Introduction

Statistics is numerical data relating to an aggregate of items (e.g. individuals) or numbers derived from the data (e.g. averages , rates) .

Statistics is the science of collecting , summarizing , presenting , organizing and analyzing data. As well as drawing valid conclusion and making reasonable decisions on the basis of such analysis .

Biostatistics

The science of biostatistics embraces those techniques pertaining to biology field. When the different statistical methods are applied in biological, medical and public health data.

Population and Sample

In collection data concerning characteristics of a group individuals or objects , such heights and weights of students in a university or numbers of defective and non-defective bolts produced in a factory on a given day , it is often impossible or impractical to observe the entire group , especially if it is large .

Instead of examine the entire group , called the **population** , one examines a small part of the group , called the **sample** .

Classification of Statistics

1- Descriptive statistics :. The phase of statistics which seeks only to describe and analyzing a given group without drawing any conclusion or inferences about a larger group is called **descriptive or deductive statistics** .

2- Statistics inference or Inductive statistics :. If a sample is representative of a population , important conclusions about the population can often be inferred from analysis of the sample. The phase of statistics dealing with

conditions under which such inference is valid is called **inductive statistics** or **statistics inference** , because such inference cannot be absolutely certain.

Definition of terms

Variable

A variable is a symbol , such as X , Y , H , ……. etc. which can assume any of a prescribed of values , called the domain of the variable. If the variable assumes are value , is called a **constant variable**.

A variable which can assume any value between two given values is called a **continuous variable**. Otherwise , is called a **discrete variable.**

Example :. The number of children in family which can assume any of the values 0 , 1 , 2 , 3 , …… is **discrete variable.**

Example :.The age A of an individual , which can be 62 years , 63.8 years or 65.83 years depending on the accuracy of measurements is called **a continuous variable.**

Data which can be described by a discrete or continuous variable are called **discrete data** or **continuous data,** respectively.

The number of children in each of 1000 families is an example of discrete data , while the heights of 100 university students is an example of continuous data. In general measurements give rise to continuous data , while enumeration or counting give rise to a discrete data.

Observation

This is an event which is seen occur , a number expressed as the value of a variable , is usually assigned to the event , and this is the measurement. The term " **observation** " is an expression of both the event and the measurement.

Measures of Central Tendency

Index or Subscript Notation

Let the symbol X_j (read X sub j) denote any of the n values X_1 , X_2 , ……., X_n .

Assumed by a variable X , which can stand for any of the numbers 1 , 2 , 3 , ……, n is called a **subscript** or **index**. Any letter other than j such as i , k , p ,q and s could have been used as well.

Summation Notation

The symbol $\sum_{j=1}^{n} Xj$ is used to denote the sum of all the X_j from j=1 to j=n , i.e. by definition:.

$\sum_{j=1}^{n} Xj = X_1 + X_2 + \ldots\ldots\ldots + X_n$.

Sometimes we use the following symbols (brief symbols / for summation)

ΣX or ΣX_j or $\Sigma_j X_j$

Averages or Measures of Central Tendency

Average

An average is a value which is typical or representative of a set of data. Since such typical values tend to lie centrally within a set of data arranged according to magnitude. Averages are also called measures of central tendency.

Several types of averages can be defined , the most common being **the arithmetic mean** (the mean) , **the mode** , **the geometric mean** and **the harmonic mean**. Each has advantages and disadvantages depending on the data and the intended purpose.

The arithmetic mean :.

The arithmetic mean (the mean) of a set of n numbers X_1 , X_2 , ……., X_n is denoted by $\overline{X}$ (read X bar) and is defined as :.

$$\overline{X} = \frac{X1+X2+\cdots+Xn}{n} = \frac{\sum Xi}{n}$$

Example :. The arithmetic mean of the numbers 8 , 3 , 5 , 12 , 10 is :.

$$\overline{X} = \frac{8+3+5+12+10}{5} = 7.6$$

If the numbers X_1 , X_2 , ……., X_n occurs F_1 , F_2 ,.., F_n times respectively lie occurs with frequencies F_1 , F_2 ,.., F_n , the arithmetic mean :.

$$\overline{X} = \frac{F1X1+F2X2+\cdots+FnXn}{F1+F2+\cdots+Fn} = \frac{\sum FiXi}{\sum Fi}$$

Example :. If 5 , 8 , 6 , 2 occurs with frequencies 3 , 2 , 4 , 1 respectively , find the arithmetic mean of this set ?

Solution :.

$$\overline{X} = \frac{(5\times3)+(8\times2)+(6\times4)+(2\times1)}{3+2+4+1} = \frac{\mathbf{57}}{\mathbf{10}} = 5.7$$

*Example :.*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

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

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

The median

The median of a set of numbers arranged in order of magnitude is the middle value or the arithmetic mean of two middle values .

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.

*Example :.*The set of numbers 3 , 4 , 4 , 5 , 6 , 8 , 8 , 8 , 10 has median 6 .

Example :. The set of numbers 5 , 5 , 7 , 9 , 11 , 12 , 15 , 18 has the median $^1/_2$ (9+11) = 10

Example:. 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.

The mode

The mode of a set of numbers is that value which occurs with the greatest frequency , i.e. *it is the most common value*. The mode may not exist , and even exist it may not be unique.

*Example :.*The set 2 , 2 , 5 , 7 , 9 , 9 , 9 , 10 has mode 9.

*Example :.*The set 3 , 5 , 7 , 9 has no mode .

*Example :.*The set 3 , 4 , 4 , 4 , 5 , 7 , 7 , 7 , 9 has two modes , 4 and 7 .

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

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

Question :. Find the mean , median and mode for the set of numbers.

A) 3 , 5 , 2 , 6 , 5 , 9 , 5 , 2 , 8 , 6 .

B) 51.6 , 48.7 , 50.3 , 49.5 , 48.9.

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:. **"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:. 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.