Statistical Analysis

I. Data Collection and Univariate Statistics

Data Collection and Preparation

Data Quality

Data analyst must have confidence in the quality of data before commencing any analysis.

Is the data set good enough?

- Natural randomness effect ➔ may need more samples!
  (Ex: Irregularly sampled data)
- Failure in measurement procedures ➔ follow proper guidelines

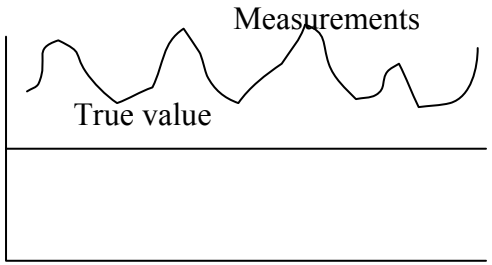
Put into consideration:

- a) Number of samples

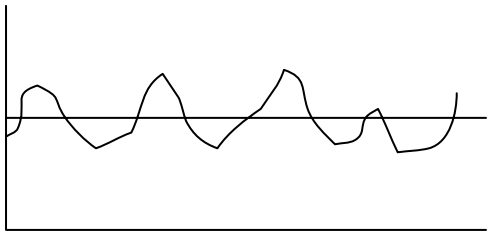
  ➔ higher # of samples = good results (cost?)

- b) Global picture of the problem (i.e. understand geology! ➔ Site investigation)

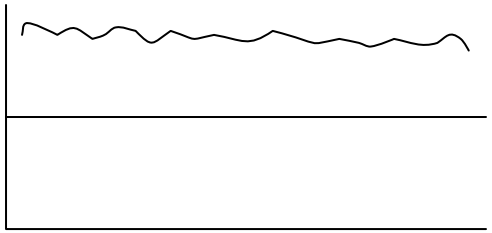
- c) Calibration of measuring devices ➔ precision and accuracy



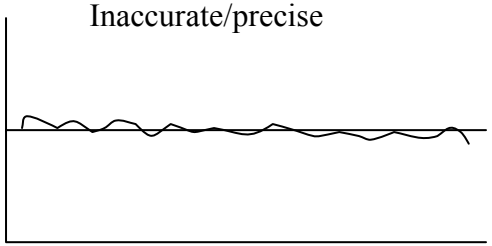
Inaccurate/non-precise



Accurate/non-precise



Inaccurate/precise



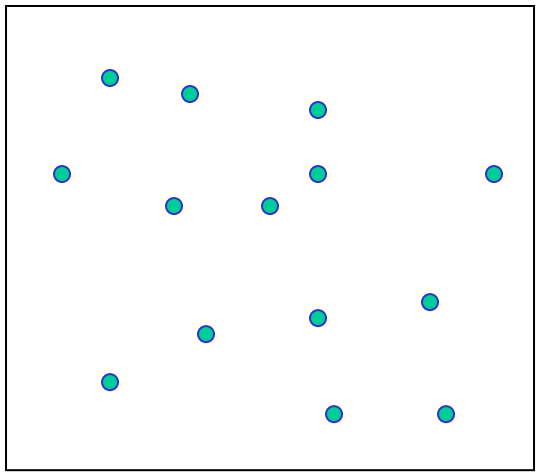
Accurate/precise

Sampling Strategies

Sample size: Could be decided on the basis of a pilot study.

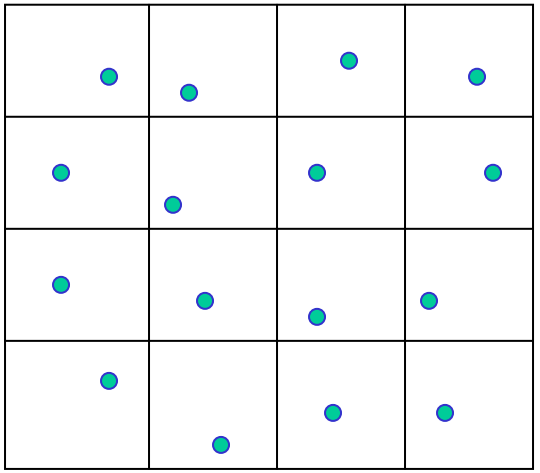
Spatial sampling schemes:

- **Random** ➔ use random number generator to decide about sampling plan

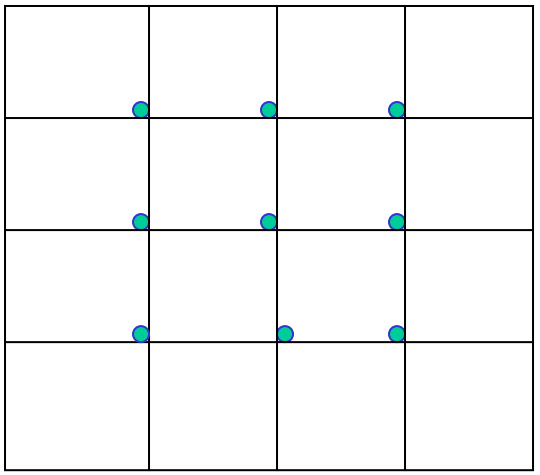


- **Uniform** ➔ use randomization sampling process within pre defined squares

  ➔ Best to reduce bias



- **Regular** ➔ use regular squares or rectangles to sample



Risk Assessment Computational Techniques

Statistics and Probability Approach

Science of collecting, analyzing, and interpreting data in order to draw a valid conclusion.

Univariate Statistics: Each variable analyzed in isolation (e.g. mean, variance, … etc.)

Bivariate Statistics: Two variables analyzed together to investigate the relationship between them (e.g. correlation coefficient, covariance, …etc.)

Spatial Analysis

The analysis of 3, 4 or more variables together where 2 or 3 of which are spatial coordinates

Examples: contour maps, GIS maps, … etc.

Analytical Methods

The description of an environmental attribute by mathematical and graphical tools.

Examples: health risk equations, RBCA tools … etc.

Flow Simulators

The **imitation** of a real natural system characteristics or response by solving space / time constrained mathematical equations (e.g. groundwater aquifers … etc.)

Examples: groundwater simulators, contaminant transport simulator … etc.

Populations and Samples

The statistical sample

The collection of available measurements that are a subset of the population of interest. The statistical sample is taken to **represent a population**.

Bias

It is any effect that deviates the statistical results from being a representative description of the population.

It is essential that the sample is an unbiased subset of the population.

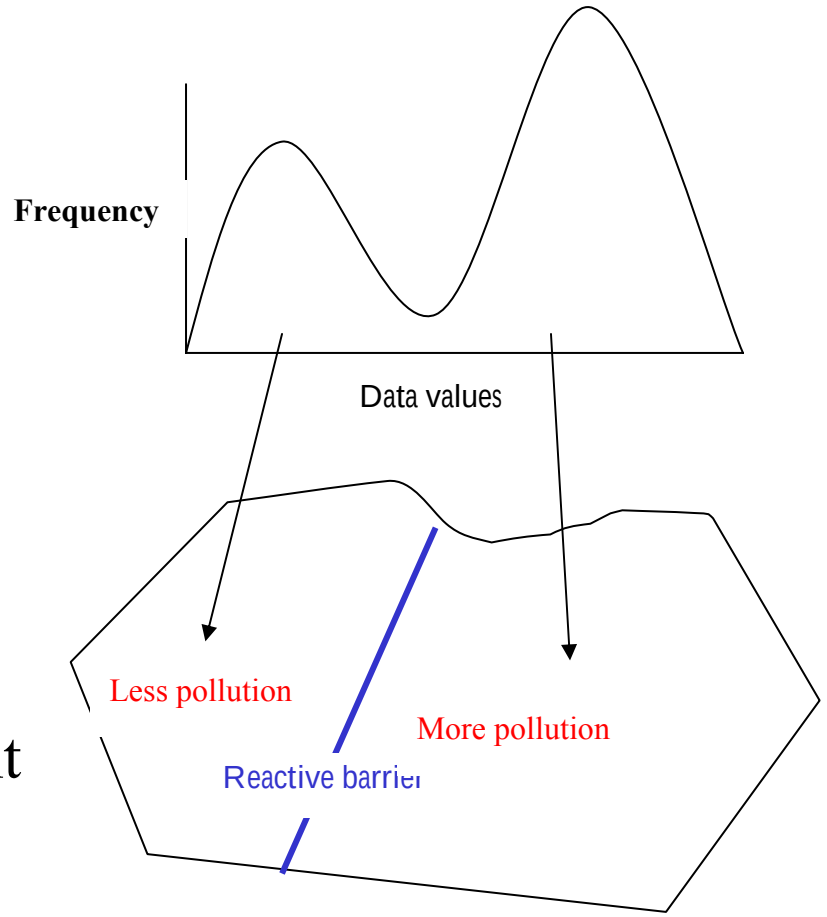
Sources of bias

a) Samples can not be collected easily (difficult accessibility)
b) High cost of sampling (less samples = more biased results!)
c) Errors in measurements (device calibration is needed!)
d) Bad sampling design

Why Statistics?

1) Quantify environmental information

- Reduction of large number of values to small number of meaningful estimates (e.g. mean … etc.)
- Characterize the pollutant

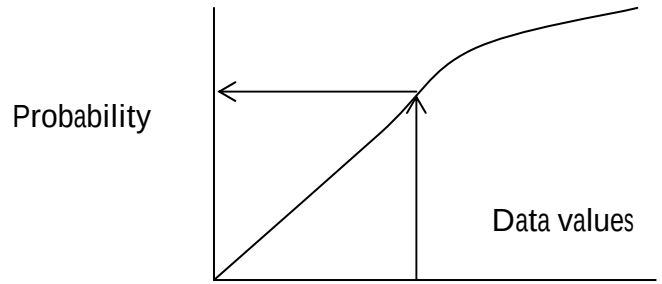


Why Statistics?

2) Conduct feasibility studies ➔ what is the probability that a pollutant concentration will exceed the safe threshold? (**Action:** remediation vs. natural attenuation)

3) Perform uncertainty analysis & risk assessment ➔ environmental site investigation (e.g. how many samples to collect)

4) Prepare a data set for further studies like the application of geostatistical methods, numerical models, … etc.



Univariate Statistics

- Univariate statistics deals with the organization, presentation, and summary of data of **ONE** variable.

- Statistical modeling is an approach for **fitting mathematical equations** to data in order to predict unknown quantities from measurements.

.

- Postulate a model that describes the data & fit its parameters
- Validate (test) the model
- Predict the unknown

Experimental Frequency Distributions

1- Construction Mechanism

- Construct a frequency table
  - Divide the data into number of classes
  - List number of observations in each class (i.e. frequency of observations)
  - Relative frequency = Frequency of each class / Total number of observations
  - % Frequency = Relative frequency X 100
  - Cumulative frequency = Adding frequencies as moving down the frequency table
  - % Cumulative frequency = Adding % frequencies as moving down the frequency table.

Experimental Frequency Distributions

2- Graphing Mechanism

- Graphical representation of data
  - Histogram or bar graph
  - Frequency curve
  - Relative frequency curve
  - Cumulative frequency bar graph
  - Cumulative frequency curve
  - % Cumulative frequency curves

- The proportion (probability of occurrence) of sample values or values that are smaller than a given value can be directly read from the above mentioned curves.

Experimental Frequency Distributions

3- Example

- Consider the following data set of benzene concentration (ug/l) collected from a groundwater aquifer:

8.1	24.9	32.2	36.9	41.3	47.6	54.5	74
12.8	26.8	33.4	37.3	41.5	48.5	54.6	77.3
14.3	27.1	33.6	37.4	41.7	49.4	56.1	
14.9	27.4	33.9	37.9	42.4	49.7	58.2	
15.7	28.2	34.1	38.5	42.9	50.1	59.3	
19.3	29.5	34.7	38.9	43.5	50.9	59.5	
20.3	29.2	35.1	39.1	43.6	51.5	59.8	
21.7	30.1	35.2	40.2	45.6	51.8	63.8	
22.6	31.8	35.6	40.8	46.1	52.7	64.1	
24.8	31.9	36.5	40.9	46.8	53.7	64.8	

Experimental Frequency Distributions

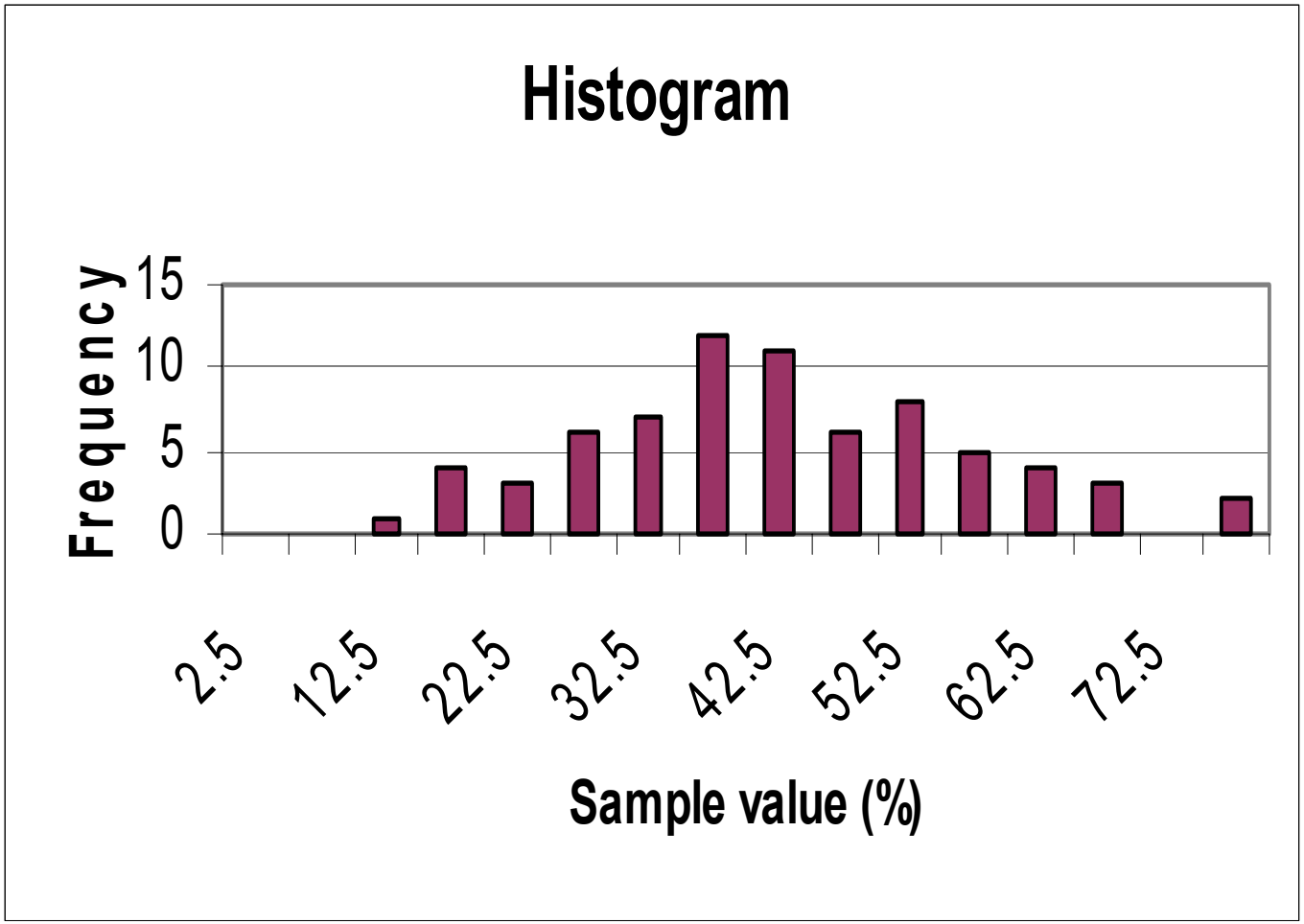
3- Example

Frequency Distribution Table

Bin(Class mid-point)	*Frequency*	*Relative Frequency*	*% Frequency*	*CF*	*%CF*
2.5	0	0	0	0	.00%
7.5	0	0	0	0	.00%
12.5	1	0.013888889	1.388888889	1	1.39%
17.5	4	0.055555556	5.555555556	5	6.94%
22.5	3	0.041666667	4.166666667	8	11.11%
27.5	6	0.083333333	8.333333333	14	19.44%
32.5	7	0.097222222	9.722222222	21	29.17%
37.5	12	0.166666667	16.66666667	33	45.83%
42.5	11	0.152777778	15.27777778	44	61.11%
47.5	6	0.083333333	8.333333333	50	69.44%
52.5	8	0.111111111	11.11111111	58	80.56%
57.5	5	0.069444444	6.944444444	63	87.50%
62.5	4	0.055555556	5.555555556	67	93.06%
67.5	3	0.041666667	4.166666667	70	97.22%
72.5	0	0	0	70	97.22%
77.5	2	0.027777778	2.777777778	72	100.00%
Sum	72	1	100		

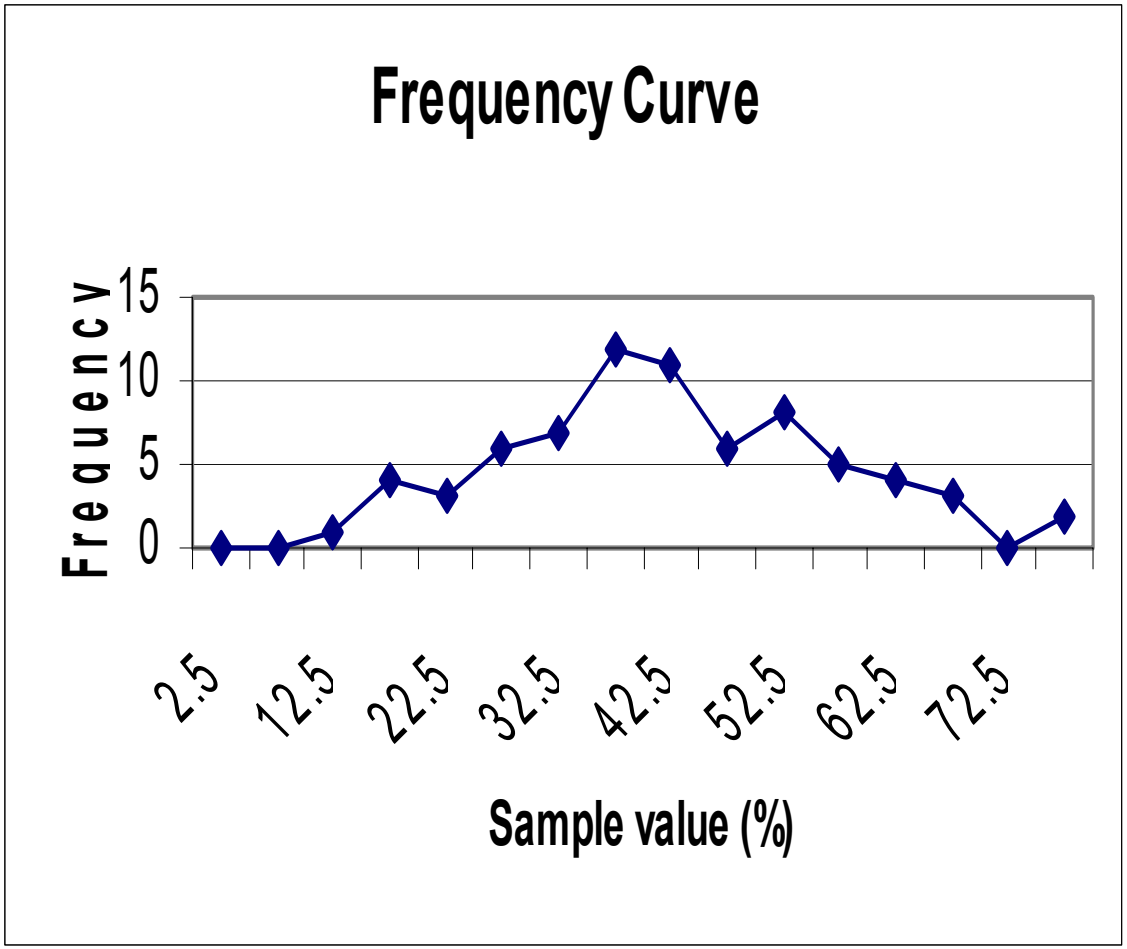
Experimental Frequency Distributions

3- Example



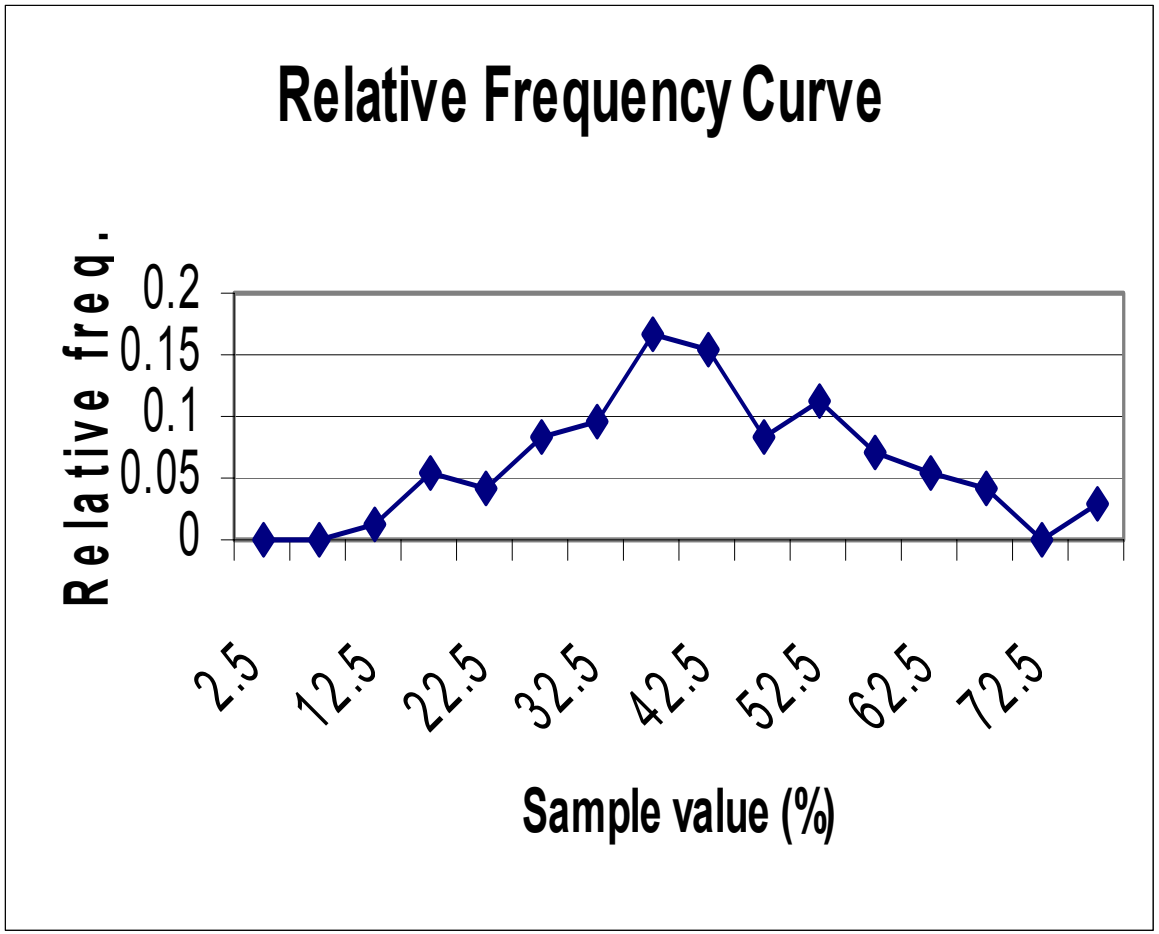
Experimental Frequency Distributions

3- Example



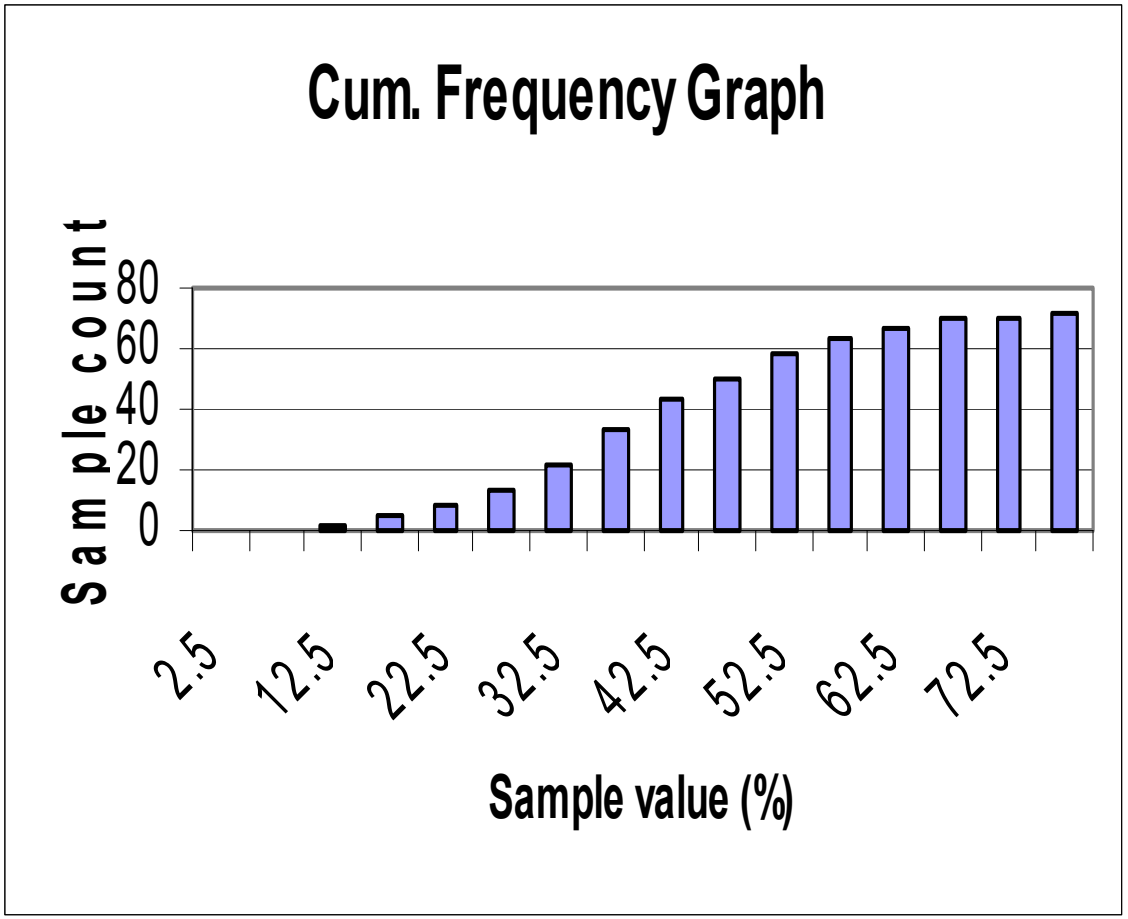
Experimental Frequency Distributions

3- Example



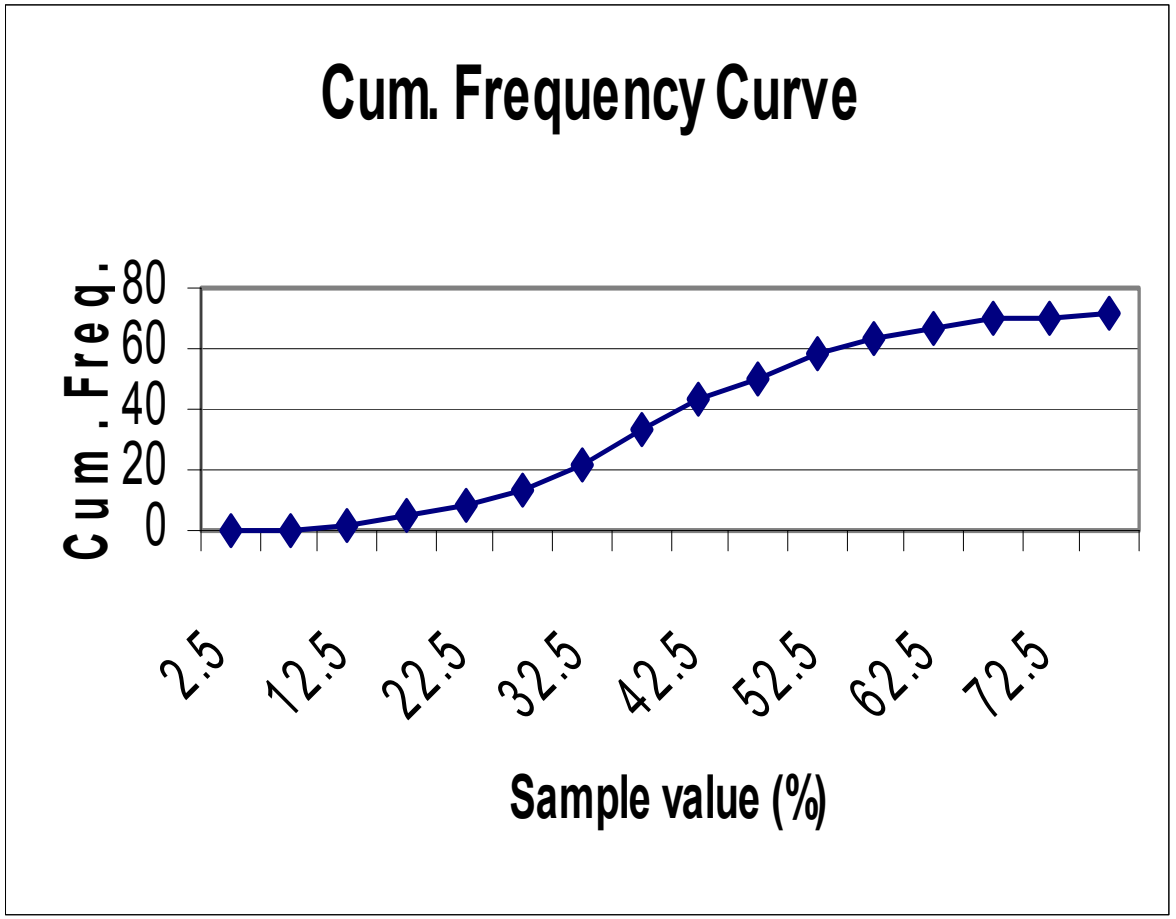
Experimental Frequency Distributions

3- Example



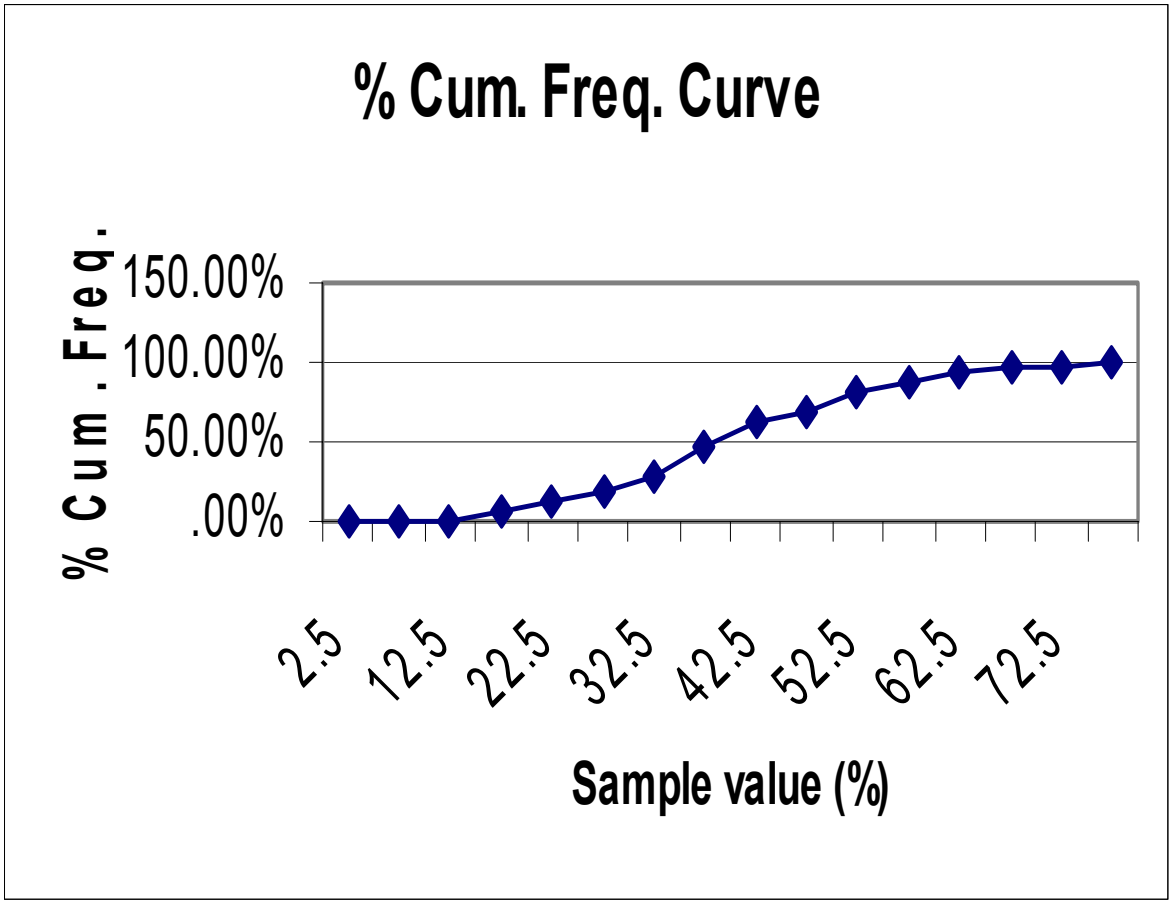
Experimental Frequency Distributions

3- Example



Experimental Frequency Distributions

3- Example



Statistical Properties & Parameters

The histogram shows important features of a data set by means of **properties** of the distribution. These histogram properties are:

1) Location: Represented by the position of a parameter along the histogram scale. Measure of center is a special case of location measurements.

2) Dispersion: Represented by the extent to which the distribution is spread out along the scale or how much values vary from some central value (i.e. average)

3) Shape: Represented by the pattern of the statistical distribution.

Measures of Center

Give an idea where the center of a data set distribution lies

- **Mean (*m*)**
  - Arithmetic average of a data set (sample)
  - It is sensitive to extreme values

$$m = \bar{v}_a = \frac{1}{n}\sum_{i=1}^{n} v_i$$

Measures of Center

Give an idea where the center of a data set distribution lies

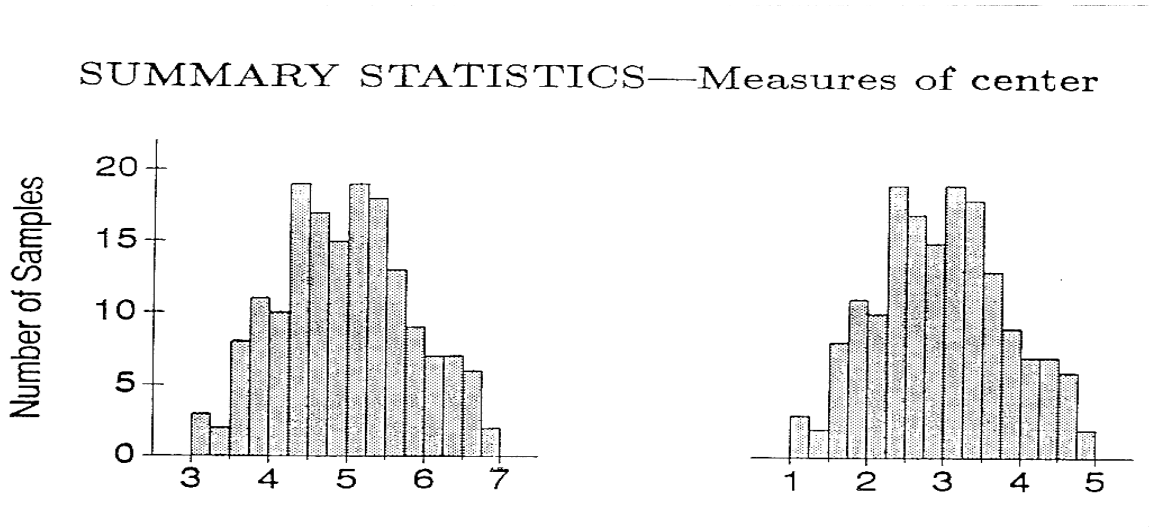
- **Median (M)**
  - A midpoint of the observed values if they are arranged in an increasing order.
  - Half of the values are below the median and half of them are above it
  - It is not sensitive to extreme values
  - It is sensitive to gaps in the middle of a data set.

$$M_{odd} = x_{\frac{n+1}{2}} \qquad M_{even} = \left(\frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} \right)$$

Measures of Center

Give an idea where the center of a data set distribution lies

- **Mode**
  - The observation that occurs most frequently.



Measures of Location

Give an idea where the location of a specific observation in a data set distribution lies

- **Minimum (*min*):** the smallest value in the data set.
- **Maximum (*max*):** the largest value in the data set.
- **Lower or First Quartile (Q_1):** an observation value below which quarter of data falls.
- **Upper or Third Quartile (Q_3):** an observation value above which quarter of data falls.

Measures of Location

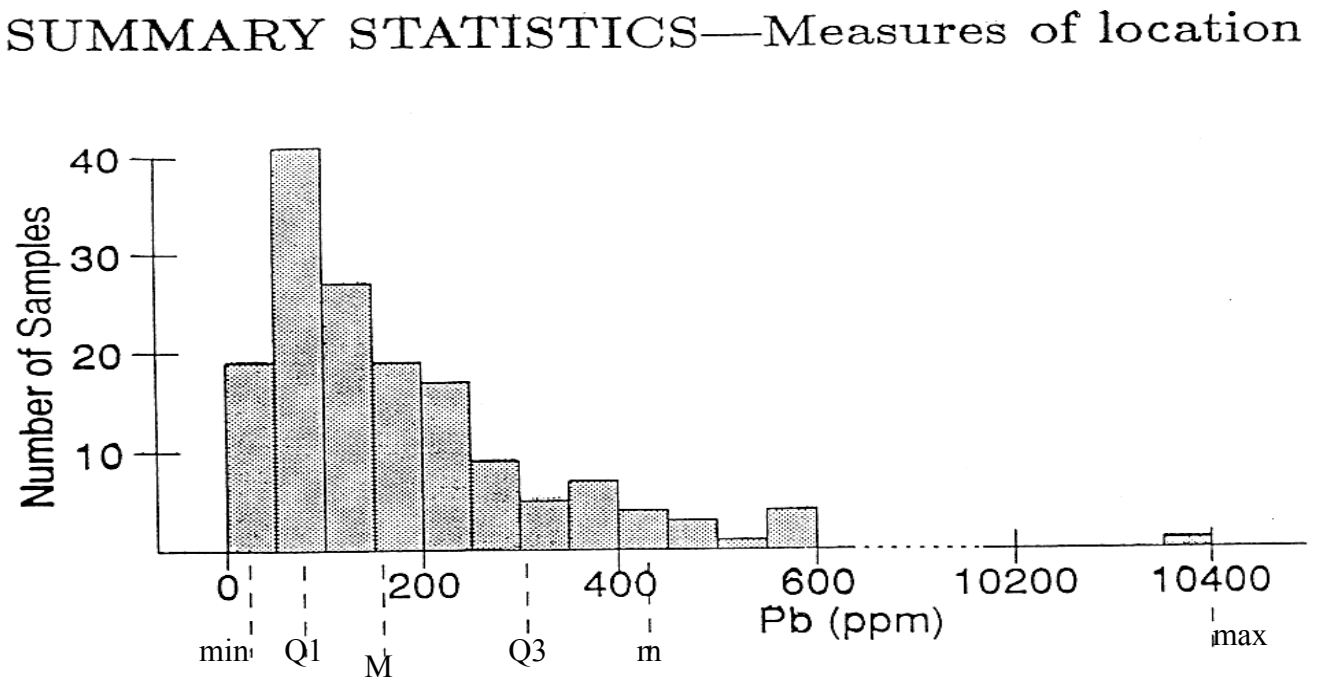
Give an idea where the location of a specific observation in a data set distribution lies

- **Quantile (q_p):** a general expression that describes an observation value below which a *percentage or fraction* quantity of data falls.

  - $Min = q_0$
  - $Q_1 = q_{0.25}$
  - $M = q_{0.50}$
  - $Q_3 = q_{0.75}$
  - $max = q_1$

Measures of Location

Give an idea where the location of a specific observation in a data set distribution lies



Measures of Spread

Describe the variability of the data values

- **Variance (S^2):** is the average squared difference of the observed values from their mean.
  - It is sensitive to extreme value
  - **Standard Deviation (S):** is the square root of variance. It measures the uncertainty of the estimated mean value, for example. It is, also, sensitive to extreme values.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(v_i - m\right)^2$$

Measures of Spread

Describe the variability of the data values

- **Interquartile Range (*IQR*):** is the difference between upper and lower quartiles.
  - It is not sensitive to extreme value
  - It is a rough measure of spread of data values.

$$IQR = Q_3 - Q_1$$

Measures of Spread

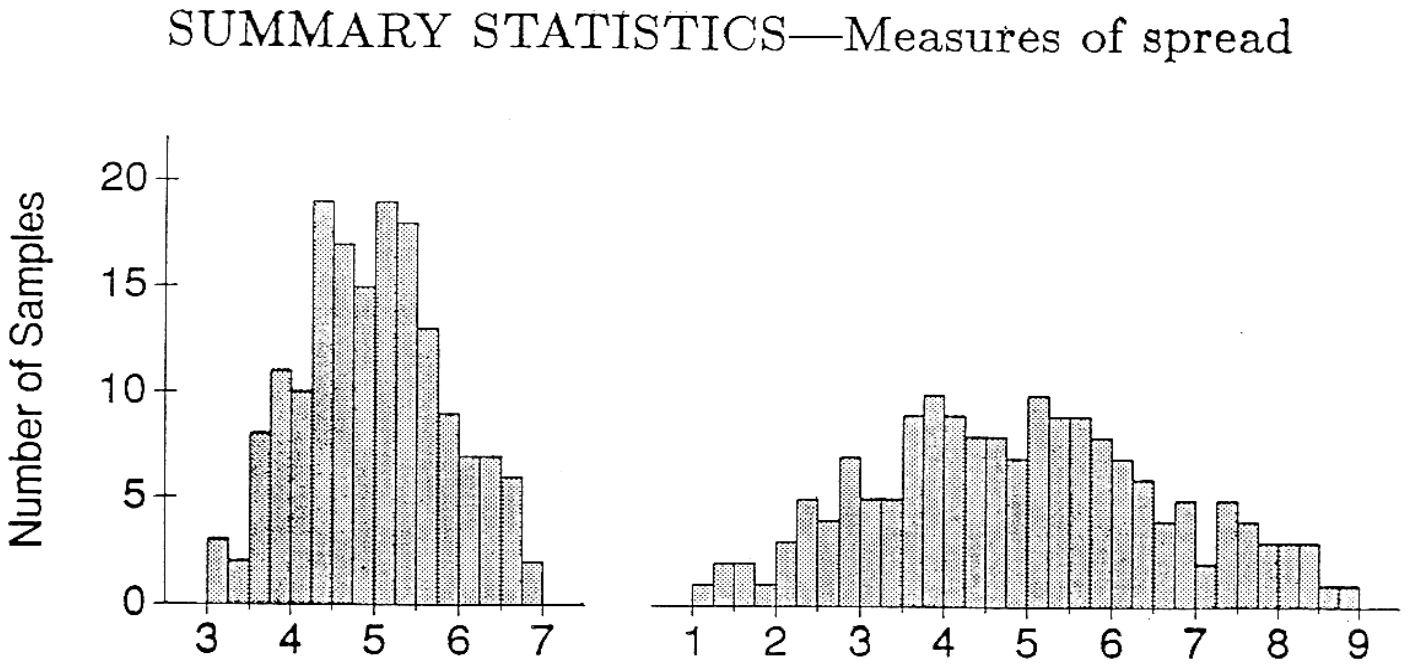
Describe the variability of the data values

- **Coefficient of Variation (*CV*):** is a measure of how significant is the impact of the presence of high values on the final estimates.

$$CV = \frac{S}{m}$$

Measures of Spread

Describe the variability of the data values



Measures of Shape

Describe the shape of distribution of the data values

- **Coefficient of Skewness (*CS or g*):** is the measure of symmetry of data values distribution.

$$CS = g = \frac{\left(\frac{1}{n} \sum_{i=1}^{n} (v_i - m)^3 \right)}{S^3}$$

Measures of Shape

Describe the shape of distribution of the data values

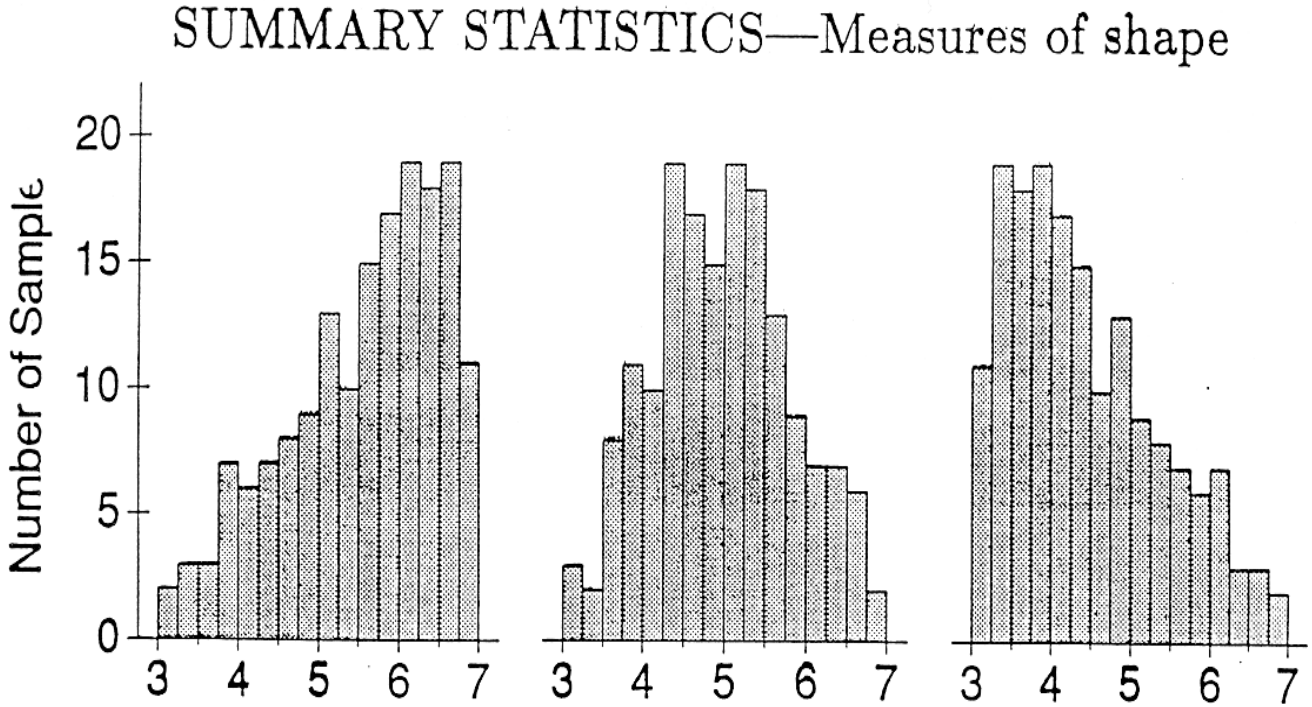
- It is very sensitive to extreme value

- Its range is:

» SC = 0	the data values are symmetrical around the mean value (M=m).
» SC = +ve	larger number of observations with low values in the data set (M<m).
» SC = -ve	Larger number of observations with high values in the data set (M>m).

Measures of Shape

Describe the shape of distribution of the data values



Measures of Shape

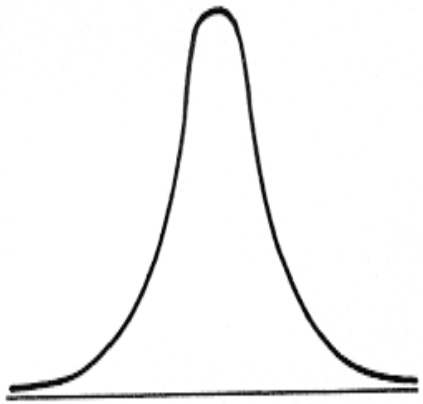
Describe the shape of distribution of the data values

- **Kurtosis (κ) or (K):** is a measure of peakedness of a distribution.

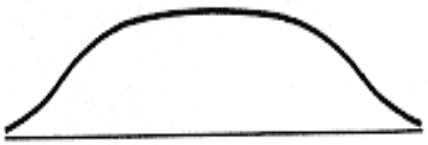
  - Kurtosis = 0 ==> Normal distribution with moderate peak and systematic shape (Mesokurtic)
  - Kurtosis > 0 ==> Distribution with sharply high peak (Leptokurtic)
  - Kurtosis < 0 ==> Distribution with flat top (Platykurtic)

Measures of Shape

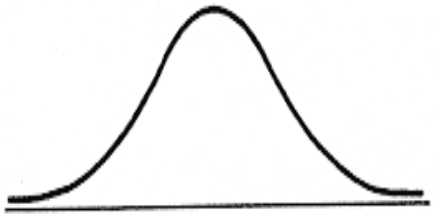
Describe the shape of distribution of the data values



(a) Leptokurtic



(b) Platykurtic



(c) Mesokurtic

Frequency Distribution Models

Probability Density Function (PDF)

- **What is a PDF model?**

  A mathematical equation that describes the frequency curve or probability distribution of a data set.

- **Why modeling?**

  – It represents and summarizes the statistical distribution of the entire "environmental phenomenon" and helps in making predictions.

  – Statistical characteristics and parameters can be derived and calculated easily from the model. Such parameters reflect the behavior of the "environmental phenomenon".

Frequency Distribution Models

Probability Density Function (PDF)

- **Types of models**

  - **Normal distribution**

    - § Gaussian (Gauss), Laplace, or Bell-shaped distribution.
    - § Values are symmetrically distributed around a central value.
    - § The mean is the most representative value of the distribution (i.e. use arithmetic average to estimate the unknown from a set of "uncorrelated random variables").
    - § The variance is the well-defined measure of the spread of observations.

Frequency Distribution Models

Probability Density Function (PDF)

Normal distribution

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2\right]$$

$$F(x_0) = \int_{-\infty}^{x_0} f(x)dx$$

Frequency Distribution Models

Probability Density Function (PDF)

- **Types of models**

  - **Non-normal distribution**

    - § Lognormal distributions (natural or base-10 logarithms of measured observations).
    - § Values are asymmetrically distributed around a central value.
    - § The mean is not the most representative value of the distribution. (i.e. do not simply use arithmetic average to estimate the unknown from a set of "uncorrelated random variables"). However, other averages or median value might work!

Frequency Distribution Models

Probability Density Function (PDF)

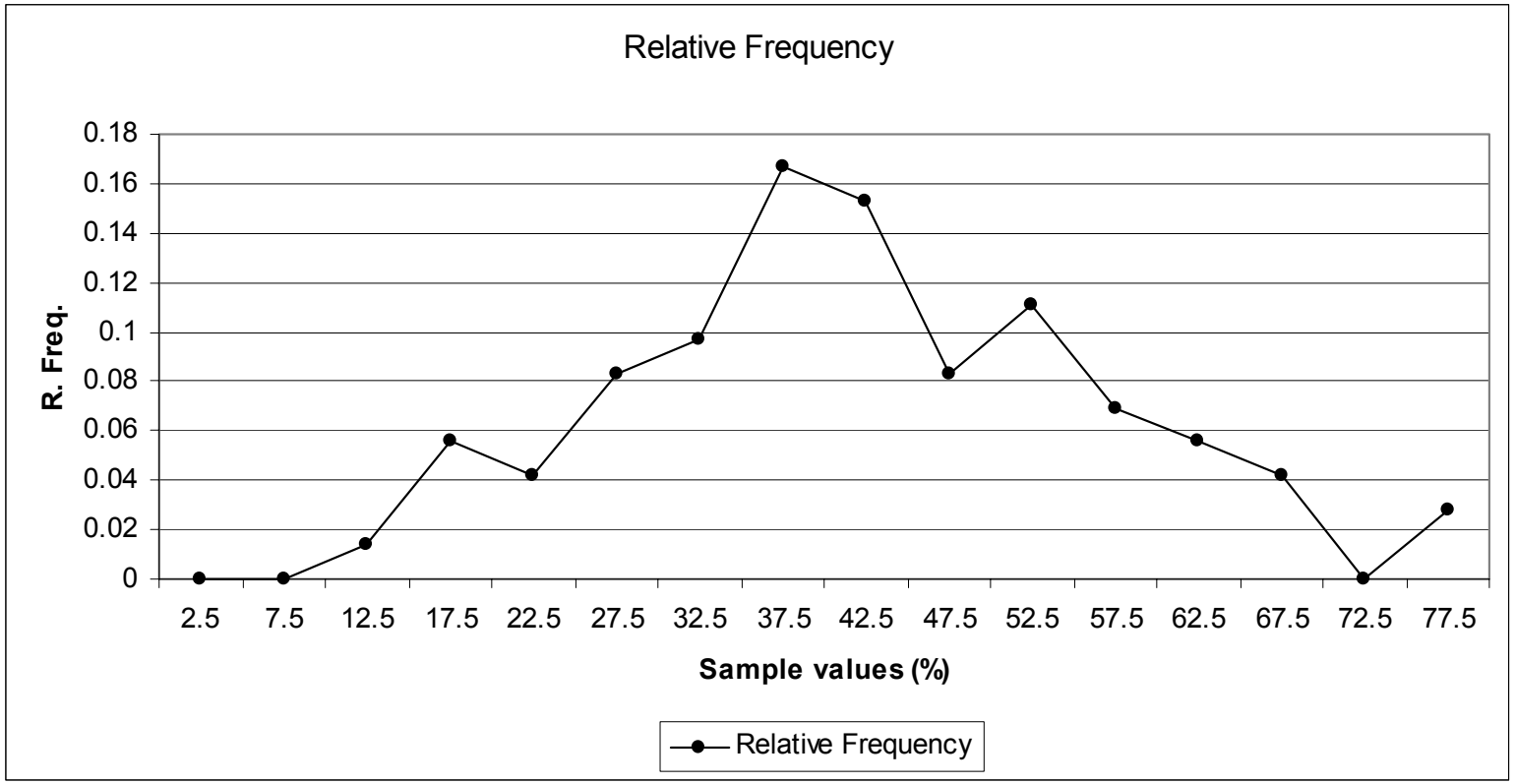
Non-normal distribution

$$f(x) = \frac{1}{x\beta\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln \gamma - \ln x}{\beta} \right)^2 \right]$$

$$\gamma = e^{\alpha} \qquad \beta = S_{\log aritms}$$

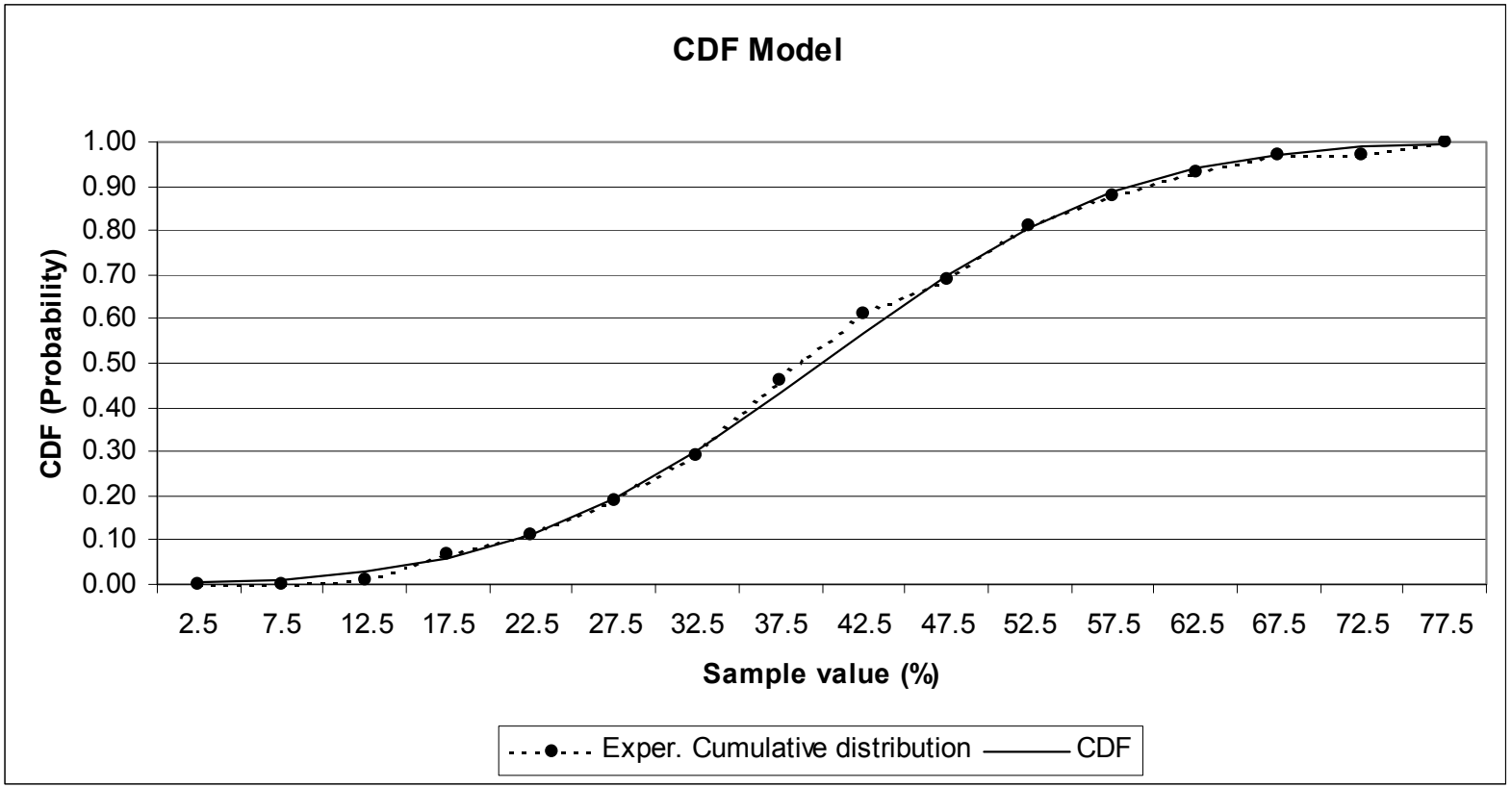
Frequency Distribution Models

Probability Density Function (PDF): Example of a normal case



Frequency Distribution Models

Cumulative Distribution Function (CDF): Example Example of a normal case



Frequency Distribution Models

Skewed Distributions

- **What if a distribution is skewed?**

  Use power transformation techniques to transform original data values into a defined power function.

- **Why and how?**

  – Distribution of transformed data is much easier to describe than the distribution of original skewed data.

  – Common power transforms:

$$y = (z^p - 1) / p \qquad\qquad y = Ln(z)$$

Frequency Distribution Models

Other Types of Skewed distributions

Lognormal Model