

MATH 203: PRINCIPLES OF STATISTICS 1

FALL 2024

THE GEOMETRIC AND POISSON DISTRIBUTIONS

Geometric Distribution:

The Geometric distribution is a probability distribution for a random variable derived from sequences of *binary trials*. As for the binomial distribution, each trial has ‘success probability’ p , that is, the probability that the outcome is 1 is p , and the probability that the outcome is 0 is $q = 1 - p$,

Experimental set up: We consider the experiment involving a sequence of identical binary trials with success probability p whose outcomes are independent. The experiment terminates when a success is observed. For example, we might have five experimental results of results

000010011000100000001

corresponding to the experiment terminating after 5, 3, 1, 4 and 8 trials respectively.

Random variable: To define the random variable, Y , we record the **number of trials** until the first success is obtained. If $0 < p < 1$, the set of possible values Y can take is

$$\mathcal{Y} = \{1, 2, 3, \dots\}.$$

with no upper limit on the possible values. All of these values will carry some positive probability.

Probability distribution: This definition for Y allows us to compute the probability that should be attached to each of the possible values that Y can take. We consider events of the form $(Y = y)$ for numerical values $y = 1, 2, 3, \dots$, and observe that

$$(Y = y) \quad \text{means} \quad \text{“first successes on trial } y\text{”}$$

In the examples above, the probability associated with each sequence is

00001	$qqqqp$
001	qqp
1	p
0001	$qqqp$
00000001	$qqqqqqqp$.

Therefore we have that the probability mass function (pmf) for the Geometric distribution is

$$p(y) = P(Y = y) = \begin{cases} q^{y-1}p & y \in \{1, 2, 3, \dots\} \\ 0 & \text{otherwise} \end{cases}$$

We write

$$Y \sim \text{Geometric}(p).$$

Numerical calculations: Numerical calculations for the Geometric distribution can be performed in R. For $p = 0.4$,

```
p<-0.4
q<-1-p
y<-3
prob<-q^{y-1}*p
print(paste('P(Y=3) = ',prob))

## [1] "P(Y=3) = 0.144"
```

The calculation can also be performed using the specially written function `dgeom`

```
print(paste('P(Y=3) = ',dgeom(y-1,p)))

## [1] "P(Y=3) = 0.144"
```

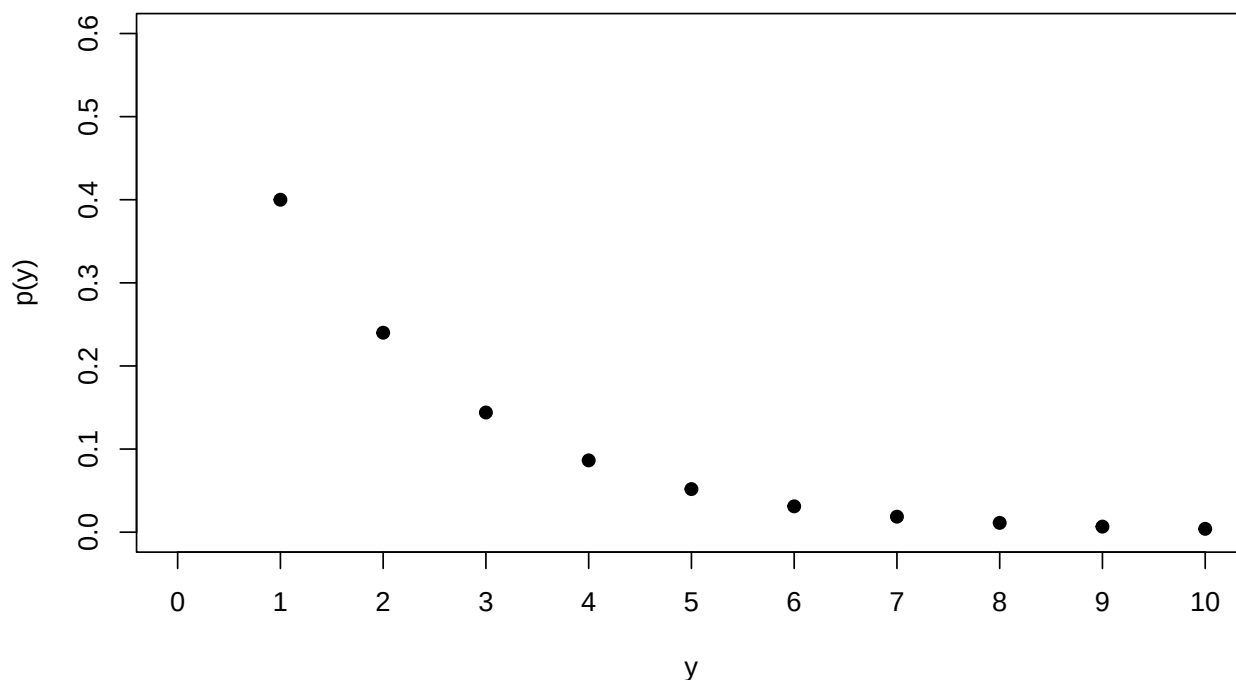
Note that you need to plug $y - 1$ (not y) into the function, due to the way the `dgeom` function is written. We can compute the complete list of values for the pmf evaluated at $y = 1, 2, \dots, 10$.

```
y<-1:10; probs<-dgeom(y-1,p)
print(cbind(y,probs))

##           y      probs
## [1,]  1 0.4000000000
## [2,]  2 0.2400000000
## [3,]  3 0.1440000000
## [4,]  4 0.0864000000
## [5,]  5 0.0518400000
## [6,]  6 0.0311040000
## [7,]  7 0.0186624000
## [8,]  8 0.0111974400
## [9,]  9 0.0067184640
## [10,] 10 0.0040310780
```

We can plot the pmf as a function of y .

```
par(mar=c(4,4,2,2))
plot(y,probs,pch=19,ylab='p(y)',xlim=range(0,10),ylim=range(0,0.6),xaxt='n')
axis(1,at=0:10)
```



Connection to samples: We can simulate data mimicking the experiment. For example, we can generate a matrix of sequences of binary trials, and then compute Y for each sequence; here for $N = 5$ sequences

```
set.seed(3)
N<-5
Yseq<-list()
for(i in 1:5){
  success<-FALSE
  bseq<-as.character()
  while(!success){
    X<-rbinom(1,1,p)
    bseq<-paste(bseq,X,sep='')
    if(X == 1) success<-TRUE
  }
  Yseq[[i]]<-bseq
  print(bseq)
}

## [1] "01"
## [1] "001"
## [1] "1"
## [1] "0001"
## [1] "00001"

Y<-as.numeric(lapply(Yseq,nchar))
Y

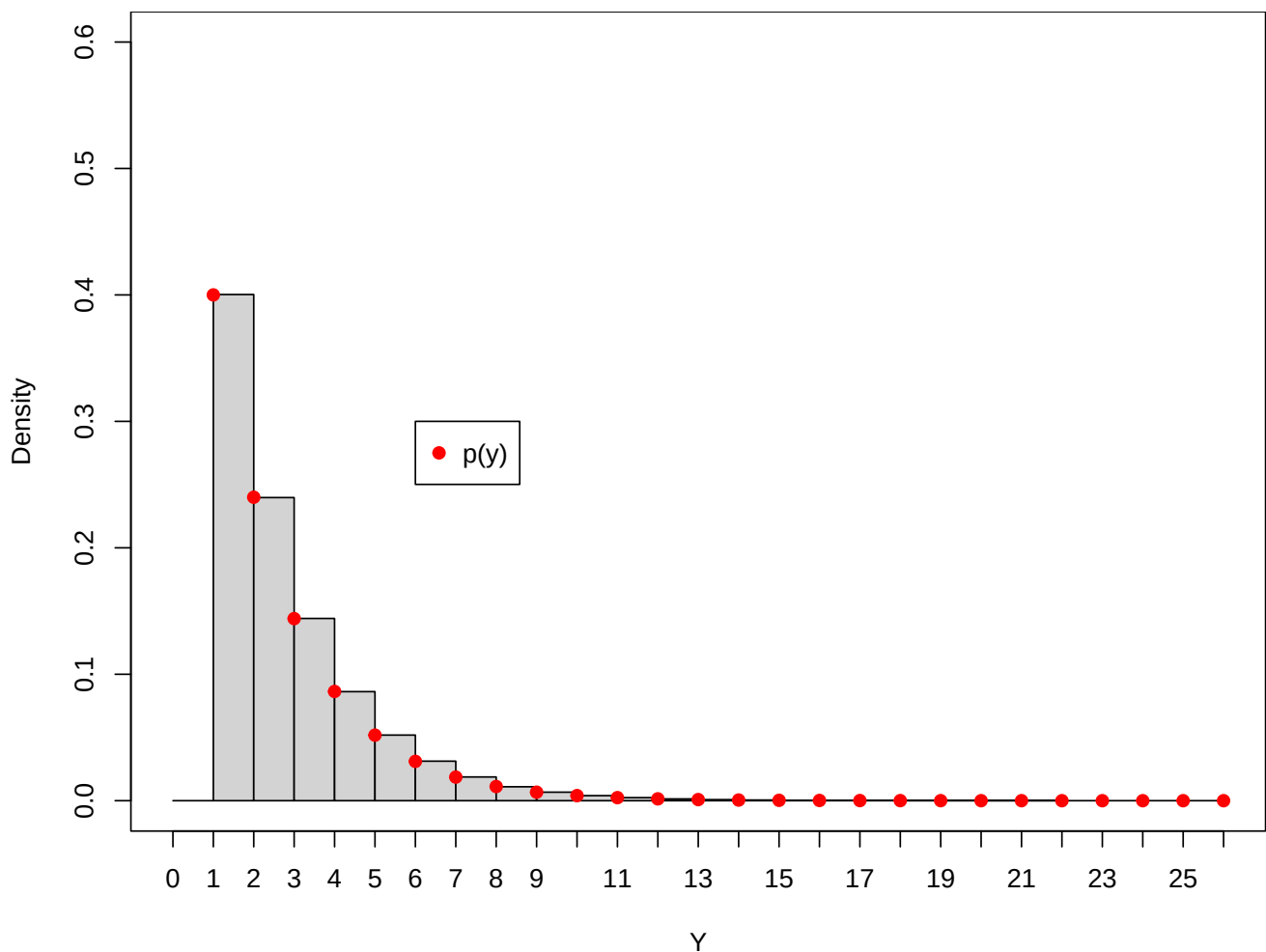
## [1] 2 3 1 4 5
```

Hence the simulated Y values are 2, 3, 1, 4, 5. We can also simulate Y directly using `rgeom`. Here are $N = 10$ values

```
N<-10;rgeom(N,p)+1
## [1] 1 3 1 9 2 1 2 3 2 1
```

Suppose we simulate a very large sample $N = 1000000$, and draw a density histogram

```
N<-1000000
Y<-rgeom(N,p)+1
par(mar=c(4,4,0,0))
hist(Y,breaks=0:max(Y),right=FALSE,include.lowest=TRUE,main='',
      xaxt='n',freq=FALSE,ylim=range(0,0.6))
axis(1,at=c(0:max(Y)))
box()
y<-1:max(Y)
probs<-dgeom(y-1,p)
points(y,probs,pch=19,col='red')
legend(6,0.3,c('p(y)'),pch=19,col='red')
```



The pmf values are superimposed in red. The pmf can therefore be interpreted as the limiting values of the density histogram heights in a very large sample.

In the sample of 10^6 from the *Geometric*(0.4) distribution, we have the following counts of the different possible values.

```
table(Y)

## Y
##      1      2      3      4      5      6      7      8      9     10     11
## 400313 239728 144124 86252 51891 31257 18756 11025 6728 4000 2400
##      12     13     14     15     16     17     18     19     20     21     22
##  1399    838    528    295    187    108    64     48     21     16      8
##      23     24     25     26
##       7      3      2      2
```

The relative frequencies are

```
table(Y)/N

## Y
##      1      2      3      4      5      6      7      8
## 0.400313 0.239728 0.144124 0.086252 0.051891 0.031257 0.018756 0.011025
##      9     10     11     12     13     14     15     16
## 0.006728 0.004000 0.002400 0.001399 0.000838 0.000528 0.000295 0.000187
##      17     18     19     20     21     22     23     24
## 0.000108 0.000064 0.000048 0.000021 0.000016 0.000008 0.000007 0.000003
##      25     26
## 0.000002 0.000002
```

which closely match the pmf values

```
round(cbind(table(Y)/N,probs),6)

##      probs
## 1  0.400313 0.400000
## 2  0.239728 0.240000
## 3  0.144124 0.144000
## 4  0.086252 0.086400
## 5  0.051891 0.051840
## 6  0.031257 0.031104
## 7  0.018756 0.018662
## 8  0.011025 0.011197
## 9  0.006728 0.006718
## 10 0.004000 0.004031
## 11 0.002400 0.002419
## 12 0.001399 0.001451
## 13 0.000838 0.000871
## 14 0.000528 0.000522
## 15 0.000295 0.000313
## 16 0.000187 0.000188
## 17 0.000108 0.000113
## 18 0.000064 0.000068
## 19 0.000048 0.000041
## 20 0.000021 0.000024
```

```
## 21 0.000016 0.000015
## 22 0.000008 0.000009
## 23 0.000007 0.000005
## 24 0.000003 0.000003
## 25 0.000002 0.000002
## 26 0.000002 0.000001
```

Expectation and Variance: If $Y \sim \text{Geometric}(p)$, then the (population) mean (or expectation) and variance are

$$E(Y) = \frac{1}{p} \quad \text{Var}(Y) = \frac{(1-p)}{p^2} = \frac{q}{p^2}$$

```
c(1/p,q/p^2)    #Expectation, Variance
```

```
## [1] 2.50 3.75
```

That is, the expectation is 2.5 and the variance is 3.75. These values describe aspects of the distribution, and the function $p(y)$. In the very large sample of $N = 100000$ values, we have sample mean and sample variance which are very close to the population expectation and variance

```
mean(Y)
```

```
## [1] 2.498248
```

```
var(Y)
```

```
## [1] 3.735123
```

The Poisson distribution:

The Poisson distribution arises as a limiting case of the Binomial distribution when n is large and p is small.

Experimental set up: We consider the experiment involving a sequence of n identical binary trials with success probability p whose outcomes are independent. Suppose that n is large and p is small, and let $\lambda = np$. We consider the limiting case where $n \rightarrow \infty$ and $p \rightarrow 0$ but where λ is held fixed: that is, we fix λ to be some positive value, and let

$$p = \frac{\lambda}{n}.$$

Random variable: We record the **total number** of successes in the sequence. The set of possible values Y can take is

$$\mathcal{Y} = \{0, 1, 2, 3, \dots\}.$$

where n is so large that we might as well consider it infinite.

Probability distribution: We consider events of the form $(Y = y)$ for numerical values $y = 0, 1, 2, \dots$, and observe that

$$(Y = y) \quad \text{means} \quad "y \text{ successes}"$$

It can be shown that

$$p(y) = P(Y = y) = e^{-\lambda} \frac{\lambda^y}{y!} \quad y \in \{0, 1, 2, \dots\}$$

with $p(y) = 0$ otherwise. We write

$$Y \sim \text{Poisson}(\lambda).$$

Numerical calculations: Numerical calculations for the Poisson distribution can be performed in R. We use the settings $\lambda = 2$ and $n = 100000$

```
lambda<-2
n<-100000
p<-lambda/n
y<-3
prob<-choose(n,y)*(p^y)*((1-p)^(n-y)) #choose(.,.) computes C_n^r.
print(paste('P(Y=3) =',round(prob,6)))

## [1] "P(Y=3) = 0.180449"
```

The calculation can also be performed using `dbinom`

```
print(paste('P(Y=3) =',round(dbinom(y,n,p),6)))

## [1] "P(Y=3) = 0.180449"
```

In fact, there is a specially written function `dpois` to deal with this case

```
print(paste('P(Y=3) =',round(dpois(y,lambda),6)))
```

```
## [1] "P(Y=3) = 0.180447"
```

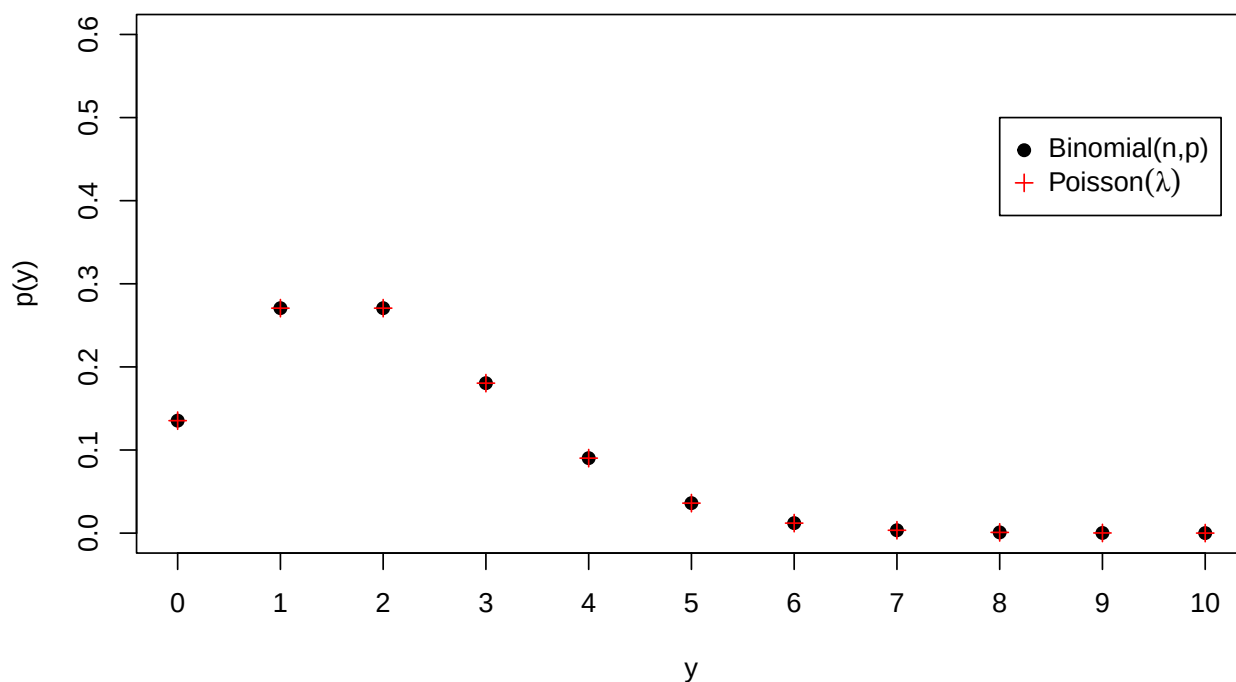
We can compute the complete list of values for the pmf evaluated at $y = 0, 1, 2, \dots, 10$.

```
y<-0:10; probs<-dbinom(y,n,p); probsPois<-dpois(y,lambda)
print(cbind(y,round(probs,6), round(probsPois,6)))
```

```
##      y
## [1,] 0 0.135333 0.135335
## [2,] 1 0.270671 0.270671
## [3,] 2 0.270673 0.270671
## [4,] 3 0.180449 0.180447
## [5,] 4 0.090224 0.090224
## [6,] 5 0.036089 0.036089
## [7,] 6 0.012029 0.012030
## [8,] 7 0.003437 0.003437
## [9,] 8 0.000859 0.000859
## [10,] 9 0.000191 0.000191
## [11,] 10 0.000038 0.000038
```

We can plot the pmf as a function of y .

```
par(mar=c(4,4,2,2))
plot(y,probs,pch=19,ylab='p(y)',ylim=range(0,0.6),xaxt='n');axis(1,at=0:10)
points(y,probsPois,pch=3,col='red')
legend(8,0.5,c('Binomial(n,p)',expression(Poisson(lambda))),
      pch=c(19,3),col=c('black','red'))
```

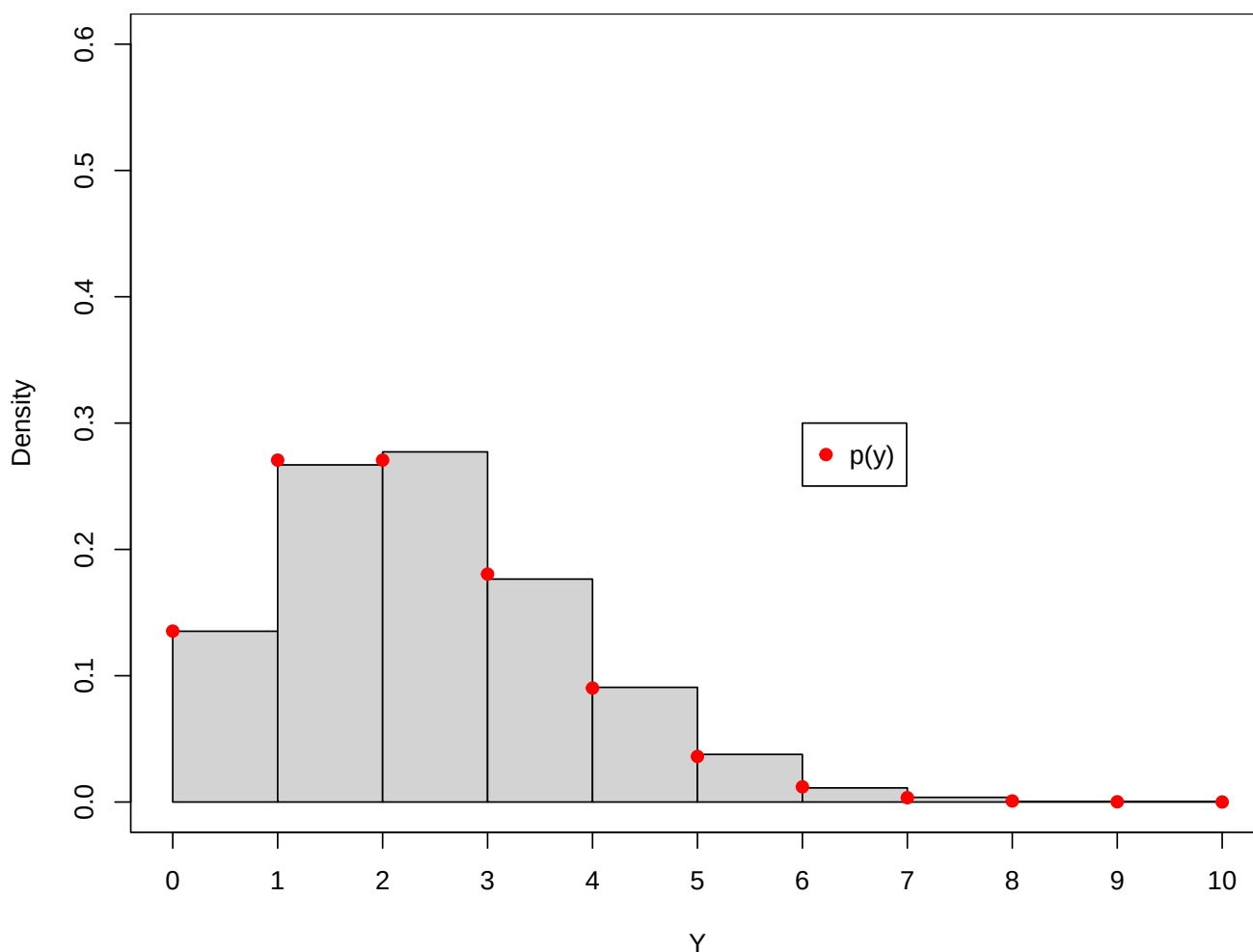


Connection to samples: We can also Y directly using `rpois`. Here are $N = 10$ values

```
N<-10;rpois(N,lambda)
## [1] 2 2 0 0 2 2 2 2 1 3
```

Suppose we simulate a very large sample $N = 10000$, and draw a density histogram

```
N<-10000
Y<-rpois(N,lambda)
par(mar=c(4,4,0,0))
hist(Y,breaks=0:10,right=FALSE,include.lowest=TRUE,main='',
      xaxt='n',freq=FALSE,ylim=range(0,0.6))
axis(1,at=c(0:10))
box()
points(y,probs,pch=19,col='red')
legend(6,0.3,c('p(y)'),pch=19,col='red')
```



The pmf values are superimposed in red. The pmf can again be interpreted as the limiting values of the density histogram heights in a very large sample.

In the sample of 10^4 from the $Poisson(2)$ distribution, we have the following counts of the different possible values.

```
table(Y)

## Y
##   0    1    2    3    4    5    6    7    8    9   10
## 1352 2669 2772 1765  908  377 112  37   5   2   1
```

The relative frequencies are

```
table(Y)/N

## Y
##   0    1    2    3    4    5    6    7    8    9   10
## 0.1352 0.2669 0.2772 0.1765 0.0908 0.0377 0.0112 0.0037 0.0005 0.0002 0.0001
```

which closely match the pmf values

```
round(cbind(table(Y)/N, probsPois), 6)

##           probsPois
## 0  0.1352  0.135335
## 1  0.2669  0.270671
## 2  0.2772  0.270671
## 3  0.1765  0.180447
## 4  0.0908  0.090224
## 5  0.0377  0.036089
## 6  0.0112  0.012030
## 7  0.0037  0.003437
## 8  0.0005  0.000859
## 9  0.0002  0.000191
## 10 0.0001  0.000038
```

Expectation and Variance: If $Y \sim \text{Poisson}(\lambda)$, then the (population) mean (or expectation) and variance are

$$E(Y) = \lambda \quad \text{Var}(Y) = \lambda$$

That is, the expectation and variance are both 2. In the very large sample of $N = 100000$ values, we have sample mean and sample variance which are very close to the population expectation and variance

```
mean(Y)

## [1] 2.0024

var(Y)

## [1] 1.992794
```