Математическая статистика

Домашняя работа № 2

Основные понятия математической статистики

Попов Юрий, СКБ-172

ОГЛАВЛЕНИЕ

Предисловие

Все графики, которые в дальнейшем будут вставлены в эту работу, были сконструированы с помощью различных библиотек, основные которые - это matplotlib и numpy в Jupyter Notebook

К работе приложены 2 основных файла: $DZ-2-Geom$ и $DZ-2-Exp$, в которых указаны расчеты соотвественно геометрического и экспоненциального распределения

Также есть файл $get-quantile-geom$, в котором расписан процесс нахождения квантиля.

Большая часть определений, которые представлены в этой работы взять с лекций нашего курса.

Также некоторые определения взяты из источника Г.И. Ивченко, Ю.И. Медведев "Ведение в математическую статистику"

Задание 2.1 *Моделирование выбранных случайных величин*

Реализация выборки

Определение 1. *Реализация выборки* - это набор из n наблюдений $\hat{x} = (x_1, x_2, \ldots, x_n)$

2.1.1 Геометрическое распределение

Используя итоговый из задания 1 для моделирования случайных величин, подчиняющих геометрическому распределению, были построены по 5 реализаций выборок различных объемов(5, 10, 100, 1000, 10**5)

Вот, пример по 5 реализаций выборки объема 5 и 10 соответственно:

```
Реализация выборки объема 5:

1): [5, 3, 14, 0, 1]
2): [8, 7, 13, 13, 5]
3): [5, 0, 4, 31, 0]
4): [21, 13, 2, 3, 8]
5): [3, 8, 3, 20, 2]
```

Рис. 1: n= 5

```
Реализация выборки объема 10:

1): [6, 29, 1, 4, 19, 0, 4, 3, 6, 3]
2): [2, 18, 10, 2, 4, 5, 12, 17, 5, 4]
3): [6, 16, 5, 1, 2, 0, 25, 12, 11, 6]
4): [9, 1, 9, 3, 23, 11, 28, 5, 5, 12]
5): [27, 14, 2, 6, 8, 8, 9, 25, 0, 18]
```

Рис. 2: n = 10

2.1.2 Экспоненциальное распределение

Используя итоговый из задания 1 для моделирования случайных величин, подчиняющих геометрическому распределению, были построены по 5 реализаций выборок различных объемов(5, 10, 100, 1000, 10**5)

Вот, пример по 5 реализаций выборки объема 5 и 10 соответственно:

```
Реализация выборки экспоненциального распределени объема 5:

1): [6.8, 8.87, 4.05, 7.33, 8.5]
2): [2.47, 4.11, 8.82, 12.82, 2.63]
3): [21.41, 5.13, 14.58, 14.41, 1.86]
4): [26.62, 5.43, 25.14, 4.77, 6.53]
5): [14.35, 1.79, 4.8, 2.95, 2.96]
```

Рис. 3: n= 5

```
Реализация выборки экспоненциального распределени объема 10:

1): [13.09, 1.96, 2.08, 3.38, 6.32, 10.56, 19.55, 4.73, 1.99, 5.26]
2): [8.61, 12.9, 12.63, 2.0, 13.55, 24.12, 2.99, 10.89, 0.94, 5.52]
3): [14.83, 0.92, 6.12, 10.23, 9.63, 11.85, 1.71, 2.68, 2.23, 1.92]
4): [30.82, 3.75, 4.27, 4.03, 4.74, 12.96, 0.34, 12.99, 10.11, 9.32]
5): [6.34, 3.49, 8.41, 34.44, 2.45, 1.67, 8.66, 6.01, 23.6, 1.64]
```

Рис. 4: n = 10

Задание 2.2 *Построение эмпирической функции распределения*

Эмпирическая функция распределения

Определение 2. Для произвольного числа $x \in R$ рассмотрим случайную величину

$$\mu_n(x) = \sum_{i=1}^{n} Ind(X_i \leq x)$$

равную числу элементов выборки меньших или равных x. Тогда функция $\hat{F}(x) = \frac{\mu_n(x)}{n}$ называется *эмпирической функцией распределения(э.ф.р)*

Эмпирическая функция распределения принимает значения $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$

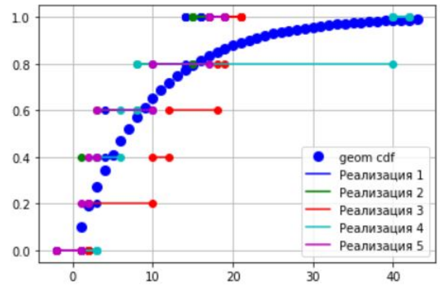
$$P(\hat{F}(x) = \frac{k}{n}) = C_k^m F^k(x)(1 - F(x))^{n-k}$$

2.2.1 Геометрическое распределение

Для каждой выборки построим эфр. На каждом графике приведены графики э.ф.р 5 реализаций одинакового объема, а также график функции распределения, помеченный синим цветом. Для его построения использовался модуль scipy.stats

Эфр для выборок объема 5:

```
In [118]: make_cdf(0.1)
          make_efr(M, 5)
```

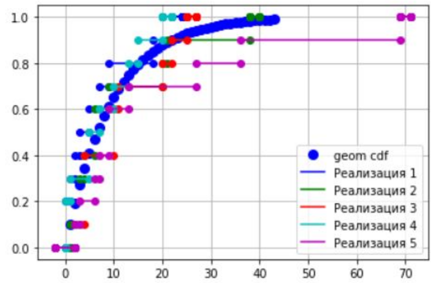


a)

Эфр для выборок объема 10:

```
In [119]: M = make_vibor(10)
```

```
In [120]: make_cdf(0.1)
          make_efr(M, 10)
```

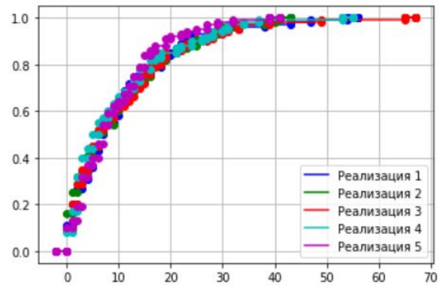


b)

Эфр для выборок объема 100:

```
In [121]: M = make_vibor(100)
```

```
In [122]: make_efr(M, 100)
```

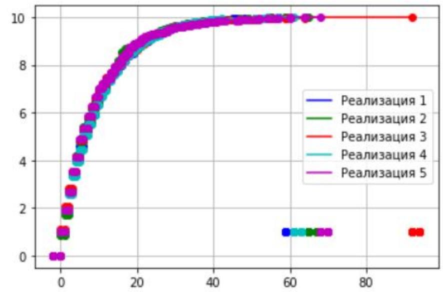


c)

Эфр для выборок объема 1000:

```
In [123]: M = make_vibor(1000)
```

```
In [124]: make_efr(M, 100)
```

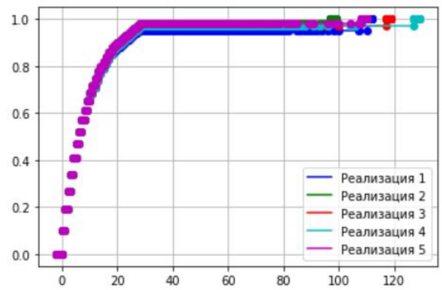


d)

Эфр для выборок объема 10**5:

```
In [125]: M = make_vibor(10**5)
```

```
In [126]: make_efr(M, 10**5)
```



e)

Теперь найдем точную верхнюю границу разности каждой пары эмпирических функций распределения:

Для этого выберет какое-то значение на оси x(например x = 4), и посмотри чему равно $f(x)$ э.ф.р для данного икса.

Приведем результаты для выборок различного объема:

Для n = 5:

```
M = make_vibor(5)
```

```
upper_limit(M, 5, 4)
```

```
Точная верхняя граница выборки объема 5:  0.2
```

n = 5

Для n = 10:

```
M = make_vibor(10)
```

```
upper_limit(M, 10, 4)
```

```
Точная верхняя граница выборки объема 10:  0.3
```

n = 10

Для n = 100:

```
M = make_vibor(100)
```

```
upper_limit(M, 100, 4)
```

```
Точная верхняя граница выборки объема 100:  0.15999999999999998
```

n = 100

Для n = 1000:

```
M = make_vibor(1000)
```

```
upper_limit(M, 1000, 4)
```

```
Точная верхняя граница выборки объема 1000:  0.08000000000000007
```

n = 1000

Для n = 10**5:

```
M = make_vibor(10**5)
```

```
upper_limit(M, 10**5, 4)
```

```
Точная верхняя граница выборки объема 100000:  0.07
```

n = 100000

2.2.2 Экспоненциальное распределение

Для каждой выборки построим эфр. На каждом графике приведены графики э.ф.р 5 реализаций одинакового объема, а также график функции распределения, помеченный синим цветом. Для его построения использовался модуль scipy.stats

Теперь найдем точную верхнюю границу разности каждой пары эмпирических функций распределения:

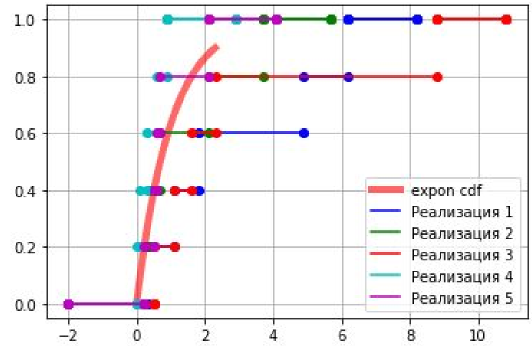
Для этого выберет какое-то значение на оси x(например x = 4), и посмотри чему равно $f(x)$ э.ф.р для данного икса.

Приведем результаты для выборок различного объема:

Эфр для выборок объема 5:

```
M = make_vibor_exp(0.8, 5)
make_cdf_expon()
make_efr_expon(M, 5)
```

[]

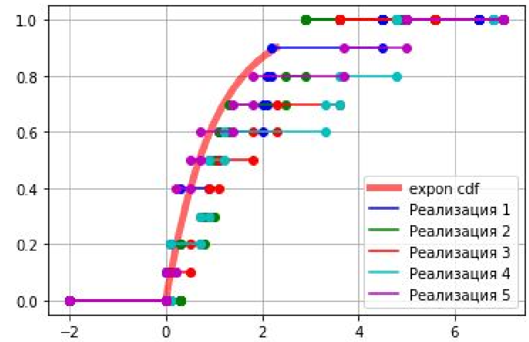


a)

Эфр для выборок объема 10:

```
M = make_vibor_exp(0.8, 10)
make_cdf_expon()
make_efr_expon(M, 10)
```

[]

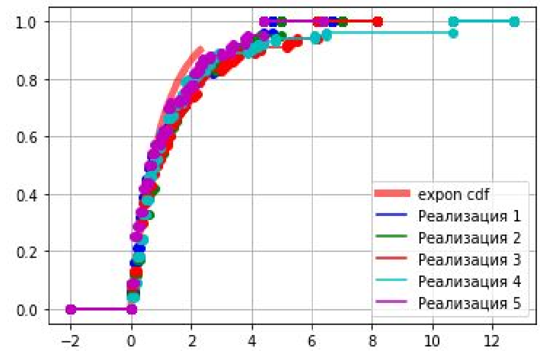


b)

Эфр для выборок объема 100:

```
M = make_vibor_exp(0.8, 100)
make_cdf_expon()
make_efr_expon(M, 100)
```

[]

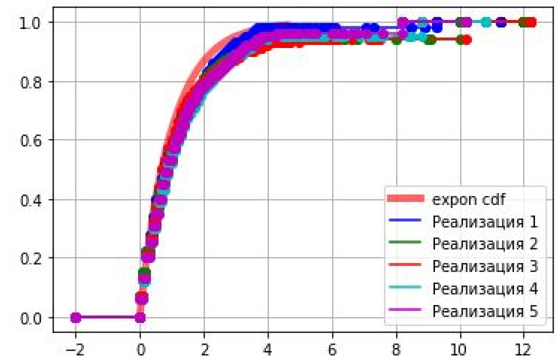


c)

Эфр для выборок объема 1000:

```
M = make_vibor_exp(0.8, 1000)
make_cdf_expon()
make_efr_expon(M, 1000)
```

[]

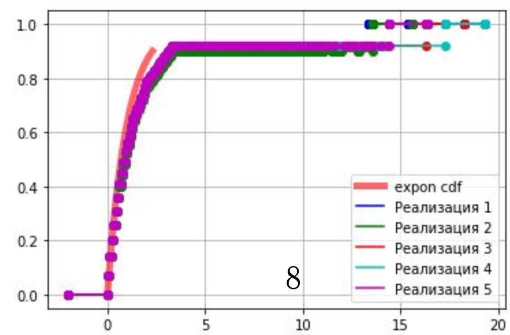


d)

Эфр для выборок объема 10**5:

```
M = make_vibor_exp(0.8, 10**5)
make_cdf_expon()
make_efr_expon(M, 10**5)
```

[]



e)

Для n = 5:

```
M = make_vibor_exp(0.8, 5)
upper_limit_expon(M, 5, 4)
```

```
Точная верхняя граница выборки объема 5:  0.2
```

n = 5

Для n = 10:

```
M = make_vibor_exp(0.8, 10)
upper_limit_expon(M, 10, 4)
```

```
Точная верхняя граница выборки объема 10:  0.1
```

n = 10

Для n = 100:

```
M = make_vibor_exp(0.8, 100)
upper_limit_expon(M, 100, 4)
```

```
Точная верхняя граница выборки объема 100:  0.04
```

n = 100

Для n = 1000:

```
M = make_vibor_exp(0.8, 1000)
upper_limit_expon(M, 1000, 4)
```

```
Точная верхняя граница выборки объема 1000:  0.03
```

n = 1000

Для n = 10**5:

```
M = make_vibor_exp(0.8, 10**5)
upper_limit_expon(M, 10**5, 4)
```

```
Точная верхняя граница выборки объема 100000:  0.01
```

n = 100000

Задание 2.3 *Построение вариационного ряда выборки*

Вариационный ряд выборки

Определение 3. Пусть есть
$$\vec{X} = (X_1, \ldots, X_n),$$

где $X_i, i = \overline{1, n}$—независимые одинаково распределенные случайные величины из распределения ξ. И $\vec{x} = (x_1, \cdots, x_n)$ является реализацией имеющейся выборки $\vec{X}$. Отсортируем вектор $\vec{x}$ по возрастанию:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$$

Тогда $x_{(1)} = \min(x_1, x_2, \ldots x_n)$, а $x_{(n)} = \max(x_1, x_2, \ldots, x_n)$. Через $X_{(i)}$ обозначают случайную величину, которая для каждой реализации выборки принимает значение $X_{(i)}$. Вектор $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ называют вариационным рядом выборки.

Квантиль

Определение 4. *Квантилью уровня* $\alpha \in (0, 1)$ функции распределения $F(x)$ называется величина $\zeta_\alpha = \sup\{x : F(x) \leq p\} = F^{-1}(p)$.

Выборочный квантиль

Определение 5. *Выборочными квантилями* называют квантили выборочного распределения.

2.3.1 Геометрическое распределение

Для каждой выборки построим вариационный ряд. Вот, примеры вариационных рядов для выборок объема n = 5 и для n = 10:

```
Вариационные ряды для 5 реализаций выборки, объема  5:

X:    0    1    5    15
------------------------------------------------
f:    1    2    1    1


X:    0    2    6    9    25
------------------------------------------------
f:    1    1    1    1    1


X:    1    4    5    7    8
------------------------------------------------
f:    1    1    1    1    1


X:    0    8    13   17   21
------------------------------------------------
f:    1    1    1    1    1


X:    0    1    3    7    18
------------------------------------------------
f:    1    1    1    1    1
```

Рис. 5: n = 5

```
Вариационные ряды для 5 реализаций выборки, объема  10:

X:    0    1    2    5    7    14   16   19   39
-----------------------------------------------------------
f:    1    1    1    2    1    1    1    1    1


X:    0    1    2    3    6    8    23
-----------------------------------------------------------
f:    1    4    1    1    1    1    1


X:    0    1    2    3    6    9    12   13   15   33
-----------------------------------------------------------
f:    1    1    1    1    1    1    1    1    1    1


X:    0    1    3    5    16   28
-----------------------------------------------------------
f:    4    2    1    1    1    1


X:    0    1    3    4    5    6    8    12   17
-----------------------------------------------------------
f:    1    1    1    2    1    1    1    1    1
```

Рис. 6: n = 10

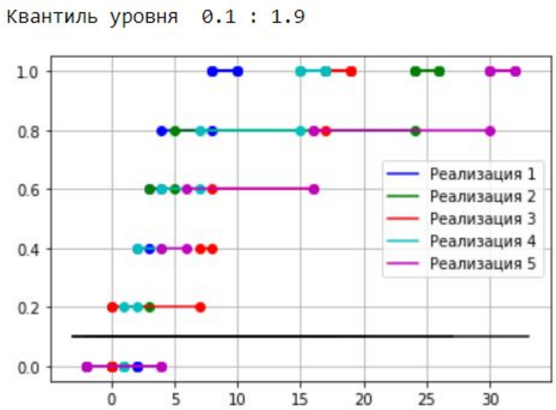
Перейдем к нахождению выборочной квантили.

Поясним простыми словами на примере уровня квантиля 0.1, что такое выборочная квантиль:

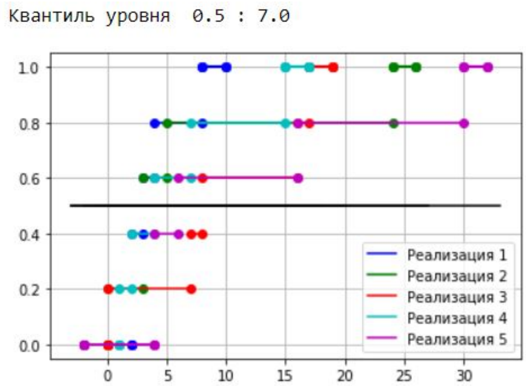
Выборочная квантиль уровня 0,1 - это точка, левее которой (включая её саму) лежит не менее 10% всей выборки, и правее которой (снова включая её саму) - не менее $100\% - 10\% = 90\%$ выборки.

Будем искать выборочную квантиль графическим способом: проведем прямую $y =$(уровень квантиля) до пересечения с графиком. И определим x, при котором прямая пересекает график. Если точка пересечения одна, то именно это значения x и будет выборочной квантиль. А если пересечение проходит по отрезку, то квантилей будет много. Например, если отрезок от 3 до 4 является прямой y, то квантилью будет любое число от 3 до 4. ля определённости в практической статистике в таких случаях выбирают по какому-то правилу одно из чисел "отрезка квантилей". Например, середину отрезка - в данном случае 3,5.

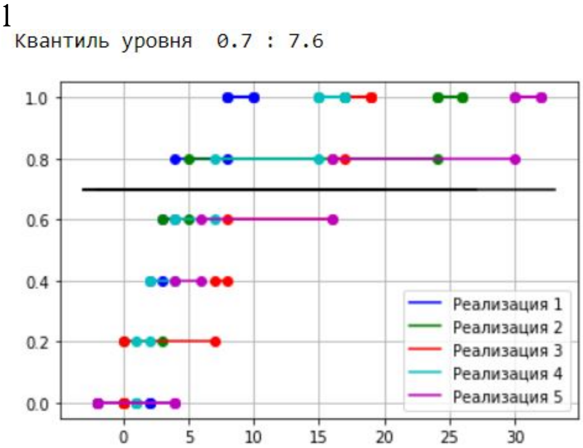
Ниже показано в первой строчке расчет выборочной квантильи уровня 0.1, 0.5, 0.7, а затем графическое представление для геометрического распределения. Жирной, серой чертой как раз и проведена линия уровня, по которой мы находим квантиль



a = 0.1



a = 0.5



a = 0.7

2.3.2 Экспоненциальное распределение

Для каждой выборки построим вариационный ряд. Вот, примеры вариационных рядов для выборок объема n = 5 и для n = 10:

```
Вариационные ряды для 5 реализаций выборки, объема  5:

X:      0.0     0.4     0.7     1.9     3.8
--------------------------------------------------
f:      1       1       1       1       1


X:      0.9     1.1     2.0     11.7
--------------------------------------------------
f:      1       1       2       1


X:      0.1     0.2     0.4     1.2     2.2
--------------------------------------------------
f:      1       1       1       1       1


X:      0.3     0.5     0.7     2.6
--------------------------------------------------
f:      1       1       2       1


X:      0.2     0.3     0.5     0.9     1.1
--------------------------------------------------
f:      1       1       1       1       1
```

Рис. 7: n = 5

```
Вариационные ряды для 5 реализаций выборки, объема  10:

X:      0.0     0.1     0.2     0.3     0.4     0.5     0.8     0.9     1.2
------------------------------------------------------------------------------------------
f:      1       1       1       1       2       1       1       1       1


X:      0.2     0.6     0.8     1.0     1.7     1.8     2.0     3.2
------------------------------------------------------------------------------------------
f:      1       2       1       1       1       2       1       1


X:      0.3     0.4     0.7     1.0     1.6     2.0     2.2     2.4     3.1
------------------------------------------------------------------------------------------
f:      1       1       1       2       1       1       1       1       1


X:      0.2     0.3     0.6     0.9     1.2     1.4     1.5     1.7     2.5     3.3
------------------------------------------------------------------------------------------
f:      1       1       1       1       1       1       1       1       1       1


X:      0.1     0.2     0.3     1.3     1.5     3.1     3.6
------------------------------------------------------------------------------------------
f:      1       1       4       1       1       1       1
```

Рис. 8: n = 10

Задание 2.4 *Построение гистограммы и полигон частот*

Гистограмма

Определение 4.Для непрерывной случайной величины ξ, обладающей непрерывной плотностью $f(x)$, также можно построить по соответствующей выборке $X = (X_1, \dots, X_n)$ статистический аналог $\hat{f}_n(x)$ для плотности $f(x)$, который называется *гистограммой*

Полигон частот

Определение 5. Наряду с гистограммой, в качестве приближения для неизвестной теоретической плотности $f(x)$ можно использовать кусочно-линейный график, называемый *полигоном частот*, и который строится так: если построена гистограмма $\hat{f}_n(x)$, то ординаты, соответствующие серединам интервалов группировки, последовательно соединяют отрезками прямых.

2.4.1 Геометрическое распределение

2.4.2 Экспоненциальное распределение

Литература

[1]

[2] ссылка1

[3] ссылка2

[4] // ссылка3

[5] // ссылка4