И. ВЫСОЦКИЙ,
г. Москва

Часть 18. Окончание

ЗАДАЧИ С УЛИЦЫ

Задача экзаменатора и прочие случайности

Это последняя статья цикла «Задачи с улицы». Она полностью посвящена вопросам статистического моделирования в MS Excel. Обычный процессор электронных таблиц не уступает специализированным статистическим пакетам в части моделирования выборок с определенными свойствами. Думаю, что он даже во многом их превосходит благодаря разнообразию и гибкости встроенных функций.

Случайная перестановка экзаменационных билетов

Легко жилось профессору в старые добрые времена, когда достаточно было разложить экзаменационные билеты на столе лицом вниз. Каждый студент сам выбирал билет, и можно было не беспокоиться о том, что двоим попадется одно и то же. А что делать, если зачет дистанционный? Как случайным образом распределить билеты без повторений? Иными словами, как, имея персональный компьютер, устроить случайную перестановку без специальных программ?

Задача 1. Задача экзаменатора. Предположим, в группе 10 студентов и для них заготовлено 12 билетов. Числа мы взяли небольшие, чтобы все данные поместились на рисунках. Как организовать случайную перестановку в MS Excel?

Решение. 1. Расположим на листе исходные данные, например, в столбцах B–D (рис. 1,*а*).

	A	B	C	D
1				
2		Задача экзаменатора. Случайная перестановка		
3				
4			Список группы	Ном.билетов
5		1	Алексеев	1
6		2	Баранова	2
7		3	Владимиров	3
8		4	Георгиев	4
9		5	Докучаева	5
10		6	Евсеева	6
11		7	Железнов	7
12		8	Закурдаева	8
13		9	Иващенко	9
14		10	Колокольцева	10
15				11
16				12

а)

	A	B	C	D	E
1					
2		Задача экзаменатора. Случайная перестановка			
3					
4			Список группы	Сл.числа	Билеты
5		1	Алексеев	0,043137	1
6		2	Баранова	=СЛЧИС()	2
7		3	Владимиров	0,473531	3
8		4	Георгиев	0,69749	4
9		5	Докучаева	0,430652	5
10		6	Евсеева	0,192041	6
11		7	Железнов	0,789862	7
12		8	Закурдаева	0,737071	8
13		9	Иващенко	0,65568	9
14		10	Колокольцева	0,395768	10
15				0,340422	11
16				0,505709	12

б)

Рис. 1

2. Вставим пустой столбец D так, чтобы номера билетов передвинулись в столбец E. В новый столбец D поместим 12 случайных чисел с помощью функции СЛЧИС (рис. 1,*б*).

3. Выделим данные в диапазоне D5:E16 и отсортируем их по возрастанию с помощью вкладки *Сортировка и фильтр* (рис. 2). Случайные числа в столбце D расположатся по возрастанию, а вместе с ними переставятся номера билетов в столбце E (рис. 3). Если до

Есть дополнительные материалы на сайте raum.math.ru.

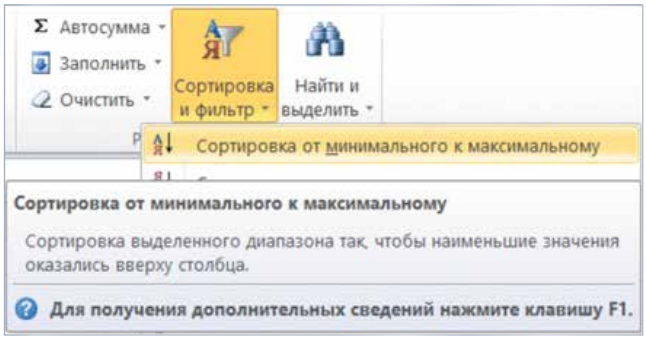


Рис. 2

сортировки числа в столбце D были расположены в порядке некоторой случайной перестановки τ, то теперь номера билетов расположились в порядке обратной перестановки τ^{-1}. Алексеев получит билет 1, Баранова — билет 4 и т.д.

С тем же успехом можно было сортировать по убыванию: на степень случайности это не влияет.

Комментарии

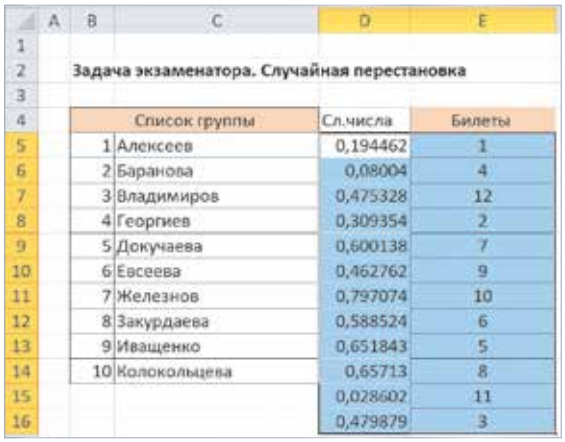
1. Функция СЛЧИС выдает (говорят «возвращает») случайное число, которое равномерно распределено на интервале от 0 до 1. Равномерное распределение мы здесь понимаем в обычном смысле: вероятность попадания этого числа в любой интервал, заключенный между 0 и 1, пропорциональна длине этого интервала.

2. Функции СЛЧИС не нужны аргументы. Поэтому ее синтаксис требует пустых скобок:

=СЛЧИС().

Генератор случайных целых чисел

Иногда для урока вероятности или статистики нужно смоделировать многократное бросание монеты, кубика или какой-нибудь другой генератор случайных целых чисел. Для этого можно использовать готовые приложения. Например, на сайте «Вероятность в школе» (ptlab.mccme.ru/node/187) можно скачать несложные программки для Windows, имитирующие бросание монет или кубиков. Но как быть, если нужен вполне определенный дизайн опыта, а готовое приложение устроено иначе? Кроме того, что может быть полезнее и интереснее в учебном смысле, чем самоделка?



Задача экзаменатора. Случайная перестановка

	Список группы	Сл.числа	Билеты
1	Алексеев	0,194462	1
2	Баранова	0,08004	4
3	Владимиров	0,475328	12
4	Георгиев	0,309354	2
5	Докучаева	0,600138	7
6	Евсеева	0,462762	9
7	Железнов	0,797074	10
8	Закурдаева	0,588524	6
9	Иващенко	0,651843	5
10	Колокольцева	0,65713	8
		0,028602	11
		0,479879	3

Рис. 3

Пусть, например, нужно всесторонне исследовать многократное бросание игрального кубика.

Задача 2. Генератор случайных чисел. Смоделировать в MS Excel 50 бросаний симметричного игрального кубика, представить на диаграмме последовательность выпавших чисел, а также последовательности их сумм и средних.

Решение. Сделаем универсальную таблицу, которая может моделировать не только кубик, но любой генератор случайных целых чисел от a до b.

Используем функцию СЛМЕЖДУ(a;b). Она возвращает случайное целое число от a до b. На рисунке 4 эти границы указаны в ячейках G6 и G7, и при этом их можно менять. В столбец A поместим номера опытов (бросаний) k,

Генератор случайных целых чисел

Номер опыта k	Случ. Число X_k	Сумма S_k	Среднее S_k/k	Диапазон		Значение X_k Мат. ожид.	Станд. откл	Нижн. гран.	Верхн. гран.	Сумма S_k Мат. ожид.	Станд. откл	Нижн. гран.	Верхн. гран.	Среднее S_k/k Станд. откл	Нижн. гран.	Верхн. гран.
1	3	3	3,00	от	1	5,5	2,8723	2,6277	8,3723	5,5	2,8723	2,6277	8,3723	2,8723	2,6277	8,3723
2	10	13	6,50	до	10	5,5	2,8723	2,6277	8,3723	11	4,062	6,938	15,062	2,031	3,469	7,531
3	2	15	5,00			5,5	2,8723	2,6277	8,3723	16,5	4,9749	11,525	21,475	1,6583	3,8417	7,1583
4	9	24	6,00	m=	1	5,5	2,8723	2,6277	8,3723	22	5,7446	16,255	27,745	1,4361	4,0639	6,9361
5	3	27	5,40			5,5	2,8723	2,6277	8,3723	27,5	6,4226	21,077	33,923	1,2845	4,2155	6,7845
6	8	35	5,83			5,5	2,8723	2,6277	8,3723	33	7,0356	25,964	40,036	1,1726	4,3274	6,6726

Рис. 4

в столбцах B–D пусть будут результаты X_k, их накапливающиеся суммы

$$S_k = X_1 + X_2 + \ldots + X_k$$

и средние полученных значений от 1-го до k-го

$$\bar{X}_k = \frac{S_k}{k}.$$

В столбцах I–S вычисляются характеристики каждой из этих трех случайных величин: математическое ожидание, стандартное отклонение и верхняя и нижняя границы выбранного интервала значений (краевые тенденции), полученные как математическое ожидание плюс или минус m стандартных отклонений. Число m вводится в ячейку G9.

Для вычисления стандартного отклонения величины X_k (оно общее для всех k) в ячейках столбца J используется формула

$$\sqrt{\mathrm{D}X_k} = \sqrt{\frac{1^2 + \ldots + n^2}{n} - \left(\frac{1 + \ldots + n}{n}\right)^2} = \sqrt{\frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4}} = \sqrt{\frac{n^2-1}{12}},$$

где n — количество возможных значений X_k, которое получается равно G7 – G6 + 1. Например, для обычного кубика это 6 – 1 + 1 = 6, а для числа орлов, выпавших при бросании монетки, количество возможных значений n = 1 – 0 + 1 = 2: орлов либо нет, либо один.

Справа на том же листе построим диаграммы последовательных значений величин X_k, S_k и $\bar{X}_k$ для k = 1, ..., 50 (рис. 5). Каждую диаграмму снабдим осью (график математического ожидания) и краевыми тенденциями. Например, на диаграмме значений величины X_k (рис. 5,*а*)

Рис. 5

получаются три горизонтальные прямые на уровнях:

$$\mathrm{E}X_k = \frac{a+b}{2},\quad \mathrm{E}X_k - m\sqrt{\mathrm{D}X_k}\ ,\ \mathrm{E}X_k + m\sqrt{\mathrm{D}X_k}.$$

Математические ожидания и стандартные отклонения величин S_k и $\bar{X}_k$ получаются из $\mathrm{E}X_k$ и $\sqrt{\mathrm{D}X_k}$ с помощью известных свойств ожидания и дисперсии:

$$\mathrm{E}S_k = k\mathrm{E}X_k,\quad \sqrt{\mathrm{D}S_k} = \sqrt{k\mathrm{D}X_k}$$

и

$$\mathrm{E}\bar{X}_k = \mathrm{E}X_k,\quad \sqrt{\mathrm{D}\bar{X}_k} = \sqrt{\frac{\mathrm{D}X_k}{k}}.$$

Комментарий. При каждом изменении содержимого любой ячейки процессор пересчитывает всю таблицу; в частности, случайные числа в столбце B генерируются заново. Это можно с успехом использовать. Поставьте курсор в любую пустую ячейку (например, в A1) и нажимайте клавишу *Delete*. При каждом нажатии вы будете получать новые серии случайных чисел, их сумм и средних, диаграммы также будут обновляться. Это дает возможность визуально оценить изменчивость каждой из этих трех величин и частоту попадания их в выбранные интервалы.

Статистическое моделирование дискретного распределения

Предположим, что нам нужно получить выборку значений случайной величины, подчиняющейся некоторому известному распределению. Простейший случай мы уже рассмотрели: генератор случайных целых чисел дает выборку из равномерного дискретного распределения. До сих пор мы строили эту выборку в виде последовательности чисел, а также последовательностей их сумм и средних, но не строили распределение частот получившихся значений (*эмпирическое распределение*).

Пусть, например, мы хотим бросить игральный кубик 500 раз и посмотреть частоты выпадения единиц, двоек и т.д., а также сравнить полученное эмпирическое распределение с теоретическим. Это выборочное исследование объемом 500.

Другой пример: стрелок попадает в мишень в тире с вероятностью $p = 0{,}2$. Мы хотим посмотреть на выборке объемом 500, как ведут себя частоты событий «ни разу не попал», «попал один раз», «попал два раза» и т.д., если стрелок стреляет в мишень, скажем, 50 раз. Для этого нам нужно заставить этого стрелка сделать 500 серий по 50 выстрелов. Полученное распределение частот можно сравнить с теоретическим биномиальным распределением числа успехов в серии испытаний Бернулли длиной $n = 50$ и с вероятностью успеха $p = 0{,}2$.

Точно так же можно моделировать геометрическое распределение: попросить этого стрелка сделать 500 серий до первого попадания в мишень и смотреть на частоты событий «потребовался один выстрел», «потребовалось два выстрела» и т.д.

В общем случае задача такая: нам нужно статистически смоделировать дискретное конечное или бесконечное распределение

$$\begin{pmatrix} x_1 & x_2 & \dots & x_k & \dots \\ p_1 & p_2 & \dots & p_k & \dots \end{pmatrix}$$

посредством выборки объемом $N = 500$ (например). При этом нужно привлечь минимальные средства, желательно универсальные, позволяющие почти не думать, какое именно распределение мы моделируем. Можно ли сделать это с помощью подручных средств в электронной таблице?

Задача 3. Абстрактная. Получить случайную выборку объемом 500 из распределения

$$R = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 0{,}1 & 0{,}15 & 0{,}25 & 0{,}25 & 0{,}15 & 0{,}1 \end{pmatrix},$$

для сравнения построить на одном поле диаграмму распределения частот значений (гистограмму) и диаграмму распределения.

Решение. Разобьем интервал (0; 1) на шесть интервалов, длины которых равны вероятностям из распределения R (рис. 6). Границы интервалов образуют последовательность накопленных вероятностей:

$$0{,}1;$$
$$0{,}1 + 0{,}15 = 0{,}25;$$
$$0{,}25 + 0{,}25 = 0{,}5;$$

и т.д. до единицы (правая граница j-го интервала равна сумме вероятностей значений, не превосходящих j).

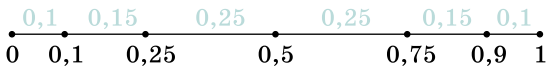


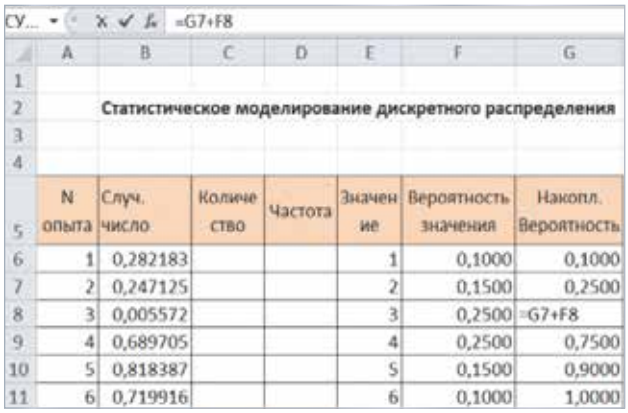
Рис. 6

Бросим в интервал (0; 1) случайное число x. Вероятность того, что x попадет в какой-то из полученных интервалов, равна длине этого интервала. Поэтому значение 1 в распределении R имеет ту же вероятность, что и событие $(0 < x < 0{,}1)$; значение 2 имеет ту же вероятность, что и событие $(0{,}1 < x < 0{,}25)$, и так далее.

Если бросить независимо друг от друга 500 таких чисел x, мы сможем подсчитать частоту попадания чисел x в интервал с номером j. Эта ча-

стота и будет частотой, с которой случайная величина, подчиняющаяся распределению R, принимает значение j.

Откроем чистый лист MS Excel. В столбец A внесем номера опытов от 1 до 500 (рис. 7). В столбец B поместим случайные независимые числа с помощью известной нам функции СЛЧИС. Распределение R запишем вертикально в столбцах E и F. Нам нужны границы интервалов, как на рисунке 6. Впишем их в ячейки от G6 до G11, они получаются суммированием вероятностей из столбца F.



	A	B	C	D	E	F	G
1							
2		Статистическое моделирование дискретного распределения					
3							
4							
5	N опыта	Случ. число	Количество	Частота	Значение	Вероятность значения	Накопл. Вероятность
6	1	0,282183			1	0,1000	0,1000
7	2	0,247125			2	0,1500	0,2500
8	3	0,005572			3	0,2500	=G7+F8
9	4	0,689705			4	0,2500	0,7500
10	5	0,818387			5	0,1500	0,9000
11	6	0,719916			6	0,1000	1,0000

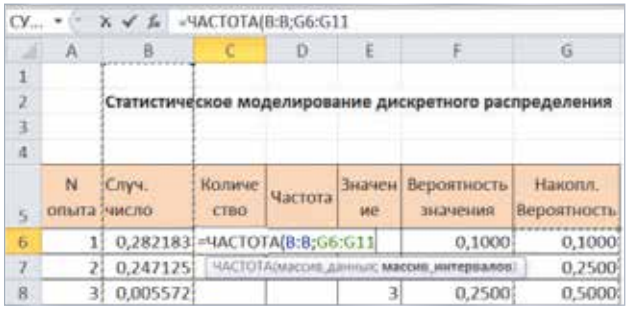
Рис. 7

Осталось подсчитать, сколько в столбце B случайных чисел, попавших в интервал от 0 до 0,1? в интервал от 0,1 до 0,25? и т.д. Задачу можно решить с помощью функции СЧЕТЕСЛИ, но удобнее использовать функцию ЧАСТОТА, которая сделает все сразу[1].

Эта функция является многозначной, поскольку возвращает не одно число, а сразу массив чисел. Многозначные функции создаются специальным образом.

Отведем столбец C (с 6 по 11 строку) для количества чисел, попавших в нужные нам шесть интервалов (рис. 8). Напишем в ячейке C6 формулу

=ЧАСТОТА(B:B;G6:G11).



	A	B	C	D	E	F	G
1							
2		Статистическое моделирование дискретного распределения					
3							
4							
5	N опыта	Случ. число	Количество	Частота	Значение	Вероятность значения	Накопл. Вероятность
6	1	0,282183	=ЧАСТОТА(B:B;G6:G11			0,1000	0,1000
7	2	0,247125					0,2500
8	3	0,005572			3	0,2500	0,5000

Рис. 8

[1] Внимание! Функция ЧАСТОТА вычисляет не частоту каждого значения, а количество значений, попавших в каждый интервал. Поэтому настоящие частоты придется вычислять в отдельном столбце (столбец D).

В результате в ячейке C6 окажется число, показывающее количество чисел в столбце B, которые меньше, чем число в ячейке G6, то есть 0,1. В нашем примере получилось 52 (рис. 9).

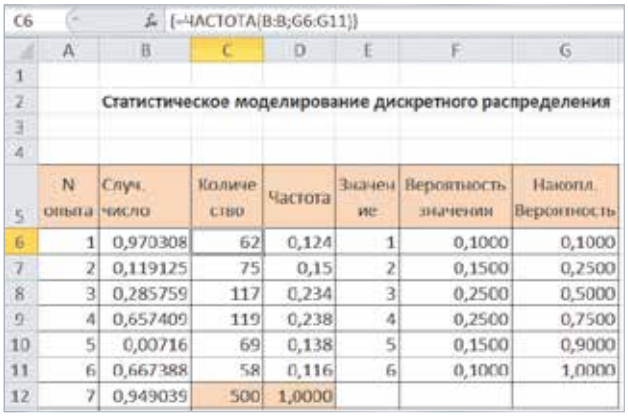
	N опыта	Случ. число	Количество	Частота	Значение	Вероятность значения	Накопл. Вероятность
6	1	0,846295	52		1	0,1000	0,1000
7	2	0,526649			2	0,1500	0,2500
8	3	0,325129			3	0,2500	0,5000
9	4	0,314892			4	0,2500	0,7500
10	5	0,606315			5	0,1500	0,9000
11	6	0,080356			6	0,1000	1,0000

Рис. 9

Теперь нужно распространить действие функции ЧАСТОТА на прочие интервалы. Выделим ячейку C6 и скопируем ее содержимое в буфер (комбинация Ctrl+C). Растянем выделенную область на ячейки от C6 до C11, нажмем F2 (сообщим Excel, что собираемся выделенные ячейки использовать как значения многозначной функции), нажмем комбинацию клавиш Ctrl+Shift+Enter.

В результате в ячейках C6–C11 будет подсчитано, сколько случайных чисел из столбца B попали в интервалы, правые границы которых указаны в ячейках G6–G11 (рис. 10).

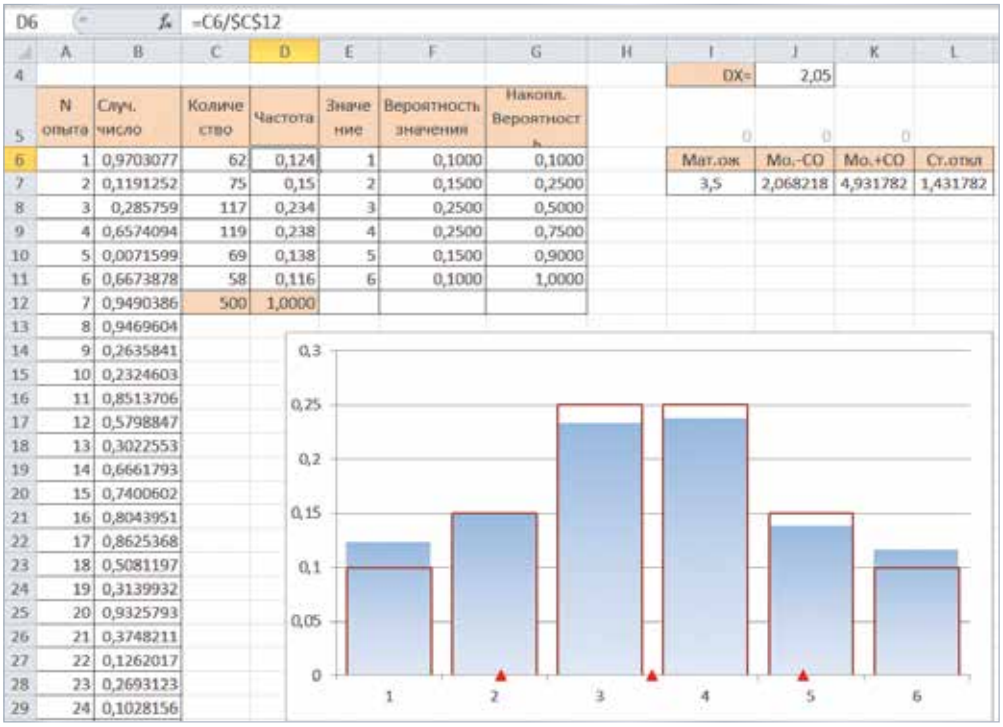
Комментарий. На рисунке 10 показан результат работы функции ЧАСТОТА в столбце C. Фигурные скобки в строке формул MS Excel поставил сам, когда понял, что функция введена как многозначная.



	A	B	C	D	E	F	G
1							
2		Статистическое моделирование дискретного распределения					
3							
4							
5	N опыта	Случ. число	Количество	Частота	Значение	Вероятность значения	Накопл. Вероятность
6	1	0,970308	62	0,124	1	0,1000	0,1000
7	2	0,119125	75	0,15	2	0,1500	0,2500
8	3	0,285759	117	0,234	3	0,2500	0,5000
9	4	0,657409	119	0,238	4	0,2500	0,7500
10	5	0,00716	69	0,138	5	0,1500	0,9000
11	6	0,667388	58	0,116	6	0,1000	1,0000
12	7	0,949039	500	1,0000			

Рис. 10

Теперь в столбец D поместим частоты — значения из столбца C, деленные на общее число опытов 500, а в строке 12 на всякий случай подсчитаем контрольные суммы. Общее количество значений равно 500, суммарная частота равна 1 (рис. 11). После этого на общее поле поместим диаграммы частот и вероятностей, построенные по числам из столбцов D и F.



N опыта	Случ. число	Количество	Частота	Значение	Вероятность значения	Накопл. Вероятность
1	0,9703077	62	0,124	1	0,1000	0,1000
2	0,1191252	75	0,15	2	0,1500	0,2500
3	0,285759	117	0,234	3	0,2500	0,5000
4	0,6574094	119	0,238	4	0,2500	0,7500
5	0,0071599	69	0,138	5	0,1500	0,9000
6	0,6673878	58	0,116	6	0,1000	1,0000
7	0,9490386	500	1,0000			
8	0,9469604					
9	0,2635841					
10	0,2324603					
11	0,8513706					
12	0,5798847					
13	0,3022553					
14	0,6661793					
15	0,7400602					
16	0,8043951					
17	0,8625368					
18	0,5081197					
19	0,3139932					
20	0,9325793					
21	0,3748211					
22	0,1262017					
23	0,2693123					
24	0,1028156					

Рис. 11

Для полноты картины в ячейках J3 и J4 вычислены математическое ожидание и дисперсия распределения R. Эти числа использованы для построения центра распределения R и точек $ER \pm \sqrt{DR}$ (красные треугольники на диаграмме).

Где же искомая выборка объемом 500? Она считывается из столбцов E и C. В выборке, показанной на рисунках 10 и 11, значение 1 случилось 62 раза, значение 2 — 75 раз и т.д.

Значение	1	2	3
Количество	62	75	117
Частота	0,124	0,15	0,234

Значение	4	5	6
Количество	119	69	58
Частота	0,238	0,138	0,116

Моделирование выборок из биномиального и геометрического распределений

Вместо абстрактного распределения R можно использовать любое дискретное распределение, если только мы умеем находить его вероятности.

Задача 4. Биномиальное распределение. Смоделировать 500 серий по $n = 50$ выстрелов в мишень, если вероятность попадания при каждом отдельном выстреле равна $p = 0{,}2$. Построить эмпирическое распределение частот и теоретическое распределение на одном поле.

При моделировании биномиального распределения

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

нужно учесть, что значения начинаются не с 1, а с 0. Придется чуть-чуть модифицировать построение (рис. 12). Вероятности в столбце F вычислены с помощью функции БИНОМРАСП. Необходимые параметры n (длина серии испытаний) и p (вероятность успеха) задаются в ячейках J2 и J3.

Напомним, что для случайной величины X, имеющей биномиальное распределение $Bi(n; p)$.

$$EX = np, \ DX = np(1-p).$$

Эти характеристики вычислены в ячейках M2 и M3.

Задача 5. Геометрическое распределение. Смоделировать 500 серий выстрелов до первого попадания в мишень, если вероятность попадания при каждом отдельном выстреле равна $p = 0{,}2$. Построить эмпирическое распределение частот и теоретическое распределение на одном поле.

Геометрическое распределение

$$P(X = k) = q^{k-1}p$$

бесконечно. В этом единственное отличие от задачи 3: в столбцах C–G придется взять *достаточно много значений* и их вероятностей для того, чтобы накопленная в столбце G вероятность *почти* достигла единицы. Если вероятность успеха $p = 0{,}2$, то в реальности число попыток пре-

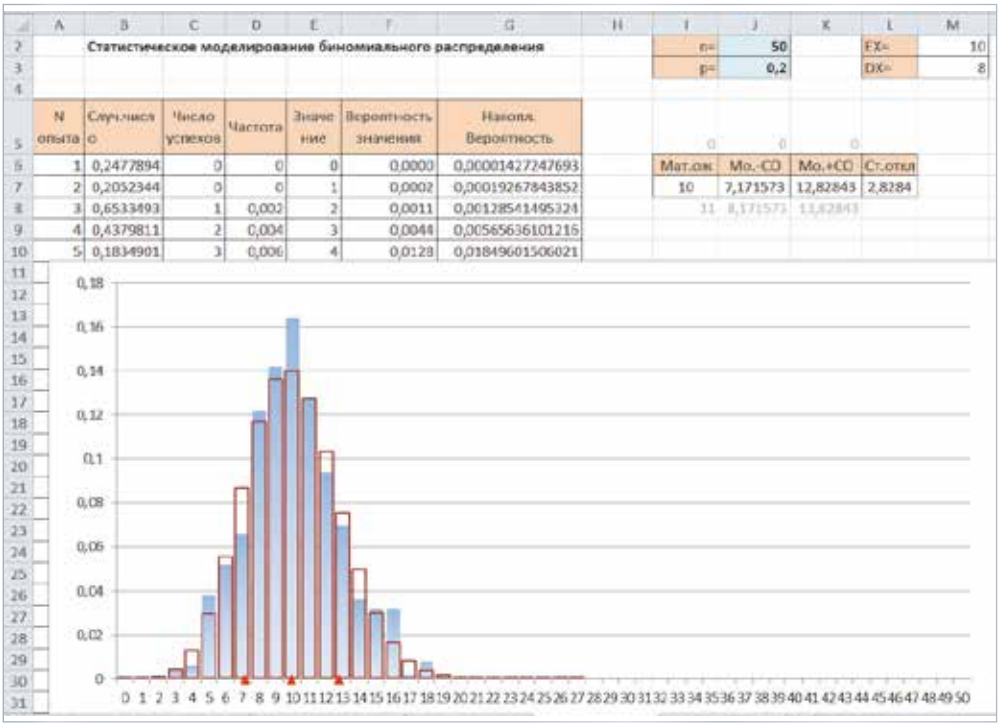


Рис. 12

взойдет 50 с пренебрежимо малой вероятностью. В столбце F вычислим вероятности, например, до $k = 100$ (с очень большим запасом), а строить диаграмму можно до $k = 50$ (рис. 13).

Напомним, что для случайной величины X, имеющей геометрическое распределение $G(p)$,

$\mathrm{E}X = \frac{1}{p}$, $\mathrm{D}X = \frac{1-p}{p^2}$ (ячейки M2 и M3).

Как получить нормальную выборку?

Задача 6. Нормальная выборка. Сделать для урока статистики правдоподобную выборку величины «рост шестиклассника (мальчика)» объемом 500, считая, что рост шестиклассников подчиняется приблизительно нормальному распределению.

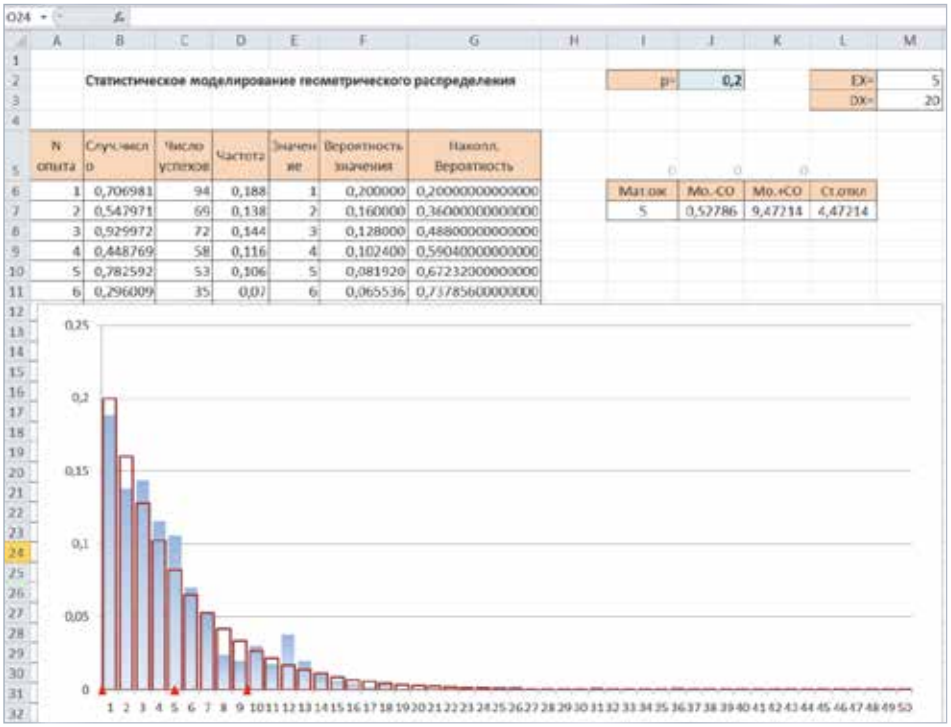


Рис. 13

Перцентиль	5	10	25	50	75	90	95
Рост (см)	142,9	145,8	150,5	156,5	161,8	167,0	169,8

Найти выборочное среднее, выборочное стандартное отклонение, проверить, близки ли эти характеристики к тем, что были заданы при моделировании.

Математически решение задачи мало отличается от решения задач 3–5. Нужно взять 500 случайных чисел и вычислить в каждом из них функцию, обратную функции нужного распределения. Для дискретных величин функцию распределения мы строили, накапливая вероятности, а потом «стреляли» случайными числами в множество ее значений. В непрерывном случае нам нужно сделать то же самое, но с функцией распределения, имеющий непрерывное множество значений. К сожалению, функция нормального распределения $N(\mu;\ \sigma^2)$ с математическим ожиданием μ и дисперсией σ^2 выглядит не очень симпатично:

$$F(x)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{x} e^{-\frac{(u-\mu)^2}{2\sigma^2}}\,du.$$

Она не является элементарной, а потому ее значения найти непросто, не говоря уже о значениях обратной. Но, как и всегда в таких случаях, в MS Excel есть готовое решение:

$$=\text{НОРМ.ОБР}(x,\mu,\sigma).$$

Чтобы воспользоваться этой функцией, нужно понять, чему равны средний рост шестиклассников μ и стандартное отклонение роста σ, то есть какое именно нормальное распределение нам нужно. Пока будем считать его *неизвестным*.

Определив шестиклассников как тринадцатилетних мальчиков из европейской части России, найдем в интернете *перцентильную*[2] таблицу[3] роста мальчиков в возрасте 13 лет в европейской части РФ по данным ВОЗ (см. вверху страницы).

Подзадача 1. С помощью таблицы узнать параметры неизвестного распределения: среднее и дисперсию.

Данных даже слишком много. Для простоты будем использовать не все, а только 5-й перцентиль и медиану (50-й перцентиль)[4]. Медиана нормального распределения совпадает с его математическим ожиданием. Поэтому считаем, что $\mu = 156{,}5$ (единицы измерения — сантиметры, мы это помним, но будем опускать для краткости и общности).

Осталось узнать стандартное отклонение σ. Используем стандартное нормальное распределение $N(0;\ 1)$, у которого стандартное отклонение равно 1. Для легкости восприятия нарисуем график его плотности (рис. 14) и посмотрим, чему у него равен 5-й перцентиль, то есть точка a, которая отсекает слева 5% площади, заключенной под графиком. Вся площадь под графиком равна 1, поэтому отсеченная площадь равна 0,05.

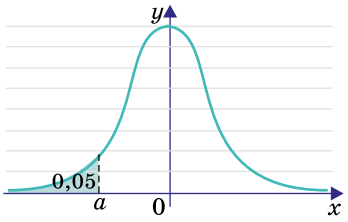


Рис. 14

Чтобы найти a, возьмем пустой лист MS Excel и воспользуемся функцией НОРМ.СТ.ОБР. В качестве аргумента нужно указать 0,05 (рис. 15).

Рис. 15

Получилось $a = -1{,}64485$. Это и есть 5%-я точка стандартного нормального распределения.

Дальше нужна линейная функция, которая переведет стандартное нормальное распределение в неизвестное. Ее нужно построить по двум найденным точкам: (–1,644 85; 142,9) и (0; 156,5). Угловой коэффициент будет равняться в точности σ, поскольку искомая функция $f(x) = kx + b$ обладает свойством

$$f(0) = m \text{ и } f(1) = m + \sigma,$$

и потому

$$b = m \text{ и } k = f(1) - f(0) = \sigma.$$

Найти угловой коэффициент прямой по двум точкам несложно с помощью карандаша и бумаги, но MS Excel и для этого предоставляет удобное средство. Стандартная функция ЛИНЕЙН тоже является многозначной, поскольку может возвращать сразу оба коэффициента k и b. Но нам требуется только k, поэтому сложный ввод не нужен.

[2] Перцентили (процентили, центили) — центральные меры массива данных или случайной величины. 50-й перцентиль — это медиана. k-й перцентиль можно определить как такое число, что хотя бы k% чисел массива не больше и хотя бы $(100 - k)$% чисел массива не меньше, чем это число.

[3] Разумеется, данные, опубликованные в разные годы и в разных источниках, могут различаться.

[4] Можно использовать все данные, но технически это сложнее, а результаты отличаться будут мало.

Аргументы функции ЛИНЕЙН[5]: сначала значения y, затем значения x (рис. 16).

Рис. 16

Подзадача 2. С помощью найденных оценок среднего и стандартного отклонений прежде неизвестного распределения создать выборку из 500 значений.

Пронумеруем наблюдения от 1-го до 500-го, например, в столбце G. Тогда в следующем столбце H во все ячейки впишем одну и ту же формулу НОРМ.ОБР (рис. 17).

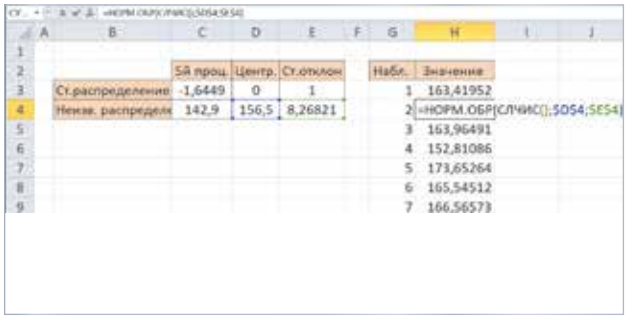


Рис. 17

Полученная выборка, как мы помним, динамическая — при любом изменении любой ячейки в таблице все значения генерируются еще раз. Чтобы избежать этого, понравившуюся выборку можно скопировать в чистый столбец, используя опцию «только значения» (рис. 18). Красным овалом обведен параметр «только значения» в меню *Вставить*. Числа, скопированные таким образом в столбец I, меняться уже не будут.

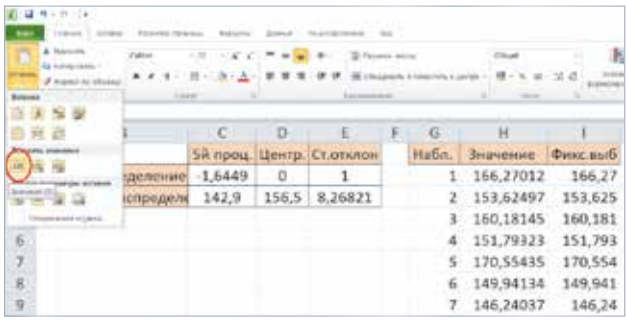


Рис. 18

[5] Если строить прямую не по двум, а по нескольким точкам, то получится прямая, построенная методом наименьших квадратов. Мы же используем функцию ЛИНЕЙН для построения прямой по двум точкам, поэтому получается прямая, проходящая в точности через эти две точки.

Подзадача 3. Найти выборочные характеристики и сравнить их с параметрами распределения, использованными при построении выборки.

Будем работать теперь только со столбцом I, в котором мы получили понравившуюся выборку[6] объемом 500 из нормального распределения со средним 156,5 и стандартным отклонением 8,268 21 (ячейки D4 и E4 на рисунке 16). Заведем две ячейки для выборочных характеристик C7 и C8 (рис. 19).

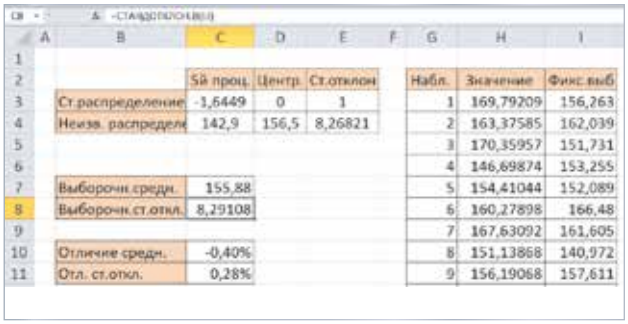


Рис. 19

Выборочное среднее — это просто среднее арифметическое всех значений в выборке. Оно является несмещенной[7] оценкой математического ожидания распределения, из которого сделана выборка, и в MS Excel используем функцию СРЗНАЧ.

Выборочное стандартное отклонение вычисляется как квадратный корень из *выборочной дисперсии*:

$$\sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(x_k-\bar{x}\right)^2}.$$

Обратите внимание на знаменатель. Он на единицу меньше, чем объем выборки. Смысл выборочной дисперсии в том, что она является несмещенной оценкой дисперсии распределения, из которого сделана выборка. Для вычисления выборочного стандартного отклонения можно извлечь корень из функции ДИСП.В или сразу использовать функцию СТАНДОТКЛОН.В.

Получилось выборочное среднее 155,88 и выборочное стандартное отклонение 8,29. Отличие от параметров распределения, по которому строилась выборка, составляет около 0,4% и 0,3% соответственно (ячейки C10 и C11). Можно с помощью подходящих теорем занудно проверять, достаточно ли малы такие отличия для выборки объемом 500. Но здравый смысл и так

[6] При написании статьи автору не удалось сохранить выборку, что получилась на рисунке 18. Пришлось фиксировать другую (рис. 19).

[7] Несмещенная оценка некоторой постоянной величины θ — случайная величина $\hat{\theta}$, математическое ожидание которой совпадает с θ, то есть $E\hat{\theta} = \theta$. Несмещенность является одним из желательных свойств оценок.

подсказывает, что отличия незначительны, поэтому выборка получилась, вероятно, неплохая.

Полученную неплохую выборку можно смело использовать на уроке в 7–8-х классах при изучении любых тем, связанных со средними, группировкой данных, рассеиванием, или в 11-м классе при изучении нормального распределения.

Задачи для самостоятельного решения

1. Постройте с помощью MS Excel случайную перестановку натуральных чисел от 1 до 100.

2. Смоделируйте 40 последовательных бросаний правильного октаэдра с пронумерованными гранями. Постройте последовательность случившихся значений, их сумм и их среднего арифметического.

Комментарий. Удобно использовать или модифицировать электронную таблицу, показанную на рисунке 4.

3. Распределению Пуассона подчиняется количество событий, наступивших в течение единицы времени в случайные моменты времени по одиночке и независимо друг от друга. Предположим, мы изучаем число покупателей в модном бутике. Известно, что в воскресенье с 12.00 до 16.00 в бутике в среднем бывает 6 покупателей. Модифицируйте таблицу, показанную на рисунке 12, чтобы получить выборку из 500 значений случайной величины «число покупателей в указанный период».

Комментарий. Обозначение распределения: P(λ), где параметр λ равен среднему числу событий в единицу времени. Используйте функцию ПУАССОН или ПУАССОН.РАСП.

4. *Гипергеометрическому распределению* подчиняется количество успешных объектов, которые оказались среди n объектов, случайным образом извлеченных из совокупности, в которой ровно K успешных и ровно $(N-K)$ неуспешных объектов. Можно считать, что объекты извлекаются одновременно или что они извлекаются по одному, но без возвращения. Предположим, что вы для урока статистики или теории вероятностей хотите смоделировать выбор 15 фломастеров из ящика, в котором 30 красных и 30 синих фломастеров. Модифицируйте таблицу, показанную на рисунке 11, чтобы получить выборку из 300 значений случайной величины «число красных среди вынутых 15 фломастеров».

Комментарий. Обозначение распределения: $H(n; K; N)$, где n — число извлеченных объектов, K — число успешных объектов в совокупности, N — общее число объектов в совокупности. Используйте функцию ГИПЕРГЕОМЕТ или ГИПЕРГЕОМ.РАСП.

5. Найдите в интернете перцентильную таблицу данных о нормальном артериальном давлении у девочек в возрасте 10 лет и смоделируйте измерения систолического (или диастолического) давления у выборки объемом 400 из такой совокупности девочек.

Комментарий. Для решения удобно модифицировать таблицу, показанную на рисунках 17–19.

Ответы к задачам из статьи «Задача кассира метро и формула Муавра–Стирлинга»

1. Из приближения находим, что отношение этих вероятностей близко к $\sqrt{2}$.

2. *Вывод.* В каждом столбике нужно выбрать k позиций, где в первом столбике орел, а во втором решка. Из оставшихся $(n-k)$ позиций нужно выбрать еще k позиций, где, напротив, в первом столбике решка, а во втором орел. Оставшиеся $(n-k)$ позиций будут заполнены парами орел-орел или решка-решка. Вероятность такой комбинации при каждом $k \le \frac{n}{2}$ равна

$$C_n^k (pq)^k \cdot C_{n-k}^k (qp)^k \left(p^2+q^2\right)^{n-2k}.$$

Осталось просуммировать вероятности этих несовместных событий по всем целым k от 0 до $\frac{n}{2}$.

3. *Решение.* Рассмотрим произведение под знаком предела и дважды умножим числитель и знаменатель на все четные числа от 2 до $2n$:

$$\frac{1\cdot 3}{4\cdot 1^2}\cdot\frac{3\cdot 5}{4\cdot 2^2}\cdot\frac{5\cdot 7}{4\cdot 3^2}\cdot\ldots\cdot\frac{(2n-1)(2n+1)}{4n^2}=$$

$$=\frac{1\cdot 2\cdot 2\cdot 3}{4\cdot 1^2\cdot 2^2}\cdot\frac{3\cdot 4\cdot 4\cdot 5}{4\cdot 2^2\cdot 4^2}\cdot\frac{5\cdot 6\cdot 6\cdot 7}{4\cdot 3^2\cdot 6^2}\cdot\ldots\cdot\frac{(2n-1)\cdot 2n\cdot 2n(2n+1)}{4n^2(2n)^2}=$$

$$=\frac{\left(1\cdot 2\cdot 3\cdot\ldots\cdot(2n-1)\cdot 2n\right)\cdot\left(2\cdot 3\cdot 4\cdot\ldots\cdot 2n\right)\cdot(2n+1)}{4^n\cdot(2\cdot 3\cdot 4\cdot\ldots\cdot n)^2\cdot(2\cdot 4\cdot 6\cdot\ldots\cdot 2n)^2}=$$

$$=\frac{(2n)!\cdot(2n)!\cdot(2n+1)}{4^n\cdot(n!)^2\cdot 4^n(1\cdot 2\cdot 3\cdot\ldots\cdot n)^2}=$$

$$=\frac{(2n)!\cdot(2n)!\cdot(2n+1)}{4^n\cdot(n!)^2\cdot 4^n(n!)^2}=\left(\frac{(2n)!\cdot\sqrt{2n+1}}{4^n(n!)^2}\right)^2.$$

Тогда произведение принимает вид:

$$\frac{\pi}{2}\lim_{n\to\infty}\left(\frac{(2n)!\cdot\sqrt{2n+1}}{4^n(n!)^2}\right)^2=1.$$

Осталось умножить обе части на $\frac{2}{\pi}$ и извлечь квадратный корень.