ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
Государственное образовательное учреждение высшего профессионального образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Ю.И. Галанов

Лабораторный практикум по математической статистике. Сборник программ в среде MathCad

*Рекомендовано в качестве учебно-методического пособия
Редакционно-издательским советом
Томского политехнического университета*

Издательство
Томского политехнического университета
2010

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Биномиальное распределение и его предельные формы

Цель занятия:

исследовать характер поведения ошибки аппроксимации биномиального распределения распределением Пуассона и нормальным распределением в зависимости от вероятности события.

В условиях справедливости локальной теоремы Муавра-Лапласа:

$$n \to \infty,\ p = const,\ n \cdot p,\ n \cdot p \cdot (1 - p) \to \infty$$

биномиальное распределение переходит в нормальное с параметрами:

$$a := n \cdot p \qquad s^2 := n \cdot p \cdot (1 - p)$$

Если выполняется условие: $n \to \infty, n \cdot p \to \lambda,\ \lambda = const,\ \lambda \le 10$

то биномиальное распределение хорошо аппроксимируется распределением Пуассона с параметром

$$\lambda := n \cdot p$$

Зададим функции биномиального распределения, распределения Пуассона и плотности нормального распределения и построим их графики при различных значениях параметров.

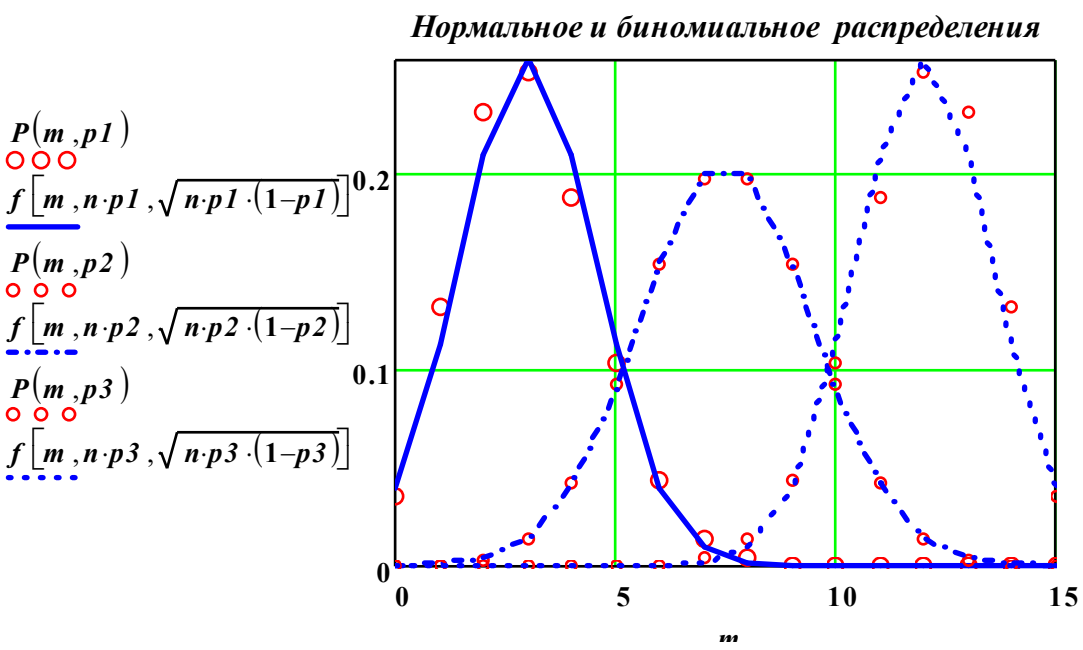
$n := 15$ $m := 0..n$ $p1 := 0.2$ $p2 := 0.5$ $p3 := 0.8$

$$P(m,p) := \frac{n!\cdot p^{m}\cdot(1-p)^{n-m}}{m!\cdot(n-m)!}$$

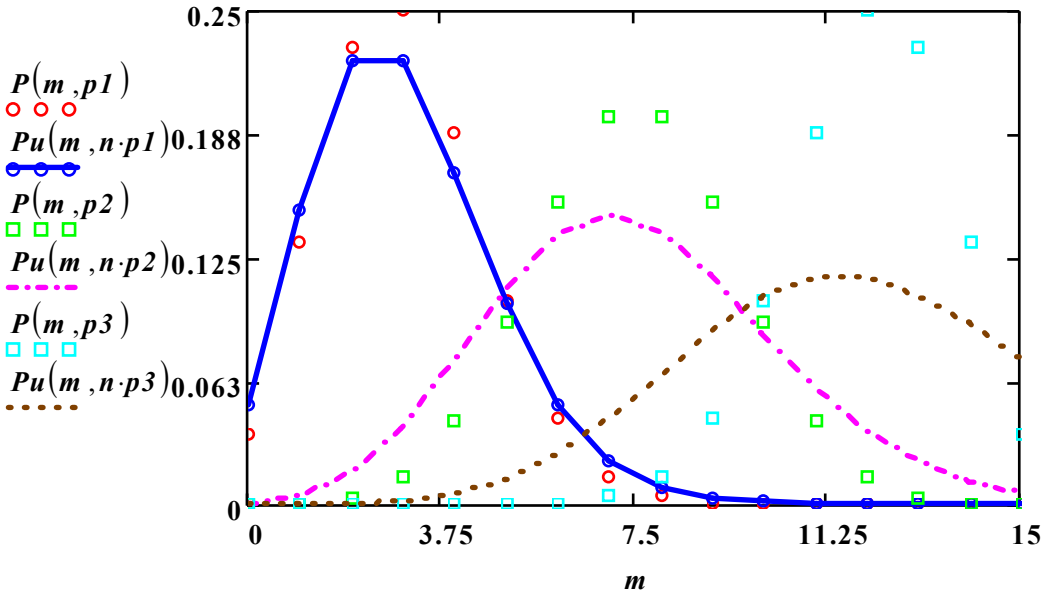
$$f(x,a,s) := \frac{exp\left[\frac{-(x-a)^2}{2\cdot s^2}\right]}{\sqrt{2\cdot\pi}\cdot s}$$

$$Pu(m,\lambda) := \frac{\lambda^{m}}{m!}\cdot exp(-\lambda)$$

Нормальное и биномиальное распределения



Биномиальное и Пуассоновское



Оценим качество нормального приближения с помощью функции качества S1(p) (остаточная дисперсия):

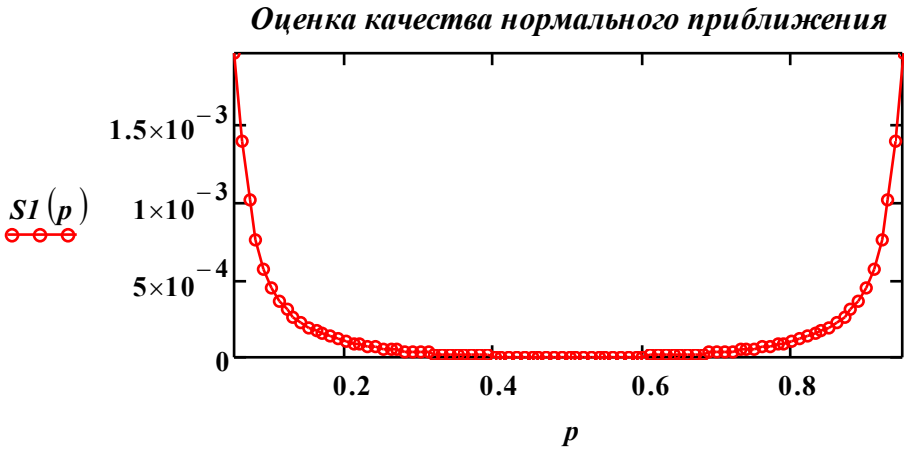
$$S1(p) := \left[\frac{1}{n}\cdot\sum_{m}\left[P(m,p)-f\left[m,(n\cdot p),\sqrt{n\cdot p\cdot(1-p)}\right]\right]^2\right]$$

Введем "функцию качества" для распределения Пуассона

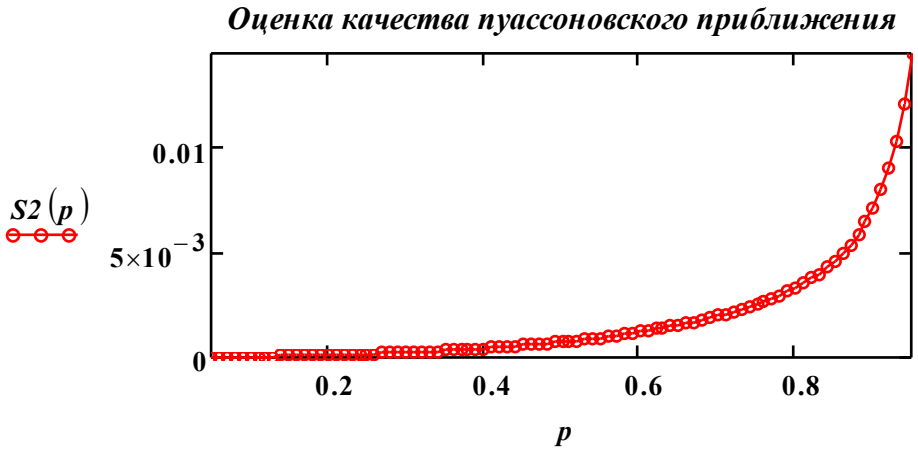
$$S2(p) := \frac{1}{n} \cdot \sum_{m} \left[P(m, p) - Pu\left[m, (n \cdot p)\right] \right]^2$$

Построим графики функций качества и оценим результаты аппроксимации

$$p := 0.05, .06 .. .95$$



Из этого графика видно, что нормальное распределение плохо работает как при малых,так и при больших значениях p.



Поведение S2(p) показывает, что Пуассоновское приближение лучше работает при малых значениях вероятности p

Указание:Изменя n (в пределах 10 - 150,)убедитесь, что ошибка аппроксимации уменьшается с ростом n.

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Моделирование случайных величин и их распределений

Моделирование случайных событий.Схема Бернулли

Опыт с двумя исходами полностью определяется заданием вероятности ***p*** события ***A***. Используя понятие геометрической вероятности, легко показать, что события {**A**} и {**rnd(1) < p**} равносильны.

Для генерирования случайных чисел, равномерно распределенных на отрезке [0,1] используем встроенную функцию ***rnd(1)***.

Зададим число ***p***, принадлежащее отрезку [0,1].

$$p := 0.75$$

Будем считать, что появление случайного числа

$$rnd(1) \le p$$

соответствует появлению события ***A***.

Введем индикатор событи ***J(p)***,равный единице, если событие произошло (т.е.если ***rnd(1)<=p***) и равный нулю, если событие не произошло (т.е. ***rnd(1)> p***)

$$J(p) := if(rnd(1) \le p, 1, 0)$$

Выборка из биномиального распределения

Смоделируем опыт, состоящий из ***n*** независимых испытаний и подсчитаем число испытаний, в которых произошло событие ***A*** в данной серии. Это число равно сумме индикаторов всех испытаний в одной серии:

$$n := 15 \quad i := 0..n \quad x_i := J(0.25)$$

$$x^T = (0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 1 \ 0 \ 1) \qquad \sum x = 5$$

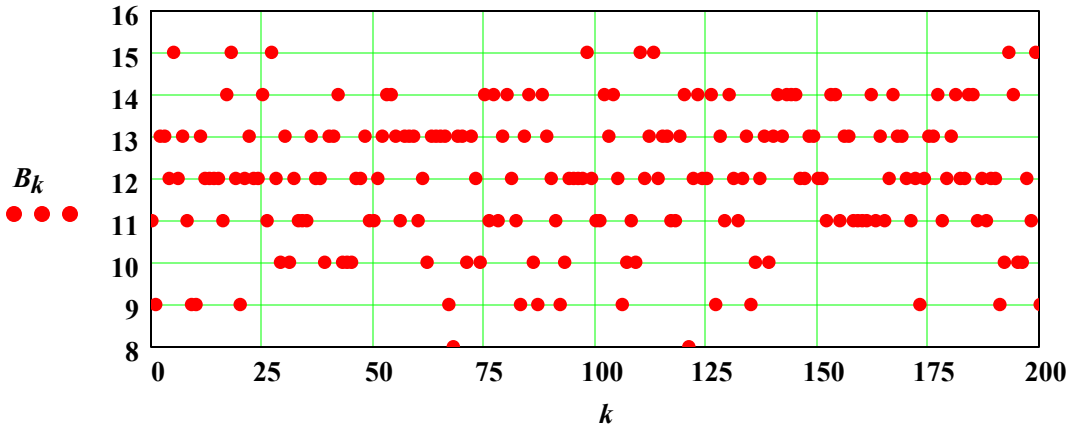
Число появлений события **A** в **n** независимых испытаниях **B** является случайной величиной, подчиняющейся биномиаальному распределению.

Создадим выборку **B**, состоящую из **N** элементов.

$$N := 201 \qquad k := 0 .. N - 1 \qquad B_k := \sum_i J(p)$$

Отображение выборки

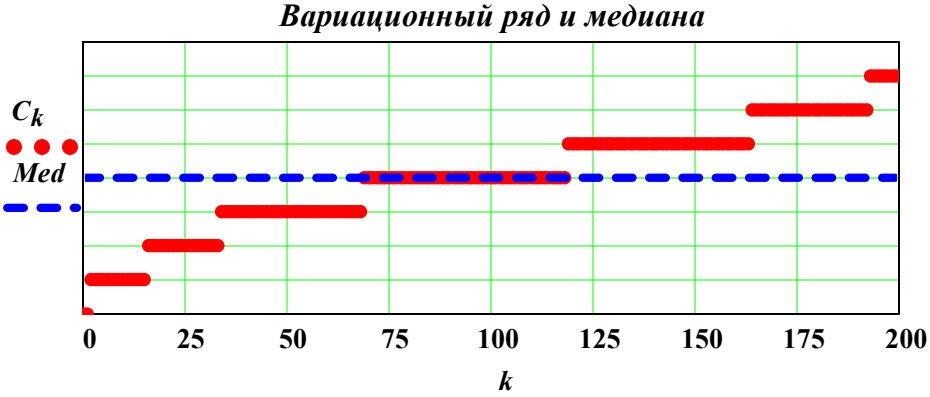
1.РЯД РАСПРЕДЕЛЕНИЯ



2. ВАРИАЦИОННЫЙ РЯД

Отсортируем выборку и получим вариационный ряд, из которого легко найти выборочную медиану и функцию распределения:

$$C := sort(B) \qquad Med := C_{\frac{N-1}{2}} \qquad Med = 12$$



<=...медиана

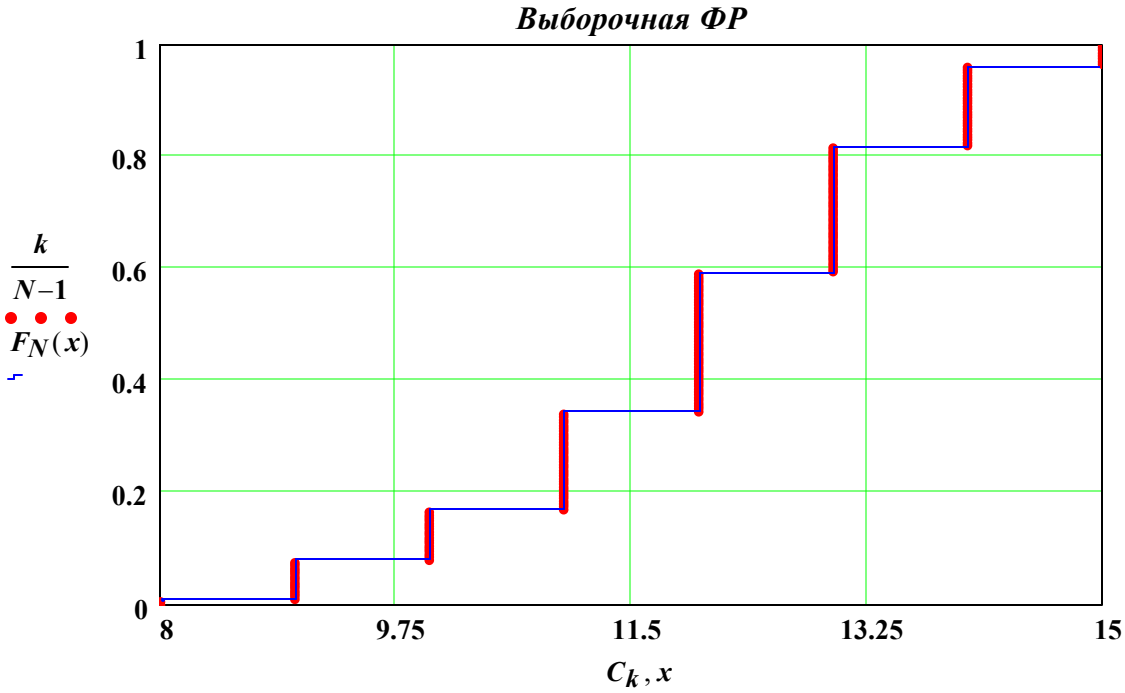
3.ВЫБОРОЧНАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Выборочную функцию распределения можно получить двумя способами:

1) задав ее в явном виде:

$$x := min(B)\,..\,n \qquad F_N(x) := \frac{1}{N}\cdot\sum_k if\left(B_k \le x, 1, 0\right) \qquad F_N(80) = 1$$

2)или используя вариационный ряд: => $F_N(k) := \frac{k}{N}$



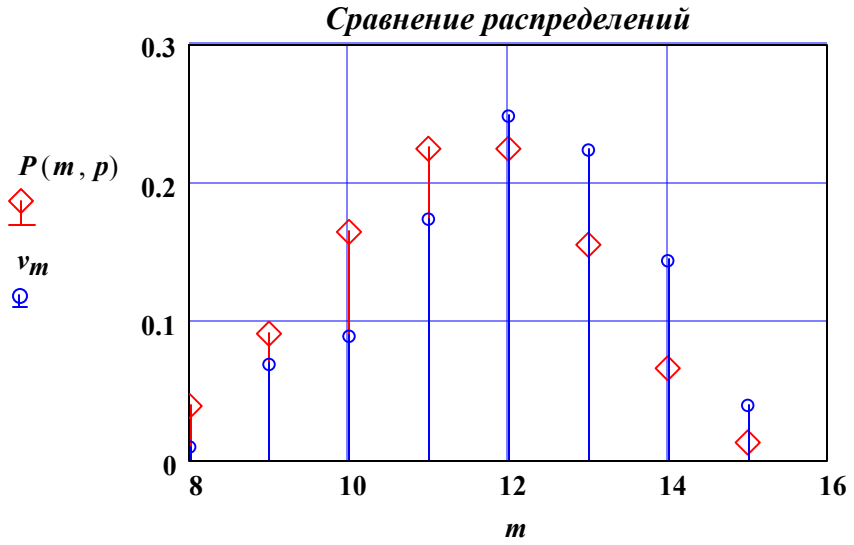
4.Сравним выборочное частотное распределение (статистический ряд) и теоретическое, рассчитанное по формуле Бернулли:

$$P(m, p) := \frac{n!\cdot p^m\cdot(1-p)^{n-m}}{m!\cdot(n-m)!}$$

Вводим индикатор равенства двух чисел.
Затем для каждого значения m подсчитываем, сколько элементов выборки принимает данное значение

$$Ident(x, y) := if(x \neq y, 0, 1) \qquad m := C_0 .. C_{last(C)}$$

$$v_m := \frac{1}{N} \cdot \left(\sum_k Ident(m, C_k) \right)$$



5. ГИСТОГРАММА И ПОЛИГОН ЧАСТОТ

В случае *дискретных* случайных величин строятся их частотные распределения. Гистограмма и полигон частот служат оценками плотности распределения для *непрерывных* случайных величин. В данном примере мы рассмотрим порядок построения гистограммы и полигона частот с использованием функции **hist**, встроеенной в MathCad, которая автоматически подсчитывает число точек, попавших в каждый полуинтервал.

1. Зададим полуинтервал **D**, полностью покрывающий область значений выборки (размах выборки)[**a**,**b**]. Разоб"ем его на **M** одинаковых полуинтервалов $[d_i, d_{i+1})$:
так,чтобы выполнялось условие $a < d_0 < d_1 < d_2 < \blacksquare .. b < d_M$

Число интервалов группировки зададип по формуле Старджеса:

$$M := 1 + trunc(log(n, 2)) \qquad l := 0 .. M$$

2. Зададим шаг дискретизации

$$h := \frac{max(B) - min(B)}{M} \cdot 1.0001$$

3. Создадим массив **d**, содержащий координаты точек разбиения области **D**:

$$d_l := min(B) + h \cdot l \qquad min(B) = 8 \qquad max(B) = 15$$

$$d^T = (8 \quad 9.75 \quad 11.5 \quad 13.251 \quad 15.001)$$

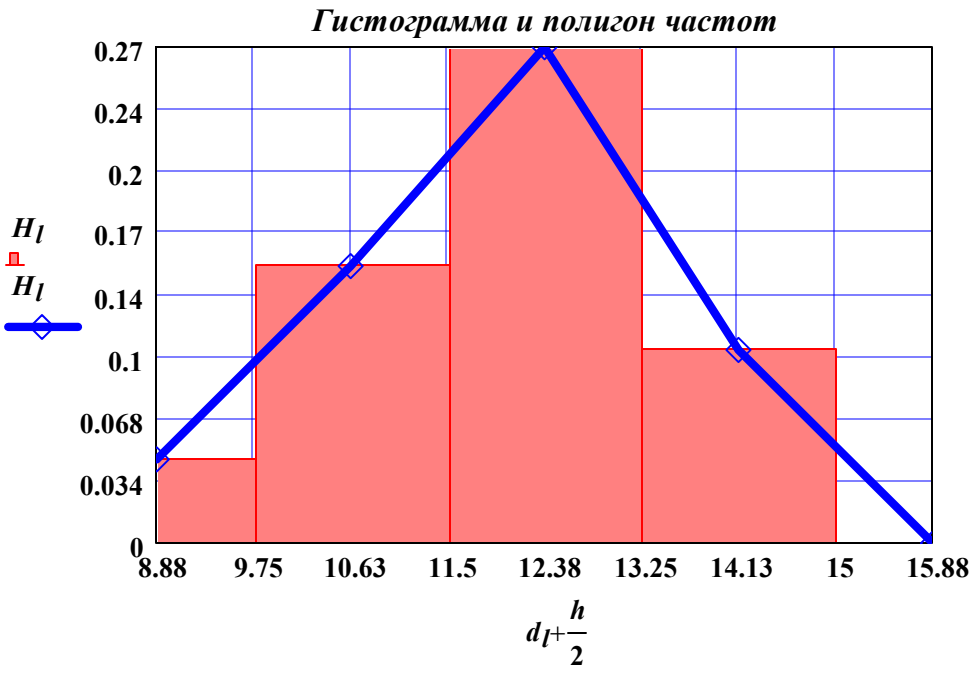
При построении гистограммы используем системную функцию **hist(vx,vy)**:

$$H := \frac{hist(d, B)}{N \cdot h} \qquad H_{last(H)+1} := 0 \qquad \sum_l H_l \cdot N \cdot h = 201$$ <= Проверочная сумма

Поскольку число элементов у массива **H** на единицу меньше, чем у массива **d**, то мы добавли ему еще один нулевой элемент. При построении графиков в качестве абсцисс берем середины частичных полуинтервалов.

Гистограмма изображается в виде прямоугольников с высотой, равной значению оценки плотности распределения.

Полигон частот - соединяет точки отрезками прямых линий.



6. Оценим выборочные параметры и сравним их с теоретическими (модельными)

- математическое ожидание и выборочное среднее

$$m := n \cdot p \quad m = 11.25 \qquad MO := \frac{1}{N} \cdot \sum_k C_k \quad MO = 12.03$$

- дисперсия и выборочная дисперсия

$$Disp := n \cdot p \cdot (1 - p) \qquad Disp = 2.813 \quad S2 := \frac{1}{N} \cdot \sum_k \left(B_k\right)^2 - MO^2 \quad S2 = 2.477$$

- исправленная дисперсия: $\frac{N}{N-1} \cdot S2 = 2.489$

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Моделирование равномерного распределения

Цель занятия:

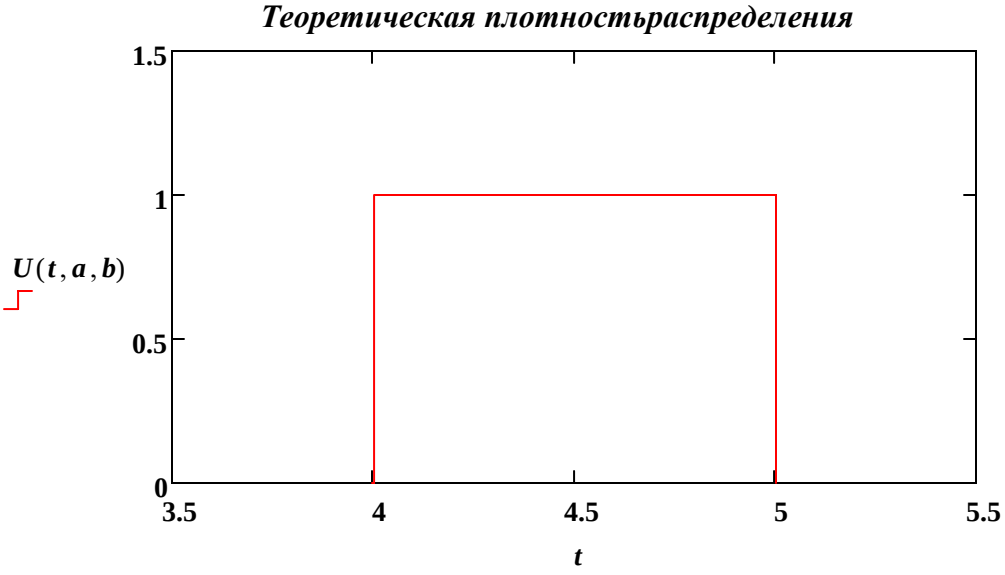
- Смоделировать выборку из равномерного распределения с заданными параметрами.
- сравнить визуально модельные и выборочные функцию и плотности распределения
- Оценить параметры распределения с помощью статистик, полученных методом моментов.

Задаем границы отрезка $a := 4$ $b := 5$

$t := a - 0.001 , a .. b + 0.001$

Теоретическая плотность распределения

$$U(x , a , b) := if\left(x < a , 0 , if\left(x > b , 0 , \frac{1}{b - a}\right)\right)$$

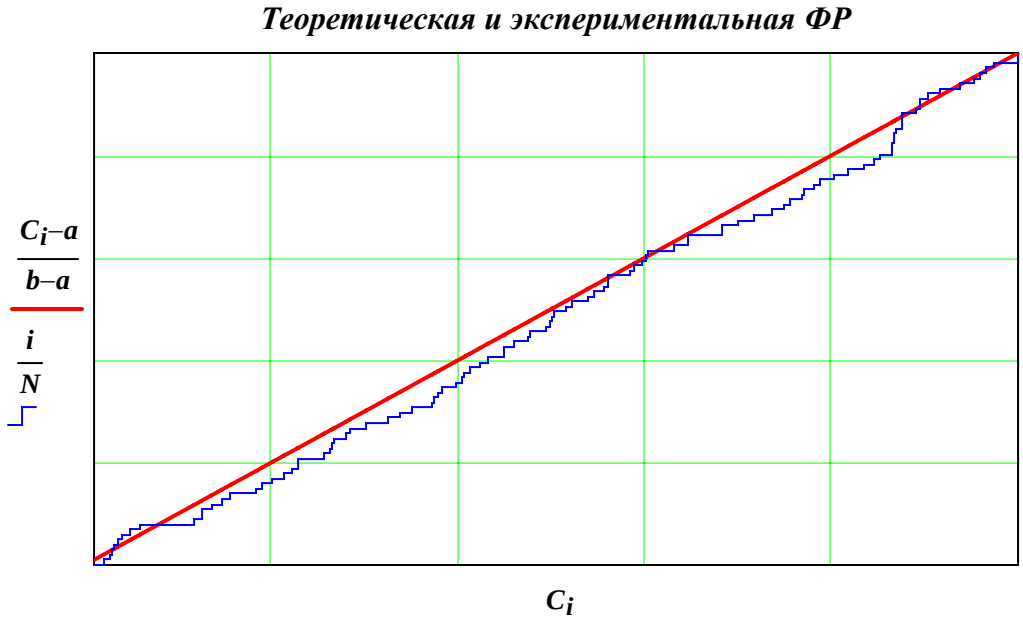


Задаем случайное число на отрезке [a, b] $Urnd(a,b) := a + rnd(b - a)$

Создаем выборку $N := 101$ $i := 0 .. N - 1$ $B_i := Urnd(a,b)$

Создаем вариационный ряд и используем его для построения эмпирической функции распределения (ФР)

$C := sort(B)$



Построение гистограммы

Находим число интервалов разбиения по формуле Старджеса

$$n := 1 + trunc(log(N, 2))$$

$$n = 7$$

Задаем шаг дискретизации

$$h := \frac{b \cdot 1.0001 - a}{n}$$

$$k := 0 .. n$$

Создаем массив точек разбиения

$$d_k := a + h \cdot k$$

Подсчитываем число точек, попавших в каждый интервал с помощью функции hist и делим на ***Nh*** (Объясните почему)

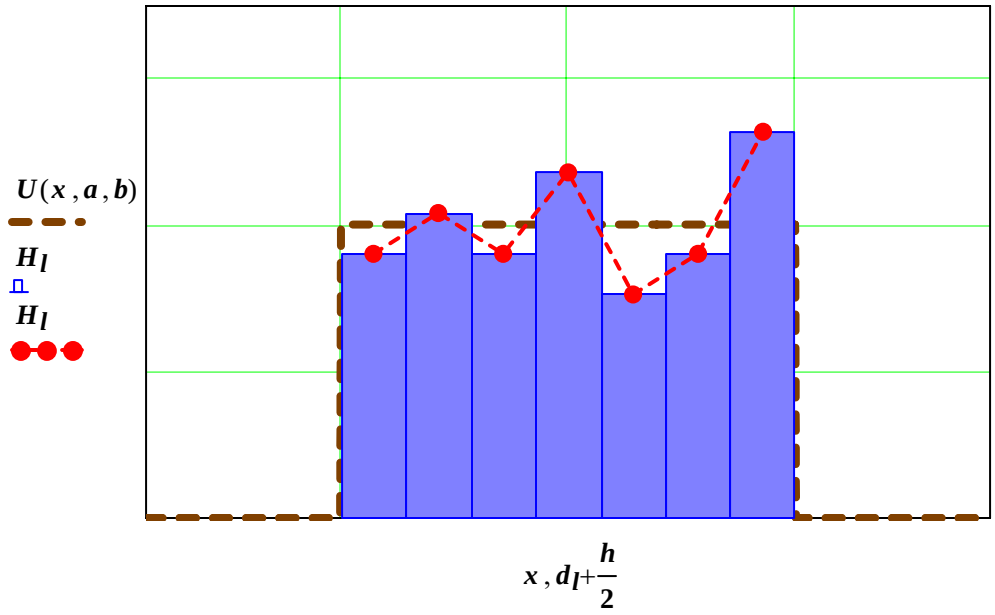
$$H := \frac{hist(d, B)}{N \cdot h}$$

$$l := 0 .. last(H)$$

Проверочная сумма =>

$$\sum_l \left[H_l \cdot (N \cdot h) \right] = 101$$

Гистограмма и полигон



Оценим выборочные параметры и сравним их с теоретическими (модельными)

математическое ожидание и выборочное среднее

$m := \frac{b+a}{2}$ $m = 4.5$ $MO := \frac{1}{N}\cdot\sum_i C_i$ $MO = 4.514968$

Дисперсия

$D := \frac{(b-a)^2}{12}$ $S2 := \frac{1}{N}\cdot\sum_i (B_i)^2 - MO^2$ $S2 = 0.082962$

$D = 0.083333$

исправленная дисперсия: $\frac{N}{N-1}\cdot S2 = 0.083791$

параметры a и $в$ находим методом моментов из системы уравнений:

$$\begin{cases} M_x = \frac{b+a}{2} \\ D_x = \frac{(b-a)^2}{12} \end{cases} \Rightarrow \begin{cases} \frac{b+a}{2} = \bar{X} \\ \frac{(b-a)^2}{12} = S^2 \end{cases}$$

$b + a = 2\cdot MO$

$b - a = \sqrt{12\cdot S2}$

$a' = 4.016083$

$a = 4$

$a' := MO - \frac{1}{2}\cdot\sqrt{12\cdot S2}$

$b' := \left(MO + \frac{1}{2}\cdot\sqrt{12\cdot S2}\right)$

$b' = 5.013852$

$b = 5$

Галанов Ю.И.

Лабораторный практикум по мат. статистике

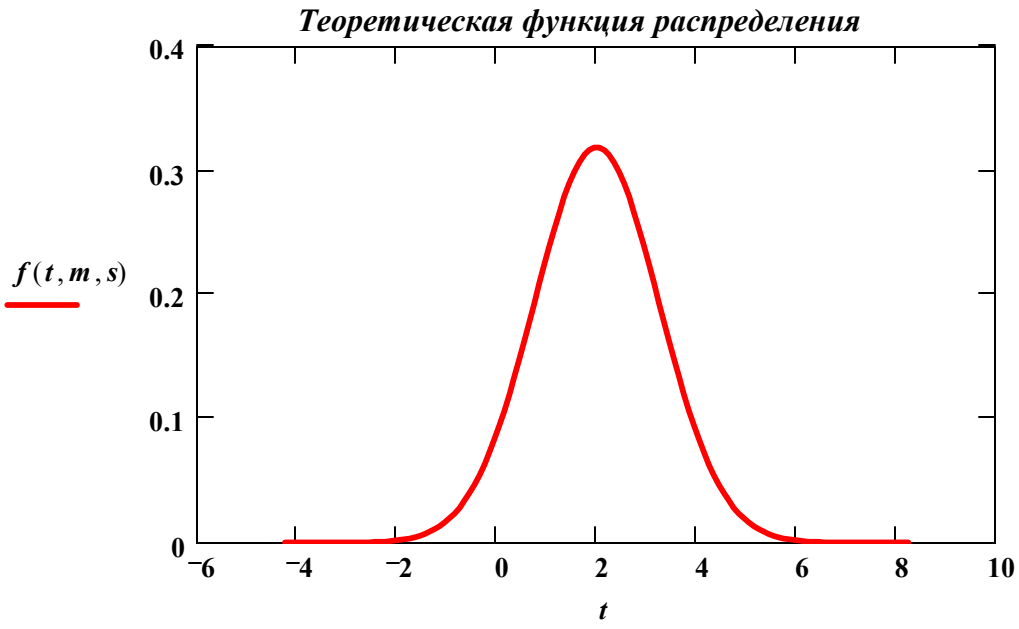
Моделирование нормального распределения

Задаем параметры распределения $m := 2$ $s := 1.25$

Теоретическая плотность распределения

$$f(x, m, s) := \frac{1}{\sqrt{2 \cdot \pi} \cdot s} \cdot exp\left[-\frac{(x - m)^2}{2 \cdot s^2} \right]$$

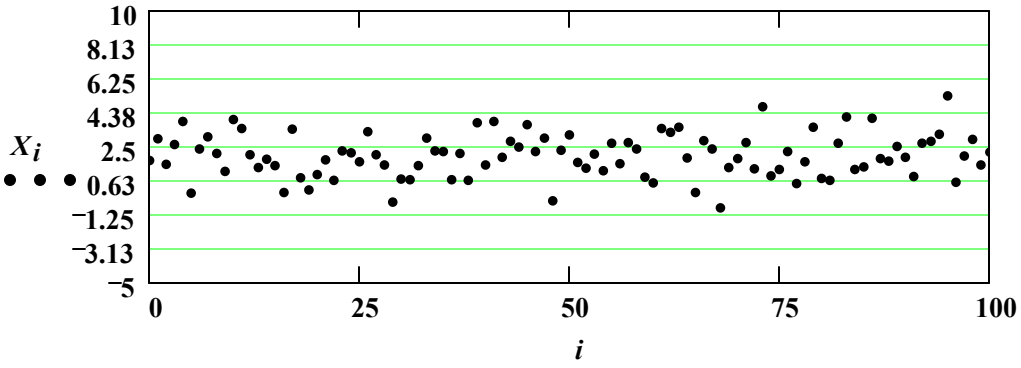
$$t := m - 5 \cdot s, m - 5 \cdot s + 0.1 .. m + 5 \cdot s$$



Создаем выборку

$$N := 101 \quad i := 0 .. N - 1$$

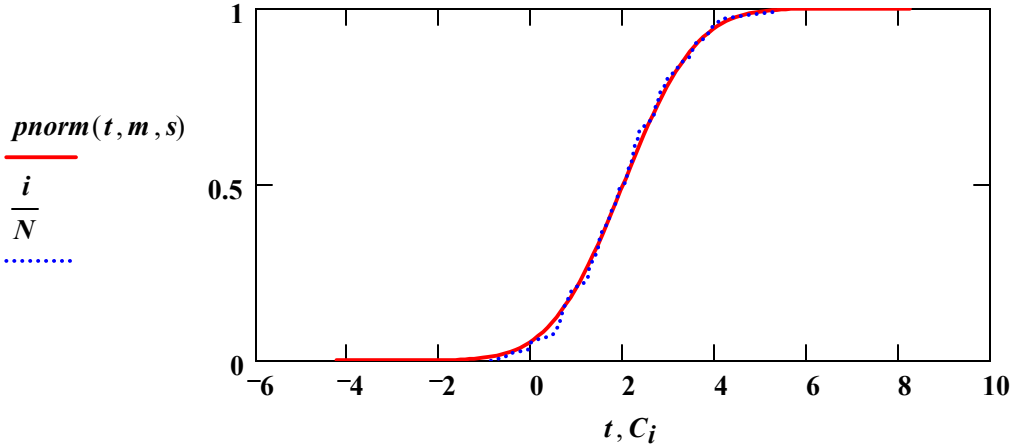
$$X_i := m + s \cdot cos(2 \cdot \pi \, rnd(1)) \cdot \sqrt{-2 \cdot ln(rnd(1))}$$



Выборка из нормального распределения

Создаем вариационный ряд и исползуем его для построения эмпирической функции распределения (ФР):

$$C := sort(X)$$



Построение гистограммы

Задаем число отрезков $n := \left(1 + trunc\left(\frac{ln(N)}{ln(2)}\right)\right)$ $n = 7$

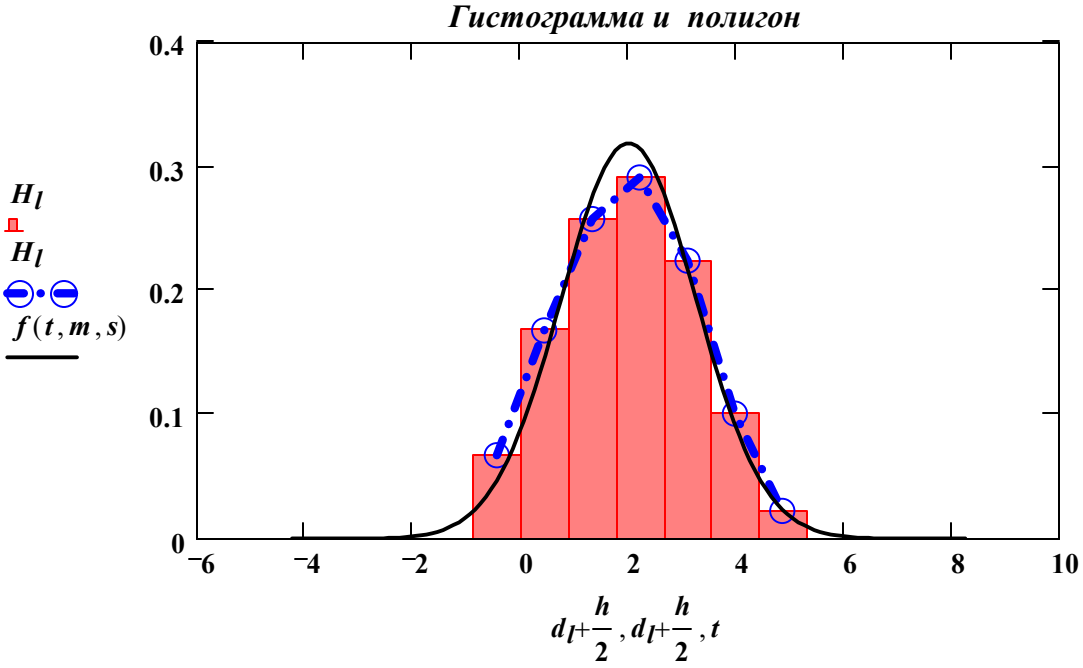
Задаем шаг дискретизации $h := \frac{C_{N-1} - C_0}{n} \cdot 1.001$ $k := 0 .. n$

Создаем массив точек разбиения $d_k := C_0 + h \cdot k$

Подсчитываем число точек, попавших в каждый отрезок
с помощью функции ***hist*** и делим на $N \cdot h$ (Объясните почему)

$$H := \frac{hist(d, X)}{N \cdot h} \qquad l := 0 .. last(H)$$

Проверочная сумма => $\sum_l H_l \cdot (N \cdot h) = 101$



Оценим выборочные параметры и сравним их с модельными

- математическое ожидание и выборочное среднее

$m = 2$ $\qquad MO := \frac{1}{N} \cdot \sum_i C_i \qquad MO = 1.984486$

- дисперсия и выборочная дисперсия:

$s^2 = 1.5625 \qquad S2 := \frac{1}{N} \cdot \sum_i \left(X_i - MO\right)^2 \qquad S2 = 1.41423$

- исправленная дисперсия: $\frac{N}{N-1} \cdot S2 = 1.428372$

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Построение доверительных интервалов для параметров нормального распределения

Задание

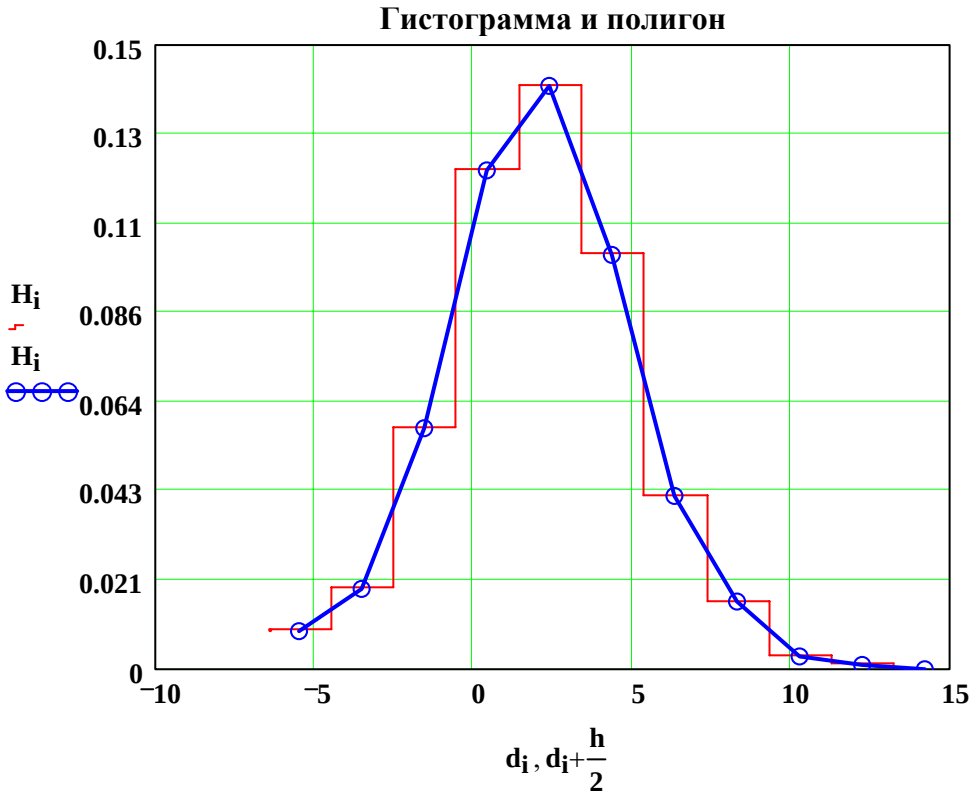
1. Смоделировать выборку из нормального распределения.
2. Построить гистограмму и полигон частот.
3. Оценить параметры распределения.
4. На уровне доверия γ построить доверительный интервал для математического ожидания и дисперсии.

$N := 500$ $k := 0 .. N - 1$ $a := 2$ $s := 3$

$X_k := a + s \cdot \sin(2 \cdot \pi \cdot \mathrm{rnd}(1)) \cdot \sqrt{-2 \cdot \ln(\mathrm{rnd}(1))}$

$M := 1 + \mathrm{ceil}\left(\frac{\ln(N)}{\ln(2)}\right)$ $i := 0 .. M$ $h := \frac{\max(X) \cdot 1.0001 - \min(X)}{M}$

$d_i := \min(X) + h \cdot i$ $H := \frac{\mathrm{hist}(d, X)}{N \cdot h}$ $H_{\mathrm{last}(H)+1} := 0$



Доверительный интервал для математического ожидания

$m := mean(X)$ $S := stdev(X)$ $\gamma := 0.95$

$T(t) := m - t \cdot \frac{S}{\sqrt{N-1}}$ $t := qt\left(\frac{1+\gamma}{2}, N-1\right)$ $t = 1.964729$

$T1 := T(t)$ $T2 := T(-t)$

$T1 = 1.823697$ $a = 2$ $T2 = 2.336328$

Доверительный интервал для дисперсии

$n := \text{rows}(X)$ $n = 500$ $D := \text{var}(X)$ $D = 8.492689$

$$t1 := \text{qchisq}\left(\frac{1-\gamma}{2}, n-1\right) \qquad t2 := \text{qchisq}\left(\frac{1+\gamma}{2}, n-1\right)$$

$t1 = 438.998025$ $t2 = 562.789493$ $T1 := \frac{D \cdot n}{t2}$ $T2 := D \cdot \frac{n}{t1}$

$T1 = 7.545174$ $s^2 = 9$ $T2 = 9.67281$

Результаты расчетов представьте в виде таблицы

γ	N	Мат. ожидание		Дисперсия	
		T1	T2	T1	T2
0.95	100				
	200				
	500				
0.995	100				
	200				
	500				

Ответьте на контрольные вопросы:

- **Какие статистики применяются для построения доверительных интервалов?**
- **Каковы их распределения?**
- **Что такое квантили распределения и как они используются при вычислении доверительных интервалов?**
- **Как изменяется величина доверительного интервала с увеличением объема выборки?**
- **Как изменяется величина доверительного интервала с увеличением доверительной вероятности?**

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Проверка гипотезы о нормальности распределения. Критерий Колмогорова

Задание.

- **Смоделировать две выборки одинакового объема: одну с нормальным распределением, а другую с равномерным распределением. Мат. ожидание и дисперсию для обеих выборок задать одинаковыми.**
- **Провести визуальную проверку гипотезы о нормальности распределений. Для чего представить стандартизованные значения вариант распределений как функции соответствующих квантилей стандартного распределения.Если распределение нормально, то должна получиться прямая линия. В противном случае экспериментальные точки не будут укладываться на прямую линию. *Визуальный метод проверки нормальности основан на способности человека отличать прямую линию от непрямой.***
- **Проанализируйте, как влияет объем выборок на на возможность их визуального распознавания.**
- **Проверьте гипотезу о нормальности выборок с помощью критерия Колмогорова. Для этого задайте и подсчитайте значения статистик критерия для обеих выборок. Критические точки возьмите из прилагаемых таблиц.**
- **Проверьте экспериментально, как влияет объем выборки и значение уровня значимости на возможность идентифицировать нормальную выборку.**
- **При составлении отчета опишите пять шагов проверки гипотезы применительно к Вашей задаче.**

Контрольные вопросы

- **Что называют статистической гипотезой?**
- **Какая гипотеза является простой?**
- **Какие ошибки возможны при проверке гипотезы?**
- **Что такое уровень значимости и каков диапазон его значений?**
- **Что такое статистический критерий и статистика критерия?**
- **Что такое мощность критерия?**
- **Что такое критическая область. Какими бывают критические области?**
- **Какова критическая область для данной задачи?**
- **От чего зависит надежность статистических выводов?**

Примерная программа проведения работы

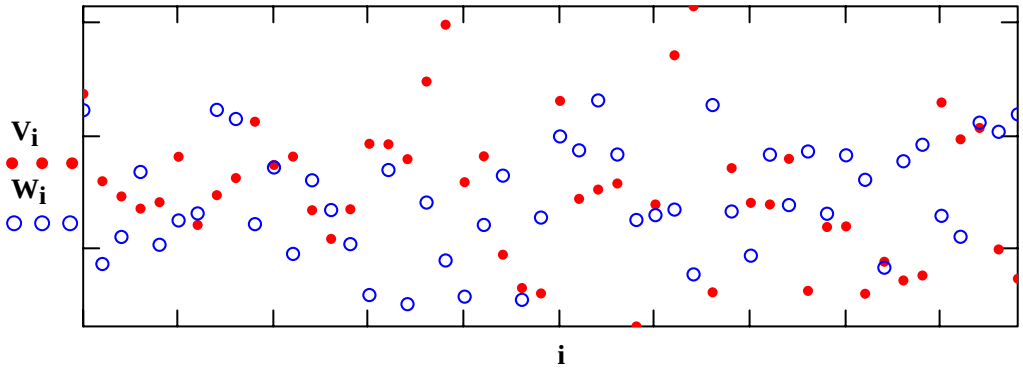
Моделируем две выбоки с различными распределениями, но одинаковыми параметрами

Нормальное **Равномерное**

$N := 50$ $i := 0 .. N - 1$ $a := 1$ $b := 5$

$m := \frac{a + b}{2}$ $m = 3$ $s := \frac{b - a}{2 \cdot \sqrt{3}}$ $s = 1.154701$

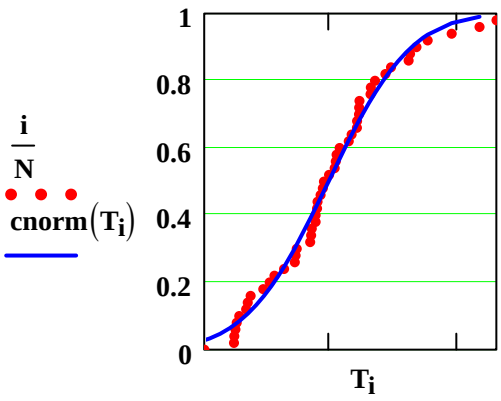
$V_i := m + s \cdot \cos(2 \cdot \pi \cdot rnd(1)) \cdot \sqrt{-2 \cdot \ln(rnd(1))}$ $W_i := a + rnd(b - a)$

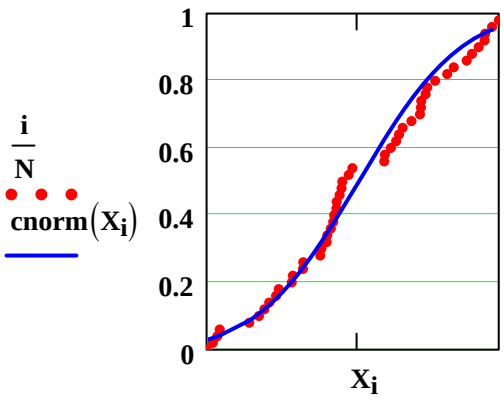


Преобразуем обе выборки в стандартные, предполагая, что справедлива гипотеза о нормальности выборок

$Z := \left(\frac{V - mean(V)}{stdev(V)} \right)$ $T := sort(Z)$ $Y := \left(\frac{W - mean(W)}{stdev(W)} \right)$ $X := sort(Y)$

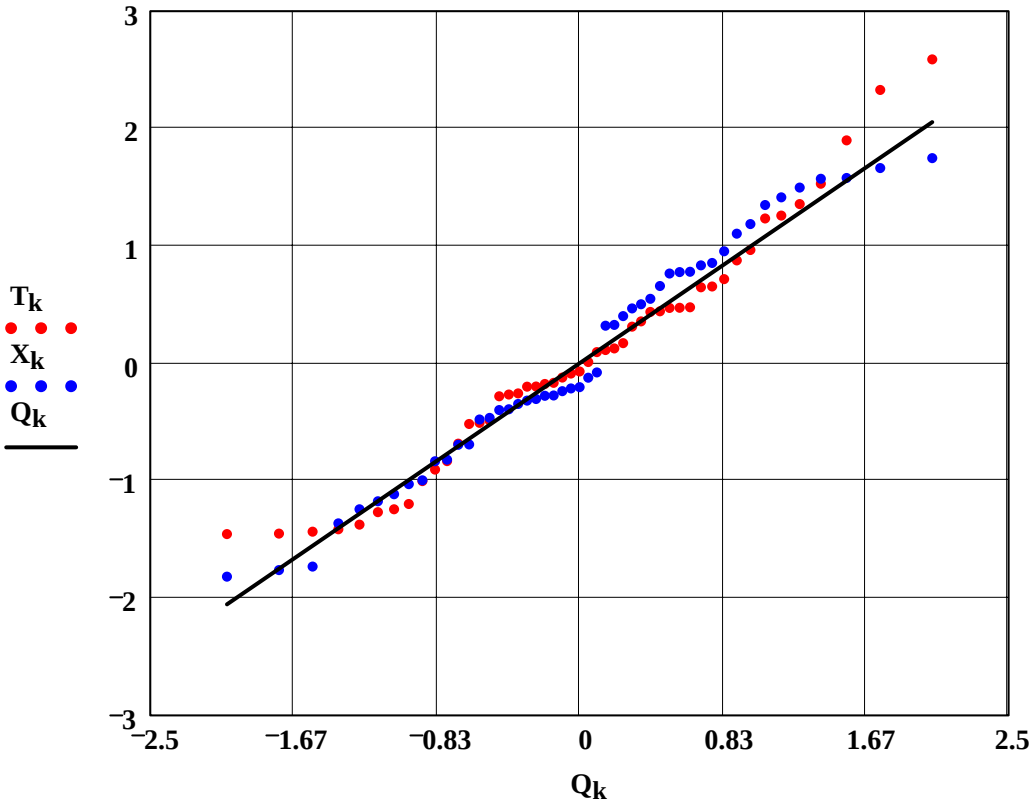
Построим выборочные функции распределения, используя вариационный ряд и сравним их с теоретическим (cnorm(x))





Визуальная проверка нормальности

$$k := 1 .. N - 1 \qquad F_k := \frac{k}{N} \qquad Q_k := qnorm(F_k, 0, 1)$$



Подсчитаем значения статистик критерия Колмогорова

$N \equiv 50$

Нормальное

$$TT_i := \left| \frac{i}{N} - cnorm(T_i) \right|$$

$$tT := \sqrt{N} \cdot max(TT)$$

$$tT = 0.48837$$

Равномерное

$$TX_i := \left| \frac{i}{N} - cnorm(X_i) \right|$$

$$tX := \sqrt{N} \cdot max(TX)$$

$$tX = 0.570625$$

Делаем выводы о принадлежности или непринадлежности выборок к нормальному распределению

УКАЗАНИЕ 1: оцените ошибку второго рода и мощность критерия; результаты представьте в виде таблицы

N	β	$1-\beta$	Критическая точка
50			
100			
200			
500			
1000			

УКАЗАНИЕ 2: Чтобы ценить ошибку 2-го рода надо выполнить:

1. **помести курсор на** $N \equiv 50$
2. **нажимай F9 10 (или 20) раз и считай, сколько раз tX примет значение, меньше критической точки . Частота этого события и есть оценка ошибки 2-го рода**
3. **Замени значение N и повтори п.2 и тд.**

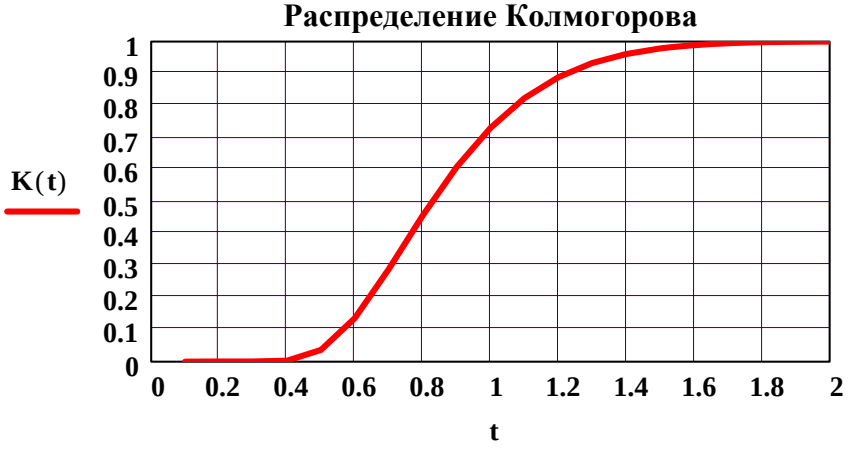
Таблица квантилей распределения Колмогорова

Уровень значимости	критическая точка
0.05	1.358
0.01	1.6276

$j := 0 .. 1001$

$$K(t) := \sum_{i=-2001}^{2001} \left[\left[(-1)^{-i} \right] \cdot e^{-\left(2 \cdot i^2 \cdot t^2\right)} \right]$$

$K(1.358) = 0.949973$ $\qquad$ $t := 0 , 0.1 .. 2$



Расчет квантилей распределения Колмогорова

$$qK(t, alfa) := root(K(t) - alfa, t) \qquad qK(1, 0.95) = 1.354618$$

$$qK(1, 0.9) = 1.223734 \qquad qK(1, 0.95) = 1.354618 \qquad qK(1, 0.99) = 1.621647$$

$$qK(1, 0.995) = 1.704622 \qquad qK(1, 0.999) = 1.871257 \qquad qK(1, 0.9995) = 1.898543$$

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Проверка гипотезы о виде распределения с помощью критерия ХИ - квадрат

Задание.

- **Для выборки X на уровне значимости 0.05 проверить гипотезу о том, что она описывается показательным распределением.**
- **Применить метод хи-квадрат.**

Указание.

- **Провести разбиение области значений функции на интервалы равной вероятности.**
- **Проверить гипотезу для двух распределений: показательного и треугольного.**

В отчете:

- **Опишите пять шагов проверки гипотезы применительно к данному случаю.**
- **Вывести услоие равенства математических ожиданий для модельных распределений.**
- **Найти оценку параметров треугольного распределения методом моментов. Сколько независимых параметров задают треугольное распредеелние?**
- **В чем суть моделирования с помощью обратных функций?**
- **Объясните особенности построения интервалов равной вероятности.**

Плотность показательного распределения $\mathbf{f}(\mathbf{x},\mathbf{L}) := \mathbf{L}\cdot\exp(-\mathbf{L}\cdot\mathbf{x})$

Функция распределения $\mathbf{F}(\mathbf{x},\mathbf{L}) := 1-\exp(-\mathbf{L}\cdot\mathbf{x})$

Обратная функция $\mathbf{qexp}(\mathbf{u},\mathbf{L}) := \frac{-1}{\mathbf{L}}\cdot\ln(1-\mathbf{u})$

Параметры распределений задаем из условия равенства математических ожиданий

$\mathbf{a} := 0.5 \qquad \mathbf{b} := \frac{2}{\mathbf{a}} \qquad \mathbf{L} := \frac{6}{\mathbf{a}\cdot\mathbf{b}^2} \qquad \mathbf{L} = 0.75$

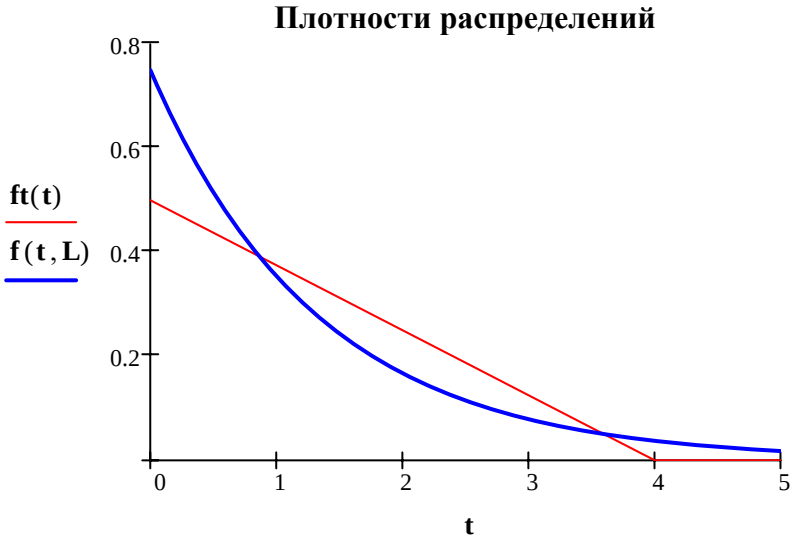
Плотность треугольного распределения $\mathbf{ft}(\mathbf{x}) := \mathbf{if}\left(\mathbf{x}<0,0,\mathbf{if}\left(\mathbf{x}>\mathbf{b},0,\mathbf{a}-\frac{\mathbf{a}}{\mathbf{b}}\cdot\mathbf{x}\right)\right)$

Функция распределения

$$Ft(x) := \text{if}\left(x < 0\,,0\,,\text{if}\left(x > b\,,1\,,a\cdot x - \frac{a}{b}\cdot\frac{x^2}{2}\right)\right)$$

Обратная функция

$$qtr(t) := b - \sqrt{\frac{2\cdot b\cdot(1-t)}{a}}$$



Моделируем выборку из показательного распределения

$i := 0 .. N - 1$ $\quad X_i := qexp(rnd(1), L)$ $\quad$ Сортируем $\quad Xs := sort(X)$

$$L1 := \frac{1}{mean(X)}$$

$L1 = 0.746$

Моделируем выборку из треугольного распределения

$Y_i := qtr(rnd(1))$ $\quad Ys := sort(Y)$ $\quad L2 := \dfrac{1}{mean(Y)}$

Задаем число интервалов группировки и формируем интервалы равной вероятности

$M := 1 + trunc(log(N, 2))$

$M = 6$ $k := 0 .. M - 1$ $l_k := \frac{k}{M}$

$dx_k := qexp(l_k, L1)$ $dx_M := max(X) \cdot 2$

$dx^T = (0 \quad 0.244 \quad 0.543 \quad 0.929 \quad 1.472 \quad 2.401 \quad 15.315)$

$dy_k := qexp(l_k, L2)$

$dy_M := max(Y) \cdot 2$ $dy^T = (0 \quad 0.246 \quad 0.547 \quad 0.935 \quad 1.481 \quad 2.416 \quad 6.955)$

Подсчитываем число точек, попавших в интервалы группировки

$v := hist(dx, X)$ $vy := hist(dy, Y)$

$v^T = (9 \quad 6 \quad 8 \quad 10 \quad 10 \quad 7)$ $vy^T = (4 \quad 10 \quad 3 \quad 13 \quad 16 \quad 4)$

Теоретическое (ожидаемое) число точек, попавших в интервалы группировки

$vt := \frac{N}{M}$ $round(vt) = 8$

Число степеней свободы

$r := M - 1 - 1$ $r = 4$

Критическая точка $C := qchisq(0.95, r)$ $C = 9.488$

Статистики критерия

$$hiX := \sum_k \frac{(v_k - vt)^2}{vt} \qquad hiX = 1.6 \qquad hiY := \sum_k \frac{(vy_k - vt)^2}{vt} \qquad hiY = 17.92$$

Функция принятия решения (для ленивых, лень - двигатель прогресса!)

$Crit(t) := if(t < C, "Ho --- Prinimaem ", "Ho --- Otwergaem")$

$N \equiv 50$ $Crit(hiX) = "Ho --- Prinimaem "$ $Crit(hiY) = "Ho --- Otwergaem"$

УКАЗАНИЕ: оцените ошибку второго рода и мощность критерия; результаты представьте в виде таблицы

N	β	$1-\beta$	Критическая точка
50			
100			
200			
500			
1000			

УКАЗАНИЕ: Чтобы ценить ошибку 2-го рода надо выполнить:

1. **помести курсор на** $N \equiv 50$
2. **нажимай F9 10 (или 20) раз и считай, сколько раз hiY примет значение, меньше критической точки (или значение функции будет** $Crit(hiY) = "Ho -- Prinimaem""$**). Частота этого события и есть оценка ошибки 2-го рода**
3. **Замени значение N и повтори п.2 и тд.**

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Проверка гипотезы о независимости

Задание

1. **Смоделировать выборку из системы 2-х случайных величин с линейной корреляционной связью.**
2. **Построить корреляционное поле.**
3. **Оценить параметры корреляционной зависимости.**
4. **Проверить гипотезу о значимости коэффициента корреляции.**
5. **Проверьте влияние объема выборки на чувствительность данного критерия к величине параметров линейной связи.**
6. **Постройте доверительный интервал для коэффициента корреляции при большом объеме выборки и значимом коэффициенте корреляции.**

Для отчета

1. **Опишите этапы проведения корреляционного анализа на данном примере.**
2. **Прокомментируйте этапы составления программы.**
3. **Изложите мотивированные выводы о влияние объема выборки и значений параметров модели на надежность статистических выводов.**
4. **Какой геометрический смысл можно придать доверительным границам коэффициента корреляции?**

1. Моделируем выборку из двухмерной системы нормально распределенных случайных величин

$N := 50$ $\quad k := 0 .. N - 1$

$w(x, y) := \sin(2 \cdot \pi \cdot x) \cdot \sqrt{-2 \ln(y)}$ $\quad k := 0 .. N - 1$

Модель линейной регрессии $Y(x, mx, my, sx, sy, \rho) := my + \rho \cdot \frac{sy}{sx} \cdot (x - mx)$

Задаем параметры модели $mx := 1$ $\quad my := 1$ $\quad sx := 100$ $\quad sy := 10$ $\quad \rho := 0.5$

Первая компонента системы $X_{k,0} := mx + sx \cdot w(rnd(1), rnd(1))$

Вторая компонента системы $X_{k,1} := Y\left(X_{k,0}, mx, my, sx, sy, \rho\right) + sy \cdot \sqrt{1 - \rho^2} \cdot w(rnd(1), rnd(1))$

ЗАМЕЧАНИЕ.При моделировании второй компоненты берем остаточную дисперсию $sy \cdot \sqrt{1 - \rho^2}$

2. Обработка экспериментальных результатов

Сортируем массив данных по первой компоненте $X := csort(X, 0)$

ОЦЕНИВАЕМ ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СОСТАВЛЯЮЩИХ СИСТЕМЫ

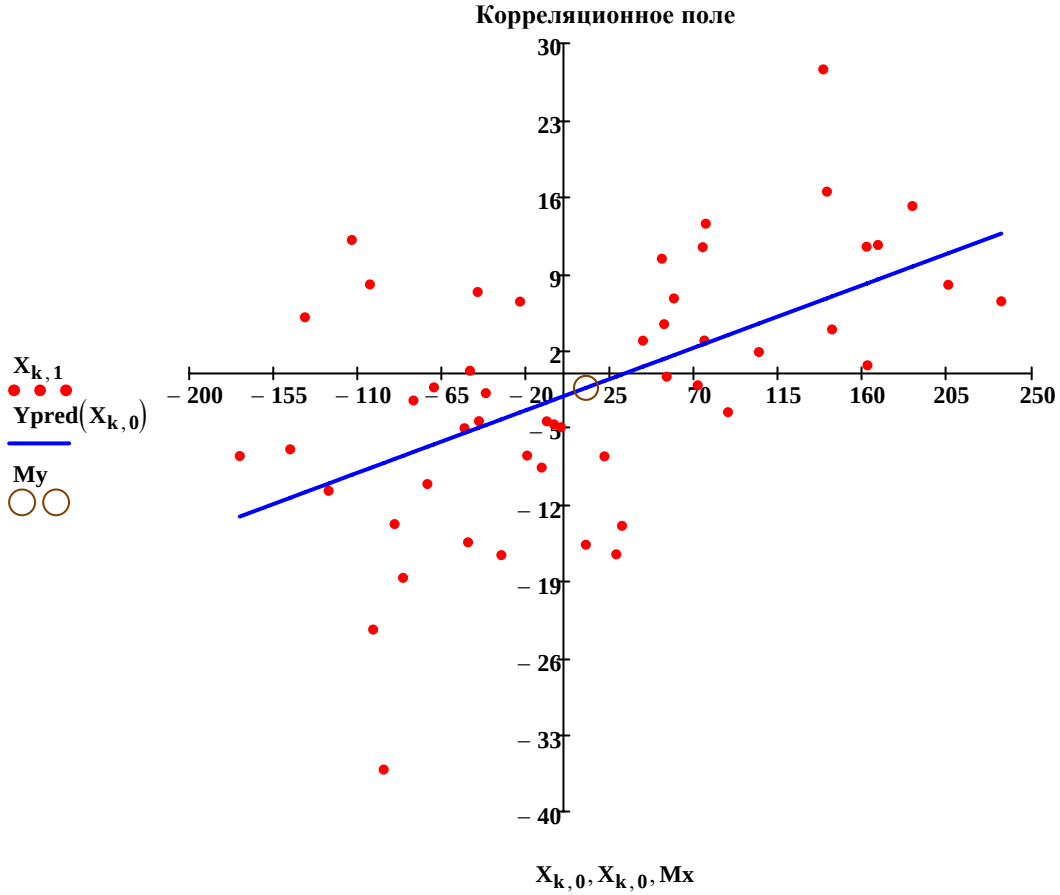
$Mx := mean(X^{\langle 0 \rangle})$ $My := mean(X^{\langle 1 \rangle})$ $Sx := stdev(X^{\langle 0 \rangle})$ $Sy := stdev(X^{\langle 1 \rangle})$

Вычисляем выборочный коэффициент корреляции

Явная формула $$r := \frac{\frac{1}{N} \cdot \sum_k \left[\left[(X^{\langle 0 \rangle})_k - Mx \right] \cdot \left[(X^{\langle 1 \rangle})_k - My \right] \right]}{Sx \cdot Sy} \qquad r = 0.548$$

Проверяем с помощью встроенной функции $corr(X^{\langle 0 \rangle}, X^{\langle 1 \rangle}) = 0.548$

Предсказанное значение $Ypred(x) := Y(x, Mx, My, Sx, Sy, r)$



Сравним модельные параметры с их оценками

$mx = 1$	$Mx = 11.989$	$my = 1$	$My = -1.394$	$\rho = 0.5$	
$sx = 100$	$Sx = 99.892$	$sy = 10$	$Sy = 11.559$	$r = 0.548$	$N = 50$

3. Проверка гипотезы о значимости коэффициента корреляции

Уровень значимости $\alpha := 0.01$

Критическая точка (двухсторонний критерий) $C := qt\left(1 - \frac{\alpha}{2}, N - 2\right)$

Статистика критерия $t := \frac{|r|}{\sqrt{1 - r^2}} \cdot \sqrt{N - 2}$

Сравниваем значения статистики критерия и критической точки.

Делаем выводы.

$t = 4.537$ $C = 2.682$

4. Доверительный интервал для коэффициента корреляции Используем Z преобразование Фишера

$$Z(r) := \frac{1}{2} \cdot \ln\left(\frac{1 + r}{1 - r}\right)$$

Статистика $\sqrt{N - 3} \cdot (Z - Mz)$ имеет стандартное распределение

$\gamma := 0.95$

$q := qnorm\left(\frac{1 + \gamma}{2}, 0, 1\right)$ $q = 1.96$

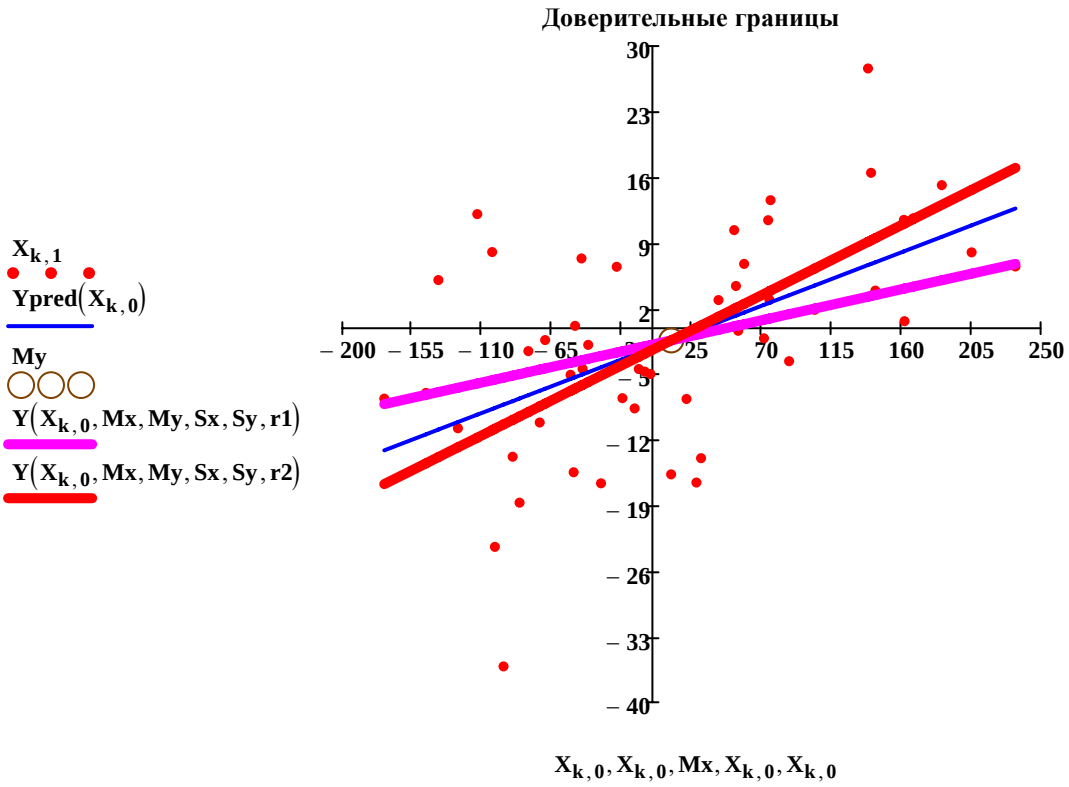
$\Delta Z := \frac{q}{\sqrt{N - 3}}$ $\Delta Z = 0.286$

Обратное преобразование Фишера $R(z) := \tanh(z)$

Нижняя граница $R(Z(r) - \Delta Z) = 0.318$ $r1 := R(Z(r) - \Delta Z)$

$r = 0.548$

Верхняя граница $R(Z(r) + \Delta Z) = 0.717$ $r2 := R(Z(r) + \Delta Z)$



Оцените ошибку 2-го рода при оценке значимости коэффициента корреляции и заполните таблицу

ρ	N=30	N=100	N=500	N=1000
	β	β	β	β
0.1				
0.25				
0.5				
0.75				

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Классический регрессионный анализ

Задание.

1. Смоделировать регрессионную зависимость в виде многочлена третьей степени.
2. Смоделировать независимую выборку для оценки дисперсии при фиксированном значении независимой переменной.
3. Оценить параметры полиномиальной модели методом наименьших квадратов.
4. Проверить гипотезу об адекватности модели с помощью критерия Фишера.
5. Оценить коэффициент множественной корреляции (детерминации) и проверить гипотезу об его значимости.
6. Проверить гипотезу о значимости коэффициентов модели.
7. Исследовать остатки.
8. Построить доверительные интервалы для предсказанных значений.

Указание.

1. Проведите проверку адекватности моделей, начиная с многочлена первой степени до многочлена 4-й степени включительно . Для этого подключайте необходимое число регрессоров в матрице F.
2. Проведите анализ остатков.
3. Сформулируйте мотивированные выводы об адекватности моделей.
4. Проверку гипотезы о значимости коэффициентов модель проведите при больших и малых значениях объема выборки.
5. Исключить незначимые коэффициенты.

Моделируем данные

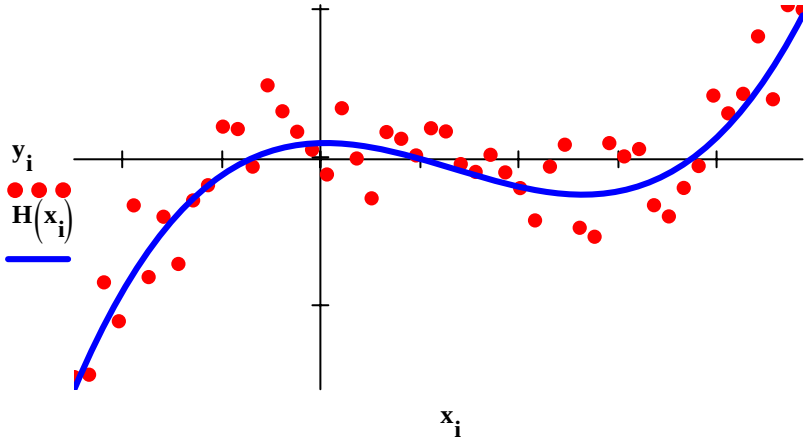
Параметры модели $b_1 := 6 \quad b_0 := 100 \quad b_2 := -40 \quad b_3 := 5$

$H(x) := b_0 + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 \quad aL := -5 \quad aR := 10$

$N := 50$ $h := \frac{aR - aL}{N}$ $i := 0 .. N - 1$ $x_i := aL + i \cdot h$

$s := 200$ $w(x, y) := \sin(2 \cdot \pi \cdot x) \cdot \sqrt{-2 \cdot \ln(y)}$

Наблюдаемые значения => $y_i := H(x_i) + s \cdot w(rnd(1), rnd(1))$



Моделируем вспомогательную выборку

$n := round\left(\frac{N}{2}\right)$ $n = 25$

$u := 0 .. n - 1$ $V_u := H(0) + s \cdot w(rnd(1), rnd(1))$

Получим независимую оценку дисперсии

$sN := Stdev(V)$ $sN = 191.452$

Оценка параметров модели

Составляем матрицу регрессоров

$F_{i,0} := 1$ $F_{i,1} := (x_i)$ $F_{i,2} := (x_i)^2$

$F_{i,3} := (x_i)^3$ $F_{i,4} := (x_i)^4$

Информационная матрица

$$G := F^{T} \cdot F$$

$$G = \begin{pmatrix} 50 & 117.5 & 1.213 \times 10^{3} & 7.256 \times 10^{3} & 6.417 \times 10^{4} \\ 117.5 & 1.213 \times 10^{3} & 7.256 \times 10^{3} & 6.417 \times 10^{4} & 4.965 \times 10^{5} \\ 1.213 \times 10^{3} & 7.256 \times 10^{3} & 6.417 \times 10^{4} & 4.965 \times 10^{5} & 4.322 \times 10^{6} \\ 7.256 \times 10^{3} & 6.417 \times 10^{4} & 4.965 \times 10^{5} & 4.322 \times 10^{6} & 3.664 \times 10^{7} \\ 6.417 \times 10^{4} & 4.965 \times 10^{5} & 4.322 \times 10^{6} & 3.664 \times 10^{7} & 3.233 \times 10^{8} \end{pmatrix}$$

Матрица ошибок

$$C := G^{-1}$$

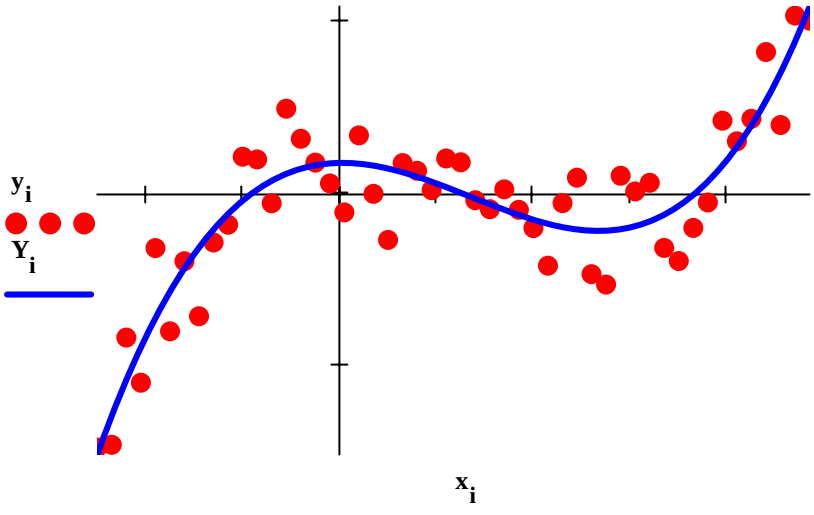
$$C = \begin{pmatrix} 0.06 & -1.585 \times 10^{-3} & -3.433 \times 10^{-3} & 2.136 \times 10^{-4} & 1.213 \times 10^{-5} \\ -1.585 \times 10^{-3} & 0.015 & -6.475 \times 10^{-4} & -6.555 \times 10^{-4} & 6.071 \times 10^{-5} \\ -3.433 \times 10^{-3} & -6.475 \times 10^{-4} & 3.953 \times 10^{-4} & 1.424 \times 10^{-5} & -5.223 \times 10^{-6} \\ 2.136 \times 10^{-4} & -6.555 \times 10^{-4} & 1.424 \times 10^{-5} & 3.572 \times 10^{-5} & -3.274 \times 10^{-6} \\ 1.213 \times 10^{-5} & 6.071 \times 10^{-5} & -5.223 \times 10^{-6} & -3.274 \times 10^{-6} & 3.483 \times 10^{-7} \end{pmatrix}$$

Оценка параметров =>

$$B := C \cdot F^{T} \cdot y \qquad B = \begin{pmatrix} 174.721 \\ 5.066 \\ -42.345 \\ 4.995 \\ 0.032 \end{pmatrix} \qquad b = \begin{pmatrix} 100 \\ 6 \\ -40 \\ 5 \end{pmatrix}$$

Предсказанные значения =>

$$Y := F \cdot B$$



Статистический анализ регрессионной модели

Число параметров модели

$k := rows(B)$ $\quad k = 5$

Средние

$y_{cp} := mean(y)$ $\quad Y_{cp} := mean(Y)$

$y_{cp} = -101.264$ $\quad Y_{cp} = -75.028$

Остатки

$e_i := y_i - Y_i$

Остаточная сумма квадратов

$Q_{ост} := \sum_i (e_i)^2$ $\quad Q_{ост} = 2.181 \times 10^6$

Число степеней свободы

$r := N - k$ $\quad r = 45$

Оценка дисперсии (остаточная дисперсия)

$d := \frac{Q_{ост}}{r}$ $\quad d = 4.847 \times 10^4$

Сумма квадратов, обусловленная уравнением регрессии

$Q_R := \sum_i (Y_i - Y_{cp})^2$ $\quad Q_R = 1.043 \times 10^7$

Дисперсия, обусловленная регрессией

$D_R := \frac{Q_R}{k - 1}$ $\quad D_R = 2.606 \times 10^6$

Полная сумма квадратов

$Q := \sum_i (y_i - y_{cp})^2$ $\quad Q = 1.264 \times 10^7$

Проверка

$Q_R + Q_{ост} = 1.261 \times 10^7$

Проверка адекватности модели

Независимая оценка дисперсии $sN^2 = 3.665 \times 10^4$

Дисперсионное отношение Фишера

$F := \frac{d}{sN^2}$ $F = 1.322$

$L := if\left(F > 1, F, \frac{1}{F}\right)$ $L = 1.322$

Критическая точка $F_c := qF(0.95, N - k, n - 1)$ $F_c = 1.876$

Функция принятия решения

$f := if\left(L \le F_c, "ADEQU", "NO\_ADEQU"\right)$

$f = "ADEQU"$

Проверка гипотезы о значимости коэффициента детерминации

Коэффициент детерминации (множественной корреляции) в случае парной регрессии (функция одной переменной) равен квадрату коэффициента корреляции между наблюдаемыми и предсказанными значениями определяемой переменной. Он является обобщенной (суммарной) мерой качества регрессионной модели.

$R := corr(y, Y)$ $R = 0.909$ $R^2 = 0.827$ $\frac{Q_R}{Q} = 0.825$

Статистика критерия -- отношение дисперсии, обусловленной регрессией к остаточной дисперсии:

$F := \frac{R^2 \cdot (N - k)}{\left(1 - R^2\right) \cdot (k - 1)}$ $F = 53.77$ **или** $\frac{D_R}{d} = 53.77$

$L := if\left(F > 1, F, \frac{1}{F}\right)$ $L = 53.77$

Критическая точка

$F_t := qF(0.95, k-1, N-k)\quad F_t = 2.579$

Функция принятия решения

$f := if(L < F_t, "NE\_ZNATHIM", "ZNATHIM$

$f = "ZNATHIM"$

Проверка гипотезы о значимости коэффициентов модели

$m := 0 .. rows(C) - 1 \qquad l := 0 .. rows(C) - 1$

Дисперсии оценок

$$SB_m := \sqrt{C_{m,m} \cdot d}$$

$SB^T = (54.143 \quad 26.674 \quad 4.377 \quad 1.316 \quad 0.13)$

Задаем уровень значимости и находим критическую точку

$\alpha := 0.05 \qquad t_{kr} := qt\left(1 - \frac{\alpha}{2}, N - k\right) \qquad t_{kr} = 2.014$

Рассчитываем статистику критерия для каждого коэффициента модели

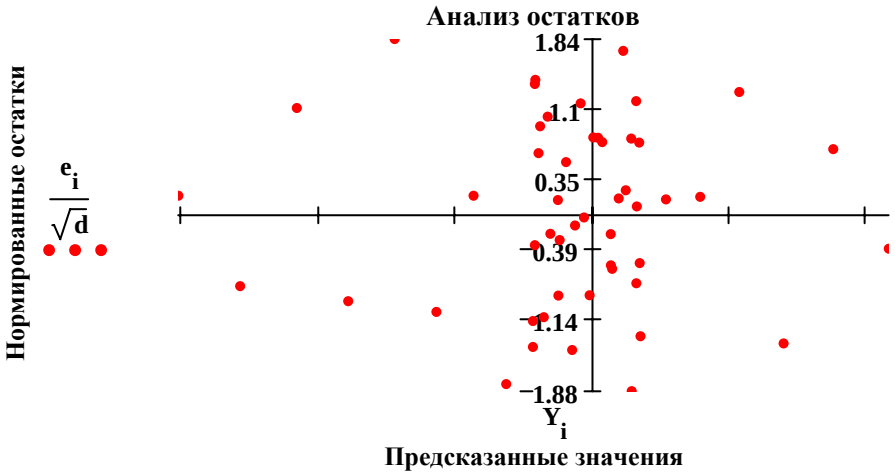
$T_m := \frac{|B_m|}{SB_m} \qquad T^T = (3.227 \quad 0.19 \quad 9.673 \quad 3.796 \quad 0.246)$

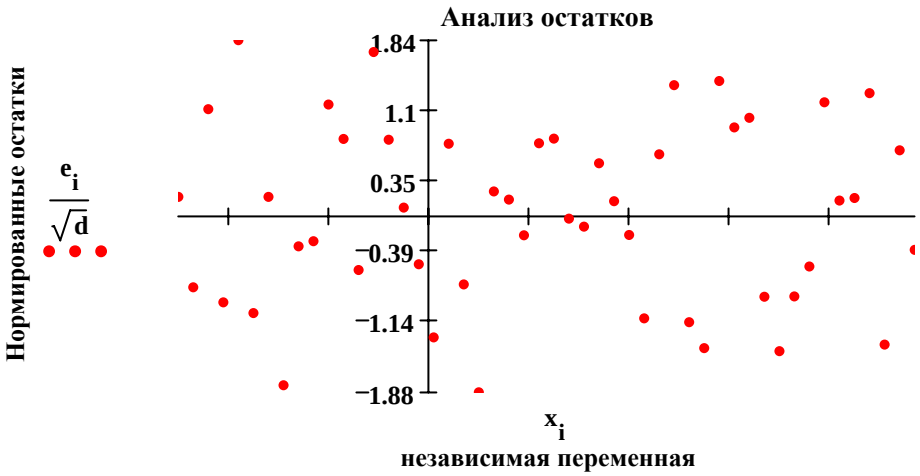
Вводим функцию принятия решения и находим ее значение, на основании которой делаем выводы о значимости коэффициентов. Если коэффициент незначим, то соответствующий член в модели надо исключить.

$Kr_m := if(T_m < t_{kr}, "NeZnathin", "Znathim")$

$$Kr = \begin{pmatrix} "Znathim" \\ "NeZnathin" \\ "Znathim" \\ "Znathim" \\ "NeZnathin" \end{pmatrix}$$

Анализ остатков





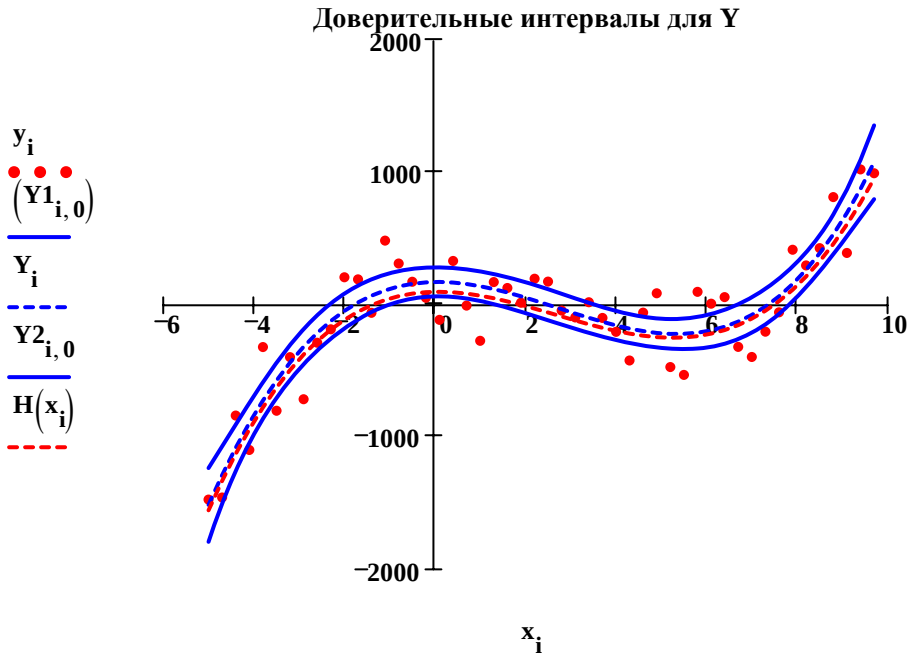
Корреляционная матрица коэффициентов модели

$$K_{l,m} := \frac{C_{l,m}}{\sqrt{C_{l,l} \cdot C_{m,m}}} \qquad K = \begin{pmatrix} 1 & -0.053 & -0.702 & 0.145 & 0.084 \\ -0.053 & 1 & -0.269 & -0.905 & 0.849 \\ -0.702 & -0.269 & 1 & 0.12 & -0.445 \\ 0.145 & -0.905 & 0.12 & 1 & -0.928 \\ 0.084 & 0.849 & -0.445 & -0.928 & 1 \end{pmatrix}$$

$$f(x) := \begin{bmatrix} 1 \\ x \\ (x)^2 \\ (x)^3 \\ x^4 \end{bmatrix} \qquad f(x) := \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \end{pmatrix}^{\blacksquare}$$

$$Y2_i := Y_i + t_{kr} \cdot \sqrt{\left(d \cdot \left(f\left(x_i\right)\right)^T \cdot C \cdot f\left(x_i\right)\right)_{0,0}}$$

$$Y1_i := Y_i - t_{kr} \cdot \sqrt{\left(d \cdot \left(f\left(x_i\right)\right)^T \cdot C \cdot f\left(x_i\right)\right)_{0,0}} \qquad \sqrt{d} = 220.169$$



Галанов Ю.И.

Лабораторный практикум по мат. статистике

Дисперсионный факторный анализ

Задание.

1. Смоделировать K выборок из нормального распределения
2. Проверить гипотезу H_0 о равенстве групповых средних с помощью дисперсионного факторного анализа.
3. Оцените вероятность выявить наличие эффекта обработки (мощность критерия) от силы связи (задается параметром **к**) и общего объема выборки, задаваемого параметром **n0.**

Контрольные вопросы:

1. Как влияет шкала измерений на проведение факторного анализа?
2. Перечислите условия применимости дисперсионного анализ?

Программа расчета

1. Формируем набор выборок

Число выборок $K := 10$ $j := 0 .. K - 1$

Минимальный объем выборки $n0 := 20$

Объемы выборок $n_j := n0 + round(rnd(n0))$ $N := \sum_j n_j$ $N =$

$n^T =$

Фиксируем объемы выборок и блокируем их изменение

$n_j :=$

20
24
32
27
36
23
34
26
22
23

$N := \sum_j n_j$ $N = 267$

эффект обработки $k := 0.05$ $a_j := 50 + k \cdot j$ $S := 2$

$$X^{\langle j \rangle} := \left| \begin{array}{l} x \leftarrow 0 \\ \text{for } i \in 0 .. n_j - 1 \\ \quad x_i \leftarrow a_j + S \cdot \sin(2 \cdot \pi \cdot rnd(1)) \cdot \sqrt{-2 \cdot \ln(rnd(1))} \\ x \end{array} \right.$$

X =

	0	1	2	3	4	5	6	7
0	50.973	49.56	50.148	49.246	50.25	52.808	49.583	51.427
1	51.12	50.005	52.096	49.142	52.596	52.075	48.462	43.595
2	53.299	49.908	51.661	50.084	49.774	50.914	52.634	50.414
3	51.842	47.867	53.418	51.59	53.652	44.506	49.077	50.942
4	51.537	48.514	51.659	50.298	52.622	50.799	49.491	49.648
5	54.153	50.715	49.879	51.399	52.482	50.667	51.376	49.17
6	49.063	45.496	54.46	52.231	48.074	48.172	51.935	50.991
7	49.843	49.011	50.55	49.771	47.435	51.725	46.114	47.544
8	51.142	51.353	46.503	50.73	51.933	49.08	50.443	45.654
9	44.997	52.84	49.461	46.413	45.094	49.27	48.629	52.081
10	50.936	49.476	49.339	50.809	51.223	48.222	50.293	50.121
11	47.441	50.314	50.382	48.366	49.698	46.382	49.043	50.033
12	51.71	49.155	51.268	51.386	50.647	50.229	55.193	49.711
13	47.83	50.967	47.203	51.253	47.755	46.763	49.203	50.976
14	50.143	49.21	53.156	48.964	49.256	49.157	49.119	53.431
15	51.161	49.612	50.112	52.31	49.753	47.677	48.661	53.623
16	47.237	46.555	52.384	48.863	46.901	53.898	51.262	51.027
17	48.854	49.981	48.837	49.541	52.029	48.355	49.298	49.981
18	51.547	51.006	51.067	54.337	50.766	52.089	49.654	50.009
19	50.252	50.786	50.11	49.486	48.95	49.974	52.517	50.338
20	0	52.367	50.077	47.928	49.79	48.387	55.201	47.062
21	0	45.125	49.084	50.754	52.526	52.907	49.327	...

1. Обработка результатов

Оцениваем внтригрупповую дисперсию D1

Групповые средние

$$M_j := \frac{1}{n_j} \cdot \sum_{i=0}^{n_j - 1} X_{i,j}$$

$M^T =$

	0	1	2	3	4	5	6	7	8
0	50.254	49.512	50.511	50.346	50.138	49.788	49.958	49.926	...

Число степеней свободы R для расчета внутригрупповой дисперсии

$$R := \sum_j \left(n_j - 1\right) \qquad R = 257$$

$$D_j := \frac{1}{n_j}\cdot\sum_{i=0}^{n_j-1}\left(X_{i,j} - M_j\right)^2 \qquad D1 := \frac{1}{R}\cdot\sum_{j}\left(n_j\cdot D_j\right) \qquad D1 = 3.994$$

Общее среднее $X00 := \frac{1}{N}\cdot\sum_{j}\left(n_j\cdot M_j\right)$

Доверительная "ошибка" для общего среднего (согласно H0):

$\gamma := 0.999$ $\qquad \Delta := \sqrt{D1}\cdot\frac{qt\left(\frac{1+\gamma}{2}, R\right)}{\sqrt{N}} \qquad \Delta = 0.407$

Оцениваем межгрупповую дисперсию D2

$$D2 := \frac{1}{K-1}\cdot\sum_{j}\left[n_j\cdot\left(M_j - X00\right)^2\right] \qquad D2 = 2.564$$

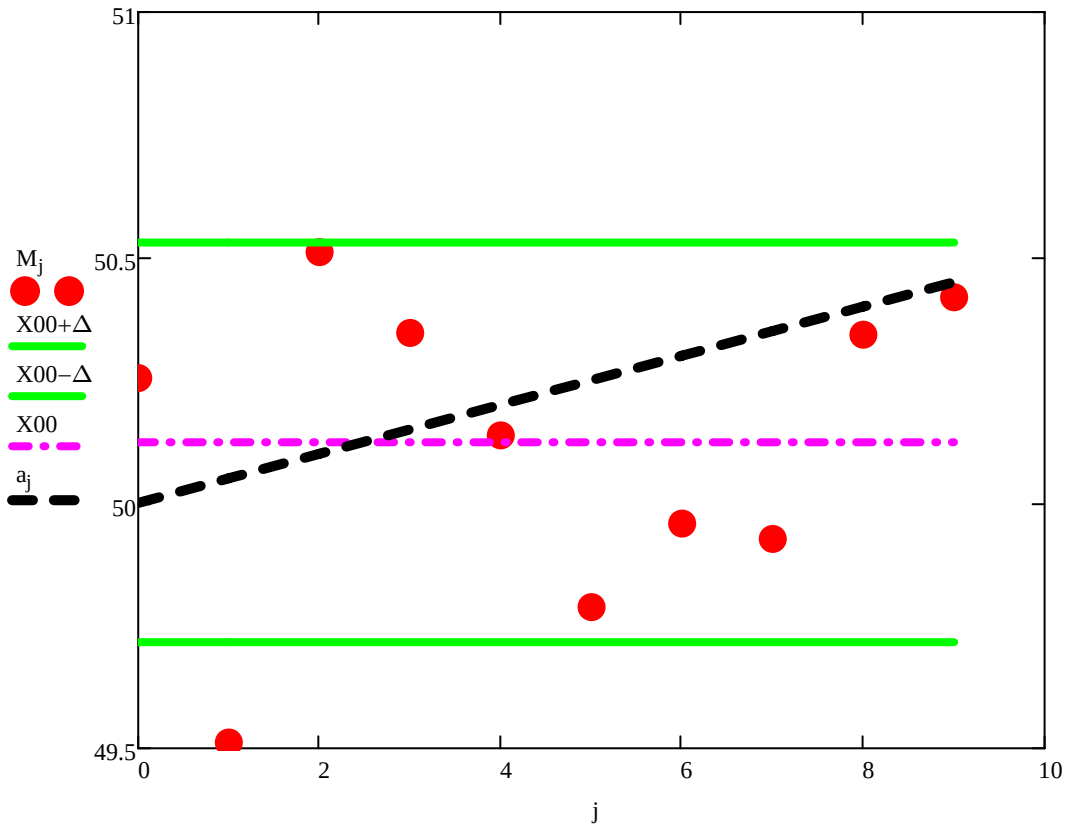
Вычисляем значение статистики критерия F (дисперсионное отношение Фишера)

$D1 = 3.994$ $\qquad D2 = 2.564$ $\qquad F := \frac{D2}{D1}$ $\qquad F = 0.642$

Находим границу критической области (критическую точку C)

$C := qF(0.95, R, K-1)$ $\qquad F = 0.642$ $\qquad C = 2.726$

Сравнивая C и F, делаем выводы о принятии или отклонении нулевой гипотезы



Указания

1. Содержание отчета:

- Теоретическое введение: Задачи и стратегия факторного анализа, двеоценки дисперсии, дисперсионное отношение Фишера, порядок проверки гипотезы.
- Программа расчетов с подробным комментарием последовательности и результатов выполняемых операций.
- Обсуждение результатов и выводы

2. Оцените вероятность выявить наличие эффекта обработки (мощность критерия) от силы связи (задается параметром **к**) и общего объема выборки, задаваемого параметром **n0.**
(10 раз перезапустите вычисления и подсчитайте число положительных исходов. Результаты занесите в таблицу).
3. Параметры, подлежащие изменениям, переместите в удобное место и переопределите с помощью "глобального" определения (▮ ≡ ▮) -- клавиша "~" -тильда (Ё).

n0	k			
	0.05	0.1	0.25	0.5
10				
20				
50				

4. Таблицу можно вставить в MathCad -документ в виде объекта MsWord, OpenOffice, Excel и тп., те. тех программных продуктов, *УСТНОВЛЕННЫХ НА ВАШЕМ КОМПЬЮТЕРЕ*, в которых можно построить таблицу. Еще проще -- объявите матрицу без имени необходимой размерности и заполняйте клетки! Текст вносите в кавычках **""** как символьную переменную (Правда, здесь нельзя объединять ячейки).

"Столбец 1"	"Столбец 1"	"Столбец 1"	"Столбец 1"	"Столбец 1"
"Строка1 "	111	▮	▮	▮
"Строка2 "	▮	222	▮	▮
"Строка3 "	▮	▮	33	▮
"Строка4 "	▮	▮	▮	44

В 2001-м MCD : Insert-Component-Input Table.

Галанов Ю.И.

Лабораторный практикум по мат. статистике

Однофакторный анализ

ПРИМЕР.

Для выяснения влияния денежного стимулирования на производительность труда шести однородным группам из 5 человек каждая были предложены задачи одинаковой трудности. Задачи предлагались каждому испытуемому независимо от всех остальных. Группы отличаются между собой величиной денежного вознаграждения за решаемую задачу.

В таблице приведено число решенных задач членами каждой группы.

Проверяем гипотезу об отсутствии влияния денежного вознаграждения на число решенных задач.

Вводим данные:

$$M := READPRN("krask.prn")$$

$$M = \begin{pmatrix} 10 & 8 & 12 & 12 & 24 & 19 \\ 11 & 10 & 17 & 15 & 16 & 18 \\ 9 & 16 & 14 & 16 & 22 & 27 \\ 13 & 13 & 9 & 16 & 18 & 25 \\ 7 & 12 & 16 & 19 & 20 & 24 \end{pmatrix}$$

$$k := 0 .. cols(M) - 1$$

$$j := 0 .. cols(M) - 1$$

$$l := 0 .. rows(M) - 1$$

$$i := 0 .. rows(M) - 1$$

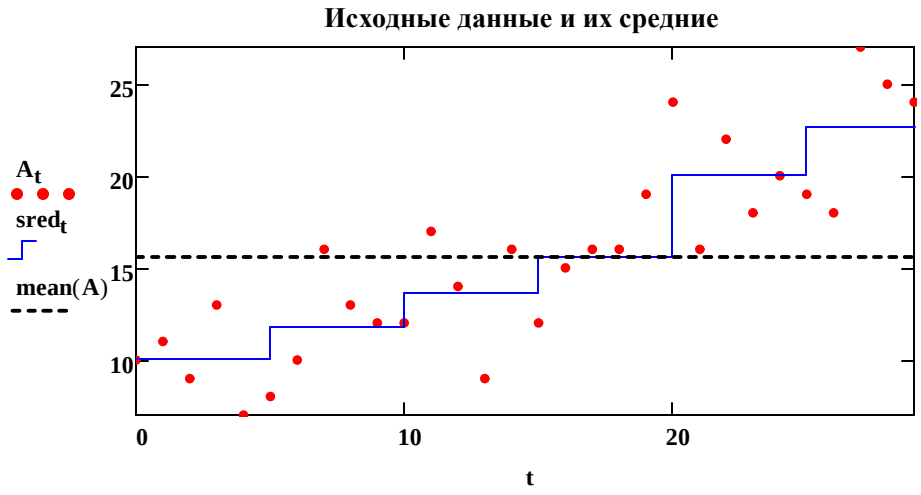
Представим их в виде одной выборки

$$A_{i+5\cdot j} := M_{i,j}$$

Посчитаем средние для каждого столбца:

$$c_j := \frac{1}{5}\cdot\sum_i M_{i,j} \qquad t := 0 \,..\, last(A)$$

$$sred_t := if\left(t < 5, c_0, if\left(t < 10, c_1, if\left(t < 15, c_2, if\left(t < 20, c_3, if\left(t < 25, c_4, if\left(t < 30, c_5, c_5\right)\right)\right)\right)\right)\right)$$



Уже на данной стадии анализа виден сдвиг средних при переходе от первой группы к последней, однако необходимо убедиться в том, что эти сдвиги значимы.

Перейдем к ранговому представлению данных:

$$R_{i,j} := \sum_l \sum_k if\left(M_{l,k} - M_{i,j} > 0, 0, 1\right)$$

Получим таблицу рангов

$$M = \begin{pmatrix} 10 & 8 & 12 & 12 & 24 & 19 \\ 11 & 10 & 17 & 15 & 16 & 18 \\ 9 & 16 & 14 & 16 & 22 & 27 \\ 13 & 13 & 9 & 16 & 18 & 25 \\ 7 & 12 & 16 & 19 & 20 & 24 \end{pmatrix} \qquad R = \begin{pmatrix} 6 & 2 & 10 & 10 & 28 & 24 \\ 7 & 6 & 20 & 14 & 19 & 22 \\ 4 & 19 & 13 & 19 & 26 & 30 \\ 12 & 12 & 4 & 19 & 22 & 29 \\ 1 & 10 & 19 & 24 & 25 & 28 \end{pmatrix}$$

Повторяющиеся ранги заменяем средними рангами

$R_{0,0} := 5.5$ $R_{1,1} := 5.5$ $R_{3,0} := 11.5$ $R_{3,1} := 11.5$

$R_{0,5} := 23.5$ $R_{4,3} := 23.5$ $R_{2,0} := 3.5$ $R_{3,2} := 3.5$

$R_{1,5} := 21.5$ $R_{3,4} := 21.5$ $R_{0,4} := 27.5$ $R_{4,5} := 27.5$

$R_{4,1} := 9$ $R_{0,2} := 9$ $R_{0,3} := 9$

$R_{2,1} := 17$ $R_{2,3} := 17$ $R_{4,2} := 17$ $R_{3,3} := 17$ $R_{1,4} := 17$

$$R = \begin{pmatrix} 5.5 & 2 & 9 & 9 & 27.5 & 23.5 \\ 7 & 5.5 & 20 & 14 & 17 & 21.5 \\ 3.5 & 17 & 13 & 17 & 26 & 30 \\ 11.5 & 11.5 & 3.5 & 17 & 21.5 & 29 \\ 1 & 9 & 17 & 23.5 & 25 & 27.5 \end{pmatrix}$$

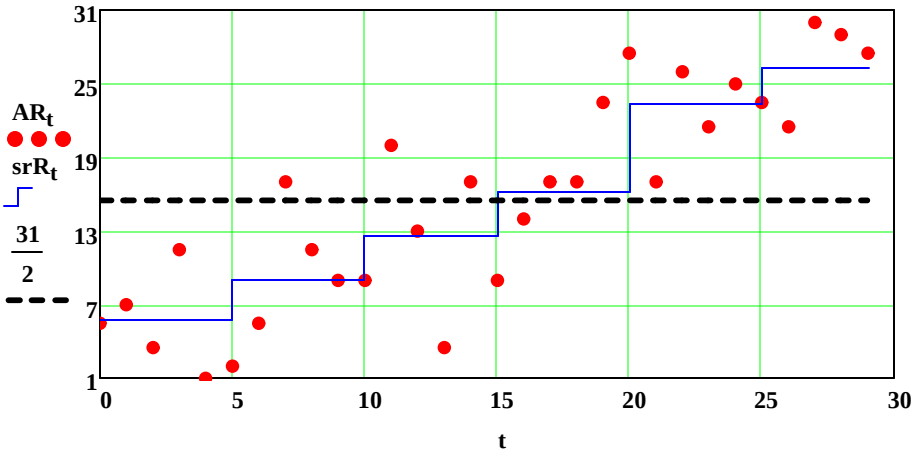
средние по столбцам ранги

$$Rj_j := \frac{1}{5} \cdot \sum_i R_{i,j}$$

$Rj^T = (5.7 \quad 9 \quad 12.5 \quad 16.1 \quad 23.4 \quad 26.3)$

$srR_t := \text{if}(t < 5, Rj_0, \text{if}(t < 10, Rj_1, \text{if}(t < 15, Rj_2, \text{if}(t < 20, Rj_3, \text{if}(t < 25, Rj_4, \text{if}(t < 30, Rj_5, c_5))))))$

$AR_{i+5 \cdot j} := R_{i,j}$



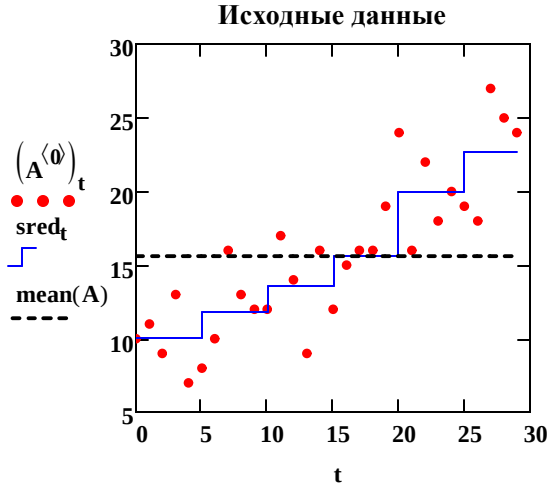
Проверим нулевую гипотезу:

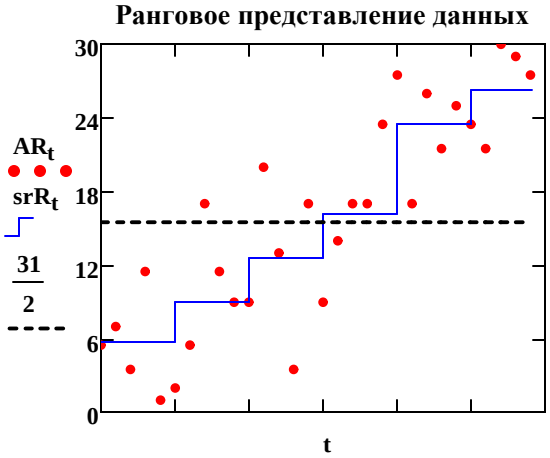
Данные, представленные в таблице принадлежат одной выборке.

Посчитаем статистику Краскела-Уоллеса:

$$H := \frac{12 \cdot 5}{30 \cdot 31} \cdot \sum_j \left(Rj_j - \frac{31}{2} \right)^2 \qquad H = 21.077419 \qquad qchisq(0.999, 5) = 20.515006$$

На уровне значимости 0.001 нулевая гипотеза отвергается





При переходе от количественой шкалы к порядковой наблюдается некоторая потеря информации, которая проявляется в большем рассеивании данных. Однако это не сказывается на окончательных результатах.

Особенность ранговых методов состоит в том, что распределение ранговых статистик не зависит от распределения исходных данных.

Поэтому они позволяют обрабатывать данные, изначально представленные в порядковых или номинальных шкалах.

Эта функция возвращает исправленное значене рангов

```
f3(x) :=
  n ← last(x)
  for i ∈ 0..n
    x_{i,1} ← i
  x ← csort(x, 0)
  for i ∈ 0..n
    x_{i,2} ← i + 1
  i ← 0
  while i < n
    if x_{i,0} ≠ x_{i+1,0}
      x_{i,3} ← 1
      i ← i + 1
    otherwise
      k ← 1
      M ← x_{i,2}
      while x_{i,0} = x_{i+1,0}
        k ← k + 1
        i ← i + 1
        M ← M + x_{i,2}
      d ← k
      for j ∈ 0..k − 1
        x_{i−j,2} ← M/d
      i ← i + 1
  x ← csort(x, 1)
  x^⟨2⟩
```

$R1 := f3(A)$ $\quad M1_{i,j} := R1_{i+5\cdot j}$

$$M1 = \begin{pmatrix} 5.5 & 2 & 9 & 9 & 27.5 & 23.5 \\ 7 & 5.5 & 20 & 14 & 17 & 21.5 \\ 3.5 & 17 & 13 & 17 & 26 & 30 \\ 11.5 & 11.5 & 3.5 & 17 & 21.5 & 29 \\ 1 & 9 & 17 & 23.5 & 25 & 27.5 \end{pmatrix}$$

$$R = \begin{pmatrix} 5.5 & 2 & 9 & 9 & 27.5 & 23.5 \\ 7 & 5.5 & 20 & 14 & 17 & 21.5 \\ 3.5 & 17 & 13 & 17 & 26 & 30 \\ 11.5 & 11.5 & 3.5 & 17 & 21.5 & 29 \\ 1 & 9 & 17 & 23.5 & 25 & 27.5 \end{pmatrix}$$

Проверочные суммы

$$\sum_t R1_t = 465 \qquad 30\cdot\frac{31}{2} = 465$$