Applied Statistics for Data Scientists with R

Class 14: Probability Distributions

What is Probability?

- Probability is a measure of how likely an event is to occur.
- Lies within 0 and 1. (0% to 100%)

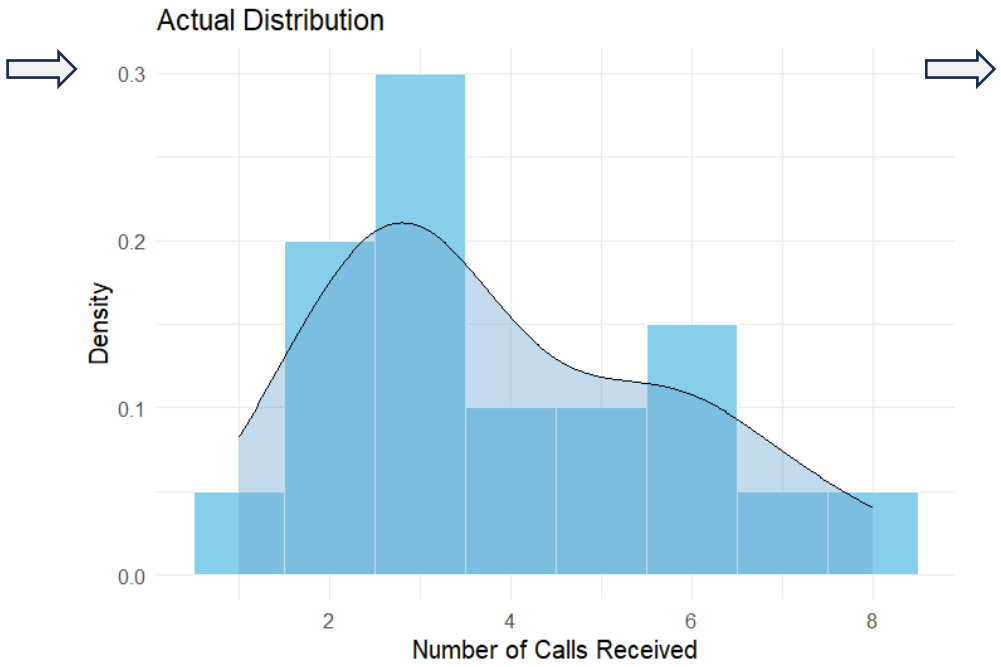
Probability Distribution

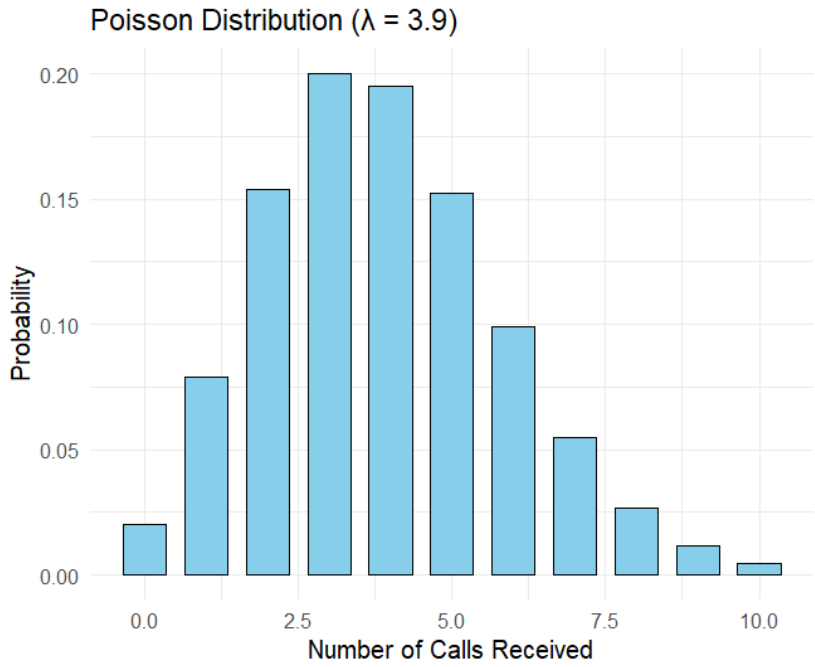
- Two types –
  1. Discrete distribution
  2. Continuous distribution

- Discrete probability distributions can be found for discrete variables.
  - Number of phone calls received at a call center in an hour.
  - Number of customers who make a purchase out of 20 entering a store.
- Continuous probability distributions can be found for continuous variables.
  - Random number generated between 0 and 1 (equal likelihood for all values).
  - IQ scores of a population.

Poisson Distribution

Day	Hour	Calls Received (Y)
1	1	3
1	2	3
1	3	2
1	4	4
1	5	6
1	6	7
1	7	6
1	8	5
1	9	3
1	10	2
2	1	1
2	2	3
2	3	3
2	4	2
2	5	5
2	6	8
2	7	6
2	8	4
2	9	3
2	10	2





Rate, $\lambda = \frac{total\ number\ of\ calls}{number\ of\ hours} = \frac{78}{20} = 3.9$

Questions that can be answered:

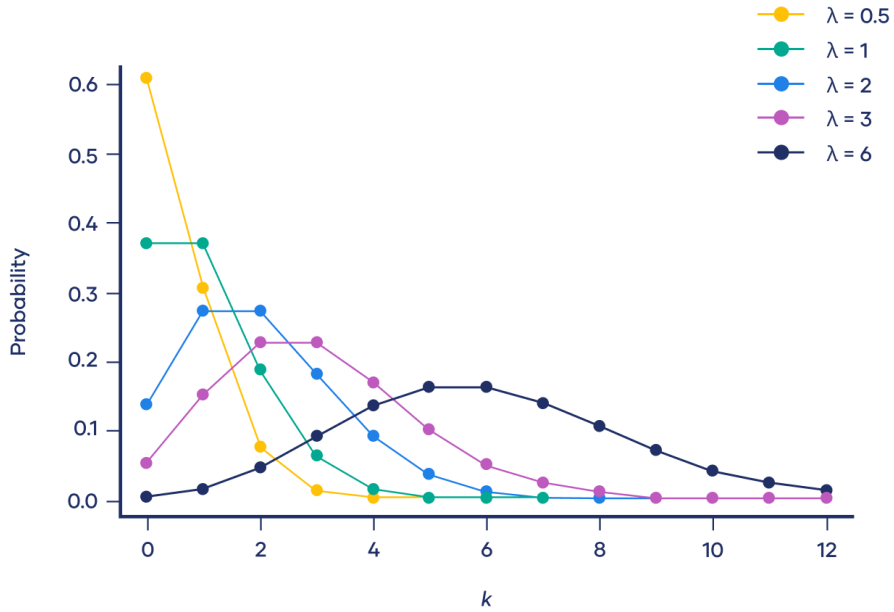
- What is the probability of getting 5 calls in an hour? (dpois)
- How many calls are expected to receive in 3rd hour? (GLM)

Poisson Distribution: Key Points

- Models the number of events occurring in a fixed interval of time, space, or area.
- Has only one parameter, λ(Lambda) = Average number of occurrences in the given interval.
- Mean and variance of the distribution is the same (λ)
- Probability mass function:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

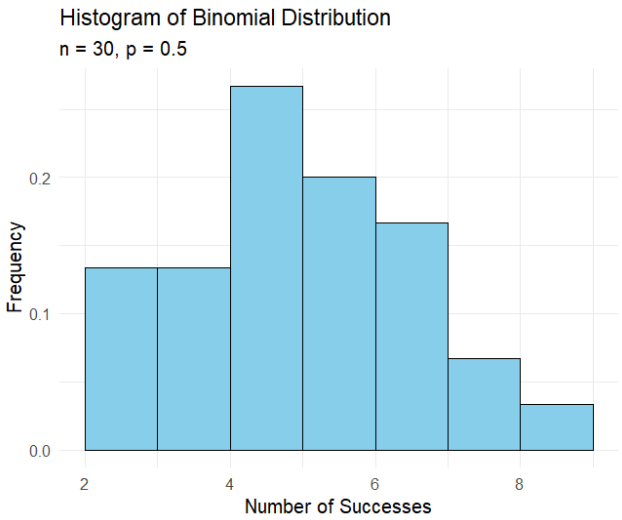
- For large λ, Poisson distribution approximately follows normal distribution.

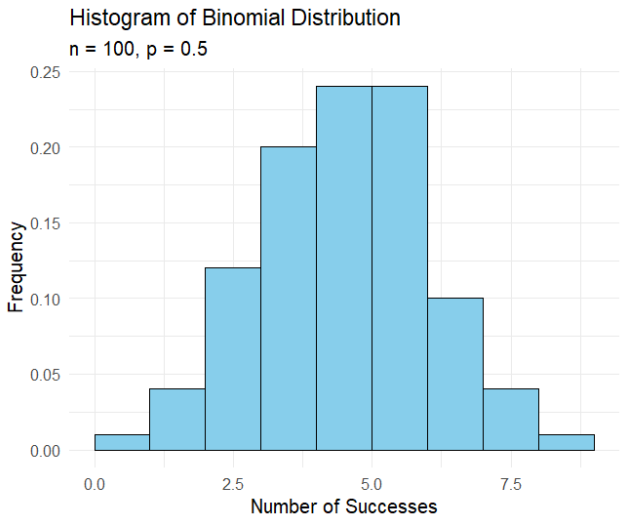


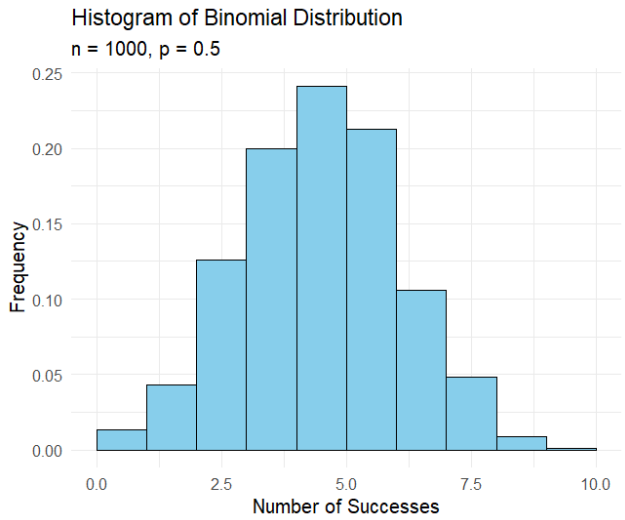
When to use?

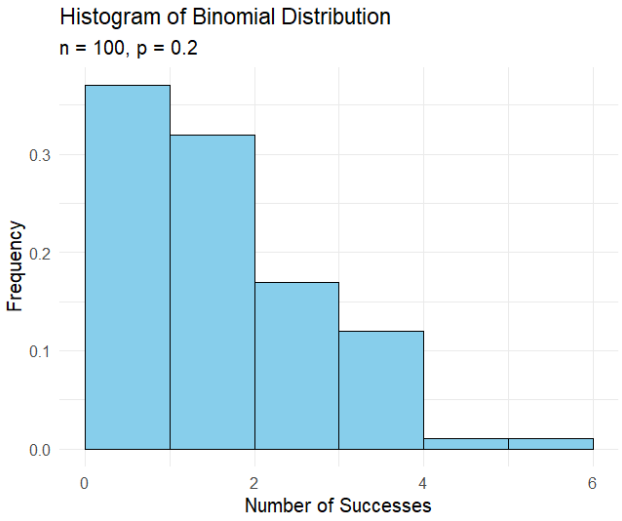
- Events are rare but possible in larger scale.
- Occurrences are independent with a constant average rate.

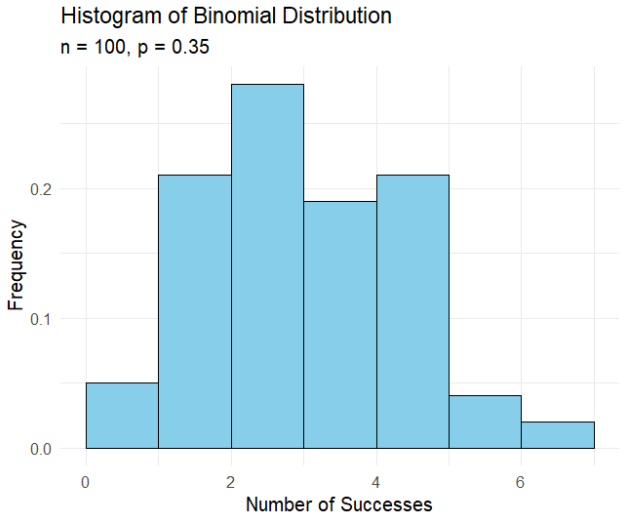
Binomial Distribution

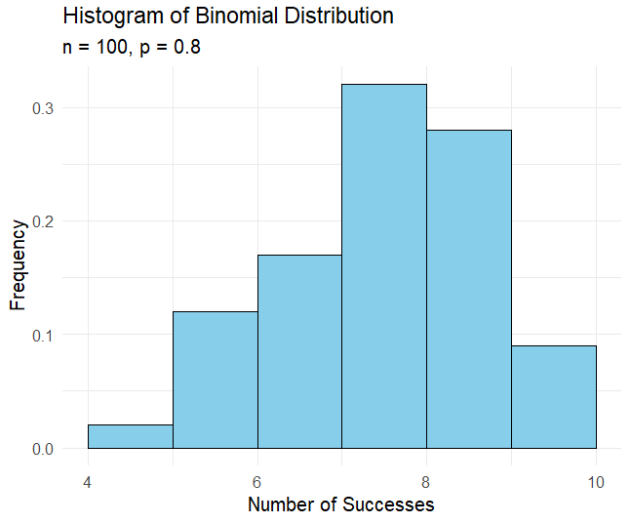












Binomial Distribution: Key Points

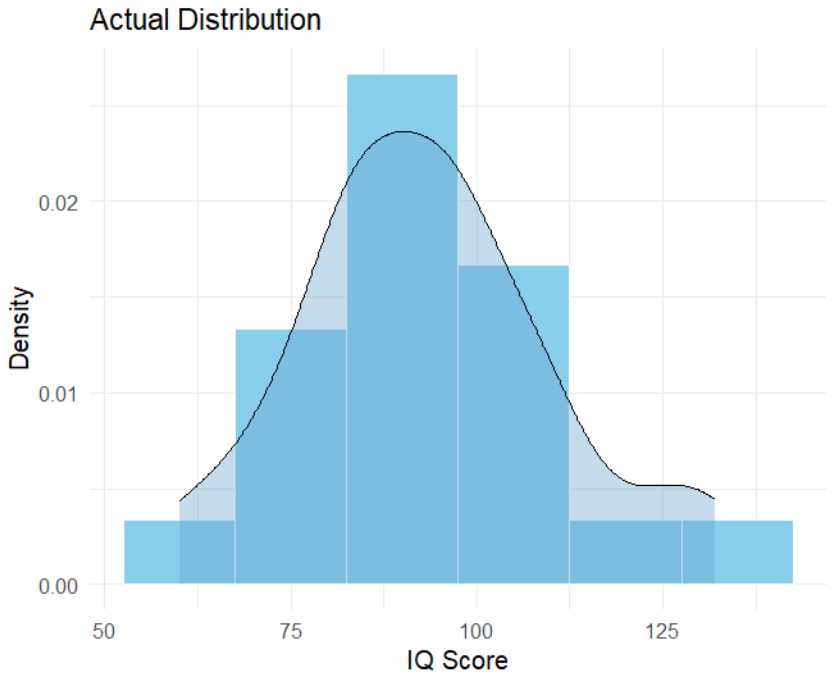
- Models the number of **successes** in a fixed number of independent **Bernoulli trials**.
- **Bernoulli trials** has two possible outcomes: **success** (probability = *p*) or **failure** (probability = *1 - p*).
- Has two parameters:
  - n: Number of trials.
  - p: Probability of success in each trial.
- Probability mass function:

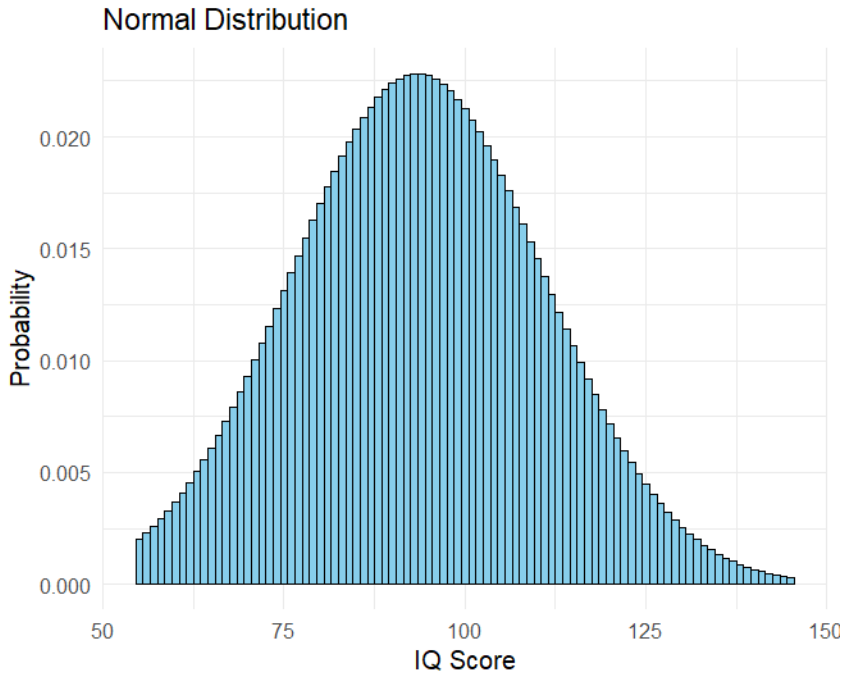
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- For large n and p=0.5, Binomial distribution tends to normal distribution.
- Binomial distribution with only 1 trial is Bernoulli distribution.
  It is the link function used in GLM for fitting a Logistic regression.

Normal Distribution

ID	IQ
1	70
2	132
3	81
4	82
5	60
6	95
7	94
8	87
9	98
10	109
11	84
12	74
13	84
14	100
15	97
16	86
17	124
18	107
19	110
20	94





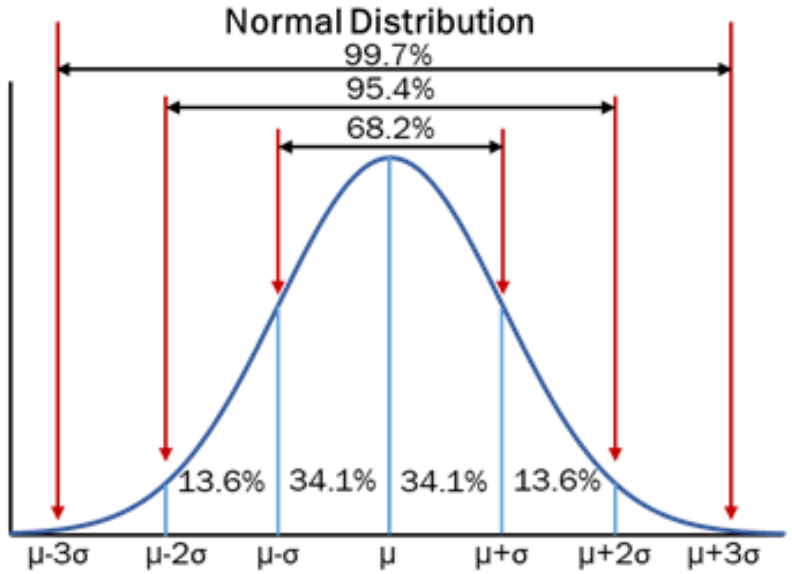
Normal Distribution: Key Points

- Has two parameters: mean and standard deviation.
- Mean determines the center of the distribution and SD determines the spread or dispersion of the distribution.
- Mean = Median = Mode in a perfectly normal distribution.
- When mean = 0 and SD = 1, then it is called standard normal distribution.
- Use z scores to standardize data:

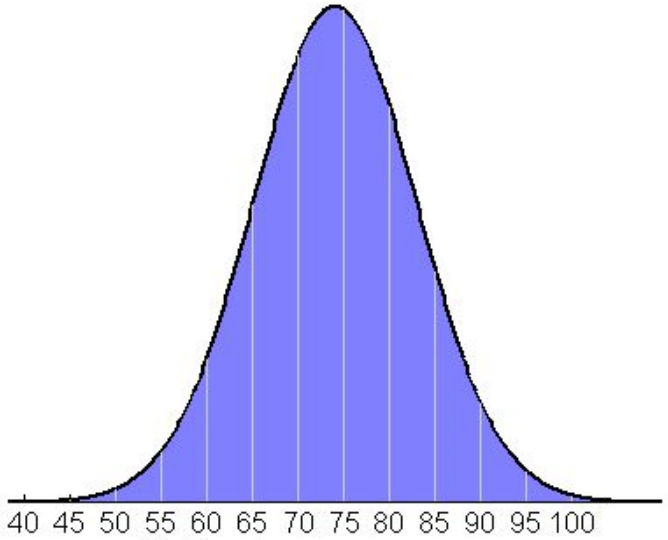
$$Z = \frac{X - \mu}{\sigma}$$

- Probability density function:

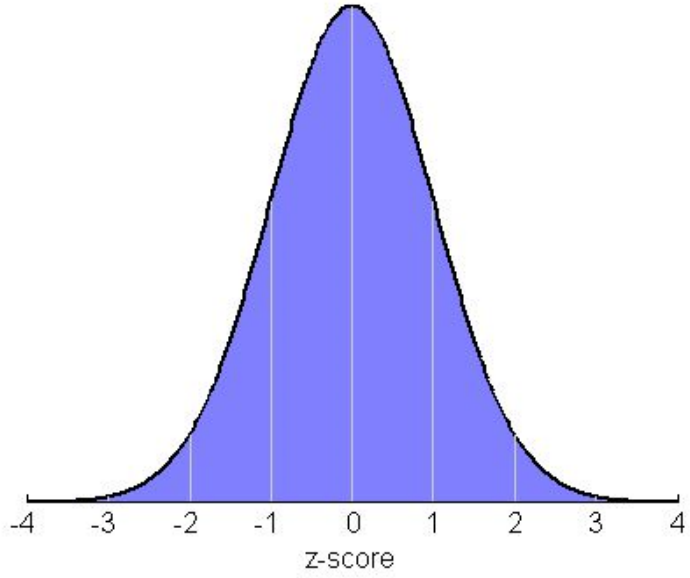
$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Normal Distribution: Standard Normal Distribution



Population (X)



Standard normal (z)