Probability Distributions

Random Variables

A **random variable** is a variable that can take on more than one value.
When dealing with random variables, it is often useful to think about the **sample space** for that variable.

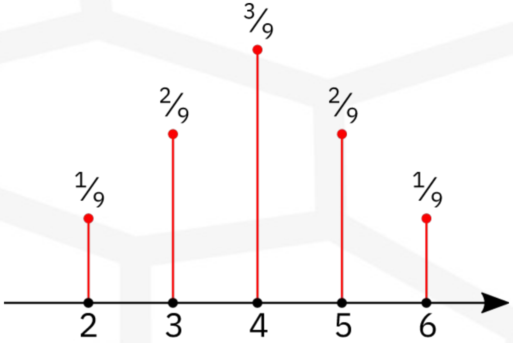
Examples

- Rolling a die
- Height of men

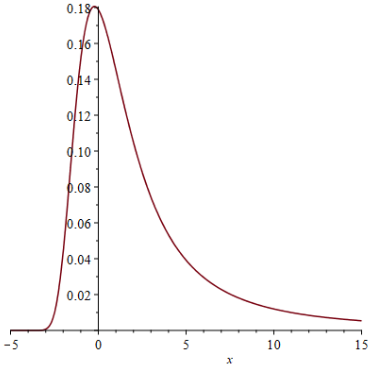
A **probability distribution** is a function that provides the likelihood of seeing each value in a random variable's sample space.

Discrete vs. Continuous Probabilities

Discrete random variables are defined by a **probability mass function.**



Continuous random variables are defined by a **probability density function**.



Why Are Probability Distributions Important in Data Science?

- Probability lets us model uncertainties.
- Probability lets us make statements about a population.
- Many ML methods depend on distributional assumptions
  - Linear Regression assumes the residuals are normally distributed
  - Gaussian Naive Bayes assumes the predictors are normally distributed
- Probability distributions can be useful to create ‘toy’ datasets to test out methods.

Common Probability Functions

Discrete

- Bernoulli
- Binomial
- Poisson
- Uniform

Continuous

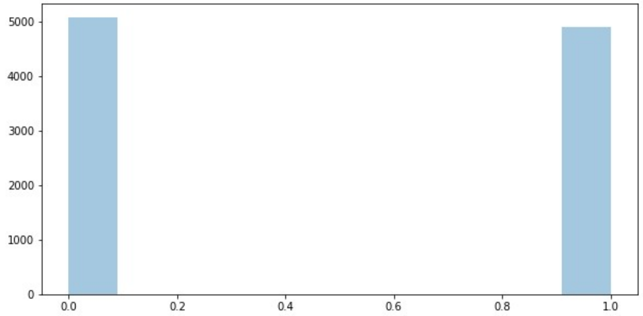
- Normal
- Lognormal
- Beta
- Uniform

Probability functions have a functional form that is defined by a set of parameters.

Discrete: Bernoulli & Binomial

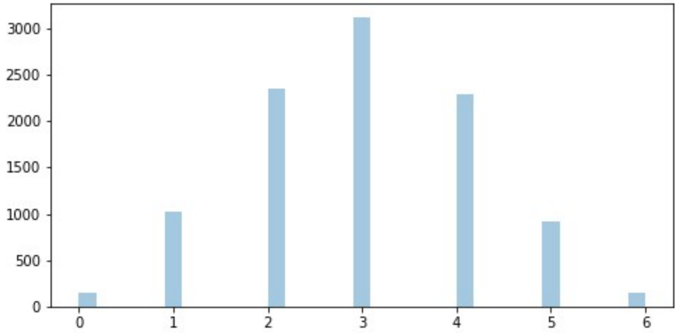
- **Bernoulli distribution**: probability of a random variable with two possible outcomes
- Example: flipping a coin

$$f(x) = p^x(1-p)^{1-x}$$



- **Binomial distribution**: probability of repeated trials of a random variable with two possible outcomes.
- Example: number of heads when flipping a coin 6 times.

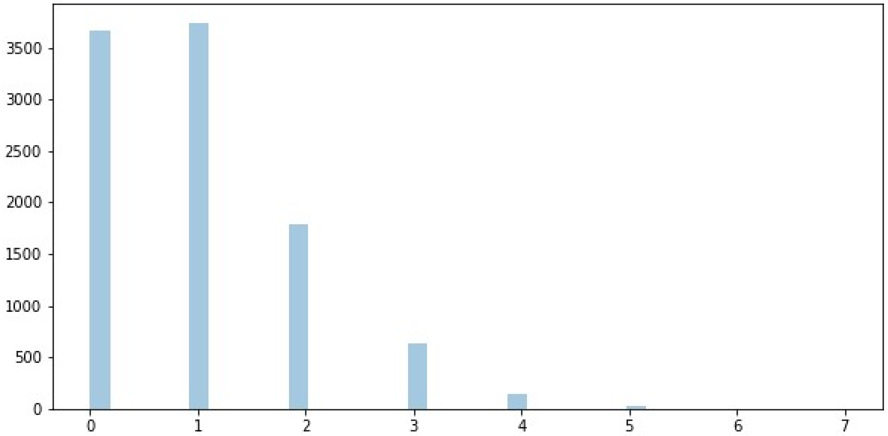
$$f(x) = \binom{n}{x} p^x(1-p)^{n-x}$$



Discrete: Poisson

- **Poisson** models the number of events in a given time period.
- Example: the number of car accidents in a day.

$$f(x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

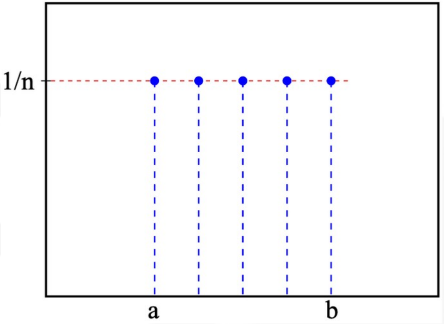


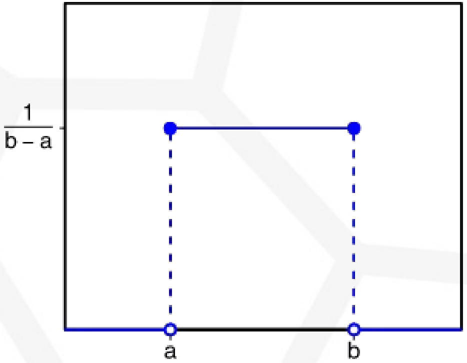
Lambda is the average number of events in the given time period.

Discrete & Continuous: Uniform

- **Uniform** models a random variable with equal probability across its sample space.
- Can be continuous or discrete.
- Example: drawing from a deck of cards

$$f(x) = \frac{1}{B - A}$$



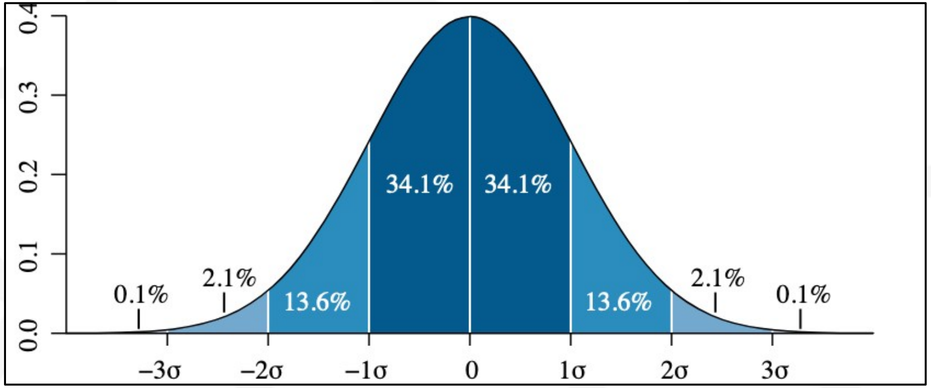


Continuous: Normal Distribution

- Completely determined by mean and standard deviation
- Also called Gaussian distribution
- Mean, median and mode are the same
- Unbounded
- Symmetric

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

μ is the sample mean and
σ is the sample standard deviation

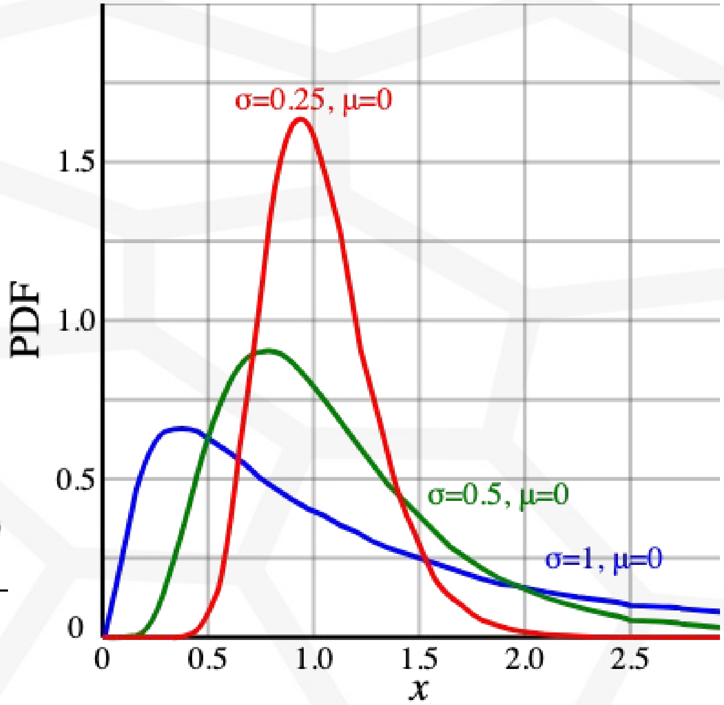


Normal (Gaussian) Distribution

Continuous :Lognormal

- Completely determined by mean and standard deviation
- Used for positive, skewed data
- Example: amount of rainfall, length of comments posted on internet forums

$$f(x) = \frac{e^{-((ln((x-\theta)/m))^2/(2\sigma^2))}}{(x-\theta)\sigma\sqrt{(2\pi)}}$$

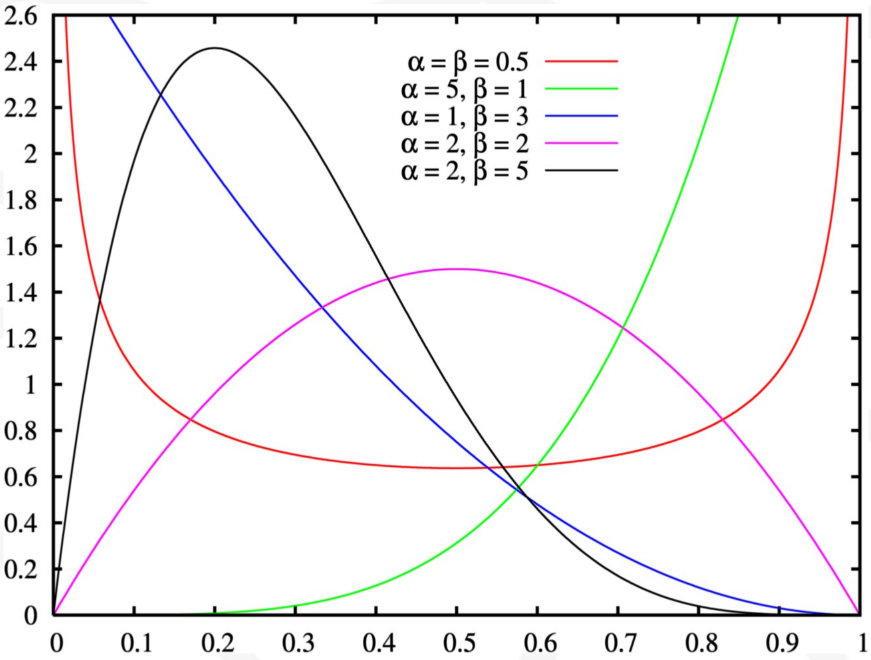


Continuous: Beta

- Bounded by [0,1]
- Very flexible in shape
- Often used to model probabilities

$$f(x) = \frac{(x)^{p-1}(1-x)^{q-1}}{B(p,q)}$$

p and q are shape parameters



How do you choose a probability distribution?

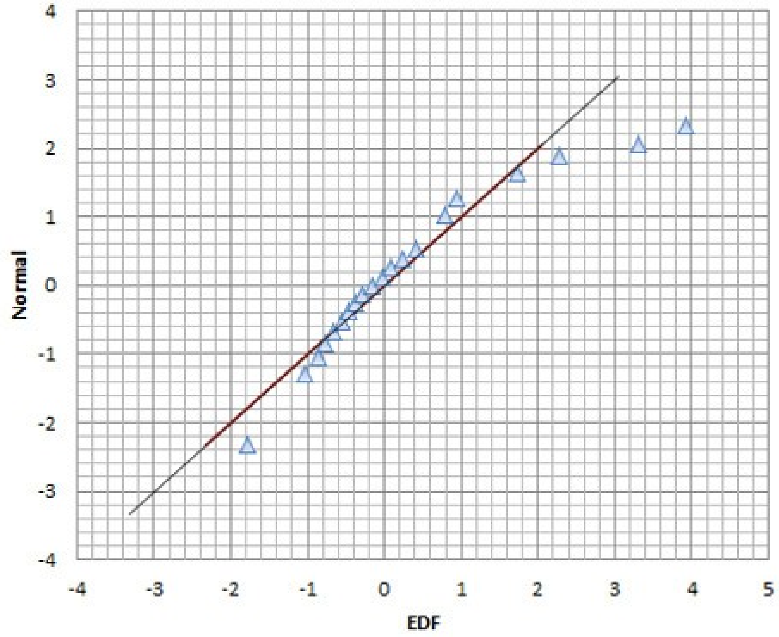
Ask yourself the following questions:

- Is your data continuous vs discrete
- Is your data symmetric or asymmetric?
- Is your data bounded (aka finite support) or unbounded?
- How likely are outliers?

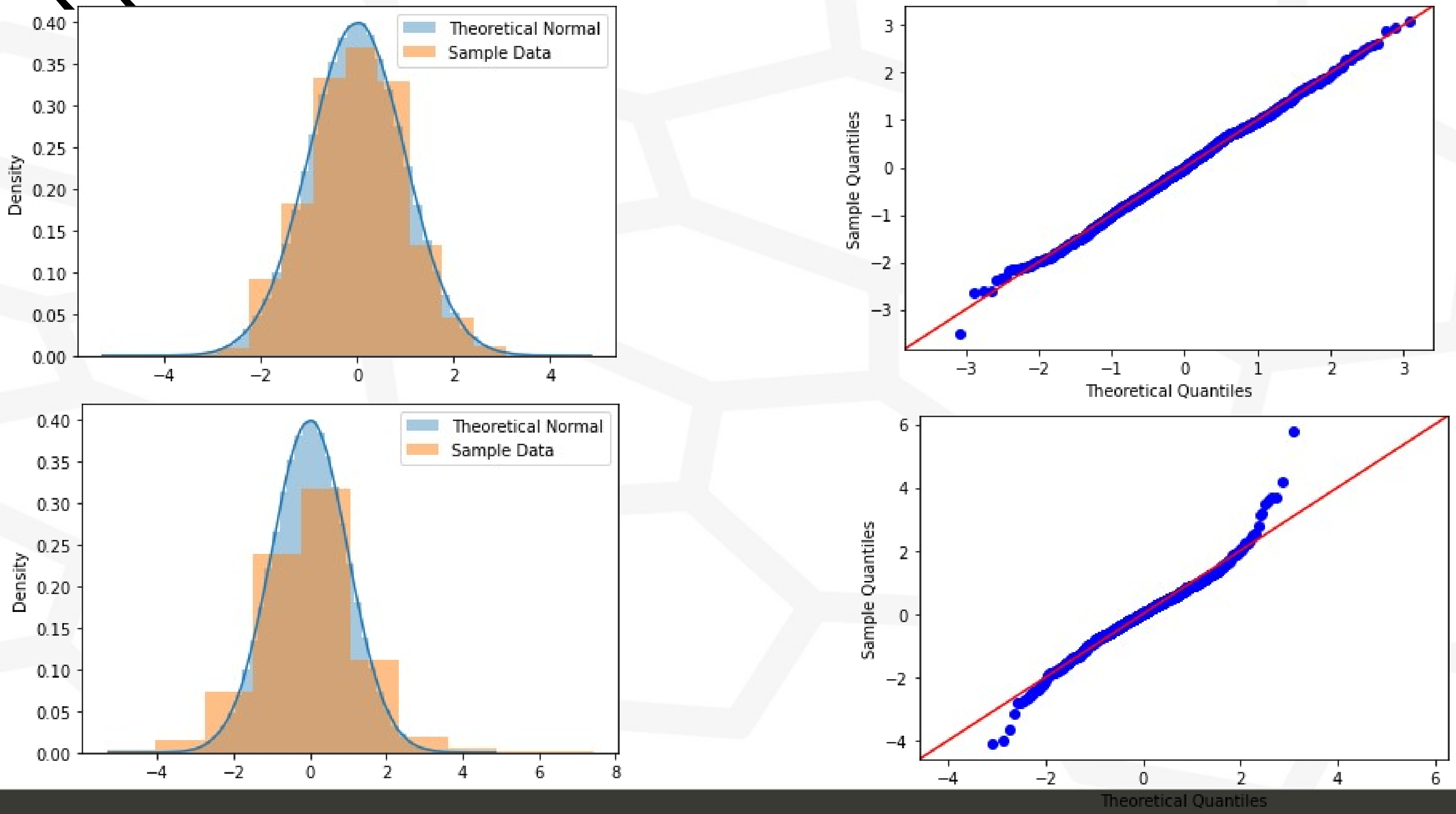
Warning: The results of an analysis can depend heavily on the choice of probability distribution. Therefore, you should take care in choosing the distribution if the results are sensitive to the choice.

Assessing Distribution Fit to Data

At times, you may want to fit a probability distribution to your data. To assess the fit, you can use a Quantile-Quantile (QQ) Plot.



QQ Plots


Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
Theoretical Normal
Sample Data
Sample Quantiles
3
2
1
0
−1
−2
−3
Theoretical Quantiles
Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−4
−2
0
2
4
6
8
Theoretical Normal
Sample Data
Sample Quantiles
6
4
2
0
−2
−4
Theoretical Quantiles