Распределения дискретных и непрерывных случайных величин и знакомство с ними в R

К.Г.Грибанов

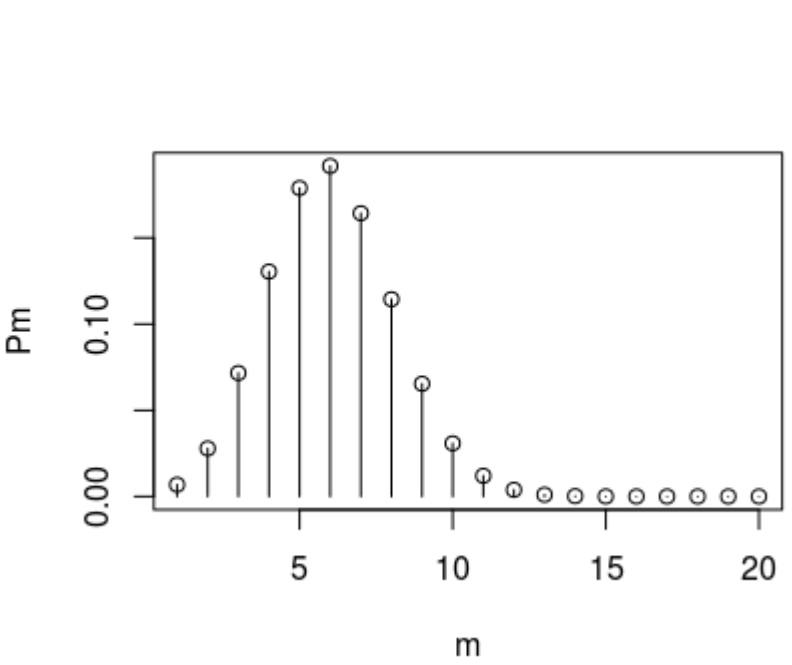
для курса «Современные методы обработки данных»

Биномиальное распределение описывает распределение числа ***m*** появлений случайного события при проведении ***n*** независимых испытаний, когда при каждом наблюдении событие может наступить с вероятностью ***q***. Описывает бросание монеты, число отказов при контроле качества продукции. Вычисляется по формуле

$$P(m) = \frac{n!}{m!(n-m)!} q^m (1-q)^{n-m} = C_n^m \, q^m (1-q)^{n-m}$$

Посмотрим на него в R:

```
m = 0:20
n = 20
q = 0.3
Pm = dbinom(m,n,q)
plot(m,Pm,type=«h»,lwd=2
points(m,Pm,lwd=2)
```

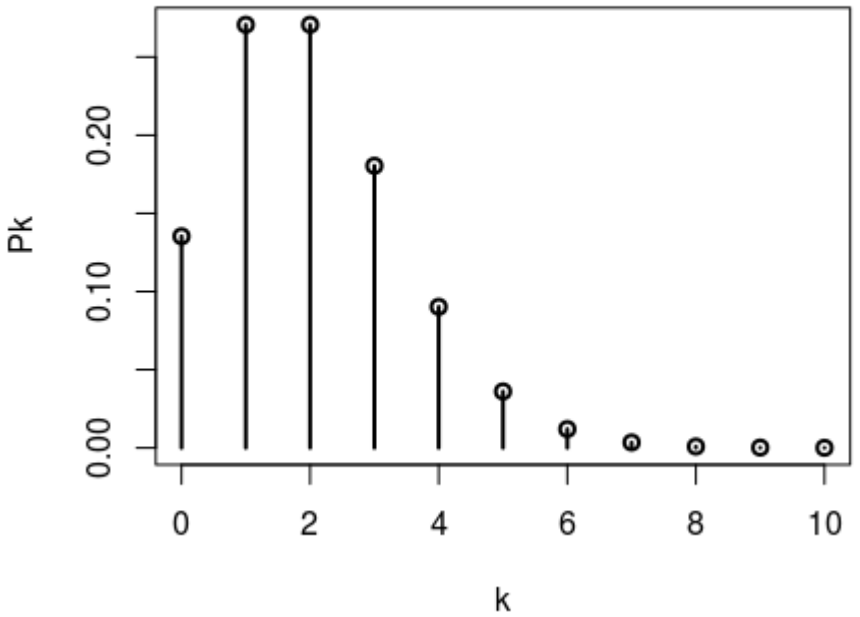


Распределение Пуассона описывает распределение случайной величины *K* со значениями *k=0,1,2,…* Используется при анализе потока заявок на обслуживание, при оценке вероятности появления *k* однородных событий в фиксированные промежутки времени. Имеет всего один параметр *A*, который совпадает и с математическим ожиданием, и с дисперсией, т. е. $M_K = D_K = A$

$$P(k) = \frac{A^k}{k!} e^{-A}, \quad k = 1,2,3,\ldots$$

R:

```
k <- 0:10
A <- 2
Pk <- dpois(k,A,log=FALSE)
plot(k,Pk,type=«h»,lwd=2)
points(k,Pk,lwd=2)
```



Имена статистических функций в R

Распределение	Базовое имя в R
Биномиальное	binom
Пуассона	pois
Бета-распределение	beta
Распр-е Вейбулла	weibull
Нормальное	norm
Логарифмически нормальное	lnorm
Равномерное	unif
...	...

Собственно функций с именами выше нет, тип функции определяет буква, которая добавляется вначале, **d** задает имя плотности распределения (например dnorm()), **p** задает имя закона распределения (интегральная функция вероятности), **q** — определяет функцию вычисления квантилей, **r** — генератор случайных чисел с заданной плотностью.

Бета-распределение описывает распределение непрерывной случайной величины *X*, ограниченной на отрезке [0;1]. Используется для описания частоты появления случайного события, для оценивания относительного времени, оставшегося до завершения какого-то процесса, для анализа суточного производства продукции и т. п. Зависит от 2-х параметров формы $\Lambda > 0$ и $\sigma > 0$:

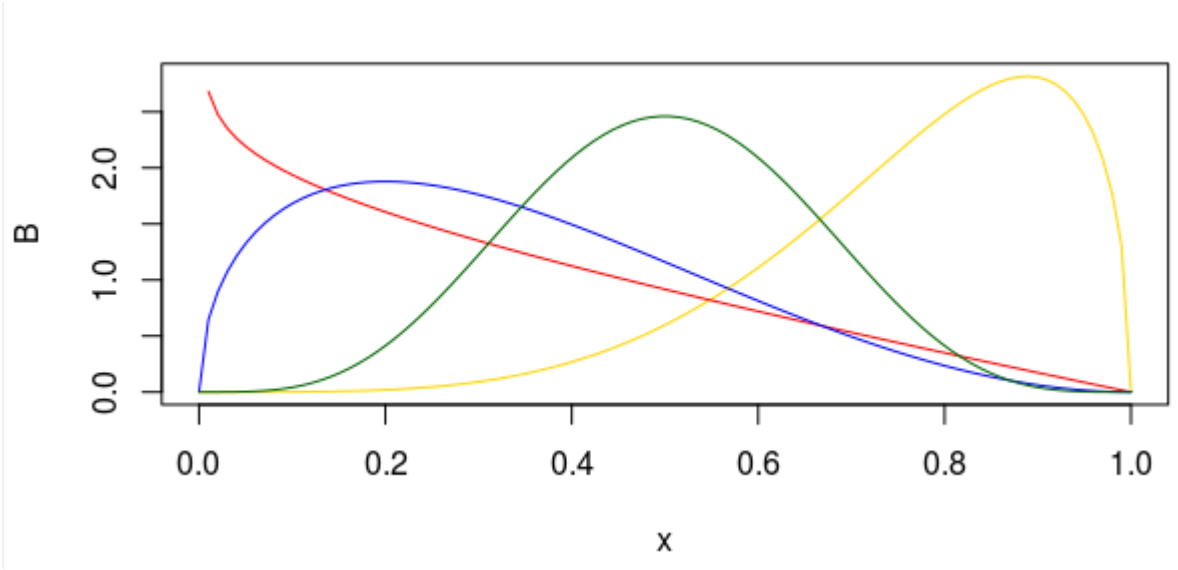
$$f(x) = \frac{1}{\mathrm{B}(\Lambda, \sigma)} x^{\Lambda-1}(1-x)^{\sigma-1}, \qquad x \in [0;1], \qquad \mathrm{B}(\Lambda, \sigma) = \frac{\Gamma(\Lambda)\Gamma(\sigma)}{\Gamma(\Lambda+\sigma)}$$

$$M_X = \frac{\Lambda}{\Lambda+\sigma}, \qquad D_X = \frac{\Lambda\sigma}{(\Lambda+\sigma)^2(\Lambda+\sigma+1)}$$

$\Gamma(z)$ - гамма-функция, специальная функция, определяется сложно, определена в общем случае для комплексных чисел, иногда говорят, что она расширяет понятие факториала на поле комплексных чисел… Однако пользователям R не о чем беспокоиться, все сделано за них :-)

Бета-распределение в R:

```
x = seq(0,1,length=100)
L = c(0.9, 1.5, 5.0, 5.0)
S = c(2.0, 3.0, 5.0, 1.5)
colors = c("red","blue","darkgreen","gold","black")
B = dbeta(x, L[4], S[4], ncp=0, log=FALSE)
plot(x,B,type="l",col=colors[4])
for (i in 1:3) {
  lines(x, dbeta(x,L[i],S[i],ncp=0,log=FALSE),col=colors[i])
}
```



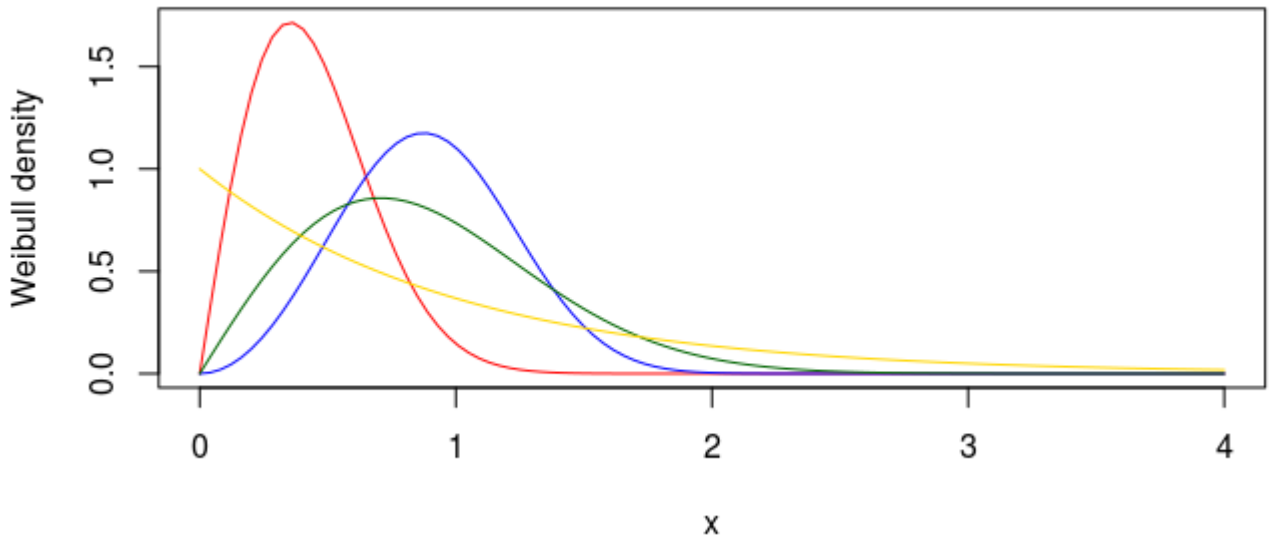
Распределение Вейбулла описывает распределение положительной случайной величины и используется в теории надежности для анализа времени безотказной работы некоторых радиотехнических элементов и технических систем, состоящих из элементов с разными интенсивностями отказов. Зависит от параметра формы $\Lambda > 0$ и параметра масштаба $\sigma > 0$:

$$f(x) = \frac{\Lambda}{\sigma}(x/\sigma)^{\Lambda-1}\exp\left(-(x/\sigma)^{\Lambda}\right)$$

$$M_X = \sigma\Gamma\left(\frac{1}{\Lambda}+1\right); \qquad D_X = \sigma^2\left\{\Gamma\left(\frac{2}{\Lambda}+1\right)-\left[\Gamma\left(\frac{1}{\Lambda}+1\right)\right]^2\right\}$$

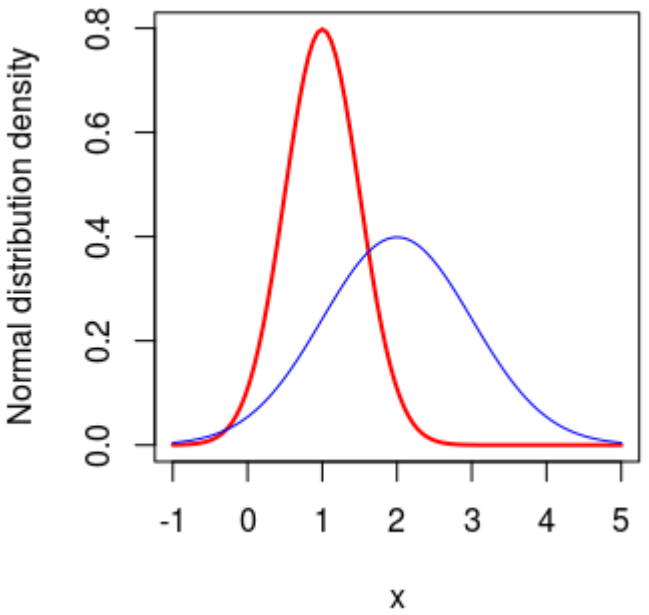
Распределение Вейбулла в R:

```
x = seq(0,4,length=100)
L = c(2.0, 3.0, 2.0, 1.0)
S = c(0.5, 1.0, 1.0, 1.0)
colors = c("red","blue","darkgreen","gold","black")
f = dweibull(x, L[1], S[1], log=FALSE)
plot(x,f,type="l",col=colors[1],ylab="Weibull density")
for (i in 2:4) {
  lines(x, dweibull(x,L[i],S[i],log=FALSE),col=colors[i])
}
```



Нормальное распределение характеризует распределение непрерывной случайной величины $X \in [-\infty; \infty]$ и используется для описания широкого класса случайных величин, на которые влияет большое количество случайных факторов. Плотность распределения удачным образом зависит от математического ожидания и дисперсии:

$$f(x) = \frac{1}{\sqrt{2\pi D_X}} \exp\left[-\frac{1}{2} \frac{(x - M_X)^2}{D_X}\right]$$



R:

```
x <- seq(-1,5,length=100)
f <- dnorm(x,1.0,sqrt(0.25),log=FALSE)
plot(x,f,type="l",lwd=2,col="red",ylab="Normal distribution density")
lines(x,dnorm(x,2.0,sqrt(1),log=FALSE),col="blue")
```

Логарифмически нормальное распределение представляет распределение положительной случайной величины $X \in [0;\infty]$ и применяется для описания размера частиц при их случайном дроблении (размер градин например), многомерное логнормальное распределение описывает распределение вертикальных профилей водяного пара в атмосфере. Если СВ имеет логнормальное распределение, то ее логарифм имеет нормальное распределение. Плотность вероятности зависит от параметра положения μ и параметра масштаба σ :

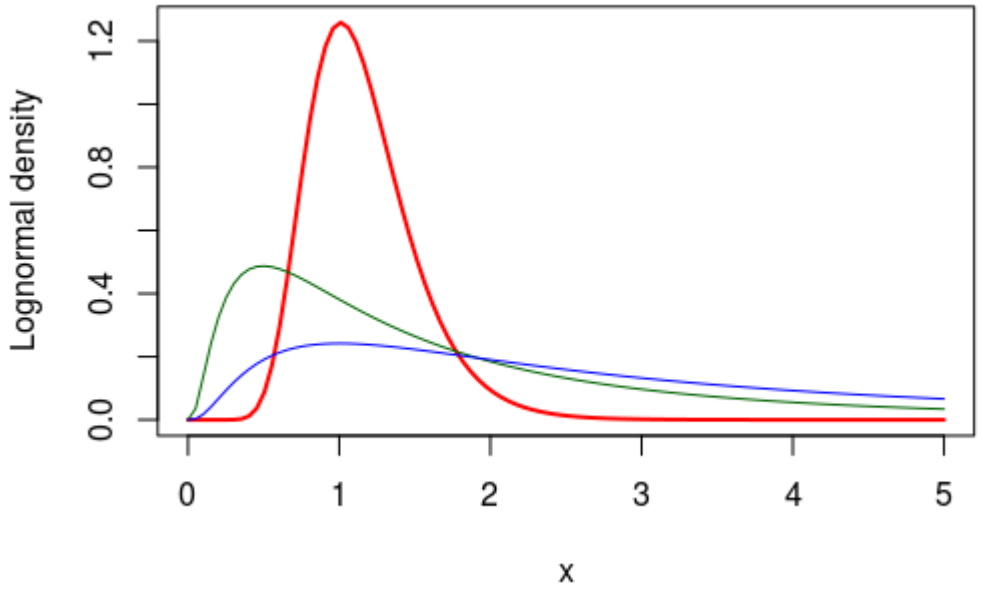
$$f(x) = \frac{1}{\sqrt{2\pi}\, x\, \sigma} \exp\left[-\frac{1}{2\sigma^2}(\ln x - \mu)^2\right]$$

$$M_X = \exp(\mu + 0.5\,\sigma^2)$$

$$D_X = \left[\exp(\sigma^2) - 1\right] \exp(2\mu + \sigma^2)$$

Логарифмически нормальное распределение в R:

```
x <- seq(0,5,length=100)
mu <- c(0.1, 0.3, 1.0)
sigma <- c(0.3, 1.0, 1.0)
f <- dlnorm(x,mu[1],sigma[1],log=FALSE)
plot(x,f,type="l",lwd=2,col="red",ylab="Lognormal density")
lines(x,dlnorm(x,mu[2],sigma[2],log=FALSE),col="darkgreen")
lines(x,dlnorm(x,mu[3],sigma[3],log=FALSE),col="blue")
```



Распределение Стьюдента (t-распределение) с α $(\alpha > 0)$ степенями свободы описывает поведение СВ $X \in [-\infty; \infty]$:

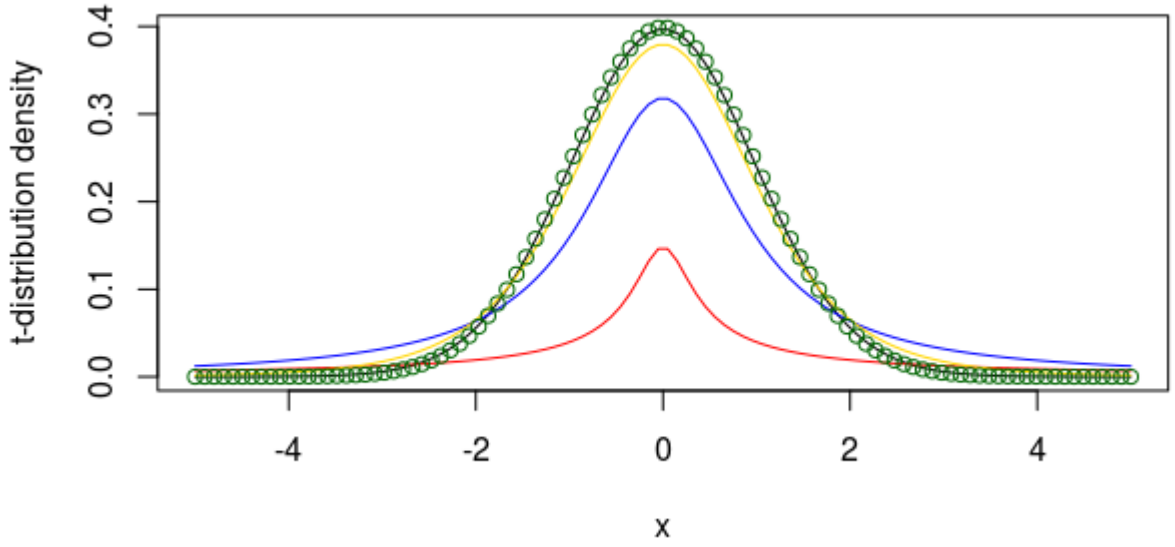
$$f(x) = \frac{\Gamma\left((\alpha+1)/2\right)}{\sqrt{\alpha\pi}\,\Gamma\left(\alpha/2\right)} \left(1+\frac{x^2}{\alpha}\right)^{-(\alpha+1)/2}$$

$$M_X = 0; \quad D_X = \alpha/(\alpha-2), \quad \alpha>2; \quad D_X = \infty, \ \alpha \le 2$$

Распределение появляется при проверке гипотезы о среднем нормально распределенной генеральной совокупности при неизвестной дисперсии. При больших α асимптотически приближается к стандартному нормальному распределению.

Распределение Стьюдента в R:

```
x <- seq(-5,5,length=100)
alpha <- c(0.1,1.0, 5.0, 50.0)
f <- dt(x,alpha[4],ncp=0,log=FALSE)
colors = c("red","blue","gold","black")
plot(x,f,type="l",ylab="t-distribution density",col=colors[4])
for (i in 1:3) {
 lines(x,dt(x,alpha[i],ncp=0,log=FALSE),col=colors[i])
}
points(x,dnorm(x,0,1,log=FALSE),col="darkgreen")
```



Метод максимального правдоподобия основан на использовании функции правдоподобия, которая строиться на основе выборки СВ и известной функции распределения:

$$L(\boldsymbol{\theta}) = P\{\{x_i\}, \boldsymbol{\theta}\} = \prod_{i=1}^{n} P\{x_i, \boldsymbol{\theta}\}$$

для дискретной величины или

$$L(\boldsymbol{\theta}) = f(\{x_i\}, \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

для непрерывной СВ. Сущность метода в максимизации функции $L(\boldsymbol{\theta})$ любым способом, от применения аналитических формул до использования численных алгоритмов оптимизации.

ММП в R:

```
library(MASS)
mymean = 3.22
mysd = 0.25
x <- rnorm(500,mean=mymean,sd=mysd)
print("True mean and sd:")
print(c(mymean,mysd))
fd <- fitdistr(x,"normal")
hist(x,probability=TRUE,main=NULL,ylab="f",xlab="x")
xvals <- seq(min(x),max(x),length=200)
f <- dnorm(xvals,fd$estimate[1],fd$estimate[2],log=FALSE);
lines(xvals,f)
print("Estimated parameters:")
print(fd$estimate)
```