Отделение рекламы и связей с общественностью, 2014-15 уч. год
Теория вероятностей и статистика
Лекция 4. Математическое ожидание и дисперсия (20.09.2014)
И. А. Хованская, Р. Я. Будылин, И. В. Щуров, Д. А. Филимонов, К. И. Сонин (РЭШ)

1 Случайные величины

В случайных экспериментах нас часто интересуют такие показатели, которые имеют числовое выражение. Например, у каждого человека имеется много числовых характеристик: рост, возраст, вес и т.д. Если мы выбираем человека случайно, то случайными будут и значения указанных характеристик. Чтобы подчеркнуть, что выражаемый числом показатель или измеряемая в ходе опыта числовая характеристика зависит от случайного исхода, и потому сама является случайной, ее называют случайной величиной. Случайными величинами являются выигрыш в казино, прибыль страховой компании, число орлов выпавших при 5 бросаниях монетки и т.д.

Определение 1. Случайная величина – это функция от элементарного исхода эксперимента.

Пример 1. Рассмотрим случайную величину – число орлов выпавших в 3 бросках. Можно проверить, что мы получим следующую функцию на элементарных исходах:

ООО	РОО	ОРО	ООР	РРО	РОР	ОРР	РРР
3	2	2	2	1	1	1	0

2 Математическое ожидание

Рассмотрим выигрыш в какой-то азартной игре. Тогда это конечно-же случайная величина. Что прежде всего интересует игрока? Размер ожидаемого выигрыша в многократно повторяемой игре.

Пример 2. Пусть мы участвуем в лотерее, про которую известно, что каждый четвертый билет выигрышный. Пусть билет стоит 100 рублей, а размер выигрыша 300 рублей. Выгодно ли нам участвовать в подобной лотерее? В среднем за 4 игры мы выигрываем 1 раз, значит, мы получаем в среднем 300 рублей за 4 игры. При этом мы покупаем билеты за 400 рублей. Получается, что за 4 игры в среднем мы теряем 100 рублей, то есть по 25 рублей за игру. Таким образом, участвовать в такой лотерее невыгодно.

Каким образом, определить среднее ожидаемое значение случайной величины? Рассмотрим еще один пример.

Пример 3. Пусть инвестор хочет оценить в новое предприятие какой из двух отраслей стоит вкладываться. Тогда стоит посмотреть какая из этих отраслей в среднем более прибыльна (см. таблицы 1 и 2). Как мы видим в первой области средняя доходность равна 120 рублям, а во второй 115. Поэтому в отсутствии иной информации стоит вкладываться в первую область.

Таблица 1: доходность предприятий в первой отрасли

номер фирмы	1	2	3	4	5
доход со 100 вложенных рублей	110	115	120	125	130

Таблица 2: доходность предприятий во второй отрасли

номер фирмы	1	2	3	4	5
доход со 100 вложенных рублей	105	110	110	115	135

Итак, если различные элементарные исходы a_i равноправны, то в качестве ожидаемого выигрыша нужно брать среднее арифметическое всех значений случайной величины, то есть сумму $\frac{1}{n}X(a_1)+\frac{1}{n}X(a_2)+\dots\frac{1}{n}X(a_n)$. Но так как, $\frac{1}{n}$ в данном случае есть $p(a_i)$, то это можно переписать как

$$p(a_1)X(a_1)+p(a_2)X(a_2)+p(a_3)X(a_3)+\dots+p(a_n)X(a_n) \tag{1}$$

Определение 2. В общем случае определим математическое ожидание случайной величины X формулой (1) и будем обозначать его $E(X)$(от Expected value).

Пример 4. Пусть мы бросаем монету три раза. Пусть случайная величина X – число выпавших орлов. Тогда ее математическое ожидание равно

$$3p(OOO)+2p(POO)+2p(OPO)+2p(OOP)+1p(OPP)+1p(POP)+1p(PPO)+0p(PPP)=$$
$$=3*\frac{1}{8}+2*\frac{1}{8}+2*\frac{1}{8}+2*\frac{1}{8}+1*\frac{1}{8}+1*\frac{1}{8}+1*\frac{1}{8}+0*\frac{1}{8}=\frac{3}{2}.$$

Заметим, что в сумме выше было удобно сгруппировать слагаемые следующим образом:

$$3p(OOO)+2(p(POO)+p(OPO)+p(OOP))+1(p(OPP)+p(POP)+p(PPO))+0p(PPP).$$

Но, что такое $p(POO)+p(OPO)+p(OOP)$ – это p(выпало 2 орла), а $p(OPP)+p(POP)+p(PPO)$– это p(выпал 1 орел). Таким образом, всю сумму можно переписать так:

$$3p(\text{выпало 3 орла})+2p(\text{выпало 2 орла})+1*p(\text{выпал 1 орел})+0*p(\text{выпало 0 орлов}).$$

В общем случае, также удобно сгруппировать члены формулы (1) с одинаковыми $X(a_i)$. Считая, что случайная величина X принимает значения $x_1,x_2,\dots,x_k$, получим формулу:

$$E(X)=x_1p(X=x_1)+x_2p(X=x_2)+\dots+x_kp(X=x_k). \tag{2}$$

На практике формула (2) применяется чаще формулы (1). Перечислим свойства математического ожидания.

1. Если мы вместо случайной величины X рассмотрим новую случайную величину $c*X$, где c – это константа, то в формуле (1) каждый член умножится на c, так что $E(cX)=cE(X)$.

2. Если сложить две случайных величины X и Y, то в формуле (1) каждый член будет суммой $X(a_i) * p(a_i) + Y(a_i) * p(a_i)$, так что всю сумму можно разбить на две суммы и получить что $E(X+Y) = E(X)+E(Y)$. Это свойство вместе с первым часто называют свойством линейности мат.ожидания.
3. Если известно, что $X > Y$ (то есть случайная величина X на любом исходе больше величины Y), то $E(X) > E(Y)$. Например, если случайная величина X на любом исходе больше 100, то и $E(X) > 100$. В качестве случайной величины Y мы взяли постоянную случайную величину, всегда равную 100.

Особенно хотелось бы поговорить о свойстве 2. Оно выглядит тривиально, но тем не менее очень часто позволяет упростить вычисление математического ожидания.

Пример 5. Пусть мы бросаем монету n раз, и рассматриваем случайную величину X, равную числу выпавших орлов. Рассмотрим случайную величину X_i равную числу выпавших орлов в i-м броске. То есть X_i равна 1, если в i-м броске орел выпал, и 0 иначе. Тогда X можно представить как сумму случайных величин

$$X = X_1 + X_2 + \cdots + X_n.$$

По свойству линейности мат.ожидания

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

Но $E(X_i)$ равны между собой и их легко вычислить, действительно, $E(X_i) = \frac{1}{2} * 0 + \frac{1}{2} * 1 = \frac{1}{2}$. Таким образом, $E(X) = \frac{n}{2}$. Если мы будем считать, что монета со смещенным центром тяжести и вероятность орла p, то получим математическое ожидание числа орлов в n бросках равным np. Хочется отметить, что нам не пришлось вычислять вероятности того, что число орлов будет равным чему-то (вероятности, фигурирующие в формуле (2)). Этим как-раз и удобно свойство линейности, оно позволяет сводить вычисление мат. ожидания к вычислению мат. ожидания простых величин.

В следующем примере без свойства линейности мат.ожидания попросту не обойтись.

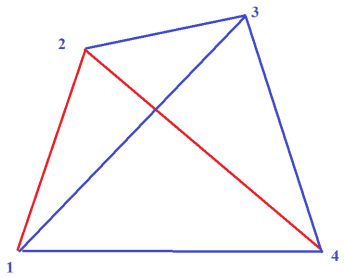

Рис. 1: В данном случае один одноцветный треугольник.

Пример 6. Рассмотрим четырехугольник с вершинами 1, 2, 3 и 4. Пусть его ребра случайным образом окрашиваются в красный или синий цвет с одинаковой вероятностью. Пусть X – случайная величина, равная числу одноцветных треугольников. Как можно вычислить

$E(X)$? Вычислять вероятность, что $X = k$ в этой задаче очень сложно, число треугольников сильно зависит от рисунка. Так что же остается рассмотреть все $2^6 = 64$ возможных рисунка, и для каждого подсчитать число одноцветных треугольников? Вместо этого воспользуемся линейностью мат.ожидания. Всего возможных треугольников 4: 123, 124, 134, 234. Давайте рассмотрим величину X_{ijk}, которая равна 1, если треугольник ijk одноцветный, и 0 иначе. Несложно понять, что $X = X_{123}+X_{124}+X_{134}+X_{234}$. Таким образом, остается вычислить $E(X_{ijk})$, например, $E(X_{123})$, ведь они все равны между собой. Треугольник 123 бывает одноцветным в двух случаях: он весь синий или весь красный. Вероятность каждого из этих двух случаев $\frac{1}{8}$, так как у нас 3 стороны, и каждая имеет данный цвет с вероятностью $\frac{1}{2}$. Тогда $E(X_{123}) = \frac{1}{4}$, а значит, $E(X) = 4 * \frac{1}{4} = 1$.

3 Дисперсия

Кроме среднего значения случайной величины интересно также то насколько сильно она в среднем отличается от своего среднего значения.

Пример 7. Допустим, мы хотим понять насколько хороша винтовка. Для этого поступают следующим образом: фиксируют винтовку (иногда просто держат ее в положении лежа, но лучше в тисках) на определенном расстоянии от мишени (20-25 м) и много раз стреляют по мишени. При этом винтовку фиксируют так, что идеальная линия выстрела точно смотрит в центр мишени. Обозначим через X, Y – координаты места попадания пули. Можно рассматривать X, Y как случайные величины. Тогда средние значения X, Y можно считать очень близкими к координатам центра мишени (x_0, y_0), то есть $E(X) = x_0, E(Y) = y_0$. Но тем не менее пули все равно попадают в центр не точно, просто влево улетает столько же, сколько и вправо, а вверх столько же сколько и вниз. Данный выстрел тем точнее чем меньше расстояние от места попадания до центра мишени. И в качестве меры точности винтовки принимают среднее расстояние до мишени в длинной серии выстрелов. Эта характеристика называется кучностью винтовки.

А вот другой пример, когда нам интересно не только среднее значение случайной величины, но и то насколько она от него отклоняется.

Пример 8. Допустим, что нужно выбрать одно из двух предприятий, в которое мы бы вложили пенсионные деньги. В такого рода вложении нужно максимально избегать риска и вкладывать в наиболее стабильное предприятие. Пусть доходности предприятий на протяжении пяти лет такие как в таблице.

Таблица 3: доходность предприятий во второй отрасли

год	2008	2009	2010	2011	2012
доходность первой фирмы	95	110	105	120	120
доходность второй фирмы	115	104	112	111	108

Среднее значение доходности у обоих предприятий равно 110, но видно, что в первом случае доходность сильнее колеблется вокруг среднего значения, иногда доходность очень велика, иногда очень низка. Поэтому менее рискованным будет вложение во второе предприятие.

Ясно, что мера разброса случайной величины связана с тем, как в среднем ведет себя разность $X - E(X)$. Может быть взять $E(X - E(X))$? Но это будет $E(X) - E(X) = 0$ по свойству линейности. Нужно чтобы отрицательные разности не компенсировались положительными. В этом смысле лучше $E(|X - E(X)|)$. Но проблема в том, что с функцией $|x|$ не очень удобно работать, потому что всегда нужно думать какой знак у x. Поэтому за меру разброса принимают $E((X - E(X))^2)$.

Определение 3. Дисперсия случайной величины X это $E((X - E(X))^2)$. Дисперсия X обозначается $D(X)$.

Совсем неформально можно сказать, что дисперсия – мера разброса случайной величины вокруг среднего значения. Говоря точнее, это среднее значение квадрата разности случайной величины со средним значением.

4 Независимые случайные величины и свойства дисперсии

Первое свойство дисперсии заключается в том, что если X умножить на константу c, то $D(cX) = c^2 D(x)$.

Чему равна $D(X + Y)$? Распишем подробнее.

$$\begin{aligned} D(X+Y) &= E((X+Y-E(X+Y))^2) = E((X-E(X)+Y-E(Y))^2) = \\ &= E((X-E(X))^2 + (Y-E(Y))^2 + 2*(X-E(X))*(Y-E(Y))) = \\ &= D(X) + D(Y) + 2*E(XY - XE(Y) - E(X)Y + E(X)E(Y)) \end{aligned}$$

Теперь заметим, что $E(X)$ и $E(Y)$ – константы, и их можно выносить из под мат. ожидания. Так что выражение выше можно переписать так:

$$\begin{aligned} D(X) + D(Y) + 2*(E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y)) = \\ = D(X) + D(Y) + 2*(E(XY) - E(X)E(Y)). \end{aligned}$$

Определение 4. Величина $E((X - E(X))(Y - E(Y))) = E(XY) - E(X)E(Y)$ называется ковариацией случайных величин X и Y и обозначается $\mathrm{cov(X, Y)}$.

Таким образом, получаем формулу:

$$D(X + Y) = D(X) + D(Y) + 2 * \mathrm{cov(X, Y)} \tag{3}$$

Так что, вообще говоря, неверно считать, что дисперсия суммы – это сумма дисперсий. Например, если взять пару X и X, то получим что $D(2X) = 4D(X)$, и $\mathrm{cov(X, X)} = \mathrm{D(X)}$. Но сейчас мы рассмотрим случай, когда для дисперсии свойство сложения верно.

Определение 5. Случайные величины X и Y называются независимыми, если события $X = x_k$ (то есть событие, что случайная величина X принимает значение x_k) независимо от событий вида $Y = y_l$.

В случае если величины X и Y независимы $p(X = x, Y = y) = p(X = x)p(Y = y)$. Для таких X и Y получаем, что

$$\begin{aligned} E(XY) = \sum_{k,l} x_k y_l p(X = x_k, Y = y_l) = \sum_{k,l} x_k p(X = x_k) y_l p(Y = y_l) = \\ = \sum_k x_k p(X = x_k) \sum_l y_l p(Y = y_l) = E(X)E(Y) \end{aligned}$$

То есть $\mathrm{cov(X, Y)} = 0$, если X и Y независимы. А значит, если X и Y независимы, то $D(X + Y) = D(X) + D(Y)$. Пользуясь этим свойством можно вычислять дисперсию во многих важных случаях.

Пример 9. Рассмотрим случайную величину X – количество выпавших в n бросках орлов для монеты с вероятностью выпадения орла в одном броске p. Тогда как мы уже видели $X = X_1 + X_2 + \cdots + X_n$, где X_i – количество выпавших орлов в i-м броске. $D(X_i) = p * (1 - p)^2 + (1 - p) * p^2 = p(1 - p)$. Значит, $D(X) = np(1 - p)$, так как X_i – независимые случайные величины и их дисперсии складываются.