SESSION 4: DATA DISTRIBUTIONS

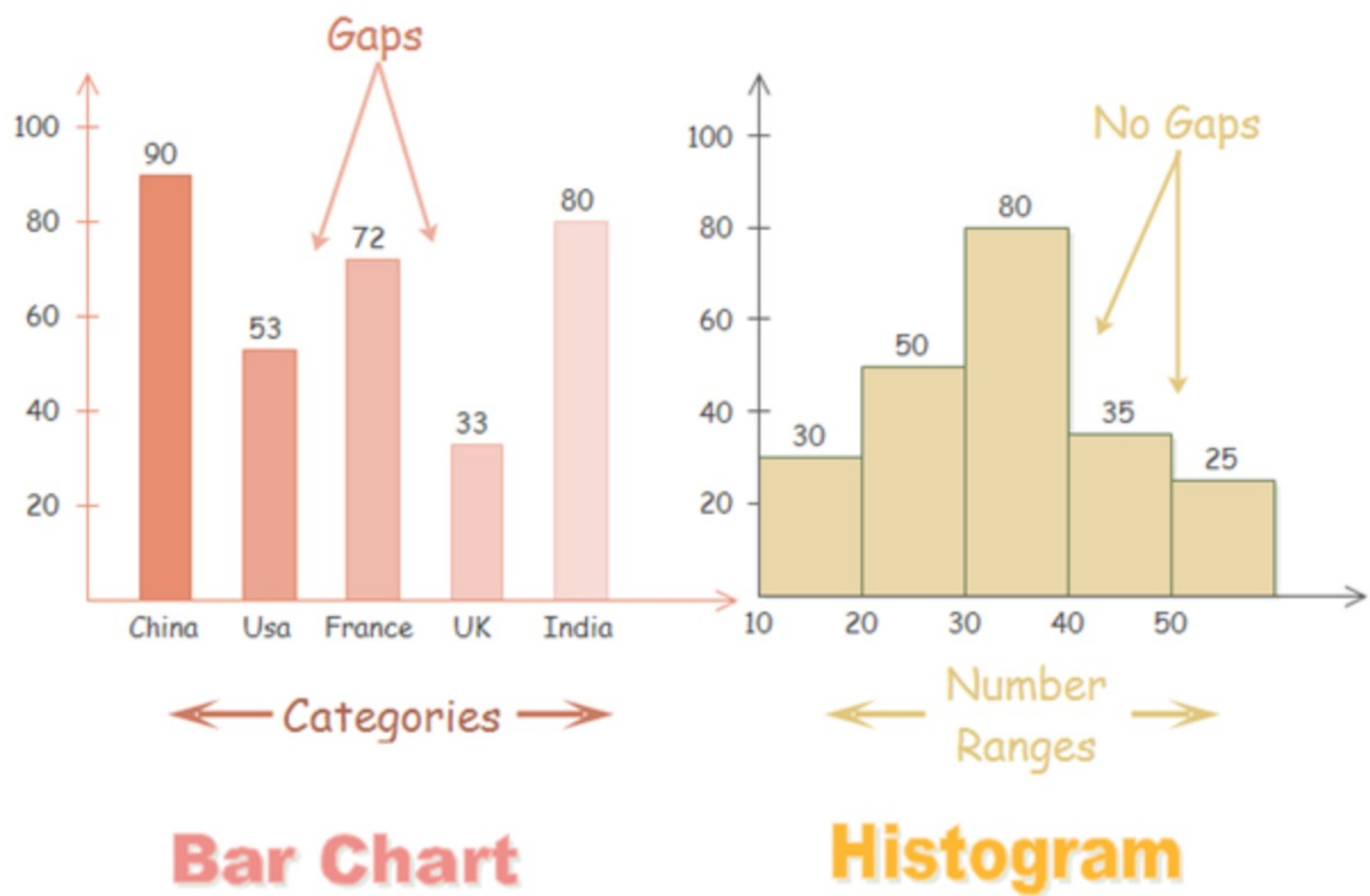
Session 4

Accounting & Statistics

Data Distributions: Why they matter!

- While data descriptive statistics provide you with useful information about location, dispersion and symmetry in your data, data distributions provide the same information in fuller and usually visual form.
  - In its simplest and least complicated form, data (in either discrete or continuous form) can be converted into a histogram or a bar chart.
  - That histogram/chart provides the same information that data descriptive statistics provide
- When the distributions approach or resemble classic statistical distributions, you gain the benefit of being able to draw on their established distributional properties to extrapolate from the data.

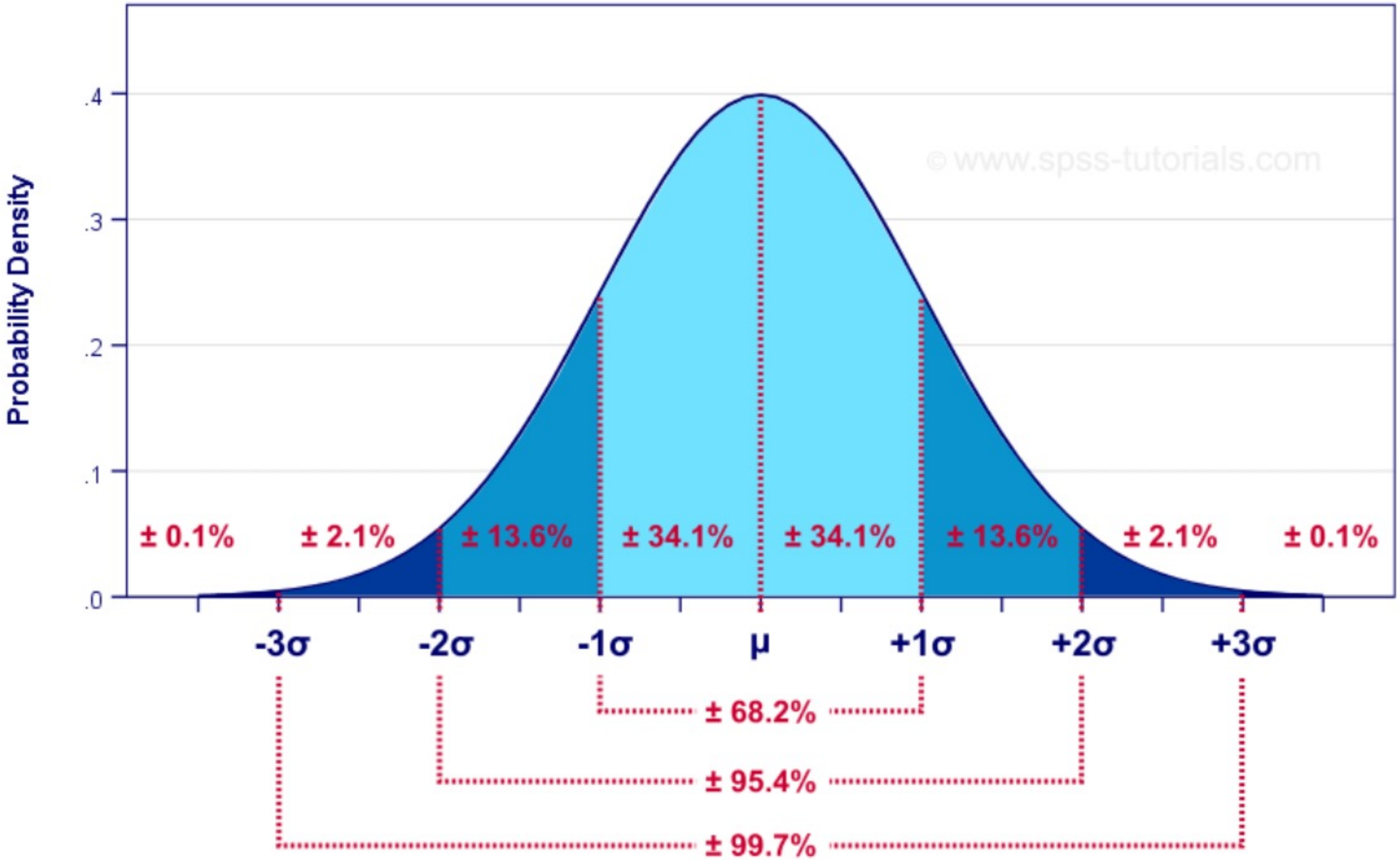
Histogram: Visualizing Data



Statistical Distributions

- While histograms summarize data in visual form, there is an advantage to replacing them with one of many standardized statistical distributions (normal, exponential, uniform etc.).
  - You have to find a distribution that best fits your data in terms of symmetry, skewness and tails.
  - The benefit of using a standardized distribution is that you can draw on their established distributional properties to extrapolate from the data or to make probabilistic statements relating to the data.

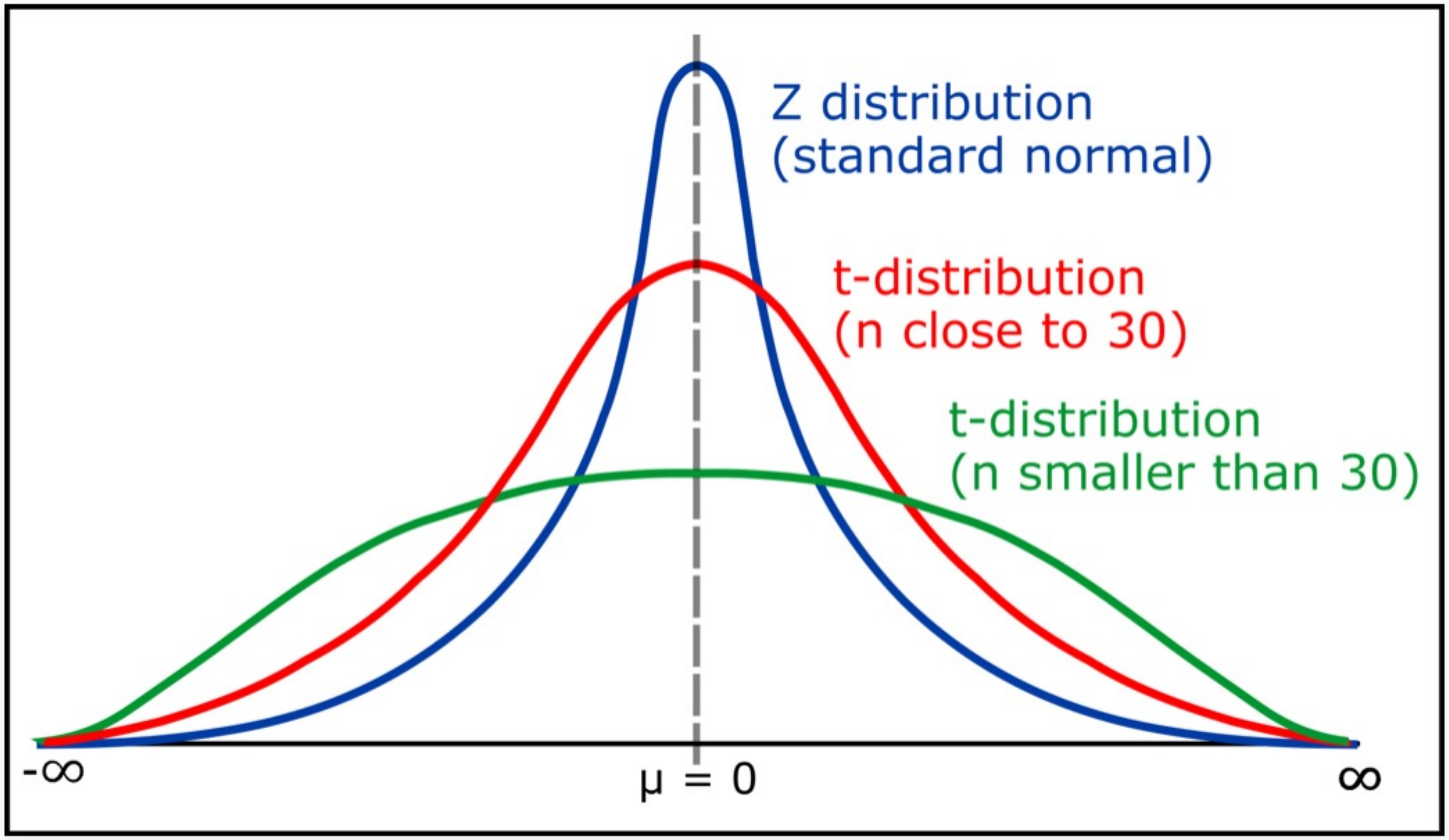
The Normal Distribution



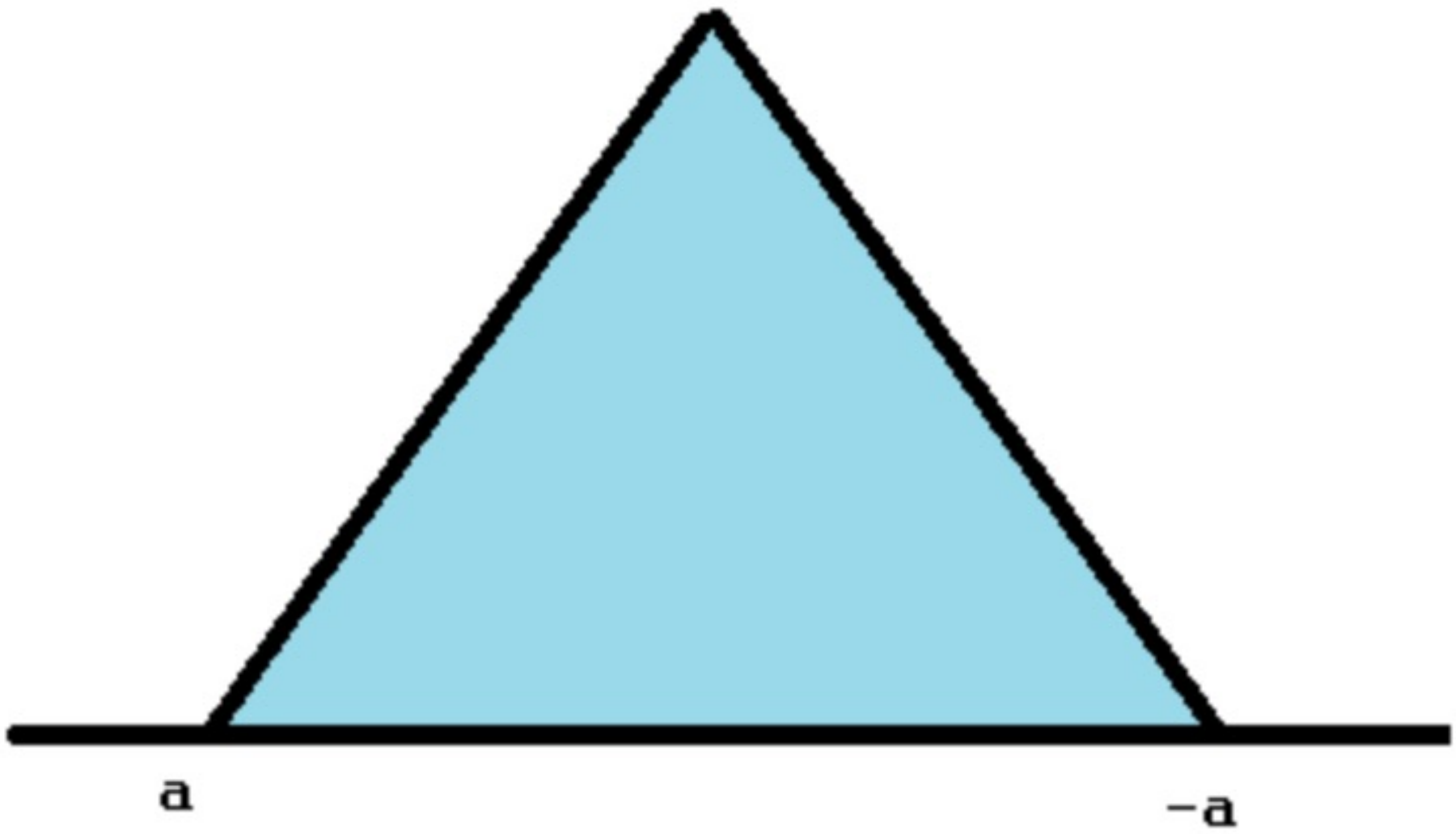
And its properties…

- Dispersion: The standard deviation in a normal distribution measures the spread around the mean, but with links to probabilities of a number occurring in the sample falling within or out of that spread.
- Skewness: A normal distribution is symmetric and has no skewness.
- Kurtosis: A variable that is normally distributed can take on values from minus infinity to plus infinity, but the likelihood of extreme values is constrained. The kurtosis for a normal distribution is three, which becomes the standard against which other distributions are measured.

A Close Relative: The t distribution



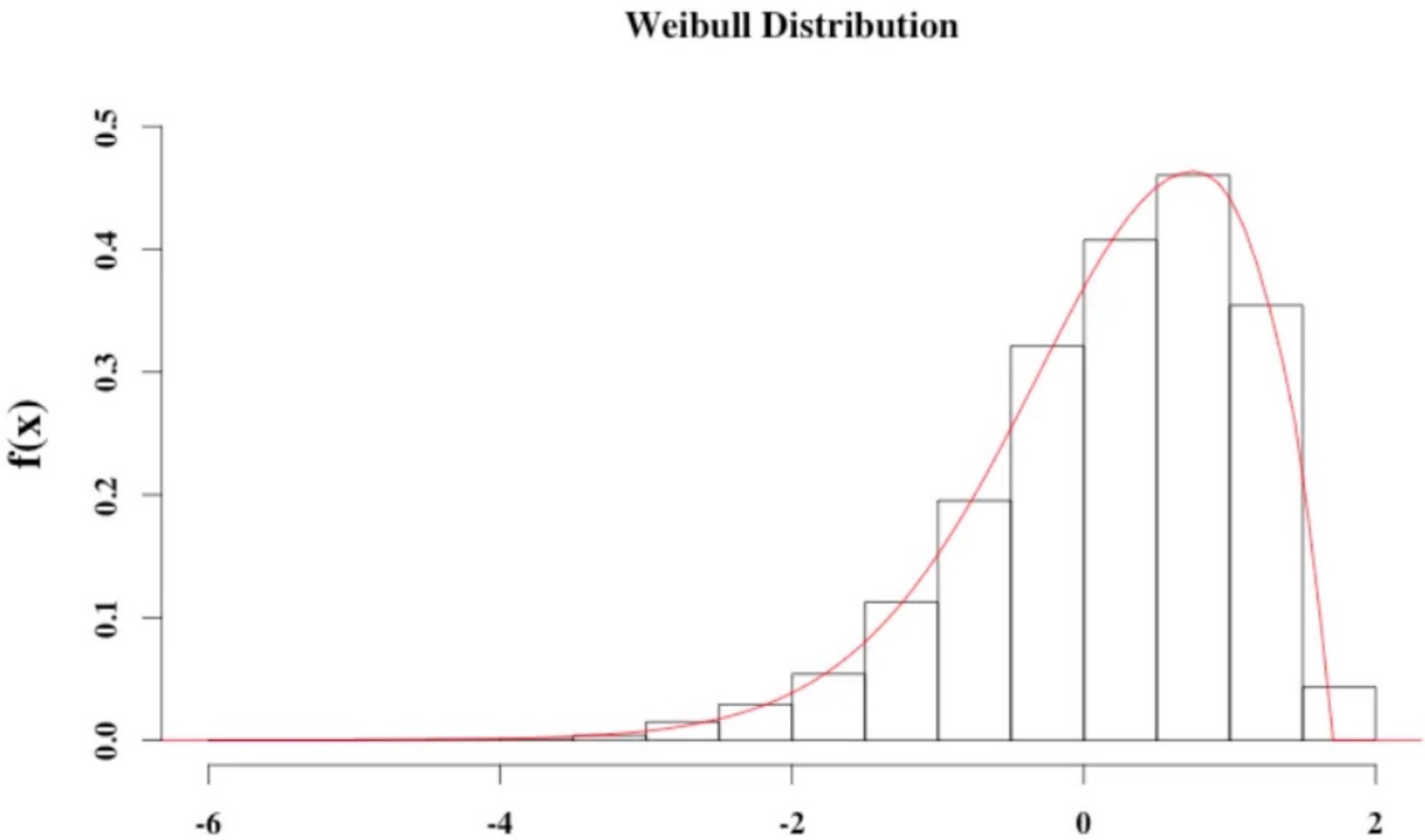
A more distant relative: The (Symmetric) Triangular Distribution



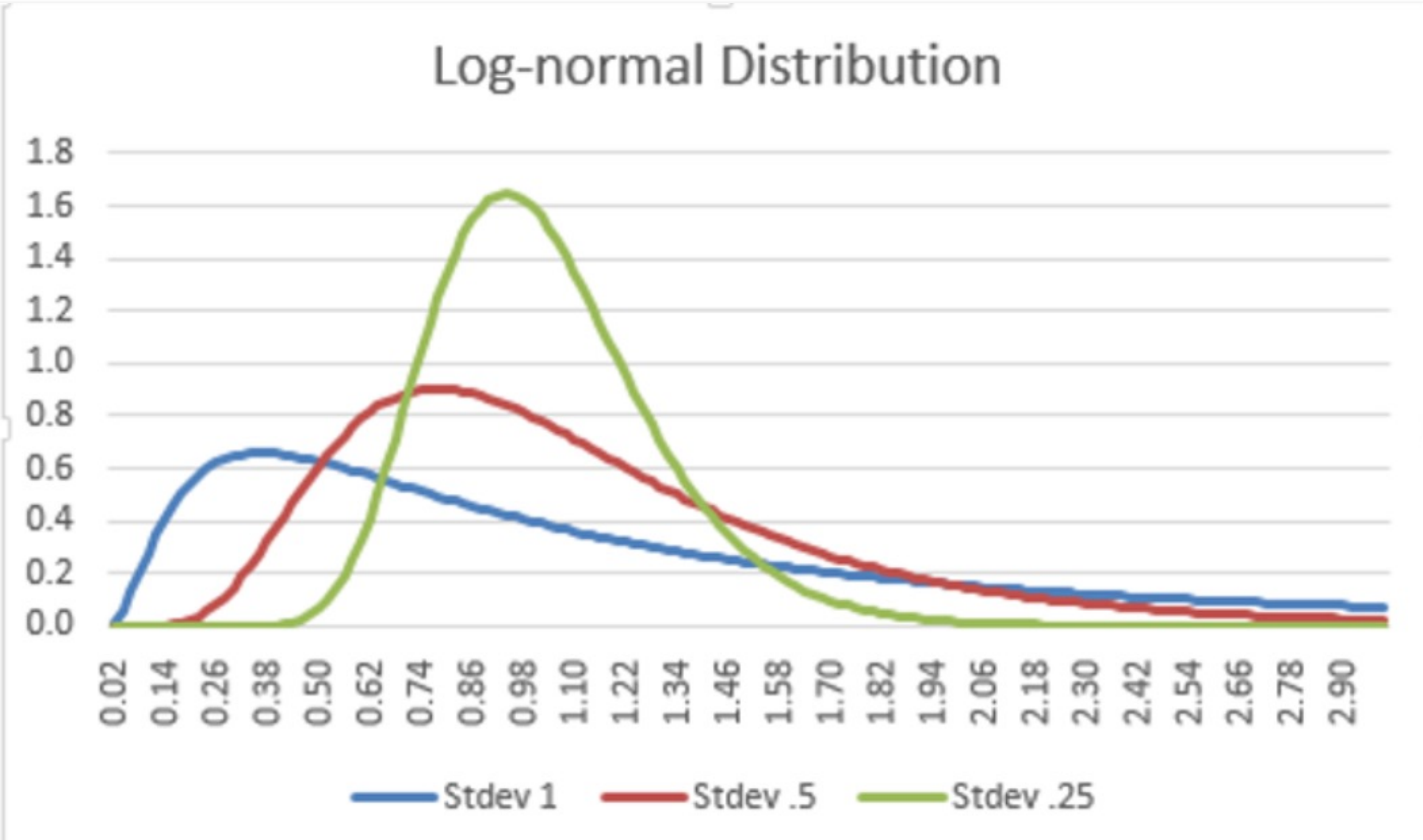
A Uniform Distribution

$$mean : \mu = E(X) = \frac{b+a}{2}$$

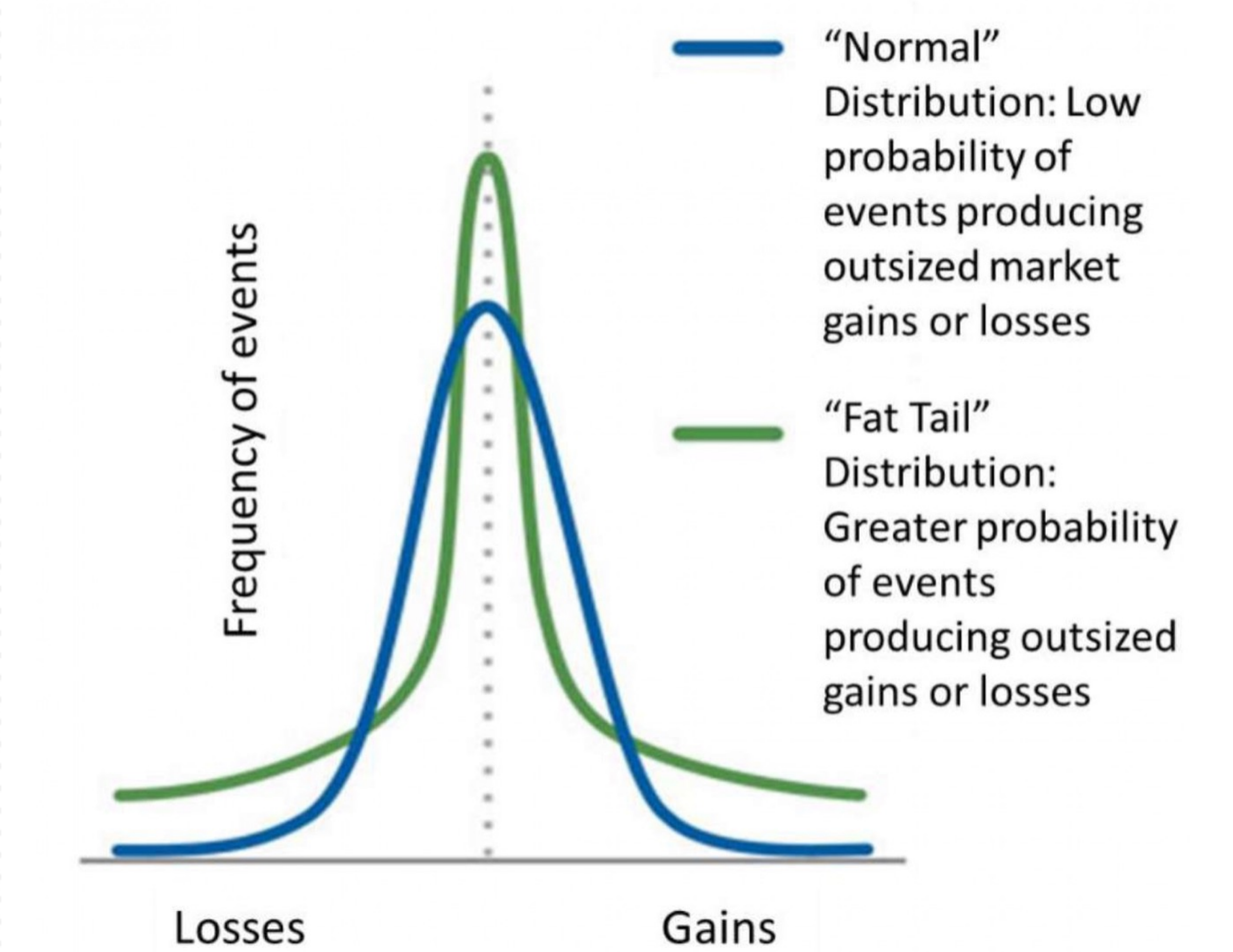
Negative Skew: Minimum Extreme Value



Positive Skew: Log Normal Distribution

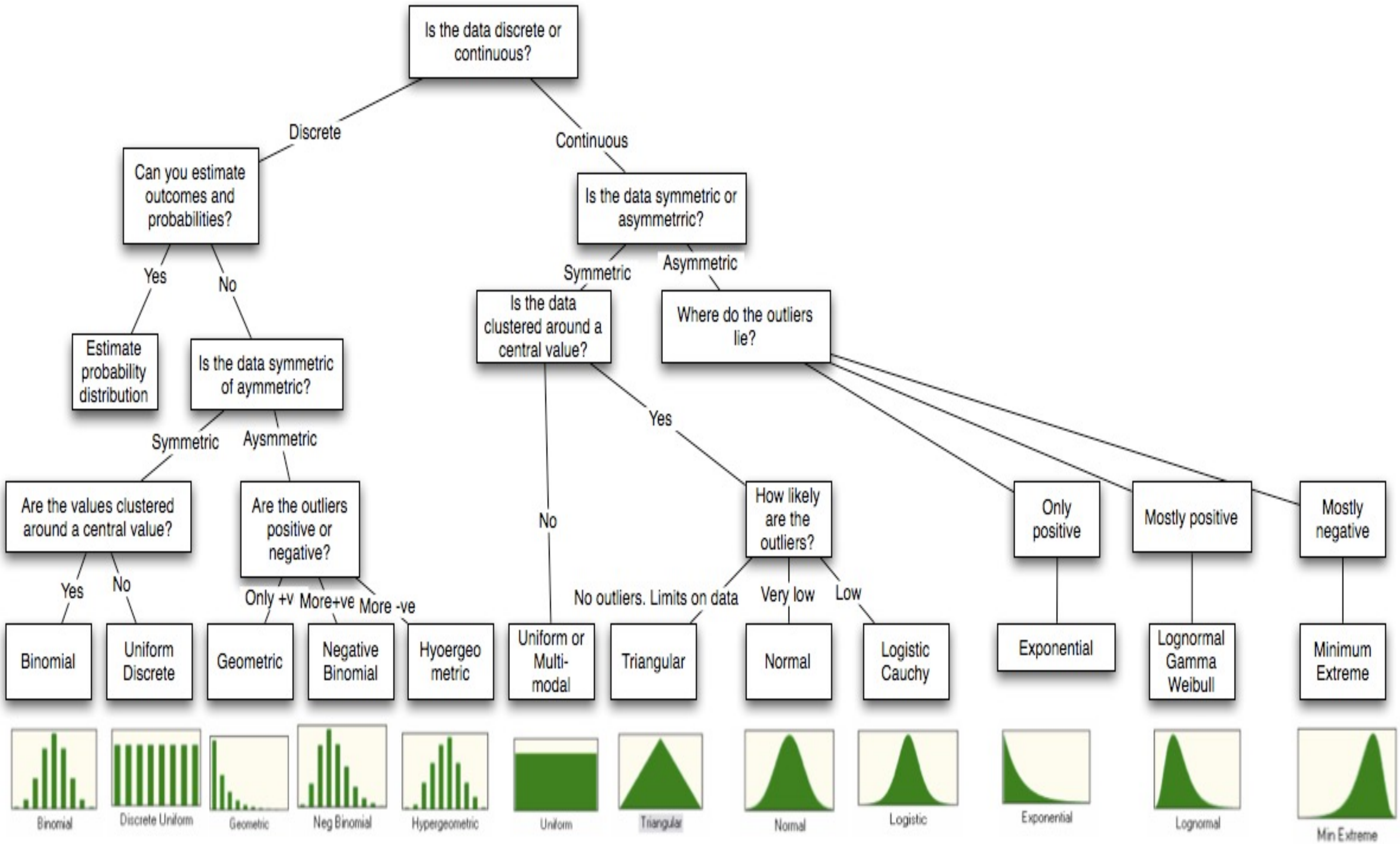


Thin tails and Fat tails…



Measured with kurtosis…

- Kurtosis is a measure of the combined weights of the tails, relative to the rest of the distribution.
- Most often, **kurtosis** is measured against the **normal distribution**. Pearson’s kurtosis is the excess kurtosis over three
  - If the Pearson **kurtosis** is close to 0, then a **normal distribution** is often assumed. These are called mesokurtic **distributions**.
  - If the Pearson **kurtosis** is less than 0, then the **distribution** has thin tails and is called a platykurtic **distribution**. (Uniform distribution is a good example)
  - If the Pearson **kurtosis** is greater than 0, then the **distribution** has fat tails and is called a leptokurtic **distribution**.



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