ГОСУДАРСТВЕННЫЙ КОМИТЕТ СВЯЗИ, ИНФОРМАТИЗАЦИИ И ТЕЛЕКОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ РЕСПУБЛИКИ УЗБЕКИСТАН

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

ФАКУЛЬТЕТ ТЕЛЕКОММУНИКАЦИОННЫХ ТЕХНОЛОГИЙ

ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методическое пособие

Ташкент 2013

ВВЕДЕНИЕ

Почти каждому современному работнику, где бы он не работал (в науке, в банке и др.), в процессе производственной деятельности необходимо производить различного рода измерения и обрабатывать их. При небольших объемах данных и несложных моделях изучаемого объекта можно проводить обработку вручную. Однако современные потоки данных, их объем и скорость поступления заставляют разрабатывать не только технику, но и методы обработки и соответствующее программное обеспечение. И во всех этих случаях человек, обрабатывающий измерения должен располагать соответствующими знаниями, которые позволили бы получит оптимальный результат.

При работе со случайными измерениями (а таковыми являются все без исключения эксперименты) исследователь должен знать стандартные методы оценки погрешности. Оценивать тип плотности вероятности распределения ошибок и проводить экспертную оценку результатов измерений на полноту и достоверность.

В настоящее время достаточно много разработано программных средств для обработки статистических и экспериментальных данных. Поэтому исследователь должен ориентироваться в различных версиях и модификациях этих средств.

Цель методического пособия - ознакомить магистрантов с основами теории обработки экспериментальных данных. Для этого необходимо уделить внимание на изучении различных методов обработки данных. Подготовить к решению различных практических задач с использованием программных средств.

ПРАКТИЧЕСКАЯ РАБОТА №1
ЭМПИРИЧЕСКАЯ ФУНКЦИЯ И ГИСТОГРАММА

Цель работы – изучение основ теории случайных величин и обработки экспериментальных данных.

Основные теоретические сведения

Различают *дискретные* и *непрерывные* случайные величины. Например, количество отказов системы есть дискретная случайная величина ξ. Ошибка при измерении тока или напряжения - пример непрерывной случайной величины. Совокупность всех возможных значений x_i дискретной случайной величины и соответствующих вероятностей $p_i = p(\xi = x_i)$ называют рядом распределения. Как дискретная, так и непрерывная случайные величины могут быть заданы функцией распределения [1]

$$F(x) = p(\xi < x). \tag{1.1}$$

Функция F(x) монотонно возрастает на всей числовой оси, причем $F(-\infty) = 0, F(+\infty) = 1$. Плотностью распределения случайной величины ξ называют функцию

$$f(x) = F'(x). \tag{1.2}$$

Если закон распределения случайной величины ξ неизвестен, то его можно приближенно определить (оценить) опытным путем. С этой целью над величиной ξ проводят ряд независимых испытаний (измерений). Вся мыслимая (бесконечная) совокупность этих измерений называется *генеральной совокупностью*. А каждый конкретный ряд измерений $(x_1, x_2, \ldots, x_n)$ называют *простой случайной выборкой.*

Если простую выборку упорядочить по возрастанию, то ее называют *вариационным рядом*. Если для каждого неповторяющегося элемента вариационного ряда x_i указать относительную частоту его появления $p_i^* = \frac{m_i}{n}$, то такой вариационный ряд называют *статистическим рядом* распределения случайной величины ξ. Здесь m_i – число повторений x_i (абсолютная частота появления элемента), а n –общее число измерений, или *объем выборки*.

Имея вариационный ряд, легко построить *эмпирическую (статистическую) функцию распределения*

$$F_n(x) = \frac{m_x}{n}. \qquad (1.3)$$

Здесь m_x – число членов вариационного ряда, лежащих левее от x, а m_x/n – частота попадания выборочного значения левее x; $F_n(x)$ – ступенчатая неубывающая функция, заданная на всей числовой оси, со скачками в точках x_i. Величина скачка равна частоте p_i^*. Поскольку сумма абсолютных частот $\sum_{i=1}^{n} m_i = n$, то сумма относительных частот $\sum_{i=1}^{n} p_i^* = 1$. Можно доказать, что $F_n(x) \xrightarrow{P} F(x)$ при $n \to \infty$. Отсюда ясно, что эмпирическую функцию распределения можно использовать как оценку теоретической функции распределения $F(x)$.

При большом объеме выборки вычисления становятся громоздкими и, с целью упрощения вычислений, элементы выборки объединяют в группы (разряды). Для этого интервал, содержащий все множество элементов выборки, разбивают на r непересекающихся интервалов. При этом правый конец каждого интервала исключают из соответствующего множества, а левый включают. Ради простоты интервалы обычно выбирают одинаковой длины $h = R/r$, где $R = x_{max} - x_{min}$ –*размах выборки*. Если m_i –число элементов выборки в i-м разряде, то m_i/n –его частота.

Совокупность разрядов или их середин и соответствующих частот называют *группированным статистическим рядом*. Геометрически его изображают в виде *группированной статистической функции распределения* или в виде *гистограммы*. Гистограмма строится следующим образом. По оси абсцисс откладывают интервалы и над каждым интервалом, как на основании, строят прямоугольник, высота которого равна значению плотности распределения для данного интервала m_i/nh. Таким образом, площадь каждого прямоугольника гистограммы равна его частоте, а общая площадь равна единице.

С увеличением объема выборки n и уменьшением длины интервала гистограмма будет стремиться к кривой плотности распределения $f(x)$, поэтому гистограмму используют в качестве оценки для плотности распределения.

Определение параметров выборки

Основными параметрами экспериментальных данных являются выборочные математическое ожидание $\tilde{m}_x$, дисперсия $\tilde{D}_x$ и среднеквадратичное отклонение $\tilde{\sigma}_x$:

$$\tilde{m}_x = \frac{1}{n}\sum_{i=1}^{n} x_i, \tag{1.4}$$

$$\tilde{D}_x = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \tilde{m}_x)^2, \tag{1.5}$$

$$\tilde{\sigma}_x = \sqrt{\tilde{D}_x}. \tag{1.6}$$

Элементарная статистическая обработка данных в массиве обычно сводится к нахождению их среднего значения, медианы (срединного значения) и стандартного отклонения. Для этого в системе MATLAB определены следующие функции [5]:

mean (A) — возвращает арифметическое среднее значение элементов

массива, если A — вектор; или возвращает вектор-строку, содержащую средние значения элементов каждого столбца, если A — матрица. Арифметическое среднее значение есть сумма элементов массива, деленная на их число;

median (A) — возвращает медиану, если A — вектор; или вектор-строку медиан для каждого столбца, если A — матрица;

std(X) — возвращает стандартное отклонение элементов массива, вычисляемое по формуле если X — вектор. Если X — матрица, то std(X) возвращает вектор-строку, содержащую стандартное отклонение элементов каждого столбца

sort (A) — в случае одномерного массива A сортирует и возвращает элементы по возрастанию их значений; в случае двумерного массива происходит сортировка и возврат элементов каждого столбца.

Пример. При испытаниях технических средств телекоммуникаций проводятся измерения мощности несущей (carrier power), при этом проводится не менее двадцати измерений, результаты которых заносятся в рабочий журнал испытательной лаборатории.

```
>> x = [21.8, 22.8, 23.0, 22.5, 22.1, 22.7, 21.7, 22.3, 22.7, 22.4, 22.6, 21.9, 22.3,
       22.2, 22.4, 22.8, 22.5, 22.6, 22.3, 22.4];
```

Определение среднего значения [mean_value]:

```
>> mean_value = mean(x);
mean_value =
  22.4000
```

Определение среднеквадратического отклонения [std_value]:

```
>> std_value = std(x);
std_value =
```

0.3418

Определение дисперсии [dispersion_value]:

```
>> dispersion_value = (std_value)^2;
dispersion_value =
   0.1168
```

Построение гистограммы выборки

```
>> [m, xout] = hist(x, 21:0.5:23);
>> bar(xout, m/length(x))
>> grid on
```

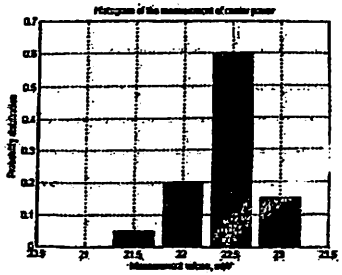

Рис.1.1. Гистограмма экспериментальных данных

Содержание отчета

Отчет должен содержать :

- теоретическое введение;
- номер варианта;
-исходные экспериментальные данные к работе (заданные преподавателем);
-текст программы для определения параметров и построения гистограммы экспериментальных данных;
- выводы о проделанной работе.

Контрольные вопросы

1. Дайте определение функции и плотности распределения случайных величин?
2. Дайте определение генеральной совокупности, выборки, размаха выборки и объема выборки.
3. Что мы называем вариационным и статистическим рядом, функцией распределения и статистической функцией распределения ?
4. Какими свойствами обладает статистическая функция распределения?
5. Дайте определение группированного статистического ряда. Как строится гистограмма ?
6. Основные функции MATLAB для обработки экспериментальных данных.

ПРАКТИЧЕСКАЯ РАБОТА №2
ОПРЕДЕЛЕНИЕ ЗАКОНА РАСПРЕДЕЛЕНИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Цель работы – изучение основных теоретических законов распределения и определения закона распределения экспериментальных данных визуальным способом.

Основные теоретические сведения

Для выполнения данной работы необходимо знать основные теоретические законы распределения случайных величин [3].

Плотность и функция случайной величины с равномерным распределением в интервале (a,b)

$$f_x(x)=\begin{cases}0; & x\notin[a,b]\\ \dfrac{1}{b-a}; & x\in[a,b]\end{cases}\quad F_x(x)=\begin{cases}0; & x\notin[a,b]\\ \dfrac{x-a}{b-a}; & x\in[a,b]\end{cases} \tag{2.1}$$

Математическое ожидание $M=\dfrac{a+b}{2}$ и дисперсия $D=\dfrac{(b-a)^2}{12}$.

Случайная величина с экспоненциальным рапределением

$$f_x(x)=\begin{cases}0; & x<0;\\ \lambda e^{-\lambda x}; & x\geq 0;\end{cases}\quad F_x(x)=\begin{cases}0; & x<0;\\ 1-e^{-\lambda x}; & x\geq 0.\end{cases} \tag{2.2}$$

Математическое ожидание $M=\dfrac{1}{\lambda}$ и дисперсия $D=\dfrac{1}{\lambda^2}$.

Случайная величина с нормальным распределением

$$f_x(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{(x-m)^2}{2\sigma^2}};\quad F_x(x)=\int_{-\infty}^{x}f_x(t)dt=\Phi\left(\frac{x-m}{\sigma}\right)+0.5\,, \tag{2.3}$$

где $\Phi(u)$ – интеграл Лапласа, m - математическое ожидание, $\sigma=\sqrt{D}$ - среднеквадратичное отклонение.

Случайная величина с Рэлеевским распределением

$$f_x(x)=\begin{cases}0; & x<0;\\ \dfrac{x}{\sigma^2}e^{\frac{x^2}{2\sigma^2}}; & x\geq 0;\end{cases}\quad F_x(x)=\begin{cases}0; & x<0;\\ 1-e^{\frac{x^2}{2\sigma^2}} & x\geq 0.\end{cases} \tag{2.4}$$

Датчик случайных чисел с равномерным распределением вероятностей в интервале 0-1.

Гистограмма для данного распределения

```
hist(rand(10000, 1), 0:0.1:1)
grid on
```

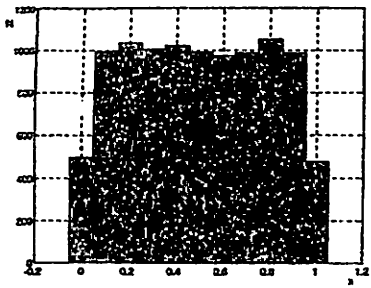

Рис.2.1. Гистограмма случайных чисел с равномерным распределением вероятностей.

Гистограмма нормального закона распределения (Гаусса закон распределения)

MU = 0; SIGMA = 1;

hist(normrnd(MU, SIGMA, 10000, 1), -3.9:0.1:3.9)

grid on

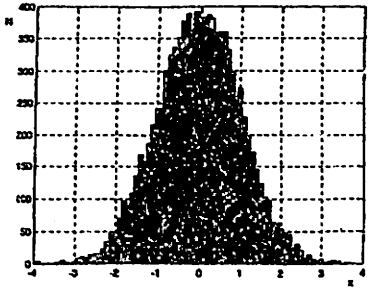

Рис.2.2. Гистограмма случайных чисел с нормальным законом распределения.

Плотность распределения нормального закона распределения

```
plot(-3.9:0.1:3.9, normpdf(-3.9:0.1:3.9, MU, SIGMA), '.')
grid on
```

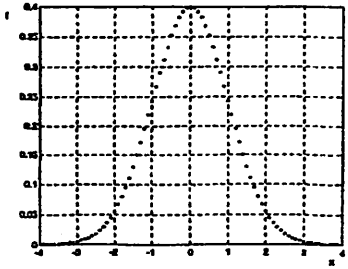

Рис.2.3. Плотность нормального закона распределения вероятностей случайных чисел.

Функция распределения нормального закона распределения

```
stairs(-3.9:0.1:3.9, normcdf(-3.9:0.1:3.9, MU, SIGMA))
grid on
```

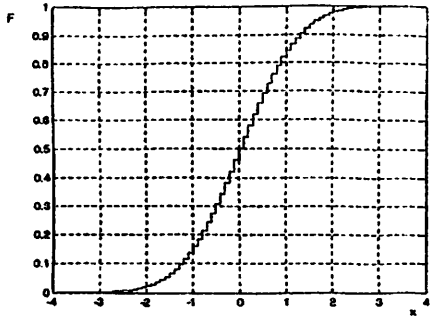

Рис.2.4. Функция нормального закона распределения вероятностей случайных чисел.

Гистограмма закон распределения Пуассона

```
LAMBDA = 2;
hist(poissrnd(LAMBDA, 10000, 1))
grid on
```

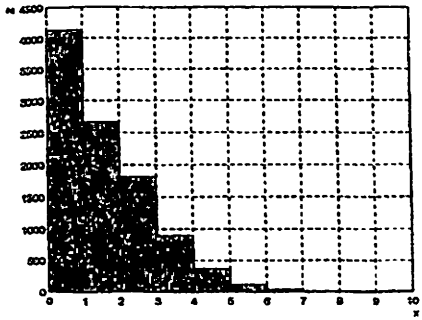

Рис.2.5. Гистограмма закон распределения Пуассона

Плотность распределения закона Пуассона

```
plot(0:15, poisspdf(0:15, LAMBDA)
grid on
```

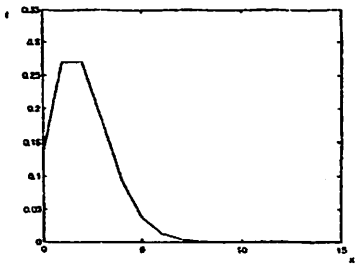

Рис.2.6. Плотность распределения закона Пуассона

Функция распределения закона Пуассона

```
stairs(0:15, poisscdf(0:15, LAMBDA))
grid on
```

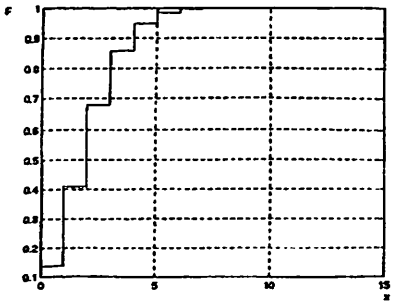

Рис.2.7. Функция распределения закона Пуассона.

Определение закона распределения экспериментальных данных визуальным способом

Для начала рисуем гистограмму (рис.2.8).

```
[m, xout] = hist(x, 10);
plot_bargraph = bar(xout, m/length(x));
```

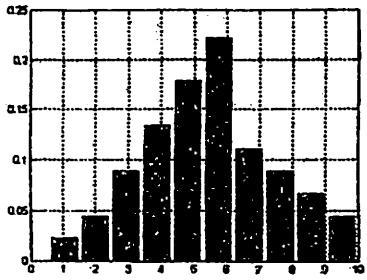

Рис.2.8. Гистограмма экспериментальных данных.

Далее поверх гистограммы рисуем следующие законы распределения (рис.2.9):

Нормальный (Гаусса) закон распределения;

Экспоненциальный закон распределения;

Равномерный закон распределения;

Релея закон распределения.

Параметры, входящие в выражение для функции и плотности теоретического распределения, найдём исходя из принципа максимума правдоподобия: так, чтобы вычисленные по этим параметрам математическое ожидание (для 1-параметрических законов) или математическое ожидание и дисперсия (для 2-параметрических законов) совпали с выборочными. Так, для нормального распределения параметры m и σ берём равными соответственно выборочным математическому ожиданию и дисперсии:

$$m=\tilde{m}_x;\quad \sigma=\tilde{\sigma}_x.$$

Для показательного распределения параметр λ находим:

$$\lambda=\frac{1}{\tilde{m}_x}.$$

Параметры равномерного распределения a и b будут равны:

$$a=\tilde{m}_x-\tilde{\sigma}_x\sqrt{3};\quad b=\tilde{m}_x+\tilde{\sigma}_x\sqrt{3}.$$

Параметр σ для Рэлеевского распределения равен:

$$\sigma=\tilde{m}_x\sqrt{\frac{2}{\pi}}.$$

```
% Perform Normal Distribution
[MU, SIGMA] = normfit(x);
y = normpdf(x, MU, SIGMA);
plot_normdst = plot(x, y, '--ko', 'MarkerEdgeColor', 'k', 'MarkerFaceColor', 'g',
'MarkerSize', 6);

% Perform Exponential Distribution
MU = expfit(x);
y = exppdf(x, MU);
plot_expdst = plot(x, y, '-.r*');

% Perform Uniform (Continuous) Distribution
[A, B] = unifit(x);
y = unifpdf(x, A, B);
plot_unifdst = plot(x, y, 'b', 'LineWidth', 2);

% Perform Rayleigh Distribution
B = raylfit(x);
y = raylpdf(x, B);
plot_rayldst = plot(x, y, '.m');

grid on
hold off
```

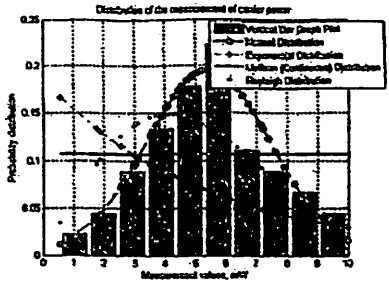

Рис.2.9. Эмпирическое и теоретические законы распределения.

Анализируем полученные результаты (рис.2.9). Какое теоретическое распределение лучше всего согласуется с эмпирическим: нормальное, показательное, равномерное или Релеевское? Выберем наиболее подходящее из них.

Содержание отчета

Отчет должен содержать :

- теоретическое введение;
- номер варианта;
- исходные экспериментальные данные, заданные преподавателем;
- текст программы для определения вида распределения экспериментальных данных;
- выводы о проделанной работе.

Контрольные вопросы

1. Функция и плотности распределения теоретических законов распределения.

2. Математическое ожидание и дисперсия случайных величин с теоретическими законами распределения.

3. Функции MATLAB для построения функции и плотности распределения случайных величин.

4. Принцип визуального определения вида распределения экспериментальных данных.

ПРАКТИЧЕСКАЯ РАБОТА №3
ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Цель работы – проверка правильности подбора теоретического распределения экспериментальных данных с помощью критериев Пирсона и Колмогорова.

Основные теоретические сведения

В предыдущей работе на основе гистограммы экспериментальных данных визуальным способом подбирали наиболее подходящий вид теоретического распределения. В данной работе определим правильность нашего решения по выбору теоретического распределения с помощью критериев Пирсона и Колмогорова.

В критерии согласия Пирсона сравниваются между собой теоретические и эмпирические числа попаданий экспериментальных данных в интервалы гистограммы. Эмпирические числа попаданий в эти интервалы n_j мы сравниваем с теоретическим числом попаданий np_j, где p_j – вероятность попадания нашей величины в j-й интервал. Теоретическое распределение можно считать подобранным верно на уровне значимости p, если суммарная квадратичная относительная разность между теоретическим и практическим числом попаданий в каждый интервал будет не очень большой: должно

выполняться условие

$$\sum_{j=1}^{k}\frac{(n_j-np_j)^2}{np_j}\le\chi^2_{1-p}(k-3), \qquad (3.1)$$

где χ^2- критерий Пирсона (хи-квадрат), $np_j\ge5$.

Критерий согласия Колмогорова применяется для проверки правильности подбора теоретического распределения. Для его применения нужно найти максимальную по модулю разность между теоретической (подобранной) функцией распределения $F_x(x)$ и выборочной (эмпирической) $\tilde{F}_x(x)$:

$$D=\max_{\forall x}\left|F_x(x)-\tilde{F}_x(x)\right|, \qquad (3.2)$$

а по ней вычислить $\lambda=D\sqrt{n}$, которую сравнить с квантилем λ-распределения Колмогорова. Если величина λ не очень большая (не превосходит квантиля λ_p), то на уровне значимости p статистическую гипотезу можно принять. Если же $\lambda>\lambda_p$, то теоретическое распределение подобрано неверно.

Проверка гипотезы о нормальном распределении экспериментальных данных

Рассмотрим критерий Пирсона χ^2. Требуется проверить, согласуются ли экспериментальные данные с гипотезой о том, что случайная величина X имеет данный закон распределения (заданный функцией распределения или плотностью). Назовем этот закон распределения "теоретическим".

Пользуясь теоретическим нормальным законом распределения с параметрами [Mx_asterisk] и [SIGMA_asterisk]

```
[m, xout] = hist(x, 10);
 for i = 1:length(m)
   Mx_asterisk(i) = xout(i)*(m(i)/length(x));
```

```
   Mx_asterisk = sum(Mx_asterisk);
end
 for i = 1:length(m)
   ALPHA2_asterisk(i) = xout(i)^2*(m(i)/length(x));
   ALPHA2_asterisk = sum(ALPHA2_asterisk);
end
 Dx_asterisk = ALPHA2_asterisk - Mx_asterisk^2;
SIGMA_asterisk = sqrt(Dx_asterisk),
```

находим вероятности попадания в разряды [m]:

```
for i = 1:length(m)
   u(i) = (xout(i)-Mx_asterisk)/SIGMA_asterisk;
end

f_theoretical = (length(x)/SIGMA_asterisk)*normpdf(u, 0, 1);
```

Определяем значение меры расхождения [chi2_empirical]:

```
for i = 1:length(m)
   chi2_empirical(i) = ((m(i)-f_theoretical(i))^2)/f_theoretical(i);
end

chi2_empirical = sum(chi2_empirical);

chi2_empirical =
   2.0537
```

Находим критическую точку [chi2_critical] правосторонней критической области по уровню значимости $\alpha = 0{,}01$.

```
chi2_critical = chi2inv(0.99, length(m) - 3);

chi2_critical =
  18.4753
```

Так как

$$\chi^2_{empirical} = 2{,}0537 < 18{,}4753 = \chi^2_{critical}$$

то гипотезу о нормальном распределении генеральной совокупности принимаем.

Рассмотрим критерий согласия Колмогорова. В качестве меры расхождения между теоретическим и эмпирическим распределениями будем рассматривать максимальное значение абсолютной величины разности между эмпирической функцией распределения [Fx_empirical] и соответствующей теоретической функцией распределения [Fx_theoretical]. Рисуем обе эти функции (рис 3.1).

```
% Calculate and plot the empirical (Kaplan-Meier) cumulative distribution
function
[Fx_empirical, xx] = ecdf(x, 'alpha', 0.01); % Dependencies on 'ecdf'
stairs(xx, Fx_empirical, 'r');
hold on
 % Calculate and plot the teoretical (normal) cumulative distribution function
MU = mean(x); SIGMA = std(x); alpha = 0.01;
Fx_teoretical = normcdf(x, MU, SIGMA, alpha); % Dependencies on 'normcdf'
stairs(x, Fx_teoretical, ':b');
grid on
hold off
```

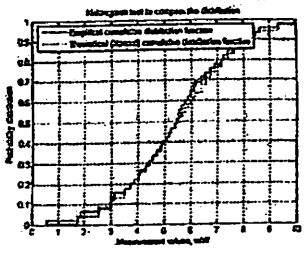

Рис.3.1. Теоретическое и эмпирическое распределения вероятностей экспериментальных данных.

Определяем меру расхождения между теоретическим и эмпирическим распределением [D] и вычисляем величину [LAMBDA_empirical]:

```
Fx_empirical = Fx_empirical'; Fx_empirical(:,1) = []; % Observe cumulative distribution function

D = max(abs(Fx_empirical - Fx_teoretical));

D =
   0.0499

LAMBDA_empirical = D*sqrt(length(x));

LAMBDA_empirical =
   0.3349
```

Определяем критическое значение критерия Колмогорова по уровню значимости $\alpha = 0{,}01$:

```
LAMBDA_critical = kolminv(0.99);
```

LAMBDA_critical =

1.6276

Так как

$$\lambda_{emprical} = 0{,}3349 < 1{,}6276 = \lambda_{critical},$$

то гипотезу о нормальном распределении экспериментальных данных принимаем.

Содержание отчета

Отчет должен содержать :

- теоретическое сведение;
- номер варианта;
- исходные эксперементальные данные, заданные преподователем в работе № 2;
- текст программы проверки гипотезы о законе распределения экспериментальных данных;
- выводы о проделанной работе.

Контрольные вопросы

1. Что такое критерий согласия?
2. Какие критерии согласия Вы знаете?
3. Опишите схему применения критериев согласия Колмогорова и Пирсона.
4. Функции MATLAB для определения критериев согласия.

ПРАКТИЧЕСКАЯ РАБОТА №4
ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Цель работы – изучение методов оценки параметров распределения экспериментальных данных, оценка параметров нормального распределения методом наибольшего правдоподобия.

Основные теоретические сведения

Наиболее часто применяемыми числовыми характеристиками случайной величины ξ являются начальные и центральные моменты различного порядка. Для дискретной случайной величины моменты порядка k определяются следующими формулами:

$$\alpha_k = \sum_{i=1}^{n} x_i^k p_i, \mu_k = \sum_{i=1}^{n} (x_i - m_\xi)^k p_i, \quad (4.1)$$

для непрерывной случайной величины ξ

$$\alpha_k = \int_{-\infty}^{\infty} x^k f(x)dx, \mu_k = \int_{-\infty}^{\infty} (x - m_\xi)^k f(x)dx. \quad (4.2)$$

Чаще всего используется первый начальный момент $\alpha_1 = m_\xi$, называемый *математическим ожиданием случайной величины* ξ, и второй центральный момент $\mu_2 = D_\xi$, называемый *дисперсией.* Матожидание - это среднее значение случайной величины, его называют еще центром распределения, дисперсия характеризует разброс случайной величины

относительно центра распределения. Часто вместо дисперсии используют среднее квадратичное отклонение $\sigma_\xi = \sqrt{D_\xi}$.

Если закон распределения случайной величины неизвестен, то мы не сможем вычислить числовые характеристики. В этом случае их заменяют оценками, полученными как функции выборки $x = (x_1, x_2, \ldots, x_n)$. Всякую функцию $t_n(x)$ от выборки называют статистикой. Подходящую статистику используют в качестве оценки числовой характеристики. Чаще всего оценками начальных и центральных моментов служат соответствующие выборочные начальные и центральные моменты

$$a_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k ; m_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - Mx)^k . \qquad (4.3)$$

Таким образом, оценкой математического ожидания служит выборочное среднее

$$Mx = \frac{1}{n}\sum_{i=1}^{n} x_i . \qquad (4.4)$$

Пусть закон распределения известен, но зависит от одного или нескольких неизвестных параметров . Например, $f(x,\theta)$ - известная плотность распределения, а $\theta = (\theta_1, \theta_2, .., \theta_n)$ - неизвестный параметр. Требуется по выборке $x = (x_1, x_2, \ldots, x_n)$ оценить параметр θ .

Существует несколько методов оценки параметра θ. Мы рассмотрим два из них - метод моментов и метод функции правдоподобия [4].

Метод моментов заключается в том, что теоретический момент k-го порядка $\alpha_k = \alpha_k(\theta)$ приравнивают к соответствующему выборочному моменту a_k. Из полученного уравнения $\alpha_k(\theta) = a_k$ находят неизвестный параметр θ. Например, случайная величина ξ (время безотказной работы радиоаппаратуры) распределена по экспоненциальному закону

$$f(t) = \frac{1}{T} e^{-\frac{t}{T}}, t \geq 0, \qquad (4.5)$$

где T - неизвестный параметр. Оценим его по методу моментов. Для этого найдем первый начальный момент

$$\alpha_1 = \int_{-\infty}^{\infty} t f(t) dt = \frac{1}{\mathrm{T}} \int_0^{\infty} t e^{-\frac{t}{\mathrm{T}}} dx = \mathrm{T}. \qquad (4.6)$$

Так как первый выборочный момент равен Mx, то из равенства $\alpha_1 = a_1$ получим $T = Mx$. Таким образом, оценкой неизвестного параметра T, найденной по методу моментов, является среднее выборочное Mx.

Пусть $L(u,\theta)$- плотность распределения выборочного вектора $x = (x_1, x_2, .., x_n)$, $\theta = (\theta_1, \theta_2, .., \theta_s)$ – неизвестный параметр. $L(u,\theta)$– функция двух аргументов, неслучайного θ и случайного $x = (x_1, x_2, .., x_n)$ называется функцией правдоподобия. Так как $L(u,\theta)$– плотность распределения, то оценка параметра θ, доставляющая максимум функции правдоподобия, является наиболее вероятной. Отсюда

$$\frac{\partial L(x,\theta)}{\partial \theta} = 0 \text{ или } \frac{\partial}{\partial \theta}[\ln L(x,\theta)] = 0 \qquad (4.7)$$

есть необходимые условия существования максимума.

Пусть $x = (x_1, x_2, .., x_n)$ - случайная выборка из генеральной совокупности, распределенной по нормальному закону

$$f(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-a)^2 / 2\sigma^2), \qquad (4.8)$$

$$f(t) = \frac{1}{T} e^{-\frac{t}{T}}, t \geq 0, \qquad (4.5)$$

где T - неизвестный параметр. Оценим его по методу моментов. Для этого найдем первый начальный момент

$$\alpha_1 = \int_{-\infty}^{\infty} t f(t) dt = \frac{1}{T} \int_0^{\infty} t e^{-\frac{t}{T}} dx = T. \qquad (4.6)$$

Так как первый выборочный момент равен Mx, то из равенства $\alpha_1 = a_1$ получим $T = Mx$. Таким образом, оценкой неизвестного параметра T, найденной по методу моментов, является среднее выборочное Mx.

Пусть $L(u,\theta)$- плотность распределения выборочного вектора $x = (x_1, x_2, .., x_n)$, $\theta = (\theta_1, \theta_2, .., \theta_r)$ – неизвестный параметр. $L(u,\theta)$– функция двух аргументов, неслучайного θ и случайного $x = (x_1, x_2, .., x_n)$ называется функцией правдоподобия. Так как $L(u,\theta)$– плотность распределения, то оценка параметра θ, доставляющая максимум функции правдоподобия, является наиболее вероятной. Отсюда

$$\frac{\partial L(x,\theta)}{\partial \theta} = 0 \text{ или } \frac{\partial}{\partial \theta}\left[\ln L(x,\theta)\right] = 0 \qquad (4.7)$$

есть необходимые условия существования максимума.

Пусть $x = (x_1, x_2, .., x_n)$ - случайная выборка из генеральной совокупности, распределенной по нормальному закону

$$f(x,\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-(x-a)^2 / 2\sigma^2), \qquad (4.8)$$

где $\theta=(a,\sigma)$ - неизвестный параметр.

Запишем функцию правдоподобия. Так как x_i - независимые случайные величины, распределенные по тому же закону, а плотность распределения вектора равна произведению плотностей составляющих вектора, то функция правдоподобия будет следующей:

$$L(x,\theta)=\prod_{i=1}^{n} f(x_i,\theta)=\frac{1}{\sigma^n(2\pi)^{\frac{n}{2}}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2\right]. \qquad (4.9)$$

Оценка параметров нормального распределения

Ниже приведена программа оценки параметров нормального распределения с помощью метода максимального правдоподобия.

```
% Метод максимального правдоподобия
%Оценка параметров норм. закона распределения
x=[                    ];
x=sort(x);
n=length(x);
mu=mean(x);
sigma=std(x);
% Вычисление моментов
% 1-й начальный момент (оценка матем. ожидания)
m1=1/n*sum(x);
% 2-й центральный момент (оценка дисперсии)
m2=1/(n-1)*sum((x-m1).^2);
% оценка ср. кв. отклонения
s=sqrt(m2);
% Оценка параметров   норм. распределения
% Плотность нормального распределения
f=inline(...
```

- текст прогроаммы и результаты оценки параметров эксперементальных данных с нормальным распределением;
- выводы по проделанной работе.

Контрольные вопросы

1. Назовите выборочные числовые характеристики.
2. Методы оценки параметров.
3. Сущность метода моментов.
4. Что такое функция правдоподобия? В чем сущность метода наибольшего правдоподобия ?

ПРАКТИЧЕСКАЯ РАБОТА №5
АППРОКСИМАЦИЯ И ИНТЕРПОЛЯЦИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Цель работы – изучение методов аппроксимации и интерполяции, аппроксимация экспериментальных данных методом наименьших квадратов, изучение функции и средств MATLAB для аппроксимации и интерполяции экспериментальных данных.

Основные теоретические сведения

Аппроксимация, или приближение – математический метод, состоящий в замене одних математических объектов другими, в том или ином смысле близкими к исходным, но более простыми. Аппроксимация позволяет исследовать числовые характеристики и качественные свойства объекта, сводя задачу к изучению более простых или более удобных объектов (например, таких, характеристики которых легко вычисляются, или свойства

которых уже известны).

Метод наименьших квадратов применяется для приближенного представления заданной функции другими (более простыми) функциями и часто оказывается полезным при обработке наблюдений.

Когда искомая величина может быть измерена непосредственно, как, например, уровень шума или затухание, то, для увеличения точности, измерение производится много раз, и за окончательный результат берут арифметическое среднее из всех отдельных измерений. Это правило арифметической середины основывается на соображениях теории вероятностей; легко показать, что сумма квадратов уклонений отдельных измерений от арифметической середины будет меньше, чем сумма квадратов уклонений отдельных измерений от какой бы то ни было другой величины. Само правило арифметической середины представляет, следовательно, простейший случай метода наименьших квадратов.

Требование метода наименьших квадратов: для того, чтобы данная совокупность наблюденных значений $y_1, y_2, \ldots y_n$ была наивероятнейшей, нужно выбрать функцию $\varphi(x)$ так, чтобы сумма квадратов отклонений наблюденных значений от $\varphi(x)$ была минимальной [4]

$$\sum_{i=1}^{n}[y_i - \varphi(x_i)]^2 = \min$$

Таким образом, обосновывается метод наименьших квадратов, исходя из нормального закона ошибок измерения и требования максимальной вероятности данной совокупности ошибок.

Пример 1. В опыте зарегистрирована совокупность значений x_i, y_i (рис.5.1)

x = [0.30, 1.57, 2.84, 4.11, 5.38, 6.65, 7.92, 9.19, 10.46, 11.73];
y = [15.33, 4.55, 3.41, 2.97, 2.74, 2.60, 2.59, 2.44, 2.38, 2.34];

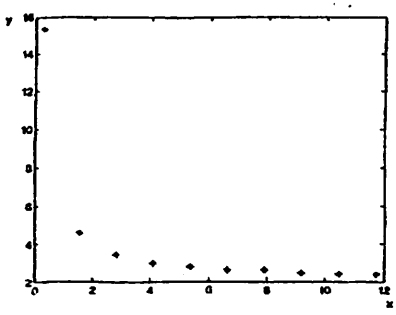

Рис.5.1. Результаты экспериментальных данных.

Требуется подобрать по методу наименьших квадратов параметры a и b функции, изображающей данную экспериментальную зависимость

$$y = \varphi(x,a,b) = a + \frac{b}{x}$$

Дифференцируя это выражение по a и b, имеем

```
syms x a b
diff(a + b./x, 'a')

ans =
1

diff(a + b./x, 'b')

ans =
1/x
```

$$\frac{\partial \varphi}{\partial a} = 1; \left(\frac{\partial \varphi}{\partial a}\right)_i = 1;$$

$$\frac{\partial \varphi}{\partial b} = \frac{1}{x}; \left(\frac{\partial \varphi}{\partial b}\right)_i = \frac{1}{x}$$

Подставляя полученные значения в основную систему уравнений,

получим два уравнения для определения a и b:

$$\sum_{i=1}^{n}\left[y_i-\left(a+\frac{b}{x_i}\right)\right]=0$$

$$\sum_{i=1}^{n}\left[y_i-\left(a+\frac{b}{x_i}\right)\right]\cdot\frac{1}{x_i}=0$$

раскрывая скобки и производя суммирование, получаем

$$n\cdot a+b\cdot\sum_{i=1}^{n}\frac{1}{x_i}=\sum_{i=1}^{n}y_i$$

$$a\cdot\sum_{i=1}^{n}\frac{1}{x_i}+b\cdot\sum_{i=1}^{n}\frac{1}{x_i^2}=\sum_{i=1}^{n}\frac{y_i}{x_i}$$

Находим корни системы линейных алгебраических уравнений по методу Гаусса:

```
% Perform Gaussian elimination
left = [length(x), sum(1./x); sum(1./x), sum(1./x.^2)];
right = [sum(y); sum(y./x)];
koef = left\right;
result = struct('a', koef(1), 'b', koef(2))

result =
    a: 2.0103
    b: 3.9954
```

Таким образом, поставленная задача решена по методу наименьших квадратов, и зависимость, связывающая y и x, имеет вид:

$$y=2{,}0103+\frac{3{,}9954}{x}$$

Отобразим найденную функцию на графике (рис.5.2)

получим два уравнения для определения a и b:

$$\sum_{i=1}^{n}\left[y_i-\left(a+\frac{b}{x_i}\right)\right]=0$$

$$\sum_{i=1}^{n}\left[y_i-\left(a+\frac{b}{x_i}\right)\right]\cdot\frac{1}{x_i}=0$$

раскрывая скобки и производя суммирование, получаем

$$n\cdot a+b\cdot\sum_{i=1}^{n}\frac{1}{x_i}=\sum_{i=1}^{n}y_i$$

$$a\cdot\sum_{i=1}^{n}\frac{1}{x_i}+b\cdot\sum_{i=1}^{n}\frac{1}{x_i^2}=\sum_{i=1}^{n}\frac{y_i}{x_i}$$

Находим корни системы линейных алгебраических уравнений по методу Гаусса:

```
% Perform Gaussian elimination
left = [length(x), sum(1./x); sum(1./x), sum(1./x.^2)];
right = [sum(y); sum(y./x)];
koef = left\right;
result = struct('a', koef(1), 'b', koef(2))

result =
    a: 2.0103
    b: 3.9954
```

Таким образом, поставленная задача решена по методу наименьших квадратов, и зависимость, связывающая y и x, имеет вид:

$$y=2{,}0103+\frac{3{,}9954}{x}$$

Отобразим найденную функцию на графике (рис.5.2)

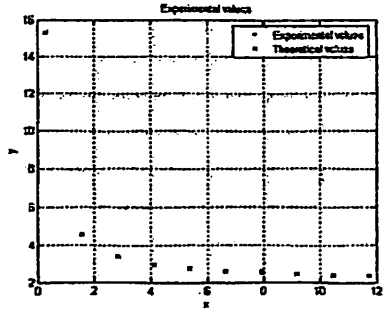

Рис.5.2. Результат аппроксимации экспериментальных данных.

Пример 2. – В условиях предыдущей задачи подобрать по методу "натянутой нити" параметры функции a и b, изображающей экспериментальную зависимость

$$y = a + \frac{b}{x}$$

В качестве отсчетных точек y и x выберем крайние точки из совокупности значений. Подставляя полученные значения в основную систему уравнений, получим два уравнения для определения a и b:

$$a + b \cdot \frac{1}{x_1} = y_1$$

$$a + b \cdot \frac{1}{x_n} = y_n$$

Находим корни системы линейных алгебраических уравнений по методу Гаусса:

```
% Perform Gaussian elimination
left = [1, 1/x(1); 1, 1/x(end)];
right = [y(1); y(end)];
koef = left\right;
result = struct('a', koef(1), 'b', koef(2))

result =
   a: 1.9991
```

Функции MATLAB для аппроксимации

Функция **polyfit** находит коэффициенты полинома заданной степени n, который аппроксимирует данные (или функцию y(x)) в смысле метода наименьших квадратов:

p = polyfit(x, y, n)

Полином третьей степени (n=3):

```
% Intervals set:
I = 0.1:0.2:1.9;
% Density set:
m = [4 126 230 260 130 120 50 30 30 20];

coef = polyfit(I, m, 3);
fun = polyval(coef, I);
figure
plot(I, m, '--*r', I, fun)

grid on

result = struct('coef', coef)

result =
    coef: [379.9048 -1.3039e+003 1.1798e+003 -101.6007]
```

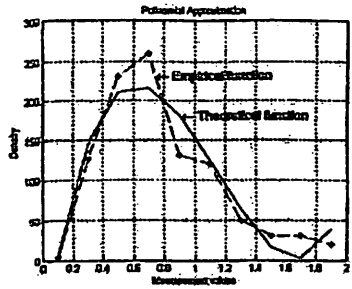

Рис.5.3. Аппроксимация экспериментальных данных полиномом третьей степени.

Соответствующее аналитическое выражение имеет вид

$$P_3(x) = 379{,}90 \cdot x^3 + 1303{,}92 \cdot x^2 + 1179{,}80 \cdot x - 101{,}60$$

Из рис.5.3 видно, что полином примерно повторяет ход экспериментальной кривой, но дает плохое согласие с экспериментом.

Полином седьмой степени (n=7) :

```
coef = polyfit(I, m, 7)
fun = polyval(coef, I);
figure
plot(I, m, '--*r', I, fun)

grid on
```

```
result =
   coef: [-1.443e+003 1.098e+004 -3.406e+004 5.479e+004 -4.761e+004 2.058e+004 -
3.276e+003 169]
```

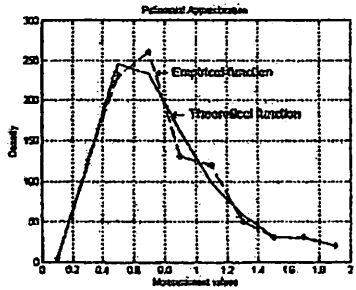

Рис.5.4. Аппроксимация экспериментальных данных полиномом седьмой степени.

В этом случае имеем

$$P_7(x) = -1{,}443 \cdot 10^3 \cdot x^7 + 1{,}098 \cdot 10^4 \cdot x^6 - 3{,}406 \cdot 10^4 \cdot x^5 + 5{,}479 \cdot 10^4 \cdot x^4$$
$$- 4{,}761 \cdot 10^4 \cdot x^3 + 2{,}058 \cdot 10^4 \cdot x^2 - 3{,}276 \cdot 10^3 \cdot x + 169$$

Визуальная оценка расчетной и экспериментальной кривых на рисунках 5.3 и 5.4 показывает, что согласие в последнем лучше. Анализ полиномов с другими степенями показывает, что с ростом , имеется тенденция к лучшему

описанию опыта полиномом.

Интерполяция экспериментальных данных

Интерполяция является процессом вычисления (оценки) промежуточных значений функций, которые находятся между известными или заданными точками. Она имеет важное применение в таких областях как теория сигналов, обработка изображений и других. MATLAB обеспечивает ряд интерполяционных методик, которые позволяют находить компромисс между точностью представления интерполируемых данных и скоростью вычислений и используемой памятью.

Функции MATLAB для интерполяции данных

griddata	Двумерная интерполяция на неравномерной сетке.
griddata3	Трехмерная интерполяция на неравномерной сетке.
griddatan	Многомерная интерполяция (n >= 3).
interp1	Одномерная табличная интерполяция.
interp2	Двухмерная табличная интерполяция.
interp3	Трехмерная табличная интерполяция.
interpft	Одномерная интерполяция с использованием быстрого преобразования Фурье.
interpn	Многомерная табличная интерполяция.
pchip	Кубическая интерполяция при помощи полинома Эрмита.
spline	Интерполяция кубическим сплайном.

Двумя основными типами одномерной интерполяции в MATLAB являются полиномиальная интерполяция и интерполяция на основе быстрого преобразования Фурье [5].

Функция interp1 осуществляет одномерную интерполяцию – важную операцию в области анализа данных и аппроксимации кривых. Эта функция использует полиномиальные методы, аппроксимируя имеющийся массив данных полиномиальными функциями и вычисляя соответствующие функции на заданных (желаемых) точках. В наиболее общей форме эта функция имеет вид

$$yi = interp1(x, y, xi, method),$$

где y есть вектор, содержащий значения функции; x – вектор такой же длины, содержащий те точки (значения аргумента), в которых заданы значения y; вектор xi содержит те точки, в которых мы хотим найти значения вектора y путем интерполяции; **method** – дополнительная строка, задающая метод интерполяции. Имеются следующие возможности для выбора метода:

• *Ступенчатая интерполяция* (**method** = 'nearest'). Этот метод приравнивает значение функции в интерполируемой точке к ее значению в ближайшей существующей точке имеющихся данных.

• *Линейная интерполяция* (**method** = 'linear'). Этот метод аппроксимирует функцию между любыми двумя существующими соседними значениями как линейную функцию, и возвращает соответствующее значение для точки в xi (метод используется по умолчанию).

• *Интерполяция кубическими сплайнами* (**method** = 'spline'). Этот метод аппроксимирует интерполируемую функцию между любыми двумя соседними значениями при помощи кубических функций, и использует сплайны для осуществления интерполяции.

• *Кубическая интерполяция* (**method** = 'pchip' или 'cubic'). Эти методы идентичны. Они используют кусочную кубическую Эрмитову аппроксимацию и сохраняют монотонность и форму данных.

При выборе метода интерполяции всегда нужно помнить, что некоторые из них требуют большего объема памяти или выполняются быстрее, чем

другие. Однако, вам может потребоваться использование любого из этих методов, чтобы достичь нужной степени точности интерполяции. При этом нужно исходить из следующих критериев.

• Метод ступенчатой аппроксимации является самым быстрым, однако он дает наихудшие результаты с точки зрения гладкости.

• Линейная интерполяция использует больше памяти чем ступенчатая и требует несколько большего времени исполнения. В отличие от ступенчатой аппроксимации, результирующая функция является непрерывной, но ее наклон меняется в значениях исходной сетки (исходных данных).

• Кубическая интерполяция сплайнами требует наибольшего времени исполнения, хотя требует меньших объемов памяти, чем кубическая интерполяция. Она дает самый гладкий результат из всех других методов, однако вы можете получить неожиданные результаты, если входные данные распределены неравномерно и некоторые точки слишком близки.

• Кубическая интерполяция требует большей памяти и времени исполнения чем ступенчатая или линейная.

Относительные качественные характеристики всех перечисленных методов сохраняются и в случае двух- или многомерной интерполяции.

Графический интерфейс подгонки кривых

MATLAB дает возможность осуществлять аппроксимацию данных наблюдений при помощи специального графического Интерфейса Подгонки Кривых (ИПК) (в английском оригинале - **Basic Fitting interface**). Используя данный интерфейс, вы можете легко и быстро решить множество задач подгонки кривых, получая при этом самую разнообразную информацию о результатах вашей подгонки. ИПК предоставляет следующие возможности:

- Аппроксимирует данные, используя сплайновый интерполянт, эрмитовый интерполянт, или же полиномиальный интерполянт до 10 порядка включительно.

- Осуществляет множество графических построений для заданных наборов данных.
- Строит графики невязок (ошибок подгонки).
- Анализирует численные результаты подгонки.
- Осуществляет интерполяцию или экстраполяцию данных подгонки.
- Аннотирует графики численными результатами подгонки и нормами ошибок аппроксимации.
- Запоминает результаты подгонки и вычислений в рабочем пространстве MATLAB.

Основываясь на ваших конкретных задачах и приложениях, вы можете использовать ИПК, возможности, предоставляемыми командным окном, или же комбинировать эти две возможности. Отметим, что ИПК предназначен только для работы с одномерными и двумерными данными.

Задание

1. Построить график экспериментальных данных по заданному варианту.
2. Выбрать опцию **Basic Fitting** из меню **Tools** вашего графического окна.
3. Нажать дважды на кнопку **More** в нижней части ИПК. В результате откроется окно с тремя панелями, а сама надпись заменится на **Less**.
4. Реализовать следующие опции ИПК:

Select data (Выбор данных) – В данном окне расположен список всех переменных, построенных на активном графике, с которым связан ИПК (на графике может быть построено несколько кривых). Используйте данный список для выбора требуемого (текущего) набора данных. Под текущим подразумевается тот набор данных, для которого вы хотите осуществить подгонку. За один раз вы можете осуществлять действия только с одним набором данных.

Center and scale X data (Центрирование и масштабирование данных X) – Если данная опция выбрана, то данные центрируются (нуль переносится в

среднее значение данных) и масштабируются к единичному стандартному отклонению (делятся на исходное стандартное отклонение). Это может потребоваться для повышения точности последующих математических вычислений. Если подгонка приводит к результатам, которые могут быть неточными, соответствующее предупреждение выводится на экран.

Plot fits (Подгонка кривых) – Эта панель позволяет визуально просмотреть результаты одной или более подгонок текущего набора данных.

- **Check to display fits on figure** (Отметьте методы для вывода на график) – Выберите методы подгонок, которые вы хотели бы использовать и *вывести* на график. Здесь имеются две основные возможности – выбор интерполянтов и выбор полиномов. Сплайновый интерполянт использует для аппроксимации сплайны, тогда как эрмитовый интерполянт использует специальную функцию **pchip** (Piecewise Cubic Hermite Interpolating Polynomial - Кусочно-кубический эрмитовый интерполяционный полином). Полиномиальная подгонка использует функцию **polyfit**. Вы можете одновременно выбрать любые методы подгонки для аппроксимации ваших данных. Если ваш набор данных содержит N точек, вам следует использовать для аппроксимации полиномы с не более чем N коэффициентами. В противном случае, ИПК автоматически приравнивает избыточное число коэффициентов нулю, что приводит к недоопределенности системы. Укажем, что при этом на дисплей выдается соответствующее сообщение.
- **Show equations** (Показать уравнения) – При выборе данной опции, уравнение подгонки выводится на ваш график.
- **Significant digits** (Значащие разряды) – Выберите число значащих разрядов для вывода на дисплей.
- **Plot residuals** (Построить графики разностей (невязок)) – При выборе данной опции, на график выводятся разности подгонок. Под разностью подгонки понимается разность между исходными данными и результатами подгонки для каждого значения аргумента исходных данных. Вы можете

построить графики невязок как столбчатую диаграмму (bar plot), как график рассеяния (scatter plot), или же как линейный график. Построения можно осуществлять как в том же графическом окне.

- **Show norm of residuals** (Показать норму разностей) – При выборе опции, на график выводятся также значения норм разностей. Норма разности является мерой качества подгонки, где меньшее значение нормы соответствует лучшему качеству. Норма рассчитывается при помощи функции **norm(V,2)**, где **V** есть вектор невязок.

Numerical results (Численные результаты) – Данная панель позволяет изучать численные характеристики каждой отдельной подгонки для текущего набора данных, без построения графиков.

- **Fit** (Метод подгонки) – Выберите метод подгонки. Соответствующие результаты будут представлены в окне под меню выбора метода. Заметим, что выбор метода в данной панели не оказывает воздействия на панель **Plot fits**. Поэтому, если вы хотите получить графическое представление, следует выбрать соответствующую опцию в панели **Plot fits**.
  - **Coefficients and norm of residuals** (Коэффициенты и норма невязок) – В данном окне выводятся численные выражения для уравнения подгонки, выбранного в **Fit**. Отметим, что при первом открытии панели **Numerical Results** , в рассматриваемом окне выдаются результаты последней подгонки, выбранной вами в панели **Plot fits**.
  - **Save to workspace** (Запомнить в рабочем пространстве) – Вызывает диалоговое окно, которое позволяет запомнить в рабочем пространстве результаты вашей подгонки.

Find Y = f(X) – Данная панель дает возможность произвести интерполяцию или экстраполяцию текущей подгонки.

- **Enter value(s)** (Введите данные) – Введите любое выражение, совместимое с системой MATLAB для оценки вашей текущей подгонки в

промежуточных или выходящих за пределы заданных аргументов точек. Выражение будет вычислено после нажатия кнопки **Evaluate** (Вычислить), а результаты в табличной форме будут выведены в соответствующее окно ниже. Метод текущей подгонки при этом указан в меню **Fit**.

- **Save to workspace** (Запомнить в рабочем пространстве) – Вызывает диалоговое окно, которое позволяет запомнить в рабочем пространстве результаты вашей интерполяции.
- **Plot results** (Построить графики) – При выборе данной опции, результаты интерполяции выводятся в графической форме на график данных.

Содержание отчета

Отчет должен содержать :

- теоретическое введение;
- номер варианта;
- исходные экспериментальные данные , заданные преподователем;
- текст программы аппроксимации экспериментальных данных методом наименших квадратов;
- результаты аппроксимации и интерполяции экспериментальных данных с помощью функции MATLAB;
- результаты задания, выполненного в графическом интерфейсе MATLAB для подгонки кривых.
- выводы о проделанной работе.

Контрольные вопросы

1. Сущность аппроксимации и интерполяции экспериментальных данных.
2. Различие между аппроксимации и интерполяции данных.
3. Методы аппроксимации и интерполяции данных.
4. Сущность метода наименщих квадратов.

5. Функции MATLAB для аппроксимации и интерполяции данных.

5. Возможности графического интерфейса MATLAB для подгонки кривых.

ПРАКТИЧЕСКАЯ РАБОТА №6
КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Цель работы – изучение методов корреляционного и регрессионного анализа, оценка значимости корреляции и определение уравнения регрессии экспериментальных данных.

Основные теоретические сведения

1. Корреляционный анализ

Многие объекты исследования характеризуются множеством параметров, и по результатам наблюдения за их функционированием формируются многомерные совокупности (матрицы) экспериментальных данных

$$\mathbf{X}=\begin{vmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{vmatrix}$$

Строки такой матрицы соответствуют результатам регистрации всех наблюдаемых параметров объекта в одном эксперименте, а столбцы содержат результаты наблюдений за одним параметром (фактором, вариантой) во всех экспериментах. Обозначим количество параметров через m (m>1), а количество наблюдений – через n.

Таким образом, объектом исследования в многомерном анализе является многомерная случайная величина, представленная выборкой конечного объема. К такой выборке применимы все методы и оценки,

рассмотренные при обработке одномерных экспериментальных данных.

Параметры, характеризующие объект исследования, имеют разный физический смысл, и матрица данных существенно изменяется, если изменяются шкалы, в которых измеряются те или иные параметры. Матрицу данных еще до проведения анализа целесообразно привести к стандартному виду. Стандартизованную матрицу будем обозначать через U. Переход от исходной к стандартизованной матрице осуществляется следующим образом [2]: вычисляются оценки математического ожидания

$$\mu_1(x_j)=\frac{1}{n}\sum_{i=1}^{n}x_{ij}$$

и дисперсии

$$\mu_2(x_j)=\sigma^2(x_j)=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\mu_1(x_j))^2$$

каждой варианты $j=\overline{1,m}$;

вычисляются элементы стандартизованной матрицы

$$u_{ij}=(x_{ij}-\mu_1(x_j))/\sigma(x_j),\ i=\overline{1,n},\ j=\overline{1,m}.$$

Элементы матрицы U являются безразмерными величинами. Именно матрица U будет являться объектом последующей обработки.

Более важным частным случаем статистической зависимости является *корреляционная* зависимость, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения.

Корреляционная зависимость определяется различными параметрами, среди которых наибольшее распространение получили показатели, характеризующие взаимосвязь двух случайных величин (парные показатели):

корреляционный момент, коэффициент корреляции.

Оценка *корреляционного момента* (*коэффициента ковариации*) двух вариант x_j и x_k вычисляется по исходной матрице X

$$\xi_{jk}=\frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\mu_1(x_j))(x_{ik}-\mu_1(x_k))$$

Этот показатель неудобен для практического применения.

Коэффициент ковариации r_{jk} нормированных случайных величин называют *коэффициентом корреляции*, его оценка

$$\rho_{jk}=\frac{1}{n}\sum_{i=1}^{n}u_{ij}u_{ik}$$

Значение коэффициента корреляции лежит в пределах от –1 до +1. Если случайные величины U_j и U_k независимы, то коэффициент r_{jk} обязательно равен нулю, обратное утверждение неверно. Коэффициент r_{jk} характеризует значимость линейной связи между параметрами:

Используя понятие коэффициента корреляции, матрице ЭД можно поставить в соответствие квадратную матрицу оценок коэффициентов корреляции (корреляционную матрицу)

$$\rho=\begin{vmatrix}\rho_{11} & \rho_{12} & \cdots & \rho_{1m}\\ \rho_{21} & \rho_{22} & \cdots & \rho_{2m}\\ \cdot & \cdot & \cdot & \cdot\\ \rho_{m1} & \rho_{m2} & \cdots & \rho_{mm}\end{vmatrix}$$

К числу характерных свойств корреляционной матрицы относят: симметричность относительно главной диагонали, $\rho_{jk}=\rho_{kj}$, $i,k=\overline{1,m}$; единичные значения элементов главной диагонали, $\rho_{kk}=1$ (ρ_{kk} соответствует дисперсии стандартизованного параметра u_k), $k=\overline{1,m}$.

Требуется оценить значимость выборочной величины коэффициента или,

в соответствии с постановкой задач проверки статистических гипотез, проверить гипотезу о равенстве нулю коэффициента корреляции. Если гипотеза H_0 о равенстве нулю коэффициента корреляции будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением. Если гипотеза H_0 будет принята, то оценка коэффициента не значима, и величины линейно не связаны друг с другом. Проверка гипотезы о значимости оценки коэффициента корреляции требует знания распределения этой случайной величины. Распределение величины ρ_{ik} изучено только для частного случая, когда случайные величины U_j и U_k распределены по нормальному закону.

В качестве критерия проверки нулевой гипотезы H_0 применяют случайную величину [2]

$$t = |\rho_{ik}| \sqrt{n-2} / \sqrt{1-\rho_{ik}^2}.$$

Если модуль коэффициента корреляции относительно далек от единицы, то величина t при справедливости нулевой гипотезы распределена по закону Стьюдента с $n - 2$ степенями свободы. Конкурирующая гипотеза H_1 соответствует утверждению, что значение ρ_{ik} не равно нулю (больше или меньше нуля). Поэтому критическая область двусторонняя.

Проверка гипотезы H_0 о равенстве нулю генерального коэффициента парной корреляции двумерной нормально распределенной случайной величины осуществляется в следующей последовательности:

вычисляется значение статистики t;

при уровне значимости α для двусторонней области определяется критическая точка распределения Стьюдента $t_{кр}(n-2; \alpha)$;

сравнивается значение статистики t с критическим значением $t_{кр}(n-2; \alpha)$. Если $t < t_{кр}(n-2; \alpha)$, то нет оснований отвергнуть нулевую гипотезу, иначе гипотеза

H_0 отвергается (коэффициент корреляции значим).

Для статистической обработки в MATLAB имеются две основные функции для вычисления ковариации и коэффициентов корреляции:

- **cov** – В случае вектора данных эта функция выдает дисперсию, то есть меру распределения (отклонения) наблюдаемой переменной от ее среднего значения. В случае матриц это также мера линейной зависимости между отдельными переменными, определяемая недиагональными элементами.
- **corrcoef** – Коэффициенты корреляции – нормализованная мера линейной вероятностной зависимости между переменными.

2. Регрессионный анализ

Одной из типовых задач обработки многомерных ЭД является определение количественной зависимости показателей качества объекта от значений его параметров и характеристик внешней среды.

Постановка задачи регрессионного анализа формулируется следующим образом. Имеется совокупность результатов наблюдений (матрица X). В этой совокупности один столбец соответствует показателю, для которого необходимо установить функциональную зависимость с параметрами объекта и среды, представленными остальными столбцами. Будем обозначать показатель через y^* и считать, что ему соответствует первый столбец матрицы наблюдений. Остальные m–1 ($m > 1$) столбцов соответствуют параметрам (факторам) $x_2, x_3, \ldots, x_m$.

Требуется: установить количественную взаимосвязь между показателем и факторами. В таком случае задача регрессионного анализа понимается как задача выявления такой функциональной зависимости $y^* = f(x_2, x_3, \ldots, x_m)$, которая наилучшим образом описывает имеющиеся экспериментальные данные.

Решение задачи регрессионного анализа целесообразно разбить на несколько этапов[2]:

предварительная обработка ЭД;

выбор вида уравнений регрессии;

вычисление коэффициентов уравнения регрессии;

проверка адекватности построенной функции результатам наблюдений.

Предварительная обработка включает стандартизацию матрицы ЭД, расчет коэффициентов корреляции, проверку их значимости и исключение из рассмотрения незначимых параметров (эти преобразования были рассмотрены в рамках корреляционного анализа). В результате преобразований будут получены стандартизованная матрица наблюдений $\mathbb{U}$ (через y будем обозначать стандартизованную величину y°) и корреляционная матрица ρ.

Задача определения функциональной зависимости, наилучшим образом описывающей ЭД, связана с преодолением ряда принципиальных трудностей. В общем случае для стандартизованных данных функциональную зависимость показателя от параметров можно представить в виде

$$y = f(u_1, u_2, \ldots, u_p) + \varepsilon ,$$

где f – заранее неизвестная функция, подлежащая определению; ε - ошибка аппроксимации ЭД.

Указанное уравнение принято называть выборочным уравнением регрессии y на u. Функция f должна подбираться так, чтобы ошибка ε в некотором смысле была минимальна.

На практике широко применяется полином первой степени или уравнение линейной регрессии

$$y = a_0 + \sum_{j=2}^{m} a_j u_j + \varepsilon$$

Применяя метод наименьших квадратов применительно к линейной регрессии стандартизованных величин, находим

$$\rho_{y,k} - \sum_{j=2}^{m} a_j \rho_{j,k} = 0, \quad k = \overline{2,m}$$

Итак, получено $m-1$ линейных уравнений, что позволяет однозначно вычислить значения $a_2, a_3, \ldots, a_m$.

Когда имеется только один параметр, уравнение линейной регрессии для исходных величин примет вид

$$\hat{y} = x_1 = \mu_1(x_1) - \rho_{y,2}\mu_1(x_2)\frac{\sigma(x_1)}{\sigma(x_2)} + \rho_{y,2}\frac{\sigma(x_1)}{\sigma(x_2)}x_2$$

Качество полученного уравнения регрессии оценивают по степени близости между результатами наблюдений за показателем и предсказанными по уравнению регрессии значениями в заданных точках пространства параметров. Если результаты близки, то задачу регрессионного анализа можно считать решенной. В противном случае следует изменить уравнение регрессии (выбрать другую степень полинома или вообще другой тип уравнения) и повторить расчеты по оценке параметров.

Пример. Пусть в результате эксперимента получена матрица экспериментальных данных, соответствующая таблице 5.1.

Таблица 5.1

Матрица экспериментальных данных

№	X1	X2	X3	X4	X5
1	26.3	16.5	17.6	41.9	22.8
2	28.0	15.4	17.1	43.8	23.2
3	27.8	17.5	15.3	42.8	24.5
4	31.6	16.9	18.3	47.2	26.5
5	23.5	15.6	18.3	38.7	26.2
6	21.4	17.0	17.8	35.1	27.5

7	21.3	17.3	17.9	35.1	27.8
8	16.9	16.7	21.4	32.7	25.6
9	18.8	15.2	17.3	32.8	23.4
10	25.7	15.2	30.4	40.5	25.7

Основным параметром является Y=X1. Необходимо определить оценки коэффициентов корреляции и оценить их значимость для следующих пар параметров: (Y,X2) , (Y,X3), (Y,X4) и (Y,X5).

Для определения оценок коэффициентов корреляции и значимости заданных пар параметров можно воспользоваться следующей программой:

```
    ExpData=[26.3 16.5 17.6 41.9 22.8;
    28.0 15.4 17.1 43.8 23.2;
    27.8 17.5 15.3 42.8 24.5;
    31.6 16.9 18.3 47.2 26.5;
    23.5 15.6 18.3 38.7 26.2;
    21.4 17.0 17.8 35.1 27.5;
    21.3 17.3 17.9 35.1 27.8;
    16.9 16.7 21.4 32.7 25.6;
    18.8 15.2 17.3 32.8 23.4;
    25.7 15.2 30.4 40.5 25.7];
[n,m]=size(ExpData);
r=corrcoef(ExpData);
p=r(1,2:m);
t_stat=abs(p).*sqrt(n-2)./sqrt(1-p.^2);
disp('-----------------------------------');
for I=2:m
  fprintf('   X%g  ', I);
end
clear I;
disp(' ');
disp(' -----------------------------------');
disp('   Коэффициенты корреляции        ');
disp(p);
disp('   Статистика критерия Стьюдента    ');
disp(t_stat);
disp(' ');
t_crit=tinv(0.99,n-2);
fprintf('Критическое значение критерия Стьюдента равно %g,следовательно\n',t_crit);
for I=1:m-1
  if t_stat(I)>t_crit
    index=I+1;
    break
  end
end
fprintf('оценка значима только для коэффициентов корреляции
p(1,%d)=%1.4g\n\n\n',index,p(index-1))
```

Результаты программы:

```
-------------------------------------
   X2        X3        X4        X5
-------------------------------------
  Коэффициенты корреляции
  0.0756   -0.0733    0.9888   -0.1720

  Статистика критерия Стьюдента
  0.2144    0.2079   18.7665    0.4938
```

Критическое значение критерия Стьюдента равно 2.89646, следовательно, оценка значима только для коэффициентов корреляции p(1,4)=0.9888 (для параметра x_4). Остальные коэффициенты следует признать равными нулю.

Для основного показателя $y = x_1$ необходимо установить функциональную зависимость с параметром x_4

$$y = a_0 + a_1 \cdot x_4$$

Для определения функциональной зависимости между двумя параметрами можно воспользоваться следующей программой:

```
a=polyfit(ExpData(:,index),ExpData(:,1),1);
rExpData=a(2)+ExpData(:,index).*a(1);
error=ExpData(:,1)-rExpData;
RegTabl(:,1)=ExpData(:,index);
RegTabl(:,2)=ExpData(:,1);
RegTabl(:,3)=rExpData;
RegTabl(:,4)=error;
disp(' Результаты регрессионного анализа ');
disp('-------------------------------');
fprintf(' X%d      X1     X1*      e\n',index);
disp('-------------------------------');
disp(RegTabl);
x=RegTabl(:,1);
y1=RegTabl(:,2);
y2=RegTabl(:,3);
plot(x,y1,'bo',x,y2,'r')
```

В ходе работы указанной программы получены следующие результаты:

$$y = -11{,}2518 + 0.9058 \cdot x_4$$

Результаты регрессионного анализа

X4	X1	X1*	e
41.9000	26.3000	26.7026	-0.4026
43.8000	28.0000	28.4236	-0.4236
42.8000	27.8000	27.5178	0.2822
47.2000	31.6000	31.5035	0.0965
38.7000	23.5000	23.8039	-0.3039
35.1000	21.4000	20.5429	0.8571
35.1000	21.3000	20.5429	0.7571
32.7000	16.9000	18.3689	-1.4689
32.8000	18.8000	18.4595	0.3405
40.5000	25.7000	25.4344	0.2656

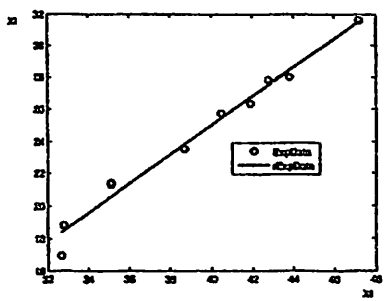

Рис.5.1. Функциональная зависимость y от x_4.

Содержание отчета

Отчет должен содержать :

- теоретическое введение;
- номер варианта;
- исходные экспериментальные данные , заданные преподователем;
- тексты программ для корреляционного и регрессионого анализа;
- результаты программы;

- выводы о проделанной работе.

Контрольные вопросы

1.Сущности корреляционного и регрессионого анализа экспериментальных данных.

2. Различие между коэффициентами ковариации и корреляции.

3. Этапы регрессионного анализа.

4. Проверка значимости коэффициента корреляции.

5. Функции MATLAB для корреляционого и регрессионного анализа.

СПИСОК ЛИТЕРАТУРЫ

1. Ходасевич Г.Б. Обработка экспериментальных данных на ЭВМ. Часть 1. Обработка одномерных данных. 220200: Учеб.пособие/ СПбГУТ. –СПб, 2002.

2. Ходасевич Г.Б. Обработка экспериментальных данных на ЭВМ. Часть 2. Обработка многомерных данных. 220200: Учеб.пособие/ СПбГУТ. –СПб, 2002.

3. Гмурман В.Е. Теория вероятностей и математическая статистика.- М.: Высшая школа, 1999.

4. Фирсов И.П., Никитина А.В., Бутенков С.А. Методические указания к практическим занятиям по математической статистике с применением ЭВМ.- http://www.exponenta/ru.

5. Дьяконов В.П. MATLAB 6/6.1/6.5+Simulink 4/5/. Основы применения. М.: Солон-Пресс, 2004.

ОГЛАВЛЕНИЕ

«Обработка экспериментальных данных»
Методическое пособие к практическим занятиям для магистрантов

Рассмотрено и рекомендовано
к изданию учебно-методическим советом
факультета телекоммуникационные технологии ТУИТ.
Протокол №8 от 9 апреля 2013 г.

Составитель Амирсандов У.Б.

Ответственный редактор Джаббаров Ш.Ю.

Корректор Хамдам-Зода Л.Х.

Формат 60x84 1/16
Заказ № - 112. Тираж - 50

Отпечатано в Издательско полиграфическом
центре «ALOQACHI» при ТУИТ
Ташкент ул. Амир Темура, 108