Лекция. Представление данных (таблицы, диаграммы, графики), генеральная совокупность, выборка, среднее арифметическое, медиана.

Одним из основных понятий теории вероятностей является понятие **случайной величины**.

Случайная величина — такая величина, которая в результате эксперимента может принимать различные значения, причем заранее неизвестно, какие именно.

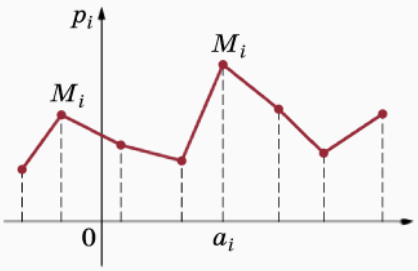
Классическая вероятность, в основе которой лежат комбинаторные подсчеты числа вариантов, позволяет решить многие практические задачи. Однако возможных исходов может быть бесконечно много, и применить к ним комбинаторные подсчеты нельзя. В этих случаях используется аппарат теории функций, при котором каждому возможному исходу сопоставляется некоторое число, а вероятность исхода определяется как функция с определенными свойствами.

1. Дискретная случайная величина. Для использования аппарата теории функций потребуем, чтобы результаты события, его исходы, выражались числами. При бросании игральной кости в качестве исхода получаем число очков, нанесенное на его грань. Для задач комбинаторики это число является просто меткой, которая позволяет различить грани кубика (вместо цифры можно было бы использовать цвет), а для задач теории функций важно само значение числа (величины).

Нагляднее всего это можно себе представить, рассматривая величины, которые принимают конечное число различных значений (их можно включить в общий класс так называемых **дискретных величин**). Если занумеровать значения от 1 до n, то можно сказать, что задана функция, определенная на конечном множестве натуральных чисел от 1 до n и принимающая **различные** числовые значения $a_1, \ldots, a_n$.

Свяжем с каждым значением случайной величины вероятность его появления. Построим график (несколько отличный от того, который строился в примере бросания двух игральных костей), на горизонтальной оси отложим возможные значения случайной величины a_k, а на вертикальной оси — вероятности появления этих значений $p_k = p(a_k)$. Точки M_k с координатами (a_k, p_k) соединим отрезками. Полученный график представляет собой **закон распределения вероятностей** данной случайной величины. Если дискретная случайная величина представлена в виде таблицы, в которой перечислены ее значения и соответствующие им вероятности, то получим другое представление закона распределения дискретной случайной величины.

Закон распределения дискретной случайной величины



2. Непрерывная случайная величина. Существуют случайные величины, которые могут принимать любые значения из некоторого промежутка. Такие величины называют **непрерывными**. Например, дальность выстрела из орудия можно считать числом, принимающим любые значения в пределах дальности стрельбы этого орудия. Приписывать вероятность точному значению дальности бессмысленно — надо оперировать с промежутками, интервалами, куда попадает результат испытания. Правильнее задать функцию $y = f(x)$, с помощью которой можно вычислить вероятность попадания значения случайной величины в заданный интервал.

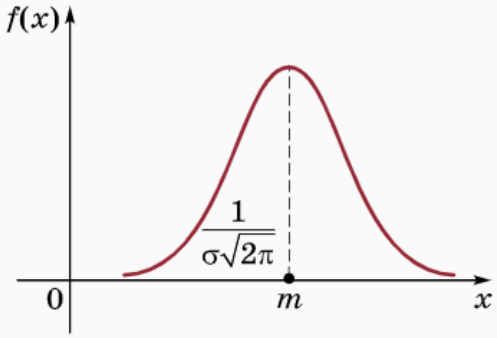
Для непрерывной случайной величины нельзя построить таблицу, в которой могут быть перечислены все ее возможные значения. Поэтому вместо вероятности конкретного значения непрерывной случайной величины X рассматривают вероятность того, что случайная величина X принимает значения, меньшие данного числа x, т. е. рассматривается функция $F(x) = p(X < x)$. Если возможные значения случайной величины заполняют промежуток $[a; b]$ (в общем случае он может быть и бесконечным), то $F(a) = 0$, $F(b) = 1$, причем функция $F(x)$ возрастает от 0 до 1.

Производная этой функции $f(x)$ характеризует **плотность** распределения вероятностей непрерывной случайной величины. Она обладает следующими свойствами:

$$f(x) \geq 0,\ F(x) = \int_a^x f(t)dt,\ \int_a^b f(x)dx = 1.$$

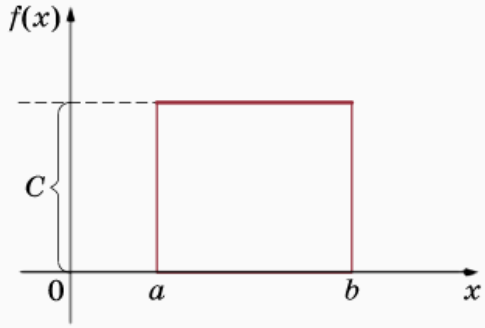
Функция $F(x)$ называется **интегральной функцией распределения**, или **интегральным законом распределения случайной величины**.

Плотность нормального распределения случайной величины

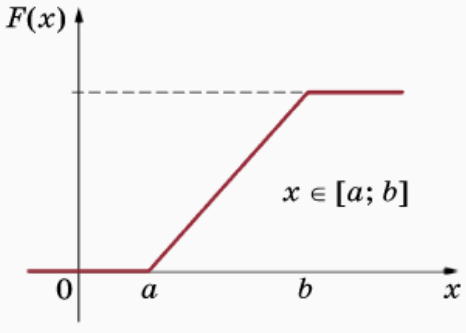


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-m)^2}{2\sigma^2}}$$

Плотность равномерно распределенной случайной величины



Равномерное распределение случайной величины (интегральная функция распределения)



$$f(x) = F'(x)$$

$$F(x) = \int_a^x f(t)dt$$

$[a; b]$ — область значений случайной величины.

Какие числовые характеристики можно связать со случайной величиной?

Математическое ожидание случайной величины является аналогом среднего арифметического. Если случайная величина принимает значения a_1, ..., a_n с одинаковой вероятностью $\frac{1}{n}$, то ее математическое ожидание M и будет средним арифметическим:

$$M = \frac{a_1 + a_2 + \ldots + a_n}{n}.$$

Если нам известны вероятности p_1, ..., p_n принимаемых значений, то в качестве математического ожидания берется взвешенное среднее арифметическое:

$$M = p_1 a_1 + \ldots + p_n a_n.$$

Пример

Вычислим математическое ожидание суммы очков при бросании двух игральных костей.

$$M = \frac{1}{36} \cdot 2 + \frac{2}{37} \cdot 3 + \frac{3}{36} \cdot 4 + \ldots + \frac{2}{36} \cdot 11 + \frac{1}{36} \cdot 12 = 7.$$

Заметим, что математическое ожидание числа очков при бросании одной кости будет равно $M = \frac{1 + 2 + \ldots + 6}{6} = \frac{1 + 6}{2} = \frac{7}{2}$.

Видим, что математическое ожидание суммы очков при бросании двух костей оказалось вдвое бóльшим.

Этот результат нетрудно обобщить для любого числа независимых испытаний.

Кроме математического ожидания вводят и другие числовые характеристики случайной величины, например дисперсию (математическое ожидание квадрата отклонения случайной величины от ее математического ожидания); среднее квадратичное отклонение (показывает разброс значений случайной величины вокруг ее математического ожидания) и др.

Мы приступим к изучению элементов математической статистики, в которой разрабатываются научно обоснованные методы сбора статистических данных и их обработки.

1. Генеральная совокупность и выборка. Пусть требуется изучить множество однородных объектов (это множество называется *статистической совокупностью*) относительно некоторого качественного или количественного признака, характеризующего эти объекты. Например, если имеется партия деталей, то качественным признаком может служить стандартность детали, а количественным — контролируемый размер детали.

Лучше всего произвести сплошное обследование, т.е. изучить каждый объект. Однако в большинстве случаев по разным причинам это сделать невозможно. Препятствовать сплошному обследованию может большое число объектов, недоступность их. Если, например, нужно знать среднюю глубину воронки при взрыве снаряда из опытной партии, то, производя сплошное обследование, мы уничтожим всю партию.

Если сплошное обследование невозможно, то из всей совокупности выбирают для изучения часть объектов.

Статистическая совокупность, из которой отбирают часть объектов, называется *генеральной совокупностью.* Множество объектов, случайно отобранных из генеральной совокупности, называется *выборкой.*

Число объектов генеральной совокупности и выборки называется соответственно *объемом генеральной совокупности* и *объемом выборки*.

Пример 10.1. Плоды одного дерева (200 шт.) обследуют на наличие специфического для данного сорта вкуса. Для этого отбирают 10 шт. Здесь 200 — объем генеральной совокупности, а 10 — объем выборки.

Если выборку отбирают по одному объекту, который обследуют и снова возвращают в генеральную совокупность, то выборка называется *повторной.* Если объекты выборки уже не возвращаются в генеральную совокупность, то выборка называется *бесповторной.* На практике чаще встречается бесповоротная выборка. Если объем выборки составляет небольшую долю объема генеральной совокупности, то разница между повторной и бесповоротной выборками незначительна.

Свойства объектов выборки должны правильно отражать свойства объектов генеральной совокупности, или, как говорят, выборка должна быть *репрезентативной* (представительной). Считается, что выборка репрезентативна, если все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку, т.е. выбор производится случайно. Например, для того чтобы оценить будущий урожай, можно сделать выборку из генеральной совокупности еще не созревших плодов и исследовать их характеристики (массу, качество и пр.). Если вся выборка будет сделана с одного дерева, то она не будет репрезентативной. Репрезентативная выборка должна состоять из случайно выбранных плодов со случайно выбранных деревьев.

2. Статистическое распределение выборки. Полигон. Гистограмма. Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, x_2 — n_2 раз, x_k — n_k раз и $n_1 + n_2 + \ldots + n_k = n$ — объем выборки. Наблюдаемые значения $x_1, x_2, \ldots, x_k$ называются *вариантами*, а последовательность вариант, записанная в возрастающем порядке, — *вариационным рядом*. Числа наблюдений $n_1, n_2, \ldots, n_k$ называются *частотами*, а их отношения к объему выборки $\frac{n_1}{n} = p_1^*$, $\frac{n_2}{n} = p_2^*, \ldots, \frac{n_k}{n} = p_k^*$ — *относительными частотами*. Отметим, что сумма относительных частот равна единице: $p_1^* + p_2^* + \ldots + p_k^* = 1$.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот. Статистическое распределение можно задать также в виде последовательности интервалов и соответствующих им частот (непрерывное распределение). В качестве частоты, соответствующей интервалу, принимают сумму частот вариант, попавших в этот интервал. Для графического изображения статистического распределения используются *полигоны* и *гистограммы*.

Для построения полигона на оси Ox откладывают значения вариант x_i, на оси Oy — значения частот n_i (относительных частот p_i^*).

Пример 10.2. На рис. 10.1 изображен полигон следующего распределения:

Варианта x_i	1	2	3	5
Относительная частота p_i^*	0,4	0,2	0,3	0,1

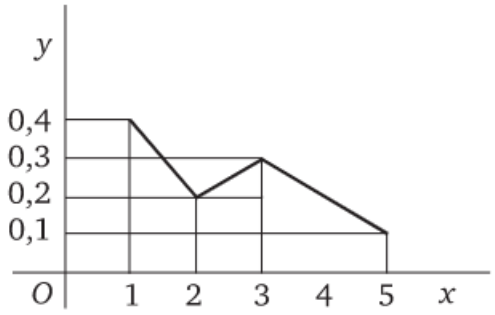


Рис. 10.1

Полигоном обычно пользуются в случае небольшого количества вариант. В случае большого количества вариант и в случае непрерывного распределения признака чаще строят гистограммы. Для этого интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько частичных интервалов длиной h и находят для каждого частичного интервала n_i — сумму частот вариант, попавших в i-й интервал. Затем на этих интервалах как на основаниях строят прямоугольники с высотами $\frac{n_i}{h}$ (или $\frac{n_i}{nh}$, где n — объем выборки). Площадь

i-го частичного прямоугольника равна $\frac{hn_i}{h} = n_i$ (или $\frac{hn_i}{hn} = \frac{n_i}{n} = p_i^*$). Следовательно, площадь гистограммы равна сумме всех частот (или относительных частот), т.е. объему выборки (или единице).

Пример 10.3. На рис. 10.2 изображена гистограмма непрерывного распределения объема $n = 100$, приведенного в следующей таблице.

Частичный интервал h	Сумма частот вариант частичного интервала n_i	$\frac{n_i}{h}$
5—10	4	0,8
10—15	6	1,2
15—20	16	3,2
20—25	36	7,2
25—30	24	4,8
30—35	10	2,0
35—40	4	0,8

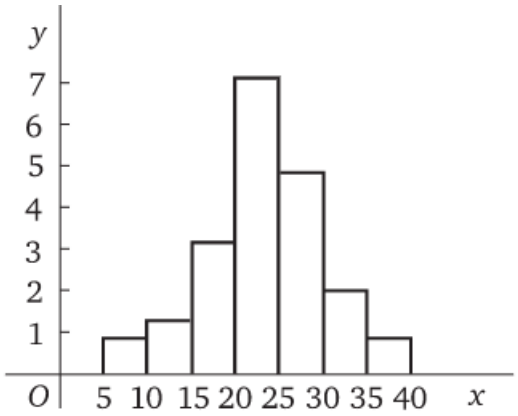


Рис. 10.2

Выборка как набор случайных величин. Пусть имеется некоторая генеральная совокупность, каждый объект которой наделен количественным признаком X. При случайном извлечении объекта из генеральной совокупности становится известным значение x признака X этого объекта. Таким образом, мы можем рассматривать извлечение объекта из генеральной совокупности как испытание, X — как случайную величину, а x — как одно из возможных значений X.

Допустим, что из теоретических соображений удалось установить, к какому типу распределений относится признак X. Естественно, возникает задача оценки (приближенного нахождения) параметров, которыми определяется это распределение. Например, если известно, что изучаемый признак распределен в генеральной совокупности нормально, то необходимо оценить, т.е. приближенно найти, математическое ожидание и среднее квадратическое отклонение, так как эти два параметра полностью определяют нормальное распределение.

Обычно в распоряжении исследователя имеются лишь данные выборки генеральной совокупности, например значения количественного признака $x_1, x_2, \ldots, x_n$, полученные в результате n наблюдений (здесь и далее наблюдения предполагаются независимыми). Через эти данные и выражают оцениваемый параметр.

Опытные значения признака X можно рассматривать и как значения разных случайных величин $X_1, X_2, \ldots, X_n$ с тем же распределением, что и X, и, следовательно, с теми же числовыми характеристиками, которые имеет X. Значит,

$$M(X_i) = M(X) \text{ и } D(X_i) = D(X).$$

Генеральная и выборочная средние. Методы их расчета. Пусть изучается дискретная генеральная совокупность объемом N относительно количественного признака X.

Определение. *Генеральной средней* $\overline{x}_г$ (или a) называется среднее арифметическое значений признака генеральной совокупности.

Определение. *Выборочной средней* $\overline{x}_в$ называется среднее арифметическое значений признака выборочной совокупности.

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} x_i n_i . \qquad (10.4)$$

Пример 10.4. Выборочным путем были получены следующие данные о массе 20 морских свинок при рождении (в г): 30, 30, 25, 32, 30, 25, 33, 32, 29, 28, 27, 36, 31, 34, 30, 23, 28, 31, 36, 30. Найти выборочную среднюю.

Согласно формуле (10.4) имеем

$$\overline{x}_в = \frac{30\cdot 5 + 25\cdot 2 + 32\cdot 2 + 33 + 29 + 28\cdot 2 + 27 + 36\cdot 2 + 31\cdot 2 + 34 + 23}{20} = 30.$$

Итак, $\overline{x}_в = 30$ г. Здесь для облегчения вычислений можно использовать калькулятор. То же следует иметь в виду и в ряде других примеров этой главы.

Медиана в статистке

Медиана — это такое значение признака, которое разделяет ранжированный ряд распределения на две равные части — со значениями признака меньше медианы и со значениями признака больше медианы. Для нахождения медианы, нужно отыскать значение признака, которое находится на середине упорядоченного ряда.

I

В ранжированных рядах несгруппированные данные для **нахождения медианы** сводятся к поиску порядкового номера медианы. Медиана может быть вычислена по следующей формуле:

$$Me = X_{Me} + i_M \frac{\frac{\sum f}{2} - S_{Me-1}}{f_{Me}}$$

где Xm — нижняя граница медианного интервала;
im — медианный интервал;
Sme— сумма наблюдений, которая была накоплена до начала медианного интервала;
fme — число наблюдений в медианном интервале.

Свойства медианы

1. Медиана не зависит от тех значений признака, которые расположены по обе стороны от нее.
2. Аналитические операции с медианой весьма ограничены, поэтому при объединении двух распределений с известными медианами невозможно заранее предсказать величину медианы нового распределения.
3. *Медиана обладает* свойством минимальности. Его суть заключается в том, что сумма абсолютных отклонений значений x, от медианы представляет собой минимальную величину по сравнению с отклонением X от любой другой величины

Глава 11 «Элементы теории вероятности и математической статистики», учебник Башмаков М.И. Математика: алгебра и начала математического анализа, геометрия: учеб. для студ. учреждений сред.проф. образования/ М.И. Башмаков. – 4-е изд.,стер. – М. : ИЦ «Академия», 2017, - 256 с.

В случае отсутствия печатного издания, Вы можете обратиться к Электронно-библиотечной системе.

Список использованных интернет-ресурсов:

1. https://23.edu-reg.ru/
2. https://urait.ru/