



# Lecture-3

## Course: Applied Data Science

### Probability & Statistics for Data Science

By

**Dr. Sibarama Panigrahi**

Assistant Professor, Department of Computer Sc. & Engineering  
National Institute of Technology, Rourkela, Odisha, 769008, India

**Mobile No.:** +91-7377302566

**Email:** panigrahis[at]nitrkl[dot]ac[dot]in  
panigrahi[dot]sibarama[at]gmail[dot]com

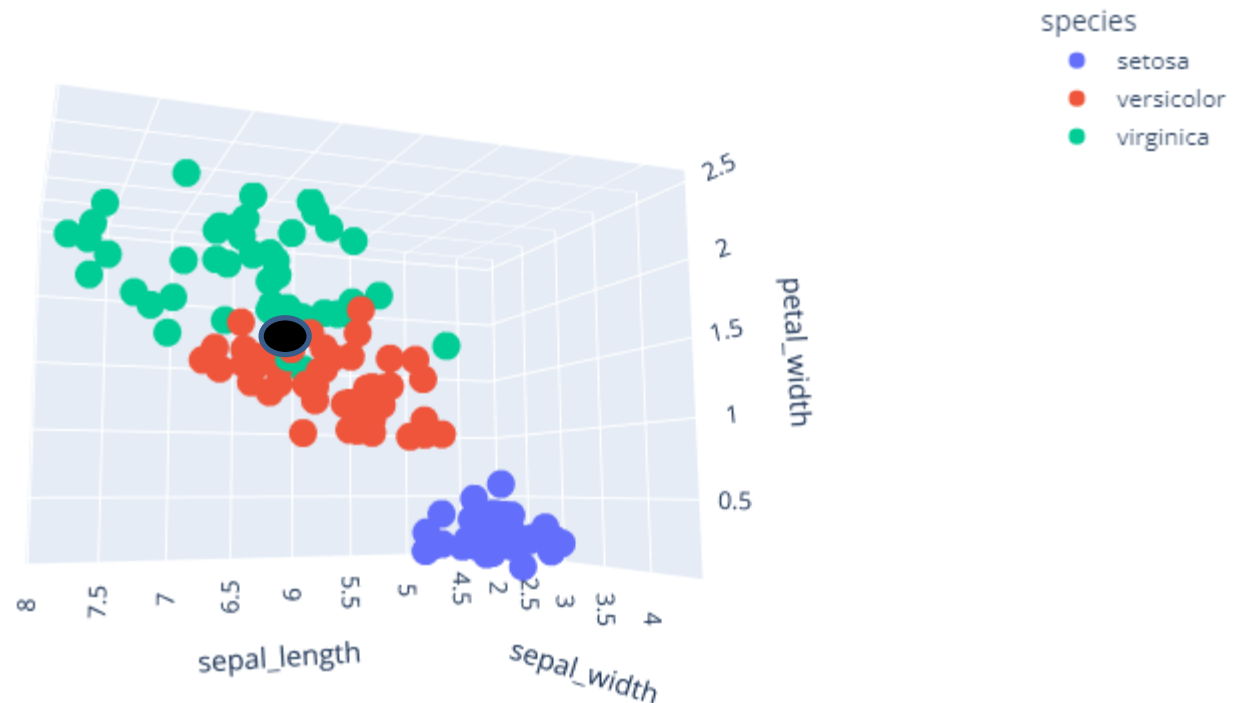
# Motivation

- Whether my query point(Black point) is setosa, versicolor or virginica.
- Setosa (Definitely No)
- May be versicolor or virginica
  - We may answer probabilistically (which is more appropriate here)
  - 0.8 versicolor
  - 0.2 virginica

## Last Session

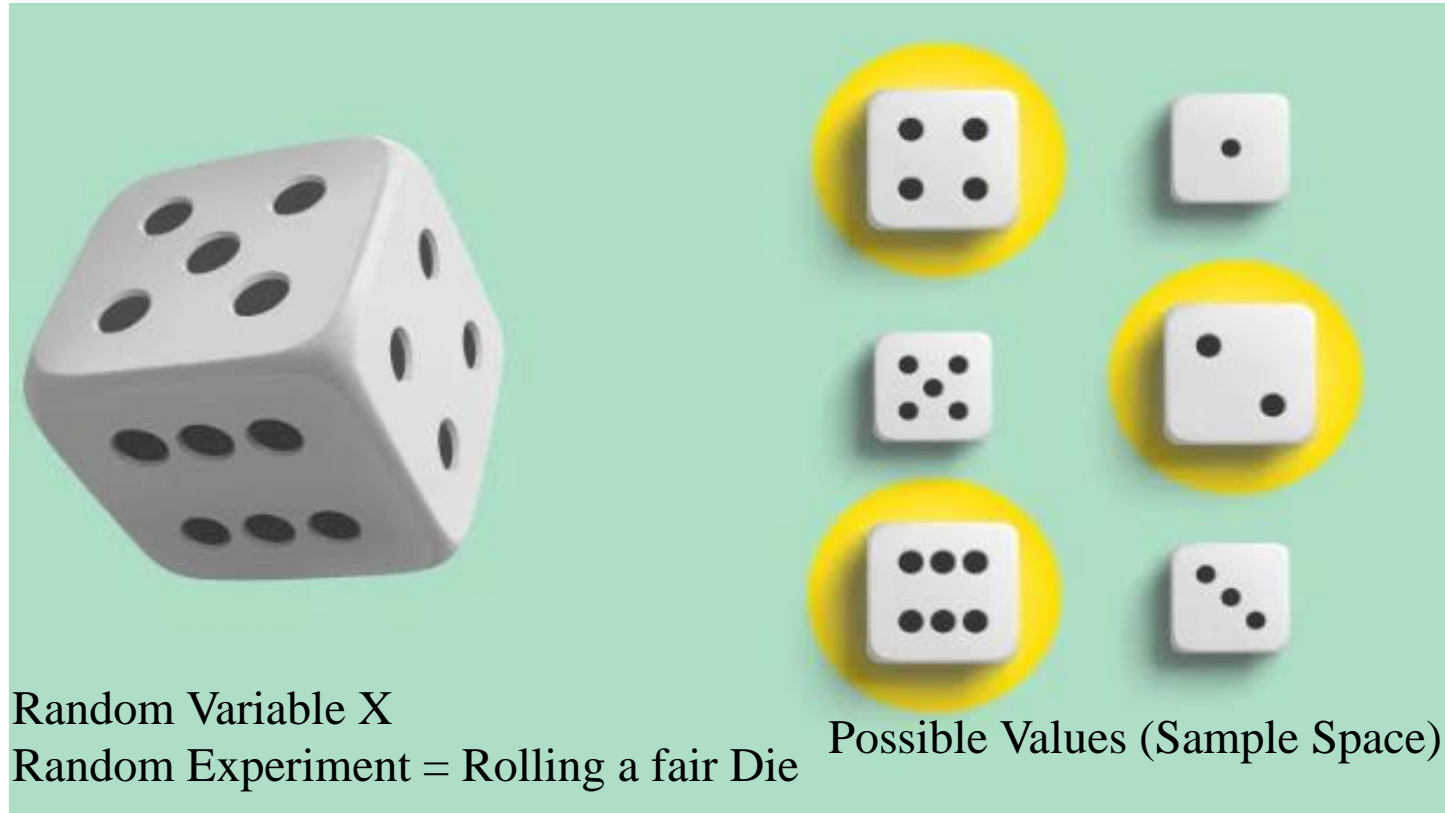
We have used the concepts like Histogram, PDF, CDF, Mean, Variance, Standard Deviation, etc.

**These are all concepts of Probability and Statistics**



# Random Variable

- A **random variable** is a variable that takes the outcomes of the random experiment as its value.
- In probability, a real-valued function, defined over the sample space of a random experiment, is called a **random variable**.



# Random Variable

- **Types of Random Variable**

- Discrete Random Variable:

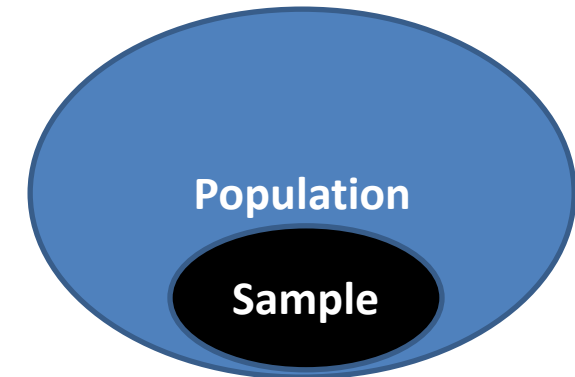
- A discrete random variable can take only a finite number of distinct values. (Sample Space has finite values)
    - Probability mass function (PMF) is used to describe the probabilities of discrete random variables.
    - e.g. Tossing a coin, Rolling a die, Gender of a person, etc.

- Continuous Random Variable:

- A continuous random variable can take infinite and uncountable set of values. (Sample Space has uncountable values)
    - Probability density function (PDF) is used to describe the probabilities of continuous random variables.
    - e.g. Height of a person, Weight of a person, etc.

# Population & Sample

- In statistics, a **population** refers to the entire group of individuals or items about which you want to gather information.
- A **sample**, on the other hand, is a subset of the population that is selected for study.
- **Q) What is the mean weight of humans?**
- **Population:** Mean weight of all humans of the world ( $\mu_{weight}$ )
- **Sample:** Mean weight of a subset of population (Let 1000 humans)  
( $\overline{x_{weight}}$ )
- As sample size increases  $\mu_{weight} = \overline{x_{weight}}$



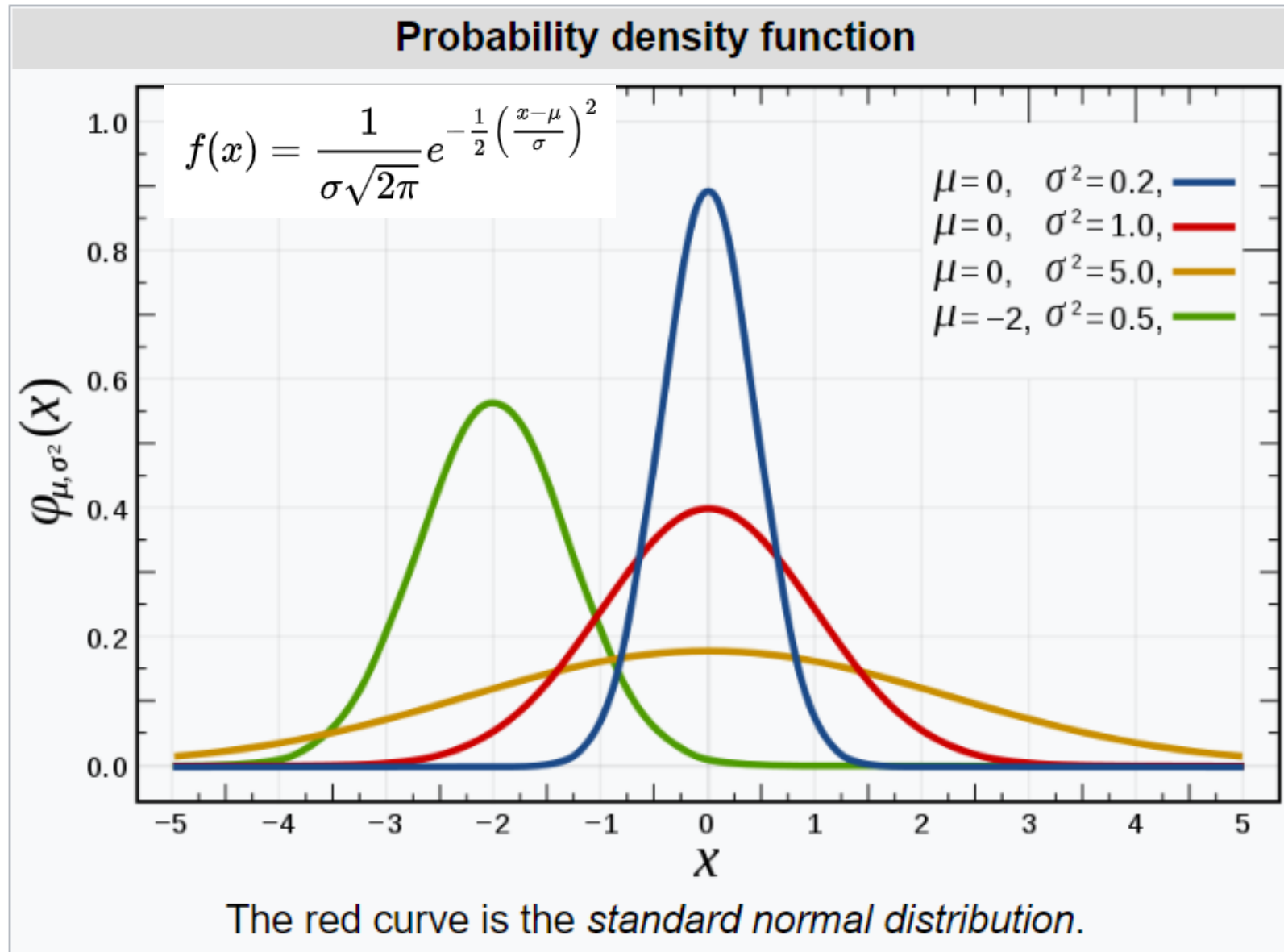
# Gaussian Distribution

- A **normal distribution** or **Gaussian distribution** is a type of continuous probability distribution for a real-valued random variable.
- The general form of its probability density function is

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

where  $\mu$  is the mean,  $\sigma$  is the standard deviation and  $f(x)$  is the probability density function.

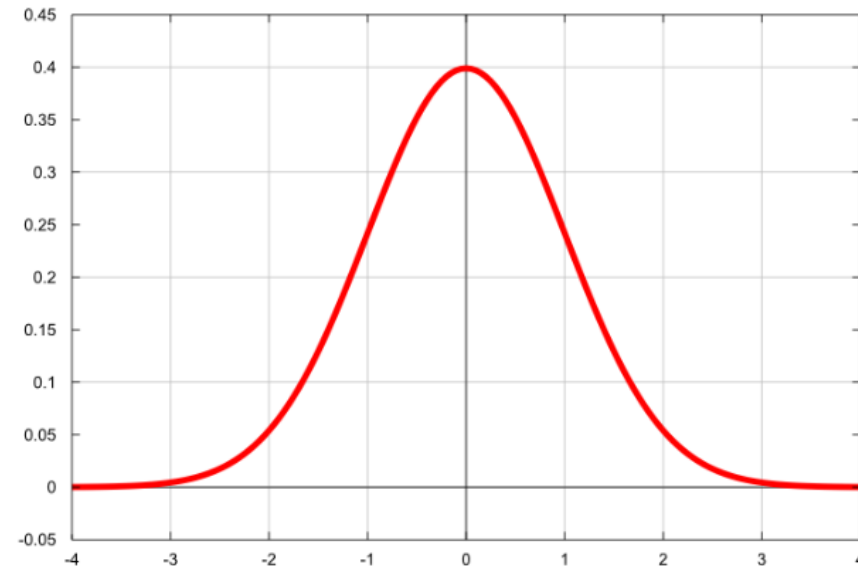
# Gaussian Distribution



# Gaussian Distribution

- Observations:
  - As  $x$  moves away from  $\mu$ , the probability reduces.
  - Normal Distribution Curve is symmetric about mean.
  - The bell shaped curve is reducing exponentially with a quadratic function.
  - Unimodal in nature with a single peak.
  - Standard Normal Distribution
    - When Mean=0, Variance=1

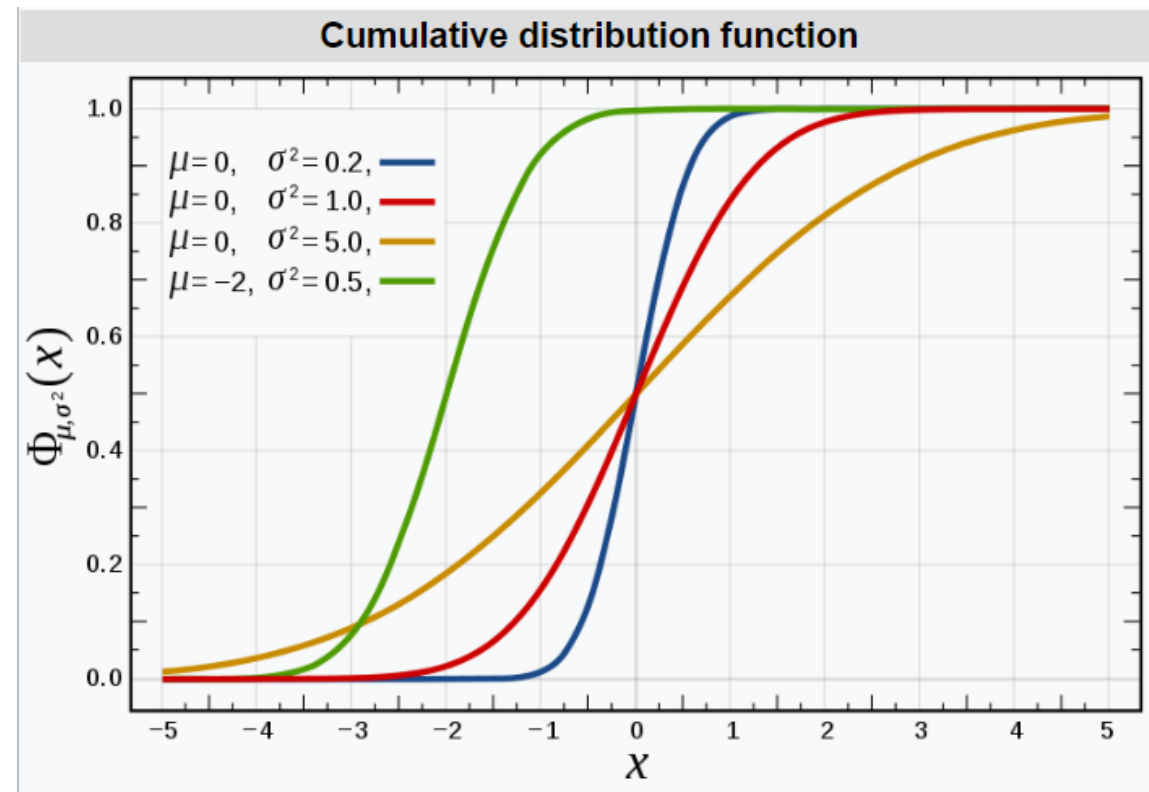
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$





# Gaussian Distribution

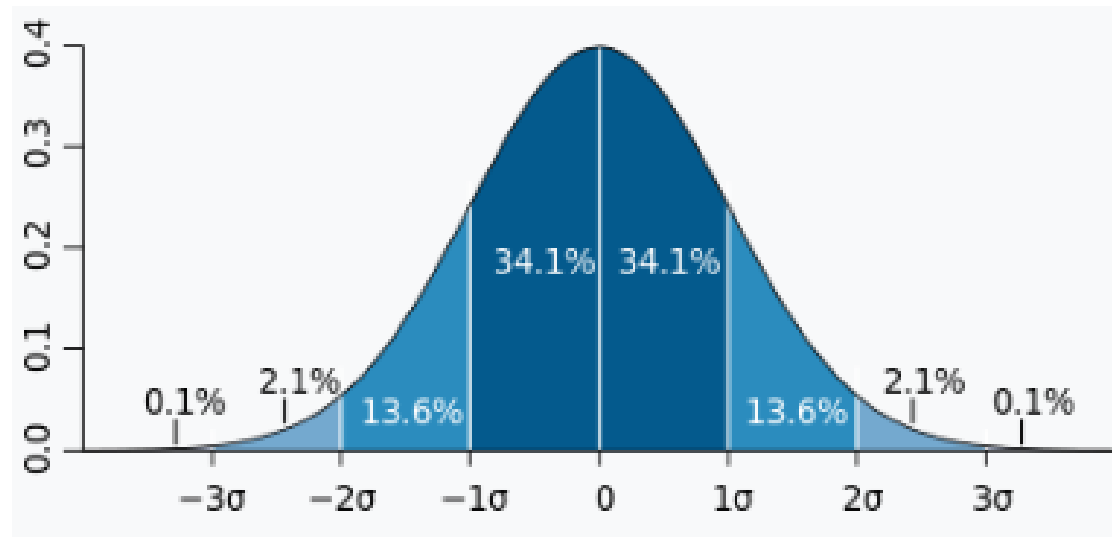
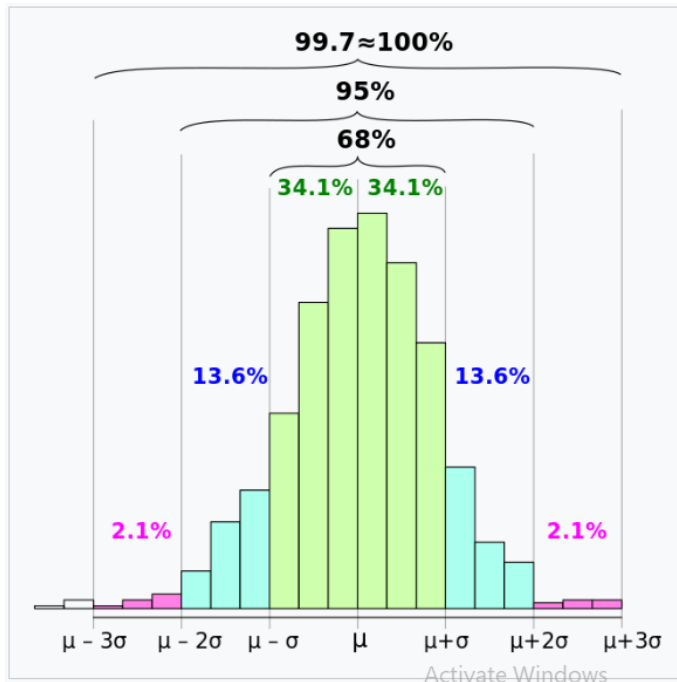
- Observation:
  - Minimum value 0, Maximum value 1
  - Value corresponding to mean is at the half of the CDF curve.
  - If variance is smaller, the CDF Curve is closer to the mean.



# Gaussian Distribution

- Observations:**

- Let  $X$  is a continuous random variable following Gaussian distribution with mean (let 0) and variance (let 1).
  - Within  $\text{mean} \pm 1 \times \text{standard deviation}$  68% points lie.
  - Within  $\text{mean} \pm 2 \times \text{standard deviation}$  95% points lie.
  - Within  $\text{mean} \pm 3 \times \text{standard deviation}$  99.7% points lie.

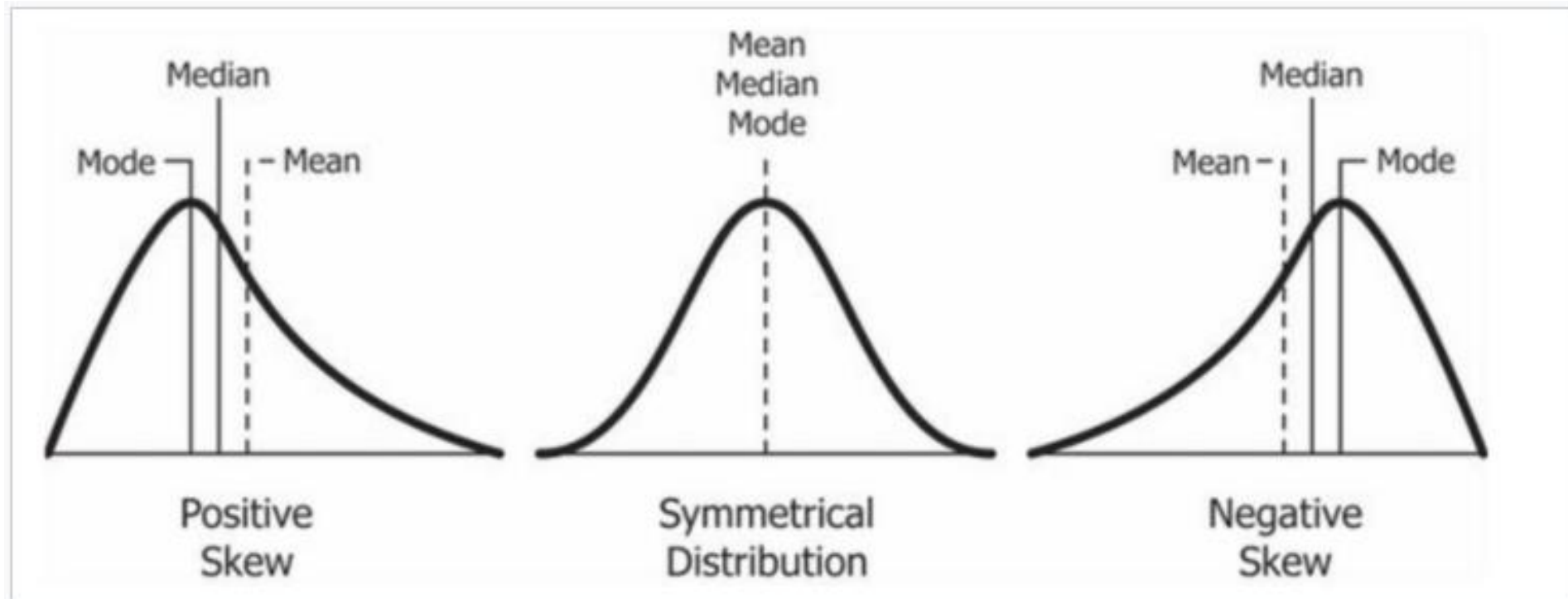


# Gaussian Distribution

- Suppose the random variable height of a person follows normal distribution with mean  $\mu = 150$  cm and variance  $\sigma^2 = 25$ .  
What can you quickly infer from this?
  - 68% persons height lie between
    - 145 cm and 155 cm i.e.  $[\mu - \sigma \text{ and } \mu + \sigma]$
  - 95% persons height lie between
    - 140 cm and 160 cm i.e.  $[\mu - 2\sigma \text{ and } \mu + 2\sigma]$
  - 99.7% persons height lie between
    - 135 cm and 165 cm i.e.  $[\mu - 3\sigma \text{ and } \mu + 3\sigma]$

# Skewness

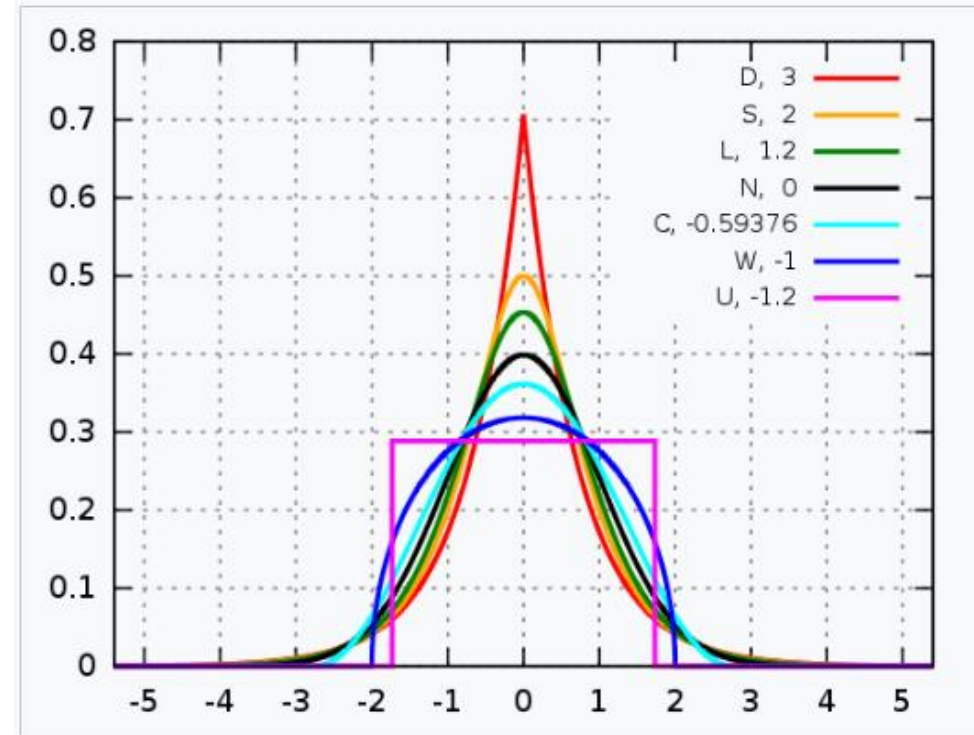
- **Skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean.
- **Positive Skew:** The right tail is longer.
- **Negative Skew:** The left tail is longer.



# Kurtosis

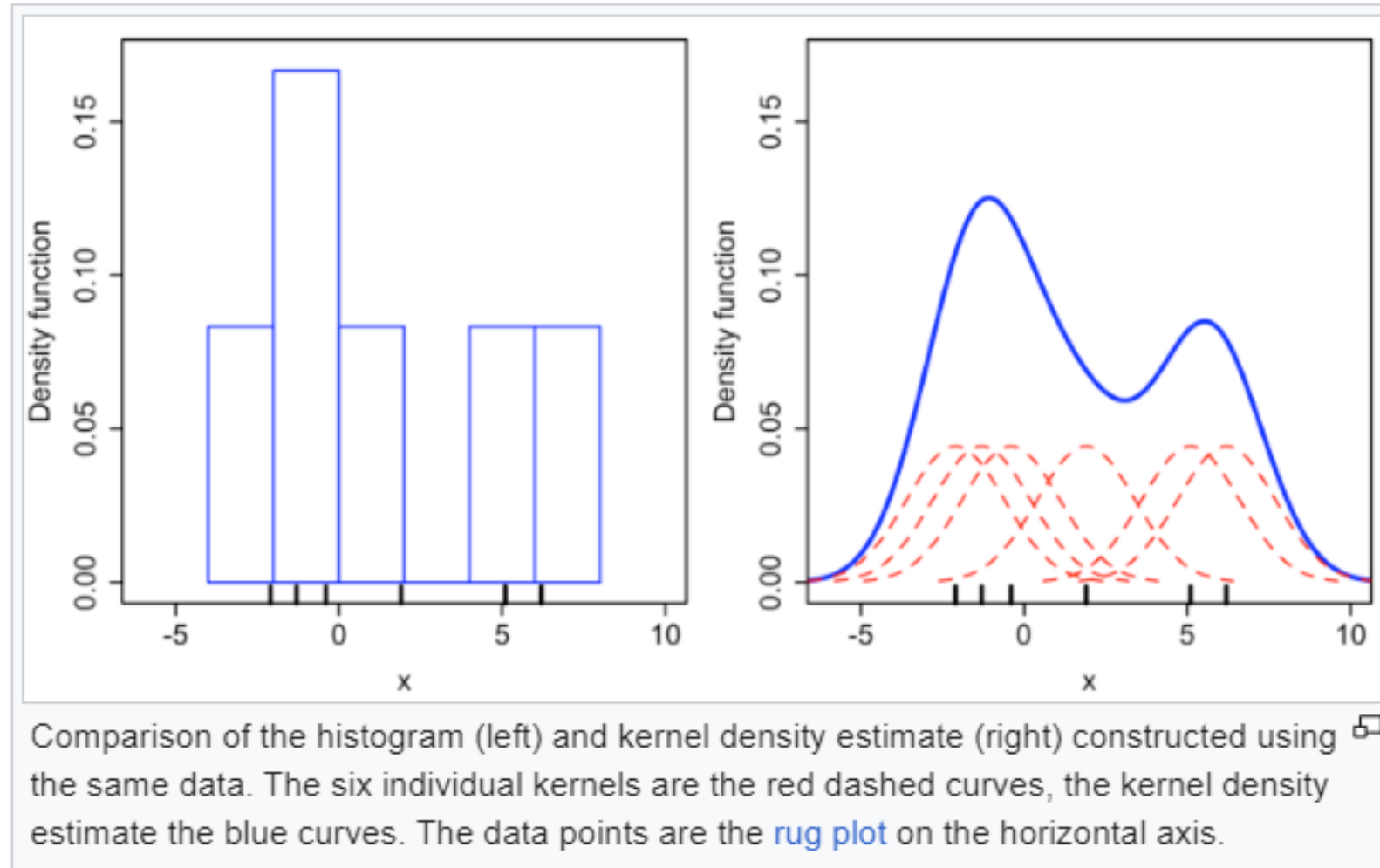
- **Kurtosis** is a measure of the "tailedness" of the probability distribution of a real-valued random variable.

- D: **Laplace distribution**, also known as the double exponential distribution, red curve (two straight lines in the log-scale plot), excess kurtosis = 3
- S: **hyperbolic secant distribution**, orange curve, excess kurtosis = 2
- L: **logistic distribution**, green curve, excess kurtosis = 1.2
- N: **normal distribution**, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- C: **raised cosine distribution**, cyan curve, excess kurtosis =  $-0.593762\dots$
- W: **Wigner semicircle distribution**, blue curve, excess kurtosis =  $-1$
- U: **uniform distribution**, magenta curve (shown for clarity as a rectangle in both images), excess kurtosis =  $-1.2$ .

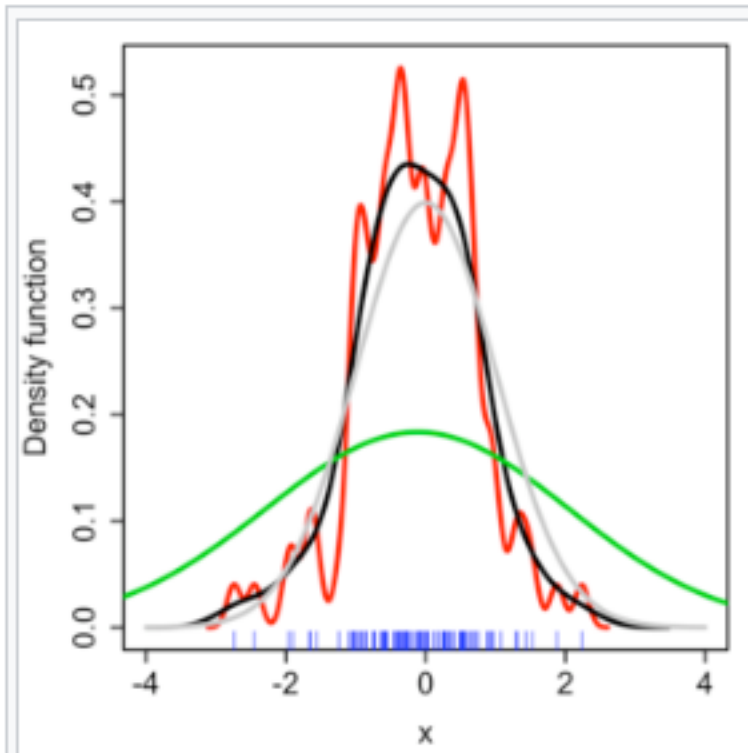


# Kernel Density Estimation

- **kernel density estimation (KDE)** is the application of [kernel smoothing](#) for [probability density estimation](#).



# Kernel Density Estimation



Kernel density estimate (KDE) with different bandwidths of a random sample of 100 points from a standard normal distribution. Grey: true density (standard normal). Red: KDE with  $h=0.05$ . Black: KDE with  $h=0.337$ . Green: KDE with  $h=2$ .

# Central Limits Theorem

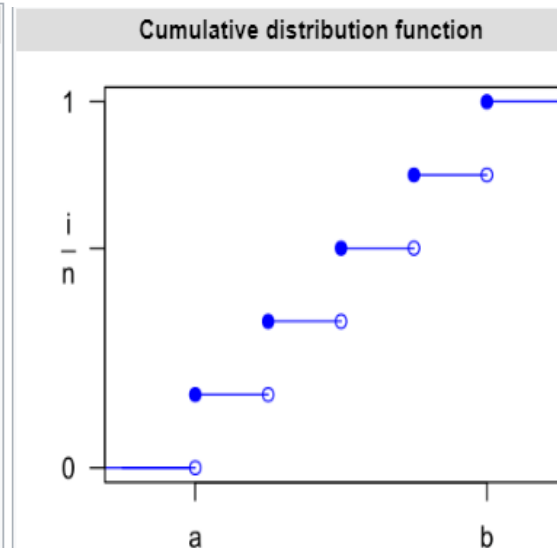
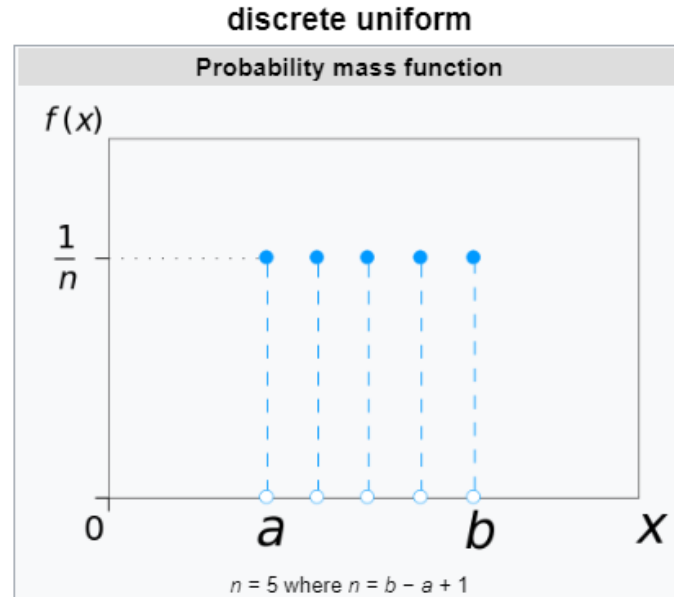
- Let  $\mathbf{X}$  be a random variable with finite population mean  $X$  and variance  $\sigma^2$  which follows any distribution (not necessarily normal distribution).
- Let  $s_1, s_2, s_3 \dots s_m$  be  $m$  random samples with each containing  $n$  observations and  $\bar{s}_1, \bar{s}_2, \bar{s}_3 \dots \bar{s}_m$  be their sample means respectively.
- Then the distribution formed by  $\bar{s}_1, \bar{s}_2, \bar{s}_3 \dots \bar{s}_m$  is called sampling distribution of sample means which follows a normal distribution with mean  $\bar{X}$  and variance  $\frac{\sigma^2}{n}$  as long as the sample size is large enough.



# Uniform Distribution

- Discrete Uniform Distribution**

- If the random variable following uniform distribution is discrete then it is called discrete uniform distribution.

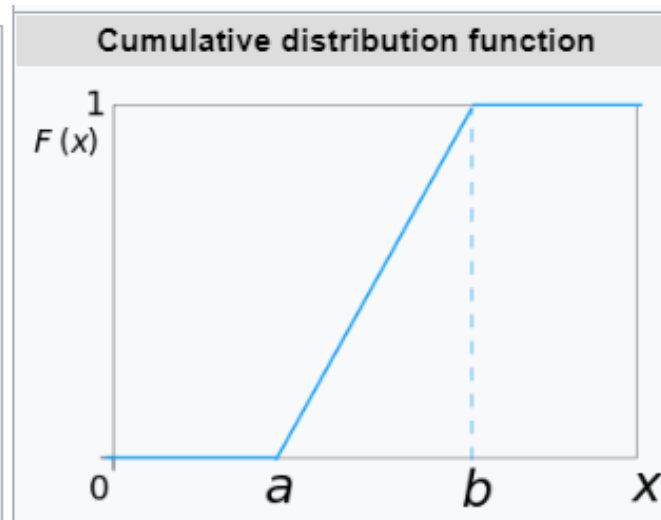
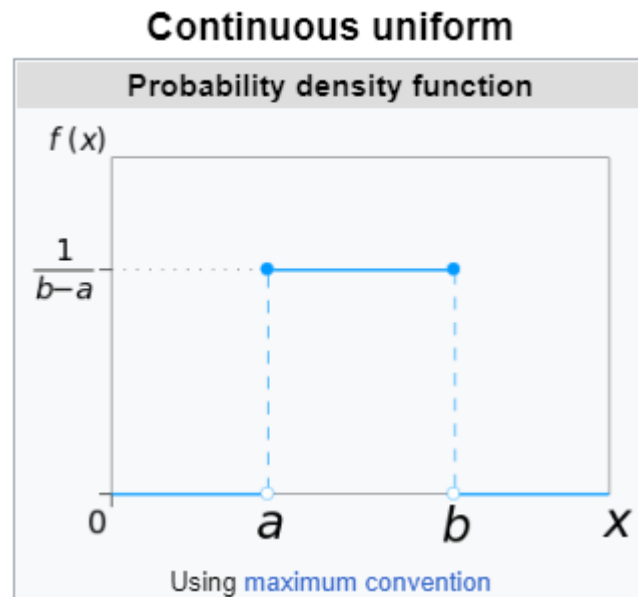


<b>Parameters</b>	$a, b$ integers with $b \geq a$ $n = b - a + 1$
<b>Support</b>	$k \in \{a, a + 1, \dots, b - 1, b\}$
<b>PMF</b>	$\frac{1}{n}$
<b>CDF</b>	$\frac{\lfloor k \rfloor - a + 1}{n}$
<b>Mean</b>	$\frac{a + b}{2}$
<b>Median</b>	$\frac{a + b}{2}$
<b>Mode</b>	N/A

# Uniform Distribution

- Continuous Uniform Distribution**

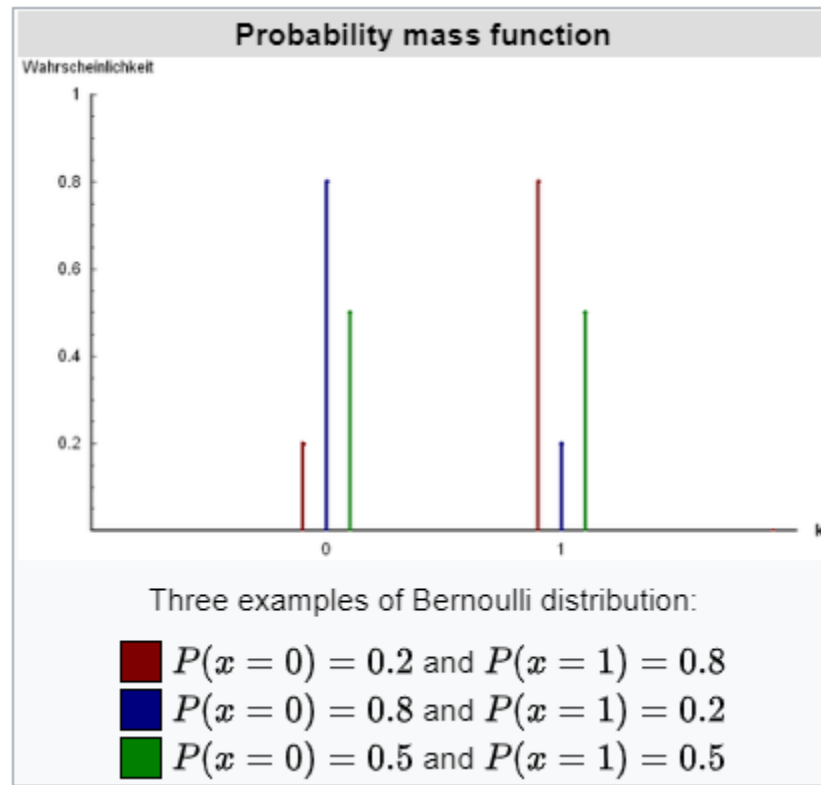
- If the random variable following uniform distribution is continuous then it is called continuous uniform distribution.



<b>Parameters</b>	$-\infty < a < b < \infty$
<b>Support</b>	$[a, b]$
<b>PDF</b>	$\begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$
<b>CDF</b>	$\begin{cases} 0 & \text{for } x < a \\ \frac{x-a}{b-a} & \text{for } x \in [a, b] \\ 1 & \text{for } x > b \end{cases}$
<b>Mean</b>	$\frac{1}{2}(a + b)$
<b>Median</b>	$\frac{1}{2}(a + b)$
<b>Mode</b>	any value in $(a, b)$

# Bernoulli distribution

- It is the discrete probability distribution of a random variable which takes the value 1 with probability  $p$  and the value 0 with probability  $q=1-p$ .



<b>Parameters</b>	$0 \leq p \leq 1$ $q = 1 - p$
<b>Support</b>	$k \in \{0, 1\}$
<b>PMF</b>	$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$
<b>CDF</b>	$\begin{cases} 0 & \text{if } k < 0 \\ 1 - p & \text{if } 0 \leq k < 1 \\ 1 & \text{if } k \geq 1 \end{cases}$
<b>Mean</b>	$p$
<b>Median</b>	$\begin{cases} 0 & \text{if } p < 1/2 \\ [0, 1] & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$
<b>Mode</b>	$\begin{cases} 0 & \text{if } p < 1/2 \\ 0, 1 & \text{if } p = 1/2 \\ 1 & \text{if } p > 1/2 \end{cases}$

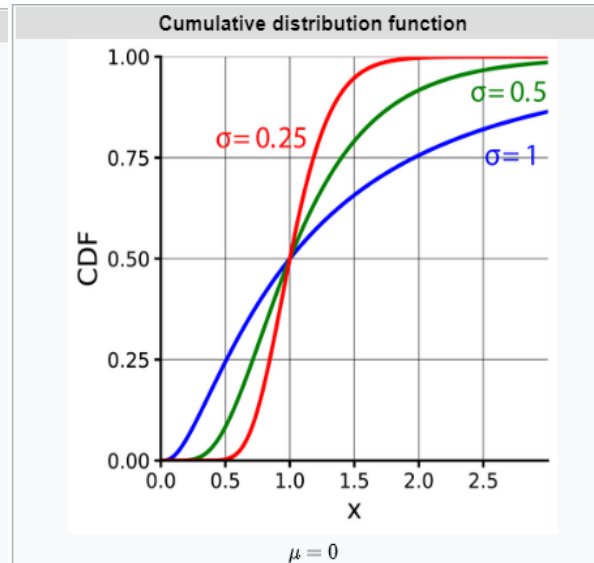
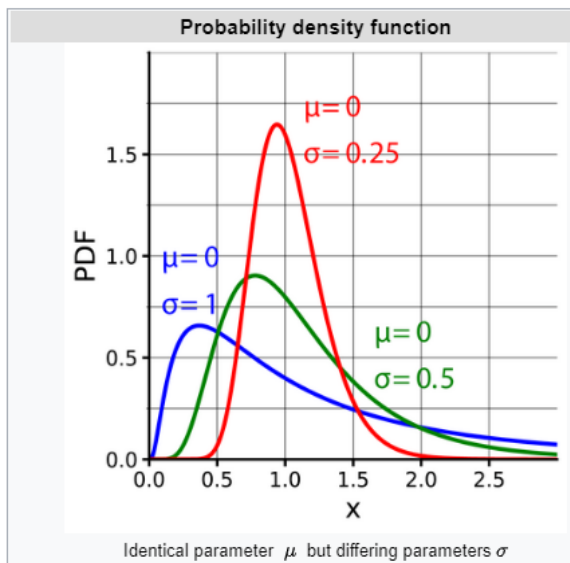
# Binomial Distribution

- The **binomial distribution** with parameters  $n$  and  $p$  is the discrete probability distribution of the number of successes in a sequence of  $n$  independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: *success* (with probability  $p$ ) or *failure* (with probability  $q=1-p$ ).

<b>Parameters</b>	$n \in \{0, 1, 2, \dots\}$ – number of trials $p \in [0, 1]$ – success probability for each trial $q = 1 - p$
<b>Support</b>	$k \in \{0, 1, \dots, n\}$ – number of successes
<b>PMF</b>	$\binom{n}{k} p^k q^{n-k}$
<b>CDF</b>	$I_q(n - \lfloor k \rfloor, 1 + \lfloor k \rfloor)$ (the <u>regularized incomplete beta function</u> )
<b>Mean</b>	$np$
<b>Median</b>	$\lfloor np \rfloor$ or $\lceil np \rceil$
<b>Mode</b>	$\lfloor (n+1)p \rfloor$ or $\lceil (n+1)p \rceil - 1$

# Log-normal distribution

- A **log-normal** (or **lognormal**) **distribution** is a continuous probability distribution of a random variable whose logarithm is normally distributed.
- Let  $X$  is a continuous random variable
- $y = \ln(X)$
- if  $y$  follows normal distribution then  $X$  follows log-normal distribution.



Notation	$\text{Lognormal}(\mu, \sigma^2)$
Parameters	$\mu \in (-\infty, +\infty)$ (logarithm of <u>scale</u> ), $\sigma > 0$
Support	$x \in (0, +\infty)$
PDF	$\frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$
CDF	$\frac{1}{2} \left[ 1 + \operatorname{erf}\left(\frac{\ln x - \mu}{\sigma\sqrt{2}}\right) \right] = \Phi\left(\frac{\ln(x) - \mu}{\sigma}\right)$

# Log-normal distribution

- **Popular Applications**

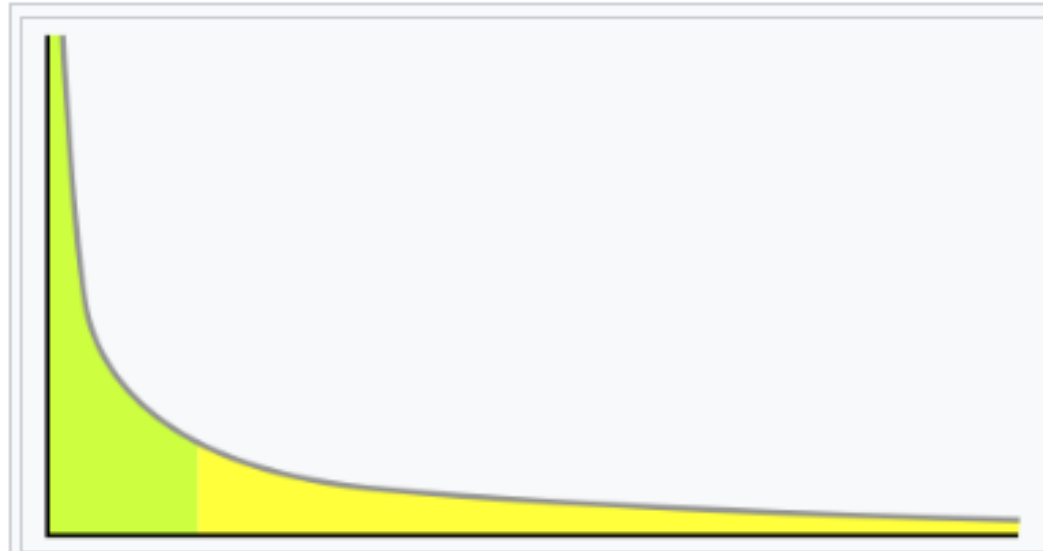
## Human behavior

- The length of comments posted in Internet discussion forums follows a log-normal distribution.
- Users' dwell time on online articles (jokes, news etc.) follows a log-normal distribution.
- The length of **chess** games tends to follow a log-normal distribution.
- Onset durations of acoustic comparison stimuli that are matched to a standard stimulus follow a log-normal distribution.

## Technology

- In **reliability** analysis, the log-normal distribution is often used to model times to repair a maintainable system.
- In **wireless communication**, "the local-mean power expressed in logarithmic values, such as dB or neper, has a normal (i.e., Gaussian) distribution." Also, the random obstruction of radio signals due to large buildings and hills, called **shadowing**, is often modeled as a log-normal distribution.
- Particle size distributions produced by comminution with random impacts, such as in **ball milling**.
- The **file size** distribution of publicly available audio and video data files (**MIME types**) follows a log-normal distribution over five **orders of magnitude**.
- File sizes of 140 million files on personal computers running the Windows OS, collected in 1999.
- Sizes of text-based emails (1990s) and multimedia-based emails (2000s).

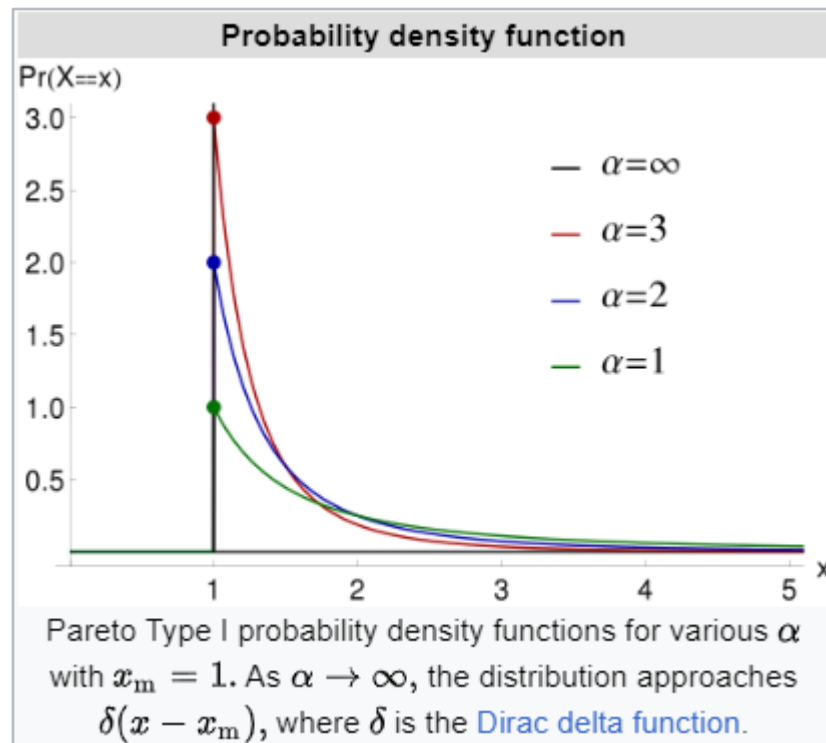
# Power Law Distribution



An example power-law graph that demonstrates ranking of popularity. To the right is the [long tail](#), and to the left are the few that dominate (also known as the [80–20 rule](#)).

# Pareto Distribution

- is a power-law probability distribution that is used in description of social, quality control, scientific, geophysical, actuarial, and many other types of observable phenomena.

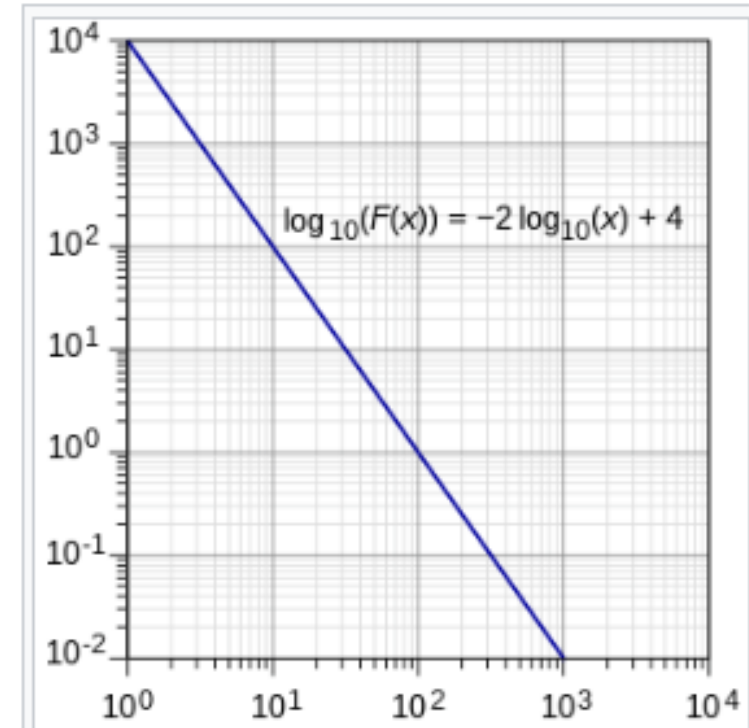


<b>Parameters</b>	$x_m > 0$ <u>scale</u> (real) $\alpha > 0$ <u>shape</u> (real)
<b>Support</b>	$x \in [x_m, \infty)$
<b>PDF</b>	$\frac{\alpha x_m^\alpha}{x^{\alpha+1}}$
<b>CDF</b>	$1 - \left(\frac{x_m}{x}\right)^\alpha$
<b>Quantile</b>	$x_m (1 - p)^{-\frac{1}{\alpha}}$
<b>Mean</b>	$\begin{cases} \infty & \text{for } \alpha \leq 1 \\ \frac{\alpha x_m}{\alpha - 1} & \text{for } \alpha > 1 \end{cases}$
<b>Median</b>	$x_m \sqrt[\alpha]{2}$
<b>Mode</b>	$x_m$



# Pareto Distribution

- Take the log of two variables
- Plot it.
- If a straight line in log-log plot appears, then it is necessary but insufficient for power laws



A straight line on a log-log plot is necessary but insufficient evidence for power-laws, the slope of the straight line corresponds to the power law exponent.

# Pareto Distribution

- Applications
  - All four variables of the household's budget constraint: consumption, labor income, capital income, and wealth.
  - The sizes of human settlements (few cities, many hamlets/villages)
  - File size distribution of Internet traffic which uses the TCP protocol (many smaller files, few larger ones)
  - [Hard disk drive](#) error rates
  - Clusters of [Bose–Einstein condensate](#) near [absolute zero](#)
  - The values of [oil reserves](#) in oil fields (a few [large fields](#), many [small fields](#))
  - The length distribution in jobs assigned to supercomputers (a few large ones, many small ones)
  - The standardized price returns on individual stocks
  - Sizes of sand particles
  - The size of meteorites

Reference: [https://en.wikipedia.org/wiki/Pareto\\_distribution#Occurrence\\_and\\_applications](https://en.wikipedia.org/wiki/Pareto_distribution#Occurrence_and_applications)

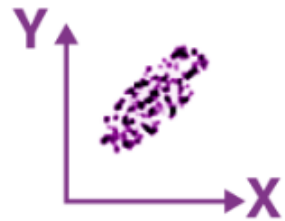
# Power Transform Box-Cox Transformation

- Suppose a random variable  $X = [x_1, x_2, \dots, x_n]$  follows pareto distribution.
- We want to apply a transformation to X, so that it will be transformed to  $Y = [y_1, y_2, \dots, y_n]$  which follows normal distribution.
- This can be achieved by using Box-Cox Transformation
- **Steps:**
  - $\text{Box-Cox}(X) = \lambda$
  - $y_i = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \lg(x_i) & \text{if } \lambda = 0 \end{cases}$

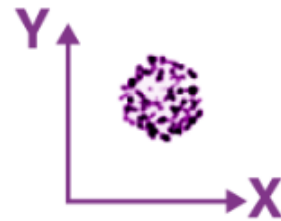
# Covariance

- **Covariance** is a measure of the *joint variability* of **two random variables**.

- $$Cov(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x) * (y_i - \mu_y)$$



$cov(X, Y) > 0$



$cov(X, Y) \approx 0$



$cov(X, Y) < 0$

- If **cov(X, Y)** is **greater than zero**, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.
- If **cov(X, Y)** is **less than zero**, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.
- If **cov(X, Y)** is **zero**, then we can say that there is no relation between two variables.

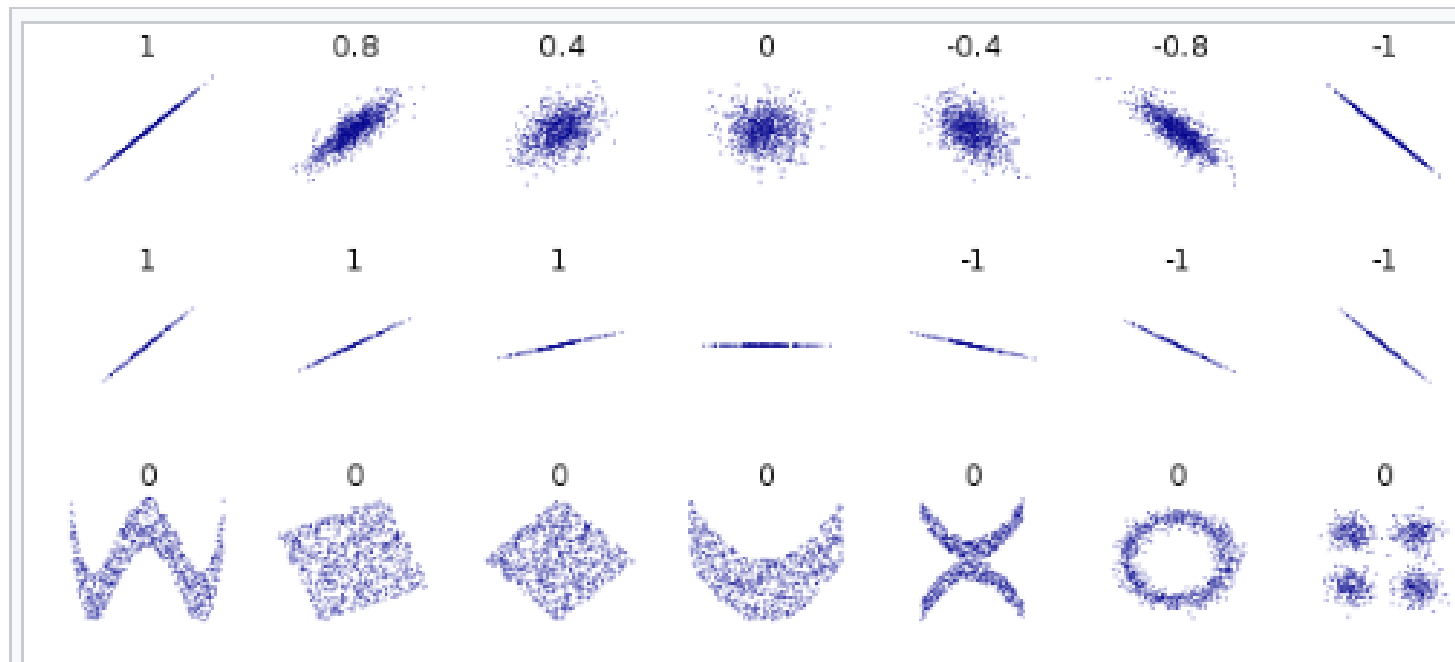
# Pearson's Correlation Coefficient

- The **Pearson correlation coefficient**( $\rho$ ) measures linear correlation between two sets of data.
- $$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$
- Where  $\sigma_x$  = Standard deviation of X
- $\sigma_y$ =Standard deviation of Y
- Minimum value -1
- Maximum value 1

# Pearson's Correlation Coefficient

Pearson correlation coefficient	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the same direction.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is no relationship between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and -1	Negative correlation	When one variable changes, the other variable changes in the opposite direction.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

# Pearson's Correlation Coefficient



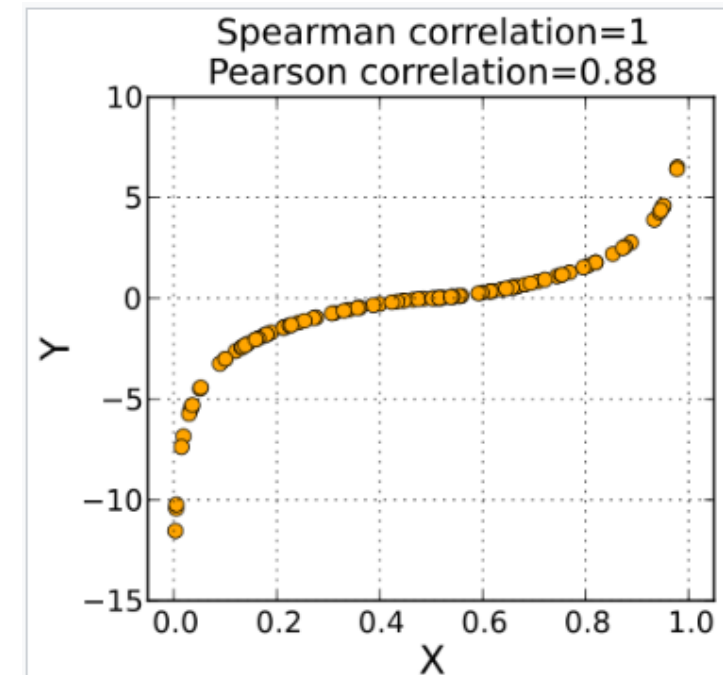
Several sets of  $(x, y)$  points, with the correlation coefficient of  $x$  and  $y$  for each set. The correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of  $Y$  is zero.

[https://en.wikipedia.org/wiki/Pearson\\_correlation\\_coefficient](https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

# Spearman's Rank Correlation Coefficient

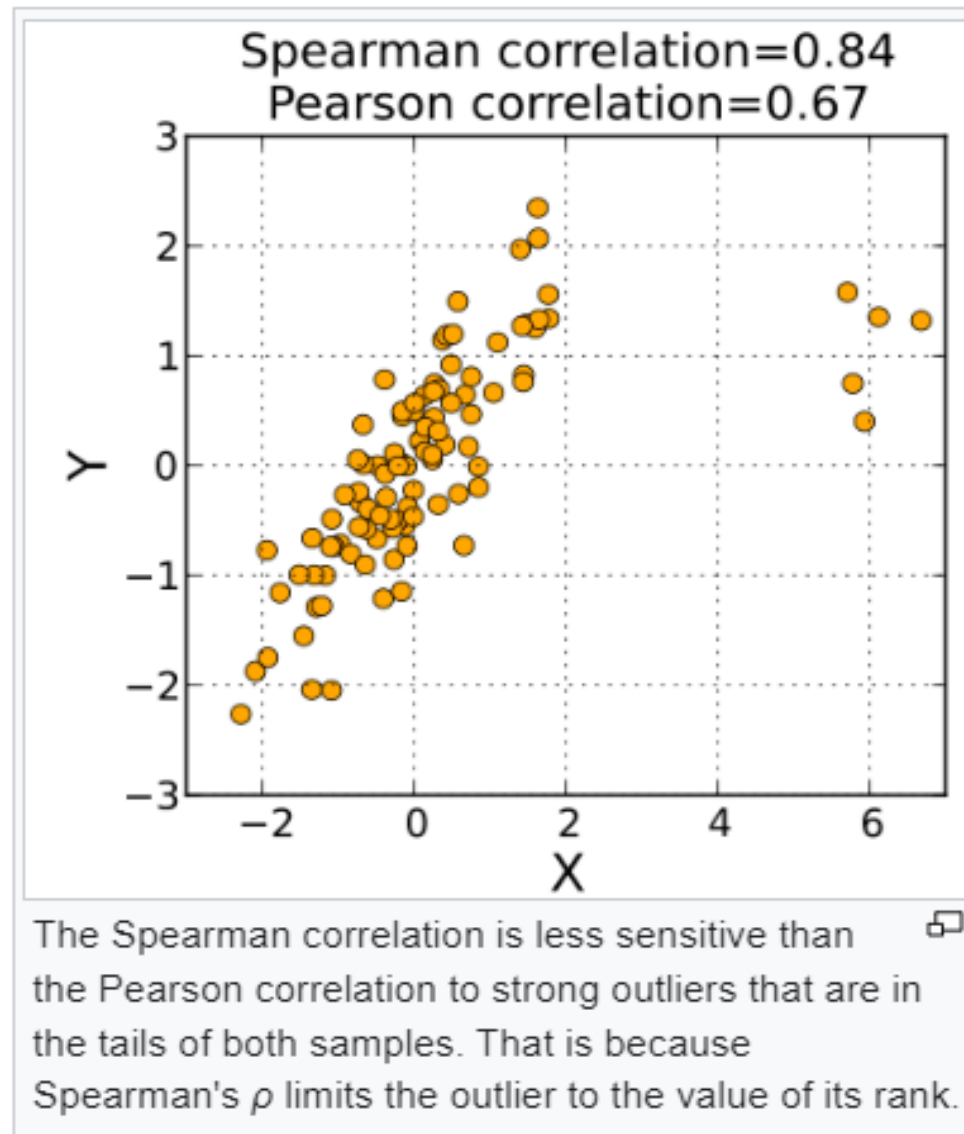
- The Spearman correlation between two variables is equal to the [Pearson correlation](#) between the rank values of those two variables.
- *Pearson's correlation* assesses *linear relationships*, **whereas**
- *Spearman's correlation* assesses *monotonic relationships* (whether linear or not)

X	Y	$R_x$	$R_y$
10	3		
15	7		
13	5		
8	1		



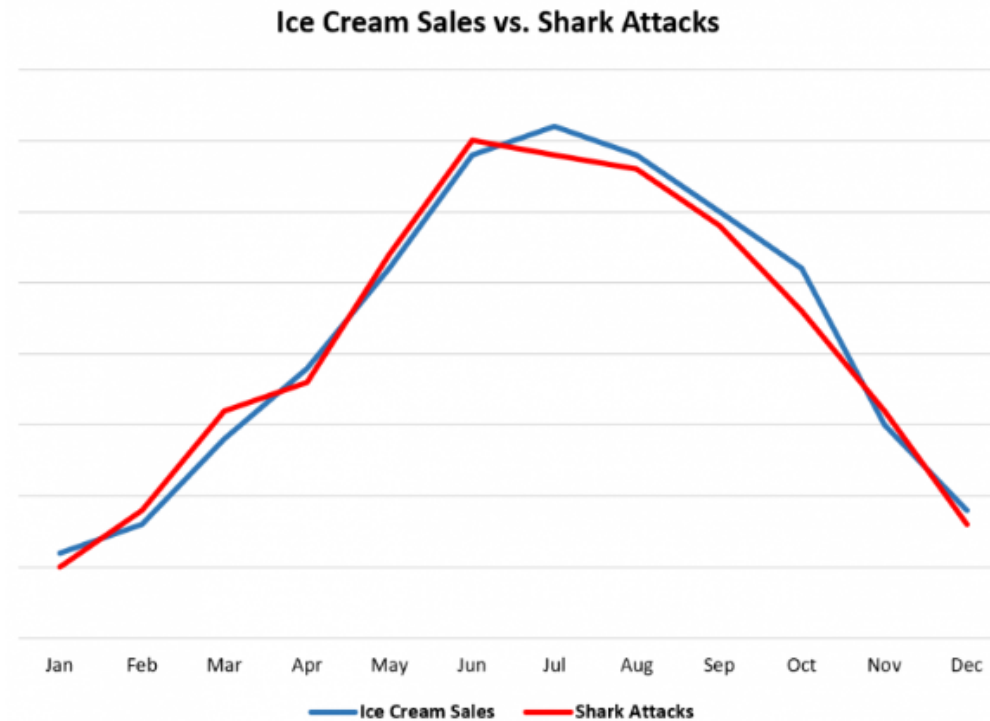


# Spearman's Rank Correlation Coefficient



# Correlation & Causation

- Correlation doesn't imply Causation

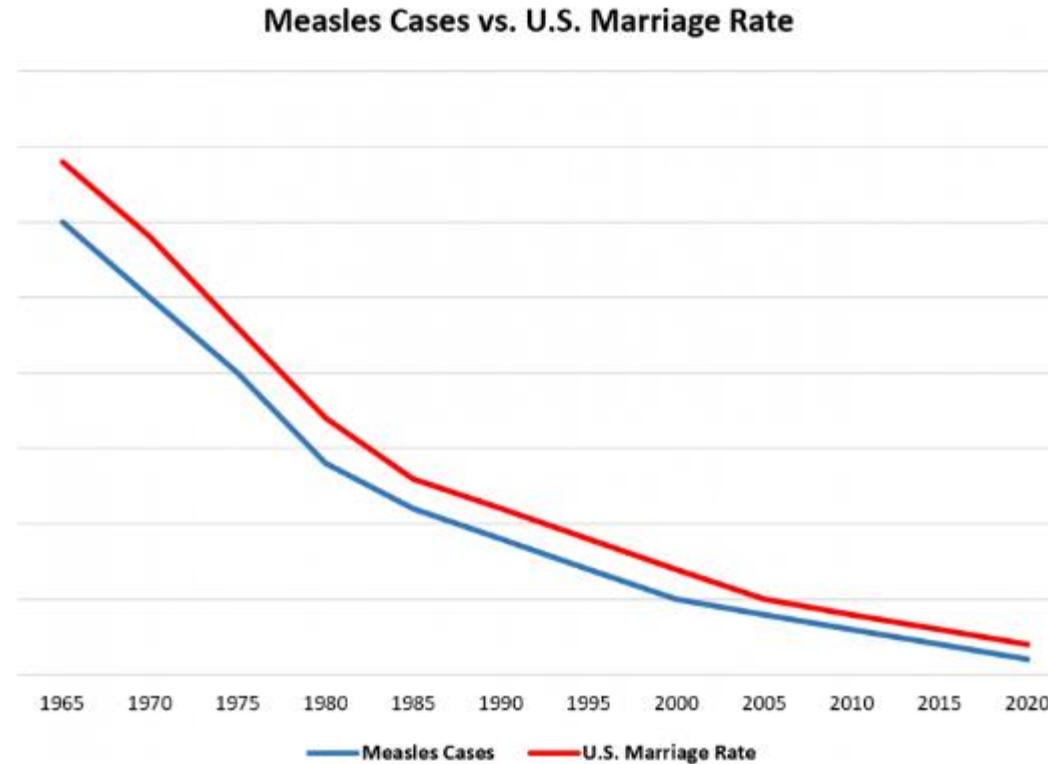


Does this mean that consuming ice cream causes shark attacks?

- Source: <https://www.statology.org/correlation-does-not-imply-causation-examples/>

# Correlation & Causation

- Correlation doesn't imply Causation

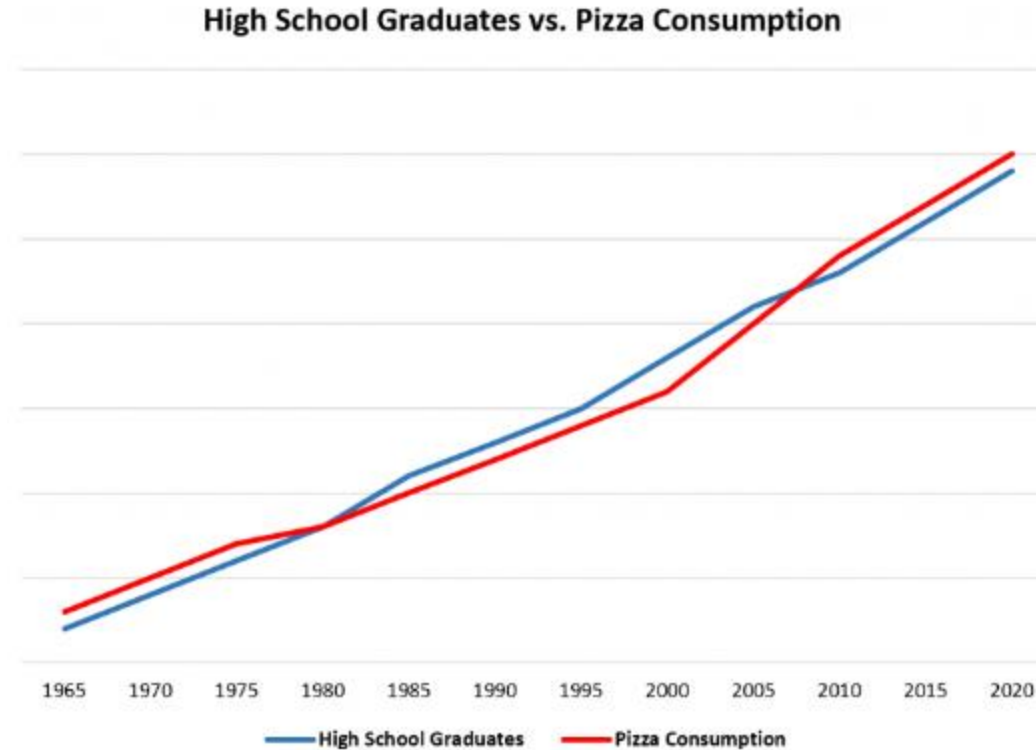


Does this mean that reduced measles cases is causing lower marriage rates?

- Source: <https://www.statology.org/correlation-does-not-imply-causation-examples/>

# Correlation & Causation

- Correlation doesn't imply Causation



Does this mean that an increased number of high school graduates is leading to more pizza consumption in the United States?

- Source: <https://www.statology.org/correlation-does-not-imply-causation-examples/>



For Your Valuable Time.