Основы статистического анализа в «R»

ЦМФ. Математические финансы

Затабулированные распределения

Название	Обозначение в R	Параметры
Нормальное	norm	mean, sd
t-распределение	t	df
Равномерное	unif	min, max
Хи-квадрат	chisq	df
F-распределение	f	df1, df2
Гамма	gamma	shape, scale
…	…	…

```
# пример со стандартным нормальным распределением
N <- 100; x <- seq(-5,5,by=0.1); alpha <- 0.95
rnorm(n=N,mean=0,sd=1)          # генератор случайных чисел
qnorm(alpha,mean=0,sd=1)        # квантиль
pnorm(x,mean=0,sd=1)            # функция распределения
dnorm(x,mean=0,sd=1)            # функция плотности
```

Гистограмма и эмпирическая плотность

```
y <- faithful$eruptions
```
исходные данные

гистограмма с диапазоном данных от 1.6 до 5.2

длина интервалов — 0.2

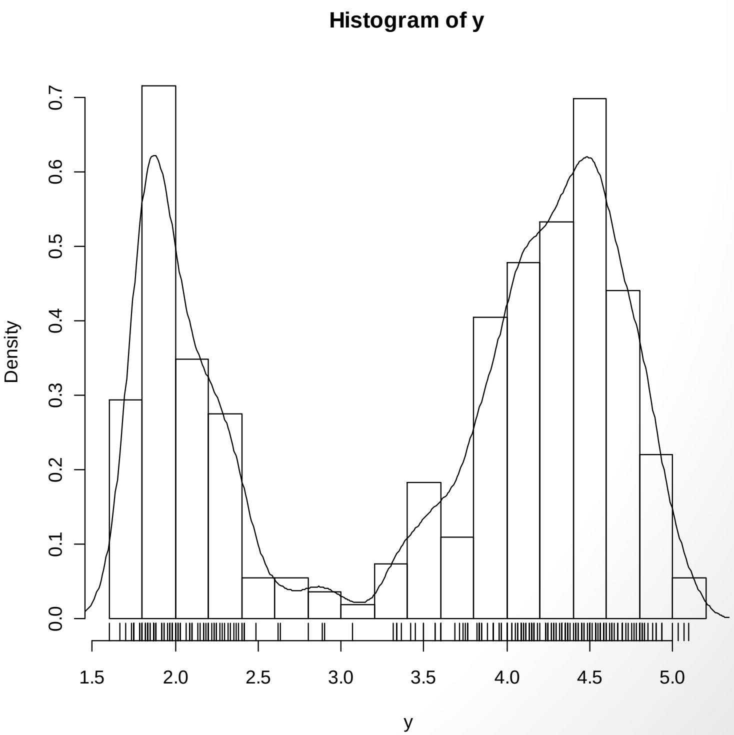
```
hist(y,breaks=seq(1.6,5.2,by=0.2),prob=TRUE)
```

добавление эмпирической плотности

```
y.pdf <- density(y,bw="ucv")
lines(y.pdf)
```

добавление исходных данных

```
rug(y)
```



Эмпирическая функция распределения

```
y.cdf <- ecdf(y)
```

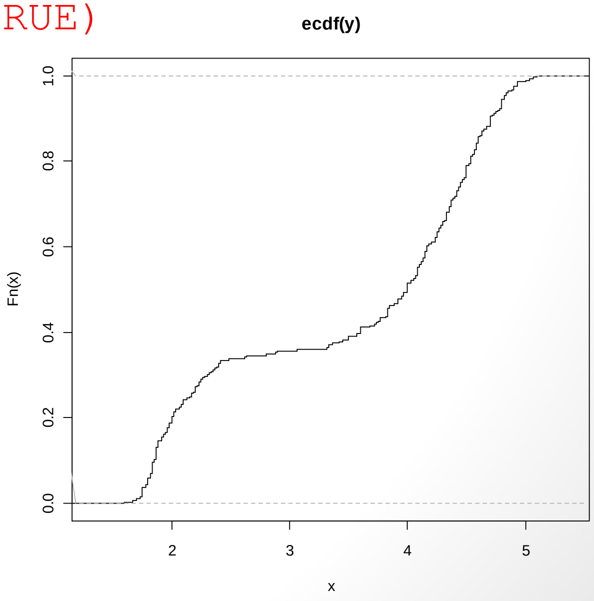
y.cdf — функция, подставляя в неё квантили, мы получаем
значения функции распределения

$y.cdf(x) = \frac{1}{n} \sum I(x_i < x)$

```
y.cdf(3)
[1] 0.3566176
```

график

```
plot(y.cdf,do.points=FALSE,verticals=TRUE)
```



Сравнение с затабулированным распределением

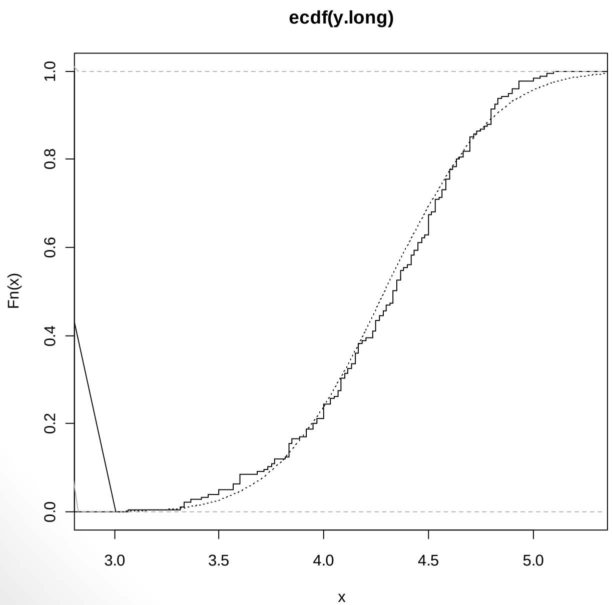
```
y.long <- y[y>3]
plot(ecdf(y.long),do.points=FALSE,verticals=TRUE)
x <- seq(3,5.4,by=0.1)
```

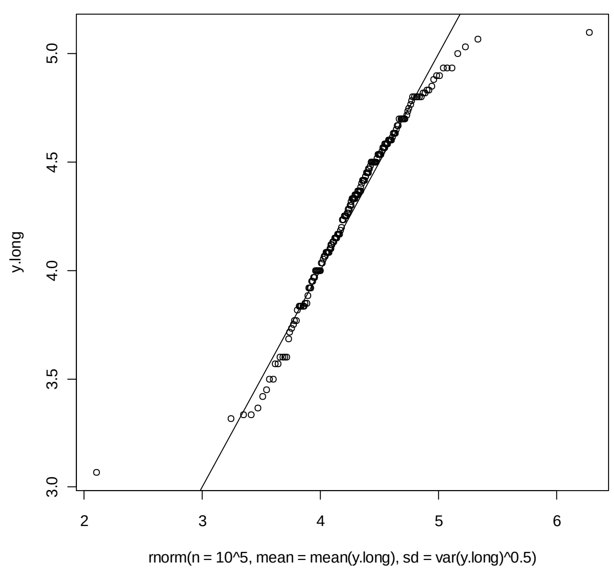
график нормального распределения

```
lines(x,pnorm(x,mean=mean(y.long),sd=var(y.long)^0.5),lty=3)
```

график квантиль–квантиль

```
qqplot(rnorm(n=10^5,mean=mean(y.long),
sd=var(y.long)^0.5),y.long); abline(0,1)
```





Тесты на нормальность

\# Шапиро–Уилка

\# гипотеза: $H_0: x \sim N(\mu, \sigma)$

\# статистика: $W = \frac{\left(\sum a_i x_{(i)}\right)^2}{\sum (x_i - \bar{x})^2}$, $(a_1, \dots, a_n) = \frac{m' V^{-1}}{(m' V^{-1} V^{-1} m)^{0.5}}$,

\# $m_i = E\left(x_{(i)} \middle| x \sim N(0,1)\right)$, $V = cov(m)$

```
shapiro.test(y.long)
```

\# Колмогорова–Смирнова

\# гипотеза: $H_0: x \sim F(x)$

\# статистика: $D = \sup_x |y.cdf(x) - F(x)|$

```
ks.test(y.long,"pnorm",mean=mean(y.long),sd=var(y.long)^0.5)
```

Сравнение двух нормальных выборок

непарный t-тест на равенство средних

гипотеза: H_0: $\bar{x}_1 - \bar{x}_2 = 0$

статистика: $t = \frac{(\bar{x}_1 - \bar{x}_2)}{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^{0.5}} \sim t(N_1 + N_2 - 2)$

```
n1 <- rnorm(n=100,mean=0,sd=1)
n2 <- rnorm(n=100,mean=0.1,sd=1.1)
t.test(n1,n2,var.equal=FALSE,conf.level=0.95)
```

F-тест на равенство дисперсий

гипотеза: H_0: $\frac{s_1}{s_2} = 1$

статистика: $F = \frac{s_1}{s_2} \sim F(N_1, N_2)$

```
var.test(n1,n2,conf.level=0.95)
```

Сравнение двух произвольных выборок

ранговый тест Уилкоксона на равенство средних

гипотеза: H_0: $\bar{x}_1 - \bar{x}_2 = 0$

статистика: $U = \min\left(R_1 - \frac{N_1(N_1+1)}{2}, R_2 - \frac{N_2(N_2+1)}{2}\right),\ R_i = \sum rank\left(x_{i,j}\right)$

```
t <- rt(n=100,df=5); n <- rnorm(n=100,mean=0,sd=1)
wilcox.test(t,n,conf.level=0.95)
```

тест Колмогорова–Смирнова

гипотеза: $H_0: F_1(x) \equiv F_2(x)$

статистика: $D = \sup_x \left|\hat{F}_1(x) - \hat{F}_2(x)\right|$

```
ks.test(t,n)
```

Домашнее задание

- скачать данные о доходности трёх акций или биржевых индексов с сайтов finam.ru, finance.yahoo.com или др.
- провести тесты на нормальность их распределения
- рассмотреть график «квантиль–квантиль» для эмпирического распределения доходностей и нормального распределения, сделать выводы о лёгкости или тяжести эмпирических хвостов
- написать комментарии