МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

УТВЕРЖДАЮ
Проректор-директор ИНК
_____В.А. Клименов
«___» _____ 2013г.

Лабораторная работа №3
Оценки параметров распределения

Томск - 2013

Оценки параметров распределения

Цель работы: Рассмотреть возможности MathCAD для решения задач теории оценивания. Для предполагаемого закона распределения получить значения точечных и интервальных оценок параметров.

1. Точечная оценка

Точечная оценка предполагает нахождение единственной числовой величины, которая и принимается за значение параметра.

Выборочное среднее значение $\bar{x}_в$ является оценкой математического ожидания генеральной совокупности.

$$\bar{x}_в = \frac{x_1 + x_2 + ... + x_n}{n}.$$

Исправленная дисперсия S^2 является несмещенной оценкой дисперсии генеральной совокупности:

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_в)^2}{n-1}$$

Способы получения точечной оценки является метод моментов и метод максимального правдоподобия.

1.1 Метод максимального правдоподобия

Наиболее вероятным значением плотность распределения $f(x, \Theta)$ оцениваемых величин будут такие, при которых совместная плотность распределения выборки, называемая *функцией правдоподобия,* достигает максимума.

$$L = \prod_{i=1}^{n} f(x_i, \tilde{\Theta}).$$

Для нахождения оценок необходимо исследовать функцию правдоподобия на экстремум. С целью упрощения вычислительной процедуры логарифм функции правдоподобия $\ln L(x, \tilde{\Theta})$. Нахождение оценок сводится к решению нелинейного уравнения (уравнений) относительного искомого параметра (параметров) вида

$$\frac{d \ln L(x, \tilde{\Theta})}{d\tilde{\Theta}} = 0.$$

1.2 Метод моментов

Сущность метода состоит в приравнивании теоретических моментов рассматриваемого распределения соответствующим эмпирическим моментам того же порядка и решении полученной системы уравнений относительно неизвестных параметров распределения. Выбирается столько эмпирических моментов, сколько требуется оценить неизвестных параметров распределения.

$$\bar{\alpha}_k = \frac{1}{n} \sum_{i=1}^{n} x_i^{\ k},$$

$$\tilde{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} x_i - \bar{x}^{\ k}.$$

Теоретическими моментами называются моменты, вычисленные аналитическим путем через плотность *f(x)* или функцию распределения *F(x)***.** Теоретические моменты являются функциями параметров распределения:

$$\alpha_k = \alpha_k \ \Theta_1, \Theta_2, ..., \Theta_n \ ,$$

$$\mu_k = \mu_k \ \Theta_1, \Theta_2, ..., \Theta_n \ .$$

Оценки параметров распределения

Закон распределения имеет один параметр, то приравняем, например, начальный теоретический момент первого порядка начальному эмпирическому $\alpha_1 = M_1$. Учитывая, что $\alpha_1 = M(X)$ и $M_1 = \bar{x}_в$, получим

$$M(X) = \bar{x}_в .$$

Математическое ожидание M (X), как видно из соотношения

$$M(X) = \int_{-\infty}^{\infty} xf(x;\Theta)dx = \varphi(\Theta),$$

есть функция от θ, поэтому $M(X) = \bar{x}_в$.можно рассматривать как уравнение с одним неизвестным θ. Решив это уравнение относительно параметра θ, тем самым найдем его точечную оценку $\overline{\Theta}$, которая является функцией от выборочной средней, следовательно, и от вариант выборки.

2. Интервальная оценка

Так как оценки являются случайными величинами, то их отклонения от оцениваемых параметров также случайны. Чтобы дать представление о точности и надежности оценки в математической статистике, пользуются так называемыми доверительными интервалами и доверительными вероятностями.

Доверительным интервалом называется интервал $[\Theta_{i1};\Theta_{i2}]$, которой с заданной *доверительной вероятностью* $P = 1 - \alpha$ накрывает истинное значение x_i, т.е.

$$P\ \Theta_{i1} < \Theta_i < \Theta_{i2}\ = 1-\alpha.$$

Чем уже интервал $[\Theta_{i1};\Theta_{i2}]$, тем точнее оценка неизвестного параметра Θ_i. В измерительной практике доверительная вероятность обычно принимается равной 0,90; 0,95; 0,98; 0,99 и реже (в особо ответственных случаях) 0,999.

Интервальная оценка имеет вид

$$\overline{\Theta} - t_p\sigma_{\overline{\Theta}} \le \Theta \le \overline{\Theta} + t_p\sigma_{\overline{\Theta}}.$$

Так в нашем распоряжении имеется только выборка, поэтому дисперсия генеральной совокупности неизвестна. Поэтому в данном случае интервальная оценка строится на основе распределения Стьюдента:

$$\overline{\Theta} - t_p S_{\overline{\Theta}} \le \Theta \le \overline{\Theta} - t_p S_{\overline{\Theta}}$$

где t_p — коэффициент распределения Стьюдента для двусторонней доверительной вероятности *P*. В MathCad коэффициент распределения Стьюдента можно найти как значение квантили соответствующего распределения qt(p,n), p – доверительная вероятность, n – число измерений (объем выборки).

3. Решение системы уравнений

Mathcad дает возможность решать системы уравнений.

Для решения системы уравнений необходимо:

- Задать начальные приближения для всех неизвестных, входящих в систему уравнений.
- Напечатать ключевое слово *Given*. Оно указывает Mathcad, что далее следует система уравнений. При печати слова Given можно использовать любой шрифт, прописные и строчные буквы.
- Ввести уравнения и неравенства в любом порядке ниже ключевого слова *Given*. Удостоверьтесь, что между левыми и правыми частями уравнений стоит символ =

(сравнение)**.** Используйте [Ctrl]= для печати символа =. Между левыми и правыми частями неравенств может стоять любой из символов <, >.

- Ввести выражение, которое включает функцию *Find*. При печати слова *Find* можно использовать шрифт любого размера, произвольный стиль, прописные и строчные буквы. *Find(z1, z2, z3, . . .*) Возвращает решение системы уравнений. Число аргументов должно быть равно числу неизвестных.

Задание 1. Используя имеющуюся выборку неизвестного закона распределения, предположить закон распределения и получить точечные и интервальные оценки его параметров.

Порядок выполнения работы.

1. С помощью функции READPRN (“file”) считать данные выборки из файла с номером, соответствующему порядковому номеру компьютера. Файлы находятся в папке «ЛБ МОРИ/ lb3» на рабочем столе ПК.
2. С помощью формулы Стерджеса (Приложение 1) определить оптимальное число интервалов для построения статистического ряда распределения.
3. Построить нормированную гистограмму любым удобным Вам способом (Приложение 1).
4. По внешнему виду гистограммы предположить неизвестный закон распределения непрерывной случайной величины. Виды и формулы законов распределения приведены в Приложении 2.
5. По выбранному закону распределения определиться с параметрами распределения.
6. Оценить параметры распределения методом максимального правдоподобия, используя возможности Mathcad для решения систем уравнений.
7. Методом моментов оценить значения параметров распределения.
    a. Рассчитать эмпирические моменты.
    b. Рассчитать моменты через параметры распределения. (Можно воспользоваться данными приложения 1), используя возможности Mathcad для решения систем уравнений.
8. Сравнить результаты, полученные разными методами.
9. Построить на графике с гистограммой плотность распределения для данной выборки на основе полученных оценок параметров распределения.
10. Сделать выводы по графикам.
11. Получить интервальные оценки для параметров распределения с доверительной вероятностью P=0,95 и P=0,99.
12. Изобразить полученные квантили на предыдущем рисунке. Сделать вывод.

В отчете необходимо указать файл выборки, предполагаемый закон распределения, оценки его параметров (точечные и интервальные), привести необходимые графики и выводы.

Литература.

Методические указания к практическим занятиям практическим занятиям по курсу «Теория вероятности и математическая статистика» / Сост.: Б.Н. Воронков, В.А Голуб, Т.М. Жукова; Воронежский гос. ун-т. – Воронеж, 1997. – 32с.

Приложение 1

Некоторые встроенные функции Mathcad

- mean(x) — выборочное среднее значение;
- max(x), min(x) — максимальное и минимальное значения выборки;
- var(x) ,stdev(x) — выборочная дисперсия и среднеквадратичное отклонение в другой нормировке;
- round(x, n) — округление числа *z* с точностью до *n* знаков после запятой;
- READPRN ("file") – чтение данных в матрицу из текстового файла;
- WRITEPRN ("file") – запись данных в текстовый файл;
- APPENDPRN ("file") – дозапись данных в существующий текстовый файл;
  • file — путь к файлу.
- CWD := "D:\tmp\" устанавливается текущий рабочий каталог.

Формула Стэрджеса

Для определения оптимального числа интервалов *l* можно использовать формулу Стэрджеса:

$$l = 1 + 3{,}322 \cdot \lg n$$

Построение гистограмм

Гистограмма с произвольными сегментами разбиения

hist,(int, x) – вектор частоты попадания данных в интервалы гистограммы;

- *int* – вектор, элементы которого задают сегменты построения гистограммы в порядке возрастания $a \le intvis_i < b$;
- *x* – вектор случайных данных.

Если вектор *int* имеет *bin* элементов, то и результат *hist* имеет столько же элементов.

Для того, чтобы построить гистограмму, нужно сначала сгруппировать выборочные данные, записанные в массиве *x,* и сохранить граничные очки интервалов группировки в векторе *int,* размерность которого равна числу интервалов

Пример 1. Построение гистограммы

N:=1000

bin:=30 ; кол-во равных сегментов, на кот. разбивается весь диапазон

x:= rnorm (N, 0, 1)

; определение границы интервала построения гистограммы

lower:= floor (min (x)) ; наибольшее цело число ≤ min(x)

upper:= ceil (max (x)) ; наименьшее цело число ≥ max(x)

$h := \frac{upper - lower}{bin}$; размер сегмента

j: = 0 .. bin

int_j *: = lower + h • j* ; массив начальных точек каждого сегмента

$f := \frac{1}{N \cdot h} \cdot hist(\text{int}, x)$

; нормирование гистограммы для удобства отображения на одном графике вместе с плотностью распределения

Обратите внимание, что в последней строке листинга осуществлена нормировка значений гистограммы, с тем, чтобы она правильно аппроксимировал плотность вероятности, также показанную на графике. Очень важно переопределение вектора ***int*** в

которое необходимо для перехода от левой границы каждого элементарного сегмента к его центру.

int := int + 0.5h

Гистограмма с разбиением на равные сегменты

histogram (bin, x) — матрица гистограммы размера bin * 2, состоящая из столбца сегментов разбиения и столбца частоты попадания в них данных;

- *bin — количество сегментов построения гистограммы;*
- *x — вектор* случайных данных.

Пример 2. Построение гистограммы (упрощенный вариант)

```
N:=100
bin:=30                    ; кол-во равных сегментов, на кот. разбивается весь
                           диапазон
x:= rnorm (N, 0, 1)
f := histogram(bin, x)
```

Приложение 2

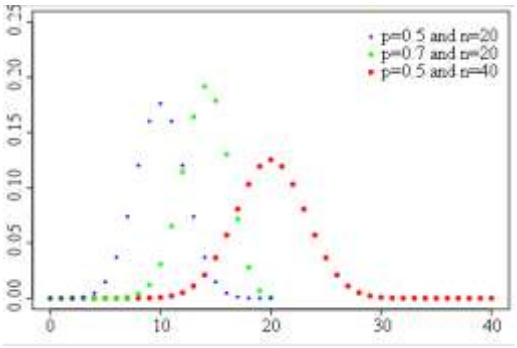
Законы распределения дискретных случайных величин

Биномиальное распределение (схема Бернулли). Пусть проводится серия из n независимых испытаний, каждое из которых заканчивается либо «успехом», либо «неуспехом». Пусть в каждом испытании (опыте) вероятность успеха *p*, а вероятность неудачи – $q = 1 - p$ С таким испытанием можно связать случайную величину *x*, равную числу успехов в серии из *n* испытаний. Эта величина принимает целые значения от 0 до *n*.

Ее распределение называется биномиальным и определяется формулой Бернулли

$$p_k = P(\xi = k) = C_n^k p^k q^{n-k},$$

где $0 < p < 1$, $q = 1 - p$, $k = 0, 1, \ldots, n$, $C_n^k = \dfrac{n!}{k!(n-k)!}$



Функция вероятности

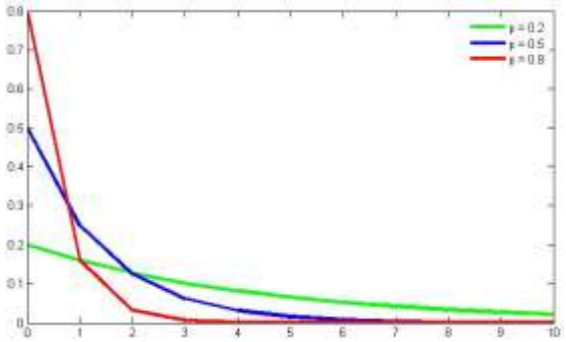
Основные характеристики биноминального распределения примут вид:

$$M(X) = np,\ D(X) = npq,\ A = \frac{q-p}{\sqrt{npq}},\ E = \frac{1-6pq}{npq}.$$

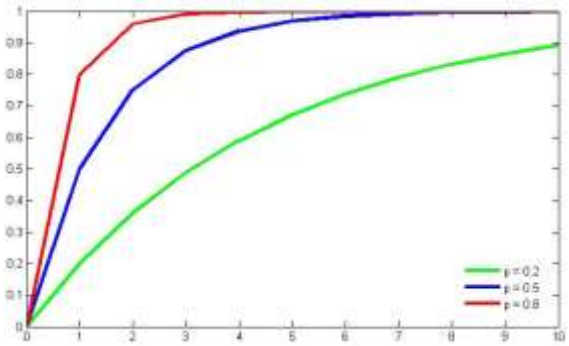
Геометрическое распределение. Со схемой испытаний Бернулли можно связать еще одну случайную величину: *h* – число испытаний до первого успеха. Эта величина принимает бесконечное множество значений от 0 до +∞, и ее распределение определяется формулой

$$p_k = P(\eta = k) = p^k q$$

где $0 < p < 1$, $q = 1 - p$, $k = 0, 1, \ldots, n$,



Функция вероятности



Функция распределения

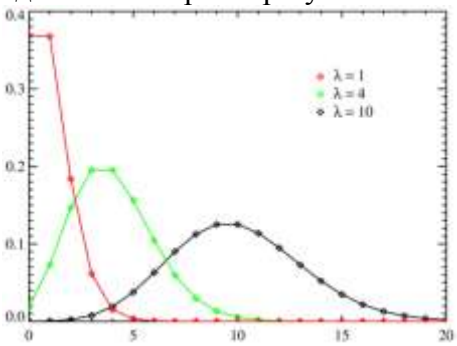
Основные характеристики геометрического распределения примут вид:

$$M(X) = \frac{1-p}{p},\ D(X) = \frac{q}{p^2},\ A = \frac{2-p}{\sqrt{1-p}},\ E = 6 + \frac{p^2}{1-p}.$$

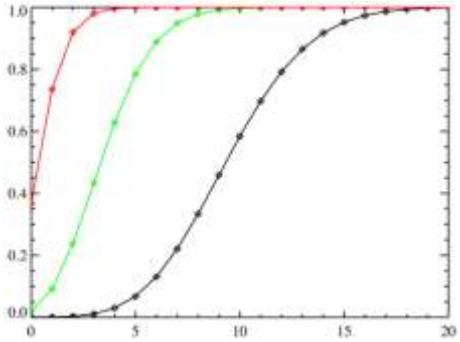
Пуассоновское распределение. Пуассоновское распределение имеет случайная величина *m*, принимающая значения *k = 0, 1, 2,…* с вероятностями

$$p_k = P(\mu = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ \text{k= 0, 1, 2, …,}$$

где $\lambda > 0$ – параметр пуассоновского распределения.



Функция вероятности



Функция распределения

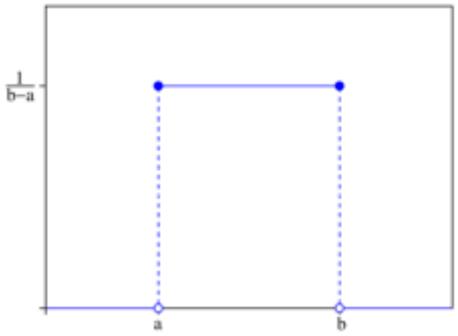
Основные характеристики пуассоновского распределения примут вид:

$$M(X) = \lambda,\ D(X) = \lambda,\ A = \lambda^{-1/2},\ E = \lambda^{-1}.$$

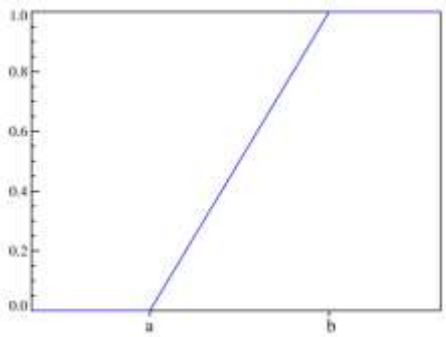
Законы распределения непрерывных случайных величин

Равномерное распределение. Непрерывная случайная величина ξ, принимающая значение на отрезке *[a,b]*, распределена равномерно на *[a,b]*, если плотность распределения $p_\xi(x)$ и функция распределения случайной величины ξ имеют соответственно вид

$$p_\xi(x) = \begin{cases} 0, & x \notin a,b \\ \dfrac{1}{b-a}, & x \in a,b \end{cases} \qquad F_\xi(x) = \begin{cases} 0, & x \le a \\ \dfrac{x-a}{b-a} & a < x \le b \\ 1 & x > b \end{cases}$$



Функция вероятности



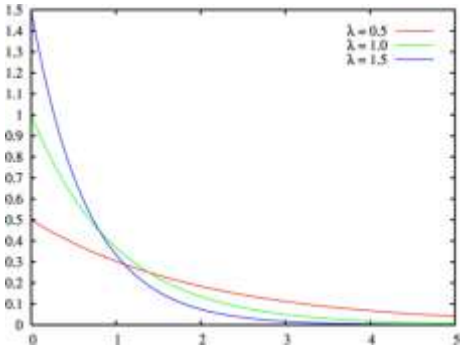
Функция распределения

Основные характеристики равномерного распределения примут вид:

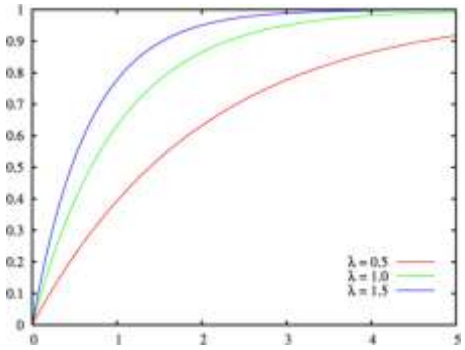
$$M(X) = \frac{a+b}{2},\ D(X) = \frac{(b-a)^2}{12},\ A = 0,\ E = -\frac{6}{5}.$$

Экспоненциальное (показательное) распределение. Непрерывная случайная величина ξ имеет показательное распределение с параметром $\lambda > 0$, если плотность распределения имеет вид

$$p_\xi(x) = \begin{cases} 0, & x < 0 \\ \lambda e^{-\lambda x}, & x \geq 0 \end{cases} \qquad F_\xi(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \lambda e^{-\lambda x}, & x > 0 \end{cases}$$



Функция вероятности



Функция распределения

Основные характеристики экспоненциального распределения примут вид:

$$M(X) = \lambda^{-1},\ D(X) = \lambda^{-2},\ A = 2,\ E = 6.$$

Нормальное распределение. Это распределение играет исключительно важную роль в теории вероятностей и математической статистике. Случайная величина ξ нормально распределена с параметрами μ и σ, $\sigma > 0$, если её плотность распределения имеет вид

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right).$$

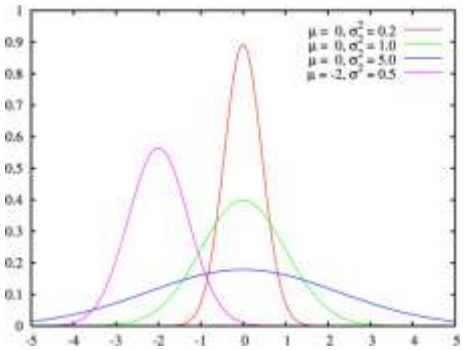
Если случайная величина ξ имеет нормальное распределение с параметрами a и σ, то будем записывать это в виде $\xi \sim N(a,\sigma)$. Случайная величина ξ имеет стандартное нормальное распределение, если $\mu = 0$ и $\sigma = 1$, $\xi \sim N(0,1)$. Плотность стандартного нормального распределения имеет вид

$$p_\xi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2} \right),$$

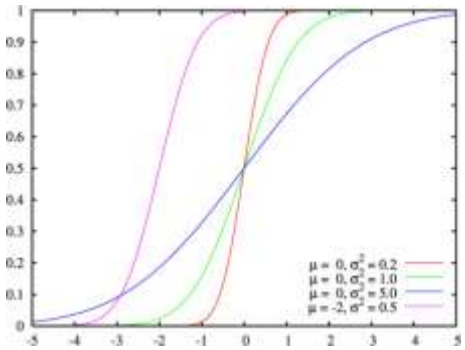
а его функция распределения $F_\xi(x) = \Phi(x)$, где $\Phi(x)$ – функция Лапласа:

$$\Phi_\xi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{z^2}{2}) dz .$$

Функция распределения нормальной величины η ~ N(μ,σ) также выражается через функцию Лапласа: $F_\eta(x) = \Phi\left(\frac{x-\mu}{\sigma} \right)$.



Функция вероятности



Функция распределения

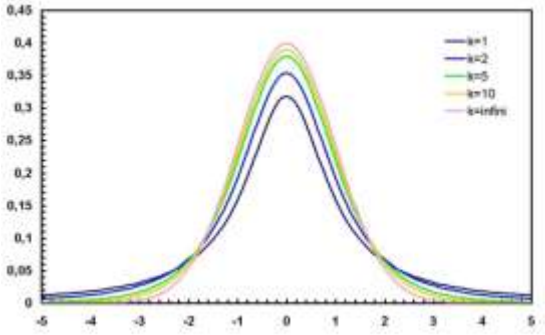
Основные характеристики нормального распределения примут вид:

$$M(X) = \mu,\ D(X) = \sigma^2,\ A = 0,\ E = 0.$$

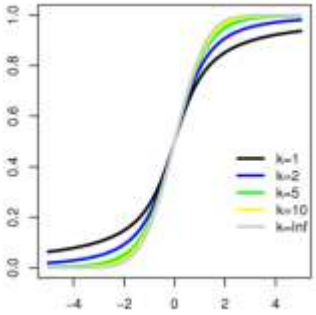
Распределение Стьюдента. Пусть случайная величина ξ имеет стандартное нормальное распределение, а случайная величина $\chi_n^2 - \chi^2$ – распределение с n степенями свободы. Если ξ и χ_n^2 независимы, то про случайную величину $\tau_n = \frac{\xi}{\sqrt{\chi_n^2 / n}}$ говорят, что она имеет распределение Стьюдента с числом степеней свободы *n*. Доказано, что плотность вероятности этой величины вычисляется по формуле

$$p_\tau n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, x \in R$$

При больших *n* распределение Стьюдента практически не отличается от N(0,1).



Функция вероятности



Функция распределения

Основные характеристики распределения Стьюдента примут вид:

$$M(X) = 0, если\ n > 1,\ D(X) = \frac{n}{n-2}, если\ n > 2,\ A = 0, если\ n > 3,\ E = \frac{6}{n-4}, если\ n > 4.$$