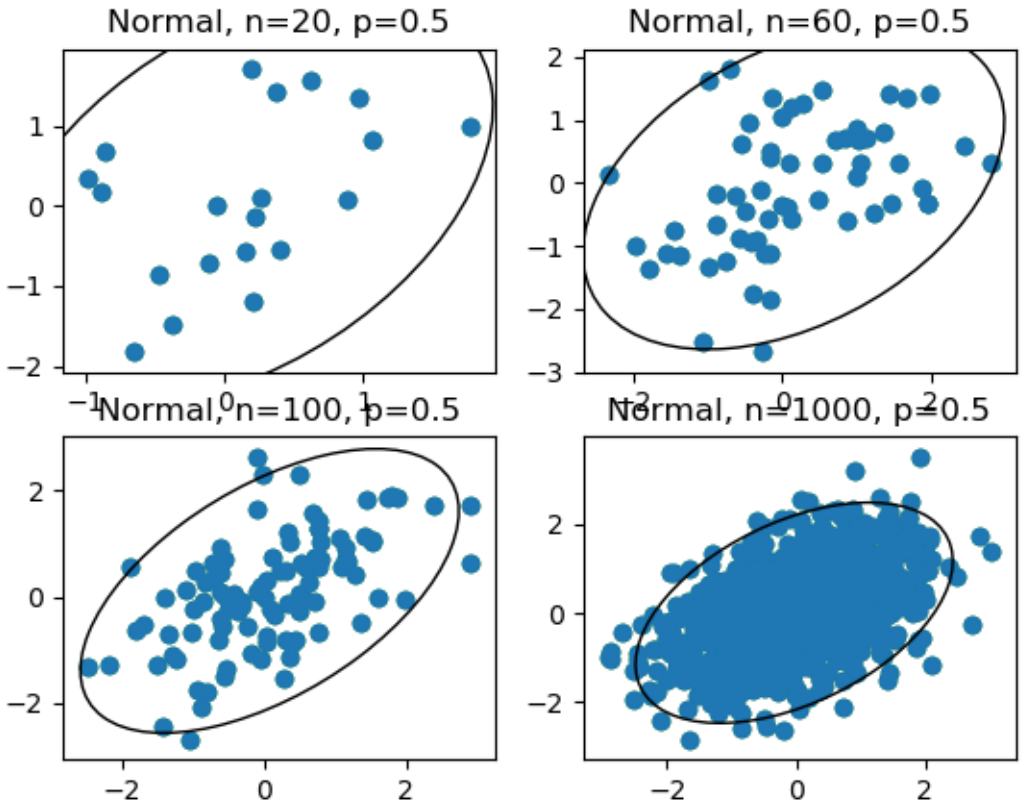

Таблица 2: Результаты для двумерного нормального распределения (2) при $p = 0.5$

Normal $n = 20$, $p = 0.5$			
	Pearson	Spearman	Quad
E	0.56562	0.53910	0.28000
E^2	0.32763	0.29964	0.12800
D	0.00770	0.00901	0.04960

Normal $n = 60$, $p = 0.5$			
	Pearson	Spearman	Quad
E	0.44858	0.44378	0.28667
E^2	0.20459	0.20291	0.10356
D	0.00336	0.00597	0.02138

Normal $n = 100$, $p = 0.5$			
	Pearson	Spearman	Quad
E	0.50044	0.49137	0.31600
E^2	0.25552	0.24642	0.10960
D	0.00509	0.00497	0.00974

Normal $n = 1000$, $p = 0.5$			
	Pearson	Spearman	Quad
E	0.49722	0.48120	0.32320
E^2	0.24757	0.23175	0.10468
D	0.00034	0.00020	0.00022

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

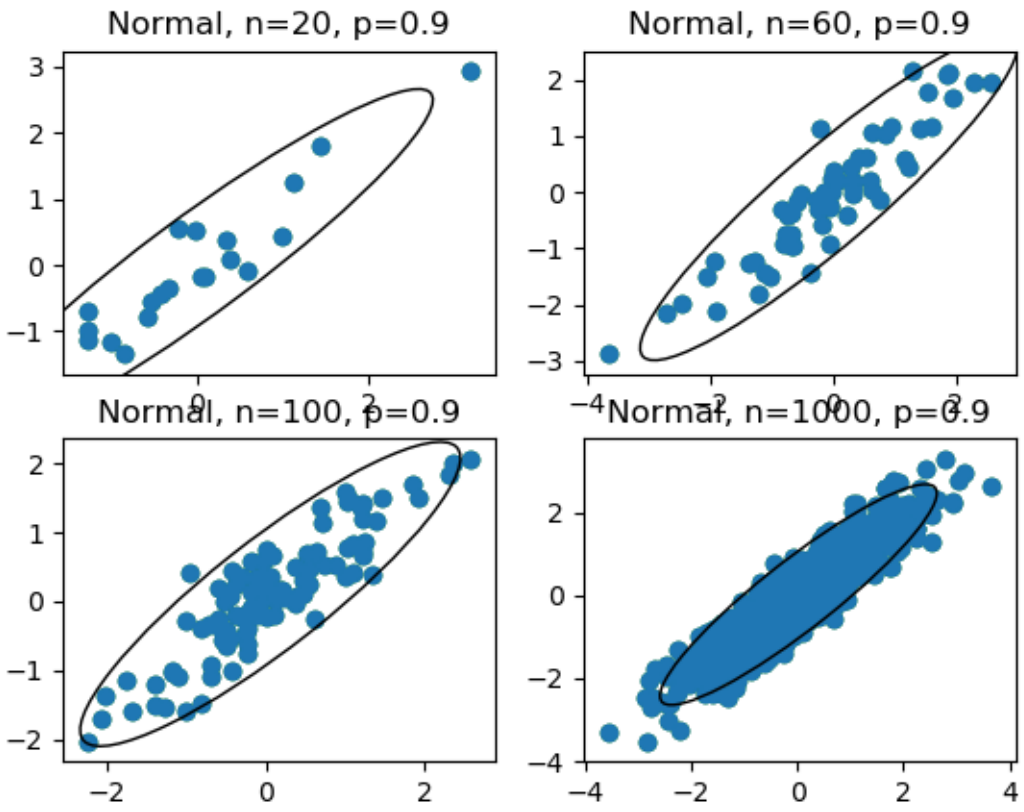

Таблица 3: Результаты для двумерного нормального распределения (2) при $p = 0.9$

Normal $n = 20$, $p = 0.9$			
	Pearson	Spearman	Quad
E	0.90608	0.88421	0.76000
E^2	0.82263	0.78440	0.58400
D	0.00165	0.00257	0.00640

Normal $n = 60$, $p = 0.9$			
	Pearson	Spearman	Quad
E	0.89748	0.87538	0.68667
E^2	0.80573	0.76683	0.48844
D	0.00027	0.00054	0.01693

Normal $n = 100$, $p = 0.9$			
	Pearson	Spearman	Quad
E	0.88816	0.87734	0.70000
E^2	0.78938	0.77048	0.49680
D	0.00055	0.00076	0.00680

Normal $n = 1000$, $p = 0.9$			
	Pearson	Spearman	Quad
E	0.90323	0.89562	0.71400
E^2	0.81584	0.80215	0.50995
D	0.00001	0.00001	0.00015

Рис. 4: Графики смеси двумерных нормальных распределений

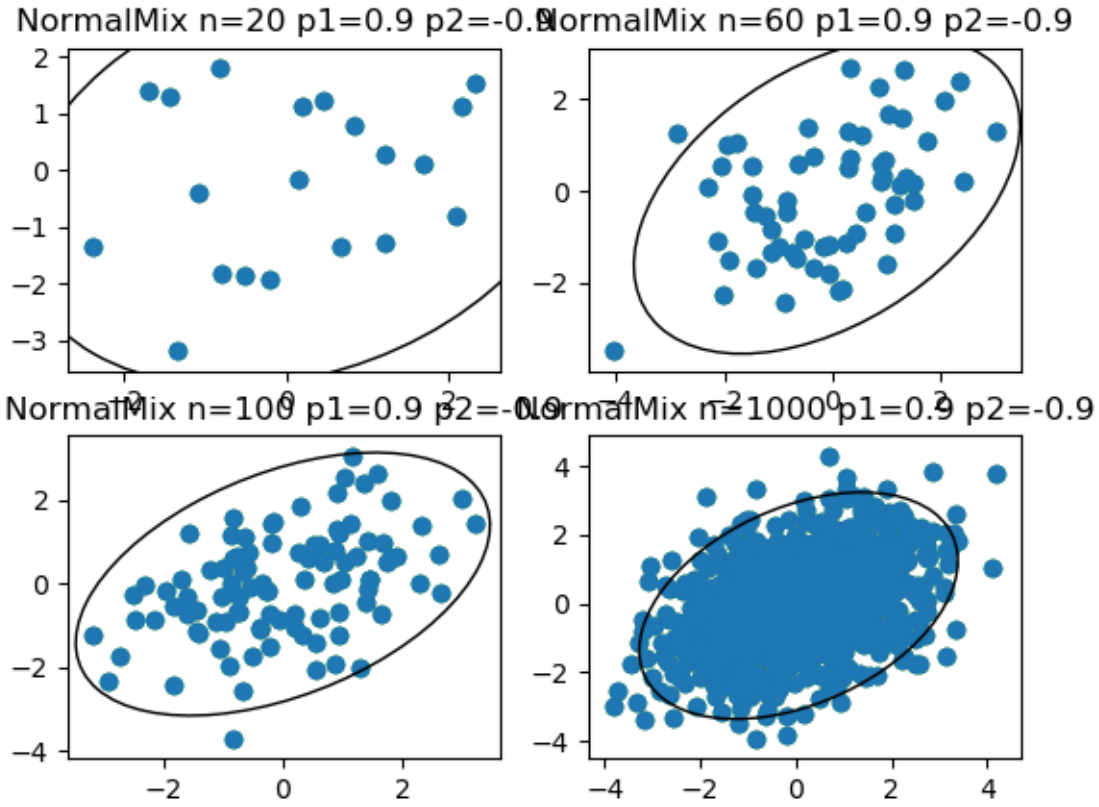

Таблица 4: Результаты для смеси двумерных нормальных распределений

NormalMix $n = 20$, $p_1 = 0.9$, $p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.31195	0.30391	0.26000
E^2	0.12850	0.10996	0.10800
D	0.03118	0.01760	0.04040

NormalMix $n = 60$, $p_1 = -0.9$, $p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.36558	0.34972	0.22000
E^2	0.14986	0.13751	0.06622
D	0.01621	0.01520	0.01782

NormalMix $n = 100$, $p_1 = 0.9$, $p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.41825	0.39421	0.30400
E^2	0.17956	0.16217	0.09792
D	0.00462	0.00677	0.00550

NormalMix $n = 1000$, $p_1 = 0.9$, $p_2 = -0.9$			
	Pearson	Spearman	Quad
E	0.37713	0.35846	0.23720
E^2	0.14274	0.12888	0.05707
D	0.00052	0.00038	0.00081

5.2 Оценка линейной регрессии

Рис. 5: Графики линейной регрессии

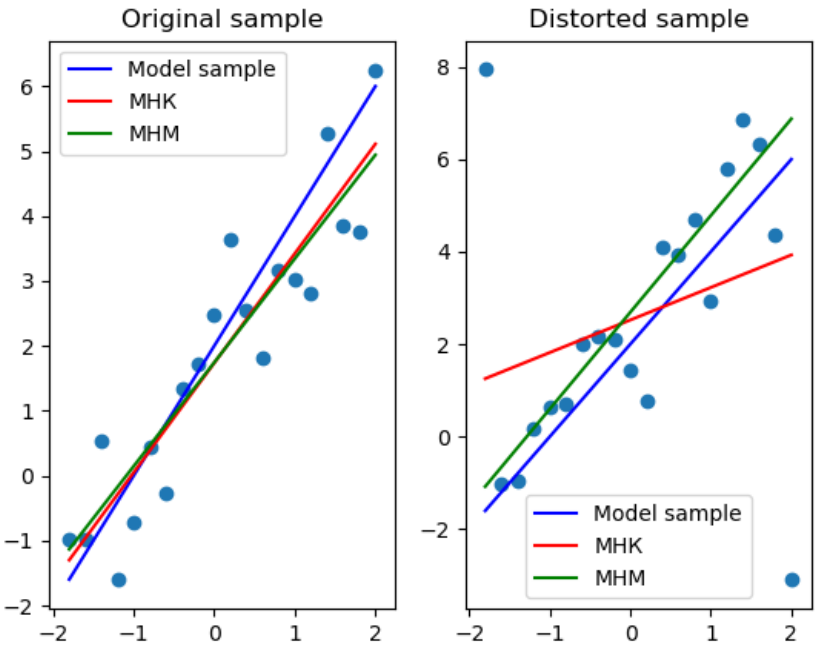

Таблица 5: Таблица оценок коэффициентов линейной регрессии без возмущёний

	$\hat{a}$	$\hat{b}$
МНК	2.000	2.000
МНМ	1.926	2.372

Таблица 6: Таблица оценок коэффициентов линейной регрессии с возмущёниями

	$\hat{a}$	$\hat{b}$
МНК	2.000	2.000
МНМ	0.755	1.938

5.3 Метод максимального правдоподобия

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

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

5.4 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	$(-\infty, -1.8018]$	4	0.0531	5.3136	-1.3136	0.3247
2	(-1.8018, -0.4727)	31	0.3025	30.2454	0.7546	0.0188
3	(-0.4727, 0.8564)	47	0.4535	45.3525	1.6475	0.0599
4	$(0.8564, \infty)$	18	0.1909	19.0886	-1.0886	0.0621
$\sum$		100	1	100	0	0.4655

$$\chi^2_B = 0.4655$$

5.5 Интервальная оценка параметров распределения

Таблица 8: Результаты интервальной оценки

Метод	n	μ	σ
На основе ММП	20	$[-0.60643, 0.16116]$	$[0.73526, 1.41287]$
	100	$[-0.06206, 0.30951]$	$[0.81800, 1.08155]$
Асимптотический	20	$[-0.64667, 0.20140]$	$[0.68468, 1.25038]$
	100	$[-0.05888, 0.30633]$	$[0.80817, 1.05515]$

6 Выводы

6.1 Вычисление коэфицентов корреляции

При увеличении объёма выборки, подсчитанные коэффициенты корреляции стремятся к теоретическим. По графикам видно, что при уменьшении корреляции эллипс равновероятности стремится к окружности, а при увеличении растягивается.

6.2 Оценки линий регрессии

По графику 5 видно, что оба метода дают хорошую оценку коэффициентов линейной регрессии, если нет выбросов. Так же видно, что выбросы больше влияют на оценку по МНК, нежели на оценку по МНМ.

6.3 Точечная оценка параметров распределения

В данной работе получено значение критерия согласия Пирсона $\chi^2_B = 0.4655$ Табличное значение квантиля $\chi^2_{1-\alpha}(k-1) = \chi^2_{0.95}(4) = 9,4877$ [8].

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

6.4 Интервальная оценка параметров распределения

При увеличении объёма выборки точность увеличивается для обоих методов.

7 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] https://en.wikipedia.org/wiki/Linear_regression

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

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

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

[8] Таблица значений χ^2 - https://ru.wikipedia.org/wiki/%D0%9A%D0%B2%D0%B0%D0%BD%D1%82%D0%B8%B8_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82

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

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

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

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

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

[14] 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