Расчетное задание 2

Статистическая обработка случайных последовательностей. Идентификация законов распределения.

Дано

В результате измерений получена выборка x1, x2, …, xN из генеральной совокупности с неизвестным законом распределения. Выборочные значения расположены в файлах (для каждой группы свой каталог, для каждого варианта файл с названием Distribtuion i), где i – номер варианта (по номеру в списке преподавателя).

Варианты

Число значений N, а также сам массив выборочных значений записаны в файле и отделены друг от друга пробелами. В случае дискретного распределения значения целые, в случае непрерывного – вещественные.

Справочная информация

Вся теоретическая часть по работе изложена в [1], а также в разделах помощи Matlab, в частности Statistic Toolbox.

В приложении 1 к данному заданию описаны основные распределения, даны формулы плотностей, функций, моментов и имеющихся аналитических оценок параметров по методу максимального правдоподобия. Также в приложении 1 представлены графики плотностей и функций основных распределений. Ими разумно пользоваться при подборе распределения под имеющуюся выборку путем сравнения графиков:

- относительной гистограммы и теоретической плотности распределения;
- эмпирической и теоретической функций распределения.

В приложении 2 приведены примеры вычисления оценок параметров распределений при подгонке параметров распределений к имеющейся выборке. Поэтому при использовании метода моментов и максимального правдоподобия целесообразно ознакомиться и разобраться в этих примерах.

В приложении 3 приведены теоретические основы трех основных статистических методов проверки гипотез о виде плотности распределения: хи-квадрат, Колмогорова-Смирнова и Мизеса.

Задание:

1. Статистическая обработка случайных последовательностей

1.1. Считать выборку X из файла. Создать на ее основе 10 подвыборок – для этого перемешать выборку (например, командой Xperm=X(randperm(length(X))) и последовательно сформировать подвыборки (Xpodv(i) = Xperm(1+(i-1)*N/10:i*N/10))

1.2. Построить выборочную функцию распределения F(x) (она должна быть ступенчатой!!!, можно воспользоваться функцией cdfplot)

1.2. Построить абсолютную и относительную гистограммы на разных графиках (функция hist строит абсолютную гистограмму; чтобы построить относительную гистограмму выборки, нужно разделить все ее значения на ее объем). Внимательно отнеситесь к выбору количества (ширины) интервалов или столбцов - оно выбирается таким образом, чтобы самый "бедный" интервал содержал **3 ÷ 5** выборочных значений. Если у распределения есть тяжелые хвосты (несколько значений в области значений, очень далеко отстоящей от скопления основной массы данных), то желательно их отбросить. Например, если 99.9 % значений, находящихся в диапазоне [-20 20] и 0.1 % значений, находящихся в диапазоне [-20000 20000], то последние 0.1 % не позволят нормально построить гистограмму и их желательно просто не учитывать при построении гистограммы (НО помнить, что они есть и характеризуют распределение как имеющее тяжелый хвост – Коши, Парето к примеру).

1.3. Определить точечные оценки:

1.3.1. моментов

- первого начального (среднее арифметическое - mean, медиана - median, средина размаха (min + max)/2)
- центральных моментов: второго-дисперсии (var), третьего, четвертого (moment) по выборочной функции распределения

Для оценки первого начального момента использовать среднее арифметическое, выборочную медиану, средину размаха. Определить моду (максимум на графике плотности).

1.3.2. асимметрии и эксцесса (функции skewness, kurtosis);

1.3.3. границ интерквантильного промежутка Jp для P=0.95 только по полной выборке (функция quantile)

1.3.4. характеристики по пп. 1.3.1-1.3.2 по подвыборкам, сформированным в п. 1.1.

Результаты представить в таблице следующей формы.

	$\overline{x}$	x_{med}	x_{cp}	s^2	s	$\dot{m}_3$	$\dot{m}_4$	As	Ex
N									
N/10									
N/10									
…									
N/10									

Представить эти же результаты графически точками на осях с указанием масштаба на этих осях по форме:

Прим. 1 Для проверки правильности результатов нужно убедиться в близости характеристик, посчитанных по полной выборке с характеристиками, посчитанными по подвыборкам.

Прим. 2. Значения характеристик по подвыборкам не должны равномерно располагаться вокруг значений характеристики по всей выборке – это свидетельствует о том, что подвыборки брались из отсортированной выборки, что в свою очередь является ошибкой.

оценки м.о.

\+ + + + + + ◆ + + + + $\bar{x}$ (◆ - оценки по **n** = n=N)

\+ + + + + ◆ + + + + + $\tilde{x}_{med}$ (+ - оценки по **n** = n=N/10)

\+ + + + ◆ + + + + + + x_{cp}

оценки дисперсии

\+ + + + + + ◆ + + + +

оценки третьего центрального момента

\+ + + + + + ◆ + + + +

оценки четвертого центрального момента

\+ + + + + + ◆ + + + +

оценки асимметрии

\+ + + + + + ◆ + + + +

оценки эксцесса

\+ + + + + + ◆ + + + +

1.4. Определить интервальные оценки с доверительной вероятностью Q=0.8:

- первого начального и второго центрального моментов (вычисления выполнить по полной выборке и по отдельным частям, как в п. 2.1.4 - по N/10 значений в каждой частичной выборке). Прим. Значения обратных функций распределения Стьюдента и Хи-квадрат удобно вычислять с помощью функций tinv и chi2inv соответственно. Нанести на эти характеристики соответствующие значения точечных оценок (для проверки правильности доверительный интервал должен располагаться вокруг точечной оценки).
- интерквантильного промежутка J для P=0.95:
  - по всей выборке с помощью непараметрических толерантных пределов, симметричных относительно среднего арифметического и относительно нуля. Прим. Количество статистически эквивалентных блоков k, отбрасываемых от выборки при нахождении непараметрических толерантных пределов, симметричных относительно среднего арифметического определяется из неравенства: $\sum_{m=n-k}^{n} C_n^m P^m (1-P)^{n-m} \le 1-Q$ (решение данной проблемы может быть выполнено последовательным увеличением k от 0 до тех пор, пока неравенство не начнет выполняться; следует учитывать, что число сочетаний C_n^m при больших n необходимо считать с применением формулы Стирлинга). Результирующий предел будет равен $\left[X_{k/2} X_{N-k/2}\right]$ при четном k или $\left[X_{(k-1)/2} X_{N-(k-1)/2}\right]$ при нечетном k. В случае если пределы симметричны относительно нуля, то необходимо преобразовать выборку, заменив отрицательные значения на их модуль и отбросить справа (k-1) эквивалентных блоков. Результирующий предел будет равен $\left[-X_{N-k+1} X_{N-k+1}\right]$.
  - по частичным выборкам с помощью параметрических толерантных пределов, считая закон распределения генеральной совокупности нормальным. Прим. Значения толерантных множителей можно найти в [1].

Результаты представить только графически аналогично тому, как описано выше – под графическим представлением соответствующей точечной оценки, предусмотрев для каждого варианта расчета отдельную ось. Графическое представление толерантных пределов — также на отдельных осях для каждого варианта. Все оси обозначить.

Сделать выводы относительно ширины доверительных интервалов. Сравнить:

а) интерквантильные промежутки с толерантными пределами

б) параметрические и непараметрические толерантные пределы, симметричные относительно среднего арифметического и относительно нуля.

2. Идентификация закона и параметров распределения

В данном задании осуществляется идентификация закона распределения исходной выборки. Для этого вначале методом проб подбирается распределение, а затем различными способами определяются параметры этого распределения. В завершении осуществляется проверка гипотез о соответствии предполагаемых законов распределения экспериментальным данным с помощью ТРЕХ критериев: "хи-квадрат", Колмогорова-Смирнова, "омега-квадрат".

Подсказка возможные распределения:

Непрерывные – арксинус, треугольное, Коши, Симпсона, Лапласа, Хи-квадрат, экспоненциальное, нормальное, равномерное, Симпсона, Стьюдента, логнормальное, гамма, Рэлея, Парето.

2.1. Начальный выбор распределения

Для начальной ориентировки в выборе закона использовать вид гистограммы, функции распределения, соотношения между моментами и полученные значения **эксцесса и асимметрии**. Удобная утилита Matlab disttool позволяет построить графики многих (но не всех!) законов (плотностей) и функций распределения, варьируя и подбирая их параметры. В результате нужно определиться с тремя основными распределениями, которые и будут идентифицироваться.

2.2. Определение параметров теоретических распределений.

Для выбранных теоретических распределений необходимо определить точные значения параметров, наиболее подходящие для описания выборки. Это необходимо сделать двумя способами:

- с помощью метода моментов, когда теоретические моменты приравниваются к выборочным и решается система уравнений по числу неизвестных параметров распределения.

- с помощью метода максимального правдоподобия – в случае, если для распределения известны аналитические ММП-оценки, можно воспользоваться ими. В общем случае необходимо найти ММП-оценки численными методами. Для этого в Matlab уже написано множество fit-функций под большое число распределений (normfit и др). В случае, если распределения нет в Matlab, его можно задать в форме встроенной функции и воспользоваться командой mle, передав туда эту функцию и начальные приближения для значений параметров (можно воспользоваться оценками метода моментов). Есть замечательная утилита Matlab – dfittool, позволяющая производить идентификацию через удобный интерфейс. Для распределений, отсутствующих в Matlab, следует использовать функцию mle (см. Приложение 1 – примеры оценки неизвестных параметров).

Сравнить оценки, полученные методом моментов и ММП. Для этого построить

- эмпирическую и теоретические функцию распределения (на 1 графике)

- гистограмму и теоретические плотности распределения (на 1 графике)

Т.о. должно быть 6 графиков (3 распределения * 2 характеристики), причем на каждом из графиков должно быть по 3-**4 зависимости** (1-эмпирическая, 2-теоретическая для оценки параметров по методу моментов, 3 и 4-теоретическая для оценки параметров по методу ММП численно и если есть, то аналитически). По графикам оценить степень сходства эмпирических и теоретических характеристик. Написать, какой метод оценки параметров дает большую точность.

2.3. Произвести проверку гипотез относительно выбранных теоретических законов распределения и их параметров (по методу ММП и моментов). Проверку провести по трем критериям - "хи-квадрат", Колмогорова-Смирнова, "омега-квадрат". Критерии можно реализовать как вручную, так и воспользоваться функциями Matlab – chi2gof – критерий Хи-квадрат, kstest – критерий Колмогорова-Смирнова. Критерий Мизеса необходимо реализовать самим и воспользоваться таблицей из [1]. Сравнить полученные статистики критериев с критическими значениями. Выбрать наиболее подходящие распределения, исходя из значениев статистики критериев.

2.4. Привести итоговую таблицу, в которой для каждого из 3 распределений приведены по 3 вида оценок (метод моментов, ММП-аналитика, ММП-численный), и

для каждого из уже 9 вариантов распределений и оценок – результаты проверки гипотез по 3 критериям – статистика критерия и критическое значение.

	Распределение 1			Распределение 2			Распределение 3		
Название									
Формула плотности									
	Мет.мом.	ММП-аналит	ММП-числ	Мет.мом.	ММП-аналит	ММП-числ	Мет.мом.	ММП-аналит	ММП-числ
Пар-р 1									
Пар-р 2									
Хи-квадрат - статистика									
Хи-квадрат – критич.знач									
Хи-Квадрат - вывод									
Колм.-Смирнова - статистика									
Колм.-Смирнова – крит.значение									
Колм.-Смирнова - вывод									
Мизеса – статистика									
Мизеса – критич.значение									
Мизеса - вывод									

Прим. 1. Вначале можно воспользоваться множеством критериев для нормального распределения – ttest, ztest, vartest.

Прим. 2. В отчете отобразить все ваши пробы относительно выбора подходящего закона распределения, а не одну последнюю (наиболее подходящую).

Литература

1. Солопченко Г.Н. Теория вероятностей и математическая статистика

Приложение 1 Формулы, характеристики и графики плотностей основных распределений

Распределения дискретных СВ

Распределение	**Плотность вероятности**	**Функция распределения**	**Числовые характеристики**	**Производящая функция моментов**	**Оценки по ММП**
Биномиальное	$P_n(k) = C_n^k p^k (1-p)^{n-k}$ $P_n(k) \approx \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$F_n(y) = \sum_{k=0}^{\lfloor y \rfloor} C_n^k p^k q^{n-k}$	$M(X) = np; D(X) = npq$ $As = \frac{1-2p}{\sqrt{npq}}; ex = \frac{1-6pq}{npq}$ $\mathrm{Mod} = \lfloor (\mathrm{n}+1)\mathrm{p} \rfloor;$	$M_X(\nu) = (pe^{\nu} + q)^n$	$p = \frac{\sum_{i=1}^{n} x_i}{mn}$
Пуассона	$P(k) = \lambda^k e^{-\lambda} / k!$	$F(k) = \frac{\Gamma(\mathrm{k}+1, \lambda)}{\lambda!}$	$M(X) = \lambda; D(X) = \lambda; As = \lambda^{-0.5}$ $Ex = \lambda^{-1}; \mathrm{Mod} = \lfloor \lambda \rfloor$	$M_X(\nu) = \exp(\lambda(e^{\nu} - 1)$	$\lambda = \overline{x_{\hat{a}}}$
Геометрическое	$P(k) = q^k p$		$M(X) = q/p; D(X) = q/p^2$	$M_X(\nu) = p/(1-qe^{\nu})$	
Равномерное	$P(k) = \begin{cases} \frac{1}{n}, a \le k \le b \\ 0, else \end{cases}$	$F(k) = \begin{cases} 0, k < a \\ (k-a+1)/n, a \le k \le b \\ 1, k > b \end{cases}$	$M(X) = (a+b)/2;$ $D(X) = (n^2-1)/12$ $As = 0$	$M_X(\nu) = \frac{e^{a\nu} - e^{(b+1)\nu}}{n(1-e^{\nu})}$	

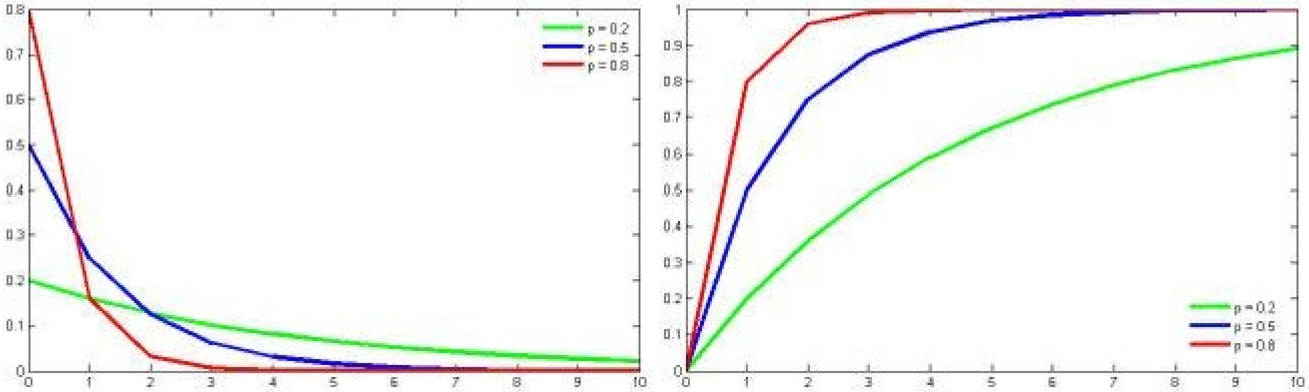
Распределения непрерывных СВ

Распределение	Плотность вероятности	Функция распределения	Числовые характеристики	Производящая функция моментов (хар.функция)	Оценки по ММП
Нормальное	$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$F(x)=0.5+\Phi((x-a)/\sigma)$ $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x\exp(-t^2/2)dt$	$M(x)=a; D(x)=\sigma^2;$ $As=0, Ex=0$	$M_x(\nu)=\exp(a\nu+\frac{\sigma^2\nu^2}{2})$ $\phi_x(\nu)=\exp(ai\nu-\frac{\sigma^2\nu^2}{2})$	$a=\overline{x};$ $\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\overline{x_в})^2}{N}$
Логнормальное	$p(x)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(x)-a)^2}{2\sigma^2}\right)$	$F(x)=0.5+\Phi((\ln(x)-a)/\sigma)$ $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x\exp(-t^2/2)dt$	$M(X)=e^{a+\sigma^2/2}; Med=e^a;$ $D(X)=(e^{\sigma^2}-1)e^{2a+\sigma^2};$		
Коши	$p(x)=\frac{\Delta}{\pi(\Delta^2+(x-c)^2)}$	$F(x)=\frac{1}{\pi}arctg(\frac{x-c}{\Delta})+0.5$ $F^{-1}(x)=c+\Delta tg(\pi(x-0.5))$	Med=Mod = c	$\phi_X(\nu)=\exp(ci\nu-\Delta\|\nu\|)$	
arcsin	$p(x)=\frac{1}{\pi\sqrt{a^2+(x-c)^2}}$	$F(x)=\begin{cases}0, x<c-a\\ \frac{1}{2}+\frac{1}{\pi}\arcsin\left(\frac{x-c}{a}\right),\\ 1, x>c+a\end{cases}$	$M(X)=Med=c;$ $D(X)=a^2/2;$ $As=0, Ex=1.5$		
Лапласа	$p(x)=\frac{\lambda}{2}\exp(-\lambda\|x-c\|)$		$M(X)=x_{0.5}=x_{\mathrm{mod}}=c;$ $D(X)=2/\lambda^2;$ $As=0, Ex=6$	$\phi_\xi(\nu)=\frac{\lambda^2}{\lambda^2+\nu^2}e^{j\nu c}$	$\hat{c}=x_{med};$ $\hat{\lambda}=N(\sum_{i=1}^{N}\|xi-\hat{c}\|)^{-1}$
Показательное (экспоненциальное)	$p(x)=\begin{cases}\lambda e^{-\lambda x}, x>=0\\ 0, x<0\end{cases}$	$F(x)=\begin{cases}1-e^{-\lambda x}, x>=0\\ 0, x<0\end{cases}$	$\mu_1(x)=1/\lambda; D(\xi)=1/\lambda^2;$ $As=2, Ex=6, \mathrm{Mod}=0,$ $\mathrm{Med}=\ln(2)/\lambda$	$M_x(\nu)=\lambda/(\lambda-\nu)$ $\phi_x(\nu)=\lambda/(\lambda-i\nu)$	$\lambda=1/\overline{x}$
Гамма-распределение (Эрланга)	$p(x)=\begin{cases}x^{k-1}\frac{e^{-x/\theta}}{\theta^k\Gamma(k)}, x\geq 0\\ 0, else\end{cases}$ $\Gamma(k)=(k-1)\Gamma(k-1);$ $\Gamma(0.5)=\sqrt{\pi}$		$M(X)=k\theta; D(X)=k\theta^2;$ $As=2/\sqrt{k}, Ex=6/k;$	$M_X(\nu)=(1-\theta\nu)^{-k}$ $\phi_X(\nu)=(1-\theta i\nu)^{-k}$	

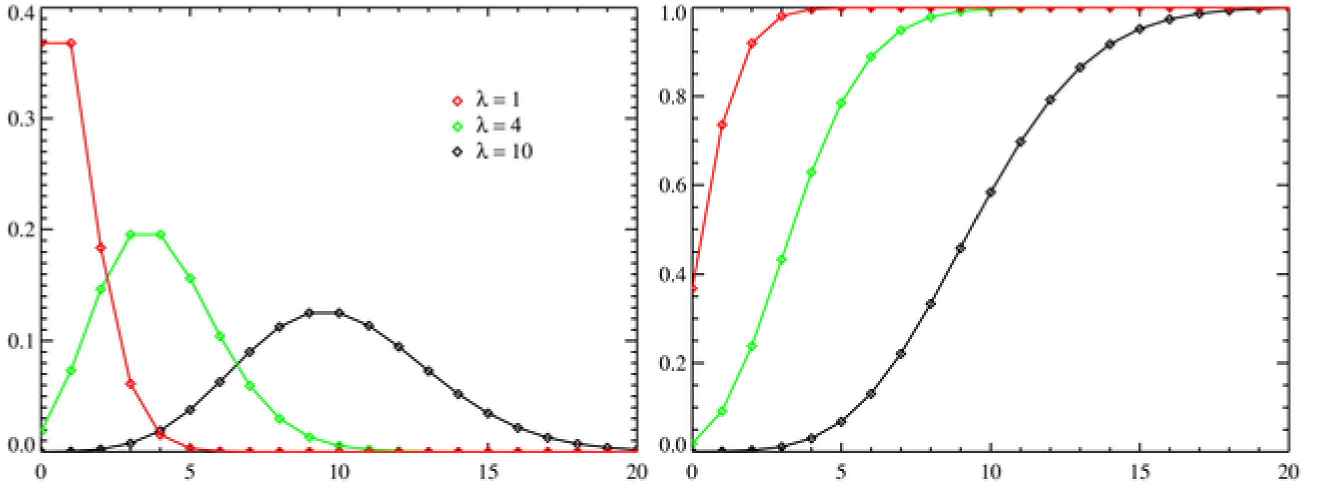
Хи-квадрат	$p(x)=\frac{(1/2)^{n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2}$	$F(x)=\frac{\gamma(n/2,x/2}{\Gamma(n/2)})$	$M(X)=n; D(X)=2n;$ $Med\approx n-2/3$ $As=\sqrt{8/n}, Ex=12/n;$	$\phi_X(\nu)=(1-2i\nu)^{-n/2}$	
Стьюдента	$p(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(n/2)}(1+\frac{y^2}{n})^{-\frac{n+1}{2}}$		$M(X)=Med=\mathrm{mod}=0;$ $D(X)=n/n-2; n>2$ $As=0, n>3,$ $Ex=(3n-6)/(n-4), n>4;$		
Равномерное	$p(x)=\begin{cases}\frac{1}{b-a}, a\le x\le b\\ 0, else\end{cases}$	$F(k)=\begin{cases}0, x<a\\ (x-a)/b-a, a\le x\le b\\ 1, x>b\end{cases}$	$M(X)=Med=(a+b)/2;$ $D(X)=(b-a)^2/12$ $As=0, Ex=-1.2$	$M_X(\nu)=\frac{e^{\nu a}-e^{\nu b}}{\nu(b-a)}$ $\phi_X(\nu)=\frac{e^{\nu ia}-e^{\nu ib}}{\nu i(b-a)}$	
Треугольное	$p(x)=\begin{cases}\frac{2a-\lvert x-c\rvert}{4a^2}, \lvert x-c\rvert\le 2a\\ 0, else\end{cases}$		$M(X)=Med=c;$ $D(X)=2a^2/3$		
Симпсона	$p(x)=\begin{cases}\frac{3a^2-\lvert x-c\rvert^2}{8a^3}, \lvert x-c\rvert\le a\\ \frac{(3a-\lvert x-c\rvert)^2}{16a^3}, a<\lvert x-c\rvert\le 3a\\ 0, else\end{cases}$		$M(X)=Med=c;$ $D(X)=a^2$		
Рэлея	$p(x)=\frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right), x\ge 0, \sigma>0$	$F(x)=1-\exp\left(-\frac{x^2}{2\sigma^2}\right), x\ge 0$	$M(X)=\sqrt{\pi/2}\sigma;$ $D(X)=(2-\pi/2)\sigma^2$		
Парето	$p(x)=\frac{kx_m{}^k}{x^{k+1}}, x\ge x_m$	$F(x)=1-\left(\frac{x_m}{x}\right)^k, x\ge x_m$	$M(X^n)=kx_m{}^n/(k-n);$ $M(X)=kx_m/(k-1)$		

Графики плотностей распределений дискретных СВ

Геометрическое

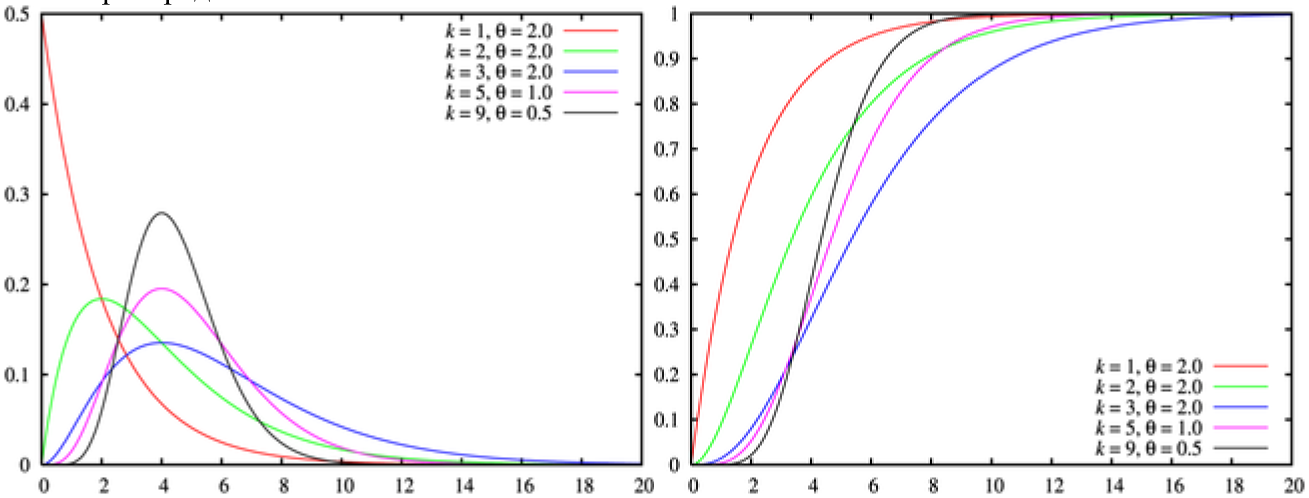


Пуассона

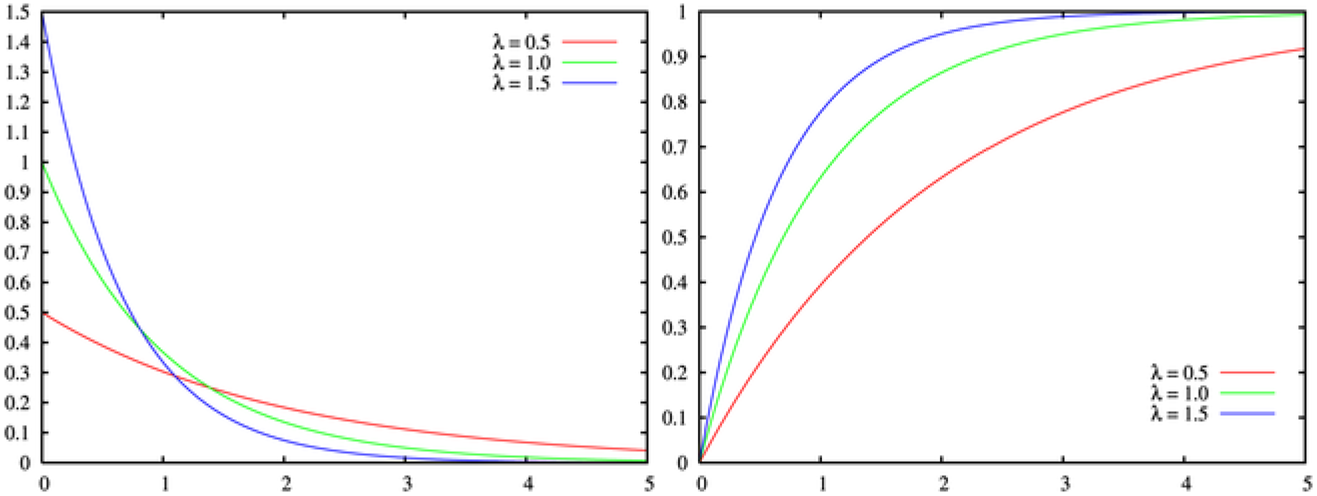


Графики плотностей распределений непрерывных СВ

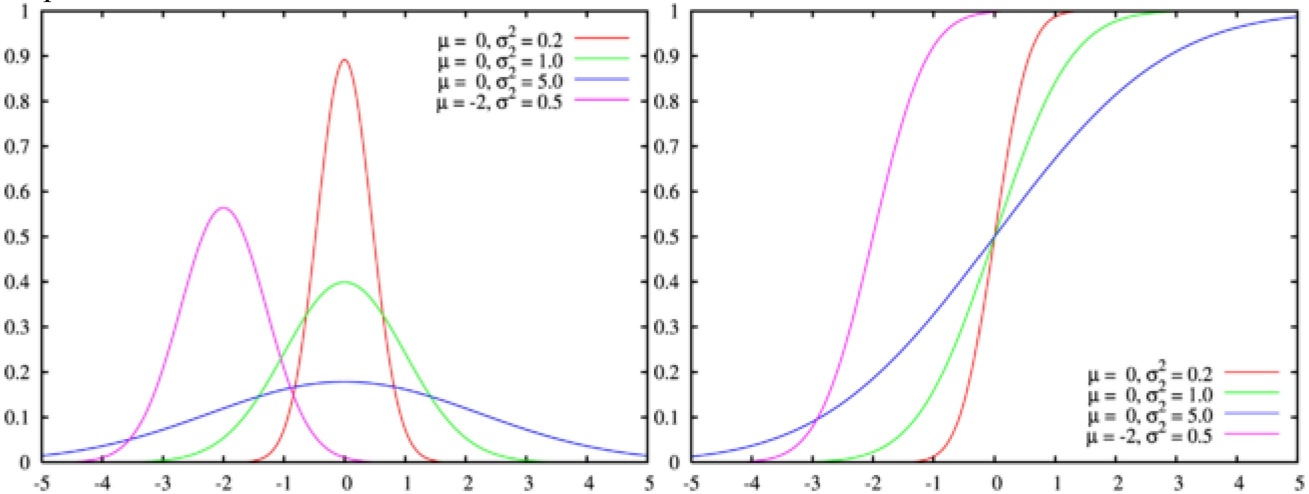
Гамма-распределение



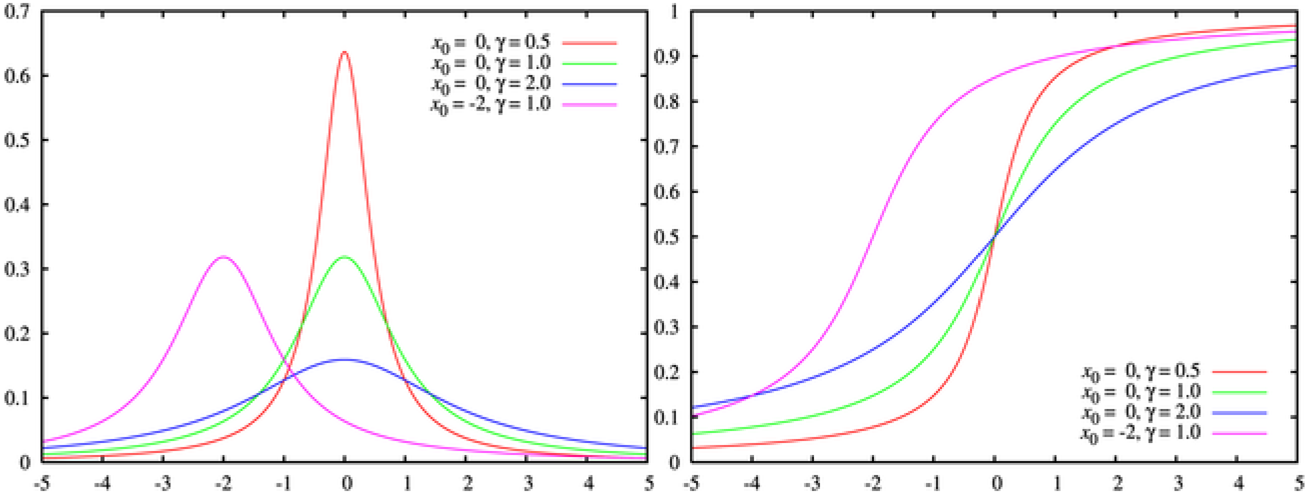
Экспоненциальное



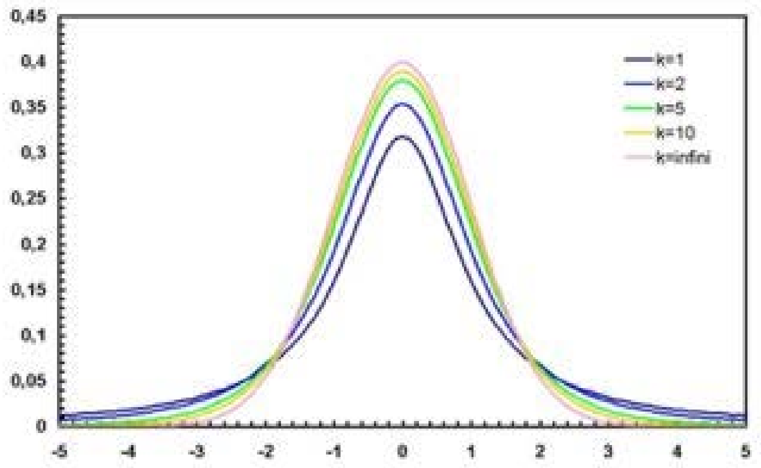
Нормальное

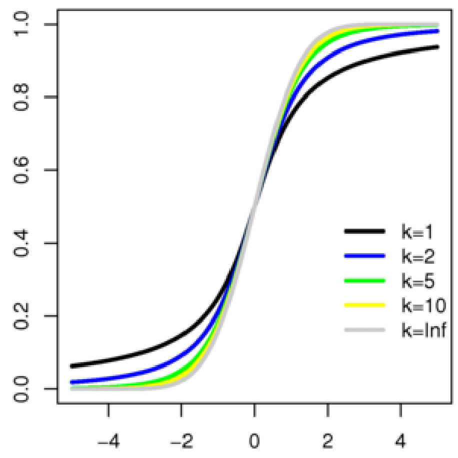


Коши

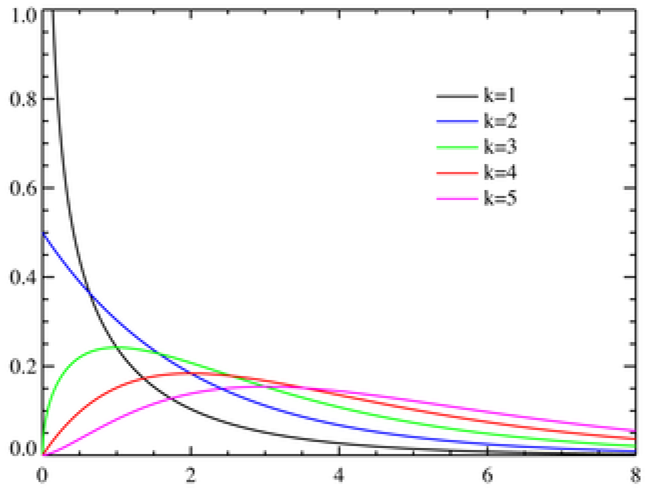


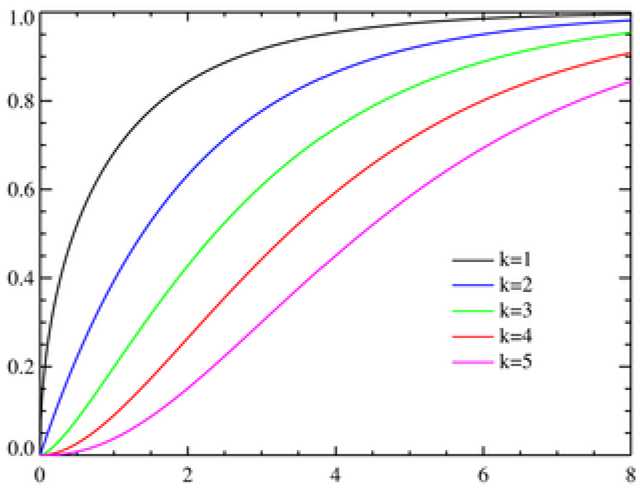
Стьюдента



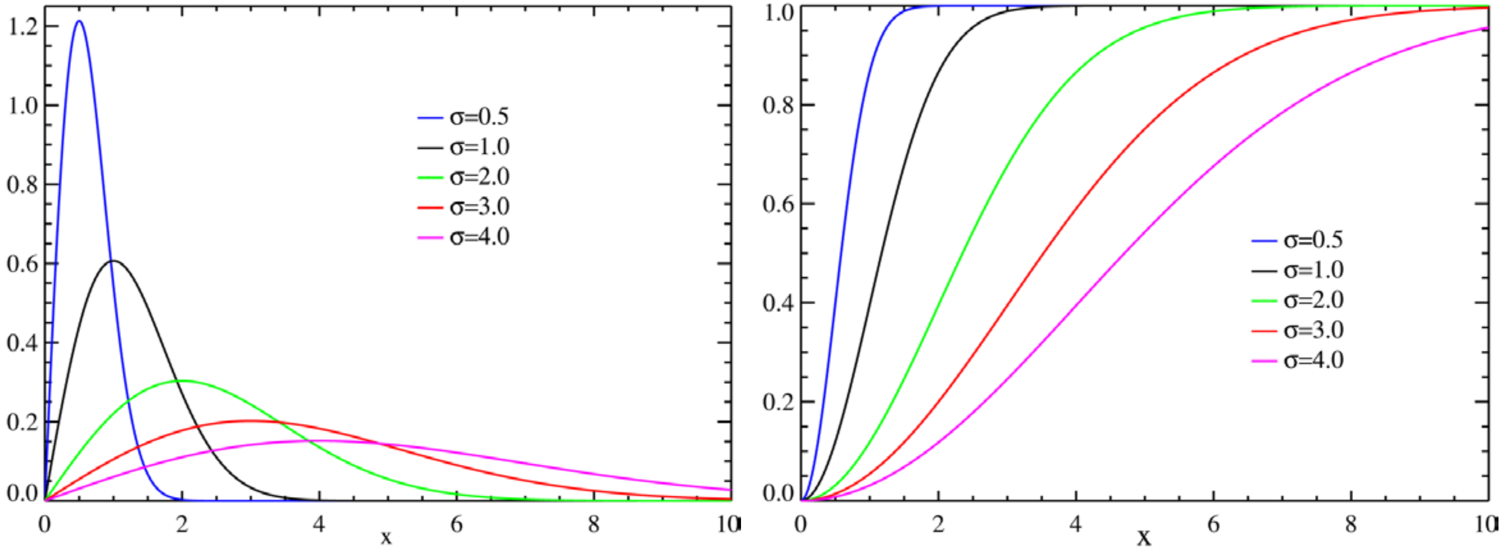


Хи-квадрат

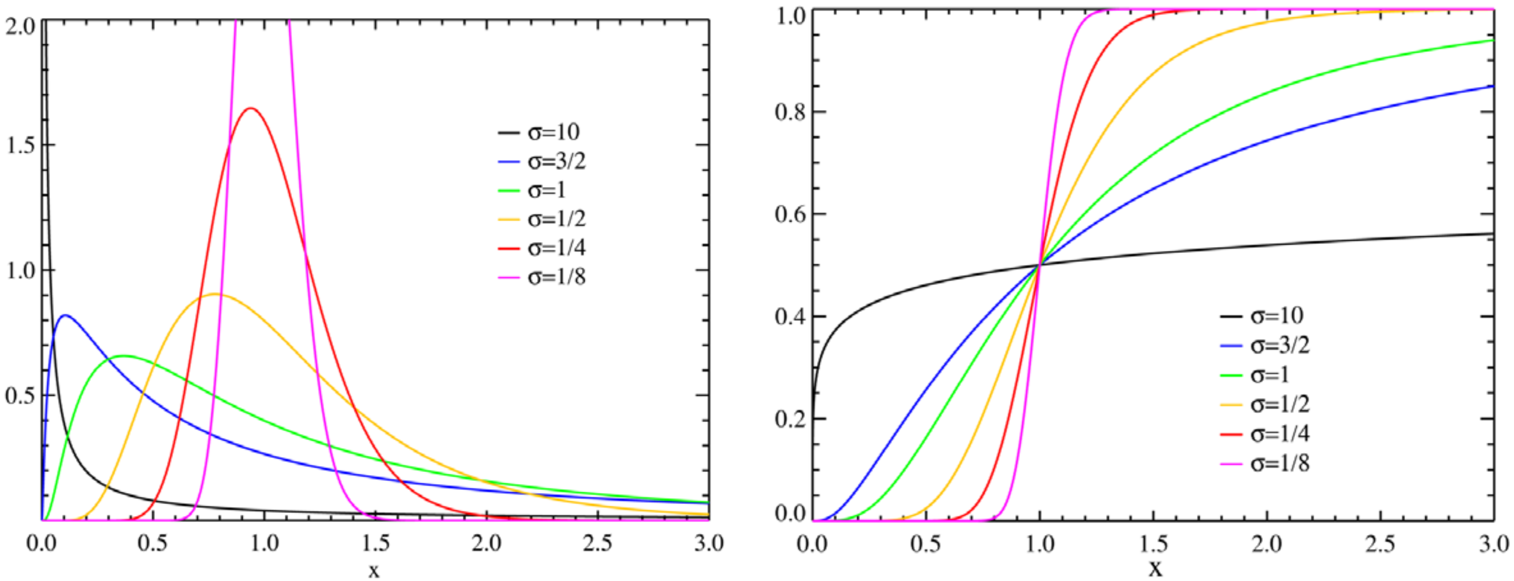




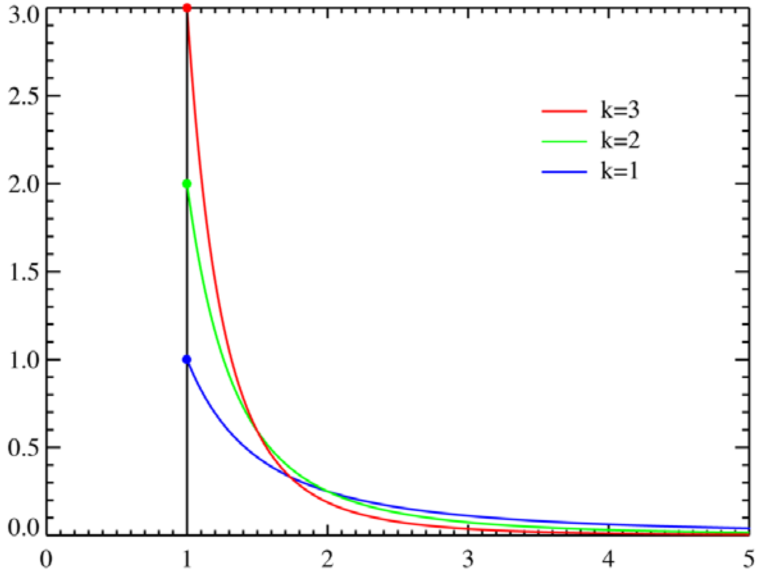
Рэлея

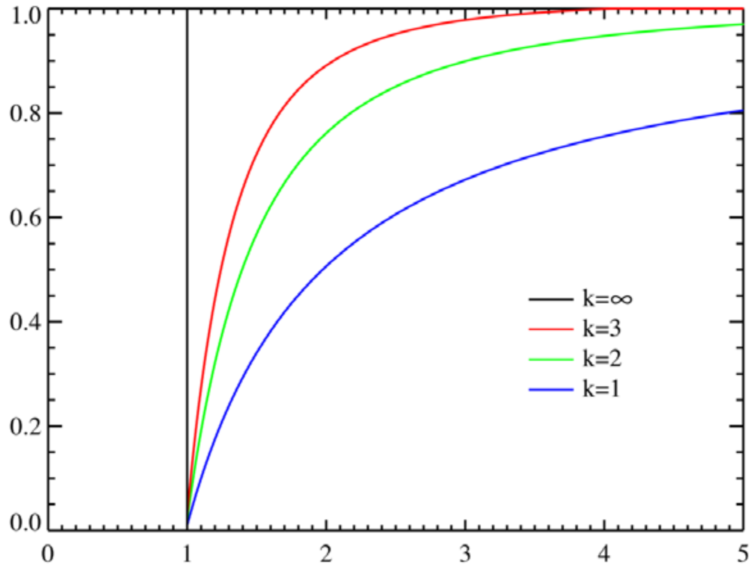


Логнормальное



Распределение Парето





Другие распределения

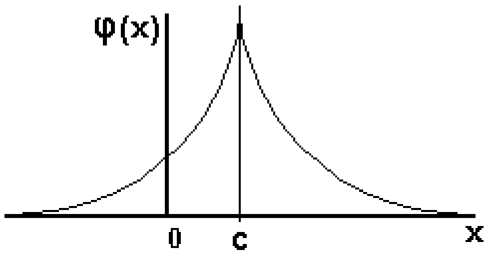


Рис. 18. Плотность распределения Лапласа

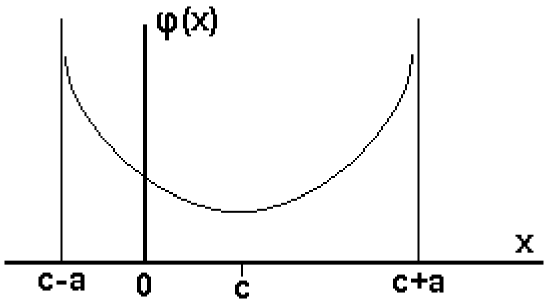


Рис. 15. Плотность распределения Arcsin

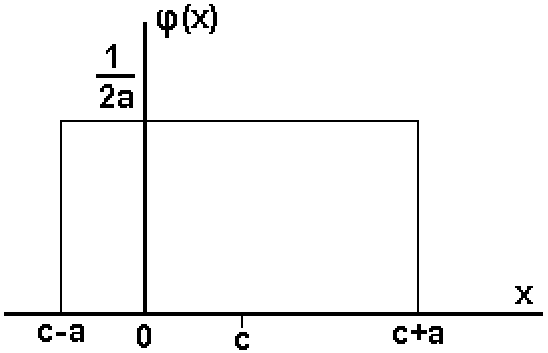


Рис. 14. Равномерная плотность распределения

Приложение 2 Пример оценки неизвестных параметров распределений

Пример 1 Нормальное распределение, метод моментов

Предположим, что мы ищем оценки параметров нормального распределения. У нормального распределения 2 параметра – центр a и разброс (СКО) σ. По методу моментов для нормального распределения находим из таблицы распределений непрерывных СВ, что a = M[x], σ^2=D[x], поэтому оценками параметров a и σ являются оценки математического ожидания и корень из оценки дисперсии соответственно:

$$\hat{a} = \hat{M}[X] = \frac{\sum_{i=1}^{N} x_i}{N} = \overline{x}$$ - среднее арифметическое

$$\hat{\sigma} = \sqrt{\hat{D}[X]} = \sqrt{s^2}$$, где $s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}$ - несмещенная выборочная оценка дисперсии.

Пример 2 Гамма-распределение, метод моментов

Предположим, что мы делаем гипотезу о гамма-распределении. У него 2 параметра – k и θ. Из таблицы находим, что M[x]=kθ, D[x]=kθ^2.

Решая систему из двух уравнений с неизвестными k и θ выражаем эти параметры через моменты: k = M^2[x]/D[x], θ = D[x]/M[x]. Далее определяем оценки параметров, используя в качестве моментов M[x] и D[x] их оценки так же, как и в примере 1. Окончательно получаем

$$\hat{k} = \frac{\hat{M}^2[x]}{\hat{D}[x]} = \frac{\overline{x}^2}{s^2} \quad \hat{\theta} = \frac{\hat{D}[x]}{\hat{M}[x]} = \frac{s^2}{\overline{x}}$$

Зам. Можно было действовать и по-другому. Например, из таблицы видно, что асимметрия и эксцесс равны $As = 2/\sqrt{k}, Ex = 6/k$;

Поэтому можно сразу выразить k как 6/Ex или 4/As^2. Далее воспользовавшись оценкой эксцесса или асимметрии можно сразу найти параметр k, а затем из одного из уравнений (для мат.ожидания или дисперсии) определить второй параметр θ. Но как правило, в методе моментов используются моменты как можно меньшего порядка для нахождения оценок параметров.

Пример 3 Нормальное распределение, метод ММП, аналитический.

Из таблицы сразу находим, что для нормального распределения существует аналитическая формула для оценки параметров по ММП. Поэтому сразу находим.

$$a = \overline{x}; \sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}}$$

Пример 4 Нормальное распределение, метод ММП, численный.

Для нормального распределения в Matlab есть функция для подгонки параметров по ММП – normfit. Ее вызов осуществляется следующим образом:

[muhat,sigmahat] = normfit(data)

где muhat – оценка центра a, sigmahat – оценка СКО σ, data – исходная выборка.

Но можно воспользоваться и более универсальной функцией подгонки по ММП – mle. Ее вызов в общем виде осуществляется следующим образом:

phat = mle(data,'pdf',pdf,'start',start)

phat – вектор оцениваемых параметров, data – выборка, pdf – указатель на функцию с плотностью распределения, start – начальное приближение для параметров

Сгенерируем выборку из 1000 точек с нормальным распределением, a = 5, σ=2:

data = normrnd(5, 2, [1000 1]);

Найдем ММП-оценку численным способом:

phat = mle(data,'pdf',@normpdf,'start',[1 1])

В качестве плотности pdf мы передали указатель на стандартную функцию плотности нормального распределения normpdf, начальное приближение мы задали равным a=1, σ=1.

Найдем аналитические оценки:

a_mle = mean(data)

sigma_mle = sqrt(var(data,1))

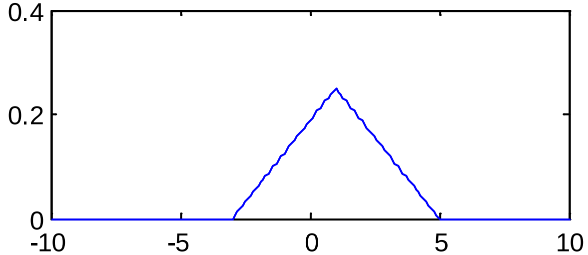
Пример 5 Распределение треугольное, метод ММП, численный.

Для начала необходимо задать плотность распределения в Matlab, поскольку ее там нет. Проще всего воспользоваться строкой-функцией:

pdf_tri = @(x,c,a)((abs(x-c)<=2*a).*((2*a-abs(x-c))/(4*a^2))+1e-10);

К функции добавляется малая константа 1e-10, чтобы плотность не обращалась в 0 – это требование для использования в дальнейшем функции mle. Убедимся, что все правильно, построим график плотности для c=1, a =2

x = -10:.1:10; y = pdf_tri(x,1,2); plot(x,y)



Интеграл от функции равен единице

integral(pdf_tri(,1,2),-10,10)

Значит все правильно и можно использовать эту функцию для нахождения плотности. Сгенерируем данные как в примере 5:

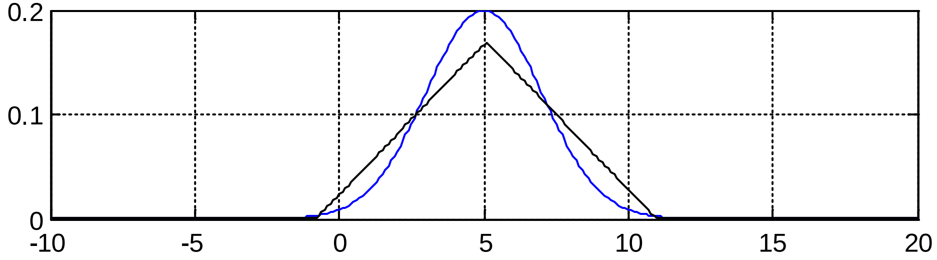
data = normrnd(5, 2, [1000 1]);

Найдем оценки для треугольного распределения по методу ММП:

phat = mle(data,'pdf',@ pdf_tri,'start',[1 1])

Функция дает ответ phat = 5.0811 2.9443. Построим на истинной нормальной плотности плотность полученного треугольного распределения:

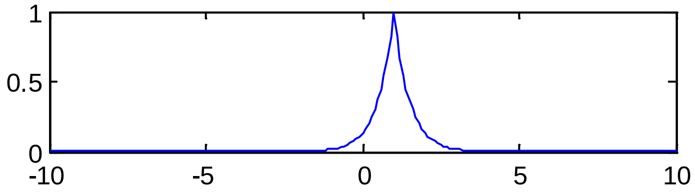
x=-10:.1:20; y1 = normpdf(x,5,2); plot(x,y1); grid on; hold on;
y2 = pdf_tri(x,phat(1),phat(2)); plot(x,y2, 'k');



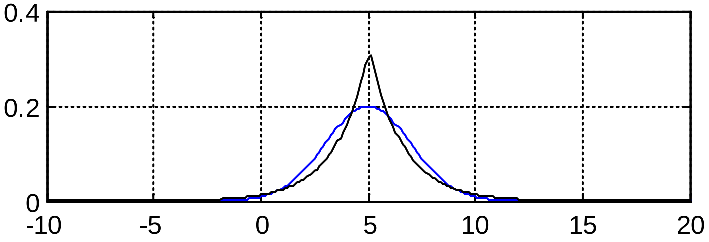
Видим, что действительно параметры треугольного распределения корректно определились под исходную выборку data.

Пример 6 Распределение Лапласа, метод ММП

Проделаем действия по аналогии с примером 5.
Задаем плотность распределения Лапласа, учитывая, что λ>0.
pdf_lapl = @(x,c,l)(l/2*exp(-l*abs(x-c))*(l>0)+1e-10);
Строим график плотности
x = -10:.1:10; y = pdf_lapl(x,1,2); plot(x,y)



Находим интеграл и убеждаемся, что он равен 1.
integral(@(x)pdf_lapl(x,1,2), -10,10)
Генерируем нормальное распределение
data = normrnd(5, 2, [1000 1]);
Находим оценки параметров распределения Лапласа
phat = mle(data,'pdf',pdf_lapl,'start',[1 1])
Строим исходную нормальную плотность и плотность распределения Лапласа с найденными параметрами
x=-10:.1:20; y1 = normpdf(x,5,2); plot(x,y1); grid on; hold on;
y2 = pdf_lapl(x,phat(1),phat(2)); plot(x,y2, 'k');



Из таблицы видно, что для распределения Лапласа параметры можно посчитать и аналитически:

$\hat{c} = x_{med}$;

$\hat{\lambda} = N(\sum_{i=1}^{N} |xi - \hat{c}|)^{-1}$

Проверяем и убеждаемся, что полученные значения совпадают с найденными функцией mle:

c_mle = median(data)

l_mle = length(data)/sum(abs(data-c_mle))

Пример 7 Программа, иллюстрирующая подгонку параметров для 4 распределений – Лапласа, треугольного, Симпсона и нормального

Чтобы изучить данную программу, необходимо создать файл с именем simp_ex.m, скопировать туда текст приведенной ниже программы, сохранить и запустить программу на выполнение. Вначале лучше выполнять программу в режиме отладки (по шагам) и контролировать изменение всех переменных.

```
function simp_ex

c = 7;
a = 3;
N = 10000;
x_range = -10:.1:20;

menu_choice = 1;
while menu_choice ~= 5
    menu_choice = menu('What to do', {'Tri', 'Sim', 'Laplace', 'Normal', 'Exit'});
    close all;
    switch menu_choice
        case 1
            tri_ex(c,a,N,x_range);
        case 2
            simpson_ex(c,a,N,x_range);
        case 3
            lapl_ex(c,a,N,x_range);
        case 4
            norm_ex(c,a,N,x_range);
    end;
end;
close all;

% Пример на треугольное распределение
function tri_ex(c,a,N,x)
```

```
% Задаем плотность треугольного распределения
tripdf = @(x,c,a)((abs(x-c)<=2*a).*((2*a-abs(x-c))/(4*a^2))+1e-10);

% Построим график плотности с параметрами c=1, a=2
y = tripdf(x,c,a);
plot(x,y)

% Убедимся, что интеграл по плотности равен 1
integral(@(x)tripdf(x,c,a),-10,10)
% Проверим, чему равны мат.ожидание и дисперсия
mean_tri = integral(@(x)tripdf(x,c,a).*x,-10,10)  % Мат.ожидание
delta1 = mean_tri - c % Должно быть равно с
var_tri = integral(@(x)tripdf(x,c,a).*(x-mean_tri).^2,-10,10) % Дисперия
delta2 = var_tri - 2/3*a^2 % Должна быть равна 2/3*a^2

% Сгенеририруем нормальное распределение N(5,2)
data = normrnd(c, a, [N 1]);

% Аппроксимируем нормальное распределение треугольным и подгоним параметры
% для треугольного
% Вначале с помощью метода ММП
phat = mle(data,'pdf',tripdf,'start',[1 1]);
c_mle = phat(1)
a_mle = phat(2)
% Затем с помощью метода моментов
c_mm = mean(data)
a_mm = sqrt(1.5*var(data))

% Далее построим на одном графике истинную плотность и две плотности
% треугольного распределения с разными
% значениями оцениваемых параметров, найденные выше
figure;
y1 = normpdf(x,c,a); % Нормальная плотность
plot(x,y1); grid on; hold on;
y2 = tripdf(x,c_mle,a_mle); % Плотность треуг. распр. с параметрами ММП
plot(x,y2, 'k');
y3 = tripdf(x,c_mm,a_mm);  % Плотность треуг. распр. с параметрами ММ
plot(x,y3, 'r');
legend({'Normal', 'Tri - MLE', 'Tri - Moment'});
% Пример на распределение Симпсона
function simpson_ex(c,a,N,x)
% Задаем плотность распределения Симпсона
simpdf = @(x,c,a) ((abs(x-c) > a) .* (abs(x-c) <= 3*a) .* (3*a-abs(x-c)).^2/(16*a^3) + (abs(x-c) <= a).* (3*a^2 - (x-
c).^2)/(8*a^3) + 1e-10);
```

```
% Построим график плотности с параметрами c=1, a=2
y = simpdf(x,c,a);
plot(x,y)

% Убедимся, что интеграл по плотности равен 1
integral(@(x)simpdf(x,c,a),-10,10)
% Проверим, чему равны мат.ожидание и дисперсия
mean_sim = integral(@(x)simpdf(x,c,a).*x,-10,10)  % Мат. ожидание
delta1 = mean_sim - c % Должно быть равна с
var_sim = integral(@(x)simpdf(x,c,a).*(x-mean_sim).^2,-10,10) % Дисперия
delta2 = var_sim - a^2 % Должна быть равна a^2

% Сгенеририруем нормальное распределение N(5,2)
data = normrnd(c, a, [N 1]);

% Аппроксимируем нормальное распределение р-м Симпсона и подгоним параметры
% для распределения Симпсона
% Вначале с помощью метода ММП
phat = mle(data,'pdf',simpdf,'start',[1 1]);
c_mle = phat(1)
a_mle = phat(2)
% Затем с помощью метода моментов
c_mm = mean(data)
a_mm = sqrt(var(data))

% Далее построим на одном графике истинную плотность и две плотности
% треугольного распределения с разными
% значениями оцениваемых параметров, найденные выше
figure;
y1 = normpdf(x,c,a); % Нормальная плотность
plot(x,y1); grid on; hold on;
y2 = simpdf(x,c_mle,a_mle); % Плотность треуг. распр. с параметрами ММП
plot(x,y2, 'k');
y3 = simpdf(x,c_mm,a_mm);  % Плотность треуг. распр. с параметрами ММ
plot(x,y3, 'r');
legend({'Normal', 'Simpson - MLE', 'Simpson - Moment'});
% Пример на распределение Лапласа
function lapl_ex(c,a,N,x)
% Задаем плотность распределения Симпсона
laplacepdf = @(x,c,a)(a/2*exp(-a*abs(x-c))*(a>0)+1e-10);

% Построим график плотности с параметрами c=1, a=2
y = laplacepdf(x,c,a);
plot(x,y)
```

```
% Убедимся, что интеграл по плотности равен 1
integral(@(x)laplacepdf(x,c,a),-10,10)
% Проверим, чему равны мат.ожидание и дисперсия
mean_laplace = integral(@(x)laplacepdf(x,c,a).*x,-10,10)  % Мат. ожидание
delta1 = mean_laplace - c % Должно быть равна с
var_laplace = integral(@(x)laplacepdf(x,c,a).*(x-mean_laplace).^2,-10,10) % Дисперия
delta2 = var_laplace - 2/a^2 % Должна быть равна 2/a^2

% Сгенеририруем нормальное распределение N(5,2)
data = normrnd(c, a, [N 1]);

% Аппроксимируем нормальное распределение р-м Симпсона и подгоним параметры
% для распределения Симпсона
% Вначале с помощью метода ММП
phat = mle(data,'pdf',laplacepdf,'start',[1 1]);
c_mle = phat(1)
a_mle = phat(2)
% Для распределения Лапласа известны аналитические оценки по ММП
c_mle_theory = median(data)
deltac = c_mle - c_mle_theory
a_mle_theory = length(data)/sum(abs(data-c_mle_theory))
delta = a_mle - a_mle_theory
% Затем с помощью метода моментов
c_mm = mean(data)
a_mm = sqrt(2/var(data))

% Далее построим на одном графике истинную плотность и две плотности
% треугольного распределения с разными
% значениями оцениваемых параметров, найденные выше
figure;
y1 = normpdf(x,c,a); % Нормальная плотность
plot(x,y1); grid on; hold on;
y2 = laplacepdf(x,c_mle,a_mle); % Плотность треуг. распр. с параметрами ММП
plot(x,y2, 'k');
y3 = laplacepdf(x,c_mm,a_mm);  % Плотность треуг. распр. с параметрами ММ
plot(x,y3, 'r');
legend({'Normal', 'Laplace - MLE', 'Laplace - Moment'});
% Пример на нормальное распределение
function norm_ex(c,a,N,x)

% Сгенеририруем нормальное распределение N(5,2)
data = normrnd(c, a, [N 1]);

% Аппроксимируем нормальное распределение нормальным и подгоним параметры
% Вначале с помощью метода ММП
phat = mle(data,'pdf',@normpdf,'start',[1 1]);
```

```
c_mle = phat(1)
a_mle = phat(2)
% Затем с помощью метода моментов
c_mm = mean(data)
a_mm = sqrt(var(data))

% Далее построим на одном графике истинную плотность и две плотности
% треугольного распределения с разными
% значениями оцениваемых параметров, найденные выше
figure;
y1 = normpdf(x,c,a); % Нормальная плотность
plot(x,y1); grid on; hold on;
y2 = normpdf(x,c_mle,a_mle); % Плотность треуг. распр. с параметрами ММП
plot(x,y2, 'k');
y3 = normpdf(x,c_mm,a_mm);  % Плотность треуг. распр. с параметрами ММ
plot(x,y3, 'r');
legend({'Normal', 'Normal - MLE', 'Normal - Moment'});
```

Приложение 3 Проверка гипотезы о виде плотности распределения

Критерий "хи - квадрат"

Из генеральной совокупности X , образованной случайной величиной ξ, извлечена выборка $x_1, x_2, ..., x_n$. Выдвигается предположение о том, что плотность распределения случайной величины есть $\varphi(\vec{\Theta}, x)$, где $\vec{\Theta}$ - вектор параметров. По выборочным данным вычисляются оценки параметров $\tilde{\vec{\Theta}}$ и проверяется сложная гипотеза

H_0: плотность распределения случайной величины ξ есть $\varphi(\tilde{\vec{\Theta}}, x)$

против альтернативы

H_1 : плотность распределения случайной величины ξ не $\varphi(\tilde{\vec{\Theta}}, x)$.

Поскольку эта гипотеза сложная, задается только вероятность ошибки первого рода α, которая в подобных случаях именуется уровнем значимости.

$$\sum_{k=1}^{K} \frac{\left(n \cdot P_k - n_k\right)^2}{nP_k} \in \chi^2(K - r).$$

$$P_k = \int_{x_{k-1}}^{x_k} \varphi(\tilde{\vec{\Theta}}, x)dx$$

если распределение нашей СВ действительно такое же как и у той СВ, с который мы его сравниваем. Ограничимся таким критическим значением, вероятность превышения которого будет не более заданного значения α. Поскольку нам известно, что при условии справедливости нулевой гипотезы статистика критерия распределена приблизительно по закону χ^2, мы можем принять в качестве критического значения $(1-\alpha)\cdot 100$ - процентную квантиль $\chi^2_{1-\alpha}(K - r)$.

Алгоритм

1. Задается уровень значимости α
2. По выборочным данным строится гистограмма в соответствии с указаниями п. 2.2. Число столбцов – K.
3. Вычисляются точечные оценки моментов.
4. Из теоретических соображений, по виду гистограммы, по соотношениям между моментами, по значениям асимметрии и эксцесса, по другим соображениям выдвигается гипотеза о виде плотности распределения $\varphi(\vec{\Theta}, x)$.
5. Вычисляются оценки $\tilde{\vec{\Theta}}$ параметров предполагаемой плотности распределения, в результате получается плотность распределения $\varphi(\tilde{\vec{\Theta}}, \mathrm{x})$.
6. С использованием $\varphi(\tilde{\vec{\Theta}}, \mathrm{x})$ вычисляются вероятности
$$P_k = \int_{x_{k-1}}^{x_k} \varphi(\tilde{\vec{\Theta}}, x)dx.$$
7. Вычисляется статистика критерия
$$\chi^2 = \sum_{k=1}^{K} \frac{\left(n \cdot P_k - n_k\right)^2}{nP_k}.$$
8. Полученное значение сравнивается с критическим значением
$$\chi^2_{1-\alpha}(K - r),$$
где r - количество оцениваемых параметров.

9. Если $\chi^2 > \chi^2_{1-\alpha}(K-r)$ делается вывод о том, что экспериментальные данные не подтверждают справедливость выдвинутой гипотезы или о том, что отсутствуют достаточные основания для того, чтобы считать нулевую гипотезу справедливой. Гипотеза пересматривается, выдвигается новая нулевая гипотеза, переход на п. 4 настоящей процедуры.

10. Если $\chi^2 < \chi^2_{1-\alpha}(K-r)$ делается вывод о том, что экспериментальные данные подтверждают справедливость выдвинутой гипотезы или о том, что имеются достаточные основания для того, чтобы считать нулевую гипотезу справедливой.

Чем больше α, тем больше шансов отклонить верную гипотезу. Наоборот, при уменьшении α граница нулевой области растет в пределе до бесконечности и повышается риск принять ложную гипотезу.

Критерий Колмогорова - Смирнова

Из генеральной совокупности X , образованной случайной величиной ξ, извлечена выборка $x_1, x_2, ..., x_n$. Выдвигается предположение о том, что функция распределения случайной величины есть $F(\vec{\Theta}, x)$, где $\vec{\Theta}$ - вектор параметров. По выборочным данным вычисляются оценки параметров $\tilde{\vec{\Theta}}$ и проверяется сложная гипотеза

H_0: функция распределения случайной величины ξ есть $F(\tilde{\vec{\Theta}}, x)$

против альтернативы

H_1: функция распределения случайной величины ξ не $F(\tilde{\vec{\Theta}}, x)$.

Поскольку эта гипотеза сложная, задается только вероятность ошибки первого рода α, которая в подобных случаях именуется уровнем значимости.

В соответствии с формулировкой гипотезы сравниваются две функции распределения: выборочная (п. 2.2) и предполагаемая, представленные на рис. 37. Различие между ними определено, как $D = \sup_i \left| \tilde{F}(x_i) - F(\tilde{\vec{\Theta}}, x_i) \right|$,

где $\tilde{F}(x_i)$ - значения выборочной функции распределения при $x = x_i$.

Статистикой критерия является величина D. Критические значения табулированы. Таблицы критических значений D_α, как функций от вероятности α, приводятся практически во всех учебниках и справочниках по математической статистике. В таблице ниже приводятся некоторые часто употребляемые критические значения.

Таблица

Критические значения критерия Колмогорова-Смирнова

α \ N	25	50	80	100
0.2	0.208	0.148	0.118	0.106
0.1	0.238	0.169	0.135	0.121
0.05	0.264	0.188	0.150	0.134

Если n > 10, для расчета критических значений можно пользоваться приближенной формулой

$$D_\alpha = \sqrt{-\frac{\ln(0.5 \cdot \alpha)}{2 \cdot n}} - \frac{1}{6n}.$$

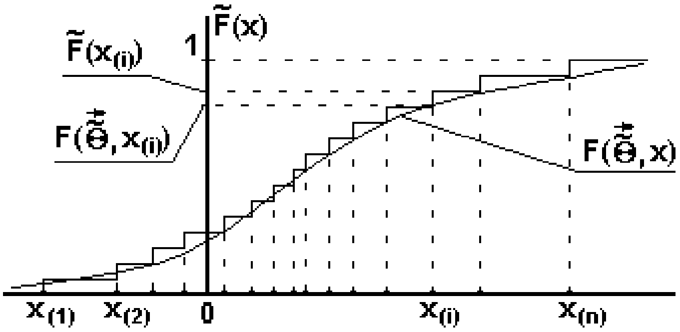


Рис. Выборочная и предполагаемая функции распределения

Процедура проверки гипотезы о виде функции распределения по критерию Колмогорова - Смирнова.

1. Задается уровень значимости α.

2. По выборочным данным строится выборочная функция распределения в соответствии с указаниями п. 2.2

3. Вычисляются точечные оценки моментов.

4. Из теоретических соображений, по виду выборочной функции распределения, по соотношениям между моментами, по значениям асимметрии и эксцесса, по другим соображениям выдвигается гипотеза о виде функции распределения $F(\vec{\Theta}, x)$ и тем самым - о виде плотности распределения $\varphi(\vec{\Theta}, x)$.

5. Вычисляется r параметров предполагаемой функции распределения и ее значения $F(\tilde{\vec{\Theta}}, x_i)$ при $x = x_i$.

6. Вычисляется статистика критерия $D = \sup_i \left| \tilde{F}(x_i) - F(\tilde{\vec{\Theta}}, x_i) \right|$

7. Полученное значение сравнивается с критическим значерием D_α.

8. Если $D > D_\alpha$ делается вывод о том, что экспериментальные данные не подтверждают справедливость выдвинутой гипотезы или о том, что отсутствуют достаточые основания для того, чтобы считать нулевую гипотезу справедливой. Гипотеза пересматривается, выдвигается новая нулевая гипотеза, переход на п. 4 настоящей процедуры.

9. Если $D \le D_\alpha$ делается вывод о том, что экспериментальные данные подтверждают справедливость выдвинутой гипотезы или о том, что имеются достаточые основания для того, чтобы считать нулевую гипотезу справедливой.

Условие корректного применения критерия Колмогорова - Смирнова: выборка $x_1, x_2, ..., x_n$. делится на две части. По одной из них определяются параметры $\vec{\Theta}$, по другой - строится выборочная функция распределения и вычисляется статистика критерия. Это позволяет избавиться от необходимости учета зависимости между выборочными значениями, которая появляется в результате вычисления параметров предполагаемой плотности распределения, как это было в случае применения критерия χ^2.

Критерий ω^2 Мизеса

Из генеральной совокупности Х , образованной случайной величиной ξ, извлечена выборка $x_1, x_2, \ldots, x_n$. Выдвигается предположение о том, что функция распределения случайной величины есть $F(\vec{\Theta}, x)$, где $\vec{\Theta}$ - вектор параметров. По выборочным данным вычисляются оценки параметров $\tilde{\vec{\Theta}}$ и проверяется сложная гипотеза

H_0: функция распределения случайной величины ξ есть $F(\tilde{\vec{\Theta}}, x)$

против альтернативы

H_1: функция распределения случайной величины ξ не $F(\tilde{\vec{\Theta}}, x)$.

В качестве статистики критерия используется

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(\tilde{\vec{\Theta}}, x_i\right) - \frac{2i-1}{2n}\right]^2$$

Критические значения $\left(n\omega^2\right)_\alpha$ табулированные в таблицах математической статистики. В таблице ниже приводятся некоторые часто употребляемые критические значения.

Таблица

Критические значения критерия Мизеса

α	0.03	0.05	0.1	0.2
$(nw^2)_\alpha$	0.55	0.4614	0.3473	0.2415

Процедура проверки гипотезы о виде функции распределения по критерию ω^2 Мизеса.

1. Задается уровень значимости α
2. По выборочным данным строится выборочная функция распределения в соответствии с указаниями п. 2.2
3. Вычисляются точечные оценки моментов.
4. Из теоретических соображений, по виду выборочной функции распределения, по соотношениям между моментами, по значениям асимметрии и эксцесса, по другим соображениям выдвигается гипотеза о виде функции распределения $F(\vec{\Theta}, x)$ и тем самым - о виде плотности распределения $\varphi(\vec{\Theta}, x)$.
5. Вычисляется r параметров предполагаемой функции распределения и ее значения $F(\tilde{\vec{\Theta}}, x_i)$ при $\mathbf{x} = \mathbf{x}_i$,.
6. Вычисляется статистика критерия

$$n\omega^2 = \frac{1}{12n} + \sum_{i=1}^{n}\left[F\left(\tilde{\vec{\Theta}}, x_i\right) - \frac{2i-1}{2n}\right]^2$$

7. Полученное значение сравнивается с критическим значением $\left(n\omega^2\right)_\alpha$.
8. Если $n\omega^2 > \left(n\omega^2\right)_\alpha$ делается вывод о том, что экспериментальные данные не подтверждают справедливость выдвинутой гипотезы или о том, что отсутствуют достаточые основания для того, чтобы считать нулевую гипотезу справедливой. Гипотеза пересматривается, выдвигается новая нулевая гипотеза, переход на п. 4 настоящей процедуры.
9. Если $n\omega^2 \le \left(n\omega^2\right)_\alpha$ делается вывод о том, что экспериментальные данные подтверждают справедливость выдвинутой гипотезы или о том, что имеются достаточые основания для того, чтобы считать нулевую гипотезу справедливой.

Критерий ω^2 Мизеса - равномерно наиболее мощный критерий проверки гипотезы о виде функции распределения.