Лекция №15. Еще раз про случайные числа

Продолжим говорить про случайные числа. Ведь самое интересное в них и самые интересные их виды все еще не изучены. В предыдущих сериях: обсудили дискретные слкучайные числа, подумали о принципах их формирования и об аналогиях в реальном мире. Сегодня же поговорим про непрерывно распределенные случайные числа.

Если представить себе дискретное случайное число несложно, то вот с непрерывными несколько сложнее. Но начнем. Если есть случайное число, дающее два значения: 1 и 2, то это явно монета: орел и решка; четыре значения — игральная кость в форме тетраэдра; шесть — обычная кубическая игральная кость; 12 — игральная кость в форме додекаэдра; 20 — игральная кость в форме икосаэдра. И так далее. Кто мешает сделать игральную кость с еще большим количеством граней? Да, собственно, никто. Соответственно, некой моделью равномерно распределенного случайного числа может быть, например, цилиндр, который мы катаем в вязкой среде (чтобы движение не было бесконечным), а результат — угол между вертикалью и неким заранее отмеченным «нулевым углом» на торцах цилиндра. То есть это формально игральная кость с бесконечным количеством граней, и вероятность выпадения каждой (одного точно заданного числа) соответственно ноль. Но тем не менее, граней бесконечно много, и какая-то одна точно выпадет, так что суммарная вероятность выпадения все еще 1.

Теперь про интересное. Возьмем, что наше число (угол) распределено в диапазоне от [0;1) (0 и 1 — совпадение метки с вертикалью). Теперь мысленно покрасим цилиндр в два цвета — по разные стороны от метки. В таком случае при прокате, очевидно, что выпадение одного цвета и другого цвета равновероятны. И так работает с любыми интервалами. В нашем случае цилиндр — модель равномерно распределенного непрерывного случайного числа. И для таких чисел вероятность выпадения числа из какого-то диапазона равна отношению ширины этого диапазона к ширине всего диапазона. В нашем случае — обе вероятности по $\frac{1}{2}$. Если раскрасить торец в несколько цветов, и секции будут заданной величины, то можно превратить цилиндр в модель любой другой игральной кости с любым ограниченным количеством

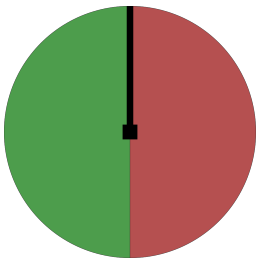

Рис. 1: Торец цилиндра

граней.

И тут встает вопрос: а все ли случайные числа распределены равномерно (вероятность выпадения равна отношению ширин интервалов)? Вопрос второй: как проверить, равномерно ли распределение?

Оба вопроса имеют очевидные ответы: 1. нет; 2. построить его. Построить можно график распределения плотности вероятности, ну точнее попытаться. Можно делить диапазон случайной величины на участки и разбрасывать множество случайных чисел. И считать количество попавших в каждый участок.

В качестве иллюстрации подобного принципа можно привести известное устройство — доску Гальтона тыц.

Так вот, я хотел бы на сегодня рассказать, как построить распределение плотности вероятности случайного числа.

Но для начала нам нужно получить непрерывное случайное число. Можно, конечно, воспользоваться готовыми генераторами, например вихрем Мерсенна mt19937. А можно и использовать встроенный генератор rand(). Для этого можно написать небольшую обертку для него.

```
double rnd() {
  return rand() / ((double) RAND_MAX + 1);
}
```

RAND_MAX это константа из библиотеки. Она определена заранее. Здесь стандартная функция дает числа в диапазоне от 0 до RAND_MAX. Для хорошей работы фмнальное случайное число никогда не должно достигать значения 1, поэтому добавляем единицу к RAND_MAX. Со способом записи знаменателя будет связан контрольный вопрос с подвохом.

Теперь есть случайное число с достаточным количеством вариантов. Нужно научиться строить распределение.

1. Нужно сгенерировать случайное число;
2. На основе числа узнать ячейку в распределении, в которую оно попадает;
3. Добавить единицу к числу попавших чисел в эту ячейку.

Для этого, во-первых, понадобится массив. Если мы хотим узнать распределение числа достаточно точно, то число ячеек в массиве должно быть относительно велико. Например, 100. Слишком много ячеек в этом массиве делать вредно — для получения показательной статистики количество запусков случайного числа растет пропорционально размеру этого массива. В качестве отправной точки количества будем считать, что в среднем в каждую ячейку массива должно попасть не менее 1000 чисел, лучше 10000.

Таким образом, количество точек для массива размером 100 это по крайней мере 1000000. Больше — лучше. Во-вторых надо научиться делить число на интервалы. То есть если у нас N интервалов в массиве, то нужно получить границы каждого интервала.

НЕ нужно писать в коде 100 условий. Это отвратительно. При смене размера массива количество условий тоже меняется.

Поэтому предложу два варианта.

1. Условие в цикле, цикл проходит по интервалам, по номеру интервала получаем его границы и спрашиваем, попало ли число в этот интервал, при попадании выходим из цикла;

2. На основе сгенерированного числа сразу получаем номер ячейки в массиве.

Первый способ прост, но нужно получать границы интервалов. Они выглядят так:

$$a_n = n\frac{b-a}{N}; \quad b_n = (n+1)\frac{b-a}{N}; \quad n \in [0:N). \tag{1}$$

Для второго способа нужно, как минимум, знать высшую математику, высшую магию и высшее программирование. Но все просто: число нормируем на интервал [0:1). Потом умножаем случайное число на количество ячеек в массиве и берем от получившегося целую часть. Это значение и есть искомый номер ячейки в выходном массиве. Подробнее объяснять не буду.

А теперь примеры.

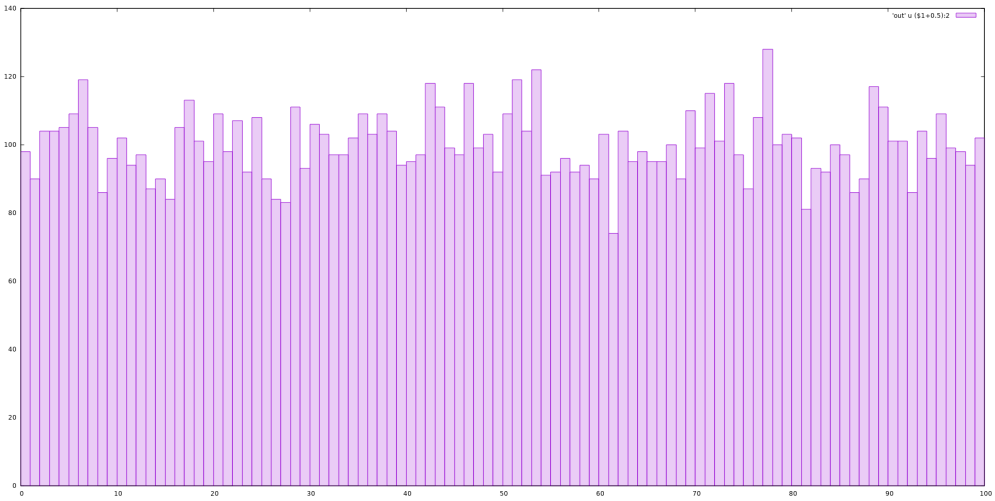

Рис. 2: Равномерно распределенное число. 100 точек на ячейку

Рапределение на последних изображениях называется биномиальным и является аппроксимацией нормального распределения.

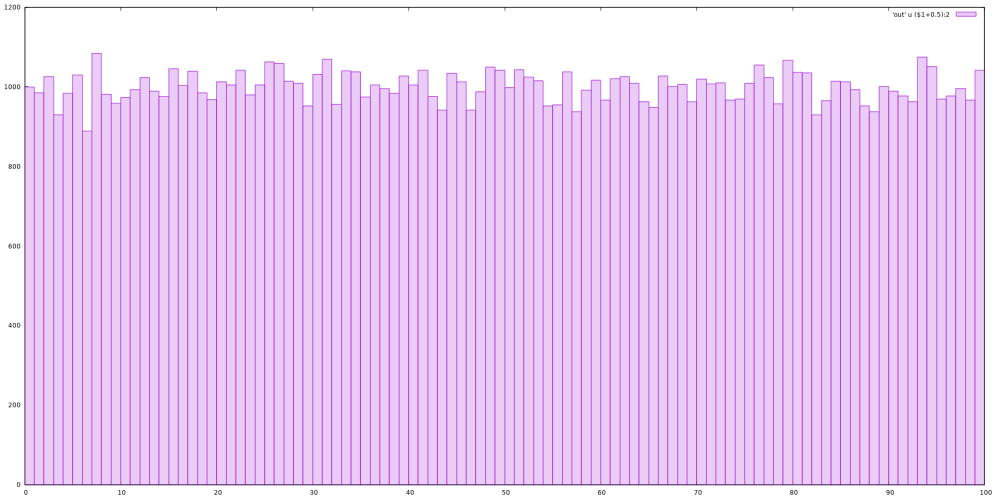

Рис. 3: Равномерно распределенное число. 1000 точек на ячейку

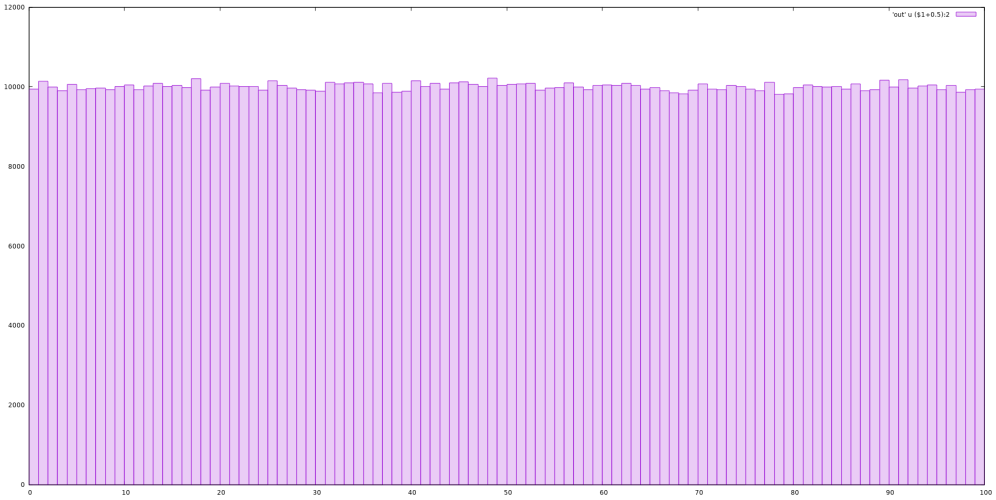

Рис. 4: Равномерно распределенное число. 10000 точек на ячейку

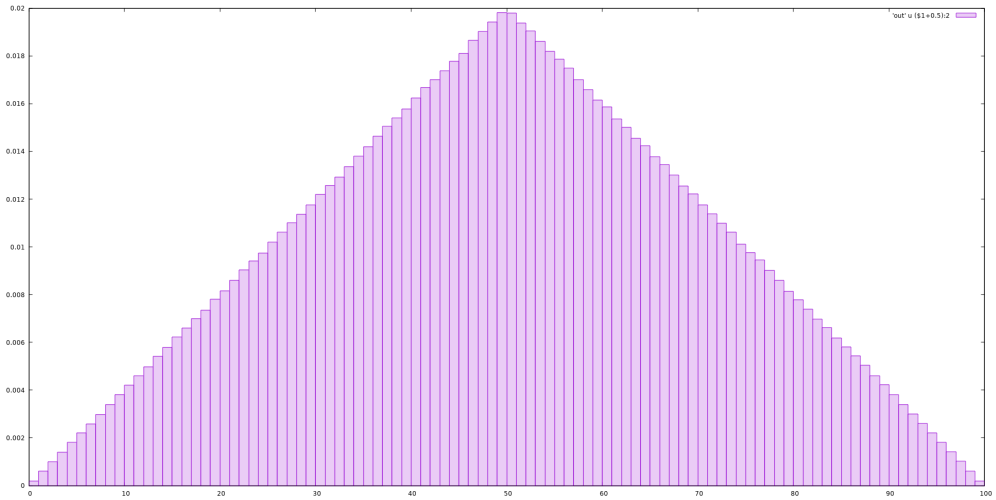

Рис. 5: Сумма двух равномерно распределенных чисел

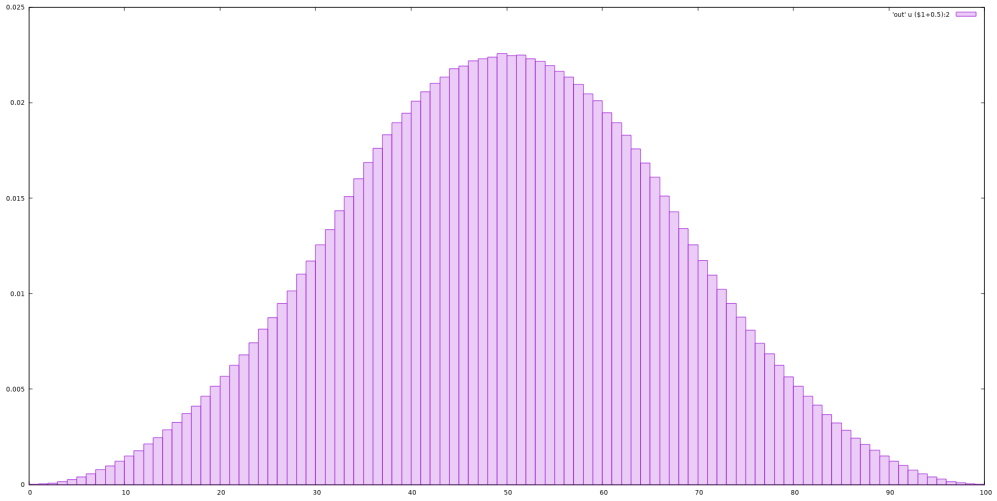

Рис. 6: Сумма трех равномерно распределенных чисел

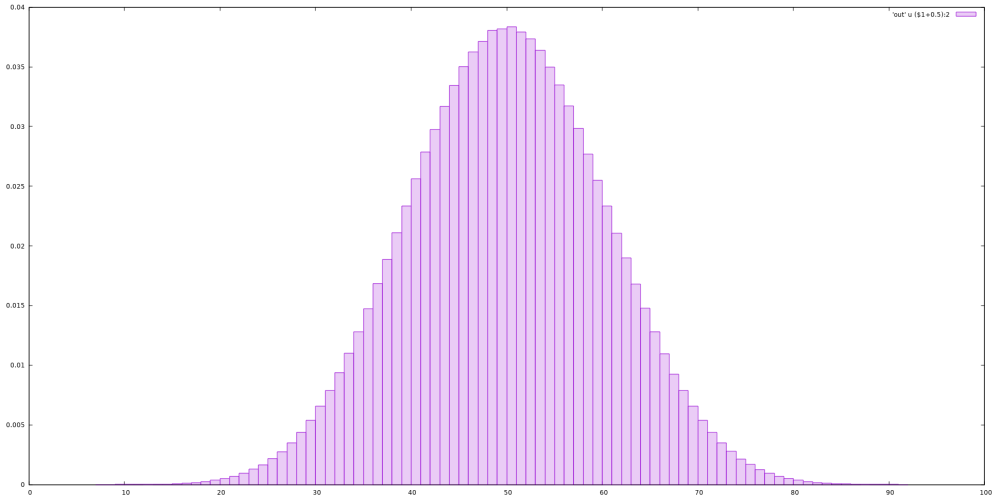

Рис. 7: Сумма восьми равномерно распределенных чисел

Контрольные вопросы к лекции.

1. Если непрерывно равномерно распределенное случайное число имеет границы области значений [3:8], то какова вероятность попадания числа в интервал [2.5:3.7]?

2. Что изменится, если убрать (double) в знаменателе в коде функции rnd()?

3. Судя по графикам, какого количества точек будет достаточно для построения графика распределения плотности вероятности равномерно распределенной случайной величины на 1237 столбцов?