электронный научно-технический журнал
ИНЖЕНЕРНЫЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51036. ISSN 2307-0595

Методика обработки результатов эксперимента с помощью системы MATLAB в курсе «Математическая статистика»

04, апрель 2016

Ахметова Ф. Х. [1], Ласковая Т. А. [1], Чигирёва О. Ю. [1,*]

УДК: 519.22

[1]Россия, МГТУ им. Н.Э. Баумана
*mkfn12@yandex.ru

Введение

При решении практических задач у исследователя возникают сложности, связанные с необходимостью обработки достаточно большого объема статистических данных. В учебной литературе приводятся методы «ручного» подсчета, однако их практическая реализация требует вычисления вспомогательных величин и составления специальных таблиц, превращая решение задачи в сложный трудоемкий процесс [1-4]. В этой связи, предлагается оригинальный опыт преподавания дисциплины, позволяющий совместить классическое изложение материала и современные компьютерные технологии. На конкретных примерах рассмотрены некоторые возможности пакета прикладных программ Statistics Toolbox системы MATLAB, позволяющие существенно повысить эффективность обработки данных [5].

Далее приведем основные понятия из курса «Математическая статистика», на которые будем опираться при обработке статистических данных.

Генеральная совокупность и выборка

Пусть имеется N объектов, каждому из которых присуще определенное значение некоторой числовой характеристики X. Совокупность этих N объектов будем называть ***генеральной совокупностью***.

Проведем испытание, состоящее в случайном выборе одного объекта из генеральной совокупности и определим для него значение характеристики X, которое обозначим через X_1. В результате проведения n испытаний получим ряд чисел $X_1, X_2, \ldots, X_n$, который будем называть ***выборкой*** объема n, а число X_k – ***k-м элементом выборки***.

Вариационный и статистический ряды

Вариационный ряд $X_1^*, X_2^*, \ldots, X_n^*$ представляет собой ту же самую выборку $X_1, X_2, \ldots, X_n$, но расположенную в порядке не убывания элементов: $X_1^* \le X_2^* \le \ldots \le X_n^*$. Элементы X_1^* и X_n^* называют ***крайними членами вариационного ряда***.

Если среди элементов выборки $X_1, X_2, \ldots, X_n$ есть одинаковые элементы, то выборку представляют в виде статистического ряда.

Статистическим рядом называют таблицу, в первой строке которой записаны в порядке возрастания различные значения $Z_1, Z_2, \ldots, Z_m$ элементов выборки $X_1, X_2, \ldots, X_n$, а во второй – числа их повторений.

Таблица 1. Статистический ряд

Z_1	Z_2	…	Z_j	…	Z_m
n_1	n_2	…	n_j	…	n_m

При этом число n_j, $j = \overline{1,m}$, показывающее сколько раз элемент Z_j встречается в выборке, называют ***частотой***, а отношение $\frac{n_j}{n}$ – ***относительной частотой*** этого значения.

Проиллюстрируем на примере методику построения вариационного и статистического рядов заданной выборки.

Пример 1. При измерении роста 25 студентов, отобранных случайным образом, получены следующие результаты (в *см*):

167;	162;	160;	164;	177;
155;	165;	169;	172;	166;
174;	164;	170;	166;	169;
166;	180;	158;	174;	157;
169;	159;	167;	162;	176.

Построить вариационный и статистический ряды этой выборки.

<u>Решение.</u> Для того чтобы получить вариационный ряд, нужно расположить элементы выборки в неубывающем порядке (повторяющиеся элементы следует указать столько раз, сколько они встречаются в выборке). Таким образом, вариационный ряд этой выборки имеет вид:

$X_1^* = 155$;	$X_2^* = 157$;	$X_3^* = 158$;	$X_4^* = 159$;	$X_5^* = 160$;
$X_6^* = 162$;	$X_7^* = 162$;	$X_8^* = 164$;	$X_9^* = 164$;	$X_{10}^* = 165$;
$X_{11}^* = 166$;	$X_{12}^* = 166$;	$X_{13}^* = 166$;	$X_{14}^* = 167$;	$X_{15}^* = 167$;
$X_{16}^* = 169$;	$X_{17}^* = 169$;	$X_{18}^* = 169$;	$X_{19}^* = 170$;	$X_{20}^* = 172$;
$X_{21}^* = 174$;	$X_{22}^* = 174$;	$X_{23}^* = 176$;	$X_{24}^* = 177$;	$X_{25}^* = 180$.

Статистический ряд представлен в таблице 2.

Таблица 2. Статистический ряд

Z_j	155	157	158	159	160	162	164	165	166
n_j	1	1	1	1	1	2	2	1	3
Z_j	167	169	170	172	174	176	177	180	Контроль: $\sum_{j=1}^{17} n_j = 25$
n_j	2	3	1	1	2	1	1	1	

Пусть выборка представлена в виде прямоугольной таблицы, записанной в текстовом файле data1.txt. Приведенная ниже программа считывает исходные данные (элементы выборки) из этого файла, вычисляет объем выборки, находит вариационный и статистические ряды, а затем полученные значения выводит на печать в Командном Окне (MATLAB Command Window).

```
%==========================================================================
%ИСХОДНЫЕ ДАННЫЕ:
% выборка представлена в виде прямоугольной таблицы,
% записанной в текстовом файле data1.txt
%==========================================================================
%РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ:
% 1. k - объем выборки;
% 2. массив X - вариационный ряд;
% 3. статистический ряд: Z - массив различных значений элементов выборки,
%                        n - массив частот.
%==========================================================================
clear
%Считывание исходных данных из файла data1.txt (разделителем является ";")
%в двумерный массив с именем A:
A=dlmread('data1.txt',';');
%Определение размерности A:
[mA,nA]=size(A);
%Нахождение числа элементов A:
k=mA*nA;
disp('*----- РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ -----*');
%Вывод на печать значения объема выборки
fprintf('Объем выборки равен %g\n',k)
%Преобразование двумерного массива A в одномерный массив X:
X=reshape(A,1,k);
%Вариационный ряд получаем в результате упорядочивания
%по неубыванию элементов массива X:
X=sort(X);
%Статистический ряд получаем из вариационного ряда
%с учетом повторений значений элементов:
i=1;
j=1;
while i <= k
  B=X(i);
  C=eq(X,B);
  Z(j)=X(i);
  n(j)=sum(C);
```

```
  i=i+n(j);
  j=j+1;
end;
jmax=j-1;
%Вывод на печать статистического ряда
disp('Статистический ряд имеет вид:');
for j=1:jmax
  fprintf('Z(%g)= %7.3f  n(%g)= %g\n',j,Z(j),j,n(j))
end;
%------------------------------------------------------------------- end
```

Эмпирическая функция распределения

Эмпирической функцией распределения называют функцию

$$F^*(x) = \frac{n_x}{n},$$

где n_x – число элементов выборки, значения которых меньше x; n – объем выборки.

Функция $F^*(x)$ обладает следующими свойствами:

1. $0 \le F^*(x) \le 1$;
2. $F^*(x)$ – неубывающая функция;
3. $F^*(x) = 0$ при $x \le X_1^*$, $\quad F^*(x) = 1$ при $x > X_n^*$.

Если выборка задана статистическим рядом, то функцию $F^*(x)$ можно записать в виде:

$$F^*(x) = \begin{cases} 0, & x \le Z_1; \\ \dfrac{1}{n}\sum\limits_{j=1}^{k} n_j, & Z_k < x \le Z_{k+1},\ k = \overline{1, m-1}; \\ 1, & x > Z_m. \end{cases}$$

Таким образом, эмпирическая функция распределения является кусочно-постоянной функцией, причем в точках Z_k, $k = \overline{1, m}$ ее значение изменяется «скачком» на величину $\frac{n_k}{n}$.

Пример 2. При проведении *n*=2608 опытов по наблюдению числа α - частиц, излучаемых радиоактивным веществом за определенный период времени (7,5 *с*). Результаты наблюдений представлены в таблице 3.

Таблица 3. Результаты наблюдений

Число наблюдаемых α - частиц	0	1	2	3	4	5	6	7	8	9	10
Число опытов с такими результатами наблюдений	57	203	383	525	532	408	273	139	45	27	16

Найти эмпирическую функцию распределения и построить ее график.

Решение. Согласно определению, эмпирическая функция распределения (см. рис.1) примет вид:

$$F^*(x)=\begin{cases}0,0000, & x\le 0;\\ 0,0219, & 0<x\le 1;\\ 0,0997, & 1<x\le 2;\\ 0,2465, & 2<x\le 3;\\ 0,4479, & 3<x\le 4;\\ 0,6518, & 4<x\le 5;\\ 0,8083, & 5<x\le 6;\\ 0,9130, & 6<x\le 7;\\ 0,9663, & 7<x\le 8;\\ 0,9835, & 8<x\le 9;\\ 0,9939, & 9<x\le 10;\\ 1,0000, & x>10.\end{cases}$$

Следующий алгоритм по заданному статистическому ряду строит график эмпирической функции распределения.

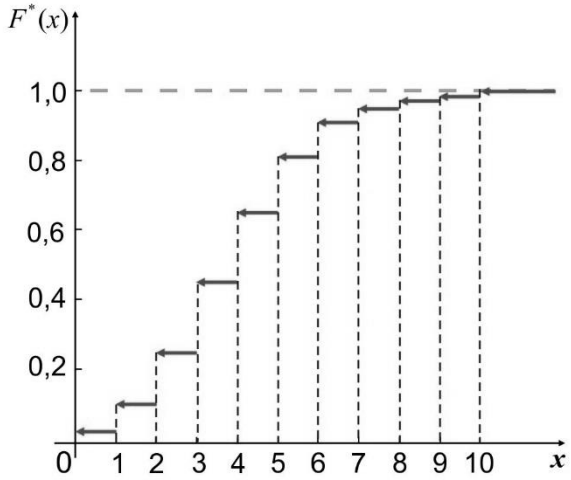

Рис. 1. График эмпирической функции распределения

```
%=========================================================================
%ИСХОДНЫЕ ДАННЫЕ:
% статистический ряд: Z - массив различных значений элементов выборки,
%                      n - массив частот.
%=========================================================================
%РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ:
% отображение графика эмпирической функции распределения
% в Графическом Окне (Figure).
%=========================================================================
for j=1:length(n);
  F(j)=sum(n(1:j));
end;
F=F'./sum(n);
stairs(Z,F,'b-')
%------------------------------------------------------------------- end
```

Гистограмма

Для выборки, заданной статистическим рядом, отрезок $\boldsymbol{J}=\left[Z_1,Z_m\right]$ разбивают на p промежутков одинаковой длины Δ: $\boldsymbol{J}=\bigcup_{i=1}^{p}\boldsymbol{J}_i$, где $\boldsymbol{J}_i=\left[x_i,x_{i+1}\right)$, $i=\overline{1,p-1}$; $\boldsymbol{J}_p=\left[x_p,x_{p+1}\right]$; $x_1\equiv Z_1$, $x_{p+1}\equiv Z_m$.

Затем для каждого промежутка $\boldsymbol{J}_i$, $i=\overline{1,p}$ находят число ν_i элементов выборки, попавших в этот промежуток, и составляют таблицу, которую называют ***интервальным статистическим рядом.***

Таблица 4. Интервальный статистический ряд

$\boldsymbol{J}_1$	$\boldsymbol{J}_2$	…	$\boldsymbol{J}_i$	…	$\boldsymbol{J}_p$
ν_1	ν_2	…	ν_i	…	ν_p

Далее переходят к построению ***гистограммы***, которая представляет собой диаграмму, составленную из прямоугольников с одинаковыми основаниями Δ и высотами, равными $h_i=\dfrac{\nu_i}{n\cdot\Delta}$, $i=\overline{1,p}$. Очевидно, что площадь всех прямоугольников, образующих гистограмму, равна единице.

Приведем один из способов построения гистограммы:

```
%=================================================================
%ИСХОДНЫЕ ДАННЫЕ:
% X - вариационный ряд;
% p - число промежутков интервального статистичесого ряда.
%==================================================================
%РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ:
% 1. массивы J и Nu определяют интервальный статистический ряд:
% Nu(i) - число элементов выборки, попавших в промежуток [ J(i),J(i+1) ),
% где i=1,2,...p-1;
% а Nu(p)- число элементов, попавших в [ J(p),J(p+1) ];
% 2. отображение гистограммы в Графическом Окне (Figure).
%==================================================================
p=10;
delta=(X(end)-X(1))/p;
J=X(1):delta:X(end);
H=histc(X,J);
H(end-1)=sum(H(end-1:end)); H(end)=0;
Nu=H(1:p);
bar(J,H/(delta*length(X)),'histc')
%------------------------------------------------------------------ end
```

Выборочные характеристики

Выборочным моментом k-го порядка называют величину, равную

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k .$$

В частности, выборочный момент 1-го порядка называют ***выборочным средним*** и обозначают $\overline{x}$.

Выборочным центральным моментом k-го порядка называют величину, равную

$$\boldsymbol{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{x})^k .$$

В частности, выборочный центральный момент 2-го порядка называют ***выборочной дисперсией*** и обозначают σ_*^2.

Выборочным средним квадратичным отклонением (стандартом) называют квадратный корень из выборочной дисперсии.

Наряду с выборочной дисперсией используют и другую характеристику разброса выборки относительно $\overline{x}$, которую называют ***исправленной выборочной дисперсией*** и вычисляют по формуле:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{x})^2 .$$

Очевидно, что $S^2 = \frac{n}{n-1}\sigma_*^2$.

Для вычисления выборочных характеристик и вывода их значений на печать используют следующие функции:

```
%==========================================================================
%ИСХОДНЫЕ ДАННЫЕ:
% X - вариационный ряд.
%==========================================================================
%РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ:
% 1. mx - выборочное среднее;
% 2. sigma - выборочное среднее квадратичное отклонение,
%    sigma2 - выборочная дисперсия;
% 3. S2 - исправленная выборочная дисперсия.
%==========================================================================
%Вычисление и вывод на печать выборочных характеристик
mx=mean(X);
sigma=std(X,1);
sigma2=sigma^2;
S=std(X,0);
S2=S^2;
disp('Выборочные характеристики:');
fprintf('а)выборочное среднее равно %7.3f\n',mx)
fprintf('б)выборочная дисперсия равна %7.3f\n',sigma2);
fprintf('выборочное среднее квадратичное отклонение равно %7.3f\n',sigma)
fprintf('в)исправленная выборочная дисперсия равна %7.3f\n',S2)
%------------------------------------------------------------------- end
```

Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

Пусть выборка $X_1, X_2, \ldots, X_n$ произведена из генеральной совокупности с неизвестной теоретической функцией распределения. Необходимо проверить, согласуются или нет эмпирические данные с гипотетическим предположением относительно теоретической функции распределения. Для этого в математической статистике используют критерии согласия, среди которых наиболее часто применяют критерий Колмогорова и критерий Пирсона (критерий «хи-квадрат»). Покажем на примере, как с помощью критерия Пирсона, проверить на уровне значимости α основную гипотезу H_0: генеральная совокупность распределена нормально.

1. Разобьем всю числовую прямую на s непересекающихся промежутков:

$$(-\infty, x_1) \cup [x_1, x_2) \cup \ldots \cup [x_{i-1}, x_i) \cup \ldots \cup [x_{s-2}, x_{s-1}) \cup [x_{s-1}, +\infty)$$

и определим числа ν_i элементов выборки, попавших в эти промежутки.

2. Найдем гипотетические вероятности P_i попадания в промежутки $[x_{i-1}, x_i)$: $P_i = F_0(x_i) - F_0(x_{i-1})$, где $F_0(x)$ – теоретическая функция распределения. Для нормального распределения:

$$P_i = \Phi_0\left(\frac{x_i - \overline{x}}{\sigma_*}\right) - \Phi_0\left(\frac{x_{i-1} - \overline{x}}{\sigma_*}\right), \quad i = \overline{2, s-1},$$

$$P_1 = \Phi_0\left(\frac{x_1 - \overline{x}}{\sigma_*}\right) + \frac{1}{2}, \quad P_s = \frac{1}{2} - \Phi_0\left(\frac{x_{s-1} - \overline{x}}{\sigma_*}\right).$$

Здесь $\Phi_0(x)$ – функция Лапласа.

3. Рассмотрим статистику

$$\chi^2 = \chi^2(X_1, \ldots X_n) = \sum_{i=1}^{s} \frac{(\nu_i - nP_i)^2}{nP_i}.$$

Доказано, что эта статистика асимптотически при $n \to \infty$ независимо о того, какому закону распределения подчинена генеральная совокупность, имеет χ^2–распределение с k степенями свободы. Число степеней свободы определяется следующим образом: $k = s - 1 - r$, где r – число параметров предполагаемого распределения, которые оценены по данным выборки. Для нормального распределения $r = 2$, так как оценивают два параметра (математическое ожидание и среднее квадратичное отклонение).

4. Применим критерий согласия Пирсона. Вычислим значение статистики $\chi^2_{набл} = \sum_{i=1}^{s} \frac{(\nu_i - nP_i)^2}{nP_i}$ и сравним его с критическим значением $\chi^2_{1-\alpha}(k)$ – квантиль уровня $1-\alpha$ χ^2–распределения с k степенями свободы.

Если $\chi^2_{набл} \le \chi^2_{1-\alpha}(k)$, то делается вывод о том, что на уровне значимости α основная гипотеза H_0 не противоречит статистическим данным и ее следует принять. Если $\chi^2_{набл} > \chi^2_{1-\alpha}(k)$, то гипотезу H_0 отвергаем.

Замечание. При разбиении числовой прямой на промежутки удобнее пользоваться построенной гистограммой или интервальным статистическим рядом. Если есть маловероятные промежутки, для которых $\nu_i < 5$, то их следует объединить с соседними; при этом уменьшится число степеней свободы k.

Пример 3. Ниже приведены результаты определения линейной плотности стальной проволоки (*г/м*):

381;	388;	384;	418;	373;	364;	376;	383;	432;	428;
440;	440;	409;	406;	416;	418;	398;	371;	391;	421;
413;	385;	425;	423;	421;	431;	429;	411;	415;	429;
390;	382;	430;	372;	430;	437;	407;	402;	400;	429;
385;	405;	363;	404;	369;	350;	421;	358;	422;	373;
418;	383;	412;	382;	383;	428;	409;	397;	427;	430;
392	376;	433;	363;	365;	395;	393;	377;	392;	379;
388;	399;	389;	362;	382;	382;	384;	415;	378;	375;
384;	375;	372;	427;	385;	410;	378;	392;	398;	398;
397;	393;	401;	405;	409;	390;	404;	402;	435;	402;
413;	412;	395;	420;	398;	421;	425;	400;	391;	403;
418;	447;	380;	353;	407;	380;	450;	418;	411;	392;
399;	391;	373;	413;	406;	395;	410;	400;	405;	412;
394;	410;	385;	370;	395;	395;	388;	361;	399;	402;
389;	403;	388;	429;	409;	406;	394;	443;	405;	417.

Для данной выборки:

1) построить интервальный статистический ряд и гистограмму;

2) найти выборочное среднее и выборочное среднее квадратичное отклонение; на уровне значимости $\alpha = 0{,}05$ основную гипотезу H_0: генеральная совокупность распределена нормально.

<u>Решение.</u> Приведем результаты обработки данных для данного примера, которые были получены при использовании программ, приведенных выше.

1. Интервальный статистический ряд имеет вид:

Таблица 5. Интервальный статистический ряд

J_i	$[350;360)$	$[360;370)$	$[370;380)$	$[380;390)$	$[390;400)$
ν_i	3	7	15	23	27
J_i	$[400;410)$	$[410;420)$	$[420;430)$	$[430;440)$	$[440,450]$
ν_i	25	20	17	8	5

Контроль: $\sum_{i=1}^{10} \nu_i = 150$.

2. Гистограмма изображена на рис. 2.

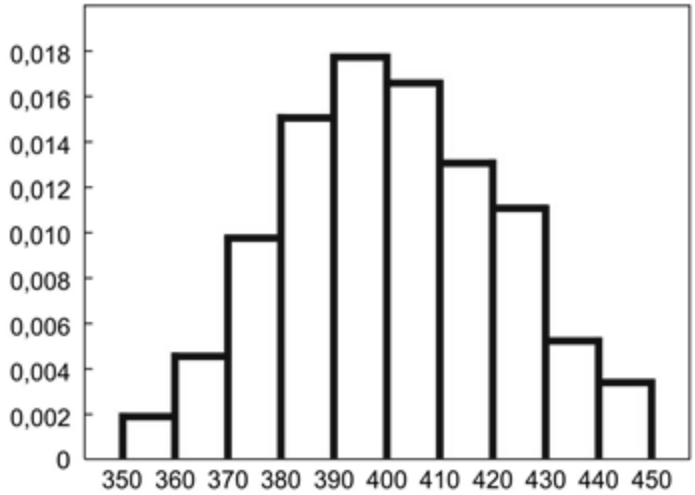

Рис. 2. Гистограмма выборки

3. Выборочные характеристики:

а) выборочное среднее $\bar{x} = 400,087$;

б) выборочное среднее квадратичное отклонение $\sigma_* = 20,987$.

Замечая, что в промежуток $[350,360)$ попало всего три элемента выборки, объединим его с соседним промежутком. Для полученных 9 промежутков вычислим вероятности P_i и запишем результаты в приведенную ниже таблицу 6.

Таблица 6. Интервальный статистический ряд

I_i	$(-\infty;370)$	$[370;380)$	$[380;390)$	$[390;400)$	$[400;410)$	$[410;420)$	$[420;430)$	$[430;440)$	$[440;+\infty)$
ν_i	10	15	23	27	25	20	17	8	5
P_i	0,0758	0,0934	0,1461	0,1830	0,1833	0,1470	0,0943	0,0484	0,0287

Вычислим значение статистики $\chi^2_{набл} = \sum_{i=1}^{9} \frac{(\nu_i - 150P_i)^2}{150P_i} = 1,4803$ и сравним его с критическим значением $\chi^2_{0,95}(6) = 12,59$. Применим критерий Пирсона. Поскольку $\chi^2_{набл} \le \chi^2_{0,95}(6)$, то на уровне значимости $\alpha = 0,05$ принимаем основную гипотезу H_0: генеральная совокупность распределена нормально.

Ниже приведен текст программы, которая позволяет вычислить наблюдаемое значение данной статистики:

```
%==========================================================================
%ИСХОДНЫЕ ДАННЫЕ:
```

```
% mx - выборочное среднее;
% sigma - выборочное среднее квадратичное отклонение;
% I(1),I(2),... - интервалы;
% Nu(k) - число элементов выборки, попавших в k-й интервал I(k).
%==========================================================================
%РЕЗУЛЬТАТЫ ВЫЧИСЛЕНИЙ:
% Hi2 - наблюдаемое значение "Хи-квадрат".
%==========================================================================
clear
mx=400.087;
sigma=20.987;
I=370:10:440;
Nu=[10 15 23 27 25 20 17 8 5];
z=(I-mx)/sigma;
FL=(erf(z/sqrt(2)))./2;
P=cat(2,FL(1)+.5,FL(2:length(I))-FL(1:length(I)-1),.5-FL(end));
Pn=P*sum(Nu);
h=(Nu-Pn).^2;
Hi=h./Pn;
Hi2=sum(Hi);
fprintf('Наблюдаемое значение "Хи-квадрат" равно %7.4f\n',Hi2);
%------------------------------------------------------------------------ end
```

Заключение

Предложенные в данной работе примеры, демонстрируют некоторые возможности пакета прикладных программ Statistics Toolbox системы MATLAB, позволяющие повысить эффективность обработки статистических данных. Показано, как полученные результаты могут быть использованы для решения задачи, состоящей в проверке гипотезы о согласии выборочного распределения с теоретическим законом распределения.

Список литературы

[1]. Бочаров П.П., Печинкин А.В. Теория вероятностей. Математическая статистика. 2-е изд. М.: Физматлит, 2005. 296 с.

[2]. Гмурман В.Е. Теория вероятностей и математическая статистика: Учеб. пособие. 12-е изд. М.: Юрайт, 2016. 479 с.

[3]. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. пособие для бакалавров: для студентов вузов. 11-е изд., перераб. и доп. М.: Юрайт. 2015. 404 с.

[4]. Ивченко Г.И., Медведев Ю.И., Чистяков П.В. Задачи с решениями по математической статистике: Учеб. пособие для вузов. М.: Дрофа, 2007. 318 с.

[5]. Потемкин В.Г. Вычисления в среде MATLAB. М.: Диалог – МИФИ, 2004. 720 с.