УДК 51 (071)

Л.П. Мироненко (канд. физ.-мат. наук, доц.)
Донецкий национальный технический университет, Донецк
кафедра высшей математики им. проф. В.В.Пака
E-mail: mironenko.leon@yandex.ua

ГИПЕРГЕОМЕТРИЧЕСКОЕ РАСПРЕДЕЛЕНИЕ И ПРЕДЕЛЬНАЯ ТЕОРЕМА

Целью статьи является вывод предельной формулы гипергеометрического распределения, оценка погрешности асимптотического разложения. При этом, показано, что гипер-распределение переходит в аналог биномиального распределения. Установлены условия перехода, получены соответствующие оценки, и расчеты сравниваются с теорией.
Ключевые слова: *вероятность, распределение, испытания, бином, предельные теоремы, случайность, независимость испытаний, нормировка.*

Введение

В основе многих расчетов компьютерных технологий используются законы и язык теории вероятностей. Особенно используются различные функции распределения, позволяющие оценивать вероятности различных процессов в реальных системах, накопление систематических ошибок, вероятности сбоев и т.д.. Гипергеометрическое распределение, особенно его предельные формы, относится к числу часто используемых в расчетах. Однако расчет систем с большим числом параметров вызывает затруднения даже при использовании компьютеров с мощными ресурсами. Возникает необходимость в приближенных, но эффективных формулах.

Гипергеометрическое распределение (в дальнейшем гипер-распределение) относится к одному из классических распределений теории вероятностей и является пробным камнем для вывода и введения других распределений. Это распределение содержит биномиальные коэффициенты, которые трудно вычислять при больших значениях числа N ($N > 100$) — количества объектов в рассматриваемой системе. Возникает необходимость в предельных теоремах.

1. Определение гипер-распределения и постановка задачи

Напомним классическую схему, как возникает гипергеометрическое распределение. Для этого рассмотрим задачу. *Предположим, что в урне N шаров, среди которых n белых и $N-n$ черных. Наугад из урны взято k шаров. Найти вероятность $P\{\xi = r\}$ случайного события ξ, состоящего в том, что из k наугад взятых шаров, окажется r белых (и, очевидно, что $k-n$ — черных).*

Число всех возможных случаев, наблюдаемых в эксперименте, равно $C_N^k = \dfrac{N!}{k!(N-k)!}$ — число способов, которыми можно выбрать k шаров из партии N шаров, $N! = N\cdot(N-1)\cdot\ldots\cdot 2\cdot 1$. Среди k шаров белых r можно выбрать C_n^r способами и черных $k-r$ шаров C_{N-n}^{k-r} способами. Согласно основному принципу комбинаторики число событий, благоприятствующих событию $\xi = r$ будет равно $C_n^r \cdot C_{N-n}^{k-r}$. Поэтому искомая вероятность равна

$$P\{\xi = r\} = \frac{1}{N}\frac{C_n^r C_{N-n}^{k-r}}{C_N^k},\ 0 \le r \le \min(n,k), \qquad (1)$$

где C_k^r — биномиальные коэффициенты, равные $C_k^r = \dfrac{k!}{r!(k-r)!}$, $1/N$ — нормировочный множитель. Этот набор вероятностей $P\{\xi = r\}$ называется ***гипергеометрическим распределением***.

При решении подобных задач предлагаем использовать схему (Рис.1)

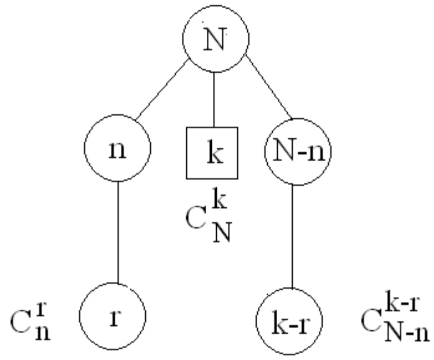

Рисунок 1 — Схема определения гипер-распределения

Условие нормировки вероятности распределения (1) имеет вид

$$\sum_{r=0}^{\min\{k,n\}} P\{\xi = r\} = \frac{1}{N} \sum_{r=0}^{\min\{k,n\}} \frac{C_n^r C_{N-n}^{k-r}}{C_N^k} = 1.$$

Априори, из постановки и решения задачи, приводящей к распределению (1), ясно, что возможны, по крайней мере, два предельных случая $N \to \infty,\ k << N$ — аналог предельной локальной теоремы Муавра-Лапласа в схеме независимых испытаний Бернулли [1]. Второй предельный случай отвечает $N \to \infty$, $k << N$ и $p_N = n/N \to 0$ — аналог предельной теоремы Пуассона [1–3]. В данной работе рассматривается только первый предельный случай, который будем называть *предельной теоремой гипер-распредения.*

2. Предельная теорема гипер-распределения

Если в гипергеометрическом распределении (1) выполняются условия $N >> 1,\ k << N$, *то имеет место формула, аналогичная формуле Бернулли в схеме независимых испытаний*

$$P\{\xi = r\} = C_k^r \left(\frac{n}{N}\right)^r \left(1 - \frac{n}{N}\right)^{k-r},\quad 0 \le r \le \min(n,k). \tag{2}$$

<u>Доказательство</u>. Преобразуем выражение (1), подставив в явном виде биномиальные коэффициенты

$$P\{\xi = r\} = \frac{C_n^r C_{N-n}^{k-r}}{C_N^k} = C_k^r \frac{C_n^r C_{N-n}^{k-r}}{C_k^r C_N^k} = C_k^r \frac{r!(k-r)!}{k!} \cdot \frac{n!}{r!(n-r)!} \cdot \frac{(N-n)!}{(k-r)!(N-n-(k-r))!} \cdot \frac{k!(N-k)!}{N!} =$$

$$= C_k^r \cdot \frac{n!}{(n-r)!} \cdot \frac{(N-n)!}{(N-n-(k-r))!} \cdot \frac{(N-k)!}{N!}.$$

Подставим формулу Стирлинга $n! = \sqrt{2\pi n} n^n e^{-n} e^{\varphi(n)}, \varphi(n) < 1/12n$ [4]. После несложных преобразований, получим

$$P\{\xi = r\} = C_k^r A_N \cdot B_N,$$

$$A_N = \sqrt{\frac{n}{(n-r)} \frac{(N-n)}{(N-n-(k-r))} \frac{(N-k)}{N}}, \tag{3}$$

$$B_N = \frac{n^n}{(n-r)^{(n-r)}} \cdot \frac{(N-n)^{(N-n)}}{(N-n-(k-r))^{(N-n-(k-r))}} \cdot \frac{(N-k)^{(N-k)}}{N^N}. \tag{4}$$

Преобразуем выражение A_N, разлагая его по степеням $1/N$ (ПРИЛОЖЕНИЕ 1 (П2))

$$A_N = \sqrt{1+\delta_{1/N}+\delta_{1/N^2}},\ \delta_{1/N} = \frac{rq_N}{Np_N},\ \delta_{1/N^2} = \delta_{1/N}^2 + \frac{(n-r)(k-r)}{N^2}, \qquad (5)$$

где $p_N = n/N,\ q_N = 1-n/N.$

Преобразуем B_N (ПРИЛОЖЕНИЕ 2 (П6))

$$B_N = p_N^r q_N^{k-r}\left(1+\Delta_{1/N}+\Delta_{1/N^2}\right),$$
$$\Delta_{1/N} = \frac{k^2 p_N q_N - (k-r)^2 p_N - r^2 q_N}{2Np_N q_N}, \qquad (6)$$
$$\Delta_{1/N^2} = \frac{(k^2 p_N q_N - (k-r)^2 p_N - r^2 q_N)^2}{4N^2 p_N^2 q_N^2} + \frac{k^3 p_N^2 q_N^2 - (k-r)^3 p_N^2 - r^3 q_N^2}{6N^2 p_N^2 q_N^2}.$$

Поскольку $\max r = k$, то имеют место оценки

$$\Delta_{1/N} \le \frac{k^2 q_N}{2Np_N} = \frac{k}{2}\varepsilon,\ \Delta_{1/N^2} \le \frac{k\left(2p_N+5\right)}{12}\varepsilon^2,\ \varepsilon = \frac{kq_N}{Np_N} << 1.$$

При $N >> 1, k << N$ гипер-распределение (1) переходит в аналог формулы Бернулли (2) с вероятностями (частотами) $p_N = n/N,\ q_N = 1-n/N.$. Точность приближения порядка $1/N$. Поправки по степеням $1/N^m,\ m = 1,2,3$ приведены в приложениях 1 и 2 и формулах (5) и (6).

Распределение (2) связано с распределением Бернулли следующим образом. Если существует предел $p = \lim\limits_{N\to\infty} n/N$, то существует предел $B = \lim\limits_{N\to\infty} B_N = p^r q^{k-r}$ и распределение (1) и его предельная форма (2) переходят в биномиальный закон распределения (распределение Бернулли в схеме независимых испытаний Бернулли)

$$P\{\xi = r\} = C_k^r p^r q^{k-r} \qquad (7)$$

с вероятностями p и q.

3. Альтернативный вывод предельной теоремы

В схеме независимых испытаний Бернулли рассматриваются независимые испытания с с двумя исходами — успех У с вероятностью p или неудача Н с вероятностью q, $p+q=1$ (термины «успех» и «неудача» употребляется условно). Вероятности p и q определены условиями задачи.

Схему Бернулли можно перенести на задачу, приводящую к гипер-распределению, если вероятности p и q заменить частотами $\nu_N = n/N$ — вытащить белый шар, а черный $1-\nu_N = 1-n/N$. Очевидно, что обе теории совпадают, если существует конечный предел $\lim\limits_{N\to\infty} \nu_N$ и этот предел равен p, $0 \le p \le 1$.

Предположим, что произведено k испытаний, в которых зафиксировано число $\xi = r$ успехов. Например, при $k=2$ пространство элементарных событий $\Omega = \{УУ, УН, НУ, НН\}$. Поскольку испытания независимые, то частоты элементарных исходов следует определить так: $УУ - \left(\nu_N\right)^2$; $УН - \nu_N\left(1-\nu_N\right),\ НУ - \left(1-\nu_N\right)\nu_N,\ НН - \left(1-\nu_N\right)^2$ [5–6]. В общем случае, пространство Ω состоит из 2^k наборов длины k, состоящих из букв *У* и *Н*. частота отдельного набора длины k равна $\left(\nu_N\right)^r\left(1-\nu_N\right)^{k-r}$, если в набор входит r букв *У*; число наборов C_k^r. Поэтому

$$P_n(r) = P\{\xi = r\} = C_k^r\left(\nu_N\right)^r\left(1-\nu_N\right)^{k-r},\ r = 0,1,2,\dots,k,$$

что совпадает с формулой (2).

Данный вывод формулы (2) является строгим, потому, что приведен для частот, в которых уже заложено приближение, которое определено формулами (3) – (5). Более точные расчеты приведены в предыдущем пункте в доказательстве теоремы.

4. Сравнение теории с расчетами

Ниже приведены расчеты по точной формуле распределения (1) и по предельной формуле (2) при $N = 100 >> 1$, $k = 10 << N$ в широком спектре изменения числа $10 \le n \le 90$ (Рис. 2). На следующем рисунке (Рис. 3) показана «динамика» приближения кривых 1 и 2 с увеличением объема выборки $10 \le k \le 50$.

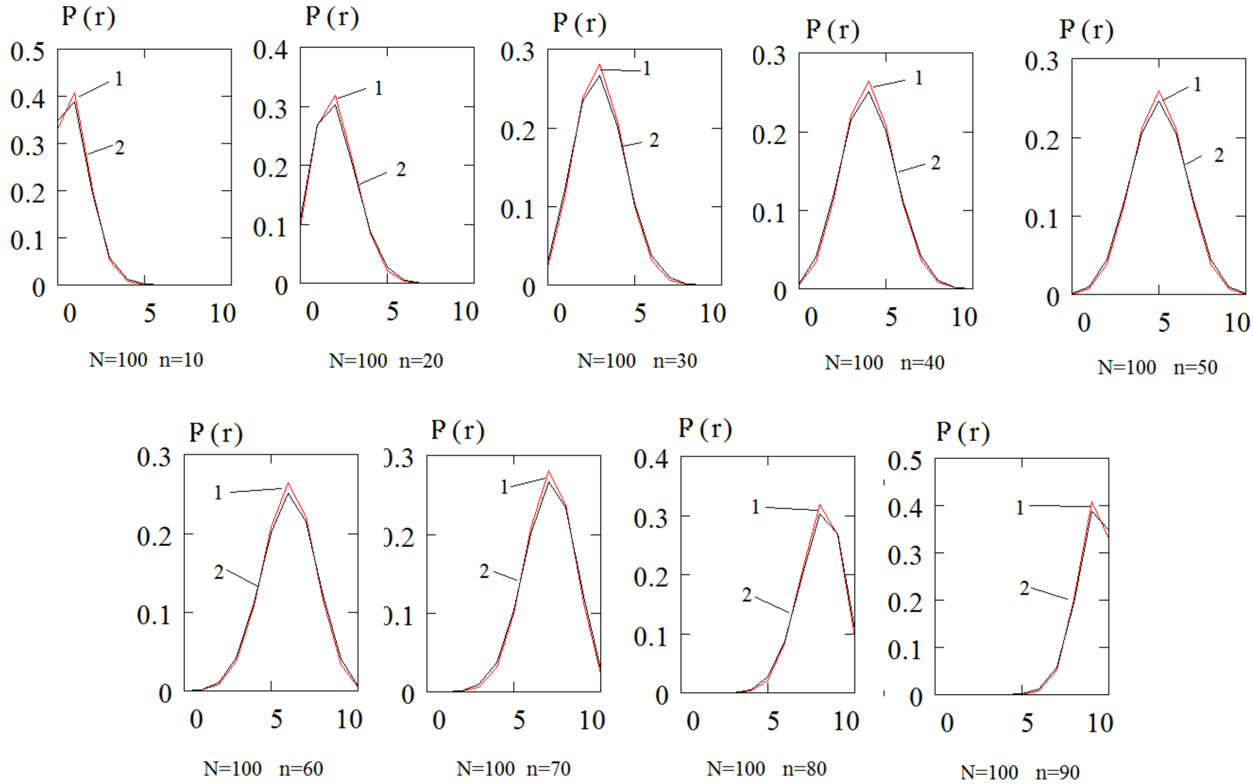

Рисунок 2 — Сравнение расчетов по формуле гипер-распределения (1) (кривые обозначены 1) и по асимптотической формуле (2) (кривые обозначены 2) при условии $k << N$

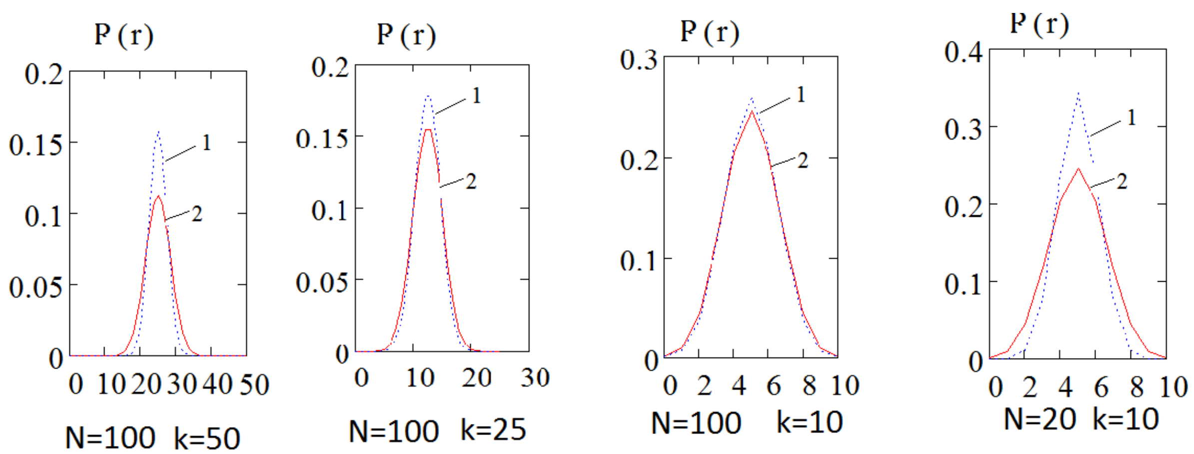

Рисунок 3 — Сравнение расчетов по формуле гипер-распределения (1) (кривые 1) и по асимптотической формуле (2) (кривые 2) при различных соотношениях между k и N, $n/N = 0.5$.

Как видно из графиков, полученная формула (2) хорошо согласуется с формулой (1) при $k << N > 25$ в широком спектре изменения числа $10 \le n \le 90$. В самом деле, нижняя граница N, как показывают расчеты, меньше, порядка $N \approx 20$. При $N < 10$ лучше формулой (2) не пользоваться.

ПРИЛОЖЕНИЕ 1 Разложение функции A_N по степеням $1/N$.

Вынесем в выражении (3) для A_N в числителе и знаменателе N за скобки и используем обозначение $n = Np_N$

$$A_N = \sqrt{\left(1-\frac{n}{N}\right)\left(1-\frac{k}{N}\right)\left(1-\frac{r}{n}\right)^{-1}\left(1-\frac{n}{N}-\frac{k-r}{N}\right)^{-1}} = \sqrt{\left(1-\frac{n}{N}\right)\left(1-\frac{k}{N}\right)\left(1-\frac{r}{Np}\right)^{-1}\left(1-\frac{n+k-r}{N}\right)^{-1}}. \tag{П1}$$

Произведем разложение этой функции по степеням $1/N$ до членов порядка $1/N^2$

$$\left(1-\frac{n}{N}\right)\left(1-\frac{k}{N}\right) = 1-\frac{n}{N}-\frac{k}{N}+\frac{nk}{N^2}, \quad \left(1-\frac{r}{Np}\right)^{-1} = 1+\frac{r}{Np}+\frac{r^2}{N^2p^2},$$

$$\left(1-\frac{n+k-r}{N}\right)^{-1} = 1+\frac{n+k-r}{N}+\frac{(n+k-r)^2}{N^2}.$$

Здесь использовано разложение $(1+x)^\alpha = 1+\alpha x+\frac{\alpha(\alpha-1)}{2!}x^2+...,$ при $\alpha=-1$ имеем $(1+x)^{-1} = 1-x+x^2+....$ тогда, выражение (П1) для A_N^2 имеет вид

$$A_N^2 = \left(1-\frac{n+k}{N}+\frac{nk}{N^2}\right)\left(1+\frac{r}{Np}+\frac{r^2}{N^2p^2}\right)\left(1+\frac{n+k-r}{N}+\frac{(n+k-r)^2}{N^2}\right) =$$

$$= 1+\frac{r(1-p)}{Np}+\frac{r^2(1-p)}{N^2p^2}+\frac{(n-r)(k-r)}{N^2}.$$

После простых преобразований, получим

$$A_N = \sqrt{1+\delta_{1/N}+\delta_{1/N^2}},$$

$$\delta_{1/N} = \frac{r(1-p_N)}{Np_N}, \quad \delta_{1/N^2} = \frac{r^2(1-p_N)}{N^2p_N^2}+\frac{(n-r)(k-r)}{N^2}. \tag{П2}$$

Проверим предельный случай $p_N = 1 \to n = N$, тогда $r = k$ и $\delta_{1/N} =_{1/N^2} = 0$. В другом предельном случае $p_N = 0 \to n = 0 \Rightarrow r = 0$ также $\delta_{1/N} =_{1/N^2} = 0$.

Поскольку $\max r = k$, то имеют место оценки $\delta_{1/N} \le \frac{kq_N}{Np_N} = \varepsilon, \ \delta_{1/N^2} \le \varepsilon^2.$

ПРИЛОЖЕНИЕ 2. Разложение функции B_N по степеням $1/N$

Подставим в функцию B_N (4) выражение

$$\frac{n^n}{(n-r)^{(n-r)}} = n^r\left(1-\frac{r}{n}\right)^{-(n-r)}, \tag{П3}$$

и аналогичные выражения в (4) для подобных членов в функции B_N, получим

$$B_N = \frac{n^r(N-n)^{(k-r)}}{N^k}\left(1-\frac{r}{n}\right)^{-(n-r)}\left(1-\frac{k-r}{N-n}\right)^{-(N-n-(k-r))}\left(1-\frac{k}{N}\right)^{(N-k)}.$$

Еще раз применим равенство (П3) к первому множителю в выражении B_N, получим

$$B_N = \left(\frac{n}{N}\right)^r \left(1-\frac{n}{N}\right)^{(k-r)} \left(1-\frac{r}{n}\right)^{-(n-r)} \left(1-\frac{k-r}{N-n}\right)^{-(N-n-(k-r))} \left(1-\frac{k}{N}\right)^{(N-k)}. \qquad (\text{П4})$$

Рассмотрим первый член и используем стандартное разложение $\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \ldots$

$$\left(1-\frac{r}{n}\right)^{-(n-r)} = \exp\ln\left(1-\frac{r}{n}\right)^{-(n-r)} = -(n-r)\ln\left(1-\frac{r}{n}\right) =$$

$$= -(n-r)\left(-\frac{r}{n} - \frac{r^2}{2n^2} - \frac{r^3}{3n^3}\right) = r - \frac{r^2}{2n} - \frac{r^3}{6n^2} - \frac{r^4}{3n^3}.$$

Аналогично,

$$\left(1-\frac{k-r}{N-n}\right)^{-(N-n-(k-r))} = k - r - \frac{(k-r)^2}{2(N-n)} - \frac{(k-r)^3}{6(N-n)^2} - \frac{(k-r)^4}{3(N-n)^3},$$

,

$$\left(1-\frac{k}{N}\right)^{(N-k)} = -k + \frac{k^2}{2N} + \frac{k^3}{6N^2} + \frac{k^4}{3N^3}.$$

Подставим эти выражения в B_N, получим

$$B = p_N^r q_N^{(k-r)} \exp\{\Delta_{1/N} + \Delta_{1/N^2} + \Delta_{1/N^3}\}, \qquad (\text{П5})$$

$$\Delta_{1/N} = \frac{k^2 p_N q_N - r^2 q_N - (k-r)^2 p_N}{2Np_N q_N},$$

$$\Delta_{1/N^2} = \frac{k^3 p_N^2 q_N^2 - r^3 q_N^2 - (k-r)^3 p_N^2}{6N^2 p_N^2 q_N^2}, \qquad (\text{П6})$$

$$\Delta_{1/N^3} = \frac{k^4 p_N^3 q_N^3 - r^4 q_N^3 - (k-r)^4 p_N^3}{3N^3 p_N^3 q_N^3}.$$

Проверим предельный случай $p_N = 1 \to n = N$, тогда $r = k$ и $\Delta_{1/N} = \Delta_{1/N^2} = \Delta_{1/N^3} = 0$. В другом предельном случае $p_N = 0 \to n = 0 \Rightarrow r = 0$ также $\Delta_{1/N} = \Delta_{1/N^2} = \Delta_{1/N^3} = 0$.

Поскольку $\max r = k$, то имеют место оценки

$$\Delta_{1/N} \le \frac{1}{2}\frac{k^2 q_N}{Np_N} = \frac{k}{2}\varepsilon,\ \Delta_{1/N^2} \le \frac{k}{6}\varepsilon^2,\ \Delta_{1/N^3} \le \frac{k}{3}\varepsilon^3,\ \varepsilon = \frac{kq_N}{Np_N} << 1.$$

Для оценки погрешности используем стандартное разложение

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots,\ x = \Delta_1 + \Delta_2,$$

получим Δ_2 в виде (6).

В завершении приведем оценку погрешности формулы (2), ограничиваясь членами $1/N$

$$P\{\xi = r\} = C_k^r \left(\frac{n}{N}\right)^r \left(1-\frac{n}{N}\right)^{k-r} (1+\delta),$$

$$\delta = \Delta_{1/N} + \delta_{1/N} = \frac{k^2 pq - r^2 q - (k-r)^2 p}{2Npq} + \frac{r(1-p)}{Np},$$

или

$$\delta \le \frac{k(k+2)q}{2Np}$$

Выводы

1. Получена предельная (асимптотическая) формула (2) гипергеометрического распределения (1), которая может быть использована при $N > 20$, $k << N$, давая абсолютную погрешность порядка 5%.
2. Предложен простой и строгий вывод предельной теоремы (2), основанный на схеме независимых испытаний Бернулли с использованием частоты появления событий. Несмотря на различия в постановке задач гипер-распределения и схемы Бернулли, оба распределения при определенных условиях имеют сходные черты и взаимозаменяемость.
3. Приведена оценка асимптотического разложения формулы (1), тем самым установлены границы применимости формулы (2). В расчетах не учтена погрешность, которую вносит приближенная формула Стирлинга. Это сделано намеренно с целью выделения поправки, обусловленной случайной величиной.
4. Полученная формула имеет перспективы, поскольку открывают возможность легкого получения предельных теорем теории вероятностей: локальную теорему Муавра-Лапласа и формулу Пуассона. Эти предельные теоремы не будут в точности совпадать с оригиналами, будут отличаться, но их гораздо проще получить из формулы (2), чем из исходного распределения (1).

Список использованной литературы

1. Колмогоров А.Н. Основные понятия теории вероятности / А.Н. Колмогоров. – М: Наука, 1974 — 120 с.
2. Ширяев А.Н. Вероятность / А.Н. Ширяев. — М: Изд. МГУ, 1979. — 575 с.
3. Феллер В. Введение в теорию вероятностей и ее приложения / В. Феллер. — М: Мир, 1964. – Т.1. — 492 с.
4. Брычков Ю.А. Интегралы и ряды. Элементарные функции / Ю.А. Брычков, А.П. Прудников, О.И. Маричев. — М: Наука, 1981. — 800 с.
5. Венцель Е.С. Теория вероятностей / Е.С. Венцель, Л.А. Овчаров. — М: Наука, 1969. — 368 с.
6. Кац М. Статистическая независимость в теории вероятностей / М. Кац. — М: ФМЛ, 1962. — 153 с.

Надійшла до редакції: 30.01.2012 р.

Рецензент: д-р физ.-мат.наук, проф. Малашенко В.В.

L.P. Mironenko. The hipher Geometrical Distribution and the Limiting Theorem. *The purpose of the paper is to obtain an asymptotic formula of the hipher geometrical distribution, an estimation of a value of a mistake of the asymptotic expansion. It is shown that the hipher-distribution transits to a formula like Bernoulli's distribution. We defined conditions of the transition to Bernoulli's formula. The theory is compared with the computations.*

Keywords: *Probability, distribution, trials, binomial, limiting theorems, causality, independent trial, standardized variable, normalization*

Л.П. Мироненко. Гіпергеометричний розподіл та гранична теорема. *Ціллю статті є отримання граничної формули гіпергеометричного розподілу, оцінка погрішності асимптотичного розподілу. При цьому, встановлено, що гіпер-розподіл переходить в аналог біноміального розподілу. Встановлені умови переходу, отримані відповідні оцінки і розрахунки порівняно з теорією.*

Ключові слова: *ймовірність, розподіл, випробування, біном, граничні теореми, випадок, незалежність випробувань, норміровка.*