1. Министерство образования и науки Российской Федерации
2. Санкт-Петербургский Политехнический Университет Петра Великого
3. —
4. Институт кибербезопасности и защиты информации

К У Р С О В А Я Р А Б О Т А

«Оценка распределения »

по дисциплине «Теория вероятностей и математическая статистика»

Выполнили:

Студенты гр. 4851004/10001 Гусева М. Е.

<подпись>

Резникова М. С.

<подпись>

Проверил:

<подпись>

Санкт-Петербург

2023

СОДЕРЖАНИЕ

|  |  |
| --- | --- |
| Содержание……………………………………………………………………...... | 2 |
| Введение…………………………………………………………………………... | 3 |
| Практическая часть………………………………………………………………. | 4 |
| 1.Исследование распределения ……………...…………………………… |  |
| 2.Вычисление статистик выборки ………………………………………….. |  |
| 3.Построение эмпирической функции распределения………..…………… |  |
| 4.Построение гистограммы………………………………………………….. |  |
| 5.Описание параметров распределения…………………………………….. |  |
| 6.Понятие точечных оценок…………………………………………………. |  |
| 7.Понятие интервальных оценок……………………………………………. |  |
| 8.Понятие статистических критериев………………………………………. |  |
| Заключение……………………………………………………………….…….... | 16 |
| Список используемых источников……………………………………………... | 17 |
|  |  |
|  |  |

##### ВВЕДЕНИЕ

**Цель работы**

Провести исследование распределения , используя различные методы и критерии.

**Задачи**

1. Изучить основные характеристики распределения .
2. Написать функции для вычисления статистик выборки.
3. Написать функцию для построения эмпирической функции распределения.
4. Написать функцию для построения гистограммы
5. Определить вид теоретической функции распределения
6. Оценить параметры распределения выборки методом моментов и методом максимального правдоподобия
7. Оценить параметры распределения выборки с помощью интервальной оценки с уровнями доверия
8. Проверить гипотезы статистическими критериями

##### ПРАКТИЧЕСКАЯ ЧАСТЬ

##### Исследование распределения

В теории вероятностей и статистике, распределение хи-квадрат (также хи-квадрат или χ -распределение ) с k степенями свободы — это распределение суммы квадратов k независимых стандартных нормальных случайных величин. Распределение хи-квадрат является частным случаем гамма-распределения и является одним из наиболее широко используемых распределений вероятностей в статистике вывода, особенно в проверке гипотез и построении доверительных интервалов

Впервые это распределение было исследовано астрономом Ф. Хельмертом в 1876 году. В связи с гауссовской теорией ошибок он исследовал суммы квадратов n независимых стандартно нормально распределенных случайных величин. Позднее Карл Пирсон дал имя данной функции распределения «хи – квадрат»

Распределение хи-квадрат имеет множество применений в выводимой статистике, например, в тестах хи-квадрат и в оценке отклонений. Оно входит в проблему оценки среднего значения нормально распределенной совокупности и проблему оценки наклона линии регрессии через ее роль в t-распределении Стьюдента, также во все задачи дисперсионного анализа благодаря своей роли в F-распределении, которое представляет собой распределение отношения двух независимых хи-квадратов случайных величин, каждая из которых делится на соответствующие степени свободы. Распределение хи-квадрат часто встречается в магнитно-резонансной томографии

Абрамович, Милтон; Стеган, Ирен Энн*,* ред. (1983) [июнь 1964]. "Глава 26". Справочник по математическим функциям с формулами, графиками и математическими таблицами. Серия "Прикладная математика". Том 55 (Девятое переиздание с дополнительными исправлениями десятого оригинального издания с исправлениями (декабрь 1972); первое изд.). Вашингтон, округ Колумбия; Нью-Йорк: Министерство торговли Соединенных Штатов, Национальное бюро стандартов; Dover Publications. стр. 940.

NIST (2006). Справочник по инженерной статистике – Распределение Хи-квадрат

 Hald 1998, стр. 633-692, 27. Выборочные распределения при нормальности.

Деккер А. Дж., Сийберс Дж., (2014) "Распределения данных на магнитно-резонансных изображениях: обзор"

1. **Вычисление статистик выборки**

Для дальнейшей работы был установлен интерпретатор языка Python, согласно инструкции, указанной на официальном сайте, и среда интерактивной разработки jupyter notebook.

Затем была считана выборка **X** из файла варианта с помощью стандартных средств языка Python и написаны функции для вычисления статистик выборки **X** из варианта.

**Сумма элементов выборки:** sum = 204.14030799999995

**Выборочное среднее:** = 0.6804676933333331

**Медиана:** 0.148249

**Мода:** 'Anti-modal sampling'

**Размах выборки:** 18.485556

**Смещенная дисперсия:** 10.219905448329307

**Несмещенная дисперсия:** 10.254085734109673

**Выборочный начальный момент k-ого порядка**: 2307641028.235439 (при k = 10)

**Выборочный центральный момент k-го порядка:** 1366211156.1902268 (при k = 10)

1. **Построение эмпирической функции распределения**

Эмпирической (опытной) функцией распределения или функцией распределения выборки называют такую функцию, которая определяет для каждого значения x частоту событий X < x и предназначена для оценки теоретической функции распределения генеральной совокупности в математической статистике.

Эмпирическая функция распределения находится по формуле:

— объем выборки;

— количество наблюдений (вариантов) меньше x.

Алгоритм:

1. Вычислить объем выборки
2. Отсортировать выборку
3. Рассчитать для каждого отрезка((, ), (, ), … , (, )) вероятность попадания туда случайной величины (из свойства эмпирической функции распределения: (x)=0 при x ≤ 1, (x)=1 при x > 299)
4. Построить график полученной функции

Была написана функция для построения эмпирической функции распределения на языке Python, а затем построены графики для исходной выборки и подвыборок из 10, 100, 200 значений .

Изображение выглядит как текст, снимок экрана, программное обеспечение, компьютер

Автоматически созданное описание

Рисунок 1 − Случайная подвыборка из 10 значений

Изображение выглядит как текст, диаграмма, линия, снимок экрана

Автоматически созданное описание

Рисунок 2 − График эмпирической функции распределения для подвыборки из 10 значений

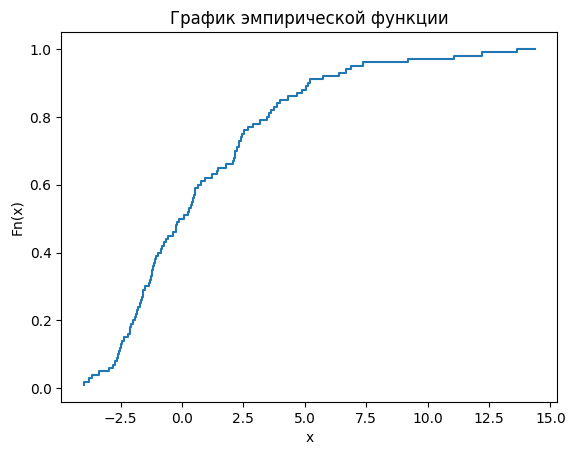


Рисунок 3 − График эмпирической функции распределения для подвыборки из 100 значений

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Рисунок 4 − График эмпирической функции распределения для подвыборки из 200 значений

Изображение выглядит как текст, диаграмма, линия, График

Автоматически созданное описание

Рисунок 5 − График эмпирической функции распределения для выборки из 300 значений

Пусть Функция – функция распределения случайно величины . Другими словами, функцией распределения случайной величины называется, определяющая для каждого наперед заданного значения вероятность того, что случайная величина примет значение меньшее, чем .

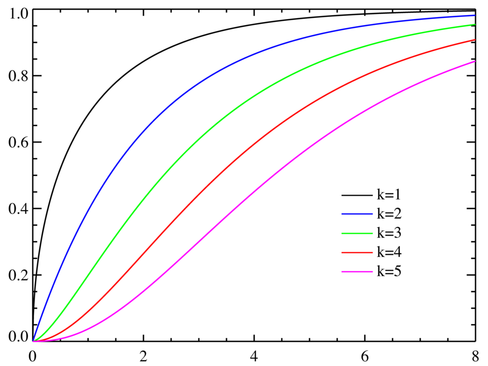


Рисунок 6 − График функции распределения для

Полученная эмпирическая функция распределения согласовывается с теоретической функцией распределения со степенями свободы равным 2. Также стоит отметить, что оно медленно стремиться к нормальному с увеличением числа степеней свободы.

1. **Построение гистограммы**
2. **Описание параметров распределения**

Параметры распределения — это его числовые характеристики, указывающие, где "в среднем" располагаются значения признака, насколько эти значения изменчивы и наблюдается ли преимущественное появление определенных значений признака.

Распределение из варианта не симметрично и существенно смещено вправо, а также имеет разную форму в зависимости от количества степеней свободы.

Следовательно, оно имеет два параметра:

df − Число степеней свободы

loc − Параметр смещения

Функция распределения имеет следующий вид:

, Где Г и γ обозначают соответственно полную и неполную гамма-функции, то есть:

Степень свободы (Degree of Freedom, df) в статистике — это количество значений или наблюдений в выборке, которые могут быть изменены независимо друг от друга без изменения ее структуры. Можно сказать, что это количество переменных, которые остаются свободными для варьирования после того, как структура выборки была определена.

Одним из важных факторов, влияющих на степень свободы, является размер выборки. Чем больше выборка, тем больше степень свободы, значит, чем больше выборка, тем менее вероятно получение ошибочных результатов в статистических тестах.

Если случайные величины независимы и все имеют стандартное нормальное распределение , тогда говорят, что случайная величина X, являющаяся суммой квадратов стандартных нормальных величин в количестве n штук, имеет распределение хи-квадрат с n степенями свободы (). Например, при построении доверительных интервалов для оценки дисперсии число степеней свободы равно df=n-1, где n – размер выборки .

*∼*

Параметр смещения определяет нижнюю границу диапазона распределения.

Затем было определено влияние параметров на график функции распределения.

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 7 – Вид функции с параметрами df = 1, loc = 0

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 8 − Вид функции с параметрами df = 15, loc = 0

Изображение выглядит как текст, линия, диаграмма, График

Автоматически созданное описание

Рисунок 9 − Вид функции с параметрами df = 3, loc = 0

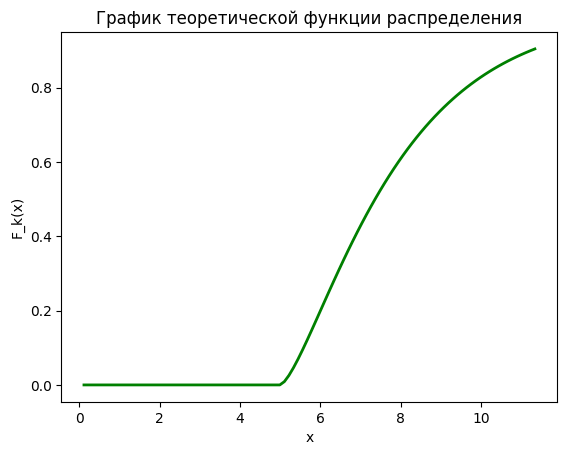


Рисунок 10 − Вид функции с параметрами df = 3, loc = 5

**Выводы о влиянии параметров на вид функции распределения:**

1. **Понятие точечных оценок**
2. **Понятие интервальных оценок**

Были оценены параметры распределения для выборки размера 300 значений с помощью интервальной оценки с уровнями доверия и

Для оценки математического ожидания получено:

Таким образом, получается интервал с уровнем доверия . (В данный интервал попадает 95% случайных величин)

Таким образом, получается интервал с уровнем доверия . (В данный интервал попадает 99% случайных величин)

Для оценки дисперсии была использована формула:

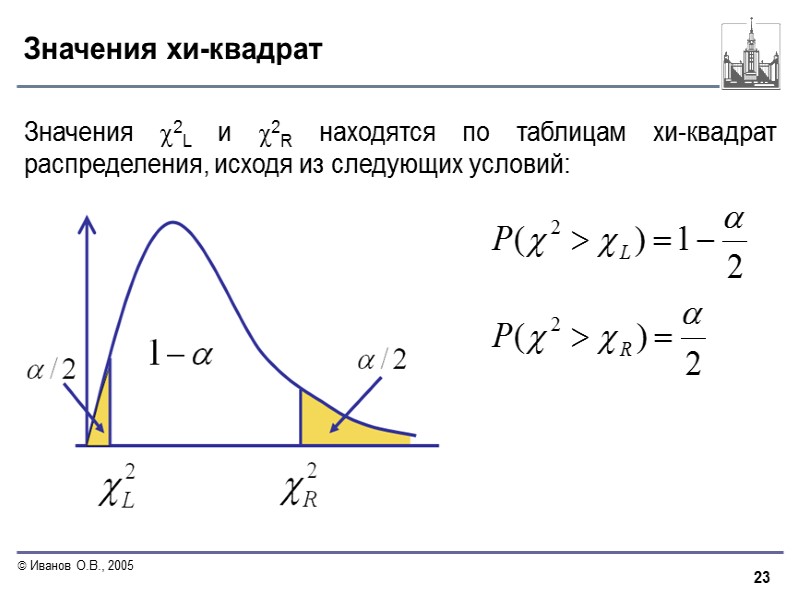


Рисунок 11 – Условия получения значений

Получены следующие интервалы:

I = (8,761; 12,078) – для уровня доверия =0,95.

I = (8,355; 12,745) – для уровня доверия =0,99.

Затем были соотнесены полученные интервальные и точечные оценки.

1. **Понятие статистических критериев**
   1. **Гипотезы о параметрах распределения**
   2. **Гипотезы о виде распределения**

Пусть выдвинута гипотеза о том, что выборка подчиняется распределению с k = 4:

**Критерий**

Основной критерий согласия – *критерий*  (или *критерий согласия Пирсона*). Он является одним из самых распространённых критериев гипотез. Данный критерий можно употреблять, когда все результаты испытаний можно разделить на конечное число категорий. Все возможные результаты независимых испытаний разделены на категорий. Обозначим вероятность того, что результат испытания попадает в ую категорию, как . А число испытаний, результаты которых попали в ую категорию, обозначим . Далее в общем случае требуется разбить область значений случайной величины на интервалов , , … , , а затем определить теоретические вероятности попадания в интервал и числа элементов выборки, попавших в эти интервалы.

Критерий Хи-квадрат предпочтителен, когда исследуются большие объемы выборок. При малых объемах выборок этот критерий практически не пригоден.

N = 157

min = 0.006333

max = 14,359223

Размер каждого интервала был рассчитан по формуле Стёрджесса:

Это означает, что число категорий не должно превосходить числа 8.

Затем выборка была разбита на 7 категорий:

Далее были вычислены теоретические вероятности и посчитан критерий .

|  |  |  |  |
| --- | --- | --- | --- |
| **Относительная частота()** | **Начало интервала** | **Конец интервала** | **Теоретическая вероятность()** |
| 0,191083 | 0 | 1,80044425 | 0,227598926 |
| 0,433121 | 1,80044425 | 3,5945555 | 0,308753724 |
| 0,261146 | 3,5945555 | 5,38866675 | 0,21395688 |
| 0,146497 | 5,38866675 | 7,182778 | 0,123151683 |
| 0,0828025 | 7,182778 | 8,97688925 | 0,064859043 |
| 0,0318471 | 8,97688925 | 10,7710005 | 0,032417886 |
| 0,0254777 | 10,7710005 | 14,359223 | 0,023029289 |

Получаем:

11,93545

Следовательно, гипотеза о подчинении выборки распределению с 4 степенями свободы принимается, а отвергается.

**Критерий Колмогорова**

Критерий Колмогорова применяют, если исследуемая случайная величина принимает бесконечно много значений. По выборке с непрерывной теоретической функцией распределения строим эмпирическую функцию распределения . Необходимо оценить, насколько далеко построенная эмпирическая функция распределения отклоняется от теоретической функции распределения.

Расстояние: .

*, α = 0,05 =>*

Таким образом, если расстояние мало, то можно сделать вывод о том, что основная гипотеза верна, данные согласуются. Стоит также отметить, что удобнее использовать не само расстояние, а нормированное расстояние. Как только мы нормируем статистику критерия Колмогорова, получаем случайную величину, имеющую распределение Колмогорова. Итак, чем больше расстояние, тем больше эмпирическая функция отклоняется от теоретической функции распределения. Соответственно, тем меньше вероятность того, что сформулированная гипотеза верна.

Для проведения теста выборка из положительных значений была разбита на 9 интервалов. Затем были рассчитаны накопленные эмпирические частоты и накопленные теоретические частоты, а также найден максимум из разности относительных и теоретических частот по модулю.

= 0.08412335092232459

Так как условие выполняется, то можно необходимо принять гипотезу о том, что данная выборка подчиняется распределению с 4 степенями свободы.

* 1. **Гипотезы об однородности выборок**

##### ЗАКЛЮЧЕНИЕ

##### СПИСОК ИСПОЛЬЗУЕМЫХ ИСТОЧНИКОВ

1. ФеллерУ. Введение в теорию вероятностей и ее приложения //Нью-Йорк: JohnWiley&SonsInc, 1971. Том II (2 изд.),стр. 704.
2. Conover W.J. Several k-sample Kolmogorov-Smirnov tests //The Annals of Mathematical Statistics, 1965. Vol. 36, No. 3. – P.1019-1026.
3. Conover W.J. Practical Nonparametric Statistics [Электронныйресурс] //Wiley, 1999. – P 584
4. Zhang J. Powerful Two-Sample Tests Based on the Likelihood Ratio //Technometrics, 2006. V. 48. No. 1. – P.95-103.
5. ЛемешкоБ.Ю., ЛемешкоС.Б., ВеретельниковаИ.В. Мощность выборочных критериев проверки однородности законов [Электронный ресурс].URL: