Big Data Visual Analytics (CS 661)

Instructor: Soumya Dutta

Department of Computer Science and Engineering

Indian Institute of Technology Kanpur (IITK)

email: soumyad@cse.iitk.ac.in

Study Materials for Lecture 13

- https://jdstorey.org/fas/index.html
- https://online.stat.psu.edu/stat500/lesson/0
- A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models
- EM Algorithm: https://stephens999.github.io/fiveMinuteStats/intro_to_em.html

Random Variables and Distributions

Random Variable

- Let S be a sample space of an experiment
  - S is associated with a probability measure P
- A random variable X is a real valued function on S
- Key property: It is a function whose values have probabilities attached with it

Random Variable: Example

- Let us flip a fair coin three times
- Sample space S = {hhh, hht, hth, htt, thh, tht, tth, ttt}
- Assume X is a function on S, so that X is the number of heads (h)
- So, we have,
  - {hhh → 3, hht → 2, hth → 2, htt → 2, thh → 2, tht → 2, tth → 1, ttt → 0}
- <u>X is a random variable</u>

Random Variable: Example

- We can answer questions like:
  - P(X=0) = P(ttt) = 1/8
  - P(X = 1) = P(htt) + P(tht) + P(tth) = 3/8
  - P(X = 2) = P(hht) + P(hth) + P(thh) = 3/8
  - P(X = 3) = P(hhh) = 1/8
- We can tabulate it:

X	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

Random Variable (RV): Example

- Rolling a fair die
- Assume a RV: X = the number that comes up
- X takes values 1,2,3,4,5,6 with probability 1/6

X	1	2	3	4	5	6
$P(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Discrete and Continuous Random Variable

- A random variable is said to be discrete if its set of possible values is a discrete set
  - Example: Rolling a fair die and measuring the value that shows up

- A random variable is said to be continuous when it can assume an uncountable number of values
  - Example: Depth of a pool, height of all the males, etc.

Expected Value and Variance of a Discrete RV

- Expected Value (mean):

$$E(X) = \sum x_i * p(x_i), \qquad p(x) = PMF$$

- Variance:

$$Var(X) = \sum \big(x - E(X)\big)^2 * p(x)$$

- Standard Deviation:

$$SD(X) = \sqrt{Var(X)}$$

Expected Value and Variance of a Continuous RV

- Expected Value (mean):

$$E(X) = \int_{-\infty}^{\infty} x * f(x) dx, \qquad f(x) = PDF$$

- Variance:

$$Var(X) = \int_{-\infty}^{\infty} \left(x - E(X)\right)^2 * f(x)\, dx$$

- Standard Deviation:

$$SD(X) = \sqrt{Var(X)}$$

Probability Distribution Function

- **A probability distribution function** is a mathematical function that provides probabilities of occurrence for the possible outcomes of a random variable

- **Probability Mass Function (PMF):** The probability distribution of a discrete random variable is called probability mass function

- **Probability Density Function (PDF):** The probability distribution of a continuous random variable is called probability density function

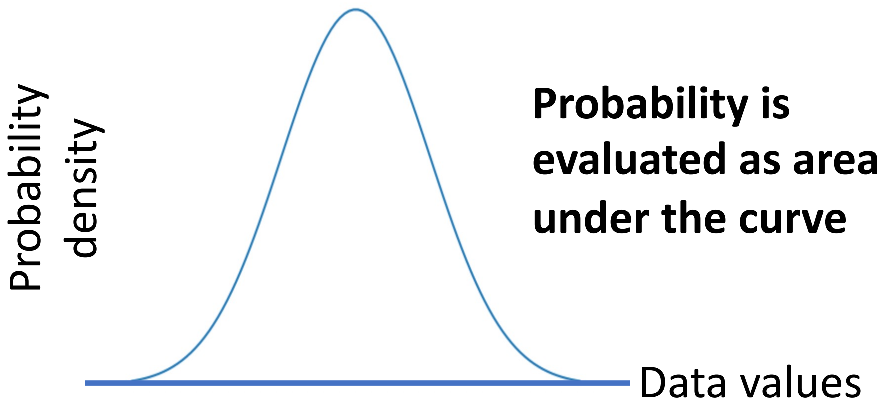
Probability Distribution Function: Properties

- Discrete case: PMF
- $p(x) = P(X = x)$

1. $p(x) \geq 0$
2. $\sum_{all\ possible\ x} p(x) = 1$
3. $p(x) = 0$ for all x outside a discrete range

- Continuous case: PDF
- $f(x)$

1. $f(x) \geq 0$
2. $\int_{-\infty}^{\infty} f(x)dx = 1$



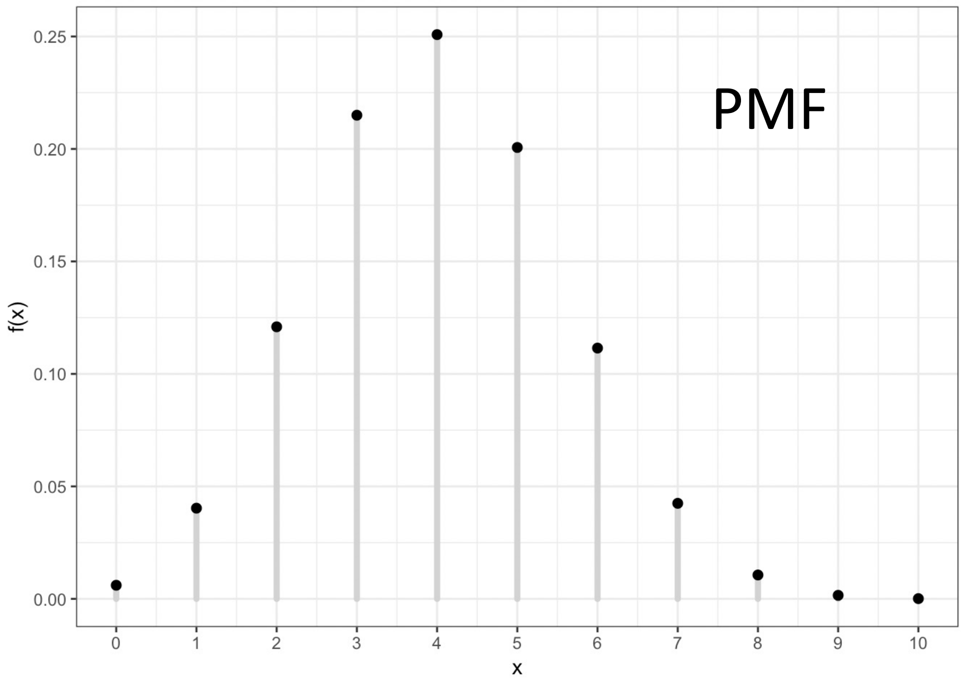
$P\ (x = c) = 0$ The probability that x takes on any individual value is zero. The area below the curve between x=c and x=c has no width, and therefore no area.

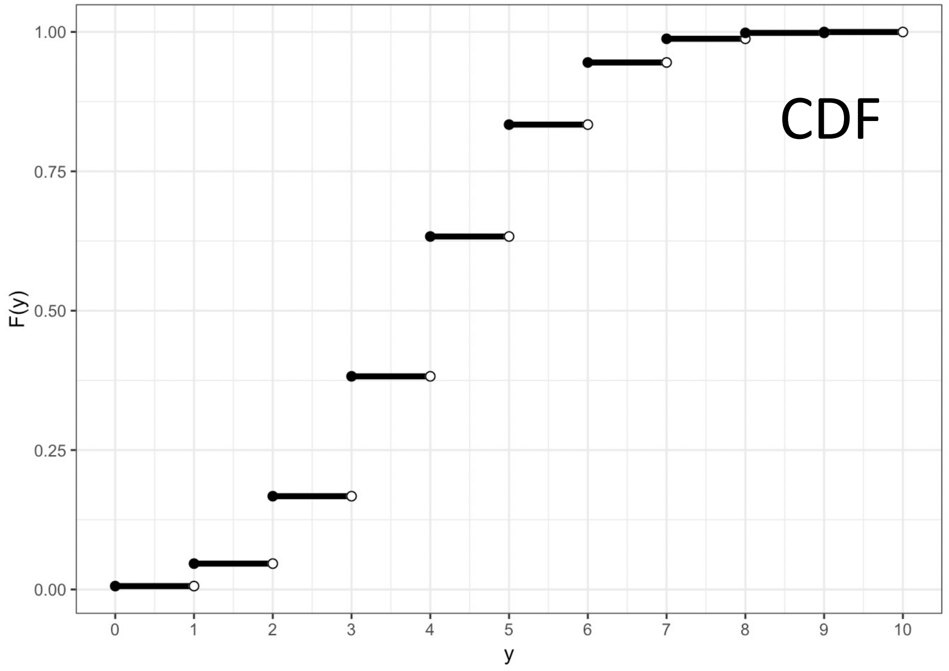
Cumulative Distribution Function (CDF)

- Discrete RV: Non decreasing function

$$F_X(x) = p(X \leq x) = \sum_{x_i \leq x} p(x_i)$$

CDF is a right continuous function for discrete RV





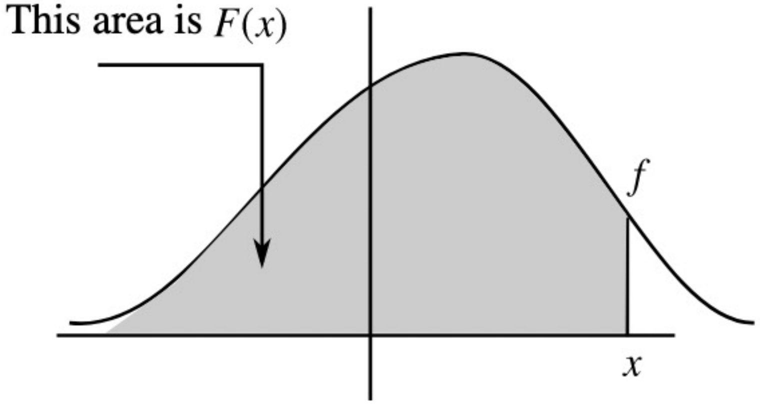
Probabilities of Events Via Discrete CDF

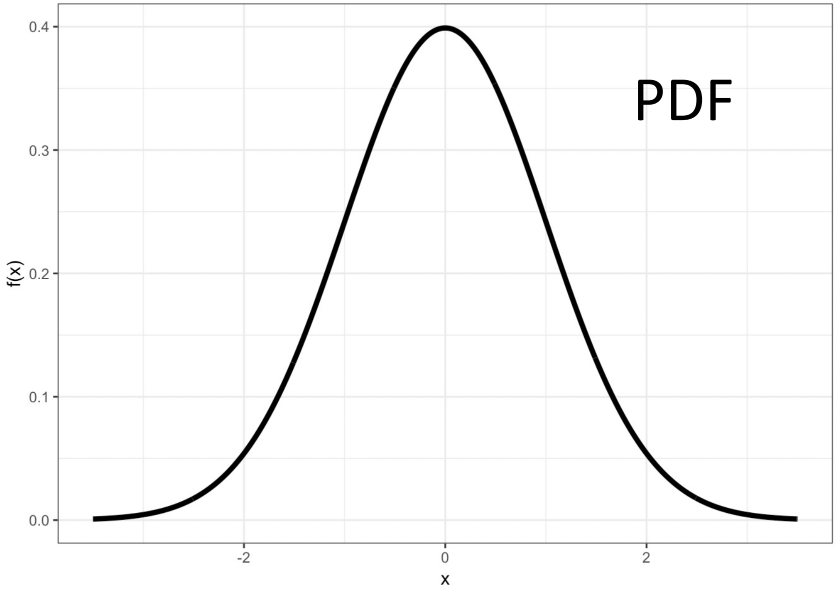
Probability	**CDF**	**PMF**
$\Pr(X \leq b)$	$F(b)$	$\sum_{x \leq b} f(x)$
$\Pr(X \geq a)$	$1 - F(a-1)$	$\sum_{x \geq a} f(x)$
$\Pr(X > a)$	$1 - F(a)$	$\sum_{x > a} f(x)$
$\Pr(a \leq X \leq b)$	$F(b) - F(a-1)$	$\sum_{a \leq x \leq b} f(x)$
$\Pr(a < X \leq b)$	$F(b) - F(a)$	$\sum_{a < x \leq b} f(x)$

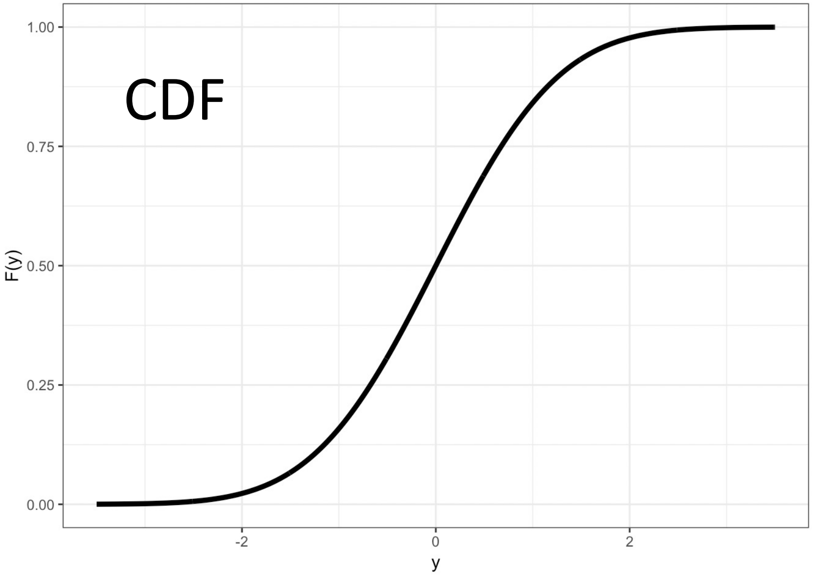
Cumulative Distribution Function (CDF)

- Continuous RV: Non decreasing function

$$F_X(x) = \int_{-\infty}^{x} f(x)dx$$







CDF is a continuous function here

https://jdstorey.org/fas/random-variables.html

Probabilities of Events Via Continuous CDF

Probability	CDF	PDF
$\Pr(X \leq b)$	$F(b)$	$\int_{-\infty}^{b} f(x)dx$
$\Pr(X \geq a)$	$1 - F(a)$	$\int_{a}^{\infty} f(x)dx$
$\Pr(X > a)$	$1 - F(a)$	$\int_{a}^{\infty} f(x)dx$
$\Pr(a \leq X \leq b)$	$F(b) - F(a)$	$\int_{a}^{b} f(x)dx$
$\Pr(a < X \leq b)$	$F(b) - F(a)$	$\int_{a}^{b} f(x)dx$

Discrete: Uniform Distribution

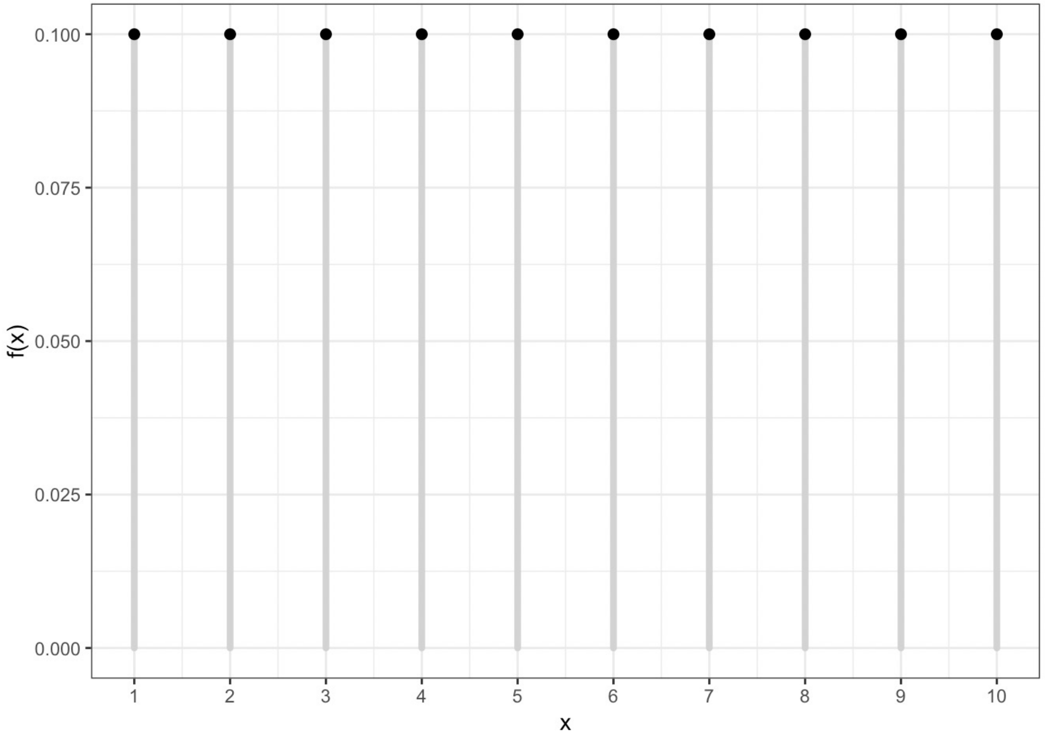
- Distribution assigns equal probabilities to a finite set of values

$$X \sim \text{Uniform}\{1, 2, \ldots, n\}$$

$$\mathcal{R} = \{1, 2, \ldots, n\}$$

$$f(x; n) = 1/n \text{ for } x \in \mathcal{R}$$

$$\text{E}[X] = \frac{n+1}{2}, \ \text{Var}(X) = \frac{n^2 - 1}{12}$$



Continuous: Exponential Distribution

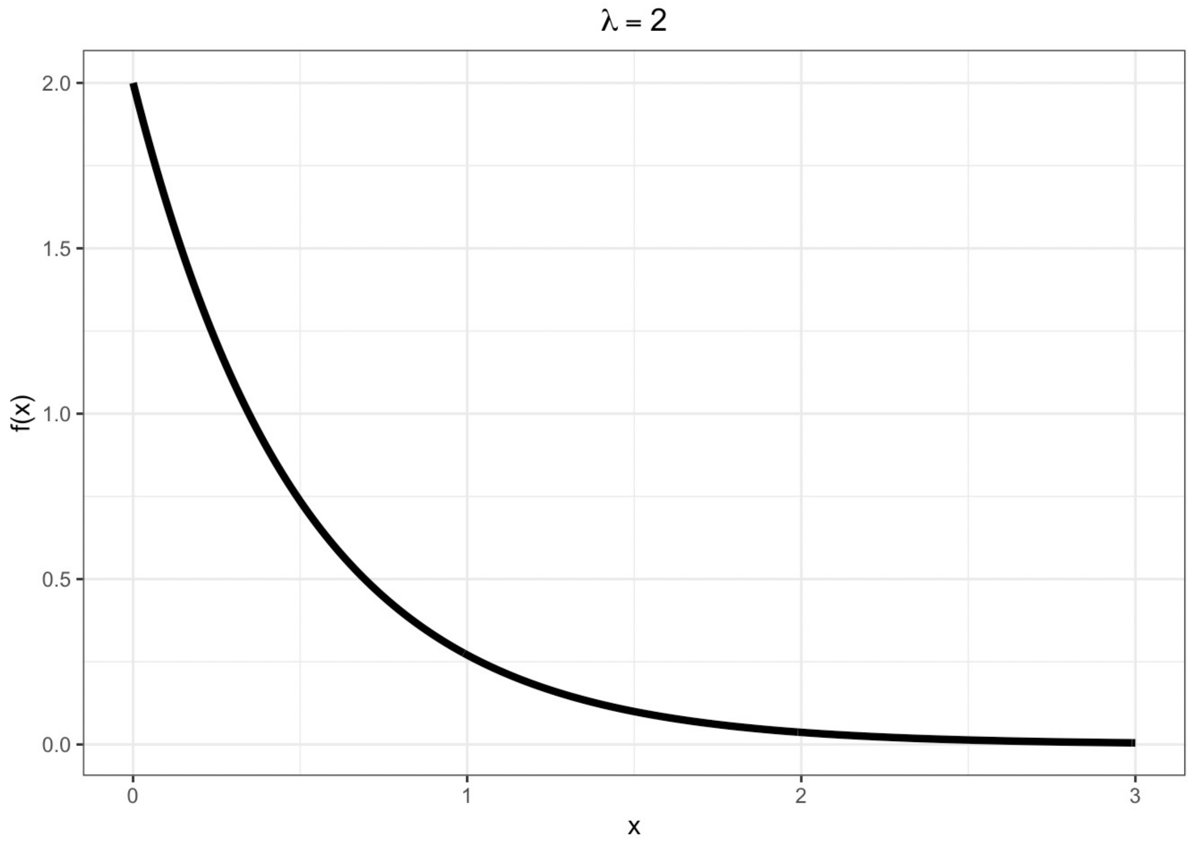
$$X \sim \text{Exponential}(\lambda)$$

$$\mathcal{R} = [0, \infty)$$

$$f(x; \lambda) = \lambda e^{-\lambda x} \text{ for } x \in \mathcal{R}$$

$$F(y; \lambda) = 1 - e^{-\lambda y} \text{ for } y \in \mathcal{R}$$

$$\text{E}[X] = \frac{1}{\lambda}, \ \text{Var}(X) = \frac{1}{\lambda^2}$$



Continuous: Beta Distribution

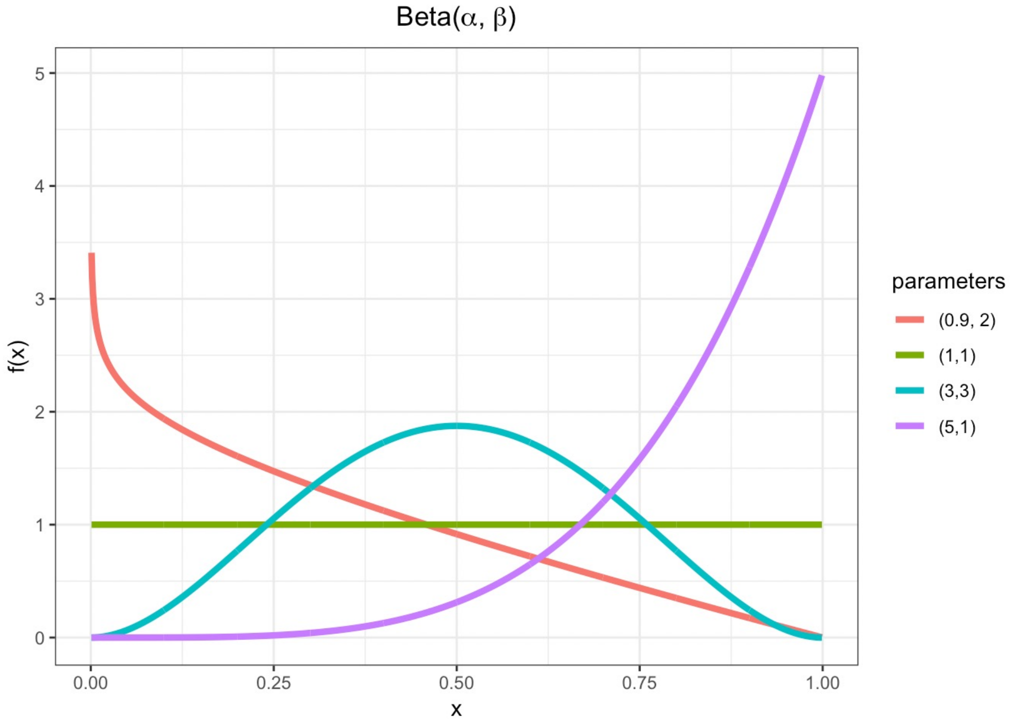
$$X \sim \text{Beta}(\alpha, \beta) \text{ where } \alpha, \beta > 0$$

$$\mathcal{R} = (0, 1)$$

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \text{ for } x \in \mathcal{R}$$

where $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}dx.$

$$\text{E}[X] = \frac{\alpha}{\alpha+\beta}, \ \text{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$



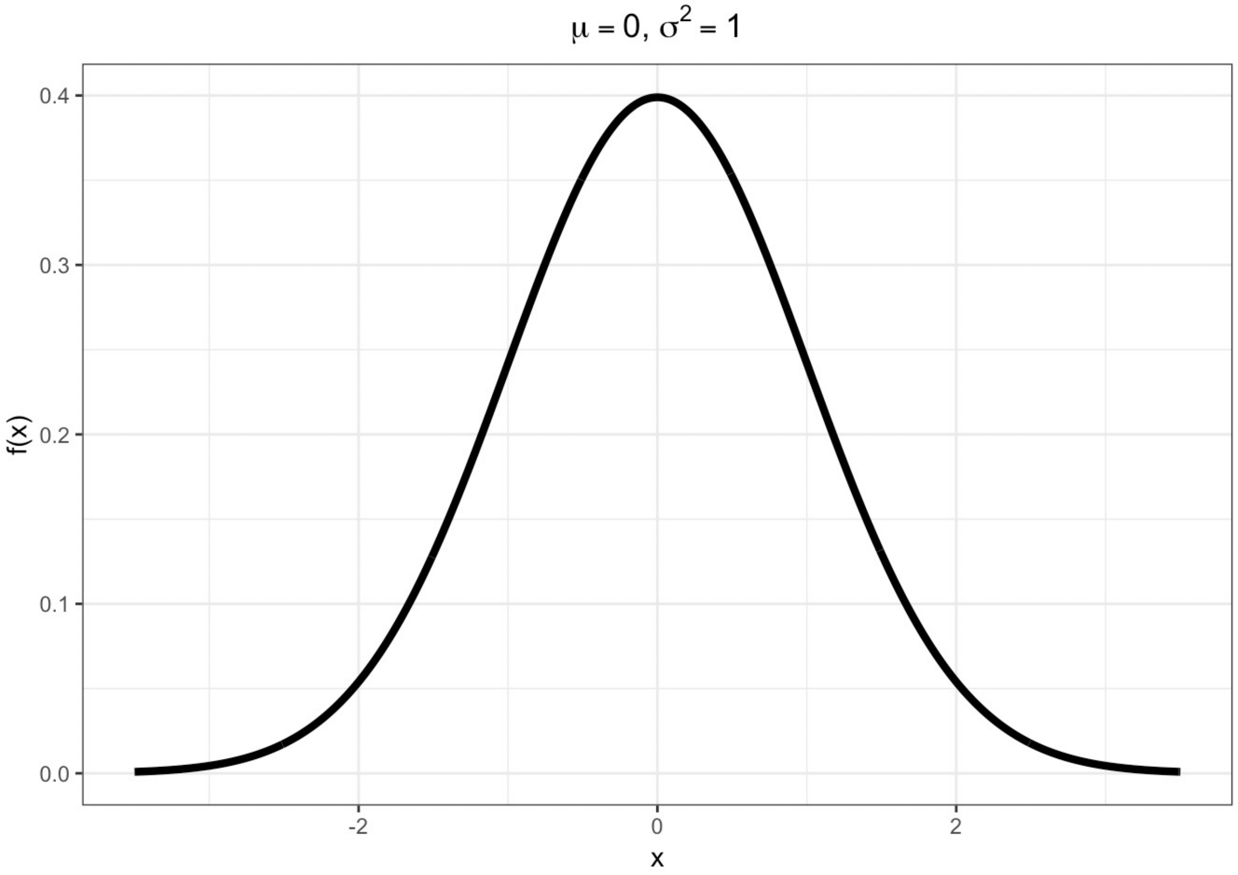
Continuous: Normal (Gaussian) Distribution

$$X \sim \text{Normal}(\mu, \sigma^2)$$

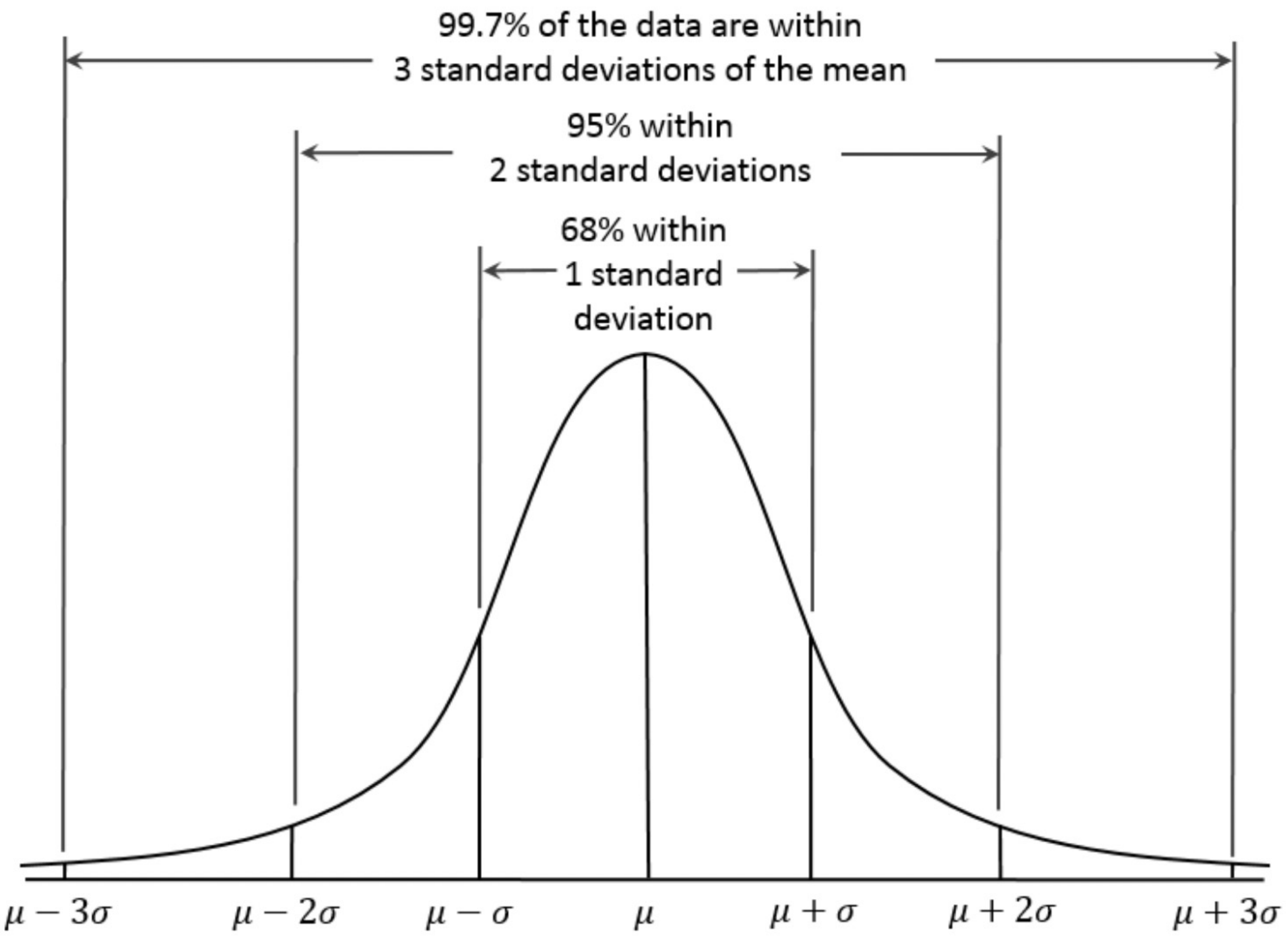
$$\mathcal{R} = (-\infty, \infty)$$

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \text{ for } x \in \mathcal{R}$$

$$\mathrm{E}[X] = \mu, \ \text{Var}(X) = \sigma^2$$



Reading a Normal (Gaussian) Distribution



Continuous: Standard Normal Distribution

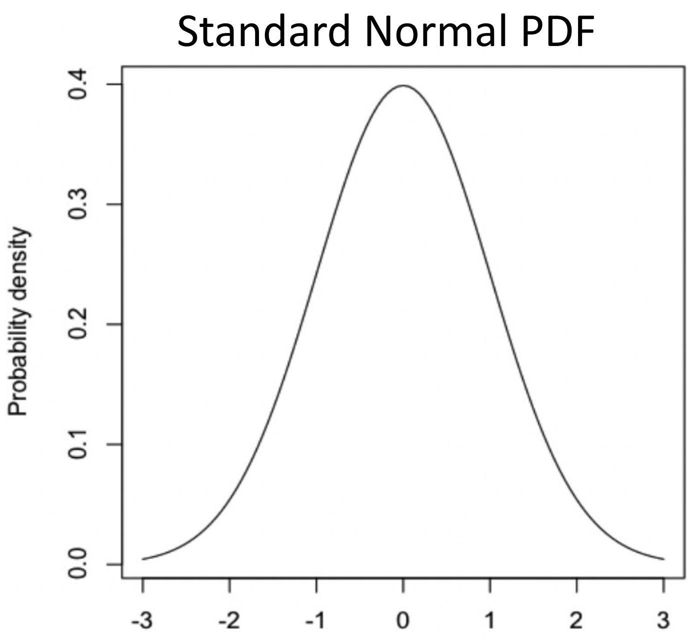
- It is the normal distribution with a mean equal to 0 and a standard deviation (also variance) equal to 1
- The standard normal distribution is often abbreviated to Z. It is frequently used to simplify working with normal distributions.

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

Continuous: Standard Normal Distribution

- It is the normal distribution with a mean equal to 0 and a standard deviation (also variance) equal to 1
- The standard normal distribution is often abbreviated to Z. It is frequently used to simplify working with normal distributions.

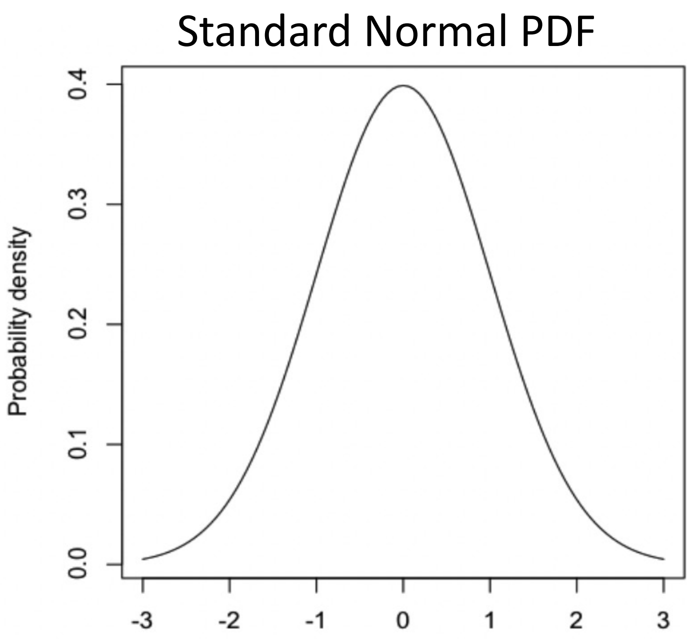
$$Z = \frac{X - \mu}{\sigma}$$

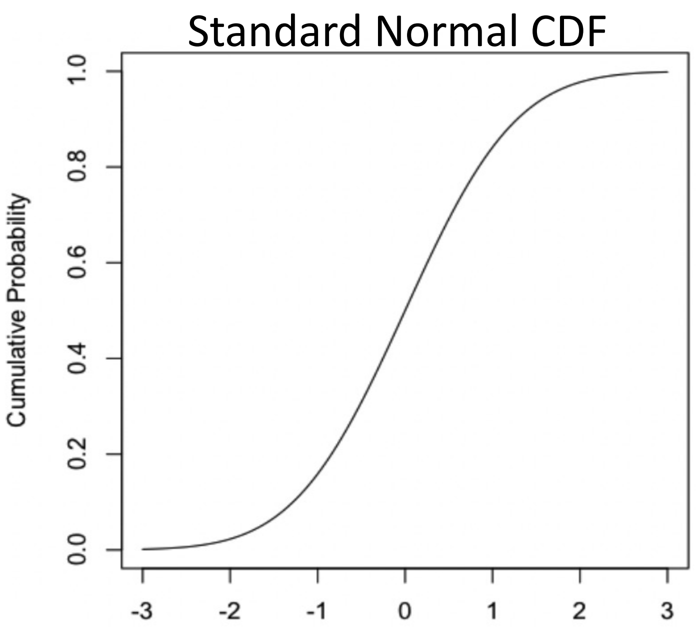


Continuous: Standard Normal Distribution

- It is the normal distribution with a mean equal to 0 and a standard deviation (also variance) equal to 1
- The standard normal distribution is often abbreviated to Z. It is frequently used to simplify working with normal distributions.

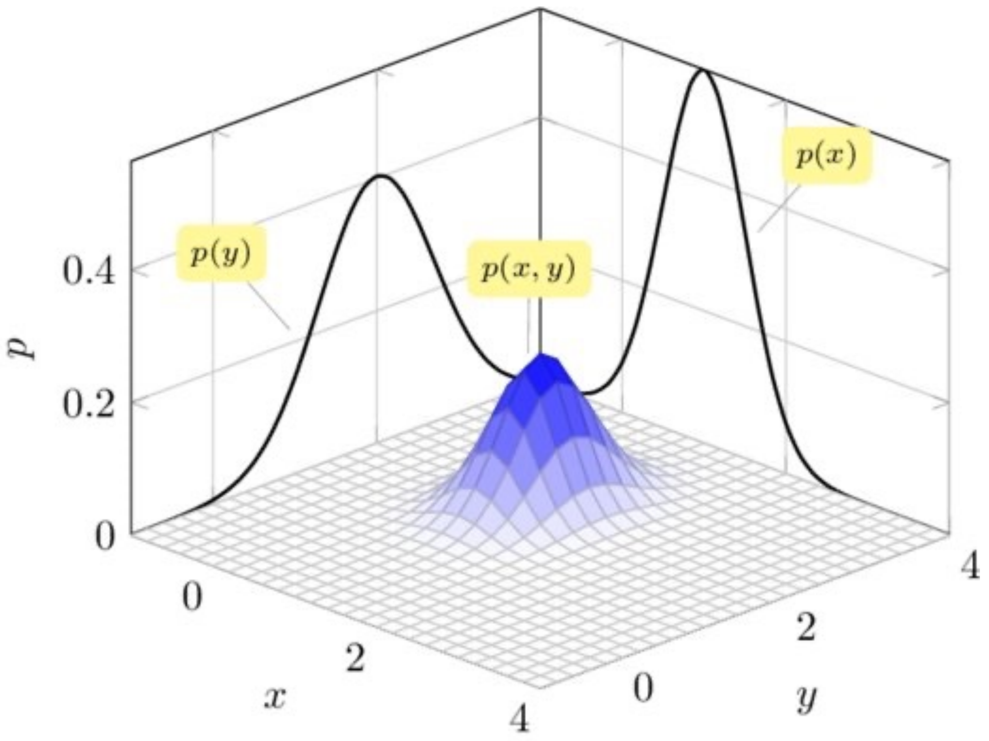
$$Z = \frac{X - \mu}{\sigma}$$





Joint Probability Distribution Function

- If we have multiple random variables, defined over the same probability space S, then the joint probability distribution is the distribution function that is defined over all possible event combinations of all the random variables
- Joint probability density function for two continuous random variables X and Y can be represented as $f_{XY}(x, y)$



Joint Probability Distribution Function

- The concept of joint probability distribution function is generalizable and goes beyond two variables: $f_{X1X2X3 \dots Xn}(x_1, x_2, x_3, \dots x_n)$
- For two variable case, $f_{XY}(x, y)$ must be a non-negative function and the following must hold:

$$\iint_{-\infty}^{\infty} f_{XY}(x, y)dxdy = 1$$

- Joint Cumulative Distribution function (CDF)

$$F_{XY}(x, y) = \mathrm{P}(X \leq a, Y \leq b) = \int_{-\infty}^{a}\int_{-\infty}^{b} f_{XY}(x, y)dxdy$$

Marginal Probability Distribution Functions

- From the joint probability distribution function, we can find the marginal probability distributions by integrating the joint distribution function $f_{XY}(x, y)$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dy, \text{ for all } x$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)\, dx, \text{ for all y}$$

- Marginal distribution functions (also known as univariate distributions) are probability distribution functions of individual random variables

Independence

- The continuous random variables are statistically independent if their joint probability distribution function factors into a product of their marginal distributions

$$f(x_1, x_2, x_3, \dots, x_n) = f(x_1)f(x_2)f(x_3)\dots f(x_n)$$

Conditional Probability and Bayes' Rule

- Conditional probability: It is the probability of an event given another event has occurred

$$f_{X|Y=y}(x) = \frac{\Pr(\{X = x\} \cap \{Y = y\})}{\Pr(Y = y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

- Bayes' Rule:

$$f_{X|Y=y}(x) = \frac{f_{Y|X=x}(y) * f_X(x)}{f_Y(y)}$$

$f_{X|Y=y}(x)$ = Conditional probability of X = x given Y = y. This is also called <u>posterior probability</u>
$f_{Y|X=x}(y)$ = Conditional probability of Y = y given X = x. This is called <u>likelihood</u>
$f_X(x)$ = marginal of X, also the <u>prior probability</u> of X= x
$f_Y(y)$ = marginal probability of Y

Representations of Distribution Functions

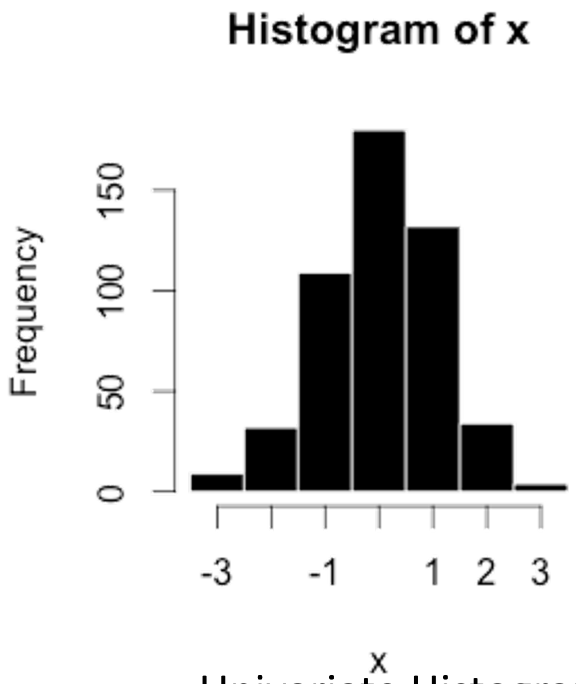
- Non-parametric model
  - Histogram
  - Kernel Density Estimation (KDE)
- Parametric models
  - Gaussian (Normal)
  - Gaussian mixture models (GMM)

Non-parametric Distributions: Histogram

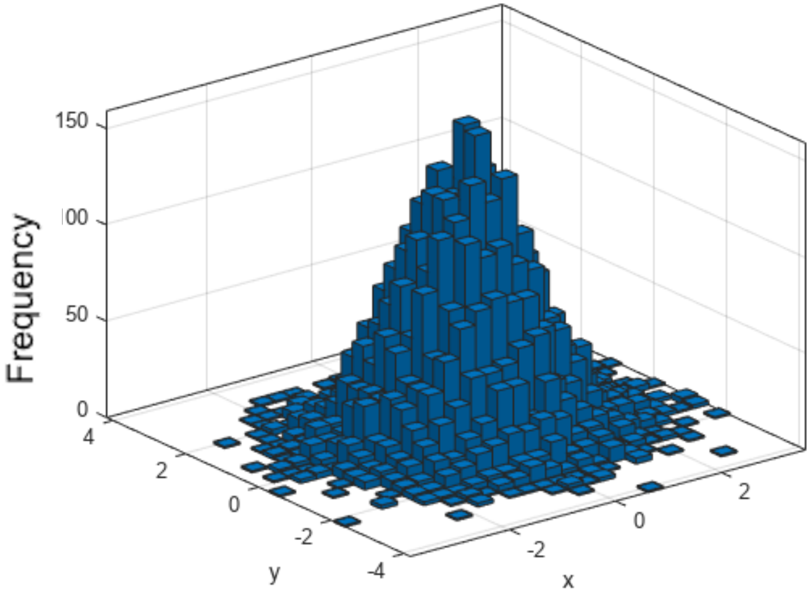
- <u>Histogram:</u> A histogram is an approximate representation of a statistical distribution. The area under a histogram can be normalized and used as a probability distribution function.

$$H(s) = \sum_i \delta(x - x_i)$$

$$\delta(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$



Univariate Histogram



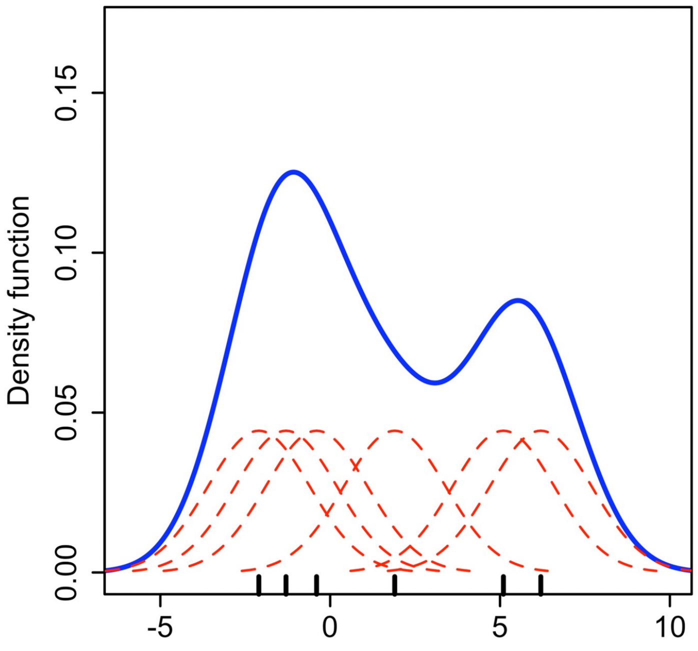
Joint Histogram

Non-parametric Distributions: KDE

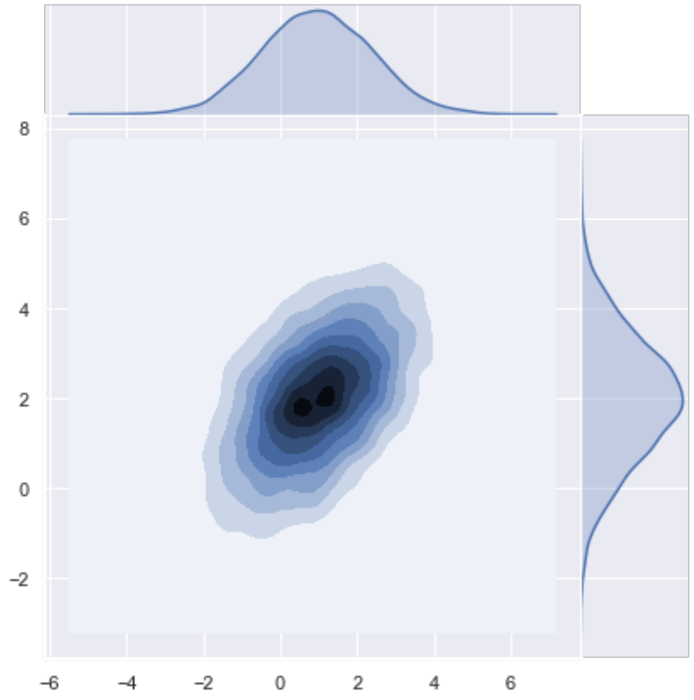
- <u>KDE:</u> Kerner Density Estimation is a popular method of distribution estimation technique from sample data. Formally it is defined as follows:

$$f(x) = \frac{1}{nb}\sum_{i=1}^{n} K(\frac{x - x_i}{b})$$

- *f(x)* is the KDE function
- *n* = number of data points
- *b* = bandwidth
- *K(.)* = Non-negative symmetric kernel function such as uniform, triangular, Gaussian etc.



Univariate KDE



Joint KDE

Parametric Distribution: GMM

- Gaussian Mixture Model (GMM): Represent a probability distribution function as a convex combination of multiple Gaussian functions

$$p(X) = \sum_{i=1}^{K} \omega_i * N(X \mid \mu_i, \sigma_i)$$

ω = Weights of the Gaussian components
K = Number of Gaussian components in the mixture model

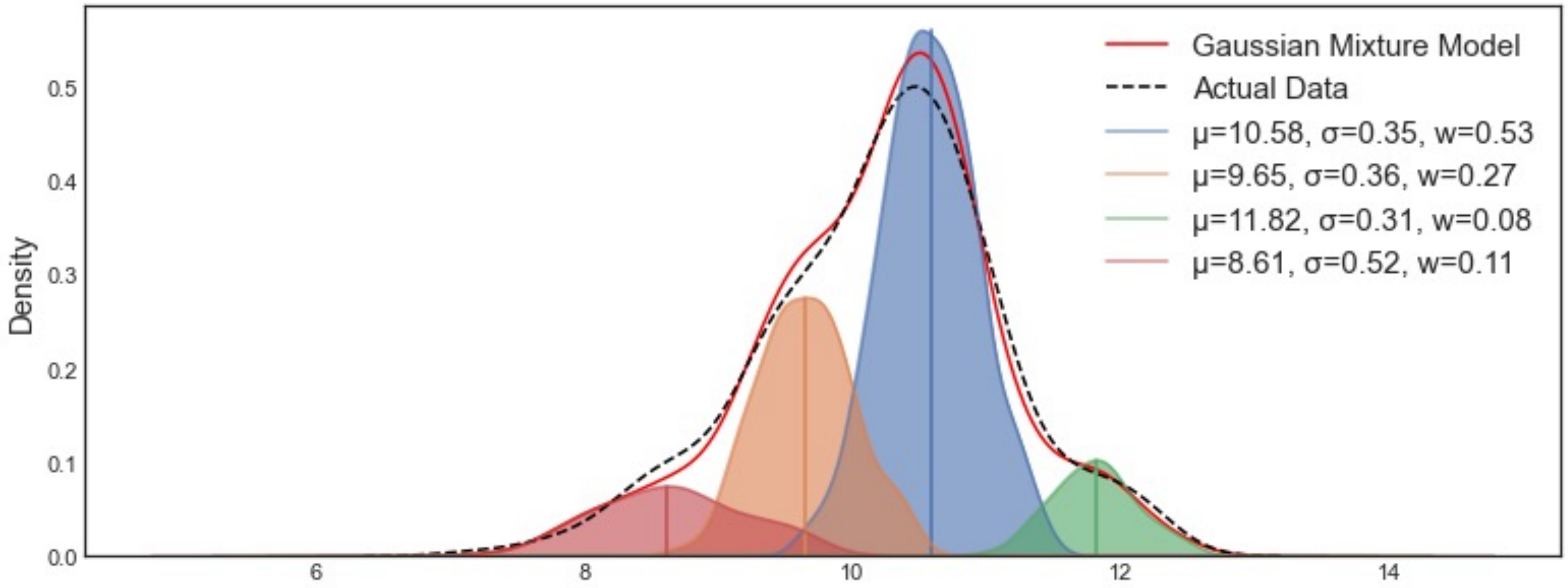


Fig. source: https://stats.stackexchange.com

Parameter Estimation Techniques

- Estimation of Gaussian distribution parameters are trivial
  - Maximum Likelihood Estimate (MLE)
  - Same as computing mean and variance
- Estimation of GMM parameters require Expectation Maximization (EM) algorithm
  - Iterative technique to fit GMM parameters
- Incremental schemes for GMM parameter estimation
  - Fast and approximate method to estimate GMM parameters
  - Can model streaming time-varying data