Probability Distribution

Digya Acharya

OUTLINE:

- General concepts of probability distribution
- Discrete Probability Distribution:
  - Bernoulli, Binomial, Categorical, Multinomial
- Continuous Probability Distribution:
  - Gaussian, Beta, Dirichlet, Gamma
- Bayes Theorem:
  - Likelihood, Prior, Posterior

Why do we study Probability?

Why study Probability?

- Most phenomena in a physical world cannot be described completely with deterministic formulas and are uncertain.
- Predict sequence of events.
- Immense importance in decision making and planning.

Example: - finding the chance of raining today.
- estimating the outcome of a coin flip.
- probability of winning lottery.
- chance of loosing bike keys.

General concepts of probability distribution

- **Probability**

- how likely something is to happen
- Two approaches to probability: Bayesian and Frequentist
- Frequentist -> suppose we tossed a coin 100 times, it came up heads 50 times so, the probability of heads is 0.5.
- Bayesian means probabilistic

start with a belief -> prior -> obtain data -> update belief -> posterior ->
new prior = old posterior

General concepts of probability distribution

- **Random variable**

- real-valued function of the experimental outcome.

- In a coin flip experiment, if we are assuming head=0 and tail=1, our random variable "X" is:

$$X = \begin{cases} 0 & (head) \\ 1 & (tail) \end{cases}$$

Why do we study Probability Distribution?

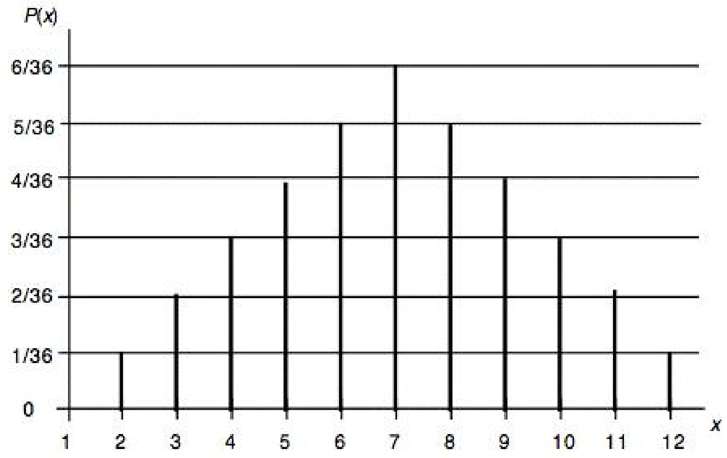
- Knowledge of sampling distribution can be very useful in making inferences about the overall population.
- Can be used to model any event. Eg: Exponential distribution to model natural disasters.

General concepts of probability distribution

- **Probability Distribution**

- provides the probabilities of occurrence of different possible outcomes in an experiment
- Types of probability distribution: Discrete and Continuous
- Discrete probability functions (probability mass functions) assumes a finite and discrete number of values. For example, coin tosses and counts of events are discrete functions.
- Continuous probability functions (probability density functions) assumes an infinite number of values between any two values. Continuous variables are often measurements on a scale, such as height, weight, and temperature.

General concepts of probability distribution



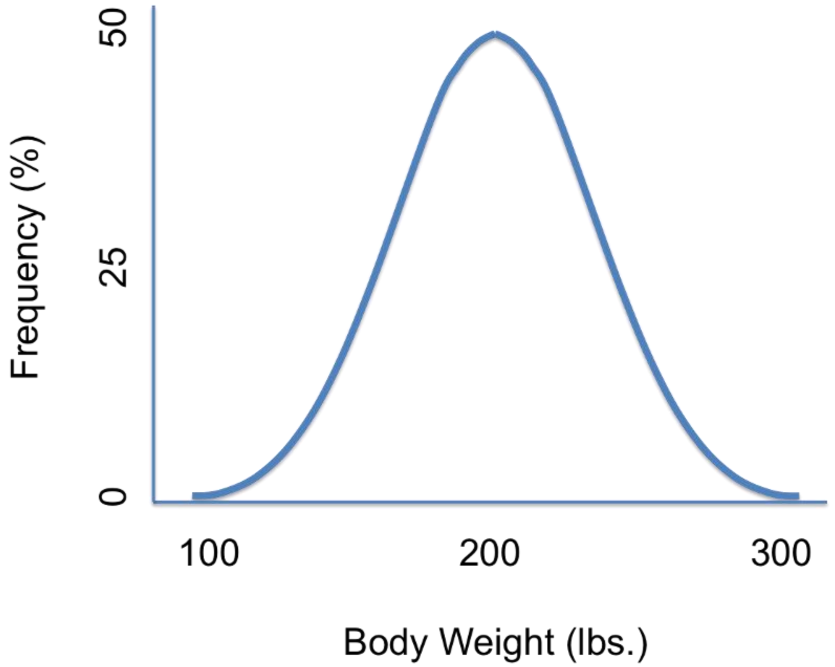


Figure 1: Discrete (left) and Continuous (right) probability distribution

Discrete Probability Distribution:

- **Bernoulli distribution:**

- Suppose ‘x’ represents the outcome of flipping an unfair coin, with ‘x=1’ representing heads and ‘x=0’ representing tails. Then, the probability of obtaining heads is given by:

$$p(x = 1|\mu) \; = \; \mu$$

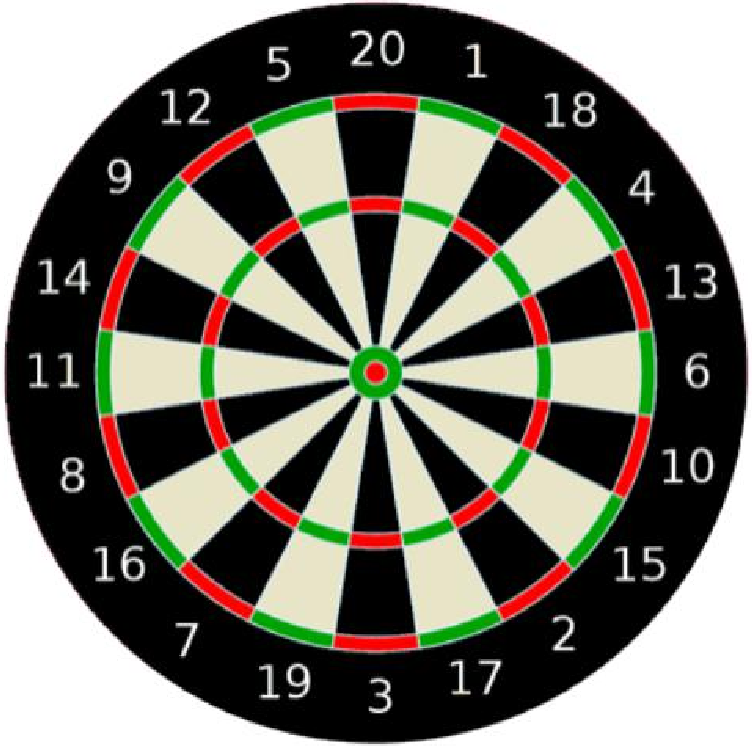
and probability of obtaining tail is given by:

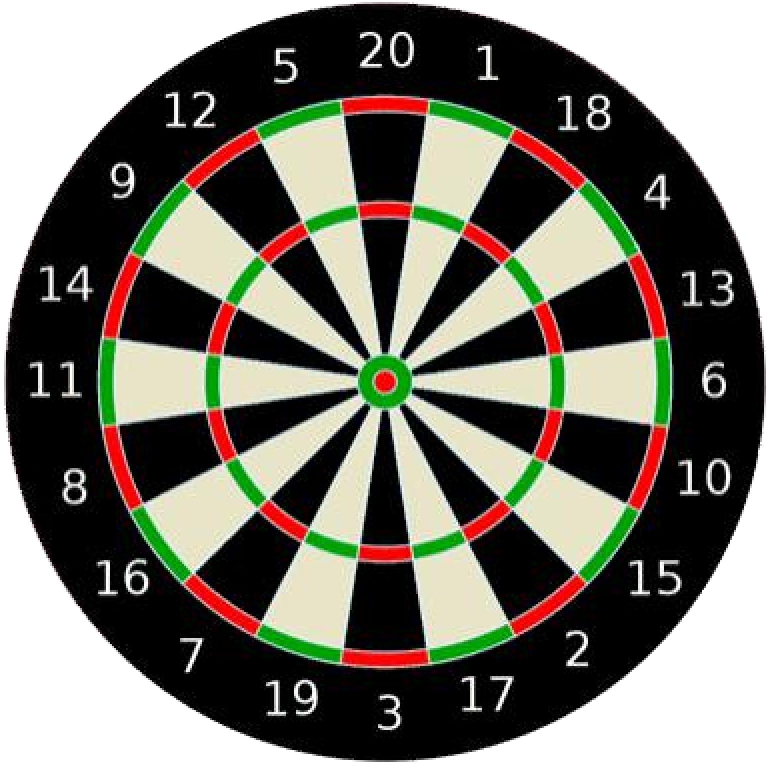
$$p(x = 0|\mu) \; = \; 1 - \mu$$

Jacob Bernoulli

Here, ‘μ’ represents the biasedness of obtaining head in the coin and $0 \leq \mu \leq 1$. The probability distribution over ‘x’ can be written as:

$$Bern(x|\mu) = \mu^{x}(1 - \mu)^{1 - x}$$


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5



[Code]

Discrete Probability Distribution:

- **Binomial distribution:**

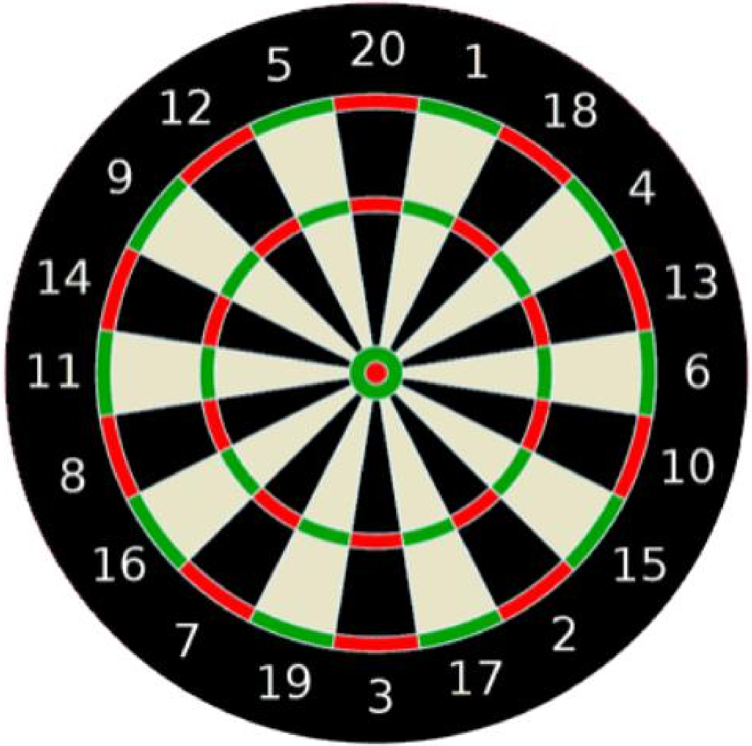
- Suppose ‘x’ represents the outcome of flipping an unfair coin, with ‘x=1’ representing heads and ‘x=0’ representing tails.

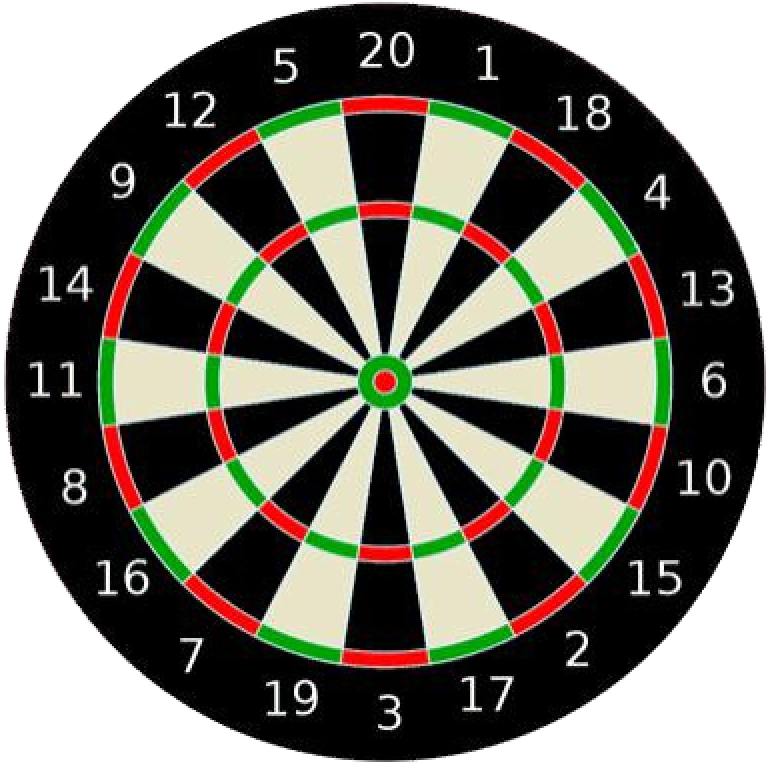
Then, the probability of obtaining heads and tails in each experiment is given by (as in Bernoulli distribution):

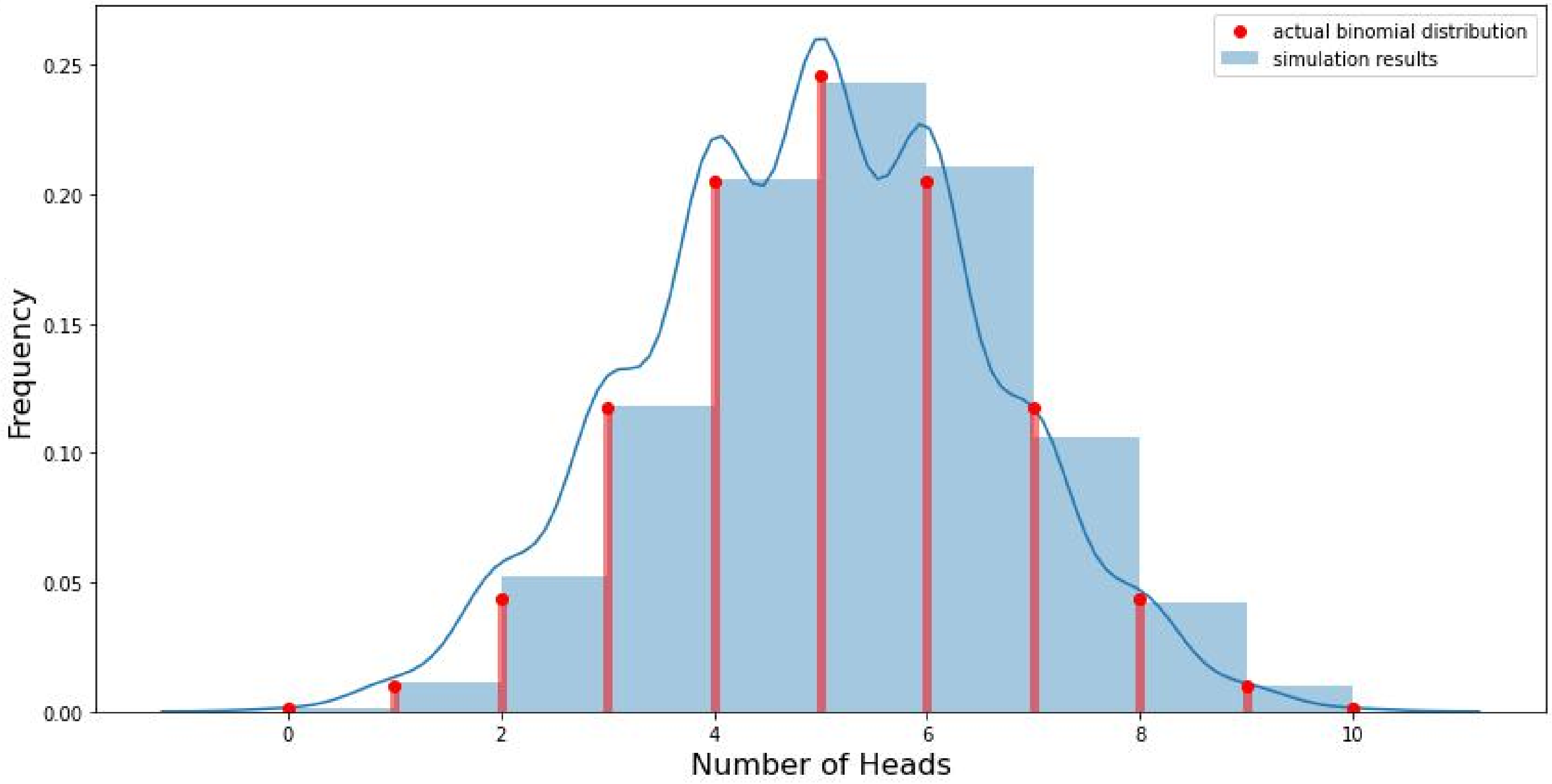
$$p(x = 1|\mu) = \mu$$
$$p(x = 0|\mu) = 1 - \mu$$

If ‘N’ is the total number of times the coin is flipped and ‘m’ is the number of times the head has occured. Then, the probability distribution can be written as:

$$Bin(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N - m}$$


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5



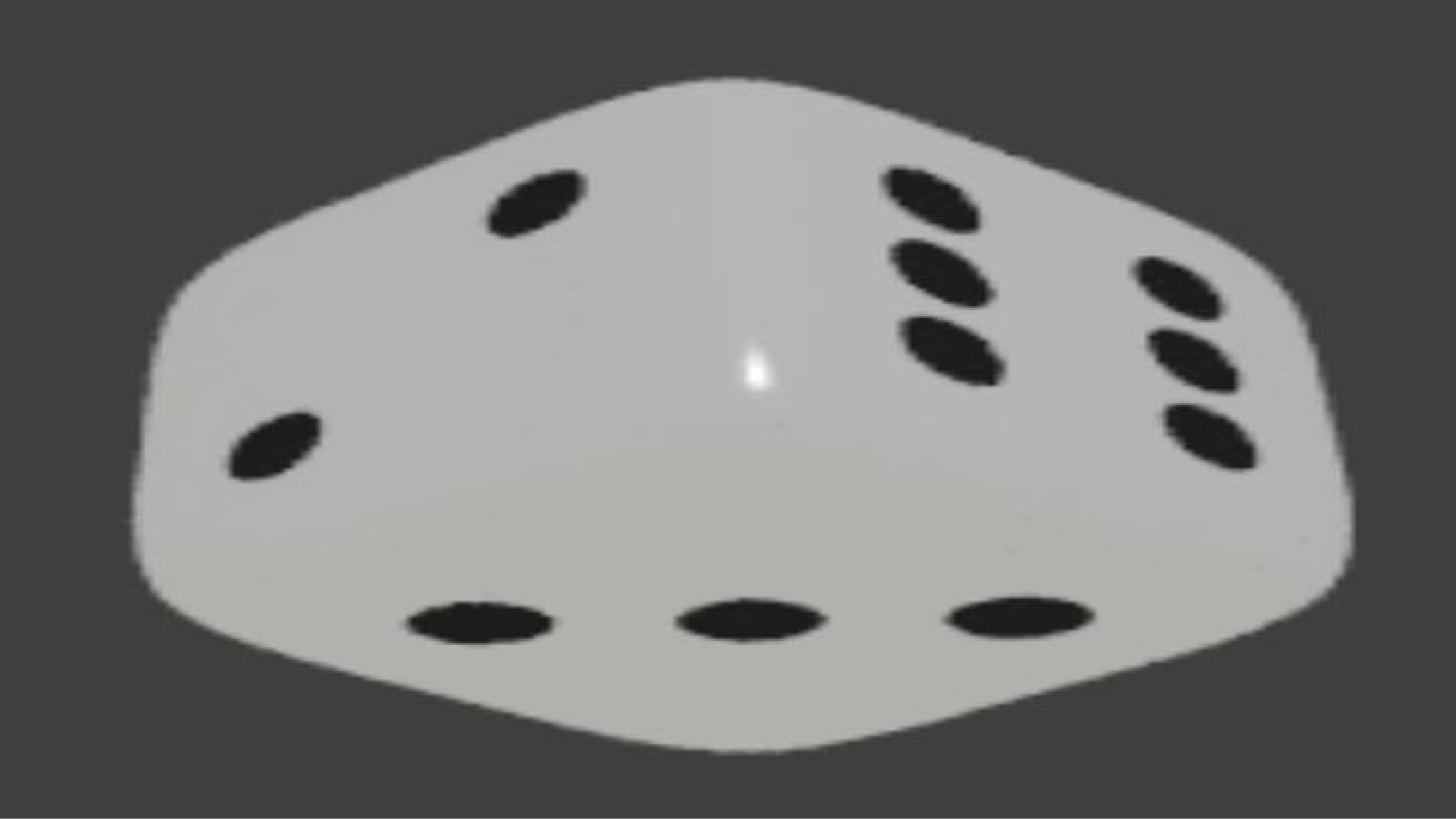
Code

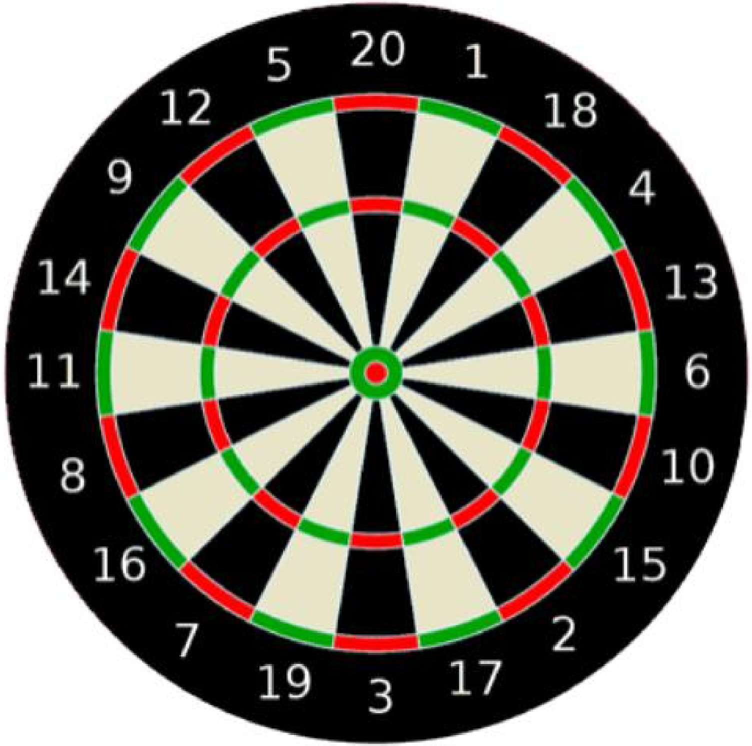
Figure 2: Simulated and actual binomial distribution while tossing a fair coin with number of trials = 1000 and 10 experiments in each trial

Discrete Probability Distribution:

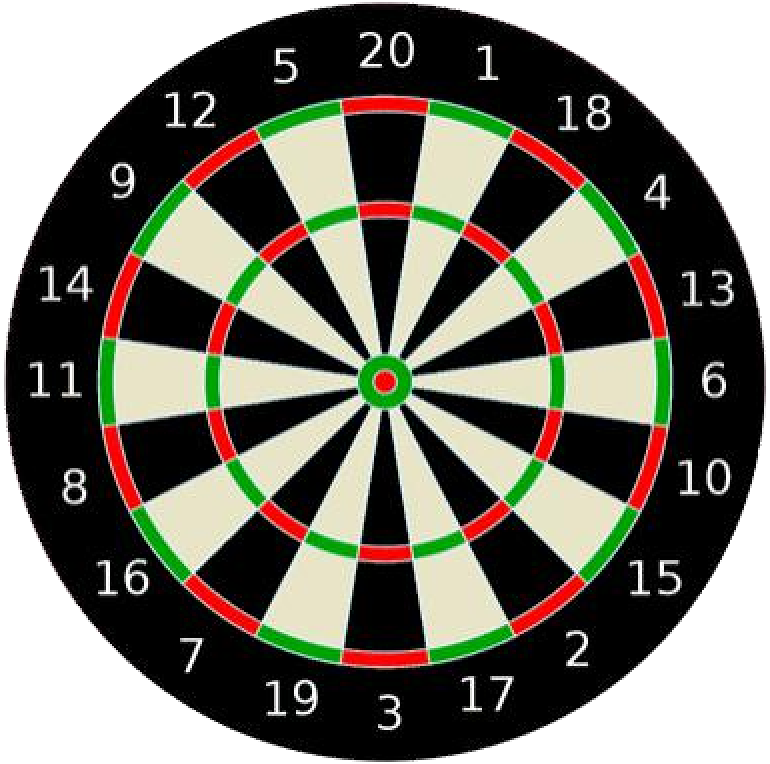
- **Categorical distribution:**

- generalization of the Bernoulli distribution for a categorical random variable.
- contains multiple possible outcomes.
- For example, a dice roll, where there are six outcomes {1,2,3,4,5,6} is a categorical distribution.





20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

Discrete Probability Distribution:

- **Multinomial distribution:**

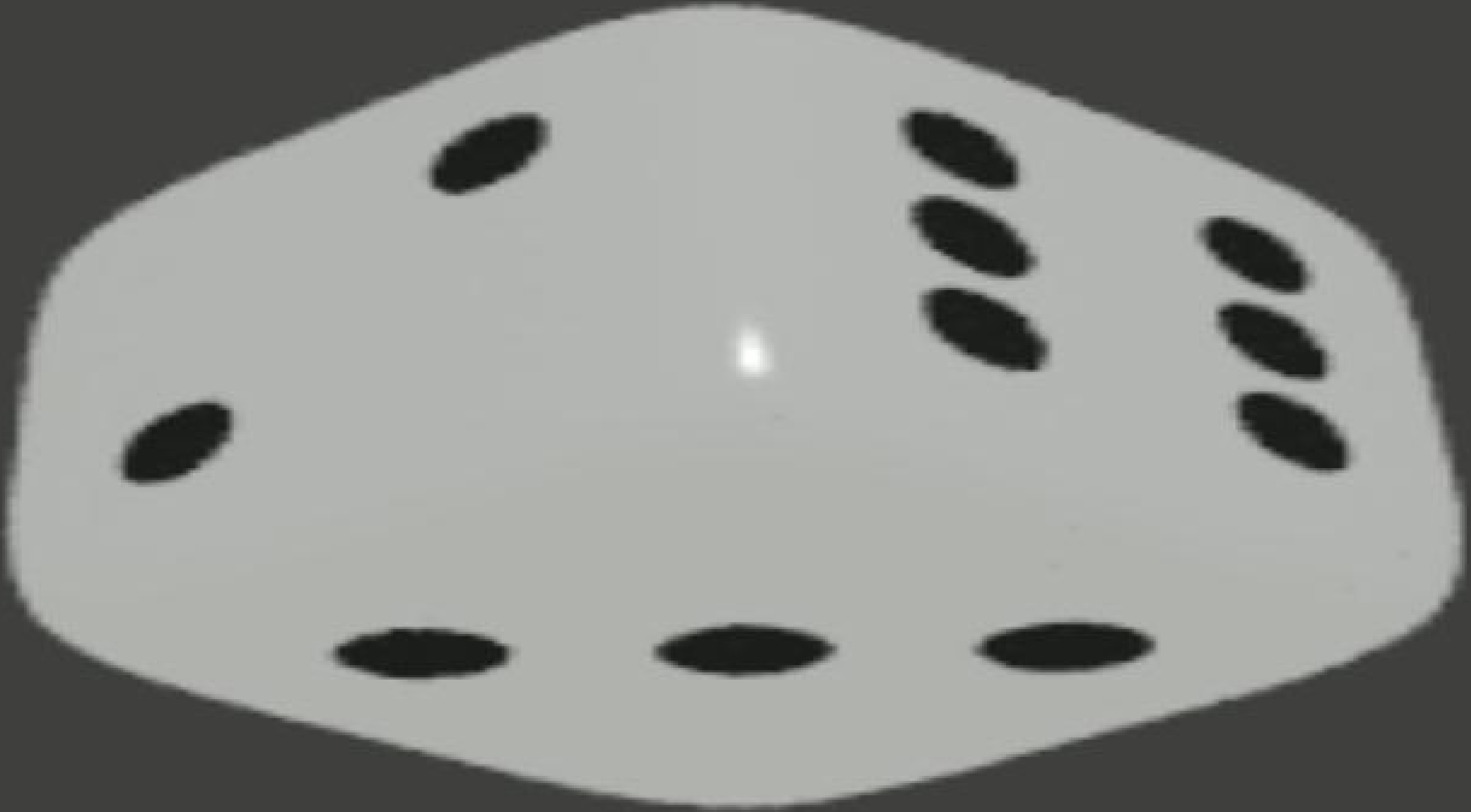
Suppose 'c' represents the number of outcome of rolling an unfair dice. If $C = \sum_i c_i$ then, the probability distribution over 'c' can be written as:

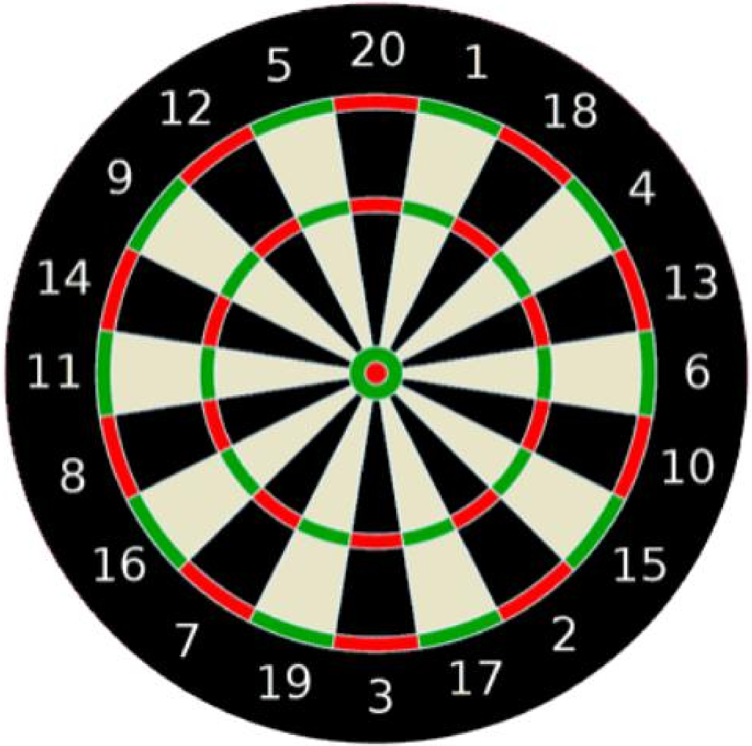
$$p(roll = c|\theta) = \frac{C!}{\prod_i c_i!} \prod_i \theta_i{}^{ci}$$

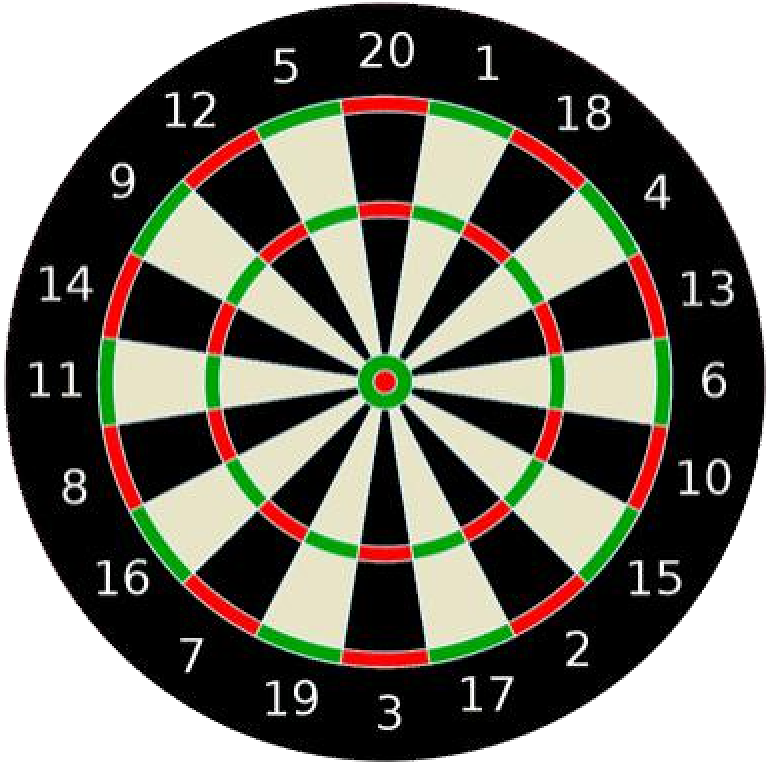
If we run an experiment in which we roll dice 500 times then, C = 500, c (count of each side) = $[c_1,...,c_6]$ and $\theta(\text{prob. of each side}) = [\theta_1, ..., \theta_6]$. Here, $\sum_i \theta_i = 1$

- Generalizes to binomial distribution when i>2.

- Equivalent to binomial distribution for i=2.




20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5


20
1
18
4
13
6
10
15
2
17
3
19
7
16
8
11
14
9
12
5

Continuous Probability Distribution:

- **Normal Distribution:**

- also known as the Gaussian distribution
- symmetric about the mean
- shows that data near the mean are more frequent in occurrence than data far from the mean
- The pmf of univariate Gaussian distribution is given by:

$$N(x|\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Here, $N(x|\mu, \sigma^2) > 0$ and $\int_{-\infty}^{\infty} N(x|\mu, \sigma^2)\mathrm{dx} = 1$

Carl Friedrich Gauss

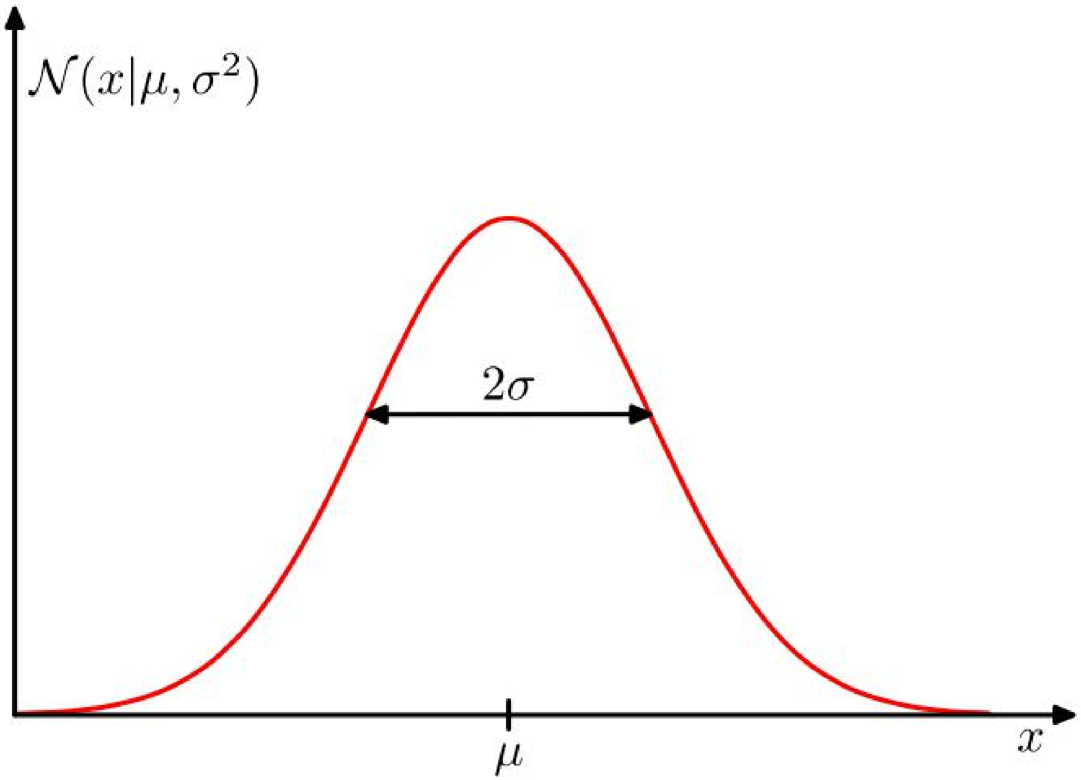


Figure 3: Univariate Gaussian Distribution

Continuous Probability Distribution:

Gamma Function

- Return the factorial of numbers.
- Interpolate the factorial of decimal values.
- For any positive integer, gamma function is defined by:

$$\tau(N) = (N-1)!$$

- For complex numbers with a positive real part, the gamma function is defined via a convergent improper integral

$$\tau(z) = \int_0^\infty x^{z-1} e^{-x} dx,$$ where, z is positive real part complex no.

Some properties of gamma function for N > 0:

- $\tau(N+1) = N!$
- $\tau(N+1) = N\tau(N)$

Code

Gamma Function

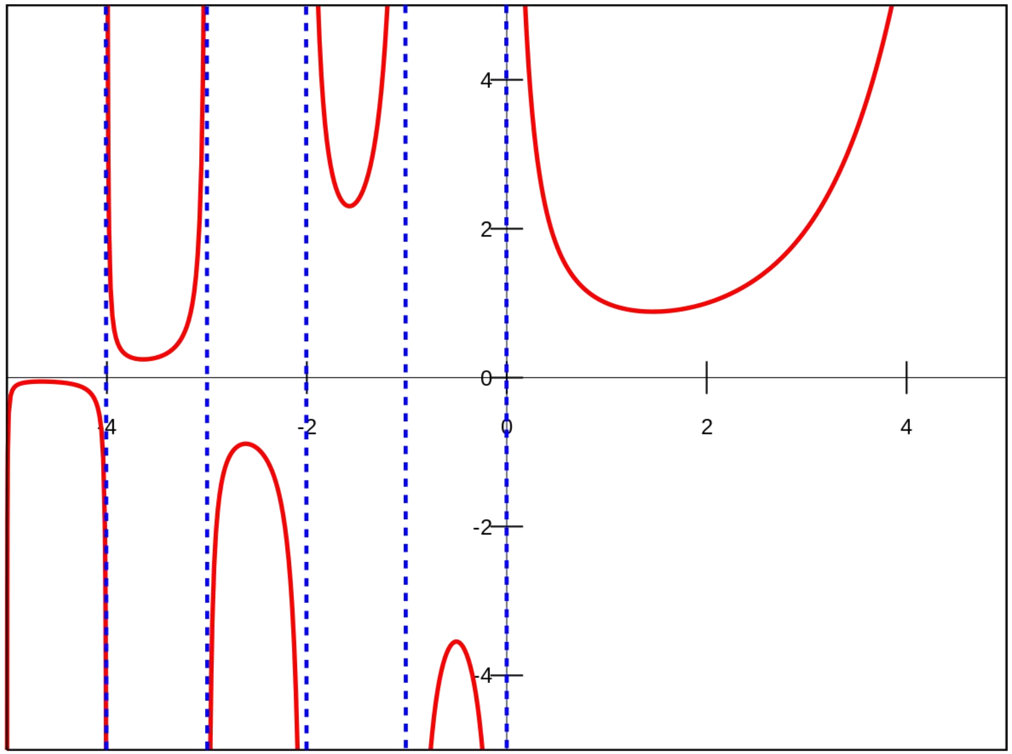


Figure 4: Gamma function along real part of the axis (Source: Wikipedia)

Continuous Probability Distribution:

- **Beta distribution:**

- Probability distribution on probabilities
- Suppose we have to model the probability of picking a coin with certain bias from a room full of coins.

Here,
Input → Bias of a coin which is a probability, i.e, a number between 0 and 1.
Output → Probability of picking a certain coin

Continuous Probability Distribution:

The beta distribution is given by:

$$p(\theta|\alpha, \beta) = \frac{\theta^{\alpha - 1}(1-\theta)^{\beta - 1}}{B(\alpha, \beta)} \sim Beta(\alpha, \beta)$$

Here, ‘θ’ is the bias, alpha and beta are the shape parameters.

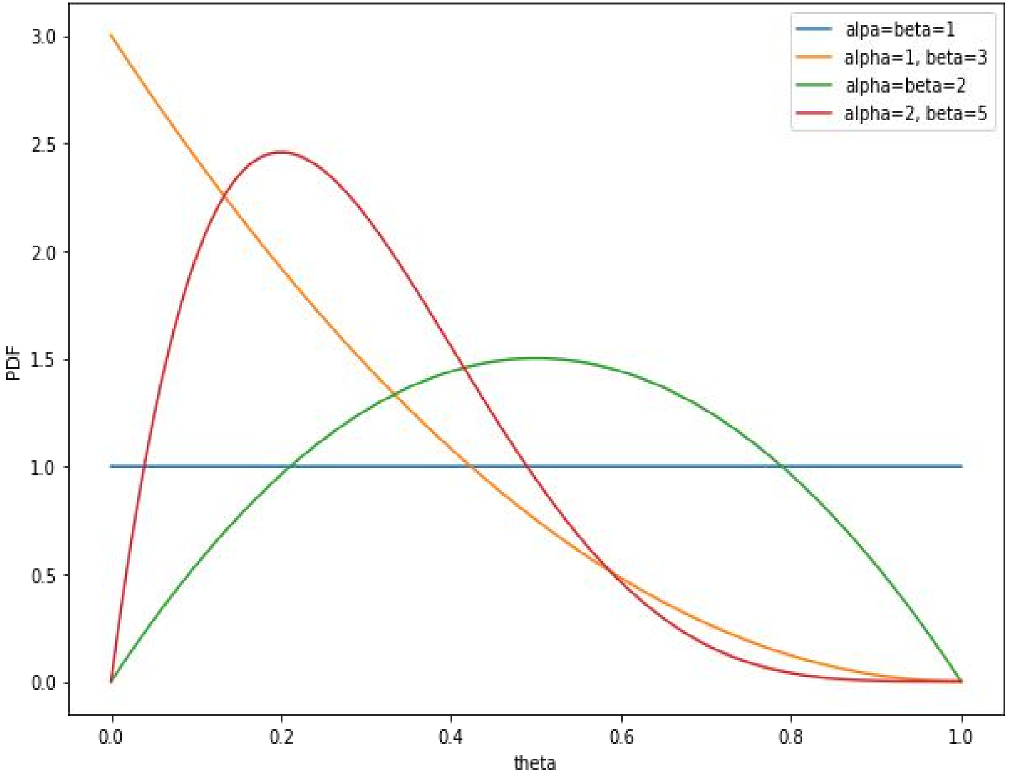
$$B(\alpha, \beta) = \int_0^1 \theta^{\alpha - 1}(1-\theta)^{\beta - 1} d\theta$$

$$= \frac{\tau(\alpha)\tau(\beta)}{\tau(\alpha+\beta)}$$ is a normalizing factor (1)

The coefficient in equation (1) ensures that the beta distribution is normalized, i.e,

$$\int_0^1 p(\theta|\alpha, \beta) = 1$$

Shape of Beta Distribution for different values



For α=β=1, the distribution is uniform with $p(\theta|\alpha, \beta)$=1. Thus, beta distribution is a generalization of uniform distribution.

[Code]

Figure 5: Pdf of beta distribution for different values of alpha and beta

Continuous Probability Distribution:

- **Dirichlet distribution:**

The dirichlet distribution is given by:

$$p(\theta|\alpha) = \frac{\tau(A)}{\prod_i \tau(\alpha_i)} \prod_i \theta_i^{\alpha i - 1} \sim Dir(\alpha_1, \ldots, \alpha_k)$$

with α = $(\alpha_1,\ldots,\alpha_\mathrm{k})$, θ=$(\theta_1,\ldots\theta_\mathrm{k})$, $\sum_i \theta_i = 1$, $A = \sum_i a_i$

Here, ‘θ’ is the bias.

- Samples from dirichlet distribution can be used to model the bias in a k-sided dice.
- Generalizes beta distribution to k-1 probability simplex.

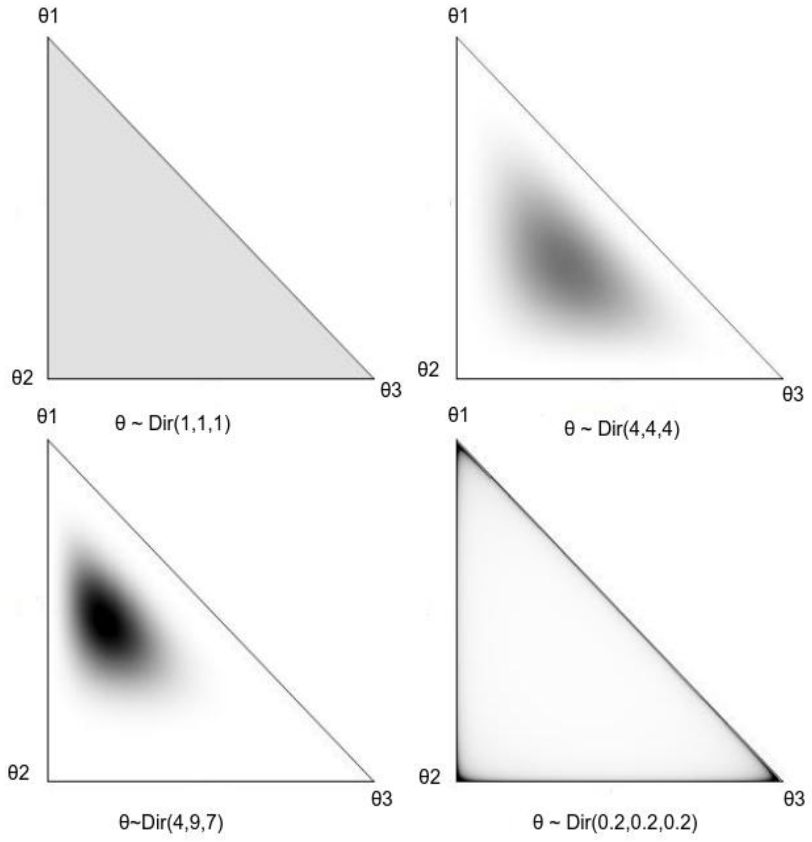


Figure 6: Pdf of dirichlet distribution for different values of alpha

Continuous Probability Distribution:

- **Inferring model parameters from data**

- Use Bayes rule to infer model parameters (θ) from data

$$p(\theta \mid D) = \frac{p(D \mid \theta) * p(\theta)}{p(D)}$$

Here, $p(D)$ is a distribution on data, also called normalizing factor

$p(\theta)$ is a prior

$p(D \mid \theta)$ is likelihood of data given model parameters

$p(\theta \mid D)$ is the posterior probability distribution of model parameters given data

Thomas Bayes

Why Bayes Rule?

- Likelihood of the occurence of any event is heavily influenced by other events occuring beforehand and is defined over time. When we toss a coin we can observe the frequency of an event (such as getting heads) occurring over time.

- "what is the probability of an effect given a cause?"
  "what is the probability of a cause given an effect?".

- Take help of simple conditional probability to compute the complex one.

- Prior helps in reducing the search space.

- **Bayes theorem: Derived from conditional Probability**

The probability of two events A and B happening, P(A∩B):

$$P(A \cap B) = P(A)P(B|A)$$

On the other hand, the probability of A and B is also equal to:

$$P(A \cap B) = P(B)P(A|B)$$

Equating the two yields:

$$P(B)P(A|B) = P(A)P(B|A)$$

Thus, $P(A|B) = \frac{P(A)P(B|A)}{P(B)}$

Applying Bayes Theorem

Video1

Video2

References:

- Manandhar Suresh. *Bayesian ML : Posterior Distributions and Mixture Models Continuous Probability Density Function*
- Bishop Christopher. *Pattern Recognition and Machine Learning*

Thank you