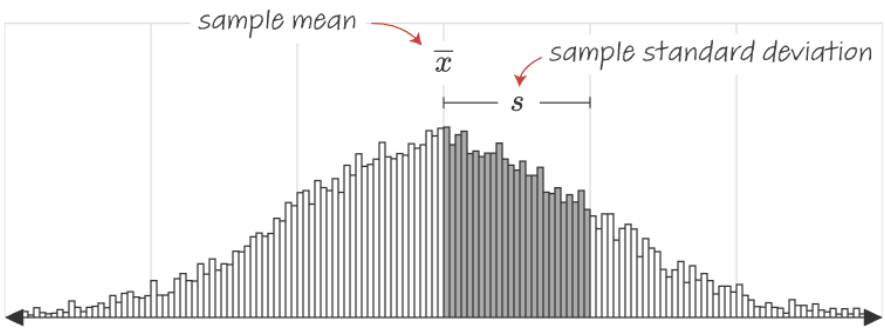
Sample Standard Deviation



$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

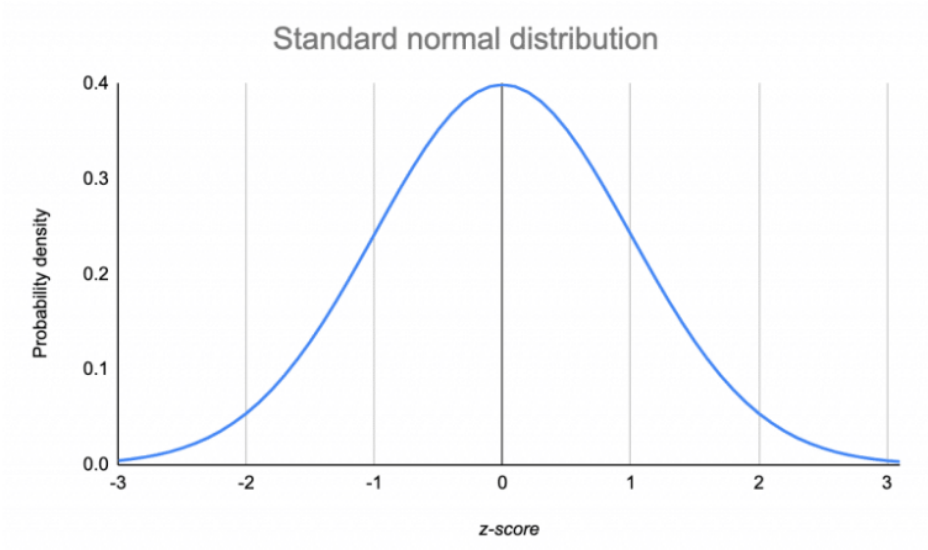
- The value for the standard deviation describes how closely the data set is to the mean. One standard deviation away from the mean on either side contains approximately 68.3% of the samples, two standard deviations contains approximately 95.4% of the samples, and so on. Because this formula calculates an approximate value for the true standard deviation there will be some discrepency between the actual normal distribution and the one modeled by the sample standard deviation.

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

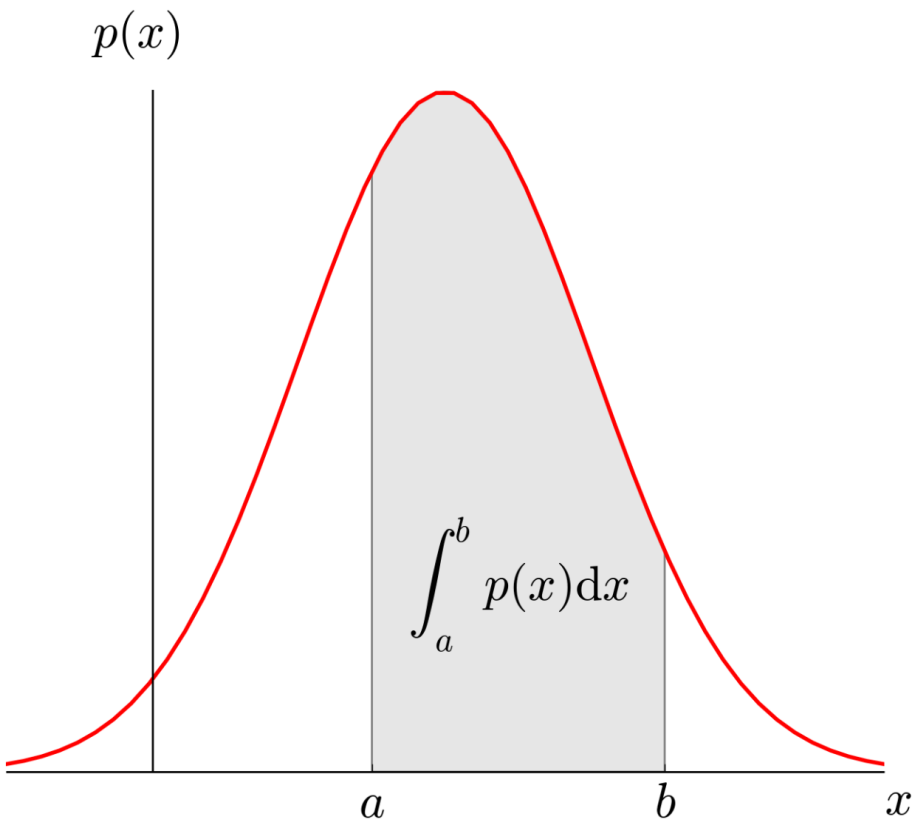
The Standard Normal Distribution

- A standard normal random variable is a normally distributed random variable with mean μ = 0 and standard deviation σ = 1. It will always be denoted by the letter Z.



The Probability Density Function(PDF)

- defines the probability function representing the density of a continuous random variable lying between a specific range of values. In other words, the probability density function produces the likelihood of values of the continuous random variable. Sometimes it is also called a probability distribution function or just a probability function.



Probability Density Function Properties

Let x be the continuous random variable with density function f(x), and the probability density function should satisfy the following conditions:

- For a continuous random variable that takes some value between certain limits, say a and b, the PDF is calculated by finding the area under its curve and the X-axis within the lower limit (a) and upper limit (b). Thus, the PDF is given by
  $P(x) = \int_a^b f(x)\,dx$
- The probability density function is non-negative for all the possible values, i.e. **f(x)≥ 0**, for all x.
- The area between the density curve and horizontal X-axis is equal to 1, i.e.
  $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- Due to the property of continuous random variables, the density function curve is **continued** for all over the given range. Also, this defines itself over a range of continuous values or the domain of the variable.

In [ ]:
```
1
```

What is a Cumulative Distribution Function?

- The Cumulative Distribution Function (CDF), of a real-valued random variable X, evaluated at x, is the probability function that X will take a value less than or equal to x. It is used to describe the probability distribution of random variables in a table. And with the help of these data, we can easily create a CDF plot in an excel sheet.
- In other words, CDF finds the cumulative probability for the given value. To determine the probability of a random variable, it is used and also to compare the probability between values under certain conditions. For discrete distribution functions, CDF gives the probability values till what we specify and for continuous distribution functions, it gives the area under the probability density function up to the given value specified.

Cumulative Distribution Function Formula

The CDF defined for a discrete random variable and is given as

$F_{X(x)} = P(X \leq x)$

Where X is the probability that takes a value less than or equal to x and that lies in the semi-closed interval (a,b], where a < b.

Therefore the probability within the interval is written as

$P(a < X \leq b) = F_X(b) - F_X(a)$

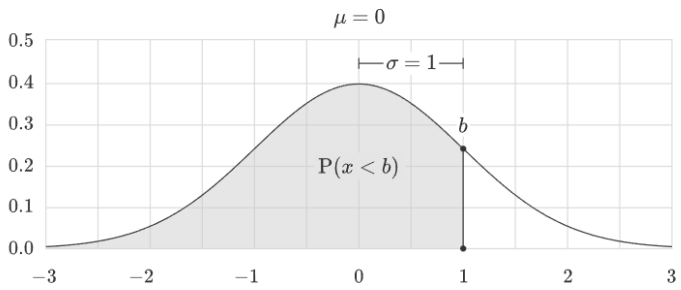
The CDF defined for a continuous random variable is given as;

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt$$

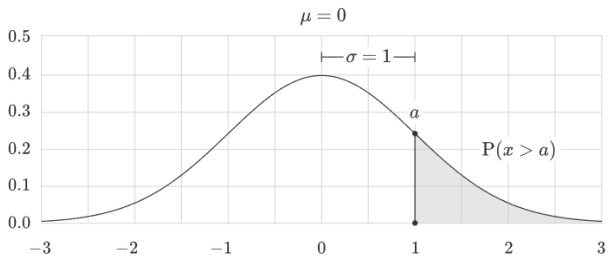
Here, X is expressed in terms of integration of its probability density function f_x.

In case, if the distribution of the random variable X has the discrete component at value b,

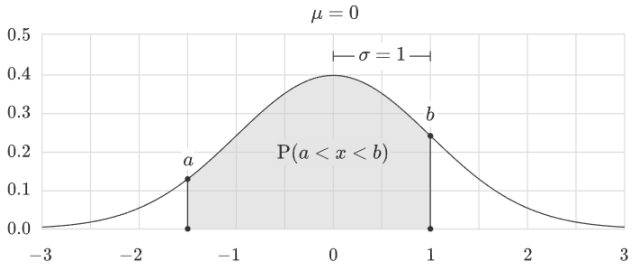
$P(X = b) = F_X(b) - \lim_{x \to b^-} F_X(x)$



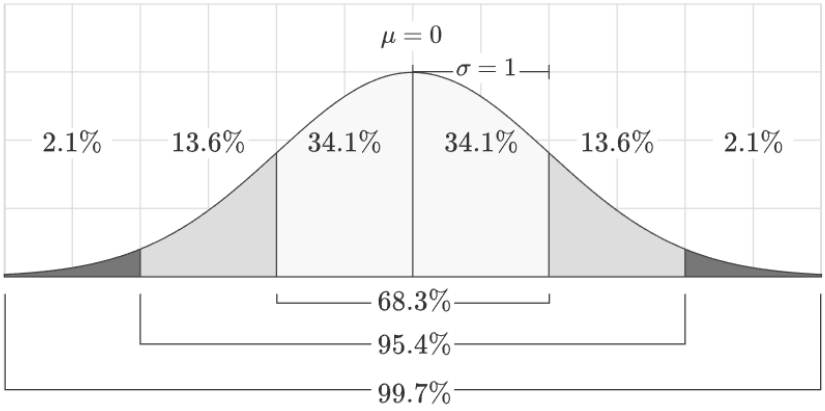
$$P(x > a) = 1 - P(x < a)$$



$$P(a < x < b) = P(x < b) - P(x < a)$$



Probability	Integral	Description
$P(x < b)$	$\int_{-\infty}^{b} f(x)dx$	The probability of an event occuring below a threshold b.
$P(x > a)$	$\int_{a}^{\infty} f(x)dx$	The probability of an event occuring above a threshold a.
$P(a < x < b)$	$\int_{a}^{b} f(x)dx$	The probability of an event occuring between a and b.



In [2]:
```
from scipy.stats import norm
```

In [3]:
```
norm.cdf(1)
```

Out[3]: 0.8413447460685429

In [4]:
```
norm.cdf(-1)
```

Out[4]: 0.15865525393145707

In [6]:
```
(norm.cdf(1) - norm.cdf(-1))*100   # P[x = -1 to x = 1] = 68.26 %      mean +/- 1σ
```

Out[6]: 68.26894921370858

In [7]:
```
(norm.cdf(2) - norm.cdf(-2))*100   # P[x = -2 to x = 2] = 95.44%       mean +/- 2σ
```

Out[7]: 95.44997361036415

In [8]:
```
(norm.cdf(3) - norm.cdf(-3))*100   # P[x = -3 to x = 3] = 99.73%       mean +/- 3σ
```

Out[8]: 99.73002039367398

Z - Scores (Standardisation)

$$z = \frac{x - \mu}{\sigma}$$

A Z-score is a numerical measurement that says , how far data point x is away from mean value.

Its is a standardized measure

If you're asked to find standardized values, use this formula to make your calculations:
Standardized Values:

calculate a standardized value (a z-score), using the above formula. The symbols are:

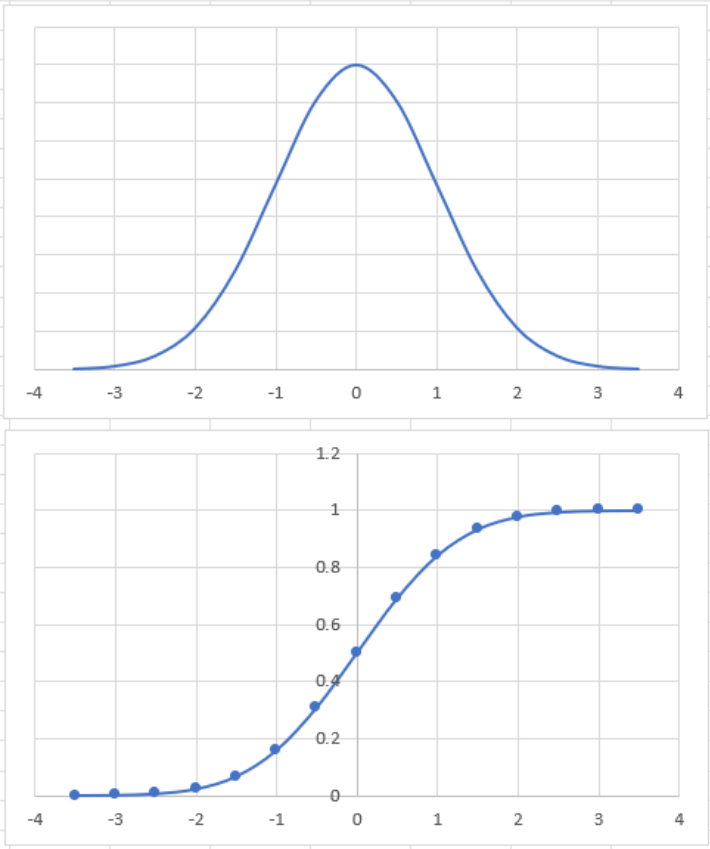
X: the observation (a specific value that you are calculating the z-score for).

Mu(μ): the mean.

Sigma(σ): the standard deviation.

In [ ]:
```
1
```

Z-score	pdf	cdf
-3.5	0.001	0.00023
-3	0.004	0.00135
-2.5	0.018	0.00621
-2	0.054	0.02275
-1.5	0.130	0.06681
-1	0.242	0.15866
-0.5	0.352	0.30854
0	0.399	0.50000
0.5	0.352	0.69146
1	0.242	0.84134
1.5	0.130	0.93319
2	0.054	0.97725
2.5	0.018	0.99379
3	0.004	0.99865
3.5	0.001	0.99977



Let X be a Gaussian random variable with mean 3 and variance 16, and let F denote the CDF of a "standard" Gaussian. What is P(X < 11)

72 users have participated

A	F(11)	17%
B	F(8)	36%
✓ C	F(2)	35%

```
1  μ = 3
2  σ = squar_root(16) = 4
3
4  P[X<=11] = (11-3)/4
5           = 8 / 4
6           = 2
7
```

In [16]:
```
norm.cdf(2)  # cummulative probability of z = 2
```

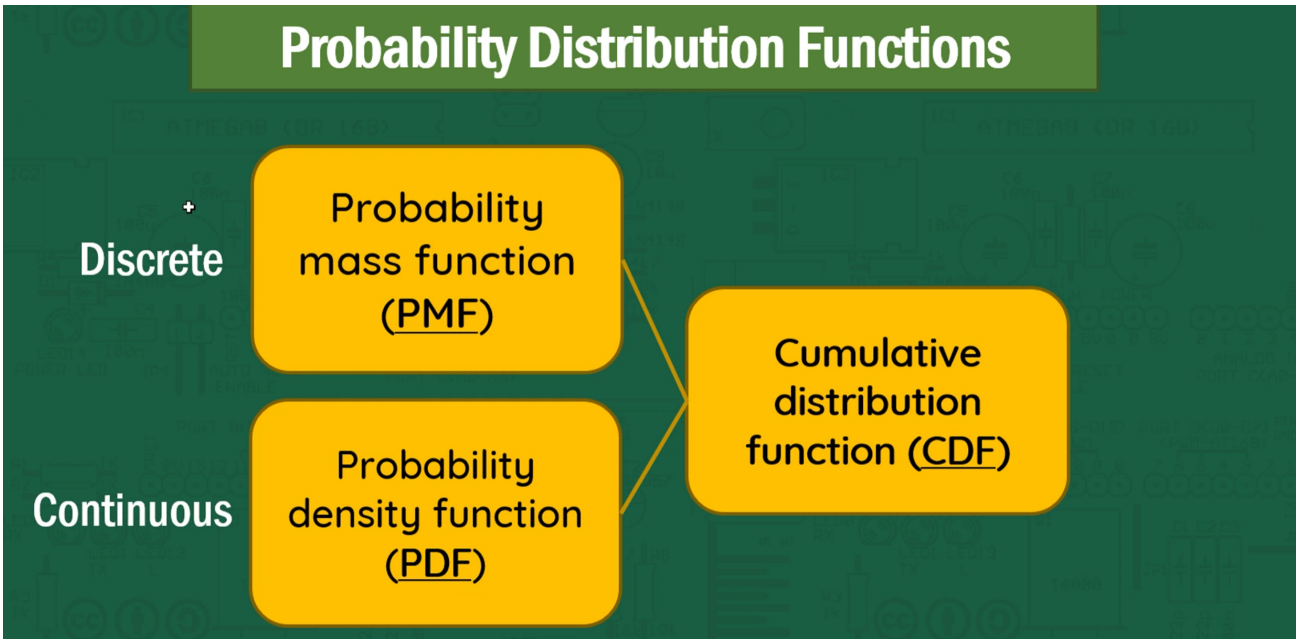
Out[16]: 0.9772498680518208

In [ ]:
```
# probability of 11 when, mean = 3 , varinace = 16
```

In [15]:
```
norm.cdf(11 , loc = 3 , scale = 4)    # P[x<=11]
                                      # = P[(x-3)/4 <= (11-3)/4]
                                      # = P[(x-3)/4 <= (8)/4]
                                      # = P[Z <= 2]
```

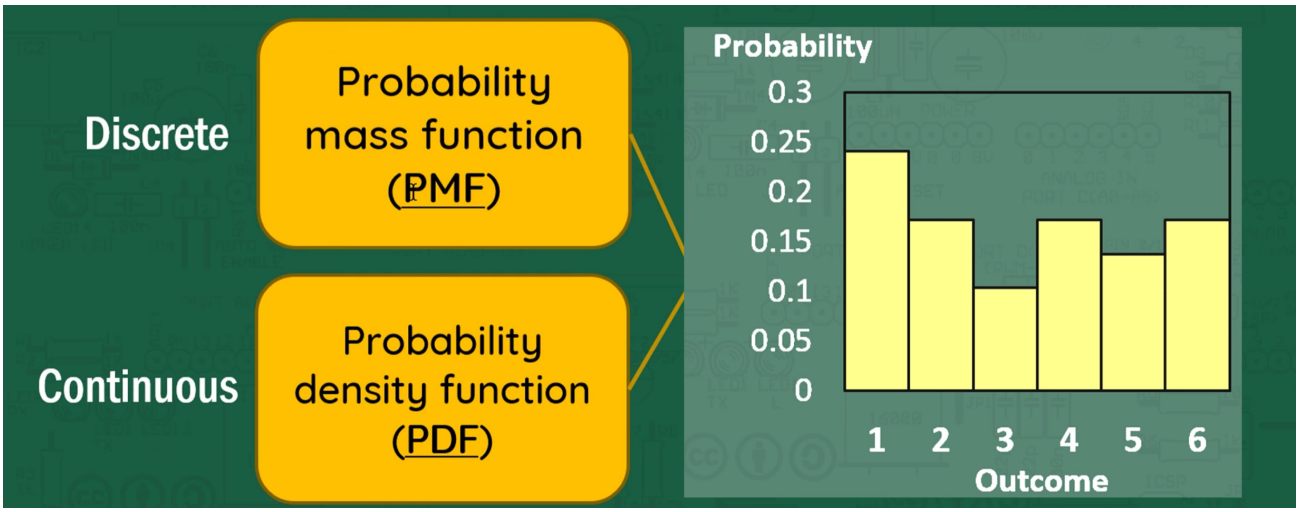
Out[15]: 0.9772498680518208

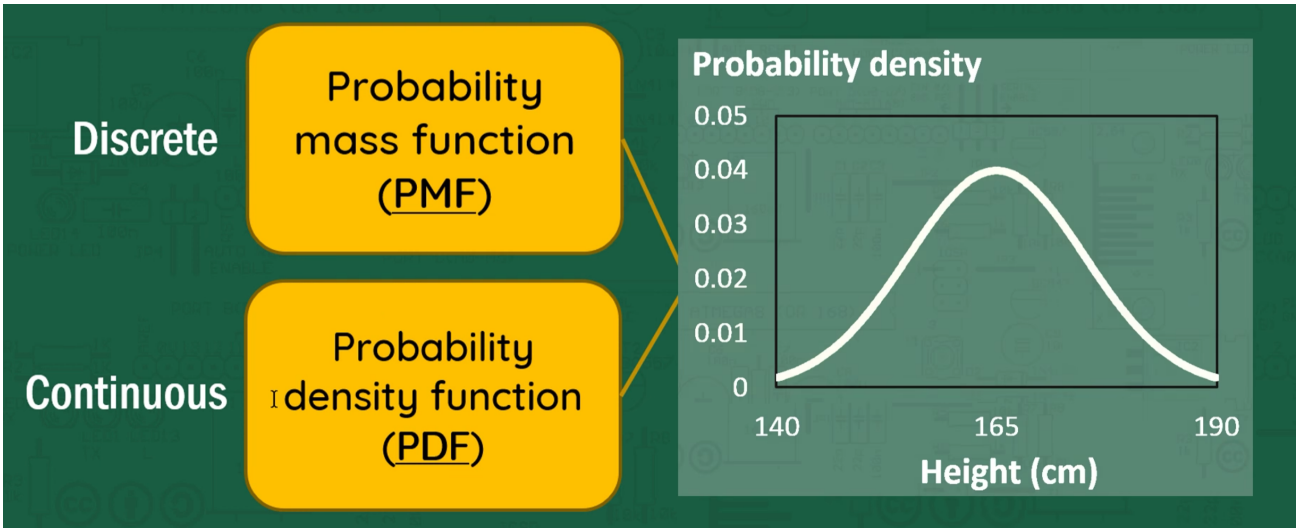
In [ ]:
```
# snapshots from : https://www.youtube.com/playlist?list=PLTNMv857s9WVzutwxaMb0YZKW7hoveGLS
```

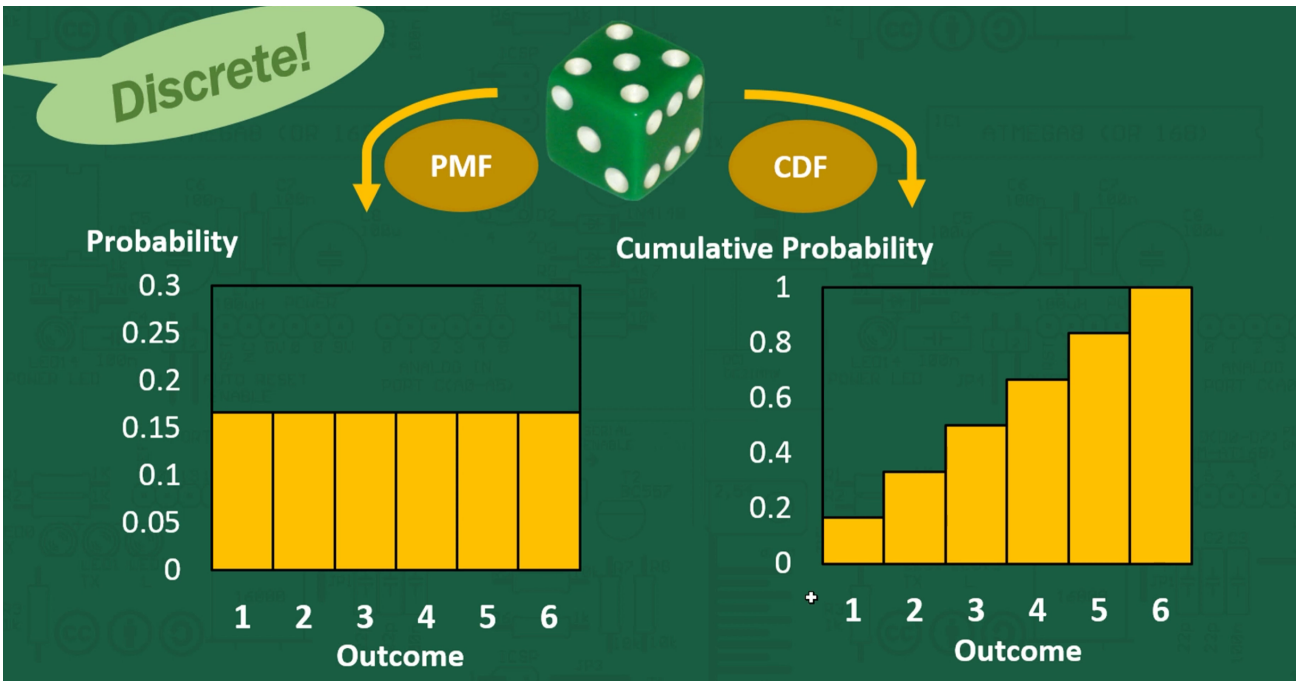


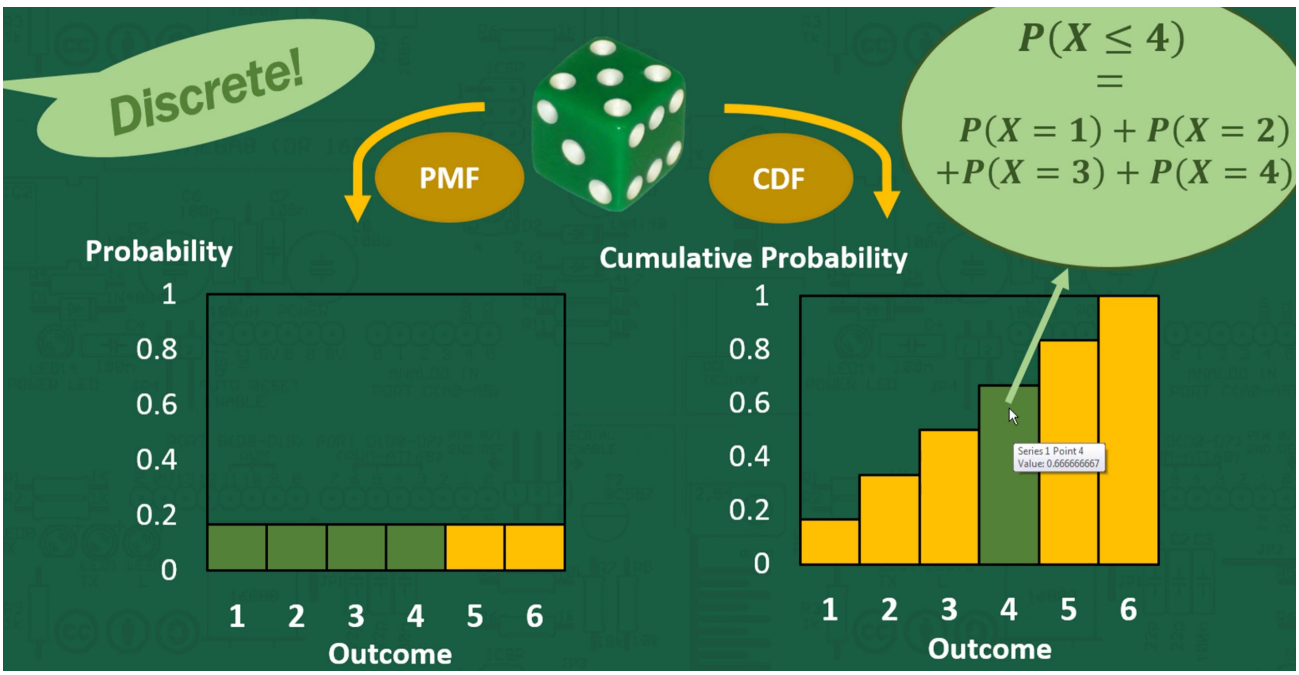
PMF : probability of each outcomes (discrete random variable)

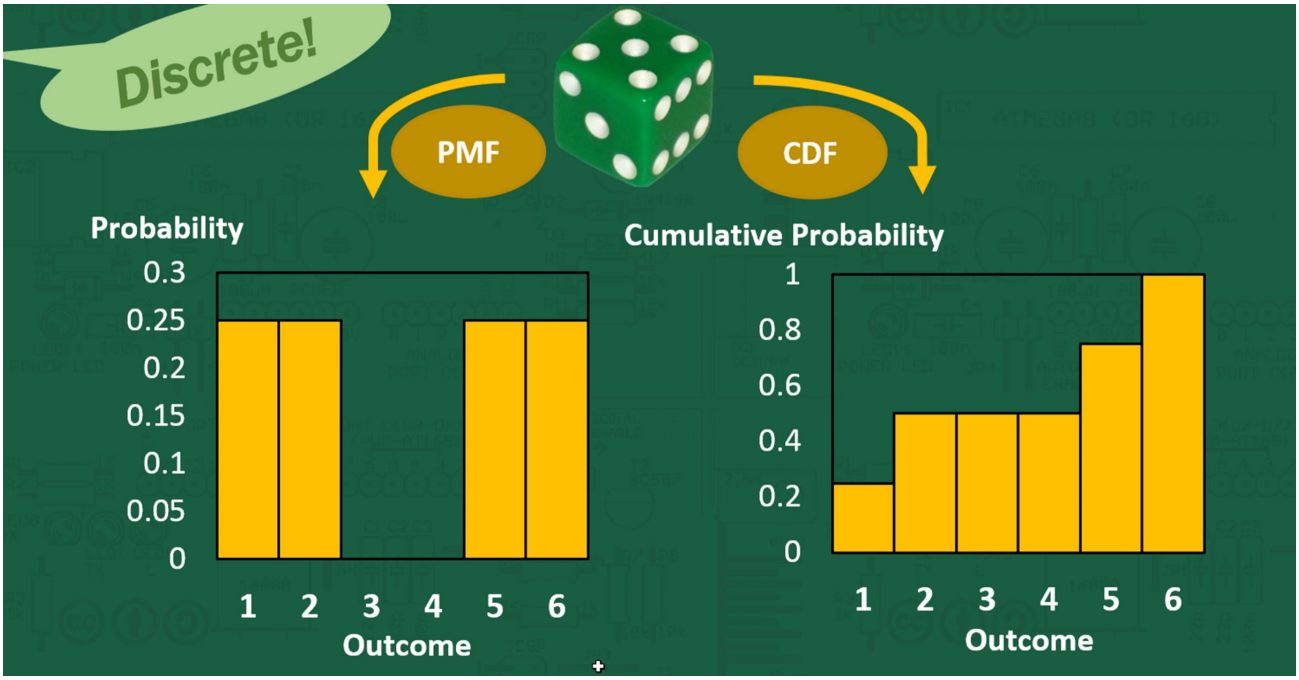
PDF : probability of continuous random variable

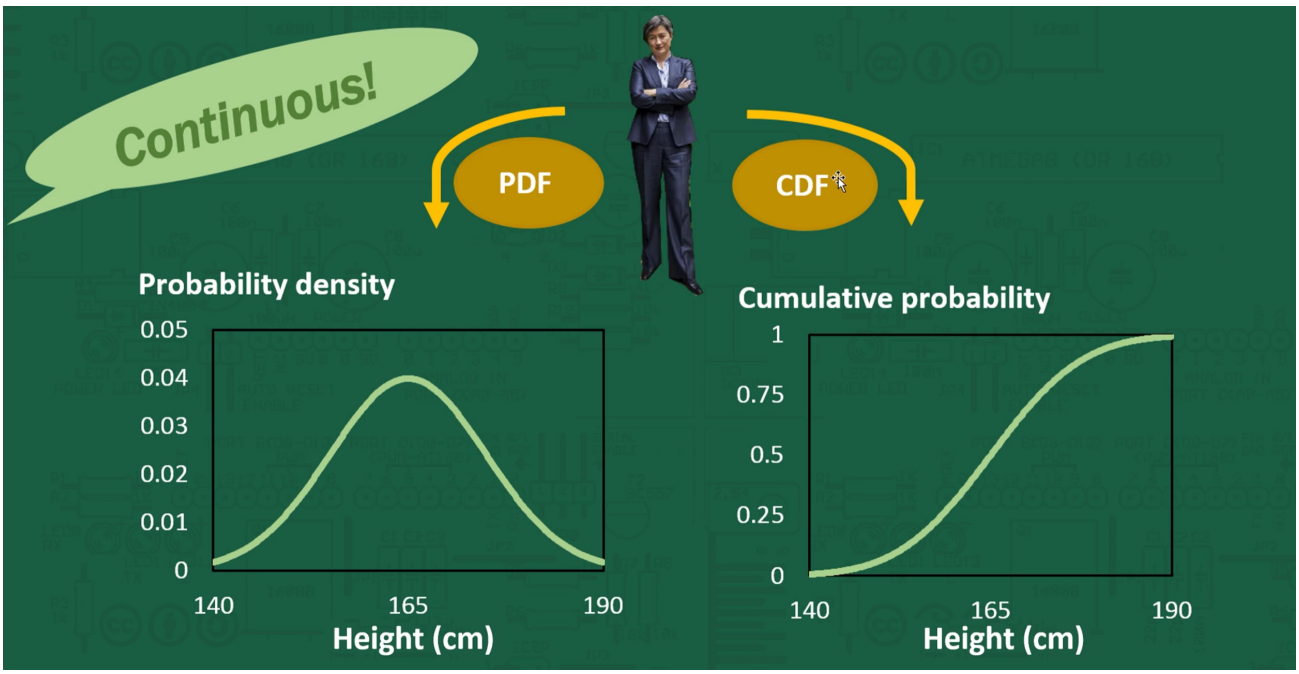


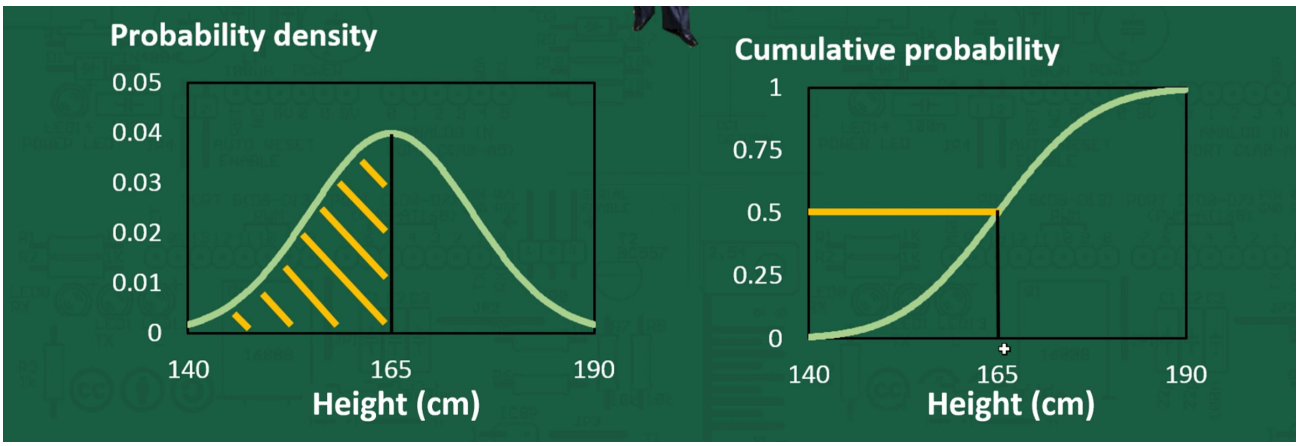

Discrete
Probability mass function (PMF)
Continuous
Probability density function (PDF)
Probability density
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Height (cm)

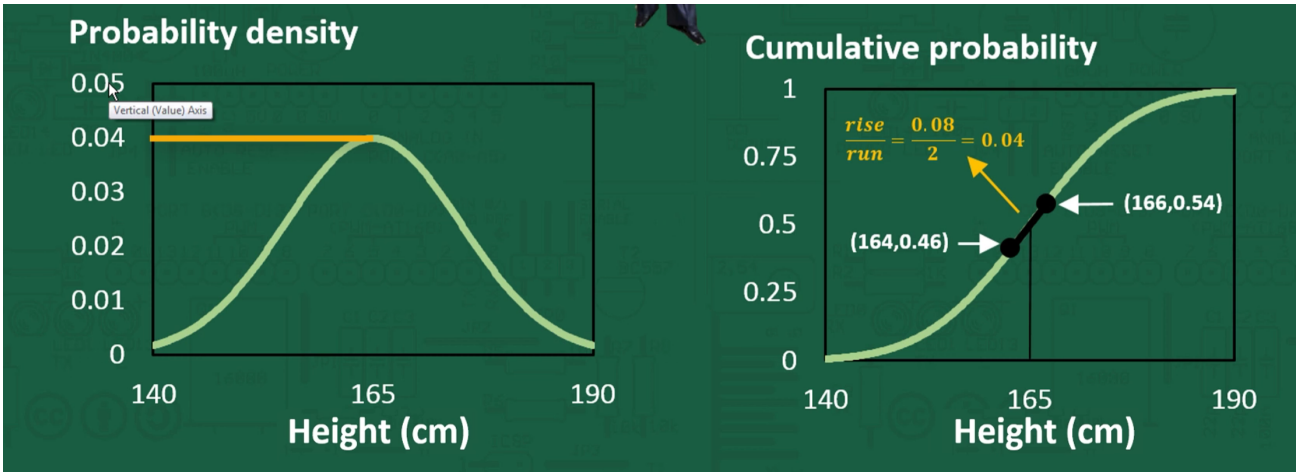

Discrete!
PMF
CDF
Probability
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6
Outcome
Cumulative Probability
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6
Outcome

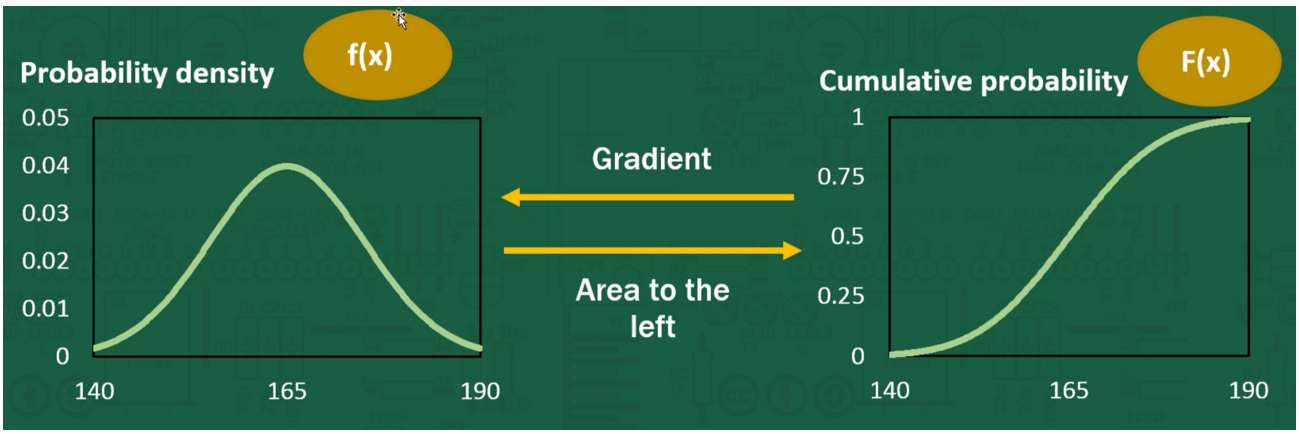

Discrete!
PMF
CDF
$P(X \leq 4)$
$=$
$P(X = 1) + P(X = 2)$
$+P(X = 3) + P(X = 4)$
Probability
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6
Outcome
Cumulative Probability
1
0.8
0.6
0.4
0.2
0
Series 1 Point 4
Value: 0.666666667
1 2 3 4 5 6
Outcome

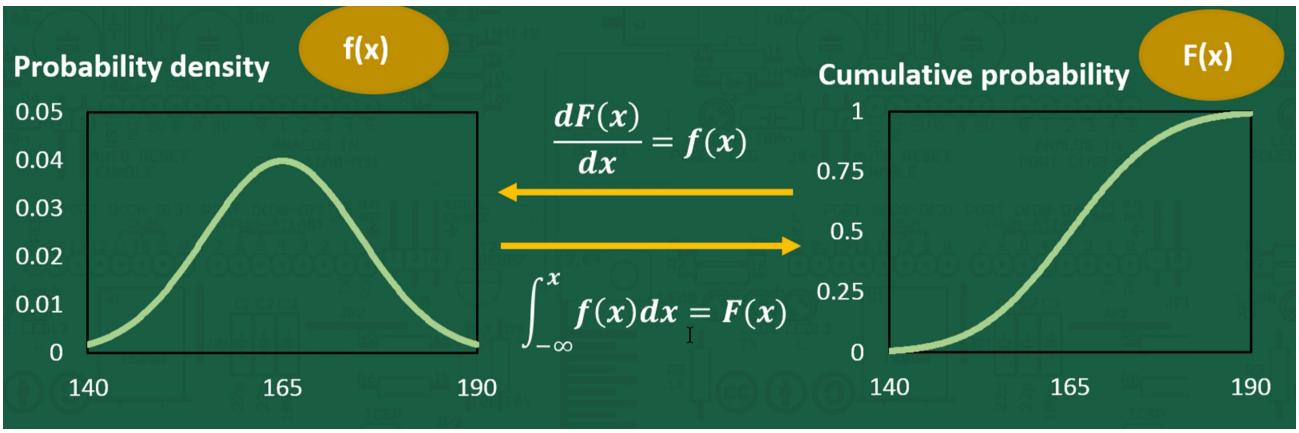

Discrete!
PMF
CDF
Probability
0.3
0.25
0.2
0.15
0.1
0.05
0
1 2 3 4 5 6
Outcome
Cumulative Probability
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6
Outcome


Continuous!
PDF
CDF
Probability density
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Height (cm)
Cumulative probability
1
0.75
0.5
0.25
0
140
165
190
Height (cm)


Probability density
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Height (cm)
Cumulative probability
1
0.75
0.5
0.25
0
140
165
190
Height (cm)







In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

Variance and its sq-root, the standard deviation are both measures of the spread or variability in data. Two or more data set could have the same mean , but very different variances .

variance is based on squared difference between each observation and mean .

the average squared deviation : How far is each data point from the center (mean) on average

Variance

- Variance is the expected value of the squared variation of a random variable from its mean value, in probability and statistics. Informally, variance estimates how far a set of numbers (random) are spread out from their mean value.
- The value of variance is equal to the square of standard deviation, which is another central tool.
- Variance is symbolically represented by σ2, s2, or Var(X).

In [ ]:

```
Expected value of variance  :
Var(X) = E[ (X −μ)2 ]
```

```
a variance is a measure of how far a set of data (numbers) are spread out from their mean (average) value.

Variance means to find the expected difference of deviation from actual value. Therefore, variance depends on the standard deviation of the given data set.

The more the value of variance, the data is more scattered from its mean and if the value of variance is low or minimum, then it is less scattered from mean. Therefore, it is called a measure of spread of data from mean.

For the purpose of solving questions, the formula for variance is given by:

Var(X) = E[ (X −μ)2 ]


Put into words; this means that variance is the expectation of the squared deviation of a random set of data from its mean value. Here,

X = Random variable

“μ” is equal to E(X) so the above equation may also be expressed as,

Var(X) = E[(X – E(X))^2]

Var(X) = E[ X2 -2X E(X) +(E(X))^2]

Var(X) = E(X2) -2 E(X) E(X) + (E(X))^2

Var(X) = E(X2) – (E(X))^2

Sometimes the covariance of the random variable itself is treated as the variance of that variable. Symbolically,

Var(X) = Cov(X, X)
```

As we know already, the variance is the square of standard deviation, i.e.,

Variance = (Standard deviation)2= σ2

The corresponding formulas are hence,

Population standard deviation σ =

$$\sqrt{\frac{\sum(X-\mu)^2}{N}}$$

and Sample standard deviation s =

$$\sqrt{\frac{\sum(x-x)^2}{n-1}}$$

Where X (or x) = Value of Observations

μ = Population mean of all Values

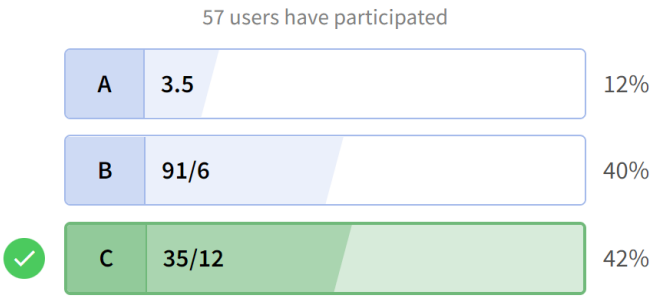
n = Number of observations in the sample set

$\bar{x}$

= Sample mean

N = Total number of values in the population

Calculate the variance of a random variable X which is the outcome of a fair dice: {1, 2, 3, 4, 5, 6}




```
Var(X) = E[ ( X - E(X) )^2 ]
       = E(X2) - (E(X))2

    X = {1,2,3,4,5,6}

E(X) = 3.5 = 7/2

X^2 = {1,4,9,16,25,36}
E(X^2) = (1*(1/6)) + (4*(1/6))  + (9*(1/6)) + (16*(1/6)) + (25*(1/6)) + (36*(1/6))

E(x^2) = 15.166 = 91/6

Var(X) = E[ ( X - E(X) )^2 ]
       = E(X2) - (E(X))2
       =  91/6 - (7/2)^2
       = 2.91
       = 35/12
```

In [17]:
```
(1*(1/6)) + (4*(1/6))  + (9*(1/6)) + (16*(1/6)) + (25*(1/6)) + (36*(1/6))
```

Out[17]: 15.166666666666666

In [19]:
```
(91/6) - ((7/2)**2)
```

Out[19]: 2.916666666666666

In [20]:
```
35/12
```

Out[20]: 2.9166666666666665

In [ ]:
```

```

In [ ]:
```
Expected values of variance :  Σ ((x - E(x))^2) * P(x)
```

In [ ]:
```

```

Expected Value and Variance of a Discrete Random Variable

Expected Value (or mean) of a Discrete Random Variable

For a discrete random variable, the expected value, usually denoted as μ or `E(x)` , is calculated using:

$$\mu = E(X) = \sum x_i f(x_i)$$

The formula means that we multiply each value, x, in the support by its respective probability, $f(x)$, and then add them all together. It can be seen as an average value but weighted by the likelihood of the value.

```
we were given the following discrete probability distribution:

0   1   2   3   4
1/5 1/5 1/5 1/5 1/5
What is the expected value?


```

$$\mu = E(X) = \sum x f(x) = 0\left(\frac{1}{5}\right) + 1\left(\frac{1}{5}\right) + 2\left(\frac{1}{5}\right) + 3\left(\frac{1}{5}\right) + 4\left(\frac{1}{5}\right) = 2$$

$$\text{Var}(X) = \left[0^2\left(\frac{1}{5}\right) + 1^2\left(\frac{1}{5}\right) + 2^2\left(\frac{1}{5}\right) + 3^2\left(\frac{1}{5}\right) + 4^2\left(\frac{1}{5}\right)\right] - 2^2 = 6 - 4 = 2$$

$$\text{SD}(X) = \sqrt{2} \approx 1.4142$$

Variance of a Discrete Random Variable

$$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 f(x_i)$$

The formula means that we take each value of x, subtract the expected value, square that value and multiply that value by its probability. Then sum all of those values.

There is an easier form of this formula we can use.

$$\sigma^2 = \text{Var}(X) = \sum x_i^2 f(x_i) - E(X)^2 = \sum x_i^2 f(x_i) - \mu^2$$

The formula means that first, we sum the square of each value times its probability then subtract the square of the mean. We will use this form of the formula in all of our examples.

Type *Markdown* and LaTeX: α^2

In [ ]:
```
1
```

Covariance

```
Covariance is a measure of the relationship between two random variables and to what extent,
they change together. Or we can say, in other words, it defines the changes between the two
variables, such that change in one variable is equal to change in another variable.
```

cov(x,y) = covariance betwen x and y variable

xi = data value of x yi = data value of x

x-bar = mean of x y-bar = mean of y N = nummber of data values

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N-1}$$

Type *Markdown* and LaTeX: α^2

```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

In [ ]:
```
1
```

Joint Distribution :

```
coin toss :
    dice  :

        h1 h2 h3 h4 h5 h6
        t1 t2 t3 t4 t5 t6
    X coin : {h,t} {h=1,t=0}
    Y dice : {1,2,3,4,5,6}

        P[x = 0 , y = 3] = 1/12

        P[y = 3] = ({y=3} and {x=0}) U ({y=3} and {x=1})
                 =  P({y=3} and {x=0}) + P({y=3} and {x=1})

```

```
# generalised :
x,y have a joint pmf:

    P[x = i , y = j]

P[x=i] =
= (P[x=i] and P[y=1]) U (P[x=i] and P[y=2]) U ......U (P[x=i] and P[y=j])
= sum( P[x=i and y=k] )

here, P[x = i] is marginal probability

```

In [ ]:
```
1
```

In [ ]:
```
1
```

```
# 3 types of batteries

NEW   WORKING   DEFECTIVE
 3       4          5

pick randomly 3 batteries .

```

```
X : no of new batteries
Y : no of working batteries

Q:

Y


3
2
1
0 1 2 3 X

X=0,y=0 means all batteries got defective
P[X=0 and y=0] =   5c3  / 12c3

P[x=0 and y=1] = 4c1 * 5c2 / 12c3
P[x=1 and y=1] = 3c1 * 4c1 5c1 / 12c3
P[x=1 and y=2] = 3c1 * 4c2 / 12c3
....
..
for entire table.


if asked P(y=1) = P(x=0,y=1) + P(x=1,y=1) + P(x=2,y=1) + P(x=3,y=1)

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```


```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [ ]:
```

```

In [90]:
```
import scipy.stats

```

In [88]:
```
import scipy.stats
scipy.stats.norm(loc=100, scale=12)
#where loc is the mean and scale is the std dev
#if you wish to pull out a random number from your distribution
scipy.stats.norm.rvs(loc=100, scale=12)

#To find the probability that the variable has a value LESS than or equal
#let's say 113, you'd use CDF cumulative Density Function
scipy.stats.norm.cdf(113,100,12)
# Output: 0.86066975255037792
#or 86.07% probability

#To find the probability that the variable has a value GREATER than or
#equal to let's say 125, you'd use SF Survival Function
scipy.stats.norm.sf(125,100,12)
# Output: 0.018610425189886332
#or 1.86%

#To find the variate for which the probability is given, let's say the
#value which needed to provide a 98% probability, you'd use the
#PPF Percent Point Function
scipy.stats.norm.ppf(.98,100,12)
# Output: 124.64498692758187
```

Out[88]:
```
124.64498692758187
```

In [91]:
```
scipy.stats.norm(loc=100, scale=12)
```

Out[91]:
```
<scipy.stats._distn_infrastructure.rv_frozen at 0x13b076a7d30>
```

In [96]:
```
scipy.stats.norm.rvs(loc=100, scale=12)
```

Out[96]:
```
86.58028716991166
```

In [98]:
```
scipy.stats.norm.cdf(113,100,12)
```

Out[98]:
```
0.8606697525503779
```

In [99]:
```
scipy.stats.norm.sf(125,100,12)
```

Out[99]:
```
0.018610425189886332
```

In [100]:
```
scipy.stats.norm.cdf(125,100,12)
```

Out[100]:
```
0.9813895748101137
```

In [101]:
```
1-0.9813895748101137
```

Out[101]:
```
0.018610425189886315
```

In [ ]:
```

```

In [ ]:
```
0.5,0.2,0.7
```

In [105]:
```
0.5*0.2*0.7
```

Out[105]:
```
0.06999999999999999
```

In [ ]:
```

```

In [174]:
```
arr = [float(x) for x in "1.764052345967664 0.4001572083672233 0.9787379841057392 2.240893199201
```

In [ ]:
```

```

In [175]:
```
arr,len(arr)
```

Out[175]:
```
([1.764052345967664,
  0.4001572083672233,
  0.9787379841057392,
  2.240893199201458,
  1.8675579901499675,
  -0.977277879876411,
  0.9500884175255894,
  -0.1513572082976979,
  -0.10321885179355784,
  0.41059850193837233,
  0.144043571160878,
  1.454273506962975,
  0.7610377251469934,
  0.12167501649282841,
  0.44386323274542566,
  0.33367432737426683,
  1.4940790731576061,
  -0.20515826376580087,
  0.31306770165090136,
```

In [176]:
```
arr = np.array(arr)
```

In [177]:
```
std = arr.std()
std
```

Out[177]:
```
1.1133678186681364
```

In [178]:
```
mean = arr.mean()
mean
```

Out[178]:
```
0.12524032797040804
```

In [179]:
```
upper = mean + (0.025774048*std)
upper
```

Out[179]:
```
0.1539363235704159
```

In [180]:
```
lower  = mean - (0.025774048*std)
```

In [181]:
```
lower,upper
```

Out[181]:
```
(0.0965443323704002, 0.1539363235704159)
```

In [182]:
```
for i in arr:
    if i < upper and i > lower:
        print("yess")
```

```
yess
yess
```

In [ ]:
```

```

In [ ]:
```

```

Central Limit Theorem:

If simple random samples of size n is selected from a distribution with mean μ and standard deviation $= \sigma$,then the distribution of sample means approach a normal distribution, as sample size increases, **irrespective of parent distribution of population** . The greater the sample size, the better is the approximation (typically $n \geq 30$).

In this condition, Sample mean follows approximately normal distribution

with Mean of sampling means:
$\mu_{\bar{x}} = \mu$

and

Standard deviation of Sample mean or Standard error:
$SE = \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Thank you

socratic.org/users/VarunStatsHelp

In [ ]:
```
1
```