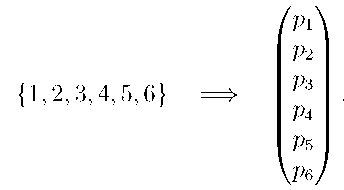
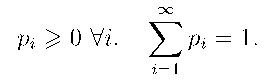
4. Дискретные случайные величины

В испытании с подбрасыванием кубика возможны шесть исходов. Эти исходы можно пронумеровать 1,2,3,4,5,6 согласно значению, выпавшему на кубике. Данному множеству исходов ставится в соответствие вектор вероятности:

image3image4image6image7

Именно так устроена дискретная случайная величина X. Она принимает счетное

множество значений

с вероятностями

Вероятности того, что

равна Pi, тогда функция вероятности имеет вид:

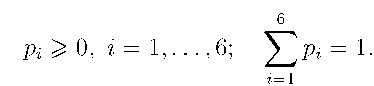
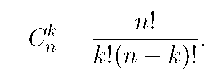
причем: 

image8Одной из наиболее часто встречающихся случайных величин является дискретная случайная величина с двумя исходами. Примером в данном случае является подбрасывание монеты. Можно обозначить выпадение решки за 1 («успех»), а выпадение орла — за 0 («неудача»). Пусть вероятность «успеха» равна р:

image9Именно так устроена бернуллиевская случайная величина:

image10Другим примером является сумма независимых бинарных случайных величин. Пусть вероятность попадания мяча в баскетбольное кольцо равна р, имеется п попыток, а число попаданий равно X. В силу независимости попыток:

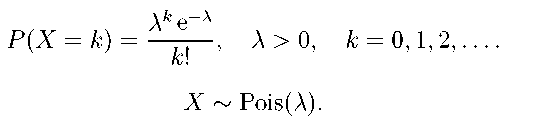
image11В общем случае:

Здесь используется биномиальный коэффициент,:

Такая случайная величина X называется биномиальной. Биномиальное распределе

image13image14ние имеет два параметра: целочисленный

Еще одним классом дискретных случайных величин являются счетчики. В качестве иллюстрации моясно рассмотреть текст романа Набокова «Приглашение на казнь». Из всех произведений Набокова составляется словарь, затем текст данного романа сверяют с этим словарем. Некоторые слова не встречаются в романе ни разу, например, слово «шахматист». Некоторые слова встречаются в романе редко, например, слово «аляповатость» (1 раз). Есть и слова, которые используются очень часто (например, «год» — 21 раз).

Пусть X — это число использований слова в тексте. Вероятность того, что X равно к, молено описать распределением Пуассона:

Распределением Пуассона описывается, например, число автобусов, которые проезжают за час мимо автобусной остановки, или число радиоактивных распадов, которые улавливает счетчик Гейгера.

5. Непрерывные случайные величины

image16Непрерывные случайные величины нельзя задавать с помощью функции вероятности в силу того, что если множество А несчетное, то вероятность события нулевая:

image17Первый способ определения таких величин — с помощью функции распределения  
(см. рис. 7):

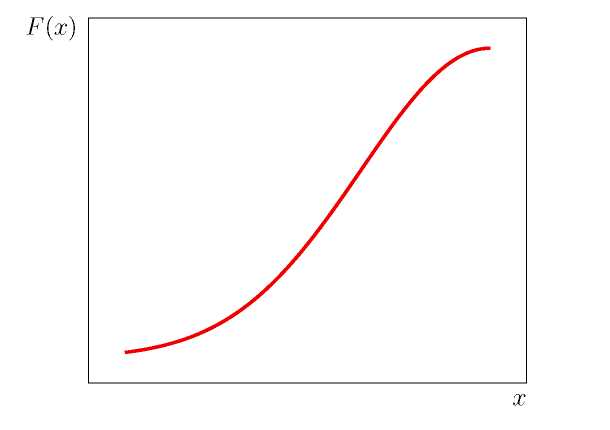
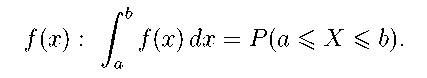
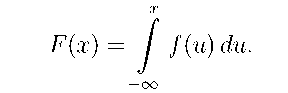
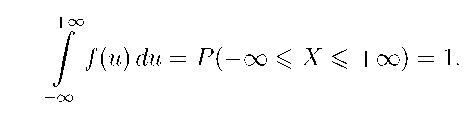


Рис. 7.

Функция распределения всегда принимает значение от 0 до 1 и не убывает по аргументу х.

Второй способ определения непрерывных случайных величин — с помощью плот- ност,и распределения:

Плотность связана с функцией распределения следующим образом:

Для непрерывной случайной величины верно равенство:

В отличие от функции распределения, плотность распределения на графике может принимать различный вид.

image22Примером непрерывной случайной величины является равномерная случайная величина. Пусть X — это время ожидания на светофоре до того, как можно будет перейти дорогу. Если на светофоре нет счетчика, то нельзя угадать, сколько именно придется ждать зеленого сигнала. Время ожидания может быть любым числом от О до, например, 30 секунд. Именно так устроено равномерное распределение — случайная величина на отрезке [а, Ь] принимает любое значение с одинаковой вероятностью:

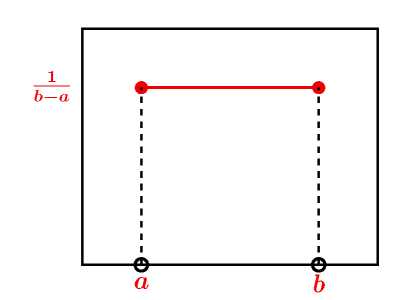
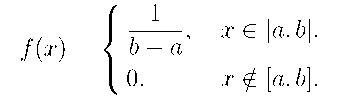


Рис. 8.

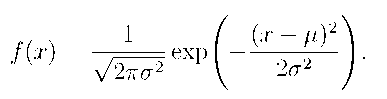
Плотность вероятности для равномерной случайной величины имеет вид (см. рис.

8):

Другим примером непрерывной случайной величины является нормальная случайная величина. Человек из примера любит приходить на работу к 11 часам, но точное время его прихода варьируется — иногда он проснется раньше, иногда он 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:

image26

В данном примере параметр ц отвечает за среднее время прихода, а параметр о определяет разброс вокруг среднего. Функция плотности вероятности нормалвного распределения имеет вид (ем. рис. 9):



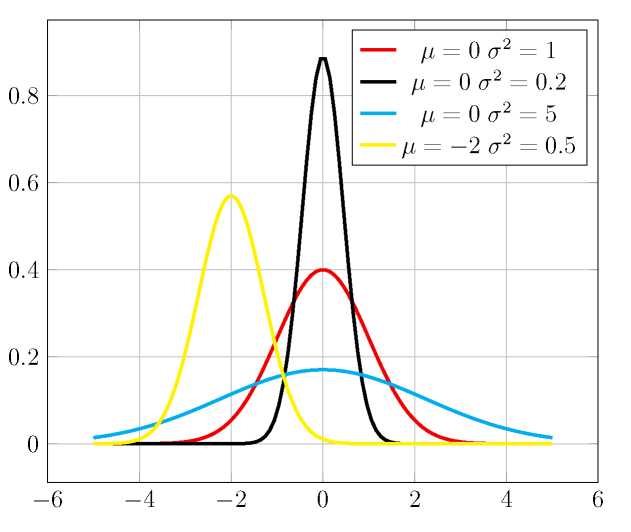


Рис. 10.

Варьируя значения параметров ц и <т, можно влиять на форму графика функции плотности вероятности нормального распределения (ем. рис. 10).

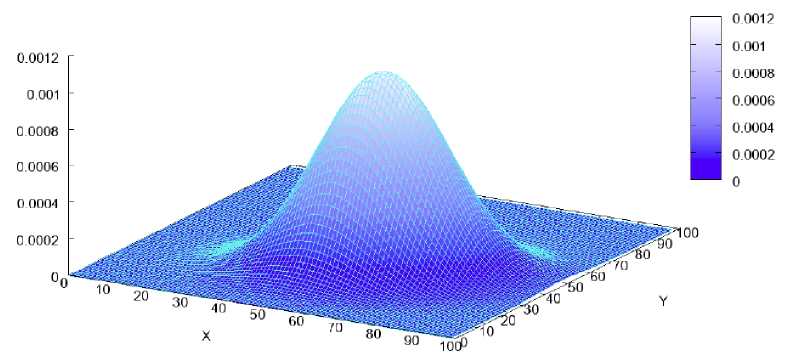
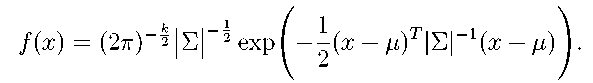


Рис. 11.

Нормальное распределение может быть многомерным. В этом случае случайная величина является не скалярной, а векторной:

image30

image31

причем матрица

является положительно определенной. Функция плотности веро

ятности в данном случае имеет вид (см. рис. 11):

Статистики

1. Оценка распределения по выборке

image33image34image35image36image37image38image39Рассматривается выборка из случайной величины

где

— объем выборки. Величины

деленные случайные величины

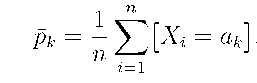
— независимые одинаково распре-

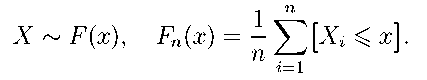
Статистикой

называется любая функция от данной выборки.

Далее будет рассмотрено, какие статистики используются для оценок по выборкам законов распределения случайных величин различных классов. Распределение дискрет,ной случайной величины задается функцией вероятности:

Для выборки из такой случайной величины лучшей оценкой для вероятностей из функции вероятностей являются частоты соответствующих событий на выборке (по закону больших чисел):



Если непрерывная случайная величина задается с помощью функции распределения, то ее можно оценить с помощью эмпирической функции распределения:

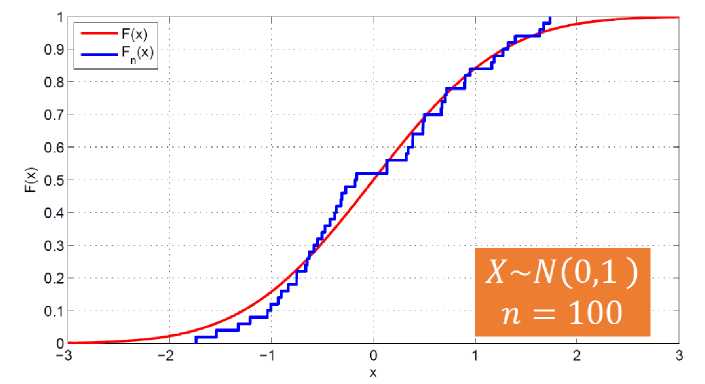


Рис. 1.

На рис. 1 красная линия соответствует теоретической функции стандартного нормального распределения (нормальное распределение со средним, равным нулю, и с дисперсией, равной 1). Синяя линия соответствует эмпирической функции распределения, построенной по выборке объема 100.

Непрерывные случайные величины также могут задаваться с помощью плотностей. Для оценки плотности можно разбить области определения случайной величины на интервалы одинаковой длины. Количество объектов выборки в каждом интервале будет пропорционально среднему значению плотности на нем. Именно так Застроена гистограмма.

На гистограмме, приведенной на рис. 2, изображена продолжительность жизни крыс на строгой диете (в днях).

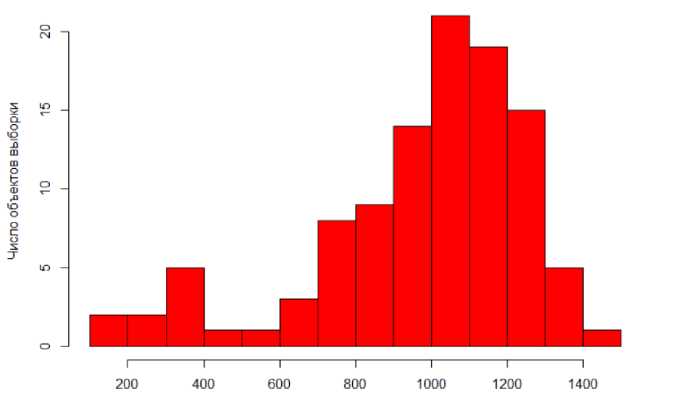


Рис. 2.

По такой гистограмме хорошо виднв1 все особенности распределения даннвгх: оно бимодалвно, основной пик приходится примерно на 1000 дней, но еств крысы, кото- рые живут существенно менвше.

Важным аспектом работы с гистограммами является правильный выбор числа интервалов.

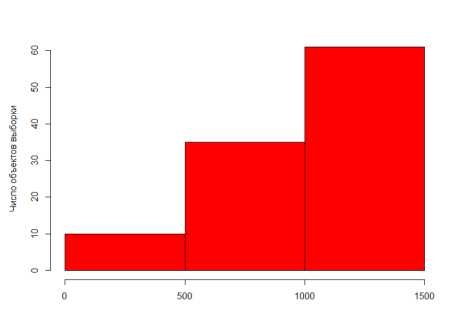


Рис. 3.

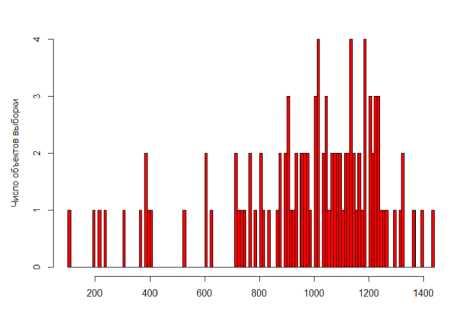
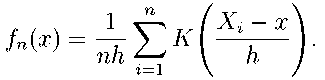


Рис. 4.

Если рассмотреть слишком мало интервалов, то они будут слишком большими, в результате гистограмма получится грубой (см. рис. 3). Аналогично в случае слишком большого количества интервалов — в большую часть из них не попадет ни одного объекта выборки (см. рис. 4). В обоих случаях построенные гистограммы не являются информативными.

image46Описанного недостатка лишены гладкие оценки плотности

Для построения такой оценки необходимо взять окно ширины h и, двигая его по числовой оси, вычислять в нем значение функции, называемой ядром. Ядерная оценка плотности имеет вид:



На рис. 5 показана оценка, построенная на тех же данных о продолжительности жизни крыс. Как и в случае гистограммы, на таком графике видны все особенности распределения данных.

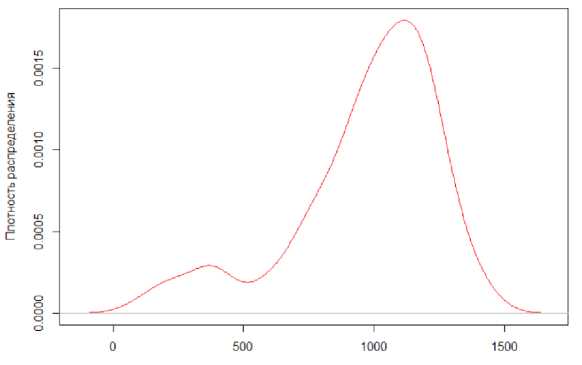
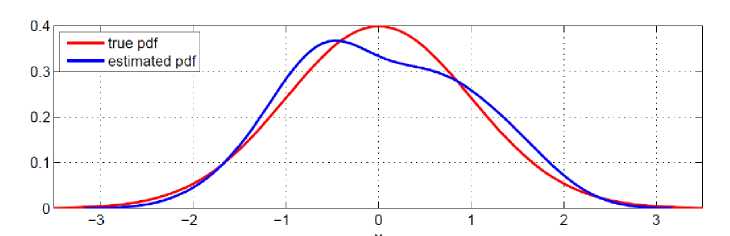


Рис. 5.

На рис. 6 продемонстрированы все виды оценок распределения для выборки из стандартного нормального распределения.



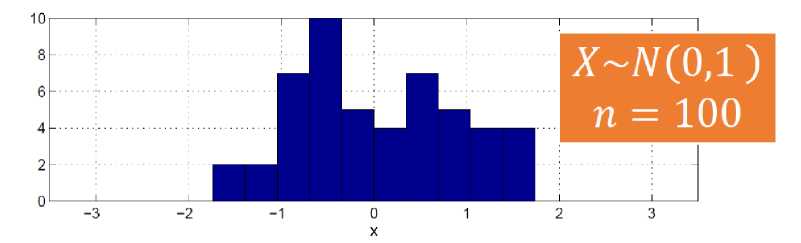


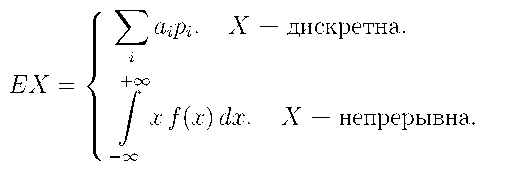
Рис. 6.

Необходимо отметить, что ни один из представленных способов оценки плотностей не является идеальным, так что рекомендуется использовать оба.

2. Важные характеристики распределений

Часто возникает необходимость оценить не всю функцию распределения, а некоторые ее параметры. Самым важным классом параметров распределения являются средние. Нестрогое определение можно сформулировать следующим образом: среднее — это значение, вокруг которого группируются все остальные.

Одним из вариантов уточнения данного определения является матожидание:



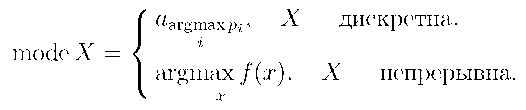
Другой характеристикой среднего является медиана. Она определяется с помощью квантиля. Квантилем порядка а £ (0,1) называется величина Ха такая, что:

image52

Медиана — это квантиль порядка 0,5:

image53

Еще одной характеристикой среднего является мода — самое вероятное значение случайной величины (в нестрогом смысле):



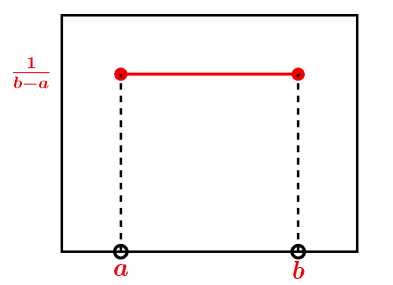
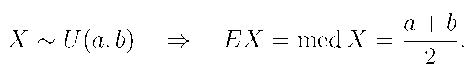


Рис. 7.

В случае нормально распределенной случайной величины ее матожидание, медиана и мода в точности совпадают:

image56

Если случайная величина X равномерно распределена на отрезке [а,Ь], то ее матожидание и медиана совпадают:



Мода такой случайной величины не определена, поскольку у плотности распределения нет максимума (см. рис. 7). Значит, модой в данном случае может быть любое число на интервале от а до Ь.

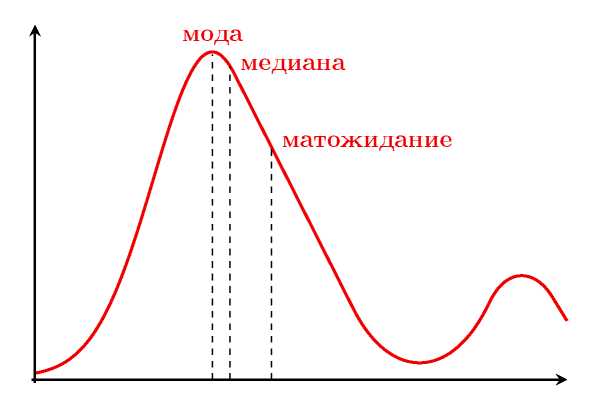
В случае бимодального распределения мода приходится на максимум плотности, а медиана и матожидание смещены в сторону второго «горба», причем смещение матожидания больше, чем смещение медианы (см. рис. 8).

image59image60Следующая рассматриваемая группа параметров распределения — это параметры, характеризующие разброс, то есть то, насколько случайная величина концентрируется вокруг своего среднего значения. Одним из наиболее важных параметров здесь является дисперсия:

Часто используется величина

называемая **среднеквадратическое откло**

**нение.**

image61Еще одна характеристика разброса — интерквантильный размах:

image62Если случайная величина X распределена по закону Пуассона с параметром А, то её дисперсия и матожидание совпадают:

Для нормально распределенной случайной величины дисперсия — это второй параметр распределения:

image63

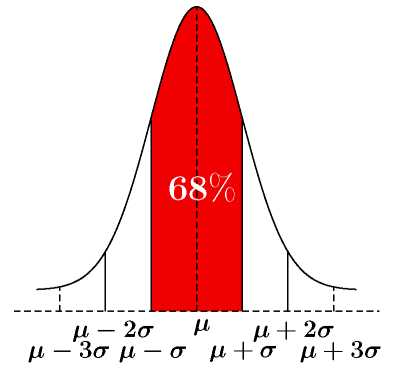


Рис. 9.

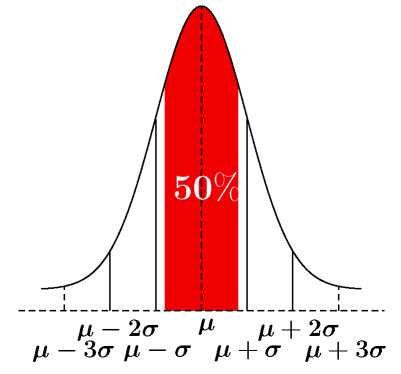


Рис. 10.

В отличие от характеристик среднего значения, характеристики разброса для нормального распределения не совпадают. Для иллюстрации можно отложить от

image66image67интервалы, соответствующие среднеквадратическим отклоне

среднего значения

ниям

image68

В интервале от

распределения (см. рис. 9). В интервал, соответствующий интерквартильному размаху вокруг среднего, попадает 50% случайной величины (ем. рис. 10).

лежит 68% вероятностной массы нормального

image69попадает примерно 95% вероятностной массы нор

В интервал от

image70мально распределенной случайной величины (см. рис. 11). Это часто используемое

на практике правило двух сигм. Другой его вариант — правило трех сигм (см. рис.

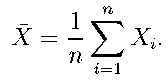
12): в интервале от

случайная величина реализуется практически

со стопроцентной вероятностью (99,7%).

3. Важные статистики

Оценка матожидания случайной величины — это выборочное среднее:



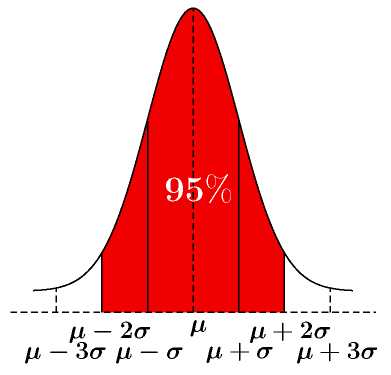


Рис. 11.

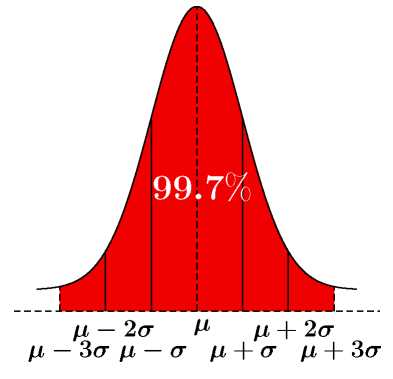
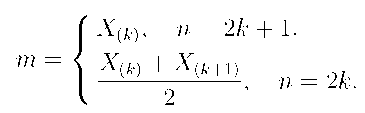


Рис. 12.

Для построения выборочной медианы необходимо составить из рассматриваемой выборки вариационный ряд:

image74

Элемент вариационного ряда г называется г-й порядковой статистикой. Выборочная медиана является центральным элементом вариационного ряда:



Выборочная мода оценивается по максимуму оценки плотности распределения.

Показателен следующий пример. Рассматривается выборка из 25 человек, для каждого из которых известен годовой доход. В выборке есть десять человек, годовой доход которых равен двум тысячам долларов, один человек с годовым доходом в три тысячи долларов, и так далее. Один человек получает сорока пять тысяч долларов в год. Среднее арифметическое годовых доходов на этой выборке — 5700 долларов. Здесь медиана составляет 3000 долларов, а мода — 2000.

Необходимо заметить, что все рассматриваемые величины называются «средними». Значит, для оптимистичного отчета по данной выборке можно воспользоваться средним арифметическим, а для пессимистичного — модой.

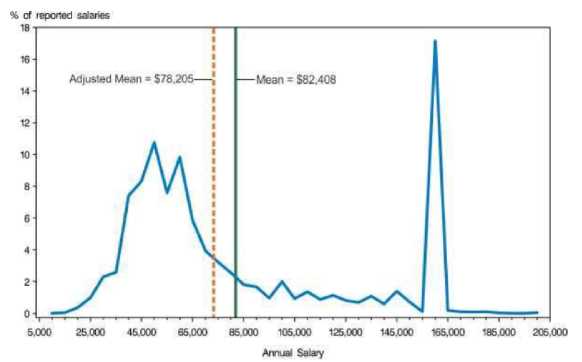


Рис. 13.

Выборочная дисперсия оценивает дисперсию и имеет следующий вид:

image78image79image80Для построения выборочной оценки интерквартального размаха необходимо опре

размах определяется следующим образом:

делить *выборочный квант,иль* порядка

торой равен целой части от

— это порядковая статистика, порядок ко- Тогда выборочный интерквартильный

В качестве следующего примера будет рассмотрено распределение годового дохода членов американской ассоциации юристов (см. рис. 13). Имеются два выраженных пика — 168 тысяч долларов и 45 тысяч. Среднее значение, посчитанное по данной выборке, равно 82 тысячам долларов. Видно, что это значение несет очень мало информации о выборке, так как крайне мало людей получают именно такой доход.

Другой показательный пример — кварт,ет Энскомба (см. рис. 14). Рассматриваются четыре искусственно сгенерированных пары выборок, характеристики которых в каждом из четырех случаев совпадают (равны выборочные средние и выборочные дисперсии). Однако, при рассмотрении диаграмм рассеяния по этим четырем парам выборок, видно, что в каждом из четырех случаев происходят абсолютно разные вещи (см. рис. 15). Таким образом, далее совокупность статистик не позволяет полностью понять данные, и рекомендуется всегда при анализе данных изучать графики, гистограммы и оценки плотности вместо одних лишь цифр.

4. Центральная предельная теорема

Рассмотрим случайную величину X с функцией распределения F(x). Пусть имеется ее выборка объема п:

image81

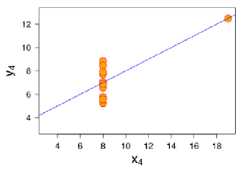
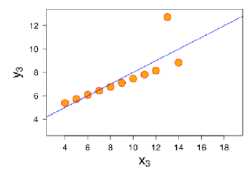
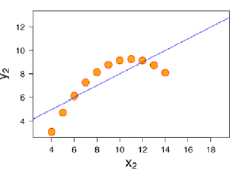
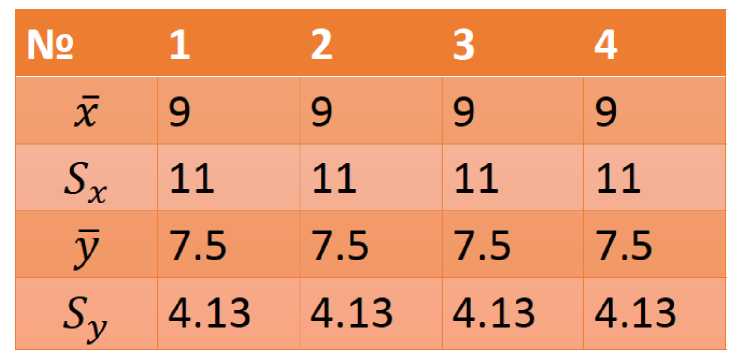
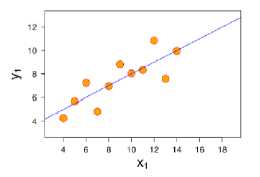
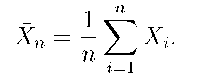


Рис. 14.



По выборке можно вычислить выборочное среднее:

Рис. 15.



Какое распределение имеет выборочное среднее, и как оно связано с исходным распределением?

Можно провести эксперимент. Берется случайная величина с распределением, показанным на рис. 16.

Из данной случайной величины можно взять выборку объема п и посчитать по ней выборочное среднее. Данное действие необходимо повторить в рамках эксперимента достаточно много раз, чтобы затем построить гистограмму полученных выборочных средних. На рис. 17 приведена гистограмма, построенные по выборкам объема п = 2. По сравнению с исходной плотностью случайной величины, данная гистограмма выглядит более гладкой. С увеличением объёма выборки процесс сглаживания продолжается (см. рис. 18 для п = 3).

При объеме выборки п = 5 гистограмма становится унимодальной (см. рис. 19). Дальнейшее увеличение выборки не влияет на форму гистограммы, она лишь становится более узкой (ем. рис. 20 для п = 30).

Можно з

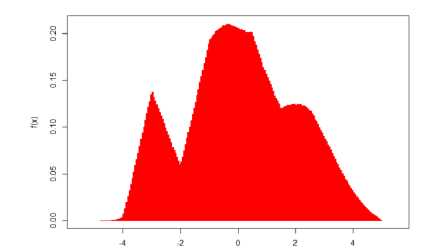


Рис. 17.

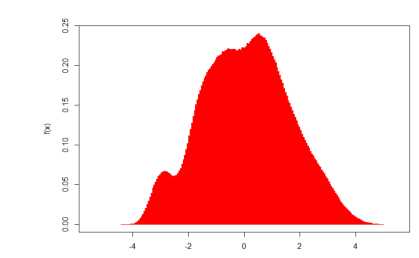
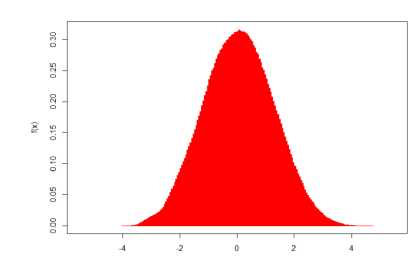
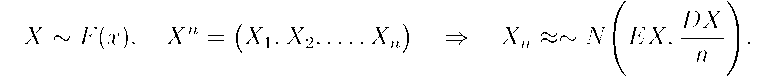


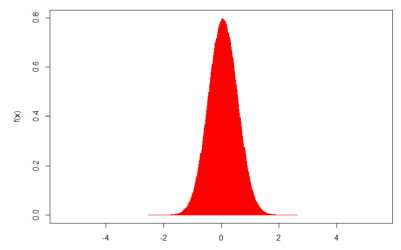
Рис. 18.

Рис. 19.

Рис. 20.

аметить, что распределение выборочных средних достаточно хорошо описывается нормальным распределением, что является утверждением центральной предельной теоремы: 

С ростом п точность нормальной аппроксимации увеличивается.

Полученный результат справедлив не только для непрерывных распределений, но и для дискретных. Это можно рассмотреть на примере биномиального распределения (см. рис. 21). Как и в предыдущем эксперименте, можно повторить данный несколько раз и построить гистограммы для различного объема выборок (см. рис. 22 для п = 2).

При увеличении объема выборок происходит то же, что и в предыдущем эксперименте — распределение становится все более гладким и все более похожим на нормальное (см. рис. 23 для п = 5 и рис. 24 для п = 30). Таким образом, центральная предельная теорема в данном случае работает.

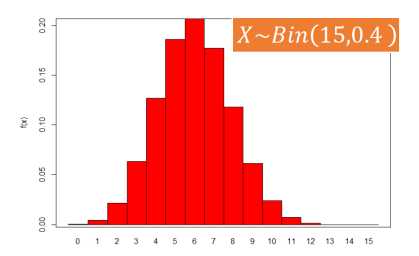


Рис. 21.

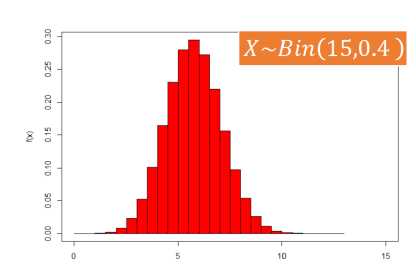


Рис. 22.

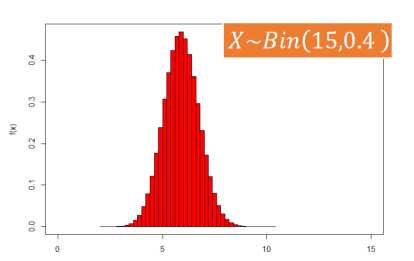


Рис. 23.

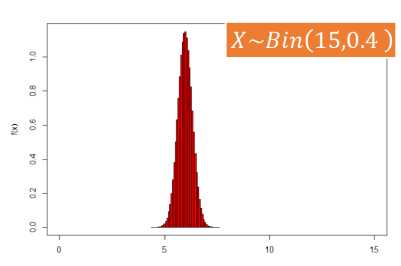


Рис. 24.

Проведём еще один эксперимент для биномиального распределения, на этот раз взяв р = 0,01. Функция вероятности данной случайной величины показана на рис. 25. На рис. 26 изображено распределение выборочных средних, построенных по выборкам объема п = 2.

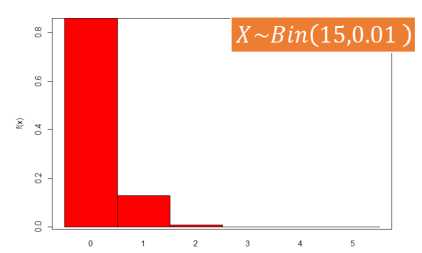


Рис. 25.

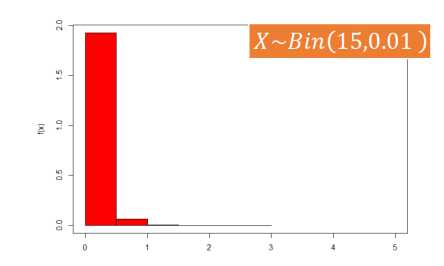


Рис. 26.

Исходное распределение не позволяет распределению выборочных средних быстро сходиться к нормальному (см. рис. 27 для п = 5 и рис. 28 для п = 30). Даже по выборке объема п = 30 гистограмма не может быть хорошо описана нормальным законом — даже относительно максимума данной гистограммы распределение несимметрично.

Центральная предельная теорема хорошо работает, если исходное распределение не слишком скошено. Существует эмпирическое правило: когда распределение X не слишком скошено, распределение Хп хорошо описывается нормальным при п ^ 30.

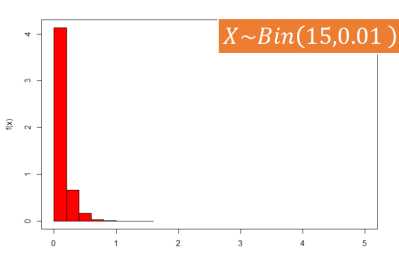


Рис. 27.

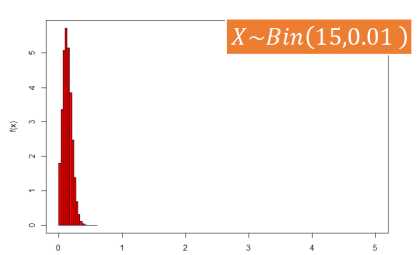


Рис. 28.

5. Доверительные интервалы

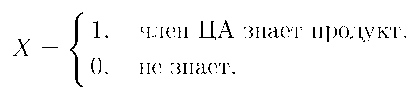
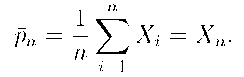
Имеется некий продукт, для которого известна его целевая аудитория. Необходимо узнать, насколько хорошо целевая аудитория знакома с данным продуктом. Введём следующую случайную величину:

image103image104image105image106image107Такая случайная величина имеет распределение Бернулли с параметром р — узна заем,остью продукта:

image108

Измерить узнаваемость продукта можно с помощью опроса. Если в опросе п

участников, то на выходе получится выборка Хп, состоящая из 0 и 1. Оценкой узнаваемости по данной выборке будет выборочное среднее:

Пусть по итогам первого опроса, в котором приняло участие 10 человек, оказалось, что 6 из них знакомы с продуктом:

п = 10, рп = 0,6.

Второй опрос дал следующий результат:

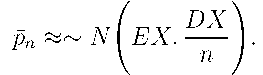
п = 100, рп = 0,44.

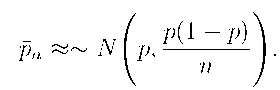
Необходимо определить, какая из двух полученных оценок лучше. Точность измеренной оценки определяется с помощью доверительных интервалов: пары статистик Сц, Си такой, что:

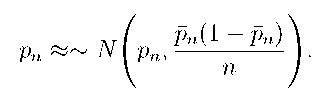
Р(СЬ ^ в ^ Си) ^ I - а.

Здесь в — это оцениваемый параметр, (1 — а) — уровень доверия, a. Cl ж Си — верхний и нижний доверительные пределы (соответственно). При бесконечном повторении эксперимента в 100(1 — а)% случаев этот интервал будет накрывать истинное значение параметра в.

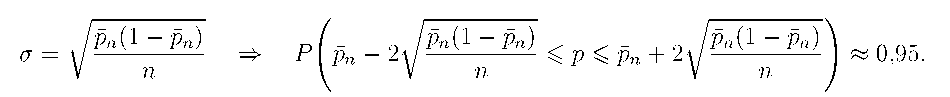
Доверительные интервалы можно построить для оценок узнаваемости продукта из рассматриваемого примера. Оценки узнаваемости являются, по сути, выборочными средними. Следовательно, можно воспользоваться центральной предельной теоремой:

image111Для случайной величины с распределением Бернулли известно, что:

тогда:

В правую часть данного выражения можно подставить рп вместо р:

Распределение стало полностью определенным. Далее необходимо воспользоваться правилом двух сигм:



Применение полученного выражения к двум опросам даст следующий результат. В первом опросе 95% доверительный интервал — (0,29, 0,91), во втором — (0,34, 0,54). Таким образом, доверительный интервал помогает в описании степени неуверенности в полученной оценке.

Доверительные интервалы не обязательно строить с помощью центральной предельной теоремы. Для конкретных распределений существуют более точные способы. Например, для распределения Бернулли наиболее точен метод Уилсона. Однако, именно центральная предельная теорема является универсальным средством построения доверительных интервалов — она работает вне зависимости от того, из какого распределения взята исходная выборка.