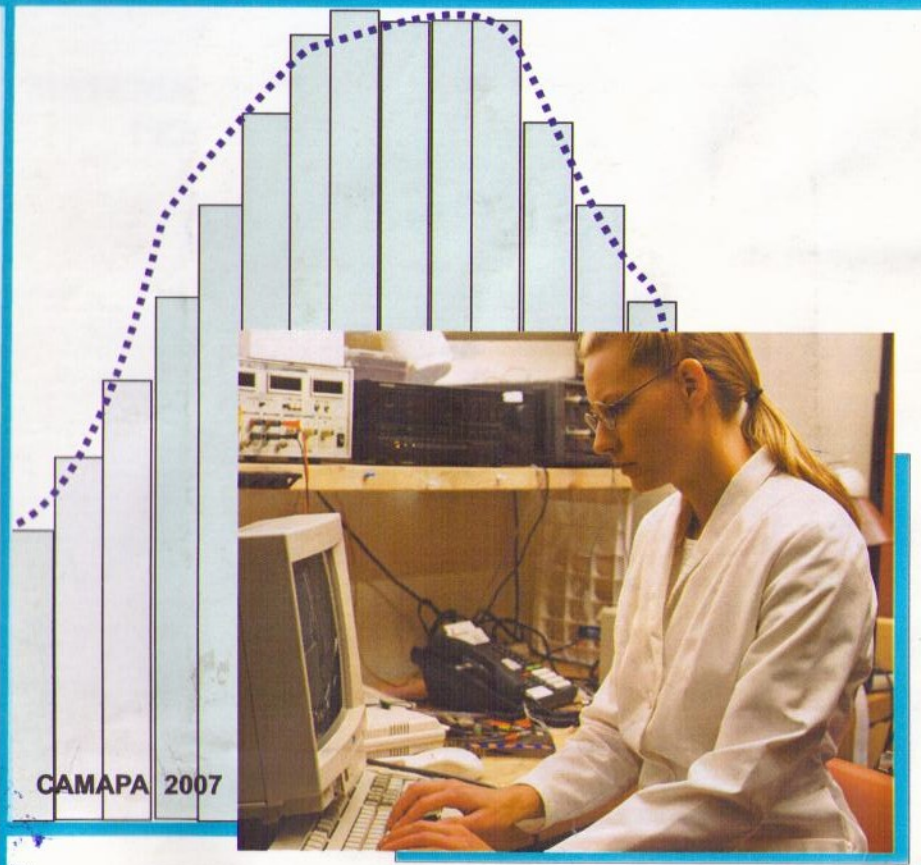
Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
"Самарский государственный
архитектурно-строительный университет"

МАТЕМАТИЧЕСКИЕ МЕТОДЫ ОБРАБОТКИ РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА



САМАРА 2007

Элементы математической статистики

Введение

Математическая статистика - раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Предметом математической статистики является изучение случайных величин (или случайных событий, процессов) по результатам наблюдений. Полученные результаты опыта, эксперимента сначала обрабатываются, упорядочиваются. Это первая задача. Затем, это уже вторая задача, оцениваются, хотя бы приблизительно, интересующие характеристики наблюдаемой случайной величины. Следующей, третьей задачей, является проверка статистических гипотез, т.е. решение вопроса согласования результатов оценивания с опытными данными.

Для обработки статистических данных созданы специальные программные пакеты (STADIA, SYATAT, STST-GRAPHICS и др.), которые выполняют трудоемкую работу по расчету различных статистик, построению таблиц и графиков.

Говорят, что математическая статистика - это теория принятия «решений в условиях неопределенности».

Курсовая работа «Анализ и обработка статистических данных» состоит из введения, двух частей и заключения. В первой части работы (I - II) описываются два алгоритма, которые применяются при обработке заданного статистического материала. Во второй части (III - IV) эти алгоритмы применяются к решению двух задач математической статистики.

В первой задаче по данным выборок x_1, x_2, x_3 объема n = 100 нужно определить законы распределения случайных величин x_1, x_2, x_3 в виде аналитических функций.

Во второй задаче по выборкам x_i, z объема n = 100 требуется построить линейную регрессию z на x. Здесь x_i - та случайная величина, которая в первой задаче распределена нормально.

Основными методами исследования в работе являются метод разрядов, метод моментов, критерий согласия Пирсона, критерии коррелированности двух случайных величин (x, z), критерий линейной зависимости случайных величин x и z и метод наименьших квадратов.

Конечные результаты, полученные нами обработкой статистического материала выборок x_1, x_2, x_3, z, сведены в таблицу 1.

В работе приводится перечень используемой литературы.

Таблица 1

X_1	X_2	X_3
Найденные законы распределения. Плотность распределения.		
Нормальная $F(x)=, m=, \sigma_x=$	Равномерная $F(x)=, a=, b=$	Экспоненциальная $f(x)=, \lambda=$
Уравнения линейной регрессии z на x		
$z-\bar{z}=S_z/S_x \cdot rxz(x-\bar{x})$	$\bar{x}=$ $z=$ $S_z=$ $rxz=$	$S_x=$

Теория

I. Исследование статистических данных одной случайной величины. Описание первого алгоритма.

Постановка задачи 1

Задача1. Провести анализ и обработку статистического материала выборок X_1, X_2, X_3 по следующему алгоритму

XN → SRN → ГN → HN → Tf/(x,a,b)

→ ПН(a,b) → Sf /(x,a*,b*) → CS,(1)

SRN - составленный для нее статистический ряд;

ГN - гистограмма по полученному ряду [1, с.137];

HN - выдвинутая гипотеза распределения [3];

Tf(x,a,b) - функция плотности вероятности;

a,b - параметры гипотезы [1, 3];

ПН (a,b) - оценка (нахождение) параметров a,b. Оценки должны быть несмещенными, состоятельными и эффективными.

Sf(x,a*,b*) - статистическая функция распределения, a*, b*- найденные параметры распределения по методу моментов;

CS - критерий согласия.

Краткая экспозиция алгоритма (1):

Обработать статические данные выборки XK с помощью нормального закона распределения.

Дано: XN =

1	X_1
2	X_2
3	X_3
4	X_4
5	X_5
.	.
.	.
.	.
n	X_n

Найти: Будет ли случайная величина X_n распределена нормально или она подчиняется другому закону распределения, т.е. определить функцию f(x) плотности вероятности.

1°. Оформим выборку XN в виде статистического ряда SRN , т.е. от одной таблицы перейдем к другой более компактной.

Для этого применим метод разрядов.

Находим x_{min}, x_{max}. Разобьем интервал $[x_{min}, x_{max}]$ на разряды (интервалы).

Шаг разбиения вычисляем по формуле

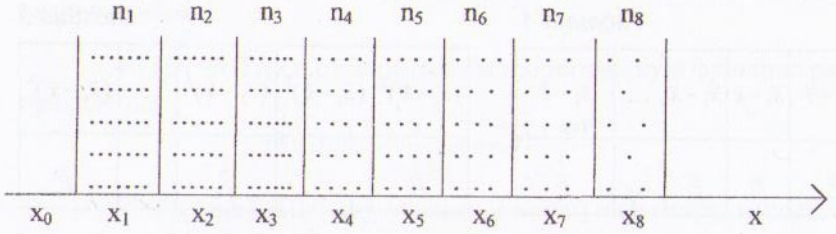
$$h = \frac{-x_{min} + x_{max}}{1 + 3.332 \cdot \lg n},$$

за x_0 выбираем число

$$x_0 = x_{min} - \frac{h}{2}.$$

Тогда $x_1 = x_0 + h,\ x_2 = x_1 + 2h,\ \ldots,\ x_k = x_0 + kh > x_{max}$.

Разносим числа выборки по разрядам:



В каждом интервале считаем количество элементов выборки. n_i - абсолютная частота попадания случайной величины в разряд $[x_{i-1}, x_i]$.

Далее считаем относительные частоты (статистические вероятности) попадания случайной величины в разряды по формуле

$$\tilde{P}\, i = \frac{n_i}{n}.$$

Заполняем таблицу 2-3:

Таблица 2

SR2 =

(x_{i-1}, x_i)	(x_0, x_1)	(x_1, x_2)	...	x_{k-1}, x_k
n_i	n_1	n_2	...	n_k
$\widetilde{P}i = n_i/n$	$\widetilde{P}_1$	$\widetilde{P}_2$	...	$\widetilde{P}k$
$\widetilde{P}i/h$				

Таблица 3

SR3 =

$\widetilde{x}_i = (x_{i-1}+x_i)/2$	$\widetilde{x}i$	$\widetilde{x}i$	...	$\widetilde{x}i$
$\widetilde{P}i = n_i/n$	$\widetilde{P}i$	$\widetilde{P}i$	...	$\widetilde{P}i$

Пользуясь таблицей 3, вычисляем статистическое среднее по формуле:

$$M(x) = m = \bar{x} = \sum_{i=1}^{k} \widetilde{x}_i \widetilde{P}_i \ .$$

Для вычисления статистической дисперсии и стандарта случайной величины составляем таблицу 4 и таблицу 5.

Таблица 4

$\widetilde{x}_i - \bar{x}$	$\widetilde{x}_1 - \bar{x}$	$\widetilde{x}_2 - \bar{x}$	...	$\widetilde{x}_\kappa - \bar{x}$
$\widetilde{P}_i$	$\widetilde{P}_1$	$\widetilde{P}_2$	...	$\widetilde{P}_k$

Таблица 5

$(\widetilde{x}_i - \bar{x})^2$	$(\widetilde{x}_1 - \bar{x})^2$	$(\widetilde{x}_2 - \bar{x})^2$	...	$(\widetilde{x}_\kappa - \bar{x})^2$
$\widetilde{P}_i$	$\widetilde{P}_1$	$\widetilde{P}_2$	...	$\widetilde{P}_k$

Эти точные оценки находим по формуле [6, с.194]:

$$D(x) = \sigma^2 = S^2 = n/(n-1)\sum_{i=1}^{k} (\widetilde{x}_i - \bar{x})^2 \cdot \ P_i \ ;$$

$$h_i = \widetilde{P}_i / h \ .$$

2°. Множество таких прямоугольников показывается гистограммой статического распределения случайной величины.

А. Предположим, что гистограмма имеет вид, изображенный на рисунке 1:

$$h_i = \widetilde{P}_i / h$$

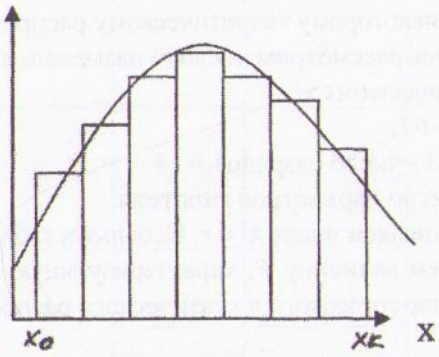


Рисунок 1

3°. ГN→ HN. По виду гистограммы выдвигается гипотеза о распределении случайной величины (в нашем случае - о нормальном распределении).

4°. HN→ Tf(x,a,b). Записываем теоретическую функцию распределения

$$f(x,a,b) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-m)^2}{2\sigma^2}} \ .$$

5°. Tf(x,a,b) >ПН(a,b). Оценить (найти) параметры гипотезы. Здесь $a=m$; $b=y$; $m=\bar{x}$; $y=s$.

Применим метод моментов. Сравниваем соответствующие статические характеристики с теоретическими.

6°. ПН(a,b) >Sf(x,a^A , b^A) Записываем статическую функцию распределения:

$$f(x) = \frac{1}{s\sqrt{2\pi}} e^{\frac{-(x-\bar{x})^2}{2s^2}} \ ; \ f(x) \approx \widetilde{f}(x)$$

7°. Sf(x,a^A ,b^A) > CS. Применяем критерий Пирсона. Введем меру расхождения x^2 между статическим и теоретическим законами распределения:

$$\chi^2 = x^2 = n\sum_{i=1}^{k}, \frac{(\tilde{P}_i - P_i)^2}{P_i}$$

где n - объем выборки; $\tilde{P}_i$ - статические вероятности, полученные по некоторому теоретическому распределению.

Далее рассмотрим число r, называемое числом степеней свободы распределения x^2:

$r=k-t-1$,

где k - число разрядов,

t- число параметров гипотезы.

Вычисляем числа x^2 и r. Используя таблицу по значениям x^2 и r, определяем величину P, характеризующую вероятность согласованности теоретического и статического распределений [4, с.74], [5, с. 413].

Если $P > 0{,}1$, то гипотеза не противоречит опытным данным. Если $P < 0{,}1$, то это означает, что теория плохо воспроизводит эксперимент.

В формуле для χ^2 в случае гипотезы нормального распределения P_i рассчитывается по формуле

$$P_i(x_{i-1} < x < x_i) = P_i = \overline{\Phi}(\frac{x_i - \overline{x}}{S}) - \overline{\Phi}(\frac{x_{i-1} - \overline{x}}{S}), i = \overline{1,k},$$

где $\overline{\Phi} = \frac{1}{\sqrt{2\pi}}\int_0^x e^{\frac{-t^2}{2}} dt$ - номерованная формула Лапласа [5, с. 411].

Для вычисления χ^2 рекомендуется составить таблицу.

Таблица 6

$x_к$	$\frac{x_k - m}{S}$	$\overline{\Phi}\left(\frac{x_k - m}{S}\right)$	$P_i = \overline{\Phi}_i - \overline{\Phi}_{i-1}$	n_i	$n_i' = nP_i$
x_0	$\frac{x_0 - m}{S}$	$\overline{\Phi}_0$	$\overline{\Phi}_1 - \overline{\Phi}_0$	n_1	$n(\overline{\Phi}_1 - \overline{\Phi}_0)$
x_1	$\frac{x_1 - m}{S}$	$\overline{\Phi}_1$	$\overline{\Phi}_2 - \overline{\Phi}_1$	n_2	$n(\overline{\Phi}_2 - \overline{\Phi}_1)$
x_2	$\frac{x_2 - m}{S}$	$\overline{\Phi}_2$	$\overline{\Phi}_3 - \overline{\Phi}_2$	n_3	$n(\overline{\Phi}_3 - \overline{\Phi}_2)$
x_3	$\frac{x_3 - m}{S}$	$\overline{\Phi}_3$	...	...	...
...	...	...			
x_{k-1}	$\frac{x_{k-1} - m}{S}$	$\overline{\Phi}_{k-1}$	$\overline{\Phi}_k - \overline{\Phi}_{k-1}$	n_k	$n(\overline{\Phi}_k - \overline{\Phi}_{k-1})$
x_k	$\frac{x_k - m}{S}$	$\overline{\Phi}_k$			

(n = 100)

Далее число χ^2 вычисляем по формуле $\chi^2 = \sum_{i=1}^{k}(n_i' - n_i)^2 / n_i'$.

Критерий Пирсона можно применять в том случае, когда $n_i \geq 5$. Если же это не так, то соседние интервалы объединяются пока $n_i \geq 5$. k - число интервалов, оставшихся после объединения соседних интервалов. Здесь $\overline{\Phi}$ - нормированная интегральная функция Лапласа [5, с. 411].

Б. Пусть гистограмма имеет вид, изображенный на рисунке 2.

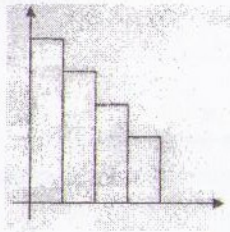


Рисунок 2

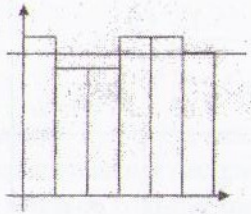


Рисунок 3

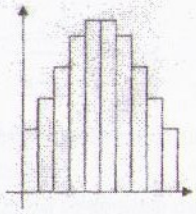


Рисунок 4

Тогда мы можем выдвинуть гипотезу об экспоненциальном распределении

$$Tf(x,\lambda)=\begin{cases}\lambda e^{-\lambda x}, x\ge 0\\ 0, x<0\end{cases}.$$

Числовые характеристики распределения будут равны

$$M=\frac{1}{\lambda};D=\frac{1}{\lambda^2}.$$

Статистические числовые характеристики найдем по формуле

$$\bar{x}=\sum_{i=1}^{k}\tilde{x}_i\tilde{P}_i, D=S^2=n/(n-1)\sum_{i=1}^{k}(\tilde{x}_i-\bar{x})^2\tilde{P}_i$$

Применяя метод моментов, найдем $\lambda=\frac{1}{\bar{x}}$.

Формула для определения теоретических вероятностей будет равна

$$P_i(x_{i-1}<x<x_i)=e^{\frac{-x_{i-1}}{\bar{x}}}-e^{\frac{-x_i}{\bar{x}}}, i=\overline{1,k}.$$

Далее применяется критерий Пирсона.

В. Пусть гистограмма имеет вид, изображенный на рисунке 3. Тогда можем выдвинуть гипотезу о равномерном распределении. Числовые характеристики теоретического распределения найдем по формуле

$$M=\frac{a+b}{2};D=\frac{(b-a)^2}{12}.$$

Применяя метод моментов, составим систему:

$$\frac{a+b}{2}=\bar{x};\frac{(b-a)^2}{12}=S^2,$$

Находим a и b. Функцию плотности вероятности определим по формуле

$$f(x)=\begin{cases}\frac{1}{b-a}, a\le x\le b\\ 0, x<a, x>b\end{cases};$$

Теоретические вероятности находим по уравнению

$P_i\cdot(x_{i-1}<x<x_i)=\frac{x_i-x_{i-1}}{b-a}=\frac{h}{b-a}$ - одно и то же число для всех разрядов.

Г. Пусть гистограмма имеет вид, изображенный на рисунке 4. Предположим, что случайная величина распределена логарифмически нормально.

Тогда $F(x,\mu,\nu)=\frac{1}{\nu_x\sqrt{2\pi}}e\frac{-(\ln x-\mu)^2}{2\nu^2}$.

Числовые характеристики теоретического распределения можно определить по формуле

$$M(x)=e^{\mu+\nu^2/2};$$

$$D(x)=e^{2\mu+\nu^2}(e^{\nu^2}-1)=M^2(e^{\nu^2}-1).$$

По методу моментов составим систему

$$e^{\mu+\nu/2}=\bar{x};e^{2\mu+\nu^2}(e^{\nu^2}-1)=S^2.$$

Решая эту систему, находим:

$$\mu=\ln\frac{\bar{x}^2}{\sqrt{\bar{x}^2+S^2}},\nu=\sqrt{\ln(\frac{\bar{x}^2+S^2}{\bar{x}^2})}.$$

Формула для определения теоретической вероятности имеет вид:

$$P_i(x_{i-1}<x<x_i)=\overline{\Phi}_i((\ln x_i-\mu)/\nu)-\overline{\Phi}((\ln x_{i-1}-\mu)/\nu).$$

II. Исследование статистических данных системы двух случайных величин. Описание второго алгоритма

Постановка задачи 2

Задача 2. Пусть *(x, z)* - система двух случайных величин, где - та случайная величина (x_1, x_2, x_3), которая распределена нормально. Определить, существует ли линейная корреляционная зависимость между этой случайной величиной и случайной величиной *z*.

Обработка статистических данных системы *(x, z)* проводится по следующему алгоритму:

$$\{\Gamma_1\Gamma_2\Gamma_3\}\to x\to SRX\left[k_x, m_x=\bar{x}, S_x=\sqrt{\widetilde{D}_{(x)}}\right]z\to$$

$$SRZ\left[k_z=k_x, m_z=\bar{z}, S_z=\sqrt{\widetilde{D}_{(z)}}\right]\to\left\|\widetilde{P}_{ij}\right\|\to Таб\ .A$$

$$\left[\widetilde{x}_i=\frac{x_{i-1}+x_i}{2}, \widetilde{z}_j=\frac{z_{j-1}+z_j}{2}\right]\to Таб\ .Б\left[\widetilde{x}_i-\bar{x}, \widetilde{z}_j-\bar{z}\right]\to$$

$$\widetilde{k}_{xz}=\frac{n}{n-1}\sum_{i-1}^{k}\sum_{j=1}^{k}(\widetilde{x}_i-\bar{x})(\widetilde{z}_j-\bar{z})\widetilde{P}_{ij}\left[аналог\ .вычисл\ .суммы\ \right]$$

$$\to\widetilde{r}_{xz}=\frac{\widetilde{k}_{xz}}{S_xS_z}\approx 1\to z-\bar{z}=\frac{S_z}{S_x}\cdot\widetilde{r}_{xz}(x-\bar{x})\,.$$

Краткая экспозиция алгоритма второго:

1°. $\{\Gamma_1, \Gamma_2, \Gamma_3\}\to x_{норм}$. Из трех случайных величин X_1, X_2, X_3 выбираем ту гипотеза о нормальном распределении которой наиболее достоверна.

2°. $x\to SRX$. Для этой случайной величины в задаче 1 был построен статистический ряд и определены числовые характеристики: k_x - число разрядов, x - статистическое среднее , S_x - статистический стандарт.

3°. $z\to SRZ$. Оформить выборку z в статистический ряд *SRZ*. Нужно положить $k_x=x_z=k$ и найти z и $S_z=\widetilde{D}_z$ по формуле

$$\widetilde{z}=\sum_{i=1}^{k}\widetilde{z}_i\cdot\widetilde{P}_i, D(z)=S^2=\sigma 2=\frac{n}{n-1}\sum_{i=1}^{k}(\widetilde{z}_i-\bar{z})^2\widetilde{P}_i\,.$$

4°. {SRX, SRZ} → ¦Pij¦. Составим закон распределения систем двух случайных величин *(x, z)* в виде таблицы с матрицей вероятностей ¦Pij¦ [SR(x,z)].Совмещая точки x_0 и z_0, построим систему координат x_0x , z_0z . В этой плоскости проводим систему вертикалей через точки x_i, i= 1,2..., k и систему параллелей $z=z_j$, j= 1,2....., k.. В результате получим k^2 двумерных разрядов (прямоугольников):

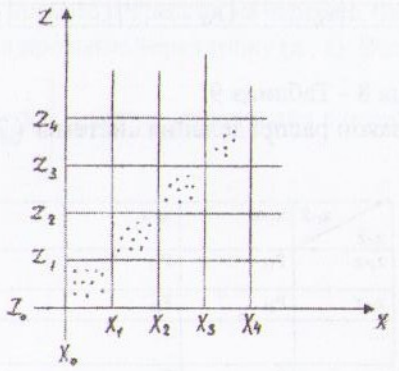


Рисунок 5

Далее разносим точки (x_i, z_j) по двумерным разрядам и подсчитываем число m_{ij}- количество точек в каждом прямоугольнике. Находим частоту попадания системы точек (x_i, z_j) в двумерный разряд с номером (i, j)

$$\widetilde{P}_{ij}=\frac{m_{ij}}{n}, n=100\,.$$

Таким образом, мы получаем матрицу вероятностей ¦ $\widetilde{P}_{ij}$ ¦.

Закон распределения системы имеет вид:

Таблица 7

SR(x,z) =

z_i \ x_i	x_1	x_2	...	x_k
z_1	P_{11}	P_{12}	...	P_{1k}
z_2	P_{21}	P_{22}	...	P_{2k}
...	...	...	...	...
z_k	P_{k1}	P_{k2}	...	P_{kk}

5°. ¦ $\widetilde{P}_{ij}$ ¦ -Таблица 8.

Запишем закон распределения системы ($\widetilde{x}_i, \widetilde{z}_j$).

Таблица 8

z_i \ x_i	x_1	x_2	…	x_k
z_1	P_{11}	P_{12}	…	P_{1k}
z_2	P_{21}	P_{22}	…	P_{2k}
…	…	…	…	…
z_k	P_{k1}	P_{k2}	…	P_{kk}

6°. Таблица 8 - Таблица 9.

Запишем закон распределения системы $(\tilde{x}_i - \bar{x}, \tilde{z}_j - \bar{z})$:

Таблица 9

$SR(\tilde{x}_i - \bar{x}, \tilde{z}_j - \bar{z}) =$

$z_i-\bar{z}$ \ $x_i-\bar{x}$	x_1-x	x_2-x	…	x_k-x
z_1-z	P_{11}	P_{12}	…	P_{1k}
z_2-z	P_{21}	P_{22}	…	P_{2k}
…	…	…	…	…
z_k-z	P_{k1}	P_{k2}	…	P_{kk}

7°. Таблица 9 - $\tilde{K}_{XZ}$

$$\tilde{k}_{xz} = \frac{n}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{k}(\tilde{x}_i - \bar{x})(\tilde{z}_j - \bar{z})\tilde{P}_{ij} .$$

По таблице 9 находим корреляционный момент связи $\tilde{K}_{XZ}$. Проверяем гипотезу коррелированности случайных величин ($k_{xz} \neq 0$).

Используем следующий алгоритм раскрытия двойной суммы: 1-й столбец таблицы 8 умножаем на 1-й столбец матрицы вероятностей $|\tilde{P}_{ij}|$. Полученную сумму умножаем на число, стоящее над первым столбцом матрицы вероятностей.

Затем 1-й столбец матрицы вероятностей и полученную сумму умножим на число, стоящее над вторым столбцом матрицы вероятностей и т.д. И, наконец, полученный результат двойной суммы умножаем на n/n-1.

Если окажется , что $\tilde{K}_{XZ} = 0$, то x и z будут некоррелированы. Если $\tilde{K}_{XZ} \neq 0$, то x и z будут зависимыми.

8°. $$\tilde{k}_{xz} \to \tilde{r}_{xz} = \frac{\tilde{k}_{xz}}{S_x S_z} \approx 1 .$$

Вычисляем коэффициент корреляции $\tilde{r}_{xz}$. Если окажется, что $|\tilde{r}_{xz}| < 1$, то это означает, что случайные величины x и z связаны линейной зависимостью.

9°. $$\tilde{r}_{xz} \approx 1 \to Z - \bar{Z} = \frac{S_Z}{S_X} r_{XZ}(X - \bar{X})$$

Построить прямую регрессии на чертеже, где изображены точки $(\tilde{X}_i, \tilde{Z}_j)$.Прямая проходит через точку (x , z). Вторую точку можно взять $\left[0, z = \frac{S_x}{S_z}\tilde{r}_{XZ}X\right]$ - точка пересечения с осью z.

Заключение

При решении различных задач, связанных со случайными явлениями, необходимо знать законы распределения фигурирующих в них случайных величин. Эти законы могут быть определены из опыта, но обычный опыт, целью которого является определение закона распределения случайной величины или системы случайных величин, оказывается и сложным, и дорогостоящим. Естественно, возникает задача свести объем эксперимента к минимуму и составлять суждение о законах распределения случайных величин косвенным образом, на основании уже известных законов распределения других случайных величин. Такие косвенные методы исследования случайных величин играют весьма большую роль в теории вероятностей. При этом обычно интересующая нас случайная величина представляется как функция других случайных величин. Зная законы распределения аргументов, можно установить закон распределения функции. В строительстве на 1-й план выдвигается проблема надежности и экономичности строительных конструкций.

Методы оценки надежности строительных конструкций отличаются от принятых подходов в оценках надежности механизмов, машин или электротехнического оборудования. Например, отсутствует этап опытно-конструкторской отработки сооружения в целом, и при использовании вероятностных методов предполагается, что законы распределения случайных факторов, учитываемых при расчете, известны заранее. Состояние конструкции в условиях эксплуатации может быть охарактеризовано конечным числом независимых переменных x. Часть этих параметров характеризует нагрузки, другие - прочность материалов, третьи - отклонение реальных условий работы конструкции от принятой расчетной схемы. Все они могут случайным образом изменяться под влиянием окружающей среды.

Библиографический список

1.Вентнцель Е.С. Теория вероятностей. - М.: Наука, 1999;

2.Гмурман В.Е. Теория вероятностей и математическая статистика. - М.: Высшая школа, 1977;

3. Гурский Е.Е. Теория вероятностей с элементами математической статистики. - М.: Высшая школа, 1971;

4. Бронштейн И.Н., Санендяев К.А. Справочник по математике. - М.: Наука, 1986;

5. Данко П.Е. Высшая математика в упражнениях и задачах. - М.: Наука, 1986;

6. Письменный Д. Конспект лекций по математической статистики. - М.: Айрис Пресс, 2004.