Summarizing Data

The basic problem of statistics can be stated as follows : Consider sample of data X_1 , …, X_n , where X_1corresponds to the first sample point and X_n corresponds to the *n*th sample point. Presuming that the sample is drown from some population *P*, what inferences or conclusion can be made about *P* from the sample ?

Before this question can be answered , the data must be **summarized** as succinctly as possible ; this is because the number of sample points is often large , and it is easy to lose track of the overall picture when looking at individual sample points. One sample. This type of measure is a **measure of location(Measures of Central Tendency**).

Measures of Central Tendency

1- The Arithmetic Mean

One measure of location is the arithmetic mean (colloquially called the average). The arithmetic mean (or mean or sample mean) is usually denoted by $\overline{X}$.

The **arithmetic mean** is the sum of all the observations divided by the number of observations. It is written in statistical terms as

$$\overline{X} = \frac{X1+X2+\cdots+Xn}{n} = \frac{\sum_{i=1}^{n} Xi}{n}$$

The sign $\sum$ (sigma) is a summation sign. The expression $\sum Xi$ is simply a short way of writing the quantity($X1 + X2 + \cdots + Xn$) .

One properly of summation signs is that if each term in the summation is a multiple of the same constant C , then C can be factored out from the summation ; that is ,

$\sum_{i=1}^{n} CXi = C\ [\sum_{i=1}^{n} Xi]$

Example 1:. If $X_1=2$ $X_2= 5$ $X_3= -4$

Find $\sum_{i=1}^{3} Xi$ $\sum_{i=2}^{3} Xi$ $\sum_{i=1}^{3} Xi^2$ $\sum_{i=1}^{3} 2Xi$

Solution :. $\sum_{i=1}^{3} Xi = 2 + 5 - 4 = 3$

$\sum_{i=2}^{3} Xi = 5-4 = 1$

$\sum_{i=1}^{3} Xi^2 = 4 + 25+ 16 = 45$

$\sum_{i=1}^{3} 2Xi = 2 \sum_{i=1}^{3} Xi = 2 \times 3 = 6$

Example 2 :.What is the arithmetic mean for the sample of birth weights in the following table

Table(1): Sample of birth weights(g) of live-born infants born at a Karbala hospital during a 1-week period

i	X_i	*i*	X_i	*i*	X_i	*i*	X_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

$\bar{X} = \frac{X1+X2+\cdots+Xn}{n} = \frac{\sum_{i=1}^{n} Xi}{n}$

$\bar{X} = (3265 + 3260 + \ldots + 2834)/ 20 = 3166.9$ g

2- The Median

An alternative measure of location, perhaps second in popularity to the arithmetic mean , is the **median** or , more precisely , the **sample median.**

Suppose there are *n* observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:.

The sample median is

(1) The [(n+1) /2] the largest observation if n is odd.

(2) The average of the [n/2]th and [(n/2) +1]th largest observations if n is even.

The rational for these definitions is to ensure an equal number of sample points on both sides of the sample median.

Example 3:. Compute the median for the sample in Table 1.

Solution :. First, arrange the sample in ascending order:.

2069 , 2581 , 2759 , 2834 , 2838 , 2841 , 3031 , 3101 , 3200 , 3245 , 3248 , 3260 , 3265 , 3314 , 3323 , 3484 , 3541 , 3609 , 3649 , 4146.

Because n is even ,

Sample median = average of the 10th and 11th largest observations = (3245 + 3248)/2 = 3246.5 g

Example 4:.**Infectious Disease.** Consider the data set in Table(2) , which consists of white-blood counts taken on admission of all patients entering a small hospital in Baghdad city on a given day. Compute the median white-blood count.

Table(2): Sample of admission white-blood counts (x 1000) for all patients entering a hospital in Baghdad city on a given day.

i	X_i	i	X_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Solution: First , order the sample as follows: 3 , 5 , 7 , 8 , 8 , 9 , 10 , 12 , 35

Because n is odd, the sample median is given by the fifth largest point, which equals 8 or 8000 on this original scale.

3- The Mode

The mode is the most frequently occurring value among all the observation in a sample.

Example 5: Consider the sample of time intervals between successive menstrual periods for a group of 500 college women age 18 to 21 years , shown in Table(3). The frequency column gives the number of women who reported each of the respective durations.

Table(3)

Value	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Frequency	5	10	28	64	185	96	63	24	9	2	7	3	2	1	1

The mode is 28 because it is the most frequently occurring value

Example 6:Compute the mode of the distribution in Table(2).

Solution: The mode is 8000 because it occurs more frequently than any other white-blood count.

Some distributions have more than one mode. In fact , one useful method of classifying distributions is called unimode ; two modes, bimodal, three mode, trimodal ; and so forth.

Example 7 :Compute the mode of distribution in Table(1).

Solution :There is no mode , because all the values occur exactly once.

4- The Geometric Mean (GM)

Many types of laboratory data, specifically data in the form of concentrations of one substance in another, as assessed by serial dilution techniques , can be expressed either as multiples of 2 or as a constant multiplied by a power of 2 ; that is, outcomes can only be of the form 2^kC , k = 0 , 1 , …, for some constant C. For example, the data in Table(4) represent the minimum inhibitory concentration (MIC) of penicillin G in the urine for *N. gonorrhoeae* in 74 patients. The arithmetic mean is not appropriate as a measure of location in this situation because the distribution is very skewed.

However , the data do have a certain pattern because the only possible values are of the form $2^k(0.03125)$ for k = 0, 1, 2, … . One solution is to work with the distribution of the logs of the concentrations. The log concentrations have the property that successive possible concentrations differ by a constant; that is,

$\log(2^{k+1}C) - \log(2^kC) = \log 2^{k+1} + \log C - \log 2^k - \log C$

$= (k+1)\log 2 - k \log 2 = \log 2$.

Thus the log concentrations are equally spaced from each other , and the resulting distribution is now not as skewed as the concentrations themselves. The arithmetic mean can then be computed in the log scale ;

Table(4) Distribution of minimum inhibitory concentration(MIC) of penicillin G for *N. gonorrhoeae* .

Concentration(µg/m)	Frequency	Concentration(µg/mL)	Frequency
$0.03125 = 2^0(0.03125)$	21	$0.250 = 2^3(0.03125)$	19
$0.0625 = 2^1(0.03125)$	6	$0.50 = 2^4(0.03125)$	17
$0.125 = 2^2(0.03125)$	8	$1.0 = 2^5(0.03125)$	3

$\overline{\text{Log X}} = 1/n \quad \sum \log X_i$

And used as a measure of location. However, it is usually preferable to work in the original scale by taking the antilogarithm of $\overline{\log X}$ to form the geometric mean , which leads the following definition :.

The geometric mean is the antilogarithm of $\overline{\log X}$, where

$\overline{\text{Log X}} = 1/n \quad \sum \log X_i$

Example 7: **"Infectious Disease"**. Compute the geometric mean for the sample in Table (4).

***Solution* :.**

(1) For convenience , use base 10 to compute the logs and antilog in this example.

(2) Compute:

$\overline{\text{Log X}}$ =[21 log(0.03125) + 6 log(0.0625) + 8 log(0.125) + 19 log(0.250) + 17 log(0.50) + 3 log(1)] / 74

= - 0.846

(3) The geometric mean=the antilogarithm of $-0.846 = 10^{-0.846} = 0.143$

The geometric mean is preferable to the arithmetic mean if the series of observations contains one or more unusually large values. The above method of calculating geometric mean is satisfactory only if there are a

small number of items. But if n is a large number, the problem of computing the nth root of the product of these values by simple arithmetic is a tedious work. To facilitate the computation of geometric mean we make use of logarithms. The above formula when reduced to its logarithmic form will be:

$GM = \sqrt[n]{(X_1)(X_2)\ldots.(X_n)} = \{ (X_1)(X_2)\ldots (X_n) \}^{1/n}$

$\text{Log GM} = \log \{(X_1)(X_2)\ldots.(X_n)\}^{1/n}$

$= 1/n \log \{(X_1)(X_2)\ldots.(X_n)\}$

$=1/n \{\log(X_1) + \log(X_2) + \ldots \log(X_n)\}$

$= \Sigma (\log X_i)/n$

The logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of individual values. The actual process involves obtaining logarithm of each value, adding them and dividing the sum by the number of observations. The quotient so obtained is then looked up in the tables of anti-logarithms which will give us the geometric mean.

***Example* 8**: The geometric mean may be calculated for the following parasite counts per 100 fields of thick films.

7	8	3	14	2	1	440	15	52	6	2	1	1	25
12	6	9	2	1	6	7	3	4	70	20	200	2	50
21	15	10	120	8	4	70	3	1	103	20	90	1	237

$GM = \sqrt[42]{7 \times 8 \times 3 \times \ldots \times 1 \times 237}$

$\log GM = 1/42 (\log 7 + \log 8 + \log 3 + .. + \log 237)$

$= 1/42 (0.8451 + 0.9031 + 0.4771 + \ldots 2.3747)$

$= 1/42 (41.9985)$

$= 0.9999 \approx 1.0000$

The anti-log of 0.9999 is 9.9992 ≈10 and this is the required geometric mean. By contrast, the arithmetic mean, which is inflated by the high values of 440, 237 and 200 is 39.8 ≈ 40.