**Лекция №2 Дискретные распределения вероятностей**

**Цель лекции:**

* познакомиться с понятием случайной величины
* изучить распределения, которые свойственны дискретной случайной величине
* рассмотреть примеры решения задач с применением законов распределений вероятностей
* научиться отличать биномиальное распределение от распределения Пуассона.

**Материал прошлого урока:**

Мы уже изучили понятие вероятности, научились рассчитывать ее с помощью формул комбинаторики, научились вычислять условную вероятность, применять формулу полной вероятности и формулу Байеса.

**План урока:**

1. Случайная величина
2. Обозначения случайной величины
3. Дискретные и непрерывные случайные величины.
4. Закон распределения вероятностей случайной величины
5. Биномиальное распределение
6. Распределение Пуассона
7. Описательная статистика
8. Описательная статистика для биномиального распределения
9. Описательная статистика для распределения Пуассона
10. Сравнение биномиального распределения и распределения Пуассона

**Случайная величина**

На прошлом уроке мы изучили расчет вероятности для конкретного события. Сегодня будем знакомиться с распределениями вероятностей. Существуют свои распределения для дискретных случайных величин и для непрерывных случайных величин.

Так что же такое вообще случайная величина? Давайте дадим определение, предлагаемое учебниками по статистики, а потом рассмотрим его на конкретном примере.

**Случайная величина  (далее СВ)**- это величина, которая в результате испытания принимает только одно возможное значение.

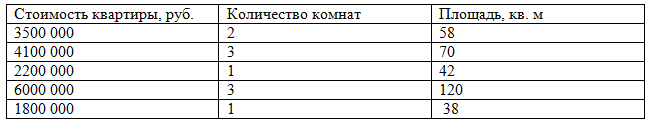
Например, мы хотим измерить рост всех людей, живущих в России. Мы создадим колонку в табличке excel, куда будем помещать рост каждого отдельного индивидуума. Этот столбец мы назовем «рост людей в России». Это и будет та случайная величина, которую мы будем измерять. Измерение каждого человека – это отдельное испытание. И в результате этого испытания случайная величина «рост людей в России» принимает только одно возможное значение, как и сказано в определении.

**Обозначения случайной величины**

Саму случайную величину (далее СВ) принято обозначать Х (совсем упрощенно говоря, это ее наименование), а вот – это те значения, которые она принимает, где – это номер испытания ( в нашем примере это порядковый номер индивидуума).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***Наименование случайной величины Х*** | ***х1*** | ***х2*** | ***х3*** | ***х4*** | ***х5*** | ***х6*** |
| ***Рост*** | ***167*** | ***173*** | ***162*** | ***168*** | ***181*** | ***158*** |

Давайте определим, сколько случайных величин на рисунке ниже?



На этом рисунке 3 СВ.

1)Стоимость квартиры

2) Количество комнат

3) Площадь квартиры.

Каждая случайная величина имеет по 5 измерений.

Иногда на лекциях приходится слышать ответ 2 СВ, а именно «количество комнат» и «площадь», а «стоимость» не считают за случайную величину, т.к. ее можно предсказать за счет двух других. Но случайные величины могут быть зависимыми и независимыми. Например, СВ «Стоимость квартиры» будет зависеть от СВ «Площадь» и СВ «Количество комнат».

**Дискретные и непрерывные случайные величины.**

Принято разделять СВ на две большие категории: дискретные и непрерывные. Дискретные принимают счетные, отдельные друг от друга значения. Например, мы собираем информацию о количестве учащихся по школам нашего города. Каждая школа присылает нам данные. В одной учатся 900 человек, в другой 587, в третьей 1008 человек и т.д. Эти цифры принимают целые, отдельные друг от друга значения. Мы не можем измерить более точно, чем уже есть. Т.е. не можем «присмотреться» и разглядеть, что там, к примеру, 900.01. И наша СВ, назовем ее «Количество учащихся в школах город N» :900, 587, 1008 …

Еще примерами дискретных СВ могут служить: количество выпавших орлов при 100 кратном подбрасывании. Т.е. при первом 100 кратном подбрасывании получили 51 орел, при следующем стократном подбрасывании получили 59 орлов и т.д. СВ: 51, 59,….

Другой пример: число метеоритов, упавших на землю за определенный год. СВ: 0, 2, 0, 1…

А **непрерывная СВ** - это случайная величина, которая лежит в некотором интервале, причем этот интервал может быть как конечным, так и бесконечным.

Например, есть производственный процесс по изготовлению определенной детали. На выходе производственного процесса находится контролер. Он измеряет длину этой детали. Если мы дадим ему линейку с ценой деления 1 мм, то он может измерить с точностью до 1 мм. Предположим, он получил 3 см 5 мм. Но если мы дадим ему более точный прибор, с меньшей ценой деления, то он может увидеть, что на самом деле там не 3,5 см, а 3,51 см.

Т.о. непрерывную случайную величину мы можем измерить с той точностью, что позволяет нам прибор или с той точностью, которая нам необходима.

**Закон распределения вероятностей случайной величины**

**Непрерывные и дискретные СВ имеют свои законы распределения.** И сегодня мы будем говорить о дискретных законах распределения вероятностей.

Изначально думали, что СВ ведет себя непредсказуемо, т.е. принимает абсолютно хаотичные значения. Но потом, изучив ее поведение, поняли, что она ведет себя определенным образом, т.е. имеет свое распределение вероятностей или, иными словами, имеет свою математическую модель поведения. Под математической моделью мы понимаем уравнение, которое описывает ее поведение. (Рассмотрим позже в этом уроке).

Давайте опять сначала дадим определение закону распределения вероятностей, а потом его поясним.

**Закон распределения вероятностей случайной величины** - это *соответствие* между возможными значениями этой величины и вероятностями, которые этим значениям соответствуют. Что это все значит?

Давайте представим, что нас интересует СВ «Число выпавших орлов при 10 кратном подбрасывании». Мы этот эксперимент проведем 10 раз. Т.е. 10 раз подбросим монетку по 10 раз. Например, при первом 10-кратном подбрасывании орел выпал 5 раз, при втором 10 кратном подбрасывании тоже 5 раз, а при третьем- 4 раза. И этот эксперимент проведем 10 раз. Занесем все эти данные в таблицу.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 5 | 5 | 4 | 5 | 3 | 4 | 5 | 4 | 5 | 4 |

Мы видим, что дискретная СВ «Число выпавших орлов при 10-кратном подбрасывании» принимает значения 5, 4, 3. Теперь составим таблицу, куда занесем те значения СВ, которые мы получили в ходе эксперимента (5,4,3) и запишем во второй столбец, сколько раз мы встретили то или иное значение, а затем посчитаем в 3-м столбце долю (пропорцию) этих значений среди всех, что мы получили.

|  |  |  |
| --- | --- | --- |
| Значение  СВ | Сколько раз встретилось значение | Доля среди всех полученных значений. |
| 5 | 5 раз | 50% |
| 4 | 4 раза | 40% |
| 3 | 1 раз | 10% |

В итоге мы видим возможные значения СВ (3,4,5) и соответствующие им вероятности (доли). И вернемся теперь опять к определению,

**Закон распределения вероятностей** – это **соответствие** между возможными значениями и вероятностями, которые этим значениям соответствуют. И мы видим, что в таблице напротив полученного значения стоит в 3м столбце соответствующая вероятность (доля).

Но здесь мы провели эксперимент всего лишь 10 раз. Т.е. 10 раз сделали 10-кратное подбрасывание. Но если бы мы проводили эксперимент бесконечное число раз, то мы могли бы увидеть и другие значения СВ, а именно от 0 до 10. У нас могло не выпасть ни одного орла (0), а могли выпасть и все 10 орлов.

Мы говорили на прошлом уроке о понятии статистическая вероятность и относительная частота. Обе эти величины экспериментальные, но для статистической вероятности свойственно огромное число испытаний. С увеличением числа испытаний вероятность стремится к какой-то постоянной величине. И так же в нашем примере. Но как же посчитать истинные вероятности?

**Биномиальное распределение**

Когда нам нужно посчитать вероятность, что событие наступит k раз из n, к тому же вероятность наступления события в отдельном испытании достаточно велика, как в нашем примере (вероятность выпадение орла при отдельном подбрасывании 0.5), то вероятнее всего мы имеем дело с биномиальным распределением.

Биномиальное распределение – это один из примеров дискретного распределения. И СВ «Число выпавших орлов при 10-кратном подбрасывании» будет следовать этому распределению.

Мы хотели ответить на вопрос, а как же посчитать истинные вероятности для возможных значений СВ. Возьмем похожий пример, только СВ «Число выпавших орлов при 3-кратном подбрасывании».

Мы воспользуемся готовой математической моделью, описывающей поведение дискретной СВ, которая следует биномиальному распределению. А именно формулой Бернулли. А потом посчитаем вероятности, используя комбинаторику. Чтобы дать представление, как это рассчитывается и не утомить вас расчетами, мы взяли не 10-кратное, а 3-кратное подбрасывание.

**Мы говорим, что случайная дискретная величина имеет биномиальное распределение с переменными и , если вероятность ее распределения задается следующим уравнением:**

* **,**число наступления события (дискретная величина из отрезка
* - число испытаний
* - вероятность наступления события в независимых испытаниях,
* . Противоположная вероятность

Выше приведенная формула для вычисления вероятностей наступления события раз из и есть **формула Бернулли.**

Воспользуемся этой формулой для расчета вероятностей для нашего примера с 3-кратным подбрасыванием. СВ может принимать значения от 0 до 3. Орел может выпасть 0 раз из 3, 1 из 3х, 2 из 3х и 3 из 3х. Т.е. , может принимать значения 0,1,2,3

А как бы мы считали, если бы у нас не было этой формулы Бернулли?

Напоминаю возможные исходы: 0 орлов, 1 орел, 2 орла, 3 орла.

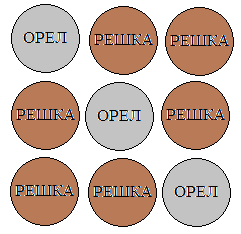
1. 0 орлов (X=k=0) это когда все три подбрасывания показали нам решку.

Т.е. 1 возможное сочетание. Математически обосновываем так:

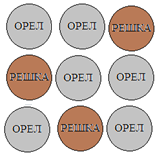


1. 1 орел (X=k=1) , т.е. один орел может быть при первом, или при втором, или при третьем подбрасывании (3 исхода)

Предвосхищая вопрос, а почему здесь порядок важен, а используем формулу сочетаний, объясняю. Что на самом деле порядок нам не важен. У нас просто есть подбрасывание первое, подбрасывание второе и подбрасывание третье. И мы хотим посчитать, а сколько способов может быть, когда одно из трех подбрасываний показывает орел. По логике и так понятно 3, но математически можем посчитать через формулу сочетаний.



1. 2 орла (X=k=1) : 3 возможных исхода.



4) 3 орла (X=k=3): 1 возможный исход.



Теперь можем посчитать число всех возможных исходов:

= 1+3+3+1 =8

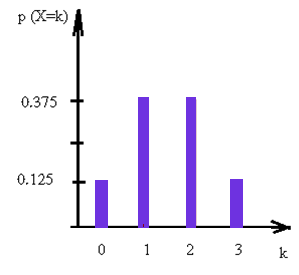
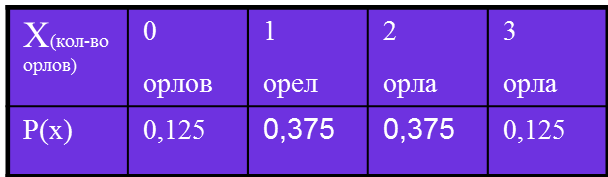
Вспоминаем, что вероятность – это отношение числа благоприятствующих исходов к общему числу исходов. Общее число исходов для этого примера мы посчитали - 8 ()

Значению ноль соответствует 1 благоприятствующий (возможный) исход, когда все решки.

Т.о. вероятность, что случайная величина примет значение ноль равна 1/8.

Найдем аналогично вероятности для остальных возможных значений.

Теперь осталось представить эти данные в читабельном виде. Часто их представляют в виде таблице или графика (графическая математическая модель)



Разберем теперь задачу.

***Пример решения задачи.*** С производственной линии для проведения контроля качества продукции случайным образом была взята выборка из 20 изделий. Если известно, что данный производственный процесс дает 3% бракованной продукции, то какова вероятность, что в выборке попадется 1 бракованное изделие.

Как здесь рассуждаем. Нам надо найти вероятность, что событие (бракованное изделие) наступит 1 раз из 20. Это дискретная случайная величина. Вероятность наступления события 0.03. Используем для расчета вероятности формулу Бернулли.

Получаем вероятность встретить бракованную деталь в выборке из 20 единиц равна 33.6%

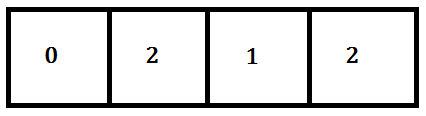
**Распределение Пуассона**

**В том случае, если вероятность наступления события очень маленькая (< 0.1)**, а число испытаний очень большое, применяется формула Пуассона. **Распределение Пуассона это также дискретное распределение и является частным случаем биномиального распределения.**

Чтобы найти вероятность наступления события за фиксированную единицу измерения или вероятность, что событие наступит m раз из n, применяем формулу Пуассона. Эта формула дает более точные значения вероятности в такой ситуации.

**Здесь появляется величина** . Это средняя интенсивность наступления события на некоторую единицу измерения. Например, число ошибок на разворот, вкраплений/ , потребителей/час.

Например, мы изготавливаем полотно и знаем, что в среднем 2 вкрапления на . Т.е. может быть 0 вкраплений, 1, 3 , но чаще всего 2 на .

******

Или же мы знаем, что при производстве какой-то продукции есть маленький процент брака. Например, 0.01. чтобы узнать, сколько в среднем на 1 партию из 1000 штук встречается бракованных изделий, нам нужно 1000\*0.01= 10 штук. Т.е. зная долю брака в устоявшемся производственном процессе, мы можем посчитать, среднюю интенсивность наступления события, т.е. в нашем примере 10 бракованных единиц на 1000 единиц (на 1 партию).

И зная эту среднюю интенсивность, иными словами, зная, что чаще всего мы встречаем 10 бракованных единиц на 1 партию, мы сможем узнать, а какая вероятность, что мы встретим не 10 (среднюю интенсивность), а 9 бракованных единиц или 12 бракованных единиц и т.д.***, используя формулу Пуассона.***

***Рассмотрим ниже 2 примера решения задач.***

Вероятность того, что среди писем, поступающих на определенный почтовый ящик, встретится письмо со спамом, составляет 0.01.  
Найдите вероятность того, что среди 1000 писем, поступивших на него за месяц, будет 11 со спамом.

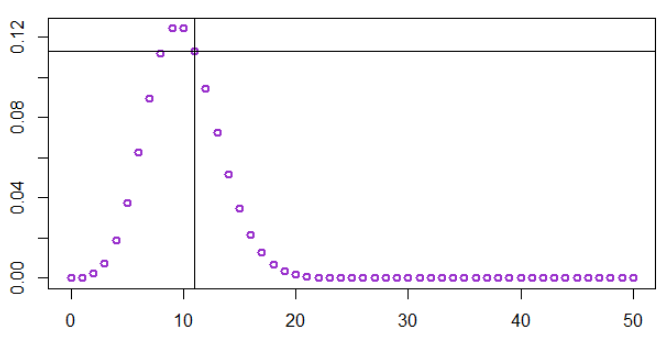
Рассмотрим решение. Мы знаем долю спам-писем во всем этом процессе. А нас интересует, а какая вероятность, что мы встретим 11 спам-писем на 1000. Но чтобы это узнать, нужно сначала выяснить, а сколько чаще всего спам-писем встречается на 1000 штук. Т.е. найти среднюю интенсивность.

λ = p\*n = 0,01 \*1000 =10 спам- писем/ 1000 штук

Зная среднюю интенсивность наступления события, можем вычислить вероятность встретить 11 спам-писем на 1000.

*0.113 или это 11,3 %*

Давайте взглянем на график распределения Пуассона для средней интенсивности 10. По графику видим, что вероятность встретить 10 спам-писем на 1000 писем  0.12 самая высокая, а для 11 писем она около 0.11

**

Еще одна задача на распределение Пуассона:

В среднем при изготовлении напольного покрытия на 1 кв.м встречается 2 вкрапления. Какова вероятность, что на 1 кв.м будет не более 1 вкрапления?

Т.е. чтобы ответить на этот вопрос, надо найти вероятности для события «встретить 0 вкраплений» и для события «встретить 1 вкрапление» и сложить их.

***P(0) + P(1)***

Т.е. с вероятностью 40.5% мы встретим не более 1 вкрапления, зная, что чаще всего встречается 2 (средняя интенсивность).

**Описательная статистика**

Когда мы работаем с какой-либо СВ, мы хотим получить как можно больше информации о ней. И на помощь приходит описательная статистика. И мы обращаемся к таким параметрам, как математическое ожидание, среднее квадратичное отклонение, дисперсия. На следующем уроке мы поговорим более подробно об этих параметрах. Сейчас будем их использовать, чтобы закончить тему с дискретными распределениями.

**Математическое ожидание (M(X))** – это среднее значение случайной величины, при числе испытаний стремящихся к бесконечности.

**Дисперсия (D(X), )** характеризует степень рассеянности значений случайной величины относительно ее математического ожидания

**Среднее квадратичное отклонение (** ) показывает, насколько далеко наблюдения могут быть "разбросаны" относительно их среднего значения .

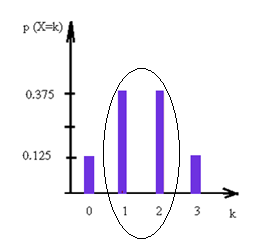
**Описательная статистика для биномиального распределения**

Рассмотрим сперва **описательную статистику для биномиального распределения.** Ниже приведены формулы характеристик, которым мы только что дали определения.

Рассчитаем эти параметры для СВ, полученной при эксперименте с трехкратным подбрасыванием монетки, где мы считали число выпавших орлов.

Несмотря на то, что мы имеем дело с дискретной СВ, математическое ожидание у нас может иметь дробную часть. Это не должно вас смущать, потому что задачей математического ожидания является показать, вокруг какого значения лежит большая доля значений СВ.

И на графике мы видим, что большая доля значений (0.375+0.375=0.75) лежит вокруг рассчитанного среднего арифметического.



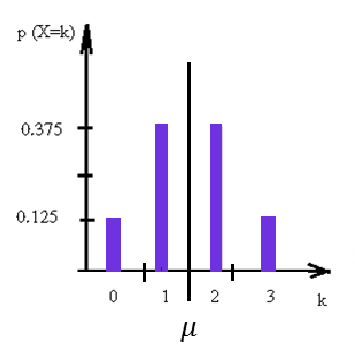
Дисперсию на этом графике мы не изобразим. Это квадратичная величина.

Но в статистике пользуются средним квадратичным отклонением или еще его называют стандартным отклонением.

Посчитаем интервал в одно стандартное отклонение

M(X)

И отметим его на графике



Получаем, что в пределах [0.634; 2.366] лежит 75% всех наблюдений.

**Описательная статистика для распределения Пуассона.**

Здесь отличительная особенность заключается в том, что математическое ожидание и дисперсия равны средней интенсивности наступления события.

М (Х) = n\*p= λ

D (X) = λ

Если в реальной жизни вы столкнулись с тем, что математическое ожидание почти такое же, как и дисперсия, то высока вероятность, что вы работаете с распределением Пуассона.

**Сравнение биномиального распределения и распределения Пуассона**

Так как же определить, какой формулой нам воспользоваться: формулой Бернулли или формулой Пуассона.

Давайте подведем итог в виде таблицы.

|  |  |
| --- | --- |
| Биномиальное распределение | Распределение Пуассона |
| * Дискретное распределение * Позволяет найти вероятность, что событие наступит k раз из n * Вероятность наступления события достаточно высока * Свойственно обычно небольшое число испытаний n | * Дискретное распределение * Позволяет найти вероятность, что событие наступит m раз из n или на единицу измерения * Вероятность наступления события очень маленькая (<0.01) * Свойственно большое число испытаний. |

На следующем уроке мы будем знакомиться с разведочным анализом, который является важной частью статистического анализа. Более подробно поговорим об описательной статистике, научимся строить и интерпретировать графики. Это даст нам возможность получить более полную информацию о СВ, и даже показать то, чего мы совсем не ожидали увидеть.