Слайд 1. Тема 5. Элементы статистики. Подготовка и исследование данных

5.1. Основы теории вероятностей

5.1.1. Основные понятия и определения

Наука о данных и машинное обучение подразумевают знания по теории вероятностей и математической статистике. Теория вероятностей и математическая статистика – два обширных раздела математики, основы которых трудно изложить даже в нескольких лекциях.

Теория вероятностей – математическая наука, изучающая закономерности в массовых случайных явлениях.

Рассмотрим некоторый эксперимент, все мыслимые исходы которого описываются конечным числом различных исходов. Любое возможное множество исходов опыта (эксперимента) мы будем называть *случайным событием*.

Случайным событием называется такое событие, которое может как произойти, так и не произойти. Примером события могут служить сделки по покупке–продаже валюты. Мы знаем, что количество сделок в разные дни и в разное время суток различается и является случайным.

По общепринятой практике события рассматриваются только в дискретные моменты времени – минута, час, день и т.д.

Случайные события, как правило, обозначаются заглавными буквами латинского алфавита: A, B, C и т.д.

При подбрасывании игрального кубика случайные события соответствуют числам, выпадающим на кубике. Например, событие A={При подбрасывании игрального кубика на верхней грани выпала двойка}.

Несколько событий образуют *полную группу событий*, если в результате опыта непременно должно появиться хотя бы одно из них. Например, появления герба или решки при подбрасывании монеты.

Несколько событий называются *несовместными* в данном эксперименте, если никакие два из них не могут появиться одновременно. Например, появления герба и решки одновременно при подбрасывании монеты.

Слайд 2. *Противоположными* событиями называются два несовместных события, образующие полную группу. Событие, противоположное событию A, обозначают $\bar{A}$. Например, появления герба и появление решки при подбрасывании монеты.

Достоверным событием называется такое событие, которое обязательно произойдет в результате опыта.

Невозможным событием называется такое событие, которое никогда не произойдет в результате опыта.

Произведением двух событий A и B называется событие C, состоящее в совместном осуществлении события A и события B. Обозначение $C=AB$.

Суммой двух событий A и B называется событие C, состоящее в осуществлении хотя бы одного из событий A или B. Обозначение $C=A+B$. При этом те исходы, которые входят и в A и в B считаются только один раз.

Для количественного сравнения между собой событий по степени их возможности вводится численная мера, называемая *вероятностью события* и характеризующая объективную возможность появления этого события. Вероятность события A обозначается $P(A)$.

Непосредственный расчет вероятностей случайного события осуществляется для событий, представляющих собой случаи.

Случаи – это события, обладающие тремя свойствами: они образуют полную группу, несовместны и равновозможны.

Случай называется *благоприятным* некоторому событию, если появление этого случая влечет за собой появление данного события.

Слайд 3. Определение. Под вероятностью $P(A)$ события A понимается отношение числа равновозможных случаев, благоприятствующих событию A, к общему числу всех равновозможных случаев.

Если m – число случаев, благоприятных событию A, а n – общее число случаев, то вероятность события A вычисляется по формуле:

$$P(A)=\frac{m}{n}.$$

Данная формула называется *классической формулой* для вычисления вероятностей.

Пример 1. В урне находится a белых и b черных шаров. Из урны наудачу вынимают один шар. Найти вероятность того, что этот шар белый.

Р е ш е н и е. Обозначим через A следующее событие:

$$A=\{\text{появление белого шара}\}.$$

Общее число случаев $n=a+b$; число случаев, благоприятных событию A, $m=a$. Следовательно,

$$P(A)=\frac{a}{a+b}.$$

Поскольку $0 \le m \le n$, то $0 \le P(A) \le 1$, т.е. вероятность любого события есть неотрицательное число, не превышающее единицы.

Слайд 4. *Основные свойства вероятности*:

1) Вероятность случайного события – величина безразмерная.

2) $0 \le P(A) \le 1$ – вероятность любого события лежит на отрезке от нуля до единицы.

3) Вероятность достоверного события, т.е. события которое в результате опыта обязательно произойдет ($m = n$), равна единице.

4) $P(\varnothing) = 0$ – вероятность невозможного события ($m = 0$) равна нулю.

5) $P(A) + P(\overline{A}) = 1$ – сумма вероятностей самого события и ему противоположного равна единице.

Основные теоремы теории вероятностей случайных событий

Основные теоремы позволяют рассчитать вероятности одних событий по вероятности других, с которыми первые связаны. Они включают в себя: теорему сложения вероятностей, теорему умножения вероятностей, формулу полной вероятности и формулу Байеса.

Введем ряд понятий.

Два события называют *независимыми*, если вероятность одного из них не зависит от появления или непоявления другого.

Пример 2. Монета брошена 2 раза. Вероятность появления герба в первом испытании (событие A) не зависит от появления или непоявления герба во втором испытании (событие B). В свою очередь, вероятность выпадения герба во втором испытании (событие B) не зависит от результата первого испытания. Таким образом, события A и B независимые.◄

Слайд 5. Два события называют *зависимыми*, если вероятность одного из них зависит от наступления или ненаступления другого события.

Пример 3. В ящике 50 деталей. Из них 30 стандартных и 20 нестандартных. Наудачу берут одну деталь, не возвращая ее в ящик.

Если появилась стандартная деталь (событие A), то вероятность извлечения стандартной детали при втором испытании (событие B) составляет $P(B) = \frac{29}{49}$. Если же в первом испытании вынута нестандартная деталь, то вероятность события B составляет $P(B) = \frac{30}{49}$.

Таким образом, вероятность появления события B зависит от наступления или не наступления события A. События A и B – зависимые.◀

Вероятность события A, вычисленная при условии, что имело место другое событие B, называется *условной вероятностью* события A и обозначается $P(A/B)$.

Теорема умножения вероятностей. Вероятность произведения двух событий равна произведению вероятностей одного из них на условную вероятность другого, вычисленную при условии, что первое имело место:

$$P(A \cdot B) = P(A) \cdot P(B/A).$$

Слайд 6. Из данной формулы можно получить формулу для вычисления условной вероятности:

$P(B|A) = \dfrac{P(A \cdot B)}{P(A)}$ – вероятность событиz B при условии, что произошло событие A.

Теорема сложения вероятностей. Вероятность суммы двух несовместных событий равна сумме вероятностей этих событий:

$$P(A+B) = P(A) + P(B).$$

Два события называют *совместными*, если появление одного из них не исключает появления другого в одном и том же испытании.

Вероятность суммы двух совместных событий выражается формулой:

$$P(A+B) = P(A) + P(B) - P(AB).$$

Пример 4. В урне 2 белых и 3 черных шара. Из урны вынимают подряд два шара. Найти вероятность того, что оба шара белые.

Р е ш е н и е. Обозначим через A следующее событие:

A={появление двух белых шаров}.

Событие A представляет собой произведение двух событий $A = A_1 \cdot A_2$, где

A_1={появление белого шара при первом вынимании};

A_2={появление белого шара при втором вынимании}.

Слайд 7. По теореме умножения вероятностей имеем:

$$P(A) = P(A_1) \cdot P(A_2/A_1) = \frac{2}{5} \cdot \frac{1}{4} = 0{,}1.$$

Пример 5. Те же условия, но после первого вынимания шар возвращается в урну и шары перемешиваются.

Р е ш е н и е. В данном случае события A_1 и A_2 независимы и

$$P(A) = P(A_1) \cdot P(A_2) = \frac{2}{5} \cdot \frac{2}{5} = 0{,}16.$$

Пример 6. В лотерее 1000 билетов; из них на один падает выигрыш 500 руб., на 10 билетов – выигрыши по 100 руб., на 50 билетов – выигрыши по 20 руб., на 100 билетов – выигрыши по 5 руб.; остальные билеты невыигрышные. Вы покупаете один билет. Найти вероятность выиграть не менее 20 руб.

Р е ш е н и е. По условию задачи событие {выиграть 5 руб. } нас не интересует. Рассмотрим следующие события:

A={выиграть не менее 20 руб.};

A_1={выиграть 20 руб.};

A_2={выиграть 100 руб.};

A_3={выиграть 500 руб.}.

Слайд 8. Приведенные события несовместны.

Очевидно, что $A = A_1 + A_2 + A_3$.

По теореме сложения вероятностей для несовместных событий имеем:

$$P(A) = P(A_1 + A_2 + A_3) = P(A_1) + P(A_2) + P(A_3) = \frac{50}{1000} + \frac{10}{1000} + \frac{1}{1000} = 0{,}061.$$

Формула полной вероятности и формула Байеса

Одним из эффективных методов подсчета вероятностей является *формула полной вероятности,* с помощью которой решается широкий круг задач. Пусть требуется определить вероятность некоторого события A, которое может произойти вместе с одной из исключающих друг друга гипотез (предположений) $H_1, H_2, ..., H_n$, образующих полную группу.

Тогда *вероятность события A равна сумме произведений вероятностей каждой из этих гипотез на соответствующую условную вероятность события A при этой гипотезе:*

$$P(A) = \sum_{i=1}^{n} P(H_i) P(A / H_i),$$

где $P(H_i)$ –вероятность гипотезы H_i, $P(A/H_i)$ – условная вероятность события A при этой гипотезе. Данная формула носит название *формулы полной вероятности*.

Слайд 9. **Пример 7.** Имеются три одинаковые на вид урны. В первой урне два белых и один черный шар; во второй – три белых и один черный; в третьей – два белых и два черных шара. Некто выбирает наугад одну из урн и вынимает из нее шар. Найти вероятность того, что этот шар белый.

Р е ш е н и е. Рассмотрим три гипотезы:

H_1 – выбор первой урны;

H_2 – выбор второй урны;

H_3 – выбор третьей урны.

Событие A={появление белого шара}. Так как гипотезы по условию задачи равновозможны, то

$$P(H_1) = P(H_2) = P(H_3) = 1/3.$$

Условные вероятности события A при этих гипотезах соответственно равны:

$$P(A / H_1) = 2/3;\ P(A / H_2) = 3/4;\ P(A / H_3) = 1/2.$$

По формуле полной вероятности

$$P(A) = 1/3 \cdot 2/3 + 1/3 \cdot 3/4 + 1/3 \cdot 1/2 = 23/36.$$

Теорема Байеса. Если до опыта вероятности гипотез $H_1, H_2, ..., H_n$, образующих полную группу несовместных событий, были известны и равны соответственно $P(H_1), P(H_2), ..., P(H_n)$, а в результате опыта произошло событие A, то новые (условные) вероятности гипотез вычисляются по формуле:

Слайд 10.

$$P(H_i / A) = \frac{P(H_i)P(A/H_i)}{\sum_{i=1}^{n} P(H_i)P(A/H_i)}, i = \overline{1,n}$$

Эта формула называется *формулой Байеса*.

Пример 8. Прибор может собираться из высококачественных деталей и из деталей обычного качества. Вообще около 40% приборов собираются из высококачественных деталей. Если прибор собран из высококачественных деталей, его надежность (вероятность безотказной работы) за время t равна 0,95; если из деталей обычного качества – его надежность равна 0,7. Прибор испытывался в течение времени t и работал безотказно. Найти вероятность того, что он собран из высококачественных деталей.

Р е ш е н и е. Возможны две гипотезы:

H_1 – прибор собран из высококачественных деталей;

H_2 – прибор собран из деталей обычного качества.

Вероятность этих гипотез до опыта $P(H_1) = 0.4$, $P(H_2) = 0.6$. В результате опыта наблюдается событие A={прибор безотказно работал время t}. Условные вероятности этого события при гипотезах H_1 и H_2 равны:

Слайд 11.

$$P(A/H_1)=0{,}95; P(A/H_2)=0{,}7.$$

По формуле Байеса находим вероятность гипотезы H_1 после опыта:

$$P(H_1/A)=\frac{0{,}4\cdot 0{,}95}{0{,}4\cdot 0{,}95+0{,}6\cdot 0{,}7}=0{,}475.$$ ◀

Примечание. *Относительной частотой случайного события* называется отношение количества случаев появления этого события m к общему числу равновозможных проведенных наблюдений n. При росте числа наблюдений ($n\to\infty$), до бесконечности, частота стремится к числу, называемому **вероятностью** случайного события: Обычно вместо относительной частоты часто употребляют слово вероятность.

Слайд 12.

5.1.2. Случайные величины и законы их распределения

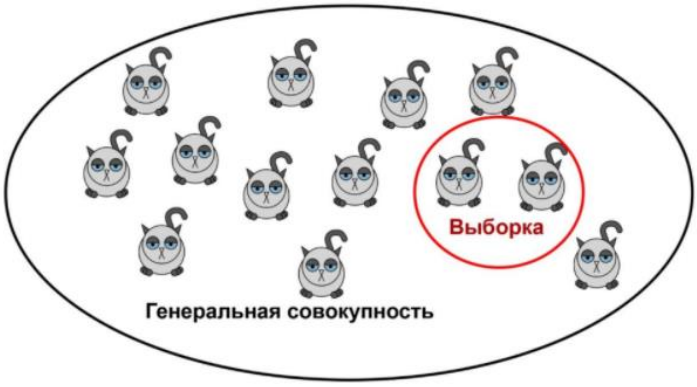
Случайной величиной называется такая величина *X*, которая в результате опыта принимает числовое значение, заранее неизвестно какое. Случайная величина характеризуется значениями, которые она может принимать, и вероятностями, с которыми эти значения принимаются.

- Случайная величина – численное выражение исхода некоторого случайного события;
- Реализация случайной величины – это появление некоторого случайного события в результате опыта;
- Набор реализаций случайной величины называется *выборкой* из нее.

Если проводить эксперимент со случайной величиной бесконечно, то каждому событию можно будет поставить в соответствие его вероятность – долю испытаний, завершившихся наступлением некоторого события.

Генеральной совокупностью называют все возможные значения, которые может принимать случайная величина. Например, на рынке ценных бумаг всегда имеют дело с выборкой из генеральной совокупности – обычно берут котировки за некоторый промежуток времени. Естественно, что статистика, вычисленная по выборке, отличается от статистики, вычисленной на генеральной совокупности, так как относительная частота отличается от вероятности, и смыслом дальнейших вычислений является оценка отличий статистик, вычисленных на выборке, от статистик, вычисленных на генеральной совокупности.

Слайд 13.



1. Дискретные случайные величины и законы их распределения

Случайная величина X называется *дискретной*, если множество ее возможных значений конечно или счетно, т.е. которые можно установить и перечислить.

Примеры дискретных случайных величин:

1) число студентов в аудитории может быть только целым положительным числом: 0, 1, 2, 3, 4… 20…

2) число событий, происходящих за одинаковые промежутки времени: частота пульса, число вызовов скорой помощи за час, количество операций в месяц с летальным исходом и т.д.

Для описания случайной величины можно использовать набор вероятностей, с которыми она принимает то или иное значение, и функцию распределения.

Соответствие между возможными значениями дискретной случайной величины и их вероятностями называется **законом распределения** этой величины.

Слайд 14.

Простейшей формой закона распределения дискретной случайной величины X является *ряд распределения* – таблица, в верхней строке которой перечислены все значения случайной величины $x_1, x_2, ..., x_n$ в порядке их возрастания, а в нижней – соответствующие им вероятности $p_1, p_2, ..., p_n$. Таблица имеет вид:

x_i	x_1	x_2	…	x_n	…
p_i	p_1	p_2	…	p_n	…

где $p_i = p(X = x_i)$; $\sum_{i=1}^{\infty} p_i = 1$.

Закон распределения вероятностей можно представить в виде функции распределения вероятностей случайной величины, которая может использоваться как для дискретных, так и для непрерывных случайных величин.

Функция распределения – это функция, характеризующая вероятность:

$$F_\xi(x) = P(\xi < x).$$

т.е. $F(x)$ есть вероятность того, что случайная величина ξ примет значение, меньшее, чем x.

Например, $F(-1) = P(x < -1), F(3) = P(x < 3)$.

Слайд 15.

Свойства функции распределения:

- $0 \le F_{\xi}(x) \le 1$,
- $F_{\xi}(x)$ – неубывающая функция,
- если $a<b$, то $P(a \le X \le b) = F(a) - F(b)$.

Функция распределения дискретной случайной величины определяется посредством равенства $F(x) = \sum_{i} p_i$, где суммирование распространяется на все индексы, для которых $x_i < x$.

График представляет собой ступенчатую ломаную линию со скачками в точках $x_1, x_2, ..., x_n$. Величины скачков равны соответственно $p_1, p_2, ..., p_n$. Левее x_1 график совпадает с осью Ox, правее x_n – с прямой $F(x) = 1$.

Пример. Дискретная случайная величина задана таблицей

x_i	0	1	2	3
p_i	0,1	0,3	0,4	0,2

Слайд 16.

Найти функцию распределения вероятностей случайной величины X и построить ее график. Чему равно значение функции распределения в точке $x = \sqrt{2}$?

Р е ш е н и е. Построим функцию распределения согласно формуле

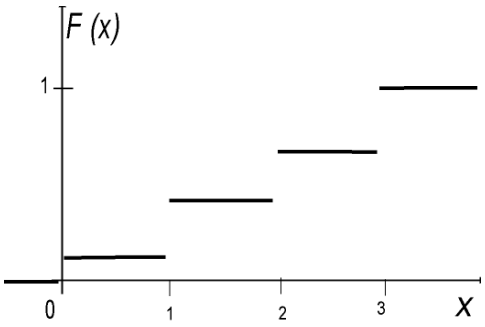
$$F(x) = P(X < x) = \sum_{x_i < x} p_i .$$

Подставляя сюда вероятности из таблицы распределения, имеем

$$F(x) = \begin{cases} 0 & , при\ x \le 0, \\ 0{,}1 & , при\ x \le 1, \\ 0{,}1 + 0{,}3 = 0{,}4 & , при\ 1 < x < 2, \\ 0{,}1 + 0{,}3 + 0{,}4 = 0{,}8 & , при\ 2 < x \le 3, \\ 0{,}1 + 0{,}3 + 0{,}4 + 0{,}2 = 1 & , при\ x > 3. \end{cases}$$

График этой функции имеет вид:

Слайд 17.



Значение функции распределения в точке $x=\sqrt{2}$ попадает в интервал $1<x<2$, следовательно, значение функции равно 0,4.

Основные дискретные распределения:

Биномиальное распределение
Геометрическое распределение
Гипергеометрическое распределение
Полиномиальное (мультиномиальное) распределение
Распределение Пуассона
Распределения Кокса

Рассмотрим некоторые наиболее часто встречающиеся на практике распределения дискретных случайных величин.

Биномиальное распределение. Дискретная случайная величина μ распределена по биномиальному закону, если она принимает значения 0, 1, 2,..., n в соответствии с рядом распределения, представленным в таблице

μ	0	1	...	i	...	n
p	q^n	$C_n^1 pq^{n-1}$	...	$C_n^i p^i q^{n-i}$	...	p^n

где $0<p<1, 0<q<1, p+q=1$.

Слайд 18.

Пуассоновское распределение. Дискретная случайная величина η распределена по закону Пуассона, если она принимает целые неотрицательные значения с вероятностями, представленными рядом распределения

η	0	1	2	...	n	...
p	$e^{-\lambda}$	$\lambda e^{-\lambda}$	$\frac{\lambda^2}{2!}e^{-\lambda}$	...	$\frac{\lambda^n}{n!}e^{-\lambda}$	...

Распределение Пуассона носит также название *закона редких событий*, поскольку оно всегда появляется там, где производится большое число испытаний, в каждом из которых с малой вероятностью происходит “редкое“ событие. По закону Пуассона распределены, например, число вызовов, поступивших на телефонную станцию; число метеоритов, упавших в определенном районе; число распавшихся нестабильных частиц и т.д.

Геометрическое распределение. Пусть ν – число испытаний, которые надо провести, прежде чем появится первый успех. Тогда ν – дискретная случайная величина, принимающая значения 0,1,2,...,*n*,... . Ряд распределения такой случайной величины задан таблицей

ν	0	1	2	...	n	...
p	p	qp	$q^2 p$	...	$q^n p$	...

Случайная величина с таким рядом распределения называется распределенной по геометрическому закону.

Числовые характеристики дискретной случайной величины

Числовые характеристики случайных величин позволяют выразить наиболее существенные особенности распределения.

Закон распределения является исчерпывающей характеристикой случайной величины, но далеко не в каждой задаче нужно знать весь закон распределения. В ряде случаев можно обойтись одним или несколькими числами, отражающими наиболее важные особенности закона распределения: например, числом, имеющим смысл “среднего значения” случайной величины, или же числом, характеризующим средний размер отклонения случайной величины от своего среднего значения, и т.д. Такого рода числа называют *числовыми характеристиками* случайной величины. Их роль в теории вероятностей чрезвычайно велика: многие задачи удается решить, оставляя в стороне законы распределения и оперируя только числовыми характеристиками.

Наиболее важное место среди числовых характеристик занимает *математическое ожидание* (или *среднее значение*) случайной величины.

Слайд 19.

Определение. *Математическим ожиданием* или средним значением дискретной случайной величины *X* с законом распределения

x_1	x_2	…	x_n	…
p_1	p_2	…	p_n	…

называется число

$$M[X] = m_x = x_1 p_1 + x_2 p_2 + \ldots = \sum_{i=1}^{\infty} x_i p_i ,$$

т.е. математическое ожидание дискретной случайной величины X равно сумме произведений возможных значений величины X на их вероятности p_i.

Смысл числа $M[X]$ заключается в том, что вокруг него колеблется среднее арифметическое значений, принимаемых величиной X, в больших сериях опытов. Число $M[\mathrm{X}]$ часто называют *центром распределения* случайной величины X.

Свойства математического ожидания

1. Математическое ожидание постоянной величины C равно ей самой:

$$M[C] = C .$$

2. Постоянный множитель C можно выносить за знак математического ожидания:

$$M[C X] = C \cdot M[X].$$

3. Математическое ожидание суммы двух случайных величин равно сумме их математических ожиданий:

$$M[Y + X] = M[Y] + M[X].$$

4. Если случайные величины X и Y независимы, то математическое ожидание их произведения равно произведению их математических ожиданий, т.е.

$$M[Y \cdot X] = M[X] \cdot M[Y].$$

Слайд 20. Введем еще одну числовую характеристику для измерения *степени рассеивания*, *разброса* значений, принимаемых случайной величиной X, вокруг ее математического ожидания.

Определение. *Дисперсией случайной величины X* называют число

$$D[X] = D_x = M[(X - m)^2] = \sum_{i=1}^{\infty} (x_i - m_x)^2 \cdot p_i ,$$

т.е. дисперсия есть математическое ожидание квадрата отклонения.

Величина

$$\sigma[X] = \sigma_x = \left| \sqrt{D[X]} \right|$$

носит название *среднего квадратического отклонения.*

Свойства дисперсии:

1. Дисперсия постоянной величины равна нулю:

$$D[C]=0.$$

2. При умножении случайной величины X на постоянное число C ее дисперсия умножается на C^2.

3. Если величины X и Y независимы, то дисперсия их суммы равна сумме их дисперсий:

$$D[X+Y]=D[X]+D[Y].$$

Слайд 21. **2. Непрерывные случайные величины и законы их распределения**

Случайная величина называется *непрерывной*, если множество ее возможных значений является несчетным и непрерывно заполняет некоторый промежуток. Например, серию измерений температуры, массу тела, массу и объем мозга можно описать непрерывной случайной величиной.

Функция распределения $F(x) = P(X < x)$ для непрерывной случайной величины определяется так же, как и для дискретной. В предыдущем разделе показано, что функция распределения дискретной случайной величины X изменяет свои значения только скачками. Функция распределения непрерывной случайной величины не имеет скачков. В связи с этим можно дать другое определение непрерывной случайной величины.

Определение. Случайная величина X называется непрерывной, если ее функция распределения $F(x)$ непрерывна при всех значениях X.

Плотностью распределения вероятностей непрерывной случайной величины называют первую производную от функции распределения

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

Свойства плотности распределения:

$$1)\ f(x) \geq 0;$$

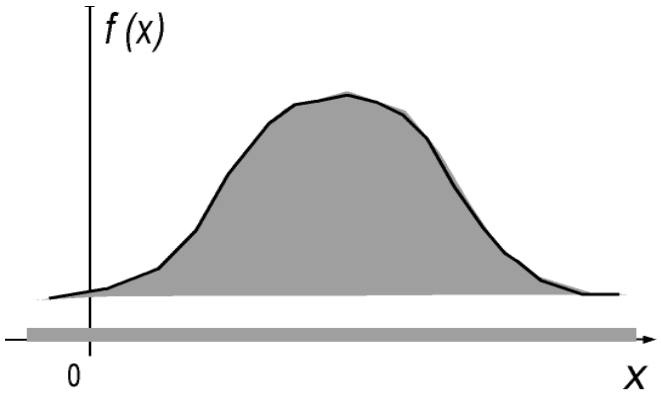
$$2)\ \int_{-\infty}^{\infty} f(x)dx = 1.$$

Плотность распределения называется также *дифференциальной формой закона распределения.* Зная плотность распределения, можно найти функцию распределения

Слайд 22.

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Функция $f(x)$ изображается графически кривой, лежащей в верхней полуплоскости и такой, что площадь, заключенная между ней и осью Ox, равна 1.



Чтобы посчитать вероятность того, что случайная величина лежит в промежутке (a, b), нужно взять интеграл от плотности распределения случайной величины:

$$P(a < X < b) = \int_a^b f(x)dx = F(a) - F(b).$$

Графически вероятность $P(a < X < b)$ равна площади криволинейной трапеции, ограниченной осью абсцисс, кривой $f(x)$ и прямыми $x = a$ и $x = b$. Кривая $f(x)$ при этом называется кривой распределения.

Если известна функция $f(x)$, то, изменяя пределы интегрирования, можно найти вероятность для любых интересующих нас интервалов. Поэтому именно задание функции $f(x)$ полностью определяет закон распределения для непрерывных случайных величин.

Для плотности вероятности $f(x)$ должно выполняться условие:

$$\int_a^b f(x)dx = 1 \text{ или } \int_{-\infty}^{+\infty} f(x)dx = 1.$$

Слайд 23. В качестве примера непрерывной случайной величины можно рассмотреть **равномерное распределение**: случайная величина на отрезке [a, b] принимает любое значение с одной и той же вероятностью. Плотность вероятности равномерного распределения задается следующим образом:

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & x \in [a; b], \\ 0, & x \notin [a; b]. \end{cases}$$

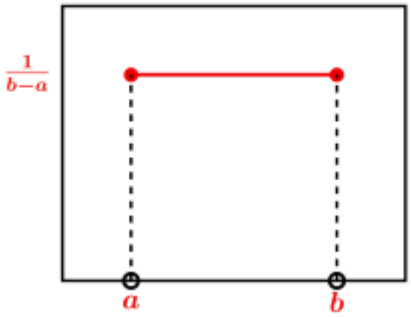


График $f(x)$ равномерного распределения

Другим примером непрерывной случайной величины является нормально распределенная случайная величина.

Пример: человек любит приходить на работу к 11 часам, но точное время его прихода варьируется – иногда он приходит раньше, иногда – опаздывает. Таким образом, точное время его прихода на работу X представляет собой результат взаимодействия большого количества слабо зависимых случайных факторов. Именно такие величины хорошо моделируются **нормальным (Гауссовым) распределением**:

Слайд 24.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$f(x)$ – функция плотности вероятности нормального распределения; μ – отвечает за среднее время прихода; параметр σ определяет разброс вокруг среднего.

Варьируя значения параметров μ и σ, можно влиять на форму графика функции плотности вероятности нормального распределения (см. рис.).

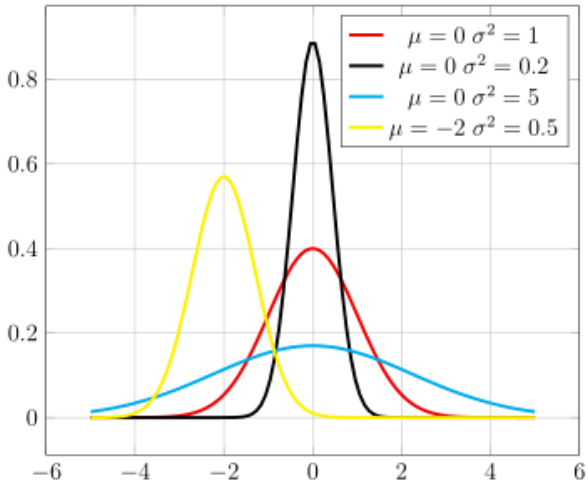


График плотности нормального распределения

Кривая распределения по нормальному закону имеет симметричный вид. Центром симметрии распределения является центр рассеивания μ.

Слайд 25. *Математическое ожидание* и *дисперсия* для непрерывной случайной величины определяются по формулам:

$$m_x = \int_{-\infty}^{\infty} x \cdot f(x)dx,$$

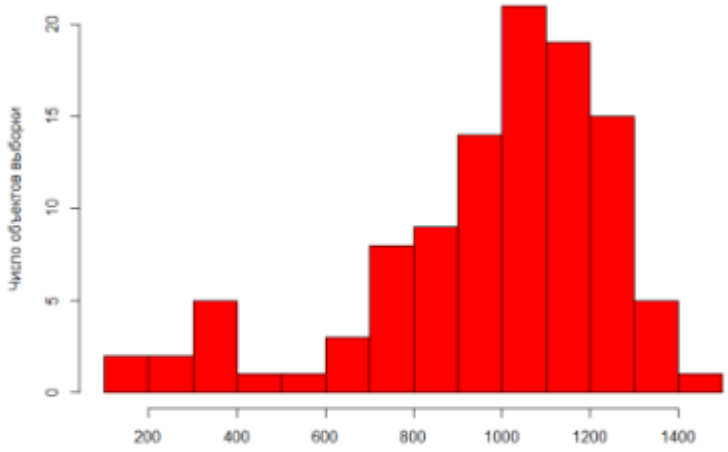
$$D_x = \int_{-\infty}^{\infty} (x - m_x)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - \left(\int_{-\infty}^{\infty} x \cdot f(x)dx\right)^2.$$

Свойства математического ожидания и дисперсии непрерывной случайной величины совпадают со свойствами этих характеристик для дискретных случайных величин.

Визуализация распределений случайных величин

Для визуализации особенностей распределения с.в. строят гистограмму. Для этого область определения случайной величины разбивают на интервалы одинаковой длины. Количество объектов выборки в каждом интервале будет пропорционально среднему значению плотности на нем.

Слайд 26.



Гистограмма

По данной гистограмме хорошо видны все особенности распределения данных: оно бимодально, основной пик приходится примерно на значение 1000.

Важным аспектом работы с гистограммами является правильный выбор числа интервалов.

Слайд 27.

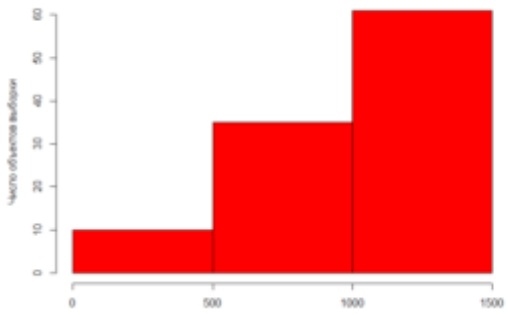


Рис. 1

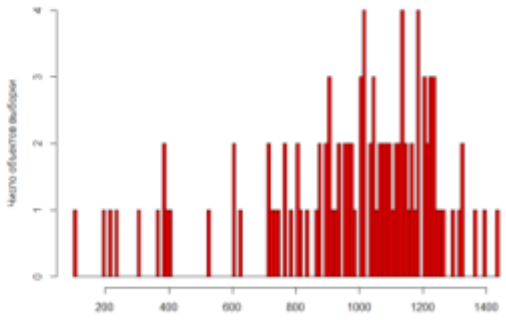


Рис. 2

Если рассмотреть слишком мало интервалов, то они будут слишком большими, в результате гистограмма получится грубой (рис. 1). Аналогично в случае слишком большого количества интервалов – в большую часть из них не попадет ни одного объекта выборки (рис. 2). В обоих случаях построенные гистограммы не являются информативными.

Число интервалов для гистограммы вычисляется по формуле:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n}.$$

Слайд 28. Ядерная оценка плотности распределения

Ядерная оценка плотности распределения является задачей сглаживания данных, когда делается заключение о совокупности, основываясь на конечных выборках данных. **Ядерная оценка плотности распределения** – это всего лишь подгонка распределения вероятности.

Пусть $(x_1, x_2, ..., x_n)$ – одномерная выборка независимых одинаково распределенных случайных величин, извлеченных из некоторого распределения с неизвестной плотностью $f(x)$. Требуется оценить форму функции $f(x)$. Данная задача решается с помощью функции плотности (1):

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \qquad (1)$$

где x – последовательность, длиной n, K – неотрицательная функция, называемая ядром, h – диапазон, сглаживающий параметр, называемый *шириной полосы*. Ядро с индексом h называется *взвешенным ядром* и определяется по формуле:

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right). \qquad (2)$$

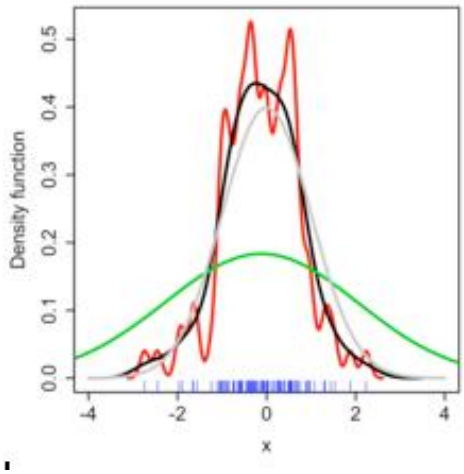
Интуитивно стараются выбрать h как можно меньше, насколько данные это позволяют.

Вследствие удобных математических свойств часто используется нормальное (Гауссово) ядро, среднее которого $K(x) = \varphi(x)$, где $\varphi(x)$ является стандартной нормальной функцией плотности:

Слайд 29.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2}. \quad (3)$$

Пример. Выбор ширины полосы *h*



Ядерная оценка плотности (*Kernel density estimate*, KDE)
с различной шириной полос случайной выборки 100 точек из стандартного нормального распределения.
Серая кривая – истинная плотность (стандартное нормальное распределение).
Красная кривая: KDE с **h=0,05**.
Чёрная кривая: KDE с **h=0,337**.
Зелёная кривая: KDE с **h=2**.

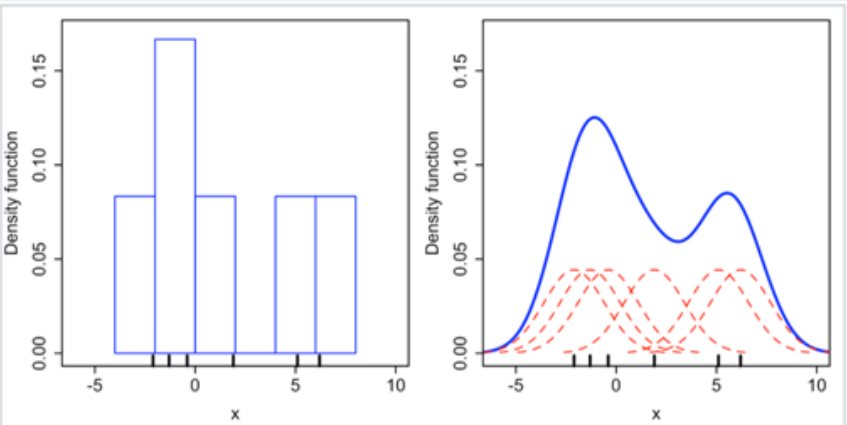
Серая кривая представляет истинную плотность (нормальная плотность со средним 0 и дисперсией 1). Для сравнения, красная кривая *недостаточно сглажена*, поскольку она содержит слишком много случайных выбросов, возникающих при использовании полосы пропускания *h*=0,05, которая слишком мала. Зелёная кривая *чрезмерно сглажена*, поскольку используемая полоса пропускания *h*=2 существенно скрывает структуру. Чёрная кривая с полосой пропускания *h*=0,337 считается оптимально сглаженной, поскольку её оценка плотности близка к истинной плотности.

Слайд 30. **Эмпирическое правило для выбора полосы пропускания**

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1{,}06 \cdot \sigma \cdot n^{-\frac{1}{5}}. \quad (4)$$

Данное правило следует применять с осторожностью, так как оно даёт сильно неточные оценки, когда плотность не близка к нормальной.

Ядерные оценки плотности тесно связаны с гистограммами, но могут быть наделены свойствами, такими как гладкость или непрерывность, путём выбора подходящего ядра. Чтобы это увидеть, сравним построение гистограммы и ядерной оценки плотности на 6 точках:



Сравнение гистограммы (слева) и ядерной оценки плотности (справа), построенных из тех же самых данных. 6 индивидуальных ядер показаны красными пунктирными линиями, ядерная оценка плотности показана синей кривой. Точки данных показаны чёрточками на ленточной диаграмме по горизонтальной оси.

Слайд 31. Основные шаги практической реализации алгоритма ядерной оценки плотности

1. Производим оценку среднего значения и стандартного отклонения входной последовательности.

2. Производим нормализацию входной последовательности. Из каждого ее значения вычитаем ранее найденное среднее и делим на величину стандартного отклонения. После такой нормализации исходная последовательность будет иметь нулевое среднее и стандартное отклонение, равное единице. Непосредственно для вычисления плотности такая нормализация не обязательна, но она позволит унифицировать результирующие графики, так как для любой последовательности на шкале X будут располагаться значения, выраженные в единицах стандартного отклонения.

3. Находим максимальное и минимальное значение в нормализованной последовательности.

4. Создаем два массива, размер которых должен соответствовать желаемому количеству отображаемых на результирующем графике точек. Например, если график предполагается строить по 200 точкам, то размер массивов должен составлять соответственно по 200 значений.

5. Оставляем один из созданных массивов для хранения результата, а в другом сформируем значения точек, для которых будет производиться оценка плотности. Для этого в диапазоне между найденными ранее

максимальным и минимальным значениями сформируем 200 (для данного примера) равноотстоящих величин и сохраним их в подготовленном массиве.

6. Используя приведенное ранее выражение, произведем оценку плотности в 200 (для нашего примера) тестовых точках, сохраняя результат в подготовленном на шаге 4 массиве.