

Statistical Description (Parameters) of Data



Satishkumar L. Varma

Professor, Department of Computer Engineering

PCE, New Panvel

www.sites.google.com/view/vsat2k

www.vsat2k.wordpress.com






www.vsat2k.moodlecloud.com

Outline

- ❄ Central tendency: mean and median
 - ❄ Arithmetic and Population Mean
- 📊 Spread: variance, standard deviation
- ❄ Population skewness and Population kurtosis
- ❄ Quantiles
- 📊 Interquartile range
- 📊 Summary and Conclusion
- 📊 Bibliography

LEARNING OBJECTIVES

 Learner should be able to:

-  Describe data by using measures of central tendency and dispersion.
-  Convert data to standardized values.
-  Use the computer to visually represent data with box plot, etc.
-  Determine measures of central tendency and dispersion for grouped data.
-  Use the coefficient of correlation to measure association between two quantitative variables.

LEARNING OBJECTIVES

🌸 Statistical methods for describing typical values in the data are:

🌸 Measures of central tendency:

🌸 A numbers describing typical data values.

🌸 e.g numerically describe the typical income of members of each group.

🌸 The primary measures of central tendency are the arithmetic mean, weighted mean, median, and mode.

🌸 Measures of dispersion

🌸 A Numbers that describe the scatter of the data.

🌸 Measures of dispersion allow us to numerically describe the scatter, or spread, of measurements.

🌸 e.g. heights of Basketball players are between 6'8" and 7'4" is not as wide as the dispersion that exists within the general population.

🌸 The measures of dispersion are the range, quantiles, mean absolute deviation, variance, and standard deviation, etc.

Measures of central tendency

📌 Measures of central tendency and dispersion use two key terms

📌 Population

📌 Sample

🔴 Information is used to make inferences about **population** from which **sample** was drawn

📌 Characteristics of the **population** are referred to as **parameters**

📌 Characteristics of the **sample** are referred to as **statistics**

📌 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

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

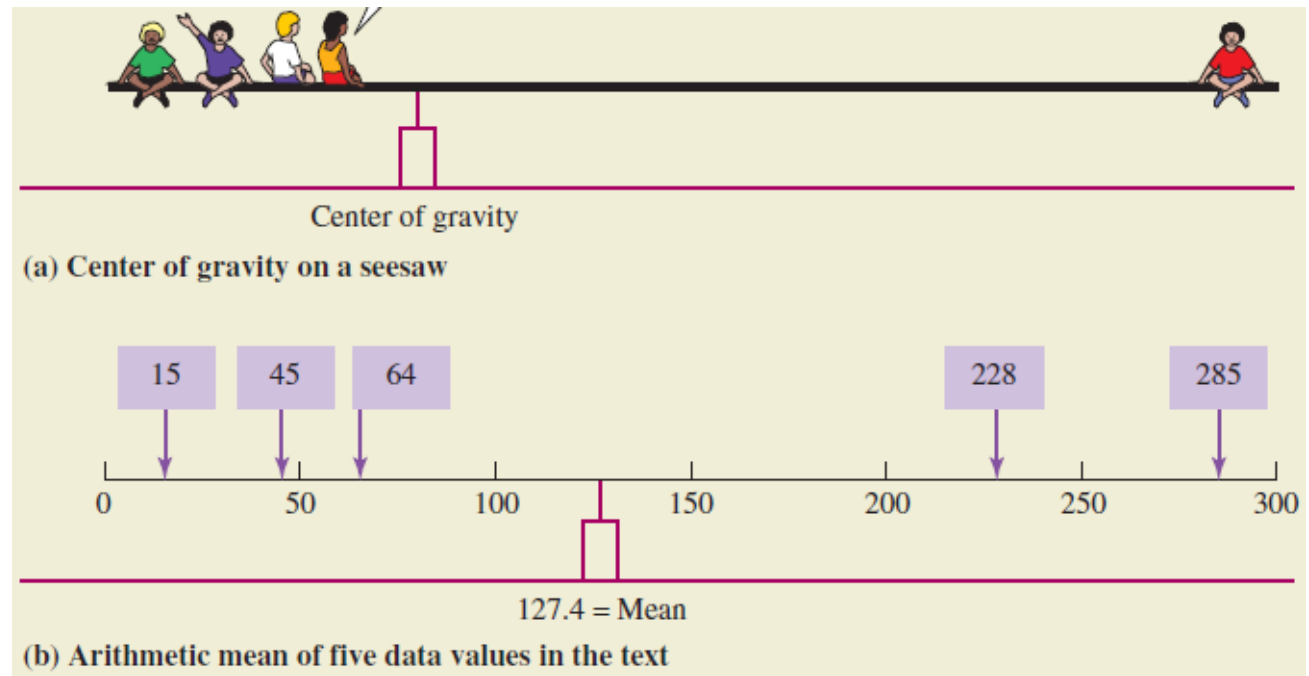
✿ On the average, each destination received 127.4 thousand bags of peanuts during the time period involved.

Example of Arithmetic mean

🌀 Potential weakness of the mean as a descriptor

🌀 AM or Avg 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.

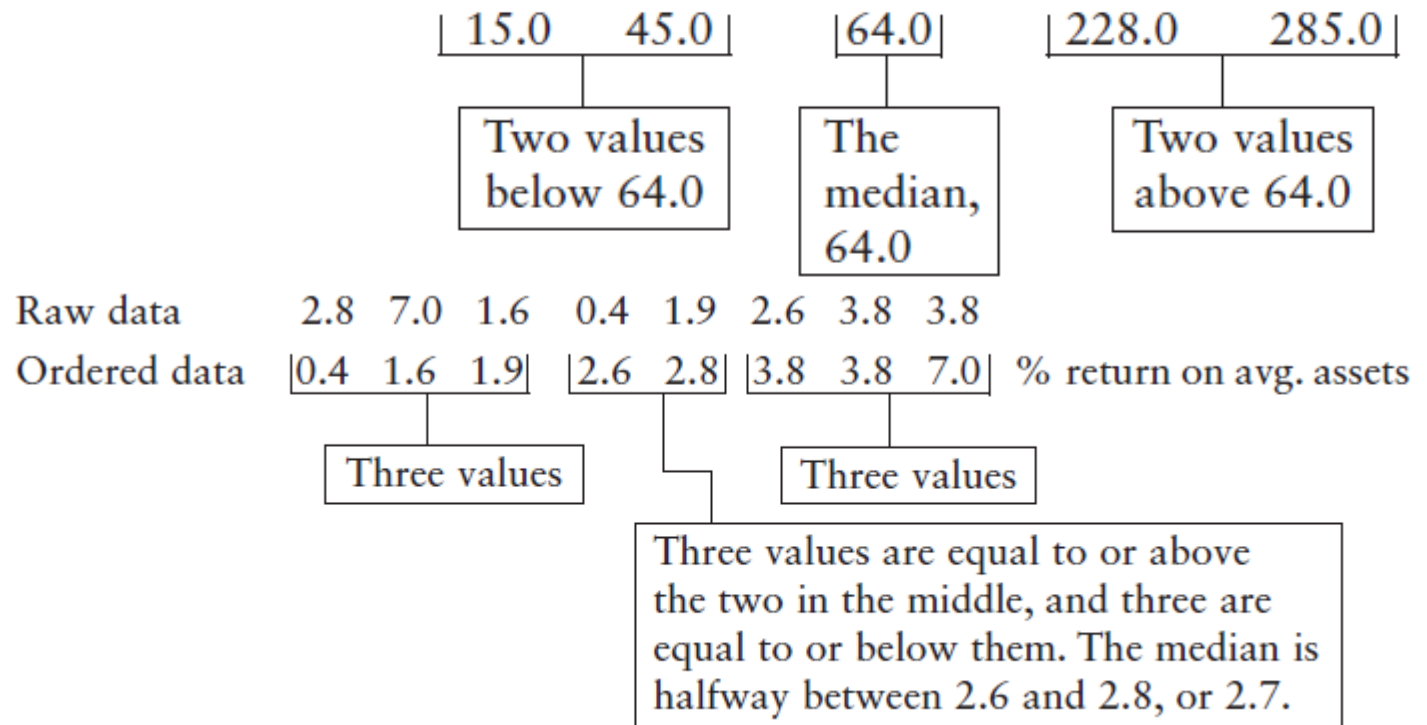
✿ A weighted mean must be calculated if we want to find the average profit per thousand bags for all shipments of peanuts

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

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.

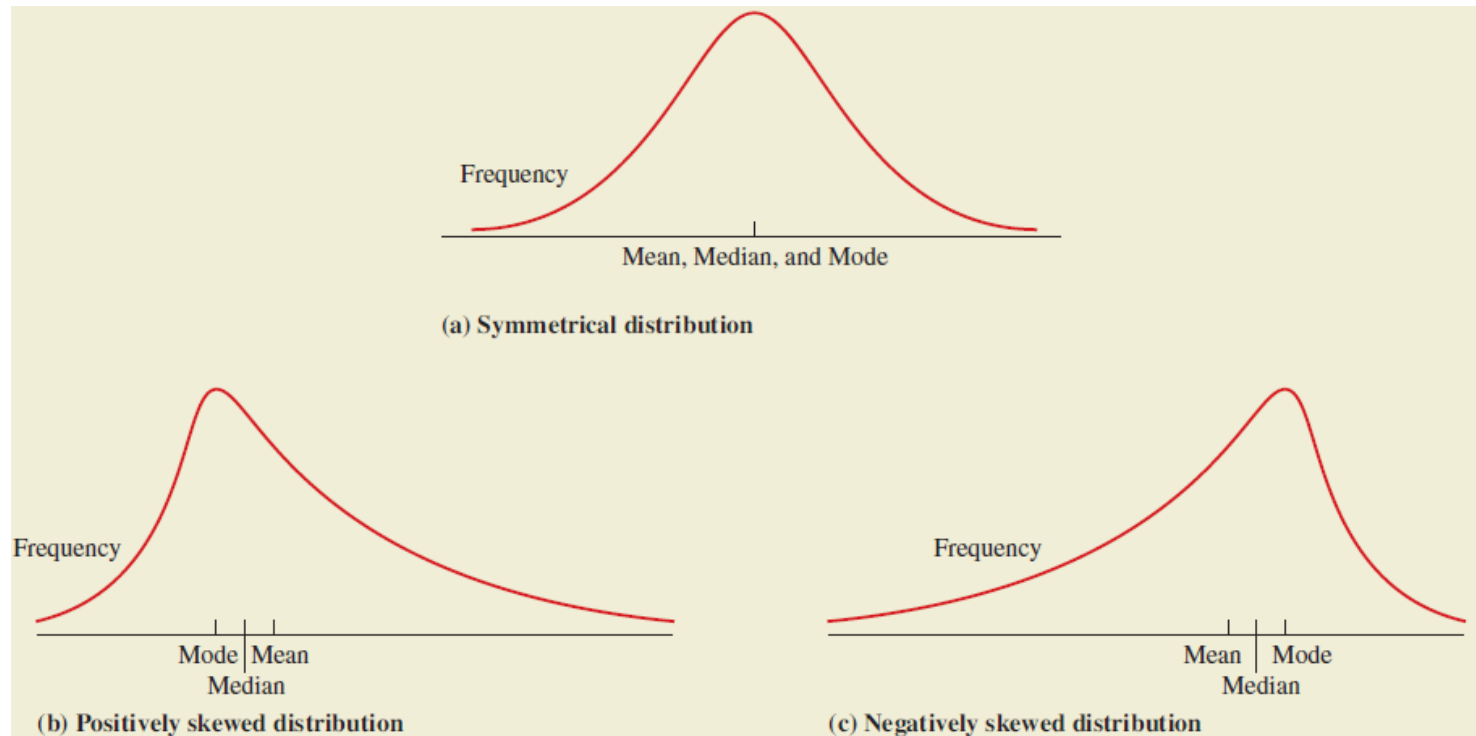
 When there are two modes, a distribution of values is referred to as bimodal.

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.

Summary

❖ Statistical Descriptors

❖ Statistical Parameters

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", TMH
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.

