Biostatistics and Computer Science (PHR 221)

Statistics

Agnik Saha

Lecturer

Department of Computer Science and Engineering

R. P. Shaha University

September 17, 2023

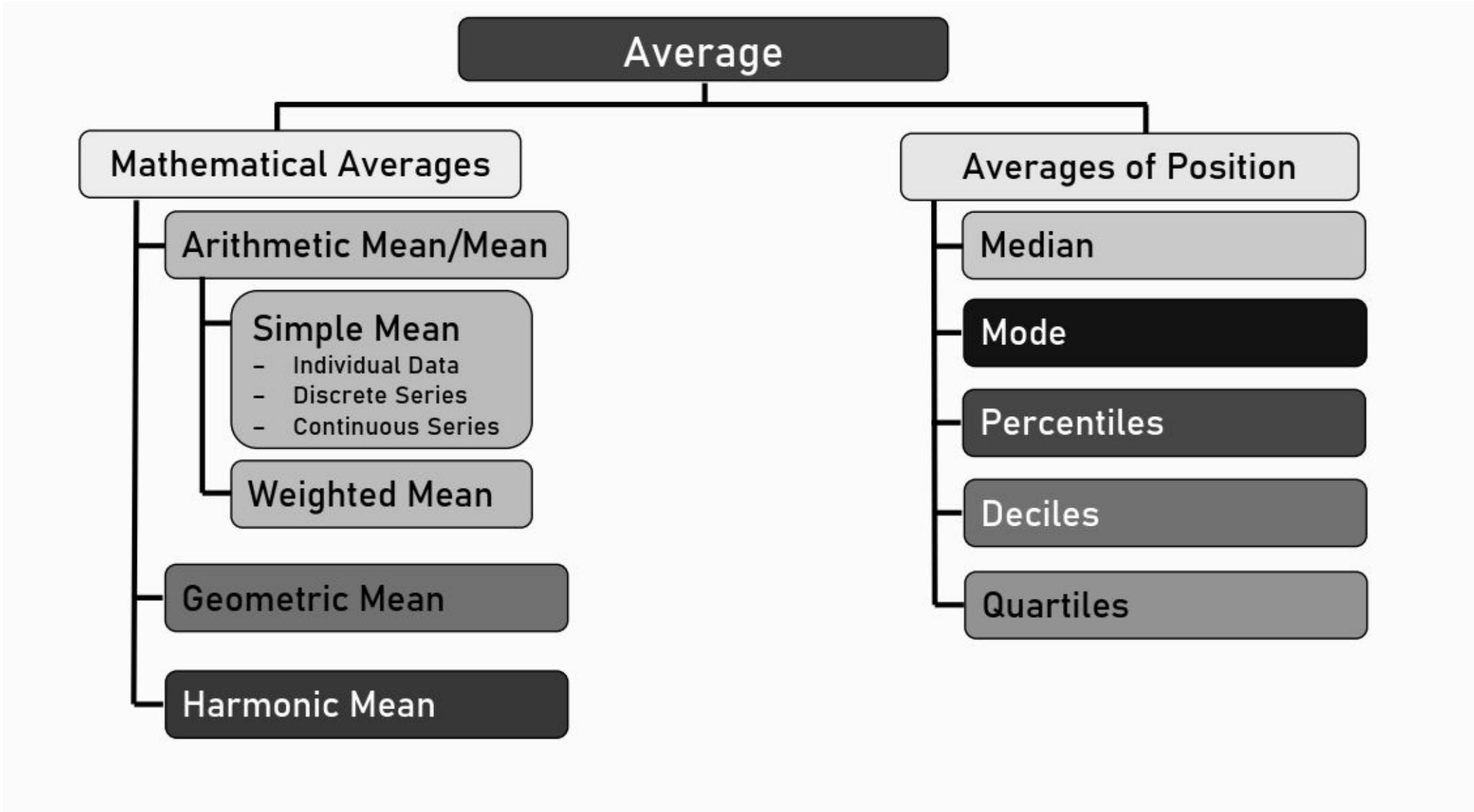
Motivation

- ➔ Statistics consists of the tools for making sense of data.
- ➔ One of the most fundamental questions we can ask about the data set is this: "What single number is most representative of the set as a whole?"
- ➔ In statistics, such numbers are called "measures of center.

Mean

- The two most important measures of center are the mean and the median.
- A mean is simply an ordinary average.
- To find the mean of, say, seven numbers on a list, we would add up the seven numbers, and then divide this sum by seven.
- In general, on a list with N entries, we add up all the entries, and then divide by N. That's the mean.

Average


Average
Mathematical Averages
Arithmetic Mean/Mean
Simple Mean
- Individual Data
- Discrete Series
- Continuous Series
Weighted Mean
Geometric Mean
Harmonic Mean
Averages of Position
Median
Mode
Percentiles
Deciles
Quartiles

Formulas

Arithmetic Mean/Mean (Simple Mean)

	Normal Formula	Short Cut Method
Individual Data	$\bar{X} = \frac{X_1 + X_2 \ldots + X_n}{N} = \frac{\Sigma X}{N}$	$\bar{X} = A + \frac{\Sigma d}{N}$ Here, d= (X-A)
Discrete Series	$\bar{X} = \frac{\Sigma fX}{N}$	$\bar{X} = A + \frac{\Sigma fd}{N}$
Continuous Series	$\bar{X} = \frac{\Sigma fm}{N}$	$\bar{X} = A + \frac{\Sigma fd}{N}$ here, d=(m-A)

Weighted Mean (Arithmetic Mean)

$$\overline{X_w} = \frac{W_1X_1 + W_2X_2 \ldots + W_nX_n}{W_1 + W_2 + \cdots Wn}$$

$$\overline{X_w} = \frac{\Sigma WX}{\Sigma W}$$

Formulas

Geometric Mean

Individual Data

$$\log GM = \frac{\log X_1 + \log X_2 \ldots + \log X_n}{N} = \frac{\Sigma \log X_n}{N}$$

$$GM = antilog\left(\frac{\Sigma \log Xn}{N}\right)$$

Discrete Series

$$GM = antilog\left(\frac{\Sigma f.\log X_n}{N}\right)$$

Continuous Series

$$GM = antilog\left(\frac{\Sigma f.\log m}{N}\right)$$

Formulas

Harmonic Mean

Individual Data

$$HM = \frac{N}{\Sigma(\frac{1}{X})}$$

Discrete Series

$$HM = \frac{N}{\Sigma(f \times \frac{1}{X})}$$

Continuous Series

$$HM = \frac{N}{\Sigma(f \times \frac{1}{m})}$$

Formulas

Median

Individual Data

$$\text{Median} = \frac{N+1}{2}\text{th}$$

Discrete Series

$$\text{Median} = \frac{N+1}{2}\text{th}$$

Continuous Series

$$\text{Median} = L + \frac{\frac{N}{2} - c.f.}{f} \times i$$

Mode

- Mode is the most frequent value of the series.
- Used when highly skewed frequency.

Individual Data

Discrete Series

Most frequent value

Continuous Series

$$M_0 = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$$

Mean, Mode, Median Relationship

Mode = 3Median – 2 Mean

Practice Questions

- In a class, 18 students took a test and had an average of 70. Alicia and Burt then took the test, and the average of all 20 students was 71. If Alicia got a 77, then what was Burt's grade?

Practice Questions

➔ In a class, 18 students took a test and had an average of 70. Alicia and Burt then took the test, and the average of all 20 students was 71. If Alicia got a 77, then what was Butt's grade?

ANS: 83

Median

- The median is the middle number on a list.
- Of course, we have to put the list in ascending order first: technically, the median is the middle number on an ordered list.
- list A = {1,2,3,4,5,6,7}
- the median =
- list B = {1, 2, 3, 4, 5, 6,7,8}
- the median is between __ and __
- the median is _____

Contd..

- ➔ Notice that the median only takes into account the number or numbers at the very center.
- ➔ We could change the numbers at either end of the list, and this change wouldn't affect the median at all.

List A = {10, 4, 7, 18}
List B = {x, 10, 4, 7, 18, 25}
If the median of List B is exactly 4 higher than the median of List A, what is the value of x?

Contd..

- ➔ Notice that the median only takes into account the number or numbers at the very center.
- ➔ We could change the numbers at either end of the list, and this change wouldn't affect the median at all.

List A = {10, 4, 7, 18}
List B = {x, 10, 4, 7, 18, 25}
If the median of List B is exactly 4 higher than the median of List A, what is the value of x? -> 15

Mode

- ➔ One final measure of center is mode.
- ➔ Mode is the most frequent number on a list. This is far less important than either mean or median.
- ➔ Some lists have a single mode: {1,2,3,3,3,4,5}
- ➔ Some lists have two modes: {2,2,3,4,5,5,7}

Contd..

- ➔ If all the numbers on the list are different from one another, as is often the case, then there simply is no mode.
- ➔ Every list has a mean.
- ➔ Every list has a median.
- ➔ Only some lists have modes. Some lists have more than one mode, and many have no mode at all.

Contd..

- ➔ We noted that changing the highest or lowest number on a list would not change the median, but it would change the mean
- ➔ Numbers far away from the center of the list are called "outliers."
- ➔ The mean is heavily influenced by outliers, and the median is entirely unaffected by outliers.

Contd..

- First of all, when are the mean and the median the same?
- If the list consists of evenly spaced numbers, then the mean equals the median.
- Consecutive integers and consecutive multiples of the same number are examples of evenly spaced lists.
- List A ={7,8,9,10, 11, 12, 13, 14, 15}
- mean = median = _____

Contd..

➔ Also, mean and median are equal whenever the list is entirely symmetrical.

Contd..

- Also, mean and median are equal whenever the list is entirely symmetrical.
- List B = {4, 8, 13, 23, 25, 27, 37, 42, 46}
- median = 25
- 23 = 25 -2 27 = 25 +2
- 13 = 25 - 12 37 = 25 + 12
- 4 = 25 -21 46 = 25 + 21
- 42 = 25 + 17
- The list is entirely symmetrical around 25, so mean = median = 25

Contd..

- When the list is asymmetrical, then the mean and median differ.
- In particular, when there's a distinct outlier or set of outliers in one direction, that pulls the mean away from the median.
- List C = {1, 2, 3, 4, 5, 6, 7}
- mean = median = 4
- List D = {1, 2, 3, 4, 5, 6, 700}
- median = 4 but mean > 4

Contd..

On a test in a class of more than 40 students, the scores had mean = median = mode = 81. Two absent students then took the test; they received grades of 83 and 47. What are the new mean & median?

(A) mean = 81 and median = 81
(B) mean < 81 and median = 81
(C) mean = 81 and median < 81
(D) mean < 81 and median < 81

Contd..

On a test in a class of more than 40 students, the scores had mean = median = mode = 81. Two absent students then took the test; they received grades of 83 and 47. What are the new mean & median?

(A) mean = 81 and median = 81

(B) mean < 81 and median = 81

(C) mean = 81 and median < 81

(D) mean < 81 and median < 81

Summary

1) If all the numbers on a list are evenly spaced, or if the list is symmetrically distributed, then mean = median.

2) Outliers pull the mean away from the median.

3) We often can compare mean & median—or infer which one got bigger or smaller—without a calculation, purely by observing the direction of outliers.

Weighted Average

Sometimes you will be asked the combined average of two different groups.

Sample problem:

On a ferry, there are 50 cars and 10 trucks. The cars have an average mass of 1200 kg and the trucks have an average mass of 3000 kg. What is the average mass of all 60 vehicles on the ferry?

Weighted Average

Sometimes you will be asked the combined average of two different groups.

Sample problem:

On a ferry, there are 50 cars and 10 trucks. The cars have an average mass of 1200 kg and the trucks have an average mass of 3000 kg. What is the average mass of all 60 vehicles on the ferry?

-> 1500 Kg

Practice Problems

- In a certain company, 70% of employees are marketers who make an average of $40,000; 20% are programmers who make an average of $80,000; and 10% are managers, who make an average of $120,000. What is the average salary of all employees at this company?

Practice Problems

➔ In a certain company, 70% of employees are marketers who make an average of $40,000; 20% are programmers who make an average of $80,000; and 10% are managers, who make an average of $120,000. What is the average salary of all employees at this company?

$ 56,000

Practice Problems

➔ At Didymus Corporation, there are just two classes of employees: silver and gold. The average salary of gold employees is $56,000 higher than that of silver employees. If there are 120 silver employees and 160 gold employees, then the average salary for the company is how much higher than the average salary for the silver employees?

Practice Problems

➔ At Didymus Corporation, there are just two classes of employees: silver and gold. The average salary of gold employees is $56,000 higher than that of silver employees. If there are 120 silver employees and 160 gold employees, then the average salary for the company is how much higher than the average salary for the silver employees?

$ 32000

Practice Problems

By weight, liquid A makes up 8 percent of solution R and 18 percent of solution S. If 3 grams of solution R are mixed with 7 grams of solution S, then liquid A accounts for what percent of the weight of the resulting solution?

Practice Problems

➔ You are a student in a university, and you have completed four subjects in a semester, each with a different number of credit hours. Your grades and the respective credit hours for each subject are as follows:

- Mathematics: Grade A (4.0) - 4 credit hours
- Chemistry: Grade B+ (3.3) - 3 credit hours
- History: Grade A- (3.7) - 2 credit hours
- English: Grade B (3.0) - 5 credit hours

Calculate your CGPA for the semester.

Practice Problems

➔ You are a student in a university, and you have completed four subjects in a semester, each with a different number of credit hours. Your grades and the respective credit hours for each subject are as follows:

- Mathematics: Grade A (4.0) - 4 credit hours
- Chemistry: Grade B+ (3.3) - 3 credit hours
- History: Grade A- (3.7) - 2 credit hours
- English: Grade B (3.0) - 5 credit hours

Calculate your CGPA for the semester.

3.45

Practice Problems

The average annual rainfall in Boynton for 1976-1979 was 26 inches per year. Boynton receive 04 Inches of rain In 1976. 30 Inches in 1977 and 15 inches in 1978. How many inches of rainfall did Boynton receive in 1979?

Range and Standard Deviation

1) Measures of spread tell us how far apart numbers on a list are from each other.
2) Range = max — min
3) If all the numbers on a list are identical, then the SD = 0
4) If all the numbers on a list are the same distance from the mean, SD = that distance
5) Lots of points close to the mean small SD; lots of points far from the mean large SD
6) List ± K doesn't change SD

Practice Problems

Y	Frequency
$\frac{1}{2}$	2
$\frac{3}{4}$	7
$\frac{5}{4}$	8
$\frac{3}{2}$	8
$\frac{7}{4}$	9

The table above shows the frequency distribution of the values of a variable Y. What is the mean of the distribution?

Practice Problems

Y	Frequency
$\frac{1}{2}$	2
$\frac{3}{4}$	7
$\frac{5}{4}$	8
$\frac{3}{2}$	8
$\frac{7}{4}$	9

The table above shows the frequency distribution of the values of a variable Y. What is the mean of the distribution?

1.29

Practice Problems

DISTRIBUTION OF THE
HEIGHTS OF 80 STUDENTS

Height (centimeters)	Number of Students
140–144	6
145–149	26
150–154	32
155–159	12
160–164	4
Total	80

. The table above shows the frequency distribution of the heights of 80 students. What is the least possible range of the heights of the 80 students?

Practice Problems

AGE DISTRIBUTION OF EMPLOYEES OF A BUSINESS

Age Interval	Number of Employees
15–24	17
25–34	24
35–44	26
45–54	21
55–64	18
Total	106

The range of the ages of the 20 oldest employees of the business

Range

- **Coefficient of Range or Coefficient of Dispersion:** The coefficient of range or coefficient of dispersion is a relative measure of dispersion and is given by:

$$Coefficient\ of\ Range = \frac{X_m - X_0}{X_m + X_0}$$

- **Numerical example of Range and Coefficient of range**

Ex # The marks obtained by 9 students are given below:

x_i	45	32	37	46	39	36	41	48	36

Contd..

- **Quartile Deviation or Semi-inter-quartile Range:** "half of the difference between the upper quartile and lower quartile is called the semi-inter quartile range or quartile deviation"i.e.

$$Quartile\ deivation = \frac{Q_3 - Q_1}{2}$$

- **Coefficient of Quartile Deviation:** The coefficient of quartile deviation is a relative measure of dispersion and is given by:

$$Coefficient\ Q.D = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Mean Deviation

Mean Deviation (M.D.)

Individual

$$\text{M. D.} = \frac{\Sigma |D|}{N}$$

Discrete Series

$$\text{M. D.} = \frac{\Sigma f |D|}{N}$$

Continuous Series

$$\text{M. D.} = \frac{\Sigma f |D|}{N}$$

Coefficient of M.D.

$$\text{Coefficient of M.D.} = \frac{Mean\ Deviation}{Median}$$

Mean Deviation

Calculate mean deviation and coefficient of mean deviation from mean in continuous grouped case, showing the weights of 60 apples.

Weights (grams)	65--84	85--104	105--124	125--144	145--164	165--184	185--204
Frequency	09	10	17	10	05	04	05

Mean Deviation

Weight (grams)	Midpoints (x_i)	Frequency (f_i)	$f_i x_i$	$\lvert x_i - \bar{x} \rvert$	$f_i \lvert x_i - \bar{x} \rvert$
65----84	74.5	09	670.5	−48.0	432.0
85----104	94.5	10	945.0	−28.0	280.0
105----124	114.5	17	1946.5	−8.0	136.0
125----144	134.5	10	1345.0		120.0
145----164	154.5	05	772.5	12.0	160.0
165----184	174.5	04	698.0		208.0
185----204	194.5	05	972.5	32.0	360.0
				52.0	
				72.0	
		$\sum_{i=1}^{n} f_i = 60$	$\sum_{i=1}^{n} f_i x_i =$ 7350.0		$\sum f_i \lvert x_i - \bar{x} \rvert$ = 1696.0

Mean Deviation

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} \Rightarrow \bar{x} = \frac{7450.0}{60} \Rightarrow \bar{x} = 122.5 \text{ grams}$$

$$M.D = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} \Rightarrow M.D = \frac{1696.0}{60} \Rightarrow M.D = 28.27 \text{ grams} \quad \text{(Answer).}$$

Measure of Dispersion

Range

- Simplest method of studying dispersion.
- Difference b/w the value of Smallest & Largest Item

Range

Range = L - S

Coefficient of Range

$$\text{Coefficient of Range} = \frac{L-S}{L+S}$$

Standard Deviation (σ)

Variance (σ^2)

Individual

$$\sigma = \sqrt{\frac{\Sigma d^2}{N} - \left(\frac{\Sigma d}{N}\right)^2}$$

Discrete Series

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2}$$

Continuous Series

$$\sigma = \sqrt{\frac{\Sigma f d^2}{N} - \left(\frac{\Sigma f d}{N}\right)^2} \times i$$

Coefficient of Variation (C.V. | % C.V.)

$$\text{C.V.} = \frac{\sigma}{\bar{X}} \times 100$$

Coefficient of Standard Deviation

$$Coefficient\ of\ S.D = \frac{Standard\ Deviation}{Mean}$$

$$C.V = \frac{Standard\ Deviation}{Mean} \times 100$$

Standard Deviation

Ex # Calculate the variance, S.D and C.V from the following marks obtained by 9 students.

x_i	45 32 37 46 39 36 41 48 36

THANK YOU