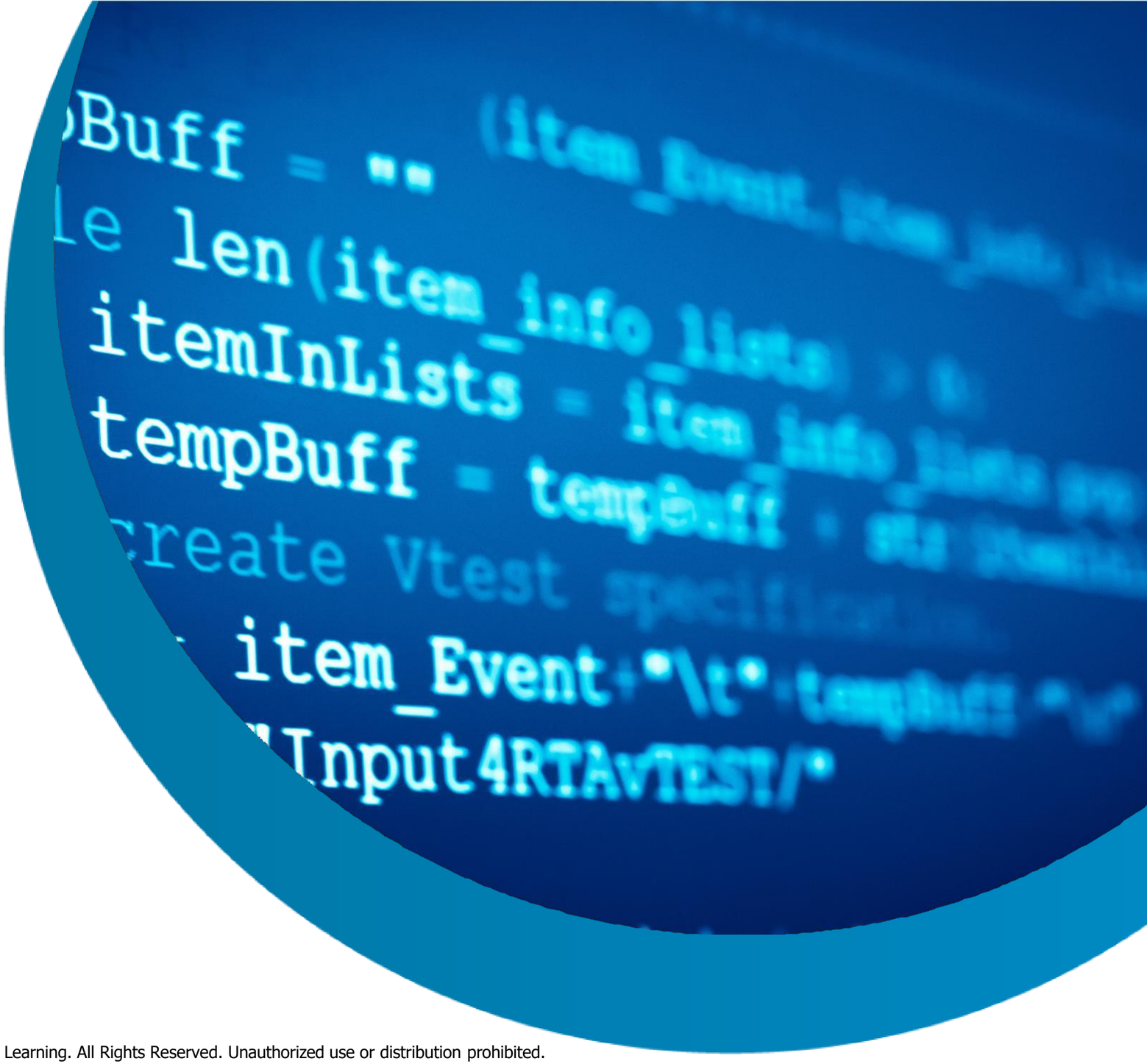
FOUNDATION OF DATA SCIENCE

"Statistics for Data Science" - WEEK 2

LEARNING OBJECTIVE OF THIS MODULE

- Basic Working proficiency in Python
- Basic Data-Manipulation using Python
- Basic Data-Visualization using Python
- Basics of Statistics



LET'S SET SOME GROUND RULES

- Come prepared for these sessions by watching the videos.
  - Concepts will be covered in the videos.

- Submit all assignments on time.

- Let's be punctual & respect each other's time.

Difference Between Continuous & Categorical Variables

- **Continuous Variable**

Numeric variable which can have infinite number of values from minus infinity to plus infinity. If a range is defined for a numeric variable e.g. salary could range from 100 USD to 100000 USD, then even between this range as well, continuous variable can have infinite number of values.

- **Categorical Variable**

Variable which has discrete categories and/or levels e.g. City Location which can have name of the city.
Categorical variable can have finite level or categories.

Mean, Median & Mode
(Measure of Central Tendency

- **Mean**
  Average of the data set, e.g. Average age of the students in a class. Mean is more often used in research, academic, sports etc.
- **Median**
  Represents Middle Value. Generally whenever average income of the country is being discussed , Median is used as a value instead of Mean, as median represents middle of a group . Mean gets impacted by very high and/or very low numbers.
- **Mode**
  Most frequently occurring value, can be applied on any data. Mean and Median can only be applied on the numeric data. E.g. eCommerce company would like to know the most frequently purchase item, they would use Mode as a measure of central tendency

Range & IQR

- **Range**

  Range in Statistics is a difference between highest and lowest value

- **IQR**

  IQR stands for Inter-Quartile-Range describes middle 50% of the data when ordered from lowest to highest. To calculate IQR, arrange data from lowest to highest, take median of the lower half of the data (Q1) and median of upper half of the data (Q3). IQR is equal to the difference between Q3 and Q1

So Range gives us the spread of the complete data whereas IQR gives us the spread of the middle half of the data.

IQR is a better measure of dispersion in comparison to Range as Range gets impacted with extreme values available either at the lower end or the upper end

Standard Deviation

- **Standard deviation (SD)** is the measure of dispersion of a data from its mean.

Formula for SD for a population is

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{n}}$$

Formula for SD for a sample is

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

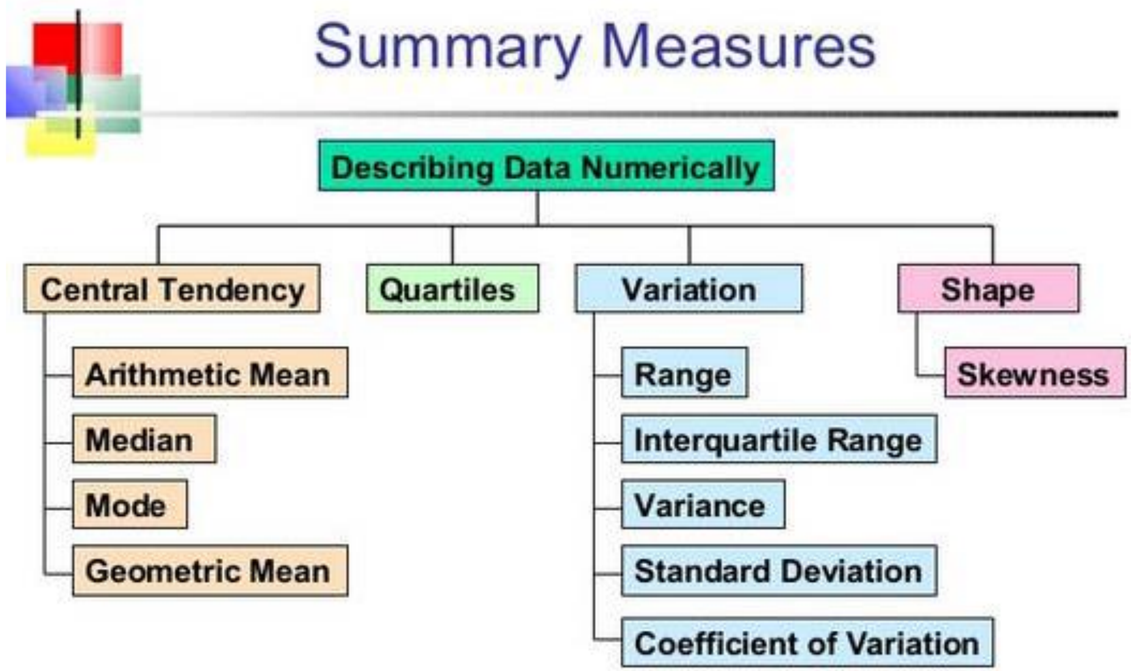
Why denominator term for Sample is n-1

Assume a population with “N” items. Suppose that we want to take samples of size “n” from that population . If we could list all possible samples of “n” items that could be selected from the population of “N” items, then we could find the SD for each possible sample.

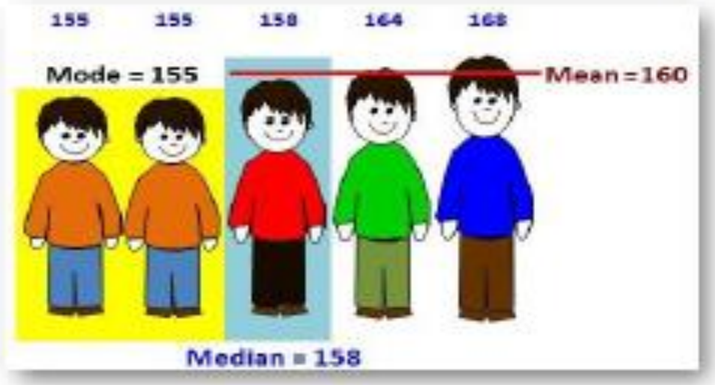
If all the sample drawn as “Unbiased” , then the average of the sample SD for all possible samples would be equal the population SD. However practically it is not possible to get all possible samples from a population.

Hence when we divide by (n −1) when calculating the sample SD , then it turns out that the average of the sample SD for all possible samples is equal the population SD. So the sample SD is what we call an unbiased estimate of the population SD. (This is also known as Bessel’s Correction)

Summarization



Measures of central tendency:



Geometric Mean = $\sqrt{ab}$

Measures of dispersion:

Range = Max – Min

IQR = Q3 – Q1

Variance : $\sigma^2 = \frac{\sum (X-\mu)^2}{N}$

Standard Deviation = σ

Coefficient of Variation = σ / μ

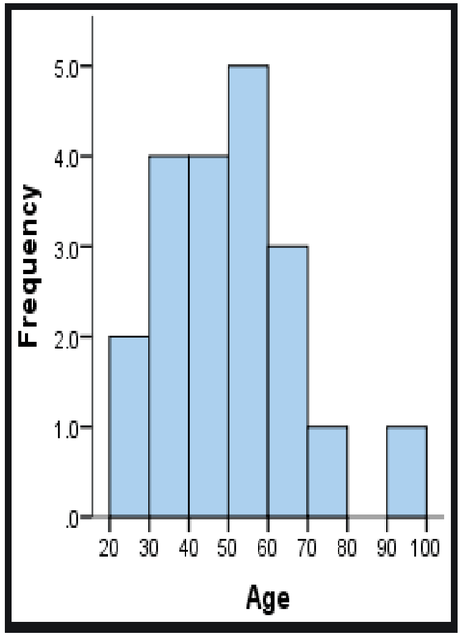
Histogram

Histogram is bar chart which represents a frequency distribution

Horizontal axis of the histogram represents the data points within an interval called as “bin” and vertical axis represents the corresponding “frequency”

How to calculate “bin” from a numeric set of data points

- Count the number of data points.
- Calculate the number of bins by taking the square root of the number of data points and round up.
- Calculate the bin width by dividing the Range (i.e. Max-Min) by the # of bins.

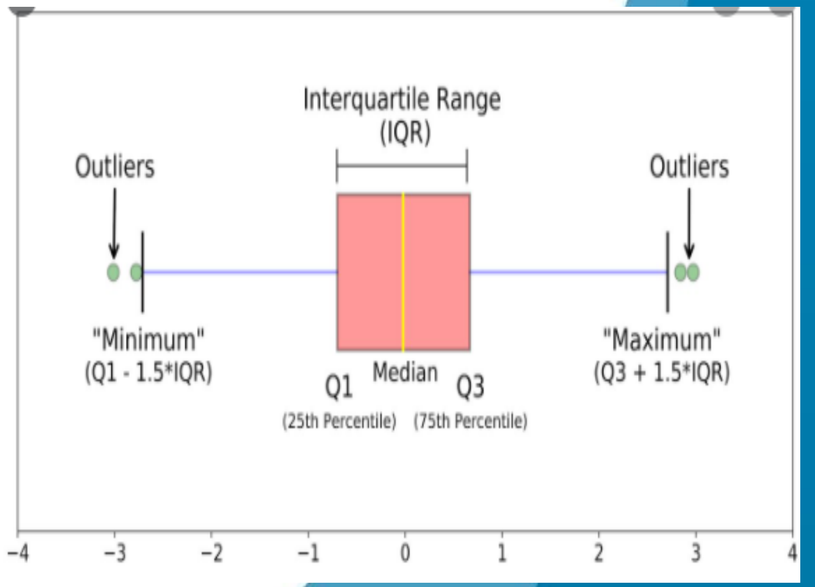


Boxplot

Box Plot or Whisker Plot displays the Five-Point summary of the data

The five-number summary is the minimum, first quartile, median, third quartile and maximum.
In a box plot, box is from the first quartile to the third quartile.

A vertical line goes through the box at the median.



Minimum value represented through a Whisker is (Q1-1.5* IQR)
Maximum value represented through a Whisker is (Q3 + 1.5* IQR)
Any point which is below the Minimum Value and/or above the maximum value is an outlier

Classical Probability

Classical Probability is a simple form of probability which measures the likelihood of an event happening, in a classic way which means that there is an equal chance of happening of every possible event

- Example of classical probability would be a fair dice roll because it is equally probable that you will land on any of the 6 numbers on the die: 1, 2, 3, 4, 5, or 6.

- Another example of classical probability would be a coin toss. There is an equal probability that your toss will yield a heads or tails result.

Formula to calculate classical probability is total number of times an event can happen (f) divided by total count of all possible outcomes (N) i.e. $P(A) = f / N$.

Probability of getting “1” if you roll an unbiased dice is 1/6 as there is only one possibility of getting “1” out of total six outcomes.

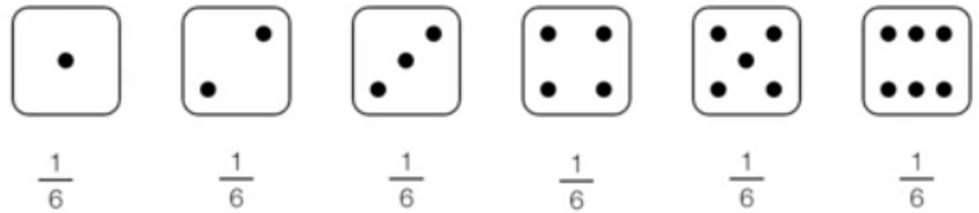
Conditional Probability

Example:

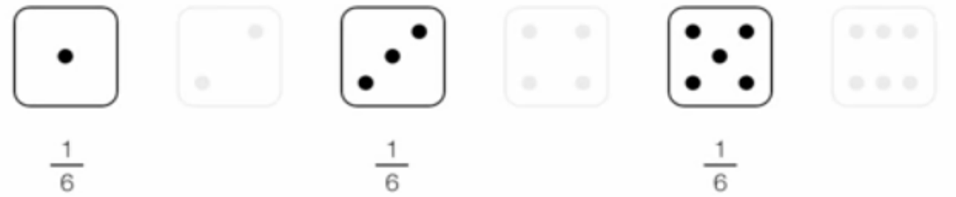
Assume you roll a biased dice, set to throw only odd numbers. What is the probability that the value is less than 4?

This situation is represented as: *Probability (Value < 4 | Value € odd)*

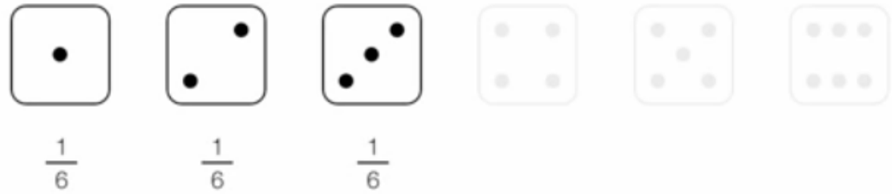
Possible outcomes on rolling a dice:



Getting an odd number (*Value € odd*)



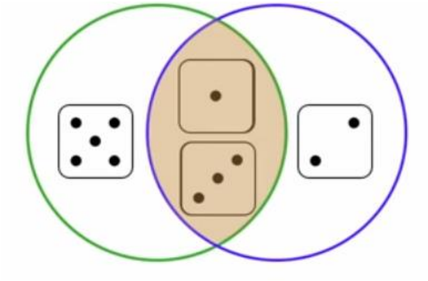
Getting a number less than 4 *(Value < 4)*



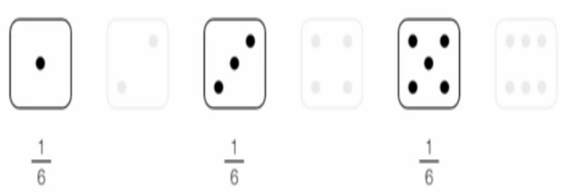
Conditional Probability continued..

Given that the first event has occurred and then the second is made to occur, the situation could be represented as:

This is given by P(*Value < 4* ∩ *Value € odd*) = 2/6 = 1/3



For this to happen, the first should necessarily have occurred in the first place:



P(*Value € odd*) = 3/6 = ½

Finally,

P (*Value < 4 | Value € odd)* = $\frac{P(Value < 4 \cap Value \in odd)}{P(Value \in odd)} = \frac{1/3}{1/2} = 2/3$

Generically, P (B | A) = $\frac{P(B \cap A)}{P(A)} = \frac{P(A \cap B)}{P(A)}$

Bayes Theorem

Bayes' Theorem (also known as the Bayes' rule) is a mathematical formula used to determine the conditional probability of events.

Bayes' theorem describes the probability of an event based on prior knowledge of the conditions that might be relevant to the event.

Formula for Bayes' theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Where:
P(A|B) – the probability of event A occurring, giv
P(B|A) – the probability of event B occurring, giv
P(A) – the probability of event A
P(B) – the probability of event B

Bayes Theorem Continued..

Example

A computer component is given scores € (A, B, C) after production.

On an average, 70% components were given a score of A, 18% were given a score of B and 12% a score of C.

It was found that 2% of the components that were given a score of A, 10% that were given a score of B and 18% that were given a score of C, eventually failed.

If you randomly pickup a failed component, what is the probability that it had received a score of B?

Assuming 10000 components:

P (A)	7000	Worked	6860
		Failed	140
P (B)	1800	Worked	1620
		Failed	180
P (C)	1200	Worked	984
		Failed	216

Probability ?
= P(B | F)

Bayes Theorem Continued..

$$P(B \mid F) = \frac{\text{Probability that quality score was B and the component failed}}{\text{Probability of any quality score and the component failed}}$$

$$P(B \mid F) = \frac{P(B \cap F)}{P(F)}$$

$$= \frac{P(B) * P(F \mid B)}{P(A) * P(F \mid A) + P(B) * P(F \mid B) + P(C) * P(F \mid C)}$$

$$= \frac{(0.18) * (0.10)}{(0.02) * (0.70) + (0.18) * (0.10) + (0.12) * (0.18)} = \frac{0.018}{0.0536} = 0.3358$$

You'll see a similar answer in the previous example:

$$P(B \mid F) = \frac{180}{140 + 180 + 216} = \frac{180}{536} = 0.3358$$

Hence, Bayes' Theorem gives the probability of an event, given an evidence event has already occurred.

Binomial Distribution

- Sequence of *n* trials / iterations.
- Each trial can have one of two possible outcomes: success or failure.
  - A single success/ failure experiment is called as a Bernoulli trial.
- Probability of success (p), remains same in each trial. Probability of failure (1-p) is also fixed in each trial.
- The trials are independent; outcome of previous trial does not influence future trials.

- *Example*:
- In an experiment involving 5 trials, random variable (discrete) represents # successes.

Binomial Distribution Contd..

#success	#failure
0	5
1	4
2	3
3	2
4	1
5	0

Trial	1	2	3	4	5
Outcomes	S	S	F	S	S
Outcomes	S	F	S	S	S
Outcomes	S	S	S	S	F
Outcomes	S	S	S	F	S
Outcomes	F	S	S	S	S

P(x) =
5 * S^4 * F

There could be 5 ways of having 4 success and 1 failure <= C (5, 4) = 5

Binomial Distribution Contd..

Formula for Binomial Distribution

$$P(x) = C(n, x)\, p^x\, (1 - p)^{n-x}$$

Example:

Jones makes an average of 10 calls per day and has a success rate of 75%. Kate makes and average of 16 calls per day but has a success rate of 45%.

What is the probability of the salespersons making 6 sales on any given day?

For Jones:

n = 10

x = 6

p = 0.75

=> P(6) = C (10,6) * $(0.75)^6$ * $(0.25)^4$ = 0.146

For Kate:

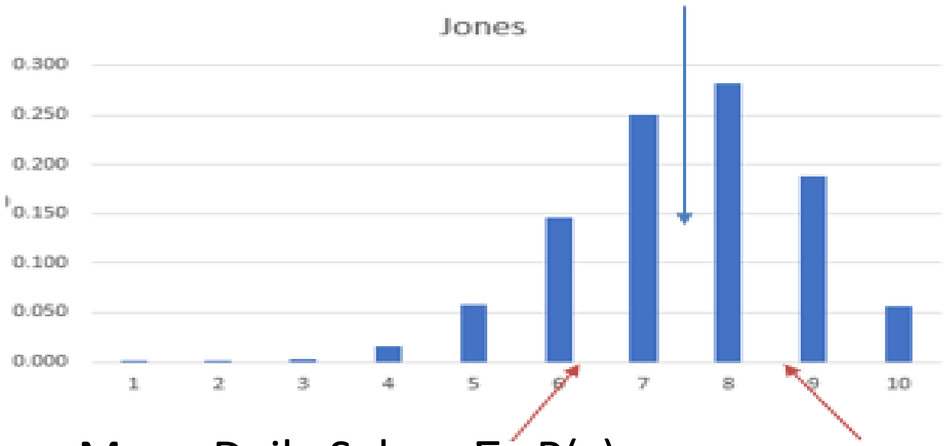
n = 16

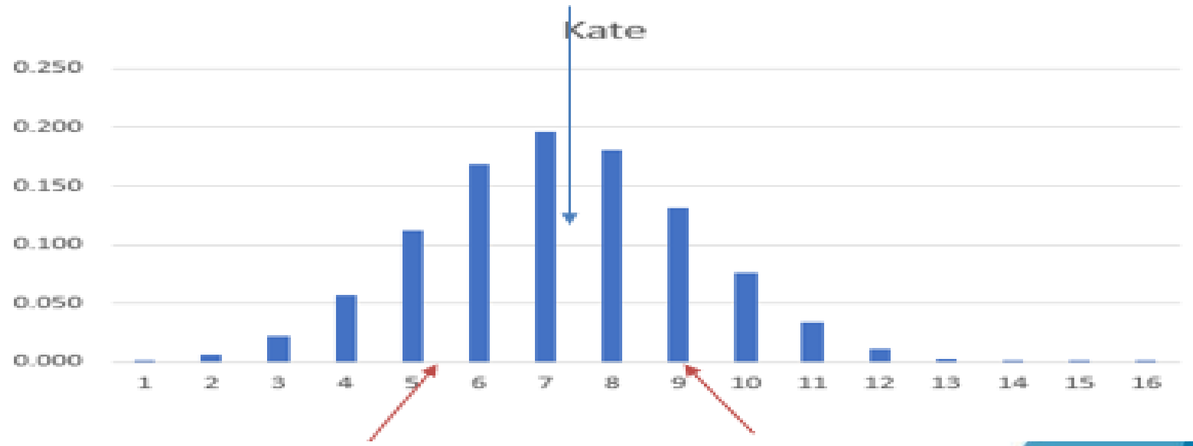
x = 6

p = 0.40

=> P(6) = C (16,6) * $(0.45)^6$ * $(0.55)^4$ = 0.168

Binomial Distribution Contd..





Mean Daily Sales : $\sum x\ P(x) = np$

Jones: $\mu = np = 10 * 0.75 = 7.5$

Kate: $\mu = np = 16 * 0.45 = 7.2$

Standard deviation in Daily Sales: $\sum (x - \mu)^2\ P(x) = npq$

Jones: $\sigma = \sqrt{(10 * 0.75 * 0.25)} = 1.369$

Margo: $\sigma = \sqrt{(16 * 0.45 * 0.55)} = 1.989$

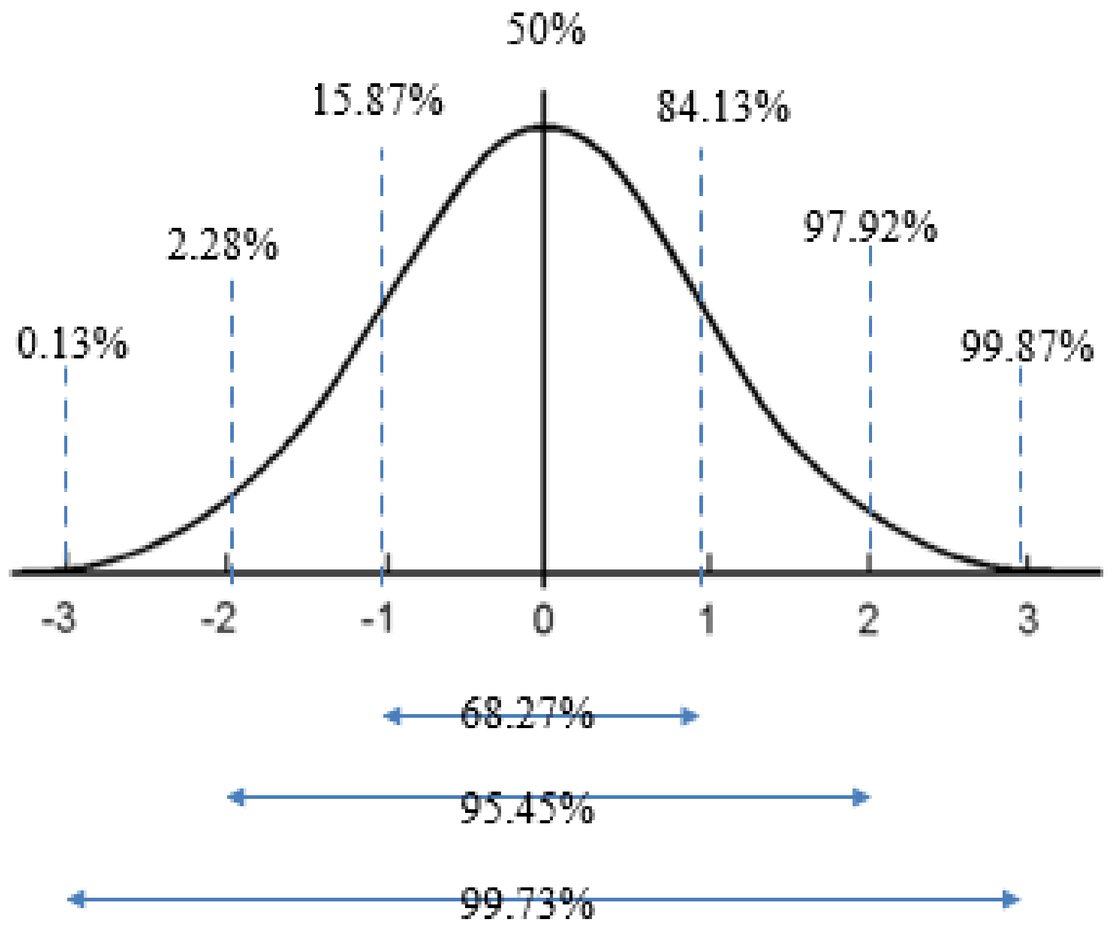
Red Arrows indicating the value μ *+1SD and* μ-1SD

n = number of trials

p = probability of success in a trial

q = probability of failure in a trial = 1-p

Normal Distribution Contd..



Standardization of the normal curve:

$\mu = 0$

$\sigma = 1$

Area under the curve = 1

Also c/a ***z-curve:***

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

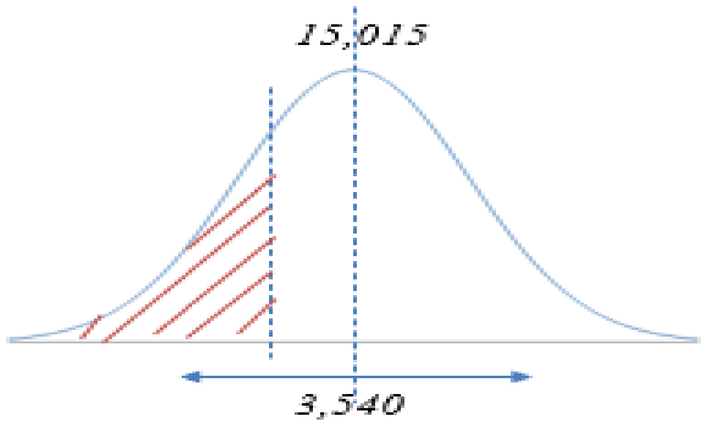
Normal Distribution Contd..

Example:

Mean debt for house buyers in a certain state of US is $15,015 with a standard deviation of $3,540. The debts are normally distributed. What is the probability that debt is less than $13,500?

NORM.DIST(13500,15015,3540,TRUE)

0.334338



Using standard normal:

z = (13500 - 15015) / 3540 = -0.428

NORM.S.DIST(-0.428,TRUE)

0.334326

Case Study- Cardio Good Fitness

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by Cardio Good Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a Cardio Good Fitness retail store during the prior three months. The data are stored in the CardioGoodFitness.csv file.

The team identifies the following customer variables to study:

- product purchased, TM195, TM498, or TM798;
- gender;
- age, in years;
- education, in years;
- relationship status, single or partnered;
- annual household income ;
- average number of times the customer plans to use the treadmill each week;
- average number of miles the customer expects to walk/run each week;
- self-rated fitness on an 1-to-5 scale, where 1 is poor shape and 5 is excellent shape.

ANY QUESTIONS

HAPPY LEARNING