

Разширени графични възможности и вероятностни разпределения

Статистическа обработка на данни с R

Пламен Петров и Тодор Балабанов

Център за обучение
Институт по информационни и комуникационни технологии
Българската академия на науките
p.petrov@iit.bas.bg todorb@iinf.bas.bg

1.VI.2020

Acknowledgments

These teaching materials are funded by Velbazhd Software LLC and it is partially supported by the Bulgarian Ministry of Education and Science (contract D01–205/23.11.2018) under the National Scientific Program “Information and Communication Technologies for a Single Digital Market in Science, Education and Security (ICTinSES)”, approved by DCM # 577/17.08.2018.

Съдържание

1 Разширени графични възможности

- Хистограми и плътности
- Диаграми на разсейване
- Графики тип кутия и цигулка
- Линейни графики
- Тематично оформление

2 Изследване на случайни величини

- Зарове за игра

3 Вероятностни разпределения

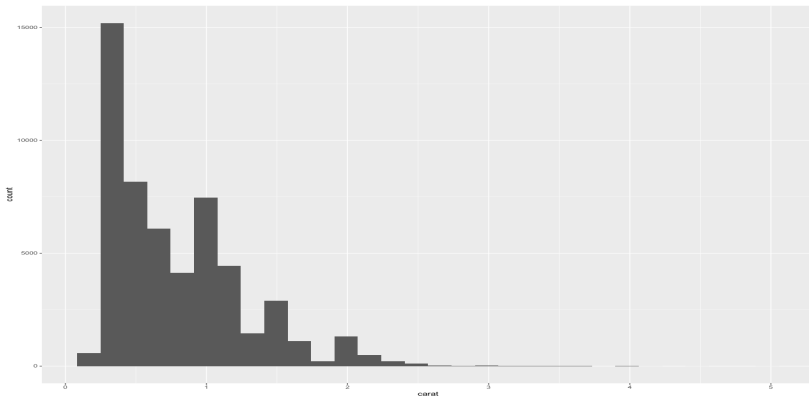
- Нормално разпределение
- Биномно разпределение
- Поасоново разпределение
- Други разпределения

4 Заключение

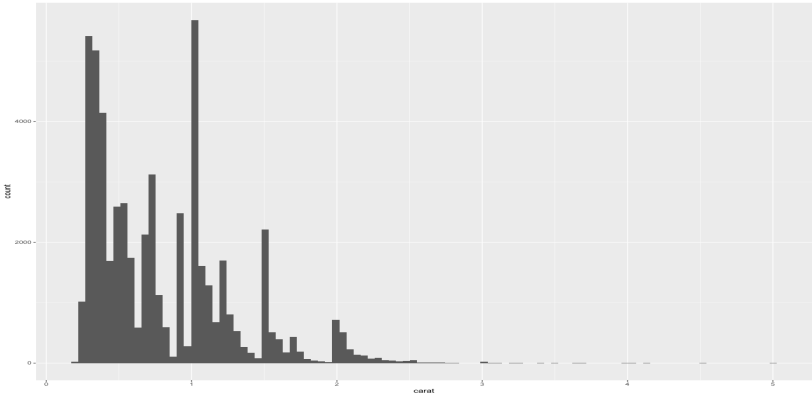
- Дискусия

Разширени графични възможности

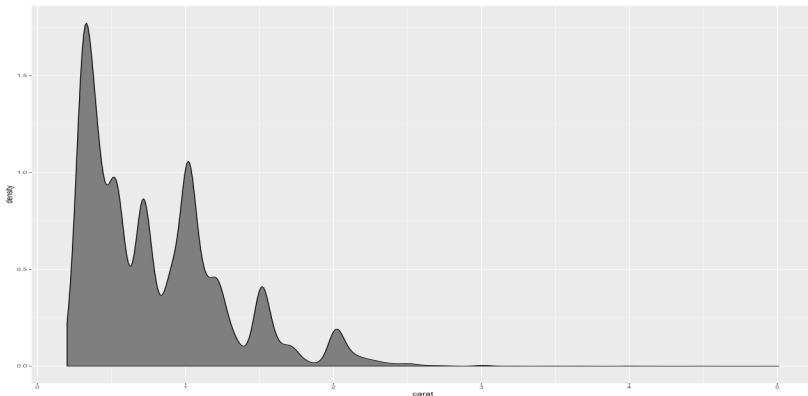
Хистограма при 30 групи



Хистограма при 100 групи



Плътностна функция

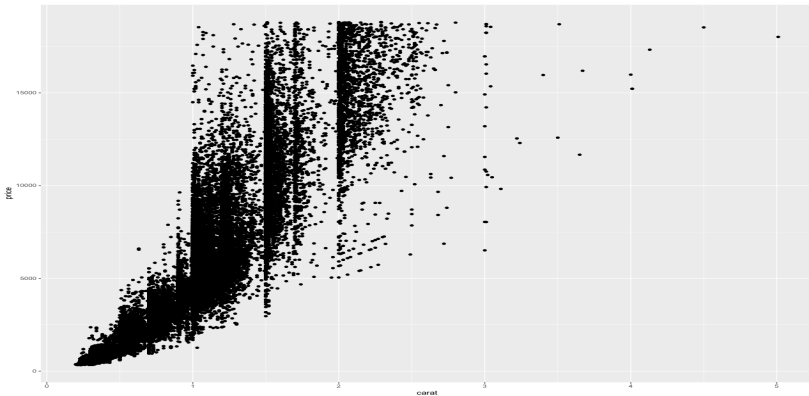


Визуализация при две променливи

Диаграма на разсейване с ggplot2

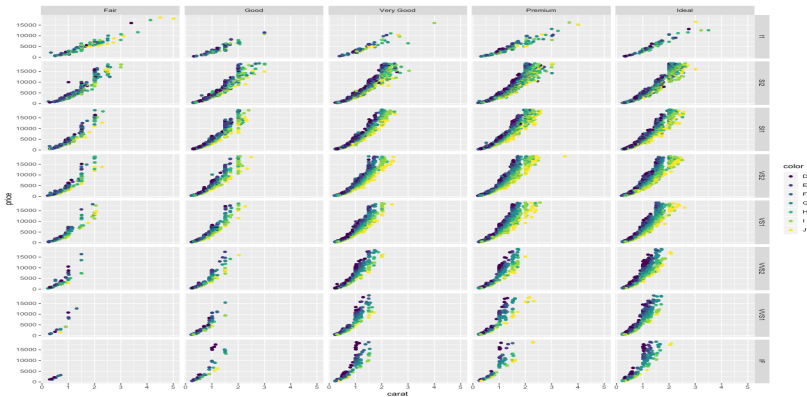
```
library( ggplot2 )
ggplot(diamonds, aes(x=carat, y=price)) + geom_point()
```

Отношение на цена към тегло



Диаграми на разсейване

Визуализация с групиране по два признака



Визуализация на хистограми с групиране

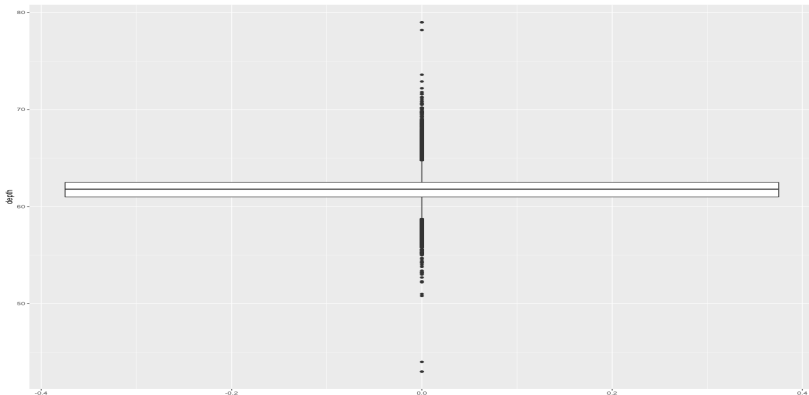


Изобразяване на квантили

Визуализация тип кутия

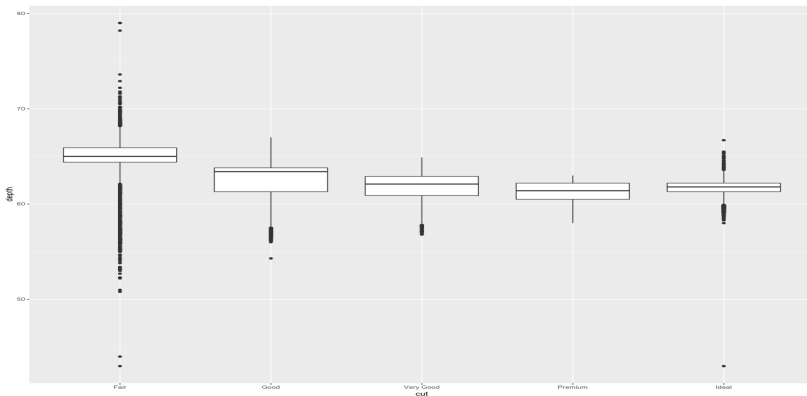
```
ggplot(diamonds, aes(y=depth)) + geom_boxplot()
ggplot(diamonds, aes(y=depth, x=cut)) + geom_boxplot()
ggplot(diamonds, aes(y=depth, x=cut)) + geom_boxplot() +
geom_violin()
ggplot(diamonds, aes(y=depth, x=cut))+ geom_point() +
geom_violin()
```

Визуализация на характеристиката за дълбочина на диамантите



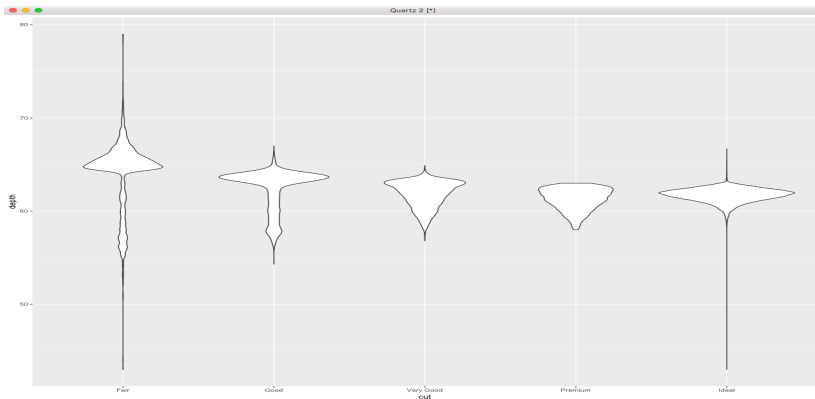
Графики тип кутия и цигулка

Дълбочина на диамантите в групи според сръза



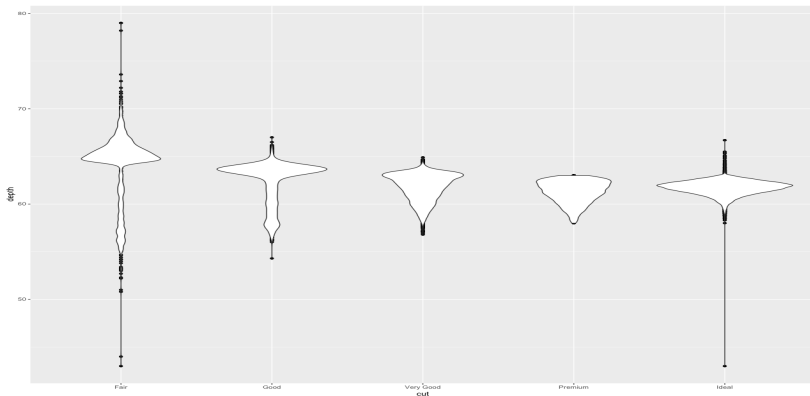
Графики тип кутия и цигулка

Графика тип цигулка



Графики тип кутия и цигулка

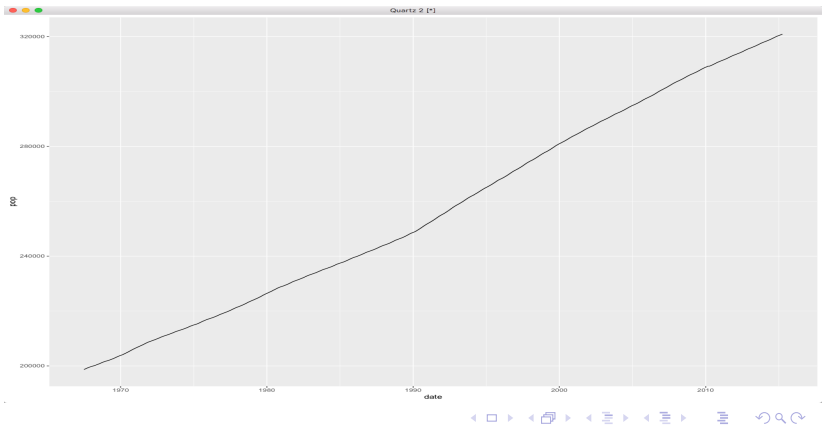
Добавяне на декорация с точки



Линейни графики

```
library( lubridate )
ggplot(economics, aes(x=date, y=pop)) + geom_line()
economics$year <- year( economics$date )
economics$month <- month(economics$date, label=TRUE)
library( scales )
ggplot(economics, aes(x=month, y=pop)) +
geom_line(aes(color=factor(year), group=year)) +
scale_color_discrete(name="Year") +
scale_y_continuous(labels=comma)+ labs(title="Population
Growth", x="Month", y="Population")
```

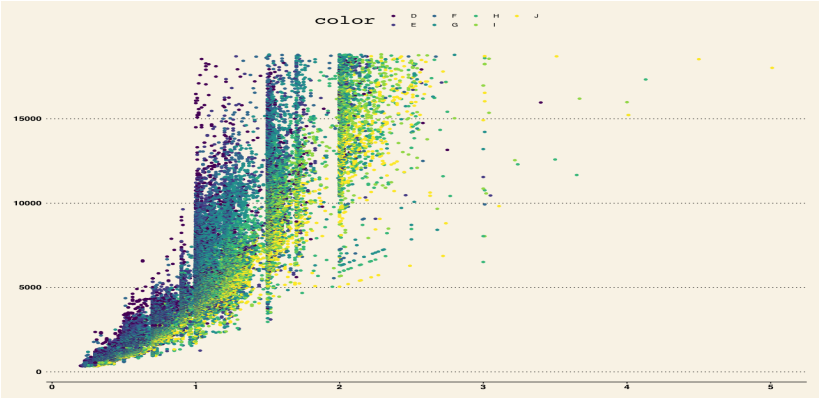
Нарастване на популацията във времето



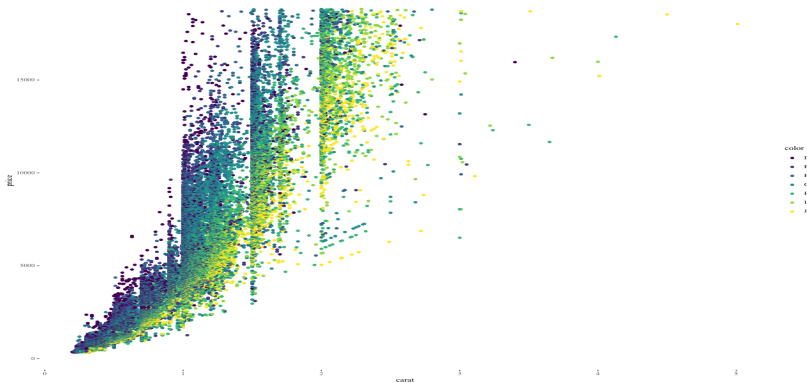
Избор на теми за визуално представяне

```
library( ggthemes )
ggplot(diamonds, aes(x=carat, y=price)) +
  geom_point(aes(color=color)) + theme_wsj()
ggplot(diamonds, aes(x=carat, y=price)) +
  geom_point(aes(color=color)) + theme_tufte()
ggplot(diamonds, aes(x=carat, y=price)) +
  geom_point(aes(color=color)) + theme_excel()
ggplot(diamonds, aes(x=carat, y=price)) +
  geom_point(aes(color=color)) + theme_economist() +
  scale_colour_economist()
```

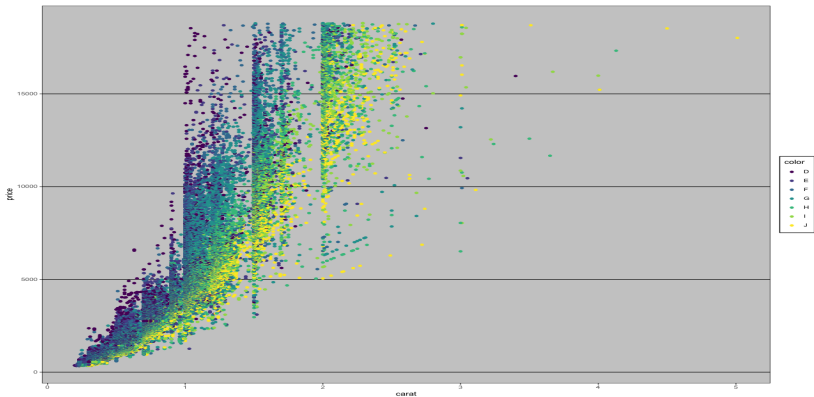
Тема Wall Street Journal



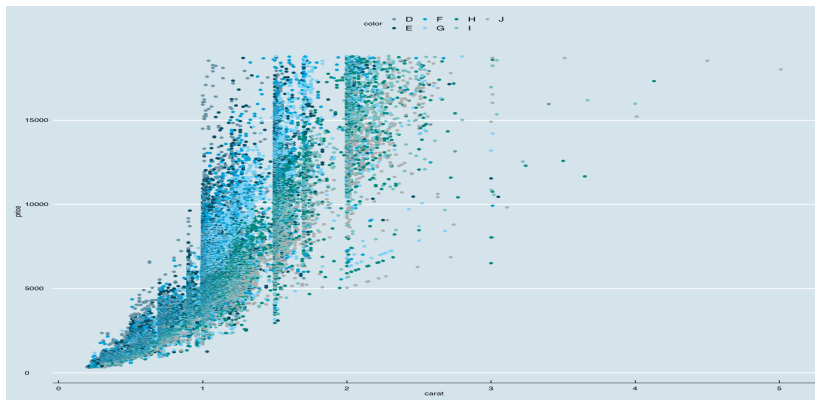
Tema Edward Tufte



Тема в стил Microsoft Excel



Tema Economist



Изследване на случайни величини

Случайни стойности от зарове

При два зара

```
x <- dice.roll(faces=6, dice=2, rolls=100000)
ggplot(data=x$results) +
  geom_histogram(aes(x=(die_1+die_2))) + ggtitle("Two dice
  rolled 100K times.") + xlab("Dice Sides") + ylab("Outcomes") +
  scale_x_continuous(breaks=round(seq(min((x$results$die_1 +
  x$results$die_2)),
  max((x$results$die_1+x$results$die_2)),by=0.5)))
```


Случайни стойности от зарове

При десет зара

```
x <- dice.roll(faces=6, dice=10, rolls=100000)
x$results$values = rowSums( x$results[,1:10] )
ggplot(data=x$results) + geom_density(aes(x=values)) +
ggtitle("Ten dice rolled 100K times.") + xlab("Dice Sides") +
ylab("Outcomes") +
scale_x_continuous(breaks=round(seq(min(x$results$values),
max(x$results$values),by=0.5))))
```


Вероятностни разпределения

Вероятностна функция на нормално разпределение

$$pdf(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \quad (1)$$

Генериране, плътност, кумулативна функция, квантили

Функции за работа с нормално разпределение

```
values <- rnorm(n=30000, mean=0, sd=0.85)
density <- dnorm( values )
cumulative <- pnorm( values )
quantile <- qnorm( cumulative )
```

Плътност, кумулативна функция, квантили

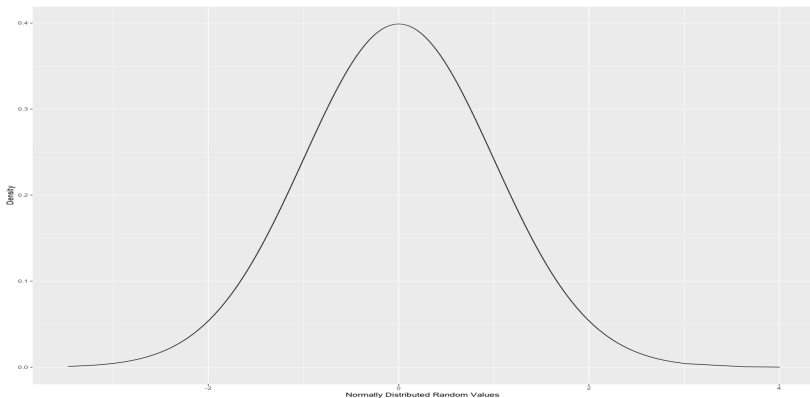
Визуализация на графики за нормално разпределение

```
ggplot(data.frame(x=values, y=density)) + aes(x=x, y=y) +  
geom_line() + labs(x="Normally Distributed Random Values",  
y="Density")
```

```
ggplot(data.frame(x=values, y=cumulative)) + aes(x=x, y=y) +  
geom_line() + labs(x="Normally Distributed Random Values",  
y="Cumulative Probability")
```

```
ggplot(data.frame(x=values, y=quantile)) + aes(x=x, y=y) +  
geom_line() + labs(x="Normally Distributed Random Values",  
y="Quantile")
```

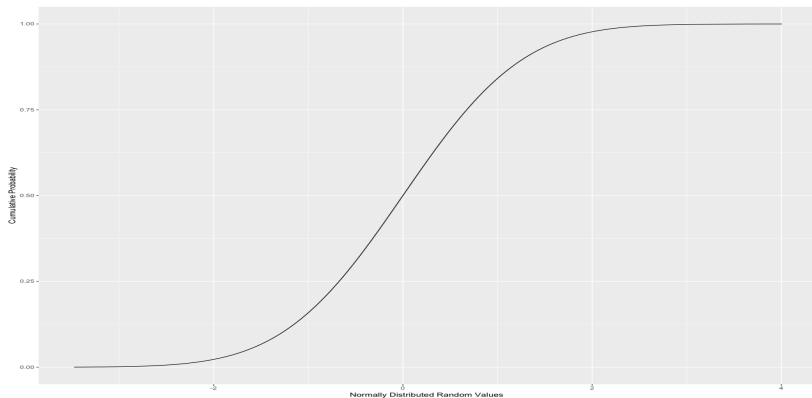

Плътностна функция на нормално разпределение



Кумулативна функция на нормално разпределение

$$cdf(x) = \int_{-\infty}^a \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} dx \quad (2)$$

Кумулативна функция на нормално разпределение



Вероятностна функция на биномно разпределение

$$pdf(x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (3)$$

Биномен коефициент

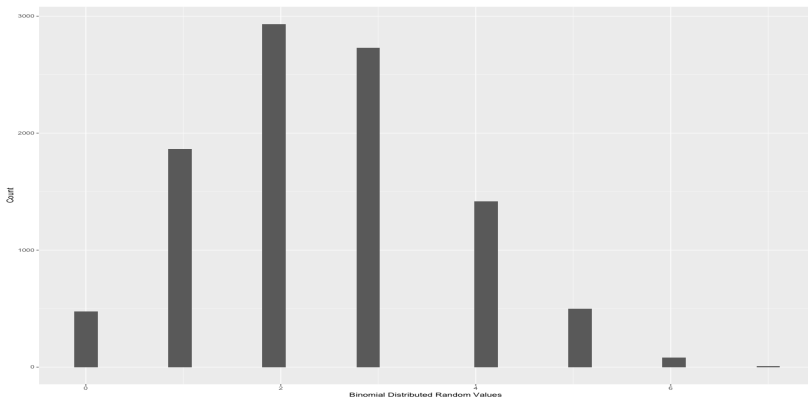
$$\binom{n}{x} = \frac{n!}{x!(n-x)!} \quad (4)$$

Плътност, кумулативна функция, квантили

Функции за работа с биномно разпределение

```
library( ggplot2 )
values <- rbinom(n=10000, size=7, prob=0.35)
density <- dbinom(x=2, size=7, prob=0.35)
cumulative <- pbinom(q=2, size=7, prob=0.35)
quantile <- qbinom(p=0.15, size=7, prob=0.35)
ggplot(data.frame(x=values)) + aes(x=x) + geom_histogram() +
labs(x="Binomial Distributed Random Values", y="Count")
```

Хистограма на биномно разпределение



Кумулативна функция на биномно разпределение

$$cdf(x) = \sum_{i=0}^a \binom{n}{i} p^i (1-p)^{n-i} \quad (5)$$

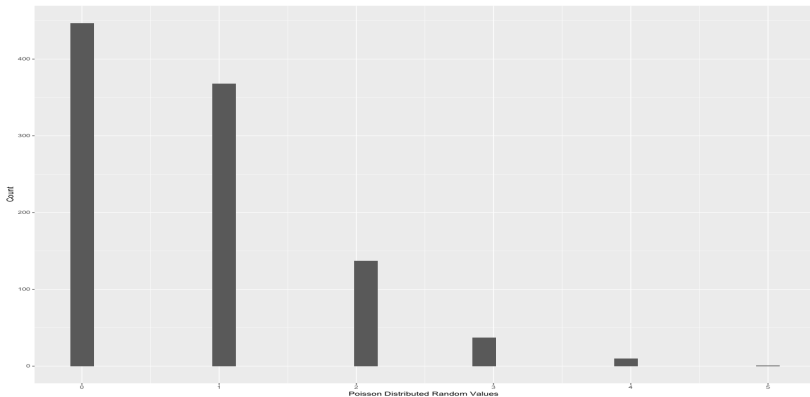
Вероятностна функция на поасоново разпределение

$$pdf(x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad (6)$$

Кумулативна функция на поасоново разпределение

$$cdf(x) = \sum_{i=0}^a \frac{\lambda^i e^{-\lambda}}{i!} \quad (7)$$

Хистограма на Поасоново разпределение



Списък с най-използваните вероятностни разпределения

Distribution	Functions			
Beta	pbeta	qbeta	dbeta	rbeta
Binomial	pbinom	qbinom	dbinom	rbinom
Cauchy	pcauchy	qcauchy	dcauchy	rcauchy
Chi-Square	pchisq	qchisq	dchisq	rchisq
Exponential	pexp	qexp	dexp	rexp
F	pf	qf	df	rf
Gamma	pgamma	qgamma	dgamma	rgamma
Geometric	pgeom	qgeom	dgeom	rgeom
Hypergeometric	phyper	qhyper	dhyper	rhyper
Logistic	plogis	qlogis	dlogis	rlogis
Log Normal	plnorm	qlnorm	dlnorm	rlnorm
Negative Binomial	pnbinom	qnbinom	dnbinom	rnbinom
Normal	pnorm	qnorm	dnorm	rnorm
Poisson	ppois	qpois	dpois	rpois
Student t	pt	qt	dt	rt
Studentized Range	ptukey	qtukey	dtukey	rtukey
Uniform	punif	qunif	dunif	runif
Weibull	pweibull	qweibull	dweibull	rweibull
Wilcoxon Rank Sum Statistic	pwilcox	qwilcox	dwilcox	rwilcox
Wilcoxon Signed Rank Statistic	psignrank	qsignrank	dsignrank	rsignrank

Заклучение

Въпроси и отговори

Благодаря за вниманието!