Министерство образования и науки Российской Федерации

Санкт-Петербургский политехнический университет Петра Великого

—

**Институт** **кибербезопасности и защиты информации**

Отчёт

по расчетному заданию

**Логнормальное распределение**

по дисциплине «Теория вероятностей и математическая статистика»

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Выполнили: | студент группы 4851003/10001 |  | | Алексеев М.И. |
|  |  | (подпись, дата) | |  |
|  | студент группы 4851003/10001 |  | | Бакулев В.М. |
|  |  | (подпись, дата) | |  |
|  |  |  | |  |
|  |  |  | |  |
| Проверил: | Ассистент преподавателя |  | | А.А Штыркина |
|  |  | (подпись, дата) | |  |
|  |  |  |  |  |

Санкт-Петербург

2023

1. **Провести исследование распределения из варианта: история его появления, его назначение, применение в научных исследованиях**

История появления распределения

Логнормальное распределение в теории вероятностей — это двухпараметрическое семейство абсолютно непрерывных распределений. Изначально во многих отраслях было достаточно применять Гауссово распределение, однако многие измерения показывают более или менее асимметричное распределение. Особенно распространены асимметричные распределения, когда средние значения малы, а дисперсии велики и значения не могут быть отрицательными, как, например, в случае численности видов, длительности латентных периодов инфекционных заболеваний, распределения полезных ископаемых в земной коре[[1]](#footnote-1).

Такие асимметричные распределения часто близко соответствуют логарифмически нормальному распределению. Логарифмически нормальные распределения обычно характеризуются с точки зрения логарифмически преобразованной переменной с использованием в качестве параметров ожидаемого значения или среднего значения ее распределения и стандартного отклонения. Эта характеристика может быть полезной, поскольку по определению логарифмически нормальные распределения снова симметричны на логарифмическом уровне.

Назначение:

Логнормальное распределение используется для описания переменных нагрузки, тогда как нормальное распределение используется для описания переменных сопротивления. Однако переменной, которая, как известно, никогда не принимает отрицательных значений, обычно назначается логарифмически нормальное распределение, а не нормальное распределение.

Применение в научных исследованиях:

Логнормальное распределение используется, например, при моделировании таких переменных, как доходы, допустимое отклонение от стандарта вредных веществ в продуктах питания и т.д. Так же при анализе поведения человека, например Длина комментариев, размещаемых на дискуссионных форумах в Интернете, подчиняется логарифмически нормальному распределению. В биологии и медицине используется для меры размеров жировой ткани, в размерах лавин разрывов в цитоскелете живых клеток показаны логарифмически нормальные распределения, причем в раковых клетках размеры значительно выше, чем в здоровых и так далее .

1. **Знакомство с jupyter notebook.**

**Определить вид функции распределения для распределения из варианта**

* Сумма элементов выборки = 5556.8950
* Выборочное среднее = 18.5230
* Медиана = -1.1890
* Мода = 0 (Все числа разные )
* Размах выборки = 1220.1618
* Смещенная дисперсия = 11746.5156
* Несмещенная дисперсия = 11785.80165
* Выборочный начальный момент 3-ого порядка = 11160811.1278
* Выборочный центральный момент 3-го порядка = 10501714.3435

1. **Понятие эмпирической функции распределения**
   1. **Определить вид функции распределения для распределения из варианта**

Пусть известно статистическое распределение частот количественного признака X. Введем обозначения:   **─** число наблюдений, при которых наблюдалось значение признака меньше x, n  – общее число наблюдений (объем выборки).

Ясно, что относительная частота события X < x  равна . Если x изменяется, то, вообще говоря, изменится и относительная частота, то есть относительная частота    есть функция от x. Так как статистическое распределение выборки находится эмпирическим (опытным) путем, то эту функцию называют эмпирической.

Эмпирической функцией распределения (функцией распределения выборки) называется функция , определяющая для каждого значения x  относительную частоту события X < x.

где    ─ число вариантов, меньших x, n  – объем выборки.

Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

Эмпирическая функция распределения имеет скачки в точках выборки, величина скачка в точке  равна , где m — количество элементов выборки, совпадающих с .

Можно построить эмпирическую функцию распределения по вариационному ряду:

Логнормальное распределение случайной величины X задается плотностью вероятности, имеющей вид

## *где > 0, loc >0, s ∈ R*

## Имеет вид при :

## S - параметр формы

## Loc - параметр смещения

## :

## 

## Рисунок 1 – График функции логнормального распределения(k=300)

## Построить эмпирическую функцию распределения для k случайных элементов выборки (k=10, 50, 200)

## 

## Рисунок 2 - Эмпирическая функция распределения(k=10)

## 

## Рисунок 3 - Эмпирическая функция распределения(k=100)

## 

## Рисунок 4 - Эмпирическая функция распределения(k=200)

* 1. **Сравнить теоретическую и эмпирическую функции распределения на графике**

Выборочная (эмпирическая) функция распределения в [математической статистике](https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%81%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D0%BA%D0%B0) — это приближение теоретической [функции распределения](https://ru.wikipedia.org/wiki/%D0%A4%D1%83%D0%BD%D0%BA%D1%86%D0%B8%D1%8F_%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), построенное с помощью выборки из него.

Рассмотрев графики, можно заметить, что эмпирическая функция повторяет форму теоретической функции распределения.

1. **Понятие гистограммы**
   1. **Определить вид функции плотности распределения для распределения из варианта**

Гистограмма — это графическое представление распределения частоты встречаемости значений в наборе данных, которое состоит из прямоугольников, где каждый прямоугольник представляет диапазон значений и высоту, пропорциональную частоте встречаемости значений в этом диапазоне.

Для того что бы определить плотность распределения необходимо взять интеграл от функции распределения. По полученной плотности был построен график плотности логнормального распределения. График представлен на рисунке 5.

## Имеет вид при :

## *где s> 0, loc > 0, s ∈ R*

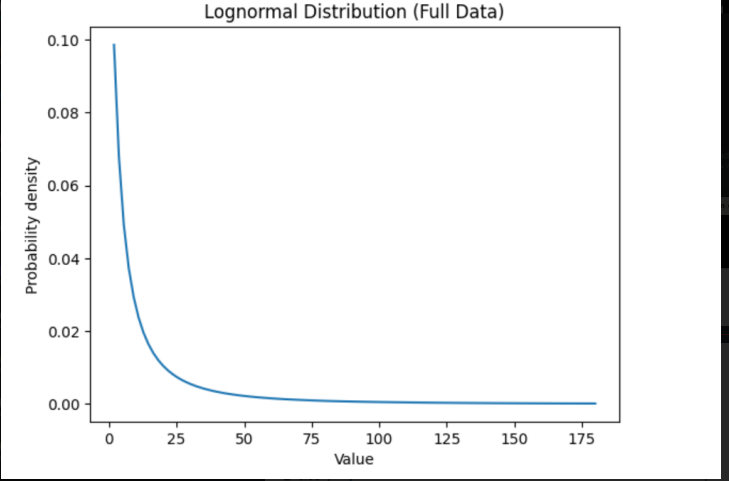


Рисунок 5.1 – График плотности логнормального распределения.

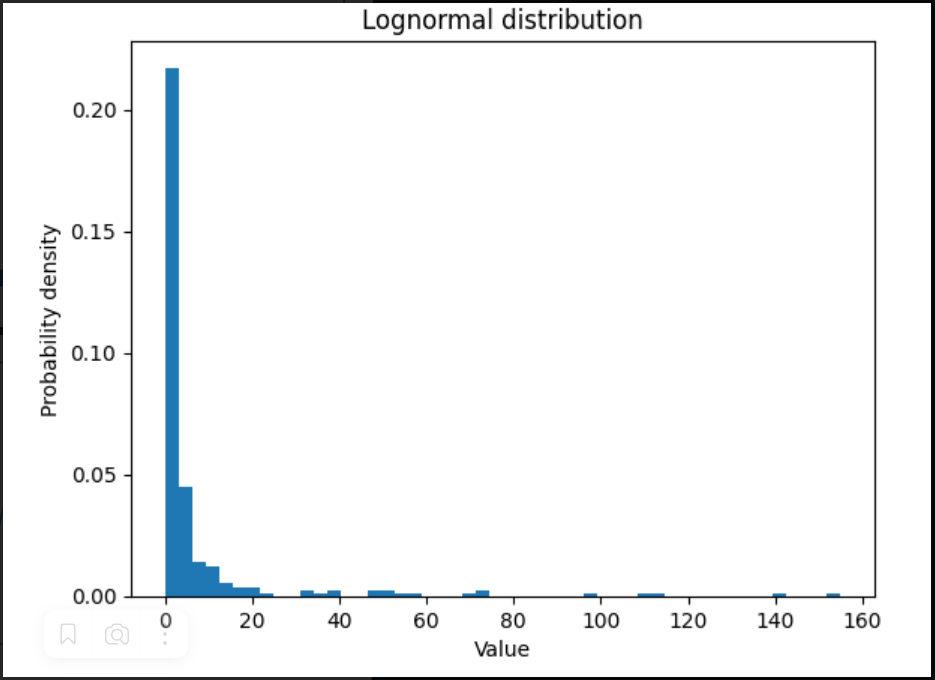


Рисунок 5.2 Гистограмма логнормального распределения

* 1. **Построить гистограмму (по относительным частотам!) для k случайных элементов выборки (k=10, 100, 200).**

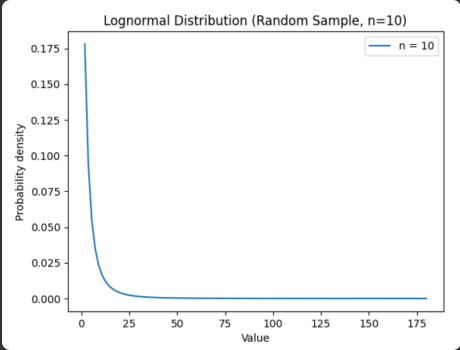


Рисунок 6.1 – Плотность распределения для k=10

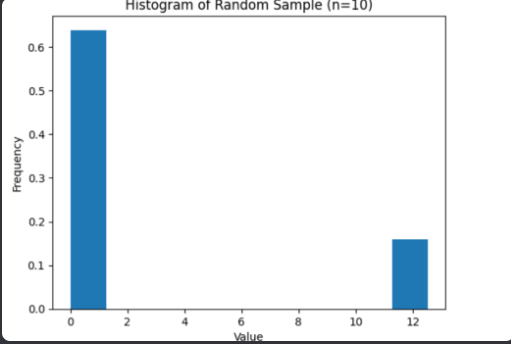


Рисунок 6.2 – Гистограмма для k=10

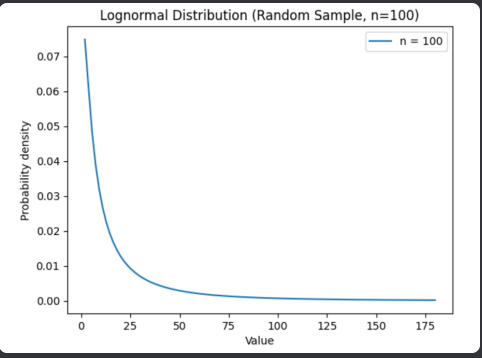


Рисунок 7.1 – Плотность распределения для k=100

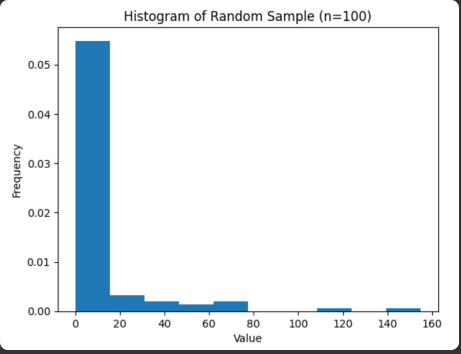


Рисунок 7.2 - Гистограмма для k=100

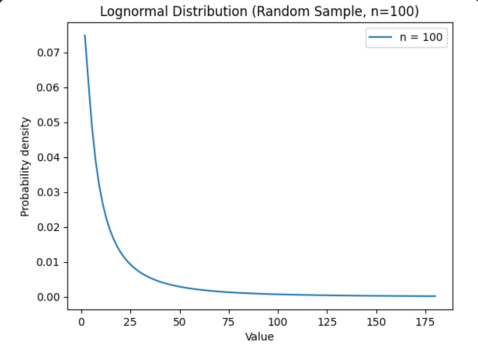


Рисунок 8.1 - Плотность для k=100

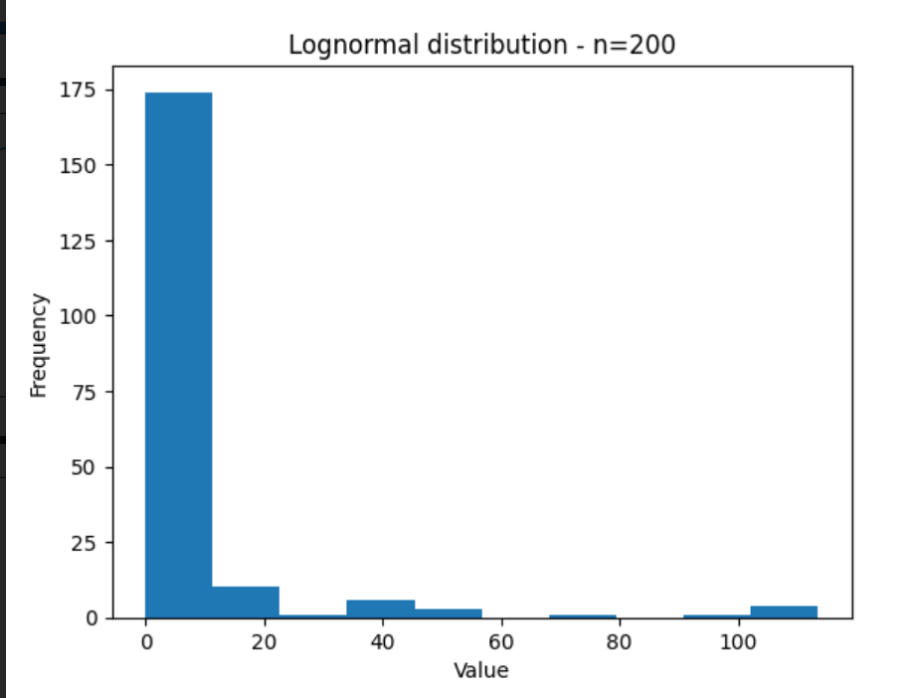


Рисунок 8.2 – Гистограмма для k=200

**4.4 Вывод о виде гистограммы**

Таким образом, в ходе выполнения данного пункта работы был определён вид функции плотности распределения для распределения из варианта. Кроме того, построены гистограммы по относительным частотам для 10, 100 и 200 случайных элементов выборки. Наконец, мы сравнили функцию плотности распределения и гистограмму.

На гистограмме можно заметить, что она достаточно похожа на форму логнормального распределения: она скошена вправо и имеет длинный правый хвост. Вероятность того, что число будет больше 15, также резко уменьшается.

**5. Описание параметров распределения**

**5.2 Вид теоретической функции распределения**

В отличие от эмпирической функции распределения для выборки, вводится понятие теоретической функции распределения для генеральной совокупности – F(x). Теоретическая функция распределения определяет вероятность события X<x. Эмпирическая функция распределения F\*(x) по вероятности стремится к теоретической функции распределения F(x) при больших количествах испытаний и обладает всеми свойствами F(x):

## 5.3 Описание параметров распределения

* **x**: Значение случайной переменной (наблюдаемая величина), для которого мы хотим вычислить функцию распределения.
* **loc**: Параметр смещения (loc), также известный как среднее или математическое ожидание распределения. Он определяет сдвиг или смещение графика распределения вдоль оси x. Значение ln(loc) используется в формуле для перевода его в логарифмическую шкалу.
* **s**: Параметр формы (s), также известный как стандартное отклонение или параметр растяжения. Он контролирует степень разброса данных в логнормальном распределении. Большие значения s приводят к более широкому распределению, тогда как малые значения s сжимают распределение вокруг значения среднего.
* Функция **erf** представляет собой функцию ошибок, которая вычисляет интеграл от стандартного нормального распределения. В данной формуле она используется для приведения значения (ln(x) - loc) / (sqrt(2) \* s) к стандартному нормальному распределению, что позволяет вычислить вероятность (**CDF**) для заданного значения x.

Таким образом, параметр loc определяет положение распределения по оси x, параметр s контролирует его форму, а функция erf выполняет приведение к стандартному нормальному распределению для вычисления вероятности.

**5.4, 5.5 Построение теоретической функции распределения**

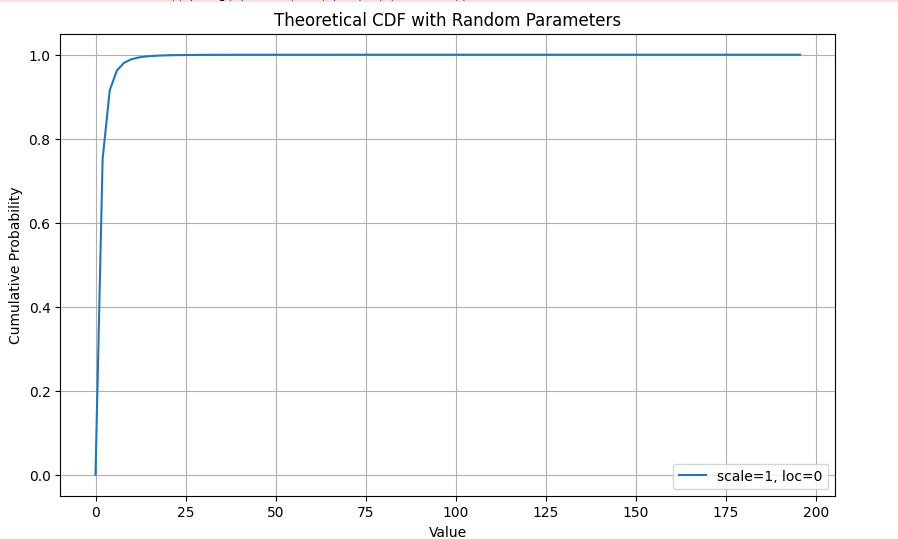


Рис. 9.1 Теоретическая функция распределения при s = 1, loc = 0

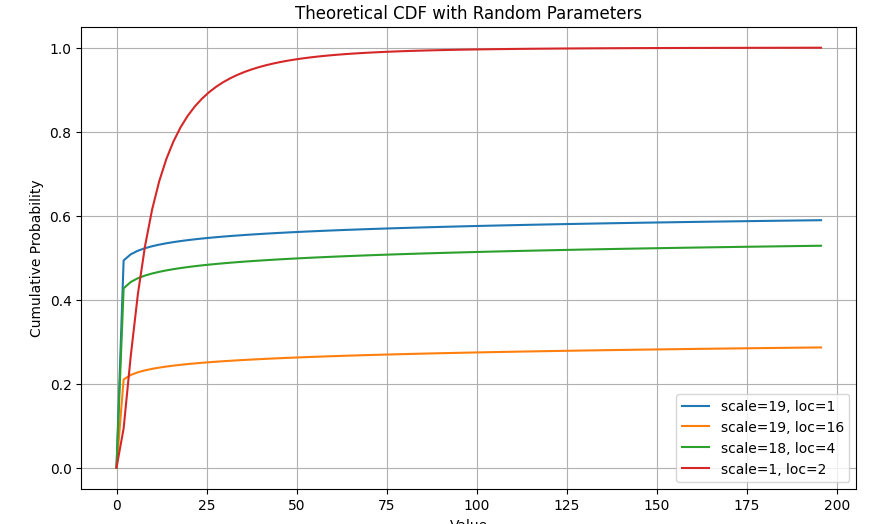


Рис. 9.2 Теоретическая функция распределения при случайных s, loc

**5.6 Построение теоретической функции распределения**

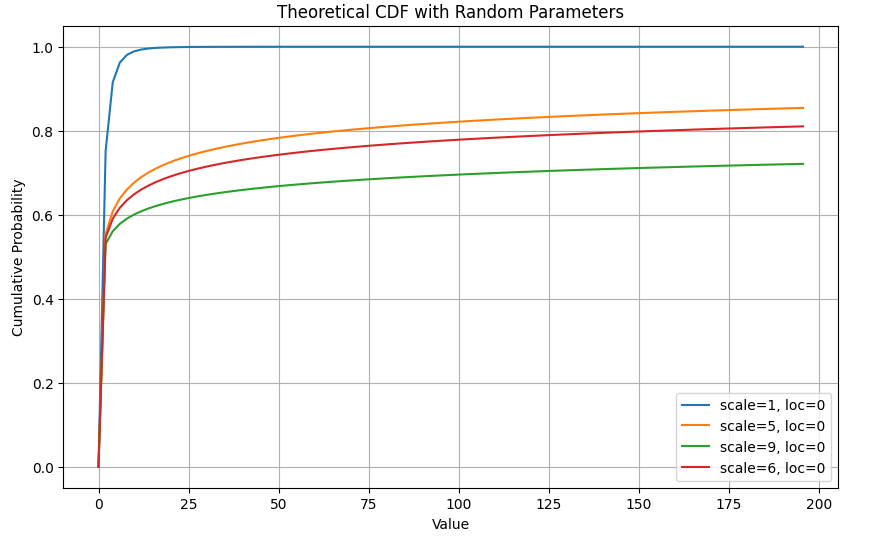


Рис. 9.3 - Теоретическая функция распределения при значении loc = 0 и случайном S

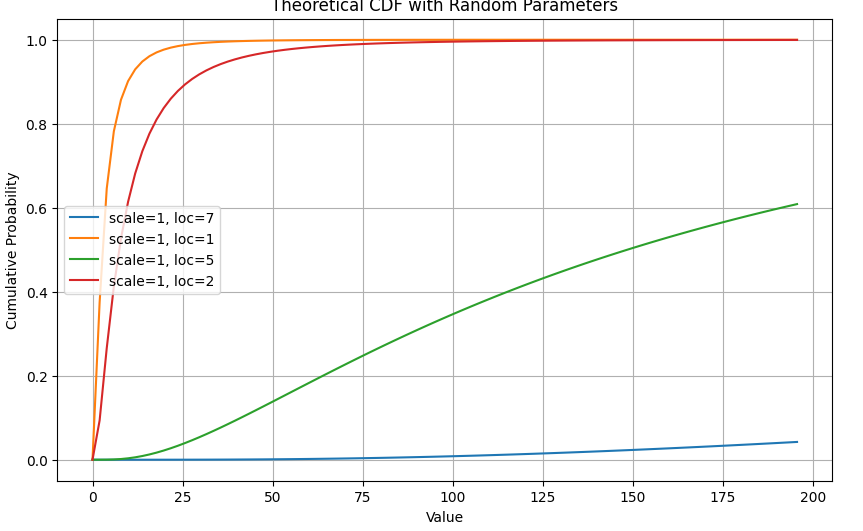


Рис. 9.4 : Теоретическая функция распределения при значении S = 1 и случайном loc

Выводы:

При изменении параметра параметра формы (s) в большую сторону видны изменения роста графика, он становится более “резким”

При изменении параметра смещения (loc) в большую сторону мы видим, что график становится более пологим и его рост становится медленнее.

**6. Понятие точечных оценок**

**6.1 Оценить параметры распределения выборки методом моментов**

Метод моментов - метод оценки неизвестных параметров распределений в математической статистике и эконометрике, основанный на предполагаемых свойствах моментов (Пирсон, 1894 г.).

Идея метода заключается в замене истинных соотношений выборочными аналогами, т. е. выразить числовые параметры теоретического распределения через моменты распределения, оценённые по выборки. Число моментов должно соответствовать числу неизвестных параметров распределения (чаще всего используют первые два момента).

Выборочные моменты первого и второго порядка имеют следующий вид:

Для оценки двух параметров распределения необходимо два теоретических момента: первого и второго порядка соответственно. В логнормальном распределении со смещением они имеют следующий вид:

Приравняем выборочные моменты с теоретическими и решим систему уравнений:

Решая данную систему, получаем оценку параметров:

Параметр смещения loc = 1.3545;

Параметр формы s = 1.8309;

Таким образом, получены оценки параметров формы и смещения логнормального распределения методом моментов.

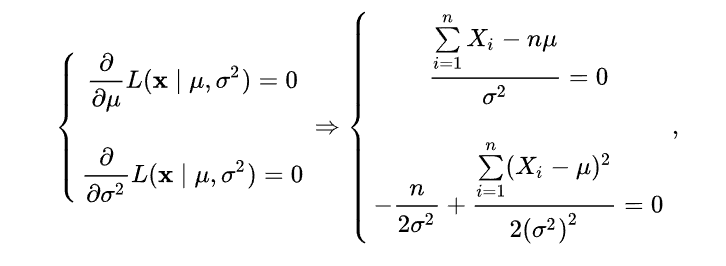
**6.2 Оценить параметры распределения выборки методом максимального правдоподобия**

**Метод максимального правдоподобия -** метод оценивания неизвестного параметра путём максимизации функции правдоподобия.

Для логнормального распределения функция правдоподобия имеет вид  
Изображение выглядит как текст

Автоматически созданное описание

Чтобы найти параметры возьмем частные производные от функции



Тогда из первого уравнения получается, что /n= *, следовательно u (loc*)= 0.0340

Подставляя по вторую производную полученное значение *s* получим  *s( (s))* = 2.5505

**6.3 Привести и обосновать свойства полученных оценок**

**Несмещенность** - это свойство, которое означает, что математическое ожидание оценочного значения какого-то параметра будет равно истинному значению этого параметра. Если простыми словами, то наиболее вероятное значение параметра, подсчитанное на основе предложенной формулы, будет равно его истинному значению. То есть ожидается, что предложенная формула максимально точно, насколько это возможно, угадывает истинное значение изучаемого параметра.

Для параметра (X), получается, что оценка , полученная методом максимального правдоподобия, является несмещенной, а для метода моментов - смещенной

= 18,5230

Для параметра несмещенная оценка равняется

= 108,5624

Следовательно, для метода максимального правдоподобия – она несмещенная, а для метода моментов – смещенной

Оценка параметра

**Состоятельность** - это свойство, которое означает, что по мере увеличения размера выборки оценочное значение будет все больше и больше приближаться к истинному. Как правило, состоятельность - это дополнительное требование, которое накладывается на несмещенную оценку. Иными словами, если несмещенность гарантирует, что при использовании заданного метода исследователь окажется около истины, то состоятельность гарантирует, что по мере увеличения выборки он будет находиться все ближе и ближе к правде.

**Эффективность** - это свойство, означающее, что данная формула оценки искомого параметра дает наименьшую погрешность из всех возможных.

**Состоятельная оценка** – это точечная оценка, сходящаяся по вероятности к оцениваемому параметру.

Таким образом, по теореме Чебышева состоятельной оценкой является выборочное среднее, значит параметр , полученный методом максимального правдоподобия, является состоятельной оценкой.

**Эффективная оценка –** это несмещенная оценка, имеющая наименьшую дисперсию из всех возможных несмещенных оценок данного параметра. Оценка из метода максимального правдоподобия является несмещенной. Проверим, является ли она эффективной. Дисперсия данной оценки равна . Приравняем ее к неравенству Рао-Крамера , где

I()=. Получается, что оценка из метода максимального правдоподобия является эффективной.

Оценку, для которой в неравенстве Рао — Крамера достигается равенство, иногда называют **R-эффективной оценкой**. Исходя из доказательства эффективности оценки из метода максимального правдоподобия можно сказать, что она является R-эффективной.

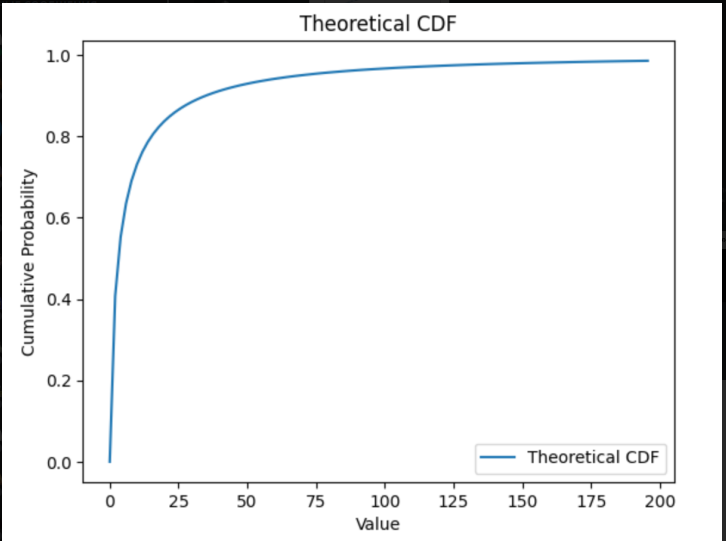
Подводя итоги, можно сказать, что только оценка из метода максимального правдоподобия подходят все вышеперечисленные свойства.  
  


Рис. 10.1: Метод моментов

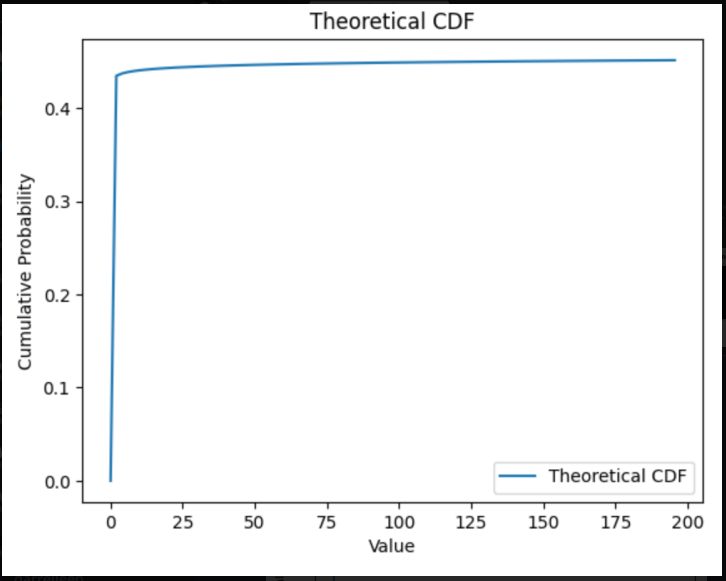


Рис. 10.2: Метод максимального правдоподобия

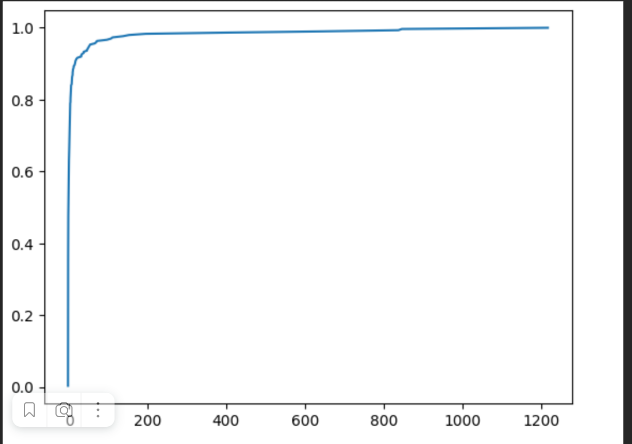


Рис. 10.3: Эмпирическая функция распределения

**Вывод**  
 Теоретические и эмпирические функции распределения дают похожий результат с некоторым отклонением, как, например, в методе моментов график более пологий, а в методе максимального правдоподобия - более резкий, в отличие от эмпирической функции распределения.

1. **Понятие интервальных оценок**

Оценим параметр смещения *loc* (x0) распределения выборки с помощью интервальной оценки с уровнями доверия , .

Поскольку нам известно значение исправленной дисперсии, то можно воспользоваться следующей формулой для вычисления интервальной оценки:

Для уровней доверия , определим значения функции распределения Стьюдента : они равны 1,96 и 2,58 соответственно.

Учитывания, что , получим:

Для уровня доверия

Для уровня доверия

Соотнесем полученные интервальные и точечные оценки. В ходе выполнения предыдущего пункта была получена следующая оценка параметра *loc*:

Можно заметить, что данное значение действительно попадает на оба полученных выше интервала для уровней доверия и , причём оно находится ближе к левому концу интервала.

**Вывод**

Таким образом, в ходе выполнения данного пункта работы удалось оценить параметр распределения *loc* выборки с помощью интервальной оценки с уровнями доверия и . Далее, сравнив полученные в предыдущем пункте точечные оценки и интервальные оценки, было замечено, что точечные оценки действительно попадают в найденные интервалы в соответствии с двумя различными уровнями доверия.

1. **Понятие статистических критериев**
   1. **Гипотезы о параметрах распределениях**

**Нулевая гипотеза** (**Н0**) — предположение о том, что между  параметрами генеральных совокупностей нет различий, то есть эти различия носят не систематический, а случайный характер.

**Альтернативная гипотеза** (**Н1**) – предположение о том, что между параметрами генеральных совокупностей есть достоверные различия.

**Ошибка первого рода** – отклонение гипотезы Н0, когда она верна. Вероятность ошибки первого рода обозначается α и называется **уровнем значимости**.

**Ошибка второго рода** – принятие гипотезы Н0, когда верна альтернативная гипотеза. Вероятность ошибки второго рода обозначается β.

В качестве уровня значимости принимаем .

Нулевая гипотеза имеет вид

**8.1.1**

**Критерий для проверки параметра формы:**

1. **Метод моментов**

Имеем две гипотезы:

Оценка формы методом моментов: s = 1.8309;

В теории статистических тестов существует несколько методов, которые можно использовать для проверки гипотез о параметрах в моделях. В нашем случае мы выбрали критерий отношения правдоподобий.

Критерий отношения правдоподобий сравнивает функции правдоподобия двух моделей - одной, где оценки параметров получены методом моментов, и другой, где оценки параметров получены методом максимального правдоподобия. Он основан на отношении максимизированных функций правдоподобия в двух моделях и предполагает, что под нулевой гипотезой (что параметры равны истинным параметрам) различие между логарифмами правдоподобия этих двух моделей распределено как хи-квадрат.

Критерий отношения правдоподобий был выбран из-за следующих причин:

* Данный критерий более предпочтителен, когда имеются оценки параметров, основанные на различных методах оценки (как в данном случае: метод моментов и метод максимального правдоподобия), и когда интересно сравнить эти методы в терминах максимизированной правдоподобия.
* Согласно лемме Неймана – Пирсона, которая утверждает, что в задаче проверки простой статистической гипотезы против простой альтернативы критерий, основанный на отношении правдоподобия, является наиболее мощным

В соответствии с леммой Неймана-Пирсона, искомая критическая область будет иметь вид:

Константу найдем из условия:

Подставив найденное значение , можно получить критическую область.

Правило проверки, задаваемое такой критической областью, является равномерно наиболее мощным критерием. Следовательно, его можно использовать при решении задачи проверки простой основной гипотезы о числовом значении параметра логнормального распределения против сложной альтернативной гипотезы .

Но в силу трудности получения значения , критическую область получить затруднительно и, следовательно, проверить гипотезы не представляется возможным. По этой же причине нахождение мощности невозможно так как она находится по формуле:

Теперь вычислим ошибки 1-го и 2-го рода.

Уровень значимости критерия проверки H0 – вероятность ошибочно отвергнуть гипотезу H0, когда она верна:

– для непрерывного распределения, – плотность распределения наблюдаемой случайной величины при условии, что верна гипотеза H0.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Принята по результатам расчета** | **Вывод** |
| **Верна на самом деле** |  |  | Вероятность правильного решения |
|  | Ошибка первого рода (ложная тревога, ложное срабатывание). Вероятность – |
|  |  | Вероятность правильного решения |
|  | Ошибка второго рода (пропуск цели, пропуск события, ложноотрицательное срабатывание).  Вероятность ошибки |

**Для уровня значимости :**

Отклонить : результаты вычислений предполагают, что .

Отказ отклонить : мы принимаем .

Ошибка I рода: с 5% риском ошибки мы примем , хотя действительный результат .

Ошибка II рода: так как у нас отсутствовали необходимые свидетельства, мы принимаем , хотя в действительности .

## 8.2 Гипотезы о виде распределения.

### **8.2 Критерий Колмогорова**

Используется только при условии, что с.в. принимают бесконечно много значений.

Выберем критическое значение используя табличные значения:

Изображение выглядит как текст, Шрифт, число, линия

Автоматически созданное описание

Рисунок 11.

Зададимся уровнем значимости следовательно:

В результате выполнения скрипта на python была получена статистика:

Критерий имеет правостороннюю критическую область, следовательно, т.к. , гипотеза о том, что генеральная совокупность имеет логнормальное распределение для полученных оценок отвергается.

### **8.2.2 Критерий Пирсона**

Процедура проверки гипотез с использованием критериев типа предусматривает группирование наблюдений. Область определения случайной величины разбивают на k непересекающихся интервалов граничными точками.

В соответствии с заданным разбиением подсчитывают число выборочных значений, попавших в i-й интервал, и вероятности попадания в интервал:

*, F – теоретическая ф-я распределения*

Статистика критерия согласия Пирсона определяется соотношением:

По аналогии с построением гистограммы, был написан скрипт на Python, который разбивает выборку на интервалы размером 1, результат разбиения ниже

Изображение выглядит как текст, Шрифт, снимок экрана, чек

Автоматически созданное описание

Рис.12  
Изображение выглядит как текст, число, Параллельный, Шрифт

Автоматически созданное описание

Рис.13

Рассчитаем число степеней свободы:

Найдем табличное значение критической точки для степени свободы **6** и уровня значимости **0.05**:

Следовательно:

В результате выполнения скрипта на python была получена статистика:

## Следовательно, с помощью этого критерия мы принимаем гипотезу о том, что параметр равен значению, которое было получено с помощью метода моментов. А параметр, полученный с помощью максимального правдоподобия, отвергаем в пользу альтернативного

## 8.3 Гипотезы об однородности выборок.

Задача: верна ли гипотеза две независимые выборки получены из одной генеральной совокупности?

То есть проверяем гипотезы вида:

.  
  
Близким по виду к логнормальному распределению является распределение Вейбулла  
  
Изображение выглядит как текст, линия, График, снимок экрана

Автоматически созданное описание

Рис.14 Логнормальное распределение рядом с распределением Вейбулла  
  
Метод обратного преобразования – способ генерации случайных величин с заданной функцией распределения, путём модификации работы генератора равномерно распределённых чисел.

Рассмотрим функцию распределения y=F(x) интересующего нас закона распределения и попробуем выразить обратную функцию:

F(x) всегда принимает значения из интервала . Тогда, если сгенерировать случайную величину , то через обратную функцию можно получить значение , принимающие все значения из области определения.

Тогда, можно вычислить обратную функцию, выглядеть она будет как:

Числа были сгенерированы с использованием генератора Mersenne Twister (numpy.random). Этот псевдослучайный генератор генерирует числа с равномерным распределением, следовательно можем его использовать.

Полученная выборка распределения из 300 элементов выглядит следующим образом:  
Изображение выглядит как текст, линия, График, диаграмма

Автоматически созданное описание

Рисунок 15. Распределение по выборке

1. **Источники информации**
2. Логарифмически нормальные распределения в разных науках: Ключи и подсказки: О прелестях статистики и о том, как механические модели, напоминающие игровые автоматы, дают ссылку на удобный способ характеристики логарифмически нормальных распределений, которые могут обеспечить более глубокое понимание изменчивости и вероятности - нормальные или логарифмически нормnальные: вот в чем вопрос. Экхард Лимперт, Вернер А. Шталь, Маркус Эббт. BioScience, Том 51, Выпуск 5, май 2001 года, Страницы 341-352. [Электронный ресурс] Режим доступа: <https://academic.oup.com/bioscience/article/51/5/341/243981?login=false>

### 2) [Анализ случайных величин и неопределенностей. Ен Бай, Вэй-Лян Цзинь, в книге "Проектирование морских конструкций" (Второе издание), 2016 г. 33.4. Выбор функций распределения.](https://www.sciencedirect.com/science/article/pii/B9780080999975000332) [Электронный ресурс] Режим доступа: <https://www.sciencedirect.com/topics/engineering/lognormal-distribution#:~:text=The%20lognormal%20distribution%20is%20used,rather%20than%20a%20normal%20distribution>

# 3) CFA - Логнормальное распределение вероятностей. [Электронный ресурс] Режим доступа: <https://fin-accounting.ru/cfa/l1/quantitative/cfa-lognormal-probability-distribution>

1. Лемма Неймана-Пирсона. [Электронный ресурс] Режим доступа: https://studfile.net/preview/9124450/page:4/

1. May, R. M. (1976). Simple mathematical models with very complicated dynamics. Nature, 261(5560), 459-467

   <https://www.researchgate.net/publication/237005499_Simple_Mathematical_Models_With_Very_Complicated_Dynamics>

   Sartwеll Р.E. The incubation period of infections diseases. Am J Hygiene, 1950; 51:310–8.

   <http://www.columbia.edu/itc/hs/pubhealth/p8462/misc/sartwell.pdf> [↑](#footnote-ref-1)