Probability and Statistics (MA-2011)

Chapter-1

1.1 Population, sample and processes

Population: Scientists and Engineers depend on a collection of facts or data. An investigation is mainly focused on a well-defined collection of objects consisting a population of interest. In fact, a large number of well-defined objects or data from which the observation or investigation is to be done is called the population.
Sample: A small amount of objects or data is selected from the population that caries the behavior of the population is called the sample. Sample is a subset of population.

For example, let a size of 10 arbitrary screws be selected from a lot of 1000 screws to study the defectiveness, then population size is 1000 and the sample size is 10.

Variable: A variable is any characteristic whose value may change from one object to another in the population. For example: for a company, the variables x, y, z are
x = the brand of the object
y = quality of the object;
z = number of defective objects etc.

Data can be observed into 3 types in the sense of variable such as

a. **Univariate:** When the observations are made on a single variable, the data set is called An univariate data.
For example The sample of pulse rates (heart beats per minute) for the patients recently admitted to an adult ICU is a numerical univariate data set: 85, 81,80, 102, 115, 95, 90, 105.
b. **Bivariate:** When the observations are made on each of two variables, the data set is called bivariate.
For example The sample of saturation (the amount of oxygen bound to hemoglobin in the blood measured in percentage of the maximum binding capacity) and pulse rates (heart beats per minute) for the patients recently admitted to an adult ICU is a numerical bivariate data set: (85, 92), (81,87), (80,85), (95,102), (94,115), (95,120), (90.95), (98,102) .

c. **Multivariate:** When the observations are made on more than one variables, the data set is called multivariate. Univariate, bivariate are particular cases of multivariate.

For example: The sample of maximum blood pressure, saturation (the amount of oxygen bound to hemoglobin in the blood measured in percentage of the maximum binding capacity) and pulse rates (heart beats per minute) for the patients recently admitted to an adult ICU is a numerical trivariate data set: (130, 85, 92), (115, 81,87), (140, 80,85), (102, 95,102), (148, 94,115), (205, 95,120), (145, 90.95), (125, 98,102) .

In general, if $X_1, X_2, X_3, \cdots$ are different type of categories of the object, then multivariate data set is of form

$$X = X_1 \times X_2 \times X_3 \times \cdots = \{(x_1, x_2, \cdots : x_i \in X_i, i = 1,2, \cdots)\}.$$

1.2 Pictorial and tabular methods in Descriptive statistics

Given a data set consisting of n observations on some variable x, the individual observations will be denoted by $x_1, x_2, \cdots, x_n$. The subscript bears not relation to the magnitude of a particular observation. Thus x_1 will not in general be the smallest observation in the set, not will x_n typically be the largest. In many applications, x_1 will be the first observation gathered by the experimenter, x_2 the second, and so on. The i^{th} observation in the data set will be denoted by x_i.

a. **Stem-and-leaf display**

Consider a numerical data set $x_1, x_2, \cdots, x_n$ for which each x_i consists at least two digits. A quick way to obtain the informative visual representation of the data set is to construct stem-and-leaf display as follows.

i. Select one or more leading digits for the stem values. The trailing digits becomes the leaves.
ii. List possible stem values in a vertical column.
iii. Record the leaf for each observation beside the corresponding stem value.
iv. Indicate the units for the stem and leaves someplace in the display.

Example: For a random sample fo lengths of golf courses (yard) that have been designated by Golf Megazine as among the most challenging in the United States. Express the sample of 40 courses

6435, 6526, 7131, 6790, 6464, 6527, 6605, 6433, 6694, 6506, 6890. 6583, 6870, 6770, 6614, 6873, 6900, 6700, 6927, 7051, 6850, 7011, 6798, 7040, 7050, 7022, 6770, 6936, 6904, 7169, 7168, 6745, 7105, 7005, 7280, 6470, 6713, 7209, 7113, 7165

by stem-and-leaf display.

Solution: Among the sample of 40 courses, the shortest is 6433 yards long, and the longest is 7280. Taking Thousand-Tundred digits for stem, and Tens-Ones digits for leaf, we obtain the stem-and-leaf display with two-digit leaves shown in Figure-(a).

64	35	64	33	70	Stem: Thousands and hundreds digits		
65	26	27	06	83	Leaf: Tens and ones digits		
66	05	94	14				
67	90	70	00	98	70	45	13
68	90	70	73	50			
69	00	27	36	04			
70	51	05	11	40	50	22	
71	31	69	68	05	13	65	
72	80	09					

(a)

Notice: Stem choice here of their a single digit (6 or 7) or three digits (643,…, 728) would yield an uninformative display, the first because of too few stems and the later because of too many.

Statistical software packages do not generally produce displays with multiple digit stems. The Minitab in Figure-(b) results from truncating each observation by deleting the one digit which is stem-and-leaf display with truncated one-digit leaves.

```
Stem-and-leaf of yardage  N = 40
Leaf Unit = 10
           4          64   3367
           8          65   0228
          11          66   019
          18          67   0147799
         (4)          68   5779
          18          69   0023
          14          70   012455
           8          71   013666
           2          72   08
```

(b)

Q.11 (Problem set-1.2)

Every score in the following batch of exam scores is in the 60s, 70s, 80s, or 90s. A stem-and leaf display with only the four stems 6, 7, 8 and 9 would nto give a very detailed description of the distribution of scores. Here we could repeat the stem 6 twice, using 6L for scores in the low 60s (leaves 0, 1,2,3, and 4) and 6H for scores in the high 60s (leaves 5, 6,7,8, and 9). Similarly other stems can be repeated twice to obtain a display consisting of 8 rows. Construct such a display for the given scores. What feature of the data is highlighted by this display

74	89	80	93	64	67	72	70	66	85	89	81	81
71	74	82	85	63	72	81	81	95	84	81	80	70
69	66	60	83	85	98	84	68	90	82	69	72	87
88												

Solution: For stem 6L, the leaves are 4 (in first row), 3 (in second row) and 0 (in third row) because leave L caries the values from 0 to 4. For stem 6H, the leaves are 7, 6 (in first row), and 9, 6, 8, 9 (in third row) because leave H caries the values from 5 to 9. Similarly we have following stem-and-leaf display with one-digit leaves.

6L	430
6H	769689
7L	42014202
7H	
8L	011211410342
8H	9595578
9L	30
9H	58

The gap in the data—no scores in the high 70s.

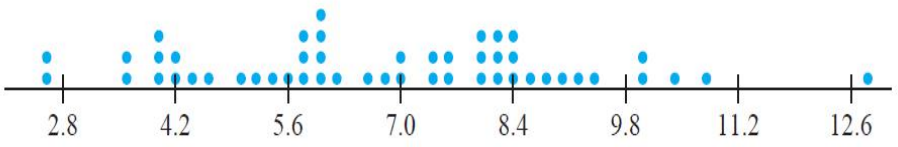
b. **Dotplots**

A dotplot is an attractive summary of numerical data when the data set is reasonably small or there are relatively few distinct data values. Each observation is represented by a dot above the corresponding location on a horizontal measurement scale. When a value occurs more than once, there is a dot of each occurrences, and these dots are stacked vertically. As with a stem-and-leaf display, a dotplot gives information about location, spread, extremes, and gaps.

Example: Sketch the dotplots of the data

10.8	6.9	8.0	8.8	7.3	3.6	4.1	6.0	4.4	8.3
8.1	8.0	5.9	5.9	7.6	8.9	8.5	8.1	4.2	5.7
4.0	6.7	5.8	9.9	5.6	5.8	9.3	6.2	2.5	4.5
12.8	3.5	10.0	9.1	5.0	8.1	5.3	3.9	4.0	8.0
7.4	7.5	8.4	8.3	2.6	5.1	6.0	7.0	6.5	10.3

Solution: In the data, number of numbers is $n = 50$. Smallest value is 2.5 and largest value is 12.8. Taking first value as 2.6, last value as 12.6 with space length 1.4, we have the dotplot



c. **Histograms**

Some numerical data is obtained by counting to determine the value of a variable (the number of traffic citations a person received during the last year, the number of customers arriving for service during a particular period), whereas other data is obtained by taking measurements (weight of an individual, reaction time to a particular stimulus). The prescription for drawing a histogram is generally different for these two cases: discrete or continuous.

Discrete and Continuous variables: A numerical variable is **discrete** if its set of possible values either is finite or else can be listed in an infinite sequence (one in which there is a first number, a second number, and so on). A numerical value is **continuous** if its possible values consist of an entire interval on the number line.

Frequency: In a discrete data set, the frequency of a variable x is the number of times that value exists.

Relative Frequency: A discrete data set (A) is obtained by some observations. The relative frequency of a value $x \in A$ is the fraction/proportion of times of the value occurs. Let number of times the value x occurs is m and number of observations in the data set is n, then

$$\text{Relative frequency of a value } x = \frac{\text{number of times the value occurs}}{\text{number of observations in the data set}} = \frac{m}{n} = f_x.$$

Theoretically sum of relative frequencies of the values is 1.

Example: A data set consists of 200 observations on x = the number of courses a college student is taking this term. If 70 of these x values are 3, then frequency of the x value 3 is 70 and relative frequency of the x value 3 is 70/200=0.35, i.e.. 35% of the students fo the sample are taking the three courses.

Note: In general sum of the relative frequencies differs slightly from 1 because of rounding.

Frequency Distribution: A frequency distribution is a tabulation of the frequencies and/or relative frequencies.

1.

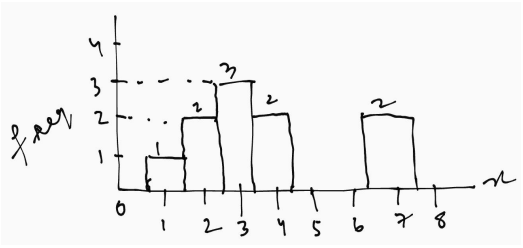
A rectangle can be plot at the point (x, f_x) with height f_x at value x such that the area of the rectangle is proportional to f_x of the value x. A **histogram** is an approximate representation of the distribution of numerical data which was introduced by Karl Pearson.

Q. Find the relative frequency of each value of the data 1, 3, 4, 8, 3, 7, 2, 7, 4, 2, 3 and draw the histogram.

Solution:

x	frequency	Relative frequency
1	1	0.1
2	2	0.2
3	3	0.3
4	2	0.2
7	2	0.2

Total frequency=10



Histogram

Chapter-2

2.1 Sample spaces and Events

2.2 Axioms, interpretations and properties of probability

2.4 Conditional Probability

2.5 Independence

Chapter-3

3.1 Random variables

3.2 Probability distribution of discrete random variable

3.3 Expected Values

3.4 Binomial Probability distribution

3.5 Hyper geometric, Negative binomial Probability distribution

3.6 Poisson Probability distribution