SECTION 1. Theoretical research in mathematics.

Zhunisbekov Sagat
doctor of technical Sciences, Professor,
academician of the National Engineering Academy
rector Taraz Technical Institute, Kazakhstan
Arne Jönsson
deputy CEO and manages the language technology research activities at
Sics East Swedish ICT,
Professor of Computer Science, Linköping University,
Director of undergraduate studies for the Cognitive Science program,
Deputy CEO at Santa Anna,Sweden
Shevtsov Alexandr Nikolayevich
candidate of Technical Sciences,
President, Theoretical & Applied Science, LLP
associate Professor of the Department «Applied mathematics»,
Taraz State University named after M.H. Dulati, Kazakhstan

ABOUT SOME CLOUD CHI-SQUARE CRITERION PEARSON

The purpose of this article is to study the field of application of the distribution, resulting from the distribution of Pearson assuming some errors.

Keywords: distribution, comparison, criterion.

О НЕКОТОРЫХ ОБЛАКАХ ТОЧЕК ХИ-КВАДРАТ КРИТЕРИЯ ПИРСОНА

Цель данной статьи – исследование области применения распределения, возникающего из распределения Пирсона при допущении некоторых ошибок.

Ключевые слова: распределение, сравнение, критерий.

Для исследования законов распределения случайных величин и их гипотез могут использоваться различные критерии: Z-тест**;** t-критерий Стьюдента; Критерий Фишера**;** Критерий Пирсона (Хи-квадрат)**;** Критерий согласия Колмогорова; Тест Вальда; U-критерий Манна — Уитни; Критерий Уилкоксона; Критерий Краскела — Уоллиса; Критерий Кохрена; Критерий Лиллиефорса [1, с.1]. Критерий Пирсона, или критерий χ^2 (Хи-квадрат) — наиболее часто употребляется – как критерий для проверки гипотезы о законе распределения. Во многих практических задачах точный

закон распределения неизвестен, то есть является гипотезой, которая требует статистической проверки.

При подобных проверках некоторые ученые и исследователи выбирают меру относительно экспериментальной гипотезы, хотя в критерии Пирсона четко описана процедура выбора меры именно относительно теоретической модели статистического процесса. Естественно возникает вопрос о степени применимости данного изменения в критерии χ^2. Как изменится и насколько результат? Не приведет ли данная ошибка к сильным искажениям результата сравнения гипотез? Практически мы ведь получаем уже не распределение Пирсона, а какое-то новое распределение, обозначим его как $\tilde{\chi}^2$, тогда получим:

$$\chi^2 = \sum_{j=1}^{k} \frac{(Q_j - E_j)^2}{E_j}, \qquad \text{(критерий Пирсона)}$$

где Q_j - гипотеза полученная в ходе эксперимента, E_j - гипотеза взятая из теоретичесой модели процесса, k - размерность пространства;

и новое распределение:

$$\tilde{\chi}^2 = \sum_{j=1}^{k} \frac{(E_j - Q_j)^2}{Q_j}.$$

Данные формулы представляют собой, не что иное, как Евклидово расстояние деленное на норму, только выбор нормы основан в первом случае на теоретической модели, а во втором на эксперименте.

Рассмотрим простой пример с бросанием монеты[2, с.1]. Если взять случайную выборку 100 бросаний, где количество выпадений орла и решки примерно одинаково (встречаются с одинаковой частотой), то в наблюдаемой выборке отношение количества орла и решки будет соотноситься с частотой как и во всей генеральной выборке(50/50). Пусть в наблюдаемой выборке 46 орлов и 54 решек, тогда число степеней свобод $k-1=2-1=1$ и

$$\chi^2 = \sum_{j=1}^{k} \frac{(Q_j - E_j)^2}{E_j} = \frac{(46-50)^2}{50} + \frac{(54-50)^2}{50} = 0.64,$$

$$\tilde{\chi}^2 = \sum_{j=1}^{k} \frac{(Q_j - E_j)^2}{E_j} = \frac{(50-46)^2}{46} + \frac{(50-54)^2}{54} = 0.6441223833.$$

В общем случае, для $k=2$ получим:

$$\chi^2 = \frac{(i-a)^2}{a} + \frac{(j-b)^2}{b},$$

$$\tilde{\chi}^2 = \frac{(a-i)^2}{i} + \frac{(b-j)^2}{j},$$

Тогда с учетом того, что

$$P(a)+P(b)=\Omega$$

получим следующие графики распределения отдельных точек облака, рис.1-9

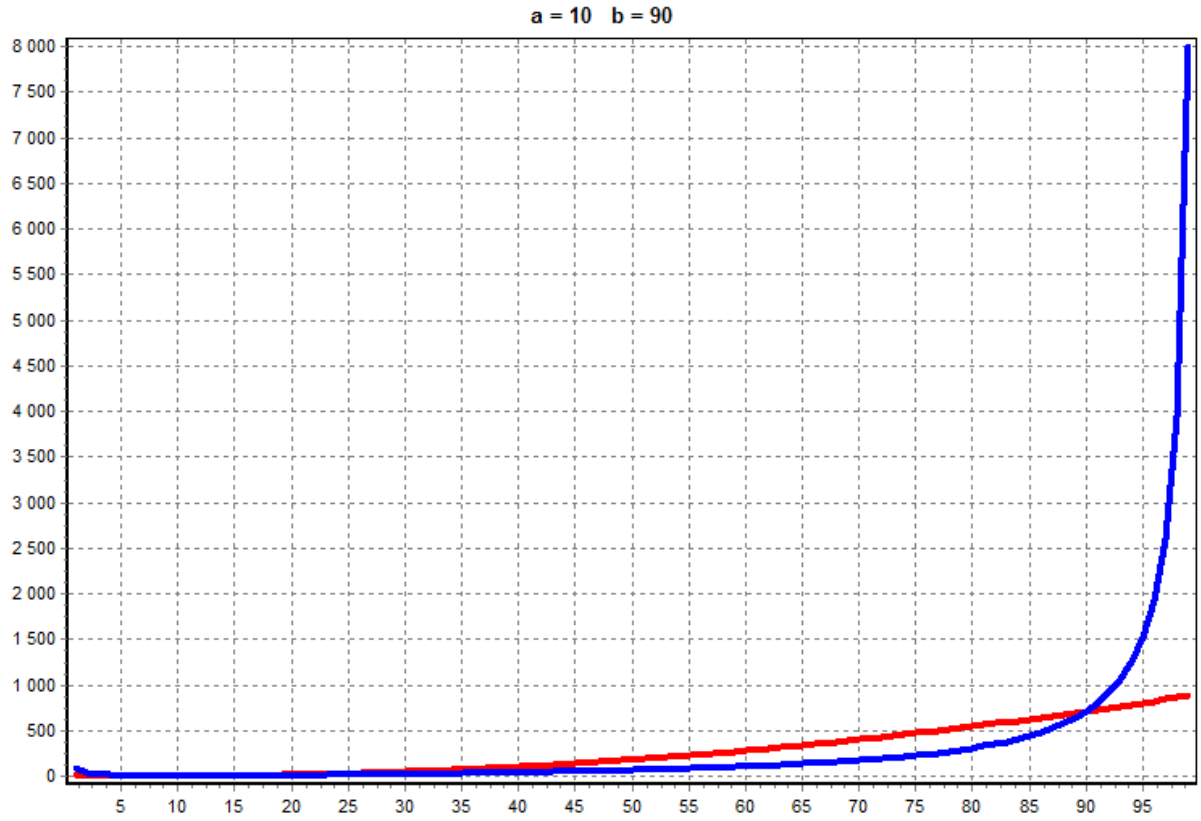


Рисунок 1 – Распределение χ^2 и $\tilde{\chi}^2$ (соответственно, красный и синий).

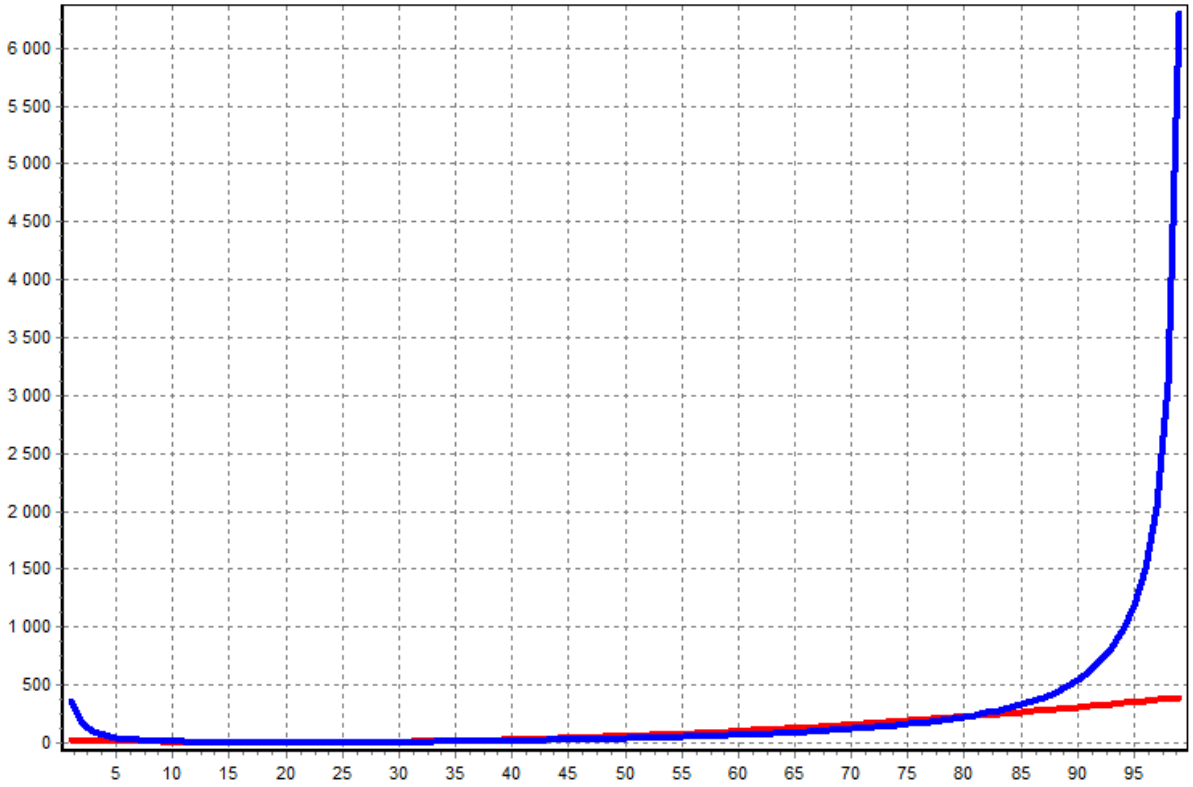


Рисунок 2 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=20, b=80$).

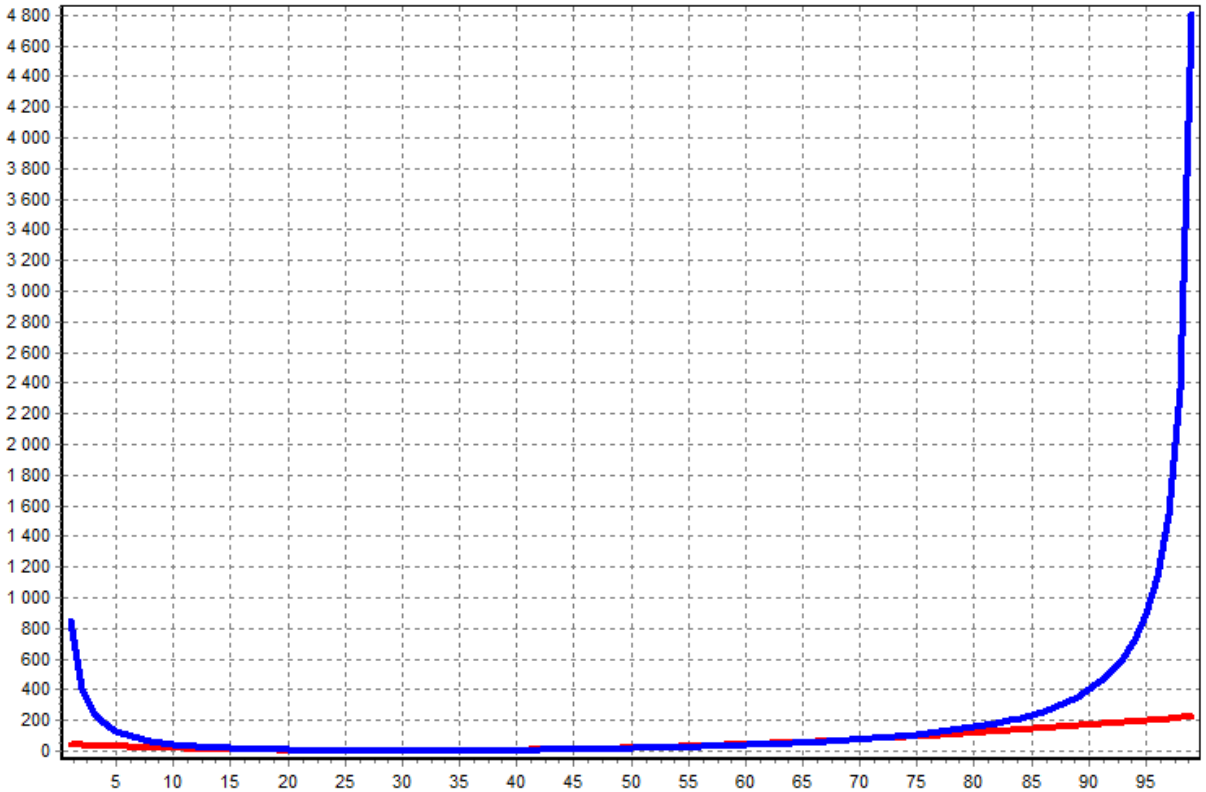


Рисунок 3 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=30, b=70$).

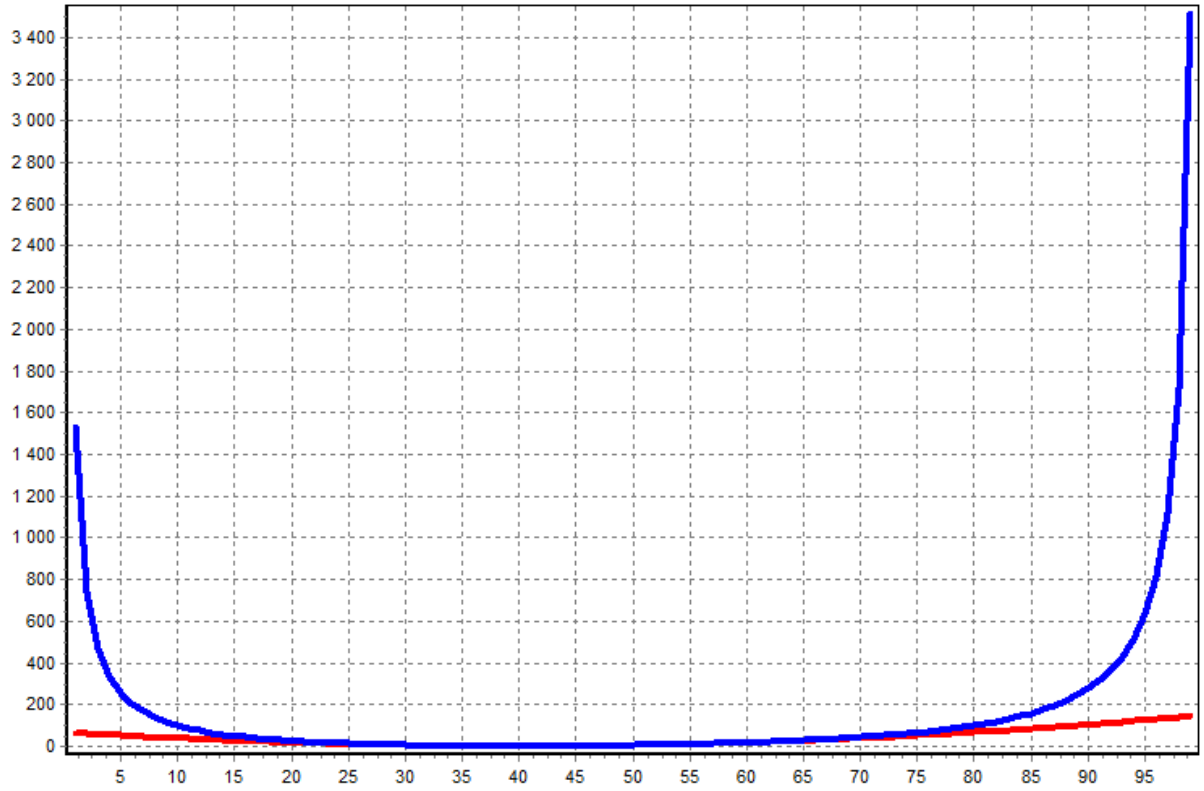


Рисунок 4 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 40, b = 60$).

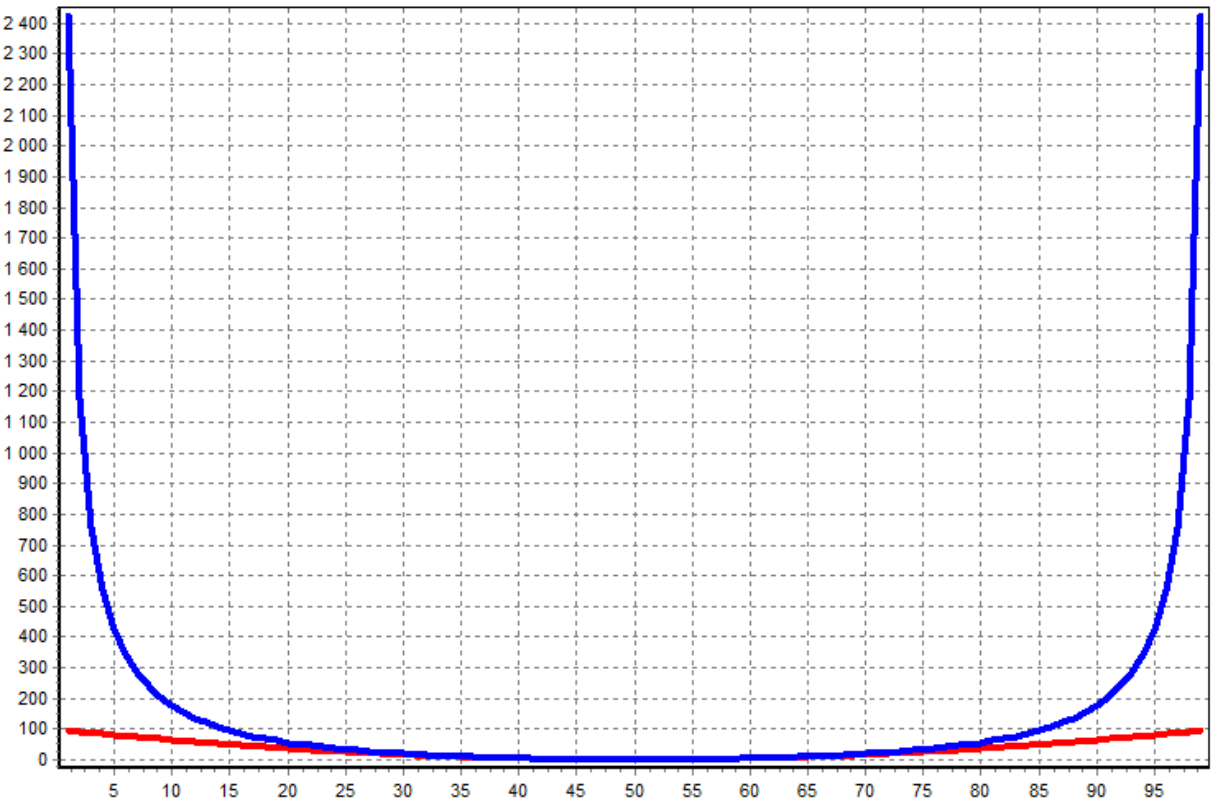


Рисунок 5 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 50, b = 50$).

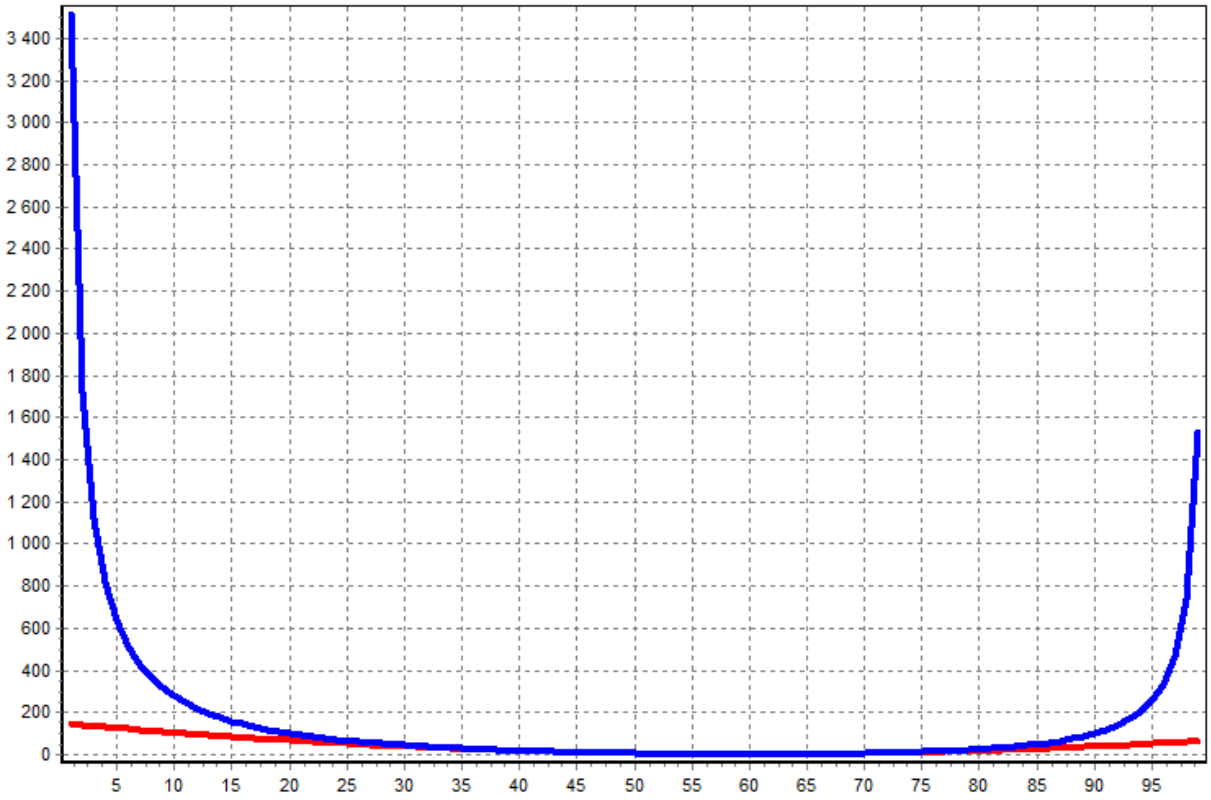


Рисунок 6 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 60, b = 40$).

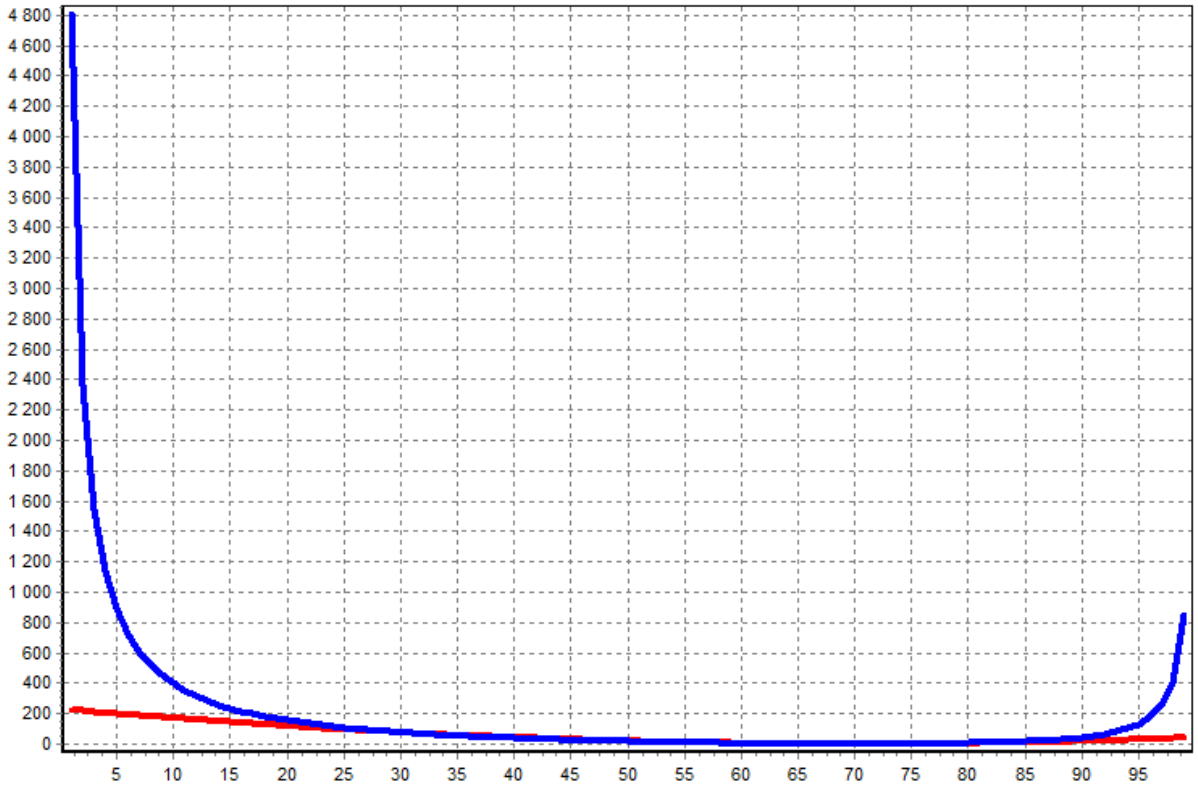


Рисунок 7 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 70, b = 30$).

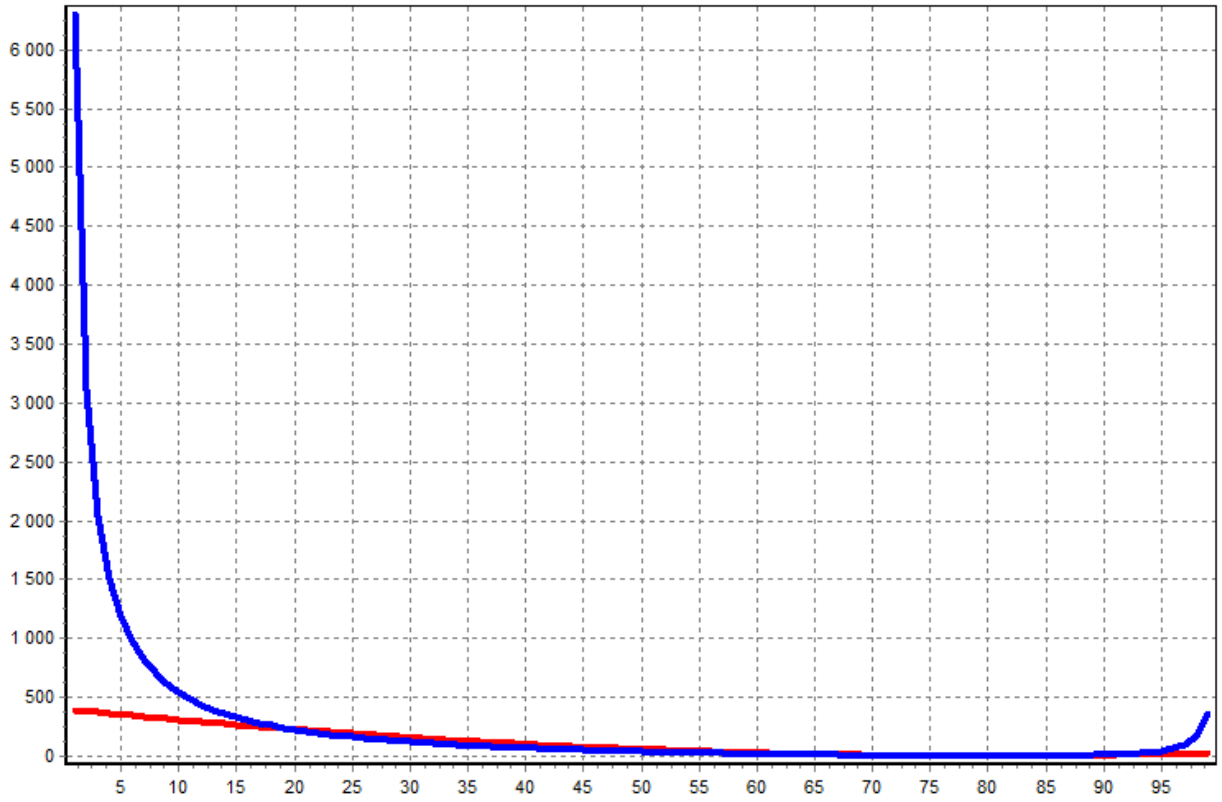


Рисунок 8 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=80, b=20$).

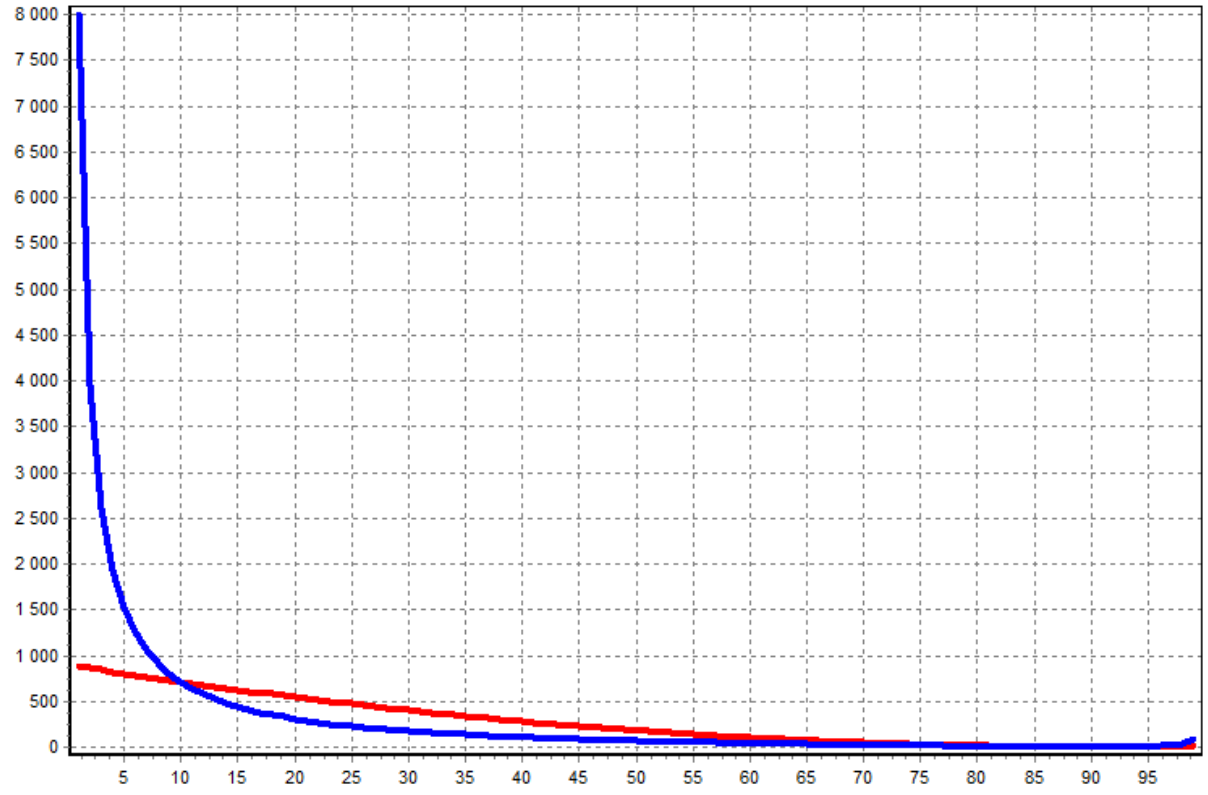


Рисунок 9 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=90, b=10$).

На границе $\tilde{\chi}^2 \to \infty$ но в зависимости от коэффициентов скорость меняется, рассмотрим графики в пределах от 10 до 90, при этом большая часть бесконечных значений будет за пределами графика.

code: Delphi

```
var
  I,j,a,b: Integer;
  hi,hi2:real;
{$R *.dfm}

procedure TForm1.Button1Click(Sender: TObject);
begin
b:=100-a;
for I := 10 to 90 do
begin
j:=100-i;

hi:=sqr(i-a)/a+ sqr(j-b)/b ;
hi2:=sqr(i-a)/i+ sqr(j-b)/j  ;

series3.AddXY(i,(int(hi*100)/100));
series4.AddXY(i,(int(hi2*100)/100));
application.ProcessMessages;
end;
chart1.Title.Caption:='a = '+inttostr(a)+'   b = '+inttostr(b);
end;

procedure TForm1.Button2Click(Sender: TObject);
var i:integer;
begin                //series3.Clear;    //series4.Clear;
for i := 1 to 40 do
begin
a:=a+2;   button1.Click;
end;
end;

procedure TForm1.FormCreate(Sender: TObject);
begin
a:=10;
button1.Click;
end;
```

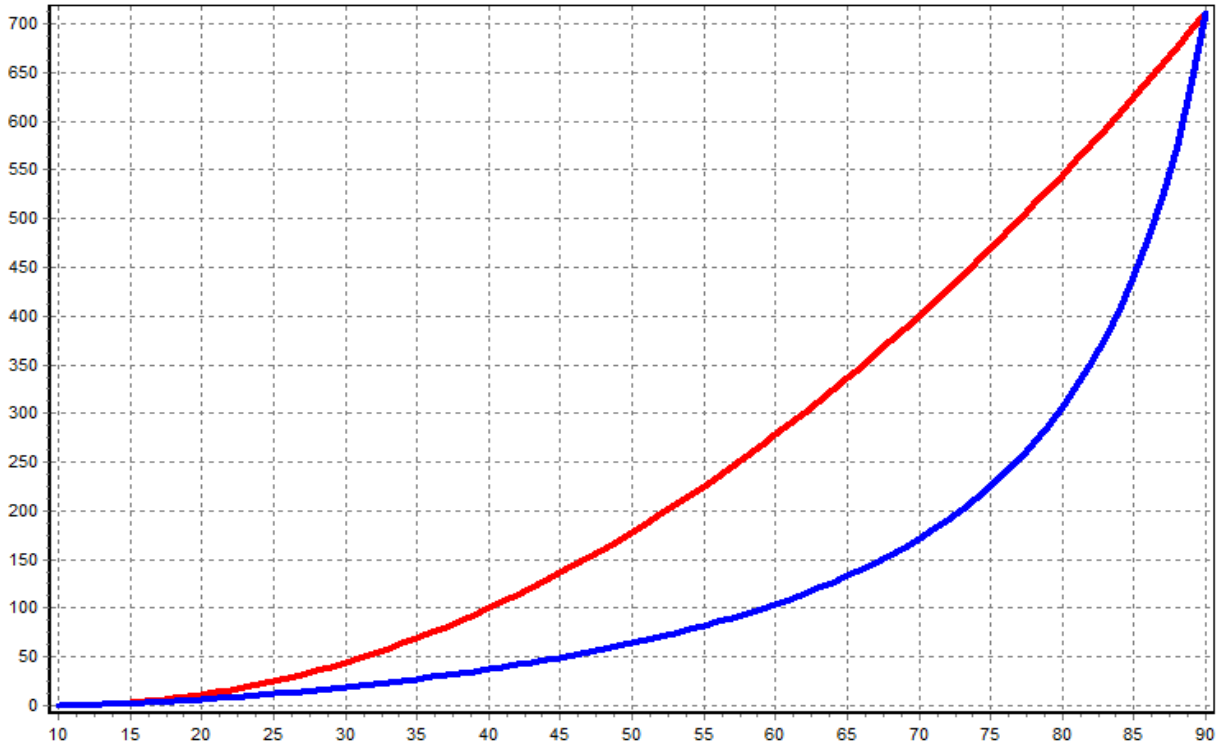


Рисунок 10 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=10, b=90$).

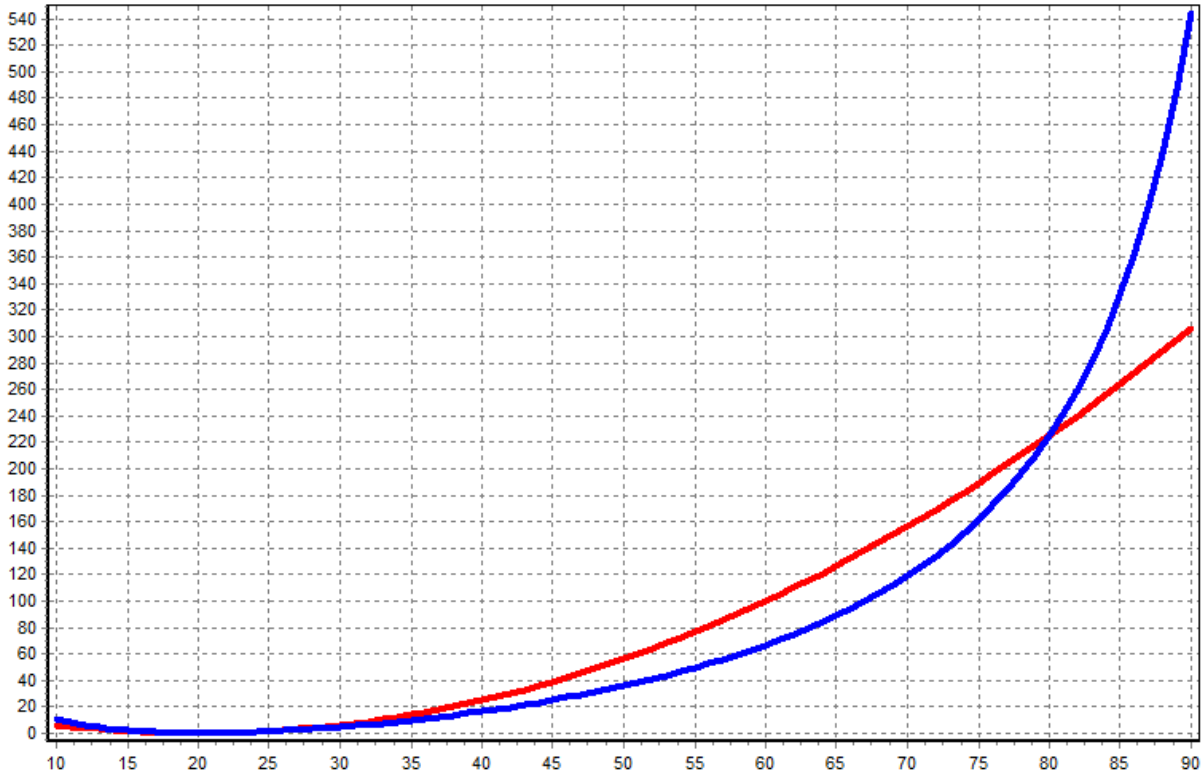


Рисунок 11 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=20, b=80$).

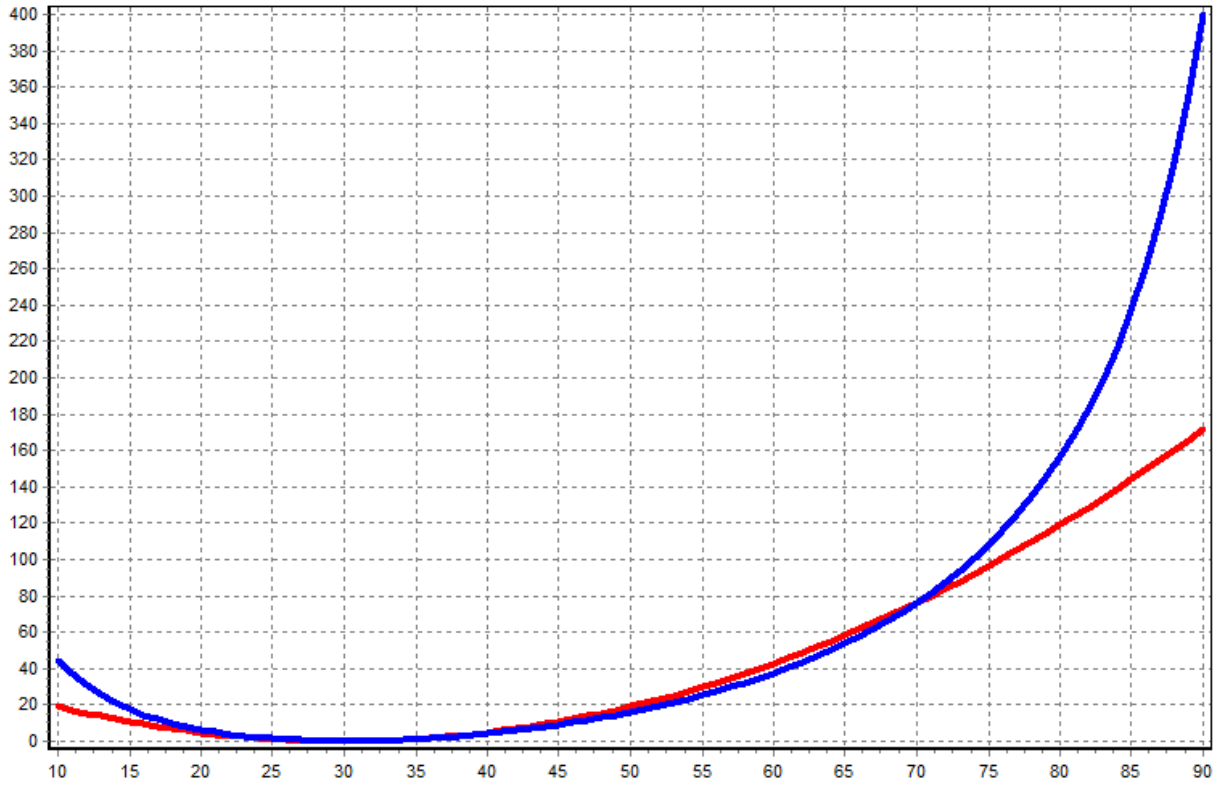


Рисунок 12 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 30, b = 70$).

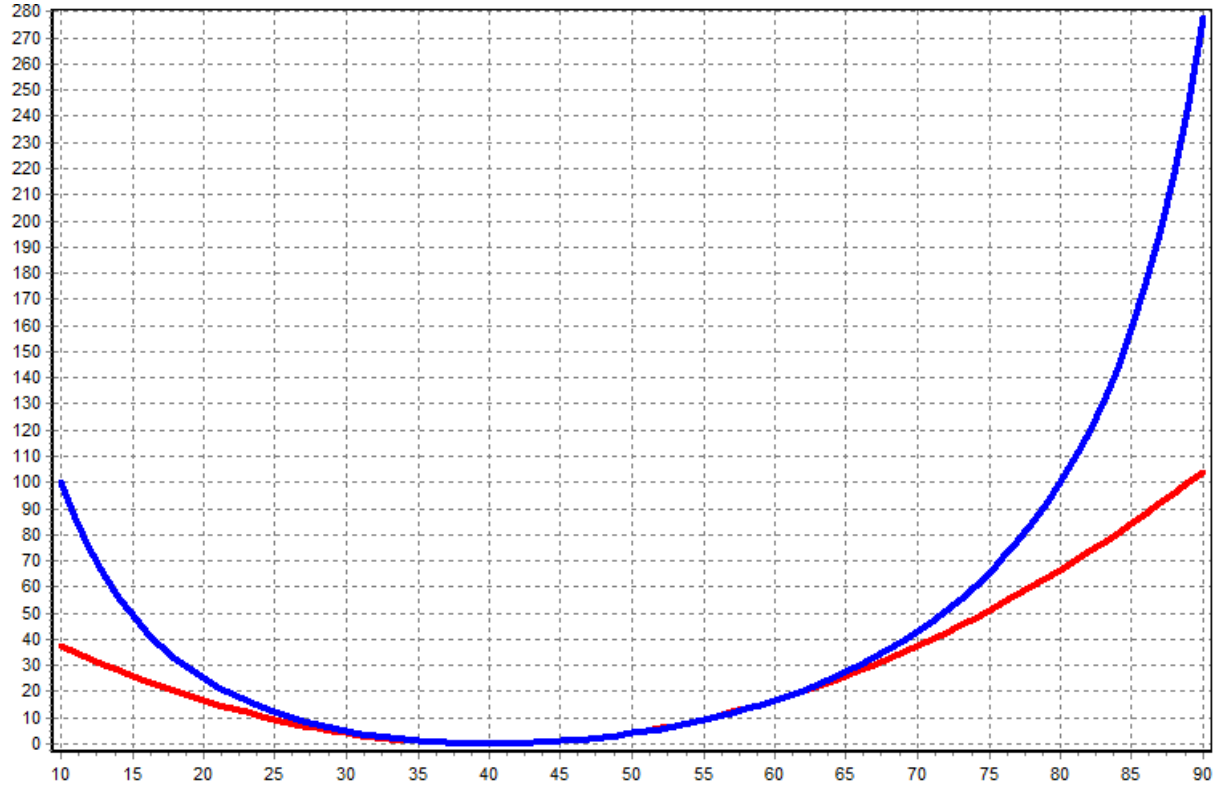


Рисунок 13 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 40, b = 60$).

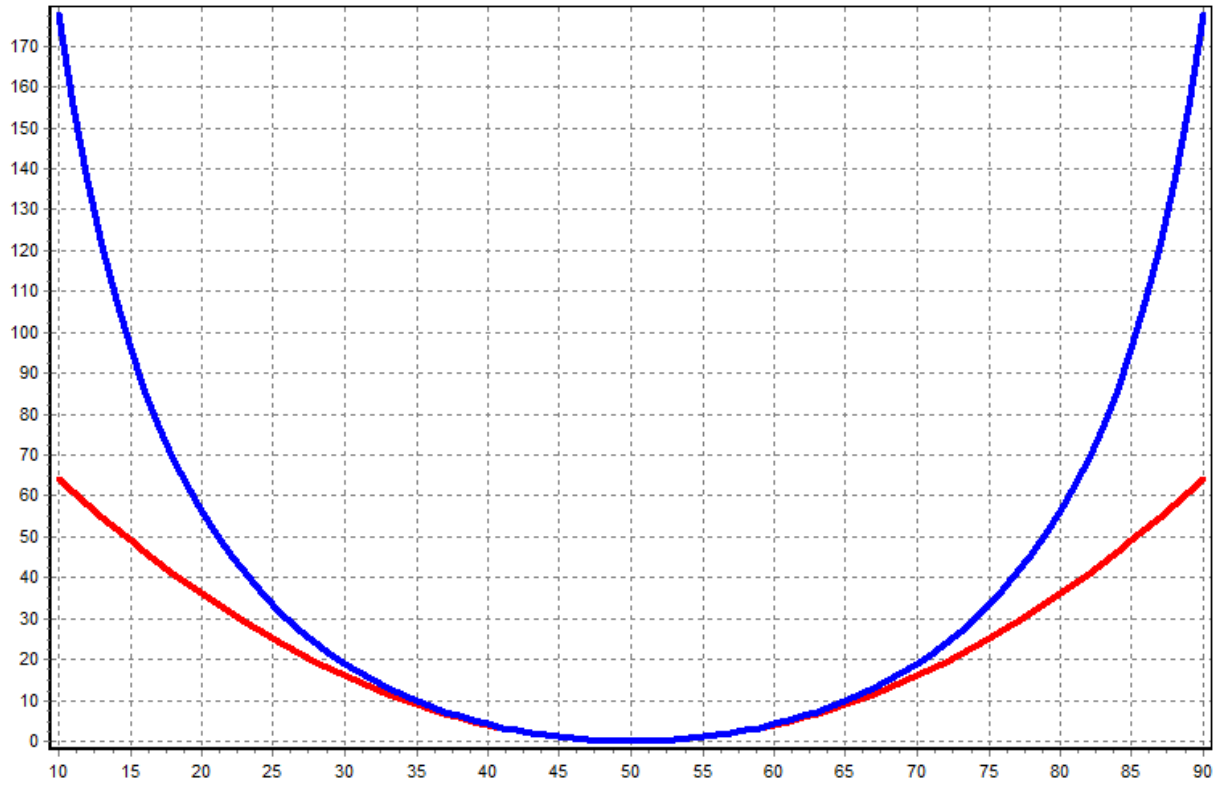


Рисунок 14 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 50, b = 50$).

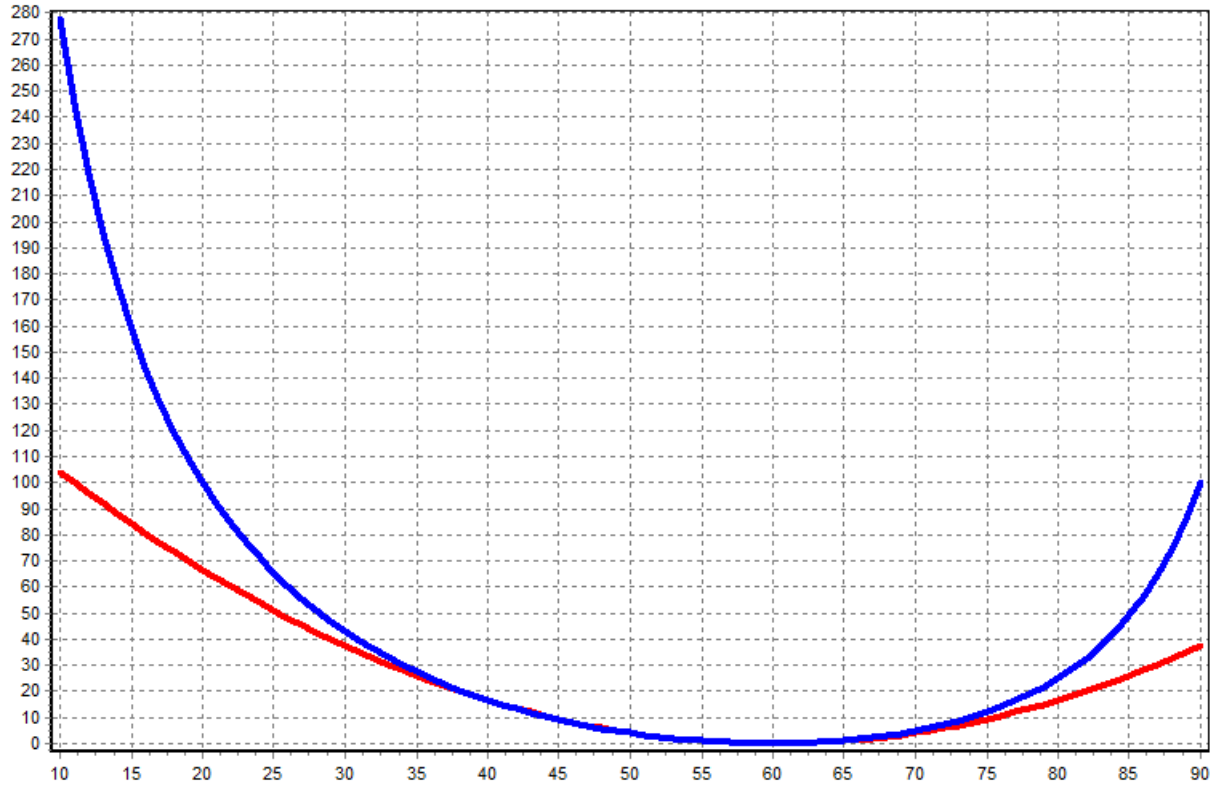


Рисунок 15 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 60, b = 40$).

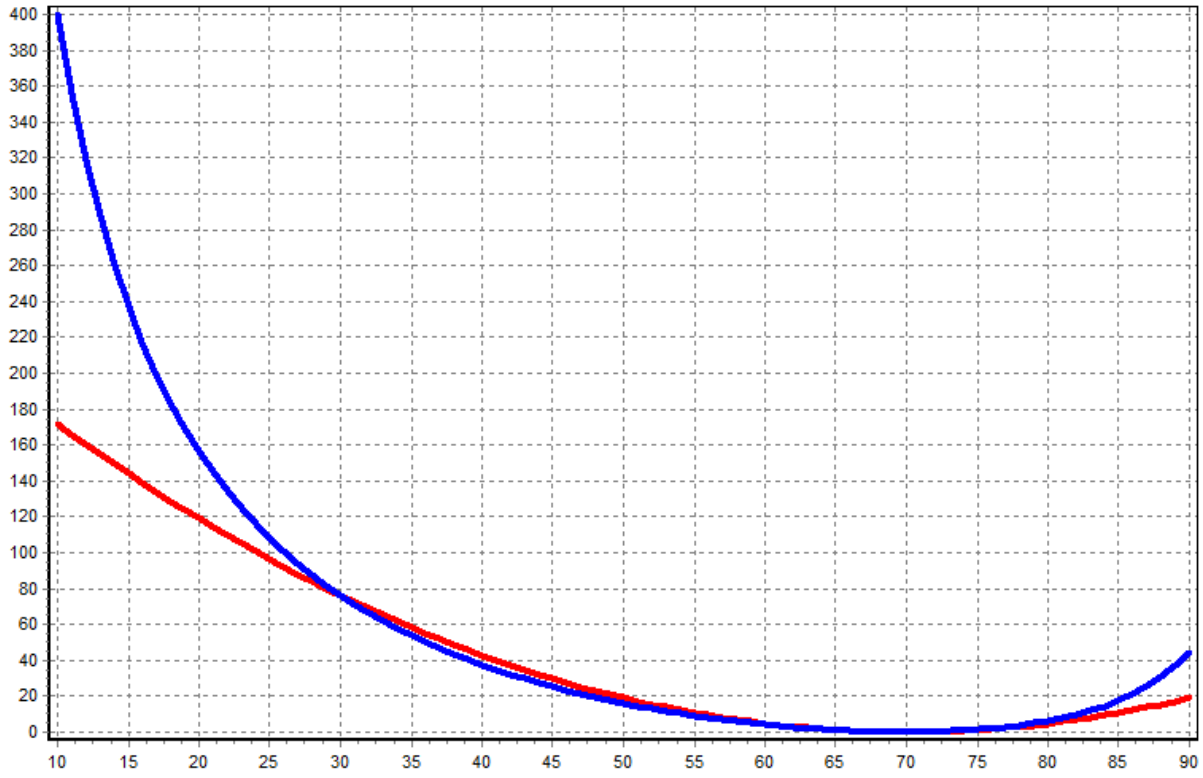


Рисунок 16 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 70, b = 30$).

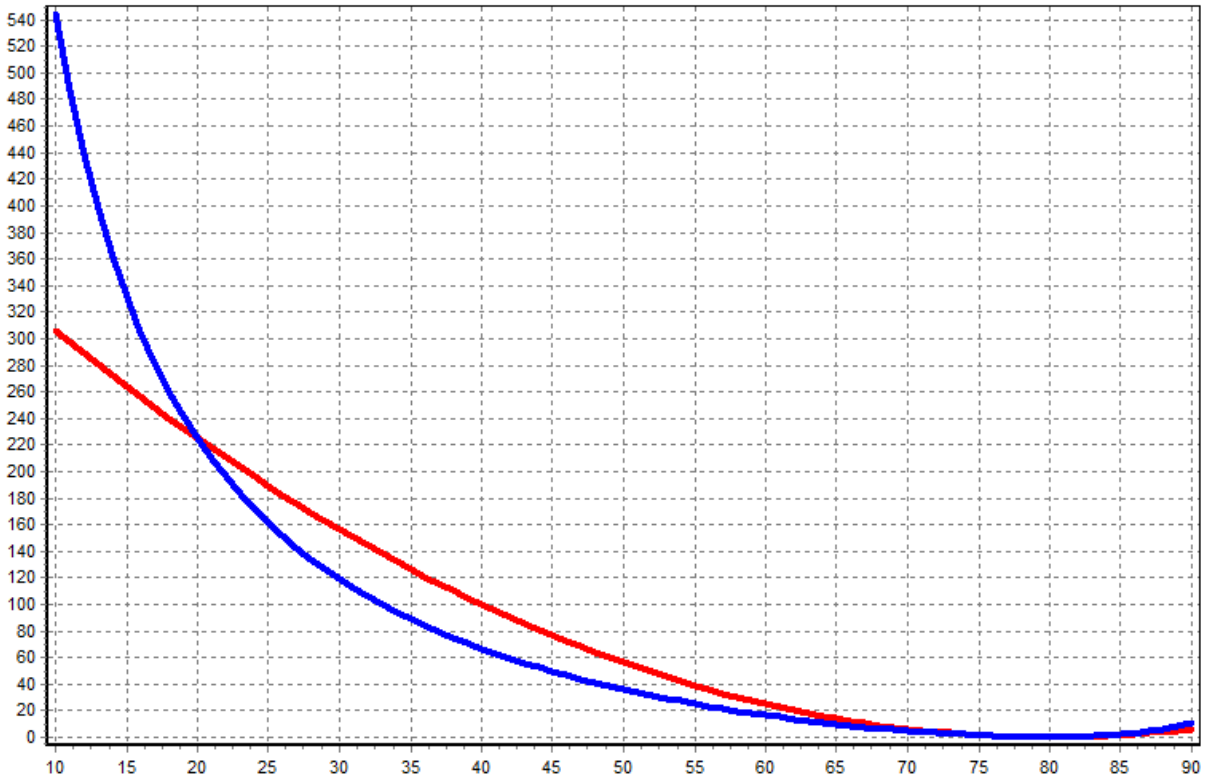


Рисунок 17 – Распределение χ^2 и $\tilde{\chi}^2$ ($a = 80, b = 20$).

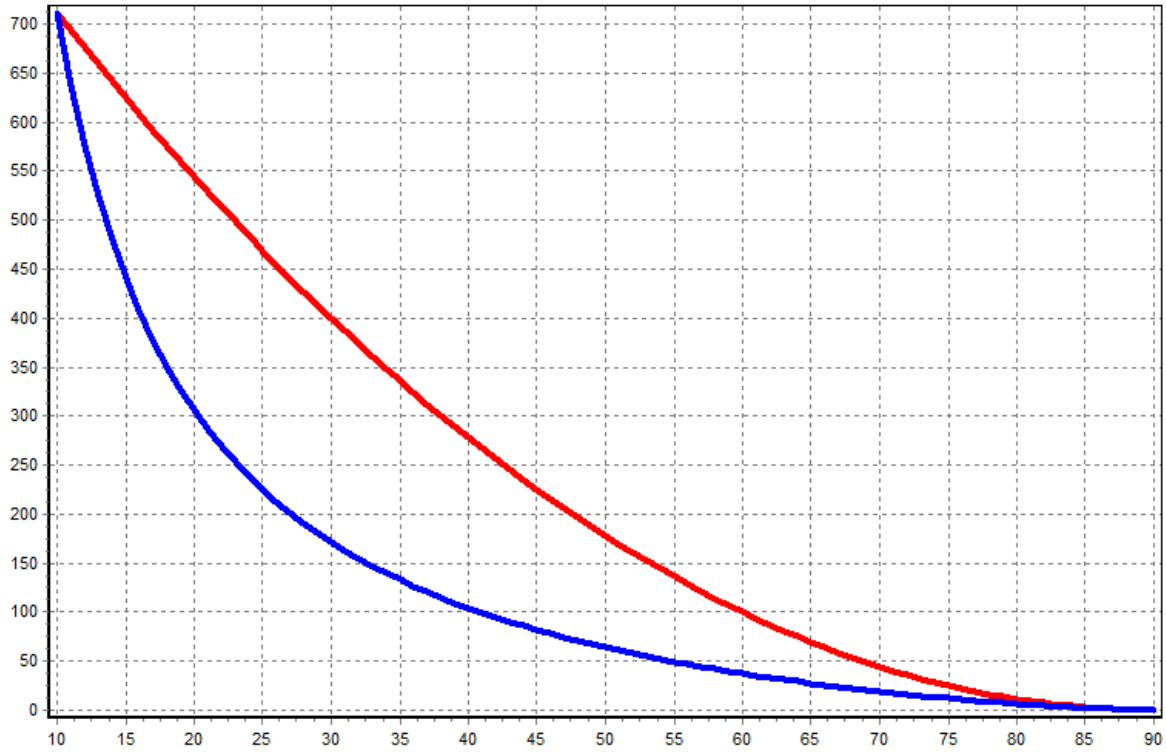


Рисунок 18 – Распределение χ^2 и $\tilde{\chi}^2$ ($a=90, b=10$).

Наибольшие концентрации точек можно отследить на рис.19.

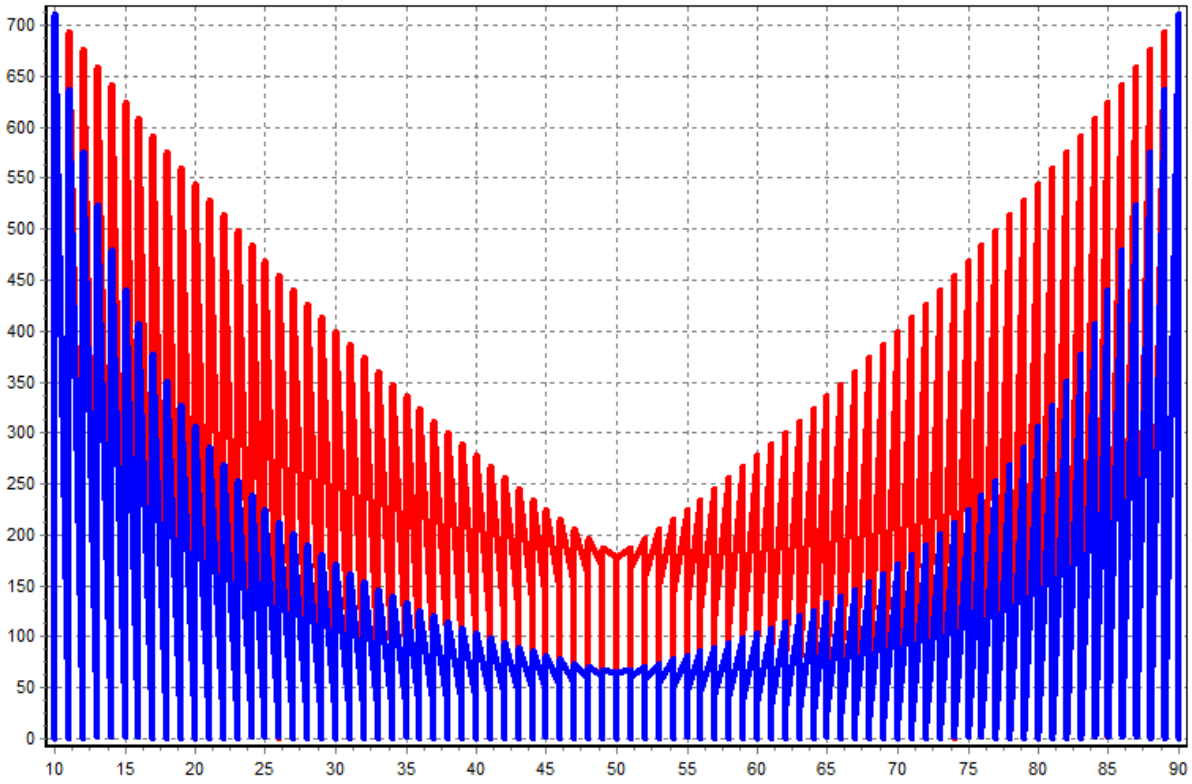


Рисунок 19 – Распределение облаков точек χ^2 и $\tilde{\chi}^2$ по плотности.

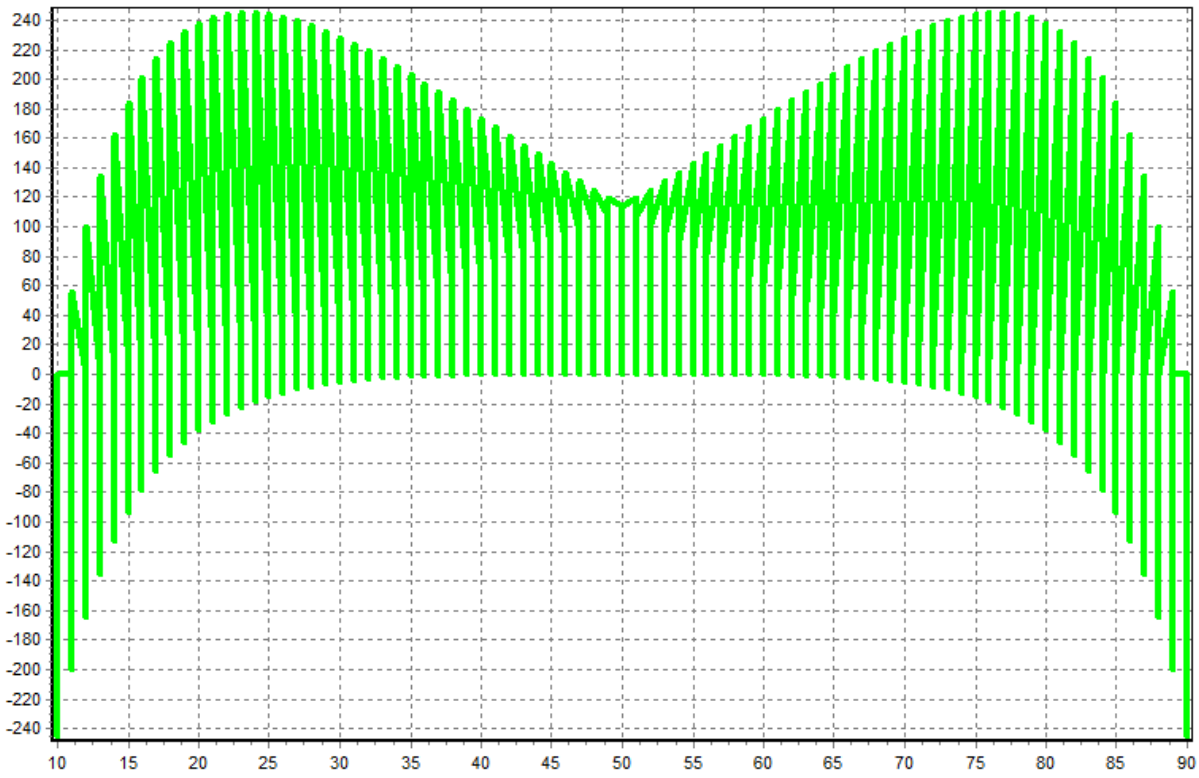
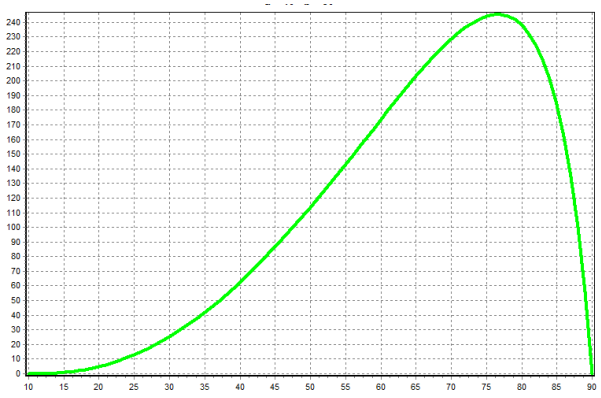
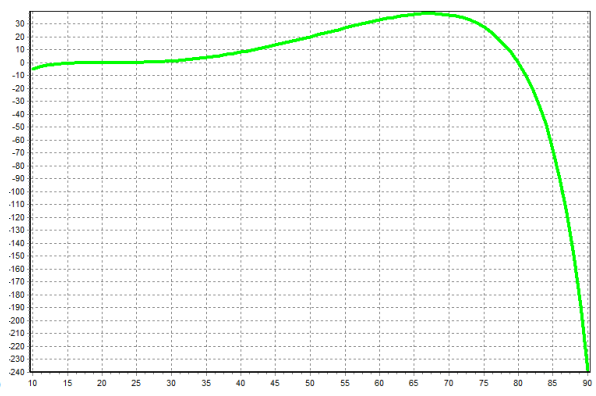


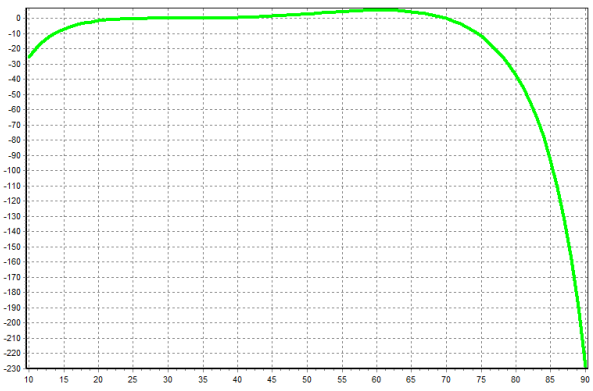
Рисунок 20 – Разность χ^2- $\tilde{\chi}^2$.



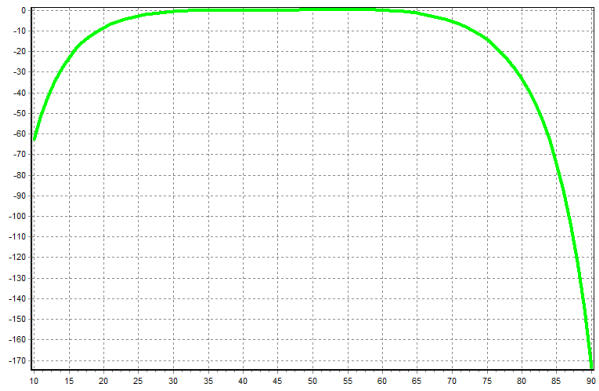
$a = 10$



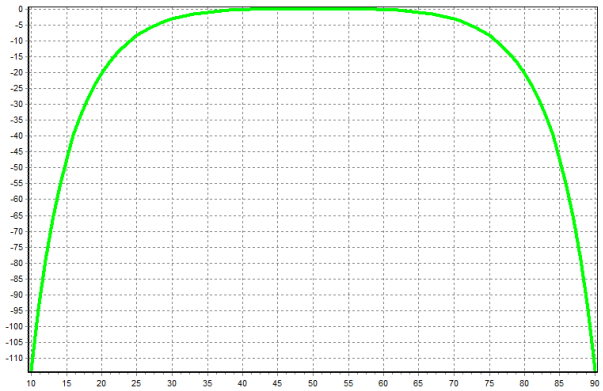
$a = 20$



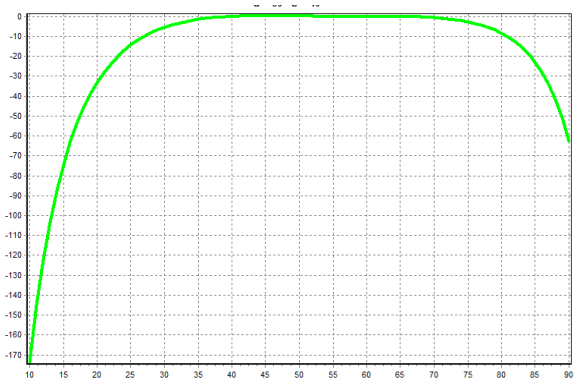
$a = 30$



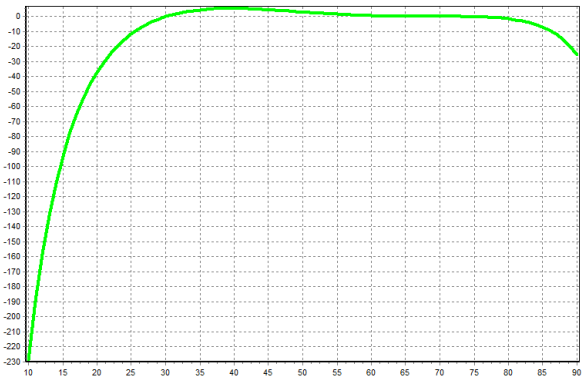
$a = 40$



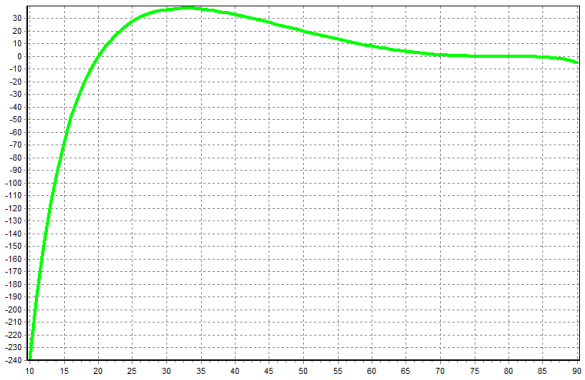
$a = 50$



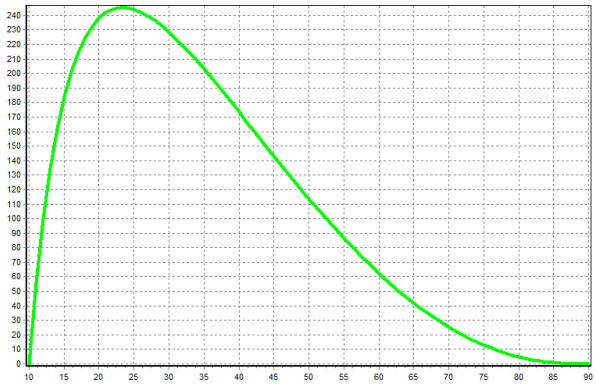
$a = 60$



$a = 70$



$a = 80$



$a = 90$

Рисунок 21 – Разность χ^2 - $\tilde{\chi}^2$ для разных a.

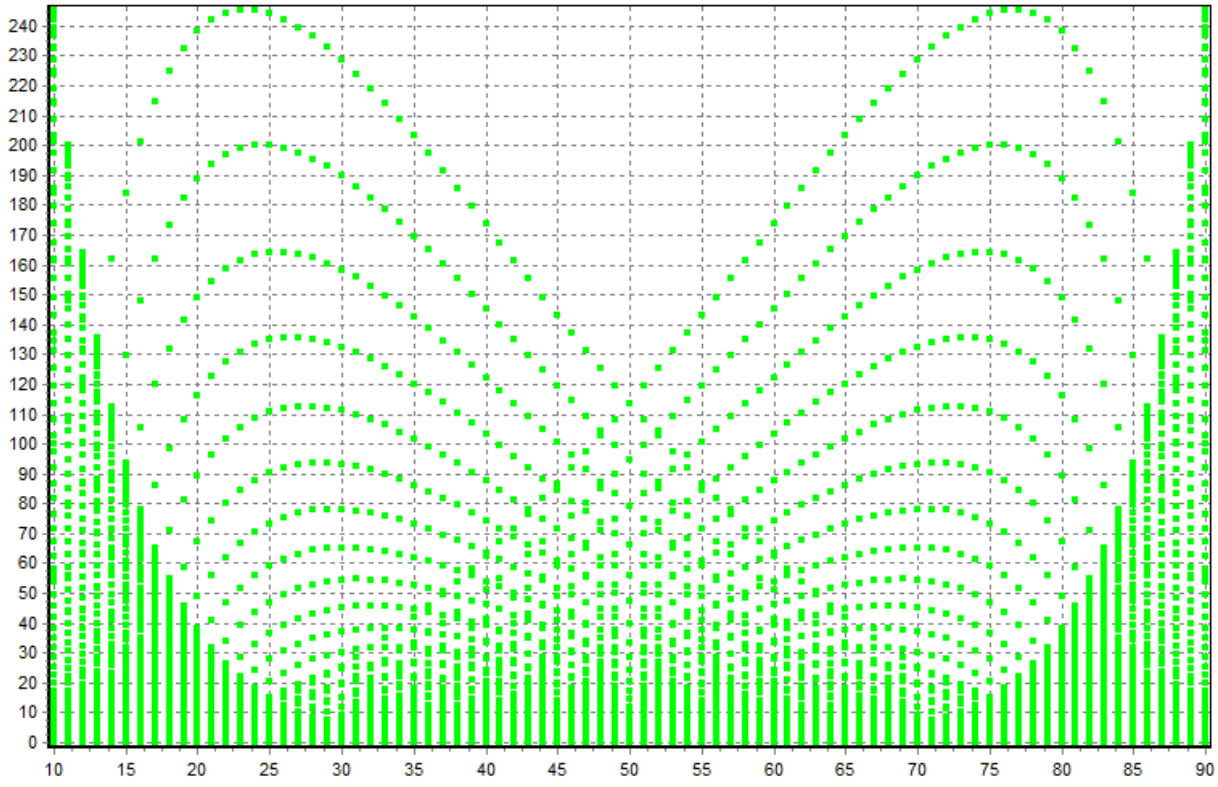


Рисунок 22 – Разность $\left|\chi^2-\tilde{\chi}^2\right|$.

Введем ограничения на $25<a<75$, $a<25$, $a>75$ тогда имеем Рис.23-25.

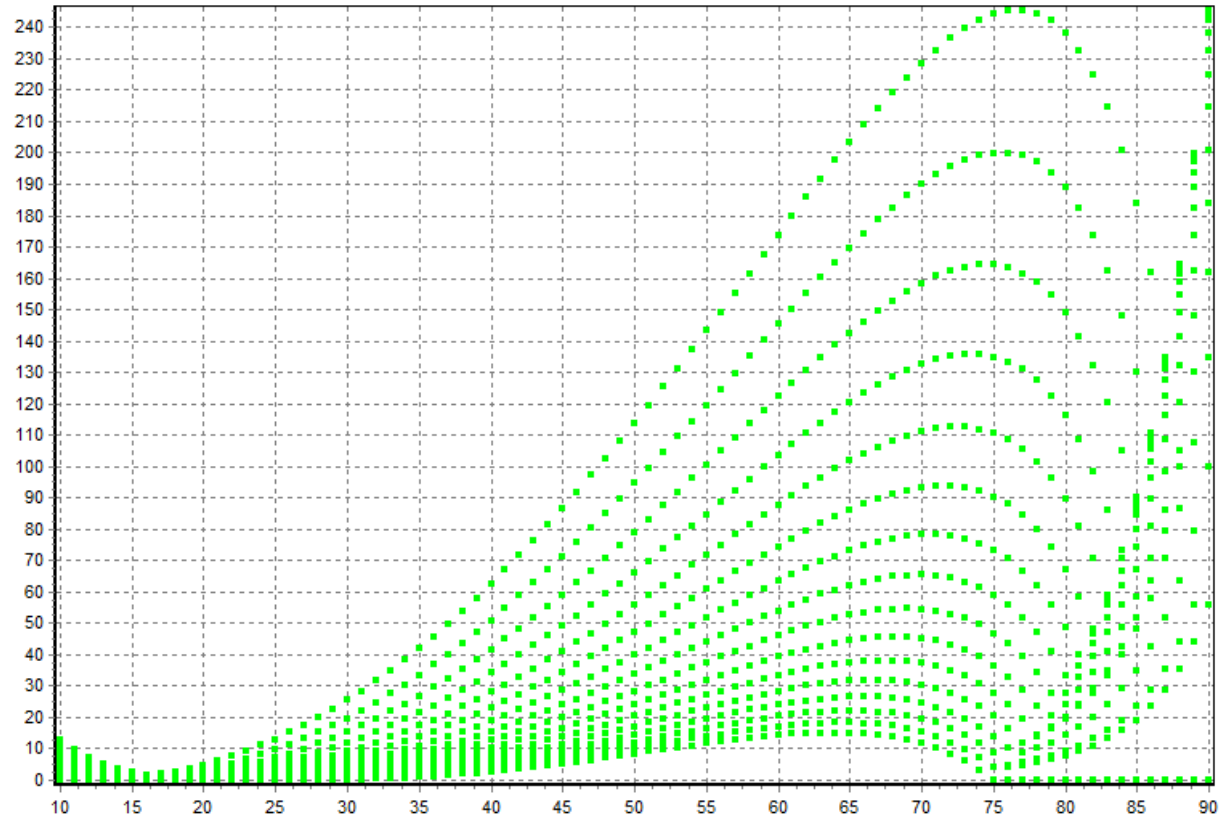


Рисунок 23 – Разность $\left|\chi^2-\tilde{\chi}^2\right|$ при $a<25$.

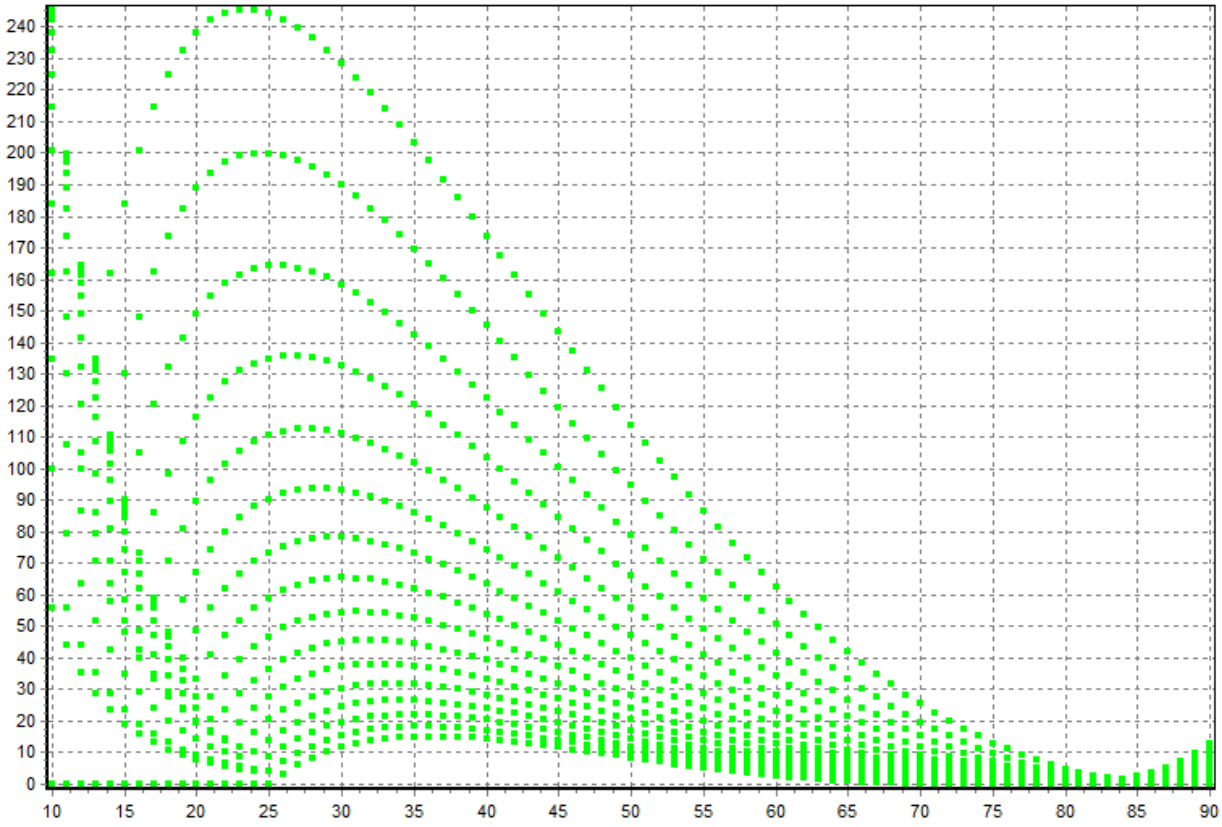


Рисунок 24 – Разность $\left|\chi^2 - \tilde{\chi}^2\right|$ при $a > 75$.

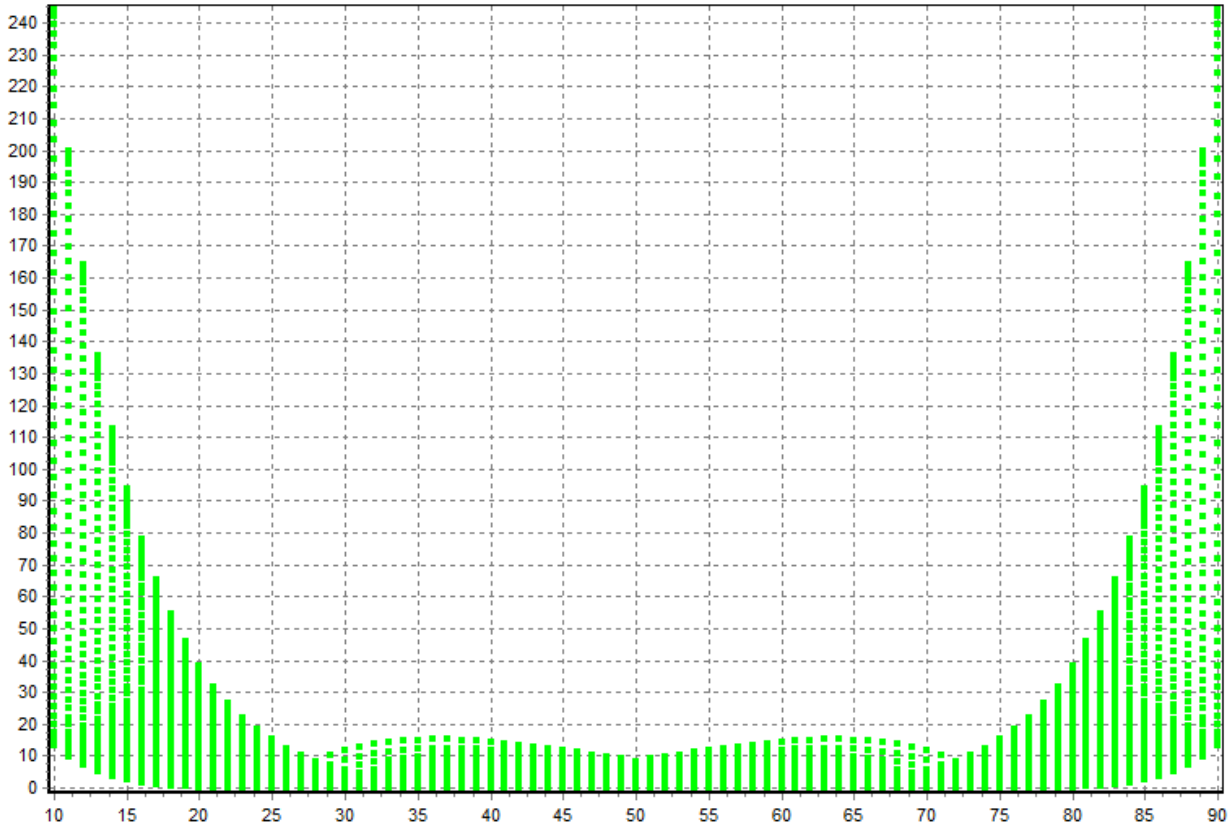


Рисунок 25 – Разность $\left|\chi^2 - \tilde{\chi}^2\right|$ при $25 < a < 75$.

Хорошо заметно, что при изменении параметров экспериментальной модели относительно теоретической в пределах $\pm 25\%$ от среднего значения, распределения Пирсона и $\tilde{\chi}^2$ отличаются не более чем на 15 пунктов.

При этом необходимо учитывать что для каждой отдельной модели рассчет ведется по своей норме, а последние значения абсолютной погрешности получены в безразмерных координатах по оси Oy (точнее с учетом размерности используемой нормы).

Для распределения χ^2 составлены таблицы [4], где указано критическое значение критерия согласия χ^2 для выбранного уровня значимости α и степеней свободы df (или ν).

Уровень значимости

α – вероятность ошибочного отклонения выдвинутой гипотезы, т.е. вероятность того, что будет отвергнута правильная гипотеза.

P — статистическая достоверность принятия верной гипотезы. В статистике чаще всего пользуются тремя уровнями значимости:

α=0,10, тогда P=0,90 (в 10 случаях из 100)

α=0,05, тогда P=0,95 (в 5 случаях из 100)

α=0,01, тогда P=0,99 (в 1 случае из 100) может быть отвергнута правильная гипотеза

Число степеней свободы df определяется как число групп в ряду распределения минус число связей: df = k –z. Под числом связей понимается число показателей эмпирического ряда, использованных при вычислении теоретических частот, т.е. показателей, связывающих эмпирические и теоретические частоты. Например, при выравнивании по кривой нормального распределения имеется три связи. Поэтому при выравнивании по кривой нормального распределения число степеней свободы определяется как df =k–3. Для оценки существенности, расчетное значение сравнивается с табличным $\chi^2_{табл}$

При полном совпадении теоретического и эмпирического распределений $\chi^2=0$, в противном случае $\chi^2>0$. Если $\chi^2_{расч} > \chi^2_{табл}$, то при заданном уровне значимости и числе степеней свободы гипотезу о несущественности (случайности) расхождений отклоняем. В случае, если $\chi^2_{расч} < \chi^2_{табл}$ то гипотезу принимаем и с вероятностью P=(1-α) можно утверждать, что расхождение между теоретическими и эмпирическими частотами случайно. Следовательно, есть основания утверждать, что эмпирическое распределение подчиняется нормальному распределению**.** Критерий согласия Пирсона используется, если объем совокупности достаточно велик (N>50), при этом, частота каждой группы должна быть не менее 5.

Рассмотрим теперь критические значения распределений. Хи-квадрат распределение задается табл.1, [3, с.25]. Степень свободы в нашем примере, по понятным причинам равна 1.

Таблица 1

Таблица критических значений для Хи-квадрат распределения

df	Вероятность ошибки										
	0.99	**0.95**	**0.9**	**0.5**	**0.25**	**0.10**	**0.05**	**0.025**	**0.01**	**0.005**	**0.001**
1	0.0002	0.004	0.02	0.46	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	0.02	0.10	0.21	1.39	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	0.12	0.35	0.58	2.37	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	0.30	0.71	1.06	3.36	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	0.55	1.15	1.61	4.35	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	0.87	1.64	2.20	5.35	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	1.24	2.17	2.83	6.35	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	1.65	2.73	3.49	7.34	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	2.09	3.33	4.17	8.34	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	2.56	3.94	4.87	9.34	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	3.05	4.57	5.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	3.57	5.23	6.30	11.3	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	4.11	5.89	7.04	12.3	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	4.66	6.57	7.79	13.3	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	5.23	7.26	8.55	14.3	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	5.81	7.96	9.31	15.3	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	6.41	8.67	10.1	16.3	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	7.01	9.39	10.9	17.3	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	7.63	10.1	11.7	18.3	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	8.26	10.9	12.4	19.3	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	8.90	11.6	13.2	20.3	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	9.54	12.3	14.0	21.3	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	10.2	13.1	14.8	22.3	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	10.9	13.8	15.7	23.3	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	11.5	14.6	16.5	24.3	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	12.2	15.4	17.3	25.3	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	12.9	16.2	18.1	26.3	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	13.6	16.9	18.9	27.3	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	14.3	17.7	19.8	28.3	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	15.0	18.5	20.6	29.3	34.8	40.3	43.8	47.0	50.9	53.7	59.7

Разность $\left|\chi^2 - \tilde{\chi}^2\right|$ существенно влияет на результаты сравнения двух гипотез! Ведь она меняется в пределах от 0 до 15, на промежутке (25,75), рис.26.

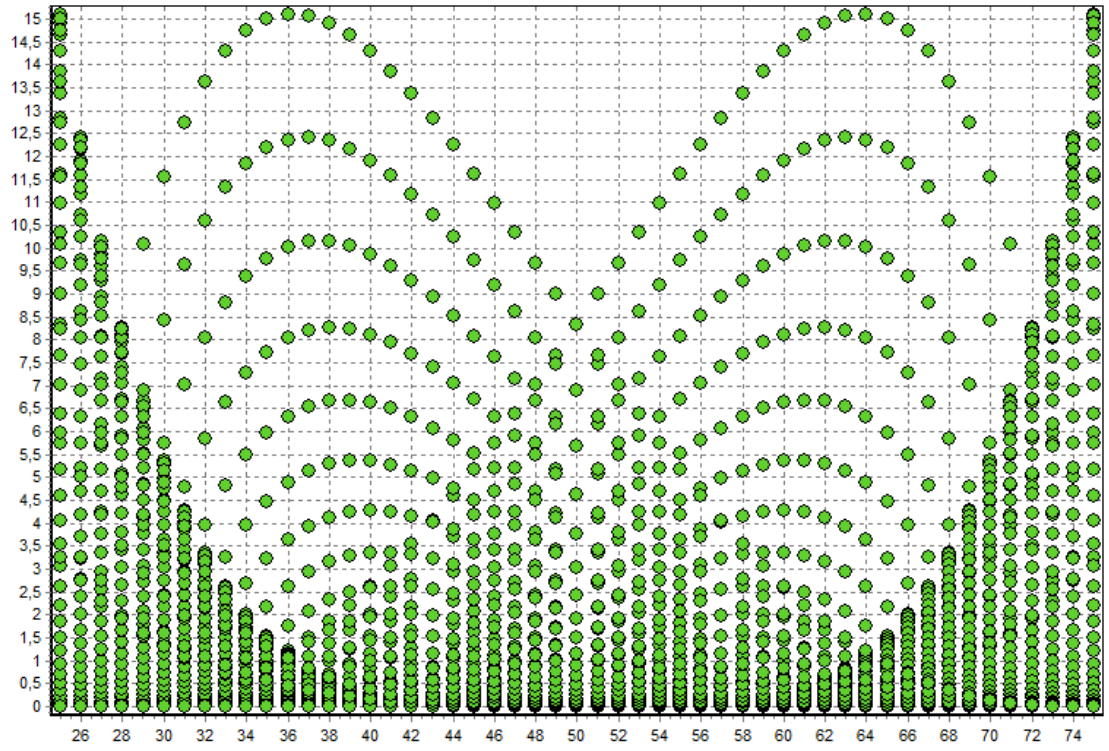


Рисунок 26 – Предельные отклонения на промежутке $25 < a < 75$.

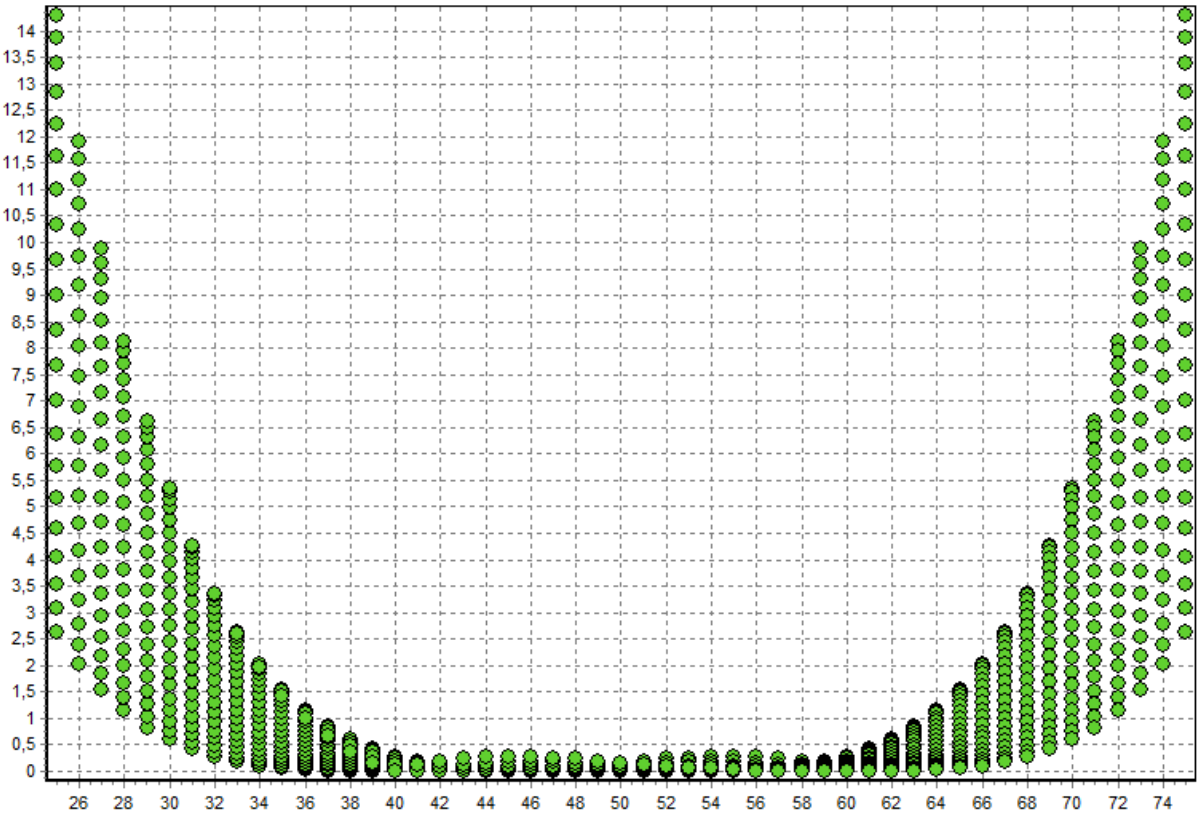


Рисунок 27 – Предельные отклонения на промежутке $40 < a < 60$.

Получаем равенство критериев χ^2 и $\tilde{\chi}^2$ только в узком промежутке отличия значений гипотез, с радиусом $\tilde{R}=10$.

Литература

1. Критерий согласия Пирсона. [Электронный ресурс]. URL : http://ru.wikipedia.org/wiki/%CA%F0%E8%F2%E5%F0%E8%E9_%F1%EE%E3%EB%E0%F1%E8%FF_%CF%E8%F0%F1%EE%ED%E0 (дата обращения: 16.08.2013).
2. Критерий хи-квадрат. [Электронный ресурс]. URL : http://www.machinelearning.ru/wiki/index.php?title=%D0%9A%D1%80%D0%B8%D1%82%D0%B5%D1%80%D0%B8%D0%B9_%D1%85%D0%B8-%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82 (дата обращения: 22.08.2013).
3. Колчинская В.Ю. Анализ данных в социологии. Методические указания по курсу. –Челябинск, 2006. -28 стр.
4. Критерии согласия.Теоретические и эмпирические частоты. Проверка на нормальность распределения. [Электронный ресурс]. URL :http://helpstat.ru/2012/09/kriterii-soglasiya/(дата обращения: 16.08.2013).
5. Попов О.А. Критерий Хи-квадрат. // Статистика в психологии и педагогике. [Электронный ресурс]. URL : http://psystat.at.ua/publ/1-1-0-29 (дата обращения: 22.08.2013).
6. Анализ двух выборок. [Электронный ресурс]. URL : http://www.tsput.ru/res/math/mop/lections/lection_6.htm (дата обращения: 22.08.2013).
7. Критерии согласия. HELPSTAT [Электронный ресурс]. URL : http://helpstat.ru/2012/09/kriterii-soglasiya/ (дата обращения: 22.08.2013).

Приложение 1

code: Delphi

```
unit Unit1;

interface

uses
  Windows, Messages, SysUtils, Variants, Classes, Graphics, Controls, Forms,
  Dialogs, StdCtrls, Grids, TeEngine, Series, ExtCtrls, TeeProcs, Chart;

type
  TForm1 = class(TForm)
    Button1: TButton;
```

```
    Chart1: TChart;
    Series1: TPointSeries;
    Series2: TPointSeries;
    Series3: TLineSeries;
    Series4: TLineSeries;
    Panel1: TPanel;
    Button2: TButton;
    Series5: TLineSeries;
    Button3: TButton;
    Series6: TPointSeries;
    Series7: TPointSeries;
    procedure Button1Click(Sender: TObject);
    procedure FormCreate(Sender: TObject);
    procedure Button2Click(Sender: TObject);
    procedure Button3Click(Sender: TObject);
  private
    { Private declarations }
  public
    { Public declarations }
  end;

var
  Form1: TForm1;
var
  I,j,a,b: Integer;
  hi,hi2:real;
implementation

{$R *.dfm}

procedure TForm1.Button1Click(Sender: TObject);

begin
b:=100-a;
for I := 25 to 75 do
begin
j:=100-i;

hi:=sqr(i-a)/a+ sqr(j-b)/b ;
hi2:=sqr(i-a)/i+ sqr(j-b)/j  ;

//stringgrid1.Cells[i,j]
//series3.AddXY(i,(int(hi*100)/100));
```

```
//series4.AddXY(i,(int(hi2*100)/100));
series7.AddXY(i,(int(abs(hi-hi2)*100)/100));

application.ProcessMessages;
end;

chart1.Title.Caption:='a = '+inttostr(a)+'   b = '+inttostr(b);

end;

procedure TForm1.Button2Click(Sender: TObject);
var i:integer;
begin
//series3.Clear;
//series4.Clear;
for i := 1 to 20 do
begin
a:=a+1;   button1.Click;

end;
end;

procedure TForm1.Button3Click(Sender: TObject);
begin
//series5.Clear;
 a:=a+10;   button1.Click;
end;

procedure TForm1.FormCreate(Sender: TObject);
begin     a:=40;
button1.Click;
end;

end.
```