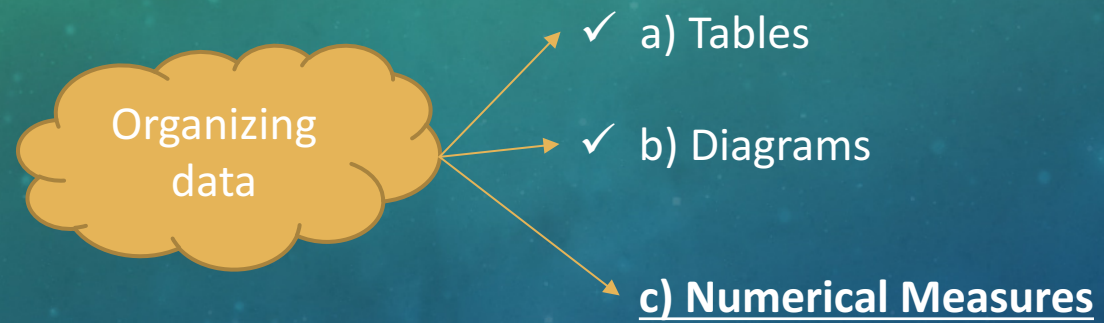


WEEK 3

# SUMMARIZING DATA: NUMERICAL MEASURES

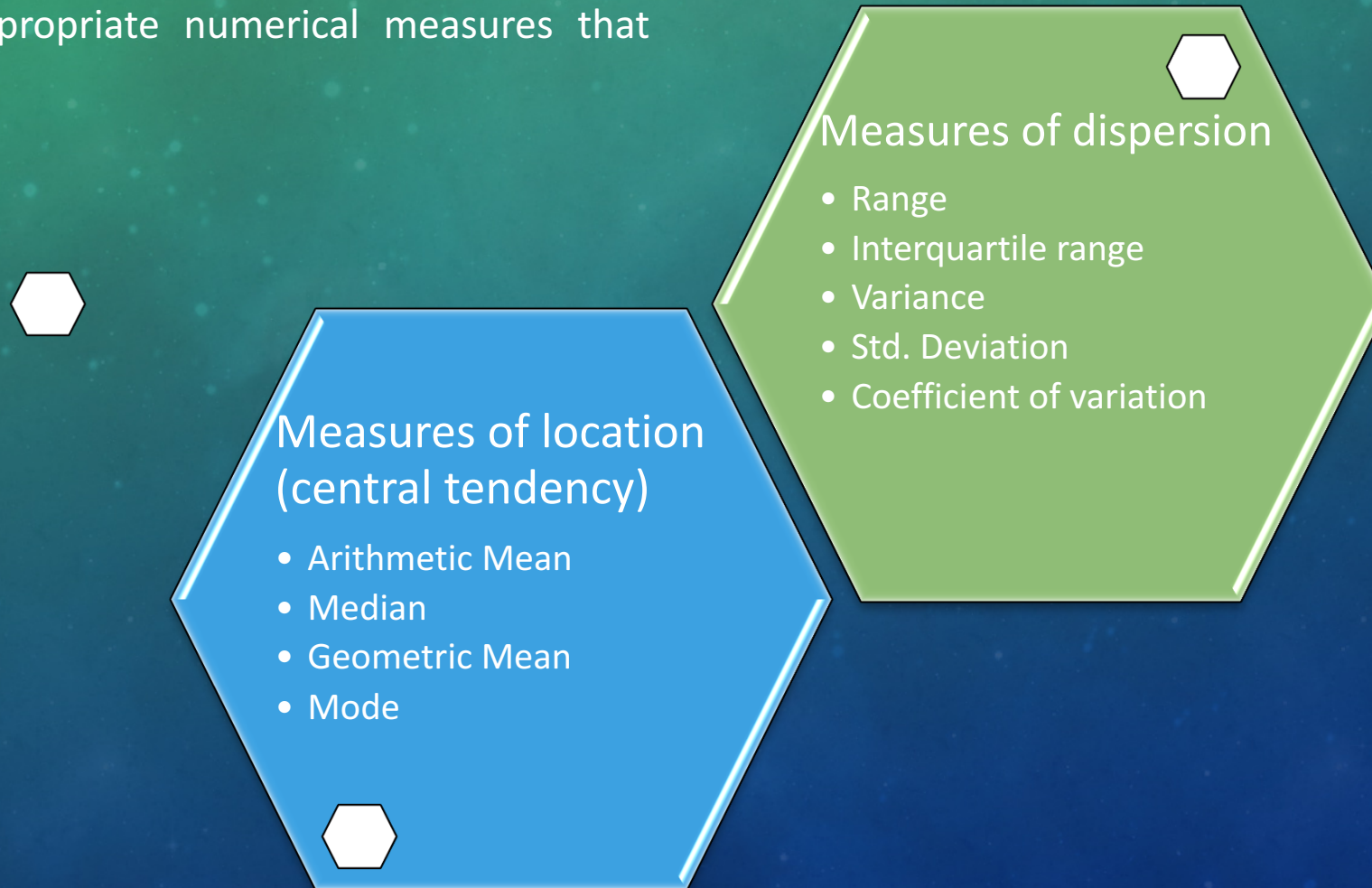
-MEASURES OF CENTRAL TENDENCY-

# SUMMARIZING DATA



# C) NUMERICAL MEASURES

It is usually essential that we supplement the visual display with the appropriate numerical measures that summarize the data.



# THE $\Sigma$ ( SIGMA) SIGN

The sign  $\Sigma$  ( sigma) is a summation sign. We can write  $(x_1+x_2+x_3+\dots x_n)$  as -->

$$\sum_{i=1}^n x_i$$

If a and b are integers and  $a < b$ , then;

$$\sum_{i=a}^b x_i$$

means  $\rightarrow x_a+x_{a+1}+x_{a+2}+\dots x_b$

**Question** : If  $x_1 = 3$ ,  $x_2 = 6$  and  $x_3 = -5$  then find the following?

a)

$$\sum_{i=1}^3 x_i = ?$$

$$\sum_{i=1}^3 x_i = 3 + 6 - 5 = 4$$

b)

$$\sum_{i=2}^3 x_i = ?$$

$$\sum_{i=2}^3 x_i = 6 - 5 = 1$$

c)

$$\sum_{i=2}^3 x_i^2 = ?$$

$$\sum_{i=2}^3 x_i^2 = 36 + 25 = 61$$

# ARITHMETIC MEAN

- Most widely used measure of central tendency !!
- Arithmetic mean is the sum of all observations divided by the number of observations.

- In statistical terms, it can be written as →

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

# LIMITATIONS OF ARITHMETIC MEAN

- In general, arithmetic mean is a natural measure of central tendency. However;
- It is oversensitive to extreme values !! 

## Arithmetic Mean

▪ Group1	:	8	7	7	7	8	8	8	26	8	8	<b>9,5</b>
▪ Group2	:	8	9	7	8	7	7	9	8	7	7	<b>7,7</b>

# MEDIAN

- An alternative measure to mean. (More precisely: sample median!)
- Suppose there are  $n$  observations in a sample. If these observations are ordered from smallest to largest, then the median can be defined as;

$\left(\frac{n+1}{2}\right)$  th largest observation if  $n$  is odd

The average of  $\left(\frac{n}{2}\right)$ th and  $\left(\frac{n}{2}\right)+1$  th largest observations if  $n$  is even.

## EXAMPLE:

The following table consist of somatic cell count measurement (x10000) of milk samples taken from 10 Holstein in a dairy farm. Compute the median value of somatic cell count.

$i$	$x_i$	$i$	$x_i$
1	11	6	8
2	21	7	9
3	18	8	110
4	14	9	12
5	13	10	20

### Solution:

*Step 1.* Order the sample from smallest to largest.

8,9,11,12,13,14,18,20,21,110

*Step 2.* Because  $n$  is even ( $n=10$ ), sample median is the average of 5<sup>th</sup> and 6<sup>th</sup> observations.

*Step 3.* Compute the sample median =  $(13+14)/2= 13,5$



# THE MODE

- The mode is the most frequently occurring value among all observations in the sample.

Value	Frequency	Value	Frequency
53	5	58	70
54	10	59	63
55	25	60	32
56	44	61	18
57	85	62	9



Please note that some distributions may have more than one mode. (e.g. unimodal or bimodal)

# THE GEOMETRIC MEAN

- Some of the laboratory data can be expressed either as multiples of 2 or as a constant multiplied by a power of 2.
- So the outcomes can be in a form of  $2^{kc}$ , where  $k=0,1,2,3,\dots$  (with a constant  $c$ )

Geometric mean = 
$$G = \sqrt[n]{x_1 + x_2 + x_3 + \dots + x_n}$$

A possible solution can be by using log-transformed observations and then taking the arithmetic mean of the observations:

$$\log G = \frac{\log x_1 + \log x_2 + \log x_3 + \dots + \log x_n}{n}$$

# EXAMPLE

- Compute the geometric mean of 3, 5, 6, 6, 7, 12 and 20.

$$\log G = \frac{\log x_1 + \log x_2 + \log x_3 + \dots + \log x_n}{n}$$

$$\log G = \frac{\log 3 + \log 5 + \log 6 + \log 6 + \log 7 + \log 12 + \log 20}{7}$$

$$\log G = \frac{0.4771 + 0.6990 + 0.7782 + 0.7782 + 0.8451 + 1.0792 + 1.301}{7} = \frac{5.9578}{7} = 0.8511$$

$$G = \text{anti log}(0.8511) = 7.097$$

Arithmetic mean = 8.43

Median = 6

Mode = 6

# A SUMMARY OF MEAN, MEDIAN AND MODE

## Mean

- the only measure whose value is dependent on the value of every core in the distribution
- more sensitive to extreme scores than the median and the mode and, hence, is not recommended for markedly skewed distributions
- not appropriate for qualitative data

# A SUMMARY OF MEAN, MEDIAN AND MODE

## Median

- widely used for markedly skewed distributions because it is sensitive only to the number rather than to the values of scores above and below it
- the most stable measure that can be used with open-ended distributions
- more subject to sampling fluctuation than the mean

## Mode

- more appropriate than the mean or the median for quantitative variables that are inherently discrete
- the only measure appropriate for unordered qualitative variables
- much more subject to sampling fluctuation than the mean and the median

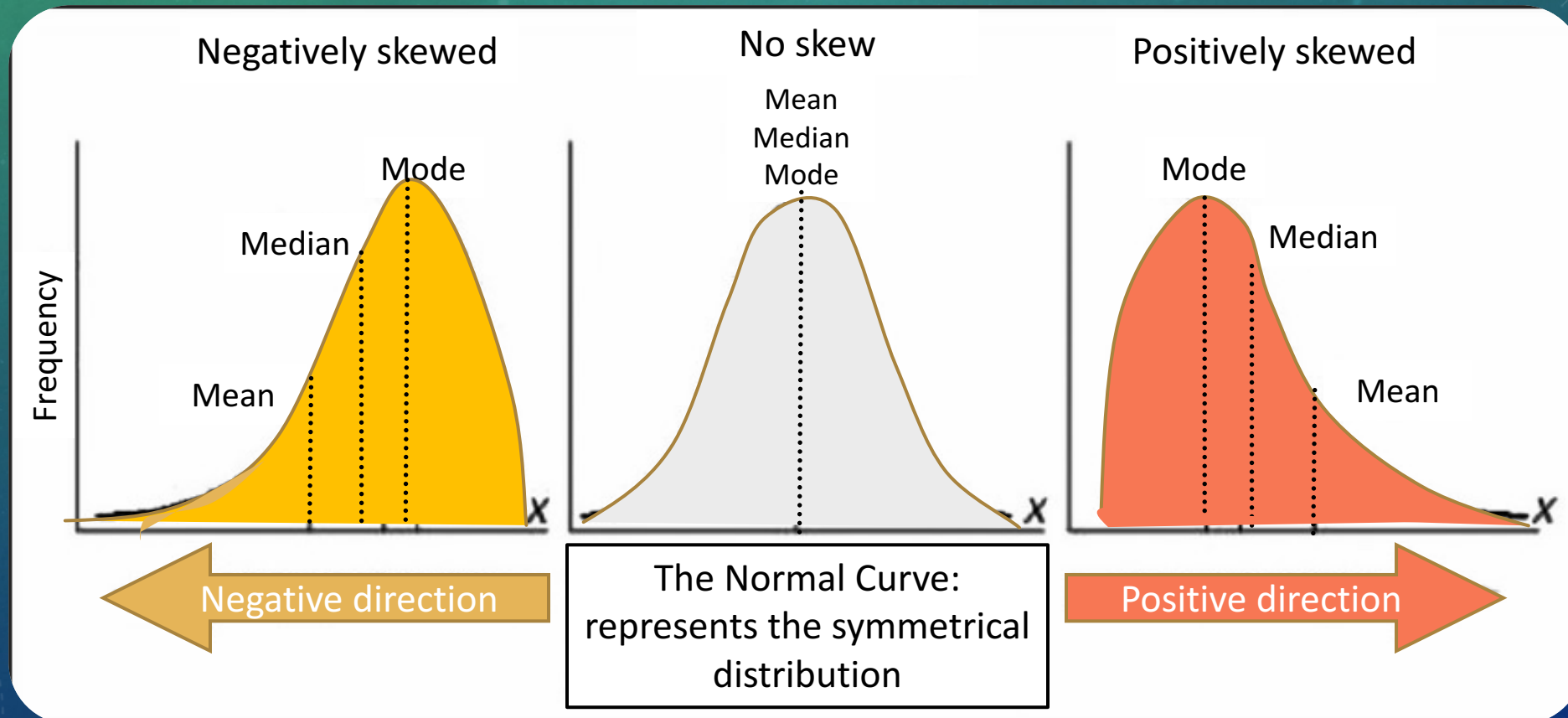
# LOCATION OF MEAN, MEDIAN AND MODE IN A DISTRIBUTION

To **skew** means to stretch in one direction.

A distribution is **skewed to the left** if the left tail is longer than the right tail.

A distribution is **skewed to the right** if the right tail is longer than the left tail.

A left-skewed distribution stretches to the left, a right-skewed to the right.





- Geometric mean is less than the arithmetic mean if the data are right-skewed.
- The geometric mean is usually equal to the mean if data are right-skewed.
- So it is preferable to use geometric mean rather than median for right skewed data.

# SUMMARIZING DATA: NUMERICAL MEASURES

-MEASURES OF DISPERSION-



• Let's say two different group of investigators collected the following data:

• Group 1: 30, 120, 130, 80, 90

• Group 2: 88, 92, 90, 86, 94



Mean = 90



To describe the data, we also need to give information about the dispersion !

## Measures of location

- Arithmetic Mean
- Median
- Geometric Mean
- Mode


## Measures of dispersion

- Range
- Interquartile range
- Variance
- Std. Deviation
- Coefficient of variation

# MEASURES OF DISPERSION

## RANGE (R)

- Defined as the distance between the largest and the smallest observations = (Max-Min)

 Range;  
can overestimate the dispersion due to extreme values.  
Tends to increase in value as the number of observations in the sample increases.

# INTERQUARTILE RANGE

- It is the difference between the first (25% -> Q1) and third quartiles (75% -> Q3).

The interquartile range;

- Is not influenced by extreme values or sample size
- Can be misleading due to ignoring most of the observations (it is calculated from only two of points)



# VARIANCE

- It is determined by calculating the *deviation* of each observation from the mean.
- This deviation will be large if the observation is far from the mean, and it will be small if the observation is close to the mean.

The sample variance is given by;

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

✓ Note that, its dimensionality is different from that of the original measurements.

# STANDARD DEVIATION (SD)

- Most important and most widely used measure of dispersion
- Equals to square root of variance !
- Unlike the variance, it is measured in the same units as the observations.
- based on every score in a distribution (so uses all observations in dataset)
- represents the square root of the mean squared distance of scores from the mean.

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



the larger the value of  $S$ , the greater is the spread or scatter of scores

# EXAMPLE

- Group 1: 30, 120, 130, 80, 90
- Group 2: 88, 92, 90, 86, 94



Mean = 90

Standard Deviation: 39,37

Mean = 90

Standard Deviation: 3,16

# COEFFICIENT OF VARIATION

- Sometimes the standard deviation is expressed as a percentage of the mean.
- It is a dimensionless quantity that can be used for comparing relative amounts of variation.

$$V = \frac{SD}{\bar{X}} * 100$$



# EXAMPLE

- Milk yielded values (lt) of first lactation of two different sheep breeds were given below;

	<u>Sakız</u>	<u>Awasi</u>
n	20	30
Mean	320.40 lt	332,5 lt
SD	48.4lt	70.4 lt

$$V = \frac{SD}{\bar{X}} * 100$$

- Compute the coefficient of variation.
- Which breed's milk yield value is more homogenously distributed?

SOLUTION:

$$a) V_{SAKIZ} = \frac{48.4}{320.4} * 100 = 15.10 \%$$

$$V_{AWASI} = \frac{70.4}{332.5} * 100 = 21.17 \%$$

b) Sakız

# DESCRIBING DATA USING MEASURES OF CENTRAL TENDENCY AND DISPERSION

Level of Measurement	Central Tendency	Dispersion
Nominal scale	Mode (most frequent category)	Number of categories
Ordinal scale	Median (data are ranked, middle value with half above and half below)	Range, Interquartile range or min-max
Continuous scale	Mean (summed and divided by number)	Standard Deviation (how much each data point deviates from the mean) or Standard Error of Mean