

Probability Distribution and Hypothesis Testing

Satishkumar L. Varma

Professor

Department of Computer Engineering,

PCE New Panvel - 410206

Theoretical

 Binomial, poisson, normal, exponential, hyper geometric, uniform distributions

Type I and II error

 Tests for equality of mean and variances of two populations

 Confidence interval

 Z test and 2 test for goodness of fit

 ANOVA (one way classification)

 Non parametric tests

 Sign test

 U test

LEARNING OBJECTIVES

After reading this chapter, you should be able to:

Understand the idea of probability distribution used in concept in statistics

How Joint and conditional probabilities are used to analyze corpus data

How probability plays an important role in statistical hypothesis testing



Probability

🌟 Two reasons why probability is important for the analysis of linguistic data:

🌟 Joint and conditional probabilities are used to analyze corpus data

🌟 Probability plays an important role in statistical hypothesis testing

🌟 Simple probability

🌟 Eg: If you toss a dice with six number (i.e. 1,2,3,4,5,6) what is the probability that you will toss a 6?

🌟 Ans: $P(6) = 1/6 = 0.1666$

🌟 Values range from 0 to 1 and total probabilities of the sample is 1

🌟 If two events are independent, the probability is the sum of their individual probabilities

🌟 Two events A and B are independent if knowing that the occurrence of A does not change the probability of the occurrence of B

Statistical hypothesis testing

🌟 Joint probability

$$\text{🌟 } P(A,B) = P(A) \times P(B)$$

$$\text{🌟 Example: } P(5,6) = (0.166) \times (0.166) = 0.0277$$

🌟 Conditional probability

$$\text{🌟 } P(A | B) = P(A \wedge B) / P(B)$$

🌟 Eg: A corpus including 2000 nouns & 500 adjectives, 50 adjectives precede a noun

🌟 What is the likelihood that a noun occurs after an adjective?

🌟 What is the likelihood that an adjective precedes a noun?

$$\text{🌟 } P(\text{ADJ} | \text{N}) = P(\text{ADJ} \wedge \text{N}) / P(\text{N})$$

$$\text{🌟 } P(\text{ADJ} | \text{N}) = P(50) / P(2000) = 0.025$$

$$\text{🌟 } P(\text{N} | \text{ADJ}) = P(50) / P(500) = 0.1$$

Probability distribution

What is the probability that you get two heads if you toss a coin twice?

0 heads = HH 25%

1 head = HT + TH 50%

2 heads = TT 25%

Sample space

Random variable

Cumulative outcome	Probability
0 = 1×	0.25
1 = 2×	0.50
2 = 1×	0.25
	$\Sigma P(x) = 1$

Binomial Distribution

🎯 Bernoulli trail:

- 🎯 two possible outcomes on each trail
- 🎯 the outcomes are independent of each other
- 🎯 the probability ratio is constant across trails

🎯 The binomial distribution has the following properties:

- 🎯 It is based on categorical / nominal data.
- 🎯 There are exactly two outcomes for each trail.
- 🎯 All trials are independent.
- 🎯 The probability of the outcomes is the same for each trail.
- 🎯 A sequence of Bernoulli trails gives us the binomial distribution.

Binomial Distribution

🌸 Example: A coin is tossed three times. What is the probability of obtaining two heads?

Sample space: HHH TTT
 HHT TTH
 HTH THT
 THH HTT

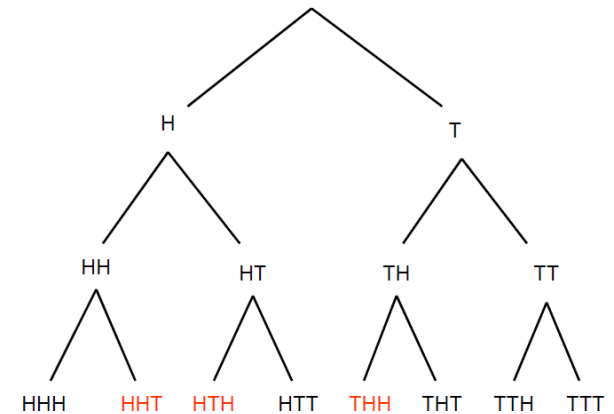
Random variables: 0 Head
 1 Head
 2 Heads
 3 Heads

0 head: 1 / 8 = 0.125

1 head: 3 / 8 = 0.375

2 heads: 3 / 8 = 0.375

3 heads: 1 / 8 = 0.125



Binomial Distribution

🌟 Example: If you toss a coin 8 times what is the probability of obtaining a score of:

🌟 0 heads

🌟 1 head

🌟 2 heads

🌟 3 heads

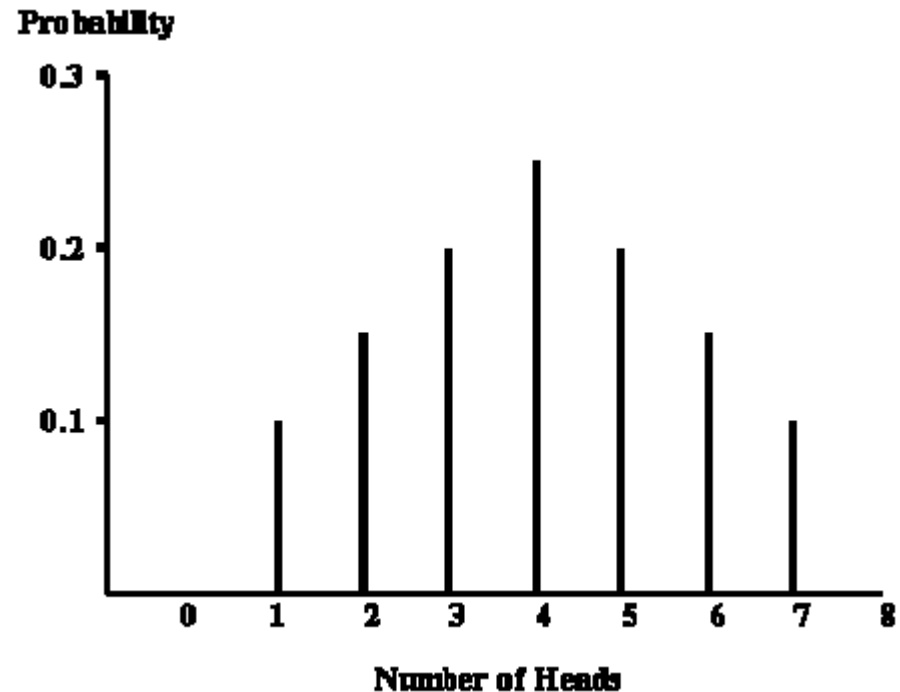
🌟 4 heads

🌟 5 heads

🌟 6 heads

🌟 7 heads

🌟 8 heads

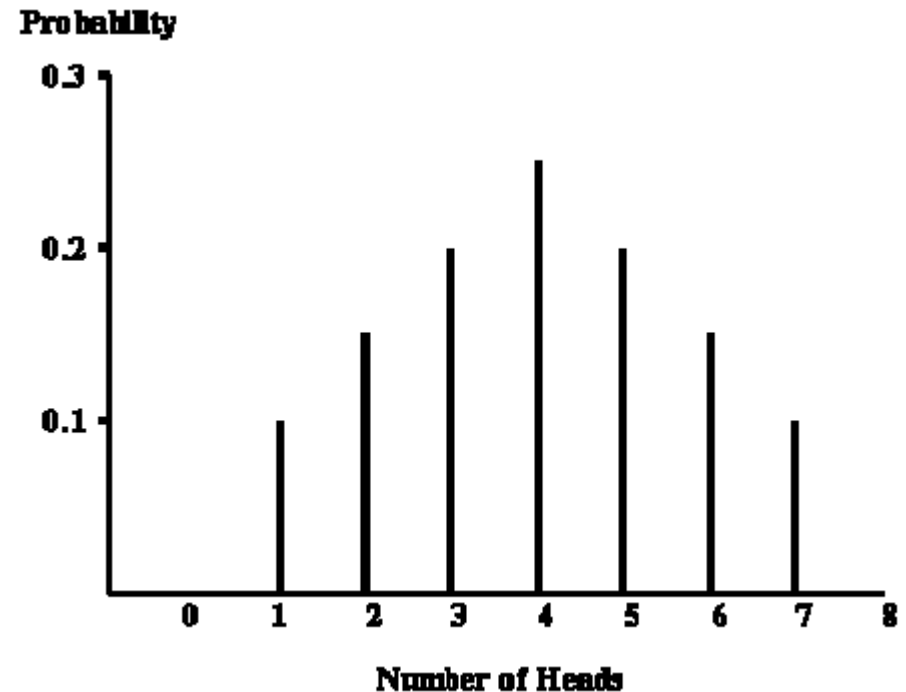


Binomial Distribution

Example: Tossing a coin one 100 times, yielded 42 heads and 58 tails. Is this a fair coin?

Heads: 42

Tails: 58



Binomial Distribution

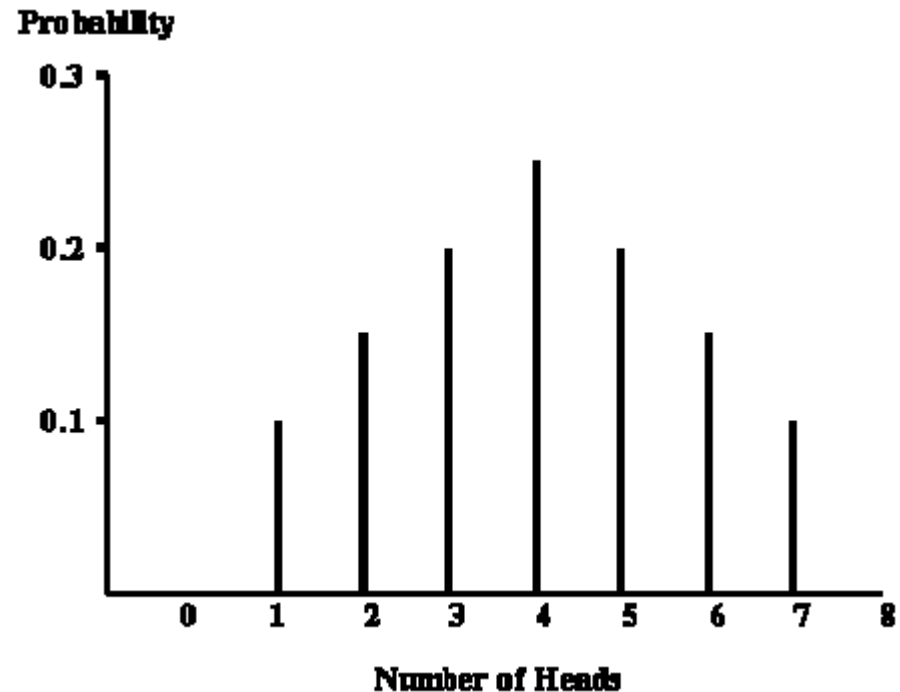
Example: Tossing a coin one 100 times, yielded 42 heads and 58 tails. Is this a fair coin?

Heads: 42

Tails: 58

Expected: 50% - 50%

Sample error?



Normal Distribution



Statistical hypothesis testing



Population (Probability) Distributions

🌱 One of the most important concepts in statistics is the idea of a probability distribution

🌱 Discrete probability distribution

- 🌱 This is the binomial probability distribution
- 🌱 Which we can plot as a histogram
- 🌱 Prominent in statistics but not much in the geosciences

🌱 Discrete probability distribution

- 🌱 Most common continuous distribution by far is the normal or Gaussian distribution
- 🌱 Normal distribution is the well-known bell shaped distribution

🌱 Feature of Normal Distribution

- 🌱 It is analytical
- 🌱 It often describes populations quite well
- 🌱 It works quite well if the populations has a lot of outliers
- 🌱 Central limit theorem: whatever the probability distribution of x , the probability distribution of means of x for repeated samples of n random samples tends to become normally distributed as n increases with

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

Population (Probability) Distributions

🏠 Measures of central tendency and dispersion use two key terms

🏠 Population

🏠 Sample

⚠ Remember:

🏠 Information is used to make inferences about the **population** from which the **sample** was drawn, and

🏠 Characteristics of the **population** are referred to as **parameters**,

🏠 while characteristics of the **sample** are referred to as **statistics**.

Parameters describing distributions

 Common parameters and their definitions as expected values

Name	Definition	Symbol
mean	$E[X]$	μ
variance	$E[(X - \mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X - \mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X - \mu)^4]/\sigma^4 - 3$	γ_2

Population (probability) distributions

✳ Population (probability) distributions of 5 different continuous random variables.

✳ **Distribution A** is a unimodal (one peak) symmetric distribution, centered around 2.0.

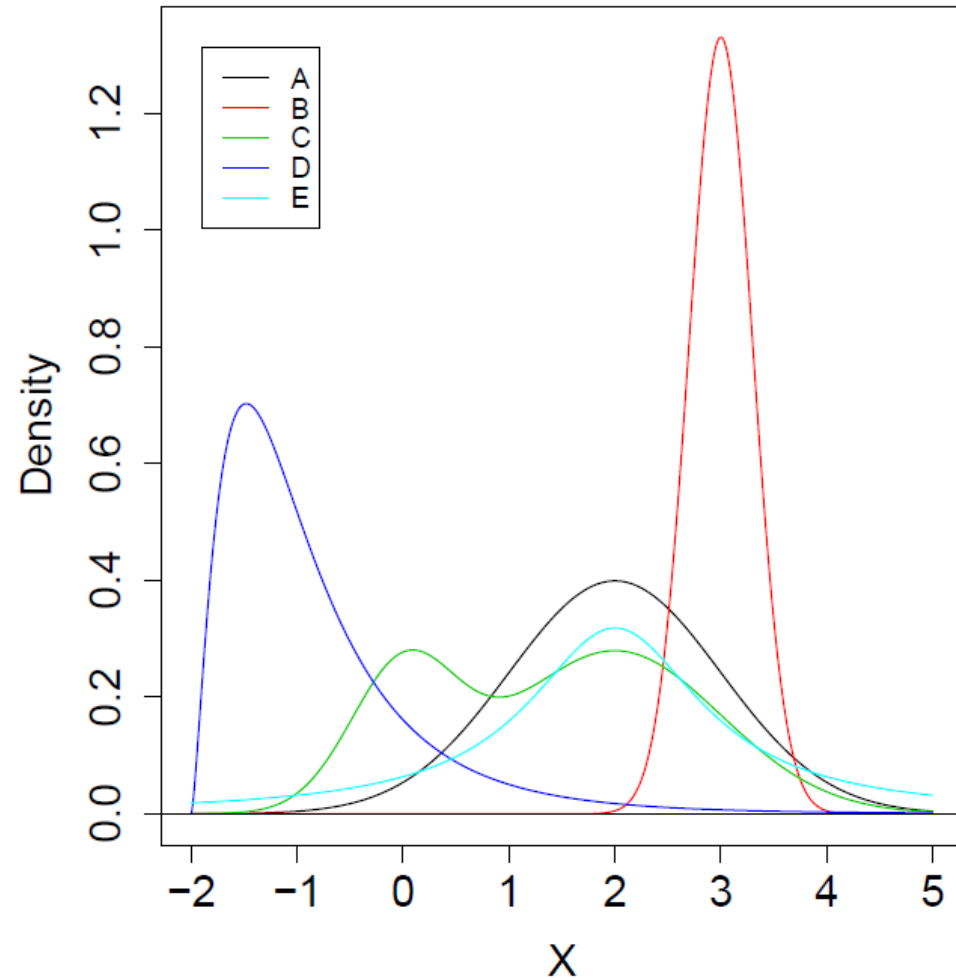
✳ It has perfect bell-shape of a Gaussian distribution.

✳ **Distribution B** is also Gaussian in shape, has a different central tendency (shifted higher or rightward), and has a smaller spread.

✳ **Distribution C** is bimodal (two peaks) so it cannot be a Gaussian distribution.

✳ **Distribution D** has the lowest center and is asymmetric (skewed to the right), so it cannot be Gaussian.

✳ **Distribution E** appears similar to a Gaussian distribution, but while symmetric and roughly bell-shaped, it has "tails" that are too fat to be a true bell-shaped, Gaussian distribution.



Measures of central tendency

🔧 Methods for representing the data with a single numerical value.

🔧 **The Arithmetic Mean** (*arithmetic average*) (*mean*):

🔧 μ (the population mean, pronounced “myew”)

🔧 \bar{x} (the sample mean, “x bar”).

🔧 The population mean μ applies when our data represent *all* of the items within the population.

🔧 The sample mean \bar{x} is applicable whenever data represent a *sample* taken from the population.

Population mean and Sample mean

Population mean

$$\mu = \frac{\sum_{i=1}^N x_i}{N} \quad \text{or simply} \quad \mu = \frac{\sum x_i}{N}$$

where μ = population mean

x_i = the i th data value in the population

Σ = the sum of

N = number of data values in the population

Sample mean

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{where } \bar{x} = \text{sample mean}$$

x_i = the i th data value in the sample

Σ = the sum of

n = number of data values in the sample

 In determining either a population mean (μ) or a sample mean (\bar{x}), the sum of the data values is divided by the number of observations.


Example of Population mean

 Arithmetic mean

 Population mean

City	Peanuts (Thousands of Bags)
Montreal	64.0
Ottawa	15.0
Toronto	285.0
Vancouver	228.0
Winnipeg	45.0

$$\mu = \frac{\sum x_i}{N} = \frac{64.0 + 15.0 + 285.0 + 228.0 + 45.0}{5} = 127.4 \text{ thousand bags}$$

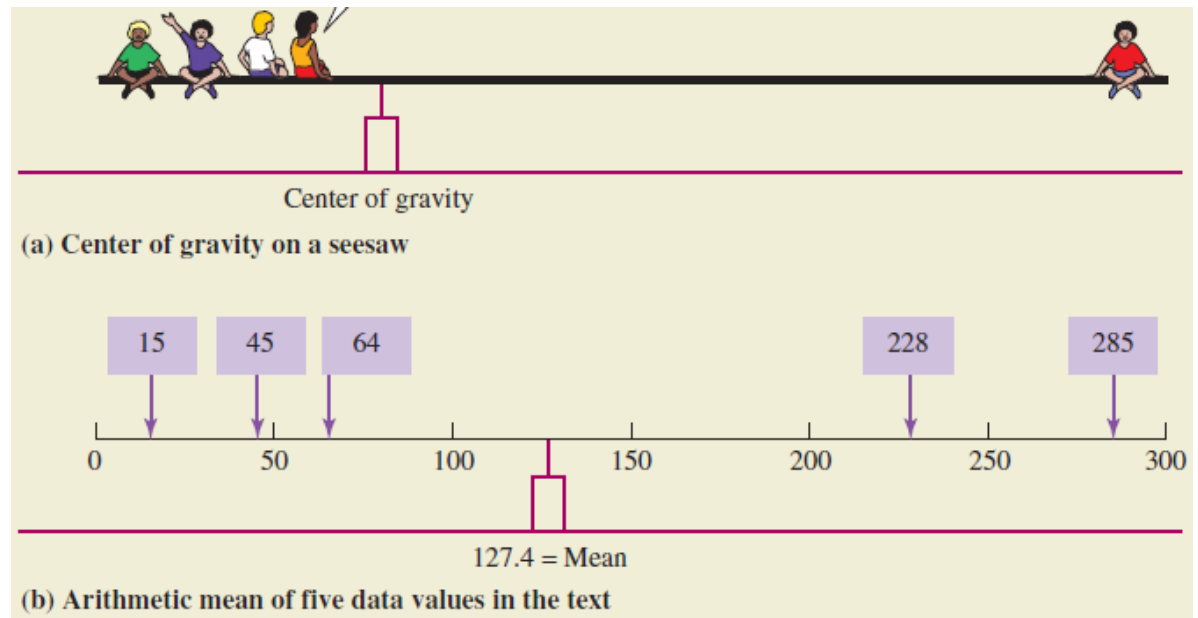
 On the average, each destination received 127.4 thousand bags of peanuts during the time period involved.

Example of Arithmetic mean

🌸 Potential weakness of the mean as a descriptor

🌸 The arithmetic mean, or average, is a mathematical counterpart to the center of gravity on a seesaw.

🌸 Although the influence of the two values that are more than 200 thousand is not quite as great as that of ptr, it causes the arithmetic mean to be greater than three of the five data values.



Example of Weighted Mean

🧠 When some values are more important than others, a **weighted mean** (sometimes referred to as a **weighted average**) may be calculated.

Weighted mean, μ_w (for a population) or \bar{x}_w (for a sample):

$$\mu_w \text{ or } \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} \quad \begin{array}{l} \text{where } w_i = \text{weight assigned to the } i\text{th data value} \\ x_i = \text{the } i\text{th data value} \end{array}$$

🧠 Continuing with the peanut example, let's assume that shipments to the respective cities will be sold at the following profits per thousand bags: \$15.00, \$13.50, \$15.50, \$12.00, and \$14.00. (Note: In this example, we are trying to determine the weighted mean for the profit;

Example of Weighted Mean

✿ The average profit per thousand bags will not be

$$(15.00+13.50+15.50+12.00+14.00)/5,$$

✿ because the cities did not receive equal quantities of peanuts.

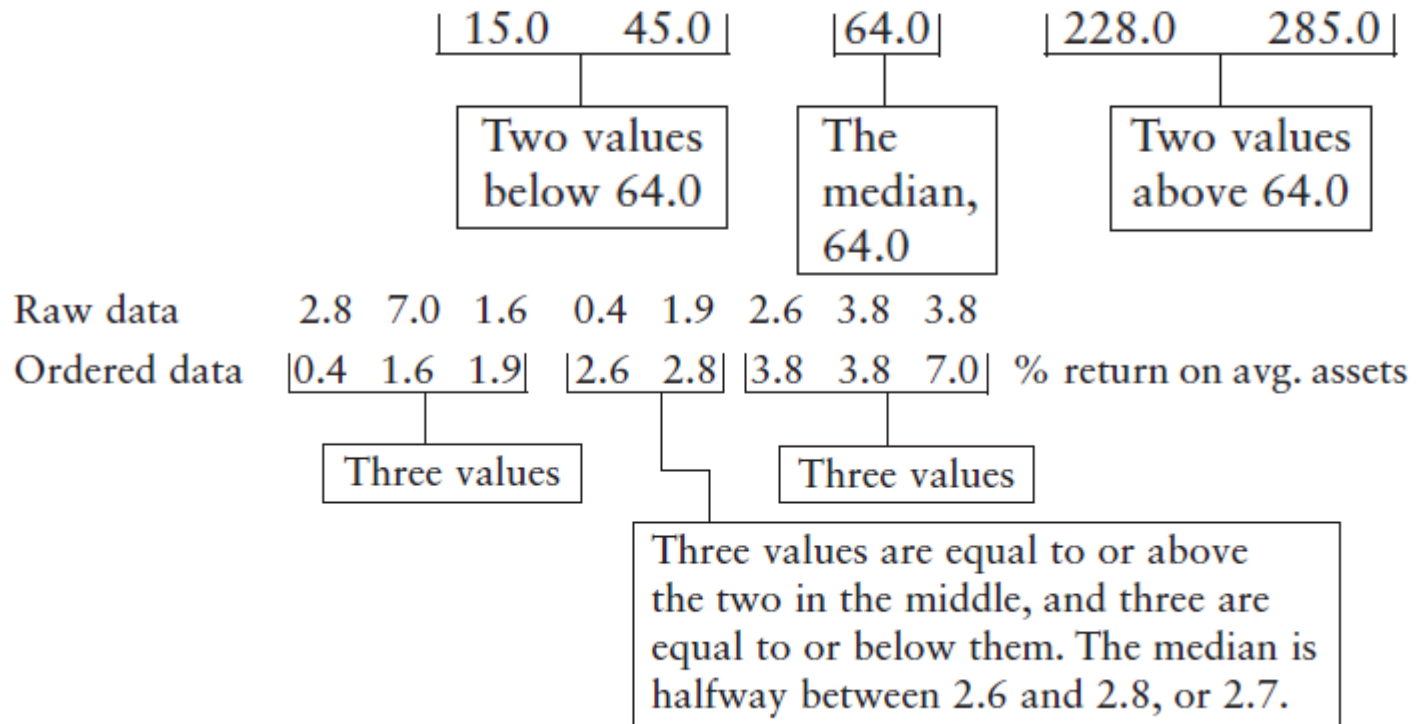
✿ A weighted mean must be calculated if we want to find the average profit per thousand bags for all shipments of peanuts

$$\begin{aligned}\mu_w &= \frac{\sum w_i x_i}{\sum w_i} \\ &= \frac{64(\$15.00) + 15(\$13.50) + 285(\$15.50) + 228(\$12.00) + 45(\$14.00)}{64 + 15 + 285 + 228 + 45} \\ &= \$14.04 \text{ per thousand bags}\end{aligned}$$

City	Peanuts (Thousands of Bags)
Montreal	64.0
Ottawa	15.0
Toronto	285.0
Vancouver	228.0
Winnipeg	45.0


Example of The Median

🌸 In a set of data, the median is the value that has just as many values above it as below it. For example, the numbers of bags of peanuts (in thousands) shipped to the five cities were




🌸 Unlike the mean, the median is not influenced by extreme high or low values in the data. For example, if the highest data value had been 600% instead of 7.0%, the median would still be 2.7%.

Example of The Mode

 In a set of data, the mode is a value that occurs with the greatest frequency.

Ordered data 0.4 1.6 1.9 2.6 2.8 3.8 3.8 7.0 % return on avg. assets

Mode = 3.8

 For these data, the mode is 3.8, since it occurs more frequently than any other value. In this case, the mode does not appear to be a very good descriptor of the data, as five of the other six values are all smaller than 3.8.

 Depending on the data, there can be more than one mode.

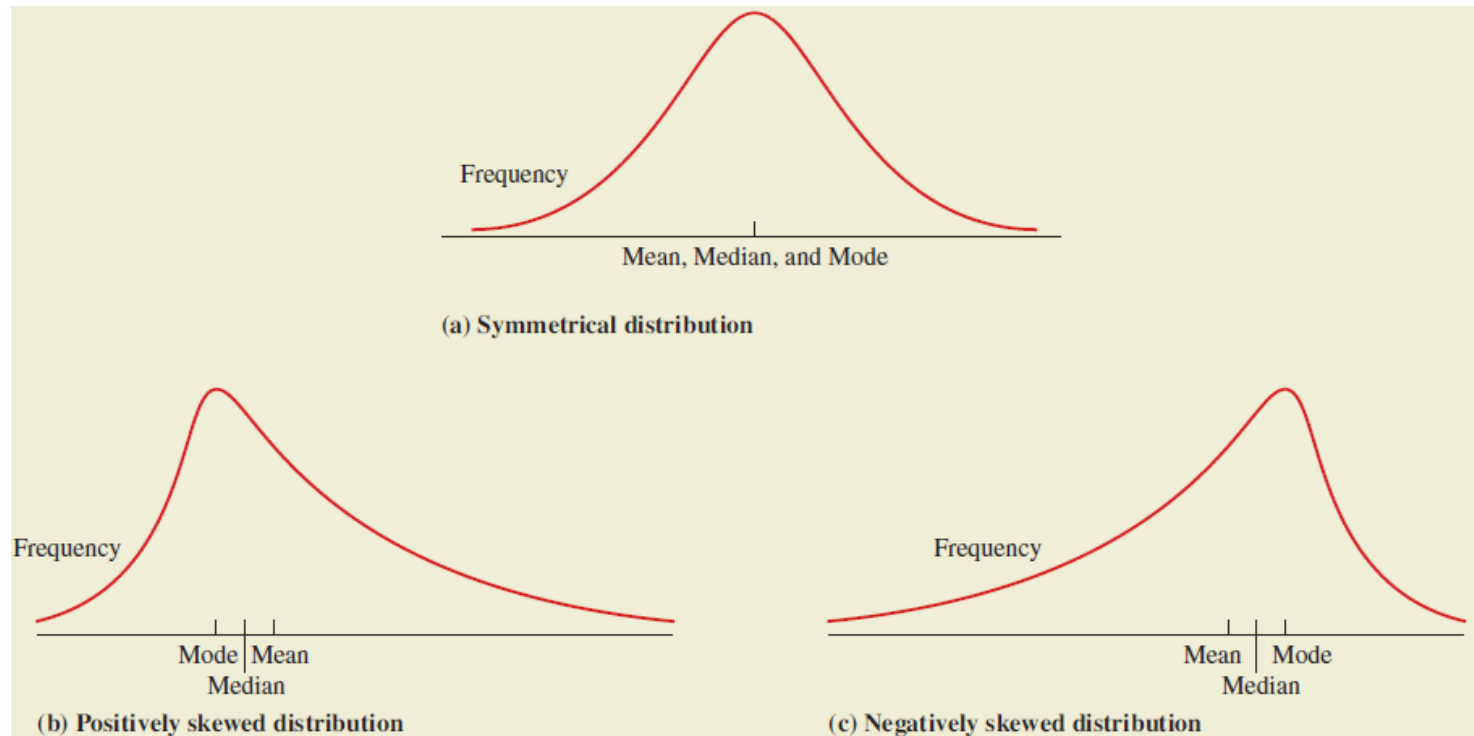
 When there are two modes, a distribution of values is referred to as bimodal.

Distribution Shape and Measures of Central Tendency

The relative values of the mean, median, and mode are very much dependent on the shape of the distribution which may be described in terms of symmetry and skewness.

Symmetrical distribution (the left and right sides of the distribution are mirror images of each other) has a single mode, is bell shaped, and is known as the normal distribution.

Skewness refers to the tendency of the distribution to “tail off” to the right or left



Range

- ✿ The simplest measure of dispersion, the **range** is the difference between the highest and lowest values.
- ✿ The **midrange**, a variant of the range, is the average of the lowest data value and the highest data value.

Quantiles

- ✿ The **median** divides data into two equal-size groups: one with values above the median, the other with values below the median.
- ✿ **Quantiles** also separate the data into equal-size groups in order of numerical value.
- ✿ There are several kinds of quantiles,

PERCENTILES divide the values into 100 parts of equal size, each comprising 1% of the observations. The median describes the 50th percentile.

DECILES divide the values into 10 parts of equal size, each comprising 10% of the observations. The median is the 5th decile.

QUARTILES divide the values into four parts of equal size, each comprising 25% of the observations. The median describes the second quartile, below which 50% of the values fall. (Note: In some computer statistical packages, the first and third quartile values may be referred to as *hinges*.)

Quantiles

- After values are arranged from smallest to largest, quartiles are calculated similarly to the median.
- It may be necessary to interpolate (calculate a position between) two values to identify the data position corresponding to the quartile.

For N values arranged from lowest to highest:

$$\text{First quartile, } Q_1 = \text{Data value at position } \frac{(N + 1)}{4}$$

$$\text{Second quartile (the median), } Q_2 = \text{Data value at position } \frac{2(N + 1)}{4}$$

$$\text{Third quartile, } Q_3 = \text{Data value at position } \frac{3(N + 1)}{4}$$

(Use N if data represent a population, n for a sample.)

Mean Absolute Deviation (MAD)

- ✿ It is also called as average deviation or the average absolute deviation.
- ✿ Consider the extent to which the data values tend to differ from the mean.
- ✿ The mean absolute deviation (MAD) is the average of the absolute values of differences from the mean and may be expressed as follows:

Mean absolute deviation (*MAD*) for a population:

$$MAD = \frac{\sum |x_i - \mu|}{N}$$

where μ = population mean
 x_i = the i th data value
 N = number of data values in the population

(To calculate *MAD* for a sample, substitute n for N and \bar{x} for μ .)

Mean Absolute Deviation (MAD): Example


✿ Calculation of mean absolute deviation for annual research and development (R&D) expenditures for Microsoft Corporation. Data are in millions of dollars.


$$\mu = \frac{4379 + 6299 + 6595 + 7779 + 6184}{5} = \$6247.2 \text{ million}$$

Year	R&D x_i	Deviation from Mean $(x_i - \mu)$	Absolute Value of Deviation from Mean $ x_i - \mu $
2001	4379	-1868.2	1868.2
2002	6299	51.8	51.8
2003	6595	347.8	347.8
2004	7779	1531.8	1531.8
2005	6184	-63.2	63.2
	<u>31,236</u>	<u>0.0</u>	<u>3862.8</u>
	$= \sum x_i$		$= \sum x_i - \mu $
Mean: $\mu = \frac{\sum x_i}{N} = \frac{31,236}{5} = \6247.2 million			
Mean absolute deviation: $MAD = \frac{\sum x_i - \mu }{N} = \frac{3862.8}{5} = \772.6			

Variance and Standard Deviation

 **Variance:** The **variance**, a common measure of dispersion, includes all data values and is calculated by a mathematical formula.

 For a population, the variance (σ^2 , “sigma squared”) is the average of squared differences between the N data values and the mean, μ .

 For a sample variance (s^2), the sum of the squared differences between the n data values and the mean \bar{x} , is divided by (n-1).

Variance for a population:

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

where σ^2 = population variance
 μ = population mean
 x_i = the i th data value
 N = number of data values in the population

Variance for a sample:

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

where s^2 = sample variance
 \bar{x} = sample mean
 x_i = the i th data value
 n = number of data values in the sample


Variance: Example


Variance:

Model	Highway mpg (x_i)	x_i^2	Residual ($x_i - \bar{x}$)	Residual ² ($x_i - \bar{x}$) ²
Saturn Outlook	23	529	3.0	9.0
Jeep Liberty	21	441	1.0	1.0
Subaru Tribeca	21	441	1.0	1.0
Land Rover LR3	17	289	-3.0	9.0
Porsche Cayenne GTS	18	324	-2.0	4.0
	<u>100</u>	<u>2024</u>		<u>24.0</u>
	$= \sum x_i$	$= \sum x_i^2$		$= \sum (x_i - \bar{x})^2$
$\bar{x} = \frac{\sum x_i}{5} = \frac{100}{5} = 20.0 \text{ miles per gallon}$				
$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{24.0}{5 - 1} = 6.0 \quad s = \sqrt{6.0} = 2.45$				
Source: U.S. Environmental Protection Agency, <i>Fuel Economy Guide</i> 2009.				

Standard Deviation

Standard Deviation:

 The positive square root of the variance of either a population or a sample is a quantity known as the standard deviation.

 The standard deviation is an especially important measure of dispersion because it is the basis for determining the proportion of data values within certain distances on either side of the mean for certain types of distributions.

	For a Population	For a Sample
Standard Deviation	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

Standard Deviation

Probability distribution or random variable

Let X be a random variable with mean value μ :

$$E[X] = \mu.$$

Here the operator E denotes the average or expected value of X . Then the **standard deviation** of X is the quantity

$$\sigma = \sqrt{E[(X - \mu)^2]}.$$

That is, the standard deviation σ (sigma) is the square root of the average value of $(X - \mu)^2$.

In the case where X takes random values from a finite data set x_1, x_2, \dots, x_N , with each value having the same probability,

the standard deviation is
$$\sigma = \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \dots + (x_N - \mu)^2}{N}},$$

or, using summation notation,

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}.$$

Standard Deviation

Consider a population consisting of the following eight values:

2, 4, 4, 4, 5, 5, 7, 9.

The eight data points have a mean (or average) value of 5:

$$\frac{2 + 4 + 4 + 4 + 5 + 5 + 7 + 9}{8} = 5.$$

To calculate the population standard deviation, first compute the difference of each data point from the mean, and square the result:

$$\begin{aligned}(2 - 5)^2 &= (-3)^2 = 9 & (5 - 5)^2 &= 0^2 = 0 \\(4 - 5)^2 &= (-1)^2 = 1 & (5 - 5)^2 &= 0^2 = 0 \\(4 - 5)^2 &= (-1)^2 = 1 & (7 - 5)^2 &= 2^2 = 4 \\(4 - 5)^2 &= (-1)^2 = 1 & (9 - 5)^2 &= 4^2 = 16\end{aligned}$$

Next divide the sum of these values by the number of values and take the square root to give the standard deviation:

$$\sqrt{\frac{9 + 1 + 1 + 1 + 0 + 0 + 4 + 16}{8}} = 2.$$

Therefore, the above has a population standard deviation of 2.

Parameters describing distributions

 Common parameters and their definitions as expected values

Name	Definition	Symbol
mean	$E[X]$	μ
variance	$E[(X - \mu)^2]$	σ^2
standard deviation	$\sqrt{\sigma^2}$	σ
skewness	$E[(X - \mu)^3]/\sigma^3$	γ_1
kurtosis	$E[(X - \mu)^4]/\sigma^4 - 3$	γ_2

Population (probability) distributions

✳ Population (probability) distributions of 5 different continuous random variables.

✳ **Distribution A** is a unimodal (one peak) symmetric distribution, centered around 2.0.

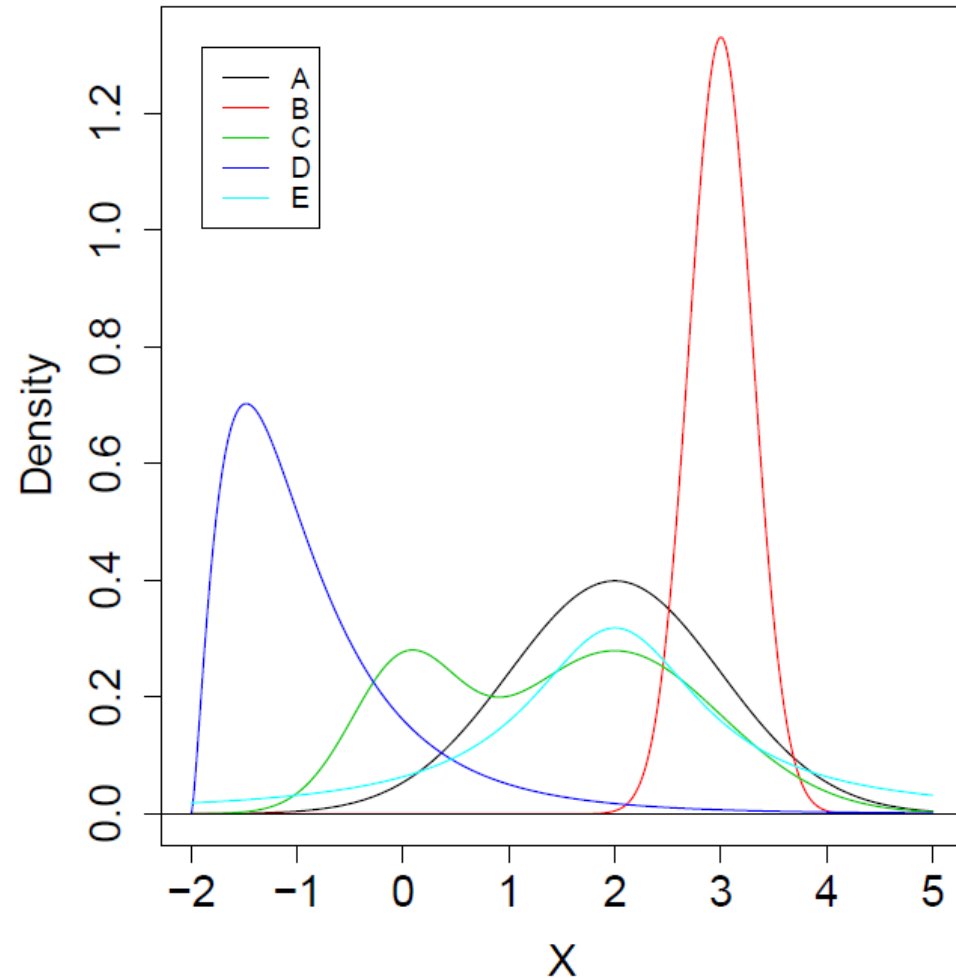
✳ It has perfect bell-shape of a Gaussian distribution.

✳ **Distribution B** is also Gaussian in shape, has a different central tendency (shifted higher or rightward), and has a smaller spread.

✳ **Distribution C** is bimodal (two peaks) so it cannot be a Gaussian distribution.

✳ **Distribution D** has the lowest center and is asymmetric (skewed to the right), so it cannot be Gaussian.

✳ **Distribution E** appears similar to a Gaussian distribution, but while symmetric and roughly bell-shaped, it has "tails" that are too fat to be a true bell-shaped, Gaussian distribution.



Conclusion



Bibliography

- [1] Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, 3rd edition
Publisher: Morgan Kaufmann; 3 edition
- [2] Business Intelligence, 2/E; Efraim Turban, Ramesh Sharda, Dursun Delen, David King;
pearson Education
- [3] Data Mining for Business Intelligence: Concepts, Techniques, and Applications in
Microsoft Office Excel with Xlminer; 2nd edition, Galit Shmueli, Nitin R. Patel and Peter C.
Bruce; John Wiley
- [4] Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management,
Author: Berry, Gordon S. Linoff, Format: Paperback, 648 pages, Edition: 3; Publisher: John
Wiley & Sons Inc.
- [5] Robert Groth, Data Mining: Building Competitive Advantage, Prentice Hall, 2000.
- [6] P. N. Tan, M. Steinbach, Vipin Kumar, “Introduction to Data Mining”, Pearson Education
- [7] Alex Berson and Smith, “Data Mining and Data Warehousing and OLAP”, McGraw Hill
Publication
- [8] E. G. Mallach, “Decision Support and Data Warehouse Systems”, Tata McGraw Hill.
- [9] Michael Berry and Gordon Linoff “Mastering Data Mining- Art & science of CRM”,
Wiley Student Edition

Thank You.

