2013-14 Spring Term

CE 204 Complementary Course Material I

LEARNING FROM and ABOUT DATA

Introduction/ Uncertainties

Many real life problems that engineers deal with are formulated under conditions of *uncertainty*. The engineering design of a physical problem may involve natural processes and phenomena that are inherently *random*. The related information may not be complete, adequate or not satisfactory on the problem of concern. So, idealized prototype of the problem and/or its mathematical model (formulated form) may involve such uncertainties together with the uncertainties related to the imperfections in modelling and parameters used. In short, uncertainties may enter the problem at the input stage *(physical uncertainties*) and at the modelling phase *(model uncertainties)* resulting output uncertainties (*statistical uncertainties*). Thus, inevitably, decisions required for planning and design are to be made or are made under conditions of uncertainty.

The above mentioned uncertainties are in general grouped into two, namely as "*aleatory*" and "*epistemic*".

Aleatory uncertainties (random/stochastic uncertainties) deal with the randomness or predictability of an event, mostly reflecting external variability in the system. They are uncertainties ascribed to physical system and/or environment under consideration. They are irreducible, inherent and stochastic. (E.g. wind speed and direction are aleatory (random) uncertainties).

Epistemic uncertainties (parameter uncertainties) reflect the possibility of errors in our general knowledge. Such uncertainties result from some level of ignorance or incomplete information about the system or surrounding environment. They are subjective and model form of uncertainties and related to the state of knowledge uncertainty. Since generally, we do not know the correct values of the parameters in the model constructed parameter uncertainties are of epistemic type. Epistemic uncertainties are of reducible nature. (E.g. I believe that the speed of the wind is less than 40 km/hr, but I am not sure of that).

The output or final uncertainty that results in the problem solution over which decisions are to be made may appear to be aleatory but usually, it may derive from both sources of uncertainty, that is both aleatory and epistemic.

Whatever the types are, the effects of uncertainty are important both in the design and planning of engineering systems and require **quantification**. For scientific quantification of uncertainties, engineers use the concepts and methods of **probability**. On the other hand, to reduce the degree of epistemic uncertainties one may require to obtain more information/data via observations, experiments and records where concepts, tools and methods of **statistics** are almost inevitable.

2013-14 Spring Term

CE 204 Complementary Course Material I

LEARNING FROM and ABOUT DATA

Introduction/ Uncertainties

Many real life problems that engineers deal with are formulated under conditions of *uncertainty*. The engineering design of a physical problem may involve natural processes and phenomena that are inherently *random*. The related information may not be complete, adequate or not satisfactory on the problem of concern. So, idealized prototype of the problem and/or its mathematical model (formulated form) may involve such uncertainties together with the uncertainties related to the imperfections in modelling and parameters used. In short, uncertainties may enter the problem at the input stage *(physical uncertainties*) and at the modelling phase *(model uncertainties)* resulting output uncertainties (*statistical uncertainties*). Thus, inevitably, decisions required for planning and design are to be made or are made under conditions of uncertainty.

The above mentioned uncertainties are in general grouped into two, namely as "*aleatory*" and "*epistemic*".

Aleatory uncertainties (random/stochastic uncertainties) deal with the randomness or predictability of an event, mostly reflecting external variability in the system. They are uncertainties ascribed to physical system and/or environment under consideration. They are irreducible, inherent and stochastic. (E.g. wind speed and direction are aleatory (random) uncertainties).

Epistemic uncertainties (parameter uncertainties) reflect the possibility of errors in our general knowledge. Such uncertainties result from some level of ignorance or incomplete information about the system or surrounding environment. They are subjective and model form of uncertainties and related to the state of knowledge uncertainty. Since generally, we do not know the correct values of the parameters in the model constructed parameter uncertainties are of epistemic type. Epistemic uncertainties are of reducible nature. (E.g. I believe that the speed of the wind is less than 40 km/hr, but I am not sure of that).

The output or final uncertainty that results in the problem solution over which decisions are to be made may appear to be aleatory but usually, it may derive from both sources of uncertainty, that is both aleatory and epistemic.

Whatever the types are, the effects of uncertainty are important both in the design and planning of engineering systems and require **quantification**. For scientific quantification of uncertainties, engineers use the concepts and methods of **probability**. On the other hand, to reduce the degree of epistemic uncertainties one may require to obtain more information/data via observations, experiments and records where concepts, tools and methods of **statistics** are almost inevitable.

Statistics

Statistics is the science and art of collecting, displaying/tabulating/compiling/summarizing and gaining insight to interpret/understand data obtained from an appropriately constructed experiment *(descriptive statistics)* or study in order to test theories and make inferences/decisions on these theories *(inferential statistics)*. In short, scientific approach in statistics first requires an underlying theory to be tested for which test data will be obtained from relevant observations or experiments. The very basic point in data analysis is to establish robust mathematical models so as to organize and gather information in an efficient way.

Before we proceed to exercises we need to define two main concepts: *population* and *samples*

- A *population* is a set of well defined distinct objects and its elements or in other words it is the whole set or group that we are interested in. Usually, we denote the population by the set S. S can be finite (if so the population size is usually denoted by N) or infinite in extent. For several reasons we may not be able to observe the population totally, but we may be able to study only a portion called the sample.
- When the observations are numerical values, the population is referred to as a *quantitative population* . If the observations are on *attributes* (type of structure, level of damage, or similar category) the population is a *qualitative population.*
- A *sample* is a subset of S (let it be denoted by A, A= $\{s_1 , \dots , s_n\}$). Usually A is a small part of the population that we draw from S in order to make observations on it so as to learn about S. The more representative sample we collect from the population , the better we may learn about the population.
- *Raw data* is the list of observations/ measurements in the sample whose values are not manipulated at all.
- Statistics is concerned with data . If the population is quantitative the data set will constitute numbers. But if the population is qualitative the observations will be non-numerical and for a statistical study numerical data can be artificially created. We call such data *nominal data* since the numbers will represent arbitrary codes.
  In engineering most of the times the data we encounter are *ratio data*, that is the basic arithmetic operations (addition, subtraction, division and multiplication) are valid for such a data.
  For some type of data only addition and subtraction are meaningful, such data are scale dependent as in the case of temperature and are called *interval data* (E.g. In Celsius 0^o , 20^o and 40^o correspond to 32^o , 68^o and 104^o Fahrenheit, respectively. The ratios in the two scales differ between the temperatures but intervals remain constant. So temperature is an interval data.
  Data where no arithmetic operations are meaningful are called *ordinal data* . The numbers in the data will represent an ordering relation in other words ranking in terms of importance, preference, strength, and etc.

The elements of the sample represented as real -valued function $X(s_i) = x_i$ that are defined on the sample A= $\{s_1 , \dots , s_n\}$ is called *data set.* Data set may be discrete or continuous depending on the physical characteristics of x_i. (If your sample is for the number of vehicles or the type of vehicles crossing a bridge you have a discrete sample but if you are interested in the weights or length of cars, the data set will be continuous).

- The number of elements in the data set is called the *sample size* and is usually denoted by n.

- Collecting sample data from a larger population and using it to make predictions and /or decisions for the entire population is called *statistical inference*. The efficiency of the inference will lie in selecting a representative, appropriate in general *a random sample.*
- The *random sample* is the one obtained by the items giving the same chance of being chosen to every item as any other item of the population.
- *Random variable:* Quantities that are measured or observed are termed variables and because of the inherent randomness they are called random variables (since the values of measured or observed quantities depend on chance) .
- *Continuous Random Variable:* Variables that can have any value on a continuous interval.
- *Discrete Random Variable:* Variables that can have countable isolated numbers (e.g. integers).
- *Distribution* refers to the variability pattern of the random variable.

Descriptive Statistics

Descriptive statistics consists of methods used to organize, display and analyse data from some population or sample.

- ***Data Collection:*** When the population size is large, it can be time consuming, expensive or impractical and/or impossible to study its set values. In practice, it is therefore more common and usually desirable to study a relatively small fraction of population, that we have defined as sample which is required to be a representative of the entire population (so to be chosen at random).
- ***Organization of Data***: We may organize a data in several different ways in order to see if a pattern exists for the characteristics of the data.
  - One basic way is to list the data in *numerical order* (ascending or descending order).

When the observations are listed in ascending order, say for a sample of size n, as

$$\{x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}\} \tag{1}$$

the elements of the set in Eq. (1) are called the order statistic and $x_{(i)}$ is called the ith order statistic. The ith order statistic $x_{(i)}$ is the $\frac{(i-0.5)}{n}$ *quantile.* Sometimes it may be useful to obtain the d th percentile value (Q_d) of the ordered data (sorted in ascending form by the following formula

$$Q_d = x_{(i)} + [\,(n+1)d - i\,](x_{(i+1)} - x_{(i)}) \tag{2}$$

where n is the sample size and the subscript i, is the largest integer such that $i \le (n+1)\,d$

Note that in general, $x_{(i)} \neq x_i$ and

$$x_{(1)} = \min\{x_1,\ldots,,x_2\}$$
$$x_{(n)} = \max\{x_1,\ldots,,x_2\} \tag{3}$$

The *range of the data* is

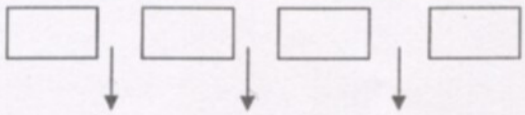
$$r = x_{(n)} - x_{(1)} \tag{4}$$

The *interquartile range (iqr)* is the length of the interval that contains the middle half of the data. The interquantile range is defined as

$$(iqr) = Q_3 - Q_1 \qquad (5)$$

where $(Q_3 - Q_1)$ is called the middle 50 percent range (Q_1 and Q_3 correspond to the first 25 percent (lower quartile) and 75 percent (upper quartile) ranges of the data, respectively).

Symbolic Representation of Quartile Ranges (each box refers to 25 % of the data)



Median of the Lower Values, Q_1 | Median, Q_2 | Median of the Upper Values, Q_3

The number of class intervals

- In case n or the range of existing numeric values are or data comes from continuous set of numbers one may arrange the data in categories or in the form of *class intervals*. The number of *class intervals k* are usually chosen of equal size and it is at least 5 (for small sized data) but not more than 20 (for large sized data) and the intervals are to be nonoverlapping . The frequency estimation in interval form is usually for continuous data but if the sample size is large one may prefer to represent the discrete set of data in intervals also.
The practical guide to choose "k" can be based on the following empirical mathematical models:

a) rule of thumb : choose closest integer to $\sqrt{n}$ as k where n is the sample size.

b) due to Sturges (1926) : $k = 1 + 3.3 \log_{10} n$ where n is the sample size.

c) due to Friedman and Diaconis (1981) : $k = \dfrac{r\, n^{1/3}}{2(iqr)}$ where r is the range, n is the size and (iqr) is the interquartile range of the sample data, respectively.

Note that the intervals will be expressed in the form [a, b) (left parenthesis is closed and the right one is open). a and b will be ***class boundaries*** or *class limits* and ***class marks*** (or mid -point) are the midpoints of each class interval.

The *number of observations* (fr*equencies*) f_i for each number in the data set or for each class interval can be counted and may be listed as frequency or class frequency in a tabular form. Such a table will give a full summary of the *population frequency distribution* or *sample frequency distribution(or empirical frequency distribution)* depending on whether the data represent the entire population or a portion of it. Usually, only samples will be available, thus population frequency distribution will remain as an unknown.

The above arrangements or organizations of the data may often be easier to summarize and to help draw conclusions from data when represented in tabular and especially in graphical forms.

- ***Graphical Representation of Data***

  - *The Histogram and Frequency Curve (Frequency Polygon),* $\hat{f}(x)$: A histogram is a bar chart in which the vertical axis represents the number of observations (frequencies) and the horizontal axis is for the observed values or corresponding class intervals. When the data is discrete and not large in size the plot of the data versus its frequency will be referred to as *line diagram* or *bar chart.* For large data or continuous data each bar showing the frequency is centered over the class mark or as a rectangle over the class interval to obtain the histogram.

    Whenever some frequencies are large the relative frequency (rel freq.)

$$\text{rel.freq.} = \frac{f_j}{\sum_{j=1}^{k} f_j} = \frac{f_j}{n \text{ or } N}$$

    where j = 1 ,2, …, k can be used on the vertical axis . Frequency distributions are the basic shapes of their histograms.

    Frequency polygon is obtained by joining the frequencies of the sample data points or class marks linearly.

    Cumulative frequency(or *relative frequency) distribution(ogive)* $\hat{F}(x)$: is constructed by adding frequency (or relative frequency) of each point or class interval of the data to the frequencies (or relative frequencies) of the lower values or classes. It gives us less than frequencies of a specific data value or interval .

- *Dot Diagram:* When the sample size is small and/or data is continuous one may sometimes use *dot diagram.* One dot per data is placed over the numerical value of the data represented on the horizontal axis .

- *Stem and Leaf Plot:* Histograms may not be effective when $n < 50$ (may not give clear indication of variability and other characteristics). In such cases we may prefer *stem and leaf plots* which yield no loss of information since all magnitudes are represented. Such a plot will also highlight extreme values and other characteristics and can be constructed easily.

- *Box Plot:* For showing the three quartiles Q_1 (lower quartile) , Q_2 (median) and Q_3 (upper quartile) on a rectangular box representation may be used to indicate variability of the data.

- *Scatter Diagram:* If there are n pairs of data (x_1 , y_1), (x_2 , y_2), …, (x_n , y_n) a preliminary indication of correlation between them is obtained by *scatter diagram* in which the horizontal axis is reserved for the independent (or the variable with least uncertainty) and the vertical axis is for the dependent or uncertain variable.

  There are several other graphical representations that are used in practice, *but in engineering bar charts , histograms , ogives, scatter diagram and sometimes stem and leaf plots* are the most commonly preferred visual tools.

You may easily note that graphic displays and frequency tables depend on the size and number of class intervals chosen (so as the stem and leaf). A good graphic description and display is mostly partly art and partly science. Unless they are accompanied by the following statistical descriptors, they may not give sound ideas about the sample and population.

• ***Numerical Statistical Descriptors***

In addition to numerical descriptors defined above like the sample range , quartiles several others exist to give locational and variability characteristics of the data.
In the following x_i represents the data value in the ungrouped data or the class mark (mid point value of the interval) in the grouped data unless .

- ***Central Value Measures (Measures of Location) :***
  - *Sample Mean (sample average or arithmetic mean):* The mean defines the centre of mass of the frequency distributions and for a set of numerical data $\{ x_1 , x_2, \ldots, x_n \}$ the sample mean $\bar{x}$ is defined by

$$\bar{x} = \frac{\sum x_i}{n} \quad \text{or} \quad \bar{x} = \frac{\sum f_j x_j}{n} \tag{6}$$

    where f_i is the frequency of x_i . x_i is either the numerical value of the data or is the class mark (class center) for the grouped data. The arithmetic mean is easy to find and it is unique for a given data set but it is highly affected by the extreme values present in the data. In such a case the sample mean may not be a good representer of the data set.
  - *Harmonic Mean:* When one needs to find the average of the reciprocals of a variable or if x_i values are very large to get a meaningful $\bar{x}$, one may compute the harmonic mean as

$$\bar{x}_h = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}\right)} \tag{7}$$

  - *Geometric Mean:* When averaging values that represent a rate of change one may use the geometric mean defined as

$$\bar{x}_g = \sqrt[n]{x_1 \, x_2 \ldots . x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i} \tag{8}$$

  - *Median:* When the data is ordered the central value is defined as the median . Median also corresponds to the second quartile Q_2 of the data set (x_{med}). So that half of the data is less than the median and the other half is larger than the median. If the number of data is odd median is middle point data as

$$x_{med} = x_{((n+1)/2)} \tag{9}$$

    If the number of data is even, the median is the average of the middle two data as

$$x_{med} = \frac{x_{(n/2)} + x_{((n/2)+1)}}{2} \tag{10}$$

- *Mode:* That value of the data set which occurs most frequently is defined as the mode of the data (x_{mod}).

One should note that if instead of the sample the data covers the all population values we have population descriptors with N instead of n, and usually population mean will be denoted by μ as

$$\mu = \frac{\sum x_i}{N} \quad \text{or} \quad \mu = \frac{\sum f_i x_i}{N} \tag{11}$$

- *Measures of Dispersion (or Spread)*

The following quantities are for measuring how far does the data spread from a central value : mean or median.

- *Mean Absolute Deviation*: It is the average distance of the data points from the central value as

$$d = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + .. |x_n - \bar{x}|}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} \quad \text{or} \quad \frac{\sum_{j=1}^{k} f_j |(x_j - \bar{x}|}{n} \tag{12}$$

- *Variance and Standard Deviation:* The variance s^2 is defined as

$$s^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2 \quad \text{for ungrouped data} \tag{13}$$

and

$$s^2 = \frac{1}{n}\sum_{j=1}^{k} f_j(x_j - \bar{x})^2 \quad \text{for grouped data} \tag{14}$$

For reasons that will be explained later the sample variance (with sample size n) is modified as

$$s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2 \tag{15}$$

and

$$s^2 = \frac{1}{n-1}\sum_{j01}^{k} f_j(x_j - \bar{x})^2 \tag{16}$$

for ungrouped and grouped data, respectively.

Computationally, the following forms for the variances defined above will be more preferable (since the number of computations will be reduced):

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2 \tag{17}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^{\,2} - \frac{n}{n-1}\bar{x}^2 \tag{18}$$

The variance is a positive quantity and physically it corresponds to mass or area moment of inertia about centroidal axis (second moment of mass or area).

The square root of the variance is defined as the *standard deviation* (s) and has the

same units as that of the data quantities. The standard deviation measures the spread from the mean values so small s may correspond to a data set clustered around the mean but since it has units s value depends on the magnitudes of the data values and it may not be easy to judge the true variation in the data set. The absolute variation can be measured by the coefficient of variation.
For the population variance or standard deviation (σ), the deviations are measured from the population mean μ in equations 13, 14 or 17.

- *Coefficient of Variation:* The unitless quantity defined as the ratio of the standard deviation s to that of the mean value is a measure of the absolute variation of the data and is called coefficient of variation

$$v = \frac{s}{\bar{x}} \tag{19}$$

If the coefficient of variation is small (approximately less than 0.20 or so) the spread in the data around the mean is small and in that case mean can represent the data more efficiently.

There are also other measures (higher ordered moments of the variations from the central values) of variability, some of them are

- *Standard Error of the Mean* is defined as

$$s.e.(\bar{x}) = \frac{s}{\sqrt{n}} \tag{20}$$

- *Coefficient of Skewness:* It is a coefficient which measures the asymmetry of the distribution and defined as

$$g_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3} \quad \text{(for ungrouped data)}$$

and (21)

$$g_1 = \frac{\sum_{j=1}^{k} f_j (x_j - \bar{x})^3}{ns^3} \quad \text{(for grouped data)}$$

If coefficient of skewness is positive it implies that the long tail of the distribution is on the right side, and so on.

- *Coefficient of Kurtosis:* This coefficient is a measure of the peakedness of the distribution and defined as

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{ns^4} \text{ (ungrouped data) and } g_2 = \frac{\sum_{j=1}^{k} f_j (x_j - \bar{x})^4}{ns^4} \text{ (grouped data)} \qquad (22)$$

Small coefficient of kurtosis implies that the tail weight of the distribution is small and data has a peak.

- *Covariance:* When dealing with pairs of data (X and Y) as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ it is usually necessary to observe how two sets of data vary together. The measure for the common variation of the smaple is given by sample covariance s_{XY}

$$s_{XY} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i y_i}{n} - \bar{x}\,\bar{y}$$

$$\text{or } s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{n-1} \qquad (23)$$

- *Correlation Coefficient*: The correlation coefficient expresses the strength of linear relation between X and Y numerically where on scatter diagrams general trend of the relations may be observed roughly. Negative values refer to negative slopes and positive values refer to positive slopes for the relation ($-1 \leq r_{XY} \leq 1$). For a sample the correlation coefficient r_{XY}

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{1}{n}\sum_{i=1}^{n}\frac{(x-\bar{x})_i}{s_X}\frac{(y_i - \bar{y})}{s_Y} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sqrt{\left[\sum x_i^2 - (\frac{1}{n})(\sum x_i)^2\right]}\sqrt{\left[\sum y_i^2 - (\frac{1}{n})(\sum y_i)^2\right]}} \qquad (24)$$

The square of the correlation coefficient gives the degree of tightness for a linear fit of the data set and is called *coefficient of determination*.

- *Outliers*

Sometimes some extreme values exist in the data which will highly affect the mean value, to detect the outliers we may use *interquartile* ranges or *z- scores* of the data:

If a data point is more than approximately 1.5 iqr, (where (iqr) is the interquartile range as defined previously) from an end of an extreme quartile value the data is considered as an *outlier* and if it is approximately 3.0 iqr from the ends than its an *extreme outlier*.

MEAN and STANDARD DEVIATIONS ARE GOOD DESCRIPTORS of the SAMPLE IN CASE THERE ARE NO OUTLIERS IN THE DATA.

Z- Score : By definition z- score of a data is the transformed variable, z as

$$z = \frac{x - \bar{x}}{s} \qquad (25)$$

Usually 68 % of the data must lie in within one standard deviation of the sample mean and 95% of the data must lie in within two standard deviations from the mean. If values fall outside such ranges , they may be considered as outliers.

When outliers are present in the data one may prefer to use *trimmed means* to get more effective means. In the ordered data T% (or outliers) of the observations are removed from each and then the sample mean of the remaining numbers are calculated. The resulting mean is the T% trimmed mean and it generally lies between the sample mean and the sample median.

NOW PLEASE GO OVER THE FOLLOWING TWO EXERCISES AND MAKE NECESSARY CORRECTIONS AND COMMENTS.

Class Exercise 1

Students' Grades Data (SGD) : n = 90

45 50 35 95 60 70 55 95 43 65 60 58 75 62 65 90 95 60 75 60 30 100
55 50 60 60 35 60 35 53 60 55 45 85 95 50 55 69 45 45 25 55 43 26
30 21 50 55 17 30 35 25 23 27 20 07 20 55 15 21 13 30 30 38 15 40
50 75 80 80 75 85 40 55 60 55 85 65 65 47 41 28 35 36 25 23 30 40
55 13

- Let us arrange the above raw data in ascending order :

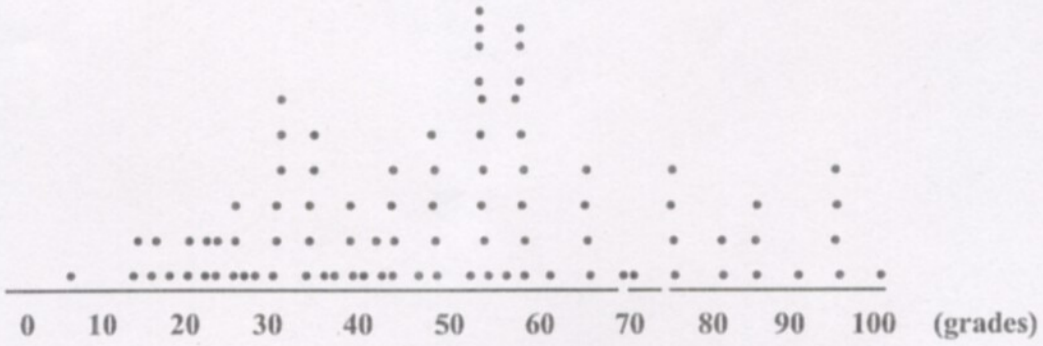
07 13 13 15 15 17 20 20 21 21 23 23 25 25 25 26 27 28 30 30 30 30
30 30 35 35 35 35 35 36 38 40 40 40 41 43 43 45 45 45 45 47 50 50
50 50 50 53 55 55 55 55 55 55 55 55 55 55 58 60 60 60 60 60 60 60
60 60 62 65 65 65 65 69 70 75 75 75 75 80 80 85 85 85 90 95 95 95
95 100

The data can be summarized on a **FREQUENCY TABLE** using class intervals (shows how many data points are around mid point value of the interval). Note that in this example we chose 11 groups from 0 to 110 to have nice interval size (as 10 here) and the smallest and largest values of the data are included in the first and last intervals, respectively.

Class Interval	Class Mark (Mid-point)	Tally (if necessary)	Frequency	Relative Frequency	Cumulative Relative Frequency	Z-score
0-10	5		1	0.011	0.011	-2.063
10-20	15		5	0.056	0.067	-1.618
20-30	25		12	0.133	0.200	-1.173
30-40	35		13	0.144	0.344	-0.727
40-50	45		11	0.122	0.467	-0.282
50-60	55		17	0.189	0.656	0.163
60-70	65		15	0.167	0.822	0.609
70-80	75		5	0.056	0.878	1.054
80-90	85		5	0.056	0.933	1.499
90-100	95		5	0.056	0.989	1.944
100-110	105		1	0.011	1.000	2.390

a) We may use one of the following graphical representations to see the frequency variations of the data:

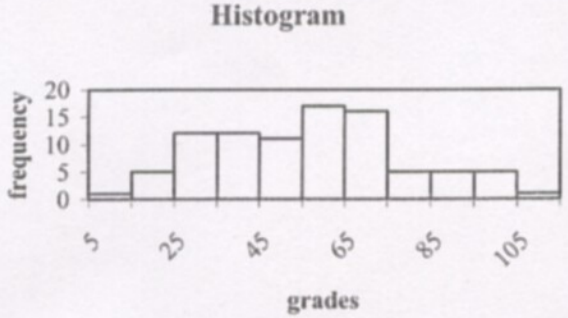
a.1) DOT PLOT (one student per dot)



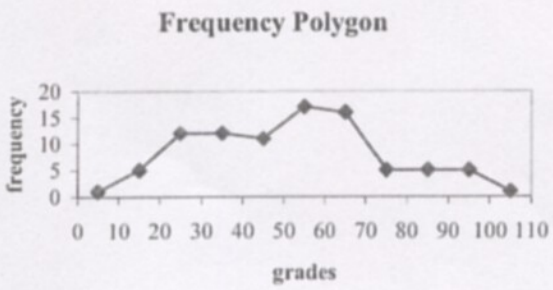
a.2) HISTOGRAM / BAR CHART (and CUMULATIVE FREQUENCY DIAGRAM)

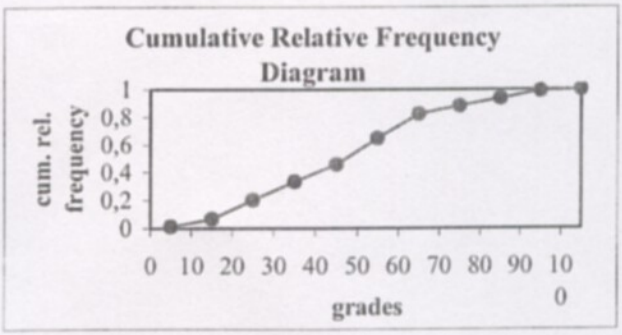
In this example we draw the following histograms from the data as given in the above table.

Histogram or frequency polygon may have the vertical frequency axis as the relative



frequencies. The basic shape of the plots will not change but relative frequencies may give better overall measure of the number of occurrences of the grades. The relative frequency polygons actually shows us the shape of the distribution of the variable in our data set.
The cumulative relative frequency diagram of the data (we assume the tabular form of the data is given) is as follows





a.3) STEM and LEAF DIAGRAM

Stem is a column of numbers and leaves are to show the weights of the data .

Stem	Leaves
0	7
10	3 3 5 5 7
20	0 0 1 1 3 3 5 5 5 6 7 8
30	0 0 0 0 0 0 5 5 5 5 5 6 8
40	0 0 0 1 3 3 5 5 5 5 7
50	0 0 0 0 0 3 5 5 5 5 5 5 5 5 5 8
60	0 0 0 0 0 0 0 0 0 2 5 5 5 5 9
70	0 5 5 5 5
80	0 0 5 5 5
90	0 5 5 5 5
100	0

a.4) BOX DIAGRAM

We divide the ordered data into four quartiles and observe how different the extreme groups are. Interquartile range is defined as, $IQR = Q_3 - Q_1$.

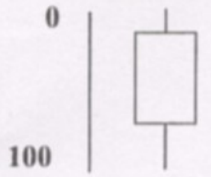
i) Assuming only grouped data is available (that is using the table) **$Q_1 = 35$, $Q_2 = 55$ (corresponds to median) and $Q_3 = 65$; IQR = 65-35 = 30.**

ii) If the original ordered data is available:

- $i \le (90+1)x\ 0.25 \rightarrow i = 22 \rightarrow Q_{0.25} = Q_1 = 30 + [(90+1)x\ 0.25 - 22](30-30) = 30$
- $i \le (90+1)x\ 0.50 \rightarrow i = 45 \rightarrow Q_{0.50} = Q_2 = 50 + [(90+1)x\ 0.5 - 45](50-50) = 50$
- $i \le (90+1)x\ 0.75 \rightarrow i = 68 \rightarrow Q_{0.75} = Q_3 = 60 + [(90+1)x\ 0.75 - 68](62-60) = 60.5$

  IQR: 60.5-30=30.5

Note that we don't have any outliers (small or large) in the SG D.



Q_1= 35 (*30*)
Q_2= 55 (*50*)
Q_3= 65 (*60*).

b) The numerical descriptors of the data are:

- **SAMPLE MEAN (ARITHMETIC AVERAGE OR AVERAGE)**

  Assuming only grouped data is available

$$\bar{x} = \frac{\sum_{1}^{11} f_i x_i}{n} = \frac{1}{90}(1\times 5 + 5\times 15 + \ldots + 1\times 105) = \frac{4620}{90} = 51.333$$

If we use the ordered original data (ungrouped data) : $\bar{x} = \frac{\sum x_i}{n} = \frac{4464}{90} = 49.6$

- **MEDIAN** : $x_{med} = \tilde{x} = \frac{x_{45} + x_{46}}{2} \cong 55$ (From the table). From the ordered data (original, ungrouped data) median is 50.

- *60^{th} percentile of the data from the ordered, original form:* $i \leq (90+1)x\ 0.60 \rightarrow i = 54$
  $Q_{60} = 55 + [(90+1)x\ 0.60 - 1](55-55) = 55$

- **MODE :** $x_{mod} = 55$ (From the table). *For the original data mode is 55 also.*

- **SAMPLE VARIANCE : s^2 for the grouped data**

$$s^2 = \frac{\sum_{j=1}^{11} f_j x^2_i}{n-1} - \frac{n}{n-1}\bar{x}^2 = \frac{282050}{89} - \frac{90}{89}(51.3333)^2 = 504.382 \text{ or}$$

$$s^2 = \frac{\sum_{j=1}^{11} f_j (x_j - \bar{x})^2}{n-1} = \frac{44890}{89} = 504.382$$

If we use the original data : $s^2 = \frac{\sum_{i=1}^{90} (x_i - \bar{x})^2}{n-1} = \frac{44171.6}{89} = 496.310$

- **STANDARD DEVIATION : s**

  $s = 22.458$ (*s= 22.278)*

- **COEFFICIENT OF VARIATION :** $v = 22.458/51.333 = 0.437$ (spread is quite large, though we don't have any outliers still mean and standard deviation may not be good statistics).

- **COEFFICIENT OF SKEWNESS:**

$$g_1 = \frac{\sum_{j=1}^{k} f_j (x_i - \bar{x})^3}{ns^3} = 0.244$$ (>0 it implies the tail of the frequency dist. is longer on the positive side)

- **COEFFICIENT OF KURTOSIS:**

$$g_2 = \frac{\sum_{j=1}^{k} f_j (x_i - \bar{x})^4}{ns^4} = 2.400$$ (sufficiently small number implying some peakedness in the distribution)

- **STANDARD ERROR OF THE MEAN:**

$$\text{s. e.}(\bar{x}) = s/\sqrt{n} = 2.367$$

- **STANDARD z- SCORES:** If you observe the z-scores on the table , approximately 75% of the data lies within one and 95 % lies within two standard deviations from the mean.

- **NUMBER OF CLASS INTERVALS :** Note that we may decide on the number of class intervals (n_c) as a rule of thumb in between 5 and 15 depending on the size of the data or

from $n_c = \sqrt{n} = \sqrt{90} = 9.49$ or $n_c = 1+3.3 \log_{10} n = 1+3.3 \log_{10} 90 = 7.45$ or

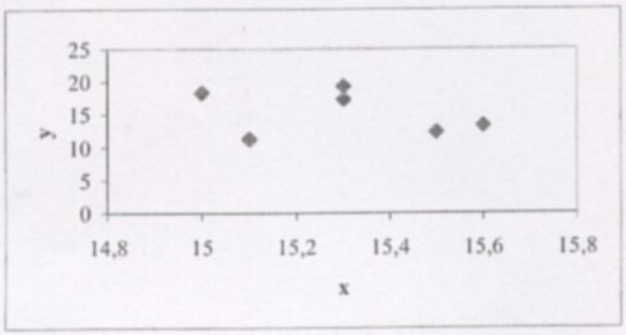
$n_c = \dfrac{m^{\frac{1}{3}}}{2(iqr)} = \dfrac{(105-5)\ 90^{\frac{1}{3}}}{2(65-35)} = 7.47$, so the choice on n_c as 11 is not a bad preference.

CLASS EXERCISE 2:

Given the following two sets of data {X,Y}:

X:	15.5	15.6	15.1	15.3	15.3	15.0
Y:	12.3	13.3	11.3	19.3	17.3	18.3

- **Scatter Plot**



- **Numerical Descriptors:**

$\bar{x} = 15.3$, $s^2_x = 0.052$, $s_x = 0.228$, $v_x = 0.015$

$\bar{y} = 15.3$, $s^2_y = 11.6$, $s_y = 3.406$, $v_y = 0.223$

Note that the two data sets have the same mean but quite different coefficient of variations !!!

- **COVARIANCE**

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{n-1} = -0.260$$

- **CORRELATION COEFFICIENT**

$$r_{XY} = \frac{s_{XY}}{s_X s_Y} = \frac{-0.260}{0.228 * 3.406} = 0.335$$

(Negative correlation implies x increases as y decreases; correlation coefficient – 0.335 gives $r_{XY}^2 = 0.112$, which implies that the strength or degree relation between X and Y is quite low).