STATISTICAL TREATMENT OF GEOCHEMICAL DATA

Disclaimer:

This presentation is a part of the assignment for MSc III Semester Mineral Exploration theory paper. This is an attempt to enable the students to collect and review the literature, prepare powerpoint presentation and present the work, independently. The data and literature used here has been taken from various sources, and duly acknowledged. This can help as a guideline, and should not be treated as final.

October 2016

By-

NAYANIKA DAS,

M.Sc. 3rd Semester,

SoS in Geology & WRM,

Pt. Ravishankar Shukla University,

Raipur

Statistics …

is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data.

Due to its application for prediction and forecasting based on data, it is applicable to a wide variety of academic disciplines, from the natural and social sciences to the humanities, government and business.

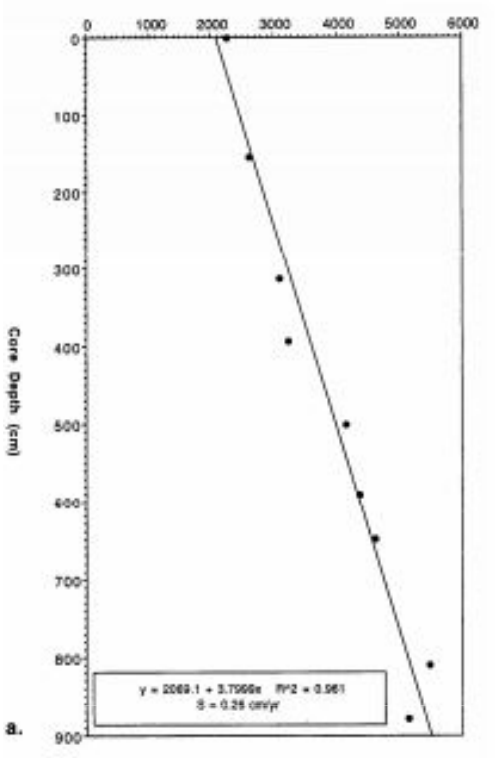
Why is Statistics Important?

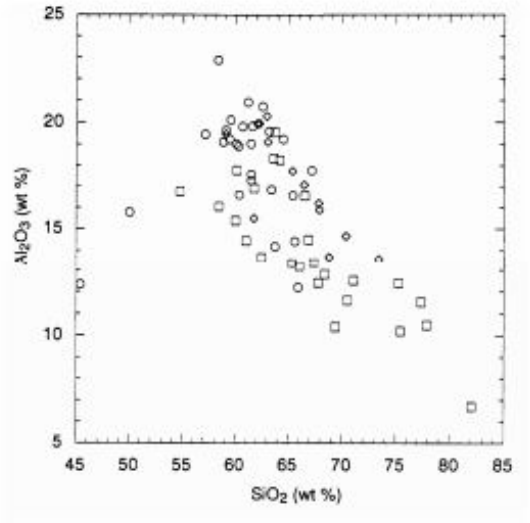
• Statistics is part of the quantitative approach to knowledge. In the past, geology has been qualitative, but is now becoming increasingly quantitative.

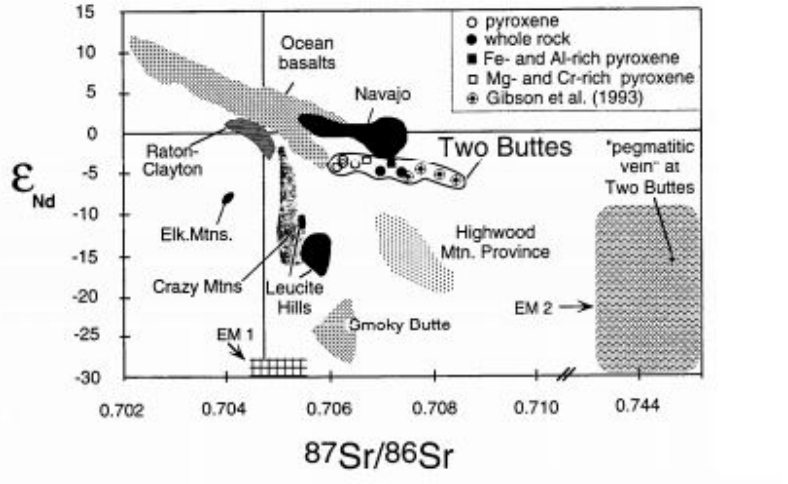
• Nowadays geologists have a bunch of numbers to deal with.

• Systematic approach or methods of statistical data analysis are required to retrieve information from the set of numbers.

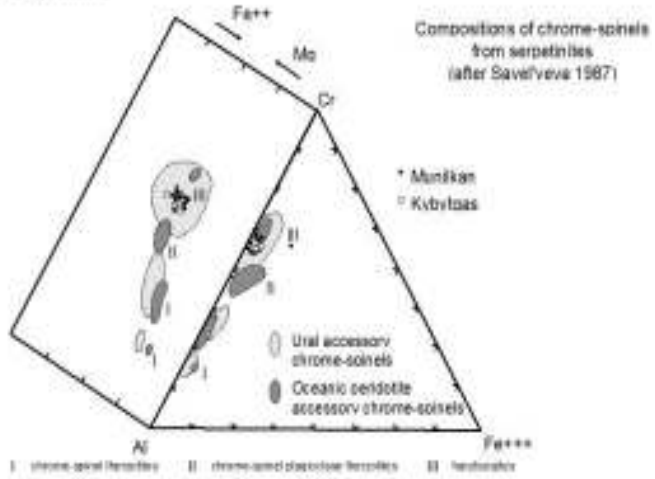
A First Approach of Data Analysis: Use of Graphs



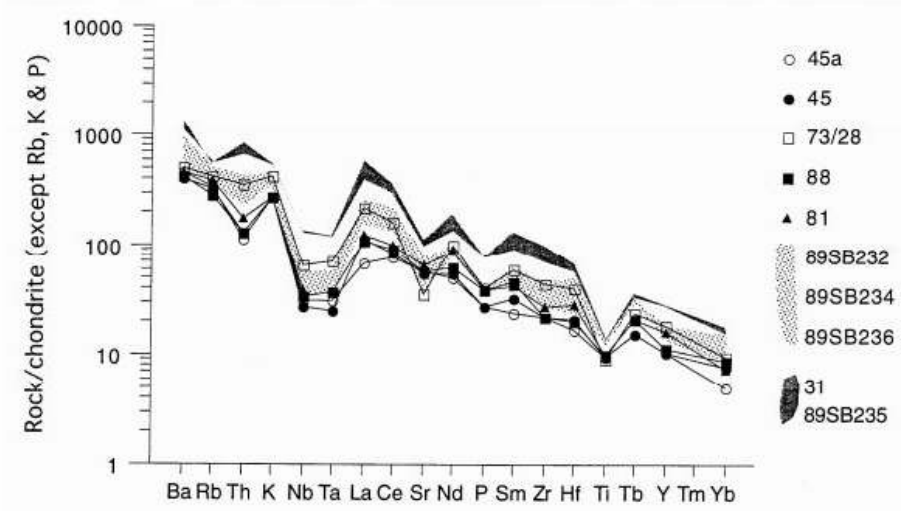




- Ternary Diagrams:



Spider Diagrams:



The graph helps us to visualise distinct fields or trends significant

A First Approach of Data Analysis: On the Use of Graph

• The use of graphs is a helpful tool to get a first grasp about the information content in the data.

• However, the following questions remain:
–What do all of these show?
–Is there any critical analysis of these curves and fields?

• We need to employ more “rigorous” scientific methods to geologic problems
–hypothesis testing

• We need to turn to statistics

What is the Field of Statistics?

• Statistics is “the determination of the probable from the possible”.

– This implies we need a rigorous definition and quantification of “probable”.

– Statistics is the quantitative study of *variance.*

• Statistics *usually* deals with data in the form of numbers.

Two common uses of the word “Statistics”:

– **Descriptive Statistics :** Numerical or graphical summary of data (what was observed).

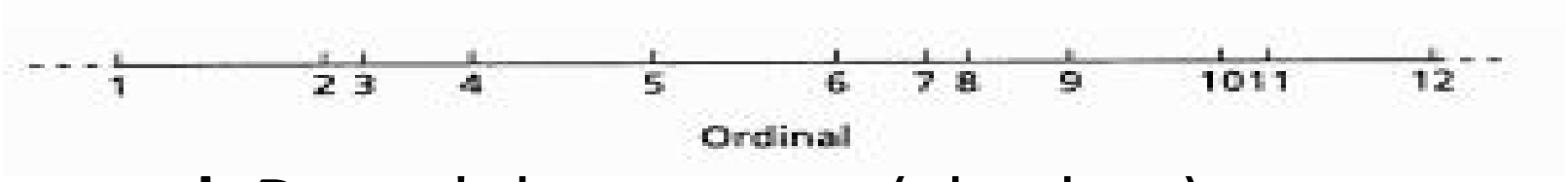
– **Inferential Statistics :** used to model patterns in the data, accounting for randomness and drawing inferences about the larger population.

• answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation), descriptions of association (correlation), or modeling of relationships (regression).

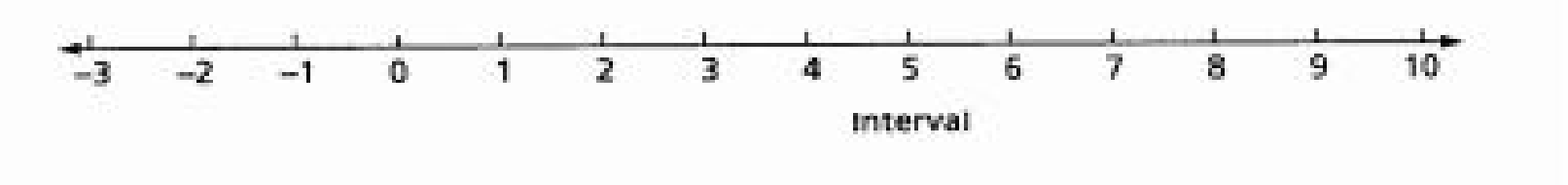
• Other modeling techniques include ANOVA, time series, and data mining

Types of Data

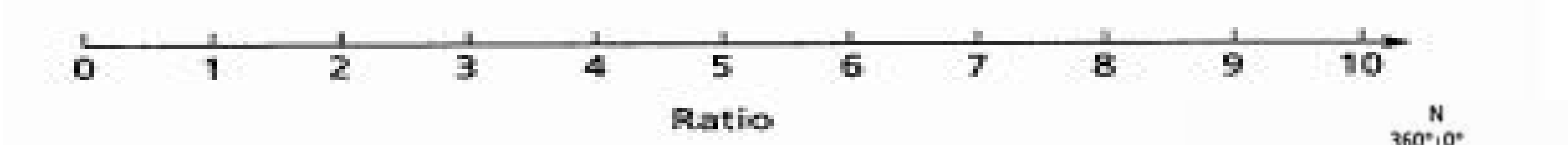
1. **Nominal:** Classification Scheme, may be numeric but could as easily be A,B,C
2. **Ordinal:** Rank order data, numeric



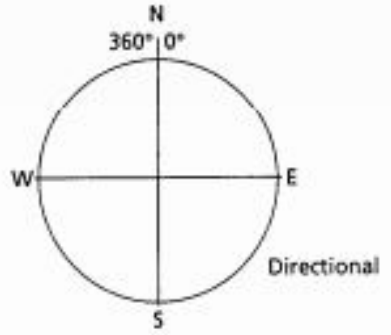
3. **Interval:** Doesn't have a true (absolute) zero



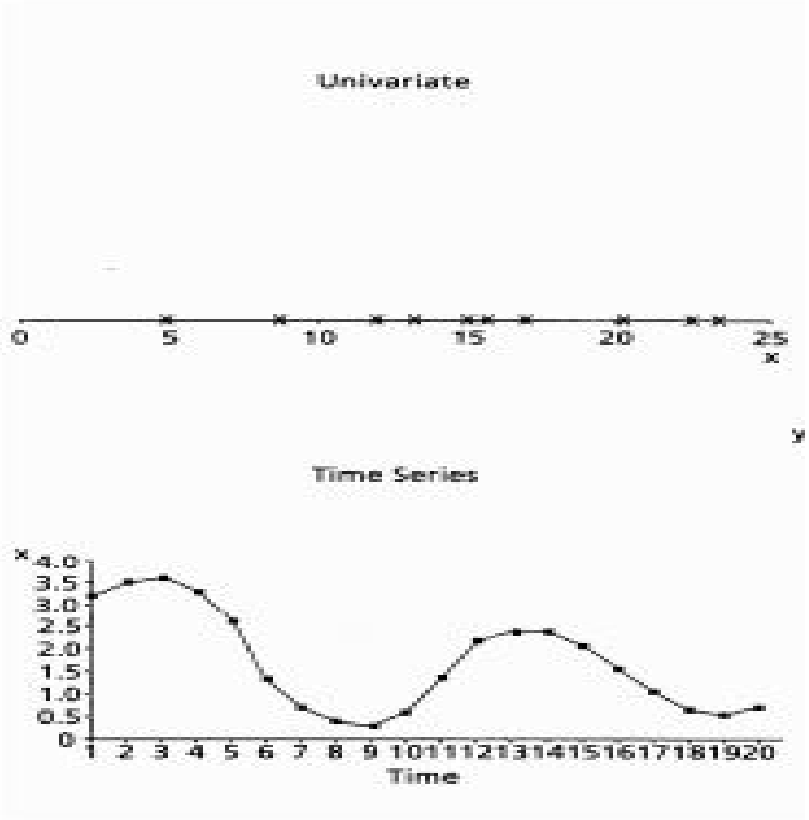
4. **Ratio**: Has true zero



5. **Angular directional**: similar to interval



Types of statistics - number of variables



1. **Univariate**

– dependant on one variable.

e.g. calculations of mean, standard deviation, median, ANOVA

2. Bivariate

– two variables x versus y

regression, correlation in geology.

e.g. Harker Diagrams

3. Multivariate

– many variables matrix manipulation is needed to handle the data.

4. Spatial: Map based, having three or four variables analyzed together. Two or three of the variables are spatial. This type of analysis is almost unique in geoscience and has led to a geology-derived field called **geostatistics**.

SOME DEFINITIONS

Population: total number of all possible specimens in a study.

Sample: geologically - a single specimen (geologic sample) – statistically - a subset of a population (group of data).

Statistic refers to something calculated from a set of data (sample).

Parameter a property of a population which *usually* cannot be measured directly but must be estimated by a statistic.

What is Variance?

Variance measures how a set of data values for a variable fluctuate around the mean of that variable.

- **Variance** is the natural error or scatter or variability in measurements or it can be thought of as the natural spread of data.

- **Variance** is also one of many quantitative measures of variability of data (assuming the data is of Gaussian nature). This is represented as σ^2 or S^2.

Why do Data Vary?

• No two measurements/samples/natural objects will ever be the same! So, a certain variation is inherent to all natural objects.

• Field sampling errors
– not getting representative sample

• Preparation errors
– contamination, final split does not represent field sample

• Analytical errors
– calibration errors (setting up the machine)
– measurement errors (fluctuations in counting)
– machine errors (properties of the machine, mass fraction).

GRAPHICAL REPRESENTATION

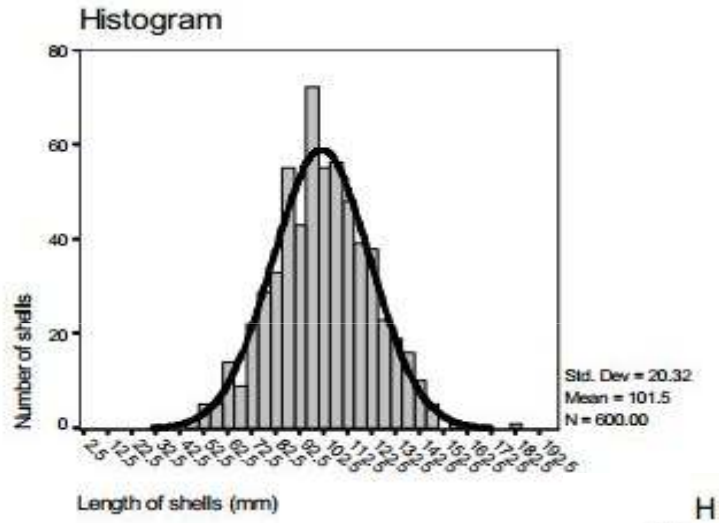
Distributions

- For almost all geologic data, the data are distributed among a range of values.
- Again, it is this variability that we want to investigate.
- We will use various statistics to describe this variability (the location, spread, shape etc) of various distributions.

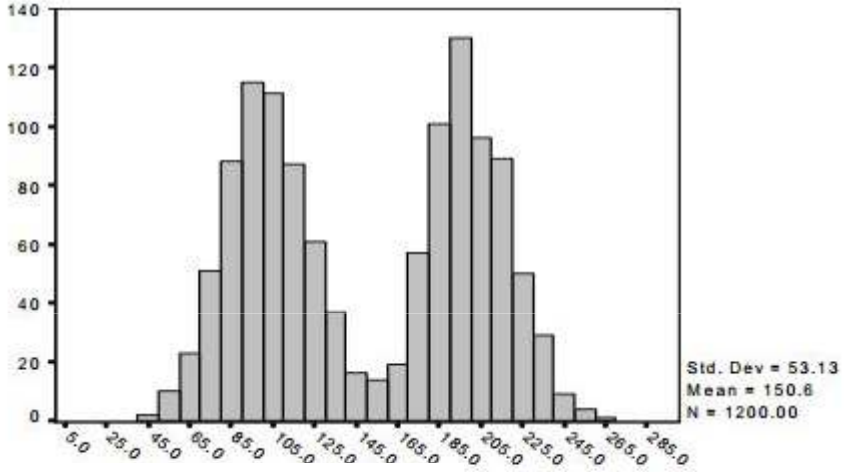
The data can be discrete or nondiscret (continuous).

Distribution can be:

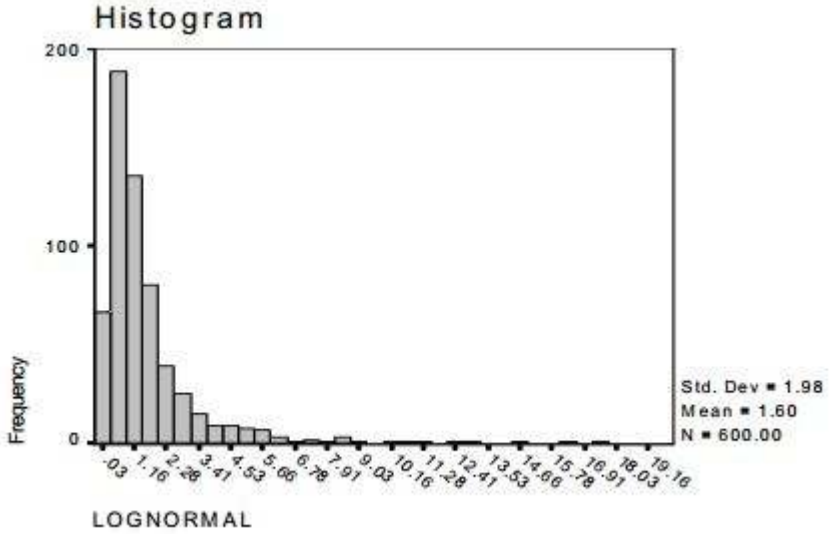
Unimodal



Bimodal:



Skewed

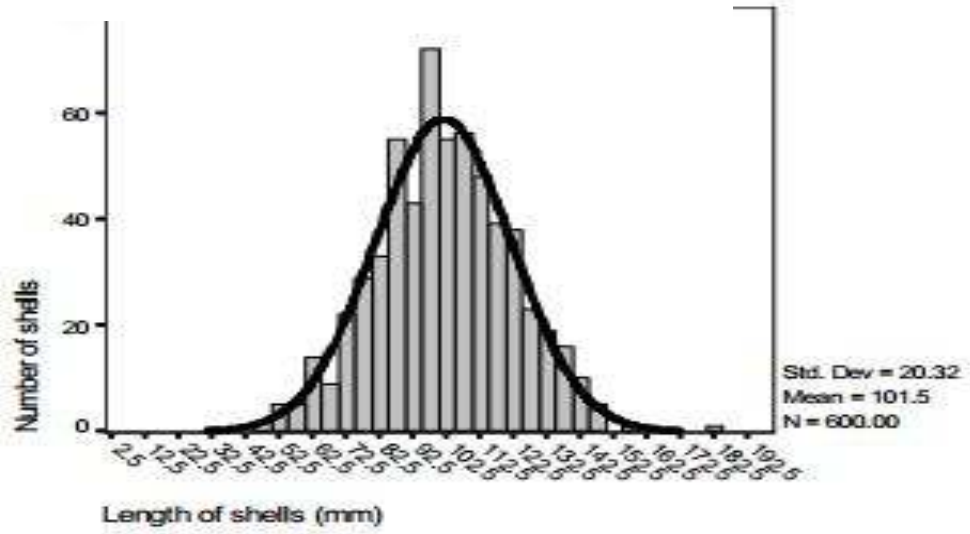


Normal Distribution:

- **Also called the Bell curve, Gaussian Distribution, Standard curve**

It is characterized by its mean (μ) and standard deviation (σ). It is a continuous function given by the equation.

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2} \quad for -\infty < X < \infty$$

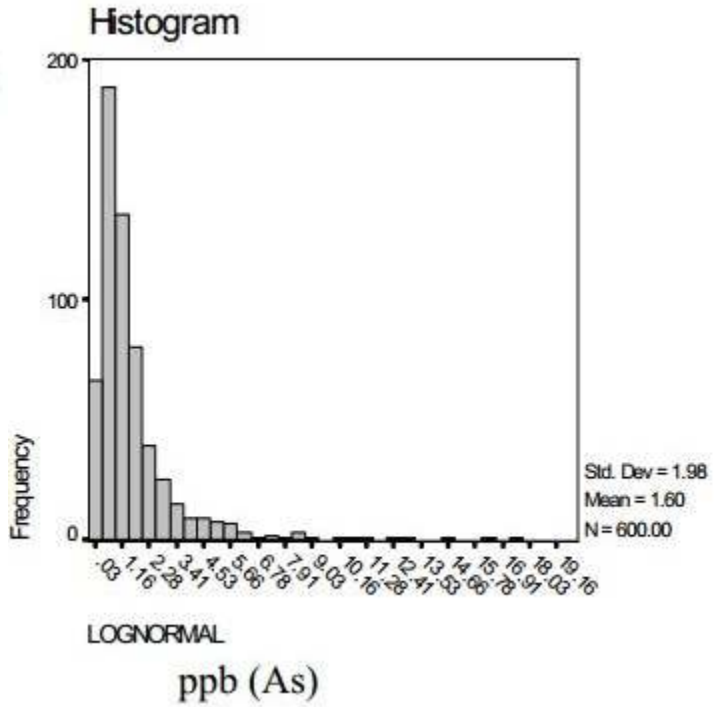


Lognormal distribution:

- A variant of the normal distribution in that the logarithms of the values form a normal distribution: A continuous distribution. Characterized by a lot of small values and a few very large values. A highly skewed distribution. Random errors are multiplicative.

The equation for this curve is

$$Y = \frac{1}{X\sigma\sqrt{2\pi}} e^{-(\log(X)-\mu)^2/2\sigma^2} \quad for -\infty < X < \infty$$



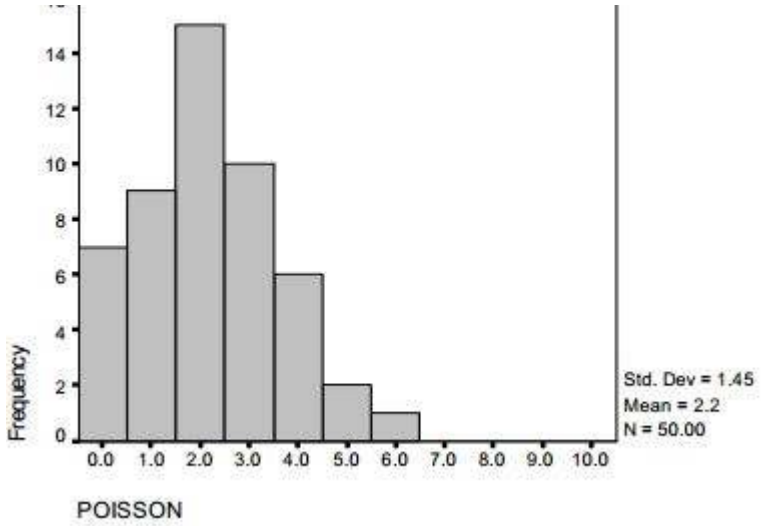
Poisson distribution

- A **discrete** distribution where the probability of an event occurring is rare and random (e.g. radioactive decays per hour, wells per square mile etc.

Defined by the equation:

$$P = \frac{(\overline{X})^r e^{-(\overline{X})}}{r!}$$

where P is the probability of r events occurring if the average number of events per unit of time or area is $\overline{X}$.



STATISTICAL AVERAGE

Median – based on rank-order statistics $\tilde{X} = X_{\frac{n+1}{2}}$

Mode – most common occurrence

Can be dependent on the binning process in data sorting

Arithmetic mean -- Related to the first moment

For a sample:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

for a population:

$$\mu = \frac{\sum_{i=1}^{n} X_i}{n}$$

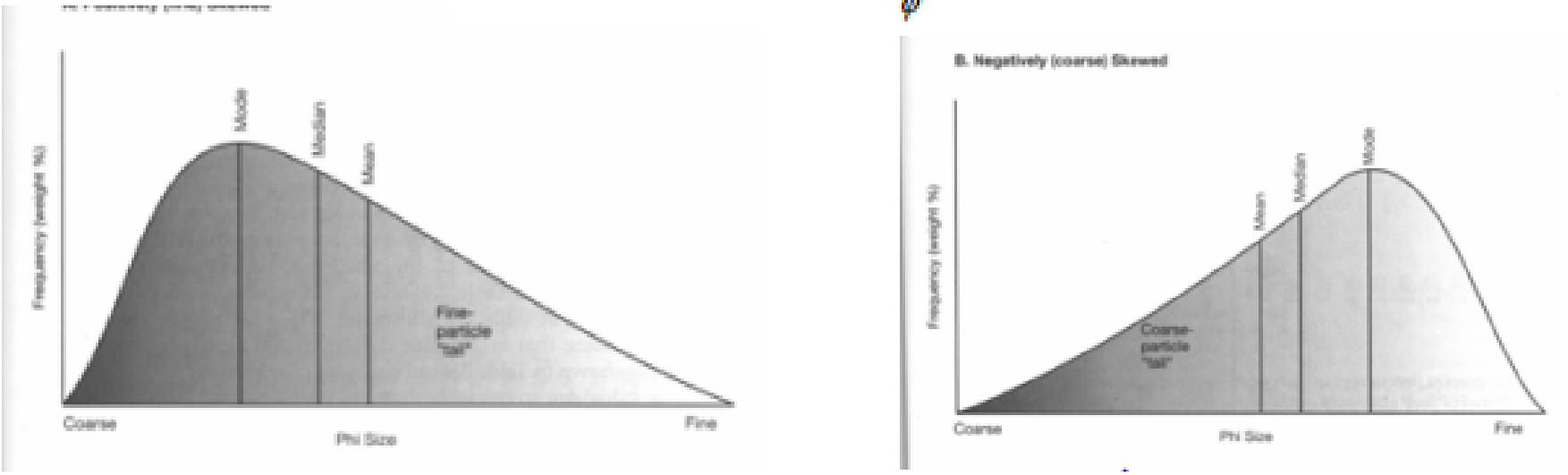
Skewness reflects sorting in the "tails of a grain-size population.

- Populations that have a tail of excess fine particles are said to be **positively skewed or fine skewed.**

- Populations with a tail of excess coarse particles are **negatively skewed or coarse skewed."**

Skewness

$$Sk_{\phi} = \frac{\sum f \cdot (m - \bar{x}_{\phi})^3}{100\sigma_{\phi}^3}$$



Kurtosis

The degree of peakedness is called kurtosis.

Kurtosis

$$K_{\phi} = \frac{\sum f \cdot \left(m - \bar{x}_{\phi}\right)^4}{100\sigma_{\phi}^4}$$

Standard deviation

- The standard deviation is a measure that is used to qualify the amount of variation or dispersion of a set of data values.
- the standard deviation tells about the sorting.

Standard Deviation

$$\sigma_{\phi} = \sqrt{\frac{\sum f \cdot (m - \bar{x}_{\phi})^2}{100}}$$

Statistics and Parameter

"Special" descriptors	Measures...	Population parameter	Sample statistic
Mean	Central tendency	μ	$\overline{X}$
Variance	Spread	σ^2	s^2
Standard deviation	Spread	σ	s
Skewness	Symmetry	$\sqrt{b_1}$	$\sqrt{b_1}$
Kurtosis	Flatness or peakedness	b_2	b_2

Measurement	Frequency (f)	Mid-value (M.V.) (X)	Cum. frequency (Cum. f.)	Dev from Assumed Mean ($X-35$) (dx)	dx^2	Product of f and dx (fdx)	Product of fdx and dx or, $f \times dx^2$ (fdx^2)
0—10	10	5	10	−30	900	−300	9,000
10—20	40	15	50	−20	400	−800	16,000
20—30	20	25	70	−10	100	−200	2,000
30—40	0	35	70	0	0	0	0
40—50	10	45	80	+10	100	+100	1,000
50—60	40	55	120	+20	400	+800	16,000
60—70	16	65	136	+30	900	+480	14,400
70—80	14	75	150	+40	1600	+560	22,400
80 and above	0	85	150	+50	2500	0	0
	$N=150$					$\Sigma fdx = +640$	$\Sigma fdx^2 = 80,800$

$$\bar{X} = A + \frac{\Sigma fdx}{N} = 35 + \frac{640}{150} = 35 + 4{\cdot}27 = 39{\cdot}27 \text{ marks approx.}$$

Median = Size of $\left(\frac{N}{2}\right)$th item

= Size of $\left(\frac{150}{2}\right)$th item = 75th item ∴ Median class is 40—50.

$$\text{Median} = l_1 + \frac{l_2 - l_1}{f}(m - c) = 40 + \frac{50-40}{10}(75-70)$$

$$= 40 + \frac{10}{10} \times 5 = 40 + \frac{50}{10} = 40 + 5 = 45 \text{ marks.}$$

$$\text{Standard Deviation } (\sigma_x) = \sqrt{\frac{\Sigma fdx^2}{N} - \left(\frac{\Sigma fdx}{N}\right)^2} = \sqrt{\frac{80,800}{150} - \left(\frac{640}{150}\right)^2}$$

$$= \sqrt{538{\cdot}67 - (4{\cdot}27)^2} \text{ approx.}$$

$$= \sqrt{538{\cdot}67 - 18{\cdot}232} = \sqrt{520{\cdot}438} \text{ approx.}$$

$$= 22{\cdot}81 \text{ marks.}$$

वैकल्पिक सूत्र

$$\text{Coefficient of Skewness } (J) = \frac{3(\bar{X} - M)}{22{\cdot}81} = \frac{3(39{\cdot}27 - 45)}{22{\cdot}81} = \frac{(3 - 5{\cdot}73)}{22{\cdot}81} = \frac{-17{\cdot}19}{22{\cdot}81}$$

$$= -{\cdot}753 \text{ approx.}$$

THANKS