Курсовая работа по курсу математической статистики

Сааков А.С. СКБ182
Версия от 08.12.2020

Содержание

1. Домашнее задание. Вероятностные распределения

In [1]:
```
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import maxwell
import scipy.stats as sts
from scipy.stats import geom
from random import random
from collections import Counter
import copy
import math
from math import *
from random import *
import pandas as pd
import calendar
import statsmodels.api as sm

plt.style.use('ggplot')  # Красивые графики
plt.rcParams['figure.figsize'] = (15, 5)  # Размер картинок
```

1.1. Геометрическое распределение

1.1.1. Описание основных характеристик распределения

Функция вероятности дискретного распределения: $P_{\xi}(x) = pq^x, x \in \{0, 1, 2, \dots\}$

Математическое ожидание:

$$M\xi = \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} kq^{k-1} = p\sum_{k=1}^{\infty} \frac{dq^k}{dq} = p\frac{d}{dq}(\sum_{k=1}^{\infty} q^k) = p\frac{d}{dq}(\frac{q}{1-q}) = p\frac{1}{(1-q)^2} = \frac{1}{p}$$

Дисперсия:

$$D\xi = M(\xi - M\xi)^2 = M\xi^2 - (M\xi)^2 = M(\xi(\xi-1) + \xi) - M\xi^2 = M(\xi(\xi-1)) + M\xi - (M\xi)^2 =$$

$$M(\xi(\xi-1)) = p\sum_{k=1}^{\infty} k^2 q^{k-1} = pq\sum_{k=0}^{\infty} \frac{d^2 q^k}{dq^2} = pq\frac{d^2}{dq^2}(\sum_{k=0}^{\infty} q^k) = pq\frac{d^2}{dq^2}(\frac{1}{1-q}) = pq\frac{2}{(1-q)^3} =$$

$$D\xi = M\xi^2 + M\xi - (M\xi)^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q-1}{p^2} + \frac{1}{p} = \frac{2q-1+p}{p^2} = \frac{2q-1+1-q}{p^2} =$$

In [2]:

```
for p in [0.1, 0.4, 0.6, 0.9]:
    geom_rv = sts.geom(p)
    sample = geom_rv.rvs(1000)
    plt.hist(sample, density = True, label='p = {}'.format(p))
    plt.legend()
    plt.show()
print('Рис. 1: 1.1.1, Гистограмма вероятностей дискретного распределения')
```

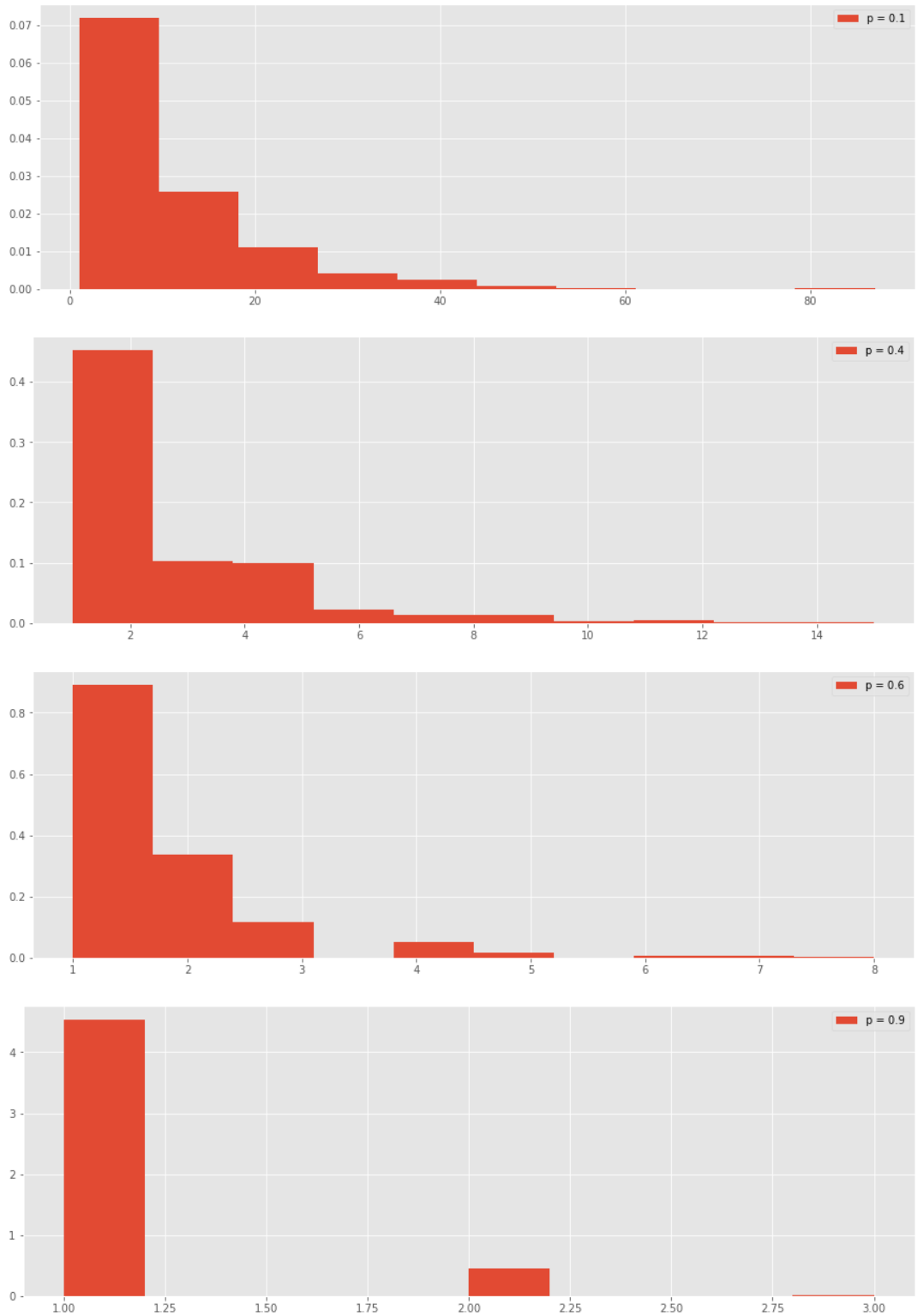


Рис. 1: 1.1.1, Гистограмма вероятностей дискретного распределения

Мода M_0 - значение во множестве наблюдений, которое встречается наиболее часто, для дискретной случайной величины определяется с помощью гистограммы вероятностей.
Из гистограмм видно, что $M_0 = 1$

In [3]:
```
for p in [0.1, 0.4, 0.6, 0.9]:
    n = np.arange(0, 8, 1)
    plt.step(n, 1-(1-p)**(n+1), label='p = {}'.format(p))
    plt.legend()
plt.show()
print('Рис. 1: 1.1.1, Гистограмма вероятностей дискретного распределения')
```

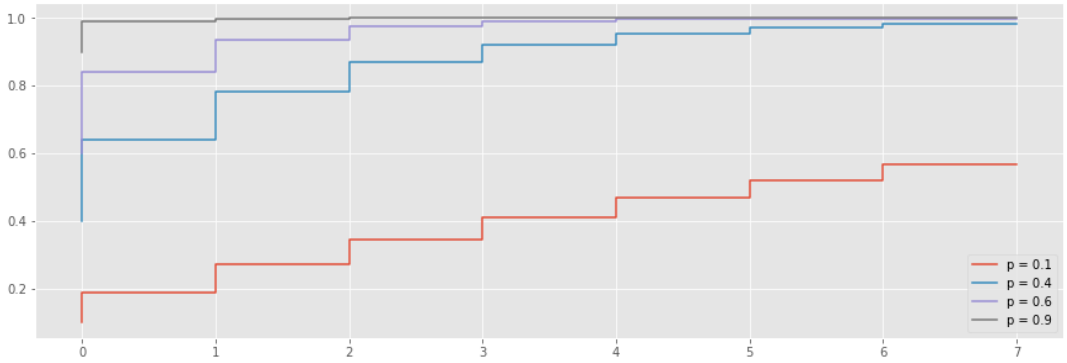


Рис. 1: 1.1.1, Гистограмма вероятностей дискретного распределения

Медиана Me находится из уравнения $P_{\xi}(x) = 0.5$

$$\begin{cases} p + qp + q^2p + \ldots + q^{Me-1}p \geq \frac{1}{2} \\ q^{Me-1}p + q^{Me}p + q^{Me+1}p + \ldots \geq \frac{1}{2} \end{cases}$$

$$\begin{cases} p\frac{1-q^{Me}}{1-q} \geq \frac{1}{2} \\ q^{Me-1}p\frac{1}{1-q} \geq \frac{1}{2} \end{cases}$$

$$\begin{cases} 1 - q^{Me} \geq 2^{-1} \\ q^{Me-1} \geq 2^{-1} \end{cases}$$

$$\begin{cases} q^{Me} \leq 2^{-1} \\ q^{Me-1} \geq 2^{-1} \end{cases}$$

$$\begin{cases} Me \cdot log_2 q \leq -1 \\ (Me-1) log_2 q \geq -1 \end{cases}$$

Отсюда $-\frac{1}{log_2 q} \leq Me \leq 1 - \frac{1}{log_2 q}$

Примеры событий, которые могут быть описаны выбранными случайными величинами

Типичные интерпретации геометрического распределения: описывает количество испытаний n до первого успеха при вероятности наступления успеха в каждом испытании p. Если n подразумевается номер испытания, в котором наступил успех, то геометрическое распределение будет описываться следующей формулой:

$$Geom_p(n) = q^{n-1}p$$

Геометрическое распределение считается дискретной версией экспоненциального распределения. Предположим, что эксперименты Бернулли проводятся через равные промежутки времени. Тогда геометрическая случайная величина X - это время, измеренное в дискретных единицах, которое проходит до того, как мы добьемся первого успеха. . Но если мы хотим смоделироватьвремя, прошедшее до того, как данное событие произойдет в непрерывном времени, то подходящим распределением для использования будетэкспоненциальное распределение. С математической точки зрения геометрическое распределение обладает тем же свойством без памяти,которым обладает экспоненциальное распределение: в экспоненциальном случае вероятность того, что событие произойдет в течениезаданного временного интервала, не зависит от того, сколько времени уже прошло, а событие не произошло; в геометрическом случаевероятность того, что событие произойдет в данный момент (дискретное) времени, не зависит от того, что произошло раньше, потому чтоэксперимент Бернулли, проведенный в каждый момент времени, не зависит от предыдущих испытаний.Геометрическое распределение полезно для определения вероятности успеха при ограниченном количестве испытаний, что очень применимо креальному миру, в котором неограниченные испытания редки. Поэтому неудивительно, что различные сценарии хорошо моделируютсягеометрическими распределениями:

- В спорте, особенно в бейсболе, геометрическое распределение полезно для анализа вероятности того, что отбивающий получит удар,прежде чем он получит три удара; здесь цель - добиться успеха за 3 испытания.
- При анализе затрат и выгод, например, когда компания решает, финансировать ли исследовательские испытания, которые в случае успеха принесут компании некоторую предполагаемую прибыль, цель состоит в том, чтобы достичь успеха до того, как затраты превысят потенциальную выгоду.
- В тайм-менеджменте цель состоит в том, чтобы выполнить задачу за установленный промежуток времени.Другие приложения, подобные вышеупомянутым, также легко создаются. Фактически, геометрическое распределение применяется наинтуитивном уровне в повседневной жизни на регулярной основе.

1.1.3 Описание способа моделирования выбранных случайных величин

Существует такой способ реализации метода обратных функций, при котором трудоемкость по крайней мере формально не зависит от p. Действительно, накопленная вероятность $s_{n+1} = p_0 + \ldots + p_n$ для геометрического распределения имеет вид

$$s_{n+1} = \sum_{i=0}^{n} p(1-p)^i = 1 - (1-p)^{n+1}$$

Поэтому событие $\{\xi = n\}$ приобретает вид

$$\{\xi = n\} = \{s_n < \alpha \leq s_{n+1}\} = \{1 - (1-p)^n < \alpha \leq 1 - (1-p)^{n+1}\} = \{(1-p)^{n+1} \leq 1 - \alpha < (1$$

$$= \{(n+1)ln(1-p) \leq ln(1-\alpha) < n \cdot ln(1-p)\} = \{n < \frac{ln(1-\alpha)}{ln(1-p)} \leq n+1\},$$

и тем самым

$$\xi = [\frac{ln(1-\alpha)}{ln(1-p)}]$$

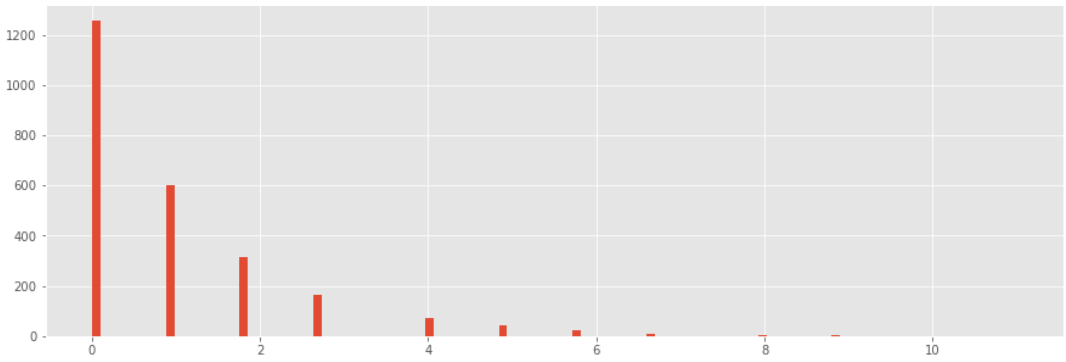
Эту же формулу можно получить по-другому. Пусть ν - случайная величина, имеющая показательное распределение с параметром λ и $\xi = [n]$. Тогда при $n \geq 0$

$$P(\xi = n) = P(n \leq \nu < n+1) = e^{-n\lambda} - e^{-(n+1)\lambda} = (1 - e^{-\lambda})e^{-n\lambda}.$$

Поскольку случайная величина $\frac{-ln(1-\alpha)}{\lambda}$ имеет показательное распределение с параметров λ, то взяв $\lambda = -ln(1-p)$, приходим к формуле $\xi = [\frac{ln(1-\alpha}{ln(1-p)}]$

In [4]:

```
def sample_(N=2500, scale = 0.5):
    for x in range(N):
        je = np.log(random())//np.log(1-scale)#Генерирование случайных чисе
    return je
def Geom(n, p=0.5):
    x=[sample_(scale=p) for x in range(n)]
    #print(x)
    return x
plt.hist(Geom(2500,0.5),25, width = 0.1)
plt.show()
```



1.2. Распределение Максвелла

1.2.1. Описание основных характеристик распределения

Математическое ожидание:

$$M\xi = \int_0^\infty x\sqrt{\frac{2}{\pi}}\frac{x^2}{\lambda^3}e^{-\frac{x^2}{2\lambda^2}}dx = \sqrt{\frac{2}{\pi}}\frac{1}{\lambda^3}\int_0^\infty x^3 e^{\frac{-x^2}{2\lambda^2}}dx = 2\lambda^4\cdot\sqrt{\frac{2}{\pi}}\frac{1}{\lambda^3} = 2\lambda\sqrt{\frac{2}{\pi}}$$

Дисперсия:

$$D\xi = M(\xi - M\xi)^2 = M\xi^2 - (M\xi)^2 = M(\xi(\xi-1)+\xi) - (M\xi)^2 = M(\xi(\xi-1)) + M\xi - (M\xi)^2$$

$$M(\xi(\xi-1)) = \int_0^\infty x^2 f(x)dx = \int_0^\infty x^2\sqrt{\frac{2}{\pi}}\frac{x^2}{\lambda^3}e^{-\frac{x^2}{2\lambda^2}} = \sqrt{\frac{2}{\pi}}\frac{1}{\lambda^3}\int_0^\infty x^4 e^{-\frac{x^2}{2\lambda^2}} = \sqrt{\frac{2}{\pi}}\frac{1}{\lambda^3}\cdot 3\lambda^5$$

$$D\xi = M\xi^2 - (M\xi)^2 = 3\lambda^2 - 4\lambda^2\cdot\frac{2}{\pi} = \frac{3\pi-8}{\pi}\lambda^2$$

Также использовались известные интегралы, который был взят из курса физики:

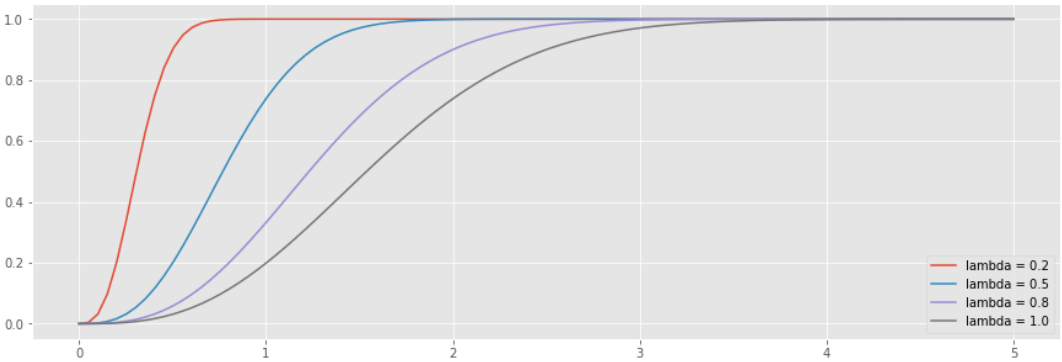
$$\int_0^\infty x^3 e^{-x^2/2\lambda^2}dx = \frac{1}{2(\frac{1}{\lambda^2})^2}\cdot 4 = 2\lambda^4$$

$$\int_0^\infty x^4 e^{-x^2/2\lambda^2}dx = \frac{3}{8}\sqrt{\pi}(\frac{1}{2\lambda^2})^{-\frac{5}{2}} = 3\lambda^5\sqrt{\frac{\pi}{2}}$$

In [5]:

```
for lambd in [0.2,0.5,0.8,1.0]:
    maxwell_rv = sts.maxwell(scale = lambd)
    x = np.linspace(0,5,100)
    cdf = maxwell_rv.cdf(x)
    plt.plot(x, cdf, label = 'lambda = {}'.format(lambd))
    plt.legend()
print('Рис.4: График функции распределения')
```

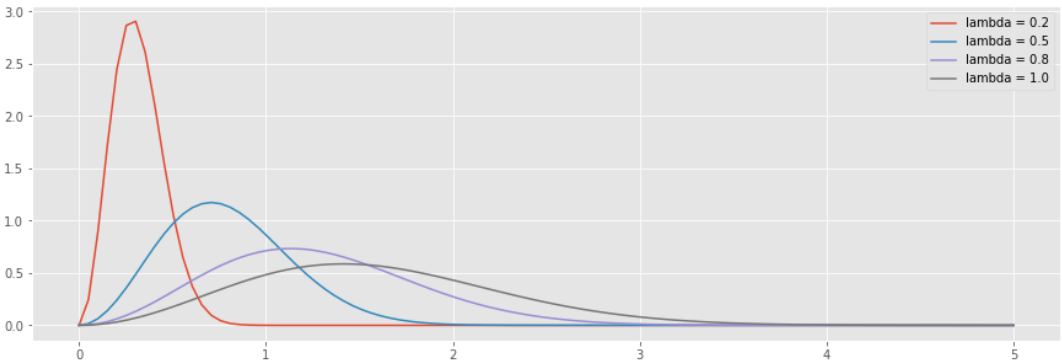
Рис.4: График функции распределения



In [6]:

```
for lambd in [0.2,0.5,0.8,1.0]:
    maxwell_rv = sts.maxwell(scale = lambd)
    x = np.linspace(0,5,100)
    pdf = maxwell_rv.pdf(x)
    k = max(pdf)
    plt.plot(x, pdf, label ='lambda = {}'.format(lambd))
    plt.legend()
print('\n')
print('Рис.5: График плотности вероятности распределения')
```

Рис.5: График плотности вероятности распределения



Модой абсолютно непрерывного распределения называют любую точку локального максимума плотности распределения. $f'(x) = \sqrt{\frac{2}{\pi}}\frac{x^2}{\lambda^3}e^{-\frac{x^2}{2\lambda^2}} = \frac{4x}{\lambda^3\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}} - \frac{2x^2}{\lambda^3\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}\frac{x}{\lambda^2} = 0$

$\frac{4x}{\lambda^3\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}} = \frac{2x^2}{\lambda^3\sqrt{2\pi}}e^{-\frac{x^2}{2\lambda^2}}\frac{x}{\lambda^2}$

$4 = 2x\frac{x}{\lambda^2}$

$x^2 = 2\lambda^2$

$x = M_0 = \lambda\sqrt{2}$

Медиана

$\int_0^{Me}\sqrt{\frac{2}{\pi}}\frac{1}{\lambda^3}x^2e^{-\frac{-x^2}{2\lambda^2}}dx = \frac{1}{2}$

$\int_0^{Me}x^2e^{-\frac{-x^2}{2\lambda^2}}dx = \frac{\lambda^3}{2}\sqrt{\frac{\pi}{2}}$

$(-\lambda^2e^{-\frac{x^2}{2\lambda^2}}x + \lambda^3\sqrt{\frac{\pi}{2}})|_0^{Me} = \frac{\lambda^3}{2}\sqrt{\frac{\pi}{2}}$

$-Me\lambda^2e^{\frac{-Me^2}{2\lambda^2}} = -\frac{\lambda^3}{2}\sqrt{\frac{\pi}{2}}$

$Me\cdot e^{\frac{-Me^2}{2\lambda^2}} = \frac{\lambda^2}{2}\sqrt{\frac{\pi}{2}}$

$Me \approx 1,5383\lambda$

1.2.2. Примеры событий, которые могут быть описаны выбранными случайными величинами

Впервые распределение было определено и использовалось для описания скоростей частиц в идеализированных газах, где частицы свободноперемещаются внутри стационарного контейнера, не взаимодействуя друг с другом, за исключением очень коротких столкновений, в которыхони обмениваются энергией и импульсом друг с другом или со своим тепловым окружением. Термин «частица» в этом контексте относится только к газообразным частицам (атомам или молекулам), и предполагается, что система частиц достигла термодинамического равновесия.Энергии таких частиц следуют так называемой статистике Максвелла – Больцмана, а статистическое распределение скоростей выводится путем приравнивания энергии частиц к кинетической энергии.Распределение Максвелла – Больцмана в основном применяется к скоростям частиц в трех измерениях, но оказывается, что оно зависит только от скорости (величины скорости) частиц. Распределение вероятности скорости частицы указывает, какие скорости более вероятны: частица будет иметь скорость, выбранную случайным образом из распределения, и с большей вероятностью будет находиться в одном диапазоне скоростей, чем в другом.

При тепловом равновесии (T=const) $u_{\text{кв}}$ молекул газа остается постоянной и равной $u = \sqrt{\frac{3kT}{m}}$

Это объясняется тем, что в газе устанавливается стационарное статическое распределение молекул по значениям скоростей, называемое распределением Максвелла:

$f(u) = \frac{dN(u)}{Ndu} = 4\pi(\frac{m}{2\pi kT})^{\frac{3}{2}} \cdot u^2 \cdot e^{-\frac{mu^2}{2kT}}$

В теории вероятностей рассматривается распределения Максвелла, в котором $x = u$ и $\frac{1}{\lambda^2} = \frac{m}{kT}$

Нетипичной интерпретацией распределения Максвелла будут данные, которые представляют время ремиссии (в месяцах) у пациентов с раком мочевого пузыря и первоначально использовались Lee и Wang.

Ремиссия (лат. remissio «уменьшение, ослабление») — период течения хронической болезни, который проявляется значительным ослаблением (неполная ремиссия) или исчезновением (полная ремиссия) её симптомов (признаков заболевания)

1.2.3. Описание способа моделирования выбранных случайных величин

Способ 1: Существует полярный метод (группа полярных методов предназначена для моделирования распределений, так или иначе связанных с двумерными распределениями, инвариантными относительно вращений), где моделируются две независимые случайные величины ξ_1, ξ_2 , каждая из которых имеет распределение N(0,1).

Полярные координаты. Каждая точка $X = (x, y)T \in R^2\backslash\{0\}$ может быть однозначно представлена в виде $X = ||X||\bar{e}$, где $\bar{e} = 1$. Полагая $s = ||X||$ и $\bar{e} = (\cos t, \sin t)^T$, где$t \in [0, 2\pi)$, получаем биекцию $\phi : (x, y)^T \to (s, t)^T$, действующую из $R^2\backslash\{0\}$ в $(0, \infty) \times [0, 2\pi)$. Конечно, переменные (s, t) являются полярными координатами вектора X, а обратное отображение $\psi : (0, \infty) \times [0, 2\pi) \to R^2\backslash\{0\}$ имеет вид $x = s\cos t, y = s\sin t$ с якобианом $det\psi'(s, t) = s$.

Если теперь рассмотреть случайный вектор $\bar{\xi} \in R^2$ с плотностью распределения $p_\xi(x, y)$ и обозначить r, φ (случайные) полярные координаты этого вектора, то, так как в этом случае n = 1, мы получим из (7.1.1), что

$$p_{r,\varphi}(s, t) = sp_\xi(s\cos t, s\sin t)1_{(0,\infty)\times[0,2\pi)}(s, t). \quad (1.2.3.1)$$

Выражение (1.2.3.1) выглядит особенно просто, если существует такая функция $f : (0, \infty) \to (0, \infty)$, что$p_\xi(x, y) = f(\sqrt{x^2 + y^2})$. В этом случае, очевидно,

$$p_{r,\varphi}(s, t) = sf(s)I_{(0,\infty)\times[0,2\pi)}(s, t) = 2\pi sf(s)I_{(0,\infty)}(s)\frac{1}{2\pi}I_{[0,2\pi)}(t). \quad (1.2.3.2)$$

Это значит, что случайные величины r и φ независимы, $\varphi \in U(0, 2\pi)$, а r имеет плотность распределения $p_r(s) = 2\pi sf(s), s > 0$

Действительно, поскольку совместная плотность распределения ξ_1, ξ_2 имеет вид

$$p(x, y) = \frac{1}{2\pi}e^{-\frac{x^2+y^2}{2}}, x, y \in R,$$

то, как следовательно из вышенаписанного, полярный радиус r и полярный угол φ случайного вектора (ξ_1, ξ_2) независимы, причем полярный угол равномерно распределен на $[0, 2\pi)$, а полярный радиус имеет распределение Рэлея ($p(x) = xe^{\frac{-x^2}{2}}$, x > 0.).

Отсюда, применяя моделирующую формулу ($\xi = \sqrt{-2ln(\alpha)}$), сразу же приходим к представлению

$$\xi_1 = \sqrt{-2ln(\alpha_1)}\cos(2\pi\alpha_2),$$
$$\xi_2 = \sqrt{-2ln(\alpha_1)}\sin(2\pi\alpha_2),$$

где $\alpha_1, \alpha_2 \in$ U(0, 1)

В итоге получаем следующий алгоритм:

In [7]:

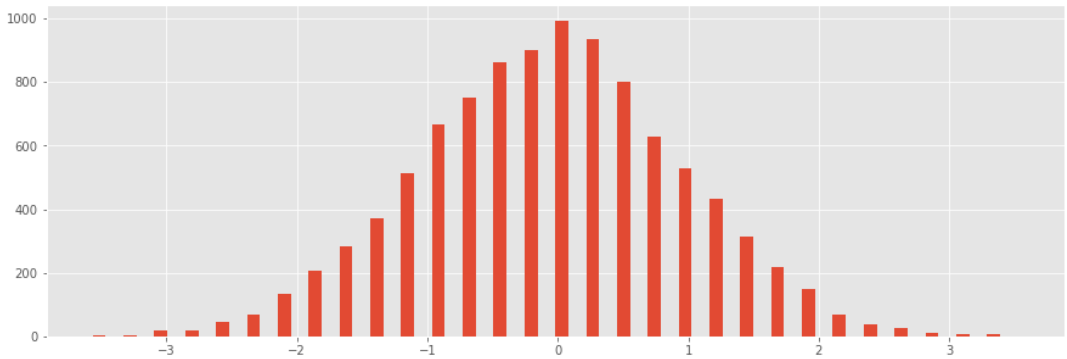
```
%matplotlib inline
import random
import pandas as pd
import math
from scipy import stats
import sys
plt.style.use('ggplot')  # Красивые графики
plt.rcParams['figure.figsize'] = (15, 5)  # Размер картинок
N=10000
random.seed(123)
epsilon = sys.float_info.epsilon

def box_muller():
    u1, u2 = 0.0, 0.0
    while u1 < epsilon or u2 < epsilon:
        u1 = random.random()
        u2 = random.random()

    n1 = math.sqrt(-2 * math.log(u1)) * math.cos(2 * math.pi * u2)
    n2 = math.sqrt(-2 * math.log(u1)) * math.sin(2 * math.pi * u2)
    return n1, n2

# Use KS to test
samples = [box_muller()[0] for x in range(N)]
test_stat, pvalue = stats.kstest(samples, 'norm', args=(0, 1), N=N)

# Plot our samples against our reference distribution
plt.hist(samples, 30, width = 0.1)
plt.show()
```



Стоит заметить, что данный способ достаточно быстр.
Я бы ещё подметил тот факт, что распределение Максвелла с параметром $\lambda = 1$ очень схоже с N(0,1):

$$f(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{x^2}{2}} x^2 - \text{Максвелла}$$

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

Способ 2: По определению, случайный вектор $\bar{\xi} = (\xi_1, \xi_2)^T$ равномерно распределен на единичной окружности S^1 с центром в нуле, если $\xi_1^2 + \xi_2^2 = 1$ с вероятностью 1 и если полярный угол φ вектора $\bar{\xi}$ равномерно распределен на $[0, 2\pi)$.

Из этого определения сразу же следует моделирующая формула для равномерного распределения на S^1:

$$\xi_1 = \cos(2\pi\alpha)$$
$$\xi_2 = \sin(2\pi\alpha)(1.2.3.3)$$

Вычисление двух тригонометрических функций, однако, может оказаться трудоемкой операцией. Стандартной альтернативой формуле (1.2.3.3) является использование метода отбора для моделирования равномерного распределения в единичном круге $B_1(0) = \{(x, y) : x^2 + y^2 < 1\}$ с центром в нуле с последующей нормировкой результата. Формальное обоснование этой процедуры представим ниже.

Аналогично полярным координатам на плоскости, каждый ненулевой вектор $X = (x, y, z)^T \in R^3$ может быть однозначно представлен в виде $X = ||X||\bar{e}$, где

$$\bar{e} = (\cos(t)\cos(u), \sin(t)\cos(u), \sin(u))^T, t \in [0, 2\pi), u \in [-\pi/2, \pi/2].$$

Это, конечно, соответствует переходу от евклидовой системы координат (x, y, z) к сферической системе (s, t, u) со сферическим радиусом $s = ||X||$, долготой s и широтой u. Хорошо известно, что якобиан обратного отображения равен $s^2\cos(u)$.

Поэтому, если случайный вектор $\bar{\xi} = (\xi_1, \xi_2, \xi_3)^T$ имеет плотность распределения $p_{\bar{\xi}}(x, y, z)$, то сферические координаты r, φ, θ этого вектора имеют совместную плотность

$$pr, \varphi, \theta(s, t, u) = p_{\bar{\xi}}(s\cos(t)\cos(u), s\sin(t)\cos(u), s\sin(u))s^2\cos(u), (1.2.3.4)$$

сосредоточенную в области $(0, \infty) \times [0, 2\pi) \times (-\pi/2, \pi/2)$. В случае, когда

$$p_{\bar{\xi}}(x, y, z) = f(\sqrt{x^2 + y^2 + z^2}), (1.2.3.5)$$

равенство (1.2.3.4) приобретает вид

$$p_{r,\varphi,\theta}(s, t, u) = 4\pi s^2 f(s^2) I_{(0,\infty)}(s) \frac{1}{2\pi} I_{(0,2\pi)} \frac{\cos(u)}{2} I_{(-\frac{\pi}{2}, \frac{\pi}{2})}.$$

Таким образом, случайные величины r, φ и θ оказываются независимыми, причем долгота φ равномерно распределена на $(0, 2\pi)$, плотность $p_r(s)$ распределения r равна $4\pi s^2 f(s^2)$, а плотность $p_\theta(u)$ распределения широты θ сосредоточена на $(-\frac{\pi}{2}, \frac{\pi}{2})$ и равна на этом интервале $0.5\cos(u)$.

Например, если $\bar{\xi} = (\xi^1, \xi^2, \xi^3)^T$ — случайный вектор с независимыми N(0, 1)-распределенными координатами, то его плотность распределения равна $(2\pi)^{-\frac{3}{2}} e^{-\frac{x^2+y^2+z^2}{2}}$, и длина $r = ||\bar{\xi}||$ этого вектора будет иметь плотность распределения $\sqrt{\frac{2}{\pi}} s^2 e^{-\frac{s^2}{2}}$.

Как уже обсуждалось при первом способе моделирования, если случайный вектор ξ имеет распределение (1.2.3.5), то вектор $\nu = \frac{\bar{\xi}}{||\bar{\xi}||}$ равномерно распределен на поверхности сферы $\{(x, y, z) : x^2 + y^2 + z^2 = 1\}$. С другой стороны, координаты ν_1, ν_2, ν_3 этого вектора выражаются через случайные величины φ и θ как

$$\nu_1 = \cos(\varphi)\cos(\theta), \nu_2 = \sin(\varphi)\cos(\theta), \nu_3 = \sin(\theta). (1.2.3.6)$$

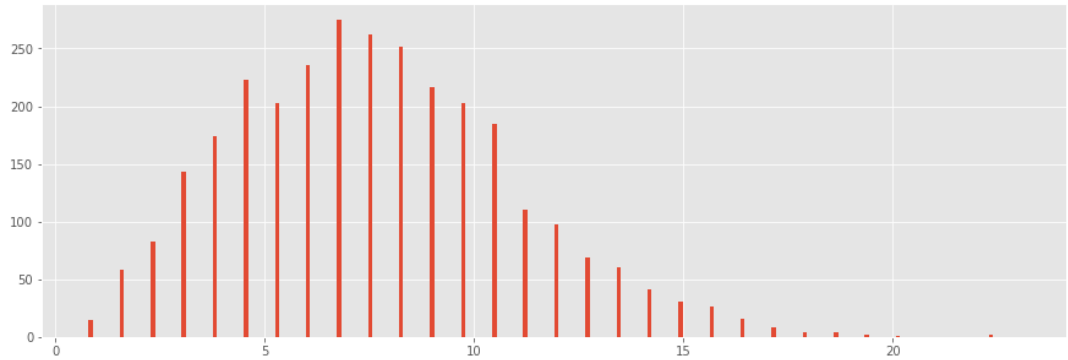
Предложение 1:

Пусть $\bar{\xi} \in R^d$ — d-мерный случайный вектор, обладающий распределением $P_{\bar{\xi}}$ с плотностью распределения $p_{\bar{\xi}}$, причем $P_{\bar{\xi}}(D) = 1$ для некоторого измеримого $D \subset R^d$. Рассмотрим измеримое отображение $\varphi : D \to R^d$ и предположим, что при $i = 1, \ldots, n$ существуют открытые попарно непересекающиеся подмножества $D_i \subset R^d$, удовлетворяющие следующим условиям.

1. Множества D и $\cup_{i=1}^n D_i$ совпадают $P_{\bar{\xi}}$ -почти всюду.
2. Обозначим $\phi_i = \phi|D_i, \phi(D_i) = \phi_i(D_i) = G_i$ и предположим, что при любом i отображение $\phi_i : D_i \to G_i$
   является биекцией с $\psi_i = \phi_i^{-1}$;
   непрерывно дифференцируемо с $det\psi_i' \neq 0 в G_i$.
   Тогда случайный вектор $\bar{\eta} = \psi(\bar{\xi})$ обладает плотностью $p_\eta(Y)$ и
   $$p_\eta(Y) = \sum_i p_{\bar{\xi}}\phi_i(Y) det\psi_i'(Y) I_{G_i}(Y).$$

–

In [8]:
```
def Maxwell(n,lambd = 1):
    x = [sample_(scale = lambd) for x in range(n)]
    y=[sample_(scale = lambd) for x in range(n)]
    z=[sample_(scale = lambd) for x in range(n)]
    l = []
    #print(x)
    for i in range(n):
        l.append(np.sqrt(x[i]**2+y[i]**2+z[i]**2))
    return l
# Our sample function of N(0,1) using Equation (2)
def sample_(N = 3000,scale = 1):
    return scale*2.0*np.sqrt(N)*(sum(randint(0,1)for x in range(N))/N-0.5)
plt.hist(Maxwell(3000,5),30, width = 0.1)
plt.show()
```



2 Домашнее задание. Основные понятия математической статистики

2.1 Геометрическое распределение

2.1.1 Моделирование выбранных случайных величин

In [66]:
```
# Создание случайной величины с геометрическим распределением, зависящим
# от параметра p
p = 0.5
geom_rv = sts.geom(p)
```

In [67]:
```
#Генерация выборки объема n = 5 с выводом
for n in [5]:
    means_5 = []
    for i in range(5):
        sample = geom_rv.rvs(n)
        means_5.append(sample)
        print(sample)
```

```
[2 2 1 3 3]
[2 2 1 2 2]
[3 1 4 1 1]
[1 2 1 1 1]
[5 2 1 1 3]
```

In [68]:
```
#Генерация выборки объема n = 10 с выводом
for n in [10]:
    means_10 = []
    for i in range(5):
        sample = geom_rv.rvs(n)
        means_10.append(sample)
        print(sample)
```

```
[1 1 2 1 4 2 1 1 2 1]
[2 1 3 1 3 1 1 1 3 1]
[1 2 2 2 2 1 4 1 1 1]
[2 2 2 2 1 1 1 2 1 1]
[4 1 1 2 1 1 1 2 3 2]
```

In [69]:
```
#Генерация выборки объема n = 100 ,без вывода
for n in [100]:
    means_100 = []
    for i in range(5):
        sample = geom_rv.rvs(n)
        means_100.append(sample)
```

In [70]:
```
#Генерация выборки объема n = 1000 ,без вывода
for n in [1000]:
    means_1000 = []
    for i in range(5):
        sample = geom_rv.rvs(n)
        means_1000.append(sample)
```

In [71]:
```
#Генерация выборки объема n = 100000 ,без вывода
for n in [100000]:
    means_100000 = []
    for i in range(5):
        sample = geom_rv.rvs(n)
        means_100000.append(sample)
```

2.1.2 Построение эмпирической функции распределения

In [31]:

```
#n=5
for a in range(5):
    b=means_5[a]
    b=sorted(b)
    print('Empirical distribution function F5(x) for sample',a+1,':')
    for i in range(4):
        if(i==0):
            n=0.
            g=1
            print(n,', x <=',b[i])
        if(b[i+1]==b[i]):
            g+=1
        else:
            n=round(n+0.2*g,1)
            g=1
            print(n,',',b[i],'< x <=',b[i+1])
        if(i==3):
            n=1.
            print(n,', x >',b[i+1])
```

```
Empirical distribution function F5(x) for sample 1 :
0.0 , x <= 1
0.4 , 1 < x <= 2
0.6 , 2 < x <= 3
0.8 , 3 < x <= 5
1.0 , x > 5
Empirical distribution function F5(x) for sample 2 :
0.0 , x <= 1
0.6 , 1 < x <= 2
0.8 , 2 < x <= 6
1.0 , x > 6
Empirical distribution function F5(x) for sample 3 :
0.0 , x <= 1
0.2 , 1 < x <= 2
0.6 , 2 < x <= 4
0.8 , 4 < x <= 5
1.0 , x > 5
Empirical distribution function F5(x) for sample 4 :
0.0 , x <= 1
0.8 , 1 < x <= 3
1.0 , x > 3
Empirical distribution function F5(x) for sample 5 :
0.0 , x <= 1
0.2 , 1 < x <= 2
0.4 , 2 < x <= 3
0.8 , 3 < x <= 4
1.0 , x > 4
```

In [32]:

```
#n=5
for a in range(5):
    b=means_5[a]
    b=sorted(b)
    v=len(b)
    N=[]
    for i in range(b[v-1]):
        N.append(b.count(i))
        x1=[]
        y1=[]
        t=0
    for i in range(b[v-1]):
        t+=N[i]
        x1.append(i)
        y1.append(t/v)
        x1.append(i+1)
        y1.append(t/v)
    x1.append(b[v-1])
    y1.append(1)
    x1.append(b[v-1]+2)
    y1.append(1)
    plt.plot(x1,y1,label="ECDF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("Эмпирическая функция выборки объема: "+str(v))
n=np.arange(1,8,1)#Построение
plt.step(n,1-(1-p)**(n-1),'k-', label='CDF')#теоретическойфункции
plt.legend()#распределения
plt.xlabel("numbers")
plt.ylabel("probability")
plt.show()
print("\n")
#n=10
for a in range(5):
    b=means_10[a]
    b=sorted(b)
    v=len(b)
    N=[]
    for i in range(b[v-1]):
        N.append(b.count(i))
        x2=[]
        y2=[]
        t=0
    for i in range(b[v-1]):
        t+=N[i]
        x2.append(i)
        y2.append(t/v)
        x2.append(i+1)
        y2.append(t/v)
    x2.append(b[v-1])
    y2.append(1)
    x2.append(b[v-1]+2)
    y2.append(1)
    plt.plot(x2,y2,label="ECDF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("Эмпирическая функция выборки объема: "+str(v))
n=np.arange(1,8,1)#Построение
plt.step(n,1-(1-p)**(n-1),'k-', label='CDF')#теоретическойфункции
plt.legend()#распределения
plt.xlabel("numbers")
plt.ylabel("probability")
plt.show()
print("\n")
#n=100
for a in range(5):
    b=means_100[a]
    b=sorted(b)
    v=len(b)
    N=[]
```

Пусть $X = (X_1, \ldots, X_n)$- выборка из дискретного распределения $\sigma(\xi)$ Величина скачка в точке j есть

$$\Delta\hat{F}_n(j) = \hat{F}_n(j) - \hat{F}_n(j-0) = \frac{\nu}{n},$$

$j = 1, \ldots, N$
Здесь $P\{\Delta\dot{F}_n(j) = 0\} = P(\nu_j = 0) = (1 - p_j)^n$ что мало при больших n, т.е. в большой выборке скачок в точке j наверняка будет иметь место. Более того, так как
$P\{\cup_{j=1}^{N}\{\Delta\hat{F}_n(j) = 0\}\} \leq \sum_{j=1}^{N}(1 - p_j)^n \to 0$, при $n \to \infty$, то в больших выборках с вероятностью, близкой к 1, скачки э.ф.р. $F_n(x)$ будут иметь место во всех точках 1,2,...,N, а случайными будут лишь величины этих скачков.
Если же теоретическая функция распределения $F_\xi = F(x)$ непрерывна, то с вероятностью 1 все элементы выборки $X = (X_1, \ldots, X_n)$ будут различны, и случайными теперь будут точки скачков, величины же скачков неслучайны и равны $\frac{1}{n}$
Таким образом, для выборок из дискретных и непрерывных распределений характер соответствующих эмпирических функций распределениябудет различным, что можно заметить на получившихся графиков для дискретного и непрерывного распределения. Тем не менее в любомслучае э.ф.р. $\hat{F}_n(x)$ с увеличением объема выборки n сближается в каждой точке x с теоретической функцией распределения F(x).
Максимальная точная верхняя граница разности пары эмпирических функций распределения - наибольшая разность между значениями вероятности двух функций в одной точке.

Верхняя граница разности э.ф.р. выборок размера n = 5 : 0.600
Верхняя граница разности э.ф.р. выборок размера n = 10 : 0.400
Верхняя граница разности э.ф.р. выборок размера n = 100 : 0.17
Верхняя граница разности э.ф.р. выборок размера n = 1000 : 0.061
Верхняя граница разности э.ф.р. выборок размера n = 100000 : 0.0049
С увеличением объема выборки верхняя граница разности уменьшается, что, впрочем, очевидно.

2.1.3 Построение вариационного ряда выборки

Определение:
Пусть $X = (X_1, \ldots, X_n)$- выборка из некоторого распределения $\sigma(\xi)$
Произвольной реализации $x = (x_1, \ldots, x_n)$ этой выборки можно поставить в соответствие упорядоченную последовательность

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

располагая $x_1, \ldots, x_n$ в порядке их возрастания, так что $x_{(1)} = min\{x_1, \ldots, x_n\}$, $x_{(2)}$ - второе по величине значение, $x_{(n)} = max\{x_1, \ldots, x_n\}$
Обозначим через $X_{(k)}$ случайную величину, которая для каждой реализации выборки X принимает значение $x_{(k)}, k = 1, \ldots, n$. Так по выборке X определяют новую последовательность случайных величин $X_{(1)}, \ldots, X_{(n)}$, называемых *порядковыми статистиками* выборки. Из определения порядковых статистик следует, что они упорядочены по возрастанию их значений, т.е. они образуют возрастающую последовательность

$$X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)},$$

которая называется *вариационным рядом* выборки X.

In [33]:
```
#Вариационный ряд для выборки объема n=5 с выводом
for a in range(5):
    b=means_5[a]
    b=sorted(b)
    print(b)
```

```
[1, 1, 2, 3, 5]
[1, 1, 1, 2, 6]
[1, 2, 2, 4, 5]
[1, 1, 1, 1, 3]
[1, 2, 3, 3, 4]
```

In [34]:
```
#Вариационный ряд для выборки объема n=10 с выводом
for a in range(5):
    b=means_10[a]
    b=sorted(b)
    print(b)
```

```
[1, 1, 1, 1, 1, 2, 2, 3, 5, 5]
[1, 1, 1, 2, 3, 3, 3, 3, 3, 4]
[1, 1, 1, 2, 2, 2, 2, 2, 3, 4]
[1, 1, 1, 1, 1, 2, 2, 2, 3, 3]
[1, 1, 2, 2, 2, 2, 3, 3, 5, 6]
```

In [35]:
```
#Вариационный ряд для выборки объема n=100 без вывода
for a in range(5):
    b=means_100[a]
    b=sorted(b)
```

In [36]:
```
#Вариационный ряд для выборки объема n=1000 без вывода
for a in range(5):
    b=means_1000[a]
    b=sorted(b)
```

In [37]:
```
#Вариационный ряд для выборки объема n=100000 без вывода
for a in range(5):
    b=means_100000[a]
    b=sorted(b)
```

Определение:
α- квантиль случайной величины ξ с функцией распределения $F(x) = P\{\xi < x\}$ — это любое число x_α, удовлетворяющее двум условиям:
$1) F(x_\alpha) \leq \alpha \; 2) F(x_\alpha + 0) \geq \alpha.$
Исходя из того, что при больших выборках э.ф.р. стремится к теоритической функции распределения, эмпирические квантили так же стремятся к теоритическим по определению.
Пусть F(x) - функция распределения. Тогда квантильная функция:
$F^{-1}(r) = min\{x \in N_+ : F(x) \geq r\} for r \in (0; 1)$
$F^{-1}(r) = [\frac{ln(1-r)}{ln(1-p}]$

In [38]:
```
k = 1
for b in [means_5[a], means_10[a],means_100[a],means_1000[a],means_100000[a
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b, 0.1))
    k += 1
```

```
n = 5
1.4
n = 10
1.0
n = 100
1.0
n = 1000
1.0
n = 100000
1.0
```

In [39]:
```
#Сравнение
np.log(1-0.1)//np.log(1-p)
```

Out[39]:
```
0.0
```

In [40]:
```
k = 1
for b in [means_5[a], means_10[a],means_100[a],means_1000[a],means_100000[a
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b, 0.5))
    k += 1
```

```
n = 5
3.0
n = 10
2.0
n = 100
1.0
n = 1000
2.0
n = 100000
1.0
```

In [41]:
```
#Сравнение
geom.median(p)
```

Out[41]:
```
1.0
```

In [42]:
```
#Сравнение
np.log(1-0.5)//np.log(1-p)
```

Out[42]:
```
1.0
```

In [43]:
```
k = 1
for b in [means_5[a], means_10[a],means_100[a],means_1000[a],means_100000[a
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b, 0.7))
    k += 1
```

```
n = 5
3.0
n = 10
3.0
n = 100
2.0
n = 1000
2.0
n = 100000
2.0
```

In [44]:
```
#Сравнение
np.log(1-0.7)//np.log(1-p)
```

Out[44]: 1.0

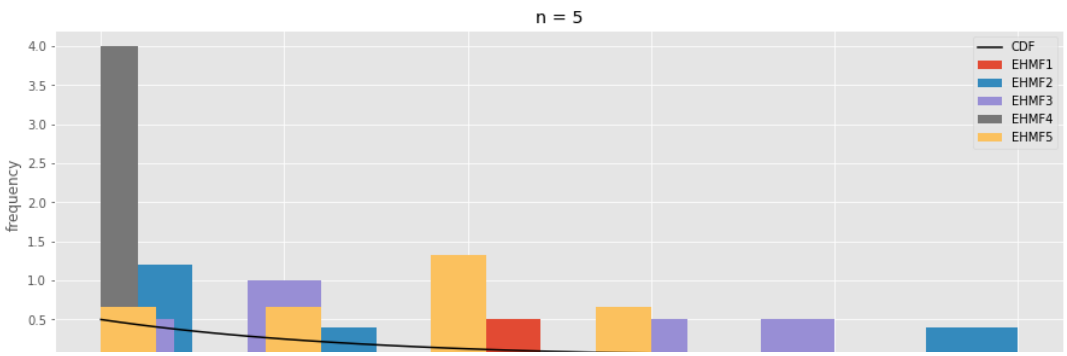
2.1.4 Построение гистограммы и полигона частот

In [47]:

```
#n=5
for a in range(5):
    b=means_5[a]
    b=sorted(b)
    x=[]
    y=[]
    c=Counter(b)
    for i in c:
        x.append(i)
        y.append(b.count(i)/5.0)
    plt.plot(x,y,label="EPMF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("n = 5")
n=np.arange(1,2,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=10
for a in range(5):
    b=means_10[a]
    b=sorted(b)
    x=[]
    y=[]
    c=Counter(b)
    for i in c:
        x.append(i)
        y.append(b.count(i)/10.0)
    plt.plot(x,y,label="EPMF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("n = 10")
n=np.arange(1,6,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=100
for a in range(5):
    b=means_100[a]
    b=sorted(b)
    x=[]
    y=[]
    c=Counter(b)
    for i in c:
        x.append(i)
        y.append(b.count(i)/100.0)
    plt.plot(x,y,label="EPMF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("n = 100")
n=np.arange(1,9,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=1000
for a in range(5):
    b=means_1000[a]
    b=sorted(b)
    x=[]
    y=[]
    c=Counter(b)
    for i in c:
        x.append(i)
        y.append(b.count(i)/1000.0)
    plt.plot(x,y,label="EPMF "+str(a+1))
```

In [48]:

```
#n=5
for a in range(5):
    plt.hist(means_5[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
n=np.arange(1,5,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.title("n = 5")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=10
for a in range(5):
    plt.hist(means_10[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
n=np.arange(1,6,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.title("n = 10")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=100
for a in range(5):
    plt.hist(means_100[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
n=np.arange(1,9,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.title("n = 100")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=1000
for a in range(5):
    plt.hist(means_1000[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
n=np.arange(1,15,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.title("n = 1000")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=100000
for a in range(5):
    plt.hist(means_100000[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
n=np.arange(1,18,0.1)#Построение
plt.plot(n,p*(1-p)**(n-1),'k-',label='CDF')#функции вероятности
plt.legend()#распределения
plt.title("n = 100000")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
```



Если наблюдаемая в эксперименте случайная величина ξ дискретна и принимает значения $a_1, a_2, ..$, то более наглядное представление о ее законе распределения дадут относительные частоты $\nu_r^* = \frac{\nu_r}{n}$, где ν_r - число элементов выборки $X = (X_1, \ldots, X_n)$, принявших значение $a_r : \nu_r = \sum_{j=1}^{n} I(X_j = a_r), r = 1, 2, ..$, т.е. ν_r^* сближается с ростом n с теоретической вероятностью $P\{\xi = a_r\}$, и потому, по крайней мере для больших выборок, относительные частоты ν_r^* можно рассматривать в качестве приближенных значений (оценок) для неизвестных вероятностей $P\{\xi = a_r\}$.
Наглядным представлением данных является полигон частот, который представляет собой ломаную с вершинами в точках $(a_r; \nu_r), r = 1, 2, \ldots$
Можно рассматривать также статистический ряд $\{\{(a_r; \nu_r)\}\}$.
На графиках выше наглядно подтверждаются наши теоретические знания.

2.2 Распределение Максвелла

2.2.1 Моделирование выбранных случайных величин

In [42]:
```
# Создание случайной величины с распределением Максвелла, зависящим
# от параметра lambda
lambd=1.0
maxwell_rv=sts.maxwell(scale=lambd)
```

In [43]:
```
#Генерация выборки объема n = 5 с выводом
for n in[5]:
    means__5=[]
    for i in range(5):
        sample=maxwell_rv.rvs(n)
        means__5.append(sample)
        print(sample)
```

```
[2.61492501 1.8185356  1.60239099 1.88471966 2.32343524]
[1.37090364 0.81484833 2.95500439 1.00735304 0.77608477]
[1.93052179 0.98916486 0.83490276 1.98262    2.28278868]
[2.21407855 2.51775003 2.5953651  1.29263856 2.0909891 ]
[3.23971372 1.82680786 2.17985755 0.31893926 2.1808736 ]
```

In [44]:
```
#Генерация выборки объема n = 10 с выводом
for n in[10]:
    means__10=[]
    for i in range(5):
        sample=maxwell_rv.rvs(n)
        means__10.append(sample)
        print(sample)
```

```
[2.12102194 2.21942606 2.75273494 0.82323637 1.82087582 1.86830689
 1.07562858 0.47693664 2.17478789 1.433096  ]
[1.66711706 2.12778881 1.59876399 2.19313351 1.76273771 1.05264685
 2.74675272 1.55595092 1.03700015 0.59029844]
[3.33341564 4.23787351 1.52515002 1.63675074 0.71231781 2.58619079
 1.33683265 1.02519597 1.97457463 2.09229573]
[1.27631799 2.25908834 1.69106005 2.57793336 0.90366426 1.35694638
 1.87914425 0.63669354 1.09633501 1.0690649 ]
[2.14434436 2.28425493 0.67972466 2.71354297 1.8494517  1.18174504
 1.4799613  0.52483374 1.61124112 1.55908243]
```

In [45]:
```
#Генерация выборки объема n = 100 без вывода
for n in[100]:
    means__100=[]
    for i in range(5):
        sample=maxwell_rv.rvs(n)
        means__100.append(sample)
```

In [46]:
```
#Генерация выборки объема n = 1000 без вывода
for n in[1000]:
    means__1000=[]
    for i in range(5):
        sample=maxwell_rv.rvs(n)
        means__1000.append(sample)
```

In [47]:
```
#Генерация выборки объема n = 100000 без вывода
for n in[100000]:
    means__100000=[]
    for i in range(5):
        sample=maxwell_rv.rvs(n)
        means__100000.append(sample)
```

In [55]:
```
#Вернёмся к медиане и убедимся, что в пункте 1.2.1 она была найдена верно
maxwell.median()
```

Out[55]: 1.5381722544550522

2.2.2 Построение эмпирической функции распределения

In [56]:

```
#n=5
for a in range(5):
    b=means__5[a]
    b=sorted(b)
    v=len(b)
    N=[]
    for i in range((v-1)):
        N.append(b.count(b[i]))
        x=[]
        y=[]
        t=0
        x.append(0.0)
        y.append(0.0)
        x.append(b[0])
        y.append(0.0)
    for i in range((v-1)):
        t+=N[i]
        x.append(b[i])
        y.append(float(t/v))
        x.append(b[i+1])
        y.append(float(t/v))
    x.append(b[v-1])
    y.append(1)
    x.append(b[v-1]+2)
    y.append(1)
    plt.plot(x,y,label="ECDF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("Эмпирическая функция выборки объема: "+str(v))
x=np.linspace(0,5,100)#Построение
cdf=maxwell_rv.cdf(x)#теоретической функции
plt.plot(x,cdf,label='CDF')#распределения
plt.legend()
plt.xlabel("numbers")
plt.ylabel("probability")
plt.show()
print("\n")
#n=10
for a in range(5):
    b = means__10[a]
    b = sorted(b)
    v = len(b)
    N = []
    for i in range((v-1)):
        N.append(b.count(b[i]))
        x=[]
        y=[]
        t=0
        x.append(0.0)
        y.append(0.0)
        x.append(b[0])
        y.append(0.0)
    for i in range((v-1)):
        t+=N[i]
        x.append(b[i])
        y.append(float(t/v))
        x.append(b[i+1])
        y.append(float(t/v))
    x.append(b[v-1])
    y.append(1)
    x.append(b[v-1]+2)
    y.append(1)
    plt.plot(x,y,label="ECDF "+str(a+1))
    plt.legend(loc='lower right')
plt.title("Эмпирическая функция выборки объема: "+str(v))
x=np.linspace(0,5,100)#Построение
cdf=maxwell_rv.cdf(x)#теоретической функции
plt.plot(x,cdf,label='CDF')#распределения
plt.legend()
```

2.2.3 Построение вариационного ряда выборки

In [39]:
```
#Вариационный ряд для выборки объема n=5 с выводом
for a in range(5):
    b=means__5[a]
    b=sorted(b)
    print(b)
```

```
[0.9048224033555651, 1.0595749548872997, 1.3361146201247887, 1.768477821152408
1, 2.064309990285417]
[1.142775449379784, 1.6322198084743014, 1.9269294058020057, 1.930731227695640
5, 2.137402279174372]
[1.168850546533639, 1.2873609279696019, 1.4055755577674152, 1.554026224592775,
2.9493391029304523]
[0.981057908261026, 1.200322004547301, 1.8319652264631978, 1.9311087818516284,
2.2702664778364174]
[1.0795029380722423, 1.103674980358039, 1.3308824708067275, 2.668729964272347
7, 2.9675918656118716]
```

In [40]:
```
#Вариационный ряд для выборки объема n=10 с выводом
for a in range(5):
    b=means__10[a]
    b=sorted(b)
    print(b)
```

```
[0.4344229664385466, 0.9873280481429386, 1.1514899800296186, 1.162078879947299
2, 1.5047110956567034, 1.6532896759587332, 1.9249137266578096, 2.0658538511219
224, 2.1461320252051195, 2.324740490202557]
[1.1220497289203735, 1.2401311800502641, 1.2435276204947807, 1.57656133708086
5, 1.6375956762713444, 1.7697569758530114, 1.8386307894229095, 1.9485280422518
474, 1.950915639549961, 1.9767833922678453]
[0.8963933247210342, 1.4705601230263092, 1.4945015272392095, 1.497442822003277
2, 1.5122209047811999, 1.6248884551477416, 1.7155281056459615, 1.9907561312818
842, 2.3686125510412857, 2.4716462283848437]
[0.39200688435666003, 0.5966409826346033, 0.9698060007821667, 1.02850758607638
97, 1.0404367899757856, 1.486851983947381, 2.019114126722199, 2.16641776220787
64, 2.641475897208004, 2.84064283839896]
[1.148593485170174, 1.1736791688497565, 1.5230401251658805, 1.553329580159769,
1.5775412015483297, 1.781091079341479, 1.941831168996058, 2.0709499700182676,
2.143143746322394, 2.9585921488894913]
```

In [41]:
```
#Вариационный ряд для выборки объема n=100 без вывода
for a in range(5):
    b=means__100[a]
    b=sorted(b)
```

In [60]:
```
#Вариационный ряд для выборки объема n=1000 без вывода
for a in range(5):
    b=means__1000[a]
    b=sorted(b)
```

In [61]:
```
#Вариационный ряд для выборки объема n=100000 без вывода
for a in range(5):
    b=means__100000[a]
    b=sorted(b)
```

Возникли сложности при вычислении теоретических значений квантилей, однако был найден справочник:"Справочник по вероятностным распределениям" Р.Н.Вадзинский. В нём была найдена таблица для приближенного решения уравнения $x_\alpha = \lambda m_\alpha$, где $x_\alpha = \lambda m_\alpha$ - квантиль порядка α распределения Максвелла

In [62]:
```
k=1
for b in [means__5[a],means__10[a],means__100[a],means__1000[a],means__1000
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b,0.1))
    k+=1
```

```
n = 5
1.2719859381522172
n = 10
1.1252696205430428
n = 100
0.7938084172369115
n = 1000
0.7374170981472817
n = 100000
0.7603097895100107
```

Сравнение со значением (с теоретическим) из таблицы:

$\alpha \approx 0.76$

In [63]:
```
k=1
for b in [means__5[a],means__10[a],means__100[a],means__1000[a],means__1000
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b,0.5))
    k+=1
```

```
n = 5
1.5059353635900166
n = 10
1.4534420207082235
n = 100
1.574535047585379
n = 1000
1.4765126474373647
n = 100000
1.537607751024725
```

In [64]:
```
#Сравнение
maxwell.median()
```

Out[64]:
```
1.5381722544550522
```

In [65]:
```
k=1
for b in [means__5[a],means__10[a],means__100[a],means__1000[a],means__1000
    if(k==1):
        print('n = 5')
    if(k==2):
        print('n = 10')
    if(k==3):
        print('n = 100')
    if(k==4):
        print('n = 1000')
    if(k==5):
        print('n = 100000')
    print(np.quantile(b,0.7))
    k+=1
```

```
n = 5
1.649424804466466
n = 10
1.7200020124901898
n = 100
1.9686707573406326
n = 1000
1.8574082725897791
n = 100000
1.9163671862147416
```

Сравнение со значением (с теоретическим) из таблицы:
$\alpha \approx 1.92$

С увеличением объема выборки э.ф.р. стремится к теоритической функции распределения, следовательно, эмпирические квантили так жестремятся к теоритическим по определению. Что и можно наблюдать выше.

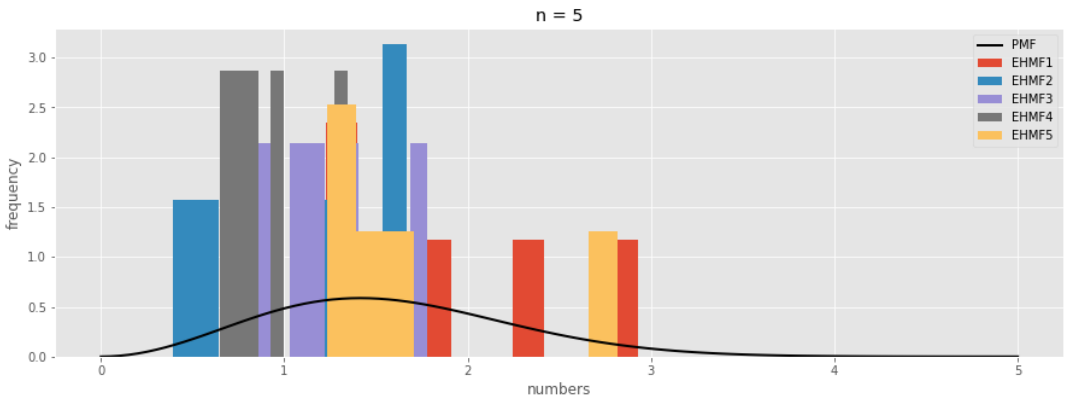
2.2.4 Построение гистограммы и полигона частот

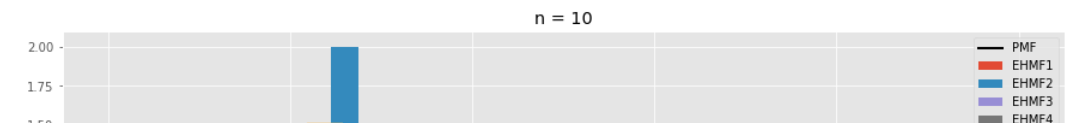
In [66]:

```
#n=5
for a in range(5):
    b=means__5[a]
    mas=list(range(1,6))
    p=[0,0,0,0,0]
    for i in range(5):
        mas[i]=b[i]
        if mas[i]>0 and mas[i]<1:
            p[0]=p[0]+1
        if mas[i]>1 and mas[i]<2:
            p[1]=p[1]+1
        if mas[i]>2 and mas[i]<3:
            p[2]=p[2]+1
        if mas[i]>3 and mas[i]<4:
            p[3]=p[3]+1
        if mas[i]>4 and mas[i]<5:
            p[4]=p[4]+1
    print()
    dob=[]
    bod=[]
    keks=0.5
    for i in range(5):
        dob.append(keks)
        bod.append(p[i]/5.0)
        keks+=1
    plt.plot(dob,bod,label='EPMF'+str(a+1))
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 5")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=10
for a in range(5):
    b=means__10[a]
    mas=list(range(1,11))
    p=[0,0,0,0,0]
    for i in range(10):
        mas[i]=b[i]
        if mas[i]>0 and mas[i]<1:
            p[0]=p[0]+1
        if mas[i]>1 and mas[i]<2:
            p[1]=p[1]+1
        if mas[i]>2 and mas[i]<3:
            p[2]=p[2]+1
        if mas[i]>3 and mas[i]<4:
            p[3]=p[3]+1
        if mas[i]>4 and mas[i]<5:
            p[4]=p[4]+1
    print()
    dob=[]
    bod=[]
    keks=0.5
    for i in range(5):
        dob.append(keks)
        bod.append(p[i]/10.0)
        keks+=1
    plt.plot(dob,bod,label='EPMF'+str(a+1))
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 10")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
```

In [67]:

```
#n=5
for a in range(5):
    plt.hist(means__5[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 5")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=10
for a in range(5):
    plt.hist(means__10[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 10")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=100
for a in range(5):
    plt.hist(means__100[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 100")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
#n=1000
for a in range(5):
    plt.hist(means__1000[a],density=True,label='EHMF{}'.format(a+1))
    plt.legend()
rv=maxwell()
x=np.linspace(0,5,100)
plt.plot(x,rv.pdf(x),'k-',lw=2,label='PMF')
plt.legend()
plt.title("n = 1000")
plt.xlabel("numbers")
plt.ylabel("frequency")
plt.show()
```





Для непрерывной случайной величины ξ, обладающей непрерывной плотностью f(x), также можно построить по соответствующей выборке $X = (X_1, \ldots, X_n)$ статистический аналог $\hat{f}_n(x)$ для плотности f(x), который называется гистограммой. Для этого используется методгруппировки, в соответствии с которым область Δ возможных значений ξ разбивается на некоторое число N непересекающихся интервалов $\Delta_1, \ldots, \Delta_N$ (так что $\Delta = \bigcup_{r=1}^{N} \Delta_r$, подсчитывают числа $\nu_1, \ldots, \nu_N$ наблюдений $X_1, \ldots, X_n$, попавших в соответствующие интервалы:
$\nu_r = \sum_{j=1}^{N} I(X_j \in \Delta_r), r = 1, \ldots, N$ (так что $\sum_{r=1}^{N} \nu_r = n$,
и строят кусочно-постоянную функцию

$$\hat{f}_n(x) = \frac{\nu_r}{n|\Delta_r|}$$

при x $\in \Delta_r, r = 1, \ldots, N$

Здесь $|\Delta_r|$ - длина интервала Δ_r. То, что построенная по такому правилу гистограмма $\hat{f}_n(x)$ действительно "похожа" на теоретическую плотность f(x), следует из закона больших чисел, согласно которому при $n \to \infty$ относительная частота $\frac{\nu_r}{n}$ сближается с теоретической вероятностью

$$P\{\xi \in \Delta_r\} = \int_{\Delta_r} f(x)dx$$

Но этот интеграл по теореме о среднем равен $f(a_r)|\Delta_r|$ где a_r - некоторая внутренняя точка интервала Δ_r(при малом Δ_r в качестве a_r можно взять, например, середину интервала), Таким образом, при больших n и достаточно "мелком" разбиении $\{\Delta_r\}\hat{f}_n(x) \approx f(a_r)$ при $x \in \Delta_r$ т.е. гистограмма $\hat{f}_n(x)$ будет достаточно хорошо приближать график плотности f(x), следовательно, $\hat{f}_n(x)$ можно рассматривать в качестве статистического аналога (оценки) для f(x).Наряду с гистограммой, в качестве приближения для неизвестной теоретической плотности f(x) можно использовать кусочно-линейный график называемый полигоном частот. Он также считается статистическим аналогом теоретической плотности. Данные на полигоне частот и гистограммах подтверждают теоретические знания: с увеличением объема выборки полигон частот и гистограммы практически совпадают с теоретической плотностью f(x).

3 Домашнее задание. Оценки

3.1. Геометрическое распределение

3.1.1 Нахождение выборочного среднего и выборочной дисперсии геометрического распределения

Наиболее важными характеристиками случайной величины ξ являются ее моменты $\alpha_k = M\xi_k$, а также цетральные моменты $\mu^k = M(\xi - \alpha 1)^k$ (когда они существуют). Их статистическими аналогами, вычисляемыми по соответсвующей выборке $X = (X_1, \ldots, X_n)$, являются выборочные моменты соотаетсвенно обычные:

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

и центральные:

$$\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\alpha}_1)^k$$

$\hat{\alpha}_1$ (принято обозначать, как $\overline{X}$) называют выборочным средним, μ^2 - выборочной дисперсией. Таким образом, выборочное среднее и выборочная дисперсия являются статистическими аналогами теоритических среднего (математического ожидания) $M\xi$ и дисперсии $D\xi$, когда они существуют.

Выборочное среднее, относящийся к выборке X, считается как:

$$\overline{X} = \hat{\alpha}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Выборочная дисперсия, относящийся к выборке X, подсчитывается как:

$$S^2 = \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Найдем математическое ожидание и дисперсию выборочного среднего и выборочной дисперсии:

$$M\overline{X} = \frac{1}{n}\sum_{i=1}^{n} MX_i = M\xi = \alpha_1$$

$$D\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n}D\xi = \frac{\mu_2}{n}$$

Для выборочной дисперсии введем обозначение: $Y_i = X_i - \alpha_1$:

$$S^2 = \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \overline{Y}^2$$

Поскольку $MY_i = 0, MY_i^2 = \mu_2$ и $MY_iY_j = MY_jY_i = 0, (i \neq j)$, то:

$$M\overline{Y}^2 = \frac{1}{n^2}\sum_{i,j=1}^{n} MY_iY_j = \frac{1}{n^2}\sum_{i=1}^{n} MY_i^2 = \frac{\mu_2}{n}$$

Отсюда следует, что

$$MS^2 = \frac{n-1}{n}\mu_2$$

Перейдём к вычислению DS^2

$$(S^2)^2 = \frac{1}{n^2}(\sum_{i=1}^{n} Y_i^2)^2 - \frac{2}{n}\overline{Y}^2\sum_{i=1}^{n} Y_i^2 + \overline{Y}^4$$

Так как случайные величины $Y_1, \ldots, Y_n$ независимы и $MY_i = 0$, то в правой части равенства

$$M\overline{Y}^4 = \frac{1}{n^4}(n\mu_4 + 3n(n-1)\mu_2^2) = \frac{\mu_4 + 3(n-1)\mu_2^2}{n^3}$$

Аналогично находим

$$\frac{1}{n^2}M(\sum_{i=1}^{n} Y_i^2) = \frac{\mu_4 + (n-1)\mu_2^2}{n} = M(\overline{Y}^2\sum_{i=1}^{n} Y_i^2)$$

С учётом этих соотношений по формуле

$$DS^2 = M(S^2)^2 - (MS^2)^2$$

получим

$$DS^2 = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} = \frac{(n-1)^2}{n^3}(\mu_4 - \frac{n-3}{n-1}\mu_2^2)$$

Аналогично можно находить моменты и более высоких порядков, хотя с увеличением порядка вид формул и их вывод усложняются.

Теперь рассмотрим свойства выборочных среднего и дисперсии при неограниченном возрастании объема выборки n, которые дадут нам ответ на вопрос, оценками каких параметров рапределений они являются. Чтобы подчеркнуть зависимость моментов $\hat{\alpha}_k, \hat{\mu}_k$ от объема выборки, будем в дальнейшем приписывать дополнительный индекс n: $\hat{\alpha}_{nk}, \hat{\mu}_{nk}$

$$M\hat{\alpha}_{nk} = \frac{1}{n}\sum_{i=1}^{n} MX_i^k = M\xi^k = \hat{\alpha}_k$$

$$D\hat{\alpha}_{nk} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i^k = \frac{1}{n}D\xi^k = \frac{1}{n}(M\xi^{2k} - (M\xi^k)^2) = \frac{\alpha_{2k} - \alpha_k^2}{n}$$

На основании неравенства Чебышева, отсюда следует, что для любого $\epsilon > 0$ при $n \to \infty$

$$P|\hat{\alpha}_{nk} - \alpha_k| < \epsilon \to 1$$

т.е. выборочный момент $\hat{\alpha}_{nk}$ сходится по вероятности при $n \to \infty$ к соответствующему теоретическому моменту α_k . Таким образом, $\hat{\alpha}_{nk}$ можно использовать в качестве оценки α_k, когда объем выборки достаточно велик. Аналогичное утверждение справедливо и для центральных моментов:

$$P|\mu_{nk} - \mu_k| < \epsilon \to 1$$

т.е. μ_{nk} можно использовать в качестве оценки μ_k , когда объем выборки достаточно велик.

1. Оценка $\hat{\theta}(X)$ параметра θ называется несмещенной, если:

$$E(\hat{\theta}(X)) = \theta$$

2. Оценка $\hat{\theta}(X) = \hat{\theta}_n(X_1, \ldots, X_n)$ параметра θ называется состоятельной, если при $n \to \infty$ соблюдается:

$$\hat{\theta}_n(X_1, \ldots, X_n) \xrightarrow{p} \theta$$

При этом для проверки состоятельности достаточно убедиться, что соблюдены следующие два условия:

$$\lim_{n\to\infty} E(\hat{\theta}_n(X_1, \ldots, X_n)) = \theta$$

$$\lim_{n\to\infty} Var(\hat{\theta}_n(X_1, \ldots, X_n)) = 0$$

Выборочное среднее является несмещенной оценкой для теоретического математического ожидания.

$$\hat{\alpha_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$M\hat{\alpha_1} = M(\frac{1}{n}\sum_{i=1}^{n} x_i) = \frac{1}{n} \cdot n \cdot MX = MX$$

$$M\hat{\alpha_1} = MX$$

Выборочное среднее является состоятельной оценкой для теоретического математического ожидания.

$$\hat{\alpha_1} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\lim_{n\to\infty} \hat{\alpha_1} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i = \lim_{n\to\infty} \frac{1}{n} \cdot n \cdot MX = MX$$

Выборочная дисперсия S^2 является состоятельной и несмещенной оценкой для теоретической дисперсии.

$$S^2 = \hat{\mu_2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{n}\sum_{i=1}^{n}(\hat{\alpha_1})^2 = MX^2 - \frac{1}{n}n(\hat{\alpha_1})^2 = \hat{\alpha_2} - (\hat{\alpha_1})^2$$

$$MS^2 = M(\hat{\alpha_2} - (\hat{\alpha_1})^2) = M(\frac{1}{n}\sum_{i=1}^{n} X_i^2) - M((\frac{1}{n}\sum_{i=1}^{n} X_i))^2 = MX^2 - (MX)^2 = DX$$

Продолжим исследование свойств выборочных моментов для больших выборок и рассмотрим теперь асимптотическое поведение их выборочных распределений.
Если распределение случайной величины ν_n сходится при $n \to \infty$ к распределению случайной величины ν и при этом $\zeta(\nu) = N(\mu, \sigma^2)$, то будем писать $\zeta(\nu_n) \to N(\mu, \sigma^2)$. Далее иногда будем говорить, что случайная величина ν_n асимптотически нормальна $N(\mu_n, \sigma_n^2)$, и записывать это следующим образом:
$\zeta(\nu_n) \approx N(\mu_n, \sigma_n^2)$, если $\zeta(\frac{\nu_n - \mu_n}{\sigma_n}) \to N(0, 1)$.
Найдем сначала асимптотические распределения выборочных моментов $\hat{\alpha}_{nk}$. Величина $n\hat{\alpha}_{nk} = \sum_{i=1}^{n} X_i^k$ является суммой независимых одинаково распределенных случайных величин. Если конечен момент $\alpha_{2k} = M\xi^{2k}$, то к этой сумме можно применить центральную предельную теорему теории вероятностей. Так как $MX_i^k = \alpha^k$, $DX_i^k = \alpha_{2k} - \alpha_k^2$, то величина

$$\frac{n\hat{\alpha}_{nk} - n\alpha_k}{\sqrt{n(\alpha_{2k} - \alpha_k^2)}} = \frac{\hat{\alpha}_{nk} - \alpha_k}{\sqrt{\frac{\alpha_{2k} - \alpha_k^2}{n}}}$$

асимтотически нормальна $N(0, 1)$. Таким образом справедлива следующая теорема:
Если конечен теоритический момент α_{2k} , то при $n \to \infty$ выборочный момент $\hat{\alpha}_{nk}$ асимптотически нормален $N(\alpha_k, \frac{\alpha_{2k} - \alpha_k^2}{n})$
Из теоремы следует, что если существует теоретическая дисперсия, то выборочное среднее $\hat{\alpha}_{n1}$ асимптотически нормально $N(\alpha_1, \frac{\mu_2}{n})$
Из теоремы об асимптотической нормальности функций от выборочных моментов следует, что асимптотически нормальными являются и центральные выборочные моменты $\hat{\mu}_{nk}$, поскольку они являются непрерывными функциями (многочленами) от обычных выборочных моментов.

Выборочное среднее

In [68]:
```
#n=5
vs_5 = []
for i in range(5):
    vs_5.append(np.mean(means_5[i]))
print(vs_5)
#n=10
vs_10 = []
for i in range(5):
    vs_10.append(np.mean(means_10[i]))
print(vs_10)
#n=100
vs_100 = []
for i in range(5):
    vs_100.append(np.mean(means_100[i]))
print(vs_100)
#n=1000
vs_1000 = []
for i in range(5):
    vs_1000.append(np.mean(means_1000[i]))
print(vs_1000)
#n=100000
vs_100000 = []
for i in range(5):
    vs_100000.append(np.mean(means_100000[i]))
print(vs_100000)
```

```
[2.4, 2.2, 2.8, 1.4, 2.6]
[2.2, 2.4, 2.0, 1.7, 2.7]
[2.02, 2.02, 1.91, 2.17, 1.98]
[1.978, 2.011, 2.058, 1.959, 1.954]
[2.0093, 2.0101, 2.01032, 1.99982, 2.00051]
```

$M\xi = \frac{1}{p}$, при p = 0.5 $M\xi = 2$

In [69]:
```
#Сравнение
p = 0.5
geom.stats(p, moments = 'm')
```

Out[69]: array(2.)

Выборочная дисперсия

In [70]:
```
#n=5
vd_5 = []
for i in range(5):
    vd_5.append(round(np.var(means_5[i]),6))
print(vd_5)
#n=10
vd_10 = []
for i in range(5):
    vd_10.append(round(np.var(means_10[i]),6))
print(vd_10)
#n=100
vd_100 = []
for i in range(5):
    vd_100.append(round(np.var(means_100[i]),6))
print(vd_100)
#n=1000
vd_1000 = []
for i in range(5):
    vd_1000.append(round(np.var(means_1000[i]),6))
print(vd_1000)
#n=100000
vd_100000 = []
for i in range(5):
    vd_100000.append(round(np.var(means_100000[i]),6))
print(vd_100000)
```

```
[2.24, 3.76, 2.16, 0.64, 1.04]
[2.36, 1.04, 0.8, 0.61, 2.41]
[2.5196, 2.2396, 1.6819, 2.3811, 2.8596]
[1.955516, 2.090879, 2.188636, 1.845319, 1.699884]
[2.033354, 2.020038, 2.033233, 2.02872, 1.99785]
```

$$D\xi = \frac{q}{p^2} = \frac{1-p}{p^2}$$

При p = 0.5, $D\xi = 2$

In [71]:
```
#Сравнение
geom.stats(p,moments = 'v')
```

Out[71]: array(2.)

Как видно из полученных значений, чем больше объём выборки, тем менее отличаются выборочное среднее от теоретического математического ожидания и выборочная дисперсия от теоретической дисперсии

3.1.2 Построение доверительного интервала для выборочного среднего

Определение: γ - доверительным интервалом для g называется такой случайный интервал $(T_1(X), T_2(X)), T_1(X) < T_2(X)$, который содержит внутри себя (накрывает) неизвестное значение g с вероятностью, не меньшей γ:

$$P\{T_1(X) < g < T_2(X)\} \geq \gamma$$

Здесь $T_1(X)$ и $T_2(X)$ - некоторые статистики (функции от выборки), называемые соответственно нижней и верхней доверительными границами, а γ - задаваемый заранее доверительный уровень, который обычно выбирается близким к 1. Длина доверительного интервала характеризует точность локализации оцениваемой характеристики g, а величина γ является показателем надежности доверительного интервала.

В сформулированной ранее теореме [Если конечен теоритический момент α_{2k} , то при $n \to \infty$ выборочный момент $\hat{\alpha}_{nk}$ асимптотически нормален $N(\alpha_k, \frac{\alpha_{2k}-\alpha_k^2}{n})$] можно заменить асимптотическую дисперсию $\frac{\alpha_{2k}-\alpha_k^2}{n}$ ее оценкой $\frac{\hat{\alpha}_{n,2k}-\hat{\alpha}_{nk}^2}{n}$. Это дает искомый асимптотический γ-доверительный интервал для момента α_k вида:

$$(\hat{\alpha}_{nk} \mp c_\gamma \sqrt{\frac{\hat{\alpha}_{n,2k} - \hat{\alpha}_{nk}^2}{n}})$$

Полагая здесь k = 1, получим соответствующий интервал для теоретического среднего $\alpha_1 = M\xi$:

$$(\overline{X} \mp \frac{c_\gamma S}{\sqrt{n}})$$

Чтобы построить асимптотический γ-доверительный интервал для теоретической дисперсии $\mu_2 = D\xi$, надо просто воспользоваться результатом теоремы об асимптотической нормальности выборочной дисперсии [$\zeta(\frac{\sqrt{n}(S^2-\mu_2)}{\sqrt{\hat{\mu}_{n4}-S^4}}) \to N(0, 1)$]: искомый интервал есть

$$(S^2 \mp c_\gamma \sqrt{\frac{\hat{\mu}_{n4} - S^4}{n}})$$

Положим $\gamma = 0.95$ и найдем доверительный интервал для выборочного среднего.

$$\Phi(c_\gamma) = \frac{\gamma}{2} = 0.475$$

Из таблицы значений функции Лапласа $c_\gamma \approx 1.96$

In [77]:

```
#n=5
print('n = 5')
for i in range(5):
    print('(', vs_5[i], '-+ 1.96 *', np.sqrt(vd_5[i]/5), ') = (', vs_5[i],
#n=10
print('n = 10')
for i in range(5):
    print('(', vs_10[i], '-+ 1.96 *', np.sqrt(vd_10[i]/10), ')= (', vs_10[i
#n=100
print('n = 100')
for i in range(5):
    print('(', vs_100[i], '-+ 1.96 *', np.sqrt(vd_100[i]/100), ') = (', vs_
#n=1000
print('n = 1000')
for i in range(5):
    print('(', vs_1000[i], '-+ 1.96 *', np.sqrt(vd_1000[i]/1000), ') = (',
#n=100000
print('n = 100000')
for i in range(5):
    print('(', vs_100000[i], '-+ 1.96 *', np.sqrt(vd_100000[i]/100000), ')
```

```
n = 5
( 1.8 -+ 1.96 * 0.33466401061363027 ) = ( 1.8 -+ 0.655941 )
( 1.6 -+ 1.96 * 0.35777087639996635 ) = ( 1.6 -+ 0.701231 )
( 2.6 -+ 1.96 * 0.8294576541331088 ) = ( 2.6 -+ 1.625737 )
( 1.4 -+ 1.96 * 0.35777087639996635 ) = ( 1.4 -+ 0.701231 )
( 2.4 -+ 1.96 * 0.4560701700396552 ) = ( 2.4 -+ 0.893898 )
n = 10
( 2.0 -+ 1.96 * 0.4 )= ( 2.0 -+ 0.784 )
( 3.2 -+ 1.96 * 0.7720103626247513 )= ( 3.2 -+ 1.51314 )
( 2.1 -+ 1.96 * 0.7543208866258444 )= ( 2.1 -+ 1.478469 )
( 1.6 -+ 1.96 * 0.322490309931942 )= ( 1.6 -+ 0.632081 )
( 1.5 -+ 1.96 * 0.291547594742265 )= ( 1.5 -+ 0.571433 )
n = 100
( 1.97 -+ 1.96 * 0.13817018491700733 ) = ( 1.97 -+ 0.270814 )
( 1.72 -+ 1.96 * 0.10007996802557444 ) = ( 1.72 -+ 0.196157 )
( 2.45 -+ 1.96 * 0.17342145196024625 ) = ( 2.45 -+ 0.339906 )
( 1.94 -+ 1.96 * 0.1412940196894405 ) = ( 1.94 -+ 0.276936 )
( 1.94 -+ 1.96 * 0.13024592124131948 ) = ( 1.94 -+ 0.255282 )
n = 1000
( 2.027 -+ 1.96 * 0.04624144245154989 ) = ( 2.027 -+ 0.090633 )
( 1.943 -+ 1.96 * 0.04330994112210267 ) = ( 1.943 -+ 0.084887 )
( 1.998 -+ 1.96 * 0.04409077000915271 ) = ( 1.998 -+ 0.086418 )
( 2.001 -+ 1.96 * 0.04506660626228694 ) = ( 2.001 -+ 0.088331 )
( 1.987 -+ 1.96 * 0.04618258329716951 ) = ( 1.987 -+ 0.090518 )
n = 100000
( 1.99964 -+ 1.96 * 0.004461636471072021 ) = ( 1.99964 -+ 0.008745 )
( 1.99584 -+ 1.96 * 0.004471759161672283 ) = ( 1.99584 -+ 0.008765 )
( 2.00301 -+ 1.96 * 0.004486001560409893 ) = ( 2.00301 -+ 0.008793 )
( 1.99926 -+ 1.96 * 0.004447649941261115 ) = ( 1.99926 -+ 0.008717 )
( 2.0108 -+ 1.96 * 0.004515952834120392 ) = ( 2.0108 -+ 0.008851 )
```

*Округлено до 6 знаков после запятой.

3.1.3 Нахождение оптимальности рассматриваемых оценок

Для построения теории оптимального оценивания прежде всего надо договориться о мере точности оценок, т.е. уточнить смысл приближенного равенства $T(X) \approx g$. Если статистика T(x) используется для оценивания g, то одной из разумных мер расхождения между ними является $(T(X) - g)^2$, или квадратичная ошибка. Но так как это величина случайная используется среднеквадратичная ошибка (с. к. о.) $\Delta(T) = M(T(X) - g)^2$.

Определение: Оценка минимизирующая с. к. о. в данном классе оценок T_g называется оптимальной в среднеквадратичном смысле и обозначается T^* :

$$T^* = argmin_{T \in T_g} \Delta(T)$$

Пусть требуется оценить заданную параметрическую функцию $\tau(\theta)$ в модели $F = F(x; \theta), \theta \in \Theta$ по соответствующей выборке $X = (X_1, \ldots, X_n)$. Обозначим τ_τ класс всех несмещенных оценок $T = T(X)$ для $\tau(\theta)$ и предположим, что он не пуст. Дополнительно предположим, что дисперсии всех оценок из класса τ_τ конечны: $D\theta T = M_\theta(T - \tau(\theta))2 < \infty$, в этом случае мерой точности оценок является их дисперсия.

Утверждение: Для несмещенных оценок среднеквадратичное отклонение совпадает с ее дисперсией, а для смещенной оценки больше ее дисперсии.

Доказательство:

$$M\theta(T - \tau)^2 = M(T - MT + MT - \tau)^2 = M(T - MT)^2 + M(MT - \tau)^2 + 2M((T - MT)(MT$$

$$b^2 = 0 \Leftrightarrow MT = \tau$$

Теорема Рао-Блэкуэлла-Колмогорова: Оптимальная оценка, если она существует, является функцией от достаточной статистики.

По определению достаточная статистика T = T(X) называется полной, если для всякой функции $\varphi(T)$ из того, что

$$M_\theta \varphi(T) = 0, \forall \theta$$

следует $\varphi(t) \equiv 0$ на всем множестве значений статистики T.

Теорема: Если существует полная достаточная статистика, то всякая функция от нее является оптимальной оценкой своего математического ожидания.

Итак, пусть существует полная достаточная статистика T = T(X) и требуется оценить заданную параметрическую функцию $\tau(\theta)$. Тогда:

1)Если существует какая-то несмещенная оценка $\tau(\theta)$, то существует и несмещенная оценка, являющаяся функцией от T; можно так же сказать, что если нет несмещенныхоценок вида H(T), то класс несмещенных оценок τ_τ для $\tau(\theta)$ пуст;

2)оптимальная (н.о.р.м.д.) оценка когда она существует, всегда является функцией от T и она однозначно определяется уравнением $M_\theta H(T) = \tau(\theta)$

3)оптимальную оценку τ^* можно искать по формуле:

$$\tau^* = H(T) = M_\theta(T_1|T)$$

исходя из любой несмещенной оценки T_1 функции $\tau(\theta)$. Условие регулярности(Рао-Крамера): Параметрическое семейство $F = \{F_\theta, \theta \in \Theta\}$ называется регулярным, если выполнены следующие условия(условия регулярности):

1) $L(\overline{x}, \theta) > 0$ для всех значений $\overline{x} \in B$ из выборочного пространства и дифференцируема по $\theta, \forall \theta \in \Theta$

2) Случайная величина V(X;θ), называемая функцией вклада выборки и определенная равенством

$$V(X; \theta) = \frac{\partial ln(L(\overline{x}; \theta))}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial ln(f_\theta(\overline{x}_j)}{\partial \theta}$$

имеет ограниченную дисперсию:

$$0 < M_\theta V^2(X; \theta) < \infty$$

при этом значение $\frac{\partial ln(f_\theta(\overline{x}_j)}{\partial \theta}$ будем называть вкладом i-ого наблюдения выборки.

3) $\forall \theta \in \Theta$ статистики T(x) верно равенство:

$$\frac{\partial}{\partial \theta} \int_{R^n} T(\overline{x}) L(\overline{x}; \theta) d\overline{x} = \int_{R^n} T(\overline{x}) \frac{\partial L(\overline{x}; \theta)}{\partial \theta} d\overline{x}$$

Условия регулярности:

1) Так как $0 < \theta < 1$, то $L(\overline{x}, \theta) > 0$. Функция дифференцируема по θ.

2) Воспользуемся равенством:

$$i_n(\theta) = M_\theta V^2(X; \theta) = n * i(\theta)$$

Найдём $i(\theta)$:

$$i(\theta) = -M_\theta \frac{\partial^2(ln f_\theta(x_1))}{\partial\theta}$$

$$\frac{\partial^2(ln(f(x1, \theta)))}{\partial\theta} = \frac{\partial^2(xln((1-\theta)\theta))}{\partial\theta}$$

Первая производная:

$$\frac{x\theta + \theta - 1}{(\theta - 1)\theta}$$

Вторая производная:

$$-\frac{\theta^2(x+1) + 2\theta - 1}{(\theta-1)^2\theta^2}$$

$$M(x) = \frac{1-\theta}{\theta}$$

$$M_\theta(\frac{\theta^2(x+1+2\theta-1)}{(\theta-1)^2\theta^2}) = M_\theta(\frac{\theta^2 x}{(\theta-1)^2\theta^2}) + \frac{\theta + 2\theta - 1}{(\theta-1)^2\theta^2} = (\frac{\theta^2 M(x)}{(\theta-1)^2\theta^2}) + \frac{3\theta-1}{(\theta-1)^2\theta^2} = (-\frac{6}{(6}$$

$$+ \frac{3\theta-1}{(\theta-1)^2\theta^2} = -\frac{\theta^2 - 4\theta + 1}{(\theta-1)^2\theta^2}$$

В силу того, что $0 < \theta < 1$, то информация Фишера конечна, следовательно, свойство выполняется.

3) Выборочное среднее не зависит от параметра θ, то свойство выполняется.

$$f(x) = (1-p)^x p$$

Запишем логарифмическую функцию правдоподобия:

$$L(\overline{x};\theta) = \prod_{i=1}^{n}(1-\theta)^{x_i}\theta$$

Прологарифмируем обе части:

$$lnL(\overline{x};\theta) = \sum_{i=1}^{n} ln(1-\theta)^{x_i}\theta = \sum_{i=1}^{n} ln(\theta) + \sum_{i=1}^{n} ln((1-\theta)x_i)$$

$$\frac{\partial(ln(L(x,\theta)))}{\partial\theta} = \sum_{i=1}^{n}(\frac{1}{\theta}) + \sum_{i=1}^{n}(-\frac{x_i}{1-\theta})$$

Так как

$$\sum_{i=1}^{n}(\frac{x_i}{1-\theta}) = \sum_{i=1}^{n}(\frac{1}{\theta}), \theta \neq 0; 1,$$

то

$$\frac{1}{1-\theta}\overline{X} = \frac{1}{\theta}$$

$$\overline{X} = \frac{1-\theta}{\theta}$$

$$M(\frac{1-\theta}{\theta}) = M\overline{X} = M\xi$$

А значит, оценка несмещенная. Также оценка является состоятельной, в силу того, что

$$M\overline{X} \overset{P}{\underset{n\to\infty}{\to}} \overline{X}$$

Теперь приступим к оцениванию эффективности данной оценки:

Вспомним, что

$$-M_\theta\frac{\partial^2(ln(f_\theta(X_1)))}{\partial\theta} = i_1(\theta)$$

f(x) геометрического распределения можно дважды продифференцировать.

$$\frac{\partial^2(ln(f_\theta(X_1)))}{\partial\theta} = -\frac{1}{\sigma^2}$$

$$e(T) = \frac{(r'(\theta))^2}{i_n(\theta)D_\theta T}$$

3.2 Распределения Максвелла

3.2.1 Нахождение выборочного среднего и выборочной дисперсии геометрического распределения

Выборочное среднее

In [72]:
```
#n=5
vs__5 = []
for i in range(5):
    vs__5.append(round(np.mean(means__5[i]),6))
print(vs__5)
#n=10
vs__10 = []
for i in range(5):
    vs__10.append(round(np.mean(means__10[i]),6))
print(vs__10)
#n=100
vs__100 = []
for i in range(5):
    vs__100.append(round(np.mean(means__100[i]),6))
print(vs__100)
#n=1000
vs__1000 = []
for i in range(5):
    vs__1000.append(round(np.mean(means__1000[i]),6))
print(vs__1000)
#n=100000
vs__100000 = []
for i in range(5):
    vs__100000.append(round(np.mean(means__100000[i]),6))
print(vs__100000)
```

```
[1.915438, 1.095802, 1.249879, 0.910554, 1.714316]
[1.577273, 1.727918, 1.602861, 1.788271, 1.522487]
[1.664325, 1.623293, 1.442467, 1.652924, 1.605711]
[1.599327, 1.606045, 1.588832, 1.574616, 1.550436]
[1.590289, 1.593489, 1.59538, 1.593763, 1.594673]
```

In [73]:
```
#Сравнение
2*np.sqrt(2/np.pi)
```

Out[73]: 1.5957691216057308

$$M\xi = 2\lambda\sqrt{\frac{2}{\pi}}$$

При $\lambda = 1.0$ $M\xi = 2 \cdot \sqrt{\frac{2}{\pi}} \approx 1.5957691216057308$

Выборочная дисперсия

In [74]:
```
#n=5
vd__5 = []
for i in range(5):
    vd__5.append(round(np.var(means__5[i]),6))
print(vd__5)
#n=10
vd__10 = []
for i in range(5):
    vd__10.append(round(np.var(means__10[i]),6))
print(vd__10)
#n=100
vd__100 = []
for i in range(5):
    vd__100.append(round(np.var(means__100[i]),6))
print(vd__100)
#n=1000
vd__1000 = []
for i in range(5):
    vd__1000.append(round(np.var(means__1000[i]),6))
print(vd__1000)
#n=100000
vd__100000 = []
for i in range(5):
    vd__100000.append(round(np.var(means__100000[i]),6))
print(vd__100000)
```

```
[0.406101, 0.287325, 0.097846, 0.057745, 0.327642]
[0.251981, 0.210423, 0.524976, 0.650905, 0.281158]
[0.359419, 0.474391, 0.422773, 0.447515, 0.395477]
[0.457396, 0.431927, 0.44121, 0.44778, 0.477308]
[0.451604, 0.45236, 0.453762, 0.453338, 0.456103]
```

In [75]:
```
#Сравнение
(3*np.pi-8)/np.pi
```

Out[75]:
```
0.4535209105296745
```

$$D\xi = \frac{3\pi - 8}{\pi} \cdot \lambda$$

При $\lambda = 1.0$ $D\xi = \frac{3\pi-8}{\pi} \approx 0.4535209105296745$

Как видно из полученных значений, чем больше объём выборки, тем менее отличаются выборочное среднее от теоретического математического ожидания и выборочная дисперсия от теоретической дисперсии.

3.2.2 Построение доверительного интервала для выборочного среднего

Положим $\gamma = 0.95$ и найдем доверительный интервал для выборочного среднего.

$$\Phi(c_\gamma) = \frac{\gamma}{2} = 0.475$$

Из таблицы значений функции Лапласа $c_\gamma \approx 1.96$

In [76]:

```
#n=5
print('n = 5')
for i in range(5):
    print('(', vs__5[i], '-+ 1.96 *', np.sqrt(vd__5[i]/5), ') = (', vs__5[i
#n=10
print('n = 10')
for i in range(5):
    print('(', vs__10[i], '-+ 1.96 *', np.sqrt(vd__10[i]/10), ')= (', vs__1
#n=100
print('n = 100')
for i in range(5):
    print('(', vs__100[i], '-+ 1.96 *', np.sqrt(vd__100[i]/100), ') = (', v
#n=1000
print('n = 1000')
for i in range(5):
    print('(', vs__1000[i], '-+ 1.96 *', np.sqrt(vd__1000[i]/1000), ') = ('
#n=100000
print('n = 100000')
for i in range(5):
    print('(', vs__100000[i], '-+ 1.96 *', np.sqrt(vd__100000[i]/100000), '
```

```
n = 5
( 1.915438 -+ 1.96 * 0.284991578822954 ) = ( 1.915438 -+ 0.558583 )
( 1.095802 -+ 1.96 * 0.23971858501167573 ) = ( 1.095802 -+ 0.469848 )
( 1.249879 -+ 1.96 * 0.13988995675172683 ) = ( 1.249879 -+ 0.274184 )
( 0.910554 -+ 1.96 * 0.10746627377926528 ) = ( 0.910554 -+ 0.210634 )
( 1.714316 -+ 1.96 * 0.2559851558196295 ) = ( 1.714316 -+ 0.501731 )
n = 10
( 1.577273 -+ 1.96 * 0.15873909411357998 )= ( 1.577273 -+ 0.311129 )
( 1.727918 -+ 1.96 * 0.14505964290594403 )= ( 1.727918 -+ 0.284317 )
( 1.602861 -+ 1.96 * 0.22912354745857091 )= ( 1.602861 -+ 0.449082 )
( 1.788271 -+ 1.96 * 0.2551283990464409 )= ( 1.788271 -+ 0.500052 )
( 1.522487 -+ 1.96 * 0.16767766696850242 )= ( 1.522487 -+ 0.328648 )
n = 100
( 1.664325 -+ 1.96 * 0.05995156378277384 ) = ( 1.664325 -+ 0.117505 )
( 1.623293 -+ 1.96 * 0.06887604808639938 ) = ( 1.623293 -+ 0.134997 )
( 1.442467 -+ 1.96 * 0.06502099660878784 ) = ( 1.442467 -+ 0.127441 )
( 1.652924 -+ 1.96 * 0.06689656194454241 ) = ( 1.652924 -+ 0.131117 )
( 1.605711 -+ 1.96 * 0.06288696208277197 ) = ( 1.605711 -+ 0.123258 )
n = 1000
( 1.599327 -+ 1.96 * 0.021386818370201774 ) = ( 1.599327 -+ 0.041918 )
( 1.606045 -+ 1.96 * 0.020782853509564082 ) = ( 1.606045 -+ 0.040734 )
( 1.588832 -+ 1.96 * 0.021004999404903586 ) = ( 1.588832 -+ 0.04117 )
( 1.574616 -+ 1.96 * 0.021160812838830177 ) = ( 1.574616 -+ 0.041475 )
( 1.550436 -+ 1.96 * 0.021847379705584834 ) = ( 1.550436 -+ 0.042821 )
n = 100000
( 1.590289 -+ 1.96 * 0.0021250976448154092 ) = ( 1.590289 -+ 0.004165 )
( 1.593489 -+ 1.96 * 0.002126875642815066 ) = ( 1.593489 -+ 0.004169 )
( 1.59538 -+ 1.96 * 0.0021301690073794615 ) = ( 1.59538 -+ 0.004175 )
( 1.593763 -+ 1.96 * 0.0021291735485863992 ) = ( 1.593763 -+ 0.004173 )
( 1.594673 -+ 1.96 * 0.0021356568076355336 ) = ( 1.594673 -+ 0.004186 )
```

*Округлено до 6 знаков после запятой.

3.2.3 Нахождение оптимальности рассматриваемых оценок

Для простоты предположим, что все частоты $p_i, i = \overline{1,n}$ равны единице.
Запишем функцию максимального правдоподобия для закона Максвелла.

$$L(\lambda) = (\frac{2}{\pi})^{\frac{n}{2}} \frac{\prod_{i=1}^{n} x_i^2}{\lambda^{3n}} e^{-\frac{1}{2\lambda^2}\sum_{i=1}^{n} x_i^2}$$

$$ln(L(\lambda)) = 2\sum_{i=1}^{n} lnx_j - 3nln\lambda - \frac{n}{2}ln\frac{2}{\pi} - \frac{1}{2\lambda^2}\sum_{i=1}^{n} x_i^2$$

$$\frac{\partial(lnL(\lambda))}{\partial\lambda} = -\frac{3n}{\lambda} + \frac{\sum_{i=1}^{n} x_i^2}{\lambda^3} = 0$$

Переходя к статистическому ряду (не все p_i равны 1, $i = \overline{1,n}$), получим уравнение для нахождения λ:

$$\lambda = \sqrt{\frac{\sum_{i=1}^{m} p_i x_i^2}{3n}}$$

Оценка методом моментов:
Поскольку по выборке оценивается лишь один параметр, то для нахождения λ используемся оценку математического ожидания.

$$M[X] = \overline{X} = 2\lambda\sqrt{\frac{2}{\pi}},$$

где $\overline{X} = \frac{1}{n}\sum_{i=1}^{m} x_i p_i, \sum_{i=1}^{m} p_i = n$
Отсюда $\lambda = \sqrt{\frac{\pi}{8}}\overline{X}$

3.3 Работа с данными

В данном задании проведем анализ реальных данных. Воспользуемся нетипичной интерпретацией геометрического распределения - игра бейсбол. В бейсболе геометрическое распределение полезно для анализа вероятности того, что отбивающий получит удар, прежде чем он получит три удара; здесь цель - добиться успеха за 3 испытания.
Рассмотрим финальную серию чемпионата МЛБ-2020 между Лос Анджелес Доджерс и Тампа-Бэй Рэйс. Проанализируем статистику бэттеров первых 5 матчей, чтобы оценить какие бэттеры должны чаще выходить на поле в последнем матче серии и сравним с тем, как на самом деле это было.
Почему принято решение выбрать именно такой подход? Всё из-за того, что чем выше процент отбивания у бэттера, тем выше вероятность того, что произойдет успех. Докажем это:
Пусть вероятность того, что бэттер отобьёт удар равна p. Тогда рассчитаем вероятность успеха:
$P(X = 0) + P(X = 1) + P(X = 2) = q^0 p + q^1 p + q^2 p = p + (1-p)p + (1-p)^2 p = p + p - p^2$

Пройдемся по циклу и убедимся, что при увеличении p увеличивается и вероятность успеха.

In [77]:
```
for p in [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]:
    print(p*(p**2-3*p+3))
```

```
0.271
0.488
0.6569999999999999
0.784
0.875
0.9359999999999999
0.9730000000000002
0.9919999999999999
0.9989999999999999
```

Рассмотрим статистические показатели, которые нам пригодятся:
AB - At Bats = PA - BB - IBB - HBP - CI - SF - SH (На бите): Выходы на биту бэттера, за исключением уоков, ударов мячом, пожертвованных ударов, умышленных помех со стороны защиты или других препятствий.
H - Hits (Хиты): общее количество хитов (1B, 2B, 3B и HR). Хит - удар, давший возможность выйти на базу. При этом защита не совершила ошибку.
BA - Batting average = H / AB (он же AVG - средний коэффициент результативности отбивания): число хитов, деленное на число выходов на биту.

In [78]:
```
match_1 = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-1.csv',',
match_2 = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-2.csv',',
match_3 = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-3.csv',',
match_4 = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-4.csv',',
match_5 = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-5.csv',',
```

In [79]:
```
match_1
```

Out[79]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	2	0	0.000
1	Arozarena R.	TAM	3	0	0.000
2	Barnes A.	LAD	4	0	0.000
3	Bellinger C.	LAD	4	1	0.250
4	Betts M.	LAD	4	2	0.500
5	Brosseau M.	TAM	1	1	1.000
6	Choi Ji-Man	TAM	0	0	0.000
7	Diaz Y.	TAM	4	1	0.250
8	Hernandez E.	LAD	2	1	0.500
9	Kiermaier K.	TAM	3	2	0.667
10	Lowe B.	TAM	4	0	0.000
11	Margot M.	TAM	4	1	0.250
12	Meadows A.	TAM	2	0	0.000
13	Muncy M.	LAD	4	2	0.500
14	Pederson J.	LAD	2	0	0.000
15	Renfroe H.	TAM	2	0	0.000
16	Seager C.	LAD	2	0	0.000
17	Taylor C.	LAD	3	2	0.667
18	Turner J.	LAD	4	1	0.250
19	Wendle J.	TAM	4	1	0.250
20	Zunino M.	TAM	3	0	0.000
21	Smit U.	LAD	5	1	0.200

In [80]:
```
1 match_2
```

Out[80]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	4	1	0.167
1	Arozarena R.	TAM	3	1	0.167
2	Barnes A.	LAD	1	0	0.000
3	Bellinger C.	LAD	3	0	0.143
4	Betts M.	LAD	3	0	0.286
5	Brosseau M.	TAM	2	0	0.333
6	Choi Ji-Man	TAM	3	1	0.333
7	Diaz Y.	TAM	1	1	0.400
8	Hernandez E.	LAD	1	0	0.333
9	Kiermaier K.	TAM	4	0	0.286
10	Lowe B.	TAM	5	2	0.222
11	Margot M.	TAM	3	2	0.429
12	Meadows A.	TAM	3	1	0.200
13	Muncy M.	LAD	3	0	0.286
14	Pederson J.	LAD	1	0	0.000
15	Phillips B.	TAM	0	0	0.000
16	Pollock A.	LAD	2	0	0.000
17	Renfroe H.	TAM	0	0	0.000
18	Rios E.	LAD	2	0	0.000
19	Seager C.	LAD	4	2	0.333
20	Taylor C.	LAD	4	1	0.429
21	Turner J.	LAD	4	1	0.250
22	Wendle J.	TAM	3	1	0.286
23	Zunino M.	TAM	4	0	0.000
24	Smit U.	LAD	4	1	0.222

In [81]:
```
1 match_3
```

Out[81]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	3	1	0.222
1	Arozarena R.	TAM	4	1	0.200
2	Barnes A.	LAD	3	1	0.125
3	Bellinger C.	LAD	4	1	0.182
4	Betts M.	LAD	5	2	0.333
5	Choi Ji-Man	TAM	4	0	0.143
6	Hernandez E.	LAD	1	0	0.250
7	Kiermaier K.	TAM	2	0	0.222
8	Lowe B.	TAM	4	0	0.154
9	Margot M.	TAM	3	1	0.400
10	Meadows A.	TAM	4	1	0.222
11	Muncy M.	LAD	4	2	0.364
12	Pederson J.	LAD	3	1	0.167
13	Seager C.	LAD	3	1	0.333
14	Taylor C.	LAD	4	0	0.273
15	Tsutsugo Y.	TAM	1	0	0.000
16	Turner J.	LAD	5	2	0.308
17	Wendle J.	TAM	3	0	0.200
18	Zunino M.	TAM	2	0	0.000
19	Perez M.	TAM	0	0	0.000
20	Smit U.	LAD	4	0	0.154

In [82]:
```
1 match_4
```

Out[82]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	4	1	0.231
1	Arozarena R.	TAM	4	3	0.357
2	Bellinger C.	LAD	4	0	0.133
3	Betts M.	LAD	5	0	0.235
4	Brosseau M.	TAM	2	1	0.400
5	Choi Ji-Man	TAM	0	0	0.143
6	Diaz Y.	TAM	3	0	0.250
7	Hernandez E.	LAD	4	1	0.250
8	Kiermaier K.	TAM	4	2	0.308
9	Lowe B.	TAM	4	1	0.176
10	Margot M.	TAM	2	0	0.333
11	Meadows A.	TAM	2	0	0.182
12	Muncy M.	LAD	4	1	0.333
13	Pederson J.	LAD	2	2	0.375
14	Phillips B.	TAM	1	1	1.000
15	Pollock A.	LAD	2	1	0.250
16	Renfroe H.	TAM	4	1	0.167
17	Seager C.	LAD	5	4	0.500
18	Taylor C.	LAD	5	1	0.250
19	Tsutsugo Y.	TAM	1	0	0.000
20	Turner J.	LAD	5	4	0.444
21	Wendle J.	TAM	1	0	0.182
22	Zunino M.	TAM	2	0	0.000
23	Smit U.	LAD	4	1	0.176

In [83]:
```
1 match_5
```

Out[83]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	4	0	0.176
1	Arozarena R.	TAM	4	1	0.333
2	Barnes A.	LAD	2	0	0.100
3	Bellinger C.	LAD	4	1	0.158
4	Betts M.	LAD	5	1	0.227
5	Brosseau M.	TAM	0	0	0.400
6	Choi Ji-Man	TAM	0	0	0.143
7	Diaz Y.	TAM	3	2	0.364
8	Hernandez E.	LAD	1	0	0.222
9	Kiermaier K.	TAM	3	2	0.375
10	Lowe B.	TAM	4	0	0.143
11	Margot M.	TAM	3	2	0.400
12	Meadows A.	TAM	2	0	0.154
13	Muncy M.	LAD	3	2	0.389
14	Pederson J.	LAD	2	1	0.400
15	Renfroe H.	TAM	1	0	0.143
16	Seager C.	LAD	3	1	0.471
17	Taylor C.	LAD	4	0	0.200
18	Tsutsugo Y.	TAM	1	0	0.000
19	Turner J.	LAD	4	0	0.364
20	Wendle J.	TAM	4	0	0.133
21	Zunino M.	TAM	2	0	0.000
22	Perez M.	TAM	0	0	0.000
23	Smit U.	LAD	4	0	0.143

In [84]:
```
1 itog = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Итого.csv',',',pa
2 itog
```

Out[84]:

	Игрок	Команда	AB	H	BA
0	Adames W.	TAM	17	3	0,176470588235294
1	Arozarena R.	TAM	18	6	0,333333333333333
2	Barnes A.	LAD	10	1	0,1
3	Bellinger C.	LAD	19	3	0,157894736842105
4	Betts M.	LAD	22	5	0,227272727272727
5	Brosseau M.	TAM	5	2	0,4
6	Choi Ji-Man	TAM	7	1	0,142857142857143
7	Diaz Y.	TAM	11	4	0,363636363636364
8	Hernandez E.	LAD	9	2	0,222222222222222
9	Kiermaier K.	TAM	16	6	0,375
10	Lowe B.	TAM	21	3	0,142857142857143
11	Margot M.	TAM	15	6	0,4
12	Meadows A.	TAM	13	2	0,153846153846154
13	Muncy M.	LAD	18	7	0,388888888888889
14	Pederson J.	LAD	10	4	0,4
15	Phillips B.	TAM	1	1	1
16	Pollock A.	LAD	4	1	0,25
17	Rios E.	LAD	2	0	0
18	Seager C.	LAD	17	8	0,470588235294118
19	Taylor C.	LAD	20	4	0,2
20	Turner J.	LAD	22	8	0,363636363636364
21	Smit U.	LAD	21	3	0,142857142857143
22	Renfroe H.	TAM	7	1	0,142857142857143
23	Wendle J.	TAM	15	2	0,133333333333333
24	Zunino M.	TAM	13	0	0
25	Tsutsugo Y.	TAM	3	0	0

А теперь попробуем спрогнозировать кто из бэттеров будет подходить к бите больше в своей команде, анализируя приведенные выше статистические данные по итогам первых 5 матчей финальной серии. Выведем данные в порядке убывания.

In [85]:
```
prognoz_TAM = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Итого_TAM.
prognoz_TAM
```

Out[85]:

	Игрок	Команда	AB	H	BA	Комментарий
0	Margot M.	TAM	15	6	0,4	Лучший BA, один из лучших AB
1	Arozarena R.	TAM	18	6	0,333333333333333	Один из лидеров по AB и отличный BA
2	Adames W.	TAM	17	3	0,176470588235294	Один из лидеров по AB и средний BA
3	Lowe B.	TAM	21	3	0,142857142857143	Лучший AB, но значимо хуже BA
4	Kiermaier K.	TAM	16	6	0,375	Очень хороший BA и неплохой AB
5	Wendle J.	TAM	15	2	0,133333333333333	Примерно равные средние показатели
6	Meadows A.	TAM	13	2	0,153846153846154	Примерно равные средние показатели
7	Diaz Y.	TAM	11	4	0,363636363636364	Отличный BA, но все же низкий AB
8	Zunino M.	TAM	13	0	0	Приличный AB, но худший BA
9	Choi Ji-Man	TAM	7	1	0,142857142857143	Низкие AB и BA
10	Renfroe H.	TAM	7	1	0,142857142857143	Низкие AB и BA
11	Brosseau M.	TAM	5	2	0,4	Один из худших AB, но отличный BA
12	Tsutsugo Y.	TAM	3	0	0	Очень низкие AB и BA
13	Phillips B.	TAM	1	1	1	Подходил к бите всего лишь раз

А теперь посмотрим, сколько в итоге было подходов к бите у игроков TAM. Выведем данные в порядке убывания.

In [86]:
```
match_6_TAM = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-6_TAM
match_6_TAM
```

Out[86]:

	Игрок	Команда	AB	H
0	Margot M.	TAM	4	0
1	Arozarena R.	TAM	4	2
2	Adames W.	TAM	4	0
3	Lowe B.	TAM	3	0
4	Kiermaier K.	TAM	3	1
5	Wendle J.	TAM	3	0
6	Meadows A.	TAM	3	1
7	Zunino M.	TAM	3	1
8	Choi Ji-Man	TAM	2	0
9	Diaz Y.	TAM	1	0
10	Renfroe H.	TAM	1	0
11	Brosseau M.	TAM	1	0

In [87]:
```
prognoz_LAD = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Итого_LAD.
prognoz_LAD
```

Out[87]:

	Игрок	Команда	AB	H	BA	Комментарий
0	Turner J.	LAD	22	8	0,363636363636364	Лучший AB, один из лучших BA
1	Betts M.	LAD	22	5	0,227272727272727	Лучший AB, высокий BA
2	Seager C.	LAD	17	8	0,470588235294118	Высокий AB, лучший BA
3	Muncy M.	LAD	18	7	0,388888888888889	Высокий AB, один из лучших BA
4	Taylor C.	LAD	20	4	0,2	Высокий AB, неплохой BA
5	Smit U.	LAD	21	3	0,142857142857143	Высокий AB, неплохой BA
6	Bellinger C.	LAD	19	3	0,157894736842105	Высокий AB, неплохой BA
7	Pederson J.	LAD	10	4	0,4	Средний AB, высокий BA
8	Hernandez E.	LAD	9	2	0,222222222222222	Средние показатели
9	Barnes A.	LAD	10	1	0,1	Средние показатели
10	Pollock A.	LAD	4	1	0,25	Низкий AB, неплохой BA
11	Rios E.	LAD	2	0	0	Всего лишь 2 подхода к бите

А теперь посмотрим, сколько в итоге было подходов к бите у игроков LAD. Выведем данные в порядке убывания.

In [88]:
```
match_6_LAD = pd.read_csv('/home/alexander/Рабочий стол/Курсовая/Матч-6_LAD
match_6_LAD
```

Out[88]:

	Игрок	Команда	AB	H
0	Betts M.	LAD	4	2
1	Muncy M.	LAD	4	0
2	Turner J.	LAD	3	0
3	Seager C.	LAD	3	0
4	Taylor C.	LAD	3	1
5	Smit U.	LAD	3	1
6	Bellinger C.	LAD	3	0
7	Barnes A.	LAD	3	1
8	Pollock A.	LAD	2	0
9	Hernandez E.	LAD	1	0
10	Pederson J.	LAD	0	0
11	Rios E.	LAD	0	0

Оценка того, кто больше из игроков будет подходить к бите в своей команде оказалось достаточно точной, несмотря на то, что данные были взяты лишь по 5 последним играм, что подтверждает важность этих показателей у игроков. Конечно же, по данным за такой короткий период корректировку могли внести такие ситуации как, например, травмы игроков, из-за которых они не смогли бы участвовать в 6 финальном матче или же участвовать в нём не в полной мере. Обычно в бейсболе статистику игроков оценивают по всему прошедшему сезону, что даст более высокую точность.

Как можно заметить, рассмотренные статистические показатели у игроков команды LAD лучше, чем у игроков команды TAM. Отсюда неудивительно, что LAD выиграли не только последний матч у TAD со счетом 3:1, но и всю финальную серию со счетом 4:2 и стали чемпионами MLB-2020.

Ссылка на данные: https://www.scoreboard.com/ru/baseball/usa/mlb-2020/ (https://www.scoreboard.com/ru/baseball/usa/mlb-2020/)

4 Домашнее задание. Проверка статистических гипотез

1.Критерий согласия Колмагорова
Пусть дана выборка $X = (X_1, \ldots, X_n)$ из распределения $L(\xi)$и$F\xi$ - неизвестное распределение.

- $H_0 : F\xi = F(x)$ - простая гипотеза
- $H_1 :$ не$F(x)$

Критерий Колмогорова основан на теореме Колмогорова:

$$D_n = D_n(x) = sup|\hat{F}_n(x) - F(x)|_{x \in R}$$

где D_n - это отклонение эмпирической функции распределения от теоретической функции распределения.

$\hat{F}_n$ - оптимальная несмещенная состоятельная оценка для F(x).

Замечание: D_n не должно сильно отклоняться от 0.

По т. Колмогорова:

$$P(nD_n \geq \lambda_\alpha | H_0) = 1 - K(\lambda_\alpha) = \alpha$$

по $\alpha \to \lambda_\alpha$

Проверяем, выполняется ли неравенство: $nD_n \geq \lambda_\alpha$

Известно, что

$$X_1 = \{x : D_n(x)\sqrt{n} \geq \lambda_\alpha\}$$

Следовательно, H_0 отвергается $\Leftrightarrow nD_n \geq \lambda_\alpha$

По Долошеву: $\frac{6nD_n}{6\sqrt{n}}$ сходится к распределению Колмогорова, причем $\sqrt{n}D_n \in \frac{1}{6\sqrt{n}}$

Способ вычисления $D_n = sup|F_n(x) - F(x)|_{x \in R}$. Вычисление супремума функции не является тривиальной задачей. Однако в данном случае $\hat{F}_n(x)$ принимает конечное число значений: $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\}$, что значительно упрощает задачу. Пусть у нас есть вариационный ряд выборки: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. Определим следующие две функции:

$$D_n^+ = max_{1 \leq k \leq n}|\frac{k}{n} - F(x_{(k)})|$$

$$D_n^- = max_{1 \leq k \leq n}|F(x_{(k)}) - \frac{k-1}{n}|$$

Тогда вычислить D_n можно следующим образом:

$$D_n = max\{D_n^-, D_n^+\}$$

Однако критерий Колмагорова обладает рядом минусов: 1.Функция $D_n = sup|F_n(x) - F(x)|_{x \in R}$ не зависит от вида функции распределения F(x), только в случае если F (x) непрерывная. Встает вопрос, что делать если F (x) имеет точки разрыва.

Пусть $Y_1, \ldots, Y_n$ - н.о.р.сл.в. $Y_i \approx R[0,1]$. А $X_1, \ldots, X_n$ - выборка из некоторого распределения, функция которого имеет точки разрыва. Построим следующую случайную величину: $U_i = F(x_i^-) + Y_i[F(x_i - F(x_i^-)]$, где $F(x_i^-) = limF(x_i - z)_{z \downarrow 0}$. Доказывается, что случайная величина $U_i \approx R[0,1]$

2.В случае сложных гипотез распределение $D_n(\theta)$, зависит как от вида априорных распределений, так и от способа получения оценок, размера выборки n, вида θ.

$$H_0 : F_\xi(x) \in F_0 = \{F_\theta(x), \theta \in \Theta\}$$

придется исследовать следующую статистику

$$b_n = sup_{x \in R}\hat{F}_n(x) - F_{\hat{\theta}(x)},$$

где $\hat{\theta}$ есть зависимость от θ. Это плохо потому что, по одной выборке считается статистический параметр $\hat{\theta}$ и вычисляется критерий. То есть, есть зависимость. В таком случае, при наличии достаточно большой выборки длины m следует выбрать границу n и использовать первую часть для оценки параметра, а вторую для вычисления критерия.

Определение:

Пусть случайные величины $\xi_1, \xi_2, \ldots, \xi_n$ имеют стандартное нормальное распределение, тогда случайная величина:

$$\chi_n^2 = \sum_{i=1}^{n} \xi_i^2$$

имеет распределение, которое называется хи квадрат с n - степенями свободы.

2. Критерий Хи-квадрат

   Пусть ξ - случайный вектор $\xi = (\xi_1, \ldots, \xi_n)$, и $\xi_i \approx N(0,1)$. И, вектор ξ имеет единичную матрицу ковариаций. Пусть также c = $(c_1, c_2, \ldots, c_n) \in R^N$, такой что |c| = 1.

   Рассмотрим проекцию $\overline{\xi}^{(c)}$ на гиперплоскость $L_{\overline{c}} = x \in R^n : (\overline{x}, \overline{c}) = 0$, которая ортагональна вектору $\overline{c}$. Тогда вектор ξ имеет математическое ожидание равное $\theta = (0, \ldots, 0)$ и матрицу ковариации

$$C(\overline{\xi}^{(c)}) = E - ||c_i c_j||_{i,j=1}^n$$

4.1 Геометрическое распределение

4.1.1 Проверка гипотез о виде распределения

Сформулируем гипотезу H_0 и H_1 : Пусть дана выборка $X = (X_1, \ldots, X_n)$ из распределения $L(\xi)$ и F_ξ - неизвестное распределение.
• H_0 : F_ξ = $Geom(x, 0.5)$ - простая гипотеза, Geom - выбранное дискретное распределение
• H_1 : не $Geom(x, 0.5)$
Для проверки гипотезы H_0 воспользуемся критерием Пирсона (хи-квадрат). Для каждой выборки объемов n = 10, 100, 1000, 100000 найдем значение критерия, границу критического множества для уровня значимости $\alpha = 0.1$ и $\alpha = 0.05$. В каждом случае возьмем N = 15. Разделим каждую выборку на равновероятностные интервалы np.arange(0.01, 1, $\frac{1}{15}$), в случае если взятый интервал меньше, чем 1, он склеивается со следующим.
ЗДЕСЬ ДОЛЖЕН БЫТЬ КОД, ВЫЧИСЛЯЮЩИЙ ЗНАЧЕНИЕ КРИТЕРИЯ, ГРАНИЦЫ ДЛЯ ЗАДАННЫХ УРОВНЕЙ И ОПРЕДЕЛЯЮЩИЙ ДЛЯ КАЖДОЙ ВЫБОРКИ ПРИНИМАЕТСЯ ЛИ ГИПОТЕЗА.

In [99]:
```
means_10
```

Out[99]:
```
[array([1, 1, 2, 1, 4, 2, 1, 1, 2, 1]),
 array([2, 1, 3, 1, 3, 1, 1, 1, 3, 1]),
 array([1, 2, 2, 2, 2, 1, 4, 1, 1, 1]),
 array([2, 2, 2, 2, 1, 1, 1, 2, 1, 1]),
 array([4, 1, 1, 2, 1, 1, 1, 2, 3, 2])]
```

In [94]:
```
def intervals_v2(X):
    vr = []
    xr = [0]
    v = len(X)
    r = 0
    Xs = sorted(X)
    a1 = Xs[0]
    an = Xs[len(X) - 1]
    d = (an - a1)/15 #длина интервала
    s = 0
    x0 =-d
    x1 = 0
    k = 0
    while s!=v: #подсчёт значений vr(попаданий в интервал)
        r += 1
        x0 += d
        x1 += d
        for z in range(v):
            if x0 <= X[z] < x1:
                k += 1
        if k >=4: #проверка, в каждый интервал не меньше 4 значений
            vr.append(k)
            xr.append(x1)
            s += k
            k = 0
        elif s+k == v:
            vr.append(k)
            xr.append(x1)
            break
    return vr, xr
```

In [95]:
```
def pearson_criterion(X):
    print("Выборки объема: ", len(X[0]))
    v = len(X[0])
    for t in range(5):
        print("Выборка: ", t+1)
        chi = 0
        vr, xr = intervals_v2(X[t])
        for i in range(len(vr)):
            v_i = vr[i]
            a = xr[i]
            b = xr[i+1]
            p_i = e**(-a*a/8)-e**(-b*b/8)
            q = ((v_i - v*p_i)**2/(v*p_i))
            chi += q
        print("Статистика критерия Пирсона Chi2:", chi)
        print("Число степеней свободы: ", len(vr)-1)
```

In [96]:
```
intervals_v2(means_1000[0])
```

Out[96]:
```
([483, 280, 112, 58, 31, 21, 8, 4, 3],
 [0,
  1.4666666666666666,
  2.1999999999999997,
  3.6666666666666665,
  4.3999999999999995,
  5.133333333333333,
  6.6,
  7.333333333333333,
  9.533333333333331,
  12.466666666666661])
```

In [97]:
```
y = [24, 27, 21, 24, 20, 28, 21, 25, 24, 24], [23, 28, 29, 17, 23, 28, 24,
```

In [98]:
```
pearson_criterion(means_10)
print()
pearson_criterion(means_100)
print()
pearson_criterion(means_1000)
print()
pearson_criterion(means_100000)
```

```
Выборки объема:  10
Выборка:  1
Статистика критерия Пирсона Chi2: 12.786544815008364
Число степеней свободы:  1
Выборка:  2
Статистика критерия Пирсона Chi2: 16.932053428952944
Число степеней свободы:  1
Выборка:  3
Статистика критерия Пирсона Chi2: 9.599063445717778
Число степеней свободы:  2
Выборка:  4
Статистика критерия Пирсона Chi2: 11.887917289322628
Число степеней свободы:  1
Выборка:  5
Статистика критерия Пирсона Chi2: 7.4846893492343
Число степеней свободы:  2

Выборки объема:  100
Выборка:  1
Статистика критерия Пирсона Chi2: 117.5990433572617
Число степеней свободы:  5
Выборка:  2
Статистика критерия Пирсона Chi2: 34.79772027043223
Число степеней свободы:  5
Выборка:  3
Статистика критерия Пирсона Chi2: 101.44634548906481
Число степеней свободы:  4
Выборка:  4
Статистика критерия Пирсона Chi2: 152.3897880295792
Число степеней свободы:  5
Выборка:  5
Статистика критерия Пирсона Chi2: 71.69093973912976
Число степеней свободы:  4

Выборки объема:  1000
Выборка:  1
Статистика критерия Пирсона Chi2: 1257.2838385174853
Число степеней свободы:  8
Выборка:  2
Статистика критерия Пирсона Chi2: 103515.17036948926
Число степеней свободы:  9
Выборка:  3
Статистика критерия Пирсона Chi2: 217.5278124946671
Число степеней свободы:  8
Выборка:  4
Статистика критерия Пирсона Chi2: 513.2244278989731
Число степеней свободы:  7
Выборка:  5
Статистика критерия Пирсона Chi2: 1483.5149838733266
Число степеней свободы:  9

Выборки объема:  100000
Выборка:  1
Статистика критерия Пирсона Chi2: 176254935.18994218
Число степеней свободы:  12
Выборка:  2
Статистика критерия Пирсона Chi2: 82414828589691.83
Число степеней свободы:  13
Выборка:  3
Статистика критерия Пирсона Chi2: 120776582.70235875
Число степеней свободы:  14
```

In [ ]:
```
1
```

Теперь проверим заведомо ложную гипотезу:
• H_0 : $F_\xi = Geom(x, 0.8)$ - простая гипотеза, Geom - выбранное дискретное распределение
• H_1 : не Geom(x, 0.8)

In [ ]:
```
1
```

Видно, что для близких по значениям параметров гипотез критерий уверенно отклоняет неверную гипотезу только при больших объемах выборки. Это говорит о том, что при малых объемах выборки критерий говорит лишь о приближенных значениях параметров.

3. Критерий хи-квадрат для сложных гипотез В общем случае сложные для полиномиального распределения, используемого в критерии χ^2 , гипотезы будут принимать следующий вид.

$$H_0 : p = p(\theta), \theta = (\theta_1, \dots, \theta_r), \theta \in \Theta, r < N - 1$$

Тогда по аналогии с предыдущим случаем можем получить статистику $\dot{X}_n^2(\theta) = \sum_{j=1}^{N} \frac{(\nu_j - np_j(\theta))^2}{np_j(\theta)}$
Эта статистика зависит от неизвестного параметра, поэтому использовать непосредственно её нельзя. Для этого параметр θ заменяеют некоторой оценкой $\hat{\theta}$ и получают в итоге статистику $\hat{X}_n^2 = X_n^2(\hat{\theta})$. Однако узнать распределение $\hat{X}_n^2$ при гипотезе H_0 представляется трудной задачей. Кроме того, величины $p_j(\hat{\theta})$ представляют собой функции от наблюдений.
Простая гипотеза H_0 заключается в том, что $p = \dot{p} = (\dot{p}_1, \dot{p}_2, \dots, \dot{p}_N)$ - заданый вероятностный вектор $(0 < \dot{p}_j < 1, j = 1, \dots, N; \dot{p}_1 + \dots + \dot{p}_N = 0)$ Р. Фишер в 1924 г. получил, предельное распределение статистики $\hat{X}_n^2$, использующая оценку максимального правдоподобия

$$\hat{\theta} = argmax_\theta \prod_{j=1}^{N} (p_j(\theta))^{\nu_j}$$

или иначе оценка по видоизменённому методу минимизации χ_2, является распределением $\chi^2(N - 1 - r)$

$$L(\hat{X}_n^2 | H_0) \to \chi^2(N - 1 - r), n \to \infty$$

Тогда можно заключить, что критерий имеет вид:
H_0 отвергается $\leftrightarrow \dot{X}_n^2 > 1 - \alpha$-квантиль распределения $\chi^2(N - 1 - r)$

Для упрощения задачи будем пользоваться оценкой параметра M , найденной методом максимального правдоподобия: $\hat{\theta}_1 = -\frac{N}{}$

Type *Markdown* and LaTeX: α^2

4.1.2 Проверка параметрических гипотез

Данное задание посвящено вопросу различения двух простых параметрических гипотез, а также закреплению основных понятий, и практическому применению методов математической статистики для решения поставленной задачи.

4.1.2.1 Выбор данных

По причине конечности данного на решение задачи времени для ускорения завершения данного домашнего задания выберем геометрическое распределение и рассмотрим две выборки с разными (но известными) параметрами. А именно, к уже сконфигурированным выборкам из распределения с параметром p = 0.5, добавим к рассмотрению выборку из распределения с параметром p = 0.8. Сгенерируем новые выборки размеров 10, 100, 1000, 10000 и запишем их в файлы по уже отработанной схеме моделирования.

4.1.2.2 Постановка задачи

Критерий однородности Смирнова используется для проверки гипотезы о принадлежности двух независимых выборок одному закону распределения, то есть о том, что два эмпирических распределения соответствуют одному и тому же закону.
Мы сопоставляем сначала частоты по первому разряду, потом по сумме первого и второго разрядов, потом по сумме первого, второго и третьего разрядов и т. д. Таким образом, мы сопоставляем всякий раз накопленные к данному разряду частоты.
Если различия между двумя распределениями существенны, то в какой-то момент разность накопленных частот достигнет критического значения, и мы сможем признать различия статистически достоверными. В формулу критерия λ включается эта разность. Чем больше эмпирическое значение λ, тем более существенны различия.
Обозначим за нулевую гипотезу H_0, что две исследуемые выборки подчиняются одному распределению случайной величины ξ . Соответственно, H_1 - исследуемые выборки не однородны.
Обозначим как X_0 - часть пространства наблюдений такая, что если $x \in X_0$, то следует принять H_0, и как Y_1 - часть пространства наблюдений такая, что если $x \in X_1$, то следует принять H_1.
Простым языком, если $x \in X_1$, а на самом деле истинна гипотеза H_0, то говорится, что допущена ошибка первого рода. Если с точностью до наоборот - это ошибка второго рода. Вероятность $P_1(X_1)$ отвергнуть гипотезу H_0, когда она действительно является ложной, называется мощностью критерия.
$P(x \in X_1|H_0) = \alpha$ - ошибка 1 рода.
$P(x \in X_0|H_1) = \beta$ - ошибка 2 рода.
Функция мощности критерия - функционал на множестве допустимых распределений F и выборке X.

$$W(F) = W(F; X_{1,\alpha}) = P(x \in X_{1,\alpha}|F),$$

где $P(x \in X_{1,\alpha}|F)$ - вероятность попасть в $X_{1,\alpha}$, если F - истинная гипотеза. Также

$$\alpha = sup_{F \in F_0} W(F)$$
$$\beta = sup_{F \in F_1} 1 - W(F)$$

4.1.3 Вычисление функции отношения правдоподобия

Рассмотрим подробнее теорию по функции отношения правдоподоия и ее применению.

$W(\theta_0, X_{1\alpha}) = P_{\theta_0}(\overline{X} \in X_{1\alpha}) \leq \alpha$

$W(\theta_1, X_{1\alpha}) = P_{\theta_1}(\overline{X} \in X_{1\alpha}) \geq \alpha$

$W(\theta_0, X_{1\alpha}) = \sum_{X_{1\alpha}} L(\overline{X}, \theta_0) \leq \alpha$

$W(\theta_1, X_{1\alpha}) = \sum_{X_{1\alpha}} L(\overline{X}, \theta_1) \geq \alpha$

$l(\overline{X}) = \frac{L(\overline{X}, \theta_0)}{L(\overline{X}, \theta_1)} = \frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_2(x_i)}$

Если $l(\overline{X}) \geq c$, то принимаем гипотезу H_0 , иначе принимаем гипотезу H_1 . Функция $l(\overline{X})$ называется статистикой отношения правдоподобия.

Рассмотрим функцию $\psi(c) = \sum_{\overline{X}: l(\overline{X}) \geq c} L(\overline{X}, \theta_0), \psi(0) = 1$. Заметим, что чем больше c, тем меньше значение $\psi(c)$, то есть заданная функция убывающая.

Теорема Леймана-Пирсона:

Пусть $\forall \alpha \exists c\ \alpha : \psi(c_\alpha) = \alpha$, тогда критическая область $X_{1\alpha}$ наиболее мощный критерий для гипотезы H_0 с уровнем значимости α относительно альтернативы H_1 , среди всех критериев с уровнем значимости α.

Предположим, что нам известно, что обе выборки подчиняются закону распределения Geom(x, θ), но θ - неизвестный параметр.

• $H_0 : F_1 = F_2 = Geom(x, \theta_1)$ - сложная гипотеза, Geom - выбранное дискретное распределение

• $H_1 : F_2 = Geom(x, \theta_2)$

Тогда $l(\overline{X}) =$

Type *Markdown* and LaTeX: α^2

Так как распределение дискретное, для равенства стоит выбрать критерий следующего вида (рандомизированный критерий - наиболее мощный для дискретных распределений):

$\phi^* = 0, \sum x_i < t_{1-\alpha}$

$\phi^* = \frac{\alpha - P(\xi_0 < t_{1-\alpha})}{P(\xi_0 < t_{1-\alpha})}, \sum x_i = t_{1-\alpha}$

$\phi^* = 1, \sum x_i > t_{1-\alpha}$

Или для удобства вычисления:

$\phi^* = 0, \frac{1}{m} \sum x_i < \frac{1}{m} t_{1-\alpha}$

$\phi^* = \frac{\alpha - P(\xi_0 < t_{1-\alpha})}{P(\xi_0 < t_{1-\alpha})}, \frac{1}{m} \sum x_i = \frac{1}{m} t_{1-\alpha}\ \phi^* = 1, \frac{1}{m} \sum x_i > \frac{1}{m} t_{1-\alpha}$

Для вычисления β для каждой выборки можно воспользоваться следующей формулой:

$$\beta = P(\xi_1 \leq t_{1-\alpha}), \xi_1 \approx Geom(\dots)$$

Применим критерий к выборкам каждого распределения объемов 10, 100 и 1000, и найдем для каждой значение ошибки второго рода при α = 0.1, 0.5.

Как видно из результатов обработки сгенерированных выборок, критерий и подобранное значение с окзываются очень точными, в следствии чего достижение равенства в нашем случае не происходит и дискретное распределение ведет себя, как непрерывное с точки зрения построения оптимального критерия.

In [ ]:
```
1
```

Минимальный необходимый объем выборки можно определить из условия $\beta \approx \alpha$

In [ ]:
```
1
```

Беря разный объем выборки, смотрим значение ошибки второго рода и получаем, что для $\alpha = 0.1$ m = ..., а для $\alpha = 0.05$ m = ... Такой объем выборки крайне мал и то, что критерий для них работает, говорит в пользу его оптимальности.

4.2 Распределение Максвелла

4.2.1 Проверка гипотез о виде распределения

Сформулируем гипотезу H_0 и H_1 : Пусть дана выборка $X = (X_1, \ldots, X_n)$ из распределения $L(\xi)$ и F_ξ - неизвестное распределение.

• H_0 : F_ξ = $Maxwell(x, 1.0)$ - простая гипотеза, Maxwell - выбранное непрерывное распределение

• H_1 : не $Maxwell(x, 1.0)$

Для проверки гипотезы H_0 воспользуемся критерием Пирсона (хи-квадрат). Для каждой выборки объемов n = 10, 100, 1000, 100000 найдем значение критерия, границу критического множества для уровня значимости $\alpha = 0.1$ и $\alpha = 0.05$. В каждом случае возьмем N = 15. Разделим каждую выборку на равновероятностные интервалы np.arange(0.01, 1, $\frac{1}{15}$), в случае если взятый интервал меньше, чем 1, он склеивается со следующим.

ЗДЕСЬ ДОЛЖЕН БЫТЬ КОД, ВЫЧИСЛЯЮЩИЙ ЗНАЧЕНИЕ КРИТЕРИЯ, ГРАНИЦЫ ДЛЯ ЗАДАННЫХ УРОВНЕЙ И ОПРЕДЕЛЯЮЩИЙ ДЛЯ КАЖДОЙ ВЫБОРКИ ПРИНИМАЕТСЯ ЛИ ГИПОТЕЗА.

In [58]:
```
pearson_criterion(means__10)
print()
pearson_criterion(means__100)
print()
pearson_criterion(means__1000)
print()
pearson_criterion(means__100000)
```

```
Выборки объема:  10
Выборка:  1
Статистика критерия Пирсона Chi2: 12.948205533236099
Число степеней свободы:  2
Выборка:  2
Статистика критерия Пирсона Chi2: 11.360567891920532
Число степеней свободы:  2
Выборка:  3
Статистика критерия Пирсона Chi2: 7.094003975477621
Число степеней свободы:  2
Выборка:  4
Статистика критерия Пирсона Chi2: 16.794661342915962
Число степеней свободы:  2
Выборка:  5
Статистика критерия Пирсона Chi2: 11.78496970099767
Число степеней свободы:  2

Выборки объема:  100
Выборка:  1
Статистика критерия Пирсона Chi2: 128.36468124889564
Число степеней свободы:  11
Выборка:  2
Статистика критерия Пирсона Chi2: 189.05692009135595
Число степеней свободы:  11
Выборка:  3
Статистика критерия Пирсона Chi2: 178.93265353183912
Число степеней свободы:  10
Выборка:  4
Статистика критерия Пирсона Chi2: 158.92587831338133
Число степеней свободы:  10
Выборка:  5
Статистика критерия Пирсона Chi2: 145.56230697737487
Число степеней свободы:  12

Выборки объема:  1000
Выборка:  1
Статистика критерия Пирсона Chi2: 1365.186737434531
Число степеней свободы:  14
Выборка:  2
Статистика критерия Пирсона Chi2: 1424.212155401219
Число степеней свободы:  15
Выборка:  3
Статистика критерия Пирсона Chi2: 1491.3398510264155
Число степеней свободы:  14
Выборка:  4
Статистика критерия Пирсона Chi2: 1465.854198192775
Число степеней свободы:  14
Выборка:  5
Статистика критерия Пирсона Chi2: 1562.4093704485408
Число степеней свободы:  12

Выборки объема:  100000
Выборка:  1
Статистика критерия Пирсона Chi2: 159925.1065809577
Число степеней свободы:  14
Выборка:  2
Статистика критерия Пирсона Chi2: 156714.40761985185
Число степеней свободы:  15
Выборка:  3
Статистика критерия Пирсона Chi2: 159240.2061818955
Число степеней свободы:  14
```

4.2.2 Проверка параметрических гипотез

4.2.2.1 Выбор данных

По причине конечности данного на решение задачи времени для ускорения завершения данного домашнего задания выберем распределение Максвелла и рассмотрим две выборки с разными (но известными) параметрами. А именно, к уже сконфигурированным выборкам из распределения с параметром λ = 1.0, добавим к рассмотрению выборку из распределения с параметром $\lambda = 1.5$. Сгенерируем новые выборки размеров 10, 100, 1000, 10000 и запишем их в файлы по уже отработанной схеме моделирования.

4.2.3 Вычисление функции отношения правдоподобия

$$l(\overline{X}) = \frac{L(\overline{X},\theta_0)}{L(\overline{X},\theta_1)} = \frac{\prod_{i=1}^{n} f_1(x_i)}{\prod_{i=1}^{n} f_2(x_i)}$$

5 Домашнее задание. Различение гипотез

5.1 Геометрическое распределение

5.1.1 Выбор данных

Необходимо найти два набора данных, соответствующих разным распределениям. Это могут быть оценки к фильмам, статистики игр разных команд и т.п.
В случае наличия сложностей можно выбрать одно из выбранных на первом домашнем задании распределений и рассмотреть две выборки с разными (но известными) параметрами.
В крайнем случае можно рассмотреть распределение Бернулли с разными вероятностями успеха p.

5.1.2 Постановка задачи

Необходимо описать постановку задачи. Что является гипотезой H 0 , что H 1 ? Что такое ошибка первого и второго рода, функция мощности?

5.1.3 Вычисление функции отношения правдоподобия.

Необходимо описать вид функции l X отношения правдоподобия, описать способ ее вычисления.

5.1.4 Вычисление критической области/количества материала

Рассмотрим один из самых сложных вопросов данной контрольной работы — вычисление критической области.
Для оценки ошибок первого и второго рода по материалу или вычислении необходимого материала при фиксированных ошибках необходимо знать распределение статистики в случае верности гипотезы H_0 — $l(X|H_0)$ и в случае верности гипотезы H_1 — $l(X|H_1)$. Для большинства распределений это сделать достаточно сложно.

В случае, если не удается вычислить распределение статистики $l(\overline{X})$ в случае верности разных гипотез, предлагается рассмотреть асимптотический подход к различению гипотез.
Прологарифмировав функцию отношения правдоподобия получим сумму одинаково распределенных независимых случайных величин вида

$$z_i = ln\frac{f_1(X_i)}{f_2(X_i)}$$

Используя Ц.П.Т. можно легко получить распределение статистики $lnl(\overline{X})$ в случае верности каждой из гипотез.
Имея две нормально распределенные случайные величины с разными параметрами задача вычисления ошибок первого/второго рода решается легко, как и вычисление минимально необходимого количества материала для достижения нужных ошибок первого и второго рода.

5.2 Распределение Максвелла

5.2.1 Выбор данных

5.2.2 Постановка задачи

5.2.3 Вычисление функции отношения правдоподобия.

5.2.4 Вычисление критической области/количества материала

6 Литература

[1] "Справочник по вероятностным распределениям" Р.Н.Вадзинский
https://fileskachat.com/view/10838_b741e0be3370efed892ccfe2b6c1358f.html (https://fileskachat.com/view/10838_b741e0be3370efed892ccfe2b6c1358f.html)
[2] "Введение в математическую статистику" (Ивченко Г.И., Медведев Ю.И.)
http://bookre.org/reader?file=1221378&pg=101 (http://bookre.org/reader?file=1221378&pg=101)
[3] Power Maxwell distribution:
https://arxiv.org/pdf/1807.01200.pdf (https://arxiv.org/pdf/1807.01200.pdf)
[4] Geometric Distribution:
https://brilliant.org/wiki/geometric-distribution/ (https://brilliant.org/wiki/geometric-distribution/)
[5]The Maxwell Distribution:
https://randomservices.org/random/special/Maxwell.html (https://randomservices.org/random/special/Maxwell.html)
[6]The Geometric Distribution:
https://randomservices.org/random/bernoulli/Geometric.html (https://randomservices.org/random/bernoulli/Geometric.html)
[7] Sampling from a Normal Distribution
http://bjlkeng.github.io/posts/sampling-from-a-normal-distribution/ (http://bjlkeng.github.io/posts/sampling-from-a-normal-distribution/)
[8] "Моделирование распределений В.В.Некруткин"
https://clck.ru/RHRdy (https://clck.ru/RHRdy)
[9] Сайт со статистикой бейсбола
https://www.scoreboard.com/ru/baseball/usa/mlb-2020/ (https://www.scoreboard.com/ru/baseball/usa/mlb-2020/)

In [ ]:
```
1
```