

Санкт-Петербургский Национальный
Исследовательский Университет Информационных
технологий, механики и оптики

Лабораторная работа #5

Генерация случайных чисел и анализ выборки данных

Выполнил: Канева
Тамара Игоревна
Группа № К3121
Проверила: Казанова
Полина Петровна

Санкт-Петербург
2021

Цель работы:

Изучить средства программы Microsoft Excel для генерации случайных чисел с требуемыми законами распределения, для построения и анализа выборок данных.

Задачи:

Изучить способы генерации случайных чисел, построения выборки данных, анализа данных (построение гистограмм, методы подсчета показателей уровня, показателей рассеивания, показателей асимметрии, использование описательной статистики).

Ход работы:

Построение распределений.

Распределение Бернулли.

Чтобы сгенерировать столбец из 100 случайных чисел, воспользуемся функцией Microsoft Excel “Генерация случайных чисел”. Для этого перейдем во вкладку “Данные” и нажмем на кнопку “Анализ данных”. В открывшемся списке выберем “Генерация случайных чисел” и увидим одноименное окно (рис. 1). В соответствующие поля введем число переменных, равное 1, число случайных чисел, равное 100, выберем распределение Бернулли и значение p , равное 0.3.

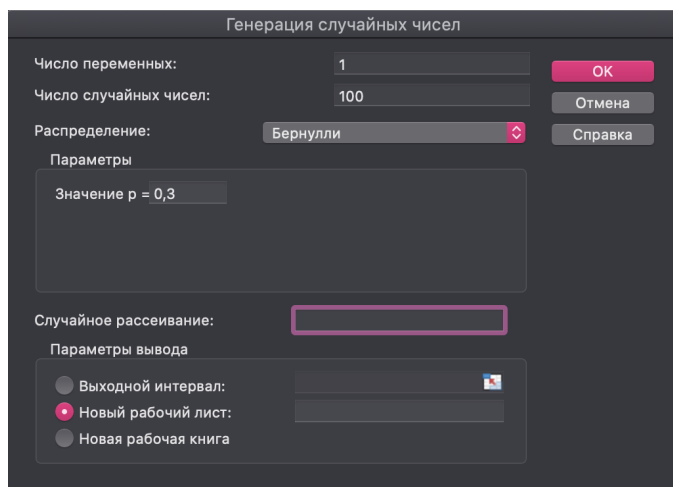


Рис. 1. Окно “Генерация случайных чисел”.

Получим столбец случайных нулей и единиц (рис. 2).

40	1
41	0
42	1
43	0
44	0
45	0
46	1
47	0
48	0
49	0
50	0

Рис. 2. Срез полученного с помощью распределения Бернулли массива.

Биномиальное распределение.

Выполняя аналогичные действия, получим массив чисел с помощью биномиального распределения. В окне “Генерация случайных чисел” выбираем биномиальное распределение, вводим значение p , равное 0.85, а число испытаний - 25. Остальные поля заполняются так же, как и в предыдущем случае. Получим столбец случайных чисел (рис. 3).

40	23
41	21
42	22
43	21
44	23
45	20
46	18
47	21
48	20
49	20
50	21

Рис. 3. Срез полученного с помощью биномиального распределения массива.

Нормальное распределение.

Выполняя аналогичные действия, получим массив чисел с помощью нормального распределения. В окне “Генерация случайных чисел” выбираем нормальное распределение, вводим значение стандартного отклонения, равное 20, а среднее - 100. Остальные поля (если они есть) заполняются так же, как и в предыдущем случае. Получим столбец случайных чисел (рис. 4).

40	102,611841
41	107,01707
42	104,719777
43	85,906652
44	84,7020717
45	134,429104
46	89,1300376
47	69,6495252
48	81,6752279
49	95,562257
50	121,936512

Рис. 4. Срез полученного с помощью нормального распределения массива.

Построение выборок.

Распределение Бернулли.

Получим случайную выборку размером 20 на основе массива, полученного с помощью распределения Бернулли. Для этого перейдем нажмём на “Анализ данных” и выберем “Выборка”, в результате чего перед нами откроется диалоговое окно (рис. 5). В качестве входного интервала выберем все ячейки, содержащие элементы массива. Метод выборки - случайный. Число выборок равно размеру выборки, то есть 20. После ввода всех данных получим столбец из 20 значений, который и является нашей выборкой (рис. 6).

Выборка

Входные данные

Входной интервал:

☐ Метки

Метод выборки

☐ Периодический

Период:

☒ Случайный

Число выборок:

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

OK

Отмена

Справка

Рис. 5. Окно "Выборка".

Построим систематическую выборку размером 20 на основе массива, полученного с помощью распределения Бернулли. В качестве входного интервала также выберем все ячейки, содержащие элементы массива. Метод выборки - периодический. Чтобы получить выборку размером 20 элементов, нужно поделить 100 (размер массива) нацело на 20 (размер выборки), то есть период будет равным 5. После ввода всех данных получим столбец из 20 значений, который и является нашей выборкой (рис. 6).

0	0
1	0
0	0
0	0
0	0
0	0
0	0
0	0
0	1
1	0
0	0
0	0
1	1
0	0
0	1
0	0
0	1
0	0
1	1
1	0
0	0
случайная	систематич

Рис. 6. Полученные для массива, построенного на основе распределения Бернулли, случайная и систематическая выборки.

Биномиальное распределение.

Выполняя абсолютно такие же действия, получим случайную и систематическую выборки для массива, построенного на основе биномиального распределения (рис. 7).

23	21
20	25
18	23
22	18
18	21
22	22
23	23
23	23
22	20
22	21
22	24
20	21
18	23
22	19
21	21
20	21
22	21
18	19
21	22
22	17
случайная	систематич

Рис. 7. Полученные для массива, построенного на основе биномиального распределения, случайная и систематическая выборки.

Нормальное распределение.

Выполняя абсолютно такие же действия, получим случайную и систематическую выборки для массива, построенного на основе нормального распределения (рис. 8).

130,024057	124,928659
100,674459	85,3508143
110,429358	99,2806352
78,1043698	103,926789
78,9317371	99,7271061
88,6974365	116,918102
97,791906	64,4729542
113,55138	102,611841
98,6297553	134,429104
99,2069661	121,936512
55,6763214	107,599647
89,1300376	160,829916
84,7020717	90,2152467
116,743752	112,494002
107,01707	110,429358
82,3469352	98,704061
94,1435362	91,765867
93,9716872	103,657965
91,765867	110,399935
88,6974365	124,176779
случайная	систематич

Рис. 8. Полученные для массива, построенного на основе нормального распределения, случайная и систематическая выборки.

Построение гистограмм.

Построим гистограмму для массива, построенного на основе нормального распределения. Для этого перейдем нажмём на “Анализ данных” и выберем “Гистограмма”, в результате чего перед нами откроется диалоговое окно (рис. 9). Не забыв поставив флажок “Вывод графика”, нажимаем “ОК”. Мы видим, во-первых, таблицу с полученными данными из исходного массива (граница кармана и частота значений, которые встречаются в исходном массиве и лежат в этом кармане), а во-вторых, саму гистограмму, являющуюся, фактически, графиком, отображающим полученную таблицу (рис. 10).

Гистограмма

Входные данные

Входной интервал:

Интервал карманов:

☐ Метки

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

☐ Парето (отсортированная гистограмма)

☐ Интегральный процент

☒ Вывод графика

Рис. 9. Окно "Гистограмма".

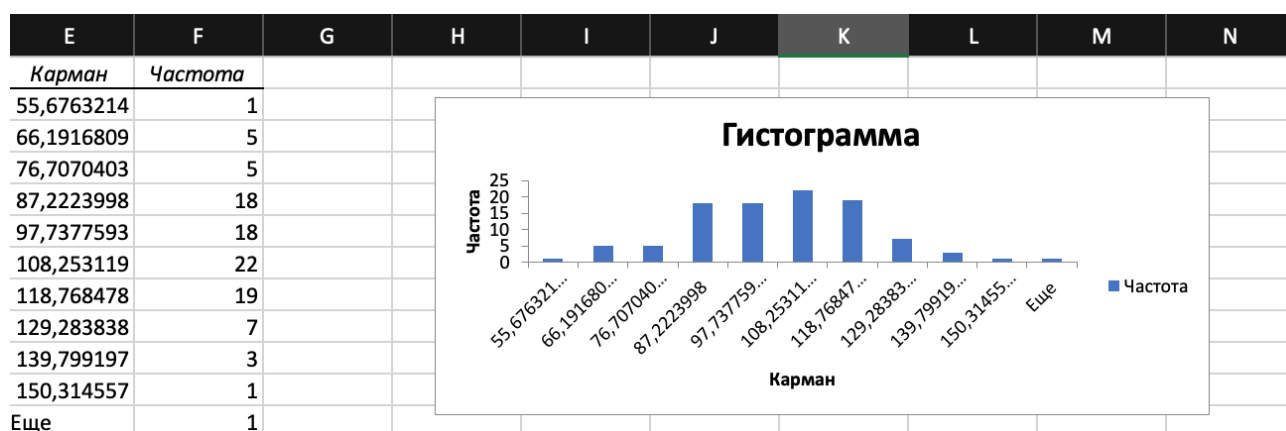


Рис. 10. Гистограмма для массива, построенного на основе нормального распределения.

Полностью аналогично построим гистограммы для случайной (рис. 11) и систематической выборок (рис. 12).

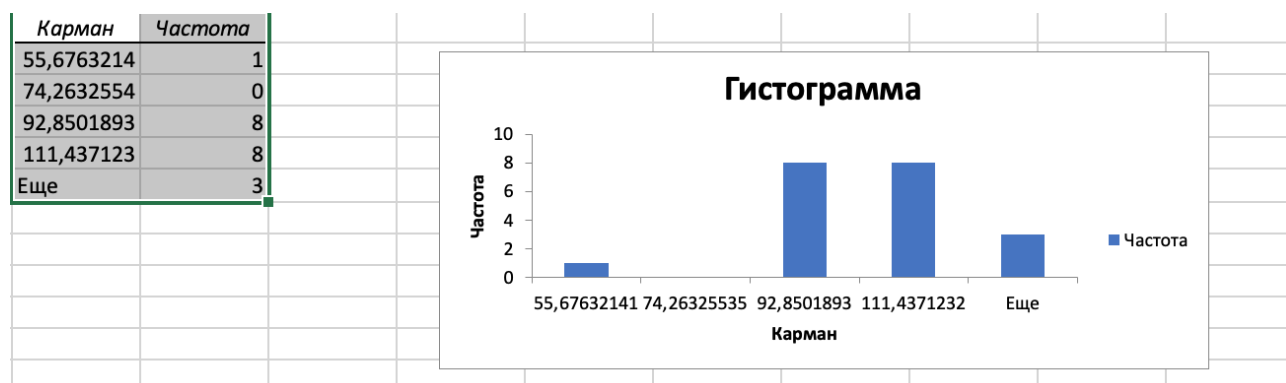


Рис. 11. Гистограмма для случайной выборки, полученной для массива, построенного на основе нормального распределения.

Карман	Частота
64,4729542	1
88,5621946	1
112,651435	12
136,740676	5
Еще	1



Рис. 12. Гистограмма для систематической выборки, полученной для массива, построенного на основе нормального распределения.

Заметим, что гистограмма для исходного массива выглядит более гладкой и равномерной, нежели гистограммы для выборок. Более того, гистограмма для систематической выборки больше похожа на гистограмму для исходного массива, чем гистограмма для случайной выборки, с точки зрения возрастания и убывания функции.

Расчет показателей анализа данных.

Рассчитаем среднее арифметическое для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “СРЗНАЧ”.

Рассчитаем среднее геометрическое для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “СРГЕОМ”.

Рассчитаем медиану для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “МЕДИАНА”.

Рассчитаем моду для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “МОДА”.

Рассчитаем минимум для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “МИН”.

Рассчитаем максимум для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “МАКС”.

Рассчитаем ранг и перцентиль для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “Ранг и перцентиль”, вызвав ее в окне диалога “Анализ данных”. В открывшемся окне “Ранг и перцентиль” введем характеристики представления нашей выборки (рис. 13) и нажмём “ОК”.

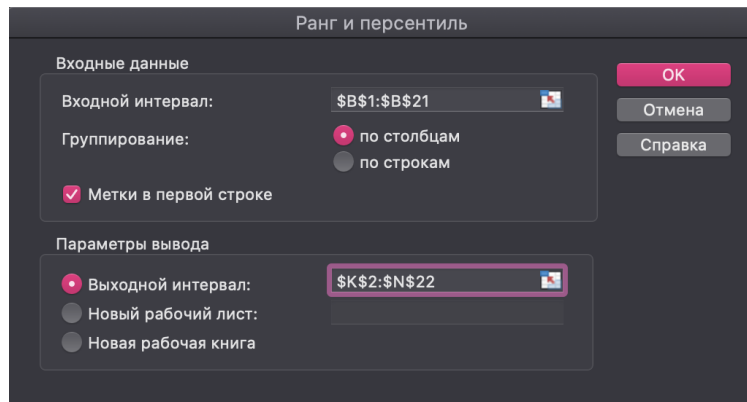


Рис. 13. Окно “Ранг и перцентиль”.

Рассчитаем размах для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого вычтем из значения максимума значение минимума.

Рассчитаем среднее линейное отклонение для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “СРОТКЛ”.

Рассчитаем среднее квадратическое отклонение для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “СТАНДОТКЛОН”.

Рассчитаем дисперсию для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “ДИСП”.

Рассчитаем эксцесс для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “ЭКСЦЕСС”.

Рассчитаем асимметрию для случайной выборки, полученной для массива построенного на основе распределения Бернулли. Для этого воспользуемся функцией “СКОС”.

Рассчитаем все те же показатели для случайной выборки, полученной для массива построенного на основе биномиального распределения и для случайной выборки, полученной для массива построенного на основе нормального распределения. Результаты занесем в таблицы (рис. 14 - 16).

Функция	Значение	Ранг и перцентиль для случайной			
		Точка	случайная	Ранг	Процент
Ср. арифметическое	0,25				
Ср. геометрическое	0	2	1	1	78,90%
Медиана	0	9	1	1	78,90%
Мода	0	12	1	1	78,90%
Минимум	0	18	1	1	78,90%
Максимум	1	19	1	1	78,90%
Размах выборки	1	1	0	6	0,00%
Ср. лин. отклонение	0,375	3	0	6	0,00%
Ср. кв. отклонение	0,444261658	4	0	6	0,00%
Дисперсия	0,197368421	5	0	6	0,00%
Экссесс	-0,496732026	6	0	6	0,00%
Ассиметрия	1,250514297	7	0	6	0,00%
		8	0	6	0,00%
		10	0	6	0,00%
		11	0	6	0,00%
		13	0	6	0,00%
		14	0	6	0,00%
		15	0	6	0,00%
		16	0	6	0,00%
		17	0	6	0,00%
		20	0	6	0,00%

Рис. 14. Показатели анализа данных для случайной выборки, полученной для массива построенного на основе распределения Бернулли.

Функция	Значение	Ранг и перцентиль для случайной			
		Точка	случайная	Ранг	Процент
Ср. арифметическое	20,95				
Ср. геометрическое	20,87657152	1	23	1	89,40%
Медиана	22	7	23	1	89,40%
Мода	22	8	23	1	89,40%
Минимум	18	4	22	4	47,30%
Максимум	23	6	22	4	47,30%
Размах выборки	5	9	22	4	47,30%
Ср. лин. отклонение	1,465	10	22	4	47,30%
Ср. кв. отклонение	1,761428846	11	22	4	47,30%
Дисперсия	3,102631579	14	22	4	47,30%
Экссесс	-0,776974966	17	22	4	47,30%
Ассиметрия	-0,750701366	20	22	4	47,30%
		15	21	12	36,80%
		19	21	12	36,80%
		2	20	14	21,00%
		12	20	14	21,00%
		16	20	14	21,00%
		3	18	17	0,00%
		5	18	17	0,00%
		13	18	17	0,00%
		18	18	17	0,00%

Рис. 15. Показатели анализа данных для случайной выборки, полученной для массива построенного на основе биномиального распределения.

Функция	Значение	Ранг и перцентиль для случайной			
		Точка	случайная	Ранг	Процент
Ср. арифметическое	95,01180696				
Ср. геометрическое	93,61828529	1	130,024057	1	100,00%
Медиана	94,05761173	14	116,743752	2	94,70%
Мода	88,69743649	8	113,55138	3	89,40%
Минимум	55,67632141	3	110,429358	4	84,20%
Максимум	130,0240572	15	107,01707	5	78,90%
Размах выборки	74,34773579	2	100,674459	6	73,60%
Ср. лин. отклонение	11,89624402	10	99,2069661	7	68,40%
Ср. кв. отклонение	16,18848082	9	98,6297553	8	63,10%
Дисперсия	262,0669111	7	97,791906	9	57,80%
Экссесс	1,201347766	17	94,1435362	10	52,60%
Ассиметрия	-0,116464876	18	93,9716872	11	47,30%
		19	91,765867	12	42,10%
		12	89,1300376	13	36,80%
		6	88,6974365	14	26,30%
		20	88,6974365	14	26,30%
		13	84,7020717	16	21,00%
		16	82,3469352	17	15,70%
		5	78,9317371	18	10,50%
		4	78,1043698	19	5,20%
		11	55,6763214	20	0,00%

Рис. 16. Показатели анализа данных для случайной выборки, полученной для массива построенного на основе нормального распределения.

Расчет показателей анализа данных с помощью инструмента “Описательная статистика”.

Получим статистический отчет для случайной выборки, полученной для массива построенного на основе распределения Бернулли, одновременно по всем основным показателям, воспользовавшись инструментом “Описательная статистика”, который находится в надстройке “Анализ данных”. При вызове функции “Описательная статистика” увидим одноименное диалоговое окно (рис. 17), в которое введем параметры нашей выборки.

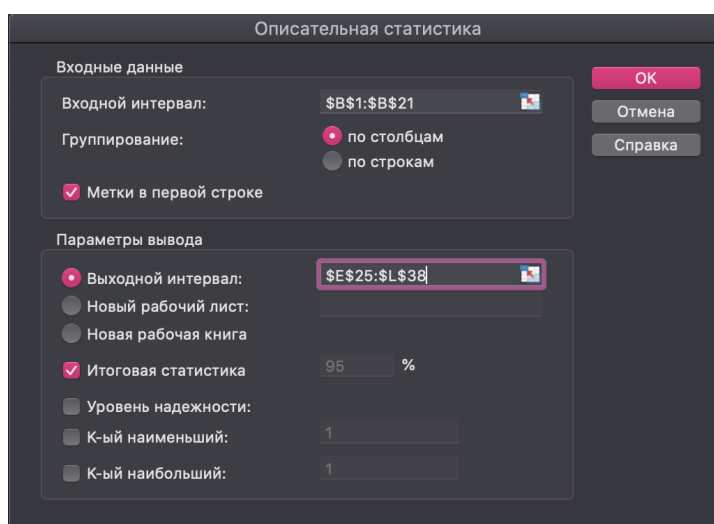


Рис. 17. Окно “Описательная статистика”.

После нажатия кнопки “ОК”, получим результат, внесенный в таблицу (рис. 18).

Описательная статистика	
Среднее	0,25
Стандартная ошибка	0,09933993
Медиана	0
Мода	0
Стандартное отклонение	0,44426166
Дисперсия выборки	0,19736842
Эксцесс	-0,496732
Асимметричность	1,2505143
Интервал	1
Минимум	0
Максимум	1
Сумма	5
Счет	20

Рис. 18. Показатели анализа данных, рассчитанные с помощью инструмента “Описательная статистика”, для случайной выборки, полученной для массива построенного на основе распределения Бернулли.

Полностью аналогично рассчитаем показатели анализа данных для случайной выборки, полученной для массива построенного на основе биномиального распределения (рис. 19) и случайной выборки, полученной для массива построенного на основе нормального распределения (рис. 20).

Заметим, что результаты, полученные с помощью инструмента “Описательная статистика” и без его использования, совпадают. Этот инструмент может помочь

сэкономить много времени и кликов мышкой. Тем не менее, он неудобен при необходимости подсчитать лишь один - два показателя.

Описательная статистика	
Среднее	20,95
Стандартная ошибка	0,39386746
Медиана	22
Мода	22
Стандартное отклонение	1,76142885
Дисперсия выборки	3,10263158
Эксцесс	-0,776975
Асимметричность	-0,7507014
Интервал	5
Минимум	18
Максимум	23
Сумма	419
Счет	20

Рис. 19. Показатели анализа данных, рассчитанные с помощью инструмента “Описательная статистика”, для случайной выборки, полученной для массива построенного на основе биномиального распределения.

Описательная статистика	
Среднее	95,011807
Стандартная ошибка	3,61985436
Медиана	94,0576117
Мода	88,6974365
Стандартное отклонение	16,1884808
Дисперсия выборки	262,066911
Эксцесс	1,20134777
Асимметричность	-0,1164649
Интервал	74,3477358
Минимум	55,6763214
Максимум	130,024057
Сумма	1900,23614
Счет	20

Рис. 20. Показатели анализа данных, рассчитанные с помощью инструмента “Описательная статистика”, для случайной выборки, полученной для массива построенного на основе нормального распределения.

Вывод:

В ходе этой лабораторной работы мы узнали и применили различные способы генерации случайных чисел, построения выборки данных, изучили основные показатели анализа данных (построение гистограмм, методы подсчета показателей уровня, показателей рассеивания, показателей асимметрии, использование описательной статистики). Эти навыки являются базовыми при подготовке данных к анализу и являются первой ступенью в изучении дисциплин, связанных с большими объемами данных.

Ответы на контрольные вопросы:

1. Законом распределения случайной величины называется всякое соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями.

2. Среди примеров, демонстрирующих закон нормального распределения можно привести следующий опыт. Необходимо выбрать некоторый интервал времени, например, 5 секунд и попытаться его измерить с помощью секундомера. Удивительно, но полученные результаты будут подчиняться этому закону, причем среднее будет равно 100, а стандартное отклонение зависит от секундомера и человека, снимающего измерения. Также нормальному распределению подчинен человеческие рост и вес.
3. Асимметрия характеризует меру несимметричности или скошенности распределения. Если коэффициент асимметрии больше нуля, то асимметрия правосторонняя, если меньше нуля – левосторонняя. Эксцесс характеризует островершинность или же плосковершинность распределения. Если эксцесс больше нуля, то распределение островершинное, если меньше нуля – плосковершинное.
4. Сначала рассчитываются значения интервалов, каждый из которых по своей длине равен остальным. Количество интервалов определяется как натуральное число, наиболее близкое к корню из количества испытаний. Нижний конец первого интервала соответствует минимуму массива, а верхний конец последнего интервала - максимум. Высота столбцов гистограммы определяется количеством значений массива, попадающих в каждый интервал. С помощью средств Microsoft Excel гистограмму можно построить почти полностью автоматически.
5. Если случайная выборка строится таким образом, чтобы каждый объект генеральной совокупности имел одинаковую вероятность быть выбранным, и при этом объекты отбираются независимо друг от друга, то для получения систематической выборки в генеральной совокупности определяют случайную начальную точку и отбирают элементы, начиная с этой точки через постоянный интервал.