Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт Кафедра «Прикладная математика»

Отчет по лабораторной работе №1 по дисциплине «Математическая статистика»

Выполнила студентка группы 5030102/00101 Проверил Доцент, к.ф.-м.н.

Еремина Ксения Игоревна

Баженов Александр Николаевич

Содержание

1	Пос	становка задачи	4
2	Teo	рия	5
	2.1	Рассматриваемые распределения	5
	2.2	Гистограмма	5
		2.2.1 Построение гистограммы	5
	2.3	Вариационный ряд	5
	2.4	Выборочные числовые характеристики	5
		2.4.1 Характеристики положения	5
		2.4.2 Характеристики рассеяния	6
	2.5	Боксплот Тьюки	6
		2.5.1 Построение	6
	2.6	Теоретическая вероятность выбросов	6
	2.7	Эмпирическая функция распределения	7
		2.7.1 Статистический ряд	7
		2.7.2 Эмпирическая функция распределения	7
		2.7.3 Нахождение э. ф. р	7
	2.8	Оценки плотности вероятности	7
	_	2.8.1 Определение	7
		2.8.2 Ядерные оценки	7
3	Pea	лизация	9
4	Рез	зультаты	10
	4.1	Гистограммы и графики плотности распределения	10
	4.2	Характеристики положения и рассеяния	11
	4.3	Боксплот Тьюки	13
	4.4	Доля выбросов	15
	4.5	Теоретическая вероятность выбросов	16
	4.6	Эмпирическая функция распределения	16
	4.7	Ядерные оценки плотности распределения	17
5	Обо	суждение	22
	5.1	Гистограммы	22
	5.2	Характеристики положения и рассеяния	22
	5.3	Доля и теоретическая вероятность выбросов	22
	5.4	Эмпирическая функция и ядерные оценки плотности распределения	22
6	Прі	иложение	23

Список иллюстраций

1	Нормальное распределение	10
2	Распределение Коши	10
3	Распределение Пуассона	
4	Равномерное распределение	11
5	Нормальное распределение	13
6	Распределение Коши	14
7	Распределение Пуассона	14
8	Равномерное распределение	
9	Нормальное распределение	16
10	Распределение Коши	16
11	Распределение Пуассона	
12	Равномерное распределение	17
13	Нормальное распределение, n=20	17
14	Нормальное распределение, n=60	18
15	Нормальное распределение, n=100	18
16	Распределение Коши, n=20	18
17	Распределение Коши, n=60	19
18	Распределение Коши, n=100	19
19	Распределение Пуассона, n=20	19
20	Распределение Пуассона, n=60	
21	Распределение Пуассона, n=100	
22	Равномерное распределение, n=20	
23	Равномерное распределение, n=60	21
24	Равномерное распределение п-100	21

Список таблиц

1	Таблица распределения
2	Нормальное распределение
3	Распределение Коши
4	Распределение Пуассона
5	Равномерное распределение
6	Доля выбросов
7	Теоретическая вероятность выбросов

1 Постановка задачи

Для 4 распределений:

- Нормальное распределение N(x, 0, 1)
- Распределение Коши C(x,0,1)
- Распределение Пуассона P(k, 10)
- Равномерное распределение $U(x,-\sqrt{3},\sqrt{3})$
- 1. Сгенерировать выборки размером 10, 50 и 1000 элементов. Построить на одном рисунке гистограмму и график плотности распределения.
- 2. Сгенерировать выборки размером 10, 100 и 1000 элементов. Для каждой выборки вычислить следующие статистические характеристики положения данных: \overline{x} , $med~x, z_R, z_Q, z_{tr}$. Повторить такие вычисления 1000 раз для каждой выборки и найти среднее характеристик положения и их квадратов:

$$E(z) = \overline{z} \tag{1}$$

Вычислить оценку дисперсии по формуле:

$$D(z) = \overline{z^2} - \overline{z}^2 \tag{2}$$

Представить полученные данные в виде таблиц.

- 3. Сгенерировать выборки размером 20 и 100 элементов. Построить для них боксплот Тьюки. Для каждого распределения определить долю выбросов экспериментально (сгенерировав выборку, соответствующую распределению 1000 раз, и вычислив среднюю долю выбросов) и сравнить с результатами, полученными теоретически.
- 4. Сгенерировать выборки размером 20, 60 и 100 элементов. Построить на них эмпирические функции распределения и ядерные оценки плотности распределения на отрезке [-4;4] для непрерывных распределений и на отрезке [6;14] для распределения Пуассона.

2 Теория

2.1 Рассматриваемые распределения

Плотности:

• Нормальное распределение

$$N(x,0,1) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{3}$$

• Распределение Коши

$$C(x,0,1) = \frac{1}{\pi} \frac{1}{x^2 + 1} \tag{4}$$

• Распределение Пуассона

$$P(k,10) = \frac{10^k}{k!}e^{-10} \tag{5}$$

• Равномерное распределение

$$U(x, -\sqrt{3}, \sqrt{3}) = \begin{cases} \frac{1}{2\sqrt{3}} & \text{при} \quad x \in [-\sqrt{3}, \sqrt{3}] \\ 0 & \text{при} \quad x \notin [-\sqrt{3}, \sqrt{3}] \end{cases}$$
 (6)

2.2 Гистограмма

2.2.1 Построение гистограммы

Множество значений, которое может принимать элемент выборки, разбивается на несколько интервалов. Чаще всего эти интервалы берут одинаковыми, но это не является строгим требованием. Эти интервалы откладываются на горизонтальной оси, затем над каждым рисуется прямоугольник. Если все интервалы были одинаковыми, то высота каждого прямоугольника пропорциональна числу элементов выборки, попадающих в соответствующий интервал. Если интервалы разные, то высота прямоугольника выбирается таким образом, чтобы его площадь была пропорциональна числу элементов выборки, которые попали в этот интервал [1].

2.3 Вариационный ряд

Вариационным ряд - последовательность элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются [2, с. 409].

2.4 Выборочные числовые характеристики

2.4.1 Характеристики положения

• Выборочное среднее

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{7}$$

• Выборочная медиана

$$med x = \begin{cases} x_{(l+1)} & \text{при} \quad n = 2l+1\\ \frac{x_{(l)} + x_{(l+1)}}{2} & \text{при} \quad n = 2l \end{cases}$$
 (8)

• Полусумма экстремальных выборочных элементов

$$z_R = \frac{x_{(1)} + x_{(n)}}{2} \tag{9}$$

• Полусумма квартилей

Выборочная квартиль z_p порядка p определяется формулой

$$z_p = \begin{cases} x_{([np]+1)} & \text{при } np \text{ дробном,} \\ x_{(np)} & \text{при } np \text{ целом.} \end{cases}$$
 (10)

Полусумма квартилей

$$z_Q = \frac{z_{1/4} + z_{3/4}}{2} \tag{11}$$

• Усечённое среднее

$$z_{tr} = \frac{1}{n-2r} \sum_{i=r+1}^{n-r} x_{(i)}, \quad r \approx \frac{n}{4}$$
 (12)

2.4.2 Характеристики рассеяния

Выборочная дисперсия

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{13}$$

2.5 Боксплот Тьюки

2.5.1 Построение

Границами ящика служат первый и третий квартили, линия в середине ящика — медиана. Концы усов — края статистически значимой выборки (без выбросов). Длину «усов»:

$$X_1 = Q_1 - \frac{3}{2}(Q_3 - Q_1), \quad X_2 = Q_3 + \frac{3}{2}(Q_3 - Q_1),$$
 (14)

где X_1 — нижняя граница уса, X_2 — верхняя граница уса, Q_1 — первый квартиль, Q_3 — третий квартиль. Данные, выходящие за границы усов (выбросы), отображаются на графике в виде маленьких кружков .

2.6 Теоретическая вероятность выбросов

Можно вычислить теоретические первый и третий квартили распределений $Q_1^{\rm \scriptscriptstyle T}$ и $Q_3^{\rm \scriptscriptstyle T}$ соответственно. По формуле (14) - теоретические нижнюю и верхнюю границы уса $X_1^{\rm \scriptscriptstyle T}$ и $X_2^{\rm \scriptscriptstyle T}$. Выбросы - величины x:

$$\begin{bmatrix}
x < X_1^{\mathrm{T}} \\
x > X_2^{\mathrm{T}}
\end{bmatrix}$$
(15)

Теоретическая вероятность выбросов:

• для непрерывных распределений

$$P_{\rm B}^{\rm T} = P(x < X_1^{\rm T}) + P(x > X_2^{\rm T}) = F(X_1^{\rm T}) + \left(1 - F(X_2^{\rm T})\right) \tag{16}$$

• для дискретных распределений

$$P_{\scriptscriptstyle \rm B}^{\scriptscriptstyle \rm T} = P(x < X_1^{\scriptscriptstyle \rm T}) + P(x > X_2^{\scriptscriptstyle \rm T}) = \left(F(X_1^{\scriptscriptstyle \rm T}) - P(x = X_1^{\scriptscriptstyle \rm T}) \right) + \left(1 - F(X_2^{\scriptscriptstyle \rm T}) \right) \tag{17}$$

Выше $F(X) = P(x \le X)$ - функция распределения.

2.7 Эмпирическая функция распределения

2.7.1 Статистический ряд

Статистическим ряд – последовательность различных элементов выборки $z_1, z_2, ..., z_k$, расположенных в возрастающем порядке с указанием частот $n_1, n_2, ..., n_k$, с которыми эти элементы содержатся в выборке. Обычно записывается в виде таблицы.

2.7.2 Эмпирическая функция распределения

Эмпирической (выборочной) функцией распределения (э. ф. р.) называется относительная частота события X < x, полученная по данной выборке:

$$F_n^*(x) = P^*(X < x). (18)$$

2.7.3 Нахождение э. ф. р.

Для получения относительной частоты $P^*(X < x)$ просуммируем в статистическом ряде, построенном по данной выборке, все частоты n_i , для которых элементы z_i статистического ряда меньше x. Тогда $P^*(X < x) = \frac{1}{n} \sum_{z_i < x} n_i$. Получаем

$$F^*(x) = \frac{1}{n} \sum_{z_i < x} n_i. \tag{19}$$

 $F^*(x)$ — функция распределения дискретной случайной величины X^* , заданной таблицей распределения

X^*	z_1	z_2	 z_k
P	$\frac{n_1}{\underline{}}$	$\frac{n_2}{\underline{}}$	 $\frac{n_k}{n_k}$
_	n	n	 n

Таблица 1: Таблица распределения

Эмпирическая функция распределения является оценкой, т. е. приближённым значением, генеральной функции распределения

$$F_n^*(x) \approx F_X(x). \tag{20}$$

2.8 Оценки плотности вероятности

2.8.1 Определение

Оценкой плотности вероятности f(x) называется функция $\widehat{f}(x)$, построенная на основе выборки, приближённо равная f(x)

$$\widehat{f}(x) \approx f(x).$$
 (21)

2.8.2 Ядерные оценки

Представим оценку в виде суммы с числом слагаемых, равным объёму выборки:

$$\widehat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^n K\left(\frac{x - x_i}{h_n}\right). \tag{22}$$

Здесь функция K(u), называемая ядерной (ядром), непрерывна и является плотностью вероятности, x_1, \dots, x_n — элементы выборки, $\{h_n\}$ — любая последовательность положительных чисел, обладающая свойствами

$$h_n \xrightarrow[n \to \infty]{} 0; \qquad \frac{h_n}{n^{-1}} \xrightarrow[n \to \infty]{} \infty.$$
 (23)

Такие оценки называются непрерывными ядерными

Гауссово (нормальное) ядро:

$$K(u) = \frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}. (24)$$

Правило Сильвермана

$$h_n = 1.06\hat{\sigma}n^{-1/5},\tag{25}$$

где $\hat{\sigma}$ - выборочное стандартное отклонение.

3 Реализация

Лабораторная работа выполнена на языке Python версии 3.8 в среде разработки PyCharm. Использовались дополнительные библиотеки:

- \bullet scipy
- numpy
- \bullet matplotlib
- \bullet math

4 Результаты

4.1 Гистограммы и графики плотности распределения

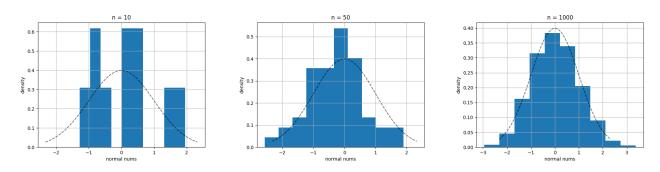


Рис. 1: Нормальное распределение

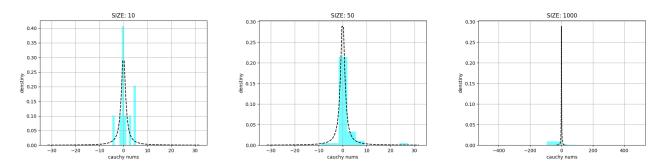


Рис. 2: Распределение Коши

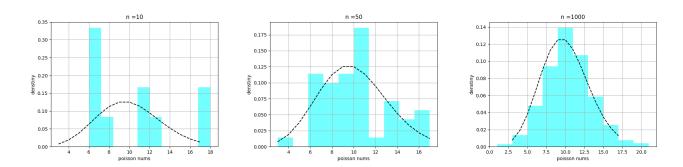


Рис. 3: Распределение Пуассона

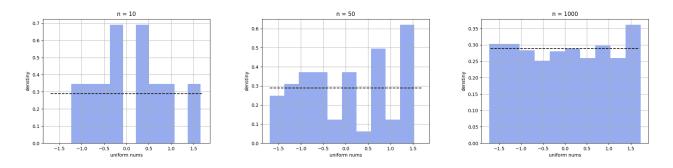


Рис. 4: Равномерное распределение

4.2 Характеристики положения и рассеяния

		1	ъ		71
	<x></x>	med x	zR	zQ	Ztr
n=10					
Normal E(z)	0.000826	0.007095	-0.002096	0.311883	0.276208
Normal D(z)	0.103509	0.142529	0.185601	0.1278	0.11815
$E(z) \pm \sqrt{D(z)}$	[-0.320902;	[-0.370435;	[-0.43291;	[-0.045608;	[-0.067522;
	0.322554]	0.384625]	0.428718]	0.669374]	0.619938]
n=100					
Normal E(z)	0.001901	-0.000416	0.008904	0.018747	0.027085
Normal D(z)	0.010657	0.017139	0.095083	0.013155	0.012952
$E(z) \pm \sqrt{D(z)}$	[-0.101332;	[-0.131332;	[-0.299451;	[-0.095948;	[-0.086722;
,	0.105134]	0.1305]	[0.317259]	0.133442]	0.140892]
n=1000					
Normal E(z)	0.000169	0.000363	0.00097	0.001855	0.002716
Normal D(z)	0.00102	0.001522	0.058933	0.001288	0.001181
$E(z) \pm \sqrt{D(z)}$	[-0.031768;	[-0.03865;	[-0.241791;	[-0.034034;	[-0.03165;
	0.032106]	0.039376]	0.243731]	[0.037744]	0.037082]

Таблица 2: Нормальное распределение

	<x></x>	med x	zR	zQ	Ztr
n=10					
Cauchy E(z)	-1.525697	-0.020663	-7.536897	1.05869	0.655965
Cauchy D(z)	739.187373	0.360723	18396.710165	5.032606	1.38404
$E(z) \pm \sqrt{D(z)}$	[-28.713698;	[-0.621265;	[-143.17137;	[-1.184657;	[-0.520487;
	25.662304]	[0.579939]	128.097576]	3.302037]	1.832417]
n=100					
Cauchy E(z)	-0.284393	-0.004505	-13.293061	0.035872	0.041822
Cauchy D(z)	275.816143	0.025083	673320.932083	0.051133	0.026515
$E(z) \pm \sqrt{D(z)}$	[-16.892106;	[-0.162881;	[-833.854413;	[-0.190254;	[-0.121012;
	16.32332]	0.153871]	807.268291]	0.261998]	[0.204656]
n=1000					
Cauchy E(z)	0.212424	0.000775	121.549193	0.004567	0.004748
Cauchy D(z)	1023.878903	0.002212	253459308.779892	0.004833	0.002351
$E(z) \pm \sqrt{D(z)}$	[-31.785684;	[-0.046257;	[-15798.856228;	[-0.064953;	[-0.043739;
, ,	32.210532]	0.047807]	16041.954614]	0.074087]	0.053235]

Таблица 3: Распределение Коши

	<x></x>	med x	zR	zQ	Ztr
n=10				-	
Poisson E(z)	9.9671	9.847	10.2465	10.916	10.7555
Poisson D(z)	0.957568	1.375591	1.820988	1.318944	1.222914
$E(z) \pm \sqrt{D(z)}$	[8.988546;	[8.674144;	[8.89706;	[9.767547;	[9.649646;
	10.945654]	11.019856]	11.59594]	12.064453]	11.861354]
n=100					
Poisson E(z)	10.00494	9.843	10.9445	9.971	9.94612
Poisson D(z)	0.105059	0.228351	1.00417	0.166659	0.125901
$E(z) \pm \sqrt{D(z)}$	[9.680812;	[9.365139;	[9.942417;	[9.562761;	[9.591295;
	10.329068]	10.320861]	11.946583]	10.379239]	10.300945]
n=1000					
Poisson E(z)	9.999381	9.994	11.678	9.997	9.863034
Poisson D(z)	0.010293	0.005464	0.684816	0.001991	0.011335
$E(z) \pm \sqrt{D(z)}$	[9.897927;	[9.920081;	[10.850464;	[9.952379;	[9.756568;
	10.100835]	10.067919]	12.505536]	10.041621]	9.9695]

Таблица 4: Распределение Пуассона

	<x></x>	med x	zR	zQ	Ztr
n=10					
Uniform E(z)	-0.009008	-0.005753	0.003201	0.301986	0.301722
Uniform D(z)	0.10032	0.220911	0.048237	0.130189	0.149299
$E(z) \pm \sqrt{D(z)}$	[-0.325741;	[-0.475765;	[-0.216428;	[-0.058831;	[-0.08467;
	0.307725]	0.464259]	0.22283]	0.662803]	0.688114]
n=100					
Uniform E(z)	0.003794	0.004227	0.000161	0.023149	0.038556
Uniform D(z)	0.009571	0.027486	0.000559	0.014503	0.018722
$E(z) \pm \sqrt{D(z)}$	[-0.094037;	[-0.161562;	[-0.023482;	[-0.097279;	[-0.098272;
	0.101625]	0.170016]	0.023804]	0.143577]	0.175384]
n=1000					
Uniform E(z)	-0.001435	-0.003582	4.3e-05	0.000131	0.001435
Uniform D(z)	0.000938	0.003066	5e-06	0.001341	0.001921
$E(z) \pm \sqrt{D(z)}$	[-0.032062;	[-0.058953;	[-0.002193;	[-0.036489;	[-0.042394;
	0.029192]	0.051789]	0.002279]	0.036751]	0.045264]

Таблица 5: Равномерное распределение

4.3 Боксплот Тьюки

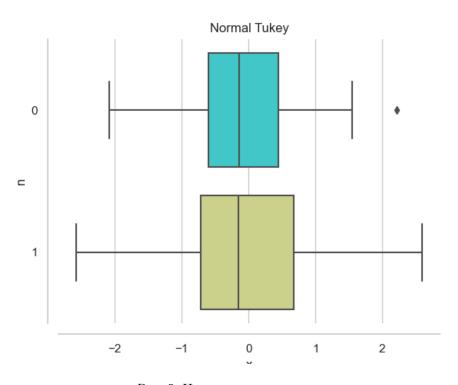


Рис. 5: Нормальное распределение

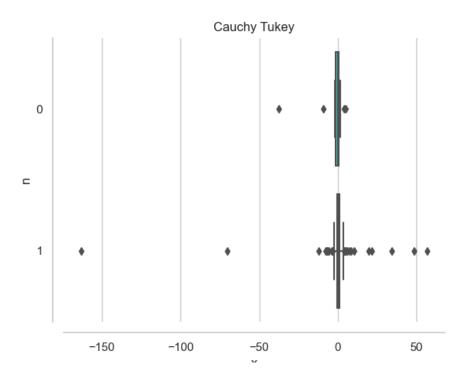


Рис. 6: Распределение Коши

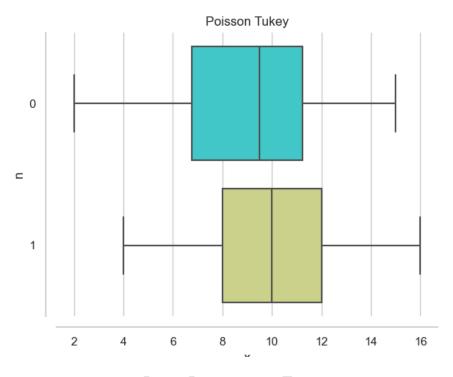


Рис. 7: Распределение Пуассона

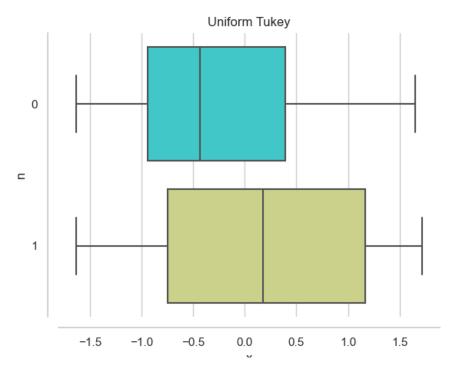


Рис. 8: Равномерное распределение

4.4 Доля выбросов

Выборка	Доля выбросов
normal n = 20	0.05
normal n = 100	0.01
cauchy n = 20	0.15
cauchy n = 100	0.13
poisson $n = 20$	0.02
poisson $n = 100$	0.01
uniform $n = 20$	0
uniform $n = 100$	0

Таблица 6: Доля выбросов

4.5 Теоретическая вероятность выбросов

Распределение	Q_1^T	Q_3^T	X_1^T	X_2^T	P_B^T
Normal	-0.674	0.674	-2.698	2.698	0.007
Cauchy	-1	1	-4	4	0.156
Poisson	8	12	2	18	0.008
Uniform	-0.866	0.866	-3.464	3.464	0

Таблица 7: Теоретическая вероятность выбросов

4.6 Эмпирическая функция распределения

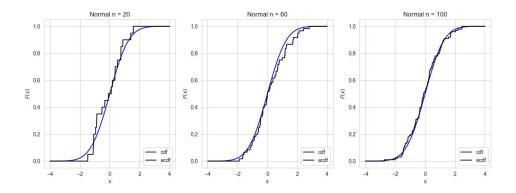


Рис. 9: Нормальное распределение

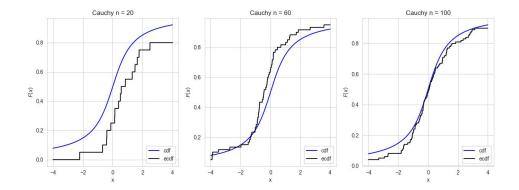


Рис. 10: Распределение Коши

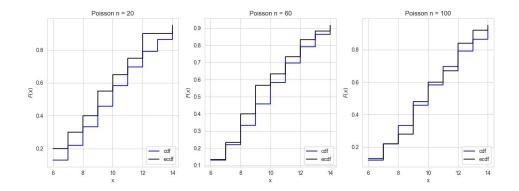


Рис. 11: Распределение Пуассона

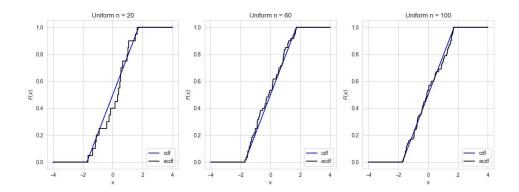


Рис. 12: Равномерное распределение

4.7 Ядерные оценки плотности распределения

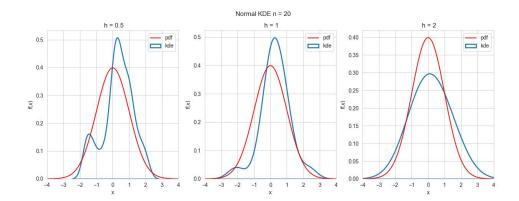


Рис. 13: Нормальное распределение, n=20

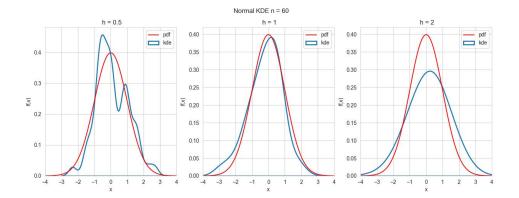


Рис. 14: Нормальное распределение, n=60

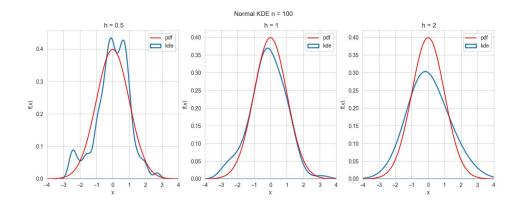


Рис. 15: Нормальное распределение, n=100

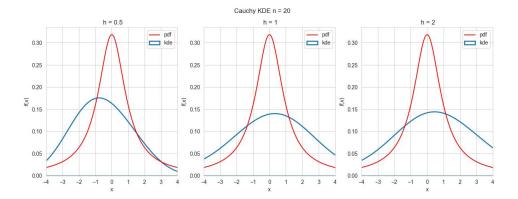


Рис. 16: Распределение Коши, n=20

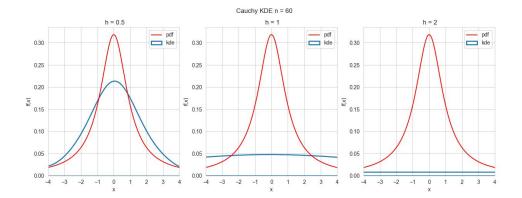


Рис. 17: Распределение Коши, n=60

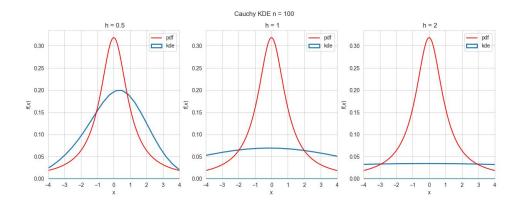


Рис. 18: Распределение Коши, n=100

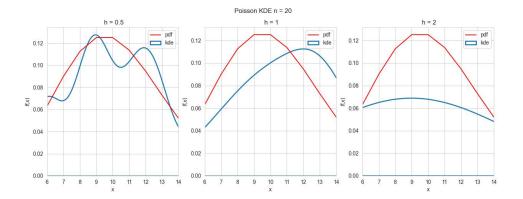


Рис. 19: Распределение Пуассона, n=20

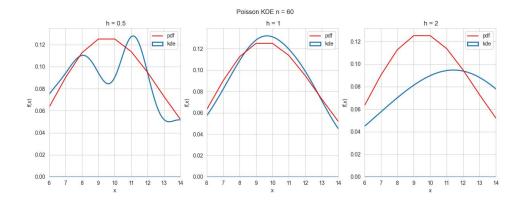


Рис. 20: Распределение Пуассона, n=60

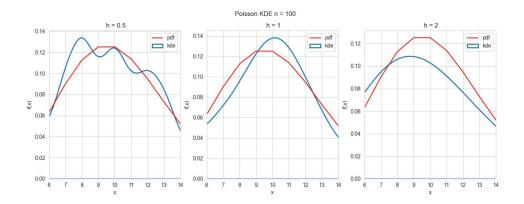


Рис. 21: Распределение Пуассона, n=100

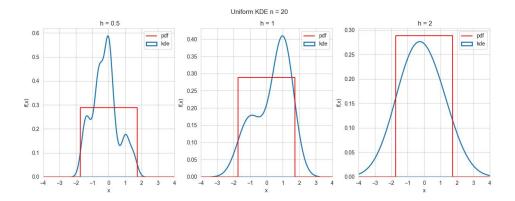


Рис. 22: Равномерное распределение, n=20

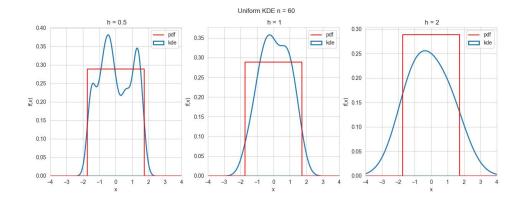


Рис. 23: Равномерное распределение, n=60

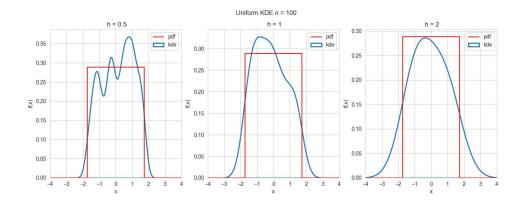


Рис. 24: Равномерное распределение, n=100

5 Обсуждение

5.1 Гистограммы

По результатам проделанной работы можем сделать вывод о том, что чем больше выборка для каждого из распределений, тем ближе ее гистограмма к графику плотности вероятности того закона, по которому распределены величины сгенерированной выборки. Чем меньше выборка, тем менее она показательна - тем хуже по ней определяется характер распределения величины. Также можно заметить, что максимумы гистограмм и плотностей распределения почти нигде не совпали. Также наблюдаются всплески гистограмм, что наиболее хорошо прослеживается на распределении Коши.

5.2 Характеристики положения и рассеяния

Исходя из данных, приведенных в таблицах, можно судить о том, что дисперсия характеристик рассеяния для распределения Коши является некой аномалией: значения слишком большие даже при увеличении размера выборки - понятно, что это результат выбросов, которые мы могли наблюдать в результатах предыдущего задания.

5.3 Доля и теоретическая вероятность выбросов

По данным, приведенным в таблице, можно сказать, что чем больше выборка, тем ближе доля выбросов будет к теоретической оценке. Снова доля выбросов для распределения Коши значительно выше, чем для остальных распределений. Равномерное распределение же в точности повторяет теоретическую оценку - выбросов мы не получали. Боксплоты Тьюки действительно позволяют более наглядно и с меньшими усилиями оценивать важные характеристики распределений. Так, исходя из полученных рисунков, наглядно видно то, что мы довольно трудоёмко анализировали в предыдущих частях.

5.4 Эмпирическая функция и ядерные оценки плотности распределения

Можем наблюдать на иллюстрациях с э. ф. р., что ступенчатая эмпирическая функция распределения тем лучше приближает функцию распределения реальной выборки, чем мощнее эта выборка. Заметим так же, что для распределения Пуассона и равномерного распределения отклонение функций друг от друга наибольшее. Рисунки, посвященные ядерным оценкам, иллюстрируют сближение ядерной оценки и функции плотности вероятности для всех h с ростом размера выборки. Для распределения Пуассона наиболее ярко видно, как сглаживает отклонения увеличение параметра сглаживания h. В зависимости от особенностей распределений для их описания лучше подходят разные параметры h в ядерной оценке: для равномерного распределения и распределения Пуассона лучше подойдет параметр $h=2h_n$, для распределения Лапласа $-h=\frac{h_n}{2}$, а для нормального и Коши $-h=h_n$. Такие значения дают вид ядерной оценки наиболее близкий к плотности, характерной данным распределениям. Также можно увидеть, что чем больше коэффициент при параметре сглаживания \hat{h}_n , тем меньше изменений знака производной у аппроксимирующей функции, вплоть до того, что при $h=2h_n$ функция становится унимодальной на рассматриваемом промежутке. Также видно, что при $h=2h_n$ по полученным приближениям становится сложно сказать плотность вероятности какого распределения они должны повторять, так как они очень похожи между собой.

6 Приложение

Код программы GitHub: https://github.com/KsenErem/MatStat