

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024
THE BINOMIAL DISTRIBUTION

The Binomial distribution is a probability distribution for a random variable derived from sequences of *binary trials*, that is, experiments that result in a numerical outcome recorded as 0 or 1. We refer to ‘successes’ (outcome is 1) and ‘failures’ (outcome is 0). Each trial has ‘success probability’ p , a numerical value between zero and one, 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 n identical binary trials with success probability p whose outcomes are independent. Let A_i represent the event that the i th trial results in a success. One possible sequence for $n = 8$ is

01010111

corresponding to the event

$$A_1^c \cap A_2 \cap A_3^c \cap A_4 \cap A_5^c \cap A_6 \cap A_7 \cap A_8$$

which under independence of the individual trials has probability

$$(1 - p) \times p \times (1 - p) \times p \times (1 - p) \times p \times p \times p = p^5(1 - p)^3 = p^5q^3.$$

There are $2^8 = 256$ binary sequences of length eight in total.

Random variable: To define the random variable, Y , we record the **total number** of successes in the sequence. If $0 < p < 1$, the set of possible values Y can take is

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

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 $y = 0, 1, 2, \dots, n$, and observe that

$$(Y = y) \quad \text{means} \quad “y \text{ successes in } n \text{ trials}”$$

Consider a sequence that results in y successes and $n - y$ failures: for example, if $n = 8$ and $y = 3$, we might have

10101000
01010001
11100000

and so on. All such sequences are equally likely under the ‘identical and independent’ assumption. The probability of any such sequence is p^yq^{n-y} , for $y \in \{0, 1, 2, \dots, n\}$; that is, in the examples above

10101000 $pqpqpqqq$
01010001 $qpqpqqqp$
11100000 $pppqqqqq$

etc. all yield p^3q^5 . For the **total** number of successes, we do not care about the order of the outcomes; we therefore must count the number of all such sequences, as the individual sequences partition the event of interest, which in the example is ($Y = 3$). By previous combinatorial arguments, we see that the number of such sequences is

$$C_y^n = \binom{n}{y} = \frac{n!}{y!(n-y)!} = \frac{8!}{3!5!} = 56.$$

Therefore we have that

$$P(Y = y) = \binom{n}{y} p^y q^{n-y} \quad y \in \{0, 1, 2, \dots, n\}$$

and $P(Y = y) = 0$ for all other y . We therefore have that the **probability mass function** (pmf) for the Binomial distribution is

$$p(y) = \binom{n}{y} p^y q^{n-y} \quad y \in \{0, 1, 2, \dots, n\}$$

with $p(y) = 0$ otherwise. The pmf is the mathematical function that describes how probability is distributed across the possible values of the random variable. We write

$$Y \sim \text{Binomial}(n, p).$$

Numerical calculations: Numerical calculations for the binomial distribution can be performed in R.

```
n<-8
p<-0.4
q<-1-p
y<-3
prob<-choose(n,y)*(p^y)*((1-p)^(n-y)) #choose(.,.) computes C_n^r.
print(paste('P(Y=3) =',prob))

## [1] "P(Y=3) = 0.27869184"
```

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

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

## [1] "P(Y=3) = 0.27869184"
```

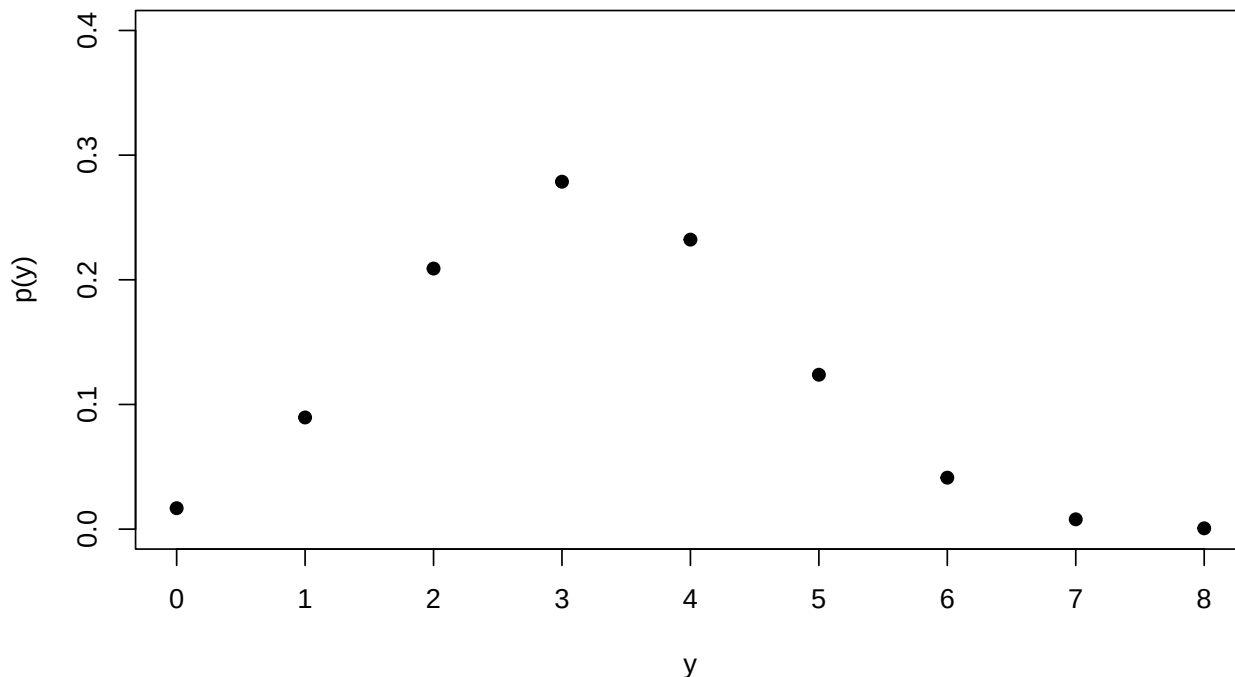
We can compute the complete list of values for the pmf evaluated at $y = 0, 1, 2, \dots, n$.

```
y<-0:n; probs<-dbinom(y,n,p)
print(cbind(y,probs))

##           y      probs
## [1,] 0 0.01679616
## [2,] 1 0.08957952
## [3,] 2 0.20901888
## [4,] 3 0.27869184
## [5,] 4 0.23224320
## [6,] 5 0.12386304
## [7,] 6 0.04128768
## [8,] 7 0.00786432
## [9,] 8 0.00065536
```

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(0,0.4),xaxt='n');axis(1,at=0:8)
```



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
X<-matrix(rbinom(n*N,1,p),nrow=5,ncol=n) #n*N is the total number of trials
rownames(X)<-paste('Sequence',1:5); colnames(X)<-paste(1:n)
Y<-rowSums(X)
cbind(X,Y)

##           1 2 3 4 5 6 7 8 Y
## Sequence 1 0 1 0 1 0 1 0 0 3
## Sequence 2 1 0 0 0 0 0 0 1 2
## Sequence 3 0 0 0 1 0 1 0 0 2
## Sequence 4 0 0 0 1 0 0 0 0 1
## Sequence 5 1 1 1 0 0 1 0 0 4
```

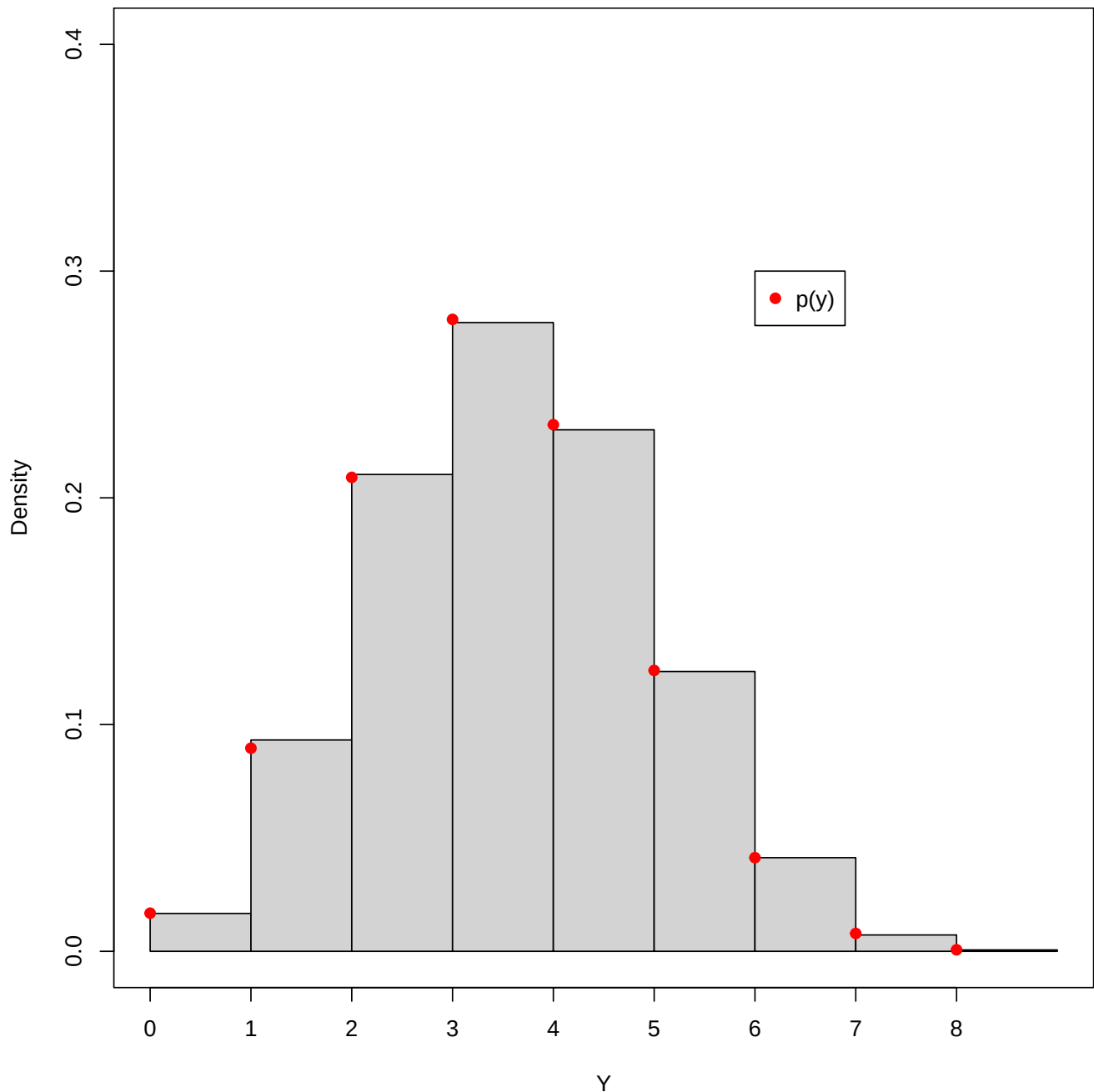
Hence the simulated Y values are 3, 2, 2, 1, 4. We can also simulate Y directly using `rbinom`. Here are $N = 10$ values

```
N<-10;rbinom(N,n,p)

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

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

```
N<-10000
Y<-rbinom(N,n,p)
par(mar=c(4,4,0,0))
hist(Y,breaks=0:9,right=FALSE,include.lowest=TRUE,main='',
      xaxt='n',freq=FALSE,ylim=range(0,0.4))
axis(1,at=c(0:8))
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 therefore be interpreted as the limiting values of the density histogram heights in a very large sample.

In the sample of 10^4 from the $Binomial(8, 0.4)$ distribution, we have the following counts of the different possible values.

```
table(Y)

## Y
##   0    1    2    3    4    5    6    7    8
## 167  932 2103 2773 2300 1234  413   72   6
```

The relative frequencies are

```
table(Y)/N

## Y
##   0    1    2    3    4    5    6    7    8
## 0.0167 0.0932 0.2103 0.2773 0.2300 0.1234 0.0413 0.0072 0.0006
```

which closely match the pmf values

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

##           probs
## 0 0.0167 0.016796
## 1 0.0932 0.089580
## 2 0.2103 0.209019
## 3 0.2773 0.278692
## 4 0.2300 0.232243
## 5 0.1234 0.123863
## 6 0.0413 0.041288
## 7 0.0072 0.007864
## 8 0.0006 0.000655
```

Expectation and Variance: If $Y \sim Binomial(n, p)$, then the (population) mean (or expectation) and variance are

$$E(Y) = np \quad Var(Y) = npq$$

```
c(n*p, n*p*q)    #Expectation, Variance

## [1] 3.20 1.92
```

That is, the expectation is 3.2 and the variance is 1.92. 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] 3.1857

var(Y)

## [1] 1.924608
```