Descriptive statistics Post-graduate-2020-2021

Prof Dr Najlaa Fawzi

Name only the type of graph(s) used to present the following data

1-The education level of 200 women with Ca breast

2- The age distribution (years) for group of patients (35 men and 30women) with Ca-colon

3- The weight reduction (kg)for 25 diabetic patients and the Calories in take / day

4- Serum cholesterol level (mg /100ml) of 120 patients with IHD , width was 3mg /100ml

5- In certain country , the main causes of deaths are: IHD, malignancy, RTA.

For each of the following select the most appropriate answer

1-Total Relative Frequency is always:
 a-2
 b-1
 c-zero
 d-half

2-Cumulative Frequency is _____ frequency
a- increasing
b- decreasing
c-fixed
d- non of these

3- A frequency polygon is constructed by plotting frequency of the class interval and the
a- upper limit of the class
b- lower limit of the class
c- mid value of the class
d- any values of the class

4- A grouping of data into mutually exclusive classes showing the number of observations in each class
is called:
 a-Frequency polygon
b- Relative frequency
c Frequency distribution
d- Cumulative frequency

5-The suitable formula for computing the number of classes is:

a-3.322 log n
b- 0.322 log n
c- 1+3.322 log n
d- 1- 3.322 log n

6-Qualitative data can be graphically represented by using

a. histogram
b. frequency polygon
c. ogive
d. bar graph

7- Fifteen percent of the students in a school of Business Administration are majoring in Economics, 20% in Finance, 35% in Management, and 30% in Accounting. The graphical device(s) which can be used to present these data is (are)

      a.   a line graph
      b.   only a bar graph
      c.   only a pie chart
      d.   both a bar graph and a pie chart

8- The most common graphical presentation of quantitative data is a

      a.   histogram
      b.   bar graph
      c.   relative frequency
      d.   pie chart

Summary Statistics
describe data in just 2 numbers

Measures of central tendency typical average score

Measures of variability typical average variation

Measures of Central Tendency

The tendency of statistical data to get concentrated at certain values is called the "Central Tendency" and   the various methods of determining the actual value at which the data tend to concentrate are called measures of central Tendency or averages.

An average is a value which tends to sumup or describe the mass of the data.

An entire distribution can be characterized by one typical measure that represents all the observations—measures of central tendency. These measures include the mode, the median, and the mean.

The goal of measures of central tendency is to come up with the one single number that best describes a distribution of scores.

Lets us know if the distribution of scores tends to be composed of high scores or low scores.

The three measures of central tendency used in medicine and epidemiology are the

mean, the median, and, to a lesser extent, the mode

All three are used for numerical data, and the median is

used for ordinal data as well.

The Mean:

Although several means may be mathemati cally calculated,

the arithmetic, or simple, mean is used most frequently

in statistics and is the one generally referred to by the term "mean."

The mean, or average, is the sum of all the elements divided by the number of elements in the distribution. It is symbolized by µ in a population and by X ("x-bar") in a sample.

Population mean: $\mu = \sum (\sum x) / N$

µ =population mean, $\sum$ = summation, & N = no. of values in population

Sample mean: m = $(\sum x) / n$

Where n = no. of value in the sample

Application & characteristics

Simple to calculate & to understand.

It is unique (single value) & covers a lot of data in a single number

, so we achieve the aim of statistics in summarization.

Takes all the values in consideration (i.e. not going to skip any single value).

At always present even if had 2 values

it is affected by extreme values (largest & smallest value).

The age in years of 8 children with fever : 1,2,3,4,5,6,2,4

Mean = ($\sum$ x) / n= 1+ (2x2) + 3+(4x2)+ 5+6= 27 / 8 = 3.375 year

Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

MEDIAN: or the 50th percentile is the value that divides the sets of data into two equal parts (i.e. the no. of values above the median equals to the no. of values below it).

The procedure for calculating the median are as follows:

we must arrange the value in ordered array then :

find the position of the median

- **Position of median = (n + 1) / 2   this if the no. of observations is *odd*.**
  In a set of data , the observations are 49

  The position of median = n+1/ 2 = 49+1/ 2 = 50/ 2 = 25th , so the value

  of median , would be the value of the observation which rank is 25th.

  Position of median = n / 2   &   (n / 2) + 1   i.e. 2 sites if the no.
  of observations is *even*.

For example, in a distribution consisting of the elements 6, 9, 15, 17, 24, the median would be 15.
If the distribution were 6, 9, 15, 17, 24, 29, the median would be 16 (the average of 15 and 17).
The median responds only to the number of scores above it and below it, not to their actual values.
If the above distribution were 6, 9, 15, 17, 24, 500 (rather than 29), the median would still be 16—
so the median is insensitive to small numbers of extreme scores in a distribution; therefore, it is a very useful measure of central tendency for highly skewed distributions.

The median is sometimes symbolized by Mdn. It is the same as the 50th centile (C50

Advantages of the median:

simple to calculate & to understand, unique & the most

important; not affected by extreme values.

Disadvantages:

it neglects all the values & takes only the median one.

Both mean & median are used only in quantitative variables & not in qualitative variables.

Q/ when shall you prefer to use the median over the mean?

MODE:

The mode is the observed value that occurs with the greatest frequency. It is found by simple inspection of the frequency distribution (it is easy to see on a frequency polygon as the highest point on the curve).
If two scores both occur with the greatest frequency, the distribution is bimodal; if more than two scores occur with the greatest frequency, the distribution is multimodal.
The mode is sometimes symbolized by Mo.
The mode is totally uninfluenced by small numbers of extreme scores in a distribution.

It is the only measure that can be used in both qualitative & quantitative variables.

Useful in describing the distributions of occurrence of observations .

Need no calculation and its simple & easy to determine .

24, 25, 26, 27, 28, 28, 29, (30,) 31, 31, |32, 32, 32,| 34, 36

↑ Median   ↑ Mode

The most frequently occurring score (mode) is Vanilla.

Flavor	*f*
Vanilla	28
Chocolate	22
Strawberry	15
Neapolitan	8
Butter Pecan	12
Rocky Road	9
Fudge Ripple	6

Unimodal Distribution
-One Mode-

Data A	
scores	f
45	1
44	2
43	3
40	5
39	2
37	1
30	3

Bimodal Distribution
–Two Modes-

Data B	
scores	f
45	1
44	4
43	3
40	5
39	2
37	5
30	1

Patient	Distance (x)	
1	5	
2	9	
3	11	
4	3	
5	12	
6	13	
7	12	
8	6	
9	13	
10	7	
11	3	
12	15	
13	12	
14	15	
15	5	
---	$\sum x = 141$	

Example: a sample of 15 patients making visits to a health center had traveled these distances (miles) :
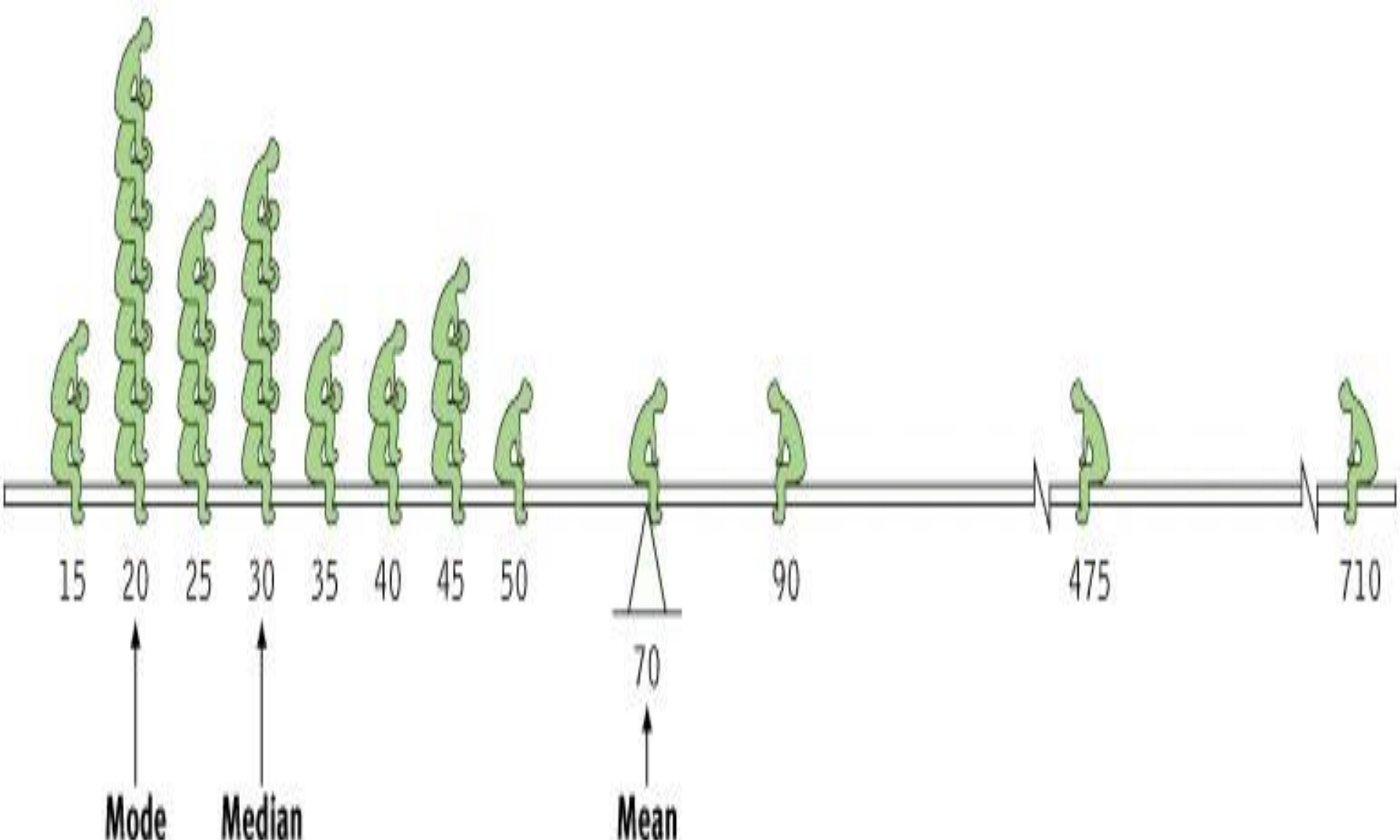
Ordered array (3,3,5,5,6,7,9,<u>11</u>,12,12,12,13,13,15,15)

Mean = m = $\sum$ x / n = 141 / 15 = 9.4 mile

The position of the median = (n +1)/2 = 8th reading &

median = 11 mile

Mode = 12 mile

15 20 25 30 35 40 45 50 90 475 710

↑Mode ↑Median ↑Mean

70

One family   Income per family in thousands of dollars

MEASURES OF CENTRAL TENDENCY IN GROUPED DATA:

we assume that the value fall into a particular class interval & all of Mean& mode are located at the midpoint of that class interval while

for the median we assume that all values are equally distributed through out the class intervals.

To find the mean:

Find the midpoint of each class interval, and then multiply it by the corresponding frequency.

Take the summation of midpoint * frequency & divide it by the summation of frequency.

$$m = \sum m p * f / \sum f$$

To find the mode: find modal class: it is the class interval having the highest frequency. Modal point is the midpoint of modal class.

In the mode we assume that all the value of a class interval in its midpoint.

To find the **median:** we assume that all the values are equally distributed through out the class interval.

The class interval containing the median is known by knowing both the position of median **(n / 2)** & the cumulative frequency.

Then the value of median is the mid point of C.I where the median is lie.

Other way to computing the value of median is by using the following Formula:

$$\boxed{\text{Median} = L + [(r / f) (U - L)]} \quad \text{OR } W$$

L is the lowest limit of class interval containing the median.

U is the upper limit of that interval

r = no. of observation in the class interval containing the median remaining to reach the median

= (n/2) – cumulative frequency of previous class interval,

F = frequency of the same class interval.

Other way for computing the median in grouped data ,from the cumulative relative frequency percent, since the median represent the 50th percentile.

The value of the median will be the midpoint of C.I where the 50% lie.

Finally the value of the median can be computed by using the cumulative relative frequency distribution curve.

Age (years)	Mid point (m.p)	Freq. (f)	Mp .f	CUM F	Cum rf %
10-19	14.5	5	72.5	5	8.77
20-29	24.5	19	465.5	24	42.1
30-39	34.5	10	345.0	34	59.64
40-49	44.5	13	578.5	47	82.44
50-59	54.5	4	218.0	51	89.46
60-69	64.5	4	258.0	55	96.48
70-79	74.5	2	149.0	57	99.99
Totals	---	57	2086.5		

Mean = m = $\sum$ m p*f / $\sum$ f = 2086.5/57 = 36.6 year

Median: position = 57/2= 28.5, so it is in the 3ʳᵈ C.I
Median =34.5 year

Median = L + [(r/ f) (U – L)] = 30 + [(4.5/10) (39-30)] = 34.05
Median = L + [(r/ f) * w = 30+ [4.5/10 x 10= 34.5 year

r=n/2- previous cum fr before the position of median
r= 28.5-24= 4.5

**Mode = midpoint of the modal class (of highest frequency) = 24.5y**

Bl .urea	frequency	mp	Rf%	Cum fr	Cum r f %
54-61	8	57.5	40	8	40
62-69	2	65.5	10	10	50
70-77	2	73.5	10	12	60
78-85	5	81.5	25	17	85
86-93	3	89.5	15	20	100
Total	20		100		

A random sample of 40 adult males with DM   presented in the following table according to their age:

Ages	Frequency
10-19	4
20-29	6
30-39	12
40-49	10
50-59	8
Total	40

1- Complete the table

2- Present this data by graph(s)

3- Competed the central tendency measurements

4- Competed the dispersion measurements

Summary of Scale Types
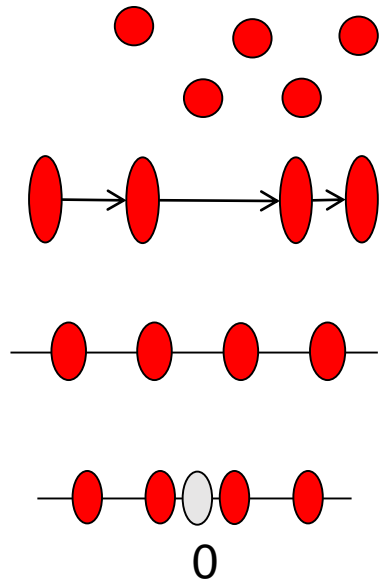
Scale		Mode	Median	Mean
Nominal		✔	✖	✖
Ordinal		✔	✖	✖
Interval		✔	✔	✔
Ratio		✔	✔	✔

MEASURES OF DISPERSION

VARIATION

**Variation is the heart of statistics
no variation, no need to do statistical analysis
the mean would describe the distribution.**

Variability is essentially a normal character . In other words, occurrence of variability is a

biological phenomenon.

Types of Variability

There are three main types of variability :
1- Biological variability
2- Real variability
3- Experimental variability

➢ **Biological Variability include the following :**

1- individual variability

2- periodical variability

3- Class, group of category variability

4- Sampling variability

➢ **Real Variability**

REAL VARIABILITY

– When the difference between two readings or observations or values of classes or samples is more than the defined limits in the universe, it is said to be real variability. The cause is external factors. e.g. significant difference in cure rates may be due to a better drug but not due to a chance.

➢ **Experimental Variability : three types-**

1- Observer error

a- subjective

b- objective

2- Instrumental error

3- Sampling defects or errors of bias

1- Observer variation

Variations in recording observations arise for several reasons including bias, errors, and lack of skill or training.

There are two principal types:

a-Inconsistency in recording repeat results – intra-observer variation

b-Failure of different observers to record the same results – inter-observer variation

Subject variation

Differences made on the same subject on different occasions may be due to several factors, including:

  Physiological changes – e.g. blood pressure, pulse

  Factors affecting response to a question – e. g relationship with the interviewer

  Changes because the participant is aware they are being studied – e.g. good manners bias, giving the answer they believe the interviewer wants to hear.

Technical limitations
Technical equipment may give incorrect results for several reasons, including:
- The method is unreliable – e.g. peak flow rate in asthma
- Faults in the test system – e.g. defective instruments, poor calibration
- Absence of an accurate test

Avoiding variation
Prior to starting on data collection, careful thought should be given to potential sources of error, bias and variation in measurements, and every effort made to minimize them. Principles include:
- Clearly defined diagnostic criteria
- Observing participants in similar biological conditions
- Training of observers
- Blinding observers and participants to the study hypothesis
- Simple equipment that is easy to use
- Standardized measurement methods
- Piloting questionnaires to identify unclear questions

The dispersion of set of observations refers to the variety that the values of the observations exhibit.

If all the values are the same , there is no dispersion ; if they are not all the same , dispersion is present in the data .

The amount of dispersion may be small , when the values though different , are close together .

If the values are widely scattered , the dispersion is greater .

Other terms used for dispersion include variation, spread ,and scatter.

There are three important measures of variability: range, variance, and standard deviation, Coefficient of variation

RANGE

The range is the simplest measure of variability. It is the difference between the highest and the
lowest scores in the distribution. It therefore responds to these two scores only.
For example, in the distribution 6, 9, 15, 17, 24, the range is (24 - 6) =18, but in the distribution
6, 9, 15, 17, 24, 500, the range is (500- 2 6) =494.

$$R = XL - XS$$

Properties of the range:

• **Simple to calculate & easy to understand.**

*It is not based on all observations.
It neglects all the values in the center & depends on
the extreme values only.

*It is not amenable for further mathematical treatment.

•**It a poor measure of dispersion.**

•It should be used in conjunction with other measures of
variability.

The most informative and frequently used measures of dispersion are :

The variance and its related function , the standard deviation.

2- VARIANCE (V) S_2:

Is the sum of the squared deviations of the observations **from their mean divided by the sample size minus one**

A) Variance of the population (δ_2) = [($\sum x_2$) – ($\sum x)_2$ /N] / N
B) Variance of sample (S_2) = $\sum(x – m)_2$ / (n – 1)

Calculating variance (and standard deviation) involves the use of **deviation scores**. The deviation score of an element is found by subtracting the distribution's mean from the element.
A deviation score is symbolized by the letter x (as opposed to X, which symbolizes an element); so the formula for deviation scores is as follows:

$$X - m = d \; ; \qquad (x - m)2 \rightarrow d2$$

$$(S2) = \sum d2 / n{-}1$$

For example, in a distribution with a mean of 16, an element of 23 would have a deviation
score of (23 - 16) =7. On the same distribution, an element of 11 would have a deviation
score of (11 - 16)= - 5.

The variance of a sample is computed by:

1- Calculating the difference between each observation (X) and the sample mean (m).

2- These differences are squared , so the negative and positive deviations will not cancel each other out.

3- The products are added together

4- The sum of the squared deviation is divided by the total number of observations minus one (n-1)[degree of freedom].

When the number of observations is large , the following equation is use:

$$\{(\sum x2) - (\sum x)2] / n\} / (n-1)$$

In calculating the S2 in grouped data , we assume that all the values falling into a particular C.I are located at the mid point of the C.I.

To find the <u>variance</u>: S2 =

$$[(\sum mp2*f) - (\sum mp*f)2] / n /(n-1)$$

Variance is sometimes known as mean square.
Variance is expressed in squared units of measurement, limiting its usefulness as a descriptive term—its natural meaning is poor.

STANDARD DEVIATION

The standard deviation remedies this problem: it is the
square root of the variance, so it is expressed
in the same units of measurement as the original data.
The symbols for standard deviation are therefore the
same as the symbols for variance, but without being raised
to the power of two, so the standard deviation of a
population is σ and the standard deviation of a sample is *S*.
Standard deviation is sometimes written as *SD*.

$$SD = \sqrt{S2} \qquad \text{for the sample}$$
$$SD = \sqrt{δ2} \qquad \text{for the population}$$

High SD indicates that the data is dispersed away from the mean.

Low SD indicates that the data points tend to be very close to mean

The standard deviation, like the mean, requires numerical data.
 Also, like the mean, the standard deviation is a very important
 statistic.
 First, it is an essential part of many statistical tests.
Second, the standard deviation is very useful in describing the
spread of the observations about the   mean value.
 Two rules when using the standard deviation are:
 1-Regardless of how the observations are distributed,
 at least 75% of the values always lie between these two numbers

The mean minus 2 standard deviations

and the mean plus 2 standard deviations.

The standard deviation is particularly useful in normal distributions because the proportion of elements in the normal distribution (i.e., the proportion of the area under the curve) is a constant for a given number of standard deviations above or below the mean of the distribution

Approximately 68% of the distribution falls within ±1 standard deviations of the mean.
• Approximately 95% of the distribution falls within ±2 standard deviations of the mean.
• Approximately 99.7% of the distribution falls within ±3 standard deviations of the mean.
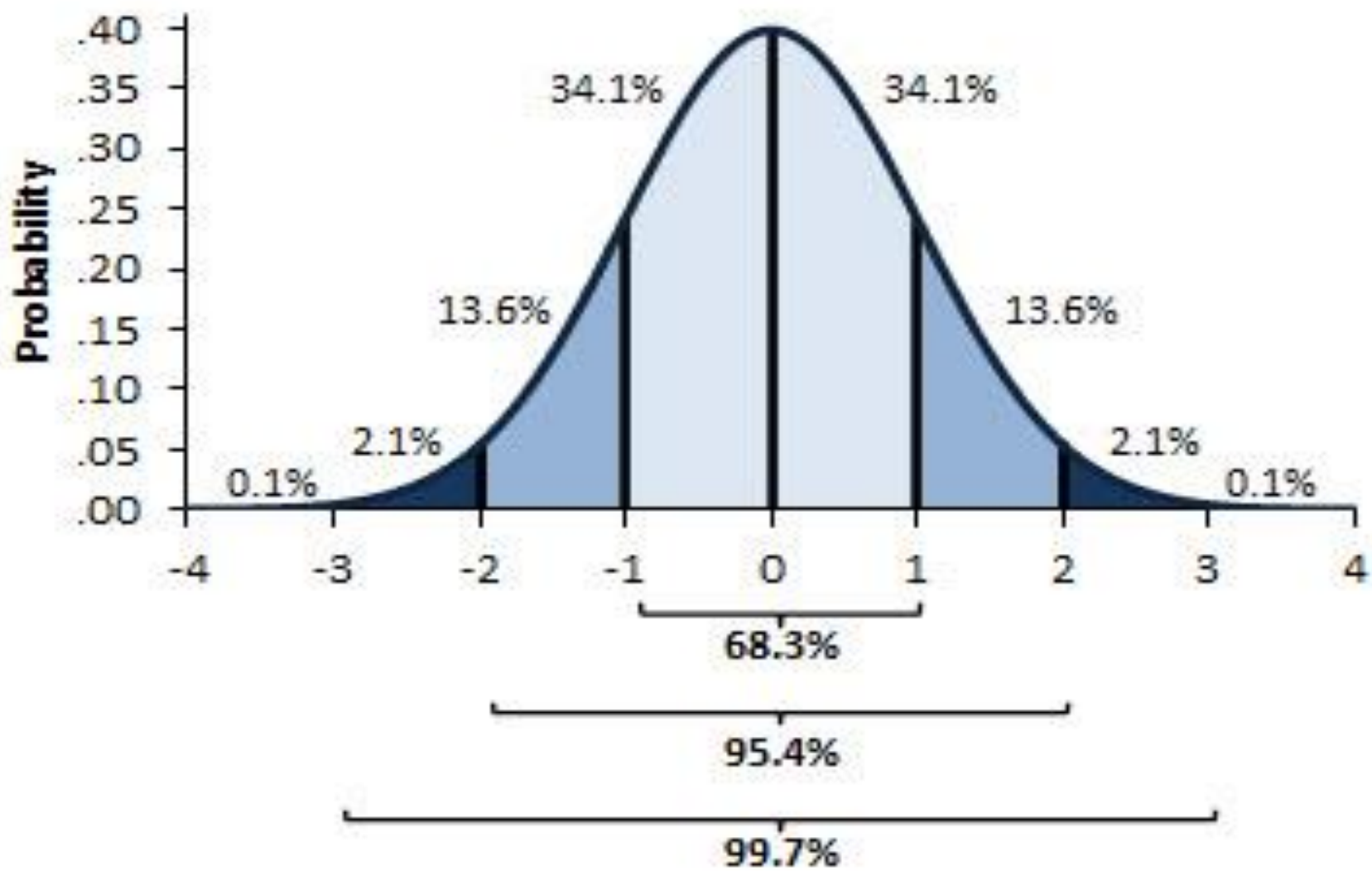
Uses of SD in biostatistics:

Summarizes the deviation of a large distribution from mean — Indicates whether the variation of difference of an individual from the mean is by chance

— Helps in finding the standard error
— Helps in finding the suitable size of sample for valid conclusions.

4- COEFFICIENT OF VARIATION (CV):

Is a useful measure of relative spread in data and is used frequently in the biological science .

The CV is defined as the SD divided by the mean times 100%

$$CV = (SD / m) * 100$$

It is useful when one desires to compared the dispersion in two sets of data, when they measured in different units .

Further more , although the same unit of measurements is used , the two means may be quite different .

What is needed in situation like these is a measure of relative variation rather than absolute variation.

The CV is also useful in comparing the results obtained by different

persons who are conducting investigation involving the same variable.

Example:

In two series of adults aged 21 years and children 3months
old following values were obtained for the height .

Persons	Mean ht	SD
Adult	160cm	10cm
Children	60cm	5cm

CV of adult = 10/ 160x 100= 6.25%

CV of children= 5/60x 100= 8.33%

STANDARD ERROR OF THE SAMPLE MEAN (SE)

it measures the variability of the sample mean around the population

mean and indicates the degree in which the sample mean reflects

the population mean.

$$SE = SD / \sqrt{n}$$

SE determines the dispersion of the sample mean from the population

mean (determines sampling error), i.e. the representative ness of

the sample to the population, so it gives us the idea for how

much the sample mean is far away from the population mean.

Diastolic Blood Pressure?

Mean = 78 mm Hg

Samples

Mean = 75

Mean = 67

Mean = 71.3
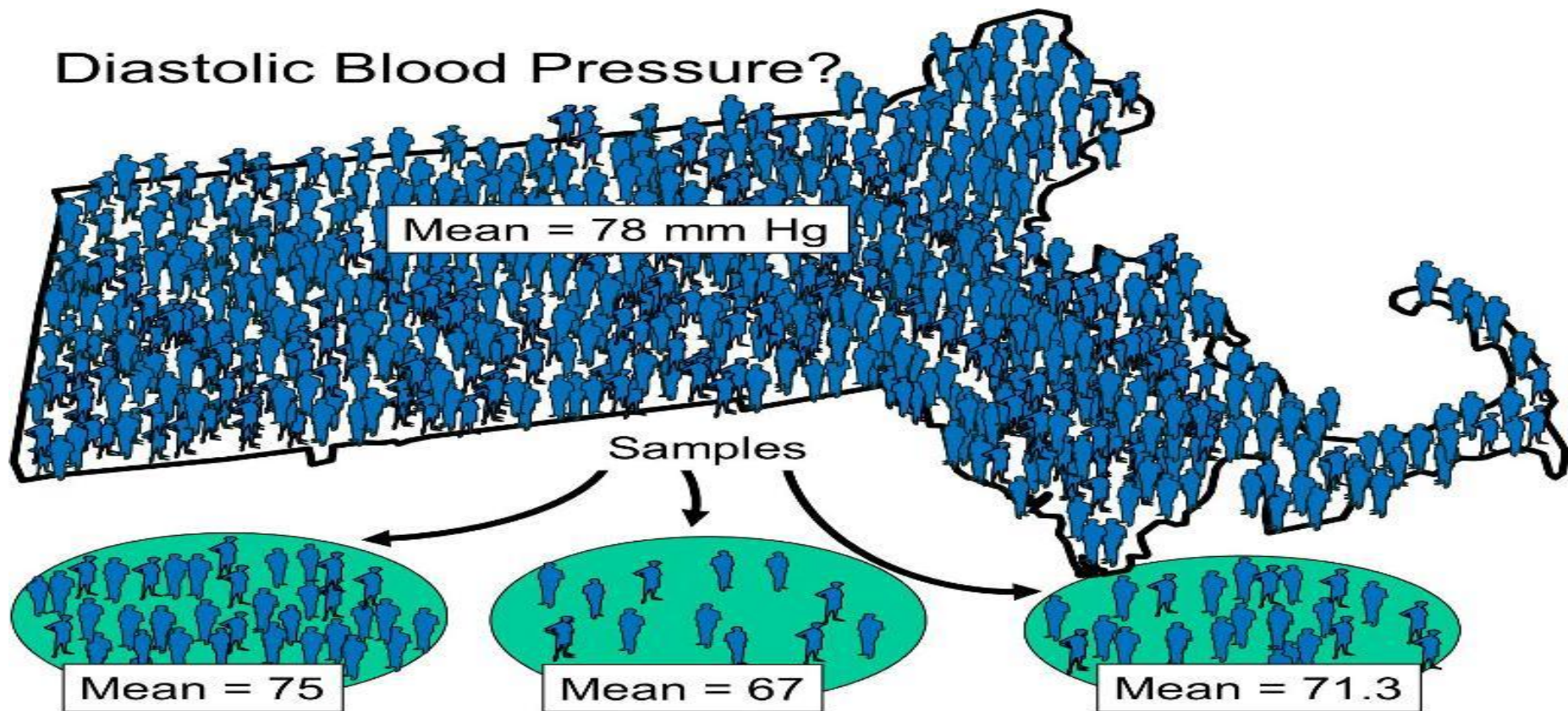
As the formula shows, the standard error is dependent on the size of the samples: standard error is inversely related to the square root of the sample size, so that the larger n becomes, the more closely will the sample means represent the true population mean.
This is the mathematical reason why the results of large studies or surveys are more trusted than the results of small ones—a fact that is instinctively obvious!

• SE is a 'measure of chance variation', and IT DOESNOT mean an error or mistake

Child no.	S f t (mm)
1	8
2	8
3	10
4	7
5	6
6	10
7	9
8	8
9	5
10	7
11	4
12	6

The fat fold at triceps (mm)

were recorded for 12 children: calculate the dispersion measures

S f t (mm) [x]	f	Xf	d =[x-m]	
4	1	4	4- 7.33= -3.33	-3.33 d2 11.08
5	1	5	5-7.33= -2.33	-2.33 d2 5.42
6	2	12	6-7.33= -1.33	-1.33x2 d2 3.52
7	2	14	7-7.33= -0.33	-0.33x2 d2 0.21
8	3	24	8-7.33= 0.67	0.67x3 d2 1.32
9	1	9	9-7.33= 1.67	1.67x1 d2 2.78
10	2	20	10-7.33= 2.67	2.67x2 d2 7.12
Total	12	88		$\sum$ **d2 = 31.45**

R = XL – XS

10- 4= 6 mm

(S2) = ∑d2/ n-1

31.45/ 11 = 2.85

SD = √ S2

SD = √2.85= 1.68mm

CV = (SD /m) *100

CV=1.68/ 7.33* 100=   22.91%

SE = SD / √ n

SE = 1.68/ √12 = 1.68/ 3.46

SE= 0.48

Bl urea	f	mp	mp2	Mp2*f
54-61	8	57.5	3306.25	26450
62-69	2	65.5	4290.25	8580.5
70-77	2	73.5	5402.25	10804.5
78- 85	5	81.5	6642.25	33211.25
86- 93	3	89.5	8010.25	24030.75
Total	20			**103077**

$\sum m p*f = 1414$

$\sum m p*f)2 = 1999396$

M=1414/20= 70.7

$S2 = [(\sum mp2*f) - (\sum m p*f)2] / n /(n-1)$

[103077 − 1999396/ 20] / 20 -1= 3107.2 /19= 163.53

SD = √ S2 = 12.788 mg

CV = (SD /m) *100 = 12.788/ 70.7x 100= 18.076%