Analyzing attributes

Probability Distribution

Histogram

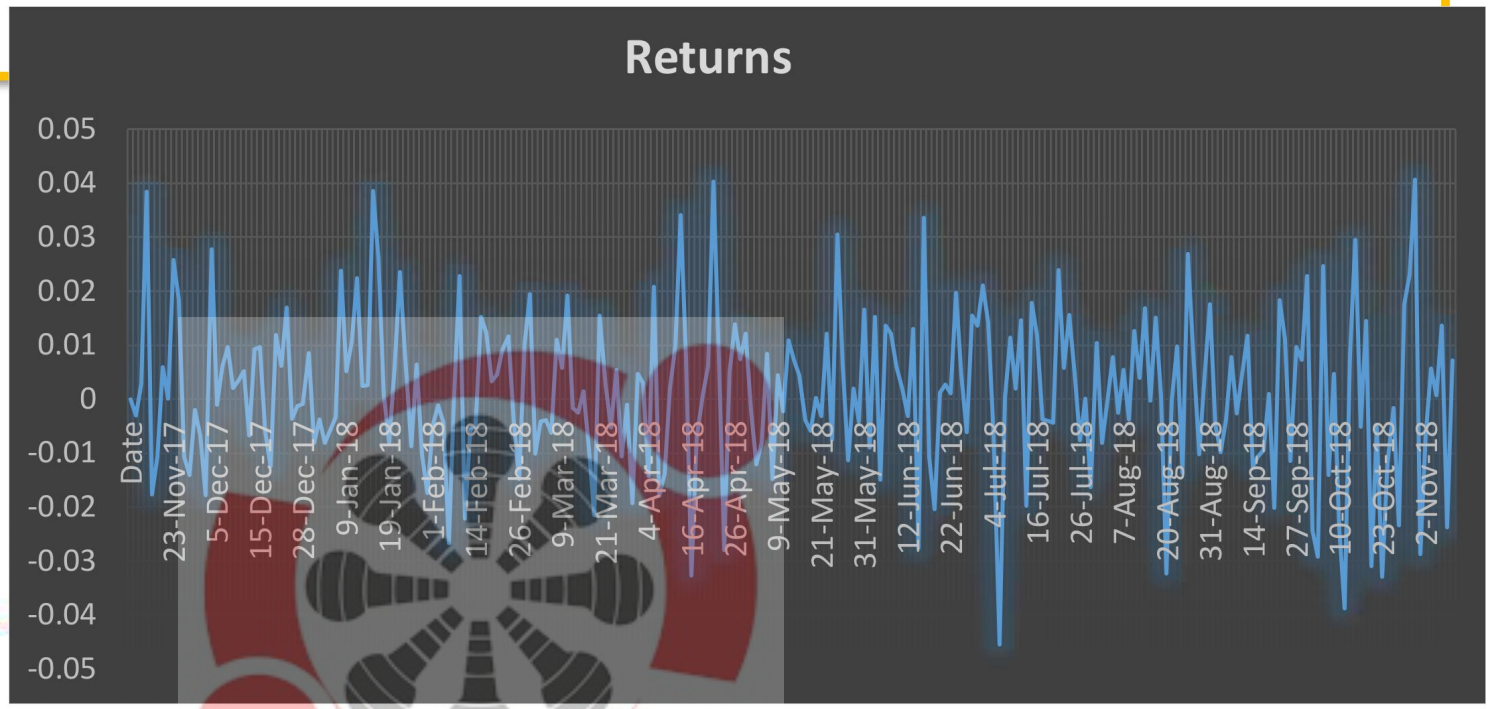
A series of contiguous rectangles that represent the frequency of data in given class intervals.

- 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

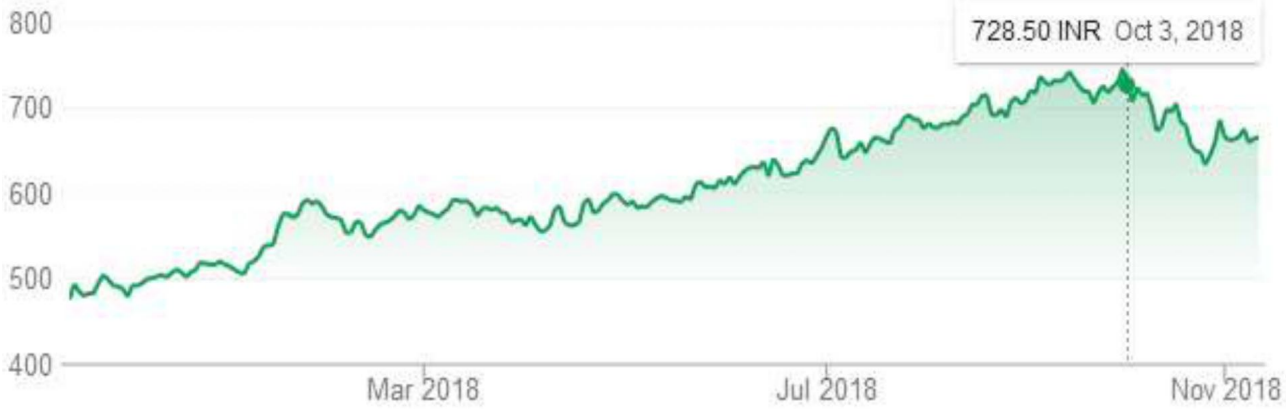
Stock Returns



Infosys Ltd
NSE: INFY

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

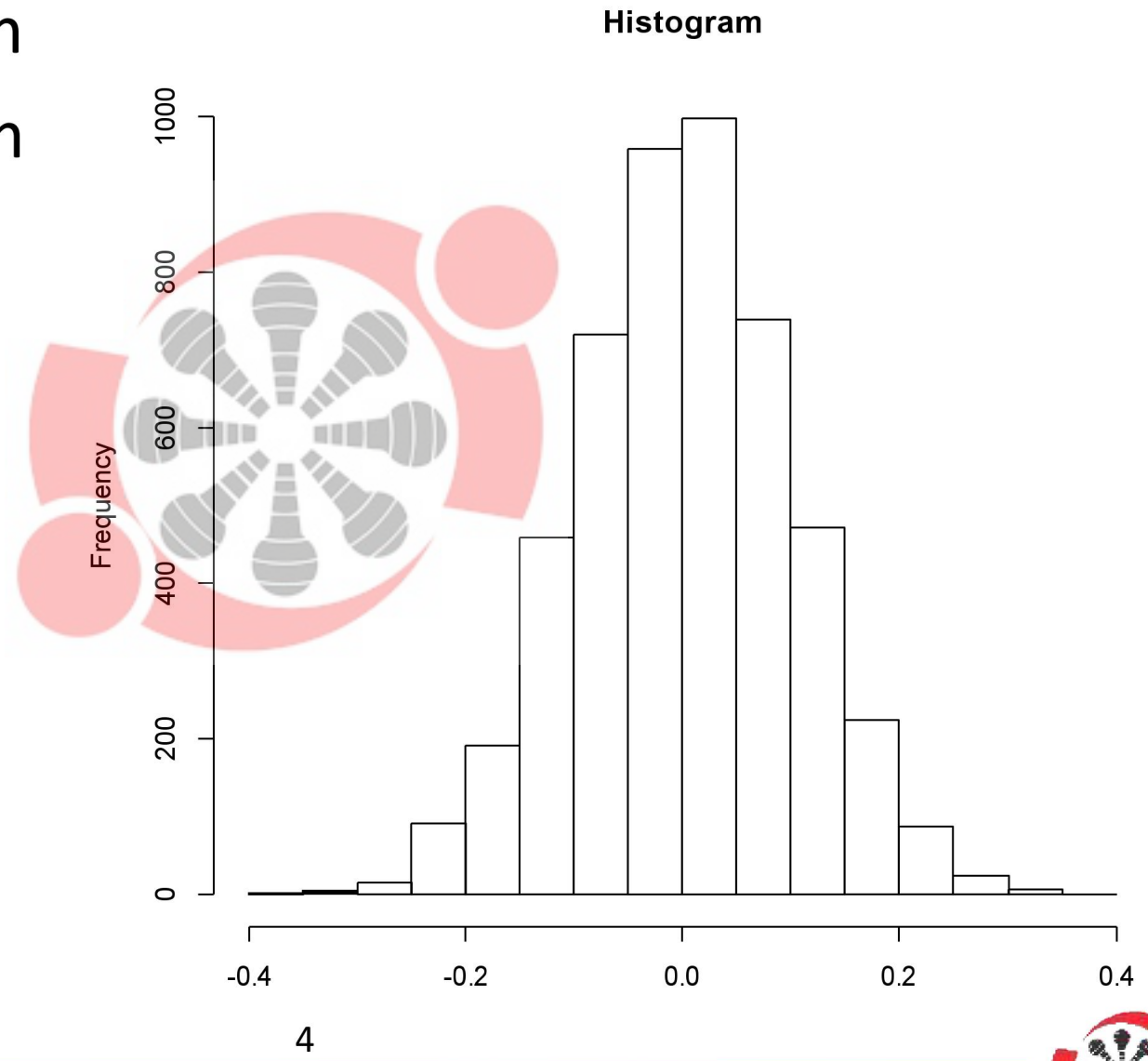
1 day | 5 days | 1 month | 6 months | YTD | 1 year | 5 years | Max



MATICS RESEARCH LAB

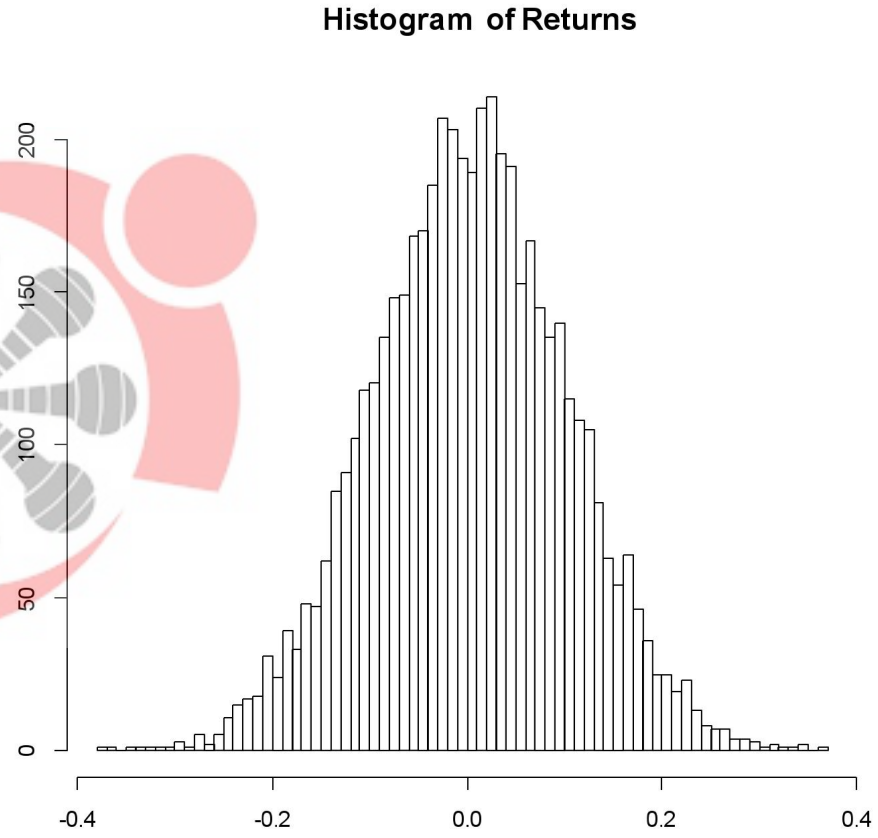
Histogram of Stock Returns

- Consider histogram of stock returns from 365 days



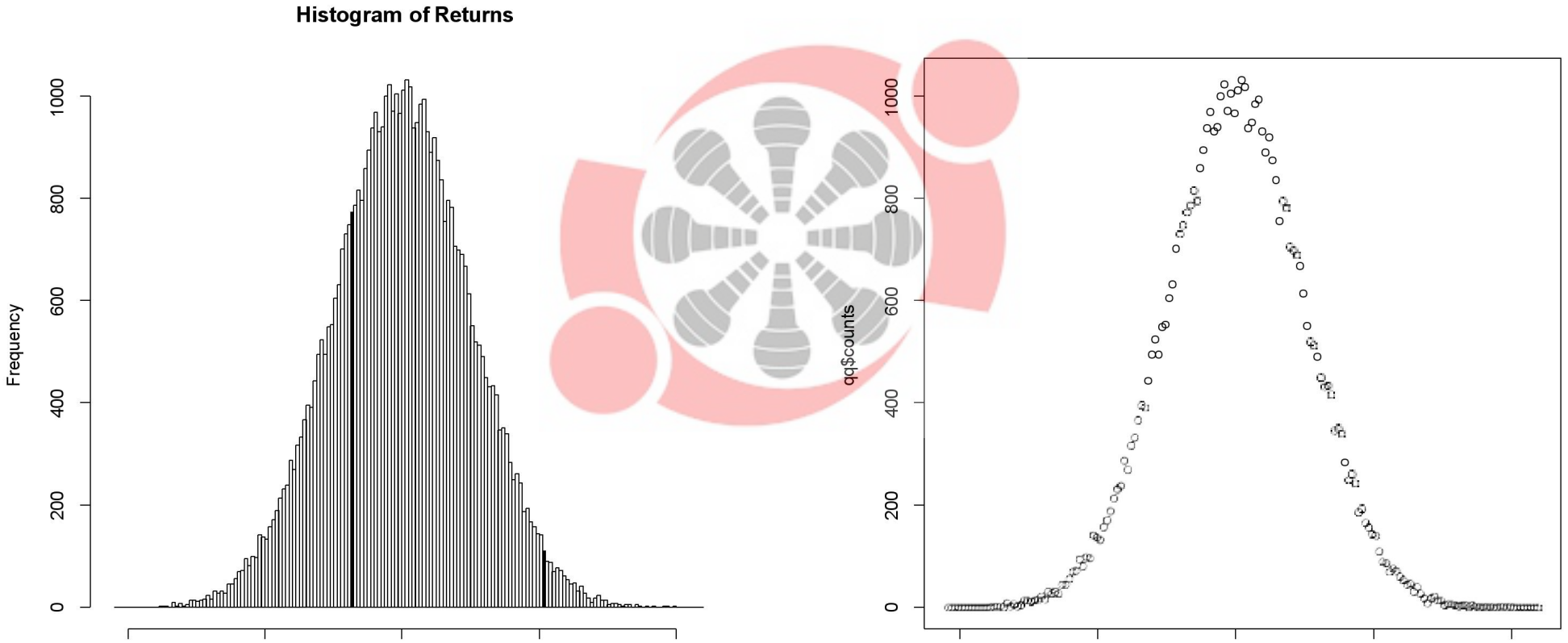
Histogram of Stock Returns

- The same histogram with larger number of bins



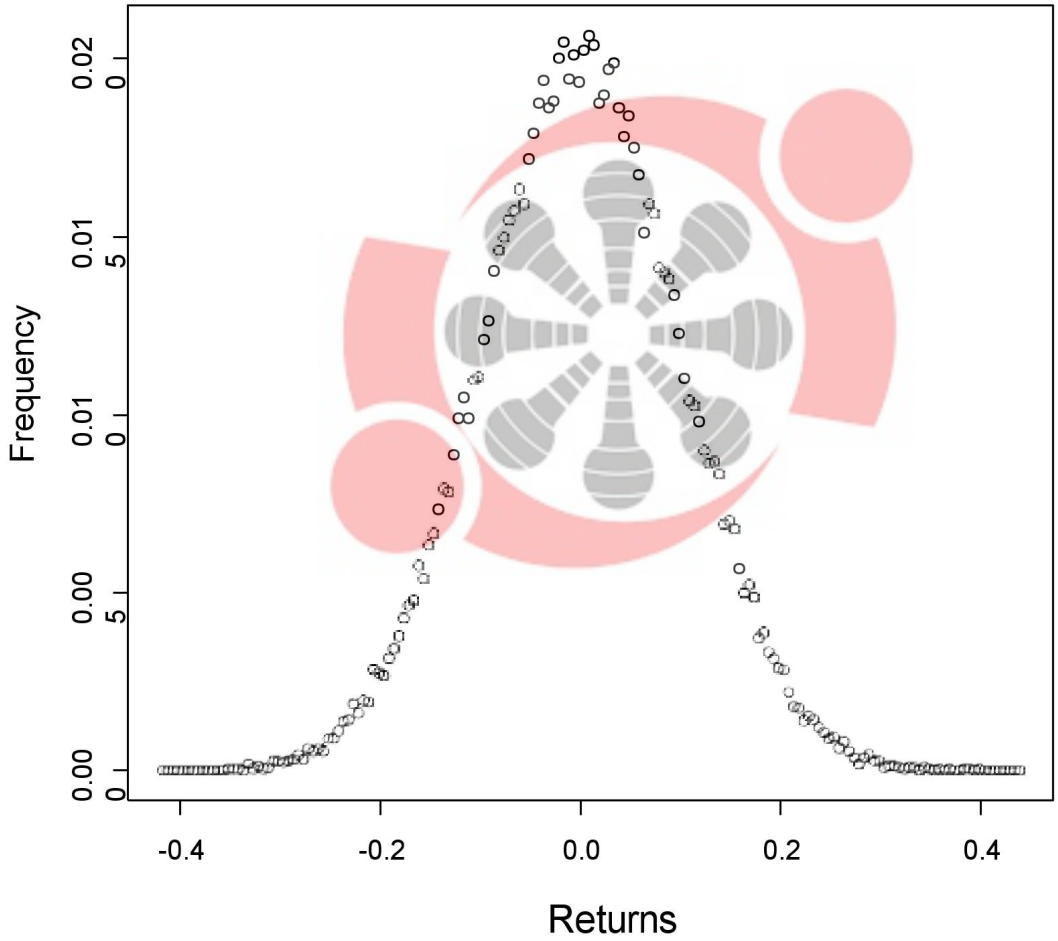
Histogram of Stock Retruns

- 500,000 data points with 2000 bins



Histogram/Probability Distribution Function

Converts the counts to frequency by dividing by 500,000

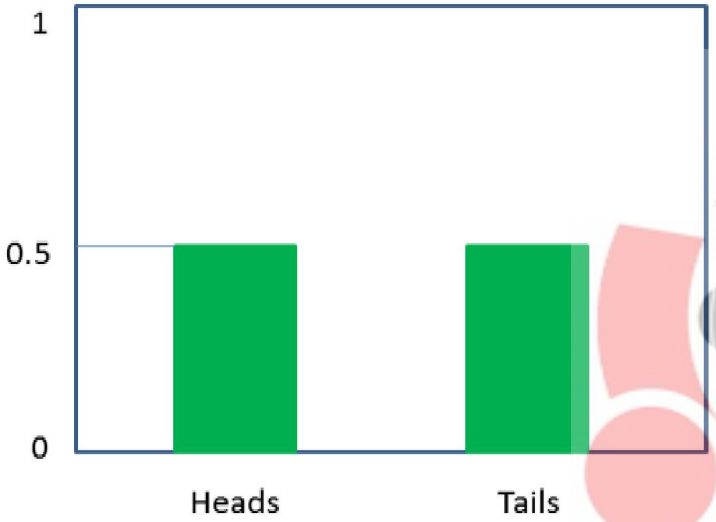


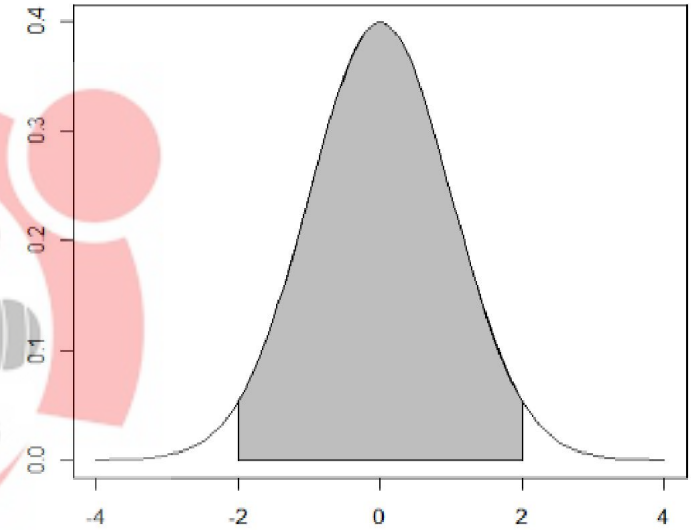
Random Variable

- 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

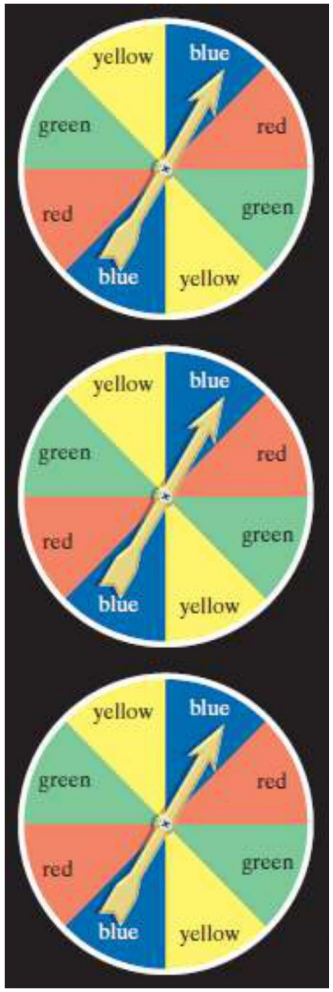
Measurable

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:

$= Rs.200$

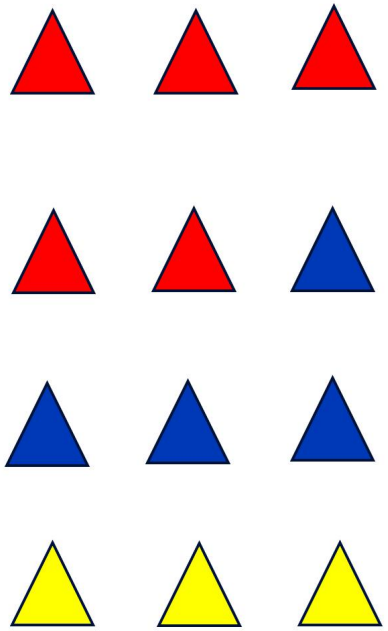
$= Rs.150\ (any\ order)$

$= Rs.50$

$= Rs.20$

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:

$= Rs.200$

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

$= Rs.50$

$= Rs.20_{13}$

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.

Review

Yoga class composition

Age (years)	19	22	23
Frequency, f	1	3	2

$$Mean, \mu = \frac{\sum x}{n} =$$

$$\frac{19 * 1 + 22 * 3 + 23 * 2}{1 + 3 + 2} \approx 22$$

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

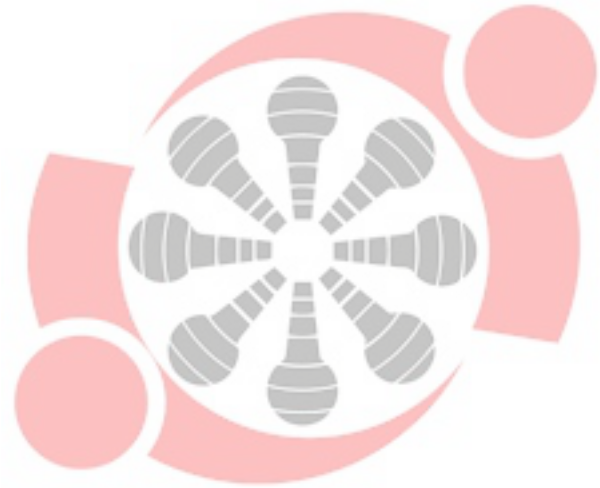
Until it does ?

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)

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

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

INNOMATICS RESEARCH LAB