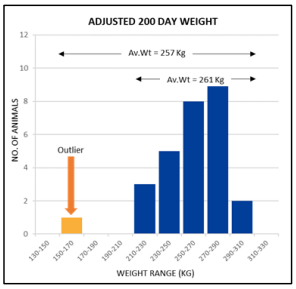
Probability & Statistics

- **Introduction to Probability and Statistics**
  - Experiment is a procedure that can be infinitely repeated and has sample space i.e. set of possible outcomes.
  - A random Experiment has more than one possible outcome, and deterministic has only one. E.g., rolling a dice is a random experiment as sample space is i.e., {1, 2, 3, 4, 5, 6}.
  - Random variable is possible outcome from a random experiment which is denoted with a capital letter.
  - The probability distribution for a random variable describes how the probabilities are distributed over the values of the random variable. There are 2 types as follows

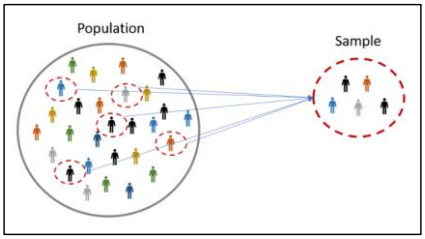
Discrete Random Variable	Continuous Random Variable
Countable set of distinct values	Any value within some interval (say 1 to 2)
Discrete data is counted	Continuous data is measured
Can take only integer values. Never include fractions or decimals.	Can take values including fractions and decimals.
Discrete data can only take certain values. Ex: Number of students in a class (you cannot have 56.5 students)	Continuous data can take any value, including decimal points (within a range) Ex: A person's height (167.54 cm) could be any value (within a range of human heights: 40 to 270 centimetres)
Examples: • Number of children in a family • Number of defective bulbs in a box of 10 • Number of ants born tomorrow • Number of classes missed last week (0,1,2..) • Toss of a coin • Number of heads in 4 flips of a coin (possible outcomes: 0,1,2,3,4) • Number of patients in hospital	Examples: • Amount of sugar in a coffee • Amount of rain in a day • Time to finish a test • Percentage of marks obtained by a student • Length of a chord of a circle (any number of decimal places) • Height of individuals • Hours spent exercising last week • Time required to finish a test
	

- **Outliers**



  - An outlier is a data point that differs significantly from other observations.
  - An outlier may be due to variability in the measurement or experimental error.
  - It corrupts the mean, standard deviation obtained.

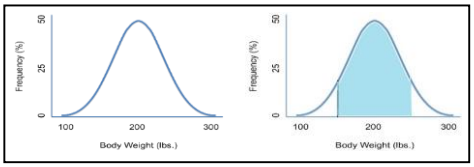
- **Population and sample**



  - Population is complete dataset having huge no of data points.
  - Sample is a subset of population data which is derived based on one or more observations.
  - As the sample size increases, $\mu_{sample}(\bar{X}) \sim \mu_{population}(\mu)$

- **Probability Mass Function**
  - If X is a discrete random variable then its range R_X is a countable set for a particular event,
    $$R_X = \{x_1, x_2, x_3, \dots\}$$
  - x_1, x_2, x_3 , … are possible values of the random variable X
  - For a discrete random variable X, we want the probabilities of $X = x_k$.
  - The event $A = \{X = x_k\}$ is defined as the set of outcomes s in the sample space S given that X is equal to x_k. $A = \{s \in S | X(s) = x_k\}$
  - The probabilities of events $\{X = x_k\}$ are shown by the probability mass function (PMF) of X. It gives probabilities of the possible values for a random variable.
  - $P_X(1)$ shows the probability that $X = 1$ for random variable X.
  - Example- fair coin is tossed twice, and let X be observed number of heads. Find the range of $X(R_X)$, PMF P_X
    - $Sample\ Space = S = \{HH, HT, TH, TT\}$
    - The number of heads will be 0,1,2. Thus $R_X = \{0,1,2\}$
    - The PMF is defined as, $P_X(k) = P(X = k)$ for $k = 0,1,2$
    - $P_X(0) = P(X = 0) = P(TT) = \frac{1}{4}$
    - $P_X(1) = P(X = 1) = P(\{HT, TH\}) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
    - $P_X(2) = P(X = 2) = P(HH) = \frac{1}{4}$

  - PMF is defined for values in the range of outcome. It is convenient to extend the PMF of X to all real numbers.
    $$x \notin R_X, P_X(x) = P(X = x) = 0.$$
  - Thus, in general we can write $P_X(x) = \begin{Bmatrix} P(X = x) & if\ x \in R_X \\ 0 & else \end{Bmatrix}$
- **Probability Density Function**



  - A continuous random variable Y takes an infinite number of possible values. E.g. Human Weights
    $$R_Y = \{180\ lb, 151.2\ lb, 201.9999999999\ lb, \cdots\}$$
  - Thus $P(Y = y) = 0$ as probability that a person will weigh exactly 180lbs? it can be $180.00001\ lb$ or $179.9999\ lb$. odds of someone weighing **exactly** 180 pounds is so tiny it's practically zero.
  - We need to find the probability that $P(a < Y < b)$. we do this using probability density function.
  - We find the shaded area in PDF; it is approximately 75 percent. So, $P\ (150 < Y < 250) = 75\%$.
  - We need to think about the probability that y is close to a single number instead of it's exactly a single number which is given with a probability density function $\rho(y)$.
  - If the probability density around a point y is large, that means the random variable Y is likely to be close to y. if $\rho(y) = 0$ in some interval, then Y won't be in that interval.
  - To convert probability density $\rho(y)$ into a probability, imagine that Iy is some small interval around the point y. The probability that Y is in that interval:
    $$P(Y \in Iy) \approx \rho(y) \times Length\ of\ Iy$$
  - The probability $P(Y \in Iy)$ approaches zero as Iy shrinks down to the point y.
  - To determine the probability that Y is in range $I(a, b)$, integrate the function $\rho(y)$ over this range
    $$Pr(Y \in I) = \int_a^b \rho(y)\,.\,dy$$

- For a function $\rho(y)$ to be a PDF, it must be non-negative, so the that integral is always non-negative, and it must integrate to one, so that the probability of Y being something is one

$$\rho(y) \geq 0 \; for \; all \; y$$

$$\int \rho(y)dy = 1$$

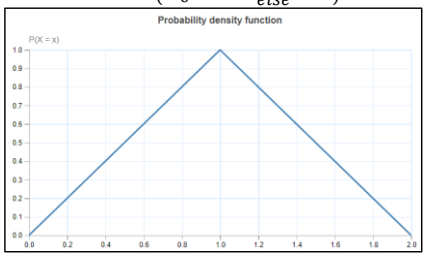
$$\rho \; is \; continuous$$

- if we aren't worrying about being too precise or about discontinuities in ρ, we may sometimes state that

$$Pr(Y \in (y, y + dy)) = \rho(y).dy$$

- Here, dy is a v.v. small number so that $(y, y + dy)$ is a v.v. small interval $I_y$around y, in which case the approximation becomes exact, at least if ρ is continuous.
- Example- Let the random variable X denote the time a person waits for an elevator to arrive. the longest one would need to wait for the elevator is 2 minutes, so that the possible values of X (in minutes) are given by the interval [0,2]. A pdf for X is given by

$$\rho(x) = \begin{Bmatrix} x & for\ 0 \leq x \leq 1 \\ 2 - x & for\ 1 < x \leq 2 \\ 0 & else \end{Bmatrix}$$



- The plot satisfies 1st and 3rd conditions stated above.
  a. From the graph, it is clear that $f(x) \geq 0$, for all $x \in R$
  b. Since there are no holes, jumps, asymptotes, we see that $f(x)$ is continuous
- Now we will check if area under curve is 1 or not.
  - ⇨ $\int_{-\infty}^{\infty} f(x).dx = \int_0^2 x.dx$
  - ⇨ $\int_0^1 x\,.dx + \int_1^2 (2 - x)\,.dx$
  - ⇨ $\left(\frac{x^2}{2}|_0^1\right) + \left(2x - \frac{x^2}{2}|_1^2\right)$
  - ⇨ $\left(\frac{1^2}{2} - \frac{0^2}{2}\right) + \left[\left(2 * 2 - \frac{2^2}{2}\right) - \left(2 * 2 - \frac{1^2}{2}\right)\right]$
  - ⇨ $0.5 + 0.5 = 1$
- probability that a person waits less than 30 seconds (or 0.5 minutes) for the elevator to arrive.

$$P(0 \leq X \leq 0.5) = \int_0^{0.5} f(x).dx = \int_0^{0.5} x.dx = \frac{x^2}{2}|_0^{0.5} = \frac{0.5^2}{2} - \frac{0^2}{2} = 0.125$$

- **Cumulative Distribution Function (CDF)**
  - CDF of a random variable X, evaluated at x, is the probability that X will take a value less than or equal to x.
  - Let X have pdf f , then the cdf F is given by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\,.dx, for\ all\ x\ \in\ \mathbb{R}$$

  - The probability that X lies in the semi-closed interval (a, b) where $a < b$

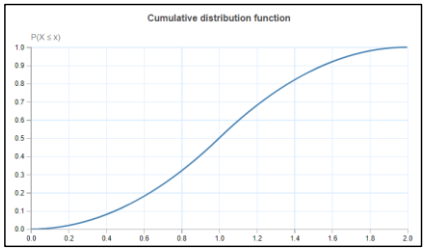
$$P(a < X \leq b) = \ F_X(a) - \ F_X(b)$$

  - The CDF for a continuous random variable is found by integrating the PDF.
  - The PDF of a continuous random variable can be found by differentiating the CDF.

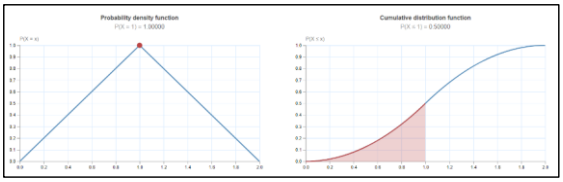
$$f(x) = \frac{d}{dx}[F(x)]$$

  - First, let's find the cdf at two possible values of X, $x = 0.5$ and $x = 1.5$
    - ⇨ $F_X(0.5) = P(X \leq 0.5) = \int_{-\infty}^{0.5} f(x)\,.dx = \int_0^{0.5} x\,.dx = \frac{x^2}{2}|_0^{0.5} = \frac{0.5^2}{2} - \frac{0^2}{2} = 0.125$
    - ⇨ $F_X(1.5) = P(X \leq 1.5) = \int_{-\infty}^{1.5} f(x)\,.dx = \int_0^1 x\,.dx + \int_1^{1.5} (2 - x)\,.dx$
    - ⇨ $\left(\frac{x^2}{2}|_0^1\right) + \left(2x - \frac{x^2}{2}|_1^{1.5}\right)$
    - ⇨ $\left(\frac{1^2}{2} - \frac{0^2}{2}\right) + \left[\left(2 * 1.5 - \frac{1.5^2}{2}\right) - \left(2 * 1 - \frac{1^2}{2}\right)\right]$
    - ⇨ $0.5 + 1.875 - 1.5 = 0.875$

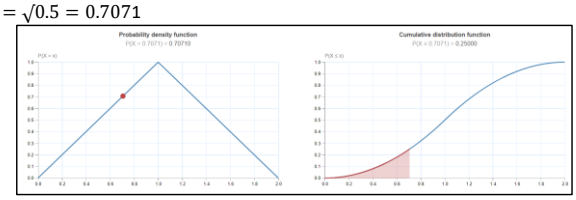
- Now we find $F(x)$ more generally, working over the intervals that $f(x)$ has different formulas:
  - ⇨ $for\ x < 0$: $F_X(x) = \int_{-\infty}^{x} 0\,.dx = 0$
  - ⇨ $for\ 0 \le x \le 1$: $F_X(x) = \int_{-\infty}^{x} x\,.dx = \frac{x^2}{2}\,|_0^x = \frac{x^2}{2}$
  - ⇨ $for\ 1 \le x \le 2$: $F_X(x) = \int_0^1 x\,.dx + \int_1^x (2-x)\,.dx = \frac{x^2}{2}\,|_0^1 + \left(2x - \frac{x^2}{2}\right)|_1^x$
  - ⇨ $0.5 + \left(2x - \frac{x^2}{2}\right) - (2 - 0.5) = 2x - \frac{x^2}{2} - 1$
  - ⇨ $for\ x > 2$: $F_X(x) = \int_{-\infty}^{x} f(x)\,.dx = 1$
- Plotting the CDF at above derived values,



- **Percentiles of a Distribution**
  - The $(100p)^{th} percentile$ ($0 \le p \le 1$) of a probability distribution with CDF F is the value $x = \pi_p$ such that
    $$F(x) = P(X < x) = p$$
  - To find the percentile π_p of a continuous random variable, which is a possible value of the random variable, we are specifying a cumulative probability p and solving the following equation for π_p
    $$\int_{-\infty}^{x} f(x)\,.dx = p$$
  - Median or 50th percentile: $\pi_{0.5} = Q2$, separates probability (area under pdf) into two equal halve. So, we have to find x such that area will be 0.5.
    - ⇨ $F(x) = 0.5$
    - ⇨ $x = \pi_{0.5} = 1$

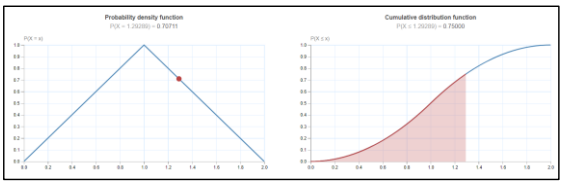


  - 1st Quartile or 25th percentile: $\pi_{0.25} = Q1$, separates 1st quarter (25%) of probability (area) from the rest. So, area must equal to 0.25. we can do this by using equation derived for $0 \le x \le 1$
    - ⇨ $\frac{x^2}{2} = 0.25$
    - ⇨ $x^2 = 0.5$
    - ⇨ $x = \pi_{0.25} = \sqrt{0.5} = 0.7071$



  - 3rd Quartile or 75th percentile: $\pi_{0.75} = Q3$, separates 3rd quarter (75%) of probability (area) from the rest. So, area must equal to 0.75. we can do this by using equation derived for $1 \le x \le 2$
    - ⇨ $2x - \frac{x^2}{2} - 1 = 0.75$
    - ⇨ $2x - \frac{x^2}{2} = 1.75$
    - ⇨ $\frac{4x - x^2}{2} = 1.75$
    - ⇨ $4x - x^2 = 3.5$

- ⇨ $-x^2 + 4x - 3.5 = 0$
- ⇨ multiplying by-1,
- ⇨ $x^2 - 4x + 3.5 = 0$
- ⇨ By quadratic equation formula, $a = 1, b = -4, c = 3.5$
- ⇨ $x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} = \frac{-4 \pm \sqrt{4^2 - (4*1*3.5)}}{2*1}$
- ⇨ $x = \frac{-4 \pm \sqrt{2}}{2}$
- ⇨ $x = 2.707, 1.29829$
- ⇨ But it can't be 2.7 as we have upper bound of 2
- ⇨ $x = \pi_{0.75} = 1.2983$

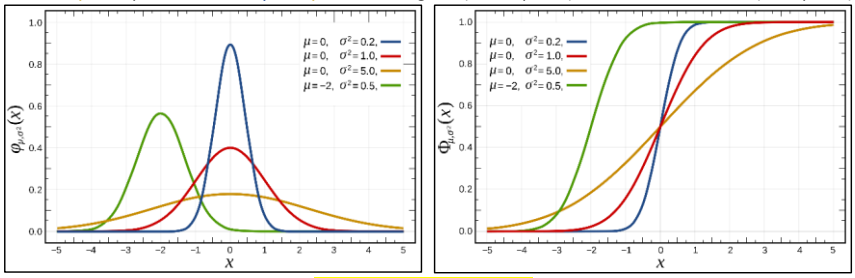


- **Expected Value (Mean) and Variance of Continuous Random Variables**
  - If X is a continuous random variable with pdf $f(x)$, then the expected value (or mean) of X is given by
    - ⇨ $\pi = \pi_X = E[X] = \int_{-\infty}^{\infty} x.f(x).dx$
  - For the variance of a continuous random variable, we can use the alternative formula,
    - ⇨ $Var(X) = E[X^2] - \pi^2 = \left(\int_{-\infty}^{\infty} x^2.f(x).dx\right) - \pi^2$
  - We will continue with previous elevator example, the mean (expected value) will be,
    - ⇨ $\pi = \pi_X = E[X] = \int_{-\infty}^{\infty} x.f(x).dx = \int_0^1 x.x.dx + \int_1^2 x.(2-x).dx$
    - ⇨ $\int_0^1 x^2.dx + \int_1^2 (2x - x^2).dx$
    - ⇨ $\frac{x^3}{3}|_0^1 + \left(x^2 - \frac{x^3}{3}\right)|_1^2$
    - ⇨ $\frac{1}{3} + \frac{2}{3} = 1$
  - To calculate variance, we will find $E[X^2]$
    - ⇨ $E[X^2] = \int_{-\infty}^{\infty} x^2.f(x).dx = \int_0^1 x^2.x.dx + \int_1^2 x^2.(2-x).dx$
    - ⇨ $\int_0^1 x^3.dx + \int_1^2 (2x^2 - x^3).dx$
    - ⇨ $\frac{x^4}{4}|_0^1 + \left(\frac{2x^3}{3} - \frac{x^4}{4}\right)|_1^2$
    - ⇨ $\frac{1}{4} + \frac{11}{12} = \frac{7}{6}$
    - ⇨ $Var(X) = E[X^2] - \pi^2 = \frac{7}{6} - 1 = \frac{1}{6}$
    - ⇨ $SD = \sqrt{Var(X)} = \sqrt{\frac{1}{6}} = 0.408$

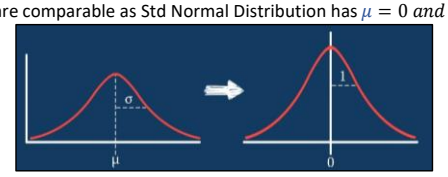
- **Normal Distributions**
  - A random variable X has a normal distribution, with parameters mean μ and standard deviation σ ,written as $X \sim normal(\mu, \sigma)$, Its pdf is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \text{for } x \in \mathbb{R}, \mu \in \mathbb{R}, \sigma > 0$$

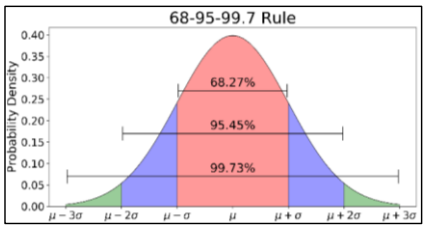
  - Mean μ gives the center of the normal pdf,
  - Standard deviation σ determines how spread out the graph is from mean
  - In following plots, the left side plot is PDF & right one is CDF. The red curve is the standard normal distribution.
  - Blue and yellow plots have same μ but yellow has larger σ (more spread) and blue has smaller σ(less spread)



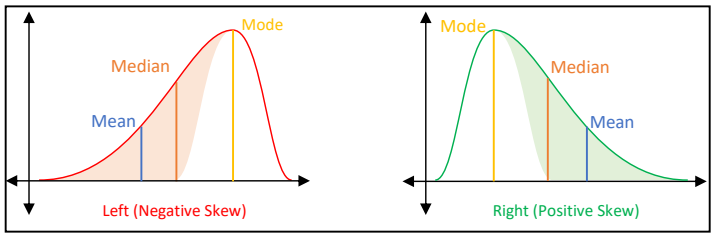
  - The normal distribution is probably the most important distribution. The population characteristics such as weight, height, and IQ distributions are modelled using normal distribution.
  - If $X \sim normal(\mu, \sigma)$
    - ⇨ $aX + b$ also follows a $normal(a\mu + b, a\sigma)$. Thus, linear transformation of a normally distributed random variable is also normally distributed.
    - ⇨ $\frac{X-\mu}{\sigma}$ follows the standard normal distribution, i.e. $\mu = 0\ and\ \sigma = 1$. We can correlate this with first property such that $a = \frac{1}{\sigma}$ and $b = -\frac{1}{\sigma}$
  - This transformation $\frac{X-\mu}{\sigma}$, is standardizing X, as resulting random variable will always have the standard normal distribution with $\mu = 0\ and\ \sigma = 1$.
  - Two normal distributions with different $\mu\ and\ \sigma$ are not comparable as they both have different scale. After standardization they are comparable as Std Normal Distribution has $\mu = 0\ and\ \sigma = 1$.



  - If Random Variable follows normal distribution, 68.2 % values lie between -1 σ and 1 σ, 95.4 % values lie between -2σ and 2σ, 99.7 % values lie between -3σ and 3σ.



- **Shape Properties of Distribution (Symmetry, Skew, Kurtosis)**



  - **Symmetric Distribution-** If both sides of PDF are mirror image, dataset is symmetrically distributed. In this case, value of PDF at particular height on left would be equal to right. The mean, median and mode are at centre.

$$f(\mu - \delta) = f(\mu + \delta)\ for\ \delta \in \mathbb{R}$$

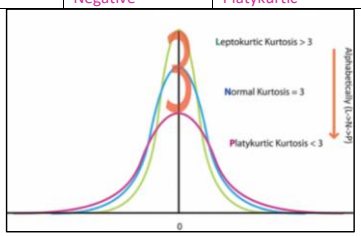
  - **Skewed Distribution-** If one tail is longer than another, the distribution is skewed.
  - Skewness is a measurement of the symmetry of a distribution. it describes how much a distribution differs from a normal distribution, either to the left or to the right.
  - For a perfect normal distribution, skewness = 0 as mean = median
  - A positive skew if the data is piled up to the left, which leaves the tail pointing to the right.
  - A negative skew if the data is piled up to the right, which leaves the tail pointing to the left.
  - This is the third standardized moment

$$S_k = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^3}{s^3}$$

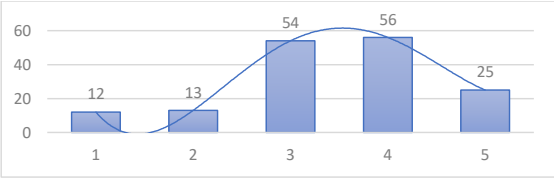
Distribution Type	Values range
Symmetric	(-0.5,0.5)
Moderated Skewed data	(-1, -0.5) and (0.5,1)
Highly Skewed data	(<-1), (>1)

  - What are some of the real-life applications of skewness? - Quora
  - **Kurtosis-** Kurtosis measures if dataset is heavy-tailed or light-tailed compared to a normal distribution.
  - High kurtosis implies heavy tails and more outliers
  - Low kurtosis implies light tails and fewer outliers.
  - Sample kurtosis S_{kr} is always measured relative to the kurtosis of a normal distribution, which is 3.
  - we have to calculate **excess** kurtosis E_{kr}.
    - ⇨ $E_{kr} = S_{kr} - 3$
    - ⇨ $S_{kr} = \frac{1}{n}\frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{s^4}$

Kurtosis	Excess Kurtosis	Distribution Type
>3	Positive	Leptokurtic
3	Zero	Mesocratic (Normal)
<3	Negative	Platykurtic

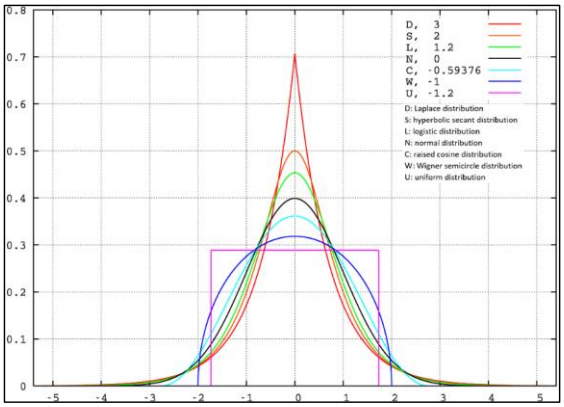


  - Example- sample $\{12 \quad 13 \quad 54 \quad 56 \quad 25\}$

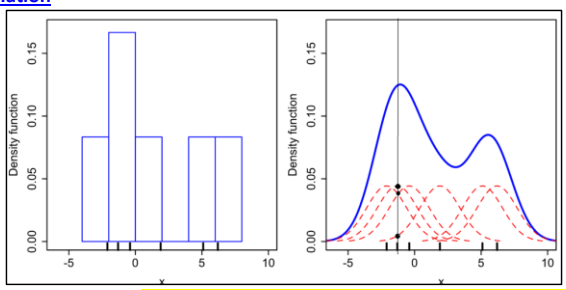


- ⇨ $\bar{X} = \frac{12+13+54+56+25}{5} = 32,$
- ⇨ $S^2 = \frac{(12-32)^2+(13-32)^2+(54-32)^2+(56-32)^2+(25-32)^2}{4} = 467.5$
- ⇨ $S = \sqrt{467.5} = 21.62$
- ⇨ $S_k = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^3}{s^3}$
- ⇨ $S_k = \frac{1}{5} * \frac{(12-32)^3+(13-32)^3+(54-32)^3+(56-32)^3+(25-32)^3}{21.63^3} = 0.1835$
- ⇨ $S_{kr} = \frac{1}{5} * \frac{(12-32)^4+(13-32)^4+(54-32)^4+(56-32)^4+(25-32)^4}{21.63^4} = 0.7861$
- ⇨ $E_{kr} = 0.7861 - 3 = -2.2139 = platykurtic$

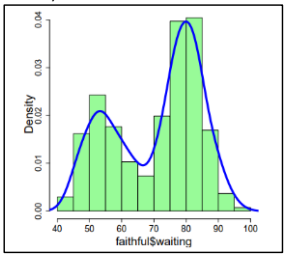
- Kurtosis and Skewness Example Question | CFA Level I - AnalystPrep
- Following are some distributions w.r.t. excess kurtosis values.



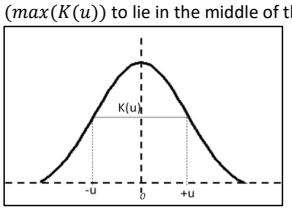
- **Kernel density estimation**



- Kernel density estimation (KDE) is a non-parametric (do not rely on any distribution) way to estimate the probability density function of a random variable.
- Here each data point is replaced with a kernel—a weighing function to estimate the pdf. The function spreads the influence of any point around a narrow region surrounding the point.
- The resulting probability density function is a summation of every kernel.
- it's like the histogram as tracing the outline of a histogram gives you a rough estimate because the area under a histogram represents 100% of the distribution.
- In histograms, number of divisions (or "bins") are considered and width for kernel densities (Stata).



- Kernel is simply a function which satisfies following three properties.
    a. Symmetry such that $K(u) = K(-u)$. The symmetric property of kernel function enables the maximum value of the function $(max(K(u))$ to lie in the middle of the curve.



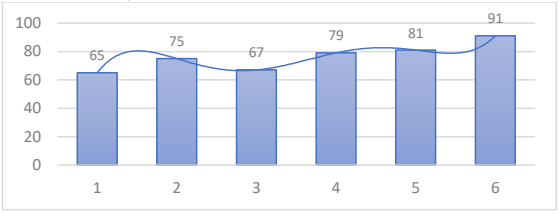
    b. The AUC of the function =1. Gaussian density function is used as a kernel function because the area under Gaussian density curve is one and it is symmetrical too.

$$\int_{-\infty}^{\infty} K(u).du = 1$$

    c. The value of kernel function(density), cannot be negative

$$K(u) \geq 0 \; for \; all \; -\infty < u < \infty.$$

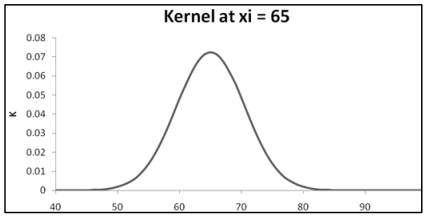
- We will use Gaussian kernel function to estimate kernel density and to optimize bandwidth
    ⇨ $K(x) = \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}$
    $x_i =$ observed data point
    ⇨ $x =$ value where kernel function is computed
    $h =$ bandwidth
- For each observation in dataset, we plot Gaussian kernels. At particular point we take sum of heights of Gaussian kernels to plot pdf.
- Example- $x_i = \{65, 75, 67, 79, 81, 91\}$



- Three inputs are required to develop a kernel curve around a data point. They are:
    1. The observation data point which is x_i
    2. The value of h
    3. A linearly spaced series of data points which ranges the observed data points $x_j = \{50,51,52 \cdots, 99\}$
- Calculation of K values for all values of x_j for a given values of x_i and h is at $x_i = 65 \; and \; h = 5.5$.

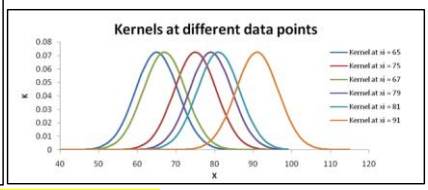
x_j	x_i	h	$A = \frac{1}{h\sqrt{2\pi}}$	$B = -0.5\left(\frac{x_j - x_i}{h}\right)^2$	$K = Ae^B$
50	65	5.5	0.072536	-3.71901	0.00175958
51	65	5.5	0.072536	-3.23967	0.002841733
52	65	5.5	0.072536	-2.79339	0.00444018
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-
99	65	5.5	0.072536	-19.1074	0.000000000365
Sum					1.000

- x_j and K are plotted below to visualize the kernel.



- Similarly, at all six data points, kernel values are estimated and plotted. It is observed that
$K(x_j) \cong 0$ for $x_j \gg x_i$. E.g., $K(x_j = 99) = 0$ when $x_i = 65$

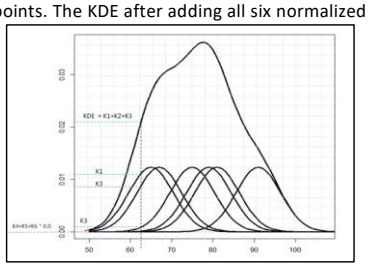
x	K(xj) Xi = 65	Xi = 75	Xi=67	Xi=79	Xi=81	Xi=91
50	0.00175958	0.0000237	0.00061093	0.0000007	0.00000001	0.00000000
51	0.00284173	0.00000532	0.00105409	0.00000017	0.00000003	0.00000000
52	0.00444018	0.00001157	0.00175958	0.00000042	0.00000007	0.00000000
53	0.00671214	0.00002433	0.00284173	0.00000102	0.00000017	0.00000000
...	...	...	...	...	...	...
78	0.00444	0.06251	0.009817	0.071347	0.06251	0.00444
79	0.002842	0.05568	0.006712	0.072536	0.067895	0.006712
80	0.00176	0.047984	0.00444	0.071347	0.071347	0.009817
81	0.001054	0.040007	0.002842	0.067895	0.072536	0.01389
...	...	...	...	...	...	...
99	0.00000	0.0000000	0.00000	0.000000	0.000342567	0.025184586



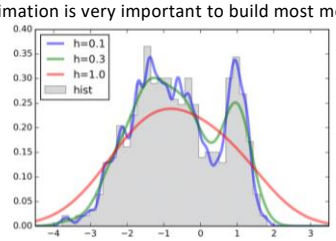
- Now, composite density values are calculated by adding the kernel values (K) from all x_j.
- The normalized sum is calculated by dividing by $n = 6$ to make the area under KDE curve =1. Therefore, the equation becomes

$$KDE_j = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-x_i}{h}\right)^2}$$

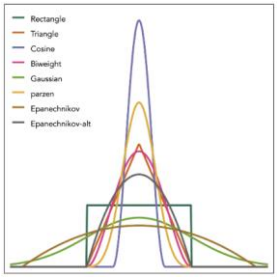
- Where n is the number of data points. The KDE after adding all six normalized kernels is shown be



- Bandwidth (h) plays an important role to fit the data appropriately.
- Low value produces high variance(overfitting) whereas high value of h produces large bias (underfitting).
- Therefore, optimal value of h estimation is very important to build most meaningful and accurate density.



- The blue one gives lot of variation in the density values which doesn't look realistic
- The red one fails to explain the actual density by hiding information.
- Here is a list of the kernels that we can fit for kernel density estimation:

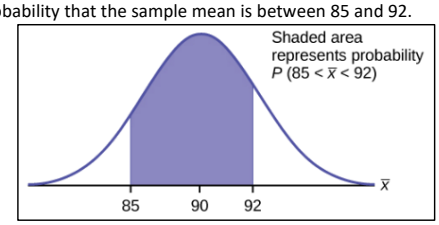


- Kernel Density Estimation - Wolfram Demonstrations Project

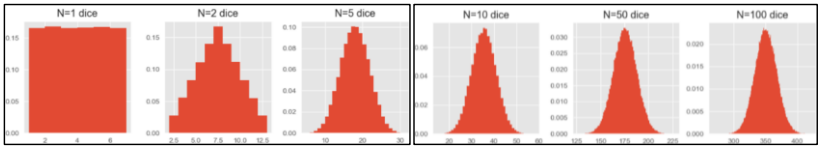
- **Sampling Distribution & Central Limit Theorem**
  - If X is a random variable with any distribution with parameters μ, σ_x
  - If random samples are drawn of size n (>30) recommended), as n increases, the random variable $\bar{X}$ which consists of sample means $\mu_{\bar{X}}$, tends to be normally distributed

$$\bar{X} \sim N\left(\mu, \frac{\sigma_x}{\sqrt{n}}\right)$$

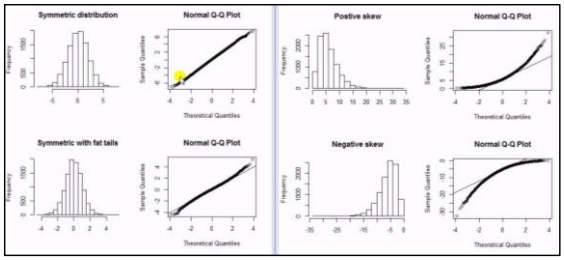
  - If you keep drawing larger and larger samples (such as rolling one, two, five, and finally, ten dice) and calculating their means, $\mu_{\bar{X}}$ form the sampling distribution with $\mu_{\bar{X}}$ and a variance $\frac{original\ var}{n}$.
  - Example: An unknown distribution has a $\mu = 90$ and a $\sigma_x = 15$, $n = 25$ are drawn randomly from population.
    - ⇨ $\bar{X} = N(90, \frac{15}{\sqrt{25}})$
    - ⇨ Find the probability that the sample mean is between 85 and 92.



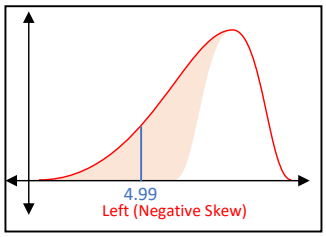
    - ⇨ Z-score when mean = 85 $z = \frac{x-\mu}{\sigma_x} = \frac{85-90}{3} = -1.667$ i.e., 1.667SD below mean
    - ⇨ Z-score when mean = 85 $z = \frac{x-\mu}{\sigma_x} = \frac{92-90}{3} = 0.667$ i.e., 0.66667SD above mean
    - ⇨ $P(-1.667 < Z < 0.667) = 0.69986$
  - Z-score Calculator
  - Central Limit Theorem - a demonstration - GaussianWaves
  - The results of a simulated 1,00,000 tosses of fair dice



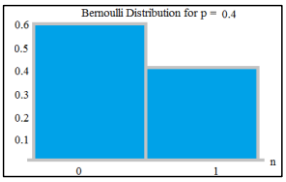
  - Sample size depends on
    - If population size is big, we can have big sample size
    - If sample collection process is expensive, we may have to reduce the sample size
    - If margin of error is small, we take samples of larger size and number of such samples are also increased. Margin of error is allowed deviation from original numbers
- **Q-Q Plot (Actual Quantiles Vs Theoretical Quantiles)**
  - Q-Q plots are used to find the type of distribution for a random variable.
    - ⇨ Suppose, we have random variable $X = x_1, x_2, x_3, \cdots, x_{1000}$
    - ⇨ Sort X in ascending order and calculate quantiles (Actual)
    - ⇨ 1000 observations from $Y \sim N\ (0, 1)$
    - ⇨ Sort Y in ascending order and calculate quantiles (Theoretical)
    - ⇨ The plot between theoretical & actual quantiles is called QQ Plot.
  - If all points lie in a straight line it signifies that sample follows a Gaussian distribution
  - The points which are away from main line shows that data is skewed.



- **Chebyshev's inequality**
  - The empirical rule (68-95-99) applies to only normal distribution.
  - Chebyshev's theorem is used to find the proportion of observations we would expect to find within k standard deviations from the mean.
  - for any population or sample, the proportion of observations is greater than $\boxed{1 - \frac{1}{k^2}}$ where k is no. of standard deviation interval i.e., 1, 2, 3
    - Example- left skewed distribution with $\mu = 4.99$ & $\sigma = 3.13$



    - $when\ k = 2$
      - ⇨ $1 - \frac{1}{2^2} = 0.75$
      - ⇨ 75% of the observations fall between $-2\sigma(-1.27)$ and $+2\sigma(11.25)$ from mean
    - $when\ k = 3$
      - ⇨ $1 - \frac{1}{3^2} = 0.9$
      - ⇨ 90% of the observations fall between $-3\sigma(-4.4)$ and $+3\sigma(14.38)$ from mean
- **Bernoulli Distribution**:
  - Bernoulli event: An event has only two possible outcomes (Success or Failure) for which the probability of occurrence is p and the probability of the event not occurring is $1 - p$
  - Bernoulli trial: It's an experiment where there are two possible outcomes (Success and Failure).
  - The two possible outcomes are labelled by $n = 1(Success)$, $n = 0(Failure)$, $P_{success} = p$ and $P_{failure} = q = 1 - p$
  - PMF $P(X = x) = \begin{cases} q\ for\ n = 0 \\ p\ for\ n = 1 \end{cases}$
  - E.g., If I toss an unbiased coin,
    - $p = P_{Heads} = 0.5$
    - $q = P_{Tails} = 1 - 0.5$

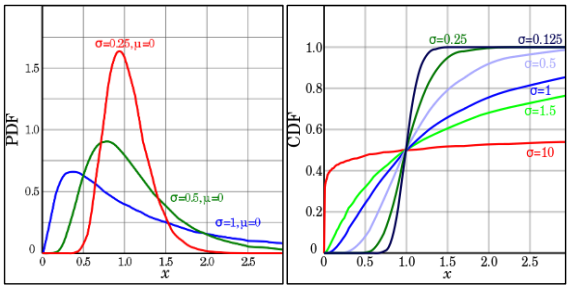


- **Binomial Distribution**
  - The binomial distribution gives PMF $P_{success}(n|N)$ of obtaining n successes out of N Bernoulli trials (only 2 possible outcomes).

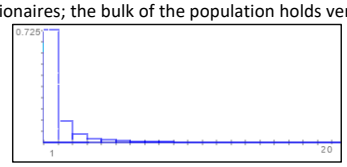
$$P_{success}(n|N) = N_{c_n} * p^n * q^{N-n}$$
$$= \frac{N!}{n!\,(N-n)!} * p^n * q^{N-n}$$

  - Example- A coin is tossed 10 times. What is the probability of getting exactly 6 heads?
    - ⇨ $N = 10, n = 6$
    - ⇨ $P_{success}(n|N) = \frac{N!}{n!(N-n)!} * p^n * q^{N-n}$
    - ⇨ $\frac{10!}{6!(10-6)!} * 0.5^6 * 0.5^4$
    - ⇨ 210 * 0.015625 * 0.0625
    - ⇨ 0.205078125

- **Log-Normal Distribution**

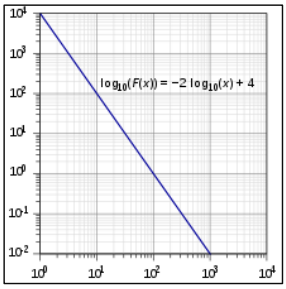


  - A random variable X is lognormally distributed if $Y = \log_e X$ is normally distributed
  - PDF- $f(x) = \frac{1}{\sigma x \sqrt{2\pi}} e^{\frac{-(\log_e x - \mu)^2}{2\sigma^2}}, x \geq 0$
  - CDF – $P(X \geq c) = 1 - (X < c)$
  - $P(X \geq c) = P\left(z \leq \left(\frac{\log_e c - \mu}{\sigma}\right)\right)$
  - Example- Suppose the lifetime of a motor has a lognormal distribution. What is the probability that the lifetime exceeds 12,000 hours if the mean and variance of the normal random variable are 11 hours and 1.3 hours?
    - ⇨ $\mu = 11, \sigma = 1.3, c = 12000$
    - ⇨ $P(X \geq 12000) = 1 - P(X < 12000)$
    - ⇨ $1 - P\left(z \leq \left(\frac{\log_e 12000 - 11}{1.3}\right)\right)$
    - ⇨ $1 - P\left(z \leq \left(\frac{9.3926 - 11}{1.3}\right)\right)$
    - ⇨ $1 - 0.109349$
    - ⇨ 0.890651
  - Quick P Value from Z Score Calculator (socscistatistics.com)
  - Skewed distributions with low mean values, large variance, and all-positive values often fit this type of distribution. Values must be positive as log(x) exists only for positive values of x.
  - The following phenomenon can all be modelled with a lognormal distribution:
    - Milk production by cows.
    - Amounts of rainfall.
    - Size distributions of rainfall droplets.
    - The volume of gas in a petroleum reserve.
- **Power Law Distribution**
  - The power law or scaling law states that a relative change in one quantity results in a proportional relative change in another. E.g., a square; $b = 2, Area = 4 \ when \ b \to 4, Area \to 16$
  - A power law distribution has the form $Y = k X^{\alpha}$, where:
    - X and Y are variables of interest,
    - α is the law's exponent,
    - k is a constant.
    - An inverse relationship like $Y = X^{-1}$ is also a power law, as change in one quantity results in a negative change in another.
  - The power law can be used to describe a phenomenon where a small number of items is clustered at the top of a distribution (or at the bottom), taking up 95% of the resources. In other words, it implies a small number of occurrences are common, while larger occurrences are rare. For example, where the distribution of income is concerned, there are very few billionaires; the bulk of the population holds very modest nest egg



  - Other Examples:
    - Distribution of income, i.e. less people would be paid heavily majority has less salaries
    - Magnitude of earthquakes,
    - Size of cities according to population,

- If we plot two quantities against each other with logarithmic axes and they show a linear relationship, this indicates that the two quantities have a power law distribution.



- **Box Cox Transformation**
  - A Box-Cox transformation is a transformation of non-normal variables into a normal shape.
  - As Many statistical tests and intervals are based on the assumption of normality.
  - Lambda(λ) is an appropriate exponent used to transform data into a "normal shape." The Lambda value indicates the power to which all data should be raised.
  - optimal λ is searched between -5 to 5
  - The Box-Cox transformation of the variable x is also indexed by λ, and is defined as

$$x' = \frac{x^\lambda - 1}{\lambda} \quad if\ \lambda \neq 0$$
$$x' = \log(x) \quad if\ \lambda = 0$$

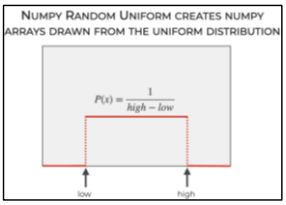
  - This test only works for positive data. The second formula can be used for negative y-values.

$$\mathrm{Y}(\lambda) = \begin{cases} \dfrac{(y + \lambda_2)^{\lambda_1} - 1}{\lambda} & \text{if } \lambda_1 \neq 0; \\ \log(y + \lambda_2) & \text{if } \lambda_1 = 0; \end{cases}$$

  - If $\lambda_2 = 0$ this is the usual boxcox transform, and the search function will estimate the two parameters λ_1, λ_2 by maximum likelihood.

Lambda value (λ)	Transformed data (Y')
-3	$Y^{-3} = 1/Y^3$
-2	$Y^{-2} = 1/Y^2$
-1	$Y^{-1} = 1/Y^1$
-0.5	$Y^{-0.5} = 1/(\sqrt{Y})$
0	log(Y)**
0.5	$Y^{0.5} = \sqrt{Y}$
1	$Y^1 = Y$
2	Y^2
3	Y^3

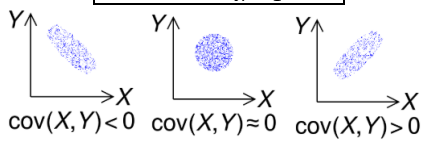
  - What are the limitations of using the Box-Cox Transformation?
  - If interpretation is your goal, then the Box-Cox transformation may be a poor choice. If lambda is some non-zero number, then the transformed target variable may be more difficult to interpret than if we simply applied a log transform.
  - A second issue is that the Box-Cox transformation usually gives the median of the forecast distribution when we revert the transformed data to its original scale. Occasionally, we want the mean (not the median) and there are ways we can do this, which I may discuss in a later article.
  - r - Box Cox Transformation with swift - Cross Validated (stackexchange.com)\
- **Applications of non-Gaussian distributions?**
  - Uniform distribution for generating random numbers.



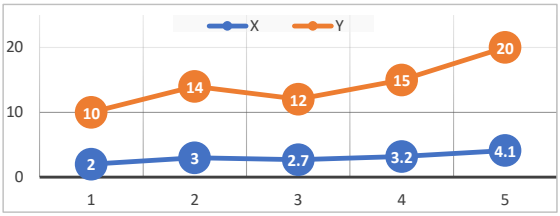
- A known distribution gives a theoretical model for the behaviour of a random variable
- Weibull distributions:
  - The upstream rainfall determines the height of a dam which stands for 100s of years without repairs; Probability of rainfall > a value is required;
  - This distribution is applied to extreme events such as annual maximum one day rainfalls and river discharges.
- probability distribution_362.pdf (csun.edu)
- Probability Distribution | Types of Distributions (analyticsvidhya.com)

- **Covariance**
  - Covariance signifies the direction of the linear relationship between the two variables.
  - Direction signifies variables are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative impact on the value of the other variable).
  - Covariance is a measure of the relationship between two random variables. It evaluates to what extent the variables change together.
  - It can range between $-\infty \; and \; \infty$. covariance only measures how two variables change together, not the dependency of one variable on another one.

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



  - The upper and lower limits for the covariance depend on the variances of x, y. These variances can vary with the scaling of the variables.
  - Change in the units of measurement can change the covariance. Thus, covariance is not useful to find the magnitude of the relationship between two variables.
  - Example: a company has a five-quarter dataset that shows quarterly gross domestic product (GDP) growth in percentages (x) and a company's new product line growth in percentages (y).

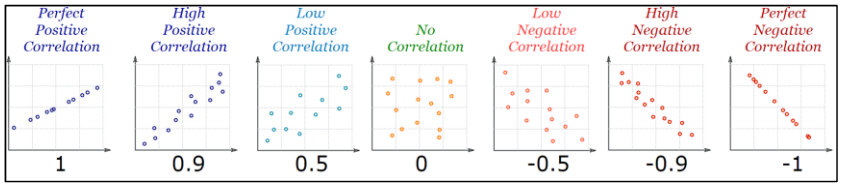


Q	X	Y	(X-X_Mean)*(Y-Y_Mean)
Q1	2	10	4.2
Q2	3	14	0
Q3	3	12	0.66
Q4	3	15	0.16
Q5	4	20	6.38
AVG	3	14.2	2.85

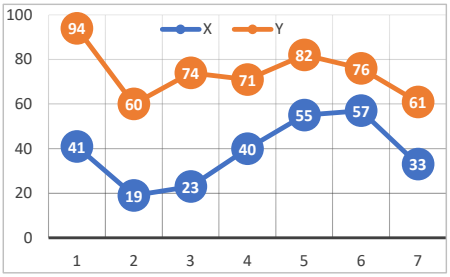
- **Correlation**
  - Correlation shows the direction and strength of the relationship. we can say the correlation values have standardized, whereas the covariance values are not standardized.
  - It can range between -1 to +1. The following formula is Pearson sample coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

  - If there is no linear relationship between x, y then $r_{xy} \approx 0$. However, there could exist other functional relationships between the variables.
    - If r_{xy} is positive, an increase in one variable also increases the other.
    - If r_{xy} is negative, the changes in the two variables are in opposite directions.
  - Covariance determines the type of interaction between two variables, while correlation determines the direction as well as the strength of the relationship between two variables. However, an important limitation is that both these concepts measure the only linear relationship.

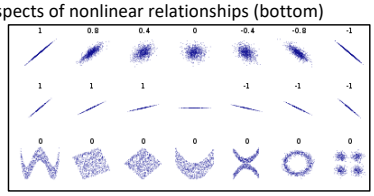


- Example: calculate the correlation for the following two data sets X: $(41, 19, 23, 40, 55, 57, 33)$
  Y: $(94, 60, 74, 71, 82, 76, 61)$

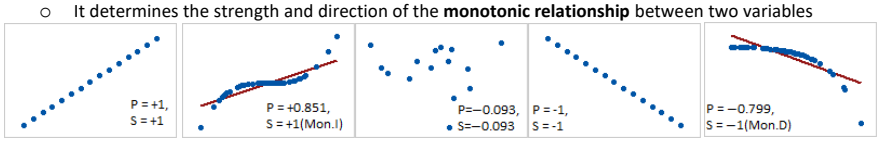


X	Y	a=x-x̄	b=y-ȳ	a*b	a^2	b^2
41	94	2.71	20	54.29	7.37	400
19	60	-19.29	-14	270.00	371.94	196
23	74	-15.29	0	0.00	233.65	0
40	71	1.71	-3	-5.14	2.94	9
55	82	16.71	8	133.71	279.37	64
57	76	18.71	2	37.43	350.22	4
33	61	-5.29	-13	68.71	27.94	169
38.29	74	corr	0.54	559.00	1273.43	842
AVG				SUM		

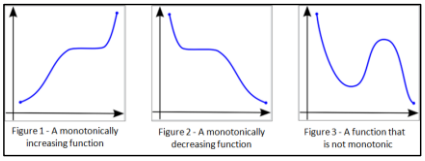
- Correlation reflects
  - Strength and direction of a linear relationship (top row),
  - Not the slope of that relationship (middle),
  - Nor many aspects of nonlinear relationships (bottom)



- **Spearman Rank Correlation Coefficient**
  - It determines the strength and direction of the **monotonic relationship** between two variables



  - A monotonic relationship one of the follows
    - Monotonically increasing - as the x variable increases the y variable never decreases
    - Monotonically decreasing - as the x variable increases the y variable never increases
    - Not monotonic - as the x variable increases the y variable sometimes decreases and sometimes increases.



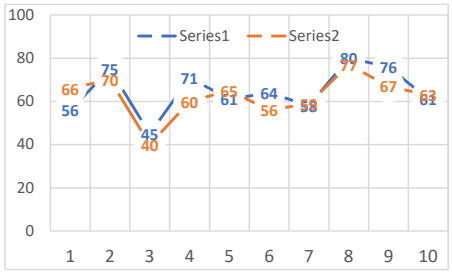
  - The coefficient can range between -1 to +1.

$$\rho = 1 - \frac{6 * \sum d_i^2}{n(n^2 - 1)}, d_i \text{ is difference between ranks}$$

  - The sign indicates the direction of association between X and Y.
    - If Y tends to increase when X increases, the Spearman correlation coefficient is positive.
    - If Y tends to decrease when X increases, the Spearman correlation coefficient is negative.
    - A Spearman correlation of 0 indicates that there is no tendency for Y to either increase or decrease when X increases.

- The Spearman correlation increases in magnitude as X and Y become closer. When X and Y are perfectly monotonically related, the Spearman correlation coefficient becomes 1.
- For perfectly monotone increasing relationship, X_i, Y_i and X_j, Y_j, $X_i - X_j$ and $Y_i - Y_j$will always have the same sign. For perfectly monotone decreasing relationship, these differences always have opposite signs.
- Example-Find correlation between English & maths marks for 10 students

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Math	66	70	40	60	65	56	59	77	67	63

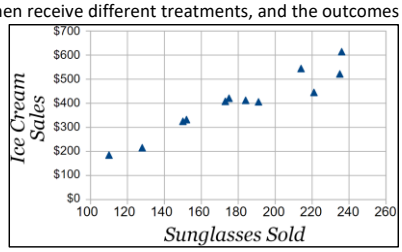


English	Math	Rank Eng	Rank Mat	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	-3	9
61	65	7	5	2	4
64	56	5	9	-4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	-1	1
62	63	6	6	0	0
		0.66			56

$$\rho = 1 - \frac{6 * \sum d_i^2}{n(n^2-1)}$$
$$= 1 - \frac{6 * 56}{10(10^2-1)} = 0.66$$

- Both of these coefficients cannot capture any other kind of non-linear relationships. Thus, if a scatterplot indicates a relationship that cannot be expressed by a linear or monotonic function, then both of these coefficients must not be used to determine the strength of the relationship between the variables.

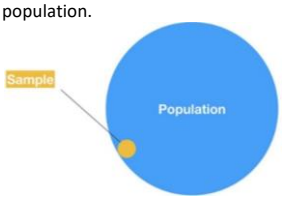
- **Correlation vs. Causation**
  - Correlation does not tell us why and how behind the relationship but it just says the relationship exists. Correlation does not imply causation, but it provides a hint.
  - Causation between random variables A and B implies that A and B have a cause-and-effect relationship with one another. A causes B or vice versa.
  - We cannot assume causation if we see two events happening together, before our eyes. Our observations are purely anecdotal
  - There are so many other possibilities for an association,
    - B actually causes A.
    - A and B are correlated, but they're actually caused by C.
    - A cause B—as long as D happens.
    - A causes E, which leads E to cause B (but you only saw that A Causes B from your own eyes).
  - Causation is assessed using a controlled study,
    - The sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.



  - From above plot, we can conclude that sales of ice cream cones and sunglasses are positively correlated.
  - We can't conclude that selling more ice cream cones causes more sunglasses to be sold. It is likely that the correlation is caused by a third factor, an increase in temperature!

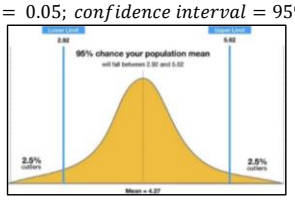
Commented [A1]: possibly not true or accurate

- **Confidence Interval**
  - When you make an estimate in statistics, there is always uncertainty around that estimate because the number is based on a sample of the population.



  - A confidence interval is a way to measure how well sample represents the population.
  - The confidence interval is the range of values that you expect your population estimate(θ) to fall between a confidence level of the time if you run your experiment again.
  - The confidence level is the percentage of times you expect to reproduce an estimate between the upper and lower bounds of the confidence interval, and is set by the a value.

$$\alpha = 0.05;\ confidence\ interval = 95\%$$



  - You can be confident that 95% confidence contains the true mean of the population will be between (2.92,5.62). Accordingly, there is a 5% chance that the population mean lies outside of (2.92,5.62).
  - Due to natural sampling variability, the sample mean (centre of the CI) will vary from sample to sample.
  - As the sample size increases, the range of interval values will narrow, this will be more accurate compared with a smaller sample.

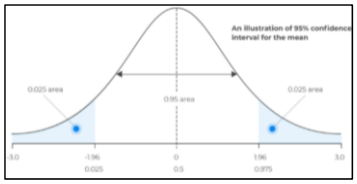
$$C.I = point\ estimate \pm reliability\ factor * standard\ error$$

Point estimate is sample statistic e.g., mean
The reliability factor depends on the sampling distribution and $(1 - \alpha)$
SE = $\frac{\sigma}{\sqrt{n}}$

  - **Calculating C.I. Normal distribution with a known variance ($\boldsymbol{\theta = \mu}$):**

$$C.I. = \bar{x} \pm z\frac{a}{2} * \frac{\sigma}{\sqrt{n}}$$

  - The reliability factor is $z\frac{a}{2}$. It is the z-score leaves a probability of $\frac{a}{2}$ on the upper tail (right-hand tail) of the standard normal distribution.



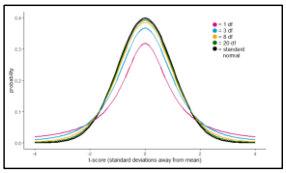
Confidence Level $(1 - a)$	a	$z\frac{a}{2}$
90%	10%	1.645
95%	5%	1.960
99%	1%	2.575

  - **Example:** boiling temperature(℃) of a liquid samples are 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2. population standard deviation is 1.2 degrees, what is the confidence interval for the population mean at a 95% confidence level?
    - ⇨ If the measurements follow a normal distribution, then the sample mean will have the distribution $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
    - ⇨ $C.I. = \bar{x} \pm 1.96 * \frac{\sigma}{\sqrt{n}}$ as both mean & SD is unknown
    - ⇨ $\bar{x} = \frac{102.5+101.7+103.1+100.9+100.5+102.2}{6} = 101.82$
    - ⇨ $C.I. = 101.82 \pm 1.96 * \frac{1.2}{\sqrt{6}}$
    - ⇨ $C.I. = (100.86, 102.78)$

- **Normal distribution with unknown variance:**
- When the variance is unknown, the z-score is replaced with the t-score. Similarly, unknown σ with S, the of the sample mean.

$$C.I. = \bar{x} \pm t_{\frac{a}{2}} * \frac{S}{\sqrt{n}}$$

- $t_{\frac{a}{2}}$ is t-score leaves a probability of $\frac{a}{2}$ on the upper tail (right-hand tail) of the t-distribution$(\mathrm{df}) = \mathrm{n} - 1$.



- **Example:** Following are heights(cm) recorded for 5 random 12-year students at school. The heights are normally distributed with unknown mean & SD. Calculate a two-tailed 95% confidence interval for the mean height.

{124,124,128,130,127}

⇨ $C.I. = \bar{x} \pm t_{\frac{a}{2}} * \frac{S}{\sqrt{n}}$ as both mean & SD is unknown

⇨ $\bar{x} = \frac{124+124+128+130+127}{5} = 126.6$

⇨ $S^2 = \frac{(124-126.6)^2+(124-126.6)^2+(128-126.6)^2+(130-126.6)^2+(127-126.6)^2}{5-1} = 6.8$

⇨ $S = 2.60768$

⇨ $\frac{a}{2} = 0.025$

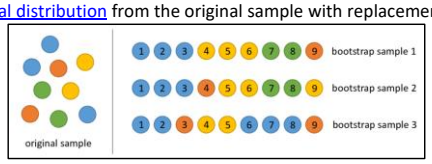
⇨ From T table, $df = n - 1 = 4$ & $one\ tail = 0.025$

⇨ $t_{\frac{a}{2}} = 2.776$

⇨ $C.I. = 126.6 \pm 2.776 * \frac{2.60768}{\sqrt{5}}$

⇨ $C.I. = (123.36, 129.83)$

- **When variance is unknown, and the sample size is large enough (any distribution):**
- By CLT, we can approximate any non-normal distribution as a normal one if sample size is large (n ≥ 30). Therefore, we can use the relevant z-score when constructing a confidence interval for the population mean.
- **Confidence interval using bootstrapping**
- Bootstrapping is a method to estimate confidence intervals which does not rely on any assumption of data distribution. bootstrap steps look as follows:
  - ⇨ Define u — statistic computed from the sample (mean, median, etc.)
  - ⇨ Sample F^*-Empirical distribution from the original sample with replacement.



    - ➢ The empirical distribution of data is simply the distribution that you see in the data.
    - ➢ Suppose we roll an 8-sided die 10 times and get the following data, written in increasing order: 1, 1, 2, 3, 3, 3, 3, 4, 7, 7.
    - ➢ Now we select a random value from above list. The full empirical distribution can be put in a probability table for each value is

$value$	1	2	3	4	7
$P(x)$	2/10	1/10	4/10	1/10	2/10

    - ➢ If true distribution is F and empirical distribution as F^*. If we have enough data, F^* should be a good approximation of F

$value$	1	2	3	4	5	6	7	8
True $P(x)$	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
Empirical $P(x)$	2/10	1/10	4/10	1/10	0	0	2/10	0

    - ➢ Because F^* is derived strictly from data we call it the empirical distribution or resampling distribution of the data. In particular, the mean of F^* is just the $\bar{x}$
  - ⇨ Repeat n times (n is bootstrap iterations).
  - ⇨ Compute u^* — the statistic calculated from each resample.
- This method approximates the difference between bootstrapped means and sample mean. There are some methods for bootstrapping.

- **Empirical bootstrapping (C.I-80%):**
  - ⇨ Start with resampling with replacement from original data n times.
  - ⇨ For each bootstrap calculate mean x^*
  - ⇨ Compute $\delta^* = x^* - \bar{x}$ for each bootstrap sample, sort them from smallest to biggest.
  - ⇨ Choose δ_1 as the 10th percentile, δ_9 as the 90th percentile of sorted list of δ^*, which gives an 80% confidence interval of $[\bar{x} - \delta_1, \bar{x} - \delta_9]$
- **Percentile bootstrap (C.I-80%):**
  - ⇨ This method uses the distribution of the bootstrap sample statistic as a direct approximation of the data sample statistic.
  - ⇨ Resample with replacement from original data n times.
  - ⇨ For each bootstrap calculate mean x^*, sort them from smallest to biggest.
  - ⇨ Choose ${x^*}_1$ as the 10th percentile, ${x^*}_9$ as the 90th percentile of sorted list of x^* ,which gives an 80% confidence interval of $[{x^*}_1, {x^*}_9]$
- Bootstrapping can't improve point estimate, the quality of bootstrapping depends on the quality of the collected data.
- If sample data is biased and doesn't represent population data well, the same will occur with bootstrap estimates. So, data collected during experimentation should be a good approximation of the whole population data.
- Bootstrapping Confidence Intervals: the basics | by Elizaveta Lebedeva | Towards Data Science
- **Example**. The sample data

30	36	37	41	42	42	43	43	43	46

Estimate the μ of the underlying distribution and give an 80% bootstrap confidence interval.

- By empirical method
  - ⇨ $\bar{x} = 40.3$
  - ⇨ Now let's return a sample data with 10 points. Each of the 20 rows in the following array is one bootstrap sample.

	1	2	3	4	5	6	7	8	9	10	$\bar{x}$	δ
1	43	43	42	37	42	36	43	41	46	42	41.5	1.2
2	36	41	43	42	36	36	37	42	42	43	39.8	-0.5
3	46	37	37	43	43	42	41	30	42	43	40.4	0.1
4	30	37	43	41	43	42	43	42	43	41	40.5	0.2
5	43	43	46	41	42	36	41	37	41	42	41.2	0.9
6	43	43	37	42	37	36	42	43	42	36	40.1	-0.2
7	43	46	36	36	42	43	43	43	30	43	40.5	0.2
8	37	36	41	42	42	41	46	42	37	30	39.4	-0.9
9	42	41	36	42	42	30	46	43	30	37	38.9	-1.4
10	42	43	43	43	46	42	36	43	42	43	42.3	2
11	43	43	41	42	30	37	43	46	43	42	41	0.7
12	37	42	36	43	43	43	42	43	42	43	41.4	1.1
13	36	41	37	41	36	41	43	30	43	41	38.9	-1.4
14	42	43	30	43	43	41	30	42	37	36	38.7	-1.6
15	43	46	46	36	43	43	41	30	37	37	40.2	-0.1
16	43	36	46	43	42	43	46	42	37	41	41.9	1.6
17	42	43	42	43	37	42	43	30	42	43	40.7	0.4
18	43	43	36	41	36	46	46	43	43	42	41.9	1.6
19	42	43	36	42	42	43	30	43	43	43	40.7	0.4
20	43	42	43	46	30	37	43	42	46	43	41.5	1.2

Bootstrap 80% confidence interval for μ is $[\bar{x} - \delta_1, \bar{x} - \delta_9]$

- ⇨ $\delta_1 = 90th\ percentile\ from\ delta = 1.6$
- ⇨ $\delta_9 = 10th\ percentile\ from\ delta = -1.4$

5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
-1.6	-1.4	-1.4	-0.9	-0.5	-0.2	-0.1	0.1	0.2	0.2	0.4	0.4	0.7	0.9	1.1	1.2	1.2	1.6	1.6	2

  - ⇨ $[40.3 - 1.6, 40.3 + 1.4]$
  - ⇨ $[38.7, 41.7]$
- By percentile method
  - ⇨ $C.I. = [{x^*}_1, {x^*}_9]$

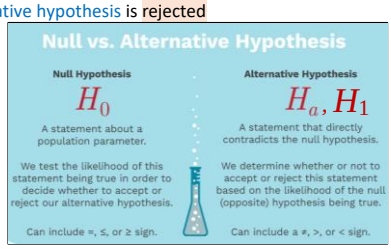
5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100
39	39	39	39	40	40	40	40	41	41	41	41	41	41	41	42	42	42	42	42

  - ⇨ $[39,42]$
- The bootstrap percentile method simple. However, it depends on the bootstrap distribution of x^* based on a particular sample being a good approximation to the true distribution of x.
- Although this direct equation of quantiles of the bootstrap sampling distribution with confidence limits may seem initially appealing, its rationale is somewhat obscure.
- class24-prep-a.pdf (mit.edu)

Commented [A2]: Points are single values, in comparison to interval estimates, which are a range of values.

- **Hypothesis Testing**
  - When we start asking questions about the data and interpret the results, statistical methods are used to provide a confidence or likelihood about the answers. These methods are called hypothesis testing, or significance tests.
  - It calculates some quantity under a given assumption. Test result interprets whether the assumption is right or has been violated.
  - examples that are used frequently in ML are:
    - A test that assumes that data has a normal distribution.
    - A test that assumes that two samples were drawn from the same underlying population distribution.
  - The null hypothesis reflects that there will be no observed effect in our experiment. This is statement that can be nullified or invalidate. If the null hypothesis is not rejected, no changes will be made.
    - If p-value is <= $significance\ level(\alpha)$ reject the null hypothesis.
    - If p-value is > $significance\ level(\alpha)$, fail to reject the null hypothesis. Failed to reject a null hypothesis does not mean that the alternative hypothesis is true.
  - The alternative hypothesis reflects that there will be an observed effect for our experiment. Accepting the alternative hypothesis will lead to changes in opinions or actions.
    - If the null hypothesis is rejected, alternative hypothesis is accepted
    - If the null hypothesis is accepted, alternative hypothesis is rejected

$H0$	$H1$
$X = Y$	$X \neq Y$
$X \geq Y$	$X < Y$
$X \leq Y$	$X > Y$



  - While accepting or rejecting a H0, two types of error may occur.
    - Type I Error(α)- a true null hypothesis is wrongly rejected.

$$\alpha = significance\ level = P\ (type\ I\ error) = P\ (rejecting\ H0\ when\ H0\ is\ true)$$

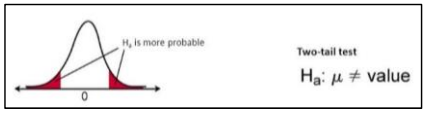
    - Type II Error(β)- a false null hypothesis is accepted i.e.; an alternative hypothesis is true.

$$\beta = P\ (type\ II\ error) = P\ (rejecting\ H1\ when\ H1\ is\ true) = P\ (accepting\ H0\ when\ H1\ is\ true)$$
$$1 - \beta = power\ of\ test = 1 - P\ (type\ II\ error) = P\ (rejecting\ H0\ when\ H1\ is\ true)$$

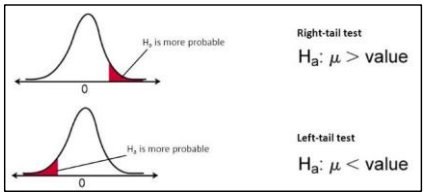
    - Correct decision: accepting a true null hypothesis
    - Correct decision: rejecting a false null hypothesis.

	Reject H_0	Fail to Reject H_0
H_0 Is True	Type I Error α (FP)	Correct 1 - α (TN)
H_a Is False	Correct 1 – β ("Statistic Power") (TP)	Type II Error β (FN)

  - In a two-sided test the null hypothesis is rejected if the test statistic is either too small or too large. Thus, the rejection region consists of one on the left and one on the right.



  - In a **two-tailed test**, critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values. If the sample being tested falls into either of the critical areas, the alternative hypothesis is accepted instead of the null hypothesis.



  - In a **one-tailed test**, critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both. If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

Commented [A3]: an idea that is suggested as the possible explanation for something but has not yet been found to be true or correct

- Test Statistic: measures how close the sample has come to the null hypothesis. It shows how closely your observed data match the distribution expected under the null hypothesis of that statistical test
- Different statistical tests will have slightly different ways of calculating these test statistics, but the underlying hypotheses and interpretations of the test statistic stay the same.

Test statistic	Null and alternative hypotheses	Statistical tests that use it
t-value	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	• *T*-test • Regression tests
z-value	**Null:** The means of two groups are equal **Alternative:** The means of two groups are not equal	• *Z*-test
F-value	**Null:** The variation among two or more groups is greater than or equal to the variation between the groups **Alternative:** The variation among two or more groups is smaller than the variation between the groups	• ANOVA • ANCOVA • MANOVA
χ^2-value	**Null:** Two samples are independent **Alternative:** Two samples are not independent (i.e. they are correlated)	• Chi-squared test • Non-parametric correlation tests

- **The p-value approach**
- The likelihood $(p\ value)$ of test statistic is compared to the $significance\ level\ (\alpha)$ of the hypothesis test.
- The $p-value$ corresponds to the probability of observing sample data at least as extreme as the actually obtained test statistic.
- Small $p\ value$ provide evidence against the null hypothesis. The smaller (closer to 0) the $p\ value$, the stronger is the evidence against the null hypothesis.

$$if\ p \leq \alpha, reject\ H0;$$
$$if\ p > \alpha\ accept\ H0.$$

p-value	Evidence against $H0$
$p > 0.1$	Weak or no evidence
$0.05 < p \leq 0.10$	Moderate
$0.01 < p \leq 0.05$	Strong
$p < 0.01$	Very Strong

- Example 1: A monthly income investment scheme exists that promises variable monthly returns. An investor will invest in it only if he is assured of an average $180 monthly income. The investor has a sample of 300 month's returns which has a mean of $190 and a standard deviation of $75. Should he invest in this scheme?
- Solution:

$H0$	$\mu = 180$
$H1$	$\mu > 180$

Commented [RS4]:

⇨ **Method 1:** Critical Value Approach.

⇨ Identify a critical value $\bar{x}_L$ for the sample mean, reject the null hypothesis if the sample mean >= critical value $\bar{x}_L$

⇨ $\alpha = significance\ level = P\ (type\ I\ error) = P\ (rejecting\ H0\ when\ H0\ is\ true)$

⇨ $P(Z > Z_a) = a$

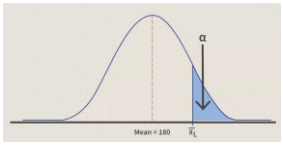
⇨ $z_a = \frac{\bar{x}_L - 180}{\frac{75}{\sqrt{300}}}$

⇨ $a = significance\ level = 5\%, Z_{0.05} = 1.645$

⇨ $1.645 = \frac{\bar{x}_L - 180}{\frac{75}{\sqrt{300}}}$

⇨ $\bar{x}_L = \left(1.645 * \frac{75}{\sqrt{300}}\right) + 180$

⇨ $\bar{x}_L = 187.12$



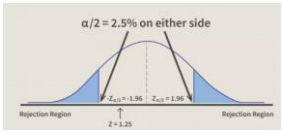
⇨ Since the $\bar{x}(190) > critical\ value\ (187.12)$, the null hypothesis is rejected, and the conclusion is that the average monthly return is indeed greater than $180, so the investor can consider investing in this scheme.

- ⇨ **Method 2**: Using Standardized Test Statistics
- ⇨ $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- ⇨ reject the null hypothesis if the $z >= z_a$
- ⇨ $z = \frac{190-180}{\frac{75}{\sqrt{300}}} = 2.3094$
- ⇨ $z_a = z_{0.95} = 1.614$
- ⇨ Since Z= 2.309 > 1.645, the null hypothesis can be rejected.
- ⇨ **Method 3**: P-value Calculation
- ⇨ We aim to identify $P\ (\bar{x} >= 190, when\ \mu = 180)$.
- ⇨ $P\left(Z >= \frac{190-180}{\frac{75}{\sqrt{300}}}\right)$
- ⇨ $P\ (Z >= 2.309) = 0.010472 = 1.04\%$
- ⇨ As per the table above this is a strong evidence against H0
- ⇨ Quick P Value from Z Score Calculator (socscistatistics.com)

- Example 2: XYZ stockbroker claims that their brokerage fees are lower than that of ABC stock broker's. Data available from an independent research firm indicates that the mean and std-dev of all ABC broker clients are $18 and $6, respectively.
  A sample of 100 clients of ABC is taken and brokerage charges are calculated with the new rates of XYZ broker. If the mean of the sample is $18.75 and std-dev is the same ($6), can any inference be made about the difference in the average brokerage bill between ABC and XYZ broker?
- Solution:

$H0$	$\mu = 18$
$H1$	$\mu \neq 18$

- ⇨ $assuming\ a = 5\%, split\ 2.5\ each\ on\ either\ side$
- ⇨ Rejection region: $Z <= -Z_{2.5}\ and\ Z >= Z_{2.5}$
- ⇨ $z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- ⇨ $z = \frac{18.75-18}{\frac{6}{\sqrt{100}}} = 1.25$



- ⇨ $p - value = P(Z < -1.25) + P(Z > 1.25)$
- ⇨ $0.10565 + 0.10565 = 0.2112 = 21.12\%$
- ⇨ This concludes that there is insufficient evidence to infer that there is any difference between the rates of your existing broker and the new broker.
- ⇨ Alternatively, this calculated Z value falls between the two limits defined by: $-Z_{2.5} = -1.96\ and\ Z_{2.5} = 1.96$. the null hypothesis cannot be rejected.

- Criticism Points for the Hypothetical Testing Method:
  - A statistical method based on assumptions
  - Error-prone as detailed in terms of alpha and beta errors
  - Interpretation of p-value can be ambiguous, leading to confusing results
- Hypothesis testing allows a mathematical model to validate a claim or idea with a certain confidence level. But it is bound by a few limitations. The use of this model for making financial decisions should be considered with a critical eye, keeping all dependencies in mind. Important limitations are as follows:
- Testing is not decision-making itself; the tests are only useful aids for decision-making. Hence "proper interpretation of statistical evidence is important to intelligent decisions."
- Test do not explain the reasons as to why does the difference exist, say between the means of the two samples. They simply indicate whether the difference is due to fluctuations of sampling or because of other reasons but the tests do not tell us as to which is/are the other reason(s) causing the difference.

- Results of significance tests are based on probabilities and as such cannot be expressed with full certainty. When a test shows that a difference is statistically significant, then it simply suggests that the difference is probably not due to chance.
- Statistical inferences based on the significance tests cannot be said to be entirely correct evidences concerning the truth of the hypothesis. This is specially so in case of small samples where the probability of drawing wrong inferences happens to be generally higher. For greater reliability, the size of samples be sufficiently enlarged.

- **Kolmogorov-Smirnov Test (K-S Test)**
  - The Kolmogorov-Smirnov Goodness of Fit Test (K-S test) compares our data with a known distribution and lets us know if they have the same distribution. It is commonly used as a test for normality to see if the data is normally distributed.
  - More specifically, the test compares a known hypothetical probability distribution (e.g., the normal distribution) to the distribution generated by our data.
  - The hypotheses for the test are:
    - Null hypothesis (H0): the data comes from the specified distribution.
    - Alternate Hypothesis (H1): at least one value does not match the specified distribution.
    - That is,
      H0: P = P0, H1: P ≠ P0.
      Where P is the distribution of our sample (i.e., the CDF) and P0 is a specified distribution.
    - The K-S test statistic measures the largest distance between the CDF $F_{data}(x)$ and the theoretical function $F_0(x)$, measured in a vertical direction .The test statistic is given by:

      $$D = supremum\ |F_0(x) - F_{data}(x)| \\ = Max(CDF_0(x) - CDF_{data}(x)$$

    - If D is greater than the critical value, the null hypothesis is rejected. Critical values i.e., c (⍰) for D are found in the K-S Test P-Value Table. Here n is no of observations & ⍰ is significance level i.e., 0.05 is 5%.

n	α 0.01	α 0.05	α 0.1	α 0.15	α 0.2
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150
OVER 50	$\frac{1.63}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.07}{\sqrt{n}}$