INNOMATICS TECHNOLOGY HUB

A sister concern of

KenMinds
...committed to solution excellence

Analyzing attributes

Probability Distribution

Histogram

A series of contiguous rectangles that represent the frequency of data in given class intervals.

- How many class intervals?
- Rule of thumb: 5-15 (not too many and not too few) Freedman-Diaconis rule:

$$No.of\ bins = \frac{\max - min}{2\ *IQR\ *n^{-\frac{1}{3}}}$$

Where the denominator is the bin - width

Histogram - Excel

Annual traffic data for 30 busiest airports in the world – 2013

Passenger Traffic 2013 FINAL (Annual)

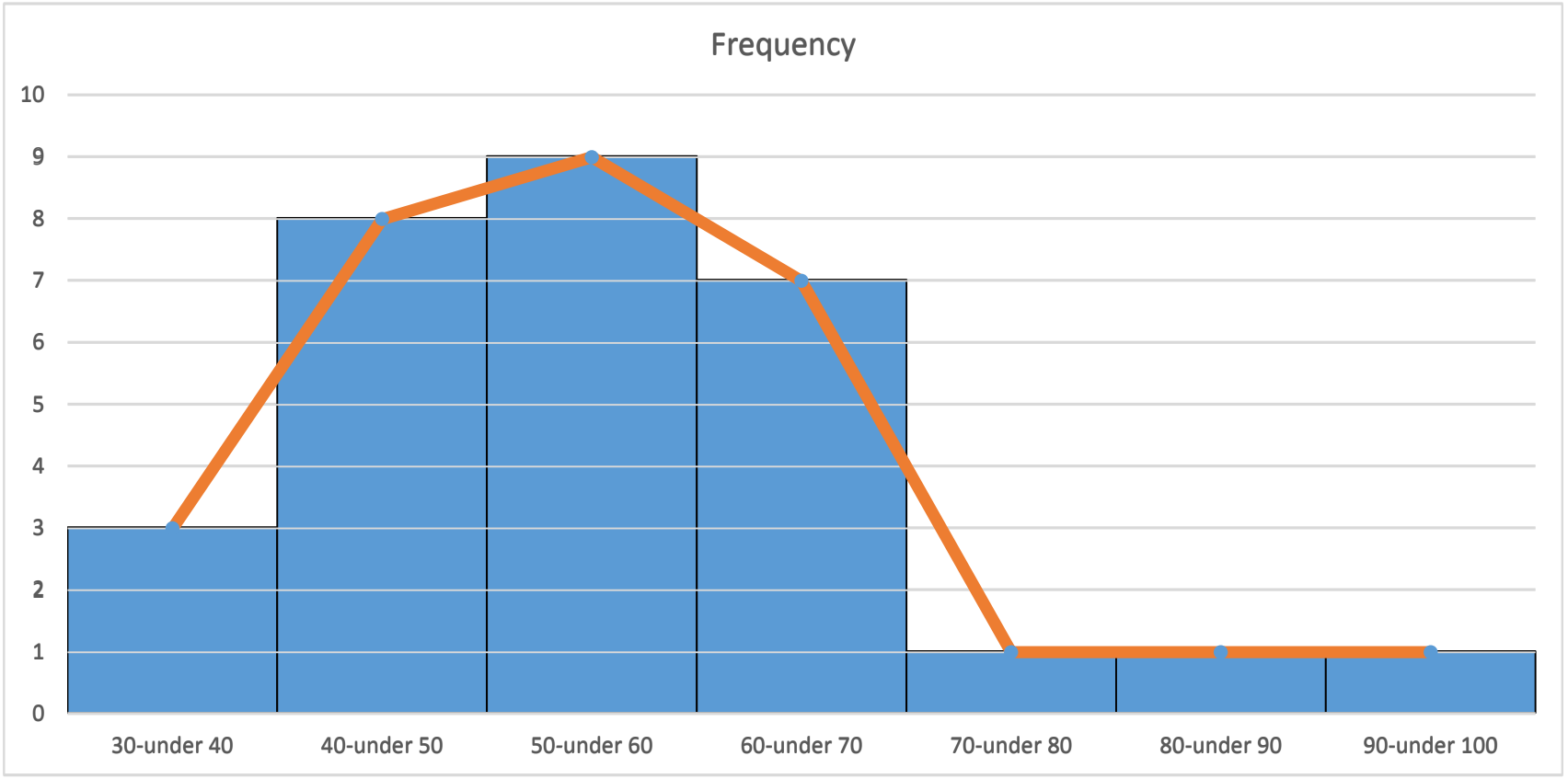
Last Update: 22 December 2014

Passenger Traffic

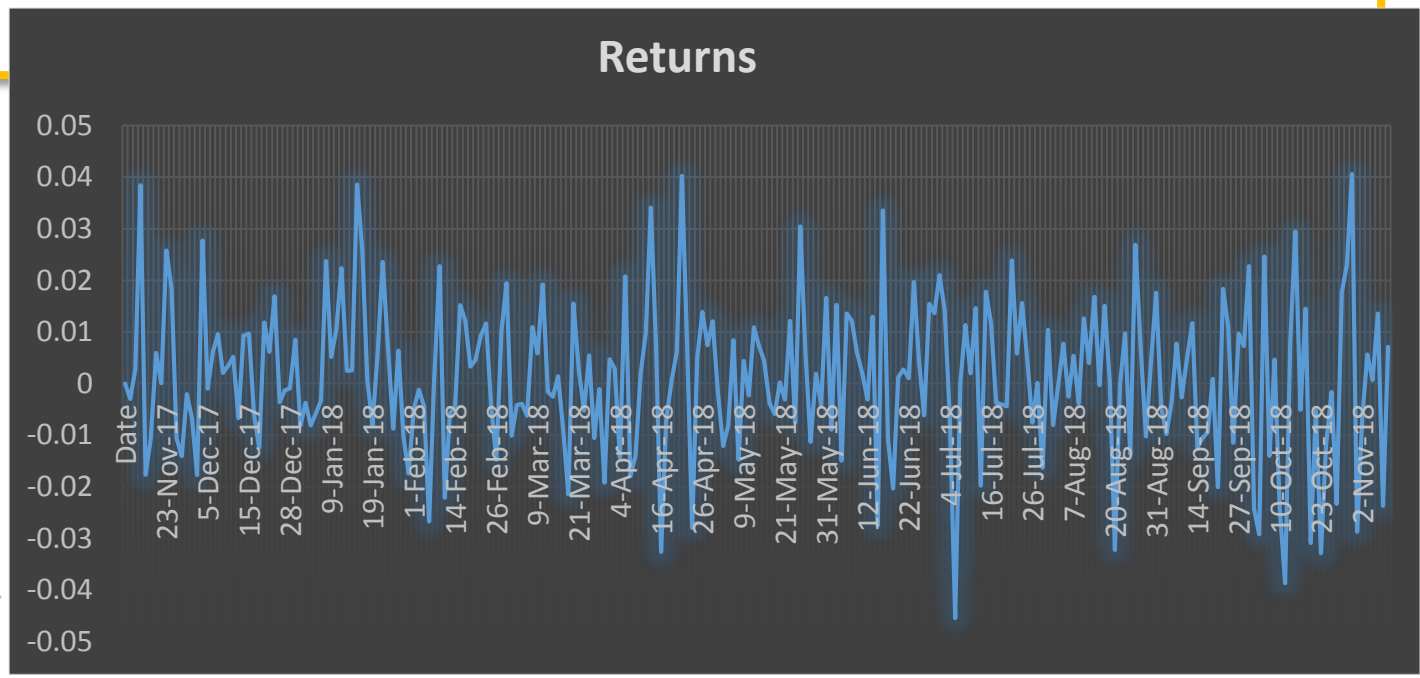
Total passengers enplaned and deplaned, passengers in transit counted once

Rank	City (Airport)	Passengers 2013	Passengers 2012	% Change
1	ATLANTA GA, US (ATL)	9,44,31,224	9,55,13,828	-1.1
2	BEIJING, CN (PEK)	8,37,12,355	8,19,29,359	2.2
3	LONDON, GB (LHR)	7,23,68,061	7,00,38,804	3.3
4	TOKYO, JP (HND)	6,89,06,509	6,67,95,178	3.2
5	CHICAGO IL, US (ORD)	6,67,77,161	6,66,29,600	0.2
6	LOS ANGELES CA, US (LAX)	6,66,67,619	6,36,88,121	4.7
7	DUBAI, AE (DXB)	6,64,31,533	5,76,84,550	15.2
8	PARIS, FR (CDG)	6,20,52,917	6,16,11,934	0.7
9	DALLAS/FORT WORTH TX, US (DFW)	6,04,70,507	5,86,20,160	3.2
10	JAKARTA, ID (CGK)	6,01,37,347	5,77,72,864	4.1
11	HONG KONG, HK (HKG)	5,95,88,081	5,60,61,595	6.3
12	FRANKFURT, DE (FRA)	5,80,36,948	5,75,20,001	0.9
13	SINGAPORE, SG (SIN)	5,37,26,087	5,11,81,804	5
14	AMSTERDAM, NL (AMS)	5,25,69,200	5,10,35,590	3
15	DENVER CO, US (DEN)	5,25,56,359	5,31,56,278	-1.1
16	GUANGZHOU, CN (CAN)	5,24,50,262	4,83,09,410	8.6
17	BANGKOK, TH (BKK)	5,13,63,451	5,30,02,328	-3.1
18	ISTANBUL, TR (IST)	5,13,04,654	4,51,23,758	13.7
19	NEW YORK NY, US (JFK)	5,04,23,765	4,92,91,765	2.3
20	KUALA LUMPUR, MY (KUL)	4,74,98,127	3,98,87,866	19.1
21	SHANGHAI, CN (PVG)	4,71,89,849	4,48,80,164	5.1
22	SAN FRANCISCO CA, US (SFO)	4,49,45,760	4,43,99,885	1.2
23	CHARLOTTE NC, US (CLT)	4,34,57,471	4,12,28,372	5.4
24	INCHEON, KR (ICN)	4,16,79,758	3,91,54,375	6.4
25	LAS VEGAS NV, US (LAS)	4,09,33,037	4,07,99,830	0.3
26	MIAMI FL, US (MIA)	4,05,62,948	3,94,67,444	2.8
27	PHOENIX AZ, US (PHX)	4,03,41,614	4,04,48,932	-0.3
28	HOUSTON TX, US (IAH)	3,97,99,414	3,98,91,444	-0.2
29	MADRID, ES (MAD)	3,97,17,850	4,51,76,978	-12.1
30	MUNICH, DE (MUC)	3,86,72,644	3,83,60,604	0.8

Annual traffic data for 30 busiest airports in the world – 2013



Stock Returns

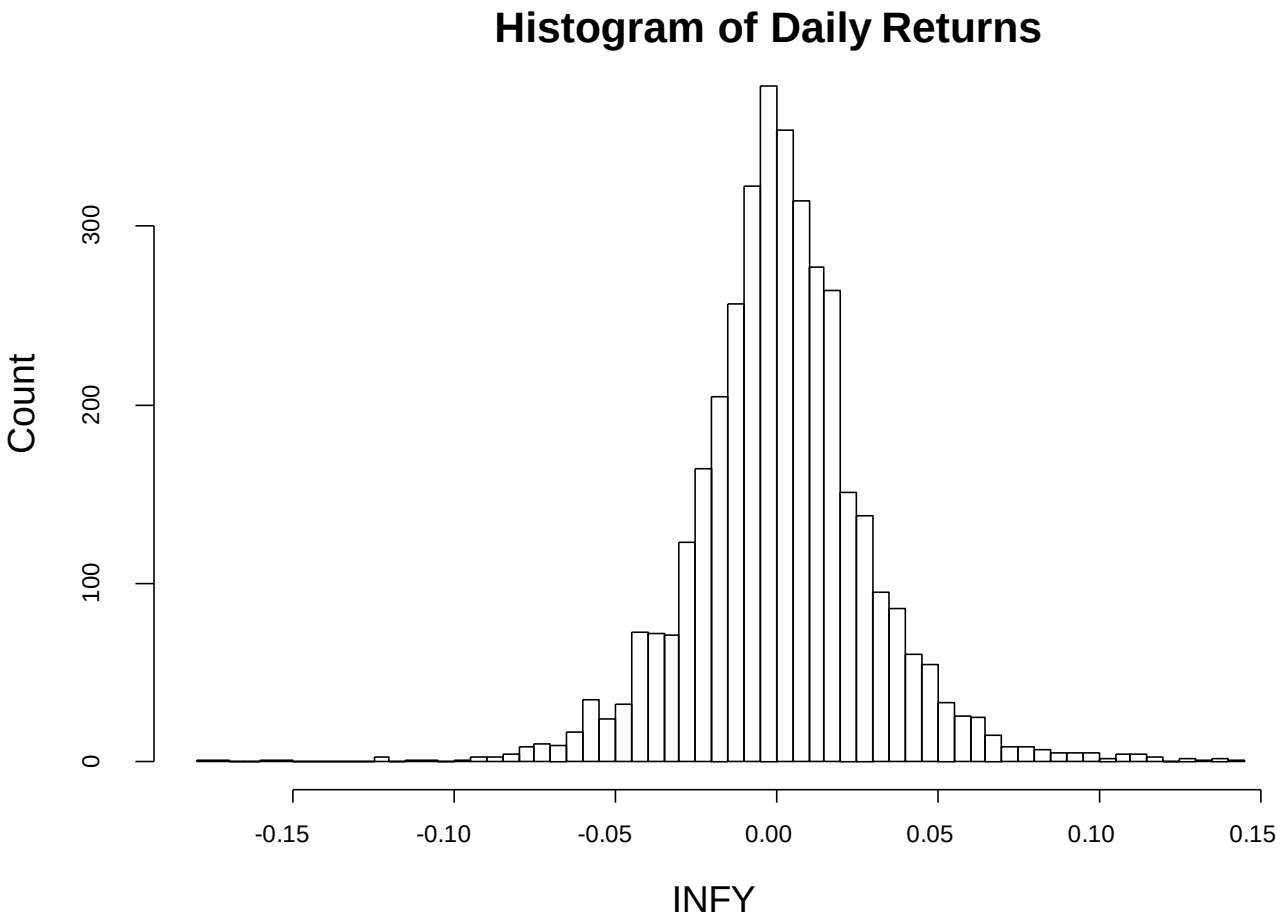


Infosys Ltd
NSE: INFY

652.40 INR −13.30 (2.00%)
14 Nov, 3:52 PM IST · Disclaimer

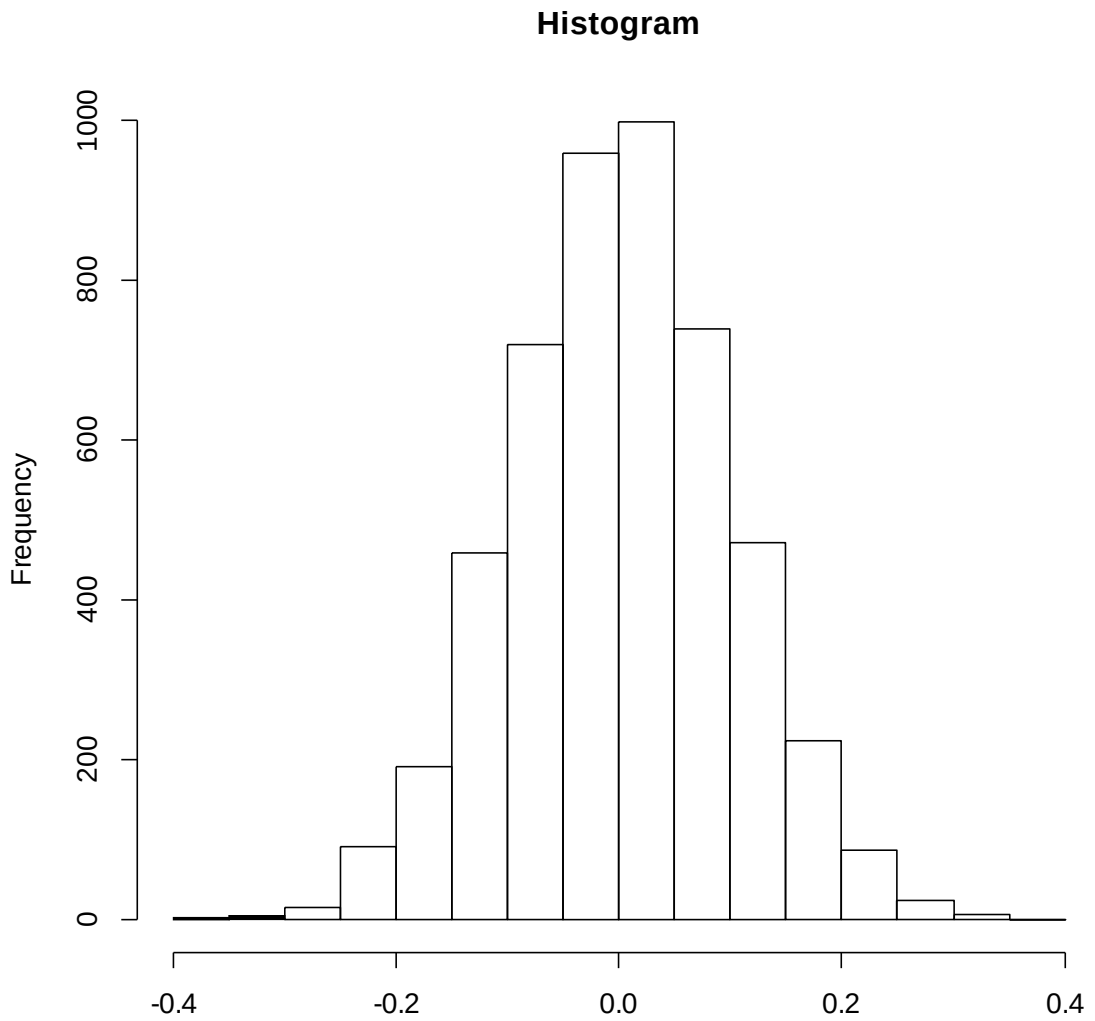
1 day | 5 days | 1 month | 6 months | YTD | 1 year | 5 years | Max

Histogram of Stock Returns



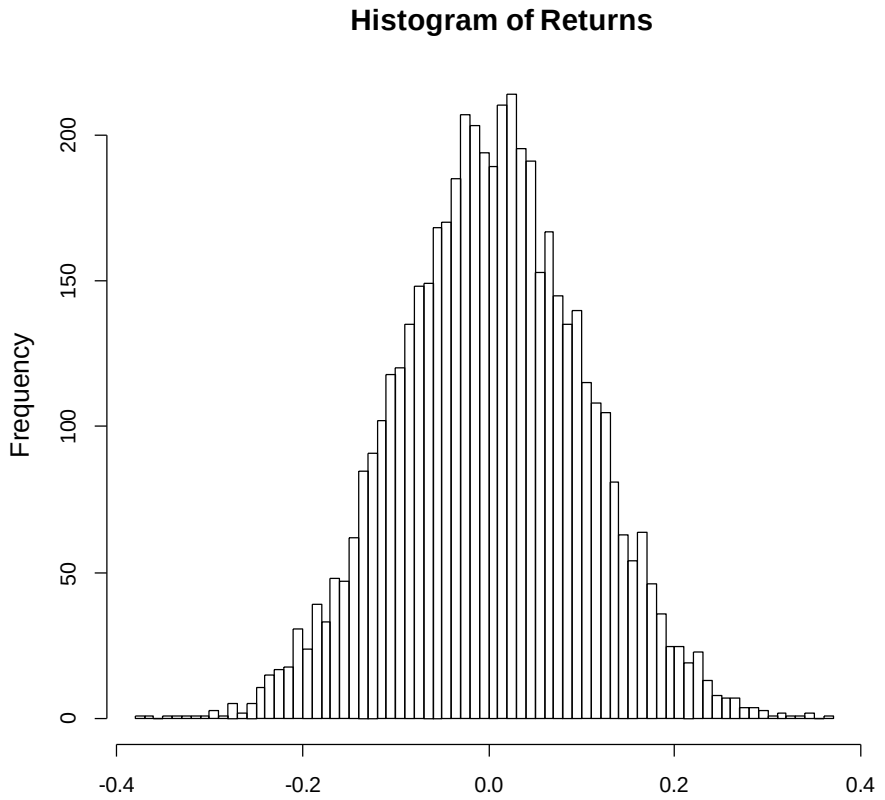
Histogram of Stock Returns

- Consider histogram of stock returns from 5000 days



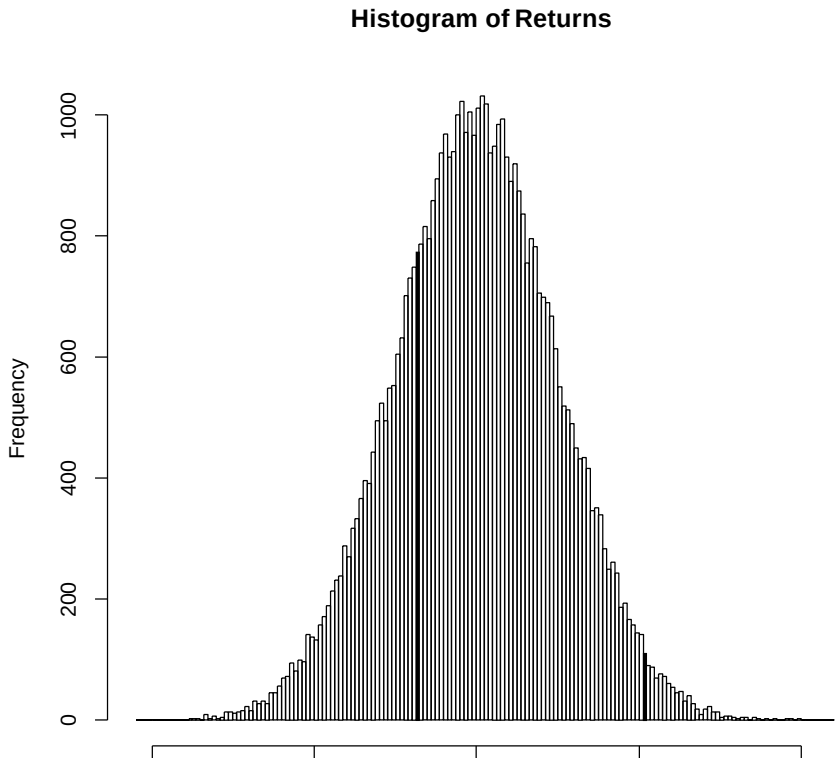
Histogram of Stock Returns

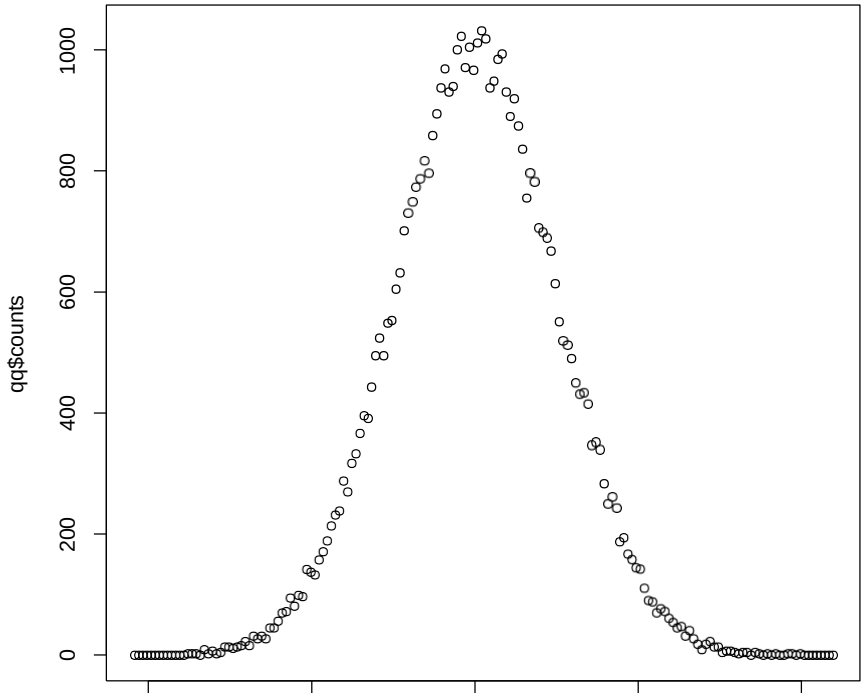
- The same histogram with larger number of bins



Histogram of Stock Retruns

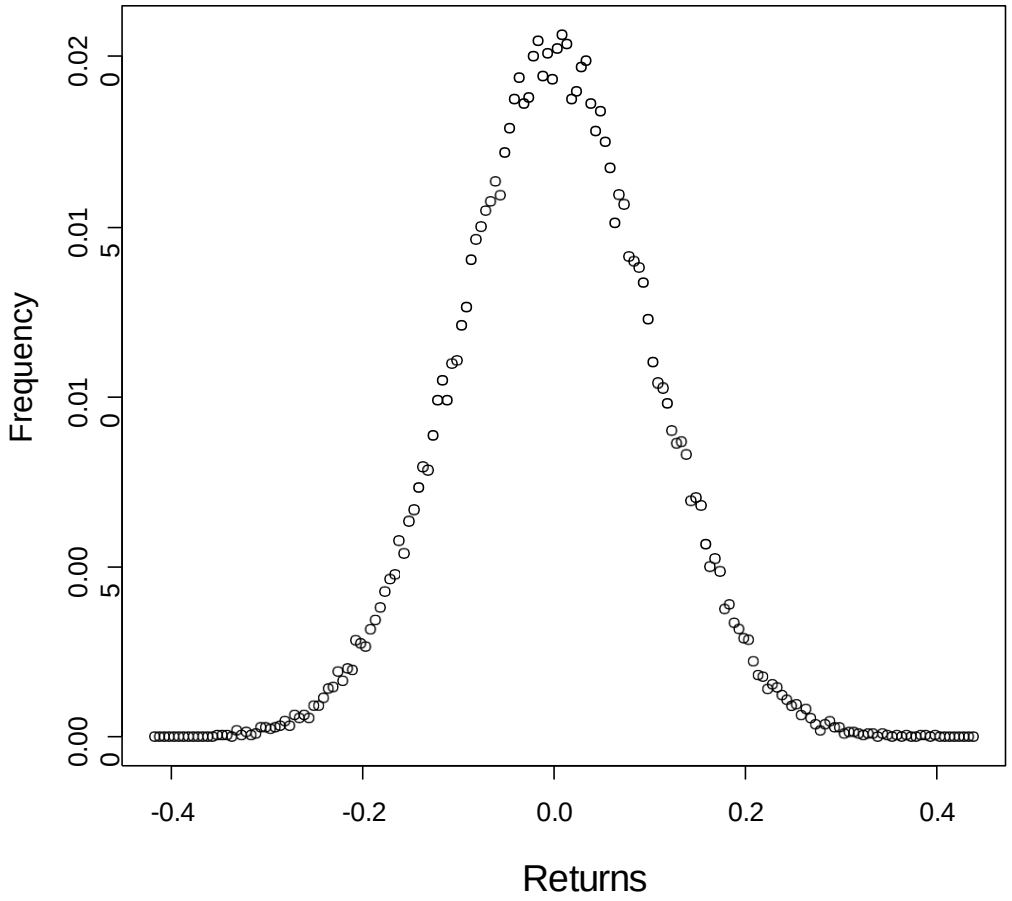
- 50,000 data points with 200 bins





Histogram/Probability Distribution Function

Converts the counts to frequency by dividing by 50,000

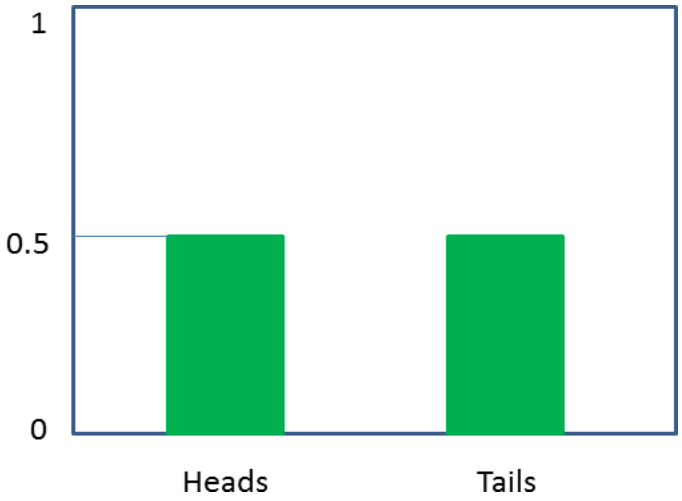


Random Variable

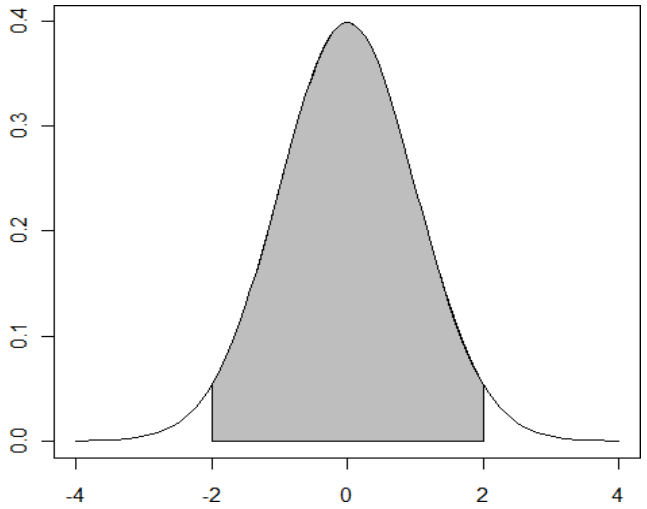
- Random Variable- a variable that can take multiple values with different probabilities.

- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

Discrete and Continuous



Countable



Measurable

Can any function be a probability distribution ?

Discrete Distributions	Continuous Distributions
Probability that X can take a specific value x is $P(X = x) = p(x)$	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$
It is non-negative for all real x	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx$
Probability Mass Function	**Probability Density Function**

Probability Distribution

Possible Outcome	$	Cherry	Lemon	Other
Probability of Outcome	0.1	0.2	0.2	0.5

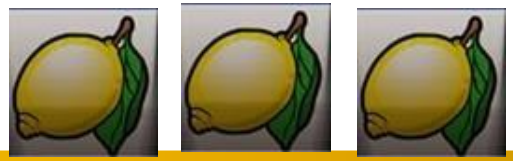
Cost: Rs.10 for each game

Winning combinations:

$= Rs.200$

$= Rs.150\ (any\ order)$

$= Rs.50$



$= Rs.20$

Probability of Winning Combination

Possible Outcome	$	Cherry	Lemon	Other
Probability of Outcome	0.1	0.2	0.2	0.5

$$= 0.1 * 0.1 * 0.1 = 0.001$$

$$= 3(0.1 * 0.1 * 0.2) = 0.006$$

$$= 0.2 * 0.2 * 0.2 = 0.008$$

$$= 0.2 * 0.2 * 0.2 = 0.008$$

No win probability?

$$= 1 - (win\ something)$$

$$= 1 - (0.001 + 0.006 + 0.008$$

Probability of Wining Combination

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	- Rs.10	Rs.20	Rs.50	Rs.150	Rs.200

Cost: Rs.10 for each game Winning combinations:

$= Rs. 200$

$= Rs. 150\ (any\ order)$

$= Rs. 50$

$= Rs. 20_{17}$

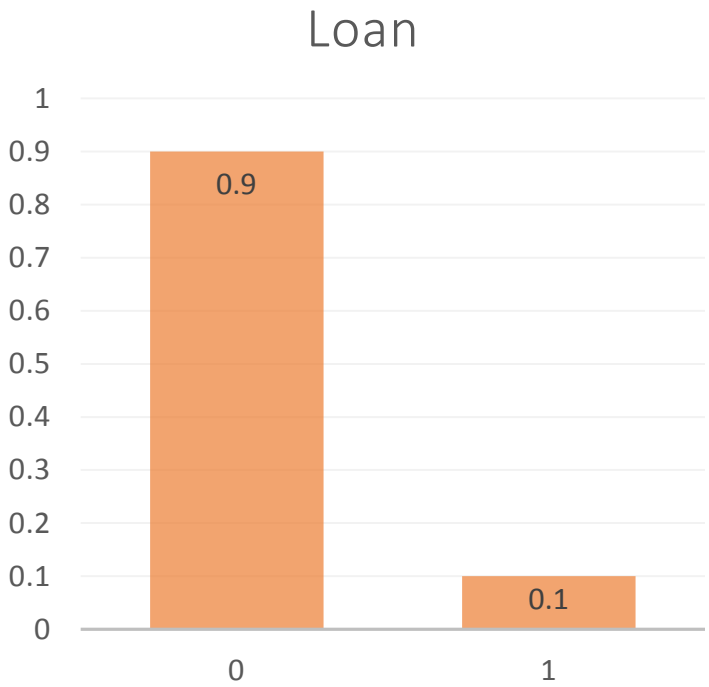
Probability Distribution of Winngings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	- Rs.10	Rs.20	Rs.50	Rs.150	Rs.200

Why do you need a probability distribution?

Once a distribution is calculated, it can be used to determine the EXPECTED outcome.

Expectation: Discrete



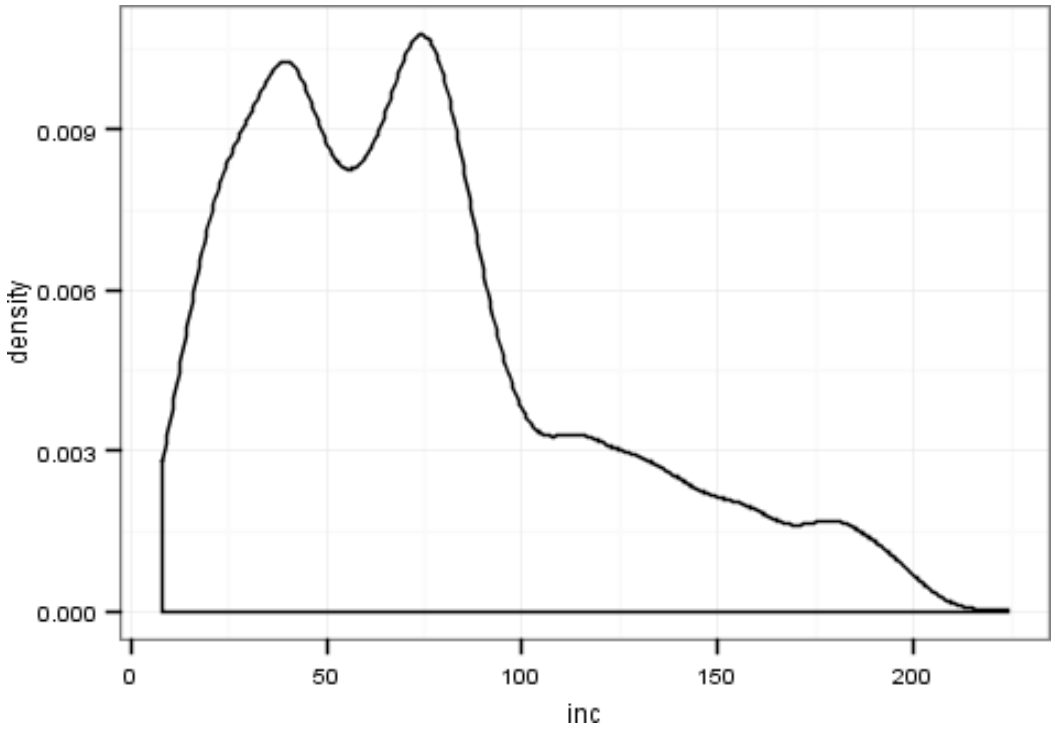
Age (years)	13	15	17
Frequency, f	1	3	2

$$Mean, \mu = \frac{\sum x}{n} = \frac{\sum xf}{\sum f} = \frac{13 * 1 + 15 * 3 * 17 * 3}{1+3+2} = 13 * \frac{1}{6} + 15 * \frac{3}{6} + 17 * \frac{2}{6}$$

Recall Assigning Probabilities using Empirical or Frequentist Method

Expectation: Continuous

-



Probability Distribution of Winnings

Combination	None	Lemons	Cherries	Dollars/Cherry	Dollars
Probability	0.977	0.008	0.008	0.006	0.001
Gain	- Rs.10	Rs.20	Rs.50	Rs.150	Rs.200

Expectation, $E(x) = \mu = \sum xP(X = x)$

$$E(x) = (-10) * 0.977 + (20) * 0.008 + (50) * 0.008 + (150) * 0.006 + (200) * 0.001$$

$$= -8.11$$

This is the amount of Rs. expected to be “gained” on each pull of lever.

So, why play ?

It never makes sense to play the slot machine or the lottery

Until it does ?

State Lottery

match all 6 numbers	1 in 9.3 million	variable jackpot
match 5 of 6	1 in 39,000	$4,000
match 4 of 6	1 in 800	$150
match 3 of 6	1 in 47	$5
match 2 of 6	1 in 6.8	free lottery ticket

Cost of ticket = $2

Jackpot value = At least $ 1 Million

$$E(x) = \frac{\$\ 1\ million}{9.3\ million} + \frac{\$4{,}000}{39{,}000} + \frac{\$150}{800} + \frac{\$5}{47} + \frac{\$2}{6.8}$$

$$= 79.8\ cents$$

State Lottery

RollDay - When the Jackpot increases to $2M, then prize money for Match 5 also increases

Prize	Chance of winning	Expected number of winners	Roll-down allocation	Roll-down per prize
match 5 of 6	1 in 39,000	12	$600,000	$50,000
match 4 of 6	1 in 800	587	$1.4m	$2,385
match 3 of 6	1 in 47	10,000	$600,000	$60

Expected value on the roll day changes dramatically.
E(x) = $5.53

Variance of the Distribution

The Width/Spread of the distribution

VARIANCE, $Var(X) = E(X - \mu)^2 = \sum(x - \mu)^2 P(X = x)$

$$\sigma = \sqrt{VarX}$$

Expectation Properties

E(X+Y) = E(X) + E(Y) e.g., Playing a game each on 2 slot machines with different probabilities of winning. This is called Independent Observation.

- E(aX+b) = aE(X)+E(b) = aE(X) + b e.g., values x have been changed. This is called Linear Transformation.

** Not all central tendencies posses this nice property*

Variance Properties

- Var(X+a) = Var(X) (Variance does not change when a constant is added)
- For Independent Observations
  - : Var(X+Y) = Var(X) + Var(Y)
  - : Var(X-Y) = Var(X) + Var(Y)

- $Var(aX) = a^2 Var(X)$

Simplifying the formula

$$E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2]$$

$$= E[X^2] - 2\mu E[X] + \mu^2$$ (we get this from previous formula as m is just a number)

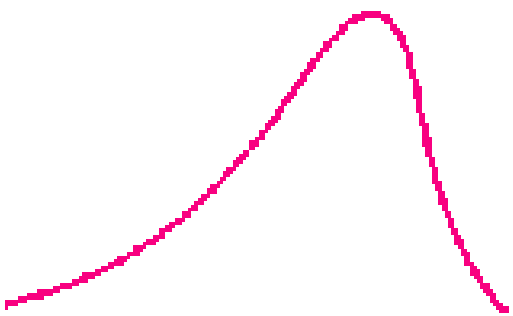
$$= E[X^2] - 2\mu^2 + \mu^2$$

$$= E[X^2] - \mu^2 = E[X^2] - [E(X)]^2$$

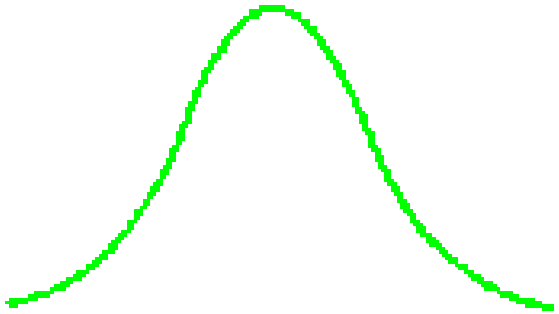
Understanding the shape of PDF - Skewness

- A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

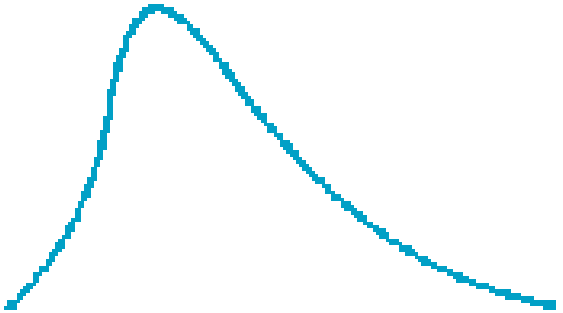
Understanding the shape of a PDF - Skewness



Negatively (left) skewed distribution



Normal skewed distribution



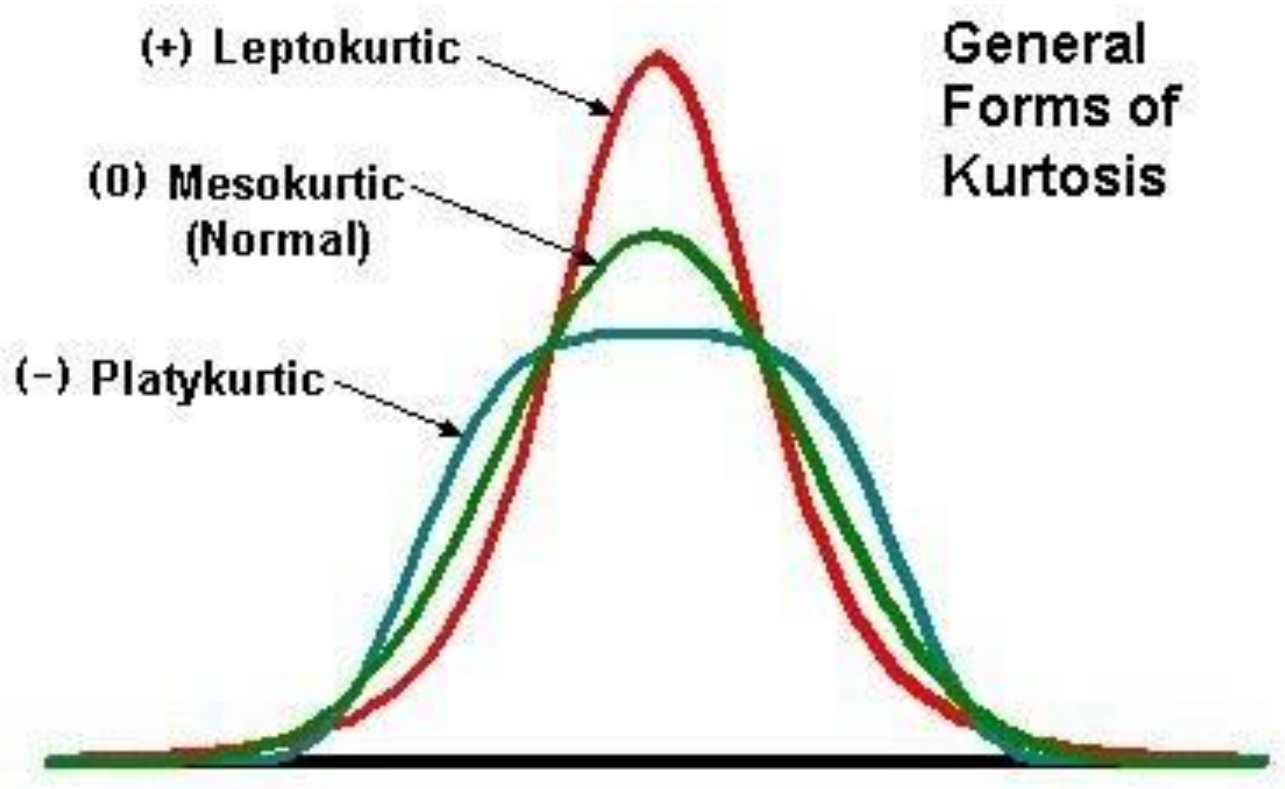
Positively (right) skewed distribution

Understanding the shape of a PDF Kurtosis

A measure of the ‘peaked’ness of the data distribution. Negative Excess-kurtosis means a flat distribution. Positive Excess-kurtosis means a peaked distribution.

$$Excess\ Kurtosis = \frac{E[(X-\mu)^4]}{\sigma^4} - 3$$

Understanding the shape of a PDF Kurtosis



Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Guide to Airline Fees in India

cleartrip

	Change fee (Domestic)	Change fee (International)	Cancellation fee (Domestic)	Cancellation fee (International)	No show charges (Domestic)	No show charges (International)
Indigo	Rs 1000 / passenger / sector	Rs 1,850 / passenger / sector	Rs 1,000 / passenger / sector	Rs 1,850 / passenger / sector	No refund	No refund
Jet Airways	Rs 250 - 997 (Premiere) Rs 500 - 1050 (Economy)	Rs 5,500 to NIL (depending on fare class)	Rs 500 - 997 (Premiere) Rs 750 - 1,050 (Economy)	Rs 8,000 to NIL (depending on fare class)	Rs 1,500 to NO REFUND (depending on fare class)	Rs 8,000 to NIL (depending on fare class)
JetKonnect	Rs 250 - 997 (Premiere) Rs 500 - 1050 (Economy)	NA	Rs 500 - 997 (Premiere) Rs 750 - 1,050 (Economy)	NA	Rs 1,500 to NO REFUND (depending on fare class)	NA
Spicejet	Rs 950 / passenger / sector	Rs 1,750 / passenger / sector	Rs 950 / passenger / sector	Rs 1,750 / passenger / sector	No refund	No refund
GoAir	Rs 950 (GoSmart) NIL (GoFlexi & GoBusiness)	NA	Rs 950 (GoSmart) Rs 350 (GoFlexi) NIL (GoBusiness, >24 hrs) Rs 750 (GoBusiness, <24 hrs)	NA	12 month credit shell for PSF + service tax	NA
Air India	Rs 750 - NIL (Economy, based on fare class); NIL (Executive / First Class)	Rs 5,000 - NIL (Economy) Rs 7,500 - NIL (Executive) Rs 5,000 - NIL (First class)	Rs 500 to NO REFUND (Economy) Rs 200 (Executive / First)	No refund (Economy Web Specials) Rs 5,000 - NIL (Economy) Rs 14,000 - NIL (Executive) Rs 5,000 - NIL (First class) + Rs 300 Refund Administration fee (all classes)	Rs 1,500 to NO REFUND (Economy); Rs 200 (Executive / First class)	Rs 5,000 - NIL (Economy) Rs 14,000 - NIL (Executive) Rs 5,000 - NIL (First class) + Rs 300 Refund Administration fee (all classes)
Kingfisher	Rs 950 (Kingfisher Red); Rs 500-950 (Kingfisher, Kingfisher First)	NA	Rs 950 (Kingfisher Red) Rs 500 - 100% of Base Fare (Kingfisher, Kingfisher First)	NA	NO REFUND (Kingfisher Red, Kingfisher); Rs 1,000 + Cancellation / change fee (Kingfisher First)	NA

Data sourced from airline websites, accurate as of 18 September 2012.
Always check fare rules before booking. Visit airline website for more details.

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Kingfisher Airlines* would like to maximize revenues by ensuring no empty seats on its flight between Bengaluru and Hyderabad. They intentionally wish to overbook the flights based on the historical data of no-shows on this sector.

- You have been hired as a statistical consultant to help formulate a solution.

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

The frequency distribution of "No-Shows" from 200 randomly selected flights on this sector is:

# of No-Shows	1	2	3	4	5	6	Total
Frequency	70	40	10	20	20	40	200

What is your advice for Kingfisher on the number of seats they should overbook on this sector?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

What is the Random Variable in the problem ?

Random Variable, X is the # of No-Shows.

What is the PMF for the frequency distribution seen in the sample ?

# of No-Shows	1	2	3	4	5	6	Total
Frequency	70	40	10	20	20	40	200

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

What is the Expectation ?

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 1 * 0.35 + 2 * 0.20 + 3 * 0.05 + 4 * 0.10 + 5 * 0.10 + 6 * 0.20 = 3$$

So, you'd advise Kingfisher to overbook 3 seats on this sector, which is the **mean** of the data in the sample.

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 1: Kingfisher tells you that it will pay you Rs 500 for your consulting and Rs 1500 as bonus for each correct prediction (prediction must be exactly correct, no more no less). Will you still go with the **mean**?

X	1	2	3	4	5	6
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 1 * 0.35 + 2 * 0.20 + 3 * 0.05 + 4 * 0.10 + 5 * 0.10 + 6 * 0.20 = 3$$

So, will you advise Kingfisher to overbook 3 seats on this sector, which is the **mean** of the data in the sample?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 1

What is the probability distribution of your earnings if you went with the expected value (or the mean)?

X (Your earnings)	500	500	2000	500	500	500
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 500 * 0.35 + 0.20 + 0.10 + 0.10 + 0.20 + 2000 * 0.05 = Rs.575$$

How much would you earn in other cases?

Would you still stick to Mean or switch to Median or Mode?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 2

Instead of a binary state for your earnings, if Kingfisher offers to pay you Rs 2000 for the consulting minus Rs 125 for each under or overbooked seat, what will be your advice now?

$$E(X) = 2000 * 0.35 + 1875 * 0.20 + 1750 * 0.05 + 1625 * 0.10 + 1500 * 0.10 + 1375 * 0.20$$

$$= Rs\ 1750$$

How much would you earn in other cases?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

Scenario 3

Instead of penalizing based on absolute magnitude of the prediction error, if Kingfisher offers to pay you Rs 2500 for the consulting minus Rs 75 times the square of the prediction error (penalizing larger errors more), what will be your advice now?

X (Your earnings)	2500	2425	2200	1825	1300	625
P(X=x)	0.35	0.20	0.05	0.10	0.10	0.20

$$E(X) = 2500 * 0.35 + 2425 * 0.20 + 2200 * 0.05 + 1825 * 0.10 + 1300 * 0.10 + 625 * 0.20$$

$$= Rs\ 1907.50$$

How much would you earn in other cases?

Central Tendencies and Probability Distribution from a Decision Perspective – A Real World Scenario

- **Conclusion**

For the same dataset, depending on the business problem, Mode was the best option in Scenario 1, Median in Scenario 2 and Mean in Scenario 3.

Moral of the story

- You should look at data carefully in the context of the business domain and problem.
- You must inculcate statistical way of thinking in all you do.
- Statistics don't lie; Statisticians may.
- In God we Trust; all others must bring data.

Reference

- **INSOFE.** www.insofe.edu.in

- Conditional probability explained visually https://www.khanacademy.org/video/conditional-probability2
- Bayes Theorem : https://youtu.be/E4rlJ82CUZI

- Creating a histogram: https://www.khanacademy.org/video/histograms-intro

- Probability Distribution Functions
- https://www.khanacademy.org/video/discrete-probability-distribution
- https://www.khanacademy.org/video/probability-density-functions