# Ch3概率与分布

本章概要

**♢ 随机抽样**

**♢ 排列组合与概率计算**

*♢***常用概率分布及其数字特征**

*♢* **R中内嵌的统计分布函数**

**3.1随机抽样**

* **函数sample(x, n, replace = FALSE, prob = NULL)：**从x的元素中抽取指定数量的一个随机样本
* 说明：x为要抽取的向量，n 为样本容量，replace = FALSE表示默认的是不放回的抽样，prob = NULL表示默认的是等概率抽样，不等概率是通过设定prob=y （y与x等长度）用于指定x中元素出现的概率。

（1）等可能的不放回随机抽样：sample(x,n)

例1：从52张扑克牌中抽取4张

> sample(1:52,4)

[1] 24 38 40 17

（2）等可能的放回随机抽样：sample(x,n,replace=TRUE)

例2：抛一枚均匀的硬币10次，

> sample(c("正","反"),10,replace=TRUE)

[1] "正" "正" "反" "正" "正" "反" "反" "反" "正" "反"

例3：掷一枚均匀的骰子10次

> sample(1:6,10,replace=TRUE)

[1] 2 2 3 1 3 4 5 2 6 3

(3)不等可能的随机抽样：sample(x,n,replace=TRUE,prob=y)

例4：一名外科医生做手术成功的概率为0.9，失败的概率为0.1，那么他做10次手术

> sample(c("成功","失败"),10,replace=TRUE,prob=c(0.9,0.1))

[1] "成功" "成功" "失败" "成功" "成功" "成功" "成功" "成功" "成功" "成功"

**3.2排列组合与概率计算**

* **函数prod(..., na.rm = FALSE)：返回所给向量元素的累乘积**
* **函数choose(n,k):计算*n(n-1)…(n-k+1) / k!***

例**1** 从一副完全打乱的52 张扑克中取4 张, 求以下事件的概率:

1) 抽取的4 张依次为红心A，方块A，黑桃A 和梅花A 的概率;

2) 抽取的4 张为红心A，方块A，黑桃A 和梅花A 的概率.

解:1) 抽取的4 张是有次序的, 因此使用排列来求解. 所求的事件(记为A) 概率为

> 1/prod(52:49)

[1] 1.539077e-07

2)抽取的4 张是没有次序的, 因此使用组合数来求解. 所求的事件(记为B) 概率为

> 1/choose(52,4)

[1] 3.693785e-06

**3.3概率分布（P72）：**概率知识的汇总，略讲

掌握常见分布符号：

**均匀分布: unif(*a，b*)**

**正态分布/高斯分布: norm(*μ，σ*2)**

**卡方(*χ*2) 分布: chisq(*n*)**

**t分布: t(*n*)**

**F分布: f(*n,，m*)**

**3.4R中内嵌的统计分布函数**

表中分别列出了R中内嵌的概率分布的英文名称、R 中的函数形式、函数中对应于分布的参数及函数所在的程序包名称。

**表3.1: R 中的概率分布及对应的R函数（教材P79）**

|  |  |  |  |
| --- | --- | --- | --- |
| 分布名称 | **R** 函数 | 参数 | 程序包 |
| Beta | \_beta | shape1, shape2 | stats |
| Binomial stats | \_binom | size, prob | stats |
| Cauchy | \_cauchy | location=0, scale=1 | stats |
| Uniform | \_unif | min=0, max=1 | stats |
| Normal | \_norm | mean=0, sd=1 | stats |
| Chi-sqaured (*\_*2) | \_chisq | df | stats |
| Student's (*t*) | \_t | df | stats |
| Fisher{Snedecor (*F*) | \_f | df1, df2 | stats |
| 。。。。。。 |  |  |  |

注：无特别申明，通常所说的卡方分布，t分布和F分布都是中心的卡方分布，t分布和F分布。所以通常ncp可以不用。

说明：

* “\_\_”表示概率分布名称前要加的前缀，有四个：
  1. 加前缀“d”（代表密度函数， density) 就得到**R** 的密度函数(对于离散分布，指分布律)

如：标准正态分布的密度函数

在R中对应dnorm()

* 1. 加前缀“p”(代表（累积）分布函数或概率，CDF) 就得到**R** 的分布函数

如：标准正态分布的（累积）分布函数

在R中对应pnorm()

* 1. 加前缀“q”(代表分位函数，quantile) 就得到**R** 的分位数函数;

如：标准正态分布的а=0.95分位数为1.96

在R中对应qnorm()

* 1. 加前缀“r”(代表随机模拟，random) 就得到**R** 的随机数发生函数

如:标准正态分布的随机数rnorm()

* 四类函数的调用格式:

1) 概率密度函数: dfunc(x, p1, p2, ...), x为数值向量

2) (累积) 分布函数: pfunc(q, p1, p2, ...), q为数值向量

3) 分位数函数: qfunc(p, p1, p2, ...), p为由概率构成的向量;

4) 随机数函数: rfunc(n, p1, p2, ...), n为生成数据的个数

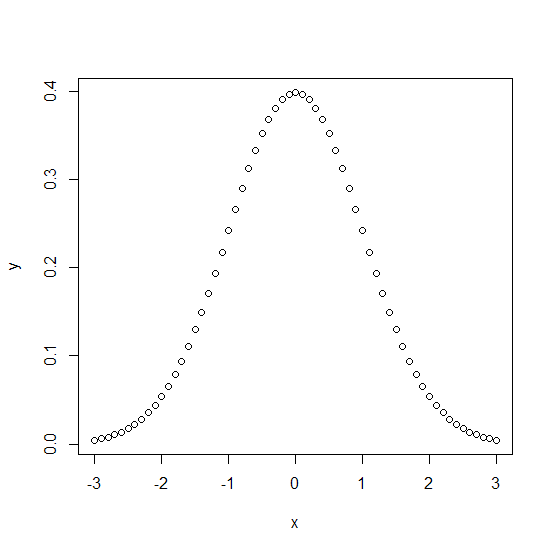
练习：用plot函数作正态分布的密度函数图形

x<-seq(-3,3,0.1);u=0;sigma=1

y<-exp(-(x-u)^2/(2\*sigma^2))/(sqrt(2\*pi)\*sigma)

data.frame(x=x,y=y)

plot(x,y) #正态分布的密度函数图形



例1：计算标准正态分布的x=0处的密度函数和分布函数

> dnorm(0) #尝试:标准正态分布改为N(2,4)后的结果 dnorm(0,mean=2,sd=2)

[1] 0.3989423

> pnorm(0)

[1] 0.5

**例2：计算标准正态a=0.5和0.05的分位数**

> qnorm(0.5)

[1] 0

> qnorm(0.05) #结果说明什么？分位数计算的是下侧（左侧）分位数

[1] -1.644854

**例3：产生100个标准正态分布的随机数，并计算其均值**

mean(rnorm(100)) #尝试:标准正态分布改为N(2,4)后的结果 mean(rnorm(100,mean=2,sd=2))

**3.5应用：中心极限定理**

**3.5.1** 中心极限定理

正态分布在概率统计中起着至关重要的作用，其中的一个原因是当独立观察(试验) 的样本容量*n* 足够大时, 那么所观察的随机变量*X*1*,X*2*, · · · ,Xn* 的和近似服从正态分布(假定*E*(*Xi*) = *μ, V ar*(*Xi*) = *σ*2存在), 即

**或**

**3.5.2** 渐近正态性的图形检验

limite.central<-function(r=runif,distpar=c(0,1),m=0.5,

s=1/sqrt(12),n=c(1,3,10,30),N=1000){

for (i in n){

if (length(distpar)==2) {

x<-matrix(r(i\*N,distpar[1],distpar[2]),nc=i)

}

else {

x<-matrix(r(i\*N,distpar),nc=i)

}

x<-(apply(x,1,sum)-i\*m)/(sqrt(i)\*s)

hist(x,col='light blue',probability=T,main=paste("n=",i),ylim=c(0,max(0.4,density(x)$y)))

lines(density(x),col='red',lwd=3)

curve(dnorm(x),col='blue',lwd=3,lt=3,add=T)

if (N>100) {

rug(sample(x,100))

}

else {

rug(x)

}

}

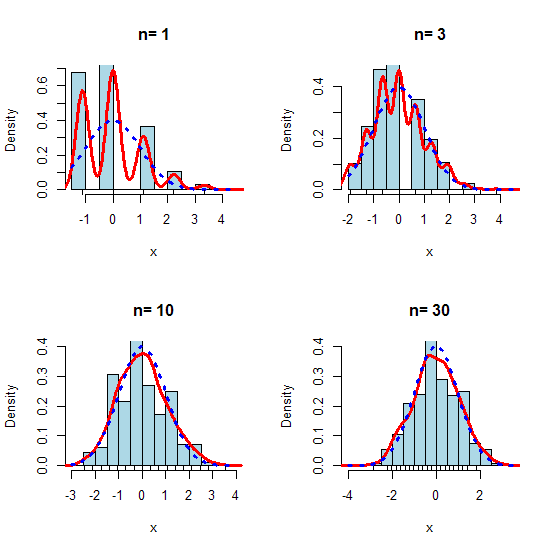
}

#二项分布

op<-par(mfrow=c(2,2))

limite.central(rbinom,distpar=c(10,0.1),m=1,s=0.9)

par(op)



本节学习的R函数小结

|  |  |  |
| --- | --- | --- |
| 第3章 概率与分布所介绍的主要函数 | | |
| 函数 | 功能 | 示例 |
|  |  |  |
|  |  |  |

**作业：练习本节所有命令**

* **3.1 3.2，3.3，3.5**