**Оценка распределения по выборке**

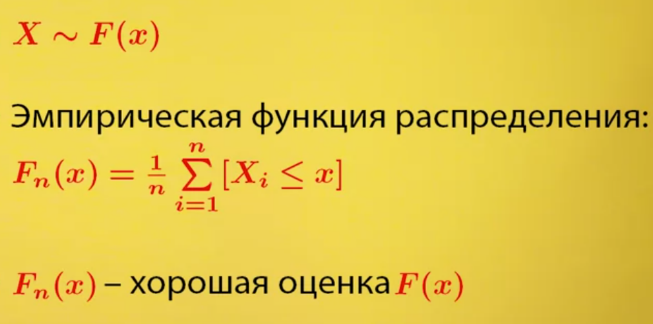
Основной инструмент статистики — это статистики. Пусть у нас есть выборка из случайной величины X объема n. Будем обознать ее за X с верхним индексом n. X1, X2, ..., Xn можно считать независимыми одинаковыми копиями исходной случайной величины X. Поэтому часто говорят, что выборка представляет собой совокупность независимых одинаково распределенных случайных величин.

В англоязычной литературе это длинное словосочетание часто заменяется аббревиатурой i.i.d. Так вот статистикой называется любая функция от этой выборки.

Если мы имеем дело *с дискретной случайной величиной*, все довольно просто.

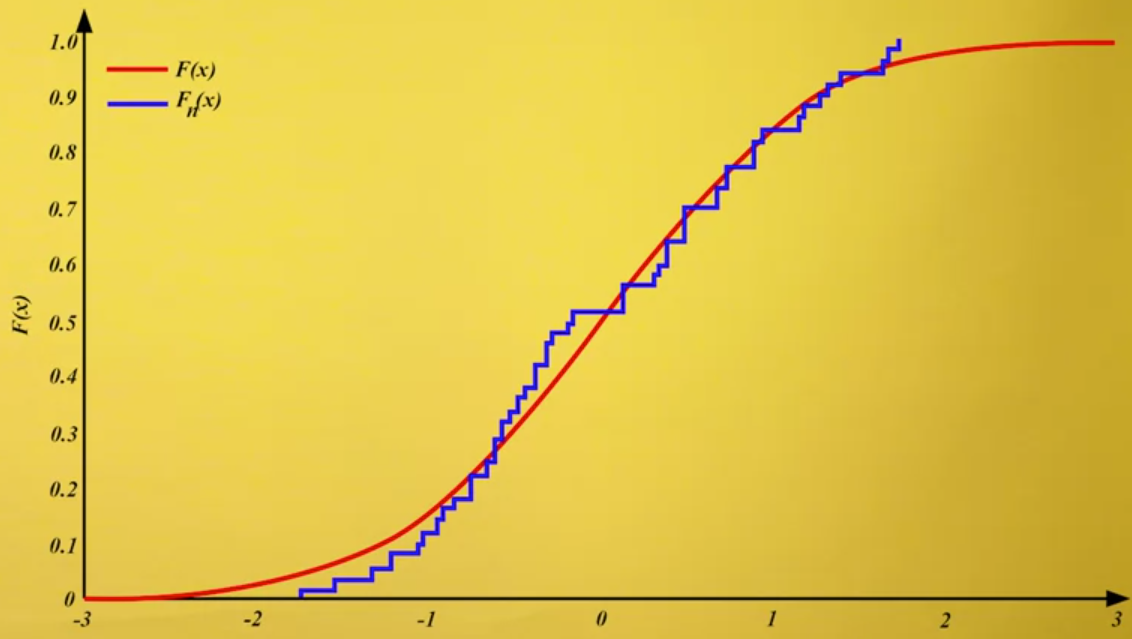
У нас есть множество ее значений, и распределение задается функцией вероятности, то есть вероятностями, с которыми дискретная случайная величина принимает все свои значения. Если у нас есть выборка из этой случайной величины, лучшие оценки для вероятностей из функции

вероятности — это *частоты* соответствующих событий на выборке.



С *непрерывными случайными величинами* все немного сложнее.   
 Если случайная величина задается с помощью функции распределения, оценить ее можно с помощью эмпирической функции распределения, которая представляет собой среднее значение по всем элементам выборки индикаторов того, что элемент выборки не превосходит аргумента функции x маленькое.

Эмпирическая функция распределения достаточно хорошо оценивает теоретическую функцию распределения, особенно если выборка большая. Естественно, чем больше выборка, тем лучше ваша оценка.

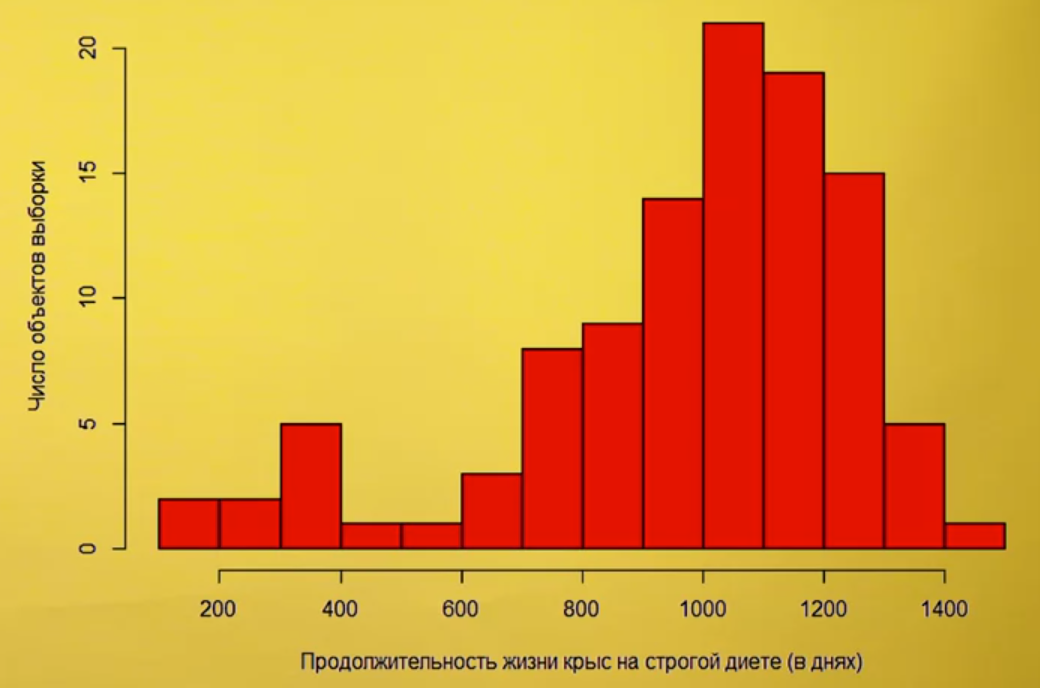
  
  
Теоретическая функция стандартного нормального распределения — красная линия. Стандартным называется нормальное распределение со средним 0 и дисперсией 1.

А синяя ступенчатая линия — это эмпирическая функция распределения, построенная по выборке объема 100.

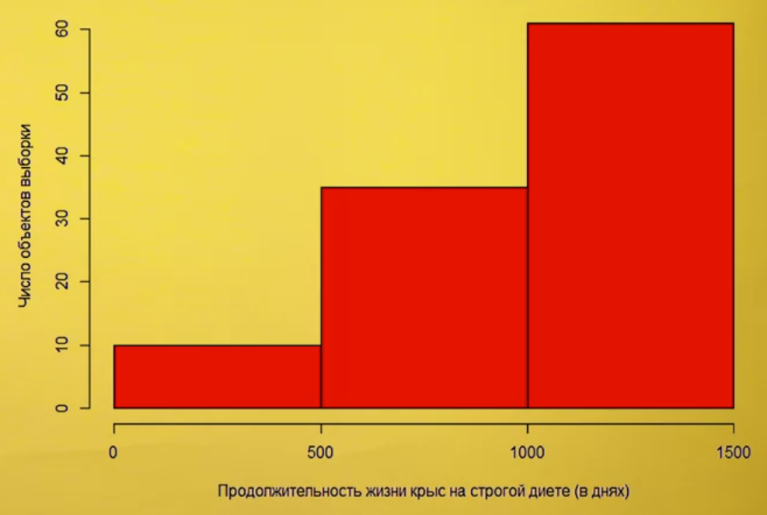
Еще непрерывные случайные величины могут задаваться своими *плотностями*. **Плотности** — это такие функции, что интеграл от них по любому отрезку от a до b равен вероятности попадания случайной величины в этот интервал. Чтобы оценить плотность, надо разбить область определения случайной величины на интервалы одинакового размера.

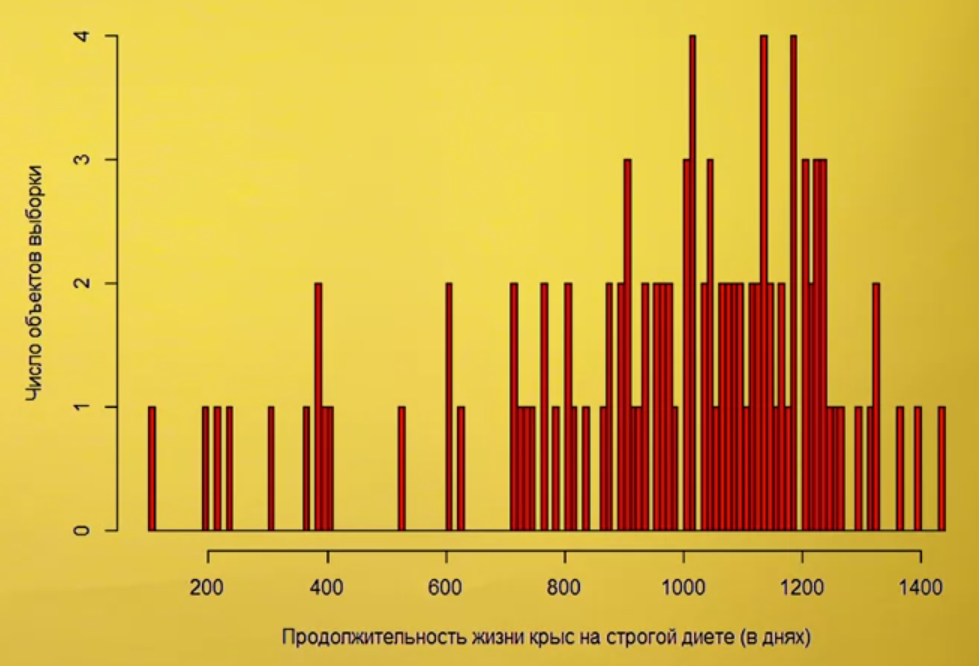
Тогда число объектов выборки в каждом интервале будет

пропорционально среднему значению плотности на этом интервале. Именно так устроена гистограмма.



А вот так выглядит гистограмма обычная. Признак, который на ней показан, — это продолжительность жизни крыс на строгой диете в днях. По гистограмме прекрасно видны все особенности распределения данных. Это распределение бимодальное. Основной его пик приходится примерно на *1000 дней*. Но есть крысы, которые живут существенно меньше — около *400 дней*.

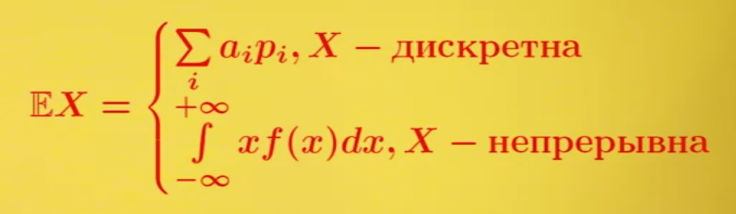
  
  
**Важный** аспект работы с гистограммами — *это правильный выбор числа интервалов*. Если мы возьмем интервалов слишком мало, они будут слишком большие и гистограмма получится грубой. По ней мы не сможем понять, что происходит в данных.

  
  
То же самое может произойти и в обратном случае. Если мы возьмем слишком много интервалов, в большую часть из них не попадет ни одного объекта выборки, и гистограмма получится разреженной и тоже не очень информативной.   
Этого недостатка лишены **гладкие оценки плотности**. *(в данном курсе не объясняется подробно что это такое*)  
  
**Важные характеристики распределения**

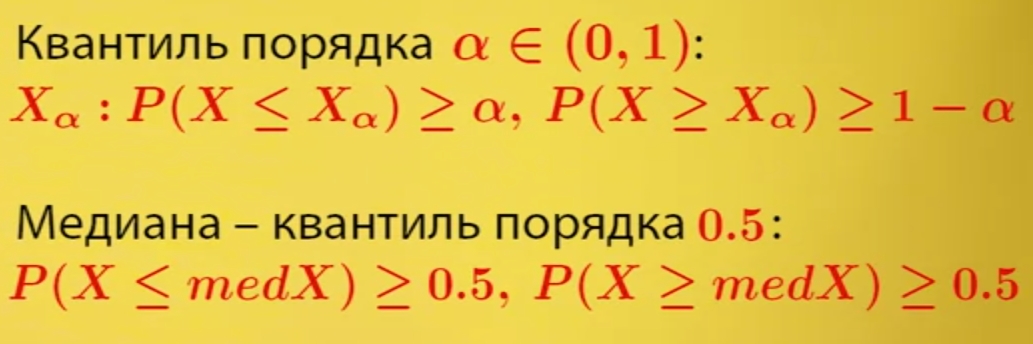
Первый и наиболее часто используемый класс параметров — это **среднее**.

Понятие **среднего** — нестрогое, его можно формализовать разными способами.

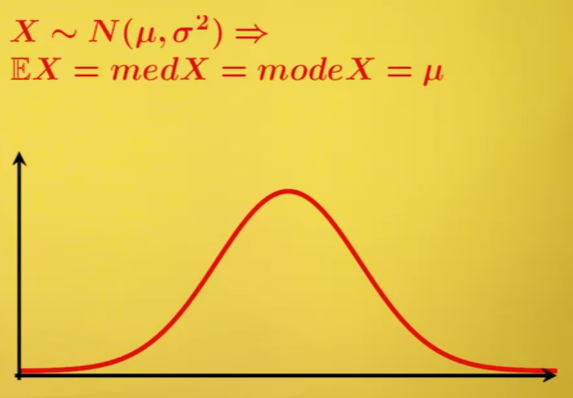
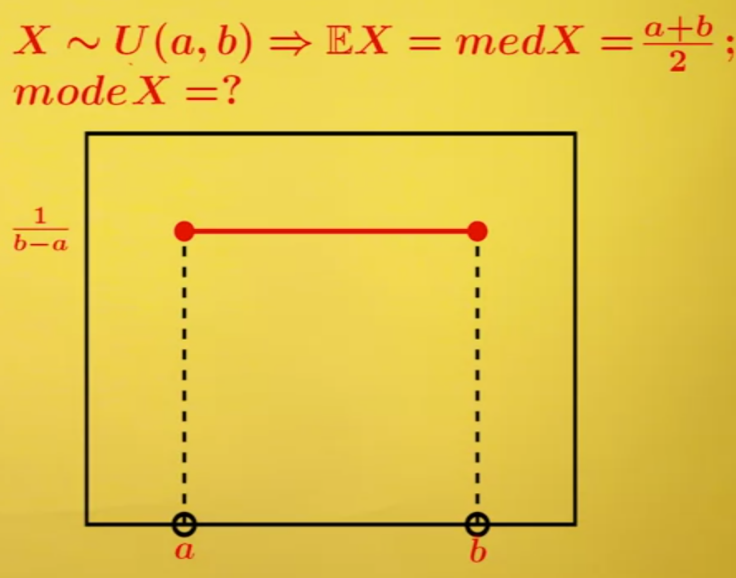
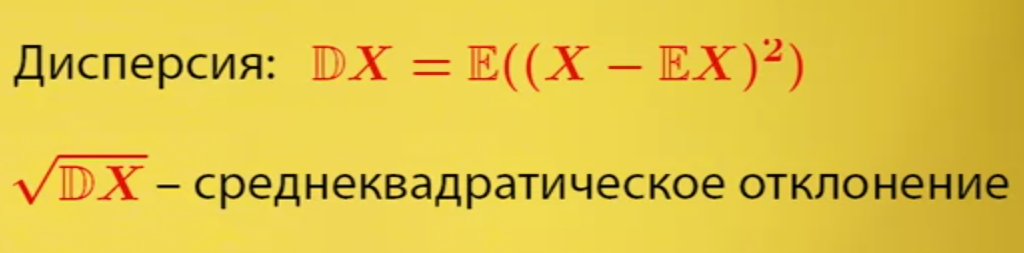
*Способ № 1 — математическое ожидание.*

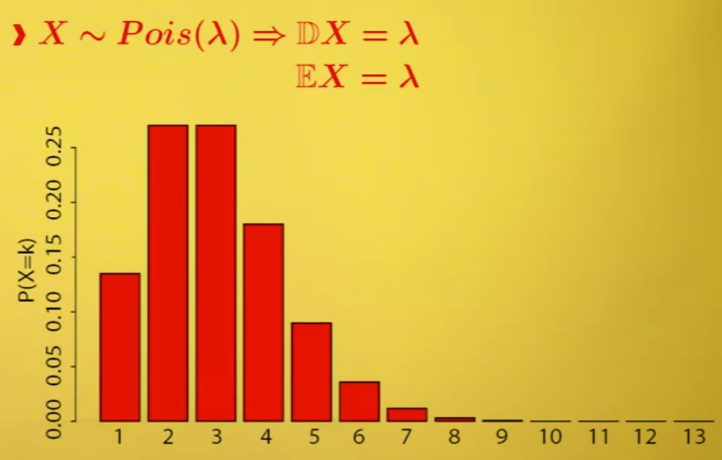
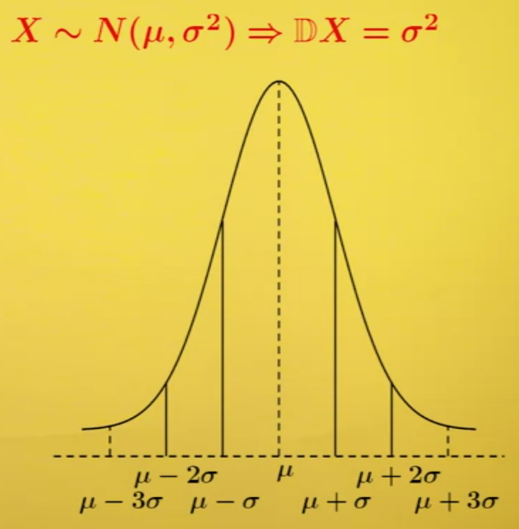
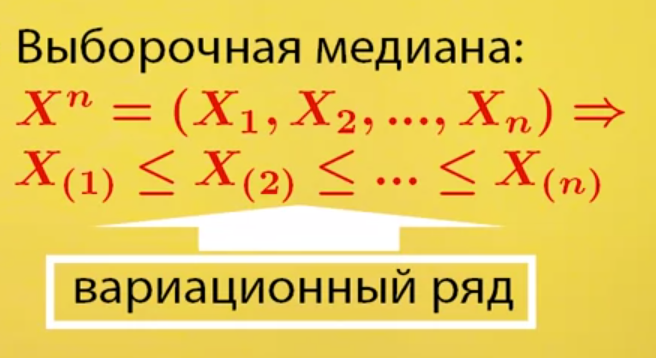


Для дискретной случайной величины математическое ожидание определяется как сумма всех значений, которые она принимает, взятая с весами, равными вероятностям этих значений. Непрерывная случайная величина принимает несчетное множество значений, поэтому ее математическое ожидание определено как интеграл по всей области определения случайной величины от x, умноженного на f(x) — плотность распределения.   
  
*Следующий способ* формализации понятия среднего — это **медиана**.



Чтобы определить медиану, введем сначала понятие квантиля. Квантилем порядка α называется такая величина Xα, что наша случайная величина X принимает значения слева от нее на числовой оси с вероятностью не меньше α и справа с вероятностью не меньше, чем (1 − α), а медиана — это вовсе не то, к чему вы привыкли, а всего лишь квантиль порядка 0,5, то есть такое значение, что наша случайная величина слева и справа от него попадает с вероятностями, близкими к 1/2.

*Третий способ* формализации понятия среднего — это **мода**.   
  
Модой дискретной случайной величины называется наиболее вероятное ее значение. То есть значение, которое она принимает с наибольшей вероятностью. Для непрерывной случайной величины мы моду так определить не можем, поскольку каждое из значений она принимает с вероятностью 0, поэтому мода — это точка, в которой максимума достигает плотность распределения непрерывной случайной величины.  
  
  
  
Если мы имеем дело со случайной величиной из нормального распределения с параметрами μ и σ квадрат, то ее математическое ожидание, медиана и мода в точности совпадают и равны μ. Это одно из волшебных свойств *нормального распределения*. Но так происходит далеко не всегда.  
  
  
  
*Например*,   
если наша случайная величина распределена равномерно на отрезке ab, то ее математическое ожидание и медиана равны (a + b) пополам, то есть середине отрезка, а мода такой случайной величины однозначно не определена, поскольку у плотности, которую вы видите перед собой, нет максимума. То есть модой может быть любая точка на отрезке от a до b.   
  
В общем случае *математическое ожидание*, *мода* и *медиана* распределения не обязаны совпадать.   
  
Еще один класс параметров распределений — это параметры, описывающие разброс случайной величины, то есть то, насколько сильно они концентрируются вокруг своего математического ожидания.   
  
  
  
Наиболее часто используемый из таких параметров — это дисперсия, определяющаяся как математическое ожидание квадрата отклонения случайной величины от своего математического ожидания. Корень из дисперсии часто используется сам по себе и имеет свое название — среднеквадратическое отклонение.   
  
*IQRn = X0.75 – X0.25*

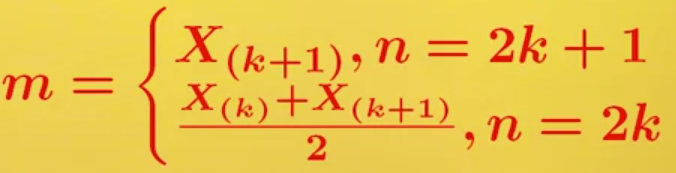
Еще один параметр, оценивающий разброс распределения, — это интерквартильный размах. Это всего лишь разность 75-процентного и 25-процентного квантилей, или квартилей, как они еще называются, потому что расположены по четвертям.  
  
  
  
Если вы имеете дело со случайной величиной, распределенной по закону Пуассона с параметром λ, то дисперсия ее равна λ. Интересно, что ее матожидание тоже равно λ. Это важное свойство распределения Пуассона — его математическое ожидание и дисперсия совпадают.  
  
  
Для нормально распределенной случайной величины с параметрами μ и σ квадрат дисперсия равна σ квадрат. Таким образом, мы теперь знаем значения обоих параметров нормального распределения. Первый отвечает за математическое ожидание, а второй — за дисперсию.  
 **Важные статистики**  
  
*Оценка матожидания* случайной величины — это выборочное среднее, то есть буквально среднее арифметическое всех реализовавшихся значений.   
  


Довольно просто.   
Чтобы построить *выборочную медиану*, которая оценивает истинную медиану, нам нужно проделать несколько более сложную манипуляцию.

Возьмем нашу выборку и составим из нее вариационный ряд,

то есть отсортируем все ее элементы по неубыванию.

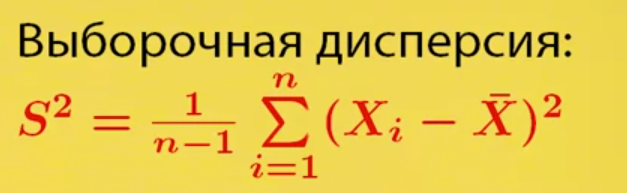
i-тый элемент вариационного ряда называется i-той порядковой статистикой.



Так вот, *выборочная медиана* — это центральный элемент вариационного ряда, то есть если мы имеем дело с выборкой, объем которой — нечетное число, то есть n представляется в виде 2k+1, то выборочная медиана — это просто k-тый элемент вариационного ряда. Если длина выборки четная, то для того, чтобы посчитать выборочную медиану, мы берем среднее арифметическое между k-тым и (k+1)-м элементом вариационного ряда.

*Выборочная мода*, в свою очередь, оценивается по максимуму оценки плотности распределения.

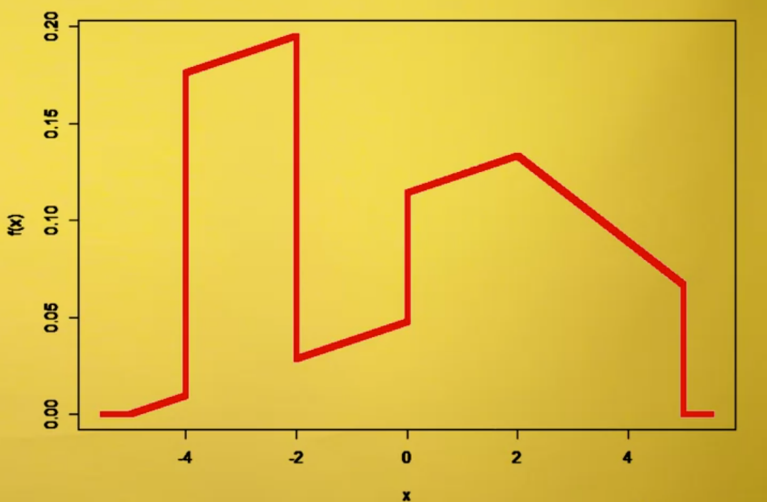
Как оценивать плотность мы говорили несколько видео назад.

*Посмотрим теперь на оценки разброса распределения.*   
  
  
  
Выборочная дисперсия оценивает дисперсию и представляет собой среднее арифметическое квадратов отклонения от выборочного среднего. Единственный нюанс заключается здесь в том, что в знаменателе перед суммой стоит не n, а n−1.   
  
Для того, чтобы построить выборочную оценку интерквартильного размаха, нам нужно определить выборочный квантиль. Выборочным квантилем порядка α будем называть порядковую статистику, порядок которой равен целой части от α\*n.   
  
*IQRn = X(|0.75|) – X(|0.25|)*  
  
Тогда выборочный интерквантильный размах — это разность соответствующих порядковых статистик.   
  
**Центральная предельная теорема**

Пусть у нас есть некая случайная величина X с функцией распределения F. И мы имеем ее выборку объема n.

Посчитаем по этой выборке выборочное среднее — X с чертой. Будем использовать нижний индекс n, чтобы подчеркнуть, что выборка именно такого объема. Какое распределение имеет эта новая случайная величина — выборочное среднее? И как оно связано с исходным распределением F?

*Можно провести следующий эксперимент*

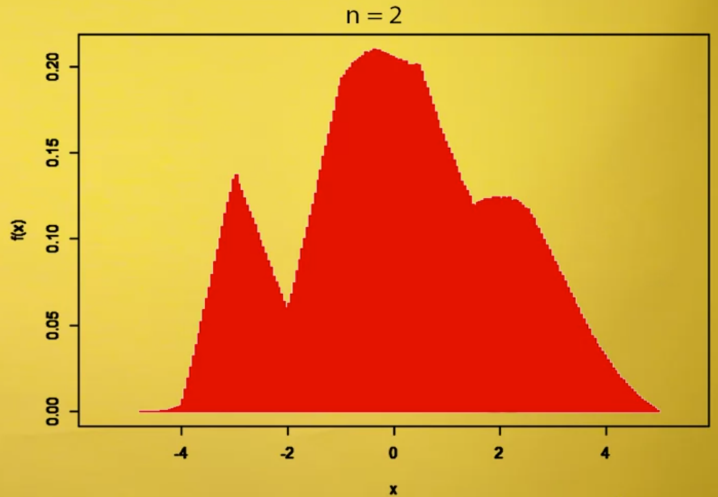


Возьмем случайную величину вот с таким странным,

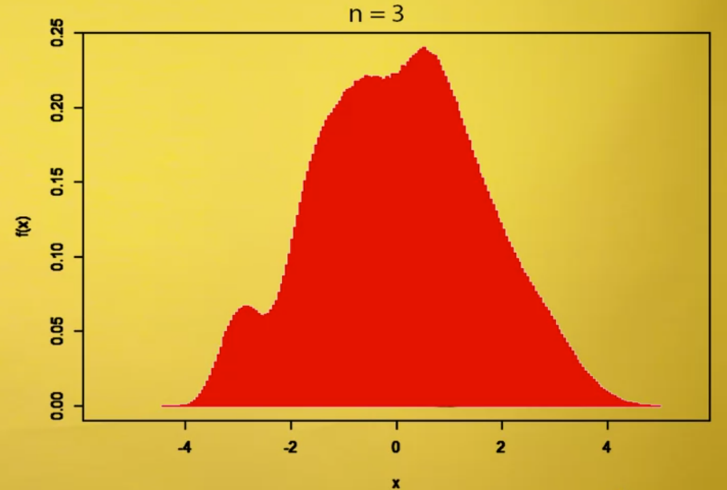
ни на что не похожим распределением. Это распределение абсолютно синтетическое, это смесь двух равномерных и одного треугольного распределения. И будем делать следующим образом. Будем из этой случайной величины брать выборку объема n, считать по этой выборке выборочное среднее, записывать это полученное значение. Повторим этот эксперимент много-много-много раз, например, миллион.

И построим гистограмму полученных выборочных средних.

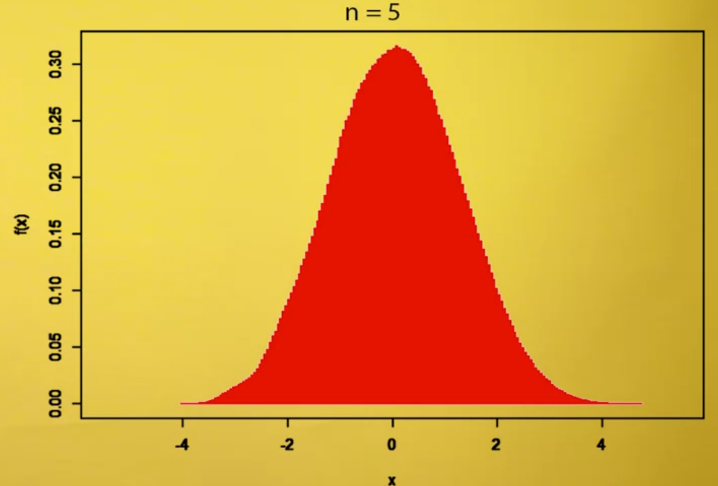
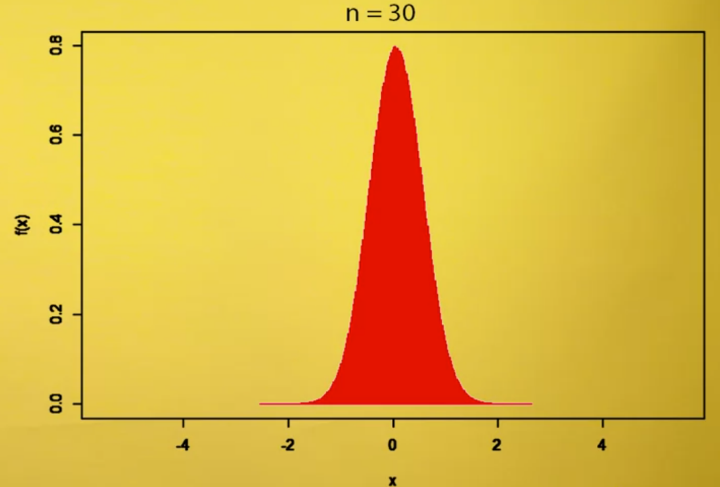
Посмотрим, на что она похожа.



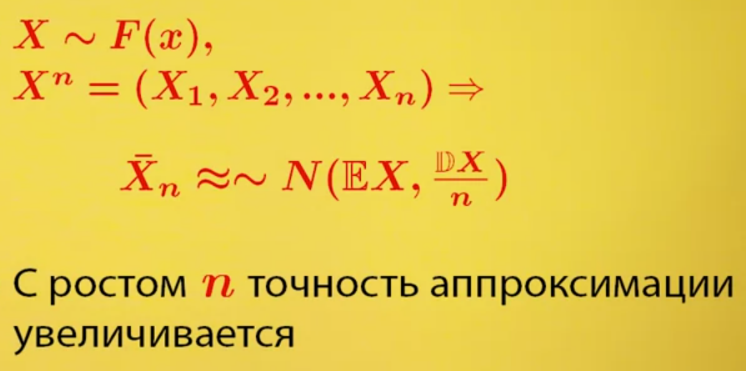
Гистограммы, построенные по выборкам объема 2. Мы видим, что по сравнению с исходной плотностью случайной величины эта гистограмма выглядит более гладкой, острые углы на ней начинают постепенно расплываться. И этот процесс продолжается с увеличением выборки.



Вот так выглядит гистограмма выборочных средних, построенных по выборкам объема 3. На ней уже практически не осталось острых углов.

  
  
При объеме выборки 5 гистограмма окончательно становится унимодальной.   
  
  
  
А дальнейшее увеличение выборки не влияет на форму гистограммы, она становится только более узкой,   
более сконцентрированной вокруг среднего значения — нуля.

Действительно, распределение выборочных средних неплохо можно описать нормальным, именно это и утверждает центральная предельная теорема.



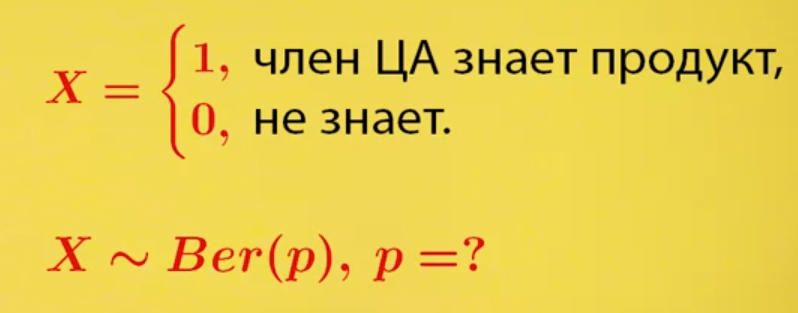
*Если у вас есть случайная величина X из практически любого распределения и у вас есть выборка объема n из нее,*

*то выборочное среднее, построенное по этой выборке,*

*можно приблизить нормальным распределением со средним значением, которое совпадает с математическим ожиданием исходной случайной величины и с дисперсией, которая равна дисперсии исходной случайной величины, поделенной на n.*

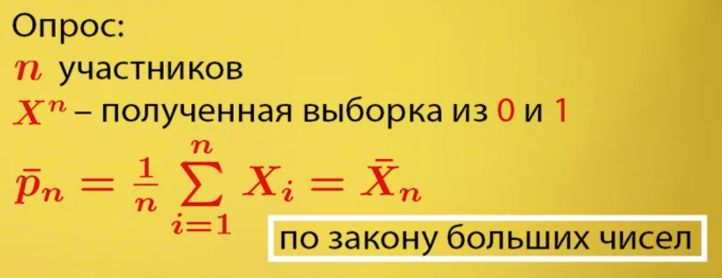
Чем больше n, тем точнее эта нормальная аппроксимация,

тем лучшее распределение описывается нормальным.  
  
*Центральная предельная теорема хорошо работает, если исходное распределение не слишком скошено. Чем сильнее оно скошено, тем больший объем выборки нужен для того, чтобы достичь хорошего качества нормальной аппроксимации.*

**Доверительные интервалы**Представим, что у вас есть некий продукт, и вы знаете, кто его целевая аудитория. Вы хотите узнать, насколько хорошо ваша целевая аудитория с ним знакома, то есть понять, какая ее доля слышала о продукте.   
  


Давайте определим вот такую случайную величину x.

Она равна единице, если член целевой аудитории знаком с вашим продуктом, и нулю, если он с ним не знаком.

Эта случайная величина имеет распределение Бернулли с параметром p. Вот этот параметр p — это и есть узнаваемость продукта, которую вы хотите каким-то образом померить.   
  


Как это можно сделать?

Можно провести опрос. Допустим, в вашем опросе n участников. Тогда на выходе вы получаете выборку Xn из нулей и единичек. И оценкой узнаваемости по этой выборке будет выборочное среднее, то есть доля единиц в полученной выборке.

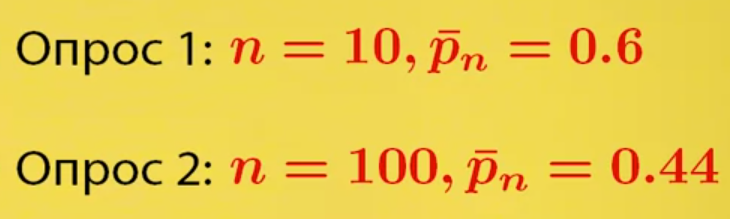
Представьте, что вы провели первый вопрос,

опросили 10 человек и узнали, что 6 из них с вашим продуктом знакомы.

Таким образом, ваша оценка параметра p — это 0,6.

Вы провели еще один опрос среди 100 членов вашей целевой аудитории. 44 из них знакомы с вашим продуктом.

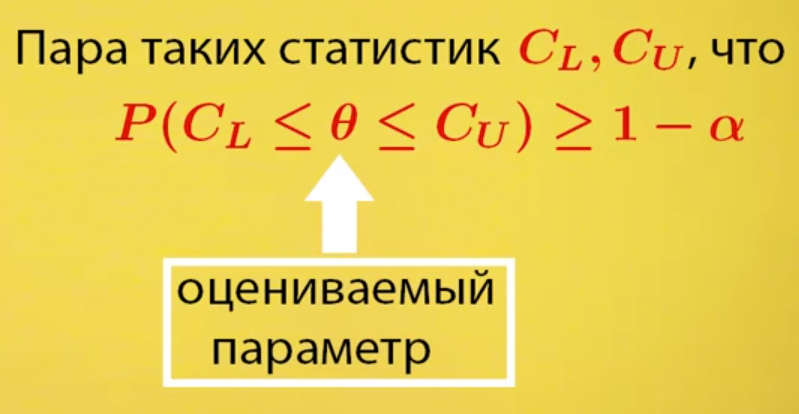
То есть ваша другая оценка параметра p — 0,44.



Какая из этих двух оценок лучше?

Интуитивно кажется, что вторая, поскольку там больше данных, значит, она каким-то образом должна быть точнее.

Но как это квантифицировать? Как показать, насколько точно ваша оценка измерена?

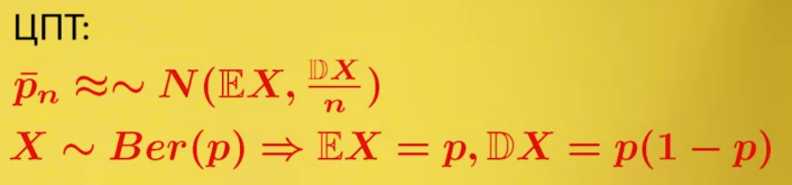


Именно это делается с помощью доверительных интервалов.

Доверительным интервалом называется такая пара статистик CL, CU, что вероятность того, что Θ лежит между этих двух статистик, не меньше, чем 1 − α. Θ здесь — это параметр, который вы оцениваете с помощью интервала. (1 − α) называется уровнем доверия, а CL и CU — соответственно нижним и верхним доверительными пределами.

Как интерпретируется доверительный интервал?

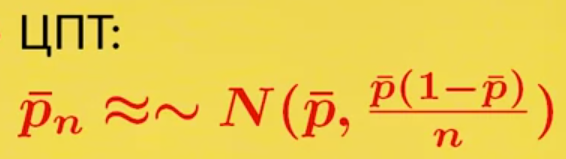
Если представить, что вы повторяете эксперимент по построению интервала бесконечно, то в (1 − α) умножить на

100 % случаев этот интервал будет накрывать истинное значение параметра Θ.   
  


Построим доверительные интервалы для узнаваемости продукта. Оценки узнаваемости, как мы видели раньше, это, по сути, выборочные средние. А значит, центральная предельная теорема говорит нам, что их распределение может быть с той или иной точностью описано нормальным.

Математическое ожидание такого распределения равно p, а дисперсия — p умноженное на (1 − p).

Подставим эти выражения в правую часть утверждения центральной предельной теоремы. 



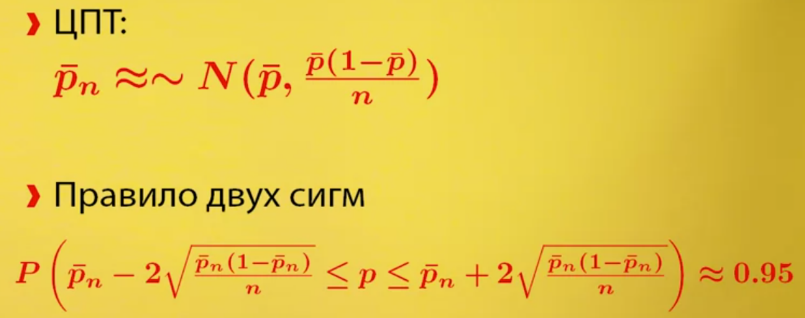
Следующая проблема. Сейчас в правой части стоит параметр p, который нам неизвестен. Что делать? Лучшее, что мы знаем о параметре p, это оценка p с чертой. Просто подставим p с чертой вместо p в правую часть. Теперь распределение, которое стоит в правой части, полностью определено. 

  
  
А дальше нам на помощь приходит правило двух сигм.

Оно утверждает, что 95 % вероятностной массы нормального

распределения лежит в интервале от μ − 2ς до μ + 2ς,

где ς, - это корень из дисперсии. Правило двух сигм очень хорошо подходит для построения 95-процентных доверительных интервалов. Именно такой интервал мы и хотим построить.



Применив правило двух сигм, мы получаем выражение для доверительного интервала.

Подставим теперь наши данные в это выражение и получим оценки. Согласно опросу номер 1, узнаваемость составляет 0,6, однако интервальная оценка узнаваемости от **0,29** до **0,91**. Эта оценка имеет довольно большую дисперсию,

то есть ее точность достаточно низкая.

Во втором опросе, где мы опросили в 10 раз больше людей,

оценка узнаваемости составила 0,44, а 95-процентный доверительный интервал для узнаваемости — от **0,34** до **0,54**.

**Доверительный интервал** — это прекрасный способ донести степень вашей неуверенности в оценке, которую вы построили, потому что любая оценка, построенная по выборке, по определению неточна. Нужно всегда явным образом утверждать, насколько она неточна, по вашему мнению.

Вообще говоря, доверительные интервалы необязательно строить с помощью центральной предельной теоремы. Если мы имеем дело с выборками из каких-то конкретных хорошо известных распределений, как правило, можно найти и более точный способ.

*Например,*

для распределения Бернулли, из которого мы брали выборку в нашем примере, наиболее точным является метод построения *доверительных интервалов Уилсона*.

*Однако именно центральная предельная теорема дает нам универсальное средство построения доверительных интервалов, которое работает независимо от того,  из какого распределения взята исходная выборка.*