Probability Distributions

Why do we need probability distributions?

- Random factors affect all areas of our life, and businesses striving to succeed in today's highly competitive environment need a tool to deal with risk and uncertainty involved

- Probability distributions are parametric models which allows us to deal with uncertainty scientifically and make informed business decisions

Distribution Fitting/Modeling for Random data

- Finding the right distribution model for given random data is the key factor. If you select and apply an inappropriate distribution model, your subsequent calculations will be incorrect, and that will certainly result in wrong decisions

- In many industries, the use of incorrect models can have serious consequences such as inability to complete tasks or projects in time leading to substantial time and money loss, wrong engineering design resulting in damage of expensive equipment.

Model fitting and Testing

- How do you fit the right distribution model for given random data?
- How do you prove that selected model is right fit for given data?

Where do we use probability distributions?

- actuarial science and insurance
- risk analysis and investment
- market research
- business and economic research
- customer support
- data mining
- reliability engineering
- chemical engineering
- image processing
- physics and medicine etc.,

Continuous Data Modeling

Continuous Data Modeling - I

How do you model the data with symmetry and having less frequent extreme values?

Normal Distribution

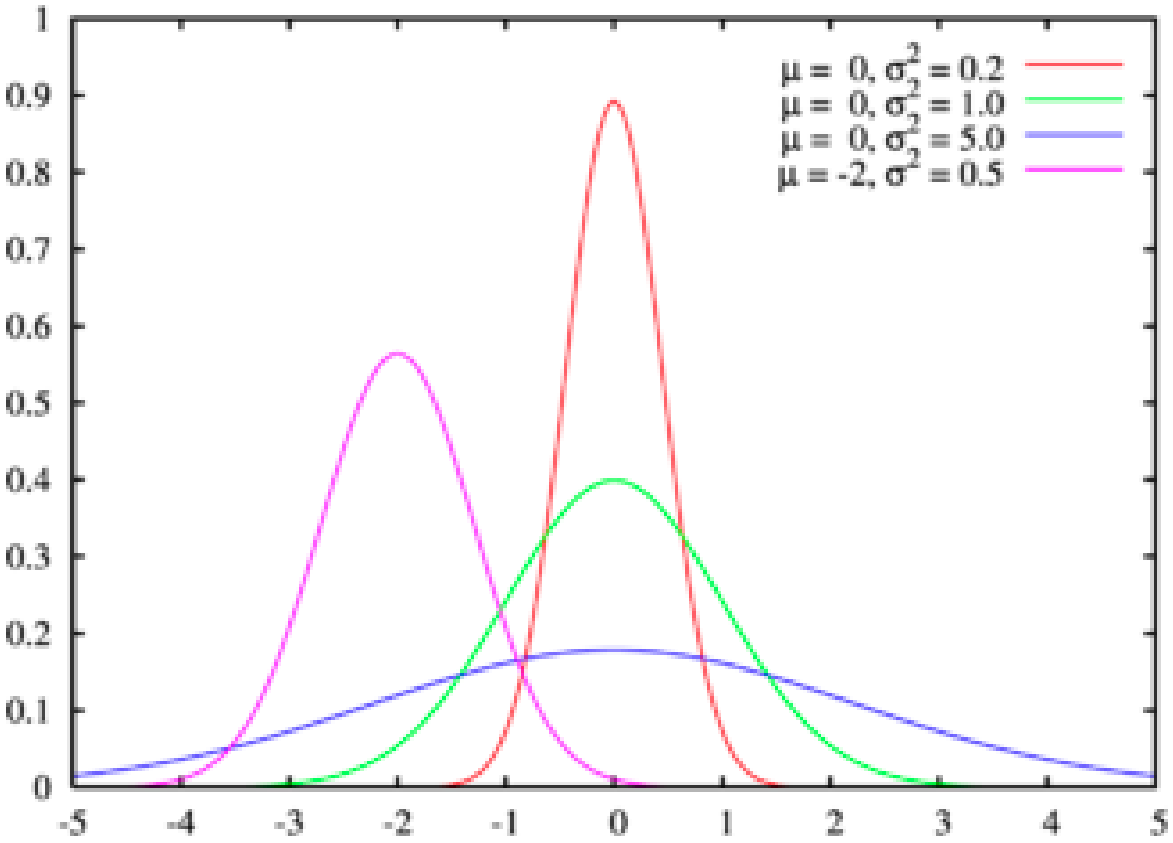
Used to model the data which have following characteristics:

- Symmetric(No skew): Strong tendency to have central value and the frequency of positive and negative deviations from this central value are equally likely
- Less likely to have outliers(No kurtosis): The frequency of the deviations falls off rapidly as we move further away from the central value.
- Unbounded: It can take values between –Inf to +Inf

Modeling examples

The following phenomenon can all be modeled with a normal distribution:

- Heights of people
- Measurement errors
- Blood pressure
- Points on a test
- IQ scores
- Salaries



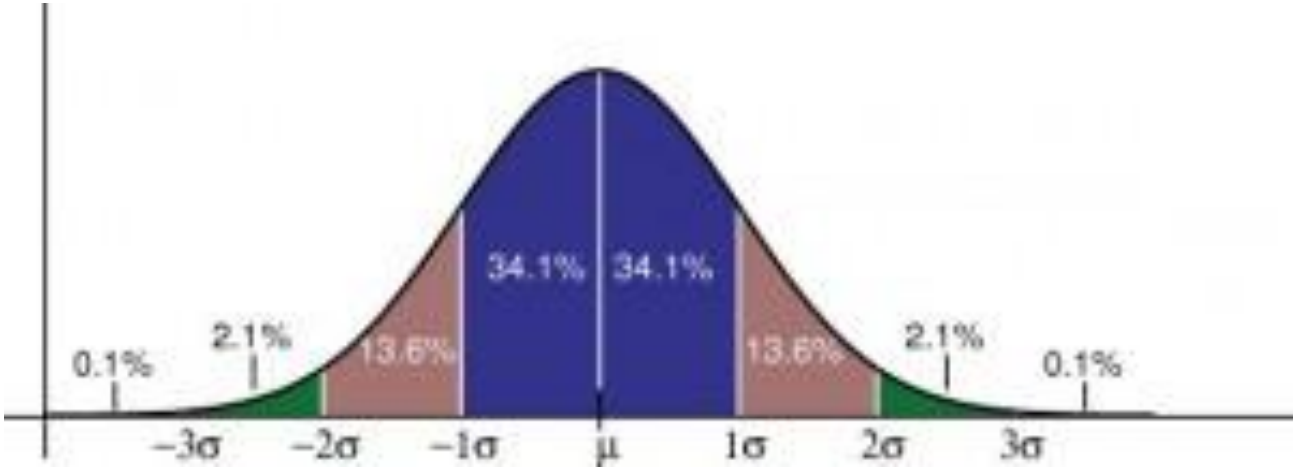
$$f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Two parameters define the shape of the distribution:

The mean parameter (μ) tells you where it's centered on the x-axis.

The spread parameter (σ) tells you what the spread is.

Interesting Insights



The empirical rule tells you what percentage of your data falls within a certain number of standard deviations from the mean:

- 68% of the data falls within one standard deviation of the mean.
- 95% of the data falls within two standard deviations of the mean.
- 99.7% of the data falls within three standard deviations of the mean.

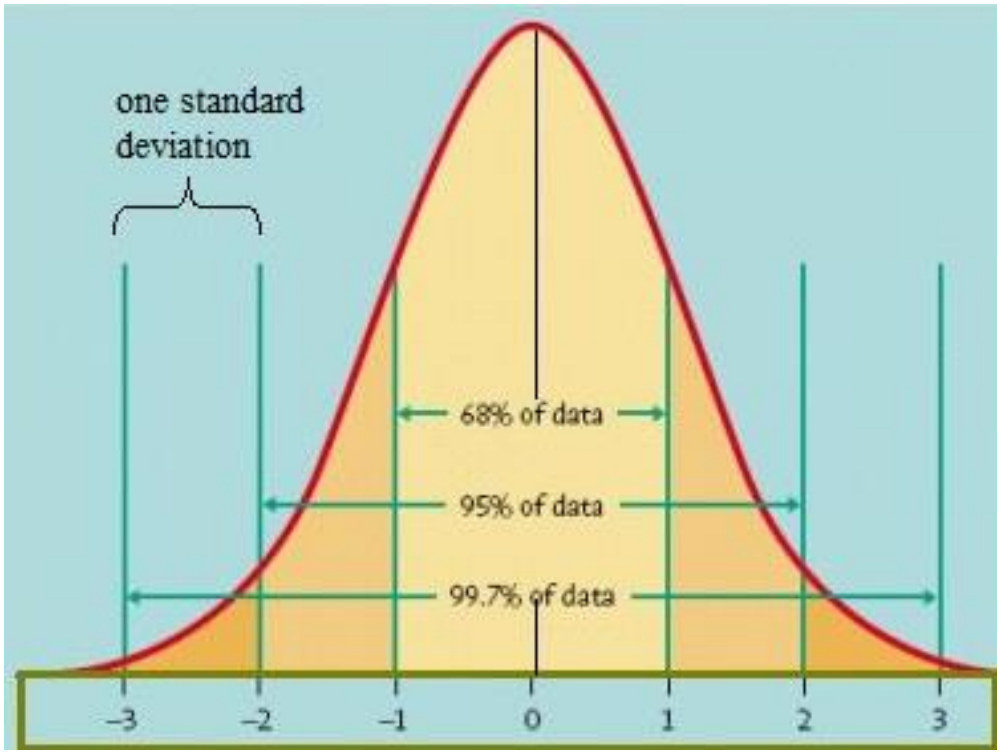
Application

The student salaries have a mean of $6,800, standard deviation of $2,500 and follows normal distribution.

a. Find the probability that a student gets salary between $7,300 and $9,000
b. Find the probability that a student gets salary above $8,000

Standard Normal Distribution

- A standard normal model is a normal distribution with a mean of 1 and a standard deviation of 1.
- It simplifies the computation of probabilities



Limitations for normal distribution modeling

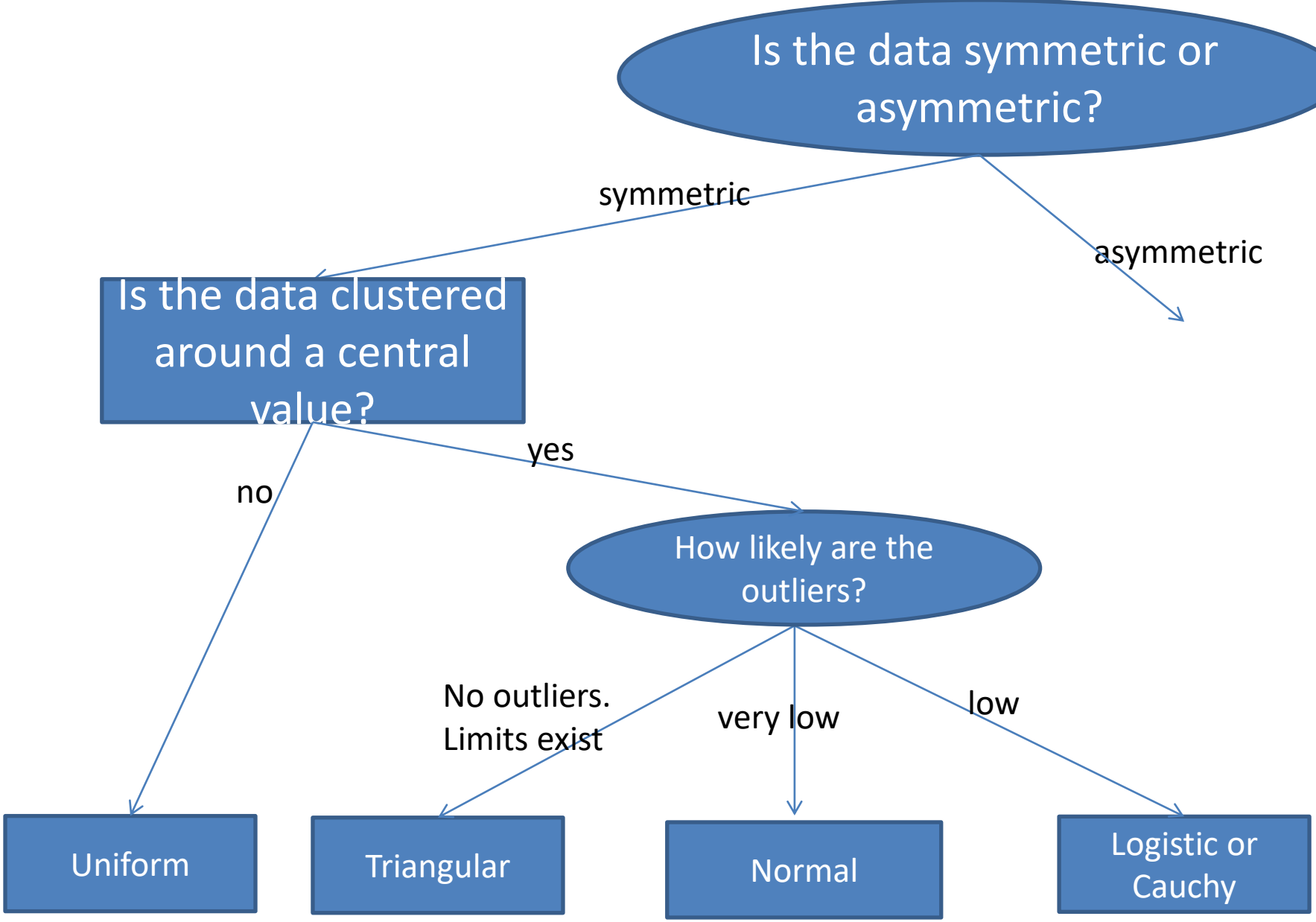
- Symmetric data with fat tails and accentuated peaks.
- Symmetric data with boundaries
- Asymmetric data

Continuous Data Modeling - II

How do you model the data with symmetry but with extreme values that occur more frequently than you would expect with a normal distribution?

How do you model the data with symmetry but without clustering around center value?

How do you model the data with symmetry but with boundaries?

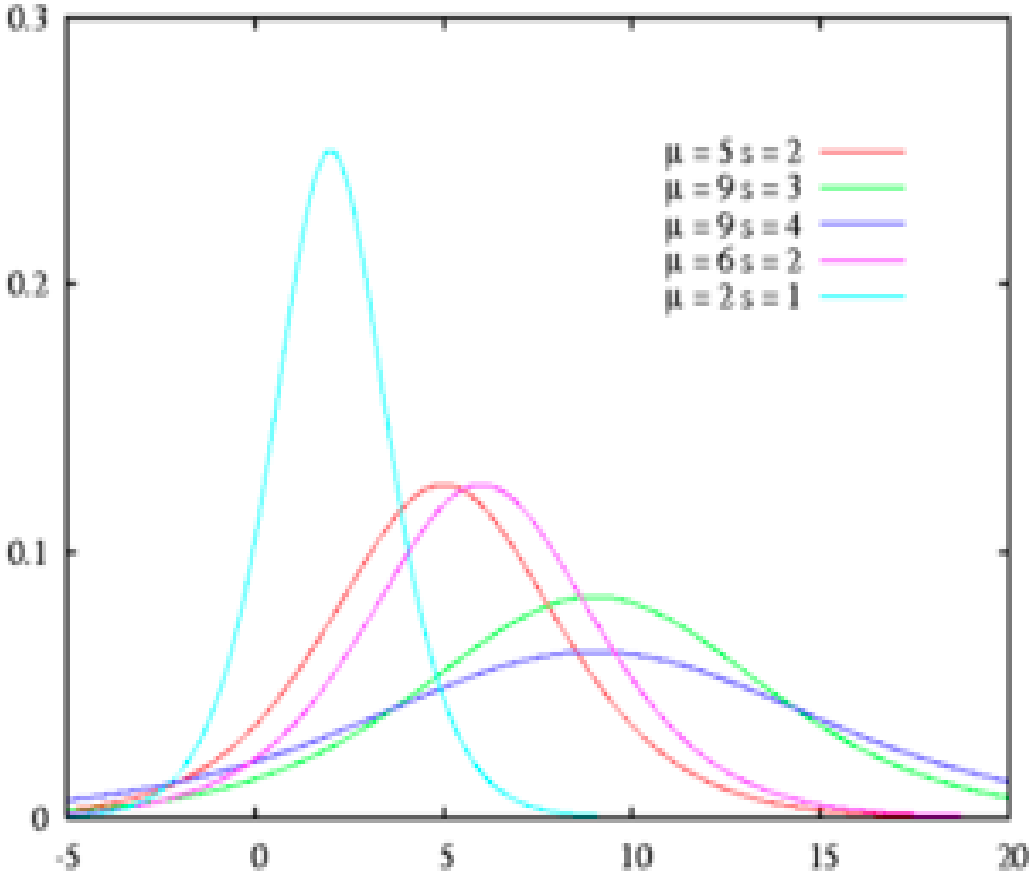

Is the data symmetric or asymmetric?
symmetric
asymmetric
Is the data clustered around a central value?
yes
no
How likely are the outliers?
No outliers. Limits exist
very low
low
Uniform
Triangular
Normal
Logistic or Cauchy

Logistic Distribution

Logistic Distribution

- Used to model the data which is symmetric around its central value and has fat tails (higher kurtosis)

- The following phenomenon can all be modeled with a logistic distribution:
  - Relative skill level of chess players



$$f(x;\mu,s) = \frac{e^{-\frac{(x-\mu)}{s}}}{s\left(1+e^{-\frac{(x-\mu)}{s}}\right)^2} \qquad -\infty < x < \infty.$$

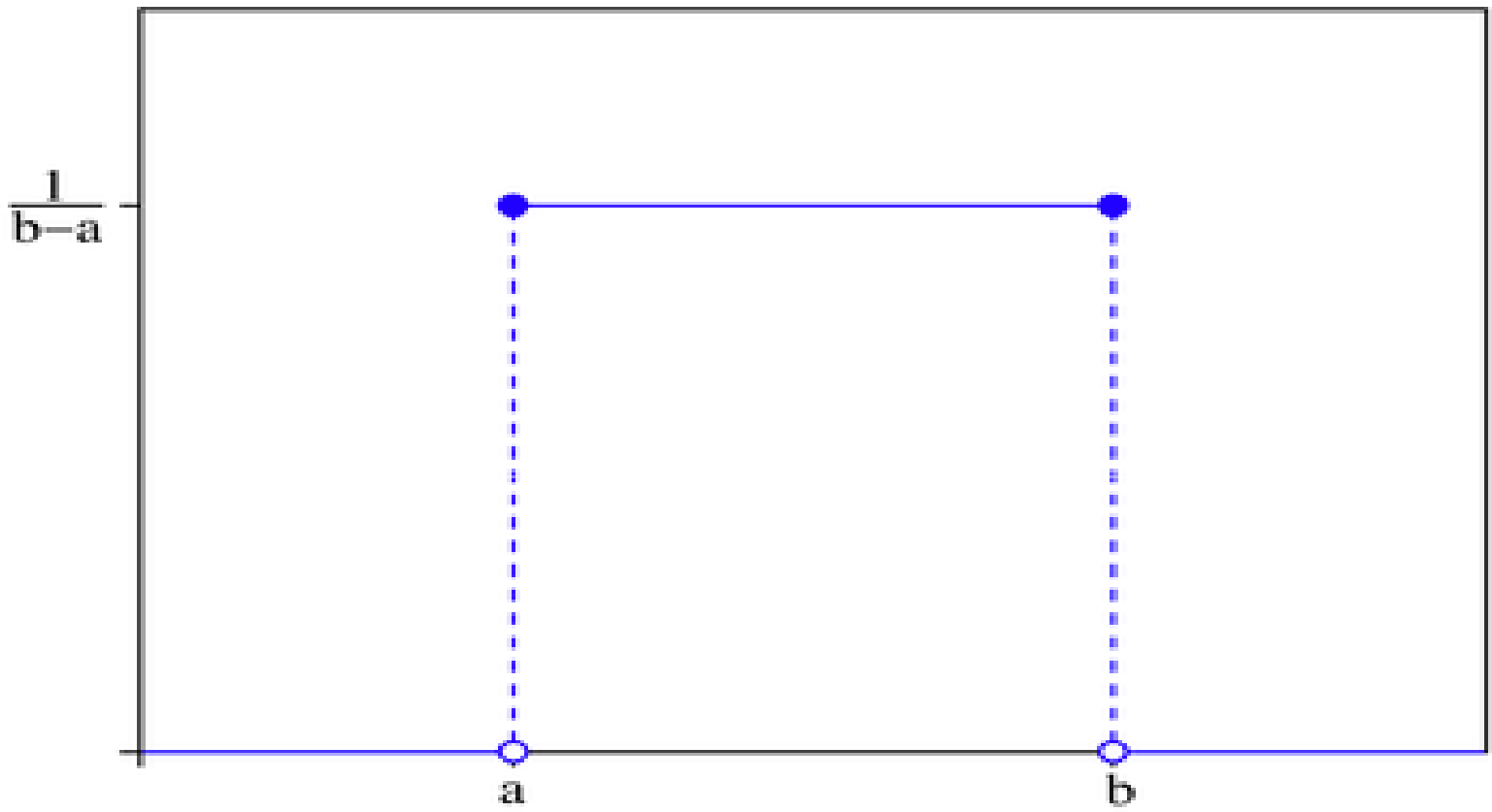
Two parameters define the shape of the distribution:

The location parameter (μ) tells you where it's centered on the x-axis.

The scale parameter (s) tells you what the spread is. It is proportional to the sd.

Uniform Distribution

Used to model data which has highest and lowest values but no real information about where within this range the value may fall. In other words, any value within that range is just as likely as any other value.


1
b−a
a
b

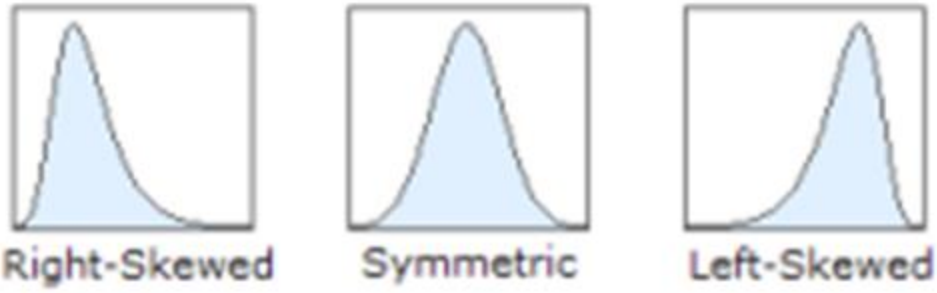
Application

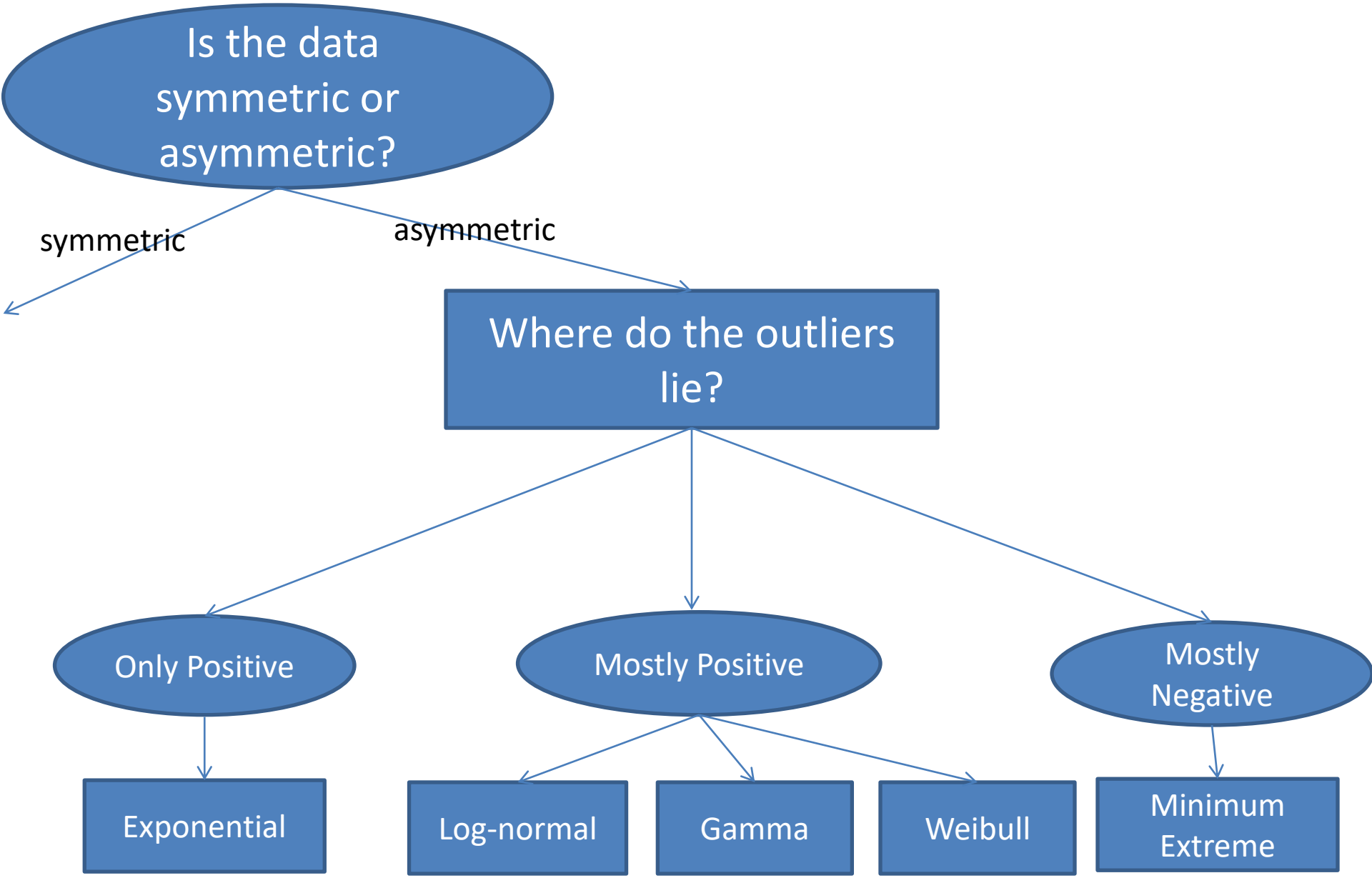
The average amount of weight gained by a person over the winter months is uniformly distributed from 0 to 30 kgs.

a. Find the probability that a person will gain between 10 and 15 kgs during the winter months.
b. Find the probability that a person will gain up to 10 kgs.

Continuous Data Modeling - III

How do you model the data with asymmetry and skews towards either very large positive or very large negative values?

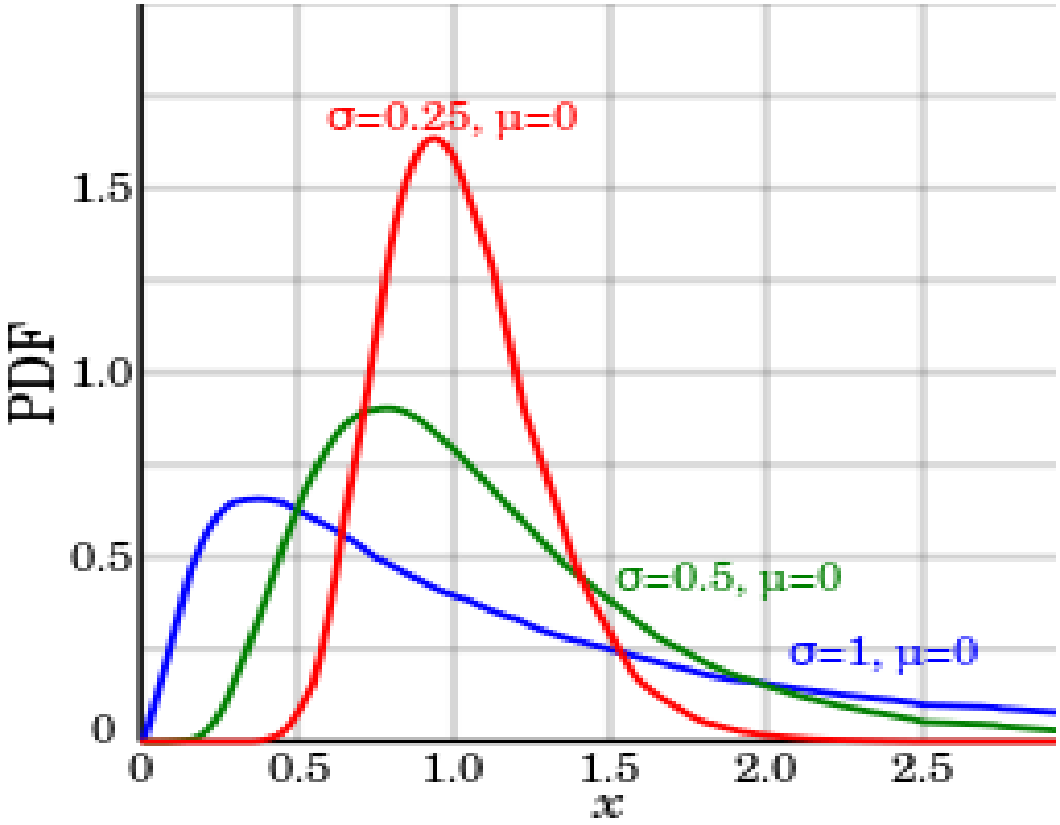



Is the data symmetric or asymmetric?
symmetric
asymmetric
Where do the outliers lie?
Only Positive
Mostly Positive
Mostly Negative
Exponential
Log-normal
Gamma
Weibull
Minimum Extreme

Lognormal Distribution

- Used to model data whose log values follow normal distribution

- The following phenomenon can all be modeled with a lognormal distribution:
  - Milk production by cows
  - Amounts of rainfall
  - The volume of gas in a petroleum reserve etc.,



$$\mathcal{N}(\ln x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(\ln x-\mu)^2}{2\sigma^2}\right], \quad x>0.$$

lognormal distribution is typically characterized by three parameters: a shape (σ or sigma), a scale (μ or median) and a shift parameter (theta).

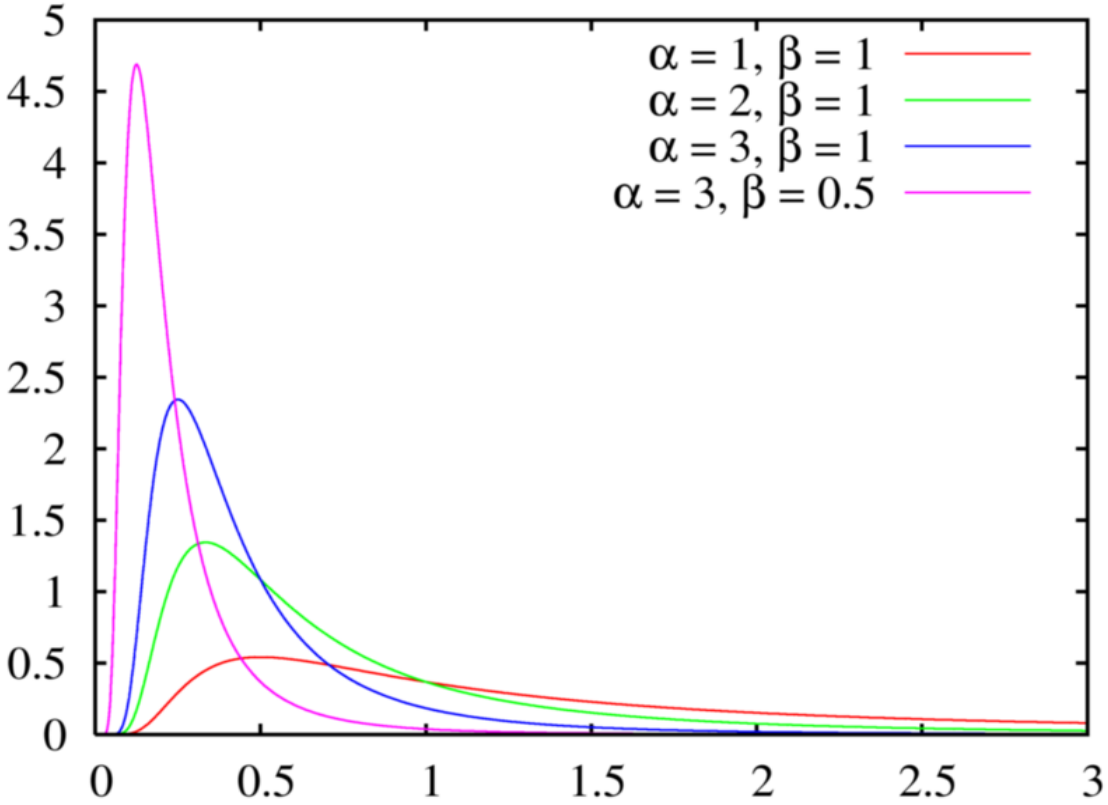
When μ = 0 and σ =1, you have the standard lognormal distribution

When theta=0, the distribution requires only scale and sigma parameters. As the sigma rises, the peak of the distribution shifts to the left and the skewness in the distribution increases.

Gamma Distribution

- Used to model data which is having mostly right skew

- The following phenomenon can all be modeled with a gamma distribution:
  - The amount of rainfall accumulated in a reservoir
  - The size of loan defaults or aggregate insurance claims
  - The load on web servers etc.,



$$f(x;\alpha,\beta) = \begin{cases} \dfrac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

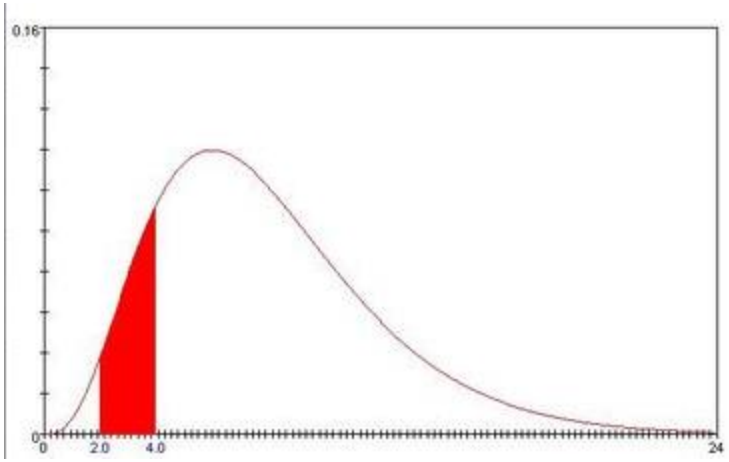
$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t}\, dt,$$

if x is a positive integer,
then $\Gamma(x) = (x-1)!$

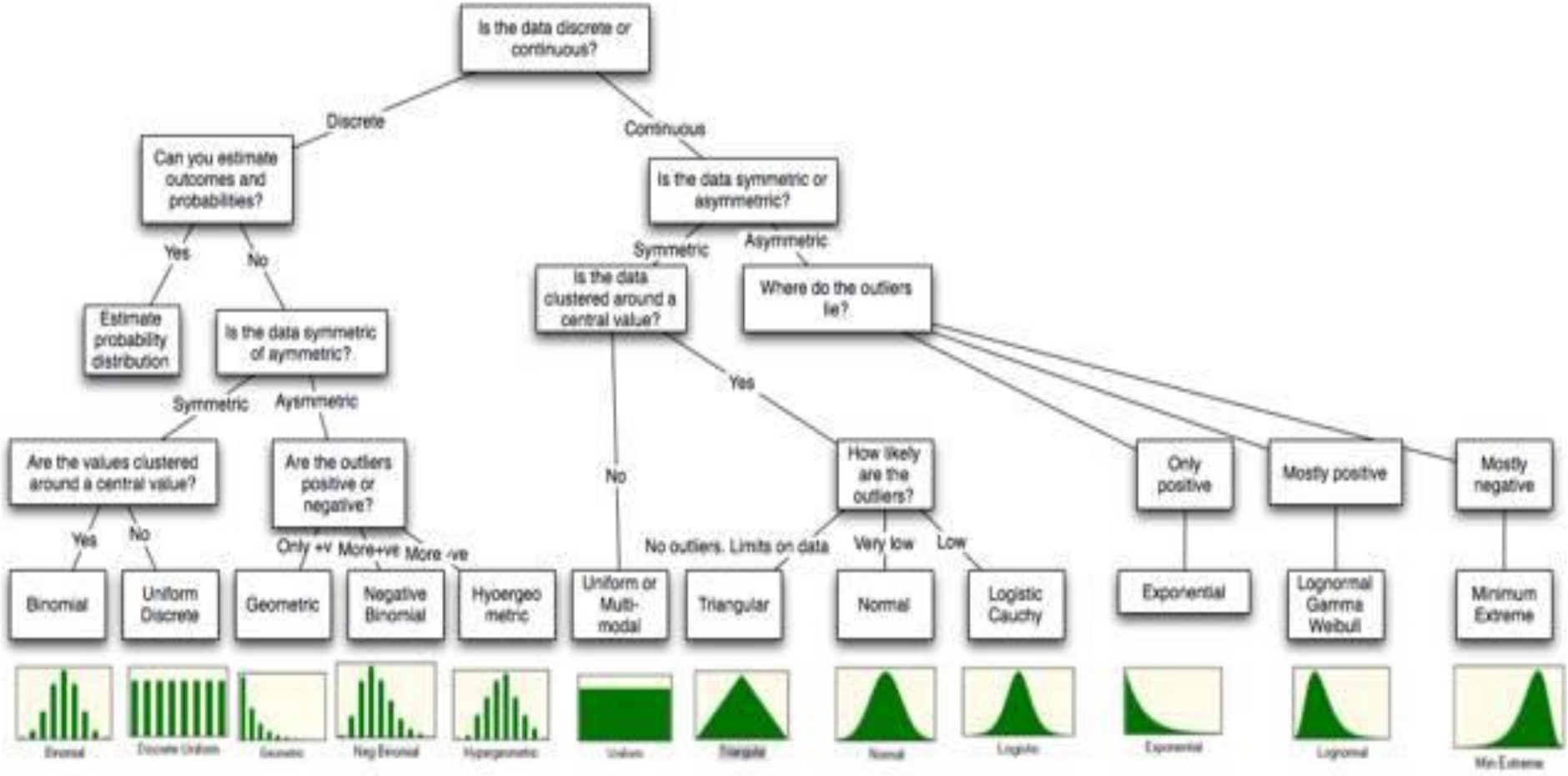
It is characterized by two parameters alpha(shape) and beta(rate) and both can change the shape of the graph. Think of α as the **number of occurrences of events** and β as the **mean number of events per time unit(equals to 1/mean time between events)**

Application

Suppose you are fishing and you expect to get a fish once every 1/2 hour. Compute the probability that you will have to wait between 2 to 4 hours before you catch 4 fishes.



Summary


Is the data discrete or continuous?
Discrete
Continuous
Can you estimate outcomes and probabilities?
Is the data symmetric or asymmetric?
Yes
No
Symmetric
Asymmetric
Estimate probability distribution
Is the data symmetric of asymmetric?
Is the data clustered around a central value?
Where do the outliers lie?
Symmetric
Asymmetric
Yes
No
Are the values clustered around a central value?
Are the outliers positive or negative?
How likely are the outliers?
Only positive
Mostly positive
Mostly negative
Yes
No
Only +v
More+ve
More -ve
No outliers. Limits on data
Very low
Low
Binomial
Uniform Discrete
Geometric
Negative Binomial
Hyoergeo metric
Uniform or Multi-modal
Triangular
Normal
Logistic Cauchy
Exponential
Lognormal Gamma Weibull
Minimum Extreme