МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Национальный исследовательский
Нижегородский государственный университет
им. Н. И. Лобачевского»**

А. В. Зорин
В. А. Зорин
М. А. Федоткин

МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН И ПРОВЕРКА ГИПОТЕЗ О ВИДЕ РАСПРЕДЕЛЕНИЯ

Учебно-методическое пособие

Рекомендовано методической комиссией
Института ИТММ для студентов ННГУ,
обучающихся по направлению подготовки
010302 «Прикладная математика и информатика»

Нижний Новгород
2017

УДК 519.21
ББК В17
З86

З86 Зорин А. В., Зорин В. А., Федоткин М. А. МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН И ПРОВЕРКА ГИПОТЕЗ О ВИДЕ РАСПРЕДЕЛЕНИЯ: Учебно-методическое пособие. — Нижний Новгород: Нижегородский госуниверситет, 2017. — 19 с.

Рецензент: д. ф.-м. н, проф. кафедры АГДМ **Н. Ю. Золотых.**

В настоящем пособии изложены методы моделирования случайных величин с заданным законом распределения дискретного и непрерывного типов. Приводятся необходимые сведения по основам статистических методов построения выборочных распределений и выборочных числовых характеристик случайных величин. Для проверки гипотезы о виде распределения предлагается критерий согласия хи-квадрат. Пособие также содержит указания по выполнению лабораторной работы и варианты заданий.

Методическая разработка создана в помощь студентам высших профессиональных учебных заведений, изучающим общий курс “Теория вероятностей и математическая статистика”.

УДК 519.21
ББК В17

Введение для любознательного студента

Впервые идея моделирования случайных процессов на электронной вычислительной машине пришла в голову Станиславу Уламу (так он сам рассказывал в своей книге «Приключения Математика») во время работы над Манхэттанским проектом. До этого инженеры пользовались печатными таблицами случайных чисел (открывая страницу наугад и выписывая приведённые там числа). Дж. фон Нейман и С. Улам использовали в своих программах простое правило [1]: предыдущее (псевдо-)случайное число возводилось в квадрат и выделялись средние цифры; так можно было получать последовательность, например, десятизначных целых чисел. Ясно, что такая последовательность не является случайной. Однако она может проявлять некоторые статистические свойства, присущие случайным последовательностям.

Другой пример получения «нерегулярной» числовой последовательности приведён в [2]. Пусть a — некоторое число из интервала $(0,\ 1)$, и образуем последовательность по формуле: $x_n = \{na\}$. Здесь $\{\cdot\}$ — дробная часть числа. Г. Вейлем была доказана замечательная теорема о том, что для иррационального x так определенная последовательность чисел равномерно распределена на $(0,\ 1)$. Под равномерной распределённостью понимается следующее: для любых величин y_1 и y_2, $0 \leqslant y_1 < y_2 \leqslant 1$,

$$\lim_{n\to\infty}\left[\frac{1}{n} \times \text{число}\{x_i\colon y_1 \leqslant x_i \leqslant y_2;\ i = 1,\ 2,\ \ldots,\ n\}\right] = y_2 - y_1.$$

Однако генератор, основанный на этом принципе, обладает двумя существенными недостатками. Во-первых, соседние члены последовательности не являются статистически независимыми. Чтобы в этом убедиться, положим $y = \{nx\}$, $z = \{(n+1)x\}$ и вычислим ковариацию между y и z для равномерно распределённого x. Заметим, что $z = x + y$ при $y \leqslant 1 - x$ и $z = y + x - 1$ при $y > 1 - x$. Поэтому

$$\mathsf{cov}(y,\ z) = \mathsf{E}(y - 1/2)(z - 1/2) = \int\limits_0^{1-x} y(y+x)\,dy + \int\limits_{1-x}^{1} y(y+x-1)\,dy - 1/4 = 1/12 - 1/2x(1-x).$$

Ковариация не обращается в нуль, если x не есть $\frac{1}{2} \pm \frac{1}{2\sqrt{3}}$. Второй же недостаток состоит в том, что при вычислении на компьютере вещественные числа представляются не точно и в определённый момент последовательность может выродиться.

Библиотеки современных сред разработки программного обеспечения содержат, как правило, генераторы псевдослучайных последовательностей це-

лых чисел из конечного множества $\{0,\ 1,\ \ldots,\ N-1\}$. Часто используется «линейный конгруэнтный метод», заключающийся в простой рекуррентности

$$x_{n+1} = (ax_n + c) \mod m,$$

с целыми $m > 0$, $0 \leqslant a < m$, $0 \leqslant c < m$. По своему построению последовательность чисел x_0, x_1, ... будет периодической. Однако можно постараться выбрать числа a, c и m так, чтобы период был как можно больший. Подробное обсуждение вопросов, связанных с выбором таких чисел, содержится в [1].

Стоит отметить также наличие в операционной системе Linux специального псевдоустройства /dev/urandom, функционирующего как датчик случайных байтов. Работа этого устройства основана на фиксировании ядром операционной системы таких случайных событий, как, например, нажатия клавиш, и поддержании так называемого «энтропийного пула», из которого извлекается последовательность случайных байт.

Плохо или хорошо то, что компьютер как правило не даёт действительно случайной последовательности чисел? Оказывается, у псевдослучайных последовательностей есть заметное достоинство: они воспроизводимы, достаточно только начать с выбранного x_0. Таким образом, возможно отлаживать работу программ на неслучайных последовательностях, и потом уже запускать отлаженную программу со случайным x_0.

1. Моделирование случайных величин

Для моделирования случайной величины η с заданным законом распределения $F_\eta(x) = \mathsf{P}(\{\eta < x\})$ применяется функциональное преобразование случайной величины с равномерным законом распределения на $(0,\ 1)$. Библиотеки современных языков программирования предоставляют датчик псевдослучайных чисел, возвращающий число U из множества $\{0,\ 1,\ \ldots,\ N-1\}$. С практической точки зрения, возвращаемые величины имеют дискретное равномерное распределение на указанном множестве. Большинство способов получения случайных чисел с заданным законом распределения требуют использования случайной величины с равномерным законом распределения на отрезке $(0,\ 1)$. Приблизить такую величину можно отношением U/N. Всюду далее будем через $\mathcal{U}$ обозначать случайную величину с равномерным распределением на $(0,\ 1)$. Для работы с датчиком случайных чисел необходимо уметь выбирать начальное значение псевдослучайной последовательности. Выбор различных начальных значений даёт возможность испытывать различные отрезки последовательности псевдослучайных чисел.

В стандартной библиотеке языка программирования C для работы с псевдослучайными числами предусмотрены функции `srand(unsigned)`, `rand()` и константа `RAND_MAX`. Как правило, для получения случайного начально-

го значения используют системные часы. Для этого один раз в начале программы помещают вызов `srand(time(0))`. После этого выполнение команды `((double) rand())/RAND_MAX` выдаёт значение случайной величины с равномерным распределением на $[0,\ 1]$.

Опишем несколько типичных функциональных преобразований [1, 2, 3].

1.1. Моделирование дискретных случайных величин

Рассмотрим сначала моделирование дискретной случайной величины η с конечным числом значений $a_1, a_2, \ldots, a_K$. Пусть p_j — вероятность значения a_j, $j = 1, 2, \ldots, K$. Разобьём отрезок $[0,\ 1]$ на полуинтервалы $\Delta_1 = [0,\ p_1)$, $\Delta_2 = [p_1,\ p_1 + p_2)$, …, $\Delta_K = [p_1 + p_2 + \ldots + p_{K-1},\ 1]$. Пусть u — значение случайной величины $\mathcal{U}$. Найдём интервал Δ_j, которому принадлежит число u. Тогда в качестве значения случайной величины η выберем a_j.

Простейшим примером такой случайной величины является индикатор $I_A = I_A(\omega)$ события $A \subset \Omega$. По определению, I_A принимает значение 1 при $\omega \in A$, и принимает значение 0 при $\omega \notin A$. Предполагая, что вероятность $p = \mathsf{P}(A)$ определена, сможем «разыграть» значение случайной величины I_A следующим образом. Получим значение u случайной величины $\mathcal{U}$. Если $u < p$, положим $I_A = 1$, иначе положим $I_A = 0$.

Для конкретных распределений схема разбиения на подинтервалы легко модифицируется. Например, для ряда распределений (пуассоновское, геометрическое) число возможных значений счётно. Заметим, что отношение $k(j) = p_j/p_{j-1}$ для этих распределений простым образом зависит от j. Поэтому нет необходимости хранить (счётное!) число точек разбиения отрезка $[0,\ 1]$, а возможно легко находить границы через рекуррентное соотношение $p_j = k(j)p_{j-1}$.

Например, рассмотрим пуассоновское распределение с параметром λ. Для $j = 1, 2, \ldots$ имеем:

$$p_0 = e^{-\lambda}, \quad p_j = \frac{\lambda^j}{j!}e^{-\lambda}, \quad k(j) = p_j/p_{j-1} = \frac{\lambda}{j}\,.$$

Пусть для определённости $\lambda = 1$ и $u = 0{,}7359$. Поскольку $p_0 = e^{-1} \approx 0{,}3679$ и $u > p_0$, вычислим p_1: $p_1 = p_0\lambda/1 \approx 0{,}3679$. Тогда $p_0 + p_1 \approx 0{,}7358$ и $u > p_0 + p_1$. Аналогично, $p_2 = p_1\lambda/2 \approx 0{,}1839$, $p_0 + p_1 + p_2 \approx 0{,}9197$. Теперь $u \in [p_0 + p_1,\ p_0 + p_1 + p_2)$ и, значит, $\eta = 2$.

Для геометрического распределения функциональное преобразование может быть и другим. Случайная величина $\left[\ln\mathcal{U}/\ln(1-p)\right]$ имеет геометрическое распределение с параметром p. Здесь $[\cdot]$ обозначает целую часть числа. Однако накладные расходы на вычисление логарифма могут сделать такой способ медленным.

1.2. Моделирование непрерывных случайных величин

Пусть непрерывная случайная величина задана своей плотностью распределения $f_{\eta}(x)$. Функция распределения $F_{\eta}(x)$ будет непрерывной. Предположим дополнительно, что $F_{\eta}(x)$ монотонно возрастает на некотором интервале $(x_{\min},\ x_{\max})$, $-\infty \leqslant x_{\min} < x_{\max} \leqslant \infty$ и постоянна вне его. Рассмотрим функцию $G(x)\colon (0,\ 1) \to (x_{\min},\ x_{\max})$, обратную к $F_{\eta}(x)$ для $x \in (x_{\min},\ x_{\max})$. Покажем, что случайная величина $G(\mathcal{U})$ имеет функцию распределения $F_{\eta}(x)$. Действительно, пусть $x \in (x_{\min},\ x_{\max})$, тогда

$$\mathsf{P}(\{G(\mathcal{U}) < x\}) = \mathsf{P}(\{\mathcal{U} < F_{\eta}(x)\}) = F_{\eta}(x).$$

Рассмотрим пример. Пусть требуется получить значение случайной величины со смещённым показательным распределением, заданным функцией распределения

$$F(x) = \begin{cases} 0, & \text{если } x < \theta, \\ 1 - e^{-(x-\theta)}, & \text{если } x \geqslant \theta. \end{cases}$$

Найдём обратную функцию. Решая уравнение $y = 1 - e^{-(x-\theta)}$, получаем: $G(x) = \theta - \ln(1-x)$. Заметим, что если $\mathcal{U}$ имеет равномерное распределение на $(0,\ 1)$, то такое же распределение имеет величина $(1-\mathcal{U})$. Окончательно имеем: $\eta = \theta - \ln \mathcal{U}$.

Такой метод работает, если обратная функция $G(x)$ имеет простое выражение. Часто это не так. Например, из одной теоремы Ж. Лиувиля (J. Liuville, Mémoir sur l'intégration d'une classe de fonctions transcendantes // Journal für die Reine und Angewandte Mathematik, Vol.13, No. 2, 1835, pp. 93–118) следует, что функция распределения стандартного нормального закона

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} e^{-t^2/2}\, dt$$

не выражается через элементарные функции в конечном виде, а значит, и нет возможности явно записать обратную к ней. Поэтому для моделирования стандартной нормальной случайной величины применяют специальные функциональные преобразования. Рассмотрим два из них.

Пусть $\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_T$ — независимые случайные величины с равномерным на $(0,\ 1)$ распределением. С учётом $\mathsf{E}\mathcal{U}_1 = 1/2$, $\mathsf{D}\mathcal{U}_1 = 1/12$, по центральной предельной теореме А. М. Ляпунова [4, с. 219],

$$\lim_{T\to\infty} \mathsf{P}\left(\left\{ \frac{\mathcal{U}_1 + \mathcal{U}_2 + \ldots + \mathcal{U}_T - T/2}{\sqrt{T/12}} < x \right\} \right) = \Phi(x).$$

Например, при $T = 12$ случайная величина $\mathcal{N}_{(0,\,1)} = \mathcal{U}_1 + \mathcal{U}_2 + \ldots + \mathcal{U}_{12} - 6$ имеет приближённо стандартное нормальное распределение. Из известных свойств нормального распределения получаем, что случайная величина $\mathcal{N}_{(a,\,\sigma)} = a + \sigma\mathcal{N}_{(0,\,1)}$ имеет приближённо нормальное распределение со средним a и среднеквадратическим отклонением σ (указанные числовые характеристики точны). Недостатком этого метода является плохое поведение $\mathcal{N}_{(0,\,1)}$ «на хвостах» распределения. Очевидно, $-6 < \mathcal{N}_{(0,\,1)} < 6$, поэтому значения, меньше минус шести, и большие шести, просто не встретятся. Второй метод позволяет получать одновременно две независимые стандартные случайные величины. Этот метод носит название метода полярных координат Бокса-Мюллера-Марсальи. Приведём только алгоритм, следуя [1].

Шаг 1. Пусть $\mathcal{U}_1$ и $\mathcal{U}_2$ — независимы и равномерно распределены на $(0,\ 1)$. Положим $V_1 = 2\mathcal{U}_1 - 1$, $V_2 = 2\mathcal{U}_2 - 1$.

Шаг 2. Вычислим $S = V_1^2 + V_2^2$.

Шаг 3. Если $S \geqslant 1$, возвращаемся к шагу 1.

Шаг 4. Вычислим $\eta_1 = V_1\sqrt{\dfrac{-2\ln S}{S}}$, $\eta_2 = V_2\sqrt{\dfrac{-2\ln S}{S}}$. Это требуемые нормально распределенные случайные величины.

1.3. Моделирование конструктивно заданной случайной величины

В случае, когда значение случайной величины определяется исходом явно заданного эксперимента, возможно моделировать непосредственно эксперимент. Рассмотрим два примера.

Пример 1. Пусть техническое устройство состоит из четырёх узлов, функционирующих независимо друг от друга. Пусть вероятность поломки i-го узла известна и равна p_i. Вероятность успешного ремонта поломанного устройства равна q_i. Случайная величина η — число работающих устройств по истечении контрольного времени. Опишем алгоритм получения значения случайной величины η. Пусть I_{i1} — индикатор события «i-е устройство сломано», I_{i2} — индикатор события «i-е устройство не отремонтировано». Тогда $\mathsf{P}(\{I_{i1} = 1\}) = p_i$, $\mathsf{P}(\{I_{i2} = 1\}|\{I_{i1} = 1\}) = 1 - q_i$, и справедливо разложение $\eta = 4 - I_{11}I_{12} - I_{21}I_{22} - I_{31}I_{32} - I_{41}I_{42}$. Установим счётчик исправных узлов в значение четыре. Получим значение величины $I_{11}I_{12}$, остальные три слагаемые получаются аналогично. По способу, описанному на с. 5, получим значение I_{11} с известном распределением. Если оно равно 0, то переходим к следующему слагаемому. В противном случае, получим значение I_{21} с известным *условным* распределением. Если оно равно 1, уменьшим на единицу значение счётчика числа исправных узлов. Значение первого слагаемого найдено.

Пример 2. Урновая схема Пойи состоит в следующем. В урне лежат a белых шаров и b чёрных. Из урны n раз извлекается наудачу шар, фиксируется его цвет и шар возвращается в урну. Если это был белый шар, то в урну добавляются α белых шаров, а если чёрный — β чёрных. Случайная величина η — число белых шаров среди n извлечённых. При моделировании полезно разложение $\eta = I_1 + I_2 + \ldots + I_n$, где I_j — индикатор события «j-й извлечённый шар белого цвета». Если среди первых $(j-1)$ шаров k белых, то в урне окажутся $a + k\alpha$ белых и $b + (j-k-1)\beta$ чёрных шаров. Тогда условное распределение I_j вполне определено и можно воспользоваться приёмом со с. 5.

Пример 3. Пусть точка наудачу выбирается в круге радиуса r. Случайная величина η — расстояние от выбранной точки до центра круга. Для «розыгрыша» значения случайной величины необходимо получить равномерное распределение в круге. С этой целью поместим концентрически круг в квадрат со стороной $2r$. Затем выберем наудачу точку $(\mathcal{U}_1,\ \mathcal{U}_2)$ внутри квадрата. Координаты этой точки независимы и будут иметь равномерное распределение на $[0,\ 2r]$. Если выбранная точка не принадлежит кругу, то есть если $(\mathcal{U}_1 - r)^2 + (\mathcal{U}_2 - r)^2 > r^2$, то выбираем новую точку. Если точка в круге, то полагаем $\eta = \sqrt{(\mathcal{U}_1 - r)^2 + (\mathcal{U}_2 - r)^2}$.

Задание по первой части работы.

1. Получить номер варианта от преподавателя и ознакомиться с текстом задачи в Приложении.

2. Изучить теоретический материал части 1 и выбрать подходящий способ решения задачи.

3. Написать первую часть программы — «розыгрыш» значений случайной величины. Эта часть должна включать в себя отображение содержания задачи, пользовательский интерфейс для ввода необходимых параметров, вывод результатов.

Если случайная величина дискретна, её различные значения $y_1 < y_2 < \ldots < y_k$ должны быть сведены в таблицу следующего вида:

y_i	y_1	y_2	$\ldots$	y_k
n_i	n_1	n_2	$\ldots$	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	$\ldots$	$\frac{n_k}{n}$

Здесь $n_j = \text{число}\{x_i\colon x_i = y_j,\ i = 1,\ 2,\ \ldots,\ n\}$

Для непрерывной случайной величины значения требуется расположить в порядке возрастания: $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$. Здесь $x_{(j)}$ — j-е по возрастанию число среди наблюдений $x_1, x_2, \ldots, x_n$.

2. Статистические характеристики случайных величин

2.1. Общие статистические характеристики

Пусть $x_1, x_2, \ldots, x_n$ — выборочные значения случайной величины η. По аналогии с функцией распределения $F_\eta(x)$ случайной величины введём выборочную функцию распределения

$$\widehat{F}_\eta(x) = \frac{1}{n} \times \text{число}\{x_i\colon x_i < x,\ i = 1,\ 2,\ \ldots,\ n\}.$$

При фиксированной выборке $x_1, x_2, \ldots, x_n$ функция $\widehat{F}_\eta(x)$ является функцией распределения. Более того, по теореме Гливенко

$$\mathsf{P}\Big(\Big\{\sup_{-\infty<x<\infty} |\widehat{F}_\eta(x) - F_\eta(x)| \to 0\Big\}\Big) = 1.$$

Полезно вычислить величину

$$D = \max_{-\infty<x<\infty} |\widehat{F}_\eta(x) - F_\eta(x)|.$$

Учитывая тот факт, что функция распределения не убывает, а выборочная функция распределения имеет конечное число скачков, величину D можно вычислять по формуле:

$$D = \max_{1\leqslant j\leqslant n}\Big(\frac{j}{n} - F_\eta(x_{(j)} + 0),\ F_\eta(x_{(j)}) - \frac{j-1}{n}\Big),$$

где $x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$ — упорядоченные по возрастанию значения $x_1, x_2, \ldots, x_n$. А.Н. Колмогоровым было показано, что для непрерывных случайных величин предельное распределение

$$\lim_{n\to\infty} \mathsf{P}(\{\sqrt{n}D < x\})$$

не зависит от вида $F_\eta(x)$.

Приведём выборочные аналоги числовых характеристик η:

$\overline{x} = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$ — выборочное среднее,

$S^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$ — выборочная дисперсия,

$\widehat{R} = x_{(n)} - x_{(1)}$ — размах выборки,

$\widehat{Me} = x_{(k+1)}$ при $n = 2k+1$ или

$\widehat{Me} = (x_{(k)} + x_{(k+1)})/2$ при $n = 2k$ — выборочная медиана.

2.2. Характеристики дискретных случайных величин

Пусть среди значений $x_1,\ x_2,\ \ldots,\ x_n$ различные суть $y_1 < y_2 < \ldots < y_k$. Для дискретной случайной величины η важной числовой характеристикой является распределение частот $n_1,\ n_2,\ \ldots,\ n_k$. По определению, $n_j = \text{число}\{x_i\colon x_i = y_j,\ i = 1,\ 2,\ \ldots,\ n\}$. В силу закона больших чисел, $\dfrac{n_j}{n} \approx \mathsf{P}(\{\eta = y_j\})$. Более того, величина n_j имеет биномиальное распределение с параметрами n и $p = \mathsf{P}(\{\eta = y_j\})$. По интегральной теореме Муавра-Лапласа, при больших n и $\varepsilon > 0$

$$\mathsf{P}\left(\left\{\frac{|n_j - np|}{\sqrt{np(1-p)}} > \varepsilon\right\}\right) \approx 2\Phi(\varepsilon) - 1.$$

Преобразование левой части этого равенства даёт:

$$\mathsf{P}\left(\left\{\left|\frac{n_j}{n} - p\right| > \varepsilon\sqrt{\frac{p(1-p)}{n}}\right\}\right) \approx 2\Phi(\varepsilon) - 1.$$

Таким образом, для увеличения *точности* $\varepsilon\sqrt{p(1-p)n^{-1}}$ оценки вероятности p при заданной *надёжности* $(2\Phi(\varepsilon) - 1)$ на порядок число наблюдений надо увеличить в сто раз.

2.3. Характеристики непрерывных случайных величин

Для непрерывной случайной величины важной эмпирической характеристикой является гистограмма. Опишем принципы построения гистограммы. Разобьём числовой промежуток, в который попали наблюдения $x_1,\ x_2,\ \ldots,\ x_n$, на примыкающие промежутки $\Delta'_1,\ \Delta'_2,\ \ldots,\ \Delta'_k$. Пусть $n_j = \text{число}\{x_i\colon x_i \in \in \Delta'_j,\ i = 1,\ 2,\ \ldots,\ n\}$. В плоскости xOy на оси Ox отметим промежутки $\Delta'_1,\ \Delta'_2,\ \ldots,\ \Delta'_k$ и над j-м из них построим прямоугольник высотой $n_i/(n|\Delta'_j|)$. Гистограммой называют фигуру, составленную из этих прямоугольников. Отметим статистический смысл гистограммы: поскольку

$$\frac{n_j}{n} \approx \mathsf{P}(\{\eta \in \Delta'_j\}) = \int\limits_{\Delta'_j} f_\eta(x)\,dx,$$

а по теореме о среднем (для непрерывной функции $f_\eta(x)$)

$$\int\limits_{\Delta'_j} f_\eta(x)\,dx \approx f_\eta(\xi)|\Delta'_j|, \quad \xi \in \Delta'_j,$$

то $n_i/(n|\Delta'_j|) \approx f(\xi)$, где $|\Delta'_j|$ — длина промежутка Δ'_j. Площадь гистограммы всегда равна 1.

Задание по второй части лабораторной работы

Определить теоретические и выборочные числовые характеристики: $\mathsf{E}\eta$, $\mathsf{D}\eta$, $\overline{x}$, S^2, $\widehat{Me}$, $\widehat{R}$. Составить таблицу:

$\mathsf{E}\eta$	$\overline{x}$	$\lvert\mathsf{E}\eta - \overline{x}\rvert$	$\mathsf{D}\eta$	S^2	$\lvert\mathsf{D}\eta - S^2\rvert$	$\widehat{Me}$	$\widehat{R}$
...	...	...	...	...	...	...	...

Построить графики теоретической $F_{\eta}(x)$ и выборочной $\widehat{F}\eta(x)$ функций распределения. Вычислить меру их расхождения D.

А) Случайная величина дискретна.

Определить из условий задачи закон распределения случайной величины η, если он не указан явно. Используя величины y_1, $_2$, ..., y_k, n_1, n_2, ..., n_k, найденные в первой части работы, вычислить теоретические вероятности $\mathsf{P}(\{\eta = y_j\})$ и найти отклонения $\left|\dfrac{n_j}{n} - \mathsf{P}(\{\eta = y_j\})\right|$. Найти максимальное отклонение $\max\limits_{j=\overline{1,\,k}}\left|\dfrac{n_j}{n} - \mathsf{P}(\{\eta = y_j\})\right|$. Результаты оформить в виде таблицы:

y_j	y_1	y_2	...	y_k
$\mathsf{P}(\{\eta = y_j\})$	$\mathsf{P}(\{\eta = y_1\})$	$\mathsf{P}(\{\eta = y_2\})$	...	$\mathsf{P}(\{\eta = y_k\})$
$\dfrac{n_j}{n}$	$\dfrac{n_1}{n}$	$\dfrac{n_2}{n}$	...	$\dfrac{n_k}{n}$

$$\max_{j=\overline{1,\,k}}\left|\frac{n_j}{n} - \mathsf{P}(\{\eta = y_j\})\right| = \ldots$$

Б) Случайная величина непрерывна.

Определить из условий задачи закон распределения случайной величины η, если он не указан явно. Организовать ввод границ промежутков Δ'_1, Δ'_2, ..., Δ'_k. Построить гистограмму. Вычислить теоретическую плотность распределения $f(z_j)$ в точке z_j — середине промежутка Δ'_j. Результаты оформить в виде таблицы:

z_j	z_1	z_2	$\dots$	z_k
$f_\eta(z_j)$	$f_\eta(z_1)$	$f_\eta(z_2)$	$\dots$	$f_\eta(z_k)$
$\dfrac{n_j}{n\lvert\Delta'_j\rvert}$	$\dfrac{n_1}{n\lvert\Delta'_1\rvert}$	$\dfrac{n_2}{n\lvert\Delta'_2\rvert}$	$\dots$	$\dfrac{n_k}{n\lvert\Delta'_k\rvert}$

$$\max_{j=\overline{1,\,k}}\left|\frac{n_j}{n|\Delta'_j|} - f_\eta(z_j)\right| = \dots$$

3. Проверка гипотезы о виде распределения

Предположение о распределении, которому подчиняются выборочные значения, будем называть гипотезой. Проверяемая гипотеза называется нулевой гипотезой и обозначается H_0. Часто вместе с нулевой гипотезой рассматривается конкурирующая гипотеза H_1. Если таковой явно не сформулировано, можно считать, что конкурирует гипотеза «данные имеют другое распределение, нежели $F_\eta(x)$». Под статистическим критерием для проверки нулевой гипотезы понимается правило (функция), которое каждому набору выборочных значений $x_1,\ x_2,\ \dots,\ x_n$ приписывает решение «принять нулевую гипотезу» или «отклонить нулевую гипотезу». Чтобы задать такое правило, достаточно задать разбиение множества наборов выборочных значений на два подмножества — область принятия гипотезы и область отклонение гипотезы. Область отклонения гипотезу называют критической областью.

С применением критерия связаны два типа возможных ошибок. Ошибкой первого рода называется отклонение верной нулевой гипотезы. Ошибкой второго рода называется принятие неверной нулевой гипотезы. Вероятность α ошибки первого рода можно найти, вычислив при нулевой гипотезе вероятность попадания наблюдений в критическую область (если гипотеза состоит из единственного распределения). Если конкурирующая гипотеза также состоит из единственного распределения, можно вычислить вероятность β ошибки второго рода — вероятность попасть в область принятия гипотезы при конкурирующем распределении. Величина α называется уровнем значимости критерия, а величина $1-\beta$ — мощностью критерия.

При построении критерия используют статистику $T = T(\eta_1,\ \eta_2,\ \dots,\ \eta_n)$, то есть неслучайную функцию $T(x_1,\ x_2,\ \dots,\ x_n)$ от случайных величин η_1, η_2, $\dots$, η_n. Критическую область при этом составляют из соответствующих значений статистики T. Так, для непрерывных случайных величин для проверки гипотезы о том, что выборочные значения действительно принадлежат

распределению $F_\eta(x)$, можно в качестве статистики выбрать величину D и в качестве критической области взять интервал больших значений $\{D > d\}$. Если гипотеза верна, то распределение статистики D есть «распределение статистики Колмогорова-Смирнова» (тавтология). Вид функции распределения этой статистики приведён в Приложении. Зная распределение, можно выбрать границу d из условия $\mathsf{P}_{H_0}(\{D > d\}) = \alpha$, то есть гарантировать уровень значимости критерия.

Рассмотрим подробнее критерий χ^2 для проверки согласия данных с распределением $F_\eta(x)$. Разобьём числовую ось на интервалы $\Delta_1'' = (-\infty,\ z_1)$, $\Delta_2'' = [z_1,\ z_2)$, $\ldots$, $\Delta_k'' = (z_{k-1},\ \infty)$, положим $z_0 = \infty$, $z_k = \infty$. Интервалы следует выбирать так, чтобы каждый содержал хотя бы одну точку, а лучше несколько. Пусть n_j — число наблюдений, попавших в Δ_j'', $q_j = \mathsf{P}_{H_0}(\{\eta \in \Delta_j''\}) = F_\eta(z_j) - F_\eta(z_{j-1})$. В качестве статистики критерия выберем величину

$$R_0 = \sum_{j=1}^{k} \frac{(n_j - nq_j)^2}{nq_j}.$$

Величина R_0 характеризует меру расхождения между наблюдавшимися частотами и ожидаемым числом попаданий в интервал при нулевой гипотезы. При справедливости нулевой гипотезы величина R_0 имеет распределение χ^2 с $k-1$ степенями свободы. Не согласующимися с нулевой гипотезой являются большие значения R_0. Выберем критическую область вида $(\chi^2_{\alpha;k-1},\ \infty)$, где число $\chi^2_{\alpha;k-1}$ является решением уравнения $\mathsf{P}_{H_0}(\{R_0 > \chi^2_{\alpha;k-1}\}) = \alpha$. По построению, такой критерий будет иметь уровень значимости α. Таким образом, возможно по α и k вычислить заранее границу $\chi^2_{\alpha;k-1}$, и потом проверять принадлежность R_0 критической области. Второй путь состоит в следующем. Пусть $F_{\chi^2_{k-1}}(x)$ — функция распределения χ^2 с $k-1$ степенью свободы. Заметим, что функция $\overline{F}(x) = 1 - F_{\chi^2_{k-1}}(x)$ — невозрастающая. Значит, $R_0 > \chi^2_{\alpha;k-1}$ тогда и только тогда, когда $\overline{F}(R_0) < \overline{F}(\chi^2_{\alpha;k-1}) = \alpha$.

Заметим однако, что при справедливой нулевой гипотезе малые значения R_0 могут быть также редки. Действительно, при $r > 1$ мода распределения χ^2 приходится на $(r-2)$, и чаще всего значения R_0 будут около того или даже больше.

Пусть случайная величина η дискретна. Тогда

$$q_j = \sum_{\{i\,:\,z_{j-1} \leqslant a_i < z_j\}} p_i.$$

Для непрерывной случайной величины

$$q_j = \int\limits_{z_{j-1}}^{z_j} f_\eta(u)\,du.$$

Метод вычисления вероятностей, связанных с нормальным распределением, приводится в Приложении.

Задание по третьей части лабораторной работы

Заключительная часть программы должна включать в себя:

1. Ввод числа k интервалов $\Delta_1'' = (-\infty,\ z_1)$, $\Delta_2'' = [z_1,\ z_2)$, $\ldots$, $\Delta_k'' = (z_{k-1},\ \infty)$. Выбор границ интервалов z_1, z_2, $\ldots$, z_{k-1}.

2. Отображение гипотезы в виде теоретических вероятностей q_1, q_2, $\ldots$, q_k.

3. Ввод уровня значимости α.

4. Отображение вычисленного значения $\overline{F}(R_0)$ и решения о принятии или отвержении гипотезы H_0.

Литература

1. Кнут, Д. Э. Искусство программирования, том 2. Получисленные алгоритмы, 3-е изд.: Пер. с англ. / Д. Э. Кнут. — М.: Издательский дом «Вильямс», 2003. — 832 с.

2. Гренандер, У. Краткий курс вычислительной вероятности и статистики / У. Гренандер, В. Фрайбергер. — М.: Наука, 1978. — 192 с.

3. Ермаков, С. М. Курс статистического моделирования. / С. М. Ермаков, Г . А. Мизайлов. — М.: Наука, 1976. — 319 с.

4. Федоткин, М. А. Основы прикладной теории вероятностей и статистики: Учебник / М. А. Федоткин. — М.: Высшая школа, 2006. — 368 с.: ил.

5. Чистяков, В. П. Курс теории вероятностей: Учебник. — 3-е издание, исправленное. — М.: Наука, 1987. — 240 с.

ПРИЛОЖЕНИЕ

Задания

1. Передаётся n сообщений по каналу связи. Каждое сообщение с вероятностью p независимо от других искажается. С.в. η — число искажённых сообщений.

2. При каждом цикле обзора радиолокатора объект (независимо от других циклов) обнаруживается с вероятностью p. С.в. η — число циклов обзора до обнаружения объекта.

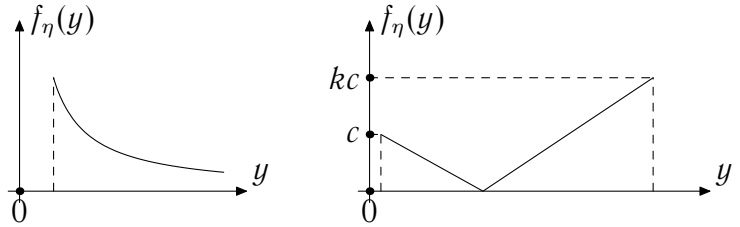
3. ЭВМ генерирует последовательность чисел до получения некоторого заданного числа. Вероятность генерации этого числа на каждом шаге независимо от других шагов равна p. С.в. η — число элементов полученной последовательности.

4. В лотерее среди N билетов M выигрышных. Игрок покупает r билетов. С.в. η — число выигрышных билетов среди купленных.

5. На автоматическую телефонную станцию поступает поток вызовов с интенсивностью λ. С.в. η — число вызовов за t минут, имеет распределение Пуассона со средним λt.

6. С.в. η — время обслуживания покупателя в кассе магазина. Пусть η распределена показательно с параметром λ.

7. В очереди к кассе стоят $N \gg 1$ человек. Сумма, которую нужно заплатить отдельному лицу, есть случайная величина со средним $\tilde{m}$ и дисперсией $\tilde{d}$. Вид плотности распределения выбрать по аналогии с приведёнными на рисунках.



   С.в. η — общая сумма сумма выплат.

8. Скорость соударения молекул — случайная величина, распределённая по закону Релея с параметром σ:

$$f_\eta(y) = \frac{y}{\sigma^2}\exp\Big\{-\frac{y^2}{2\sigma^2}\Big\}, \quad y \geqslant 0.$$

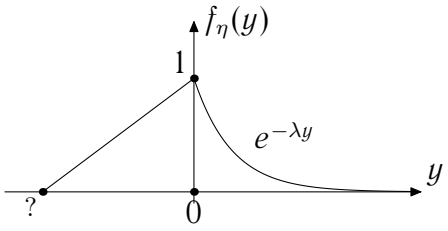
9. Каждому из трёх стрелков предоставляется возможность поразить цель с r выстрелов. Вероятность попадания в мишень для этих стрелков при j-м выстреле равна p_{jk}. При поражении мишени стрелком следующие выстрелы не производятся. С.в. η — общее число произведённых выстрелов.

10. Два баскетболиста поочерёдно бросают мяч в корзину до первого попадания одним из баскетболистов. Вероятность попадания при каждом броске для первого баскетболиста равна p_1, для второго — p_2. С.в. η — число бросков, произведённых вторым (по очереди) баскетболистом.

11. Распределение случайной величины η задано плотностью:

$$f_\eta(y) = ae^{-\lambda|y|}.$$

12. Плотность распределения с.в. η задана графически:



13. В течение некоторого времени испытываются M приборов на надёжность. Вероятность выхода из строя каждого прибора независимо от других равна p. С.в. η — число отказавших приборов.

14. Студенту на зачёте задаются вопросы, которые прекращаются, если студент на заданный вопрос не ответит. Вероятность ответа на каждый вопрос независимо от других равна p. С.в. η — число полученных ответов.

15. В партии из N лампочек M перегоревших. С.в. η — число перегоревших лампочек среди r выбранных наудачу.

16. К светофору по некоторому направлению подъезжают машины с интенсивностью λ. С.в. η — число машин, прибывших за t секунд, распределена по закону Пуассона с параметром λt.

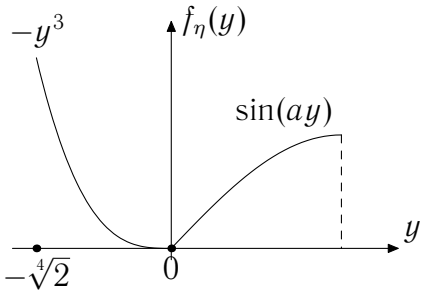
17. С.в. η — время безотказной работы ЭВМ. Пусть η распределена показательно с параметром λ.

18. Устройство состоит из $N \gg 1$ дублирующих приборов. Каждый следующий прибор включается после выхода из строя предыдущего. Время безотказной работы каждого прибора — положительная с.в. со средним

Q и дисперсией R. Плотность распределения выбрать по аналогии с приведёнными на рисунках на с. 15. С.в. η — время безотказной работы всего устройства.

19. Имеются три колоды карт. Первая из них обычная в 36 листов, из второй удалены все «вини», в третьей нет вальтов. Из каждой колоды берут по одной карте. С.в. η — общее число вынутых картинок.

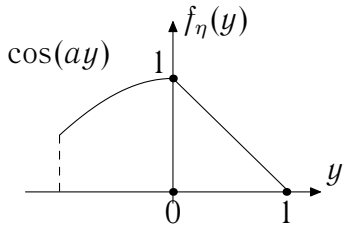
20. Плотность распределения с.в. η задана графически:



21. Вероятность изготовления годной детали из каждой отливки независимо от других равна p. С.в. η — число использованных отливок для изготовления годной детали.

22. В сбербанк внесли деньги $N \gg 1$ человек. Сумма вклада каждого из них есть с.в. η со средним A и дисперсией B. С.в. η — общая сумма вклада.

23. Плотность распределения с.в. η задана графически:



24. Точка наудачу выбирается в правильном треугольнике. Случайная величина η — расстояние до центра треугольника.

25. Точка наудачу выбирается в правильном треугольнике. Случайная величина η — расстояние до ближайшей стороны треугольника.

26. Скорость пули имеет нормальное распределение со средним a и дисперсией σ^2. Чтобы пробить фанерный лист, пуля должна иметь скорость не меньше, чем v_0. С.в. η — скорость пули, пробившей лист.

27. Девушка пригласила на свидание k молодых людей. Каждый молодой человек имеет привычку опаздывать на случайное время с показательным распределением со средним λ_k. Девушка ждёт первого пришедшего и с ним уходит. С.в. η — время ожидания девушкой.

28. Пусть с.в. θ имеет нормальное распределение со средним a и дисперсией σ^2. С.в. $\eta = e^\theta$.

29. Монета подбрасывается до тех пор, пока орёл не появится три раза. С.в. η — число подбрасываний монеты.

30. Случайная величина λ имеет показательное распределение с параметром a. При «разыгранном» значении λ случайная величина η имеет распределение Пуассона с параметром λ.

31. Число покупателей, посетивших магазин в течение дня, случайно и имеет пуассоновское распределение со средним λ. Стоимость покупок одного человека имеет нормальное распределение со средним m и дисперсией σ^2. С.в. η — дневная выручка магазина.

32. С.в. η — сумма n независимых случайных величин с плотностью распределения
$$f(x) = \begin{cases} 0, & |x| > e^{-1} \\ (2|x| \ln^2 |x|)^{-1}, & |x| < e^{-1}. \end{cases}$$

Формулы для численных расчётов

Плотность распределения χ^2 с r степенями свободы имеет вид:

$$f_{\chi^2_r}(x) = \begin{cases} 0, & x \leqslant 0, \\ 2^{-r/2}[\Gamma(r/2)]^{-1} x^{r/2-1} e^{-x/2}, & x > 0. \end{cases}$$

Исходя из вида плотности, вычислить

$$\overline{F}(R_0) = \int\limits_{R_0}^{\infty} f_{\chi^2_r}(x)\,dx = 1 - \int\limits_{0}^{R_0} f_{\chi^2_r}(x)\,dx$$

можно, например, методом трапеций:

$$\int\limits_a^b g(x)\,dx \approx \sum_{k=1}^{n} \Big(g\Big(a + (b-a)\frac{k-1}{n}\Big) + g\Big(a + (b-a)\frac{k}{n}\Big)\Big) \frac{b-a}{2n}.$$

При этом можно использовать соотношение $\Gamma(a) = (a-1)\Gamma(a-1)$ и известные значения

$$\Gamma(1) = 1, \quad \Gamma(1/2) = \sqrt{\pi}\,.$$

Приближённое вычисление функции $\Phi(x)$ можно проводить также методом трапеций или на основании формулы (см. Зубков А.М., Севастьянов Б.А., Чистяков В.П. Сборник задач по теории вероятностей, М.: Наука, 1989. с. 34):

$$\begin{aligned} P_x &= 2(1-\Phi(x)) = \frac{2}{\sqrt{2\pi}} \int\limits_x^\infty e^{-u^2/2}\,dx = \\ &= \exp\left\{-\frac{(83x+351)x+562}{703/x+165}\right\}, \quad 0 < x < 5{,}5; \\ P_x &= \sqrt{\frac{2}{\pi}} \cdot \frac{1}{x} \exp\left\{-\frac{x^2}{2} - \frac{0{,}94}{x^2}\right\}, \quad x \geqslant 5{,}5. \end{aligned}$$

Для распределения χ^2 можно также использовать разложения неполной гамма-функции

$$I(x,\ p) = \frac{1}{\Gamma(p)} \int\limits_0^x u^{p-1} e^{-u}\,du$$

при $p \leqslant x < 1$ или при $x < p$ в ряд

$$I(x,\ p) = \frac{\exp\{-x\}x^p}{\Gamma(p+1)}\Big[1 + \sum_{r=1}^{\infty} \frac{x^r}{(p+1)(p+2)\cdots(p+r)}\Big],$$

а для остальных случаев в непрерывную дробь:

$$I(x,\ p) = 1 - \frac{\exp\{-x\}x^p}{\Gamma(p)} \cfrac{1}{x + \cfrac{1-p}{1 + \cfrac{1}{x + \cfrac{2-p}{1 + \cfrac{2}{x + \ldots}}}}}.$$

Использовать равенство $F_{\chi^2_r}(x) = I(x2^{r/2},\ r/2)$.

Функция распределения статистики Колмогорова-Смирнова (с. 9) имеет вид:

$$F_{KC}(x) = \sum_{j=-\infty}^{\infty} (-1)^j e^{-2j^2x^2}.$$

Андрей Владимирович **Зорин**
Владимир Александрович **Зорин**
Михаил Андреевич **Федоткин**

МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН И ПРОВЕРКА ГИПОТЕЗ О ВИДЕ РАСПРЕДЕЛЕНИЯ

Учебно-методическое пособие

Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский
Нижегородский государственный университет
им. Н. И. Лобачевского»
603950, Нижний Новгород, пр. Гагарина, 23