Course Name :Basic Statistics using GUI-R (RKWard)
Module : Measures of Central Tendency
Week 2 Lecture : 1

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Last Class

- Plots, Data Processing

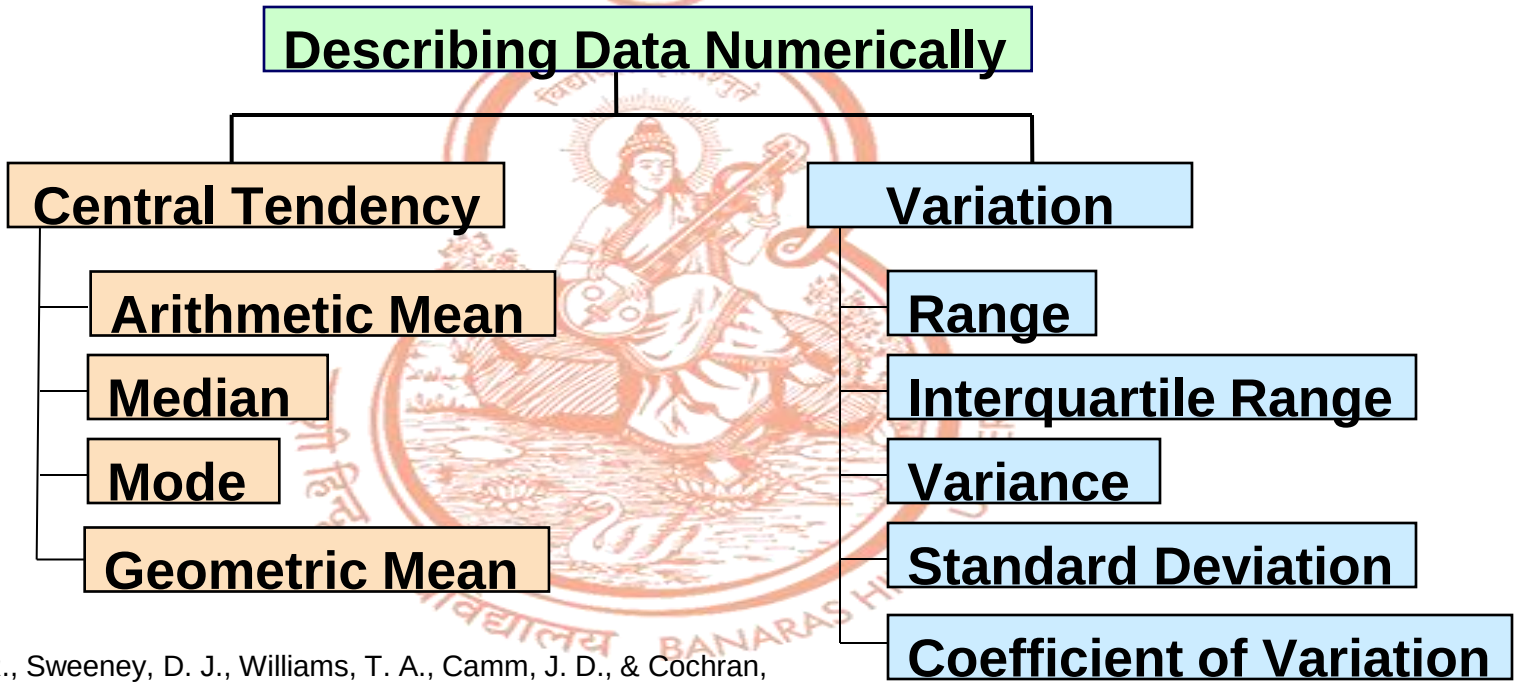
AGENDA

- Measure of Central Tendency

Why descriptive statistics is important to managers?

- Managers also need to become acquainted with numerical descriptive measures that provide very brief and easy-to understand summaries of a data collection
- There are two broad categories into which these measures fail: measures of central tendency and measures of variability

Describing Data Numerically



Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J. D., & Cochran, J. J. (2016). *Statistics for business & economics*. Cengage Learning

Requisites Of A Good Measure Of Central Tendency

- It should be rigidly defined.
- It should be easy to understand and calculate even for a non-mathematical person.
- It should be based on all the observations.
- It should be representative of the data.
- It should have sampling stability.
- It should not be affected much by extreme values.
- It should be suitable for further mathematical treatment.

Measures of Central Tendency

Overview

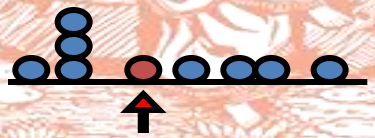
Central Tendency

Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

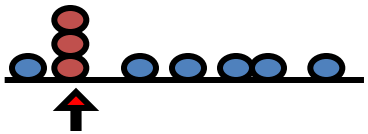
Arithmetic average

Median



Midpoint of ranked values

Mode



Most frequently observed value

Characteristics of the arithmetic mean

l. Every data set measured on an interval or ratio level has a mean.

2. The mean has valuable mathematical properties that make it convenient to use in further computations.

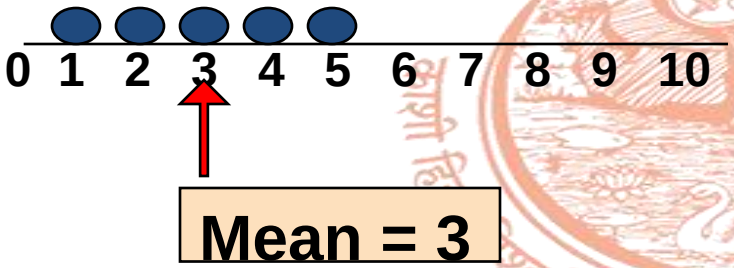
3. The mean is sensitive to extreme values.

4. The sum of the deviations of the numbers in a data set from the mean is zero.

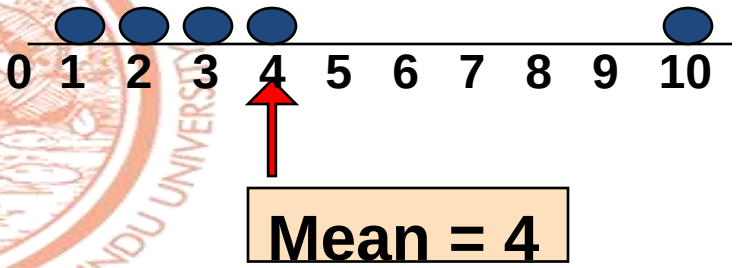
Arithmetic Mean

(continued)

- The most common measure of central tendency
- Mean = sum of values divided by the number of values
- Affected by extreme values (outliers) !!!



$$\frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$



$$\frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Median

- The numerical value in the middle when data set is arranged in order (50% above, 50% below)

- Not affected by extreme values

Finding the Median

- The location of the median:

$$\text{Median position} = \frac{n+1}{2} \text{ position in the ordered data}$$

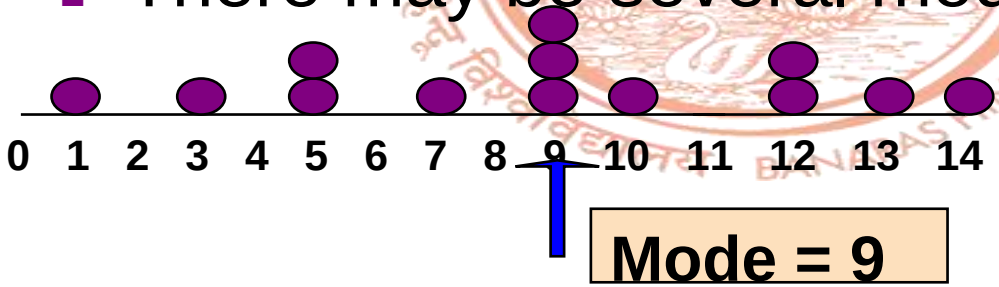
  - If the number of values is odd, the median is the middle number
  - If the number of values is even, the median is the average of the two middle numbers

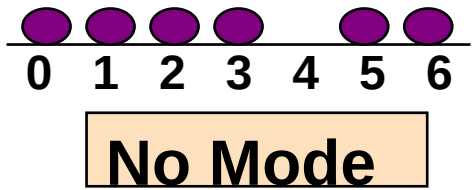
- Note that $\frac{n+1}{2}$ is not the *value* of the median, only the *position* of the median in the ranked data

The sum of the absolute deviations about the median is the minimum

Mode

- A measure of central tendency
- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical data
- There may be no mode
- There may be several modes





Which measure of location is the “best”?

- **Mean** is generally used, unless extreme values (outliers) exist
- Then **median** is often used, since the median is not sensitive to extreme values.
  - Example: Median home prices may be reported for a region – less sensitive to outliers

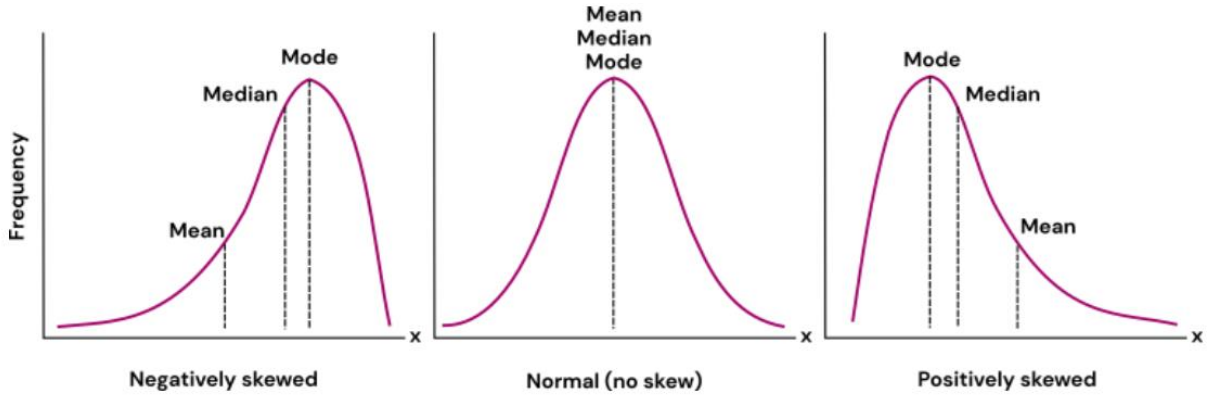
Comparison Of The Mean, Median And Mode

- ✓ Mean is an appropriate measure of the central location for interval and ratio variables.
- ✓ The median is an ordinal statistic. Its calculation is based on the ordinal properties of the data.
- ✓ Mode is a nominal statistic. Its calculation depends merely on frequency of occurrence of a particular value of the variable.
- ✓ In a symmetrical frequency distribution, the mean, median and mode coincide.

IIM Ahmedabad Placement Report: 2022 Highlights

Particulars	PGP Placement Statistics (2022)	PGP-FABM Placement Statistics (2022)	PGPX Placement Statistics (2023)
Highest package	INR 61.48 LPA	INR 33.83 LPA	INR 1.08 crore per annum
Average package	INR 32.79 LPA	INR 21.49 LPA	INR 35.68 LPA
Median package	INR 31.49 LPA	INR 20 LPA	INR 33.05 LPA
Lowest package	INR 17.50 LPA	INR 12.85 LPA	INR 18 LPA

https://timesofindia.indiatimes.com/education/news/iim-ahmedabad-placement-last-years-top-recruiters-highest-package-more/articleshow/106173746.cms


Frequency
Mode
Median
Mean
x
Negatively skewed
Mean
Median
Mode
x
Normal (no skew)
Mode
Median
Mean
x
Positively skewed

Next

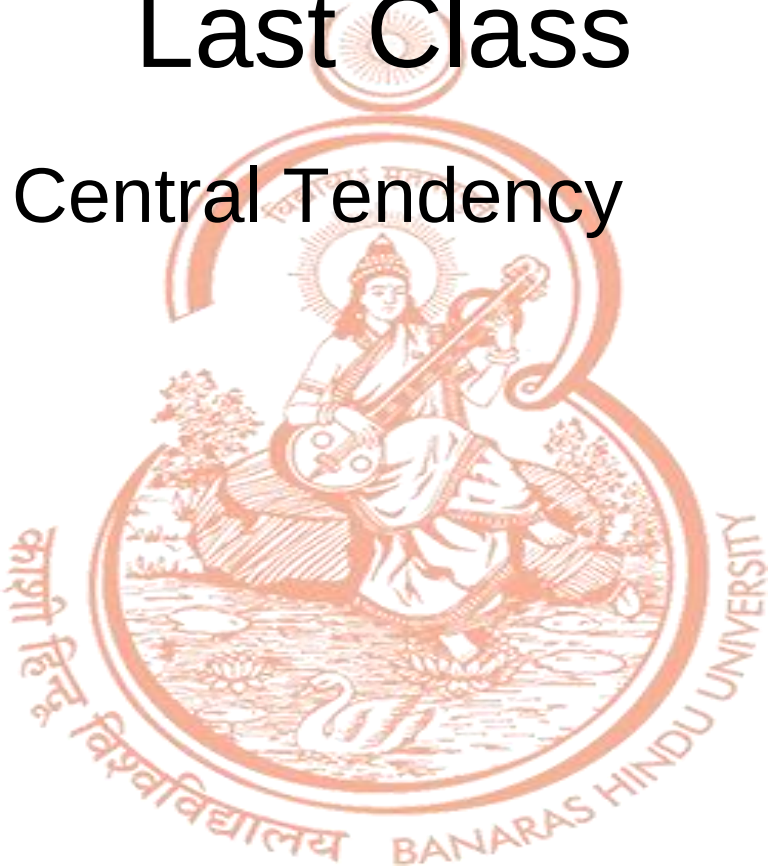
Measure of Variability

Course Name :Basic Statistics using GUI-R (RKWard)
Module : Measures of Variability
Week 2 Lecture : 2

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Last Class

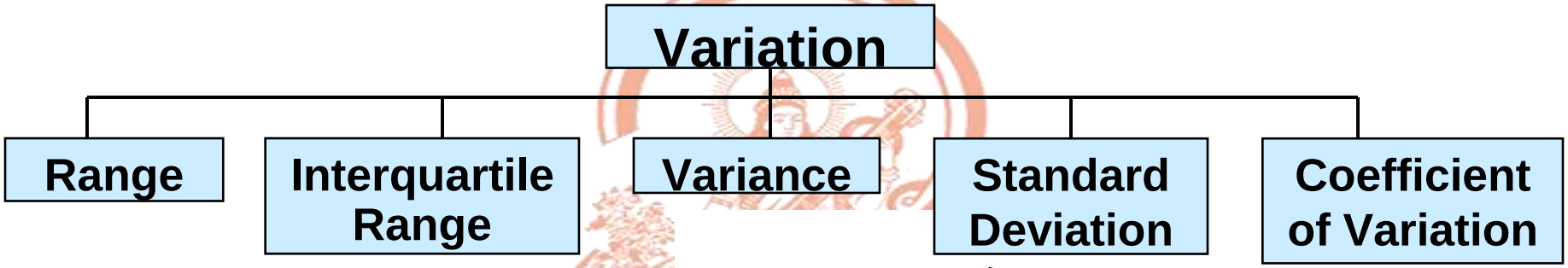
- Measure of Central Tendency



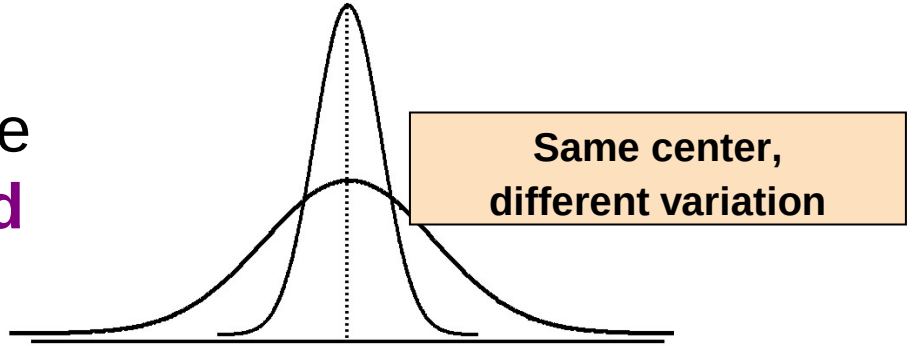
AGENDA

- Measure of Variability

Measures of Variability



- Measures of variation give information on the **spread** or **variability** of the data values.

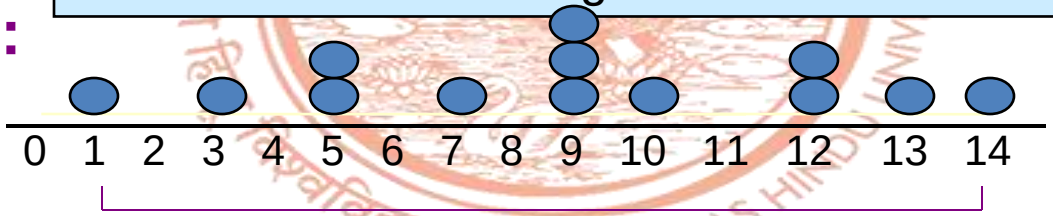


Range

- Simplest measure of variation
- Difference between the largest and the smallest observations:

$$\text{Range} = X_{\text{largest}} - X_{\text{smallest}}$$

Example:



Range = 14 - 1 = 13

Disadvantages of the Range

- Ignores the way in which data are distributed

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,5

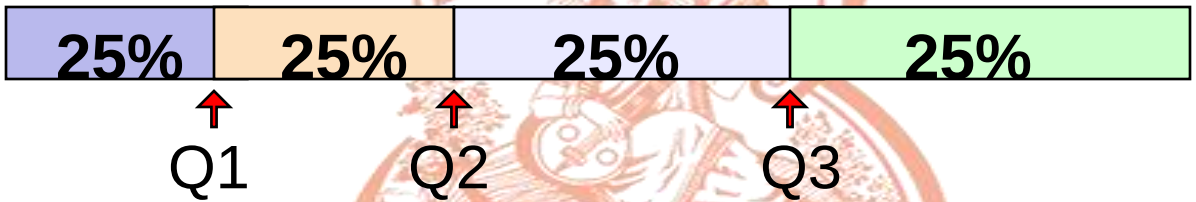
Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,120

Range = 120 - 1 = 119

Quartiles

- Quartiles split the ranked data into 4 segments with an equal number of values per segment



- The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger
- Q_2 is the same as the median (50% are smaller, 50% are larger)
- Only 25% of the observations are greater than the third quartile

Quartile Formulas

Find a quartile by determining the value in the appropriate position in the ranked data, where

First quartile position: $\mathbf{Q_1 = 0.25(n+1)}$

Second quartile position: $\mathbf{Q_2 = 0.50(n+1)}$

(the median position)

Third quartile position: $\mathbf{Q_3 = 0.75(n+1)}$

where **n** is the number of observed values

Quartiles

- Example: Find the first quartile

Sample Ranked Data: 11 12 13 16 16 17 18 21 22

(n = 9)

Q_1 = is in the 0.25(9+1) = 2.5 position of the ranked data

so use the value half way between the 2nd and 3rd values,

so **Q_1 = 12.5**

Comparing Measures Of Variability

The relative advantages and disadvantages of the four measures of variability are discussed with reference to the following factors that affect variability:

- Extreme scores
- Sample size
- Stability under sampling
- Open-ended distribution

Moments About The Mean/Origin

Moment	What it measures
(i) First moment about the origin	Mean
(ii) First moment about the mean	Average Deviation
(iii) Second moment about the mean	Variance
(iv) Third moment about the mean	Skewness
(v) Fourth moment about the mean	Kurtosis

Interquartile Range

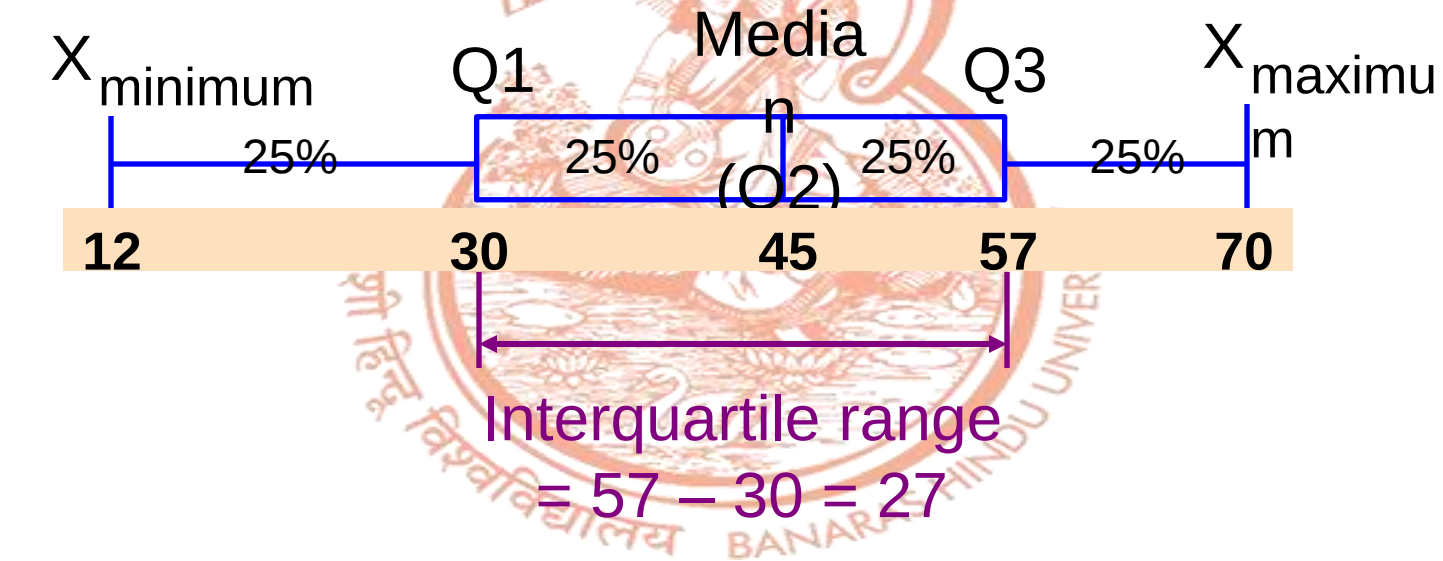
- Can eliminate some outlier problems by using the **interquartile range**
- Eliminate high- and low-valued observations and calculate the range of the middle 50% of the data
- Interquartile range = 3rd quartile – 1st quartile

$$IQR = Q_3 - Q_1$$

Interquartile Range

Five number summary –Box plot

Example:



Population Variance

- Average of squared deviations of values from the mean
  - Population variance:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \text{ (simple) or } \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2 \cdot n_i}{N} \text{ (weighted)}$$

Where μ = population mean

N = population size

x_i = i^{th} value of the variable x

Sample Variance

- Average (approximately) of squared deviations of values from the mean
  - Sample variance:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \text{ (simple) or } s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{n-1} \text{ (weighted)}$$

$\bar{x}$ = arithmetic mean

n = sample size

x_i = i^{th} value of the variable x

n_i = absolute frequency

Population Standard Deviation

- The square root of population variance
- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data

Sample Standard Deviation

- The square root of the sample variance
- Most commonly used measure of variation
- Shows variation about the mean
- Has the same units as the original data
  - Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \text{ (simple) or } s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot n_i}{n-1}} \text{ (weighted)}$$

Calculation Example: Sample Standard Deviation

Sample Data (x_i) : 10 12 14 15 17 18 18 24

n = 8 Mean = $\bar{x}$ = 16

$$s = \sqrt{\frac{(10-\bar{X})^2 + (12-\bar{x})^2 + (14-\bar{x})^2 + \square + (24-\bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \square + (24-16)^2}{8-1}}$$

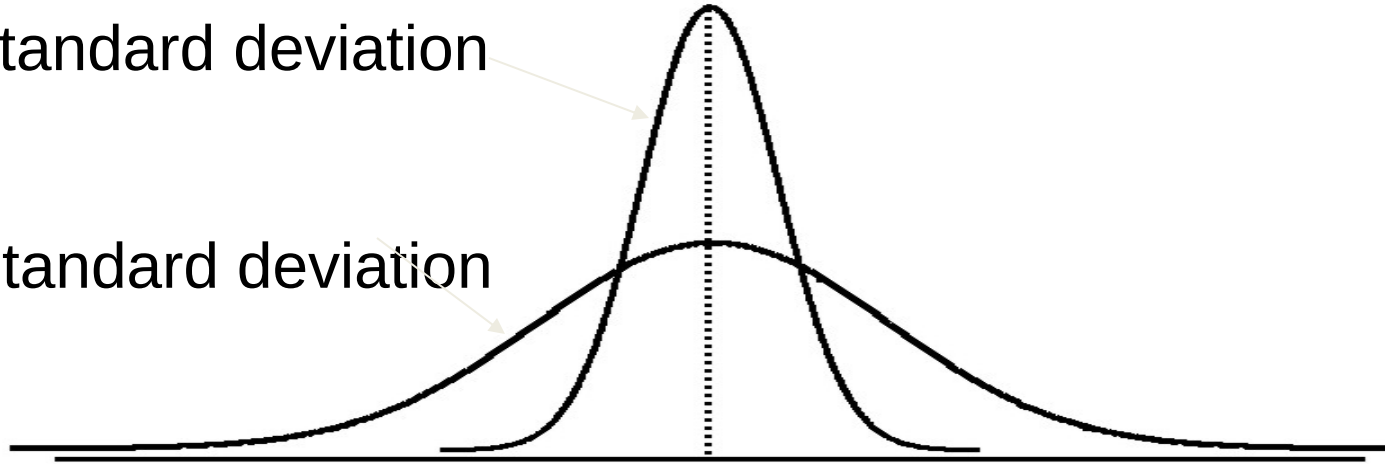
$$= \sqrt{\frac{126}{7}} = 4.2426$$

→ A measure of the "average" scatter around the mean

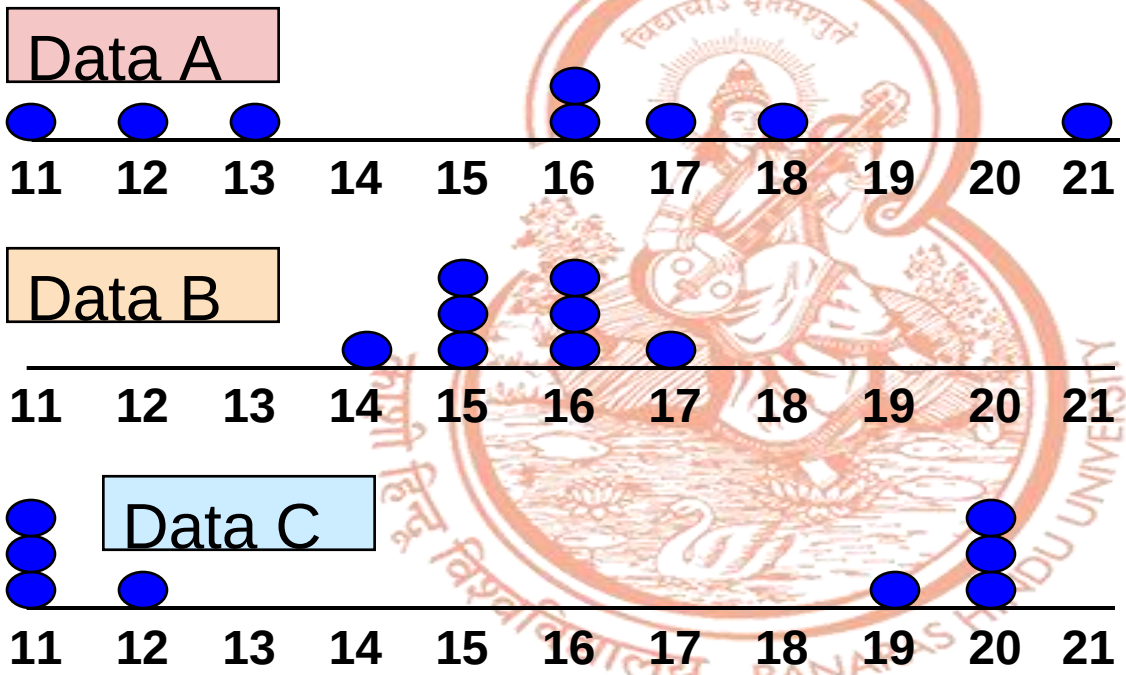
Measuring variation

Small standard deviation

Large standard deviation



Comparing Standard Deviations



Mean = 15.5
s = 3.338

Mean = 15.5
s = 0.926

Mean = 15.5
s = 4.570

Advantages of Variance and Standard Deviation

- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight

  (because deviations from the mean are squared)

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Can be used to compare two or more sets of data measured in different units

$$CV = \left(\frac{s}{\bar{x}} \right) \cdot 100\%$$

Comparing Coefficient of Variation

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$cv_A = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$cv_B = \left(\frac{s}{\bar{x}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Next

Z-Score. Standard data

References

B1. Mohanty, B., & Misra, S. (2016). Statistics for behavioural and social sciences.

B2. Pandya, K., Joshi, P., Bulsari, S., & Nachane, D. M. (2018). Statistical analysis in simple steps using R

B3. Field, A. P., Miles, J., & Field, Z. (2012). Discovering statistics using R

B4. Harris, J. K. (2019). Statistics with R: solving problems using real-world data. SAGE Publications.

Course Name :Basic Statistics using GUI-R (RKWard)
Module : Introduction to Probability For Statistics
Week 2 Lecture : 4

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Probability

Experiment Any process of observation or measurement.

Outcome The result one obtains from an experiment.

Sample Space ***(Ω)*** The set of all possible outcomes of an experiment.

Event A subset of the sample space. A collection of desirable outcomes.

Probability

- The objective of Probability is to assign to each event A a number P(A), called the probability of the event A, which will give a precise measure of the chance that A will occur.

Equally likely outcomes

- If all outcomes in a finite set Ω are equally likely, the probability of A is the number of outcomes in A divided by the total number of outcomes.

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

Probability Postulates

1. $P(B) \geq 0$

2. If $B_1, B_2, \ldots, B_n$ is a partition of B, then $P(B) = P(B_1) + P(B_2) + \ldots + P(B_n)$

3. $P(\Omega) = 1$

Postulates of Probability

- Complement Rule: $P(A^c) = 1 - P(A)$
- If A is a subset of B, then $P(A) \leq P(B)$
- Difference Rule: If A is a subset of B, $P(B \cap A^c) = P(B) - P(A)$
- Inclusion – Exclusion :

  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Equally likely outcomes

- If all outcomes in a finite set are equally likely, computing probabilities reduces to counting.
- The probability of A is the number of outcomes in A divided by the total number of outcomes.

$$P(A) = \frac{\#(A)}{\#(\Omega)}$$

Example

A fair die is rolled and the number on the top face is noted. Then another fair die is rolled, and the number on its top face is noted.

- What is the probability that the sum of the two numbers showing is 5?
- What is the probability that the second number rolled is greater than the first number?

Methods of Assigning Probabilities

- Frequency Interpretation
- Subjective Interpretation

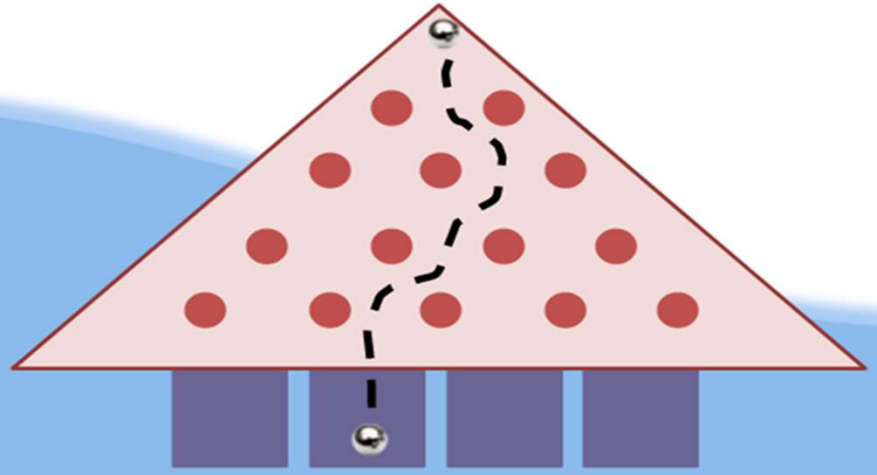
Course Name :Basic Statistics using GUI-R (RKWard)
Module : Introduction to Normal Distribution
Week 2 Lecture : 5

Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Agenda

- Distributions: Bernoulli, Binomial, Normal, X2, t, and F
- Relationship
- Characteristics (mean, variance, range, and symmetry)

The Galton box consists of pegs arranged in a triangular pattern.



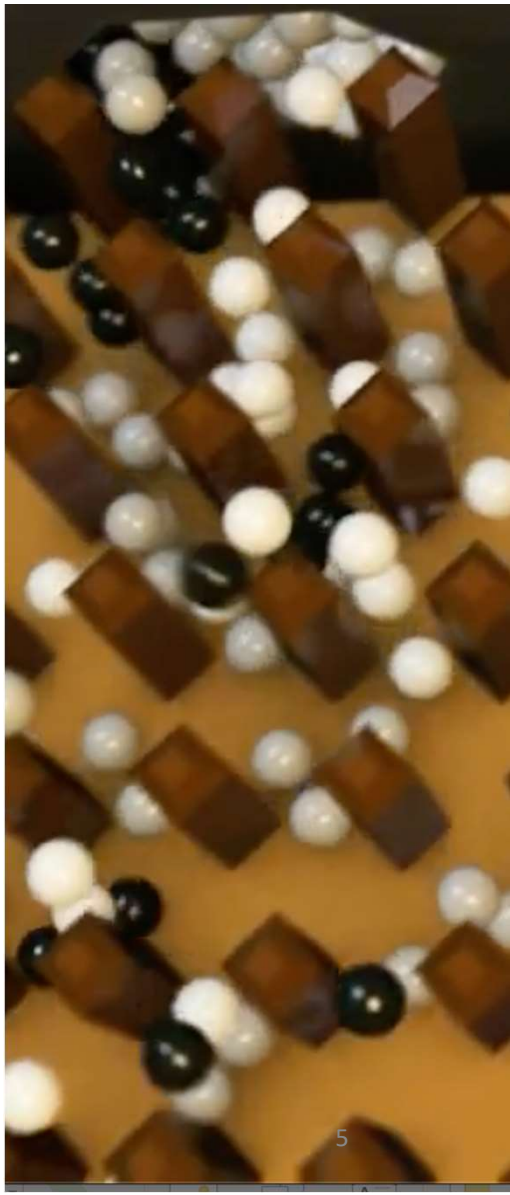
Balls (or beans) are dropped from the top of the board, bounce among the pegs, and collect in

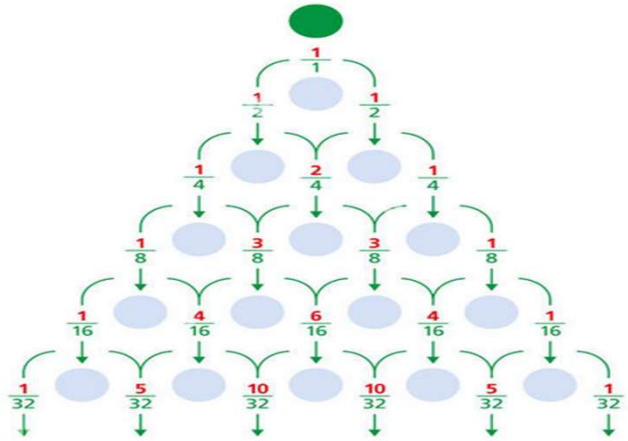
Bernoulli distribution

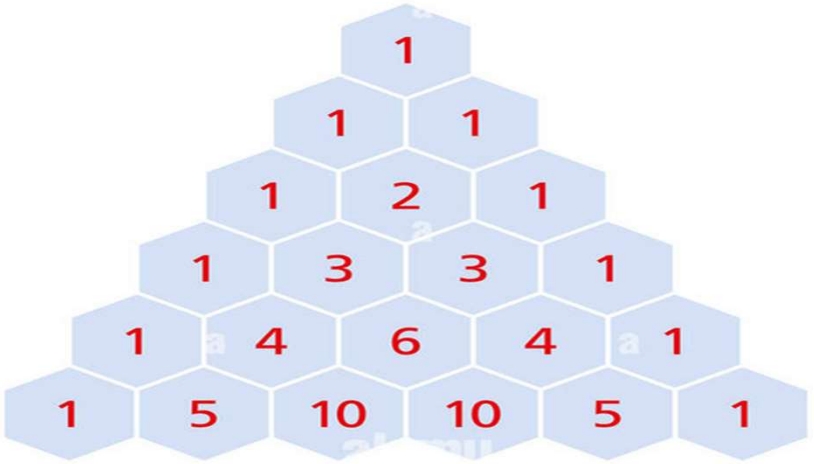
- Named after Swiss mathematician Jacob Bernoulli
- A single event (Bernoulli trial) has successful rate of p.
- Two outcomes: 1 (success) and 0 (fail)
- P(1)=p and P(0)=1-p
- Mean: p – p*1+ (1-p)*0
- Variance: p(1-p)
- Notation: b(p)

Binomial distribution

- **Multiple (n) Bernoulli trials** each has success rate of p.
- The total number of success is a random variable, x
- Range from zero to n.
- The probability is $c(n, x)p^x(1-p)^{(n-x)}$,where c(n, x) is number of combinations of choosing x from n.
- Notation: B(n, p)







Each row represents a Bernouli trial

Binomial distribution

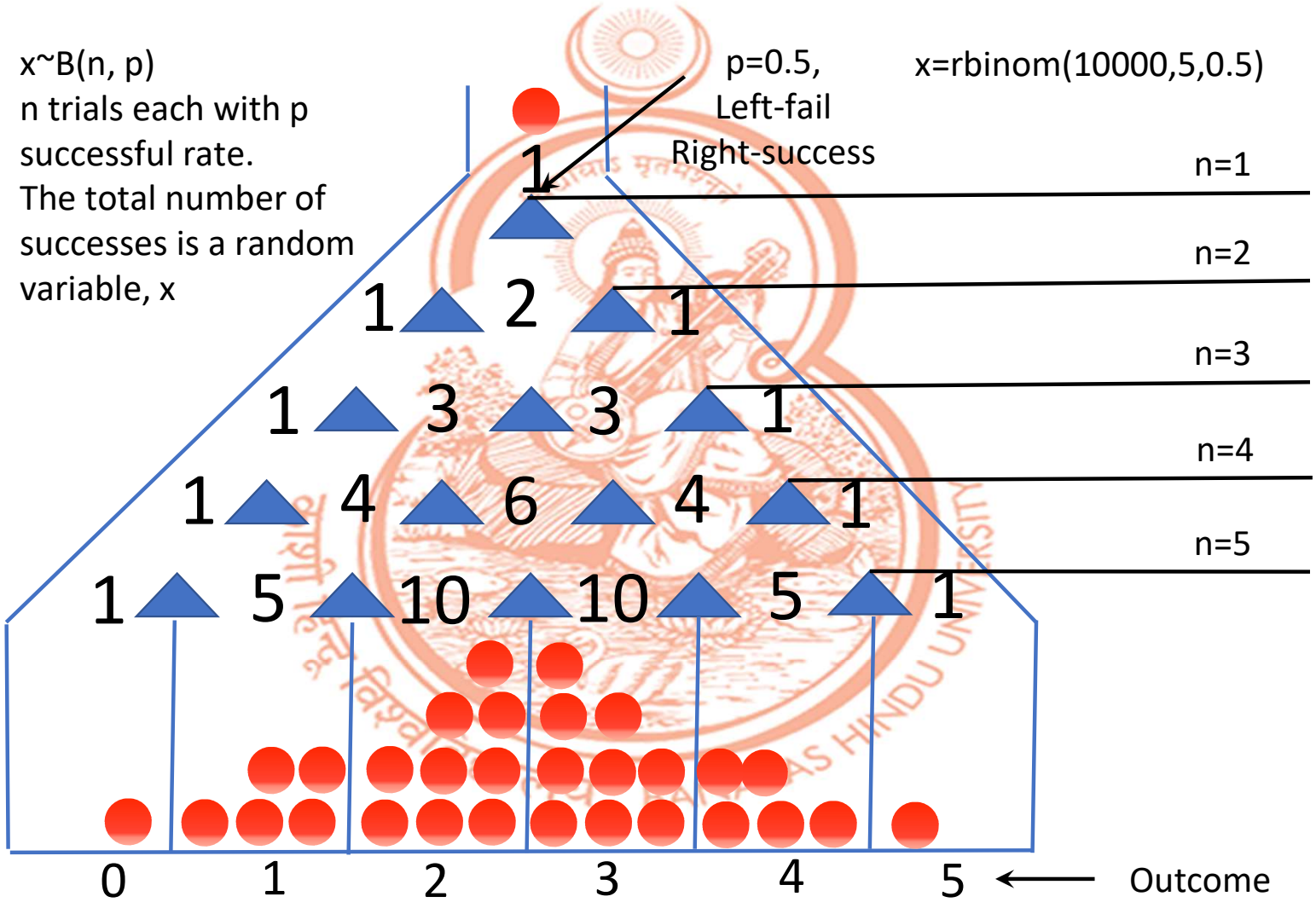
- Mean=np
- Variance=np(1-p)
- p>=0
- Symmetric only if p=.5
- When n is large, binomial is close to normal distribution
- Probability of r success from n is nCr (p^r)((1-p)^(n-r))

Binomial distribution and Galton board

x~B(n, p)
n trials each with p successful rate.
The total number of successes is a random variable, x

p=0.5,
Left-fail
Right-success

x=rbinom(10000,5,0.5)



Course Name :Basic Statistics using GUI-R (RKWard)
Module : Introduction to Normal Distribution
Week 2 Lecture : 6

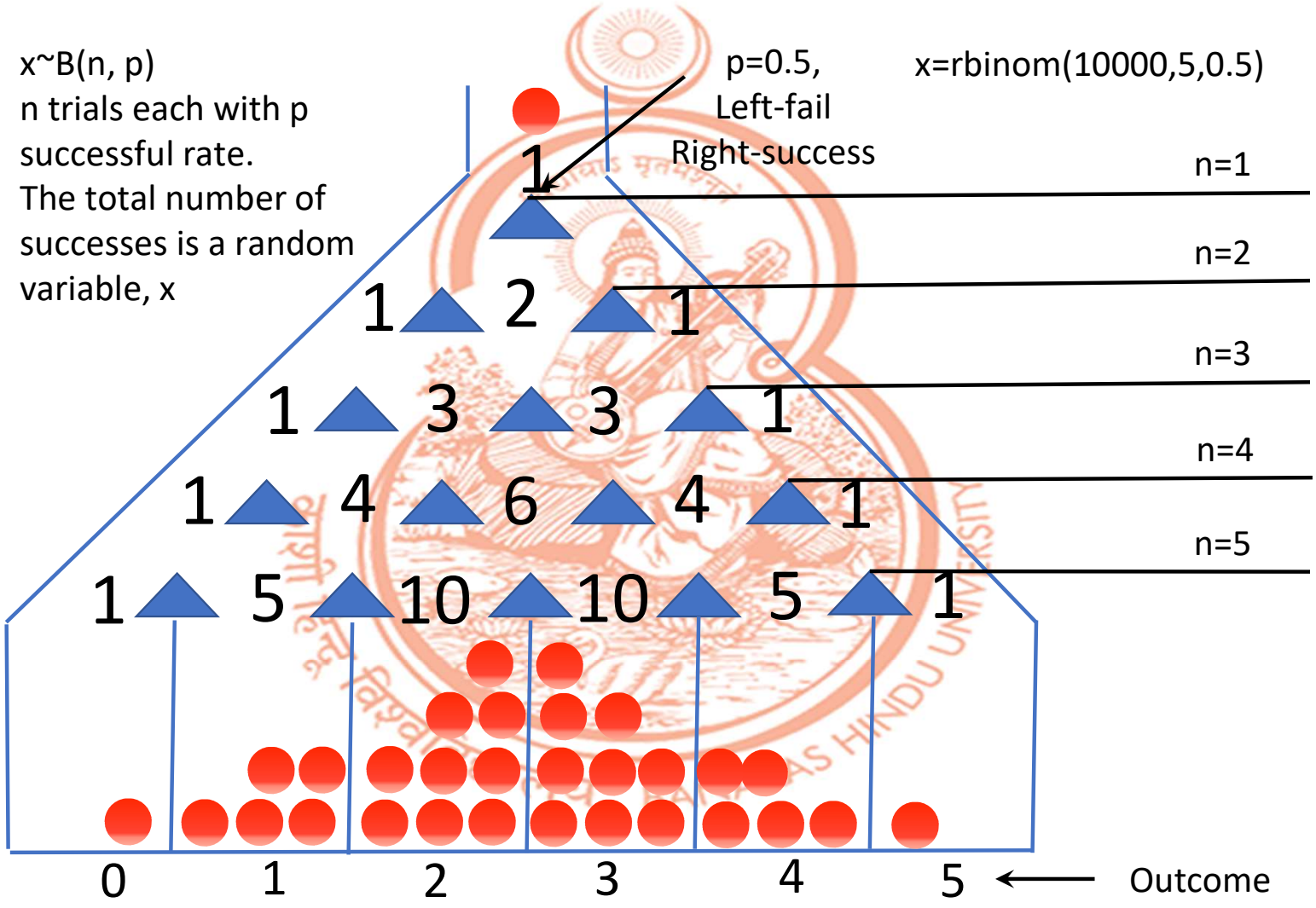
Harsh Pradhan, Assistant Professor,
Institute of Management Studies, BHU
https://bhu.ac.in/Site/FacultyProfile/1_5?FA000562

Binomial distribution and Galton board

x~B(n, p)
n trials each with p successful rate.
The total number of successes is a random variable, x

p=0.5,
Left-fail
Right-success

x=rbinom(10000,5,0.5)



Normal distribution

- Binomial distribution with large n
- Bell shape
- Probability Density Function: Exponential

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Notation: N(mean, variance)
- -infinity to +infinity
- symmetric