Расчетное задание 2а

Статистическая обработка случайных последовательностей. Идентификация законов распределения.

Исходные данные

В результате измерений получена выборка x1, x2, …, xN из генеральной совокупности с неизвестным законом распределения.

Варианты

Данная работа выполняется теми студентами, у которых в личной папке файлового хранилища есть архив Task_3a.zip с текстовым файлом "Task_3a.txt". В этом файле задано число значений N, а также сам массив выборочных значений, отделенных 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 основы статистической обработки данных и построения оценок функции и плотности распределения. Подробно описана методика применения ядерного оценивания плотности.

В приложении 3 приведены основы теории точечного и интервального оценивания. Показаны примеры вычисления оценок параметров распределений при подгонке параметров распределений к имеющейся выборке. При использовании метода моментов и максимального правдоподобия целесообразно ознакомиться и разобраться в этих примерах.

В приложении 4 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. Представить визуально оценки функции плотности распределения.

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

1.2.2 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).

1.2.3. Построить оценки плотности с применение ядерного оценивания (kernel density estimation). Рассмотреть 3-4 разных вариантов ядра и для каждого из них выбрать оптимальное значение ширины окна h. В качестве начальной оценки использовать одну из параметрических оценок h. Подробности по ядерному оцениванию плотности приведены в приложении 3.

1.3. Определить точечные оценки:

1.3.1. моментов

- первого начального (среднее арифметическое - mean, медиана - median, средина размаха $(x_{min} + x_{max})/2$)
- 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 оценки плотности - гистограммы).

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

1.3.3. границ интерквантильного промежутка J_p 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.

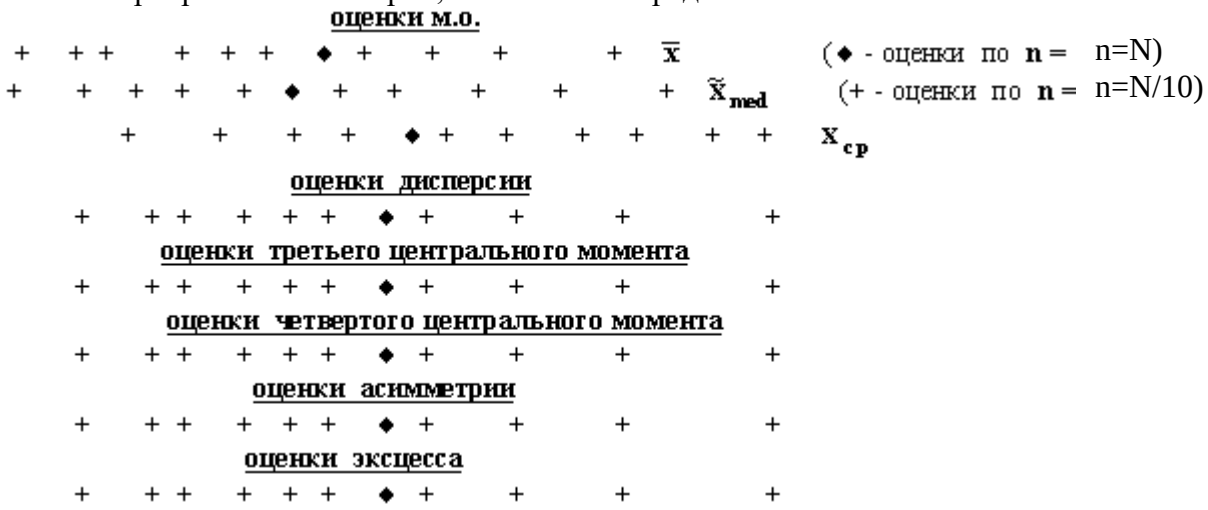
Результаты представить в таблице следующей формы.

	$\overline{x}$	x_{med}	$x_{\text{ср}}$	s^2	s	m_3	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.



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.

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 или лучше ядерную оценку плотности (на 1 графике)

Таким образом 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).

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	$P_n(k) = C_n^k p^k (1-p)^{n-k}$ $P_n(k) \approx \approx \frac{1}{\sqrt{npq}} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$	$F_n(y) = \sum_{k=0}^{y} C_n^k p^k q^{n-k}$	$M(X) = np; D(X) = npq$ $As = \frac{1-2p}{\sqrt{npq}}; ex = \frac{1-6pq}{npq}$ $\mathrm{Mod} = (n+1)p$;	$M_X(\nu) = (pe^{\nu} + q)^n$	$p = \frac{\sum_{i=1}^{n} x_i}{mn}$
Пуассона	$P(k) = \lambda^k e^{-\lambda}/k!$	$F(k) = \frac{\Gamma(k+1,\lambda)}{\lambda!}$	$M(X) = \lambda; D(X) = \lambda; As = \lambda^{-0.5};$ $Ex = \lambda^{-1}; \mathrm{Mod} = \lambda$	$M_X(\nu) = \exp(\lambda(e^{\nu} - 1)$	$\lambda = \overline{x_в}$
Геометрическое	$P(k) = q^k p$		$M(X) = q/p; D(X) = q/p^2$	$M_X(\nu) = p/(1 - qe^{\nu})$	
Равномерное	$P(k) = \begin{cases} \frac{1}{n}, a \le k \le b \\ 0, else \end{cases}$	$F(k) = \begin{cases} 0, k < a \\ (k-a+1)/n, a \le k \le b \\ 1, k > b \end{cases}$	$M(X) = (a+b)/2;$ $D(X) = (n^2-1)/12$ $As = 0$	$M_X(\nu) = \frac{e^{a\nu} - e^{(b+1)\nu}}{n(1-e^{\nu})}$	

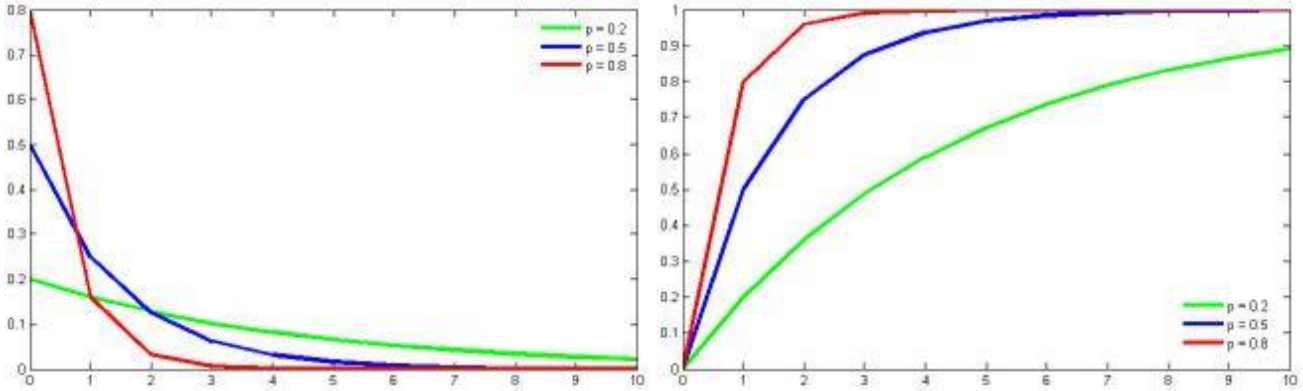
Распределения непрерывных СВ

Распределение	Плотность вероятности	Функция распределения	Числовые характеристики	Производящая функция моментов (хар.функция)	Оценки по ММП
Нормальное	$p(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-a)^2}{2\sigma^2}\right)$	$F(x)=0.5+\Phi_x((x-a)/\sigma)$ $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x \exp(-t^2/2)dt$	$M(x)=a;\ D(x)=\sigma^2;$ $As=0,\ Ex=0$	$M_x(\nu)=$ $=\exp(a\nu+\frac{\sigma^2\nu^2}{2})\phi_x(\nu)=$ $=\exp(ai\nu-\frac{\sigma^2\nu^2}{2})$	$a=\overline{x};$ $\sigma^2=\frac{\sum_{i=1}^{N}(x_i-\overline{x_в})}{N}$
Логнормальное	$p(x)=$ $=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{(\ln(x)-a)^2}{2\sigma^2}\right)$	$F(x)=0.5+\Phi_x((\ln(x)-a)/\sigma)$ $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x \exp(-t^2/2)dt$	$M(X)=e^{a+\sigma^2/2};\ Med=e^a;$ $D(X)=(e^{\sigma^2}-1)e^{2a+\sigma^2};$		
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	$\phi_X(\nu)=\exp(ci\nu-\Delta\lvert\nu\rvert)$	
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	$p(x)=\frac{\lambda}{2}\exp(-\lambda\lvert x-c\rvert)$		$M(X)=x_{0.5}=x_{\mathrm{mod}}=c;$ $D(X)=2/\lambda^2;$ $As=0,\ Ex=6$	$\phi_\xi(\nu)=\frac{\lambda^2}{\lambda^2+\nu^2}e^{j\nu c}$	$c=x_{med};$ $\lambda=N(\sum_{i=1}^{N}\lvert xi-c\rvert)$
Показательное (экспоненциальное)	$p(x)=\begin{cases}\lambda e^{-\lambda x}, x>=0\\ 0, x<0\end{cases}$	$F(x)=\begin{cases}1-e^{-\lambda x}, x>=0\\ 0, x<0\end{cases}$	$\mu_1(x)=1/\lambda;\ D(\xi)=1/\lambda^2;$ $As=2,\ Ex=6, Mod=0,$ $\mathrm{Med}=\ln(2)/\lambda$	$M_x(\nu)=\lambda/(\lambda-\nu)\phi_x(\nu)=$ $=\lambda/(\lambda-i\nu)$	$\lambda=1/\overline{x}$
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	$p(x)=\frac{(1/2)^{n/2}}{\Gamma(n/2)}x^{n/2-1}e^{-x/2}$	$F(x)=\frac{\gamma(n/2,x/2}{\Gamma(n/2)})$	$M(X)=n;\ D(X)=2n;$ $Med\approx n-2/3$ $As=\sqrt{8/n},\ Ex=12/n;$	$\phi_X(\nu)=(1-2i\nu)^{-n/2}$	

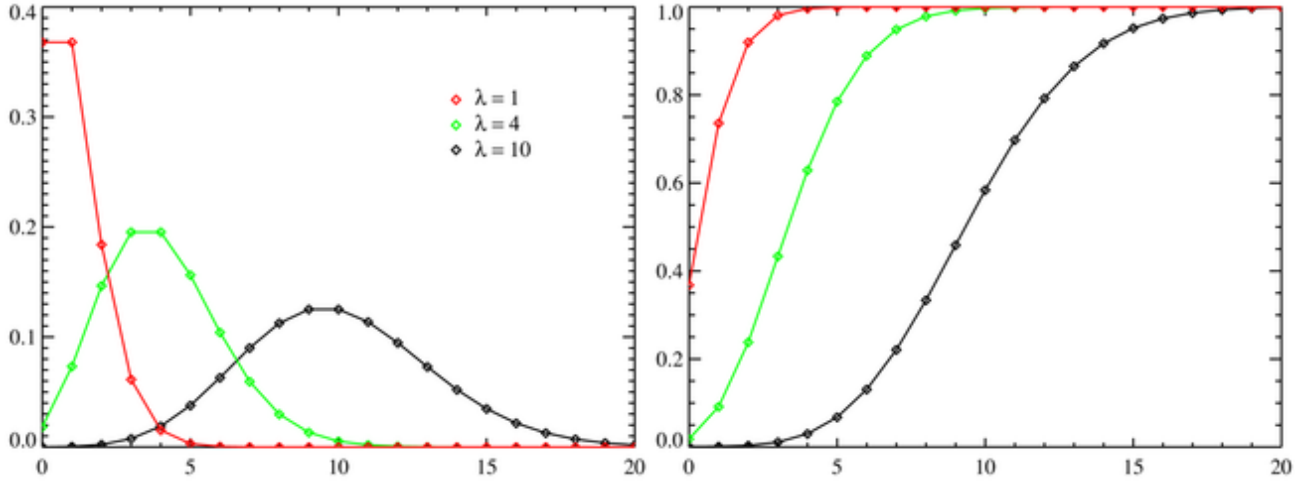
Стьюдента	$p(x)=\frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(n/2)}(1+\frac{y^2}{n})^{-\frac{n+1}{2}}$		$M(X)=Med=\text{mod}=0;$ $D(X)=n/n-2; n>2$ $As=0, n>3,$ $Ex=(3n-6)/(n-4), n>4$		
Равномерное	$p(x)=\begin{cases}\frac{1}{b-a}, a\le x\le b\\ 0, else\end{cases}$	$F(k)=$ $=\begin{cases}0, x<a\\ (x-a)/b-a,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

Графики плотностей распределений дискретных СВ

Геометрическое

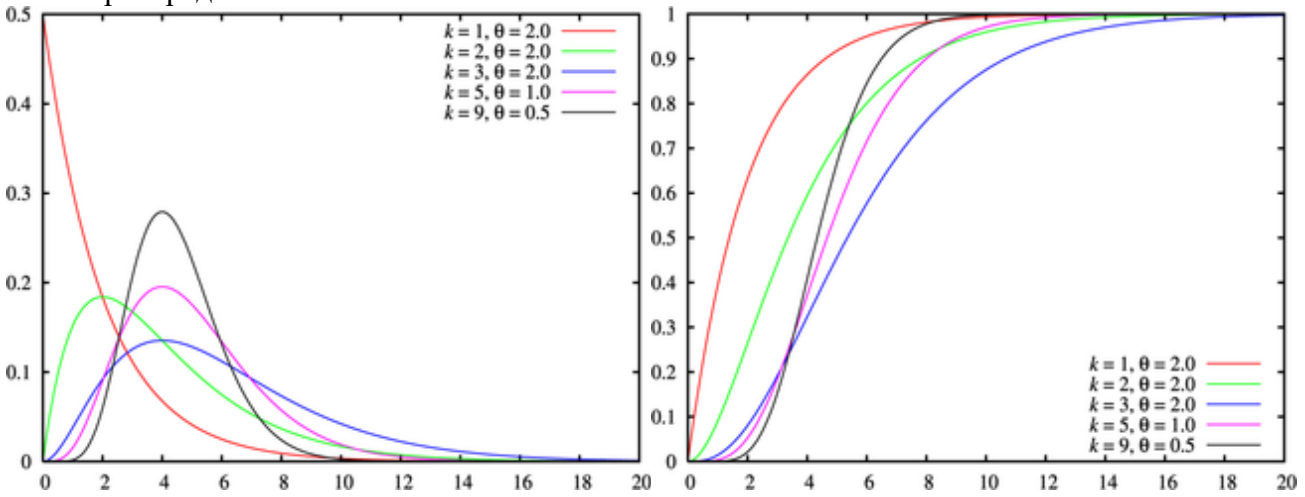


Пуассона

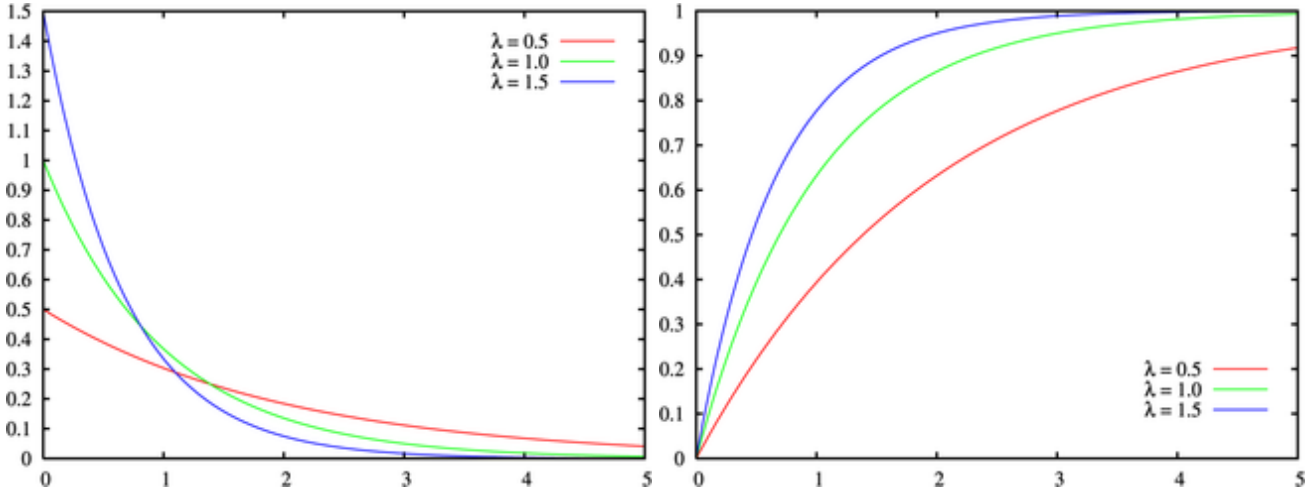


Графики плотностей распределений непрерывных СВ

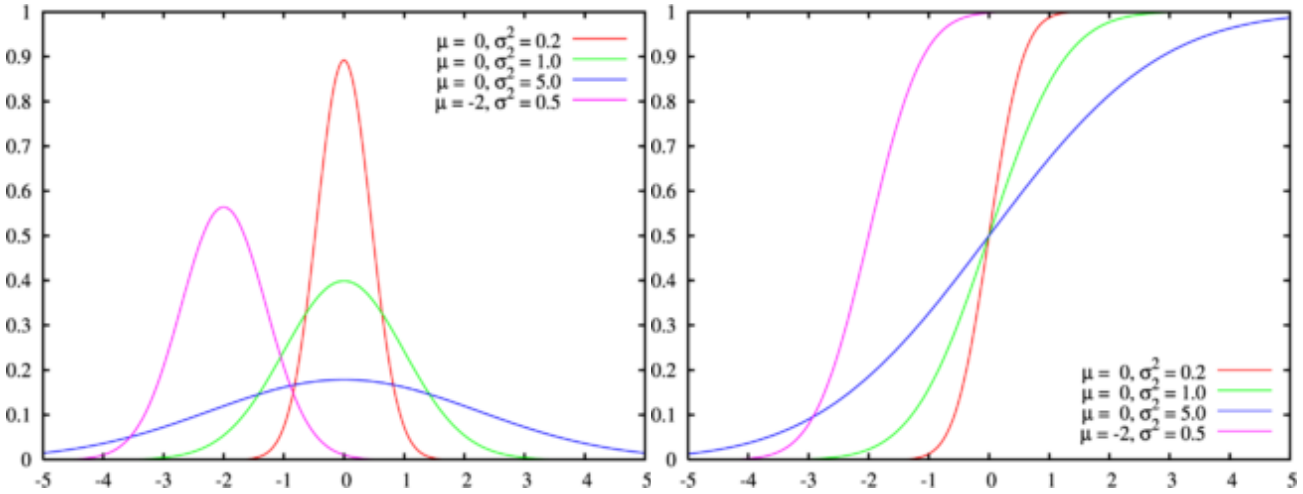
Гамма-распределение



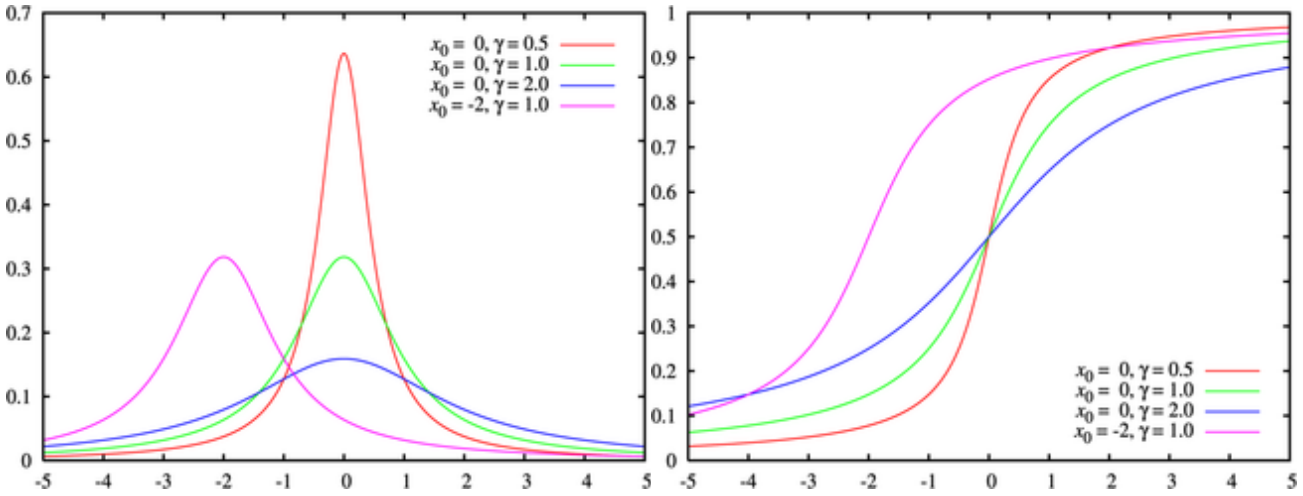
Экспоненциальное



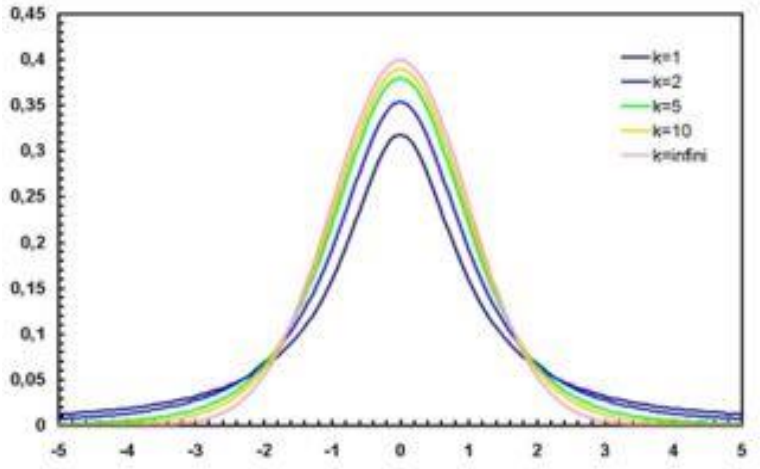
Нормальное

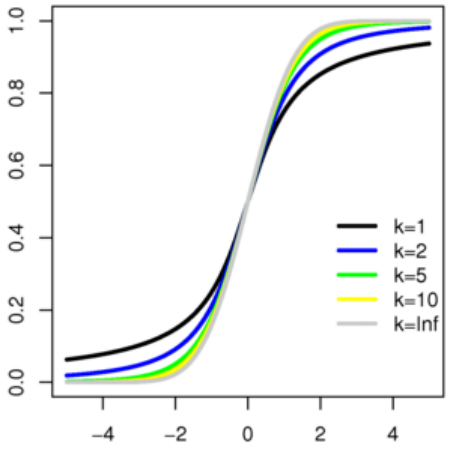


Коши

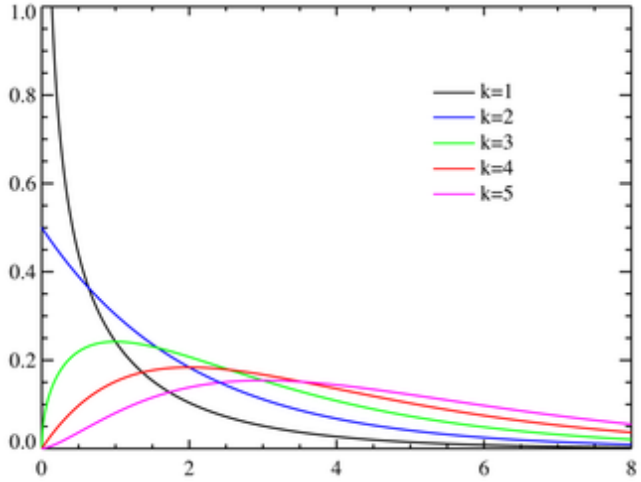


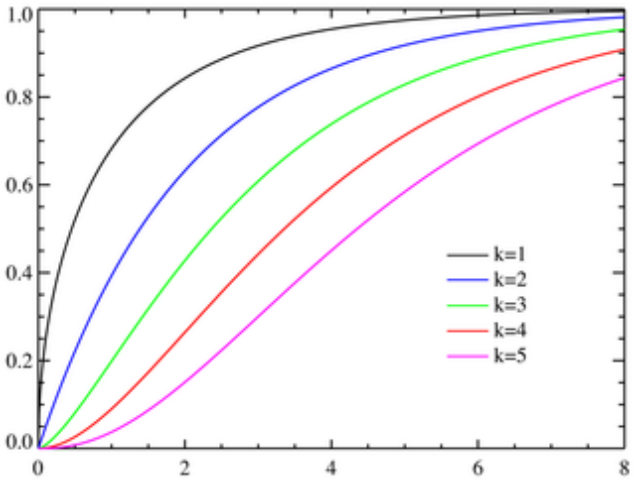
Стьюдента


0,45
0,4
0,35
0,3
0,25
0,2
0,15
0,1
0,05
0
-5
-4
-3
-2
-1
0
1
2
3
4
5
k=1
k=2
k=5
k=10
k=infini

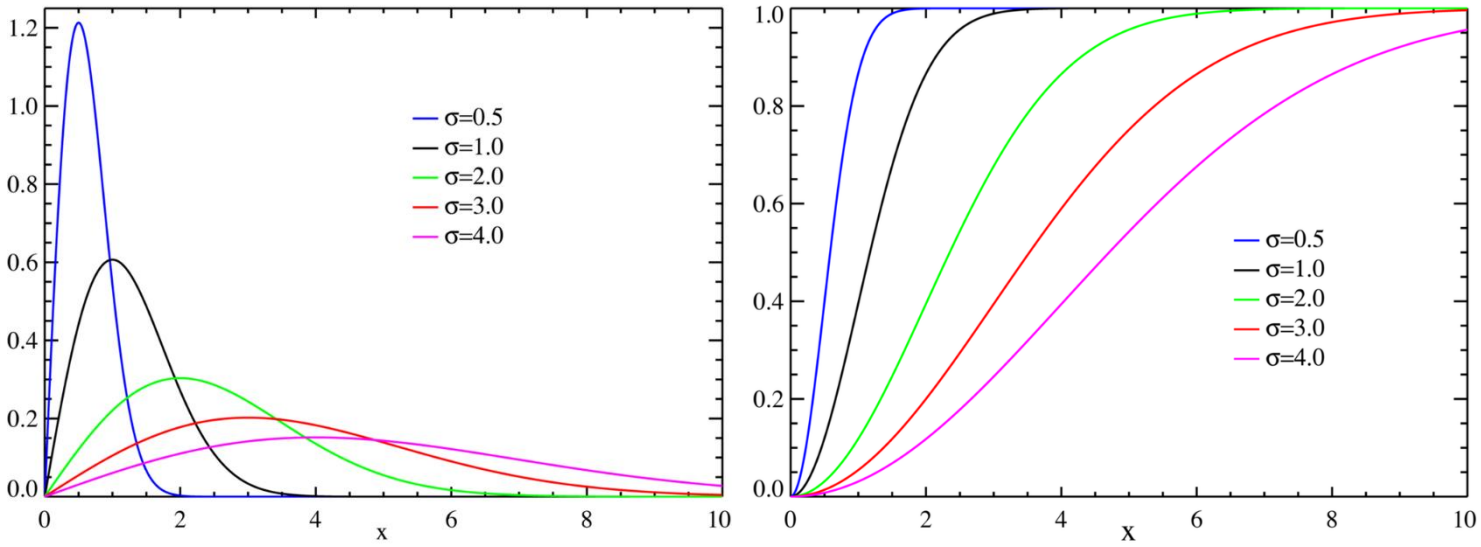

1.0
0.8
0.6
0.4
0.2
0.0
−4
−2
0
2
4
k=1
k=2
k=5
k=10
k=Inf

Хи-квадрат

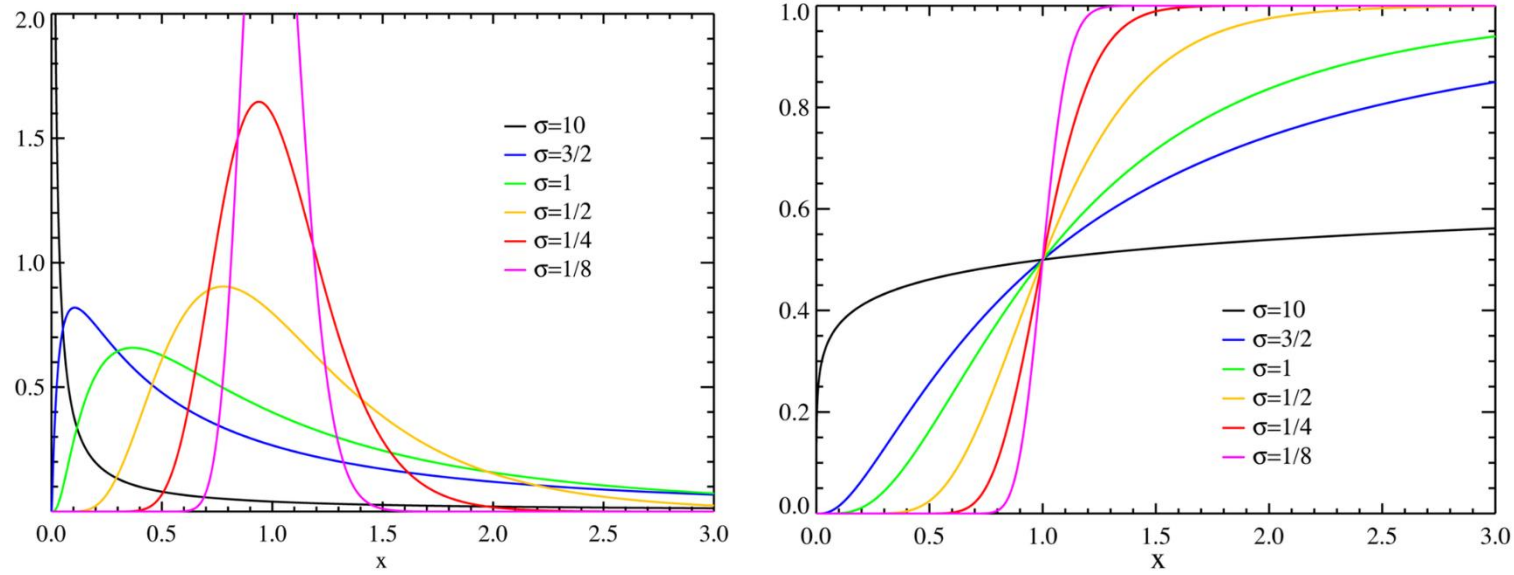

1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
k=1
k=2
k=3
k=4
k=5


1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
k=1
k=2
k=3
k=4
k=5

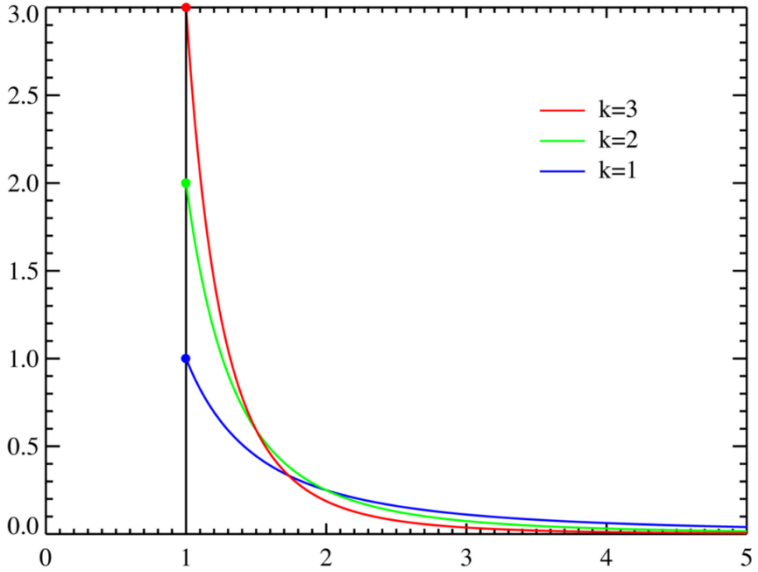
Рэлея

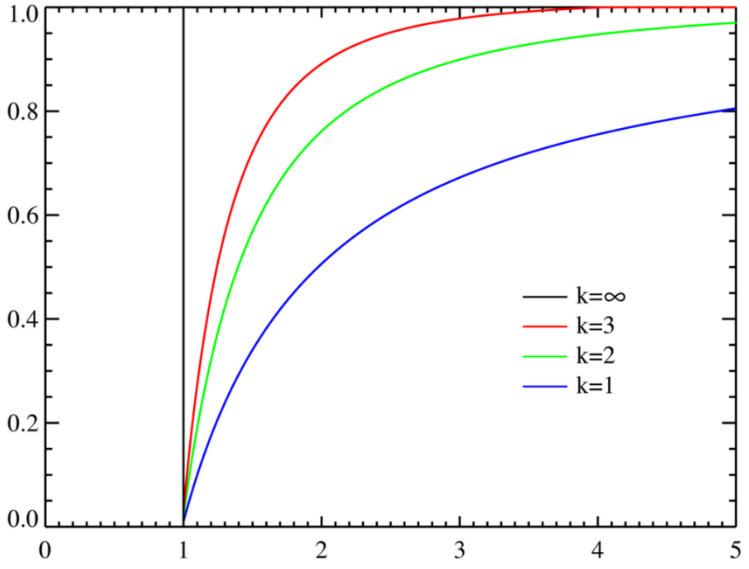


Логнормальное



Распределение Парето





Другие распределения

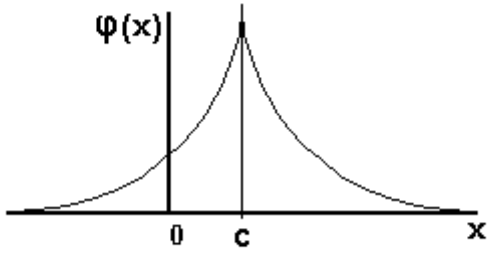


Рис. 18. Плотность распределения Лапласа

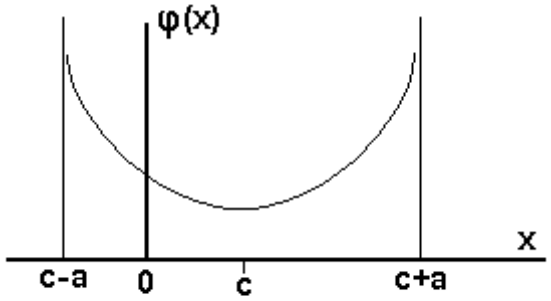


Рис. 15. Плотность распределения Arcsin

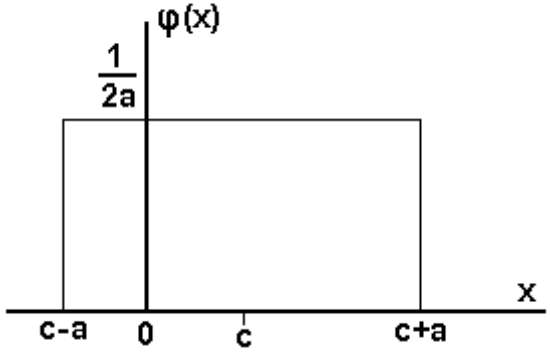


Рис. 14. Равномерная плотность распределения

Приложение 2 Обработка результатов экспериментов и основы теории оценивания

После получения выборочных значений случайной величины в хронологическом порядке первым шагом в их обработке является сортировка в порядке возрастания.

Отсортированная выборка называется вариационным рядом, отдельные его элементы - членами вариационного ряда. Первый и последний члены называются крайними членами вариационного ряда. Средний член вариационного ряда называется выборочной медианой.

Выборочная функция распределения изображается ступенчатой линией. Абсциссами каждого скачка этой линии являются выборочные значения $x_{(i)}$. Высота всех ступеней одинакова и равна $1/n$.

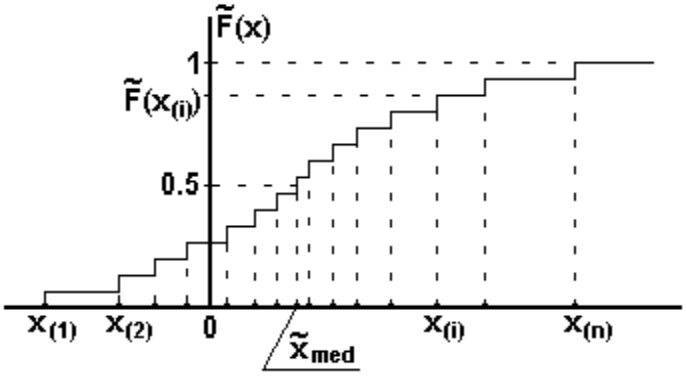


Рис. Выборочная функция распределения

Оценкой плотности распределения является относительная гистограмма. Для ее построения интервал между крайними членами вариационного ряда делится на m интервалов $(x_{m-1}, xm]$ равной длины $\Delta = x_m - x_{m-1}$. Подсчитывается количество n_m выборочных значений, попавших в каждый m-ый интервал и вычисляется отношение n_m/n, которое является оценкой вероятностной меры каждого интервала. Далее на полученных интервалах, как на основаниях строятся прямоугольники, высота каждого из которых должна быть равна

$$h_m = \frac{n_m}{n \cdot \Delta}.$$

При таком построении площадь гистограммы будет равна единице, точно так же, как и под плотностью распределения, оценкой которой и является гистограмма.

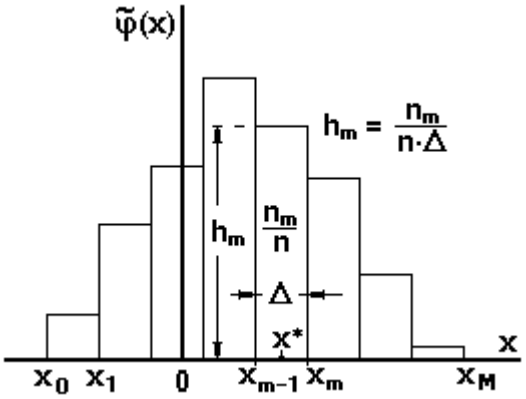


Рис. Гистограмма

Замечание. Иногда помимо рассмотренной выше *относительной* гистограммы строится еще так называемая *абсолютная* гистограмма, у которой по оси Y откладывается величина n_m – количество точек, попавших в m-й интервал. Площадь под абсолютной гистограммой равна $n \cdot \Delta$.

Количество интервалов или их ширина выбирается таким образом, чтобы самый “бедный” интервал содержал $3-5$ выборочных значений. Для удобства построения гистограммы и последующих вычислений рекомендуется округлить значение ширины интервалов до ближайшего удобного числа.

Для распределений с тяжелыми хвостами (Коши) целесообразно строить гистограмму не на всем интервале, а в некотором ограниченном диапазоне – в противном случае вид гистограммы будет ненаглядный (один столбец в центре).

У гистограммы есть ряд недостатков:

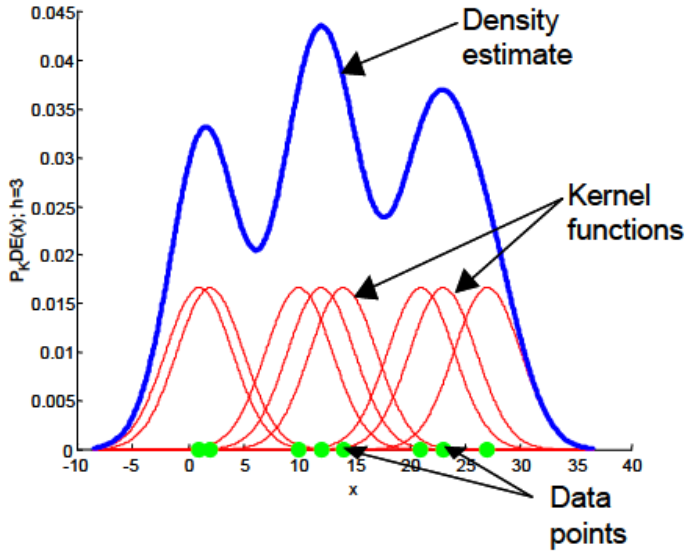
- вид гистограммы зависит от выбора начальной точки;
- не гладкая, а ступенчатая форма;
- гистограмма применима только для одномерных (максимум двумерных) СВ.

Существуют более совершенные методы оценки плотностей – ядерное оценивание плотности (kernel density estimation), модели смесей (гауссовы смеси).

Ядерное оценивание плотности

Плотность распределения можно оценить с помощью **относительной гистограммы**, либо с использованием методов **ядерного оценивания плотности** (kernel density estimation, KDE). Идея метода KDE заключается в построении вокруг каждой точки выборки некоторой затухающей функции. Затем эти функции суммируются, нормализуются и результирующая функция используется для аппроксимации плотности:

$$p_{KDE}(x) = \frac{1}{N h^D} \sum_{n=1}^{N} K\left(\frac{x - x^{(n)}}{h}\right)$$

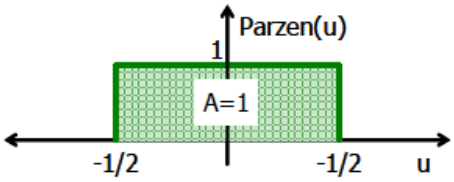


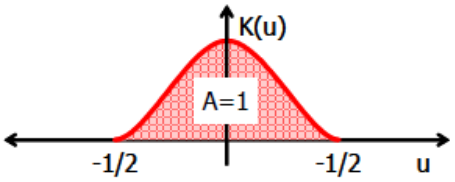
Функция $K(.)$ называется ядром и от ее выбора зависит то, каким образом будет аппроксимироваться плотность вокруг каждой точки. Для $K(.)$ должно обязательно выполняться условие:

$$\int_{R^D} K(x)dx = 1$$

D – размерность случайной величины, N – объем выборки, $x^{(n)}$ – точки выборки.

Существует большое количество типов ядер (см. таблицу ниже). Наиболее распространенными являются прямоугольные и гауссовы ядра.





Тип ядра	Уравнение (1-мерное)	Уравнение (многомерное) при независимых признаках
прямоугольное (равномерное)	$K(u) = \begin{cases} 1, & \text{при } \lvert u\rvert < 1/2 \\ 0, & \text{при} \lvert u\rvert > 1/2 \end{cases}$	$K(u) = \begin{cases} 1, & \text{при} \lvert u_j\rvert < 1/2 \quad \forall j \\ 0, & \text{иначе} \end{cases}$
гауссово	$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$	$K(u) = \frac{1}{(2\pi)^{D/2}} \exp\left(-\frac{u^T u}{2}\right)$
Экспоненциальное (Лапласа)	$K(u) = \frac{1}{2}\exp(-\lvert u\rvert)$	$K(u) = \frac{1}{2^D}\exp(-\sum_{j=1}^{D} \lvert u_j\rvert)$
Коши	$K(u) = \frac{1}{\pi}\frac{1}{1+u^2}$	$K(u) = \frac{1}{\pi^D \prod_{j=1}^{n}(1+u_j^2)}$
треугольное	$K(u) = \begin{cases} 1-\lvert u\rvert, & \text{при } \lvert u\rvert < 1 \\ 0, & \text{при} \lvert u\rvert > 1 \end{cases}$	$K(u) = \begin{cases} \prod_{i=1}^{n}(1-\lvert u_j\rvert), & \text{при } \lvert u_j\rvert < 1, \forall j \\ 0, & \text{иначе} \end{cases}$
восстанавливающий фильтр	$K(u) = \frac{1}{2\pi}\left(\frac{\sin(u/2)}{u/2}\right)^2$	$K(u) = \frac{1}{(2\pi)^D}\prod_{j=1}^{D}\left(\frac{\sin(u_j/2)}{u_j/2}\right)^2$

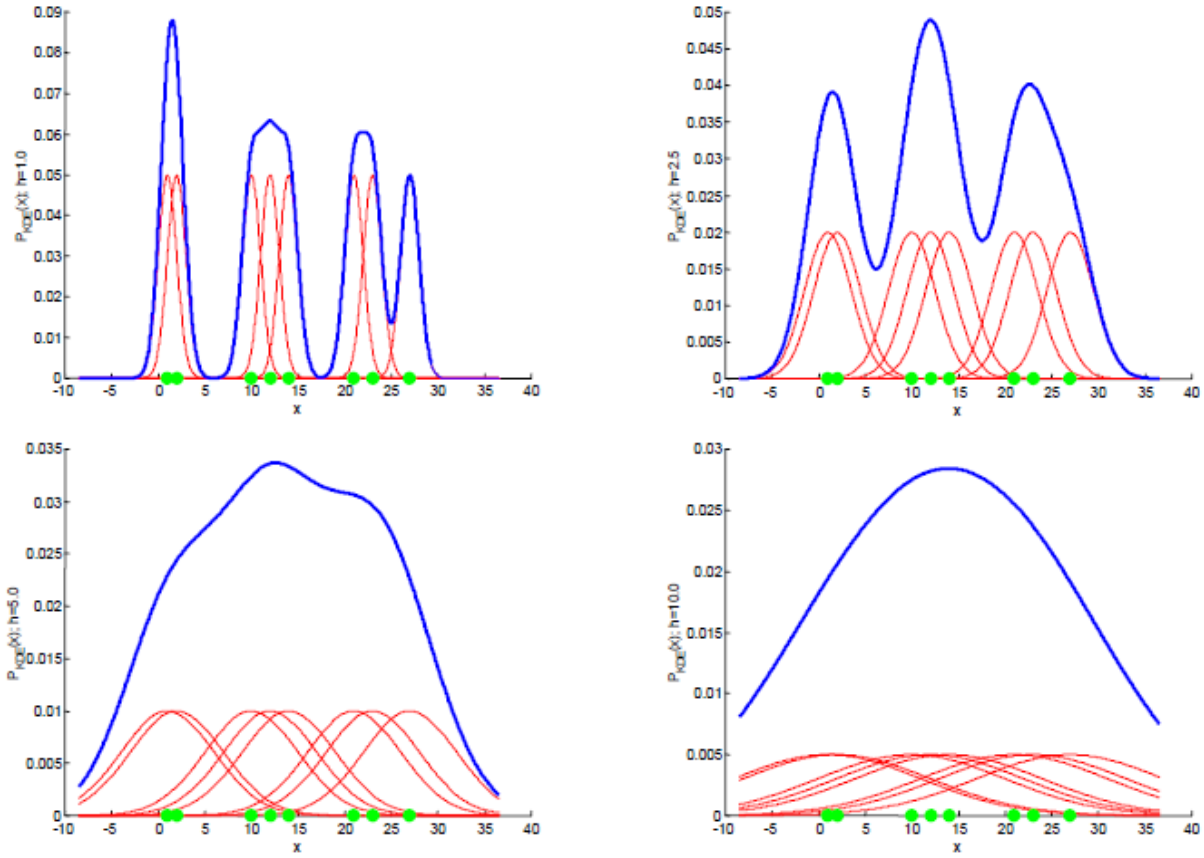
Выбор ширины окна

Параметр h является ключевым при ядерном оценивании. Он называется масштабирующим или сглаживающим параметром, шириной окна. Проблема выбора ширины окна (полосы) – ключевая в задаче оценки плотности. Маленькая ширина приведет к тому, что оценка плотности будет иметь много пиков и будет трудно интерпретируемой. Наоборот, слишком большая ширина приведет к тому, что оценка плотности будет чересчур гладкой и при этом потеряется информация о структуре данных.

Субъективный выбор

Субъективный выбор ширины окна заключается в построении множества кривых и выбора оценки, которая наиболее соответствует некоторым априорным (субъективным) идеям, предпосылкам. Например, можно оценить средний разброс σ_i по каждому признаку x_i и задать значение h_i для каждого признака пропорциональным этому разбросу.

Однако данный метод не всегда подходит для практического использования, так как чаще всего данные являются многомерными.



Параметрический выбор

Если предположить, что истинное распределение известно, то оптимальное значение h может быть найдено аналитически. Если сделать гипотезу о нормальном распределении и использовать гауссово ядро, то можно показать, что оптимальное значение ширины h равно:

$h_{opt1}=1.06\sigma N^{-1/5}$,

где σ – СКО распределения, а N – объем выборки. Из формулы следует, что с увеличением числа примеров в 10^5 раз значение h уменьшается в 10 раз.

Для лучшей аппроксимации мультимодальных плотностей можно использовать следующую формулу:

$h_{opt2}=0.9AN^{-1/5}$, где $A = \min(\sigma, IQR/1.34)$,

где IQR – ширина 50-% интерквантильного промежутка, рассчитанного по 75% (Q_3) и 25% (Q_1) квантилям: $IQR = Q_3-Q_1$, IQR выполняет функцию грубой оценки разброса.

Приложение 3 Основы теории оценивания

Точечное оценивание

Точечной статистической оценкой называется оценка числовой характеристики или параметра генеральной совокупности, имеющая такую же размерность, как и оцениваемая характеристика (1 для дисперсии, мат. ожидания, *mxm* для ковариационной матрицы, где *m* – число случайных величин и др.).

Оценивание квантилей и интерквантильного промежутка

Квантиль x_p характеризует значение случайной величины такое, что $F(x_p) = p$. Квантили оцениваются по выборочной функции распределения. Приближенно квантиль x_p можно оценить, выбрав элемент вариационного ряда с номером $k = np$.

Необходимый объем выборки для вычисления квантили определяется по формулам

$$n \geq \frac{1}{p}, p < 0.5,$$

$$n \geq \frac{1}{1-p}, p > 0.5.$$

Для медианы находится значение x_{med}, при котором $F(x_{med}) = 0.5$. Это значение соответствует среднему члену вариационного ряда. Если объем выборки *n* нечетный, то

$$x_{med} = x_{n/2}.$$

Если объем выборки *n* четный, то

$$x_{med} = 0.5\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right).$$

Интерквантильный промежуток J_p, накрывающий распределение с вероятностью *p* можно определить различными способами. Наиболее распространенный вариант – с помощью симметричных относительно границ вариационного ряда квантилей $x_{\frac{1-P}{2}}$ и $x_{\frac{1+P}{2}}$, но можно и с помощью квантилей $x_0, x_p,$или $x_1 - p, x_1$.

Т.н. интеквартильный промежуток оценивается с помощью квантилей $x_{0.25}, x_{0.75}$.

Точечное оценивание моментов

Несмещенная оценка математического ожидания – среднее арифметическое

$$m_1 = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Аналогичным образом вычисляются оценки всех начальных моментов:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

Если математическое ожидание неизвестно, то смещенная оценка дисперсии вычисляется по формуле

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

а несмещенная – по формуле

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Оценивание параметров законов распределения

Метод моментов

Данный метод базируется на том, что для известных распределений существуют соотношения между начальными (центральными) моментами и параметрами распределений:

$\upsilon_k = f_{1k}(\Theta)$,

$\mu_k = f_{2k}(\Theta)$.

Суть метода моментов – приравнять эмпирические моменты (оценки теоретических) к самим теоретическим и по ним оценить неизвестные параметры распределения.

Примеры

1. Экспоненциальное распределение:

$\nu_1 = M[X] = 1/\lambda$, $\mu_2 = D[X] = 1/\lambda^2$

По методу моментов получаем следующие оценки параметров:

$\lambda = 1/\bar{x}$ либо $\lambda = 1/s$.

2. Равномерное распределение

$\nu_1 = M[X] = (a+b)/2, \quad \mu_2 = D[X] = \sigma^2 = (b-a)^2/3.$

Решая систему получаем оценки параметров a и b:

$$\begin{cases} (a+b)/2 = \bar{x} \\ (b-a)^2/3 = s^2 \end{cases} \Rightarrow \begin{cases} a+b = 2\bar{x} \\ b-a = s\sqrt{3} \end{cases} \Rightarrow \begin{cases} a = \bar{x} - s\sqrt{3}/2 \\ b = \bar{x} + s\sqrt{3}/2 \end{cases}$$

3. Гамма-распределение

$\nu_1 = M[X] = k\theta, \quad \mu_2 = D[X] = k\theta^2.$

По методу моментов получаем следующие оценки параметров:

$$\begin{cases} k\theta = \bar{x} \\ k\theta^2 = s^2 \end{cases} \Rightarrow \begin{cases} k = \bar{x}^2/s^2 \\ \theta = s^2/\bar{x} \end{cases}$$

4. Нормальное распределение

$\nu_1 = M[X] = c, \quad \mu_2 = D[X] = \sigma^2.$

По методу моментов сразу получаем оценки параметров:

$$\begin{cases} c = \bar{x} \\ \sigma^2 = s^2 \end{cases}.$$

Метод максимального правдоподобия

Метод максимального правдоподобия точечной оценки неизвестных параметров заданного распределения сводится к отысканию максимума функции одного или нескольких оцениваемых параметров.

Дискретные случайные величины.

Пусть X – дискретная случайная величина, которая в результате n опытов приняла возможные значения x_1, x_2, ..., x_n. Допустим, что вид закона распределения величины X задан, но неизвестен параметр θ, которым определяется этот закон; требуется найти его точечную оценку $\theta = \theta(x_1, x_2, \ldots, x_n)$

Обозначим вероятность того, что в результате испытания величина X примет значение x_1 при параметре, равном θ, через $p(x_1/\theta)$. Данную вероятность еще иногда называют правдоподобием.

Функцией правдоподобия дискретной случайной величины X называют функцию, являющуюся произведением вероятностей $p(x_i,\theta)$ по всем выборочным значениям:

$$L(x_1, x_2, \ldots, x_n, \theta) = p(x_1, \theta)p(x_2, \theta)\ldots p(x_n, \theta)$$

Оценкой наибольшего правдоподобия параметра θ называют такое его значение θ^*, при котором функция правдоподобия достигает максимума.

Для удобства часто используют т.н. логарифмическую функцию правдоподобия $\ln L$.

Функции L и $\ln L$ достигают максимума при одном и том же значении θ, поэтому вместо отыскания максимума функции L ищут максимум функции $\ln L$. Это проще, т.к. логарифмическая функция позволяет от произведения перейти к сумме.

Для отыскания максимума функции правдоподобия ее производная или градиент приравнивается к нулю и из полученного уравнения (системы уравнений) находятся искомые оценки параметров:

$\frac{d\ln L}{d\theta} = 0$ или $\nabla L_\theta = 0$

Данные уравнения называются уравнения правдоподобия.

Полученные значения оценок являются максимумом функции правдоподобия, если вторая производная отрицательна (матрица вторых производных отрицательно определена).

Найденную точку максимума θ^* принимают в качестве оценки наибольшего правдоподобия параметра θ.

Непрерывные случайные величины.

Пусть X – непрерывная случайная величина, которая в результате n испытаний приняла значения $x_1, x_2, ..., x_n$. Пусть вид плотности распределения – функции $f(x)$ задан, но неизвестны параметры θ, которыми определяется эта функция. Функцией правдоподобия непрерывной случайной величины X называют функцию аргумента:

$$L(x_1, x_2, \ldots, x_n, \theta) = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$$

Оценку максимального правдоподобия неизвестных параметров распределения непрерывной случайной величины ищут так же, как в случае дискретной случайной величины. Производные (градиент) от функции правдоподобия по оцениваемым параметрами приравниваются к нулю, решается уравнение (система уравнений), далее проверяется, что вторая производная меньше нуля (матрица вторых производных отрицательно определена) в точке, соответствующей решению θ^*.

Определить ММП-оценку параметра *p* биномиального распределения

$$P_N(m) = C_N^m p^m (1-p)^{N-m}$$

Составим логарифмическую функцию правдоподобия:

$$\ln L = \ln\left(\prod_{i=1}^{n} C_N^{m_i} p^{m_i} (1-p)^{N-m_i}\right) = \sum_{i=1}^{n} \ln C_N^{m_i} + \ln p \cdot \sum_{i=1}^{n} m_i + \ln(1-p) \sum_{i=1}^{n} (N - m_i)$$

Приравняем производную от $\ln L$ по p к 0 и решим данное уравнение относительно p:

$\frac{d\ln L}{dp} = \frac{1}{p}\sum_{i=1}^{n} m_i - \frac{1}{1-p}\sum_{i=1}^{n}(N - m_i) = 0.$

$(1-p)\sum_{i=1}^{n} m_i = p\sum_{i=1}^{n}(N - m_i) \Leftrightarrow \sum_{i=1}^{n} m_i = p\sum_{i=1}^{n} N \Leftrightarrow p = \frac{1}{N}\frac{\sum_{i=1}^{n} m_i}{n} = \frac{1}{N} m.$

Оценка совпадает с оценкой по методу моментов.

Найти ММП – оценку математического ожидания *c* для нормального распределения (измерения неравноточные, СКО каждого измерения равно σ_i).

Плотность вероятности нормального распределения для выборочного значения x_i равна:

$$f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i - c)^2}{2{\sigma_i}^2}\right)$$

Составим логарифмическую функцию правдоподобия:

$$\ln L = \ln \prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(\frac{-(x_i - c)^2}{2{\sigma_i}^2}\right)\right] = \sum_{i=1}^{n} \ln\left[\frac{1}{\sqrt{2\pi}\sigma_i}\right] - \sum_{i=1}^{n} \frac{(x_i - a)^2}{2{\sigma_i}^2}$$

Приравняем производную от $\ln L$ по c к 0 и решим данное уравнение относительно c:

$$\frac{d\ln L}{dc} = \sum_{i=1}^{n} \frac{2(x_i - c)}{2{\sigma_i}^2} = 0 \Leftrightarrow c\sum_{k=1}^{n} \frac{1}{{\sigma_k}^2} = \sum_{i=1}^{n} \frac{x_i}{{\sigma_i}^2} \Rightarrow$$

$$c = \left(\sum_{k=1}^{n} \frac{1}{{\sigma_k}^2}\right)^{-1} \sum_{i=1}^{n} \frac{x_i}{{\sigma_i}^2} = \sum_{i=1}^{n} x_i w_i \,, w_i = \frac{1}{{\sigma_i}^2}\left(\sum_{k=1}^{n} \frac{1}{{\sigma_k}^2}\right)^{-1}$$

Найти ММП – оценку математического ожидания для нормального распределения (измерения равноточные, СКО равно σ).

Решение можно получить из предыдущего примера, положив $\sigma_i = \sigma, i = 1 \ldots n$. Тогда

$$w_i = \frac{1}{\sigma^2}\left(\sum_{k=1}^{n}\frac{1}{\sigma^2}\right)^{-1} = \frac{1}{n}$$

и оценка параметра c равна

$c = \sum_{i=1}^{n} x_i w_i = \frac{1}{n}\sum_{i=1}^{n} x_i = x.$

Как видно, при равноточных измерениях ММП-оценка параметра c нормального распределения совпадает со средним арифметическим и оценкой, полученной по методу моментов.

Найти ММП – оценку параметров c и σ для нормального распределения (измерения равноточные).

Для нахождения оценок параметров необходимо приравнять производные от $\ln L$ по этим параметрам к нулю:

$$\begin{cases}\frac{\partial \ln L}{\partial c} = 0\\ \frac{\partial \ln L}{\partial \sigma} = 0\end{cases}$$

Из первого уравнения находится оценка c (предыдущая задача): $c = x$.

Поэтому оценку второго параметра можно найти из второго уравнения. Составим логарифмическую функцию правдоподобия для случая равноточных измерений:

$$\ln L = \ln\prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-(x_i-c)^2}{2\sigma^2}\right)\right] = \sum_{i=1}^{n}\ln\left[\frac{1}{\sqrt{2\pi}\sigma}\right] - \sum_{i=1}^{n}\frac{(x_i-a)^2}{2\sigma^2} =$$

$$= -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mathrm{c})^2$$

Приравняем производную от $\ln L$ по σ к 0 и решим данное уравнение относительно σ:

$$\frac{d\ln L}{d\sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - c)^2 = 0 \Leftrightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Как видно, ММП-оценка σ^2 совпадает со смещенной оценкой дисперсии и совпадает с оценкой по методу моментов.

Найти ММП – оценку параметра экспоненциального распределения

Плотность вероятности экспоненциального распределения для выборочного значения x_i равна:

$f(x_i) = \lambda\exp(-\lambda x_i).$

Составим логарифмическую функцию правдоподобия:

$$\ln L = \ln\prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n}\ln(\lambda\exp(-\lambda x_i)) = n\ln\lambda - \lambda\sum_{i=1}^{n} x_k$$

Приравняем производную от $\ln L$ по λ к 0 и решим данное уравнение относительно λ:

$$\frac{\partial \ln L}{\partial \lambda} = \frac{N}{\lambda} - \sum_{i=1}^{n} x_i = 0 \Leftrightarrow \lambda = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{x}$$

Найти ММП – оценку параметров распределения Лапласа

Плотность вероятности распределения Лапласа для выборочного значения x_i равна:

$$f(x_k) = \frac{\lambda}{2}\exp(-\lambda|x_k - \mathrm{c}|)$$

Составим логарифмическую функцию правдоподобия:

$$\ln L = \ln\prod_{i=1}^{n} f(x_i) = \sum_{i=1}^{n}\ln\left[\frac{\lambda}{2}\exp(-\lambda|x_i - c|)\right] = n\ln(\lambda/2) - \lambda\sum_{i=1}^{n}|x_i - c|$$

ММП-оценки параметров находятся из системы уравнений:

$$\begin{cases}\frac{\partial \ln L}{\partial c}=0\\ \frac{\partial \ln L}{\partial \lambda}=0\end{cases}$$

Найдем вначале ММП-оценку математического ожидания c. Поскольку первое слагаемое не содержит c, его значение не влияет на положение максимума, его можно исключить. Останется одно отрицательное слагаемое, минимум модуля которого совпадает по положению с максимумом функции правдоподобия. Поэтому будем отыскивать ММП-оценку путем поиска минимума суммы $\sum_{i=1}^{n}|x_i-c|$. Эта сумма недифференцируема, и придется находить искомую оценку геометрически.

Пусть в результате эксперимента получено всего три выборочных значения x_1, x_2, x_3, которые разместились на вещественной оси так, как показано на рис.

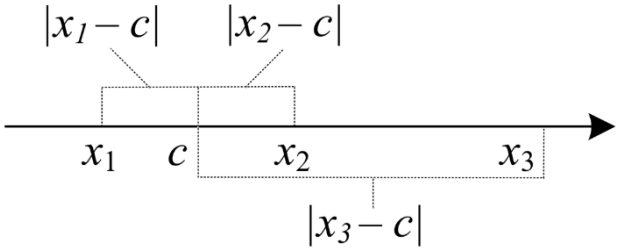


Рис. Подход к определению минимум суммы $\sum_{i=1}^{n}|x_i-c|$

Каждое слагаемое минимизируемой суммы есть расстояние от каждого выборочного значения до точки a. В ситуации, представленной на рисунке, одно из этих расстояний входит в сумму дважды. Это расстояние $|x_2-c|$. Видер, что рассматриваемая сумма достигнет минимума только тогда, когда оценка c совместится с x_2, то есть с выборочной медианой. Добавляя к этой небольшой выборке четное количество элементов, мы придем к тому же выводу, что ММП-оценкой математического ожидания случайной величины, распределенной по Лапласу, является выборочная медиана, то есть $c=x_{med}$. И по определению ММП данная оценка эффективна. Далее найдем ММП-оценку параметра λ. Приравняем производную от ln L по λ к 0 и решим данное уравнение относительно λ:

$$\frac{d\ln L}{d\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}|x_i-c|=0\Rightarrow\lambda=\frac{n}{\sum_{i=1}^{n}|x_i-c|}=\frac{n}{\sum_{i=1}^{n}|x_i-x_{med}|}.$$

Метод минимума Хи-квадрат

Ставится задача оценки параметров θ плотности распределения $f(x, \theta)$, вид которой известен. Исходными данными являются выборочные значения, по которым строится гистограмма. Идея метода заключается в подборе таких значений искомых параметров, при которых достигается минимальное отличие кривой плотности распределения от гистограммы. В качестве меры этого отличия чаще всего используется квадратичный функционал, наиболее удобный для реализации аналитических и численных методов поиска экстремума (минимума или максимума).

В данной задаче в качестве такого функционала используется сумма, обозначенная как χ^2:

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

В этой сумме K - общее количество интервалов, на которых построена гистограмма, n - объем выборки, n_k – количество выборочных значений, попавших в k-ый интервал гистограммы, P_k – вероятность случайной величине с плотностью $f(x, \theta)$ попасть в k-й интервал гистограммы (вероятностная мера k-го интервала гистограммы):

$$P_k=\int_{x_{k-1}}^{x_k}f(x,\theta)dx.$$

Таким образом, слагаемые этой суммы представляют собой квадраты разностей между вероятностными мерами k-ых интервалов, порожденными генеральной плотностью

распределения, и частотными оценками $P_k = n_k/n$ этих вероятностных мер. Знаменателем каждого слагаемого является вероятность P_k, благодаря чему в процессе поиска значений параметров θ повышается вес интервалов с низкой вероятностью и тем самым обеспечивается повышенная точность подгонки в области этих интервалов. Как правило, эти интервалы находятся на удалении от центра распределения (на так называемых “хвостах” распределений).

Вероятности P_k являются функциями от искомых параметров θ, от этих же параметров зависит и величина χ^2, и процедура оценивания параметров по методу минимума χ^2 формально записывается в виде:

$$\theta^* = \arg\min_{\theta}\chi^2 = \arg\min_{\theta}\left[\sum_{k=1}^{K}\frac{(n\cdot P_k(\theta)-n_k)^2}{nP_k(\theta)}\right].$$

В большинстве случаев этот минимум и оценки находятся численными методами. Доказано, что оценки, полученные методом минимума χ^2, обладают свойствами, сопоставимыми со свойствами ММП - оценок, а именно, эти оценки асимптотически эффективны.

Примеры определения оценок 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:

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

$\sigma = \sqrt{\hat{D}[X]} = \sqrt{\overline{s^2}}$, 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

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

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))
```

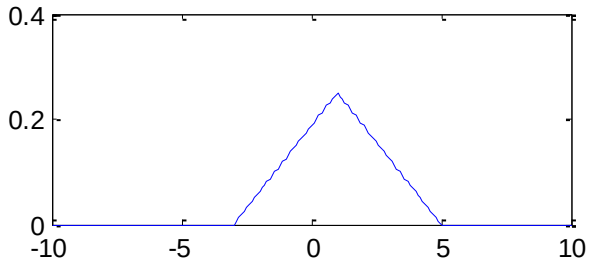
Пример 5 Распределение треугольное, метод ММП, численный.

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:

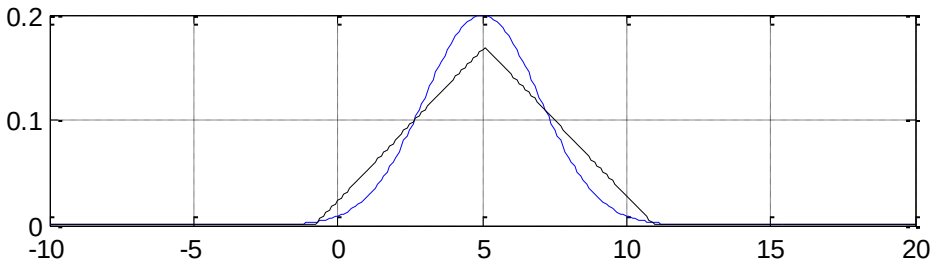
```
data = normrnd(5, 2, [1000 1]);
```

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

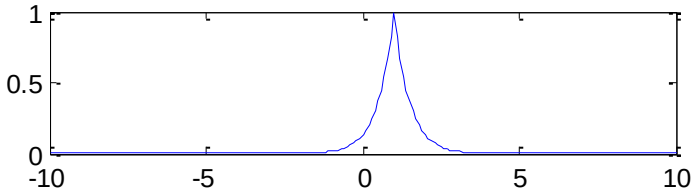
Проделаем действия по аналогии с примером 5.

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

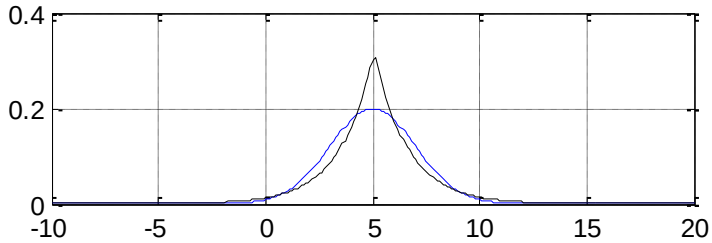
```
data = normrnd(5, 2, [1000 1]);
```

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:

$$c = x_{med};$$

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

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 c
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

% Сгенерируем 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 c
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

% Сгенерируем 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 c
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

% Сгенерируем 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)

% Сгенерируем 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'});
```

Интервальное оценивание

Целью интервального оценивания является вычисление по выборочным данным x объема n такого интервала с границами: нижней - $\underline{\theta}(x)$ и верхней $\overline{\theta}(x)$, чтобы

$$P\left(\underline{\theta}(x) < \theta < \overline{\theta}(x)\right) \geq Q,$$

где Q - вероятность, близкая к единице, например $Q = 0.8 - 0.95$.

Такой интервал называется доверительным интервалом, а вероятность Q – доверительной вероятностью.

Доверительный интервал для математического ожидания.

Если дисперсия генеральной совокупности известна и случайная величина X имеет нормальное распределение, то среднее арифметическое $\bar{x}$ также имеет нормальное распределение с параметрами $N(a, \sigma^2/n)$, где a и σ – истинные значения математического ожидания и СКО. Далее можно рассмотреть СВ

$$z = \frac{(\bar{x} - a)\sqrt{n}}{\sigma}.$$

Она имеет распределение $N(0,1)$. Для z можно построить интервал $[z_\alpha, z_{1-\alpha}]$ такой, что

$$P(z_\alpha < z < z_{1-\alpha}) = 1 - 2\alpha = Q.$$

Путем простых преобразований и учитывая, что $z_{1-\alpha} = -z_\alpha$, $\alpha = (1+Q)/2$, получаем искомый интервал:

$$P\left(\bar{x} - \frac{\sigma}{\sqrt{n}} N_{\frac{1+Q}{2}} < a \le \bar{x} + \frac{\sigma}{\sqrt{n}} N_{\frac{1+Q}{2}}\right) = Q.$$

Если дисперсия генеральной совокупности неизвестна и для ее оценки используется несмещенная оценка

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$

то можно показать, что случайная величина

$$t = \frac{\bar{x} - a}{s}\sqrt{n}.$$

имеет распределение Стьюдента с $(n-1)$ степенями свободы. Аналогично для t можно построить интервал $[t_\alpha, t_{1-\alpha}]$ такой, что

$$P(t_\alpha < t < t_{1-\alpha}) = 1 - 2\alpha = Q.$$

Путем простых преобразований и учитывая, что $t_{1-\alpha} = -t_\alpha, \alpha = (1+Q)/2$, получаем искомый интервал:

$$P\left(\bar{x} - \frac{s}{\sqrt{n}} \cdot t_{\frac{1+Q}{2}}(n-1) \; < \; a \; \le \; \bar{x} + \frac{s}{\sqrt{n}} \cdot t_{\frac{1+Q}{2}}(n-1)\right) = Q.$$

С увеличением n полученными формулами можно пользоваться и для случая если распределение X отличается от нормального (в силу центральной теоремы распределение $\bar{x}$ стремится к нормальному).

Доверительный интервал для дисперсии

При использовании несмещенной оценки дисперсии

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

в предположении, что СВ X имеет нормальное распределение можно показать, что СВ

$$z = \frac{s^2(n-1)}{\sigma^2}$$

имеет распределение χ^2 с $(n-1)$ степенями свободы. Для z можно построить интервал $[z_\alpha, z_{1-\alpha}]$ такой, что

$$P(z_\alpha < z < z_{1-\alpha}) = 1 - 2\alpha = Q.$$

Путем простых преобразований получаем искомый интервал:

$$P\left(s^2 \cdot \frac{n-1}{\chi^2_{\frac{1+Q}{2}}(n-1)} \; < \; \sigma^2 \; \le \; s^2 \cdot \frac{n-1}{\chi^2_{\frac{1-Q}{2}}(n-1)}\right) = Q.$$

Рис. Расположение квантилей распределения Хи-квадрат

Доверительные интервалы для интерквантильного промежутка

Доверительные интервалы для интерквантильного промежутка называются толерантными пределами. Толерантные пределы накрывают истинные интерквантильные промежутки J_P с заданной доверительной вероятностью Q.

Параметрические толерантные пределы

Параметрические толерантные пределы рассчитываются в предположении, что СВ X имеет нормальное распределение. В этом случае доверительный интервал J_P можно найти как

$$J_P: \left[a - \sigma \cdot N_{\frac{1+P}{2}}, a + \sigma \cdot N_{\frac{1+P}{2}}\right].$$

Тогда задачу отыскания толерантных пределов можно поставить следующим образом:

$$P\left(\underline{tol} \le a - \sigma \cdot N_{\frac{1+P}{2}}, a + \sigma \cdot N_{\frac{1+P}{2}} \le \overline{tol}\right) = Q.$$

Данная задача имеет решение, которое выражается через т.н. толерантные множители $\kappa(n, P, Q)$, протабулированные в ряде справочников и зависящие от объема выборки n, вероятности интерквантильного промежутка P и доверительной вероятности Q. Искомые параметрические толерантные пределы определяются следующим выражением:

$$\left(\bar{x} - \kappa(n, P, Q) \cdot s\ , \bar{x} - \kappa(n, P, Q) \cdot s\ \right),$$

где как и ранее $\bar{x}$ - среднее арифметическое, s^2 – несмещенная оценка дисперсии.

Табл. Значения толерантных множителей и границы J_p для $N(0,1)$

$Q\backslash n$	5	8	12	20	100	$N_{\frac{1+P}{2}}$
0.9	4.152	3.264	2.863	2.564	2.170	1.65
0.95	5.079	3.732	3.162	2.752	2.231	1.96

Непараметрические толерантные пределы

Для произвольного распределения оценивают т.н. непараметрические толерантные пределы, являющиеся доверительном интервалом для интерквантильного промежутка. Для этого на основе выборки все пространство делится на n т.н. статистически эквивалентных блоков: $(-\infty, x_1), (x_1, x_2), (x_2, x_3), \ldots, (x_n - 1, x_n)$. Вероятность каждого блока равна $1/n$ (это можно получить, рассмотрев СВ $Y = F(X)$, где F – функция распределения, $y \in R(0,1)$.

Поэтому интерквантильный промежуток можно построить из любых последовательных $m = nP$ блоков. Другими словами, вероятность того, что блок попадет в J_P равна P. Если же выбирать m блоков из n, то нужно найти количество блоков, которое накроет оценку m/n вероятности P с заданным значением Q. Для вероятности P выше была получена нижняя граница доверительной вероятности:

$$P\left(P \le \underline{p}\right) = \sum_{m=m_0}^{n} C_n^m \underline{p}^m \left(1 - \underline{p}\right)^{n-m} \le \alpha.$$

По условию нижняя граница задана и равна $\underline{p} = P$, а верхняя равна 1. Если задать $\alpha = 1-Q, m_0 = n - k$, то задача заключается в отыскании такого минимального значения k, для которого выполняется неравенство:

$$\sum_{m=n-k}^{n} C_n^m P^m (1-P)^{n-m} \leq 1 - Q,$$

где $n - k$ - число статистически эквивалентных блоков, находящихся внутри интервала между элементами вариационного ряда, которые желательно объявить толерантными пределами, иными словами, границами доверительного интервала для искомого интерквантильного промежутка, определенного при вероятности P. В теории непараметрического интервального оценивания число k именуется, как количество отброшенных статистически эквивалентных блоков.

Задача отыскания непараметрических толерантных пределов может ставиться 2 способами:

1. Задан количество k отбрасываемых статистических эквивалентных блоков. Требуется найти объем выборки n, позволяющий обеспечить заданные вероятности P, Q.

2. Зафиксирован объем выборки требуется найти количество k отбрасываемых статистических эквивалентных блоков, обеспечивающих заданные вероятности P, Q.

Важно понимать, что при $k = 0$ толерантные пределы получаются в виде $(-\infty, xn)$.
Выше был рассмотрен вариант нахождения непараметрических толерантных пределов, симметричных относительно математического ожидания. Если распределение симметрично относительно нуля, то иногда находят непараметрические толерантные пределы, симметричные относительно нуля. Отличие заключается в том, что вначале вся выборка берется по модулю (получается вариационный ряд y_i), а затем количество k отбрасываемых блоков рассчитывается без учета блока $(0, y_1)$, поскольку обе границы находятся справа. Это позволяет сэкономить один блок, а сами пределы записываются в форме $(-y_{n-k}, y_{n-k})$. При $k = 0$ в данном случае толерантные пределы получаются в виде $(-y_n, y_n)$.

Приложение 4 Проверка гипотезы о виде плотности распределения

Критерий “хи - квадрат”

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

H_0: плотность распределения случайной величины X есть $\varphi(\theta, x)$

против альтернативы

H_1: плотность распределения случайной величины X не $\varphi(\theta, x)$.

Поскольку эта гипотеза сложная, задается только вероятность ошибки первого рода – уровень значимости.

Для проверки сформулированной гипотезы естественно построить оценку плотности распределения – гистограмму и сопоставить ее с предполагаемой плотностью распределения. На рис. приведен пример гистограммы и кривая предполагаемой плотности распределения $\varphi(\theta, x)$, которая построена после того, как по выборочным значениям вычислены оценки θ ее параметров .

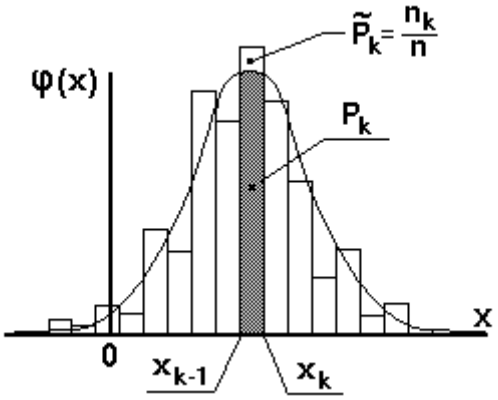


Рис. Иллюстрация к использованию критерия Хи-квадрат

Степень различия между гистограммой и предполагаемой плотностью распределения или статистика критерия выражается суммой квадратов разностей

$$\sum_{k=1}^{K}\left(P_k - P_k\right)^2 = \sum_{k=1}^{K}\left(\frac{nP_k - n_k}{n}\right)^2,$$

где

$$P_k = \int_{x_{k-1}}^{x_k} \varphi(\theta, x)dx = F(x_k) - F(x_{k-1})$$

то есть вероятность попадания значения случайной величины в интервал $(x_{k-1}, x_k]$ при условии справедливости нулевой гипотезы, $P_k = n_k/n$ - оценки этих вероятностей, где n_k - количество выборочных значений, попавших в интервал $(x_{k-1}, x_k]$, n - объем выборки, K - общее количество интервалов, на которых построена гистограмма.

При больших значениях N можно показать, что введенная статистика критерия принадлежит распределению Хи-квадрат с $K - r$ степенями свободы.

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

При заданной вероятности ошибки первого рода α (уровень значимости), критическое значение $z_{крит}$ назначается исходя из следующих предпосылок: при справедливости нулевой гипотезы маловероятно, чтобы статистика критерия оказалась слишком большой, поэтому достаточно задать критическое значение таким, что вероятность превышения его не больше α. Для этого достаточно вычислить квантиль $\chi^2_{1-\alpha}(K-r)$ и использовать ее в качестве значения $z_{крит}$. Вероятность ложного срабатывания, т.е. принятия гипотезы H_1 при условии, что на самом деле верна гипотеза H_0 будет равна α.

Полученный критерий называется критерием “хи - квадрат” (Пирсона) проверки гипотезы о виде плотности распределения (или закона распределения) генеральной совокупности по экспериментальным данным.

Процедура проверки гипотезы о виде плотности распределения по критерию “хи - квадрат”.

1. Задается уровень значимости α
2. По выборочным данным строится гистограмма в соответствии с указаниями.
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 $\varphi(\theta, x)$.
5. Вычисляются оценки θ параметров предполагаемой плотности распределения, в результате получается плотность распределения $\varphi(\theta, x)$.
6. С использованием $\varphi(\theta, x)$ или $F(\theta, x)$ вычисляются вероятности

$P_k = \int_{x_{k-1}}^{x_k} \varphi(\theta, x)dx = F(\theta, x_k) - F(\theta, x_{k-1})$.

7. Вычисляется статистика критерия

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

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.

Зам. С уменьшением вероятности α возрастает критическое значение$\chi^2_{1-\alpha}(K-r)$, а это значит, что объективно вероятность β пропуска события (т.е. принятия гипотезы H_0 при условии, что верная гипотеза H_1). Действительно, если задать α = 0, то критическое значение $\chi^2_1(K-r) = \infty$, а это означает, что нулевая гипотеза будет всегда приниматься и ошибка второго рода β будет равна 1.

Критерий Колмогорова - Смирнова

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

H_0: функция распределения случайной величины X есть $F(\theta, x)$

против альтернативы

H_1: функция распределения случайной величины X не$F(\theta, x)$.

Поскольку эта гипотеза сложная, задается только вероятность ошибки первого рода α - уровень значимости.

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

где $F(x_i)$ - значения выборочной функции распределения при $x = x_i$.

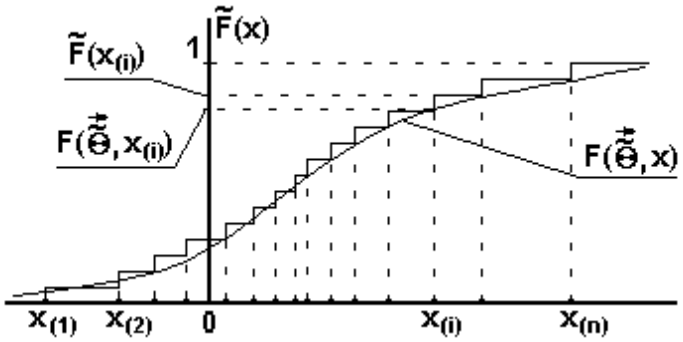


Рис. Выборочная и предполагаемая функции распределения

Статистикой критерия является величина D. Критические значения для различных значений α табулированы и приводятся практически во всех учебниках и справочниках по математической статистике. В таблице ниже приводятся некоторые часто употребляемые критические значения.

Таблица 4

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

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

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

1. Задается уровень значимости α.

2. По выборочным данным строится выборочная функция распределения.

3. Вычисляются точечные оценки моментов.

4. Из теоретических и практических соображений (вид выборочной функции распределения, гистограммы, соотношения между моментами, значениям асимметрии и эксцесса) выдвигается гипотеза о виде функции распределения $F(\theta, x)$ и тем самым - о виде плотности распределения $\varphi(\theta, x)$.

5. Оценивается r параметров θ предполагаемой функции распределения и ее значения $F(\theta, x_i)$ при $x = x_i$.

6. Вычисляется статистика критерия $D = \sup_i \left| F(x_i) - F(\theta, x_i) \right|$

7. Полученное значение сравнивается с критическим значением D_α.

8. Если $D > D_\alpha$, 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.

9. Если $D \le D_\alpha$ 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.

Зам. Для корректного применения критерия Колмогорова - Смирнова выборку $x_1, x_2, \ldots, x_n$. следует разделить на две части и по одной из них оценить параметры θ, а по другой построить выборочную функцию распределения и вычислить статистику критерия *D*. 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 w² Мизеса

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

H_0: функция распределения случайной величины *X* есть $F(\theta, x)$

против альтернативы

H_1: функция распределения случайной величины *X* не $F(\theta, x)$.

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. Различие между ними определено, как

$$\omega^2 = \int_{-\infty}^{\infty} \left| F(x) - F(\theta, x) \right|^2 \varphi(\theta, x) \cdot dx,$$

где $\varphi(\theta, x)$ - предполагаемая плотность распределения.

Этот интеграл вычисляется, как сумма интегралов по интервалам между соседними членами вариационного ряда. Если на этих интервалах предполагаемая функция распределения интерполируется прямой линией, то этот интеграл выражается суммой

$$\omega^2 = \sum_{i=1}^{n-1} \int_{x_i}^{x_{i+1}} \left| F(x) - F(\theta, x) \right|^2 \varphi(\theta, x) \cdot dx = \frac{1}{12n^2} + \frac{1}{n} \cdot \sum_{i=1}^{n} \left[F\left(\theta, x_i\right) - \frac{2i-1}{2n} \right]^2.$$

В качестве статистики критерия используется

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

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

Таблица 5

Критические значения критерия ω^2Мизеса

α	0.03	0.05	0.1	0.2
$(n\omega^2)_\alpha$	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.

3. Вычисляются точечные оценки моментов.

4. Из теоретических соображений выдвигается гипотеза о виде функции распределения $F(\theta, x)$ и тем самым - о виде плотности распределения $\varphi(\theta, x)$.

5. Оценивается *r* параметров θ предполагаемой функции распределения и ее значения $F(\theta, x_i)$ при $x = x_i$.

6. Вычисляется статистика критерия

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

7. Полученное значение сравнивается с критическим значением $(n\omega^2)_\alpha$.

8. Если $n\omega^2 > n\omega^2{}_\alpha$ 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.

9. Если $n\omega^2 \leq n\omega^2{}_\alpha$ 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.

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

Литература

Задачи

1. Свешников А.А. (ред.) Сборник задач по теории вероятностей, математической статистике. СПб.: Лань, 2008.

Matlab

1. Потемкин Система Инженерных И Научных Расчетов Matlab
2. Мэтьюс Финк Численные методы Использование Matlab
3. Чен, Джиблин, Ирвинг Matlab в математических исследованиях.djvu – есть упражнения с комментариями
4. Лазарев Начала программирования в среде Matlab

Расчетные задания

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