Лекция 2.

Примеры дискретных распределений

1. Биномиальное распределение. Случайная величина ***ξ*** имеет биномиальное распределение с параметрами $n \in \mathbb{N}$ и $p \in (0,1)$, если ***ξ*** принимает значения *k=1,2,…,n* с вероятностями $\mathrm{P}(\xi = k) = C_n^k p^k (1-p)^{n-k}$. Случайная величина с таким распределением имеет смысл числа успехов в ***n*** независимых испытаний схемы Бернулли с вероятностью успеха ***p***. Среднее значение и дисперсия для биномиального распределения имеют вид: *M* ***ξ*** = *Np*, *D* ***ξ* =** *Npq* , где *q=(1-p).*

2. Распределение Пуассона. Случайная величина ***ξ*** имеет распределение Пуассона с параметром *λ > 0*, если ***ξ*** принимает значения *k=1,2,…* с вероятностями $\mathrm{P}(\xi = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.
Математическое ожидание и дисперсия распределения Пуассона равны *λ*.

Примеры непрерывных распределений

1. Равномерное распределение.

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

Равномерное распределение на отрезке *(a,b)* задается плотностью и функцией распределения

$$F(x) = \begin{cases} 0, x < a \\ \frac{x-a}{b-a}, x \in [a,b] \\ 1, x > b \end{cases}$$

Математическое ожидание равномерного распределения равно (*b-a)/2,* дисперсия - *(b-a)*^{2}*/12* .

2. Нормальное распределение.

Нормальное распределение с параметрами ***a*** и ***σ*** задается плотностью. Обозначение: $x \in N(a,\sigma)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-a)^2}{2\sigma^2}}$$

Стандартное нормальное распределение. Стандартное нормальное распределение - это нормальное распределение с параметрами *a = 0* и *σ = 1*, т. е. его плотность примет вид . Математическое ожидание нормального распределения равно *a*, дисперсия - σ^2.

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{-y^2}{2}} dy$$

Функция распределения стандартного нормального закона называется интегралом вероятности.

Если ***x*** имеет стандартное нормальное распределение, то случайная величина ***y=x*a+σ*** распределена нормально с параметрами ***a*** и ***σ.***

3. Показательное распределение. Показательное распределение задается плотностью

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, x > 0 \\ 0, x \le 0 \end{cases}$$

и функцией распределения

$$F(x) = \begin{cases} 1 - e^{-\lambda x}, x > 0 \\ 0, x \le 0 \end{cases}$$

Математическим ожиданием показательного распределения является величина 1/λ, обратная к параметру распределения.

4. Распределение Коши. Случайная величина ξ имеет распределение Коши с параметрами положения ***a*** и масштаба ***c***, если ее функция распределения и функция плотности вероятностей имеют соответственно вид:

$$F_\xi(x) = \tfrac{1}{2} + \frac{1}{\pi} arctg \frac{x-a}{c},\ f_\xi(x) = \tfrac{1}{\pi} \frac{c}{c^2 + (x-a)^2}, -\infty < x < \infty.$$

У распределения Коши не существует ни математического ожидания, ни дисперсии.

5. Распределение Релея. Распределение Релея имеет следующую функцию плотности:

$f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}$, $0 \le x < \infty$, где σ > 0 - параметр масштаба.

Распределением Релея является распределение длины случайного вектора с компонентами, распределенными по стандартному нормальному закону.

6. Распределение хи-квадрат. Величина $\boldsymbol{\chi^2_f(x)}$ является распределением суммы ***f*** квадратов нормально распределенных случайных величин. Параметр распределения ***f*** – целое положительное число; его часто называют числом степеней свободы.

Плотность $\boldsymbol{\chi^2}$- распределения имеет вид,
где Г(x) - гамма-функция.

$$\frac{x^{(f-2)/2}\exp(-x/2)}{2^{f/2}\Gamma(f/2)}$$

Математическое ожидание $\boldsymbol{\chi^2}$ равно *f* .
Дисперсия - *2 f*.

Как генерировать на компьютере случайные последовательности с такими распределениями?

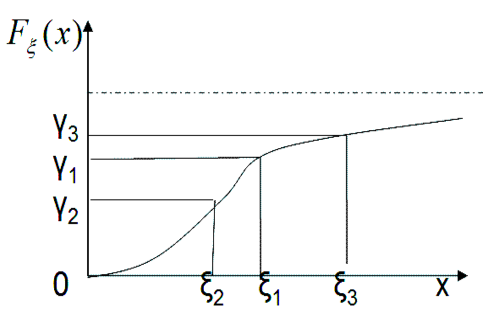
Есть несколько методов такой генерации. Они основаны на ряде теорем о распределении функций от случайных величин, а также на одном общем приеме, предложенном Дж.фон Нойманном еще в конце 40-х лет прошлого века.

Распределения функций от случайных величин.

Теорема 1.

Пусть $F_\xi(x)$ – это функция распределения некоторой случайной величины ξ, а γ – случайная величина с равномерным законом распределения на интервале [0, 1] : $F_\gamma(x)=x\ (0<x<1)$.

Тогда случайная величина $\xi = F_\xi^{-1}(\gamma)$,



где $F_\xi^{\ -1}$ – функция, обратная к $F_\xi(x)$, подчиняется закону распределения $F_\xi(x)$.
Доказательство:

$$F_{\xi}(x)=P\{\xi\le x\}=P\{F^{-1}(\gamma)\le x\}=P\{\gamma\le F(x)\}=\int_{0}^{F(x)} f_{\gamma}(y)dy=F(x)$$

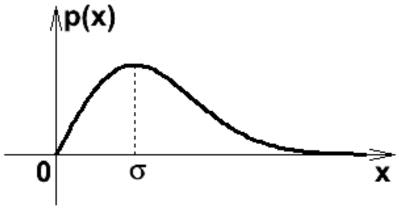
Следствие: Метод моделирования случайной величины с законом распределения *F(x)*: разыгрываем случайное число *γ*, равномерно распределенное в интервале (0,1) и вычисляем величину ***x=F⁻¹(γ)*.**

Пример применения Теоремы 1. Функция, обратная к функции распределения показательного закона распределения ***y= F(x)=1-exp(-λx)*,** будет

x = - Ln(y)/λ. Поэтому для моделирования случайной величины ξ с показательным законом распределения достаточно получить равномерно распределенное *y* с помощью встроенной функции *r*:=random; и вычислить *x* = -*Ln(r)/ λ.*

Пример 2. Моделирование случайной величины с распределением Релея.

Примером **распределения** **Рэлея** может служить распределением длины двумерного случайного вектора, каждая из компонент которого распределена по нормальному закону

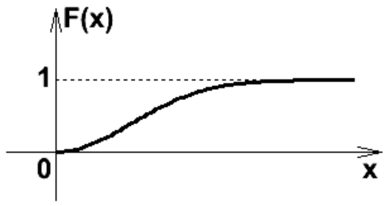


с параметрами $(0,\sigma)$. Плотность этого распределения имеет вид:

$$p(x)=\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}},\quad x\in[0;+\infty).$$

Соответственно, функция распределения Рэлея выглядит как:

$$F(x)=1-e^{-\frac{x^2}{2\sigma^2}},\ x\in[0;+\infty),\quad (*)$$



Функция, обратная к функции распределения (*) будет $x_i = \sigma\sqrt{-2\ln r_i}$ (здесь мы воспользовались тем, что число $1 - r_i$ имеет такое же распределение, как и само r_i).

Теорема 2.

Пусть $F_\xi(x)$ – это функция распределения некоторой случайной величины ξ, а функция $y=g(x)$ имеет обратную к ней однозначную функцию $g^{-1}(y)$.

Тогда случайная величина $\eta = g(\xi)$, будет распределена по закону $F_\eta(x) = F_\xi(g^{-1}(x))$. Доказательство:

$$F_\eta(x) = P\{\eta < x\} = P\{g(\xi) < x\} = P\{g^{-1}(g(\xi)) < g^{-1}(x)\} = P\{\xi < g^{-1}(x)\} = F_\xi(g^{-1}(x))$$

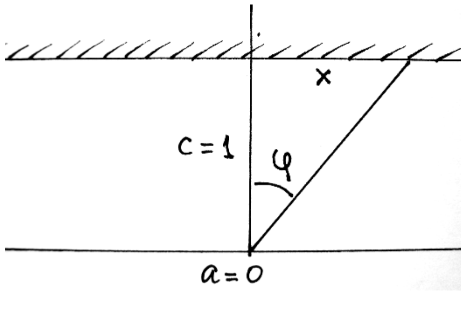
Применения Теоремы 2.

Пример 1. Пусть случайная величина ξ равномерно распределена в интервале (0,1), т.е. ее функция распределения $F_\xi(x) = x$ для $(0<x<1)$, $F_\xi(x)=0$ для $x \le 0$ и $F_\xi(x)=1$ для $x \ge 1$.

Требуется найти функцию распределения $F_\eta(x)$ случайной величины $\eta = \sqrt{\xi}$. Функция, обратная к $\sqrt{x}$, есть x^2. Поэтому по Теореме 2 $F_\eta(x) = F_\xi(x^2) = x^2$ в интервале $(0<x<1)$.

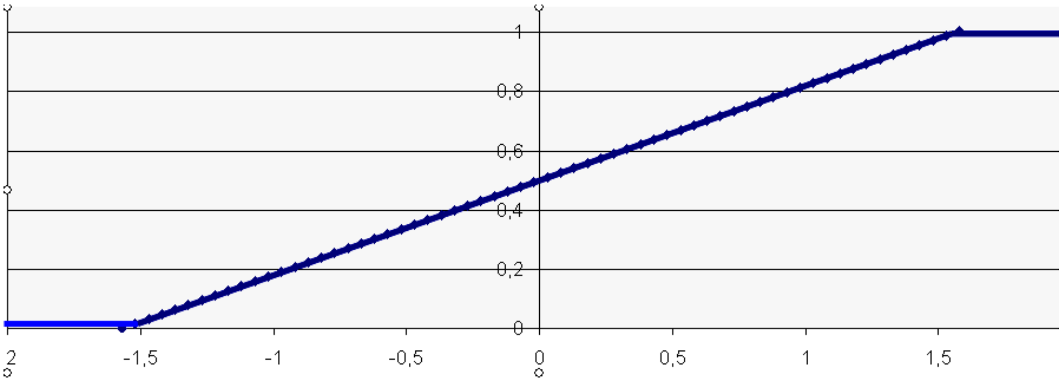
Соответственно, плотность распределения случайной величины η будет равна $f_\eta(x)=2x$ в том же интервале.

Пример 2. *Задача об облучении поверхности точечным источником.* Пусть точечный источник радиации излучает во все стороны равномерно по углу φ. Если расположить этот источник на расстоянии $c=1$ от бесконечной стены в начале координат, т.е. при значении параметра сдвига $a=0$, то как видно на рисунке слева, облучение в точке, лежащей на расстоянии x от начала



координат, будет зависеть от угла ***φ*** по формуле: $x = tg(\varphi)$, где угол ***φ*** будет распределен равномерно в отрезке $(-\pi/2 \le \varphi \le \pi/2)$.

Функция распределения угла ***φ*** будет $F_\varphi(t) = \frac{1}{\pi}(t + \frac{\pi}{2})$.



По теореме 2 для функции $x = tg(\varphi)$ отсюда получим функцию распределения $F(x) = F_\varphi(arctg(x)) = \frac{1}{\pi}(arctg(x) + \frac{\pi}{2})$. То-есть мы получили распределение Коши с параметрами c=1, a=0. Плотность этого распределения $f(x) = \frac{dF(x)}{dx} = \frac{1}{\pi(1+x^2)}$.

Вышеприведенный вывод подсказывает метод моделирования случайной величины с распределением Коши: по случайному числу $r \in (0,1)$ вычисляем случайную величину y=(r-1/2)*π, равномерно распределенную на отрезке (-π/2,π/2). Искомое случайное число с распределением Коши вычисляется как $x = tg(y)$. Для произвольных значений параметров ***a*** и ***c*** формула моделирования усложнится естественным образом: $x = c(a + tg(y))$.

Центральная предельная теорема (ЦПТ)

Сумма достаточно большого количества слабо зависимых случайных величин, имеющих примерно одинаковые масштабы (ни одно из слагаемых не доминирует, не вносит в сумму определяющего вклада), имеет распределение, близкое к нормальному.

Пусть $X_1, X_2, \ldots$ бесконечная последовательность независимых одинаково распределённых случайных величин, имеющих конечное математическое ожидание μ и дисперсию σ. Среднее арифметическое n величин обозначим

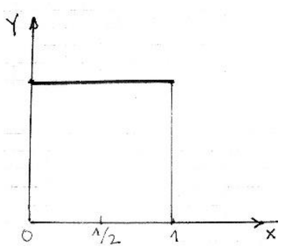
$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ $N(0,1)$ — нормальное распределение с нулевым математическим ожиданием и стандартным отклонением, равным единице.

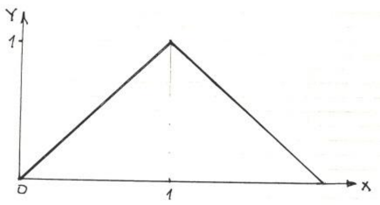
Тогда при $n \to \infty$. получим $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \to N(0,1)$

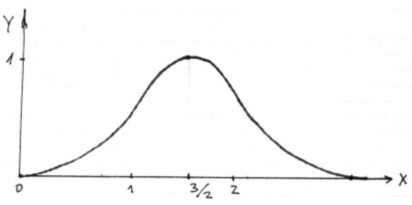
Пример приложения ЦПТ.

Алгоритм генерации случайных нормально распределенных последовательностей.

Складываем случ.числа ξ_i, $i=1,2\ldots$, равномерно распределенные в (0,1)







σ=1/12 **σ=1/6** **σ=1/4**

Наиболее рационально взять 12 слагаемых ξ_i, $i=1,12$, чтобы суммарная величина имела дисперсию 12/12=1, а для получения нулевого среднего достаточно из суммы вычесть ½*12=6. Таким образом, случайная величина $\eta = \sum_{i=1}^{12}\xi_i - 6$ будет с хорошим приближением иметь стандартное нормальное распределение

Пример 2. Приложение формулы для суммы показательно распределенных величин к генерации случайных величин с распределением Пуассона.

Будем укладывать на отрезок длины $\boldsymbol{\lambda}$ независимые случайные величины τ_i, $(i=1,2,\ldots)$ с показательным распределением $\boldsymbol{f_\tau(x)=exp(-x)}$.

Можно доказать, что количество k таких отрезков, умещающихся в отрезок длины λ, будет случайной величиной с распределением Пуассона, то-есть вероятность того, что на этом отрезке уложится ровно ***к*** таких величин, будет равна

$$p_k = P\{\sum_{i=1}^{k}\tau_i \le \lambda; \sum_{i=1}^{k}\tau_i + \tau_{k+1} > \lambda\} \quad = e^{-\lambda}\frac{\lambda^k}{k!}$$

Таким образом получена **вероятностная схема**, в которой случайное событие, состоящее в том, что сумма независимых случайные величин $\boldsymbol{\tau_i}$***, (i=1,2,…)*** с показательным распределением $\boldsymbol{f_\tau(x)=exp(-x)}$ накапливается так, что на ***(к+1)***-м шаге сумма $\sum_{i=1}^{k} y_i$ превышает величину $\boldsymbol{\lambda}$, с вероятностью, соответствующей распределению Пуассона.

Если вспомнить, что мы получаем τ_i , как $\boldsymbol{=-ln(r_i)}$***,*** то эта

схема приводит к простому алгоритму генерации случайных величин с распределением Пуассона с параметром $\boldsymbol{\lambda}$:

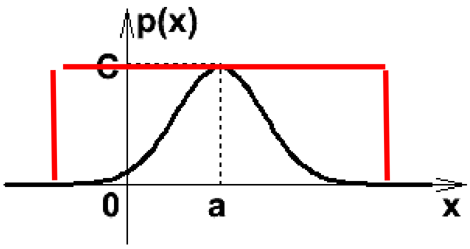
1. Готовим счетчик k=0 и произведение p=1, начинаем цикл ***while*** $\boldsymbol{(p>e^{-\lambda})}$
2. С помощью программы ***random*** получаем случайное число $\boldsymbol{x_i}$, равномерно распределенное в отрезке (0,1), умножаем $\boldsymbol{p=p^{*} x_i}$;увеличиваем счетчик k++;
3. Сравниваем $\boldsymbol{p>e^{-\lambda}}$, если больше, то целое число ***к-1*** выдается как очередное значение случайной пуассоновой величины с параметром $\boldsymbol{\lambda}$***.*** Если нет, то возвращаемся на пункт 2

Метод фон Ноймана или метод браковки

Если для функции распределения нет аналитического выражения, как например для нормального распределения, то Теорема 1 неприменима. Именно для этого случая Нойман предложил свой метод браковки.

Метод браковки:

Для генерации чисел ограничиваются некоторым заданным интервалом, например, $[a-3\sigma; a+3\sigma]$. На этом интервале строят прямоугольную область, совпадающую по высоте с максимумом кривой плотности распределения $p(x) = C\exp(-\frac{(x-a)^2}{2\sigma^2})$, где $C = \frac{1}{\sigma\sqrt{2\pi}}$. При σ=1 C≈0.4



Затем для получения каждого числа генерируют случайные точки ***(x_j;y_j)*** с равномерным распределением в области прямоугольника до тех пор, пока не выполнится условие ***y_j<p(x_j)***. Последовательность так «отбракованных» ***x_j*** и будет иметь нормальное распределение. Эффективность этого метода, т.е. отношение числа используемых равномерно распределенных точек к их общему числу пропорционально отношению площадей под кривой на рисунке к площади прямоугольника, т.е. примерно 1/(0.4*6) ≈ 40%, что много выше, чем для метода суммирования (1/12≈8%).

Однако метод браковки приблизителен и в отличие от метода суммирования не выдает редких случайных чисел, выходящих за интервал $[a-3\sigma; a+3\sigma]$.

Проверка правильности моделирования

Критерии согласия.

Критерий χ² Пирсона позволяет проверить гипотезу о соответствии выборки сгруппированной в *m* интервалов, т.е. гистограмы $h_1, h_2, \ldots, h_m$ распределению с плотностью *f(x|θ)*, где *θ* – параметр распределения. К.Пирсон предложил сравнивать значения h_k с их теоретическими значениями, т.е. для каждого интервала гистограммы Δ_kвычислять вероятность попадания в него p_k, умноженную на общее число группированых наблюдений $n = \sum_{k=1}^{m} h_k$. Находка Пирсона состояла в том, что он в качестве критерия согласия выборки с распределением предложил суммировать взвешенные квадраты разностей $(h_k - np_k)^2$, взяв в качестве весов величины, обратно пропорциональные этим теоретическим значениям $1/p_k$, и доказал, что при $n \to \infty$ полученная сумма

$$\chi^2 = \sum_{k=1}^{m} \frac{(h_k - np_k)^2}{np_k}$$

будет иметь известное распределение χ^2 с числом степеней свободы, равным $m-1$. Вероятности попадания в интервал Δ_k вычисляются как $p_k = \int_{\Delta_k} f(x \mid \vartheta) dx$, если значение параметра θ известно заранее. В противном случае параметр θ подлежит предварительной оценке по исходной выборке, так что **интеграл при вычислении вероятностей p_k должен браться при значении параметра, равным полученной оценке $\hat{\theta}$.** В случае, когда плотность распределения зависит от r параметров, число степеней свободы χ^2 будет $\tilde{m} = m - 1 - r$. Такая потеря этих степеней свободы связана с зависимостью слагаемых в формуле для χ^2. Следует также иметь в виду, что из-за асимптотического характера критерия χ^2 следует необходимость такой группировки выборки, которая должна обеспечить достаточную заполняемость интервалов наблюдениями, т.е. необходимо следить, чтобы все h_k были достаточно велики, обычно требуют, чтобы было $\min h_k \geq 5$; ($к=1,2,...m$). По заданному уровню достоверности критерия $1-\alpha$ из таблиц распределения χ^2 нетрудно найти такое критическое значение $\chi^2_{кр}$, чтобы выполнялось $P\{\chi^2 > \chi^2_{кр}\} = \alpha$. Поиск критического значения упрощается при достаточно большом значении числа интервалов группировки m $(m>20)$, когда в силу центральной предельной теоремы распределение χ^2, как суммы большого числа слагаемых, будет нормальным с большой степенью точности. И поэтому по правилу "трех сигм" можно считать, что при *α~0.001* получаем $\chi^2_{kp} = \tilde{m} + 3\sqrt{2\tilde{m}}$, где $\tilde{m}$ - число степеней свободы с учетом числа параметров плотности распределения.

Примеры применения критерия Пирсона.

1. Требуется проверить на нормальность выборку x_1, x_2, …, x_n объема n. Нормальное распределение зависит от двух параметров a и σ, оценками максимального правдоподобия которых являются выборочное среднее $\bar{x}$ и среднеквадратичное значение $s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ Для сведения задачи к проверке более простой гипотезы о том, что выборка взята из стандартного нормального распределения необходимо преобразовать каждое выборочное значение: $x_i^{'} = \frac{x_i - \bar{x}}{s}$. Далее, при группировке полученных $x_i^{'}; \mathrm{i} = 1,...,\mathrm{n}$ на m интервалов разумно разбить основную область их значений (-3,+3) на m-2 интервала с тем, чтобы в первом интервале подсчитывать число значений $x_i^{'} \le -3$, а в последнем – число случаев $x_i^{'} > 3$. Для вычисления вероятностей p_k попадания в k-й интервал гистограммы мы должны проинтегрировать плотность нормального распределения $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ по интервалу Δ_k. Однако интеграл от плотности нормального распределения не вычисляется аналитически, поэтому нам для нахождения p_k придется воспользоваться приближением, основанным на теореме о среднем, и аппроксимировать p_k как значение плотности в средней точке интервала, умноженное на его длину. Если принять, что все интервалы гистограммы имеют одинаковую длину: $\Delta=6/(m-2)$, то для p_k получается простая формул: $p_k = \int_{-3+\Delta(k-1)}^{-3+\Delta k} f(x)dx = \Delta * f(-3+\Delta(k-0.5))$.

После этого вычисляем концевые вероятности

$p_1 = p_m = 0.5(1-\sum_{k=2}^{m-1} p_k)$. Далее вычисляется значение критерия $\chi^2 = \sum_{k=1}^{m}\frac{(h_k - np_k)^2}{np_k}$ и его критическое значение $\chi^2_{kp} = \tilde{m} + 3\sqrt{2\tilde{m}}$, которое при $\tilde{m}$ *= 32* будет равно $\chi^2_{кр} \approx 53.24$.

Для проверки гипотезы, что наша выборка распределена по нормальному закону, сравниваем вычисленное по гистограмме значение χ^2 с $\chi^2_{кр.}$ и если $\chi^2 < 53.24$, то гипотеза о том, что исследуемая случайная величина подчиняется нормальному закону распределения, принимается на уровне значимости $\alpha \sim 10^{-3}$, в противном случае - гипотеза отклоняется.

Заметим, что для выполнения условия: $\min h_k \geq 5$, объем выборки n следует брать достаточно большим $(n \geq 2000)$

2. Для проверки гипотезы о распределении выборки по закону Коши с плотностью $f(x) = \frac{1}{\pi(1+(x-a)^2)}, x \in (-\infty; +\infty)$ следует воспользоваться тем, что известна функция распределения $F(t) = \frac{1}{\pi} arctg(t-a) + \frac{1}{2}$.Это позволяет вычислять значения вероятностей попадания в *к*-й интервал гистограммы по точной формуле: $p_k = F(x_{k+1}) - F(x_k)$. Из-за большого разброса чисел с распределением Коши относительно их среднего значения гистограмму можно построить лишь в некотором ограниченном интервале, определяемом оценкой центра распределения Коши по медиане выборки: $\hat{a} = x_{(N/2)}$, например, в интервале $[\hat{a}-5; \hat{a}+5]$ длины 10. Тогда для границы *k*-го интервала гистограммы x_{k+1} и x_k, можно вычислять по формуле: $x_k = \hat{a} - 5 + \Delta * (k-1)$, $k = 2, \ldots m-1$, $\Delta = 10/m$.

Как и в предыдущем примере, находим $p_1 = p_m = 0.5(1-\sum_{k=2}^{m-1} p_k)$, и потом вычисляем значение критерия $\chi^2 = \sum_{k=1}^{m}\frac{(h_k - np_k)^2}{np_k}$. Его критическое значения $\chi^2_{кр} = \tilde{m} + 3\sqrt{2\tilde{m}}$. Поскольку у закона распределения Коши только один параметр, то при 52-х интервалах гистограммы получим число степеней свободы критерия равным $\tilde{m} = 51$, что даст $\chi^2_{кр} \approx 81.2985$.

Критерий Колмогорова-Смирнова:

Для критерия Колмогорова-Смирнова выбор меры расхождения связан с эмпирической функцией распределения $F_N(x)$. А именно, рассматривается **статистика Колмогорова**:

$$D_N = \sup_x \left| F_N(x) - F(x) \right|.$$

Эта статистика D_N является случайной величиной, поскольку ее значение зависит от случайного объекта $F_N(x)$. Важное с теоретической точки зрения свойство критерия, основанного на D_N, состоит в том, что он **состоятелен** против любой альтернативы G ≠ F. Это означает, что любое отличие распределения выборки от теоретического будет с его помощью обнаружено, если наблюдения будут продолжаться достаточно долго.
Статистика Колмогорова вычисляется по формуле:

$$D_N = \max_{0 \le i \le (N-1)} \left\{ \left| \frac{i}{N} - F(x_{(i)}) \right|, \left| \frac{i+1}{N} - F(x_{(i)}) \right| \right\},$$

где $x_{(i)}$ - элементы вариационного ряда, построенного по исходной выборке.

Коснемся вопроса об асимптотическом распределении этих функций. Согласно **теореме Колмогорова**, если гипотеза H_0 верна и $F(x)$ непрерывна, то

1. распределение статистики D_N является одним и тем же для любой функции распределения $F(x)$;

2. у последовательности $\sqrt{N}D_N$ существует предельное распределение:

$$\lim_{N\to\infty} P(\sqrt{N}D_N < t) = K(t) = \sum_{j=-\infty}^{+\infty} (-1)^j e^{-2j^2t^2}.$$

Это предельное распределение носит название **распределения Колмогорова**.

Теорема Колмогорова является основой для построения соответствующего критерия согласия: если статистика $\sqrt{N}D_N$ превышает **квантиль** распределения Колмогорова K_α заданного уровня значимости α, то гипотеза H_0 отвергается. В противном случае гипотеза принимается на уровне α.

Если N достаточно велико, то K_α можно приблизительно рассчитать по формуле:

$$K_\alpha \approx \sqrt{-\frac{\ln\frac{\alpha}{2}}{2}}.$$

В частности, при $N > 100$ для 95% -го уровня значимости $K_\alpha \approx 1.36$, а для 99%-го - $K_\alpha \approx 1.63$.