Topic 3 – Discrete Distributions

ENVX1002 Introduction to Statistical Methods

Dr. Floris van Ogtrop
The University of Sydney

Feb 2024

CRICOS 00026A TEQSA PRV12057

Outline – Discrete distributions

- Example
- What is a distribution
- Binomial distribution
- Poisson distribution

Learning outcomes

At the end of this topic students should able to:

- Have a good understanding of what a distribution is
  - Definitions
  - Functions
  - Binomial and Poisson distributions
- Apply the correct model to describe data
- Demonstrate proficiency in the use of R and Excel for calculating probabilities

Types of data

Remember the types of data

Numerical

- Continuous: yield, weight
- Discrete: weeds per m^2

Categorical

- Binary: 2 mutually exclusive categories
- Ordinal: categories ranked in order
- Nominal: qualitative data

Example

- We have 5 insects which we spray with an insecticide, each insect has a 60% chance of being killed;
- $P(K) = 0.6$
- Possible questions:
  - What is the probability that all 5 insects will be killed?
  - What is the probability that at least 3 insects will be killed?
- The data is ‘binary’ and the events are mutually exclusive (either dead or alive unless it is zombie fly);
  - we can say the data is categorical or numeric discrete
  - we can use a binomial (discrete) distribution to “model” the data.

Generated using DALL.E3

What is a distribution?

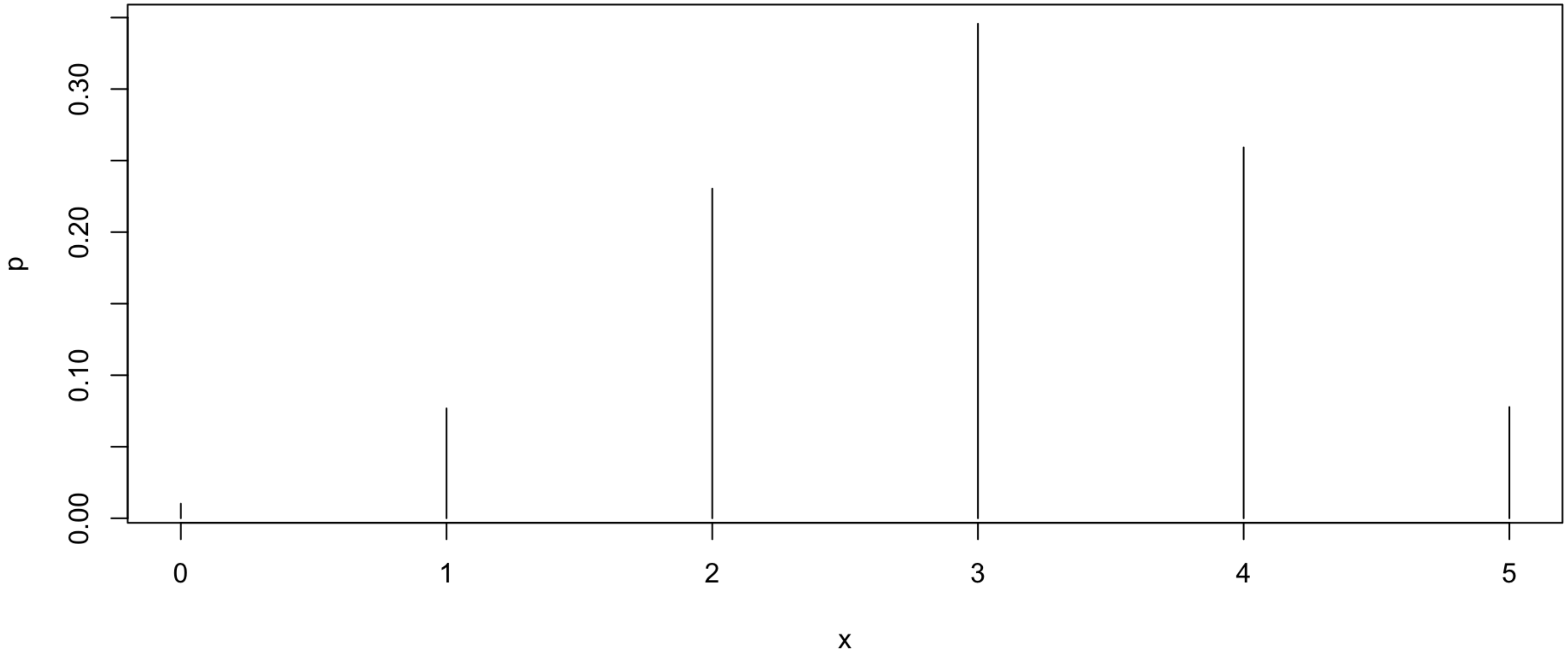
- In our case we are generally referring to a distribution function
  - This is a function (or model) that describes the probability that a system will take on value or set of values {x}
- For any variable X, we describe probabilities by
  - Discrete variables: probability distribution function $P(X = x)$
  - Continuous variables: probability density function $f(x)$
  - Discrete and Continuous variables: cumulative density function $F(x) = P(X \leq x)$

Back to our example

- We spray 5 flies with insecticide which has 60% chance of killing each insect. If X is the number of flies that die, what is the distribution of X?
- The set of possible values is $x = 0, 1, 2, 3, 4, 5$
- The likelihood of each value is
- $P(X = 0) = P(\text{no insects die}) = 0.4 \times 0.4 \times 0.4 \times 0.4 \times 0.4 = 0.4^5 = 0.01024$
- $P(X = 1) = P(\text{one insects die}) = 0.6 \times 0.4 \times 0.4 \times 0.4 \times 0.4 + 0.4 \times 0.6 \times 0.4 \times 0.4 \times 0.4 + 0.4 \times 0.4 \times 0.6 \times 0.4 \times 0.4 + 0.4 \times 0.4 \times 0.4 \times 0.6 \times 0.4 + 0.4 \times 0.4 \times 0.4 \times 0.4 \times 0.6 = 0.0768$
- $P(X = 2) = P(\text{two insects die}) = \ldots 10 \text{ different combinations} = 0.2304$
- $P(X = 3) = P(\text{three insects die}) = \ldots 10 \text{ different combinations} = 0.3456$
- $P(X = 4) = P(\text{four insects die}) = \ldots 5 \text{ combinations} = 0.2592$
- $P(X = 0) = P(\text{all insects die}) = 0.6 \times 0.6 \times 0.6 \times 0.6 \times 0.6 = 0.65 = 0.07776$

Example plot

```r
x <- c(0, 1, 2, 3, 4, 5)
p <- c(0.01024, 0.0768, 0.2304, 0.3456, 0.2592, 0.07776)
plot(x, p, type = "h")
```



Example – properties of the distribution

- Remember we have a binomial (dead or alive) distribution here
- A key property of all (discrete) distributions is that all probabilities add to

$$\Sigma_{i=0}^{5} P(X = i) = 0.01024 + 0.0768 + 0.2304 + 0.3456 + 0.2592 + 0.07776 = 1$$

- Note that all probabilities lie between 0 and 1

Binomial Distribution

- Why does a binomial distribution fit our insect data??
  - Basic element is a Bernoulli trial – each insect;
  - The outcome of each trial can be classified in precisely one of two mutually exclusive ways termed “success” (dead) and “failure” (alive);
    - We usually assign p to success and q to failure.
  - Binomial experiments consists of n Bernouilli (independent binary) trials (i.e. 5 insects);
  - The probability of a success, denoted by p, remains constant from trial to trial. The probability of a failure, $q = 1 – p$;
    - $p = 0.6$ and $q = 0.4$
  - The trials are independent; that is, the outcome of any particular trial is not affected by the outcome of any other trial;
  - The number of successes, x, is a binomial variable.

Example

- How many combinations are there for exactly 2 flies to die out of 5 flies?

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{(5\times4\times3\times2\times1)}{(2\times1\times3\times2\times1)} = 10$$

- What is the probability that exactly 2 flies will die?

$$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x} = \binom{5}{2}0.6^2(1-0.6)^{5-2}$$

$$= 10 \times 0.36 \times 0.064 = 0.2304$$

- `dbinom(2,5,0.6)`

```
dbinom(2,5,0.6)
```

```
[1] 0.2304
```

- `=BINOM.DIST(2,5,0.6,FALSE)`

Point, cumulative or interval probabilities

- We have already calculated a point probability i.e. the probability of exactly 2 flies dying.
- But what about if we wanted to know the probability of 2 or more flies dying $P(X \geq 2)$ or between 2 and 4 flies $P(2 \leq X \leq 4)$ dying??
- So how can we calculate these?
  - One way is to calculate all the point probabilities from 0-5 and add the probabilities cumulatively or in the interval

x	P
0	0.01024
1	0.0768
2	0.2304
3	0.3456
4	0.2592
5	0.07776
SUM	1

Point, cumulative or interval probabilities

$P(X \geq 2) = (X = 2) + (X = 3) + (X = 4) + (X = 5) = 0.2304 + 0.3456 + 0.2592 + 0.07776 = 0.91296$

- You guys try to calculate the following

$P(2 \leq X \leq 4) = ??$

x	P
0	0.01024
1	0.0768
2	0.2304
3	0.3456
4	0.2592
5	0.07776
SUM	1

Cumulative

$P(X \geq 2)$

```r
1 - pbinom(1,5,0.6)
```

```
[1] 0.91296
```

```r
## OR

pbinom(1,5,0.6, lower.tail = FALSE)
```

```
[1] 0.91296
```

- `=1-BINOM.DIST(1,5,0.6,TRUE)`

Interval

$P(2 \leq X \leq 4)$

```r
pbinom(4,5,0.6)-pbinom(1,5,0.6)
```

```
[1] 0.8352
```

- `=BINOM.DIST(4,5,0.6,TRUE)-BINOM.DIST(1,5,0.6,TRUE)`

Mean and variance of the binomial distribution

- Mean binomial distribution

$\mu_x = np$

$= 5 \times 0.6 = 3$ On average 3 flies die in 5 trials

- Variance binomial distribution

$\sigma_x^2 = np(1-p)$

$= 5 \times 0.6(1-0.6) = 1.2$ with a variance of 1.2 flies

Count Data

Horse kick deaths in the Prussian Army

```r
library(knitr)
kick <- read.csv("data/Kick_deaths.csv")
kable(kick[1:12,])
```

Year	GC	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C14	C15
1875	0	0	0	0	0	0	0	1	1	0	0	0	1	0
1876	2	0	0	0	1	0	0	0	0	0	0	0	1	1
1877	2	0	0	0	0	0	1	1	0	0	1	0	2	0
1878	1	2	2	1	1	0	0	0	0	0	1	0	1	0
1879	0	0	0	1	1	2	2	0	1	0	0	2	1	0
1880	0	3	2	1	1	1	0	0	0	2	1	4	3	0
1881	1	0	0	2	1	0	0	1	0	1	0	0	0	0
1882	1	2	0	0	0	0	1	0	1	1	2	1	4	1
1883	0	0	1	2	0	1	2	1	0	1	0	3	0	0
1884	3	0	1	0	0	0	0	1	0	0	2	0	1	1
1885	0	0	0	0	0	0	1	0	0	2	0	1	0	1
1886	2	1	0	0	1	1	1	0	0	1	0	1	3	0

Horse kick deaths in the Prussian Army

https://en.wikipedia.org/wiki/Ladislaus_Bortkiewicz

```r
library(tidyverse)

frequency_kick <- kick %>%
  select(-Year) %>%
  pivot_longer(cols = everything(), names_to = "Column", 
  count(Deaths) %>%
  arrange(Deaths) %>%
  mutate(Total_Deaths = Deaths*n) %>%
  mutate(Probability = "?")

kable(frequency_kick)
```

Deaths	n	Total_Deaths	Probability
0	144	0	?
1	91	91	?
2	32	64	?
3	11	33	?
4	2	8	?

- What is the Probability in any month of
  - 0 injuries by horse kick
  - 1 injuries by horse kick
  - 2 injuries by horse kick
  - 3 injuries by horse kick
  - 4 injuries by horse kick
- λ “Lambda” is the mean

Horse kick deaths in the Prussian Army

```
total_kick <- frequency_kick %>%
  summarize(n = sum(n), sum_Total_Deaths = sum(Total_Deaths))

kable(total_kick)
```

n	sum_Total_Deaths
280	196

Poisson Distribution

$X \sim Po(\lambda)$

$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $x = 0, 1, 2, \ldots$ $\lambda > 0$

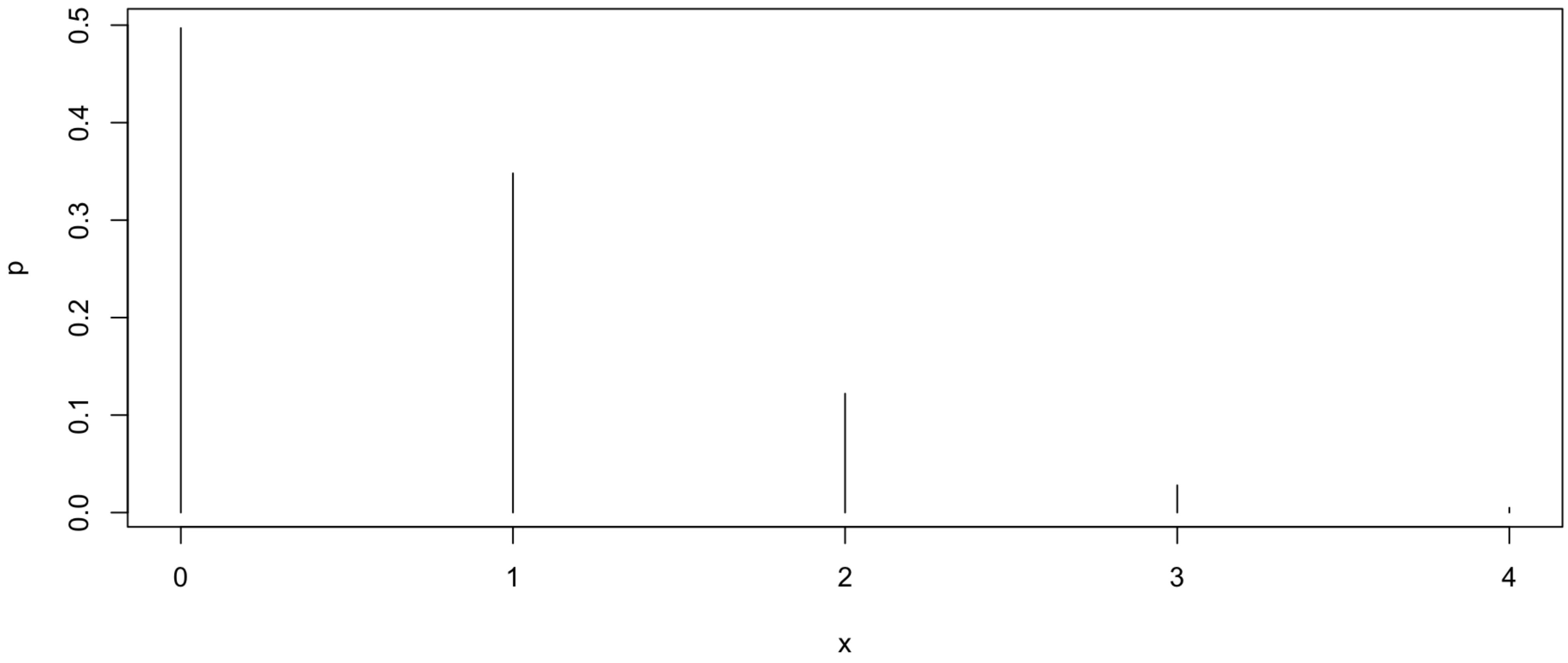
- Note that e denotes the exponential function such that
  - $e^0 = 1$
  - $e^-2 = 0.135 (3 \text{ d.p.})$
  - $e^-10 = 4.540 \times 10^{-5} (3 \text{ d.p.})$

Example with kick deaths

- We first identify the model
  - X = the number of soldiers injured by horse kick $\sim Po(\lambda)$ where λ = the average number of deaths = $196/280 = 0.7$. We can now calculate the probability of having exactly 0, 1, 2, 3, 4, 5 deaths
- $P(X=0) = \frac{0.7^0 e^{-0.7}}{0!} = \frac{1e^{-0.7}}{1} = 0.497(3 \text{ d.p.})$
- $P(X=1) = \frac{0.7^1 e^{-0.7}}{1!} = \frac{0.7e^{-0.7}}{1} = 0.348(3 \text{ d.p.})$
- $P(X=2) = \frac{0.7^2 e^{-0.7}}{2!} = \frac{0.49e^{-0.7}}{2\times 1} = 0.122(3 \text{ d.p.})$
- $P(X=3) = \frac{0.7^3 e^{-0.7}}{3!} = \frac{0.343e^{-0.7}}{3\times 2\times 1} = 0.028(3 \text{ d.p.})$
- $P(X=4) = \frac{0.7^4 e^{-0.7}}{4!} = \frac{0.2401e^{-0.7}}{4\times 3\times 2\times 1} = 0.005(3 \text{ d.p.})$

Example with kick deaths

```r
x <- c(0, 1, 2, 3, 4)
p <- c(0.497, 0.348, 0.122, 0.028, 0.005)
plot(x, p, type = "h")
```



Example with kick deaths

```r
frequency_kick1 <- kick %>%
  select(-Year) %>%
  pivot_longer(cols = everything(), names_to = "Column", values_to = "Deaths") %>%
  count(Deaths) %>%
  arrange(Deaths) %>%
  mutate(Total_Deaths = Deaths*n) %>%
  mutate(Probability = c(0.497, 0.348, 0.122, 0.028, 0.005)) %>%
  mutate(Observed_Probability = n/280)

kable(frequency_kick1)
```

Deaths	n	Total_Deaths	Probability	Observed_Probability
0	144	0	0.497	0.5142857
1	91	91	0.348	0.3250000
2	32	64	0.122	0.1142857
3	11	33	0.028	0.0392857
4	2	8	0.005	0.0071429

- note that observed probability is n divided by the total number of observations (280). For example, for 0 deaths there were 144 observed out of a total of 280 observations i.e. the observed probability of 0 deaths in a cavalry corps over 20 years of observations was 0.51 or 51%

Example with kick deaths

- So now you all can calculate what the probability is, as an example, the of having less than 2 deaths across all cavalry corps for the period of 1875-1894 $P(X < 2)$.

```
ppois(1,0.7)
```

```
[1] 0.844195
```

- `=POISSON(1,0.7,TRUE)`

Example with kick deaths

- Another example, is having exactly 2 deaths across all cavalry corps for the period of 1875-1894 $P(X = 2)$.

```r
dpois(1,0.7)
```

```
[1] 0.3476097
```

- `=POISSON(2,0.7,FALSE)`

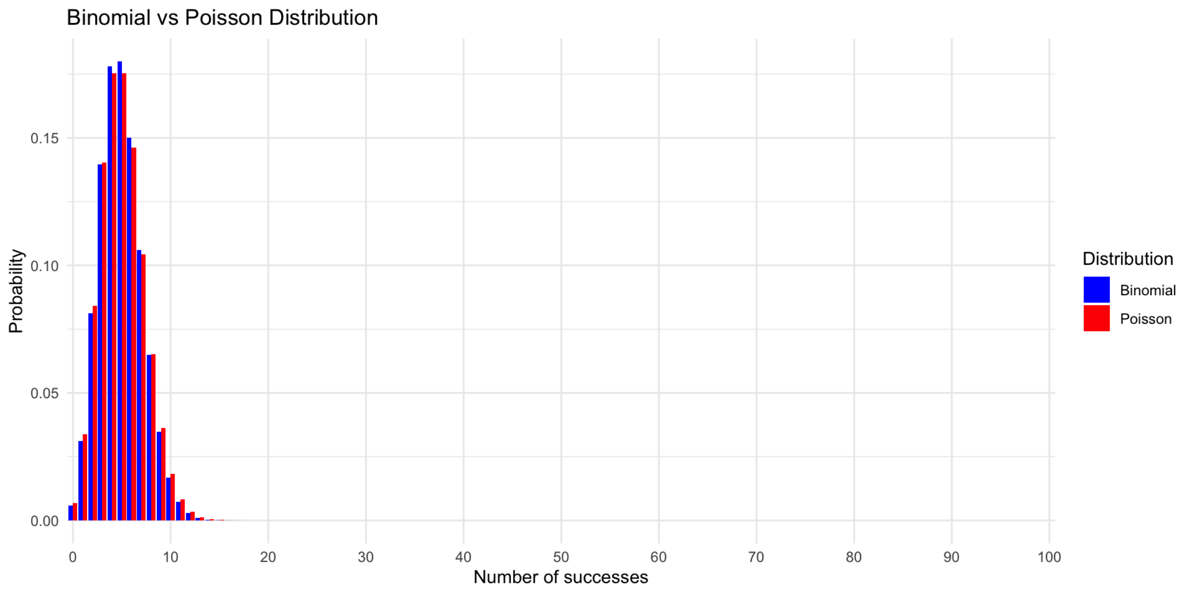
Interesting results with the binomial and Poisson distributions

- For large n and small p the Binomial distribution $X \sim Bin(n, p)$ can be approximated by the Poisson distribution $Y \sim Po(\lambda = np)$
  - The general rule is if $n > 20$ and $np < 5$
- We often say that the Poisson Distribution models rare events

Interesting results with the binomial and Poisson distributions

```r
# Parameters for the binomial distribution
n <- 100
p <- 0.05

# Calculating lambda for the Poisson approximation
lambda <- n * p

# Generate the range of values
x <- 0:n

# Data frames for plotting
data_binom <- data.frame(x = x, probability = dbinom(x, n
data_pois <- data.frame(x = x, probability = dpois(x, laml

# Combine data
data_combined <- rbind(data_binom, data_pois)

# Create the plot
p <- ggplot(data_combined, aes(x = factor(x), y = probabil
    geom_bar(stat = "identity", position = position_dodge
    ggtitle("Binomial vs Poisson Distribution") +
    xlab("Number of successes") +
    ylab("Probability") +
    scale_fill_manual(values = c("blue", "red")) +
```

```r
print(p)
```



Further reading

- Quinn & Keough (2002)
  - Chapter 1. Sections 1.5, p. 9-13
- Mead et al. (2002)
  - Chapter14. Sections 14.4-14.5, p. 339-377

Thanks!