9. Named distributions

Principles of Data Science with R

Dr. Uma Ravat

PSTAT 10

we did:

- Random Variables: Discrete or Continuous
- Discrete Random variables (By hand and using R)
  - P.m.f
  - E(X)
  - V(X)
  - C.d.f

Next we will see...

Certain distributions are so common that they get their own name!

- Named distributions
  - Discrete uniform
  - Binomial Distribution
  - Uniform Distribution
  - Normal Distribution

Discrete uniform distribution

k	1	2	3	4	5	6
$P(X = k)$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$	$1/6$

$X \sim \mathrm{DUnif}(\{1, 2, \ldots, 6\})$

Read as X follows *discrete uniform distribution on 1, 2, . . . , 6*.

More generally, $X \sim \mathrm{DUnif}(\{1, 2, \ldots, n\})$ then $P(X = k) = 1/n$

Binomial distribution

Binomial Experimental setup:

1. There are a fixed number of trials (denoted by n)
2. These n trials are independent.
3. Each trial has two possible outcomes, 0 or 1. We call an outcome of 1 a *success*.
4. The probability of success for each trial is the same and is denoted by p. Correspondingly, the probability of a failure is denoted by $1 - p$.

If $X =$ the total number of successes in the n trials of a binomial experiment then X is a binomial r.v. with a binomial probability distribution written as

$$X \sim \text{Binom}(n, p).$$

n and p are the parameters of the Binomial distrubiton.

Consider the following situations:

1. Rolling a fair die 100 times and counting the number of times we got an even number.

Verify assumptions are valid: Number of rolls is fixed, each roll has two outcomes, each roll is an *independent* trial with the *same* success probability $p = 3/6$, i.e. the probability rolling an even.

2. A bag contains 10 black marbles, 1 red marble, and nothing else. I choose a marble uniformly at random. I repeat this 100 times and count the number of black marbles I got.

Verify assumptions are valid: Number of draws is fixed, each trial has two outcomes- black or red, each draw is an *independent* trial with the *same* success probability $p = 10/11$

These are totally different experiments, but they both satisfy all the four conditions of a Binomial experiment.

Their probability distribution can be modeled as Binomial distribution (with different parameters of course).

1. Rolling a fair die 100 times and counting the number of times we got an even number.

Their probability distribution can be modeled as Binomial distribution (with different parameters of course).

1. Rolling a fair die 100 times and counting the number of times we got an even number.

$X_1 =$ number of times an even number shows up in 100 rolls of a fair die; $X_1 \sim \mathrm{Binom}(n = 100, p = ?)$

Their probability distribution can be modeled as Binomial distribution (with different parameters of course).

1. Rolling a fair die 100 times and counting the number of times we got an even number.

$X_1 =$ number of times an even number shows up in 100 rolls of a fair die; $X_1 \sim \mathrm{Binom}(n = 100, p = ?)$

2. A bag contains 10 black marbles, 1 red marble, and nothing else. I choose a marble uniformly at random. I repeat this 100 times and count the number of black marbles I got.

Their probability distribution can be modeled as Binomial distribution (with different parameters of course).

1. Rolling a fair die 100 times and counting the number of times we got an even number.

X_1 = number of times an even number shows up in 100 rolls of a fair die; $X_1 \sim \mathrm{Binom}(n = 100, p = ?)$

2. A bag contains 10 black marbles, 1 red marble, and nothing else. I choose a marble uniformly at random. I repeat this 100 times and count the number of black marbles I got.

X_2 = number of black marbles drawn in 100 draws from the bag; $X_2 \sim \mathrm{Binom}(n = 100, p = ?)$

Recall

For discrete r.v.s, $P(X = k)$ is the *probability mass function* or **pmf** of X.

It is a function of k.

Probabilities for Binomial RVs

$X \sim \text{Binom}(n, p)$

The **Binomial distribution** describes the probability of having exactly k successes in n independent trials, where probability of success in each trial is p. Here $k = 0, 1, \ldots, n$

The p.m.f of X is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{1.2.3 \cdots n}{1.2.3 \cdots k (1.2.3 \cdots n-k)}$ gives the number of ways to choose k of the n outcomes to be successes.

We'll use the `choose(n,k)` function in R to calculate $\binom{n}{k}$

- **Mean** $\mu = np$
- **variance** $\sigma^2 = np(1-p)$; **SD** $\sigma = \sqrt{np(1-p)}$

Let $X \sim \text{Binom}(10, 1/4)$.

Determine the probability of getting 4 successes ie $P(X = 4)$.

```
n <- 10
p <- 1/4
k <- 4
choose(n, k) * p^k * (1-p)^(n-k)
```

```
## [1] 0.145998
```

Use R-built in function `dbinom` for binomial probability

```
dbinom(4, size = 10, prob = 1/4)
```

```
## [1] 0.145998
```

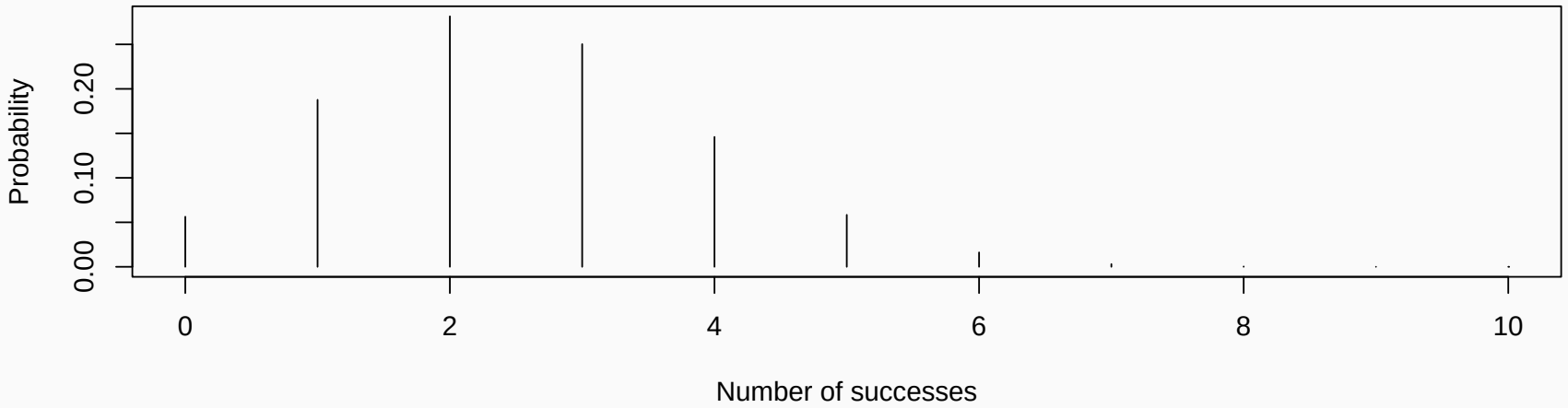
The syntax for the function is,

```
dbinom(x, size, prob, log = FALSE)
```

Let $X \sim \text{Binom}(10, 1/4)$, then the pmf of X is

```
##    0    1    2    3    4    5    6    7    8    9   10
## 0.06 0.19 0.28 0.25 0.15 0.06 0.02 0.00 0.00 0.00 0.00
```

```
plot(0:10, dbinom(0:10, 10, 1/4),
     xlab="Number of successes",
     ylab="Probability", type="h")
```



$E(X) = np = 10 * 0.25 = 2.5,$

$Var(X) = np(1-p) = 10 * 0.25 * 0.75 = 1.875,$ SD $= 1.37$

Typical values lie between 2.5 -1.37 $= 0.625$ to 2.5+1.3 $= 4.375$

Let $X \sim \text{Binom}(10, 1/4)$.

What's the probability of at most 4 successes? $P(X \leq 4) = ?$.

$$P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

```
# `dbinom` gives the pmf
sum(dbinom(0:4, size = 10, prob = 1/4))
```

```
## [1] 0.9218731
```

```
pbinom(4, size = 10, prob = 1/4)
```

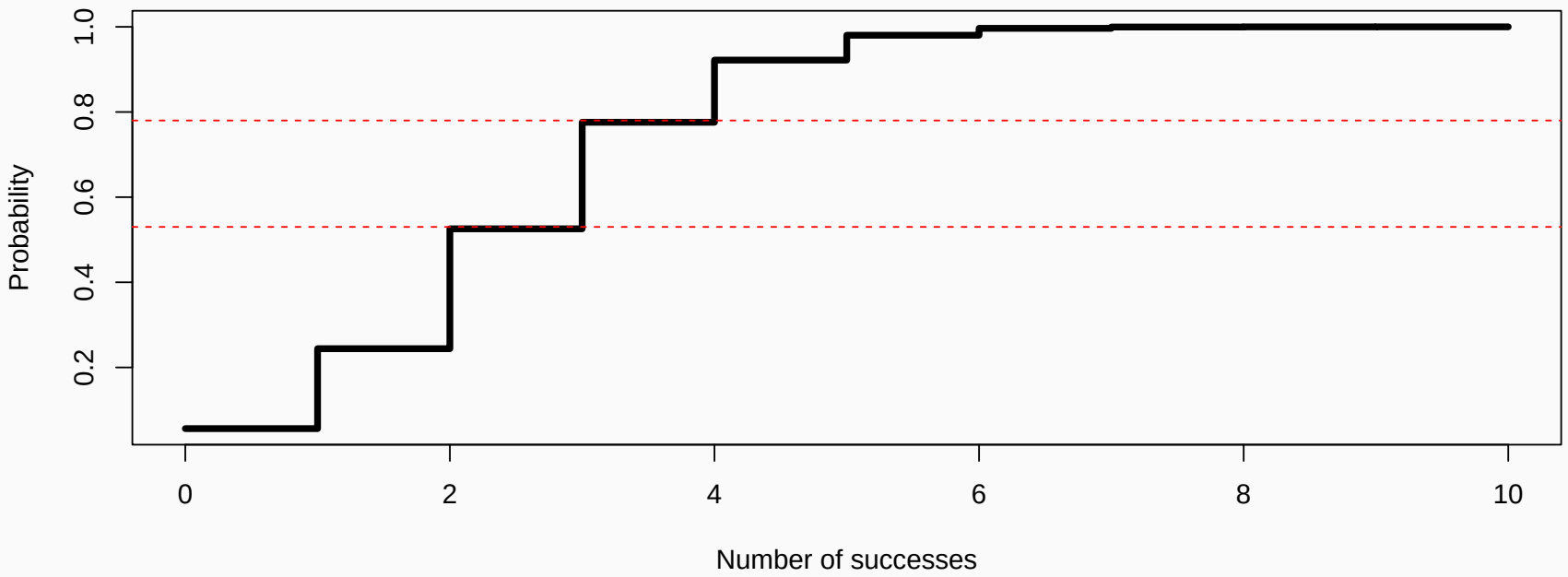
```
## [1] 0.9218731
```

`pbinom` gives the *cumulative probabilities*, the *cdf*.

Let $X \sim \text{Binom}(10, 1/4)$, then the cdf of X is

```
##    0    1    2    3    4    5    6    7    8    9   10
## 0.06 0.24 0.53 0.78 0.92 0.98 1.00 1.00 1.00 1.00 1.00
```

```
plot(0:10, pbinom(0:10, 10, 1/4),
     xlab="Number of successes",
     ylab="Probability", type="s", lwd=4)
abline(0.53, 0, lty = 2, col = "red")
abline(0.78, 0, lty = 2, col = "red")
```



Let $X \sim \text{Binom}(10, 1/4)$.

What's the probability of getting between 4 and 8 successes?

$P(4 \leq X \leq 8) = ?$

$$\underbrace{\underbrace{0 \quad 1 \quad 2 \quad 3}_{X \leq 3} \quad \underbrace{4 \quad 5 \quad 6 \quad 7 \quad 8}_{4 \leq X \leq 8}}_{X \leq 8} \quad \underbrace{9 \quad 10}_{9 \leq X \leq 10}$$

$$P(4 \leq X \leq 8) = P(X \leq 8) - P(X \leq 3)$$

```
pbinom(8, 10, 1/4) - pbinom(3, 10 , 1/4)
```

```
## [1] 0.2240953
```

```
sum(dbinom(4:8, 10,1/4))
```

```
## [1] 0.2240953
```

Let $X \sim \text{Binom}(10, 1/4)$.

What's the probability of getting more than 4 successes?

$P(X > 4) = ?$

$P(X > 4) = 1 - P(X \leq 4)$

```
1 - pbinom(4, 10, 1/4)
```

```
## [1] 0.07812691
```

Generating binomial observations

If $X \sim Bino(10, 0.8)$ then X is total number of successes in 10 trials of a binomial experiment with $p = 0.8$

```
sum(sample(0:1, 10, replace = T, prob = c(0.2, 0.8)))
```

```
## [1] 6
```

Using built in binomial generator:

```
rbinom(1, size = 10, prob = 0.8)
```

```
## [1] 8
```

Generate a sample of size 5 from this binomial distribution.

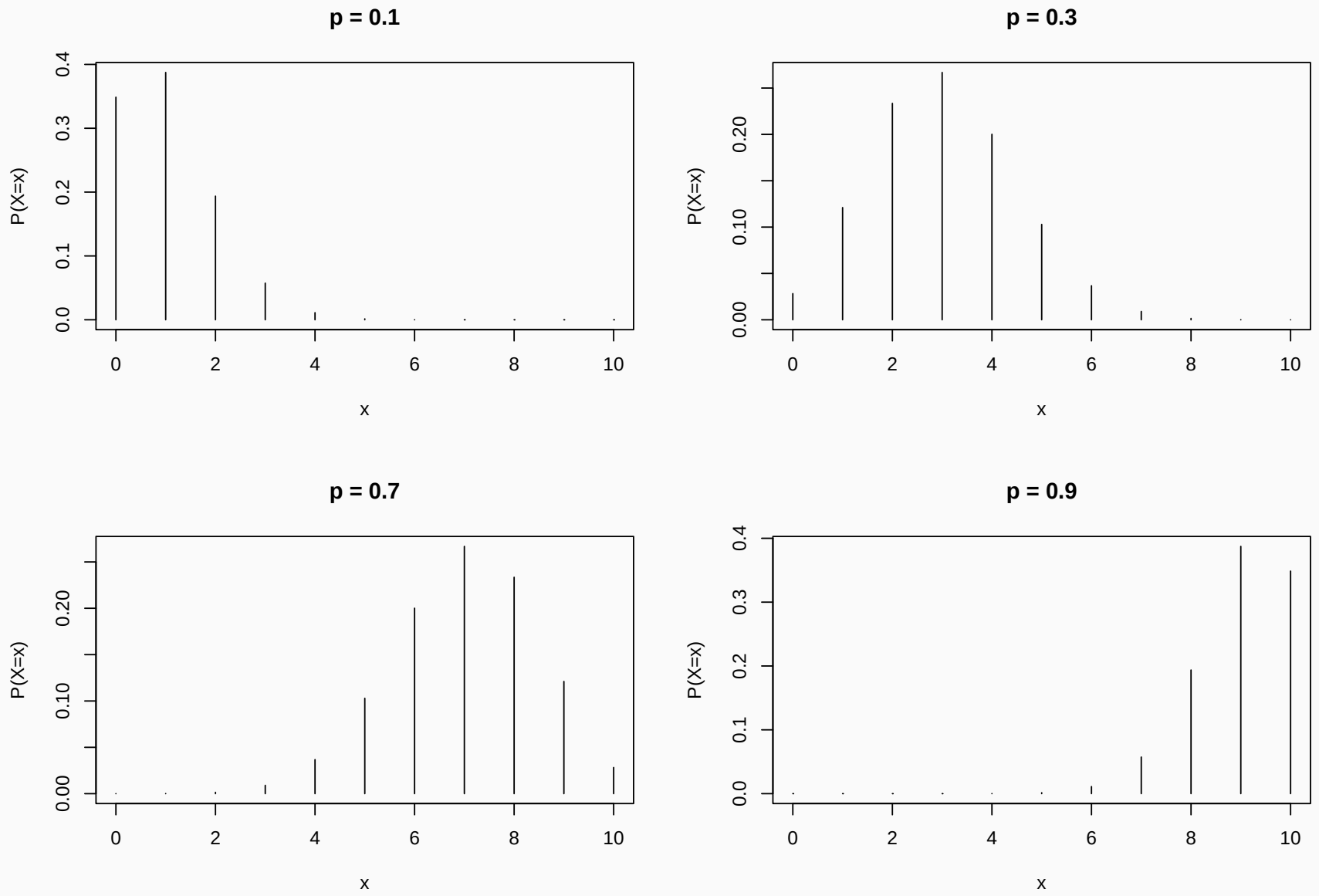
```
rbinom(5, size = 10, prob = 0.8)
```

```
## [1]  8  9  8 10  9
```

Binomial distributions for $n = 10$ and different p

```
par(mfrow = c(2,2))
plot(0:10,dbinom(0:10, 10, 0.1),
     type="h", xlab = "x", ylab = "P(X=x)",
     main = "p = 0.1")
plot(0:10,dbinom(0:10, 10, 0.3),
     type="h", xlab = "x", ylab = "P(X=x)",
     main = "p = 0.3")
plot(0:10,dbinom(0:10, 10, 0.7),
     type="h",xlab = "x", ylab = "P(X=x)",
     main = "p = 0.7")
plot(0:10,dbinom(0:10, 10, 0.9),
     type="h", xlab = "x", ylab = "P(X=x)" ,
     main = "p = 0.9")
```

Binomial distributions for $n = 10$ and different p



Distributions in R

R comes with many named distributions built in.

The functions below illustrate a pattern in R:

- `dbinom(x, size, prob)`
- `pbinom(q, size, prob)`
- `rbinom(n, size, prob)`

We will soon see

- `dunif`, `punif`, `runif`
- `dnorm`, `pnorm`, `rnorm`

Where does $\binom{n}{k}$ come from?

A Gallup survey suggests that 25% of Americans are obese. Suppose we randomly select four individuals to participate in this experiment. What is the probability that exactly 1 of them will be obese?

Where does $\binom{n}{k}$ come from?

Let's call these people Alex (A), Brian (B), Carol (C), and Dalia (D). Only four scenarios will satisfy the condition of "exactly 1 of them is obese":

A	B	C	D	Probability
$\underbrace{NO}_{0.75}$	$\underbrace{NO}_{0.75}$	$\underbrace{NO}_{0.75}$	$\underbrace{O}_{0.25}$	$0.75^3 0.25 = 0.1055$
NO	NO	O	NO	$0.75^3 0.25 = 0.1055$
NO	O	NO	NO	$0.75^3 0.25 = 0.1055$
O	NO	NO	NO	$0.75^3 0.25 = 0.1055$

The probability of exactly one 1 of 4 people being obese is the sum of all of these probabilities.

$$0.1055 + 0.1055 + 0.1055 + 0.1055 = 4 \times 0.1055 = 0.422$$

$$4 \times 0.25 \times 0.75^3 = 0.422$$

number of ways of choosing 1 slot for obese out of 4 slots $\times$ $P(O) \times P(NO)^3$

$$\binom{4}{1} \times p \times (1-p)^4 = 0.422$$

$$\binom{4}{1} = \frac{4!}{1!(4-1)!} = \frac{4.3.2.1}{1(3.2.1)} = 4$$

More generally, k successes in n trials

- Step 1: an outcome with k success and $(n-k)$ failures has probability $p^k(1-p)^{n-k}$.

$$\underbrace{\underline{S} \quad \underline{S} \quad \cdots \quad \underline{S}}_{k \text{ trials}} \quad \underbrace{\underline{F} \quad \cdots \quad \underline{F}}_{n-k \text{ trials}}$$

- Step 2: there are $\binom{n}{k}$ possible outcomes with k success and $(n-k)$ failure.
- Step 3:

$$P(\#\text{success} = k) = (\#\text{outcomes with } k \text{ successes}) \times P(\text{each outcom}$$

$$= \binom{n}{k} \times p^k(1-p)^{n-k}.$$

Summary:

- Named discrete distributions
  - Discrete uniform
  - Binomial Distribution

Next:

- Named continuous distributions
  - Uniform Distribution
  - Normal Distribution