8.1 Introduction to Statistics

DEFINITION

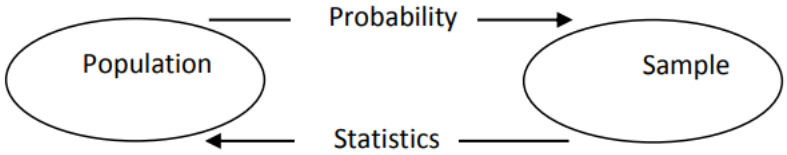
Statistics: is the branch of scientific inquiry that provides methods for organizing and summarizing data, and for using information in the data to draw various conclusions.

Descriptive Statistics: The part of statistics that deals with methods for organization and summarization of data. Descriptive methods can be used with list of all population members (a census), or when the data consists of a samples.

Population: A population of participants or objects consists of all those participants or objects that are relevant in a particular study.

Sample: A sample is any subset of the population of individuals or things under study. Probability function: is a rule, denoted by p(x) that assigns numbers to elements of the sample space.

Link between statistics and Probability



Three fundamental components of statistics

Statistical techniques consist of a wide range of goals, techniques and strategies. Three fundamental components worth stressing are:

1. Design, meaning the planning and carrying out of a study.
2. Description, which refers to methods for summarizing data.

3. Inference, which refers to making predictions or generalizations about a Population of individuals or things based on a sample of observations available to us.

8.2. Numerical Summaries of Data

✓ Summation notation

In symbols, adding the numbers $X_1, X_2, \ldots, X_n$ is denoted by

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n,$$

where ∑ is an upper case Greek sigma. The subscript i is the index of summation and the 1 and n that appear respectively below and above the symbol ∑ designate the range of the summation.

Example :

$$1.2, 2.2, 6.4, 3.8, 0.9.$$

Then

$$\sum_{i=2}^{4} X_i = 2.2 + 6.4 + 3.8 = 12.4$$

and

$$\sum X_i = 1.2 + 2.2 + 6.4 + 3.8 + 0.9 = 14.5.$$

$$\sum X_i^2 = 1.2^2 + 2.2^2 + 6.4^2 + 3.8^2 + 0.9^2 = 62.49$$

and

$$\left(\sum X_i\right)^2 = (1.2 + 2.2 + 6.4 + 3.8 + 0.9)^2 = 14.5^2 = 210.25.$$

8.3. Measures of location:

✓ The sample mean:

The first measure of location, called the sample mean, is just the average of the values and is generally labeled $\bar{X}$ The notation $\bar{X}$ is read as X bar. In summation notation,

$$\bar{X} = \frac{1}{n}\sum X_i.$$

Example 1:

You sample ten married couples and determine the number of children they have.
The results are 0, 4, 3, 2, 2, 3, 2, 1, 0, 8.

The sample mean is: $\bar{X} = (0+4+3+2+2+3+2+1+0+8)/10 = 2.5$.

Example 2

The salaries (in thousands Iraqi D) of the 11 individuals currently working at the company are:

300,250,320,280,350,310,300,360,290,2000,5000,

where the two largest salaries correspond to the vice president and president,

The average is 887, but it gives a distorted sense of what is typical!

✓ The median

Another important measure of location is called the sample median. The basic idea is easily described using the example based on the weight of trout. The observed weights were

1.1,2.3,1.7,0.9,3.1.

Putting the values in ascending order yields

0.9,1.1,1.7,2.3,3.1.

Notice that the value 1.7 divides the observations in the middle in the sense that half of the remaining observations are less than 1.7 and half are larger.

If instead we have an even number of observations, there is no middle value,

0.8, 1.3, 1.8, **2.6, 2.7**, 2.7, 3.1, 4.5

The sample median in this case is taken to be the average of 2.6 and 2.7, namely $(2.6 + 2.7)/2 = 2.65$.

8.4. Measures of variation

1. The range

The *range* is just the difference between the largest and smallest observations. In symbols, it is $X(n) - X(1)$.

2. The variance and standard deviation

The following data written in ascending order:

$$7.5,8.0,8.0,8.5,9.0,11.0,19.5,19.5,28.5,31.0,36.0.$$

The data mean is $\bar{X} = 17$, so the deviation scores are

$$-9.5,-9.0,-9.0,-8.5,-8.0,-6.0,2.5,2.5,11.5,14.0,19.0.$$

Deviation scores reflect how far each observation is from the mean, but often it is best to find a single numerical quantity that summarizes the amount of variation in our data

The average difference is always ***zero***, so this approach is unsatisfactory
The average squared difference from the mean is called the ***sample variance***, which is:

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

The ***sample standard deviation*** is the (positive) square root of the variance, S.

Example 1

The following data are the sample test results

3,9,10,4,7,8,9,5,7,8.

The sample mean is $\bar{X} = 7$,

i	X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
1	3	−4	16
2	9	2	4
3	10	3	9
4	4	−3	9
5	7	0	0
6	8	1	1
7	9	2	4
8	5	−2	4
9	7	0	0
10	8	1	1
∑		0	48

The sum of the observations in the last column is

$$\sum(Xi - \bar{X})^2 = 48.$$

So,

$$S^2 = 48/9 = 5.33.$$