Практическая работа 17.
Статистическая обработка массива данных

Цель работы: приобретение навыков статистической обработки массива данных в среде MatLab.

Теоретические сведения

Статистическая обработка массива данных

Пусть исследуется случайная величина X с функцией распределения F(X). Пусть имеем выборку объема *n*: $x_0, x_1, \ldots, x_n$. Эмпирическая функция распределения имеет вид:

$$F^*(x) = \frac{1}{N}\sum_{i=1}^{N}\theta(x - x_i),$$

где x – случайная переменная, x_i – величина случайной переменной в i-м наблюдении или опыте, N – полное число реализаций, θ(ξ) – функция Хевисайда:

$$\theta(\xi) = \frac{1}{2}\left(\frac{|\xi|}{\xi} + 1\right) = \begin{cases} 1, & \xi > 0 \\ 0, & \xi < 0 \end{cases}.$$

Функция Хевисайда представляет собой ступеньку единичной высоты, расположенную в точке с ординатой $x=a$ оси абсцисс. Будем считать, что выборку упорядочили: $x_1 \le x_2 \le \ldots \le x_n$. Для сортированного по возрастанию ряда значений данных эмпирическую функцию распределения можно построить, нормируя высоту ступенек

$$F^*(x_i) = \frac{i}{N},$$

где число i пробегает ряд значений $i = 1, \ldots, N$.

В среде Matlab для построения графика эмпирического распределения используют функцию – *cdfplot(Z)*.

В среде Matlab для построения точечного графика с визуализацией ординат значения функции, где *x* – одномерный массив значений абсцисс, представляющих номера наблюдаемой или измеряемой в опыте величины или значения входных данных, *y* – одномерный массив значений ординат, представляющих значения выходных данных или реализаций случайной переменной можно использовать функцию *stem(x,y)*.

Для представления векторных и матричных данных используются диаграммы и гистограммы. Значение элемента вектора пропорционально высоте столбика диаграммы или площади сектора диаграммы. Гистограммы используются для получения информации о распределении данных по заданным интервалам. Отображение вектора в виде столбчатой диаграммы осуществляется функцией *bar(x,y,h)*. Разметку горизонтальной оси можно задать вектором с возрастающими значениями, что

учитывается в первом аргументе функции *bar*. Выбор ширины столбцов осуществляется заданием третьего дополнительного аргумента.

В математической статистике исходная исследуемая величина называется генеральной совокупностью, а полученный из нее набор экспериментальных данных – выборочной совокупностью или выборкой. Под выборочным методом исследования понимают исследование выборки и перенесение его результатов на генеральную совокупность.

Гистограммой распределения данных называют график частот или относительных частот случайной переменной x, в зависимости от интервалов ее значений. Для построения гистограммы диапазон наблюдаемых величин x разбивают на k интервалов (бинов) Δx :

$$\Delta x = \frac{x_{\max} - x_{\min}}{k},$$

где x_{max} и x_{min} – максимальная и минимальная величина переменной x. Далее вычисляют частоту n_i значений переменной x, попавших в i-й интервал, и откладывают либо саму величину n_i или плотность относительной частоты вдоль оси ординат:

$$f_i = \frac{n_i}{N\Delta x}$$

Аналитическая форма такой зависимости имеет вид:

$$f_i = \frac{n_i}{N\Delta x}\left[\theta\left(x - x_i + \frac{\Delta x}{2}\right) - \theta\left(x - x_i - \frac{\Delta x}{2}\right)\right],$$

где x_i – середина выбранного интервала. В результате оценок для всех k интервалов получается гистограмма, для которой выполняется условие нормировки:

$$\sum_{i=1}^{k}\frac{n_i}{N} = \sum_{i=1}^{k} f_i \Delta x = 1.$$

Для построения гистограммы в Matlab и вычисления ее параметров можно использовать функции:

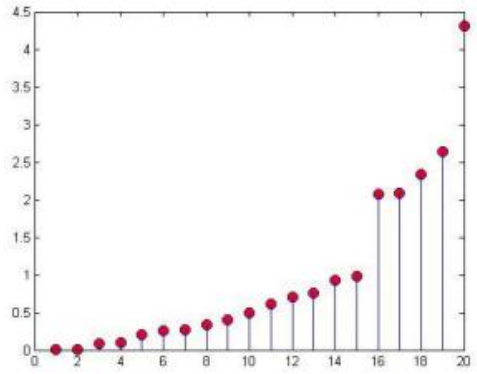
– *hist(data, Nbins)* – строит гистограмму данных, записанных в вектор *data*, по числу бинов *nbins* или по вектору середин бинов (интервалов);

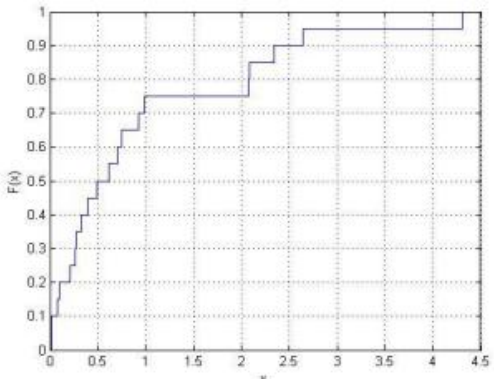
– *[n,xout]=hist(data,nbins)* – вычисляет вектор частот и вектор середин бинов (интервалов).

Например, упорядочим выборку и построим эмпирическую функцию распределения:

```
data=[0.2083 0.7519 2.6392 0.9831 0.7058 0.0131 0.9305 2.3412 0.4889 0.10];
N=length(data);
i=1:N;
st=sort(data);
figure
stem(i,st)
figure
cdfplot(data)
```

Получаем графики:





Для построения гистограммы распределения данных для количества бинов: k=5, 10, 15, 20 можно использовать функции *subplot* и *hist*:

```
j=0
for nbin=5:5:20
    j=j+1;
    subplot(2,2,j)
    hist(data,nbin);
end
```

Генеральными параметрами называют числовые параметры генеральной совокупности. Выборочными параметрами называют числовые параметры выборки: m_x^* – выборочное математическое ожидание, σ_x^* – выборочное среднеквадратичное отклонение. Выборочные параметры называют оценками соответствующих генеральных параметров.

Если случайная величина X – дискретная, n-значная и принимает значения $x_1, x_2, \ldots, x_n$ с одинаковыми вероятностями $p_i=1/n$, то математическое ожидание:

$$m_X^* = \sum_{i=1}^{n} x_i \cdot p_i = \frac{1}{n}\sum_{i=1}^{n} x_i ,$$

то есть вычисляется как среднее арифметическое (для вызова функции Matlab, например, можно указать - *mean(data)*). Дисперсия (для вызова функции Matlab, например, можно указать - *var(data)*):

$$D_X^* = \sum_{i=1}^{n} x_i \cdot p_i = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^2 .$$

Число в знаменателе выборочной дисперсии $f = n - 1$ называется числом степеней свободы выборки из n элементов. Среднеквадратичное отклонение (для вызова функции Matlab, например, можно указать - *std(data)*):

$$\sigma_x^* = \sqrt{D_x^*} .$$

Асимметрия (для вызова функции Matlab, например, можно указать - *skewness(data)*):

$$a_X^* = \frac{\sqrt{n}}{\sqrt{(n-1)^3}\cdot\left(\sigma_x^*\right)^3}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^3 ,$$

эксцесс (для вызова функции Matlab, например, можно указать - *kurtosis(data)*):

$$e_X^* = \frac{n}{(n-1)^3\cdot\left(\sigma_x^*\right)^4}\sum_{i=1}^{n}\left(x_i - m_x^*\right)^4 - 3 .$$

Медиана выборочного распределения равна среднему элементу упорядоченной выборки (в выборке нечетное число элементов) или полусумме средних элементов (если в выборке четное число элементов) - для вызова функции Matlab, например, можно указать - *median(data)*.

По виду гистограммы можно подобрать теоретический закон распределения. Наиболее часто встречаются в приложениях следующие законы распределений: нормальное распределение, показательное распределение, равномерное распределение, рэлеевское распределение:

```
x=-1:0.01:5;
ypdf1=pdf('norm',x,2,1);
figure
plot(x,ypdf1)
ypdf2=pdf('exp',x,2,0);
figure
plot(x,ypdf2)
ypdf3=pdf('unif',x,0,4);
figure
plot(x,ypdf3)
ypdf4=pdf('rayl',x,1,0);
figure
plot(x,ypdf4)
```

Нормальное распределение является двухпараметрическим, для его задания требуется инициализировать параметры m и σ. В Matlab это распределение реализуется, например, с помощью встроенной функции *normpdf* с параметрами m и σ.

Показательное распределение является однопараметрическим (параметр α). Следует отметить, что плотность показательного распределения отлична от нуля только для неотрицательных значений x. В нуле она принимает максимальное значение, равное α. В Matlab это распределение реализуется, например, с помощью встроенной функции *exppdf c* параметром – величиной, обратной к α.

Равномерное распределение – двухпараметрическое (параметры a и b). Плотность равномерного распределения отлична от нуля только в заданном интервале $[a, b]$, и принимает в этом интервале постоянное значение. В Matlab это распределение реализуется, например, с помощью встроенной функции *unifpdf* с параметрами a и b.

Рэлеевское распределение является однопараметрическим. В Matlab это распределение реализуется, например, с помощью встроенной функции *raylpdf* с параметром – величиной σ.

Параметры теоретических распределений:

– для нормального распределения:

$$m = m_x^*; \sigma = \sigma_x^*;$$

– для показательного распределения:

$$\alpha = \frac{1}{m_x^*};$$

– для равномерного распределения:

$$a = m_x^* - \sigma_x^*\sqrt{3}; b = m_x^* + \sigma_x^*\sqrt{3};$$

– для рэлеевского распределения:

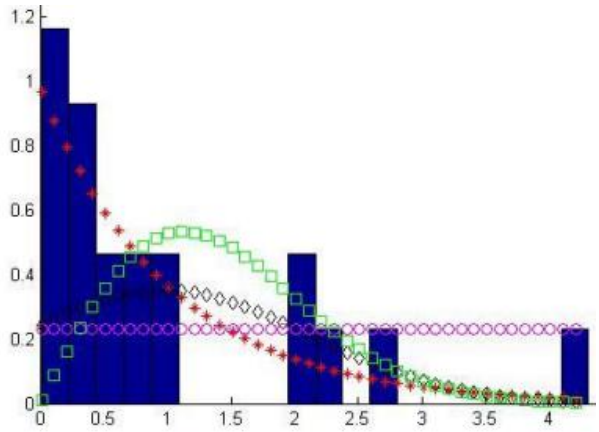
$$\sigma = m_x^*\sqrt{\frac{2}{\pi}}.$$

Ниже приведена последовательность действий для построения на одном графике теоретической и эмпирической плотности распределения (для выборки $N=20$):

```
datamin=st(1);
datamax=st(N);
Mdata=mean(data);
Ddata=var(data);
Sdata=std(data);
```

```
[nn,xx]=hist(data,20);
d=(datamax-datamin)/20;
femp=nn/(N*d);
figure
bar(xx,femp,1)
xt=datamin:0.1:datamax;
alfa=abs(1/Mdata);
ftheorexp=exppdf(xt,alfa);
ftheornorm=normpdf(xt,Mdata,Sdata);
ftheorunif=unifpdf(xt,datamin,datamax);
ftheorrayl=raylpdf(xt,Sdata);
hold on
plot(xt,ftheorexp, xt,ftheornorm, xt,ftheorunif, xt,ftheorrayl)
```

Ниже приведен результат построения:



На основе визуального анализа графического представления эмпирического и теоретических распределений сделать предположение о наиболее подходящем распределении.

На практике могут встретиться и другие виды распределений (β, χ 2 , Вейбулла и другие). Многие из них реализованы в Matlab.

Критерием согласия называется критерий проверки правильности подбора теоретического распределения на его соответствие выборке. Правильность выбора теоретического распределения можно проверить, используя критерий согласия Колмогорова.

Критерий согласия Колмогорова. Максимальная по модулю разность между выборочной и генеральной функциями распределения

$$D = \max_{\forall x \in R} \left| F^*(x) - F(x) \right|$$

является случайной величиной. Эта величина с ростом объема выборки сходится по вероятности к нулю. Колмогоров уточнил этот результат, выяснив, как именно D сходится к нулю. Он рассмотрел случайную величину

$$\Lambda = D\sqrt{n}$$

и нашел ее закон распределения.

Теорема (Колмогорова). Для любого непрерывного закона распределения генеральной совокупности X функция распределения случайной величины

$$\Lambda = D\sqrt{n}$$

при достаточно большом n имеет вид:

$$F(\lambda) = \sum_{k=-\infty}^{+\infty} (-1)^k \exp(-2k^2 \cdot \lambda^2)$$

– распределение Колмогорова. Эта теорема дает возможность проверить правильность подбора теоретического распределения. Величина $q = 1 - p$ – называется уровнем значимости, а p – доверительная вероятность.

В MatLab для проверки критерия согласия Колмогорова применяется функция *kstest*. При этом выбирается такое распределение, для которого q принимает *максимальное* значение (вероятность p разницы между функциями распределения минимальная). Величина $q = 1 - p$ – называется уровнем значимости, при этом p – доверительная вероятность.

Событие A называется абсолютно достоверным, если $P(A)=1$. Событие А называется практически достоверным, если оно практически всегда происходит на практике ($P(A) \approx 1$). Статистической гипотезой называется любое предположение о законе распределения генеральной совокупности или его параметрах. Области, не попадающие в доверительный интервал, называются критическими областями статистической гипотезы, а границы доверительного интервала – критическими числами гипотезы.

График выборочной функции распределения строит функция *cdfplot*. Ниже приведена последовательность действий для проверки критерия Колмогорова-Смирнова для нормального распределения:

```
CDFnorm=[];
CDFnorm(:,1)=data;
CDFnorm(:,2)=cdf('norm',data,Mdata,Sdata);
[hKolm1,pnorm,kKolm1,cKolm1]=kstest(data,CDFnorm,0.05,0);
disp(pnorm)
```

Переменная *pnorm* показывает критический уровень значимости для нормального распределения. Для показательного распределения последовательность действий аналогична:

```
CDFexp=[];
CDFexp(:,1)=data;
CDFexp(:,2)=cdf('exp',data,Sdata);
[hKolm2,pexp,kKolm2,cKolm2]=kstest(data,CDFexp,0.05,0);
disp(pexp)
```

Переменная *pexp* показывает критический уровень значимости для показательного распределения. Выбирая максимальное значение, можно сделать вывод, какое распределение является наиболее подходящим. В рассматриваемом примере наиболее подходящим является показательное распределение. Ниже на графике выполнена визуализация выборочной (эмпирической) и теоретической функции распределения: